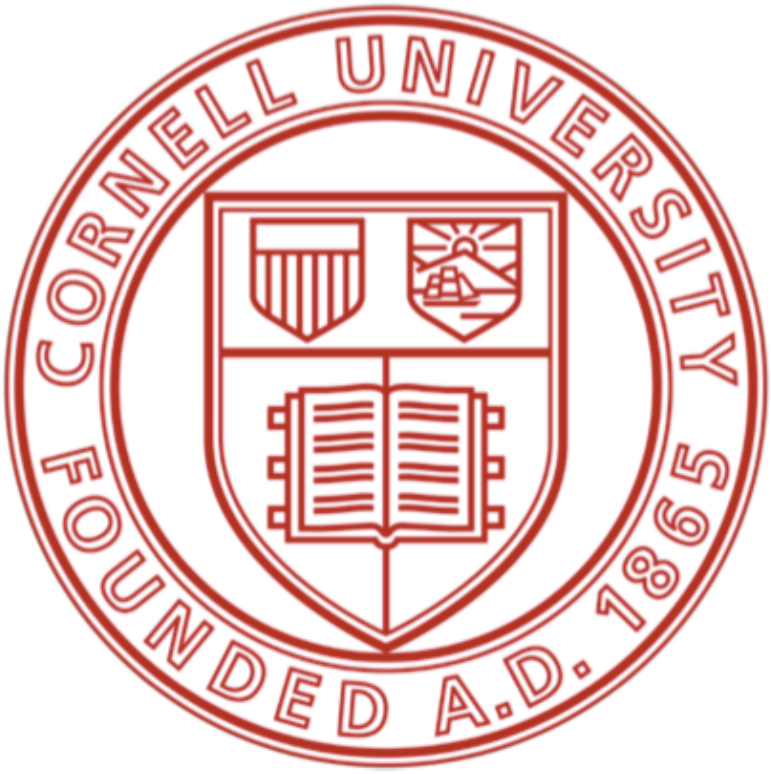
Consumer Perception of Hotel Competitive Sets



Alexa Angelica Perrucci

Cornell University School of Hotel Administration

May 2018

Consumer Perception of Hotel Competitive Sets

By

Alexa Perrucci

A thesis submitted in partial satisfaction of the requirements for the degree of

Bachelor of Science

In

Hotel Administration

In the

SC Johnson College of Business

At

Cornell University

Spring 2018

Consumer Perception of Hotel Competitive Sets

By

Alexa Perrucci

The thesis titled Consumer Perception of Hotel Competitive Sets

Presented by Alexa Perrucci is approved by:

Thesis Committee Chair Date

Research Committee Chair Date

Cornell University
May 2018

Abstract

Consumer Perception of Hotel Competitive Sets

By

Alexa Perrucci

Bachelor of Science

Cornell University

This research explores consumer perception of hotel competitive sets by analyzing TripAdvisor data collected from 11 cities internationally. The study included running regressions, generating visual displays (scatter plots and histograms), and performing K-means clustering. The results were encouraging, as the outcomes demonstrated that there is an ability to generalize consumer preference when it comes to hotel competitive sets. The research identifies a strong need for industry executives to begin focusing their attention on consumer perception when conducting competitive analysis.

Author's Biography

Alexa Perrucci is from Wood-Ridge, New Jersey and is a senior at Cornell University's School of Hotel Administration. After attending the Academy for Culinary Arts and Hospitality Administration in high school, she knew she wanted to further her hospitality education at SHA.

Throughout her time at Cornell, she has been actively involved in her Professional Business Fraternity, Phi Gamma Nu, her sorority, Kappa Kappa Gamma, and the Cornell Cheerleading team. Within SHA, she served as a teaching assistant for Business Computing and Management Communication I as well as a consultant in the Communication Center. As a volunteer for HEC, she received experience in many departments including Wines, Conference Services, Procurement, Culinary, and Employee Dining.

Alexa feels privileged to have been a part of the Hotel School, where she received incredible opportunities that helped her grow both as a person and as a professional. In the spring of her junior year, Alexa studied abroad in Barcelona, Spain. She had a once in a lifetime experience at Universitat Pompeu Fabra. While in Europe, Alexa was able to visit 8 countries and see 21 different cities. Her most memorable experience at SHA was running her Establishment night, "End of Summer Bash." She was the FOH manager and had an extremely successful and enjoyable experience.

In the fall of her junior year, Alexa was inducted into Ye Hosts Honorary Society, which is an honor society formed of the top 10% of students in the school. As a member of Ye Hosts, she was given the opportunity to begin her research in the Latin Honors Program. Moreover, Alexa was named a Merrill Presidential Scholar in the spring of her senior year. This program selects students from the top 1% percent in each of the colleges based not only on academic scholarship, but also demonstrated intellectual drive and leadership abilities.

Alexa will graduate as the 2018 First Degree Marshal with a Bachelor of Science in Hotel Administration and a minor in Real Estate. Upon graduation, she will be working as an analyst at Goldman Sachs.

Dedication

This thesis is dedicated to my family

For their endless love, support, and guidance throughout the years.

And to Father Paul

For his homilies and special blessings, which continue to strengthen my faith.

Acknowledgements

I would like to express my deep gratitude to all of those who have contributed to my personal and professional development, making it possible for me to complete this thesis.

To my advisor, Professor Anderson, for guiding me throughout this journey, despite not having me as a former student. Thank you so much for your assistance in solidifying a topic, instruction in gathering and analyzing the data, and support while researching and writing my thesis.

To all of my professors at SHA, who were always incredibly passionate about their subject matter. Thank you so much for sharing your incredible wealth of knowledge, being so willing to provide support, and guiding me to where I am today. In particular, thank you to Mark McCarthy and Professor Wolfe for providing both academic and emotional support throughout my four years at Cornell.

To Chef J, Chef B, and my ACAHA family, for assisting me in discovering my undeniable passion for hospitality at the Academy for Culinary Arts and Hospitality Administration. Thank you for introducing the idea of SHA to me, encouraging me to further my education in the field, and paving the way for my future.

To my grandparents, who provide endless amounts of love and happiness in my life. Grand, I can't put into words how much I treasure talking to you on a daily basis- our conversations are something I look forward to each and every day. Nonno, thank you for introducing me to the quote "Never, Never, Never Give Up" by Winston Churchill. I constantly wear the necklace you gave me and am reminded to constantly give my all. Grams and Pop, thank you for supporting me throughout this journey, constantly checking in on me, and making sure I am taking time to rest in the midst of all the stress. You all mean so much to me, and I am incredibly thankful to have you in my life.

Finally, to my Mom, Dad, Taylor, and Kristin (Bean), for absolutely everything. Thank you for supporting me through each and every milestone, encouraging me to pursue my goals, and making it possible for me to make my dreams become a reality. I genuinely do not know where I would be without your constant love, support, and guidance. You are the reason I am who I am today, and I can't thank you enough for everything. I love you always!

Table of Contents

Section I: Introduction

One of the greatest challenges facing hospitality companies today is the intense volume and pace of competition. The endless possibilities leave consumers with the opportunity to substitute, the ability to be price-sensitive, and the capacity to expect high levels of service (Kandampully & Suhartanto, 2000). The constant threat of substitutes and new entrants leads the hotel industry to rely heavily on competitive analysis to evaluate a property's performance and design specific competitive strategies. Although there is extensive research on pre-determined competitive sets, little research has been done to analyze hotel competitive sets from the consumer perspective.

Nowadays, it is not enough for hotels to benchmark themselves against their pre-determined competitive sets. According to Kim and Canina, although all luxury hotels belong to a specific product type, consumers may not consider all properties when making a decision and may even consider options outside of this product category (Kim & Canina, 2011). Discrepancies such as this one occur across all scales and classes.

With that said, this thesis uses TripAdvisor data to analyze consumer behavior. Which hotels did online viewers click on in the same session? Why? The study analyzes trends in target/competitor pairs across 11 cities internationally. The main focus of the research is based on the intensity of *common sessions*, which is defined as the number of sessions in which both the target property and the competitor property were viewed by a potential guest. This study determines the relationship between the intensity of common sessions and the following independent variables: scale (independent or branded), class (economy, midscale, upper midscale, upscale, upper upscale, luxury), distance (from target to competitor), and TripAdvisor score.

Overall, this thesis aims to find a generalizable model that explains consumer perception of hotel competitive sets. Which factor holds the most weight in pushing a consumer to click on a competitor? Is it the scale? Class? Location? TripAdvisor score? The answer to this question would help consumers think about which hotels they should be viewing as competitors while also showing the target hotels how they should benchmark their properties. After running regressions and performing cluster analysis, the results showed that there is an ability to generalize consumer views on competitive sets.

This thesis is organized in the following manner. Section II provides a literature review on four topics: the identification of hotel competitive sets, the significance of accurate identification, the inconsistency in the definition of competitive sets, and the objective of our competitive set research. Section III provides detailed information about the data sample. In Section IV, the regression results, visual displays, and cluster data will be revealed. Finally, Section V will discuss the significance of these results, conclude the study, and provide limitations of this research.

Section II: Literature Review

This literature review is comprised of 4 sections. They are organized in the following order: 1) The identification of hotel competitive sets 2) The significance of accurate identification 3) The inconsistency in the definition of competitive sets 4) The objective of our competitive set research.

Identification of Hotel Competitive Sets

Within the hospitality industry, hotel competitive sets play a vital role in providing an accurate depiction of a hotel's historical success and predicting future performance. According to Chen, a competitive set can be defined as "firms operating in the same industry, offering similar products, and targeting similar customers" (Chen, 1996). Competition, however, has recently become even more complex for hospitality companies due to the increasing use of the internet as a convenient and reliable search and transaction channel.

Online reviews published both on specialized websites (i.e. TripAdvisor.com), as well as on OTA websites (i.e. booking.com), are becoming an important focus of research (Filieri & McLeay, 2013). Online reviews can have a drastic effect on business due to the fact that they can positively or negatively alter a consumer's decision-making process. As a result, although hotels used to compete primarily with other hotels in close proximity, they now compete with hotels located farther if they have better online reviews or offer more appealing services, amenities, or rates (Li & Netessine, 2012).

With that said, understanding the competition structure in a market is crucial, especially when hotel executives are using the performance of the competitive set for benchmarking purposes. According to Kim and Canina, competitive set identification is required for hotel companies to generate strategy, recognize market position, and evaluate success (Kim & Canina,

2011). Interestingly enough, current practice for defining a competitive set in the hotel industry "ranges from looking across the street to identifying properties that charge the same basic rates (appealing to customers with the same price tolerance), and to weighing and scoring property attributes" (Li & Netessine, 2012).

While strategies vary, two approaches to competitor identification that have been applied throughout various industries include the following methods (Li, 2014):

- **Supply-based**- This approach is focused on the attributes of competing companies. According to the supply-based model, competitors are identified based on the product, service, resources, or strategies provided. Within the hotel industry, certain attributes may consist of room rate, location, scale, human capital resources, and organizational resources (Li, 2014).
- **Demand-based**- The demand-based model focuses primarily on the guests. More specifically, this approach looks at customers' purchasing behaviors to identify competitors. This approach tends to be more subjective, as it is based on the manager's understanding and perspective of the market and customer base. In addition, this model is more costly, as it requires research and analysis (Li, 2014).

Depending on upper-level management, certain properties may implement one or both of these strategies to define their competitors. Taking these approaches into account, it is important to note that the general process of competitive set identification in the lodging industry can typically be divided into 3 separate stages. First, managers classify their hotel's identity, which is comprised of several factors including tangible (i.e. location, size), intangible (i.e. brand image, reputation), and strategic (i.e. mission, vision). After, managers will begin market screening to find a list of properties with similar identities. This stage is where the strategies mentioned above

come into play. Finally, managers will match and select competitor hotels. This final step is where most of the subjectivity comes in, as managers end up choosing which properties will be included in competitive set (Li, 2014).

Significance of Accurate Identification

In an increasingly dynamic, consumer-driven environment, it is critical for hotels to accurately identify and immediately react to potential competitors (Webb & Zvi, 2017). The correct identification of a competitive set is essential for establishing competitive advantage and profitability, as shown below:

- **Competitive Advantage**- Competitive advantage becomes apparent when firms can offer lower prices than competitors or provide unique benefits that offset the higher price. The primary purpose of competitive analysis is to "evaluate a company's position in a market and try to keep ahead of competition through certain competitive advantages" (Li, 2014). The first step in working towards competitive advantage is getting the competitive analysis right and knowing who the competitors are. Once a competitive set is established, it is important to understand the key forces of supply and demand within a competitive environment in order to attain and sustain competitive advantage (Phillips, 1999).
- **Profitability**- When striving to increase profits, it is necessary to understand the competitive environment in order to prepare for price changes. Kim and Canina reveal that competitive pricing strategies affect both occupancy and RevPAR across market segments (Kim & Canina, 2011). In order to sustain long-term profitability, firms must respond strategically to competition (Porter, 1996). Again, defining the competitive set is the first step in order to accomplish this goal.

According to Kim and Canina, "If key competitors are left out of the analysis, the result could be misleading findings and poor strategic and competitive analysis" (Kim & Canina, 2009). Without the correct competitive set established, hotel executives will face much more of a challenge in attempting to attain success. As shown, it is crucial for hotels to accurately identify their competitive sets in order to understand the competitive atmosphere and both establish competitive advantage and achieve profitability in the industry.

Inconsistency in Definition of Competitive Sets

Due to the fact that the traditional approach for determining competitive sets typically entails a combination of methods, it becomes a subjective process. The subjectivity leaves room for error in terms of identifying and benchmarking against the competitive set. Extensive research on this discrepancy reveals the following issues:

1) **Fragmented nature of the hotel industry-** Inherently, hotels are different in terms of target market, ownership, management structure, location, size, amenities, and more (Webb & Zvi, 2017). Attempting to take all of these considerations into account leads to a very large list of potential competitors. With that said, hotel managers face difficulty in coming to a consensus on a single set of hotels for benchmarking purposes (Mohammed, 2014).
2) **Potential conflict of interest-** In the service industry, incentives are oftentimes tied to performance, and therefore, lead to deliberate oversight. When the hotel's management team selects hotels for their competitive set, they face conflicting financial and personal incentives to outperform that same competitive set. As a result, hoteliers have "acquired comp sets, changed comp sets, added comp sets, and used comp sets to understand the economic climate around them, base internal analysis,

indexes, and often performance bonuses" (Hillyard, 2011). This conflict of interest can lead to serious damage given the competitive set's crucial role as a performance assessment measure in the hotel industry.

As a result of these errors, there is oftentimes discrepancy in the makeup of competitive sets across market segments. As mentioned before, this inconsistency can be detrimental for the success of the hotel in the long run.

Objective of Competitive Set Research

As shown, most current competitive set identification methods are failing to reflect the true competitive position of any given property. The issues that result from the current processes could potentially be alleviated if consumer perception was the main focus of competitor identification. According to Kim and Canina, a hotel's position is heavily determined by the way the consumer views that property against its competition (Kim & Canina, 2009). The competitive set from a guest's perspective consists of the properties viewed as substitutes. This group of hotels that are alternatives to the initial search is exactly where the target hotel's true competition lies. While hoteliers must be aware of their hotel's position in terms of product tiers, consumer perception should be a crucial consideration when determining a competitive set (Kim & Canina, 2009).

With that said, this research aims to shed light on the idea of consumer perception of hotel competitive sets. Some initial questions that came to mind include the following: Which hotels did consumers click on after their initial search? How many competitors did they look at? Did they visit any of the competitor company websites? Answers to questions of this nature would begin to reveal which hotels customers perceive as competitors and why. Taking a

customer-centric approach, we study hotel competition through the use of TripAdvisor and STR data.

One work that is similar to ours in terms of the research goal is a paper by Jun Li and Serguei Netessine. They "not only construct the competition network from the customer perspective, but also compare it against the competition network from the hotelier perspective to examine the degree of network mismatch" (Li & Netessine, 2012). They focus on the notion that hotels should see themselves in the eyes of potential consumers because ultimately hotels are competing for customers. They argue that rather than asking themselves with whom they think they are competing with, hotel executives should ask who their customers identify as their competition. Instead of TripAdvisor and STR data, Li and Netessine analyze a similar research question with clickstream data (Li & Netessine, 2012).

Section III: Data Sample

The section below will discuss the details of the data that was used to conduct our study. It is split into the following three sections: TripAdvisor data, STR data, and data for analysis.

TripAdvisor Data

This research focuses on TripAdvisor data that was extracted in June and July of 2017. The data sample is international, as it includes information about the following 12 cities: Boston, Dallas, Denver, Dublin, Edmonton, Helsinki, London, New York City, Orlando, San Francisco, Sydney, and Warsaw. The Excel spreadsheet contains material regarding target/competitor pairs. For each target property, the data includes the top competitor properties in terms of the number of times they are viewed in the same session (up to a total of 25 competitors).

Within the set, there are 84,453 unique target/competitor pairs. There are also 23 variables, which include the following:

- t_prop_id
- t_property
- t_address
- t_country
- t_city
- c_prop_id
- c_property
- c_address
- c_country
- c_city
- **common_sessions**
- **same_sess_target_pageviews**
- **same_sess_competitor_pageviews**
- **same_sess_target_clicks**
- **same_session_competitor_clicks**
- **t_avg_score**
- **c_avg_score**
- **t_num_reviews**
- **c_num_reviews**
- **t_rank_percentile**
- **c_rank_percentile**
- **t_night_rate**
- **c_night_rate**

The list above shows terminology used by TripAdvisor for data analysis. Variable names beginning with “t” are associated with the target property and variable names beginning with “c” are associated with the competitor property. The definitions for the bolded terms are as follows:

Variable	Definition
common_sessions	Number of sessions in which both competitor property and target property were viewed
same_session_target_clicks	Number of clicks the target property received when viewed in the same session as the competitor
same_session_competitor_clicks	Number of clicks the competitor property received when viewed in the same session as the target property
same_session_target_pageviews	Number of target property pages viewed when viewed in the same session as the competitor property
same_session_competitor_pageviews	Number of competitor property pages viewed when viewed in the same session as the target property
t_avg_score, c_avg_score	Average score given in reviews in the past month
t_num_reviews, c_num_reviews	Number of reviews received in the past month
t_rank_percentile, c_rank_percentile	Average rank in the last month
t_night_rate, c_night_rate	Minimum nightly rate shown in the past month

Table 1: Definitions of TripAdvisor terms

From looking at the information, it is clear that *common sessions*, *same session target clicks*, *same session competitor clicks*, *same sessions target pageviews*, and *same session competitor pageviews* are the 5 variables that provide data about the correlation between the target and competitor property. In order to determine which metric would be most useful for our analysis, we considered the number of observations and calculated the means and standard deviations for each variable. The results can be found below:

Variable	Mean	Standard Deviation	# of Observations
common_sessions	58.57	403.74	84,453
same_session_target_clicks	401.76	1,864.88	84,453
same_session_competitor_clicks	437.97	3,694.45	84,453
same_session_target_pageviews	10.76	20.09	55,188
same_session_competitor_pageviews	10.92	19.81	58,845

Table 2: Analysis for variables that describe target/competitor pairs

After analyzing the data, it became clear that the common sessions variable was the most useful piece of information. The values for same session target pageviews and same session competitor pageviews tended to be extremely high numbers, causing the means and standard deviations to react accordingly. This is due to the inherent nature of the variable; one consumer could click on a property multiple times when deciding between different hotels. As a result, this data doesn't necessarily show the volume of consumers that are making the association between the target and competitor hotel, which is what this research is focused on.

In addition, the same session target clicks and same session competitor clicks variables did not have sufficient data across all cities, as shown by the number of observations for these two variables. Overall, common sessions proved to be the most valuable piece of information in discovering a relationship between the target and competitor properties. With that said, this study uses common sessions as the dependent variable to measure the intensity of the target/competitor association.

STR Data

STR data was used to analyze specific independent variables affecting common session intensity. The data collected included the class (economy, midscale, upper midscale, upscale, upper upscale, luxury), scale (branded or independent), location (urban, suburban, airport), price

(budget, economy, midprice, upscale, luxury), affiliation (hotel brand), and geographic position (latitude and longitude) for each target and competitor. This process shed light on the relationship between the pair and allowed for investigation on attributes in which consumers have perceived as characteristics that make properties similar.

It is important to note that by using the latitude and longitude coordinates, the distance between the target and competitor was calculated. This study displays the distance in miles. The formula that was used for this variable is as follows:

*= ((ACOS(COS(RADIANS(90-**Latitude_t**))*COS(RADIANS(90-**Latitude_c**))+SIN(RADIANS(90-**Latitude_t**))*SIN(RADIANS(90-**Latitude_c**))*COS(RADIANS(**Longitude_t-Longitude_c**)))*3959)*[1]

After calculating the distance, a holistic look at the data helped determine specific variables for the investigation. Taking the analysis approach into account and focusing on consumer perception, the following independent variables were chosen: scale (independent or branded), class (economy, midscale, upper midscale, upscale, upper upscale, luxury), distance (from target to competitor), and TripAdvisor score.

Data for Analysis

After gathering the data for all target/competitor pairs, certain factors were taken into account to ensure a strong sample. First, the data for Dublin, Ohio was deleted. The original intent in requesting data for Dublin was to capture information for Dublin, Ireland. Being that we received records for Ohio instead, the data turned out to be minimal and irrelevant to the scope of the study. Next, all targets without at least 16 competitors were deleted. This ensured that the study was focusing on target hotels with between 16 to 25 competitors listed. Moreover, all

[1] This formula was obtained from Microsoft (citation in Bibliography)

competitors with a distance of more than 60 miles were deleted to reduce the potential effect of outliers.

Finally, the common session variable was ranked for each city. In order to do this, the following formula was utilized:

=PERCENTRANK.EXC(All common session values, Value for one target/competitor pair)

To ensure standardization, the common sessions were ranked according to each city instead of across the targets or across the entire data set. When looking at the targets within each city, it became clear that certain targets had much higher common session averages than others. For example, in Sydney, the Adina Apartment Hotel Coogee had an average of **32.47** with a range of **30**. For the Amora Hotel Jamison, the average was **176.13** with a range of **221**. As a result, ranking the common sessions by target property would leave us with an inconsistent index.

In addition, when looking at the entire data set, some cities had much higher common session averages than others due to the use of TripAdvisor. For example, the common session average in NYC was **117.81** with a range of **24,050** whereas the common session average in Helsinki was **54.15** with a range of **226**. This shows that location generally has an effect on the amount that consumers use TripAdvisor. Once again, ranking common sessions by the entire data set would leave us with an inconsistent index.

After ranking the variable by city, the common session index was also converted into a qualitative measure. We broke the results down into two categories, high intensity and low intensity. If the common session rank was above 50%, it was considered a high intensity competitor, and if it was below 50%, it was considered a low intensity competitor.

Taking the changes listed above into account, the data sample consisted of the following breakdown:

City	Data Sample Size
Boston	1,767
Dallas	2,368
Denver	2,460
Edmonton	759
Helsinki	1,053
London	15,832
NYC	7,784
Orlando	4,450
San Francisco	4,371
Sydney	520
Warsaw	842
Total:	***42,206***

Table 3: Breakdown of data sample

Section V: Analysis & Results

The 3 main components of this study include regression, visual displays (scatter plots & histograms), and K-means clustering. The analysis process included moving back and forth between methods based on new findings. This section provides information about the 3 analysis components and reveals specific details about the process.

Discovering Relationships in the Data Set

In order to begin evaluating the data sample, we first applied regression analysis. This way, we could begin uncovering interaction between the independent and dependent variables.

Regression Summary

Regression analysis is a statistical method for studying the relationship between two or more variables. The main objective of the analysis is to arrive at a mathematical relationship which will predict values for one variable, called the dependent variable, based on the values of the remaining variables, called the independent variables. Regression analysis with only one independent variable is called a simple linear regression and analysis with two or more independent variables is called multiple regression. By considering more than one independent variable, multiple regressions are expected to develop a predictive equation that better fits the data than a simple linear regression equation would (Eldredge & Black, 2002). Since this study includes several independent variables, a multiple regression was utilized.

Regression Preparation

Within multiple linear regression, predictor variables may be defined quantitatively (continuous) or qualitatively (categorical) (Eldredge & Black, 2002). Since this study included both quantitative and qualitative information, we had to convert the qualitative information (STR data) into categorical values that could be put into the regression. The initial city-specific

regression focused on scale (independent or branded), class (economy, midscale, upper midscale, upscale, upper upscale, luxury), distance (from target to competitor), TripAdvisor score, and common session intensity. Common session intensity was the dependent variable and all other factors were independent variables.

Being that target scale, competitor scale, target class, and competitor class were all qualitative values, dummy coding was used to assign new values for the regression. Dummy coding assigns values "1" and "0" to reflect the presence and absence, respectively, of a specified variable (Gupta, 2008). The following steps were taken to prepare the data for regression:

Variable	Conversion
Target Scale *(Categorical)*	**Use dummy coding-** 0's represent branded target properties and 1's represent independent target properties
Competitor Scale *(Categorical)*	**Use dummy coding-** 0's represent branded competitor properties and 1's represent independent competitor properties
Competitor Class *(Categorical)*	**Use dummy coding-** *Competitor class worse* – 1's represent a competitor with a class that is "worse" than the target *Competitor class better* - 1's represent a competitor with a class that is "better" than the target *If both of these variables are a 0, then the target and competitor have the same class*
TripAdvisor ratio *(Continuous)*	Turn TripAdvisor scores into a single ratio using the following formula: $\frac{\textit{Competitor TripAdvisor Score}}{\textit{Target TripAdvisor Score}}$
Distance *(Continuous)*	Leave value as is *(formula discussed on page 12)*
Common Session Intensity *(Continuous)*	Leave value as is *(formula discussed on page 13)*

Table 4: Steps to prepare data for city-specific regression

As shown, dummy coding was very helpful in determining values for scale and class. Being that scale only had two options (independent and branded), simple dummy coding was effective. However, since there were many options for class (economy, midscale, upper midscale, upscale, upper upscale, luxury), the most efficient approach was to compare the competitor class to the target class. By doing this, we defined 3 variables: same, better, worse. As an example, if the target was an upscale property and the competitor was a midscale property, the competitor class would be *worse*. Since there are 3 values, we needed two dummy variables (worse class or better class), which were the predictors of the regression model. Each dummy variable was compared to the reference level (same class), which was coded as "0" for both dummy variables.

City-specific Regression

After preparing the data, we ran individual regressions for each city separately. The results for the coefficients of each independent variable are displayed below:

Regression #1) City-specific

	Boston	Dallas	Denver	Edmonton	Helsinki	London	NYC	Orlando	San Fran	Sydney	Warsaw
Target Scale	-0.028	0.122	0.054	-0.148	0.039	-0.110	0.008	0.132	-0.183	-0.144	-0.177
Competitor Scale	0.048	0.141	0.067	-0.054	0.059	-0.045	-0.032	0.005	-0.093	-0.130	-0.076
Competitior Class_Better	-0.143	-0.098	-0.050	-0.066	-0.154	-0.058	-0.057	-0.019	-0.065	-0.097	-0.094
Competitor Class_Worse	-0.040	-0.044	-0.025	0.018	-0.021	-0.057	-0.022	0.017	0.090	-0.048	-0.078
Trip Advisor Ratio	-0.410	-0.097	-0.112	-0.077	-0.226	-0.119	-0.240	-0.084	-0.137	-0.126	-0.269
Distance	-0.096	-0.018	-0.040	-0.018	-0.111	-0.033	-0.011	-0.011	-0.077	-0.049	-0.056

As shown, there is a large disconnect in both the signs and magnitude of the coefficients. When looking into the possible cause for this inconsistency, we started taking specific variables into account. When considering distance, we established that target and competitor hotels are typically much farther away from each other in cities like Orlando and much closer together in cities such as NYC. With that said, we realized that differences like this could be causing some variation in the data. Since the main goal of this research is to find a generalizable model that

works across all locations, we had to account for these differences in the individual cities. The following regressions attempted to account for inconsistencies across the 11 cities.

Additional Regression Attempts

Regression #2) Distance squared

	Boston	Dallas	Denver	Edmonton	Helsinki	London	NYC	Orlando	San Fran	Sydney	Warsaw
Target Scale	-0.017	0.122	0.073	-0.153	0.044	-0.099	0.007	0.135	-0.180	-0.162	-0.168
Competitor Scale	0.042	0.141	0.065	-0.051	0.070	-0.038	-0.032	0.004	-0.092	-0.117	-0.067
Competitior Class_Better	-0.145	-0.098	-0.045	-0.060	-0.165	-0.050	-0.056	-0.014	-0.065	-0.112	-0.087
Competitor Class_Worse	-0.034	-0.044	-0.011	0.025	-0.018	-0.049	-0.021	0.025	0.096	-0.056	-0.073
Trip Advisor Ratio	-0.419	-0.097	-0.125	-0.079	-0.235	-0.118	-0.240	-0.082	-0.141	-0.133	-0.264
Distance Squared	-0.014	-0.018	-0.002	-0.001	-0.023	-0.002	-0.002	0.000	-0.011	-0.002	-0.005

Regression #3) Both distance and TripAdvisor ratio squared

	Boston	Dallas	Denver	Edmonton	Helsinki	London	NYC	Orlando	San Fran	Sydney	Warsaw
Target Scale	-0.016	0.113	0.067	-0.145	0.042	-0.100	0.009	0.137	-0.182	-0.162	-0.167
Competitor Scale	0.039	0.147	0.067	-0.046	0.069	-0.038	-0.031	0.004	-0.092	-0.116	-0.067
Competitior Class_Better	-0.148	-0.101	-0.051	-0.064	-0.168	-0.053	-0.056	-0.016	-0.071	-0.112	-0.087
Competitor Class_Worse	-0.031	-0.033	-0.005	0.026	-0.017	-0.048	-0.021	0.024	0.098	-0.057	-0.072
Trip Advisor Ratio Squared	-0.160	-0.021	-0.026	-0.016	-0.083	-0.026	-0.092	-0.021	-0.032	-0.075	-0.122
Distance Squared	-0.014	-0.001	-0.002	-0.001	-0.023	-0.002	-0.002	0.000	-0.012	-0.002	-0.005

Regression #4) Distance squared and no target or competitor scales

	Boston	Dallas	Denver	Edmonton	Helsinki	London	NYC	Orlando	San Fran	Sydney	Warsaw
Competitior Class_Better	-0.147	-0.087	-0.042	-0.067	-0.160	-0.062	-0.056	-0.007	-0.077	-0.089	-0.077
Competitor Class_Worse	-0.032	-0.025	-0.009	0.005	-0.024	-0.068	-0.026	0.019	0.082	-0.065	-0.073
Trip Advisor Ratio	-0.406	-0.096	-0.122	-0.100	-0.230	-0.139	-0.237	-0.077	-0.161	-0.143	-0.281
Distance Squared	-0.014	-0.001	-0.002	-0.001	-0.023	-0.002	-0.002	-0.001	-0.011	-0.002	-0.005

Regression #5) Distance squared and no target or competitor scales or class

	Boston	Dallas	Denver	Edmonton	Helsinki	London	NYC	Orlando	San Fran	Sydney	Warsaw
Trip Advisor Ratio	-0.524	-0.106	-0.128	-0.110	-0.294	-0.144	-0.244	-0.079	-0.179	-0.150	-0.313
Distance Squared	-0.014	-0.001	-0.002	-0.001	-0.024	-0.002	-0.002	-0.001	-0.011	-0.002	-0.005

In certain cases, these attempts fix inconsistencies in the signs of the coefficients of the independent variables (i.e. more negative coefficients in #4 and #5). However, there is still a large discrepancy in the magnitudes across the cities. Take Regression #5 as an example. All coefficients are negative, but when considering the magnitudes of these coefficients, there is a disconnect. For TripAdvisor ratio, Orlando holds the maximum value (-0.079) and Boston holds

the minimum value (-.524), leaving a range of **.445**. This large range demonstrates differences across the cities.

In terms of regression statistics, the R-squared values, also known as coefficients of multiple determination, are statistical measures of how close the data fits the regression line. These values suggest that a specified percentage of the variability of the dependent variable (common sessions) can be explained by the independent variables in each of the respective regressions (Eldredge & Black, 2002). To provide more information on the regressions previously mentioned, the R-squared value for each is listed below:

	Boston	Dallas	Denver	Edmonton	Helsinki	London	NYC	Orlando	San Fran	Sydney	Warsaw
#1) City specific	0.362	0.224	0.325	0.200	0.318	0.187	0.054	0.090	0.358	0.301	0.415
#2) Distance squared	0.308	0.193	0.198	0.178	0.272	0.156	0.054	0.076	0.331	0.167	0.320
#3) Distance & TAR squared	0.308	0.190	0.189	0.174	0.269	0.153	0.059	0.078	0.326	0.169	0.321
#4) Distance squared & no scales	0.303	0.152	0.182	0.132	0.249	0.113	0.051	0.041	0.185	0.080	0.206
#5) Distance squared & no scales or class	0.255	0.124	0.176	0.117	0.171	0.100	0.043	0.039	0.143	0.063	0.189

Figure 1: R-squared value for all regressions across cities

Similar to the variable coefficients, there is inconsistency in the magnitude of these values across regression attempts. For example, in Regression #2, San Francisco holds the maximum value (0.331) and NYC holds the minimum value (0.054), leaving a range of **.277**. Again, this inconsistency highlights the variation within the data.

Taking the coefficient and R-squared values into account, we can infer that the regressions above show that while there is clearly interaction between the variables, the relationship is not linear. When applying this concept, the conclusion makes sense in terms of consumer perception. For example, in some cases, a consumer may be interested in a competitor who is extremely close by but has a lower TripAdvisor score, and in other cases, the consumer may consider a competitor who is farther away with a higher TripAdvisor score. As a result, we recognize that the relationship between these variables is intricate and complex.

Analyzing Distance and TAR Interaction

As mentioned above, all cities have negative distance and TripAdvisor ratio coefficients in the final regression. Even though the magnitudes greatly differ for TripAdvisor ratio, seeing the same sign across cities for both distance and TripAdvisor ratio is reassuring. It demonstrates that there is clearly interplay between these two variables. In order to get a better sense of the interaction between these two variables, the next step was to overlay our dependent variable, common session intensity.[2]

Since TripAdvisor ratio and distance were the two continuous independent variables in the regressions, it made most sense to display these quantitative values on a scatter plot. An example of the results for San Francisco is displayed below:

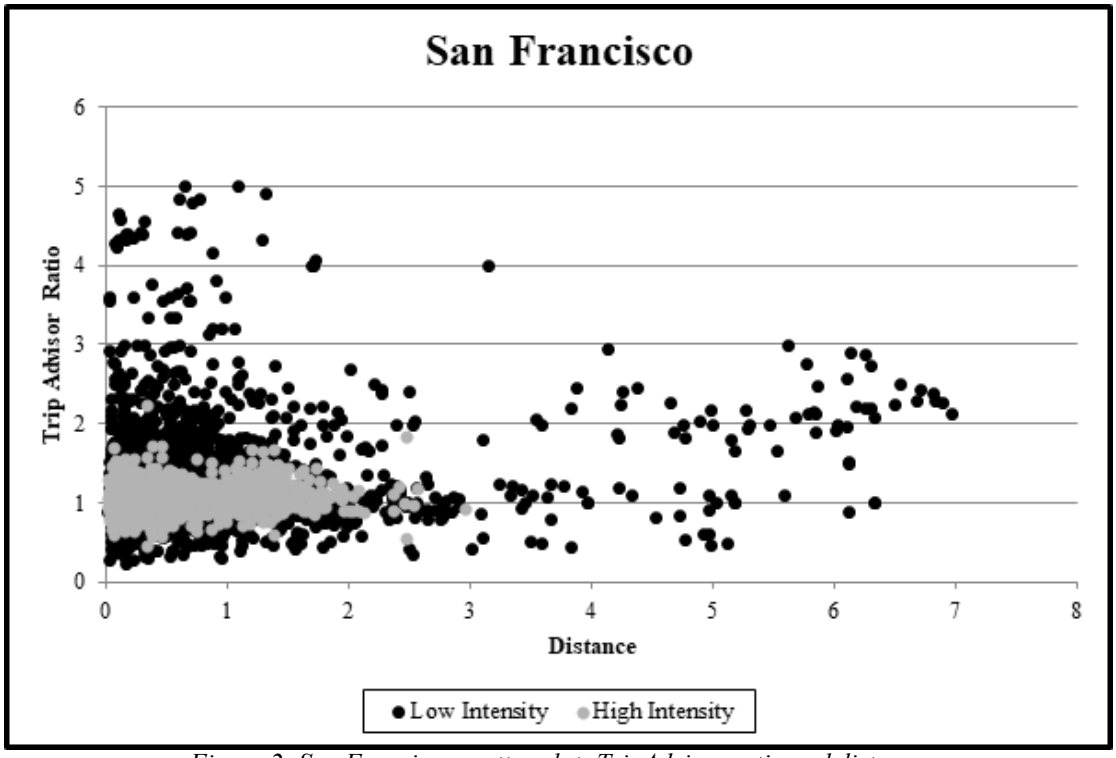


Figure 2: San Francisco scatter plot- TripAdvisor ratio and distance

[2] If the common session rank was above 50%, it was considered a high intensity competitor. If it was below 50%, it was considered a low intensity competitor.

In this scatter plot, there are 4371 total data points (2,369 for high intensity and 2,002 for low intensity). The graph shows the relationship between distance and TripAdvisor ratio for both the high and low intensity target/competitor pairs.

As shown, distance and TripAdvisor ratio are clearly impacting high and low intensity combinations in different ways. This scatter plot indicates that the high intensity common session pairs tend to have a similar TripAdvisor score (ratio close to 1) and be close by in terms of distance (<2 miles away).[3] On the other hand, while the low intensity common session pairs do have a large concentration near similar TripAdvisor score and close by in distance, these pairs are also scattered elsewhere. There is a large group of data points in the top left, indicating that the competitor has a higher TripAdvisor score (ratio >1) and is close by in distance. In addition, there is a group of data points in the bottom right area of the scatter plot. Within this group, some are similar in terms of TripAdvisor score (ratio close to 1) and some are different in terms of TripAdvisor score (ratio below 1 or well above 1), but all pairs are far away from each other compared to the majority of the sample (>3 miles away).

Overall, from looking at the plot, it is clear that as one independent variable increases (distance or TripAdvisor ratio), the number of high intensity target/competitor pairs decreases. However, if the distance goes up slightly and is still considerably close in terms of TripAdvisor ratio, there are still many high intensity competitors. This demonstrates that distance is moderated with TripAdvisor score. Overall, there is clearly interplay between TripAdvisor ratio and distance. However, it is not a linear relationship suitable for a regression.

[3] The formula for TripAdvisor ratio is *Competitor TripAdvisor Score/Target TripAdvisor Score*

Evaluating the Impact of Scale and Class

After establishing a relationship between distance and TripAdvisor ratio, we wanted to see how scale and class impacted our dependent variable, common session intensity. The histogram below displays the percentage of target hotels, low intensity competitors, and high intensity competitors by *class* (based on the full data set)[4]. For this analysis, the null hypothesis was that the distribution of class for the target hotels would look the same as the distribution of class for the competitor hotels.

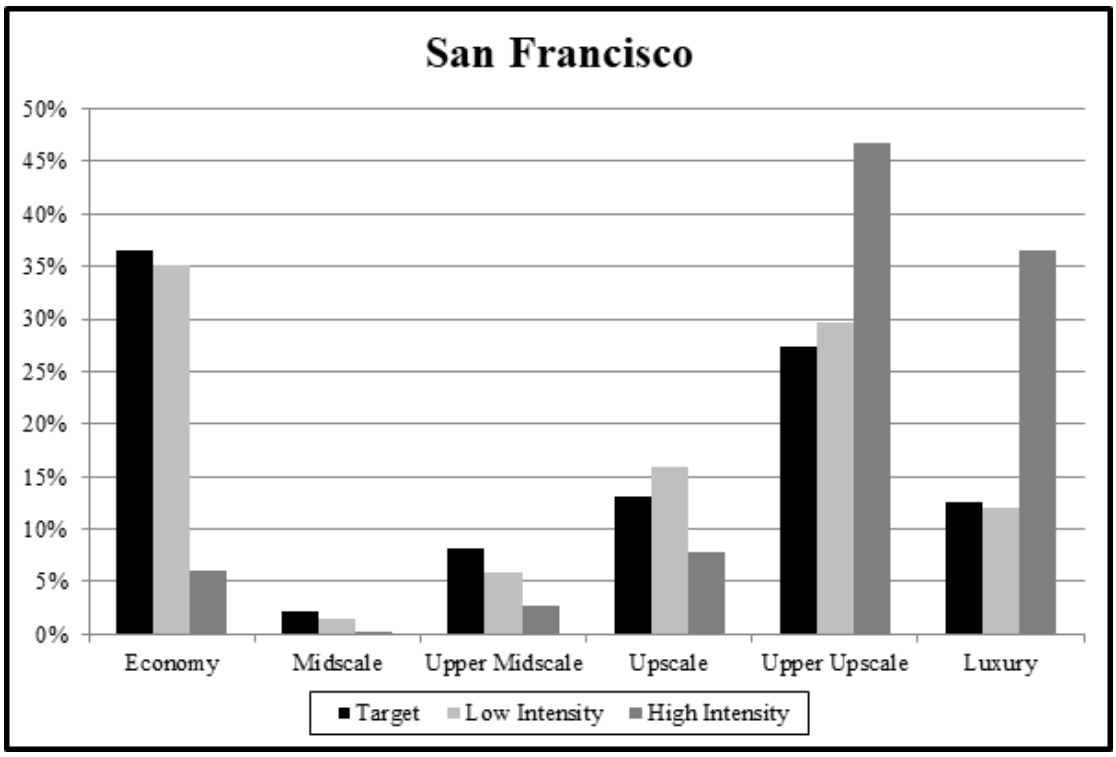


Figure 2: San Francisco histogram- class

As shown in the histogram above, the low intensity competitors closely mirror the target distribution. However, the high intensity competitors are clearly skewed toward the upper

[4] See page 14 for breakdown of the data sample

upscale and luxury properties. With that said, we would expect that the null hypothesis would be rejected in this case.

The second histogram shown below displays the percentage of target hotels, percentage of low intensity competitors, and percentage of high intensity competitors by *scale* (based on the full data set). Again, the null hypothesis was that the distribution of scale for the target hotels would look the same as the distribution of scale for the competitor hotels.

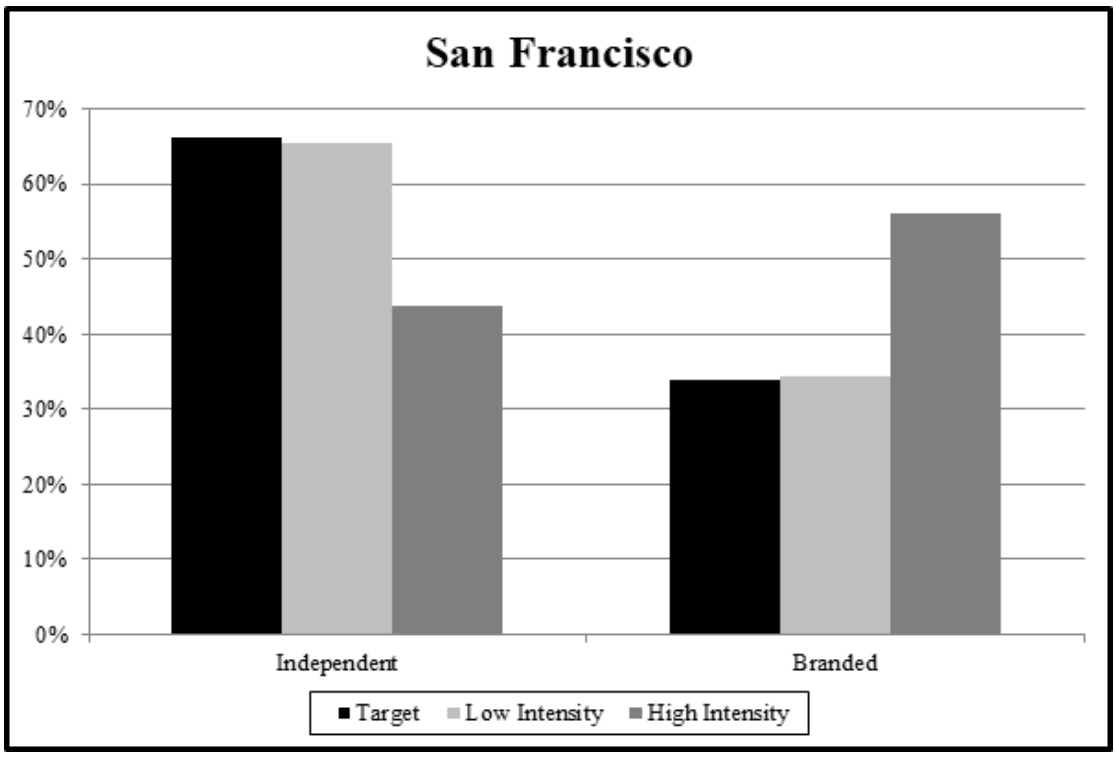


Figure 4: San Francisco histogram- scale

Here, we see that the low intensity competitors closely mirror the target distribution. However, the high intensity competitors are skewed toward banded hotels. Similar to the last example, we would expect that the null hypothesis would be rejected here as well.

In order to confirm or reject these assumptions, we calculated the chi-squared value. The chi-squared statistic is used for testing relationships between categorical variables. This test

allowed us to determine if the observed cell counts were significantly different from the expected cell counts (Eldredge & Black, 2002).

In terms of our research, we compared the actual counts of target hotel distributions (for low and high intensity competitors) to expected counts. We estimated the expected counts with following logic: *Step 1) Divide the actual distribution of target hotels by the total number of target hotels (i.e. 67/183 for economy class) Step 2) Multiply the number from step #1 by the total number of competitors for each intensity level (i.e. (67/183)*2,002 for low intensity)*. We completed this process for both low and high intensity. The chi-squared results for class in San Francisco (histogram displayed in Figure 3) are shown below:

	Actual Count			**Expected**	
Class	***Target***	***Low Intensity***	***High Intensity***	**Low Intensity**	**High Intensity**
Economy Class	*67*	*703*	*142*	**732.97**	**867.34**
Midscale Class	*4*	*28*	*6*	**43.76**	**51.78**
Upper Midscale Class	*15*	*117*	*64*	**164.10**	**194.18**
Upscale Class	*24*	*319*	*187*	**262.56**	**310.69**
Upper Upscale Class	*50*	*593*	*1,106*	**546.99**	**647.27**
Luxury Class	*23*	*242*	*864*	**251.62**	**297.74**
Total	183	2,002	2,369		
P-value				**6.60E-07**	**0.00E+00**

Table 5: Chi-squared test results for class

The chi-squared tests proved our expectation and therefore, rejected the null hypotheses. The actual and expected counts indicate dramatic differences between them with p-values < 0.001. The histograms and chi-squared results prove that variables matter, but again, the

relationship is not linear or continuous. This conclusion further demonstrates that the initial regression attempts didn't capture all of the interplay between the independent variables.

Standardizing the Data

In addition to proving that the relationship between variables is not linear, analyzing the visual displays across cities revealed inconsistencies in the continuous variables. When considering a specific example, NYC had a maximum distance of **9.22** miles and a maximum TripAdvisor ratio of **3.29**. Orlando, on the other hand, had a maximum distance of **25.91** miles and a maximum TripAdvisor ratio of **5**. This discrepancy led to very different scatter plot results between these cities, as shown below:

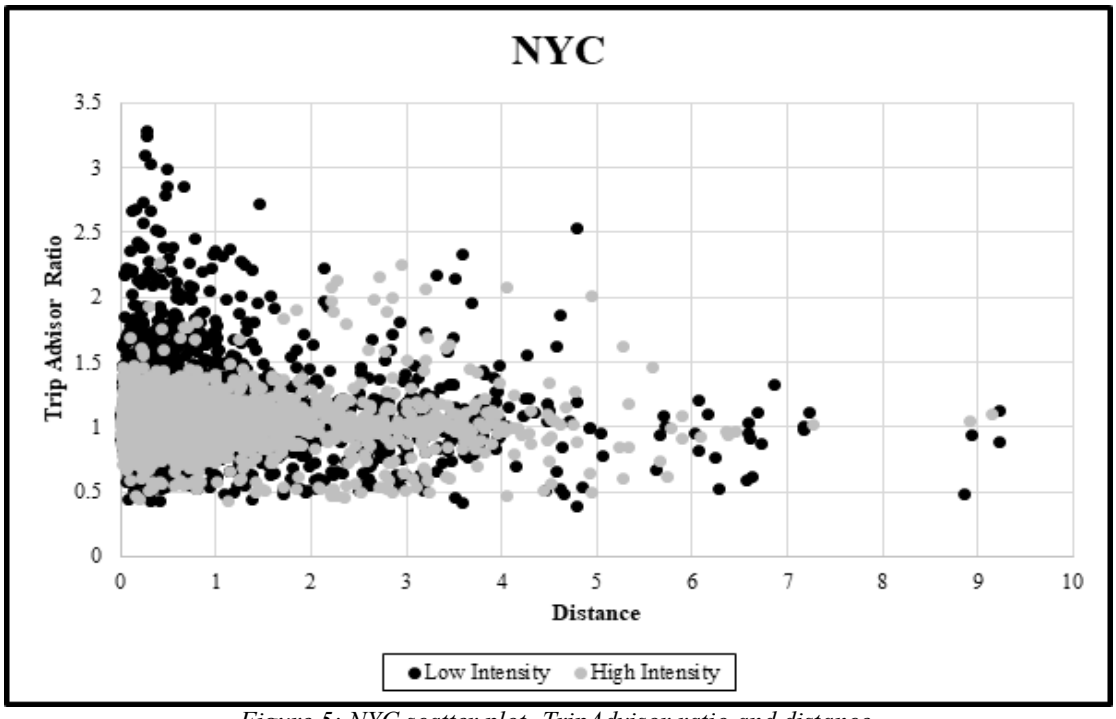


Figure 5: NYC scatter plot- TripAdvisor ratio and distance

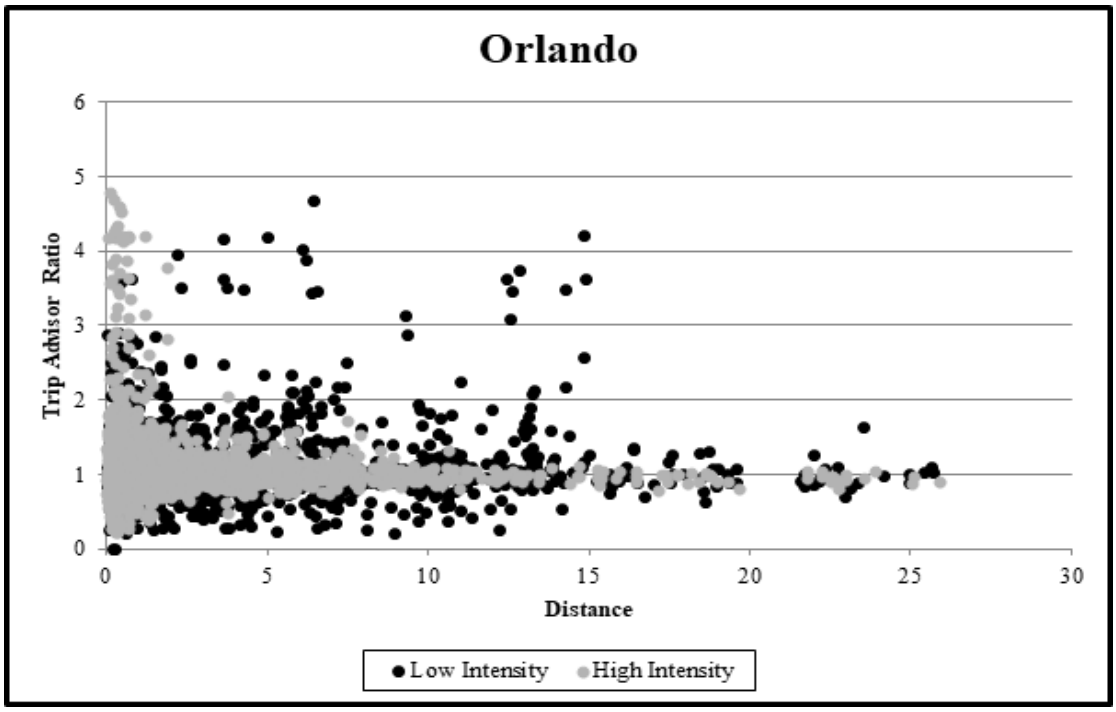


Figure 6: Orlando scatter plot- TripAdvisor ratio and distance

As shown by the maximum values on the axes labels, these two cities are on completely different pages in terms of TripAdvisor ratio and distance. In particular, the distance axis in Orlando is 3 times the distance axis in NYC. Taking this into account, we realized that in regression, it is difficult to put these cities on the same playing field.

As a result of this inconsistency, we had to standardize our continuous variables, TripAdvisor ratio and distance. For all cities, the means and standard deviations were calculated for these variables. The results are shown below:

City	TripAdvisor ratio		Distance	
	Mean	SD	Mean	SD
Boston	1.09	.20	1.18	1.15
Dallas	1.17	.59	3.82	3.82
Denver	1.18	.58	2.43	2.88
Edmonton	1.25	.78	4.46	3.58
Helsinki	1.07	.23	.88	.85
London	1.15	.48	1.49	2.38
NYC	1.07	.23	.74	.91
Orlando	1.14	.50	2.89	3.95
San Francisco	1.17	.46	.66	.83
Sydney	1.05	.14	1.59	2.67
Warsaw	1.04	.15	2.17	2.13

Table 6: Means and standard deviations of TripAdvisor ratio and distance

When looking at the means and standard deviations in NYC and Orlando, we found major differences. In particular, the mean distance for Orlando is more than 4 times the mean distance in NYC. This finding clearly demonstrates the need for data standardization. As a result, we used the following formula to standardize the TripAdvisor ratio and distance variables.

(Variable Value – Mean)/ Standard Deviation

This now puts all the cities on the same level for these two measures. After, we ran a regression with standardized TripAdvisor ratio and distance. We first ran the regressions for each of the cities and then ran a regression for all cities together. The results are shown below:

	Boston	Dallas	Denver	Edmonton	Helsinki	London	NYC	Orlando	San Fran	Sydney	Warsaw	ALL
R Square	0.307	0.157	0.306	0.140	0.234	0.118	0.043	0.056	0.169	0.210	0.263	0.113
S_TAR Coefficient	-0.105	-0.060	-0.066	-0.084	-0.063	-0.071	-0.057	-0.041	-0.081	-0.021	-0.049	-0.067
S_Distance Coefficient	-0.112	-0.064	-0.115	-0.058	-0.103	-0.063	-0.009	-0.049	-0.064	-0.132	-0.123	-0.060

Figure 7: Results from standardized regressions (city-specific and entire data set)

While the standardization process did eliminate inconsistency to a certain degree, the results for the individual cities still revealed some unique differences, as shown by the magnitudes. As a result, we came to the conclusion that the standardized regression for all cities was not sufficient in coming to a conclusion on consumer perception of hotel competitive sets.

Clustering to Model Distance and TAR Interaction

Since we couldn't capture all of the interaction between the independent variables, the regression results motivated the need to cluster variables based on similarity.

K-means Clustering Summary

Cluster analysis is a statistical method that classifies unknown groups of similar objects. It does not constrain the number of categories or predetermine the cutoff points. Instead, the number of categories and cutoff points are specific to the data sample. Put simply, cluster analysis identifies groups of homogenous objects by using underlying factors that drive the similarity (Mehra, 1996).

By summarizing the data into a small number of groups, the labels can provide a pattern of similarities and differences in the data. Statistically, a cluster is formed by minimizing the variance within a group (smaller variance implies that the objects are more similar) and maximizing the variance between the groups. In order to identify the group of homogenous objects by cluster analysis, it is important to pre-specify the factors that determine the similarity between the objects (Everitt, 2001).

Overall, cluster analysis offers several advantages to market researchers in the hospitality industry. Specifically, "the technique can be used to (a) develop typologies or classifications of customer groups, (b) define conceptual schemes for grouping customers, (c) use data to generate hypotheses about customer groups, or (d) test a concept to determine if specific types of

customers are present in a data set" (Aldenderfer et al., 1984; Romesburg, 1984). In our study, the cluster analysis mainly helped us develop classifications of consumer groups.

Cluster Analysis

In K-means clustering, the researchers have to make two major decisions: 1) which variables to use to divide the sample and 2) how many clusters are optimal (Everitt, 2001). In this study, we chose the standardized distance and standardized TripAdvisor ratio as the two variables for clustering.

In order to run the cluster analysis, an online template from *Cluster Analysis for Marketing* was downloaded and modified (Fripp, 2016). The standardized distance and standardized TripAdvisor ratio values (for target/competitor pairs across all cities) were used in the cluster model. Since we used the entire data set, we had 42,206 cases for the analysis (see page 14 for a breakdown of the total).

When choosing the number of clusters, there are a few important things to keep in mind. First, only one option should be chosen (two, three, four or five segments/clusters). Next, since the results of the cluster analysis are statistically derived, there is no connection in specific clusters across the different outputs. Each segmentation (i.e. 2 clusters, 3 clusters, etc.) is unique. Finally, it is important to look for a decent distribution across the segments (i.e. don't have 90% of all the results in one segment). While doing this, consider an output with a lower sum of squared error (doesn't need to be the lowest, but avoid the highest) (Fripp, 2016).

Taking these considerations into account, the segmentation map that made the most sense in our analysis was **3 clusters**. This cluster grouping avoided overlap and minimized the variance within groups while maximizing the variance between groups. A segmentation map for the results is displayed below:

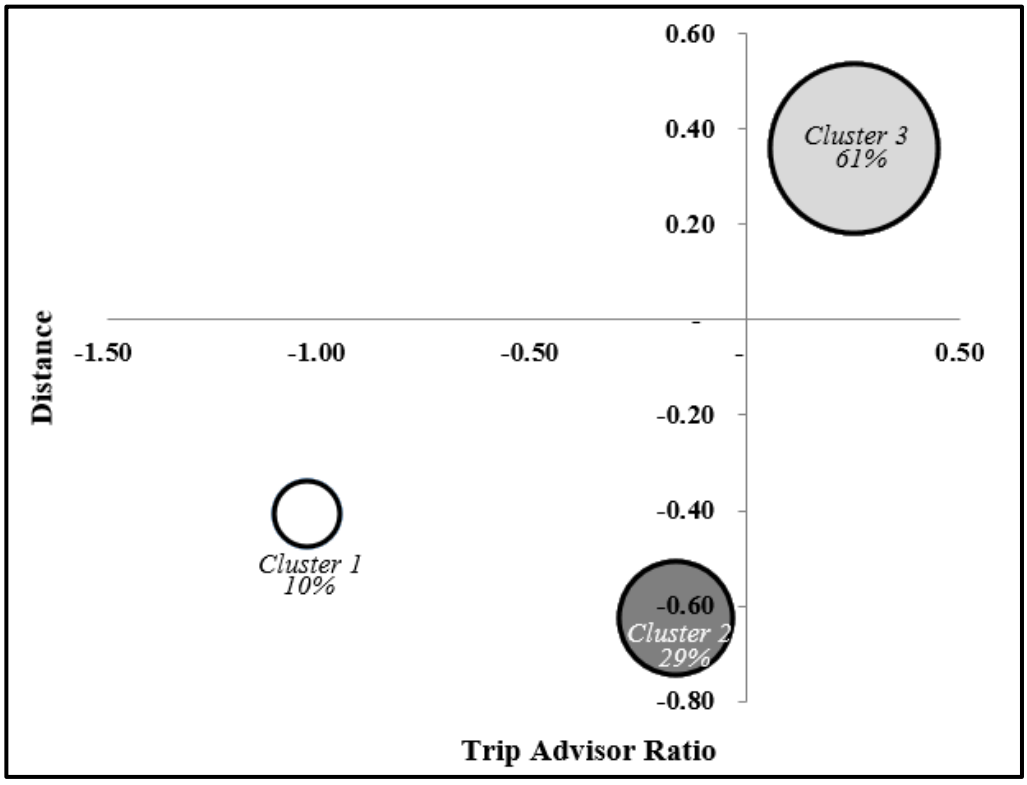


Figure 8: Cluster results

As shown from the segmentation map, cluster 1 accounts for 10% of the data, cluster 2 accounts for 29% of the data, and cluster 3 accounts for 61% of the data. Cluster 1 represents the competitors that are clearly not a match- both far away in distance and worse in terms of TripAdvisor score. Cluster 2 represents the competitors that are close by in distance and have a lower TripAdvisor score. Finally, cluster 3 represents competitors that are close by in distance and have a better TripAdvisor score.

It is important to note that when clustering, our main goal was to understand the common session index across the data sample. As mentioned above, this exercise revealed 3 distinct clusters, which provided insight into what was truly going on with the common session intensity. After clustering the data, the last step in the analysis was to run a final regression taking the cluster results into account.

Cluster-based Regression Preparation

As mentioned, with multiple linear regression, predictor variables may be defined quantitatively or qualitatively (Eldredge & Black, 2002). In this case, the cluster data had to be converted into categorical values that could be used as inputs for the regression.

The final regression used cluster, class (chain scale difference), and common session intensity. Again, common session intensity was the dependent variable and all other factors were the independent variables. The following steps were taken to prepare the data for regression:

Variable	Conversion
Cluster *(Categorical)*	**Use dummy coding-** *Cluster 1* – 1's represent a target/competitor pair in cluster 1 *Cluster 2* – 1's represent a target/competitor pair in cluster 2 *If both of these variables are a 0, then target/competitor pair is in Cluster 3*
Chain Scale Difference *(Continuous)*	Convert both target and competitor chain scale into numerical values *(economy= .5, midscale= 1, upper midscale= 1.5, upscale= 2, upper upscale= 2.5, luxury =3)* Use the following formula: *Target chain scale value – Competitor chain scale value*
Common Session Intensity *(Continuous)*	Leave value as is *(formula discussed on page 13)*

Table 7: Preparation for final regression

Cluster-based Regression

After preparing the data, we ran a regression for the entire data set (all 42,206 target/competitor pairs). The results are displayed below:

	Coefficients	Standard Error	t Stat	P-value
Intercept	0.516	0.002	309.962	0.000
Cluster 1	-0.002	0.004	-0.511	0.609
Cluster 2	0.136	0.003	47.655	0.000
CS Change	0.045	0.002	27.595	0.000

Figure 9: Final regression results[5]

The results of this regression reveal that the cluster 2 and chain scale change independent variables are significant (p-value < .05). Since cluster 2 represents competitors that are close by in distance but have a lower TripAdvisor score, the significance in this variable reveals a strong relationship between consumer perception and this bucket of competitors. The significance in the chain scale change variable shows that the number of common sessions increases as the target chain scale becomes better than the competitor chain scale.[6] These variable results are encouraging, as they begin to reveal what really matters in the eyes of consumers.

The cluster 1 variable, on the other hand, is insignificant, suggesting that this cluster is essentially no different than cluster 3 (as a result of the dummy coding applied). Taking these results into account, we began to further investigate the similarities between cluster 1 and cluster 3. As mentioned, cluster 1 represents the competitors that are both far away and worse in terms of TripAdvisor score, and cluster 3 represents competitors that are close by and have a better TripAdvisor score. Interestingly enough, these clusters correspond to the bottom right portion and top left area of the scatter plot discussed in Figure 2 respectively. With that said, clusters 1

[5] The R Square value was .071
[6] The formula for Chain Scale Change is *Target Chain Scale-Competitor Chain Scale (see page 31 for more details)*

and 3 represent the two regions in Figure 2 that are highly concentrated with purely low intensity target/competitor pairs. The regression results suggesting that these two clusters are no different is a logical argument being that our dependent variable is common session intensity.

After establishing that cluster 1 was insignificant, we ran the regression once more without this variable. The results are shown below:

	Coefficients	Standard Error	t Stat	P-value
Intercept	0.516	0.002	336.537	0.000
Cluster 2	0.136	0.003	48.941	0.000
CS Change	0.045	0.002	27.697	0.000

Figure 10: Final regression results without cluster 1[7]

Once again, the results for cluster 2 and chain scale change are significant with p-values <.05. As mentioned, the positive coefficient for cluster 2 suggests that after their initial search, consumers tend to click on properties that are nearby with a lower TripAdvisor score. The positive coefficient for chain scale change reveals that consumers click on competitors with lower chain scales than their initial search.

Interestingly enough, these conclusions go hand in hand. If consumers are looking at competitors with lower chain scales compared to the target property, one would expect that the TripAdvisor scores would be lower as well. When considering our data sample, we see that this assumption holds true. The average TripAdvisor score increases as the class gets better (economy, midscale, upper midscale, upscale, upper upscale, luxury). The count (distribution of competitors within the data sample) and average TripAdvisor score for each class are displayed below:

[7] The R Square value was .071

Class	Count	Average TripAdvisor Score
Economy	5,207	3.61
Midscale	3,325	3.64
Upper Midscale	5,492	3.89
Upscale	9,390	4.13
Upper Upscale	11,427	4.31
Luxury	7,365	4.55
Total	***42,206***[8]	

Table 8: Count of competitors and average TripAdvisor score by class

After analyzing this data, we came to the conclusion that consumers are looking at competitors who are close in distance with lower chain scales, and therefore, lower TripAdvisor scores. This finding suggests that consumers are looking for value and making price-based decisions for nearby hotels.

[8] For a breakdown of this total by city, please see page 14

Section V: Summary

Taking the analysis components into account, it is clear that individual cities have unique trends with specific attributes/independent variables. As discussed, the discrepancy in both TripAdvisor ratio and distance in NYC and Orlando demonstrated a need to standardize the data. The main reason for this standardization was to ensure that all cities were on an even playing field. This step was critical when considering the goal of this study, which was to establish one generalizable model to explain consumer perception of hotel competitive sets.

Once we standardized the data across cities, K-means clustering helped tell the story behind the sample. What was really going on with the target/competitor pairs? Was there a way to classify these combinations? The cluster analysis answered these questions and consisted of the following segments: 1) Competitors that are clearly not a match- both far away and worse in terms of TripAdvisor score 2) Competitors that are close by the target in distance but have a lower TripAdvisor score 3) Competitors that are close by the target in distance and have a better TripAdvisor score.

Although there were 3 distinct segments in the cluster exercise, the final regression showed that only one cluster, cluster 2, was significant. Taking this into account, we essentially have two main buckets across the common session index. The first bucket consists of competitors that are close by in distance and have a lower TripAdvisor score, which correlates with cluster 2. The second bucket consists of all target/competitor pairs that do not fall into the previous category; this comes from the idea that cluster 1 and cluster 3 are no different according to the regression. As a result of this finding, we came to the following conclusion: ***Consumers are most likely to consider properties that are close by in distance with a lower chain scale, and therefore, lower TripAdvisor score, after their initial search for a lodging property.***

Limitations and Future Research

Although the study results are encouraging and provide insight into consumer perception of hotel competitive sets, it is important to point out multiple limitations of our research. While addressing the restraints of our study, we also bring up opportunities for future research.

First, the study focused on a single sample of TripAdvisor data from June and July of 2017. Within the set, there were 84,453 unique target/competitor pairs. With that said, when considering the amount of data that TripAdvisor collects on a daily basis, this study's sample was relatively small.

In addition, while this research includes international data, it only accounts for 11 cities across the globe. Moreover, 6 out of the 11 cities analyzed are in the United States. With that said, the sample certainly was not as diverse as it could have been. While we did notice inconsistencies across the data (which led us to standardize the sample), other discrepancies may have been brought to our attention if the scope of the study was larger.

Moreover, while we did look into independent versus branded hotels, the study did not research how specific brands play a role in altering consumer perception. For example, if a consumer initially searches for a Marriott hotel, is he/she more likely to click on another Marriott property? This is a valuable question to consider and would be an interesting topic to investigate for future research.

In another light, this research does not provide any insight into how resorts or specified all-inclusive packages fit into this analysis. Is there a similar consumer perception consensus among these types of properties? Again, this is another interesting question that would provide additional insight into the idea of consumer perception.

Finally, this study does not analyze how environmental & green initiatives impact consumer perception of hotel competitive sets. If a consumer initially searches for a hotel that has many green initiatives in place, is he/she more likely to click on another property that offers similar benefits? Again, this is another interesting topic that would be useful to investigate in order to get a better sense of consumer perception of hotel competitive sets.

Bibliography

Aldenderfer, M. S., & Blashfield, R. K. (1984). Cluster analysis. Newberry Park, CA: Sage.

Buhalis, D. (2003). ETourism: Information Technology for Strategic Tourism Management, Financial Times Prentice Hall, Harlow.

Chen, MJ (1996). Competitor analysis and interfirm rivalry: toward a theoretical integration. Academy of Management Review 21(1): 100–134.

Eldredge, David L., and Ken Black. A Microsoft Excel Companion for Business Statistics. South-Western, 2002.

Everitt, B. (2001). Cluster Analysis (4th ed.). New York: Oxford University Press.

Filieri, R. and McLeay, F. (2013). "E-WOM and accommodation: an analysis of the factors that influence travelers' adoption of information from online reviews", Journal of Travel Research.

Fripp, Geoff. "Free Download of the Cluster Analysis Template." Cluster Analysis 4 Marketing, 2016, www.clusteranalysis4marketing.com/free-download/.

Gupta, Resmi. "Cornell Statistical Consulting Unit." Coding Categorical Variables in Regression Models: Dummy and Effect Coding, May 2008,

Hillyard C (2011) Comp Sets Revisited. Lodging. http://lodgingmagazine.com/comp-sets-revisited.

Kandampully, Jay & Suhartanto, Dwi (2000) "Customer loyalty in the hotel industry: the role of customer satisfaction and image", International Journal of Contemporary Hospitality Management, Vol. 12 Issue: 6, pp.346-351.

Kim, J. Y., & Canina, L. (2009). Product tiers and ADR clusters: Integrating two methods for determining hotel competitive sets. Cornell Hospitality Report, 9(14), 6-18.

Kim, J. Y., & Canina, L. (2011). Competitive sets for lodging properties. *Cornell Hospitality Quarterly, 52*(1), 20-34.

Lee, H., Guillet, B.D. and Law, R. (2013), "An examination of the relationship between online travel agents and hotels a case study of choice hotels international and Expedia.com", Cornell Hospitality Quarterly, Vol. 54 No. 1, pp. 95-107.

Li, Jing. "An Agency Perspective on Hotel Competitive Set Identification." Cornell University, 2014, pp. 1–48.

Li, Jun, & Netessine, Serguei (2012). Who are my Competitors? Let the Customer Decide. The Business School for the World, 1-35.

Mehra, A. (1996). Resource and market based determinants of performance in the U.S. banking Industry. Strategic Management Journal, 17(4), 307-322.

Microsoft Community, 4 Nov. 2016, answers.microsoft.com/en-us/msoffice/forum/msoffice_excel-mso_winother-mso_2016/formula-not-working-in-new-spreadsheet/b5cb7a17-8501-4294-91d6-aacb95c6a973.

Mohammed, Ibrahim. "Competitor Set Identification in the Hotel Industry: A Case Study of a Full-Service Hotel in Hong Kong." School of Hotel and Tourism Management, Hong Kong Polytechnic University, International Journal of Hospitality Management, 2014, pp. 1–12.

Phillips, Paul A. (1999) "Hotel performance and competitive advantage: a contingency approach", International Journal of Contemporary Hospitality Management, Vol. 11 Issue: 7, pp.359-365.

Porter, Michael E. “HBR’s Must-Reads on Strategy.” What Is Strategy?, Nov. 1996.

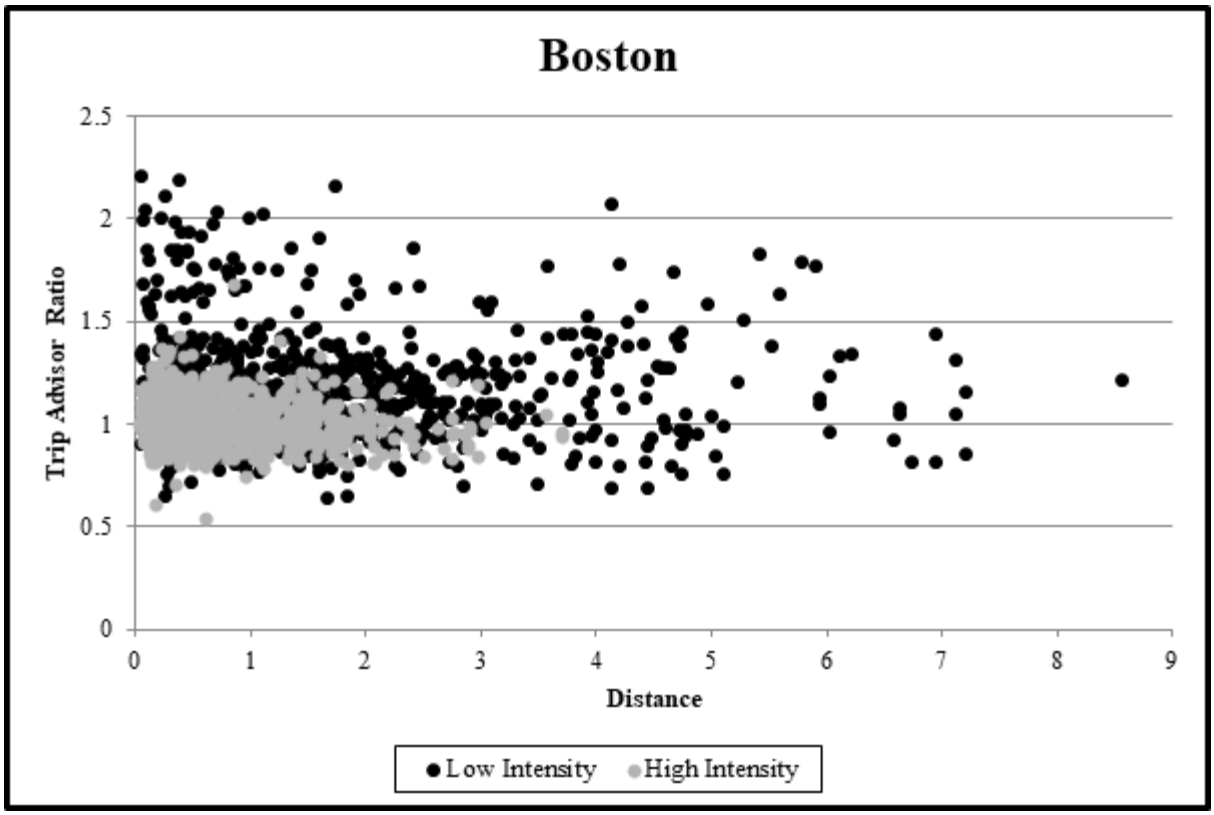
Romesburg, H. C. (1984). Cluster analysis for researchers. Belmont, CA: Wadsworth.

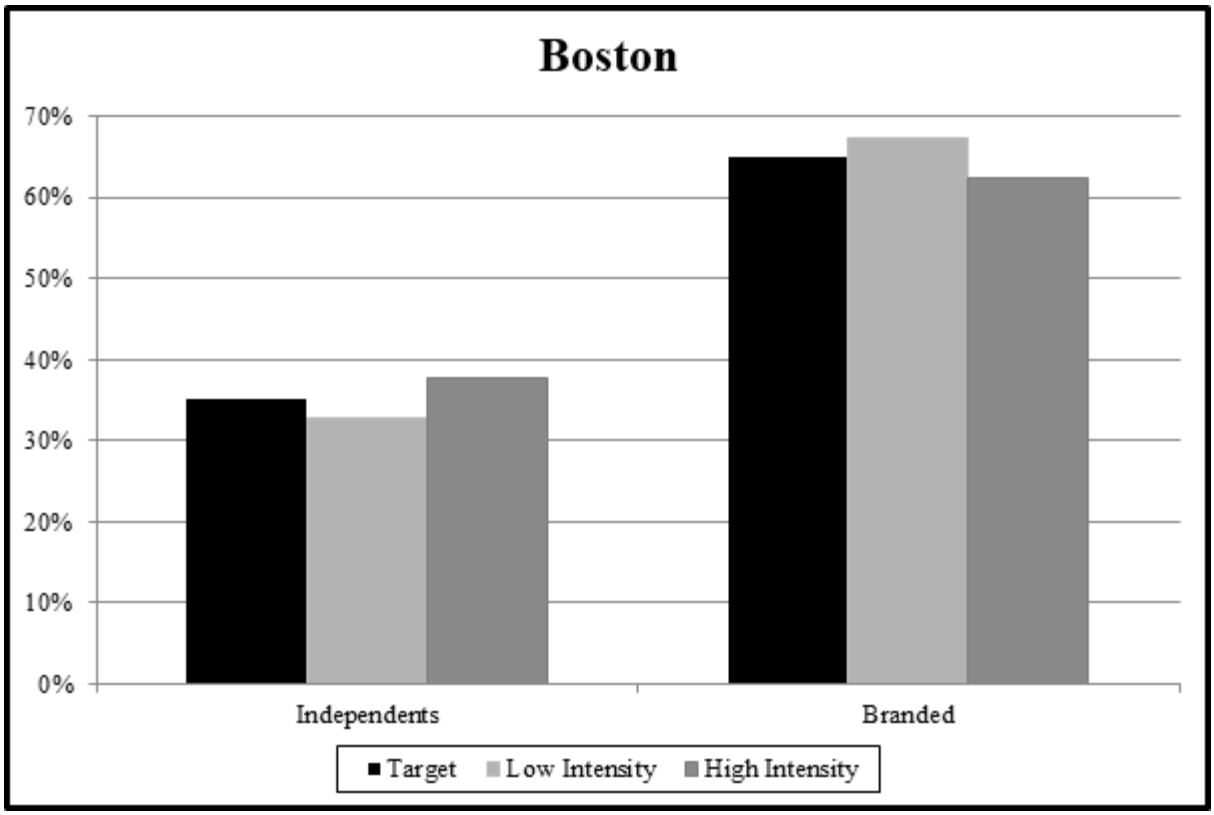
Webb, Timothy, and Zvi Schwartz. “Revenue Management Analysis with Competitive Sets: Vulnerability and a Challenge to Strategic Co-Opetition among Hotels.” *Virginia tech; University of Delaware*, Tourism Economics, 2017, pp.1-14.

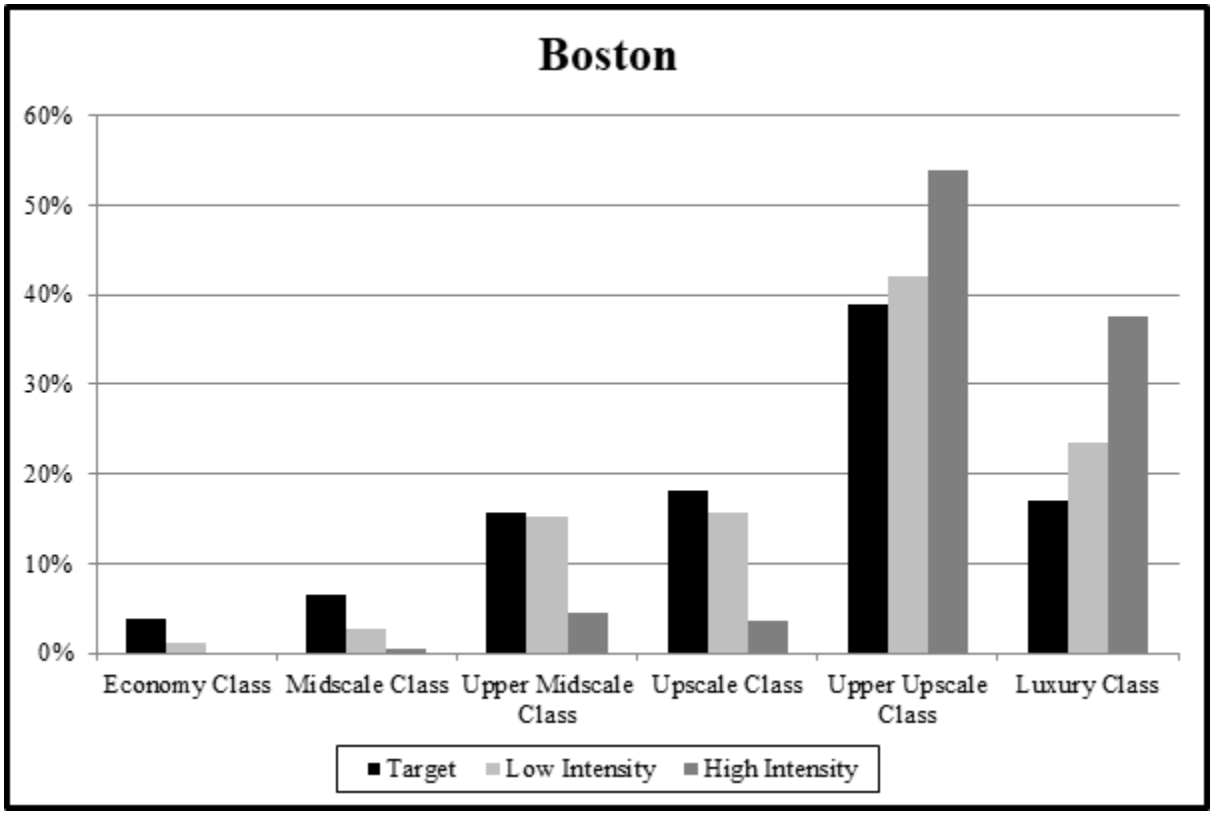
Appendix

The figures below include the scatter plots and histograms discussed in the paper (pages 19-23) for each of the 11 cities (in alphabetical order). Once again, the 11 cities include: Boston, Dallas, Denver, Edmonton, Helsinki, London, New York City, Orlando, San Francisco, Sydney, and Warsaw.

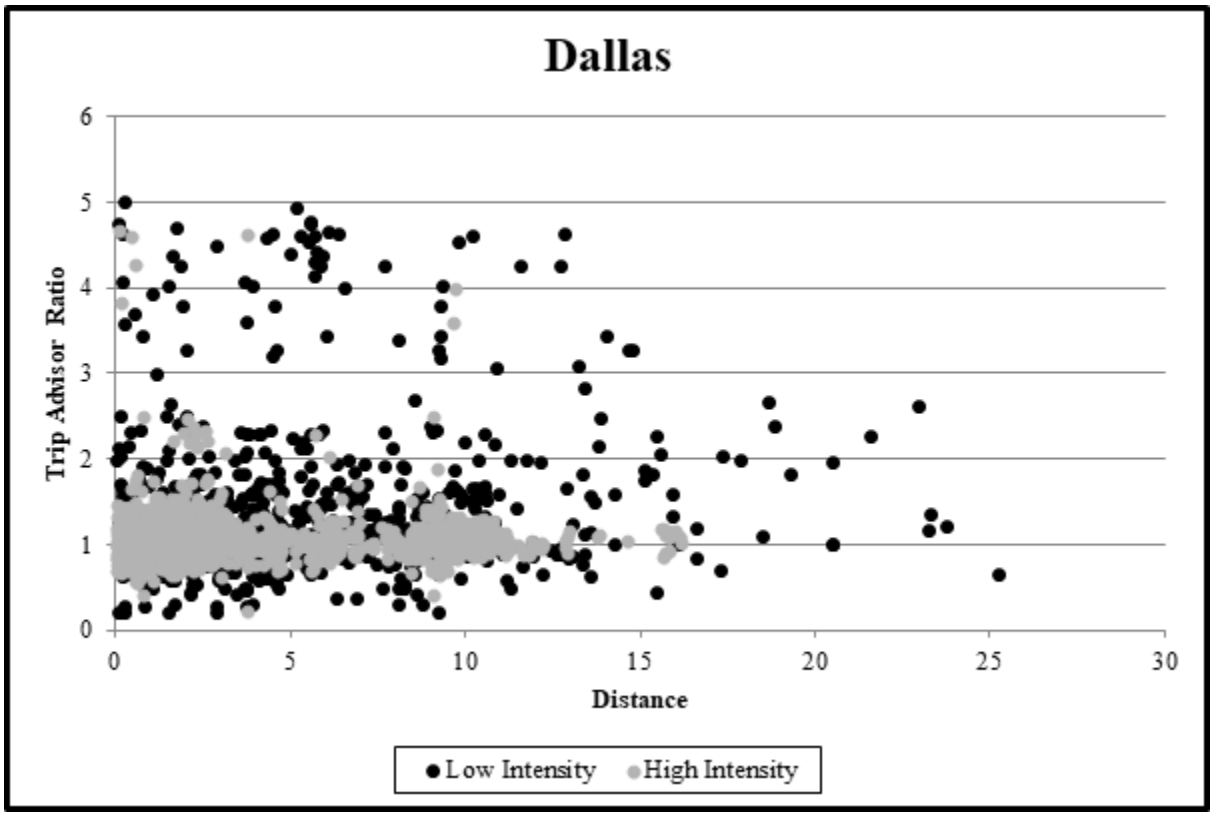
The order of the figures for each city is as follows: 1) Scatter plot representing the interaction between TripAdvisor ratio and distance 2) Histogram showing the distribution of *scale* (independent or branded) across targets, low intensity competitors, and high intensity competitors 3) Histogram showing the distribution of *class* (economy, midscale, upper midscale, upscale, upper upscale, luxury) across targets, low intensity competitors, and high intensity competitors.

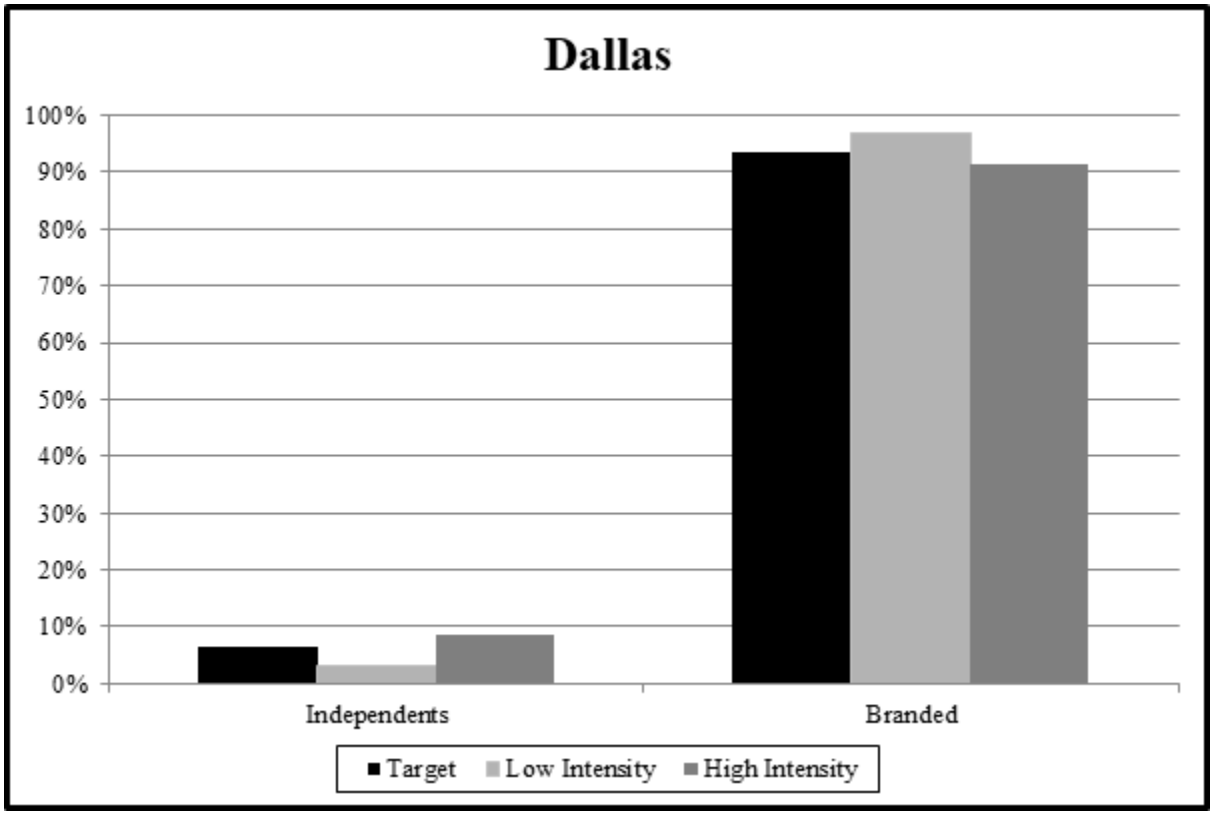


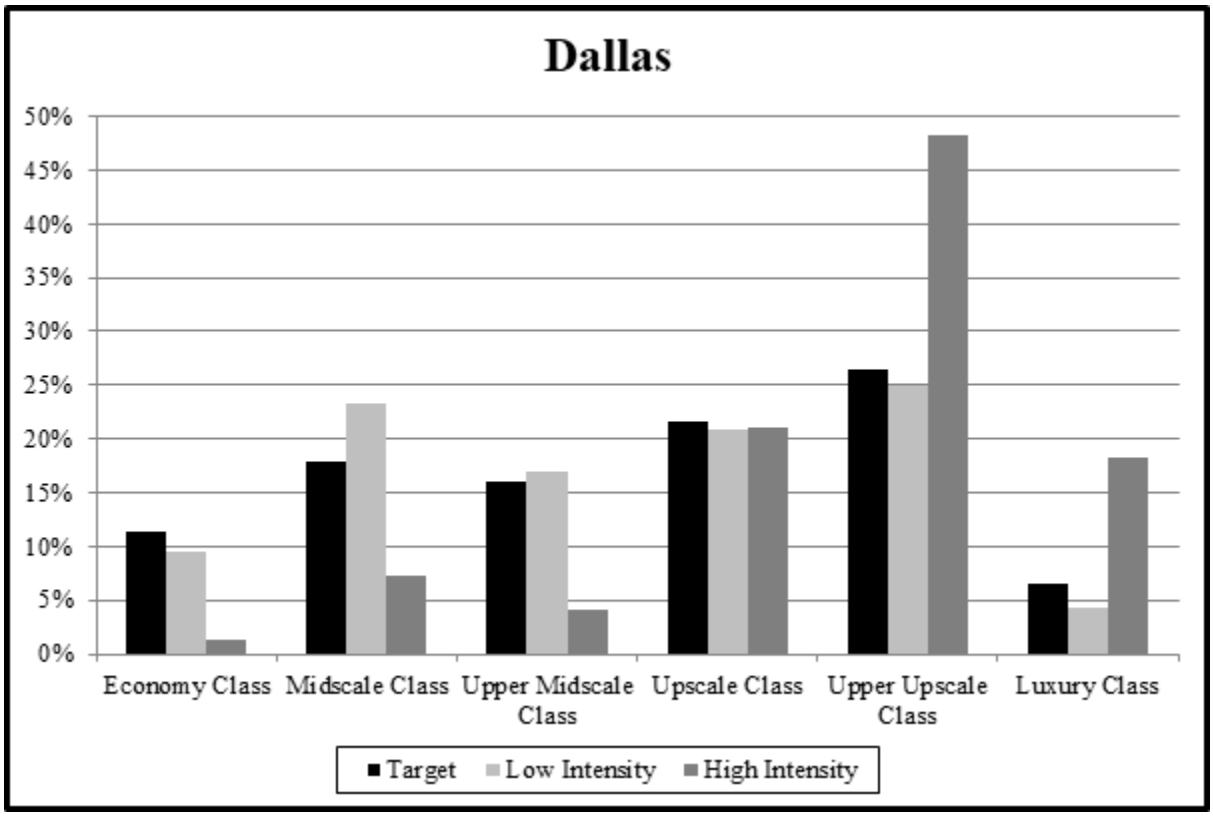

Boston
70%
60%
50%
40%
30%
20%
10%
0%
Independents
Branded
Target
Low Intensity
High Intensity

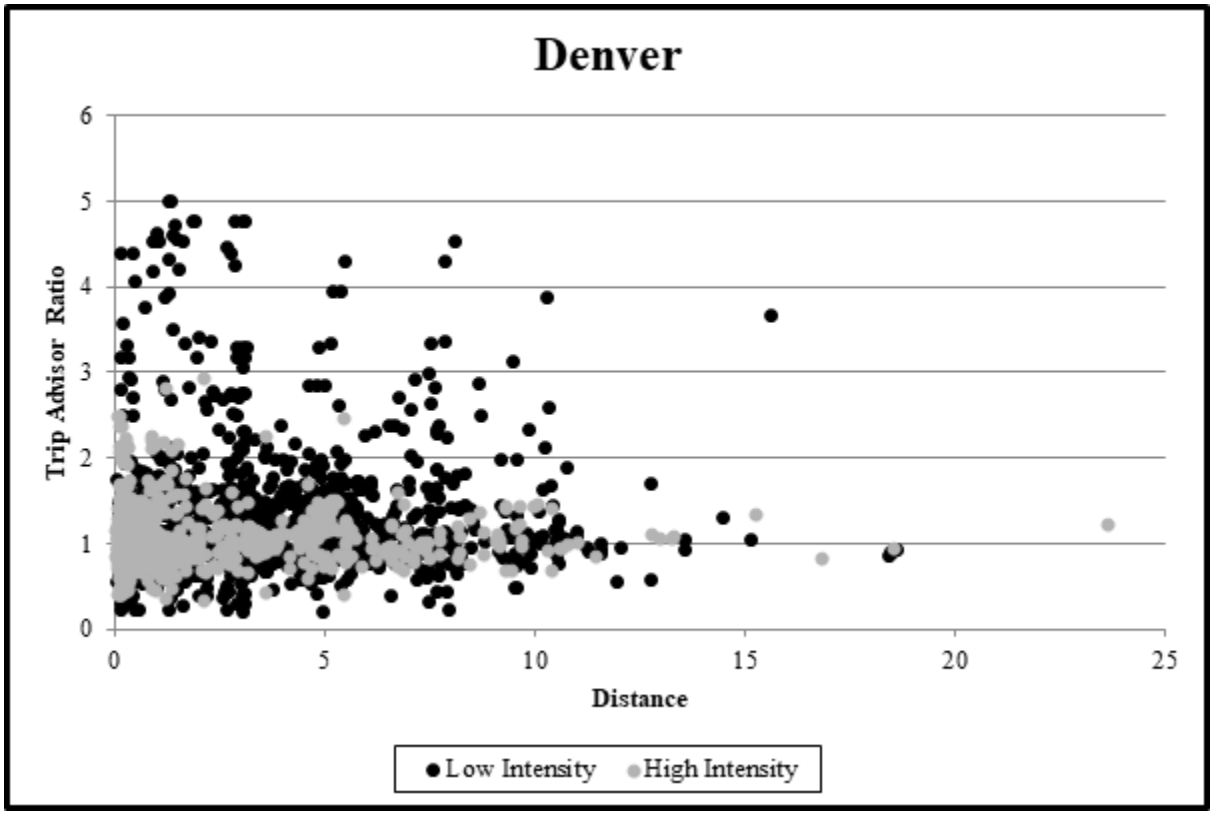

Boston
60%
50%
40%
30%
20%
10%
0%
Economy Class
Midscale Class
Upper Midscale Class
Upscale Class
Upper Upscale Class
Luxury Class
Target
Low Intensity
High Intensity

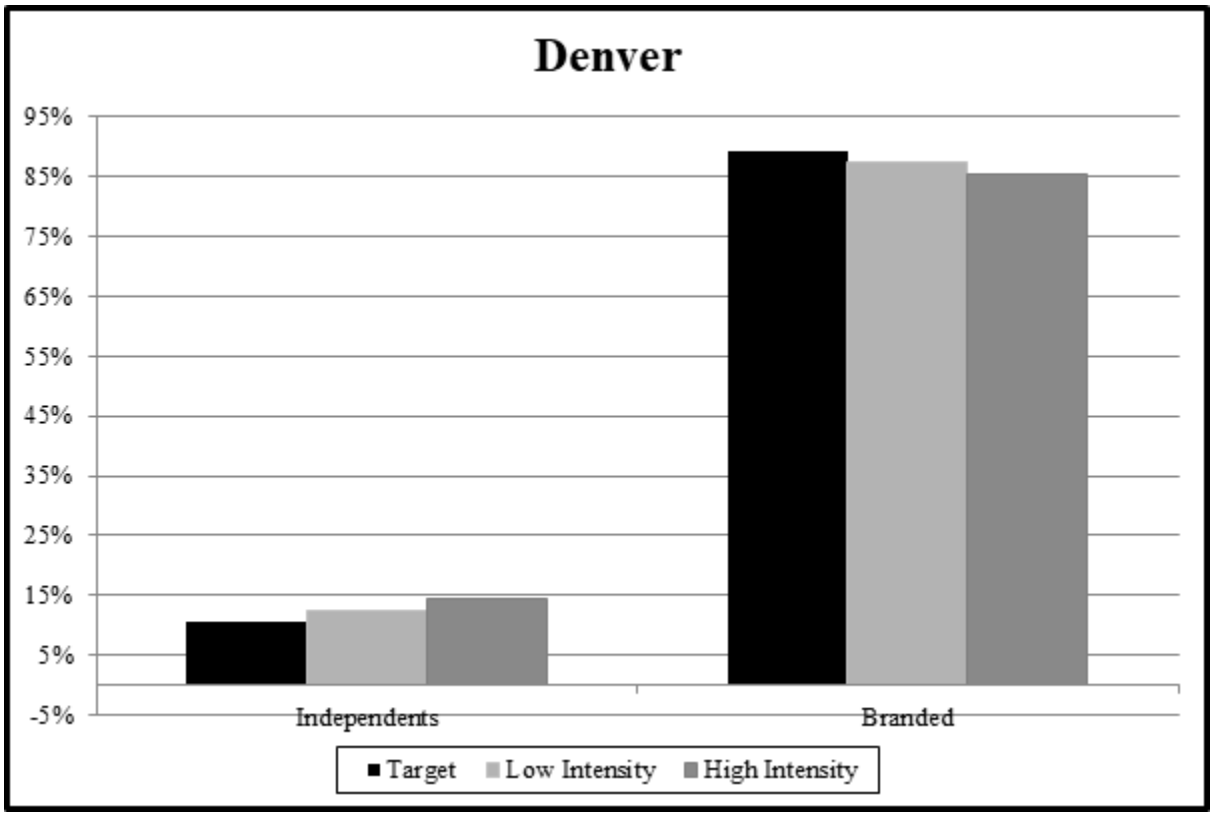
Appendix B

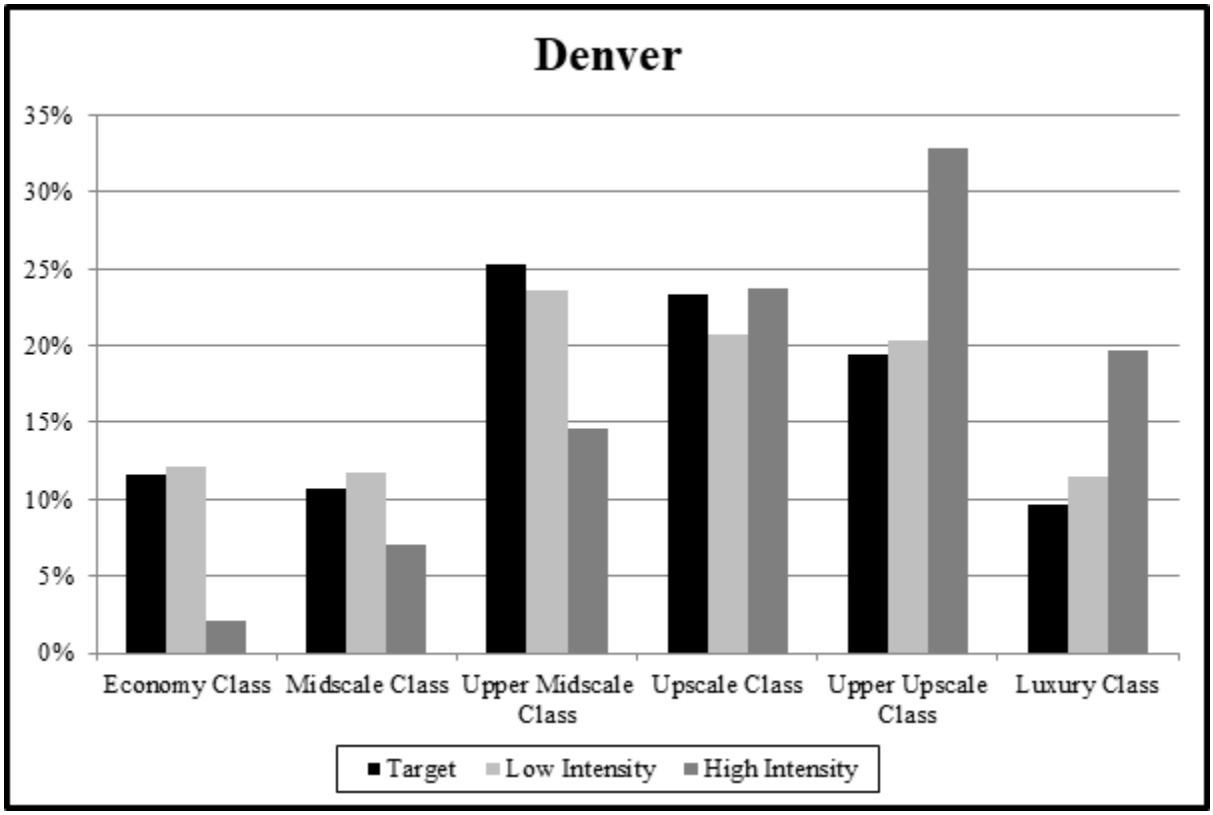

Dallas
Trip Advisor Ratio
0
1
2
3
4
5
6
0
5
10
15
20
25
30
Distance
Low Intensity
High Intensity

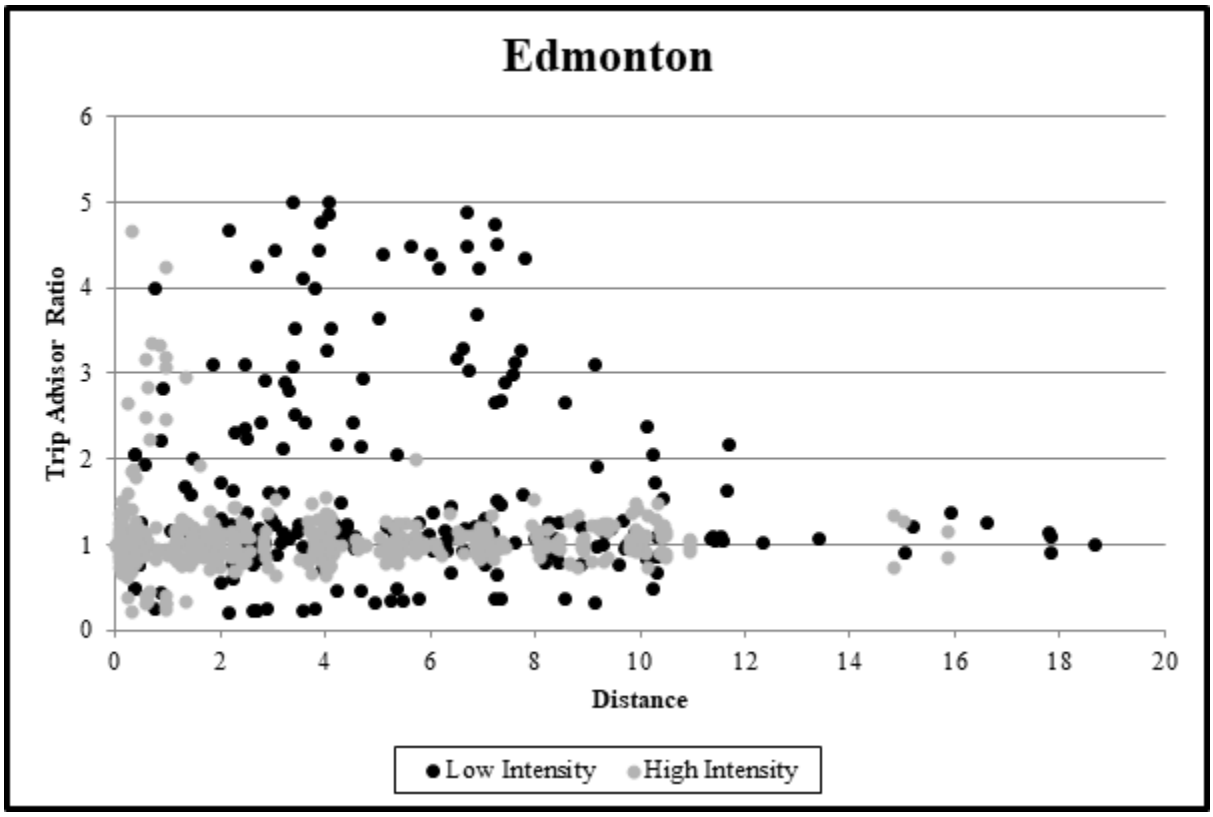

Dallas
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Independents
Branded
Target
Low Intensity
High Intensity

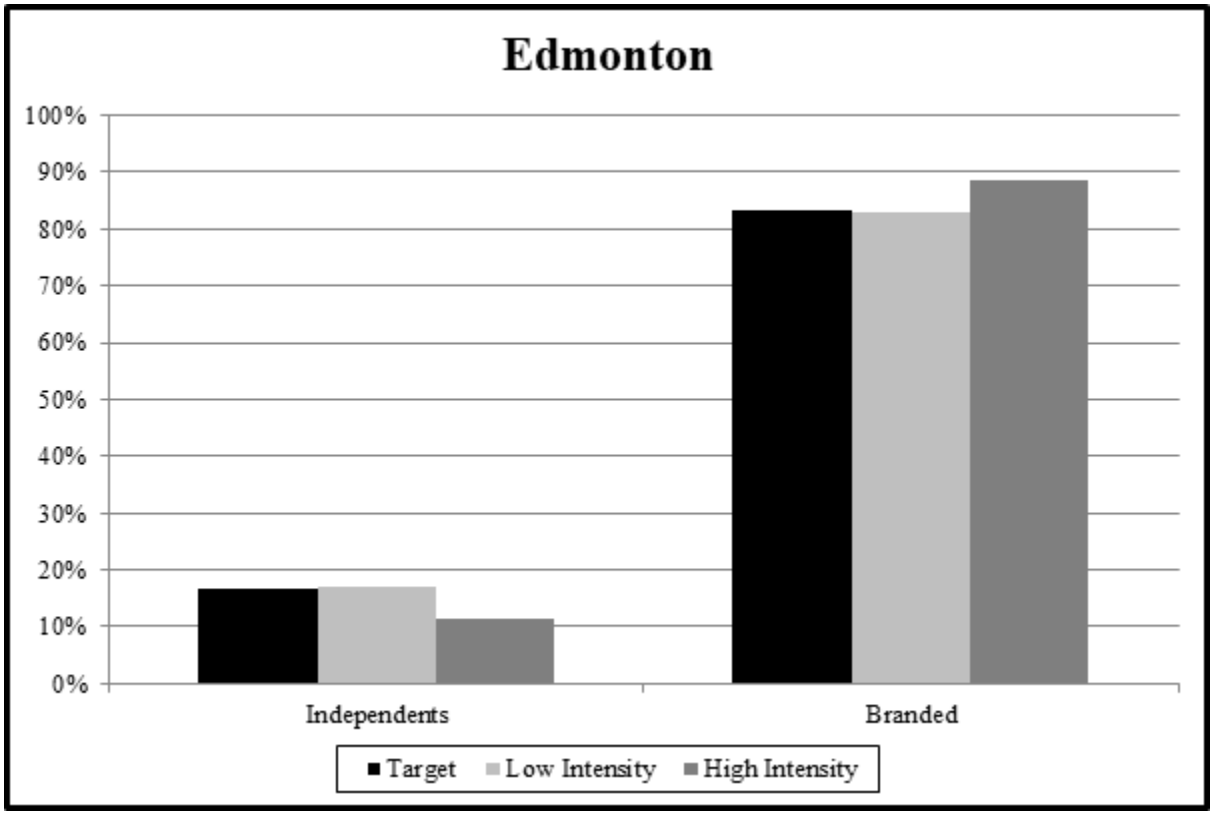

Dallas
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Economy Class
Midscale Class
Upper Midscale Class
Upscale Class
Upper Upscale Class
Luxury Class
Target
Low Intensity
High Intensity

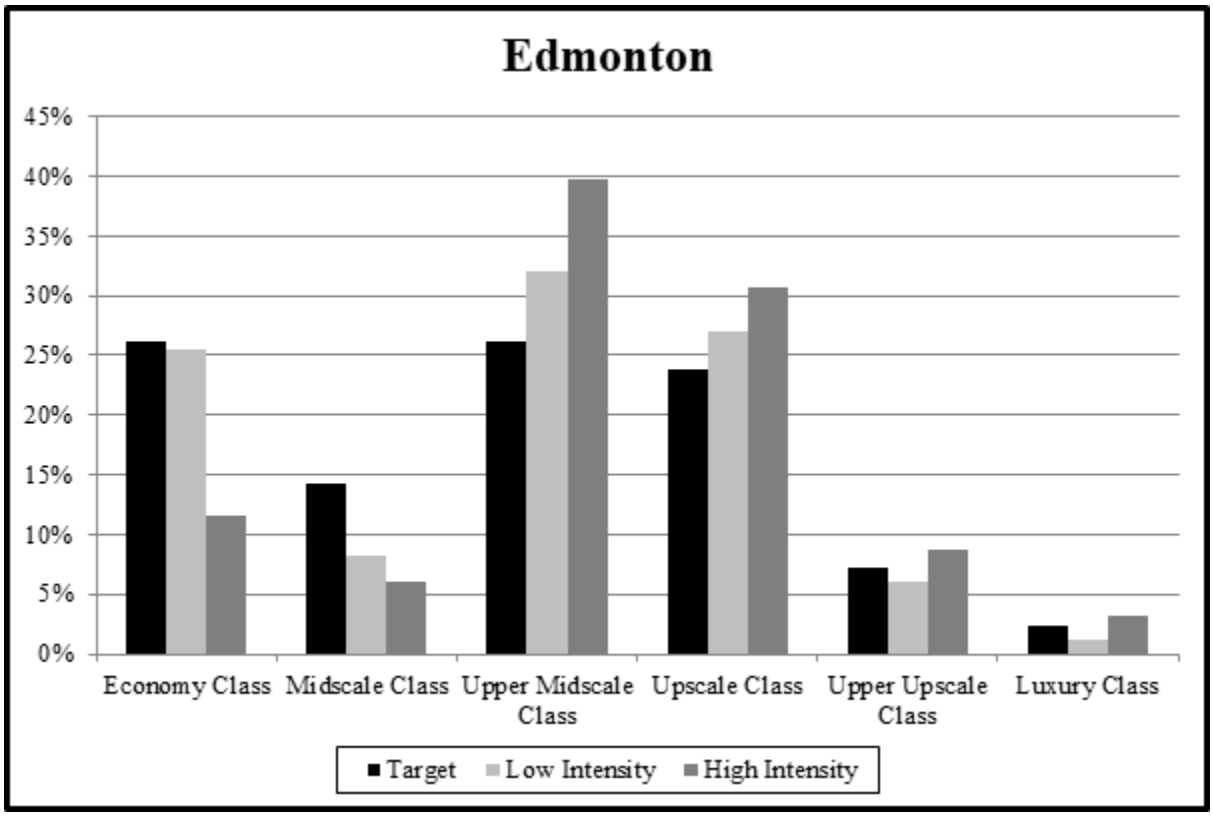

Denver
Trip Advisor Ratio
6
5
4
3
2
1
0
0
5
10
15
20
25
Distance
Low Intensity
High Intensity

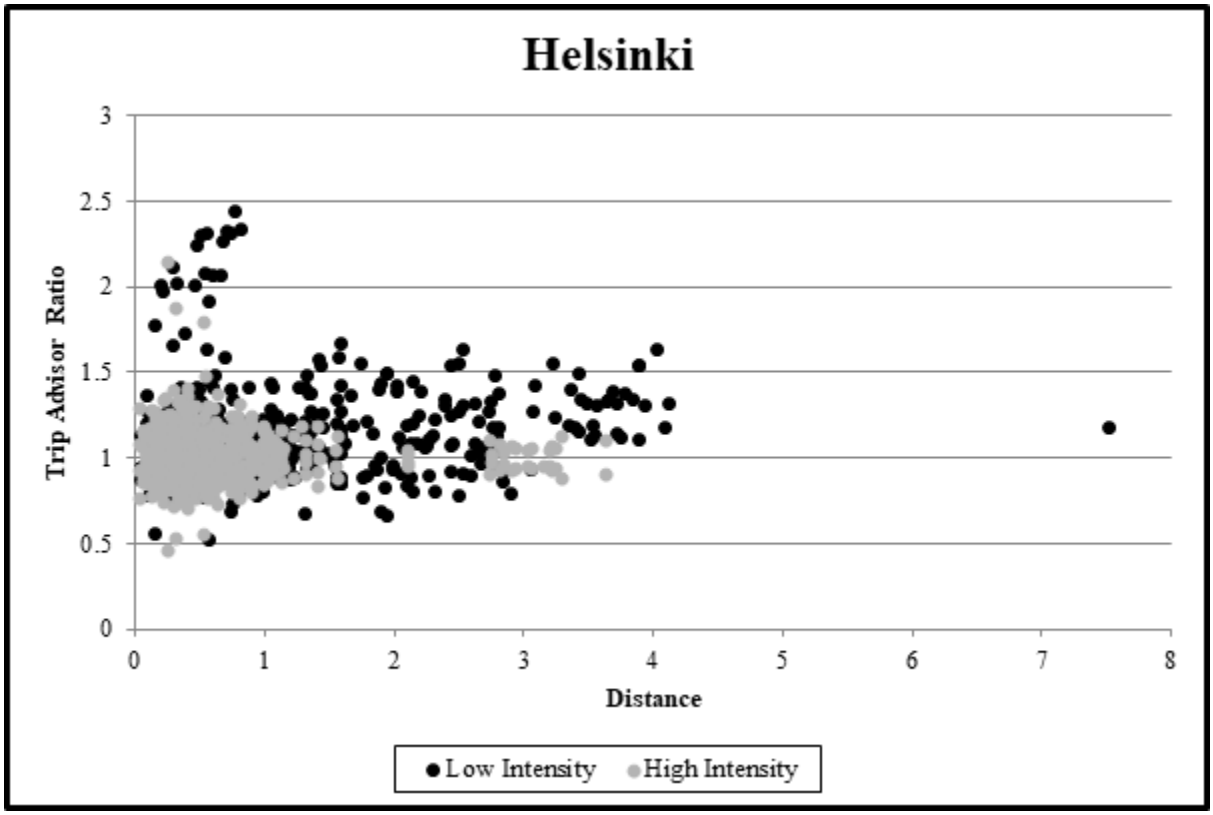

Denver
95%
85%
75%
65%
55%
45%
35%
25%
15%
5%
-5%
Independents
Branded
Target
Low Intensity
High Intensity

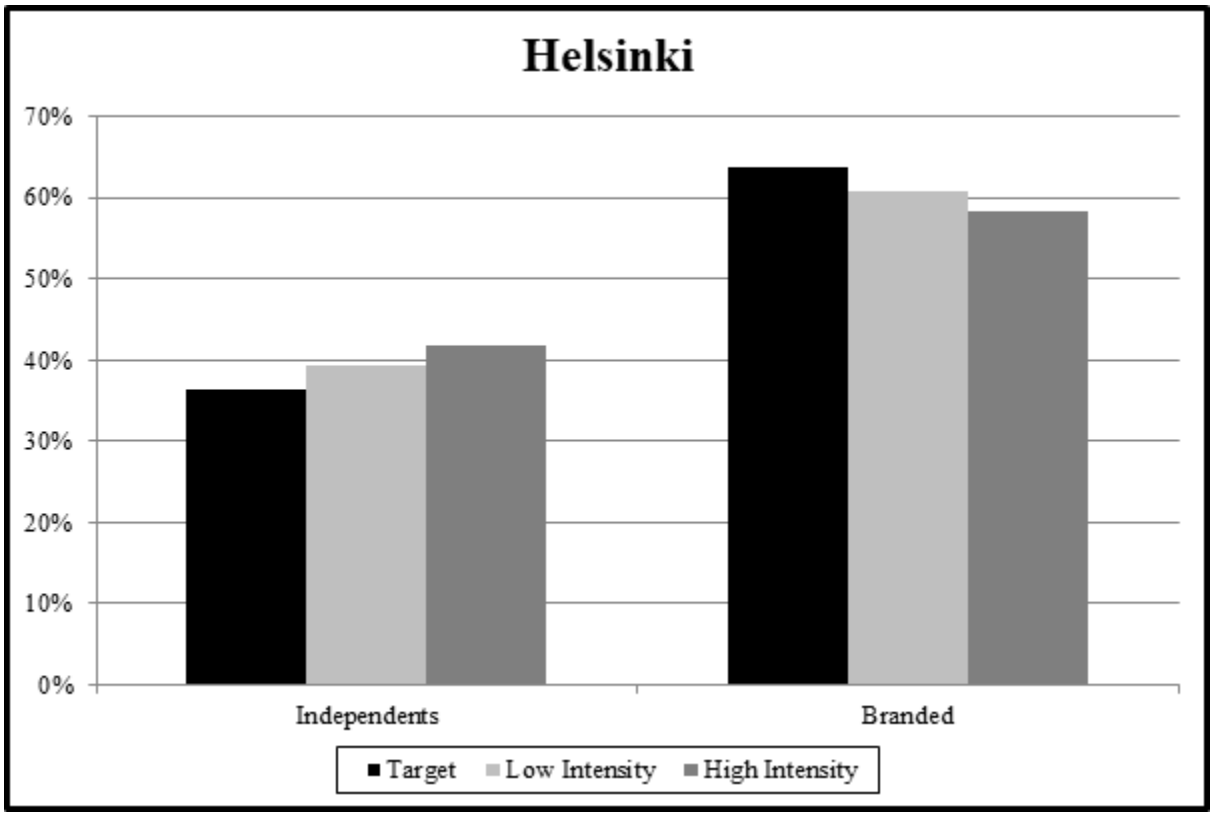

Denver
35%
30%
25%
20%
15%
10%
5%
0%
Economy Class
Midscale Class
Upper Midscale Class
Upscale Class
Upper Upscale Class
Luxury Class
Target
Low Intensity
High Intensity

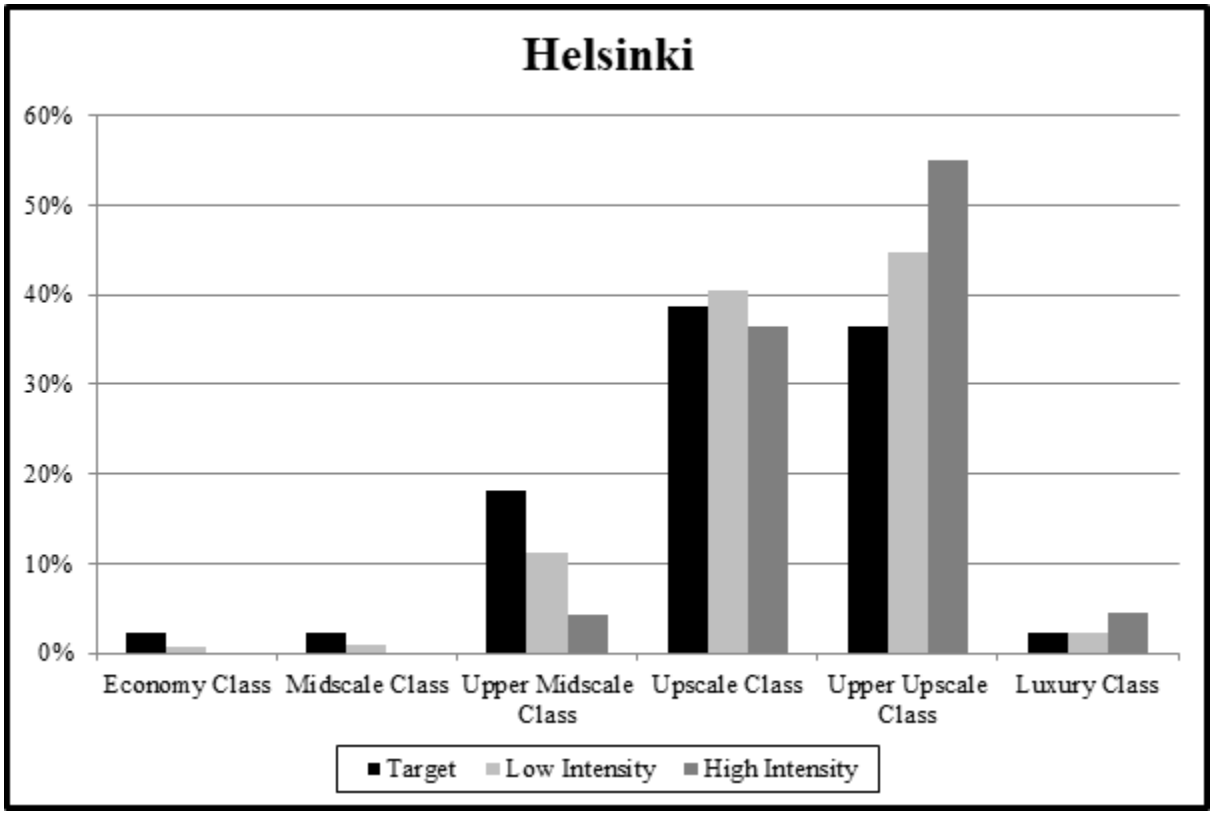

Edmonton
Trip Advisor Ratio
0
1
2
3
4
5
6
0
2
4
6
8
10
12
14
16
18
20
Distance
Low Intensity
High Intensity

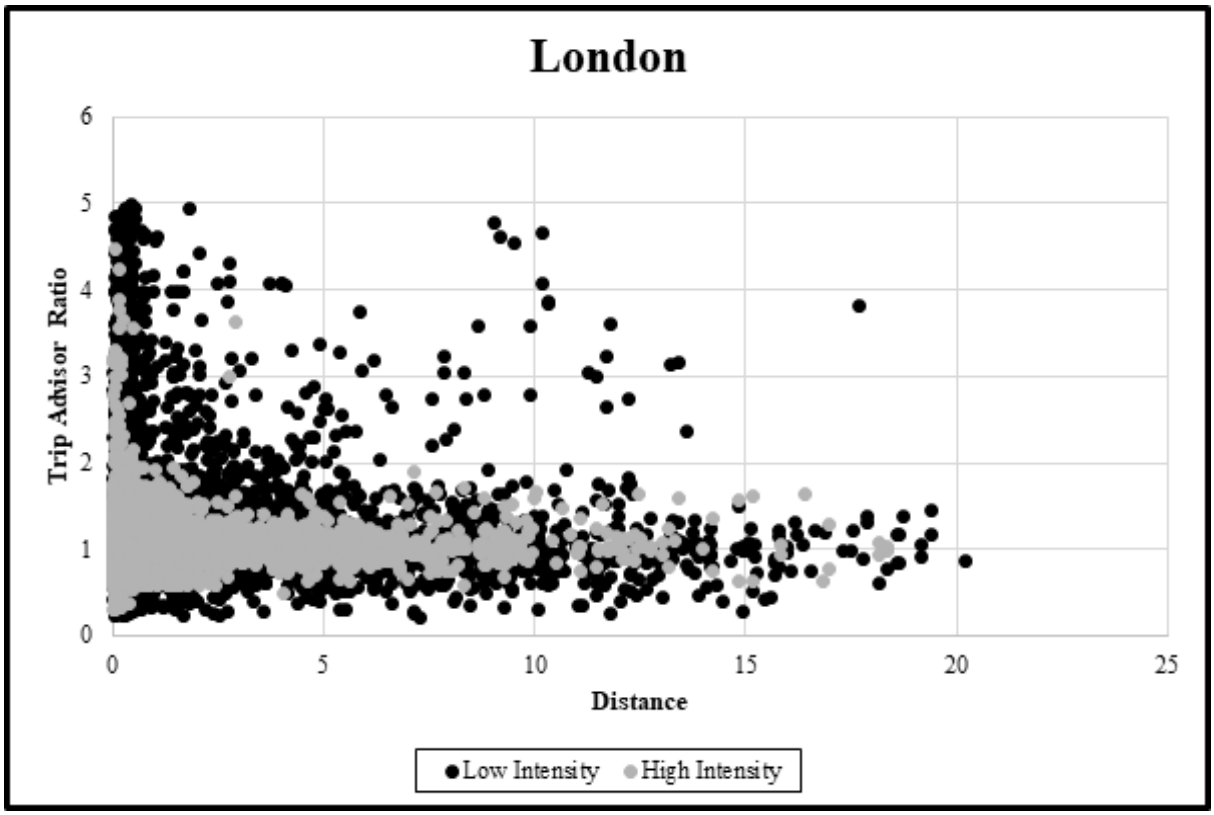

Edmonton
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Independents
Branded
Target
Low Intensity
High Intensity

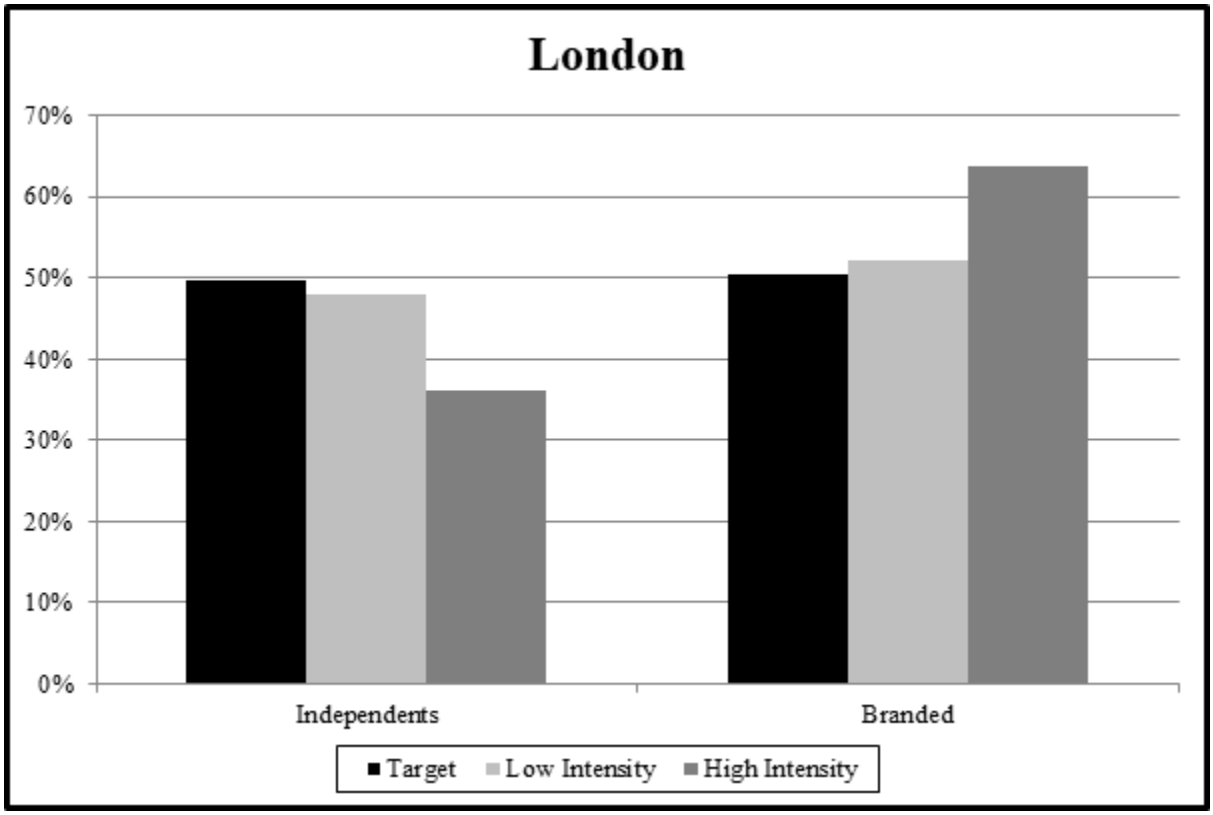

Edmonton
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Economy Class
Midscale Class
Upper Midscale Class
Upscale Class
Upper Upscale Class
Luxury Class
Target
Low Intensity
High Intensity

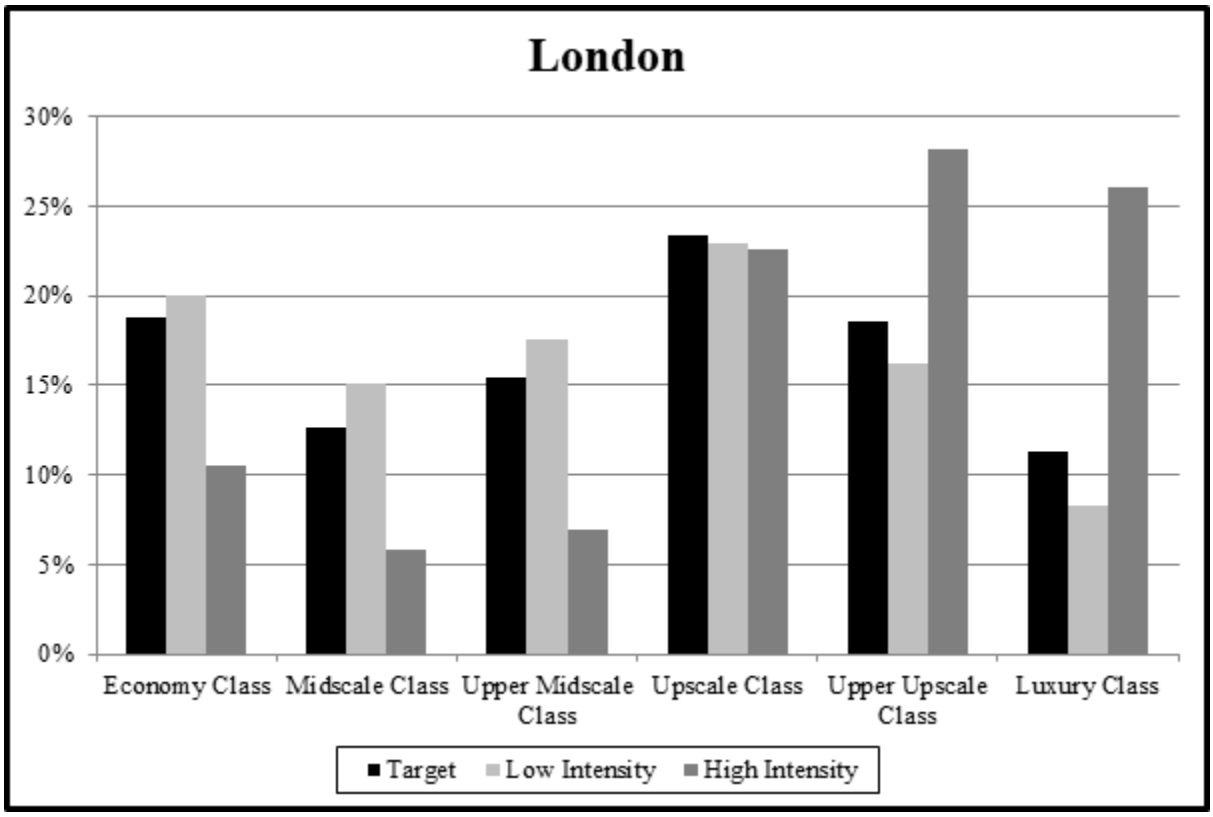

Helsinki
Trip Advisor Ratio
3
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6
7
8
Distance
Low Intensity
High Intensity

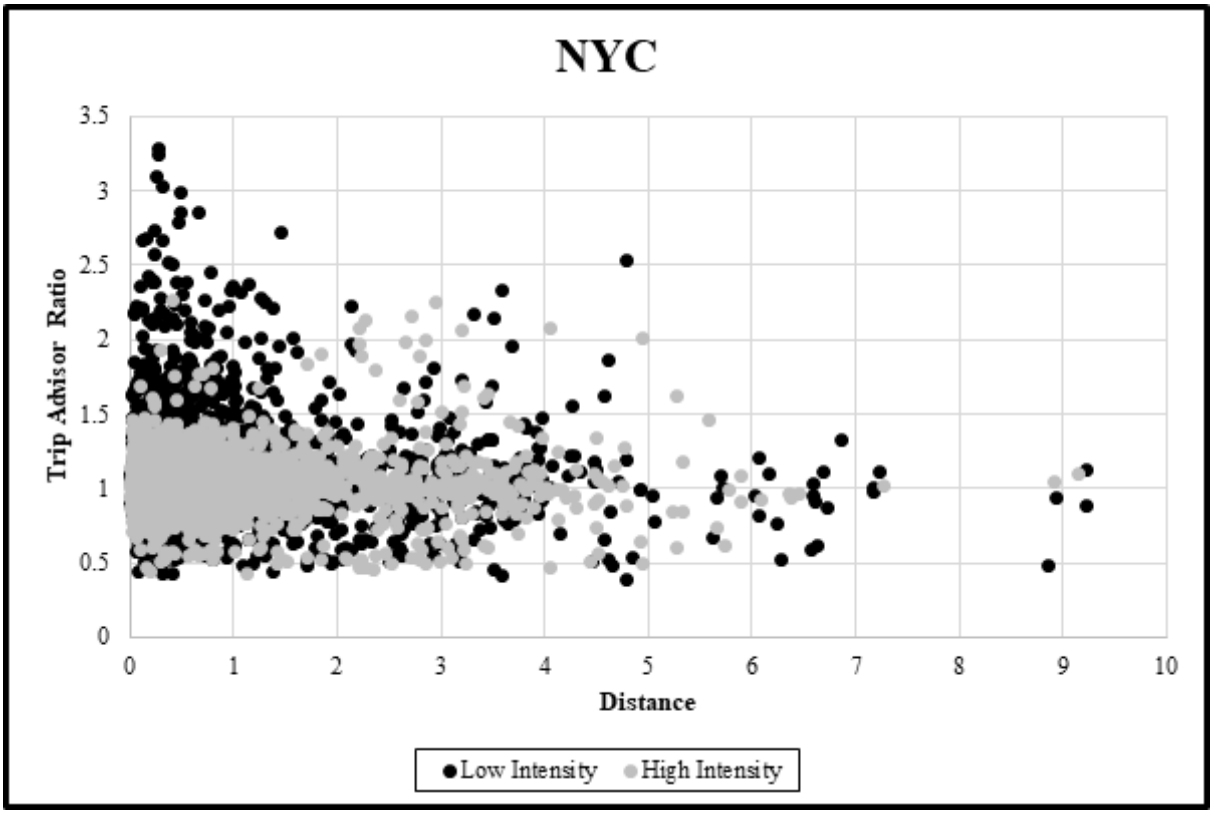

Helsinki
70%
60%
50%
40%
30%
20%
10%
0%
Independents
Branded
Target
Low Intensity
High Intensity


Helsinki
60%
50%
40%
30%
20%
10%
0%
Economy Class
Midscale Class
Upper Midscale Class
Upscale Class
Upper Upscale Class
Luxury Class
Target
Low Intensity
High Intensity


London
Trip Advisor Ratio
0
1
2
3
4
5
6
0
5
10
15
20
25
Distance
Low Intensity
High Intensity

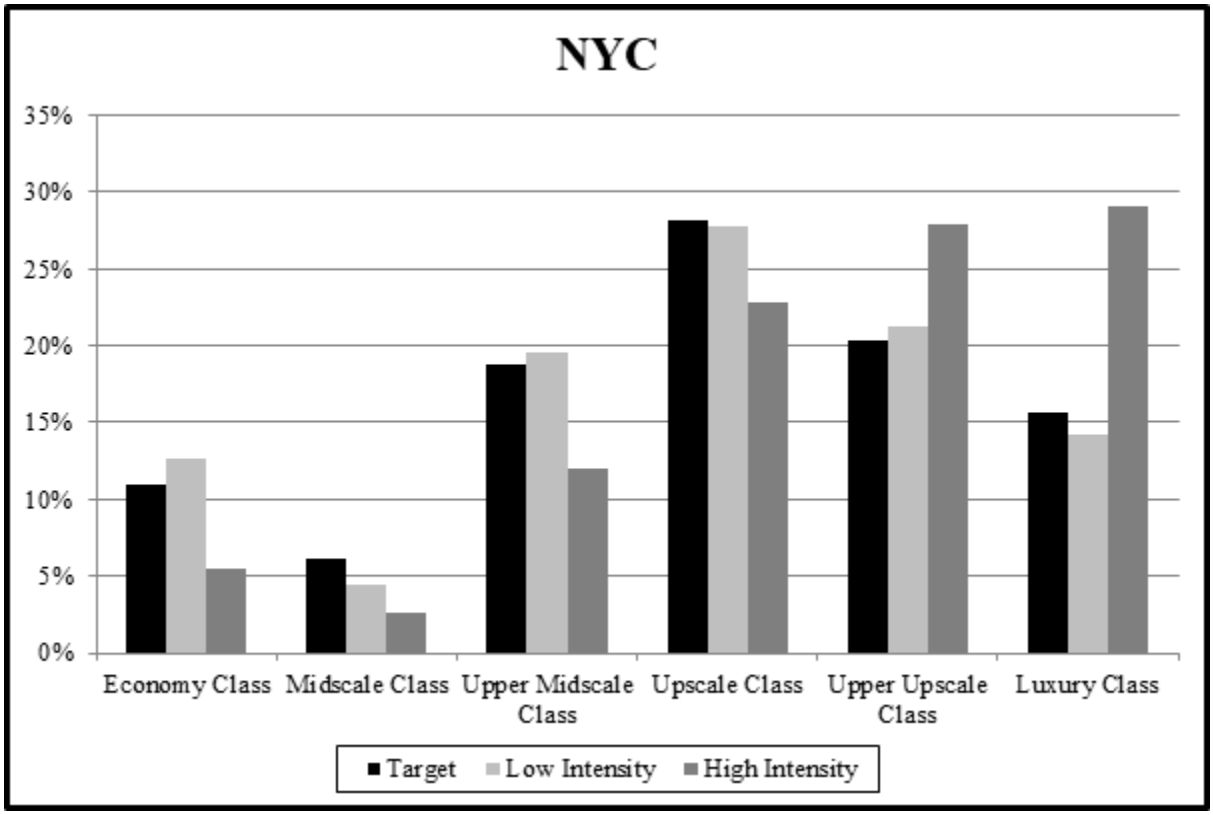

London
70%
60%
50%
40%
30%
20%
10%
0%
Independents
Branded
Target
Low Intensity
High Intensity

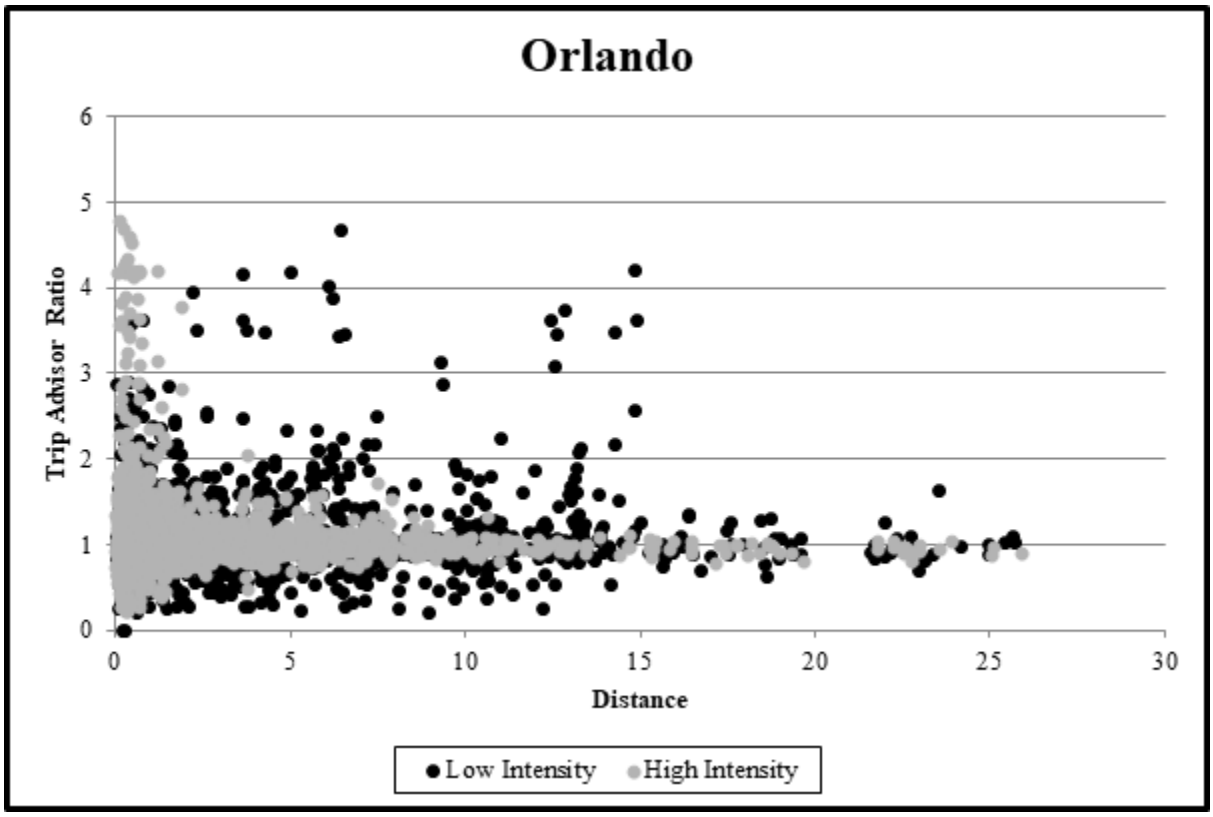

London
30%
25%
20%
15%
10%
5%
0%
Economy Class
Midscale Class
Upper Midscale Class
Upscale Class
Upper Upscale Class
Luxury Class
Target
Low Intensity
High Intensity

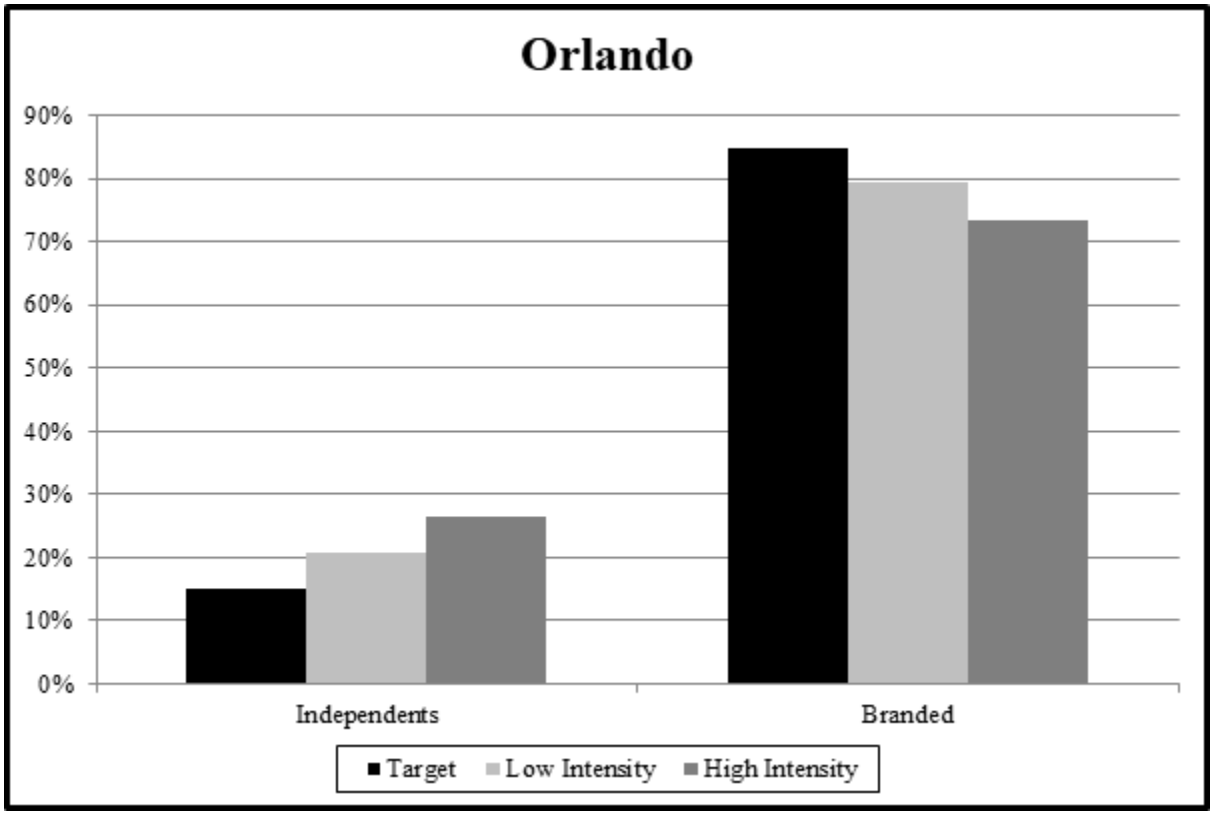

NYC
Trip Advisor Ratio
3.5
3
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6
7
8
9
10
Distance
Low Intensity
High Intensity

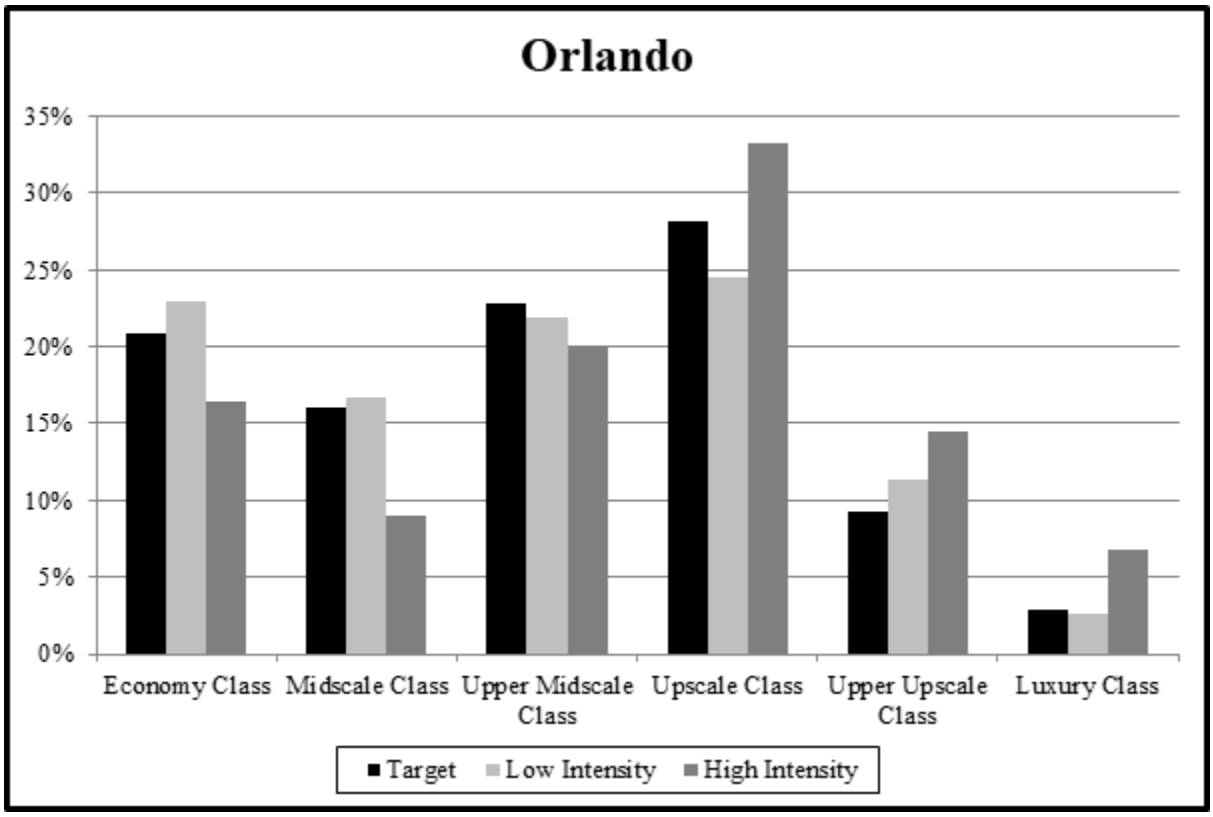
Appendix J

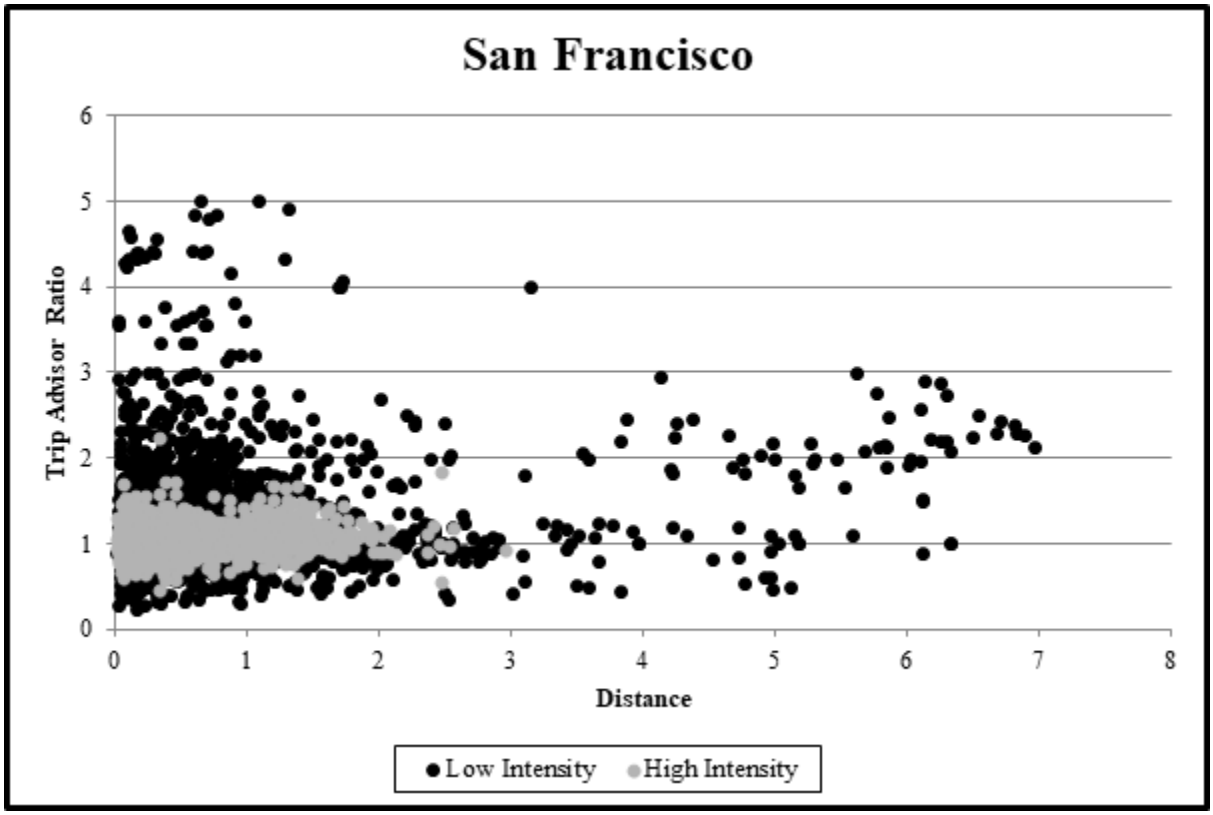

NYC
53%
52%
51%
50%
49%
48%
47%
46%
Independents
Branded
Target
Low Intensity
High Intensity

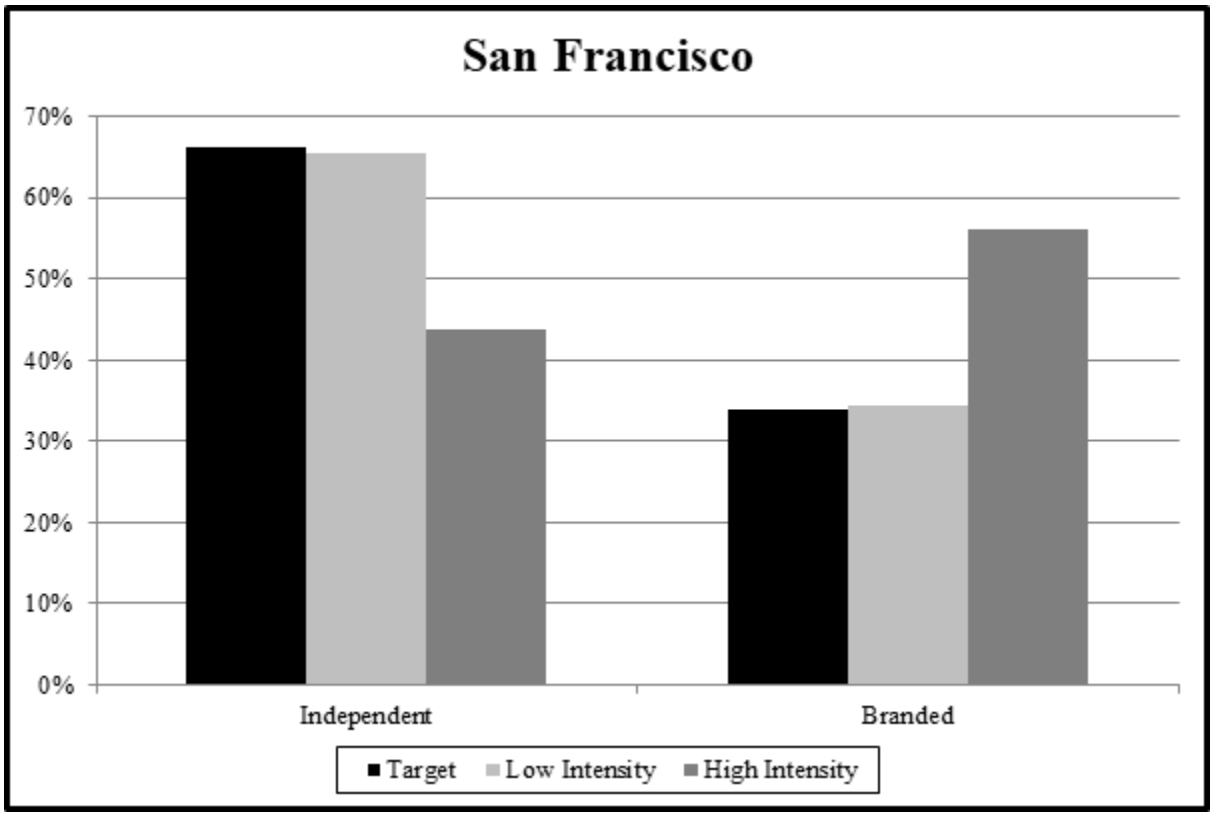

NYC
35%
30%
25%
20%
15%
10%
5%
0%
Economy Class
Midscale Class
Upper Midscale Class
Upscale Class
Upper Upscale Class
Luxury Class
Target
Low Intensity
High Intensity

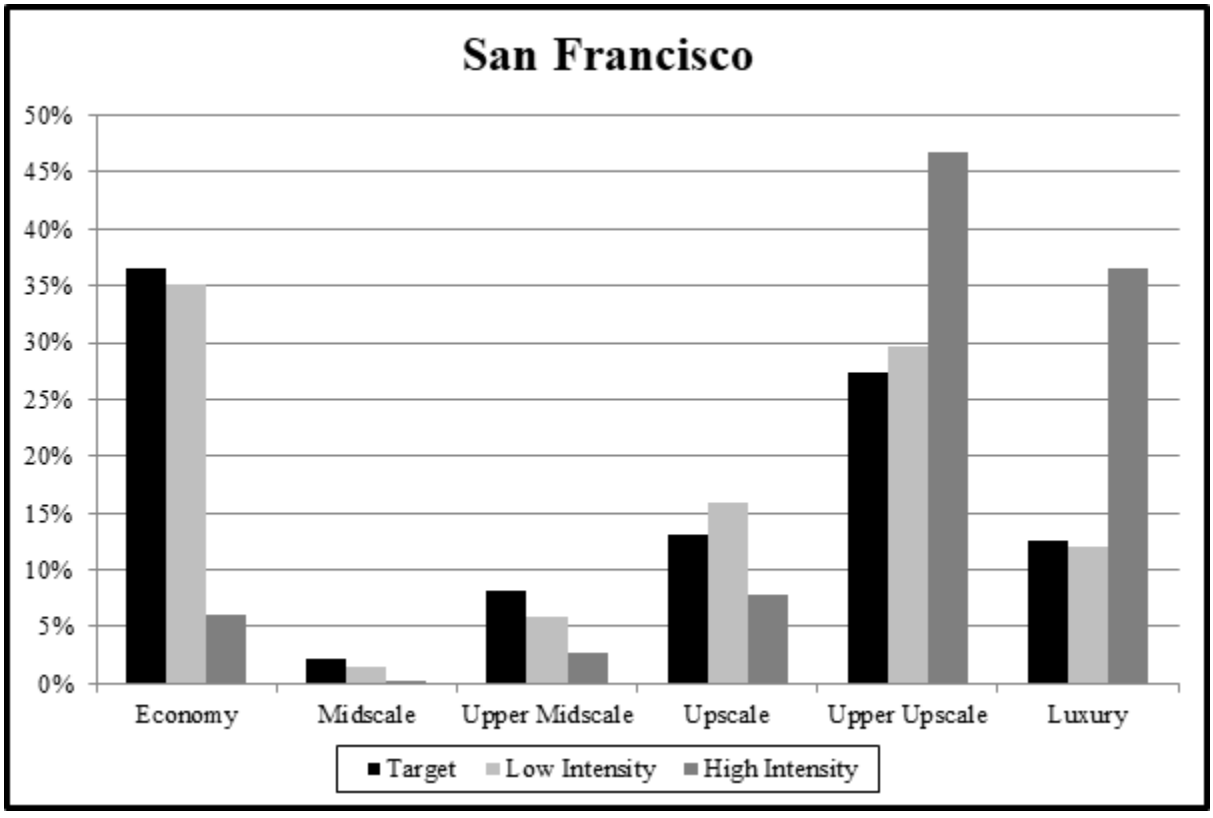

Orlando
Trip Advisor Ratio
0
1
2
3
4
5
6
0
5
10
15
20
25
30
Distance
Low Intensity
High Intensity

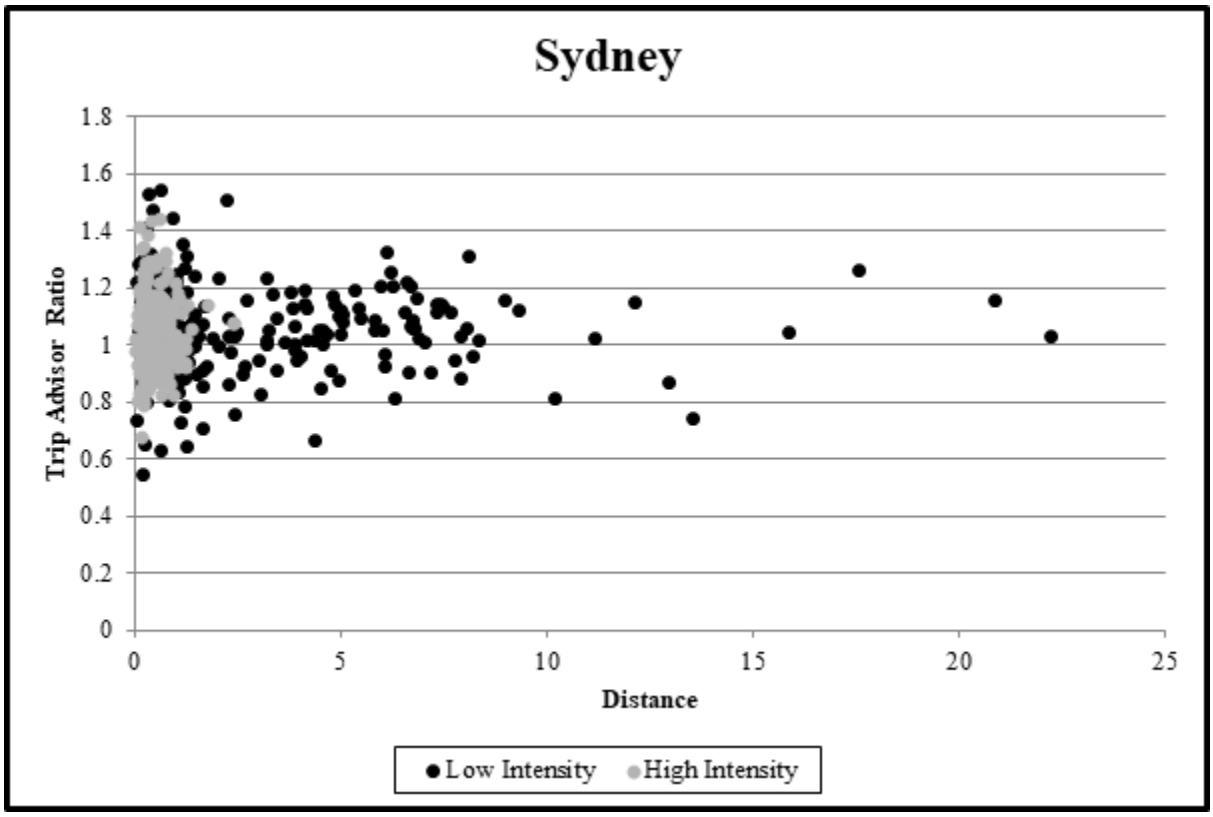

Orlando
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Independents
Branded
Target
Low Intensity
High Intensity

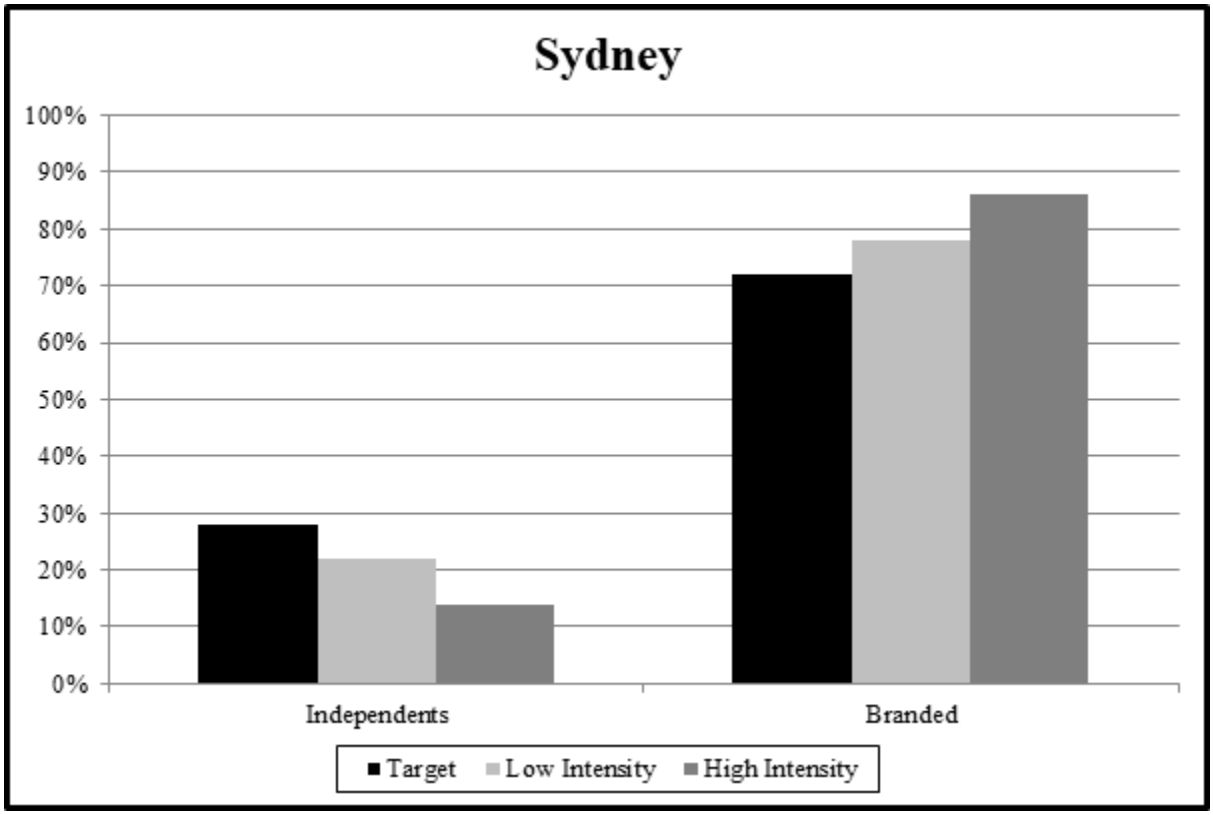

Orlando
35%
30%
25%
20%
15%
10%
5%
0%
Economy Class
Midscale Class
Upper Midscale Class
Upscale Class
Upper Upscale Class
Luxury Class
Target
Low Intensity
High Intensity

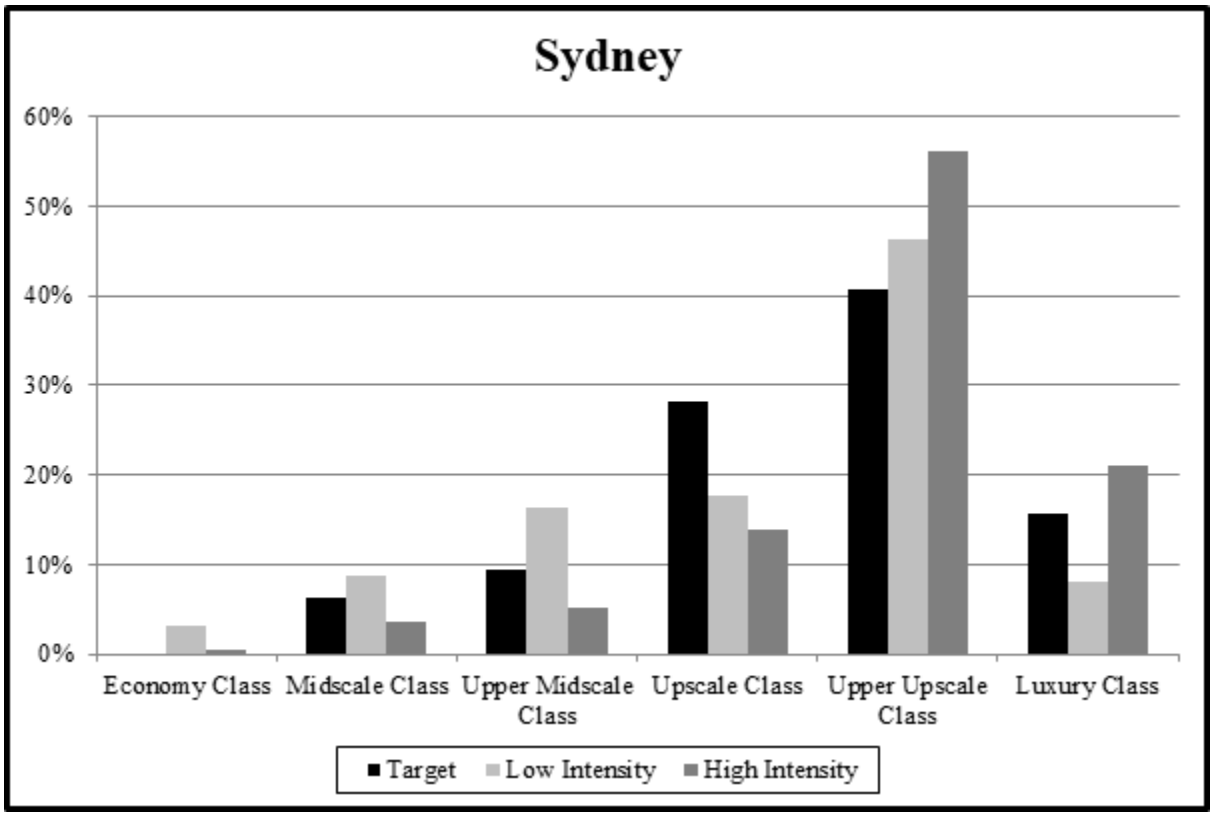

San Francisco
Trip Advisor Ratio
Distance
0
1
2
3
4
5
6
7
8
Low Intensity
High Intensity

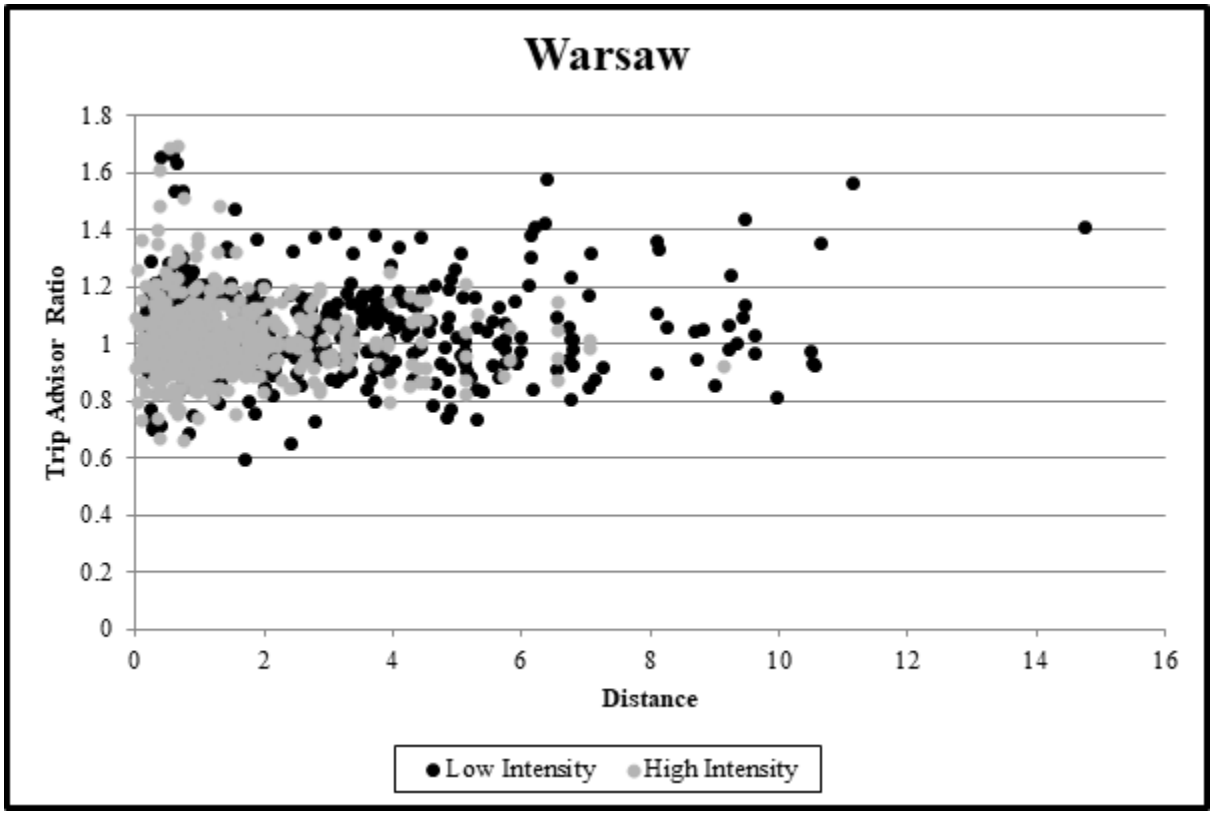

San Francisco
70%
60%
50%
40%
30%
20%
10%
0%
Independent
Branded
Target
Low Intensity
High Intensity

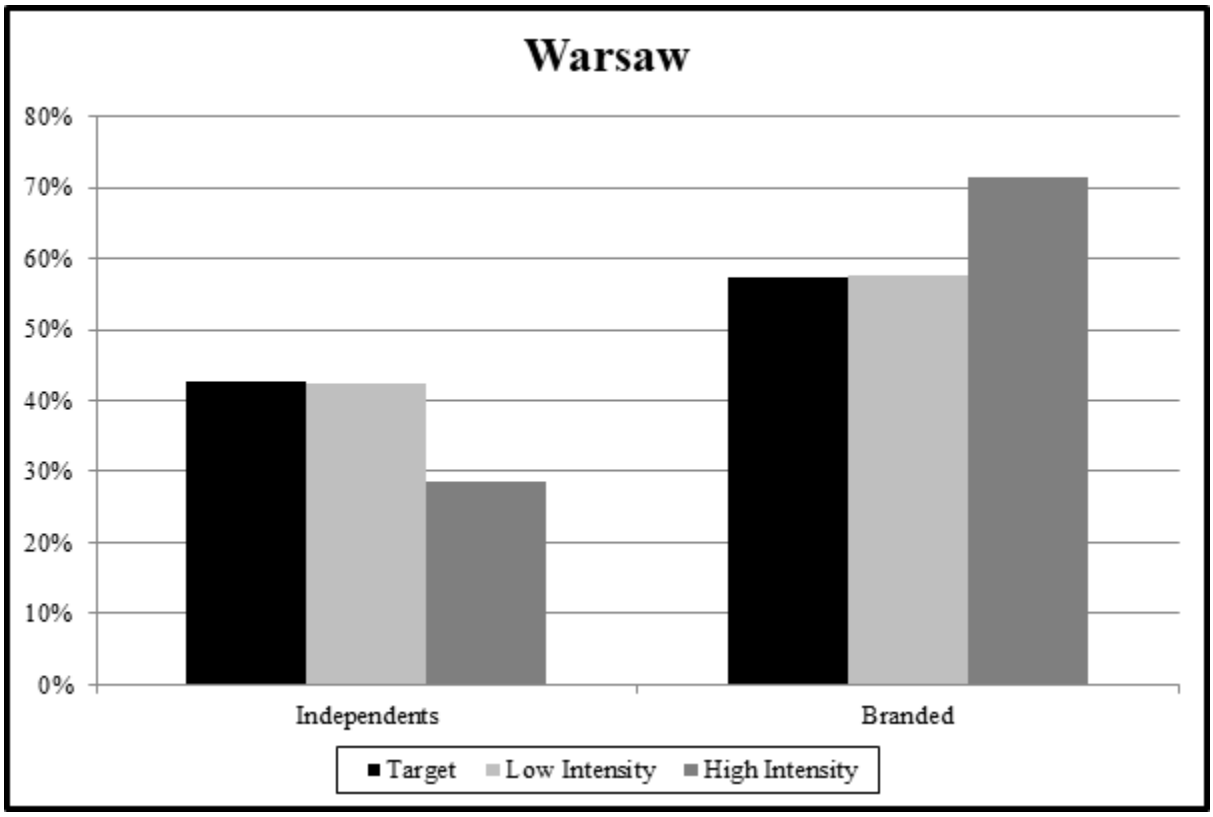

San Francisco
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Economy
Midscale
Upper Midscale
Upscale
Upper Upscale
Luxury
Target
Low Intensity
High Intensity

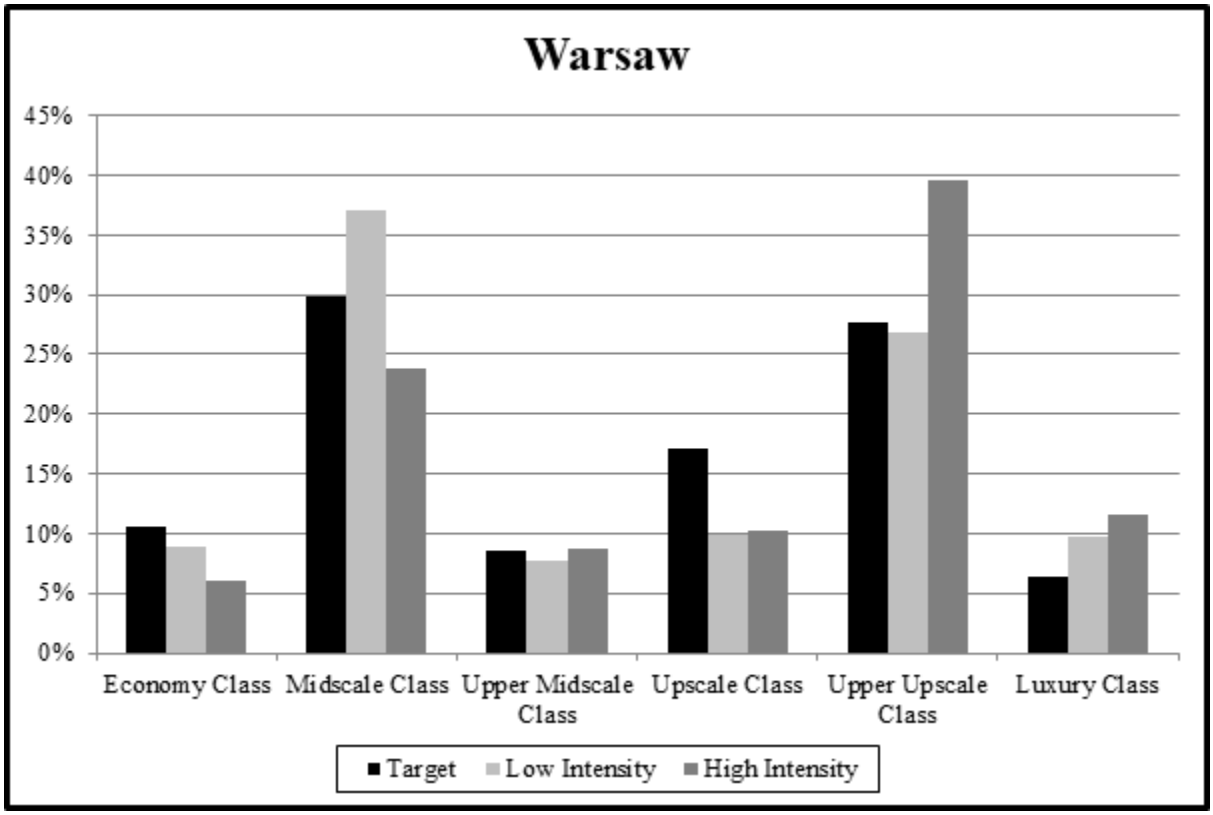

Sydney
Trip Advisor Ratio
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
5
10
15
20
25
Distance
Low Intensity
High Intensity


Sydney
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Independents
Branded
Target
Low Intensity
High Intensity


Sydney
60%
50%
40%
30%
20%
10%
0%
Economy Class
Midscale Class
Upper Midscale Class
Upscale Class
Upper Upscale Class
Luxury Class
Target
Low Intensity
High Intensity


Warsaw
Trip Advisor Ratio
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10
12
14
16
Distance
Low Intensity
High Intensity


Warsaw
80%
70%
60%
50%
40%
30%
20%
10%
0%
Independents
Branded
Target
Low Intensity
High Intensity


Warsaw
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Economy Class
Midscale Class
Upper Midscale Class
Upscale Class
Upper Upscale Class
Luxury Class
Target
Low Intensity
High Intensity

Appendix Q