Final Paper: Predicting the Establishment of Hotel Hubs in New York City

Jacqueline Cafasso, Dahlia Darwiche, Marium Sultan,
Timothy Hambridge, and Haoran Huang

Abstract

This study aims to predict the number of hotel hubs in a given zip code in New York City. Drawing on data from PLUTO and the American Community Survey (ACS), we use demographic and land-use variables to predict zip codes that have an above average number of hotels. Using a random sample of 30,000 observations from our overarching population, which contains over 30,000,000 records from 2011-2016, our analysis reveals that the mean score for hotel hub for the random sample was .155, suggesting that roughly 85% of the zip codes analyzed had a below average number of hotels. It follows that zip codes that have a below-average number of hotels, an abnormally high population, median household income, or assessed value may be prime spots to for future hotel establishments. Moreover, these initial results will serve as an enrichment tool for understanding the relationships between hotels and urban structure while also helping practitioners to identify tourism functional zones or potential sites for the establishment of new hotels.

Introduction

Over the past decade, urban tourism researchers have shifted their attention to focus on hotel location analysis and the identification of contributing factors to 'tourism functional zones.' Success, in the hotel industry, relies heavily on location factors. An ideal location (e.g. areas with a high concentration of restaurants) is associated with a number of payoffs: larger accommodation demand, higher revenue per available room, higher customer satisfaction, better performance, and lower failure rates (Yang et al, 2012, 2014). Since a hotels' location is a long term fixed investment, it is important for companies to invest in a 'ideal location.' A flawed

location may prove to be destructive to the financial health of an establishment and can be one of the main reasons for why a hotel closes up shop.

Our analysis seeks to identify the key location characteristics and patterns of hotel establishments in New York City in hopes that our study will provide policymakers and industry practitioners with an understanding of what features make a given location (i.e. zip code) more or less attractive to a consumer. Past researchers have examined patterns of hotel establishments in tourism sectors of cities and have taken into account factors that may contribute to hotel location choice (Li et al, 2015; Yang et al. 2012; Ashworth and Tunbridge, 1990; Pearce, 1987, 1995). Their findings revealed clustering and linear patterns, based on the following factors: accessibility, comparative shopping, land rent, planning restrictions, and proximity to other urban tourism phenomenons.

The aim of our study is to investigate the spatial relationships between hotel distribution, land types, and other surrounding factors in NYC. As a result, we will (1) analyze features of hotel establishment location and (2) identify how these features affect the number of establishments in a given zip code (i.e. hotel hub). Using NYC's extensive land use and geographic tool, PLUTO, and demographic data from the American Community Survey (ACS) we will study the relationship between hotels and neighborhood features from 2011 to 2016 using basic regression models. Our analysis will contribute to hotel location literature, aid in the understanding of the makeup of urban tourism space and structure, educate governments and authorities to understand the geography, contribute to industrial policies for urban tourism development, and will provide vital information to urban and regional planning efforts in New York City.

Literature Review

With more than 61 million tourists expected to visit New York City this year alone, finding an ideal location to build a hotel is a main priority for many businesses (Plitt, 2017). Although research on tourism and hotel establishment dates back to before the 1980s, there are no

meaningful studies that link success to neighborhood features (i.e. land use, comparative shopping, etc.). Research on the field of urban tourism did not evolve until after 1989 when GJ Ashworth pointed out the importance of tourism to a city, "… the failure to consider tourism as a specifically urban activity imposes a serious constraint that cannot fail to impede the development of tourism as a subject of serious study" (Ashworth 1989:33). In his work, Ashworth identified four approaches to analyzing urban tourism: (1) the facility approach (e.g. spatial analysis of location features of tourism; attractions, facilities, infrastructure, and zones); (2) the ecological approach (e.g. analysis of structures in urban areas and the identification of functional zones); (3) the user approach (e.g. marketing perspective); and (4) the policy approach (e.g. policy-based perspective; infrastructure provision, destination marketing) (1989).

Understanding the spatial characteristics of tourism in major cities is a vital component for future planning and designs as tourism "occupies substantial amounts of space within urban destinations via tourist-historic urban cores, museums of all kinds, urban waterfronts, theme parks, and specialized precincts" (Edwards et al, 2008). As noted by previous researchers, the first step to any successful hotel business plan should be to identify the determinants of location selection (Issahaku and Francis, 2013). However, our current research is lacking any interpretation of how tourism in an urban area is related to surrounding spatial elements (i.e. land use types). Additionally, few studies have gone in depth to examine the variations in the spatial relationships between hotel location and its surrounding environment. So far, studies have gone as far as to establish that distinct distribution patterns exist and are exasperated by factors, such as accessibility, land rent, the regency effect, planning restrictions, comparative shopping, and proximity to other tourism-related phenomena (Ashworth and Tunbridge, 1990; Pearce, 1987, 1995).

Related Previous Work

Hotel attributes are important when it comes to the selection of hotels. Among them one of the most crucial criteria is the location of hotel. As the father of modern hotel industry, Ellsworth

Statler once said “There are three things that make a hotel famous – location, location, location.” In Sara and Thomas (2003) reviews on previous studies regarding hotel attributes in the past two decades, among 21 studies, 18 of them looked into hotel locations as an significant factor of their studies, which again yields the importance of hotel locations.

By tracking 557 tourists movements from four different hotels in HongKong, Bob et al. (2011) implied that hotel locations have profound impacts toward subsequent tourism movement and there needs on at least four aspects: spatially concentrated activity around the hotel; places tourists are likely or unlikely to visit; volume of visitors at all but icon attractions and; diurnal visitation patterns. Lew and McKercher (2004) have suggested that urban tourist flows clearly have a tendency to spread themselves unevenly, both spatially and temporally. While popular destinations are suffered from overcrowding tourists, other facilities are under-utilized. The disproportional distribution of tourists lead to the low-efficient use of economic and social resources, which is unsustainable towards future development of the city. (Noam, Bob, Erica and Emit, 2011).

The implications of hotel locations on tourist behaviour could significantly affect local tourism and business development, especially small business, also bring reflections on social and economic resources allocations. By looking into 89 companies from the most prominent tourist destination in Macedonia, the city of Ohrid, using a linear regression model, Ljubomir and Aleksandra(2017) came to the conclusion that by being in the tourist zone, a small business could gain 0.434 growth based on a growth measurement scale from 1-5, holding all other controls including external constrains, internal constraints, owners’ educational level, the size and age of the business constant.

The locations of the hotel establishments have been a driven factor for tourists movement and tourism resources distribution, which is beneficial to surrounding local business development. However. No previous study has ever looked into developing a model of predicting the possible site selections for future hotels. Thus by building a machine learning model which is able to

predict future hotel locations can shine a light on various factors of social and economic development, and as a powerful tool for policy makers to generate related regulations to promote and guide local tourism, small business and even infrastructure construction programs.

Data Description and Methodology

We based our project on two datasets: New York City Department of City Planning's PLUTO dataset and the American Community Survey (ACS). PLUTO was valuable in providing extensive land use and geographic data at the tax lot level. This was essential when collecting data on hotel counts, facility types, and zoning information. On the other hand, ACS was useful when gathering socioeconomic information on our population.

We began by downloading PLUTO data between the years of 2011 and 2016. This data was organized in separate files for each borough for each year. We then went through and standardized our data. This involved calling up relevant column names manually through terminal to import and concatenate the files into a singular dataframe using a loop. Next, we cleaned the data, removing rows with nans in the zipcode column. We chose zip code as our unit of analysis because of its granularity and the fact that each zipcode is a unique value, unlike census tract which has repeating codes across different counties.

We downloaded ACS data for every zip code in the state of New York from the years 2011-2016 using the ACS API and our API key. Our ACS data included Population and Median Household Income (in inflation adjusted dollars) by zip code.

Next, we created binary variables from the PLUTO data. For this step we referred to the PLUTO data dictionary to understand Land Use (LandUse) and and Building Class (BldgClass) codes. For examples, hotel buildings were covered by BldgClass codes starting with K, so we created a new column called Hotel_Building and if the BldgClass code in that row began with a K, this column would get a 1, indicating the building on the plot was a hotel, if not, the column would

get a 0. We did the same for shopping centers, to check if areas with an above average amount of shopping centers correlated with areas with an above average amount of hotels. We used the LandUse column to create new binary columns for if the plot was located in a residential zone, commercial zone or manufacturing zone.

Zipcode-year is an innovative category we invented to be able to run our analysis and merges on a value that does not repeat across years. It is simply a concatenation of zip code and year. For example, the zip code 11101 shows up in reference to data from 2011 and from 2016 (and from all the years in between). If we wanted to get hotel building counts for 2011 and 2016 separately we would need to have a value that regards 11101 for 2011 and 11101 for 2016 differently. The zipcode year would read 11101_2011 or 11101_2016.

We counted up the amount of hotel buildings, shopping center buildings, residential zones, commercial zones, and manufacturing zones in each zipcode-year. We took our counts and ran them through a function we created to check if each zipcode-year had an above or below average amount of each of these variables. If the count of hotels in a certain zipcode-year was above the mean of hotel counts per zipcode-year then the hotel_hub column would have a 1, indicating the zipcode was a hotel_hub, otherwise it would have a 0.

After creating these additional variables we merged the ACS and PLUTO data on zip-year. Here we also created an average assessed total value column, using each plot's assessed total averaged out per zipyear. At this point the size of the dataframe was 30543750 by 21. We used a sample of 30000 entries from this dataframe for one of our machine learning models.

Lastly, we created a dataframe that included only one row per zip-year, using only the columns that had aggregate data, such as binary hubs. This left out values individual to any one plot. Due to server speed constraints this dataframe was created out of a random sample of 30000 entries rather than the full amount of records, which turned out to take a long time to load in. This new dataframe was 1024 by 17.

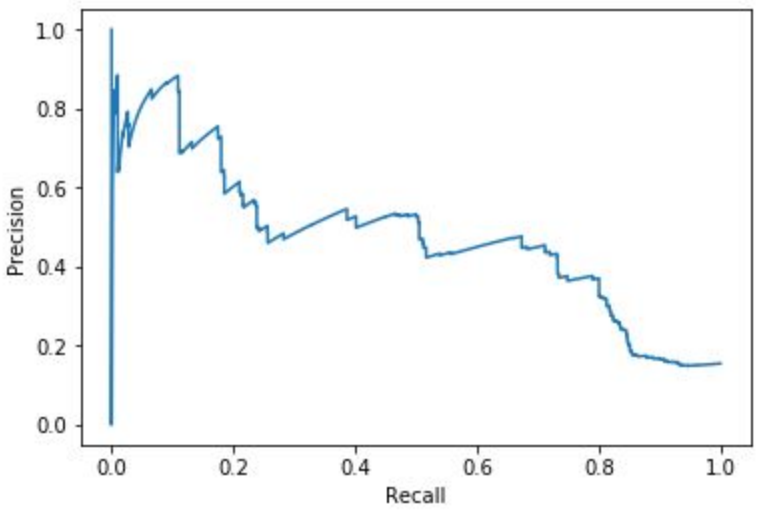
We ran two machine learning models. One included a sample of 30000 entries, and the other used the dataframe that only included one value for each zip-year, which included 1024 observations, or zipyears. The machine learning methods provided a novel way to predict hotel hubs beyond traditional statistical methods. We were able to create and train a model using a portion of each dataset and then test accuracy, precision, and recall on the remainder of each dataset. The model involved fitting and scaling, followed by reviewing the variables via a series of descriptive statistics. We then used a standard linear regression method and compared the precision results to those of other classification methods including random forest, extra trees, logistic regression, and others.

Results

We attempted to predict which zip codes in New York City will have an above average number of hotels using a random sample of 30,000 observations from the overarching population, which contains 30,000,000 records from 2011-2016. The mean score for "Hotel Hub" from the random sample was .155, suggesting roughly 85% of zip codes had a below average number of hotels.

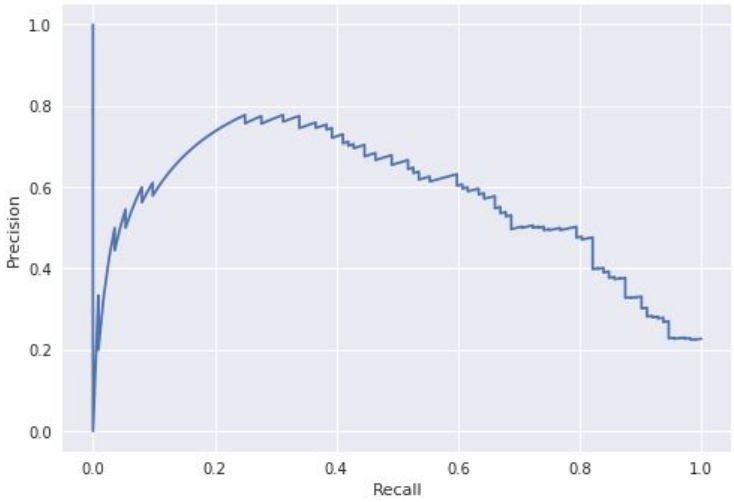
Using the independent variables averaged assessed value total, median household income, and total population for each zip code, we trained a model to predict the dependent variable, hotel hub. It is clear looking at descriptive charts that there are certain associations in the data. Total population and averaged assessed value are noticeably higher for hotel hubs. Median income has a 25^{th} to 75^{th} percentile range that is higher, but it is not as noticeable as for the aforementioned variables. All three of these independent variables had a number of outliers, requiring imputation to avoid outsized outlier effects. Across all independent variables, the distribution is concentrated at the lower end of the spectrum, making it difficult to extrapolate without drastically increasing the margin of error.

Using a standard linear regression model, our results of our model were heavily weighted toward the lower end of the probability spectrum, suggesting a much higher likelihood of a zip code not being a hotel hub. The confusion matrix for our model run on the sample shows that 12,581 instances were classified correctly as not hotel hubs and 365 were correctly classified as hotel hubs. There were 130 false positives (type II error) and 1924 false negatives (type I error). This implies that our model may have been underestimating hotel hubs on average across NYC. The model was 83.6% accurate with 74% precision and 15% recall. The model therefore sacrificed identification of hotel hubs for a higher likelihood of being correct among those hubs that were recalled. In the context of identifying hotel hubs, it may be preferable to have higher recall if we were to re-train the model given that the consequences of imprecision are not especially high. Based on the plot of precision recall for the linear regression model, it looks like we could maintain 60% precision at 20% recall, which would help expose a significantly higher number of hotel hubs. The model performs very well against the baseline (totally random), which only had a precision of 16% at 1%. However, the linear regression model does not perform well compared to the random forest classifier, for which there is hardly any sacrifice of precision as recall goes up.It is possible that we need to consider more independent variables to improve our recall without sacrificing precision. Precision and recall are charted below.



We also created a predictive model in which we condensed the number of observations to only one per unique combination of zip code and year ("zipyear"). The condensed value for each zipyear is the average of all observations for the zipyear. The independent variables are the same for this specification. There are 1024 zipyears total, with the model trained on half of them. The differences in median household income, averages assessed value, and population between hotel hubs and non-hotel hubs was much more pronounced for the zipyear analysis. It can therefore be concluded that the process of condensing observations into zipyears provided valuable clarity between the two groups.

Looking at the performance of the linear regression machine learning model for zipyear, there are many similarities to the first model. It is also heavily skewed toward the left bound when looking at the probability of predicting a hotel. Regarding the confusion matrix, there were 386 correctly identified non-hubs and 42 correctly identified hotel hubs. There were 14 false positives and 70 false negatives. This gives us an accuracy of 83.6%, a precision of 75%, and a recall of 37.5%. While the zipyear method didn't achieve results as high as the sample of the full population, it did maintain better recall, as seen in the chart below. This indicates that the zip year method may be more useful in assessing the need for additional hotels.



There are several important implications from these findings. Companies could use this predictive model when deciding where and when to open hotels. If zip codes are classified as hotel hubs by the model, it is possible that they are already saturated with hotels for given demographic levels. Zip codes with a below-average number of hotels and abnormally high population, income, or assessed value may be prime spots for new hotels. The model can therefore be used to study the relationship between demand for hotels and the existing supply.

Implications

According to a study conducted by the NYC Department of City Planning, findings suggest that NYC is one of the largest and most diverse travel and tourism markets in the Western Hemisphere, with both domestic-based and overseas visitors pouring in each day. Specifically, the number of tourists traveling to NYC grew from 47 million in 2007 to 60.7 million in 2016, an increase of almost 30 percentage points. As a result, demand for demand for hotel rooms has risen to a new high (NYC Department of City Planning, 2017). By predicting the future development of hotel establishments as the potential hotel hubs in NYC, our study has the potential to benefit not only hotel industry but also has the capacity to strengthen economic sectors and urban planning efforts.

For a hotel developer, site selection is a vital component to the industry as it contributes to a property's financeability, profitability, and long-term success. In the past decade, we have seen a consistent growth in hotel supply not just in Manhattan but also in the Brooklyn and Queens as well with the DCP reporting that the number of hotels in two boroughs doubling over the past ten years. (2017). As a result, identifying potential hotel hubs by zip codes may lead to prosperous business acquisitions for individuals in the hotel industry as properties outside of Manhattan may be less competitive.

As we previously have noted, the relationship between hotels and tourism is crucial. In many aspects, tourism shapes the hotel industry. For an example, hotels are built to fit tourists' needs

and often exist around popular destinations. Additionally, hotels play an important roles as pipelines. They generate local employment opportunities and help guide tourists to local businesses. As the available land dwindles in NYC, light manufacturing zoning districts (M1) have been a target for most hotel developers. Sadly, current land use regulations for M1 zones in NYC have not changed much since the rezoning in 1961. These zoning regulations limit the expanding and development of hotel facilities .The New York City Department of City Planning is now proposing new modified regulations towards hotel constructions in hopes that it will have a direct impact on the hotel industry and on employment. Additionally, growth of the hotel industry will benefit other industries as well. A previous study reported that hotels in tourism clusters in small-metro areas and towns performed consistently higher than their comparable segment properties out of the clusters (Angel, Maria, Luis and Rohit, 2014). Using the findings obtained from our study, policymakers will be able to foresee the trends of future hotel constructions, which in turn will allow them to provide guidance of where to build (i.e. districts who are lacking but have the potential for growth) while also boosting local employment and sales small businesses receive.

Limitations

Through the design of our machine learning model and our literature review, we have identified a few limitations with our project. One of the more obvious restraints is the exclusion of Airbnb data. We were unable to obtain data on Airbnb presence in New York City. We believe that this information is critical in making comprehensive conclusions on future hotel counts in the area. The literature reflects a crowding-out effect of Airbnb, a home-sharing platform, on New York City's housing market. Naturally, we expect Airbnb activity to have an effect on hotel counts in New York City. However, we believe that this interference is short-term, and will not have as large an effect on future projection as before the creation of regulations in New York City. This regulation, passed in 2010, prohibits Airbnb rentals of less than 30 days when the owner or tenant is not present. This regulation could minimize the effects on the hotel market. This

precaution is essential as Airbnb activity disrupts the existing hotel market and negatively affects housing affordability in New York City.

An additional limitation to consider involves the demographic information from ACS in our training dataset. This issue is twofold. We first would like to acknowledge the drawbacks of relying on ACS data. The issues of ACS data involve both the survey data collection methodology, and the use of historical data to create estimates. This is problematic as the distribution of underlying variables change over time. The secondary issue with our demographic data is that is measures characteristics of New Yorkers. This issue involves capturing and representing essential populations. This information is especially valuable when considering local housing debates. However, when determining hotel creation behavior, we must consider the backgrounds and activity of our tourist and out-of-town population. The majority of hotel decisions are made with touristic activity in mind. While it is important to consider the demographic information of New Yorkers as it is a factor in the debate, it is essential to supplement it with similar information of tourists.

A final restraint that must be considered is in regards to the machine learning methodology. While there are results that can paint a picture on the hotel debate in New York City, it is important to recognize biases and reliability issues. As the results are derived from our training set, application to a differentiated context may be difficult. A variation in context could be due to any changes in New York City, anything from economic to political circumstances, that are not accounted for in the model. The weights and design are set in the algorithm, and decision makers must understand this limitation when relying on these results.

Conclusion

The results of our linear regression machine learning models illustrate emerging value of machine learning methods as we enter the age of big data. In the context of predicting zip codes that are hotel hubs, both of our specifications significantly outperformed the the baseline

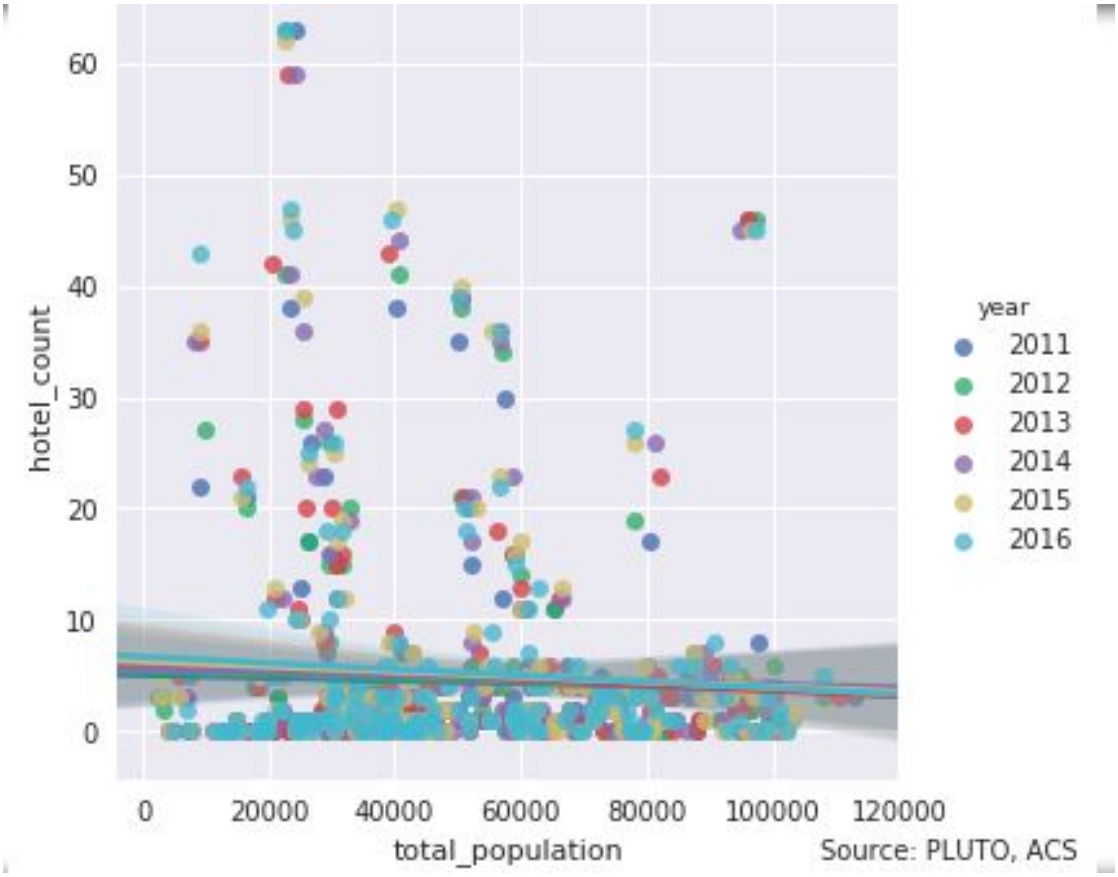
precision--the random sample method produced a precision at 1% of 74% compared to the baseline of 15% while the zipyear method produced a precision at 1% of 75% compared to the baseline of 21%. Hotels can use the information to more accurately meet demand for new lodging. These results, based on a relatively small number of variables with static observations, show how promising machine learning models that can continuously learn and improve performance will be moving forward.

Appendices

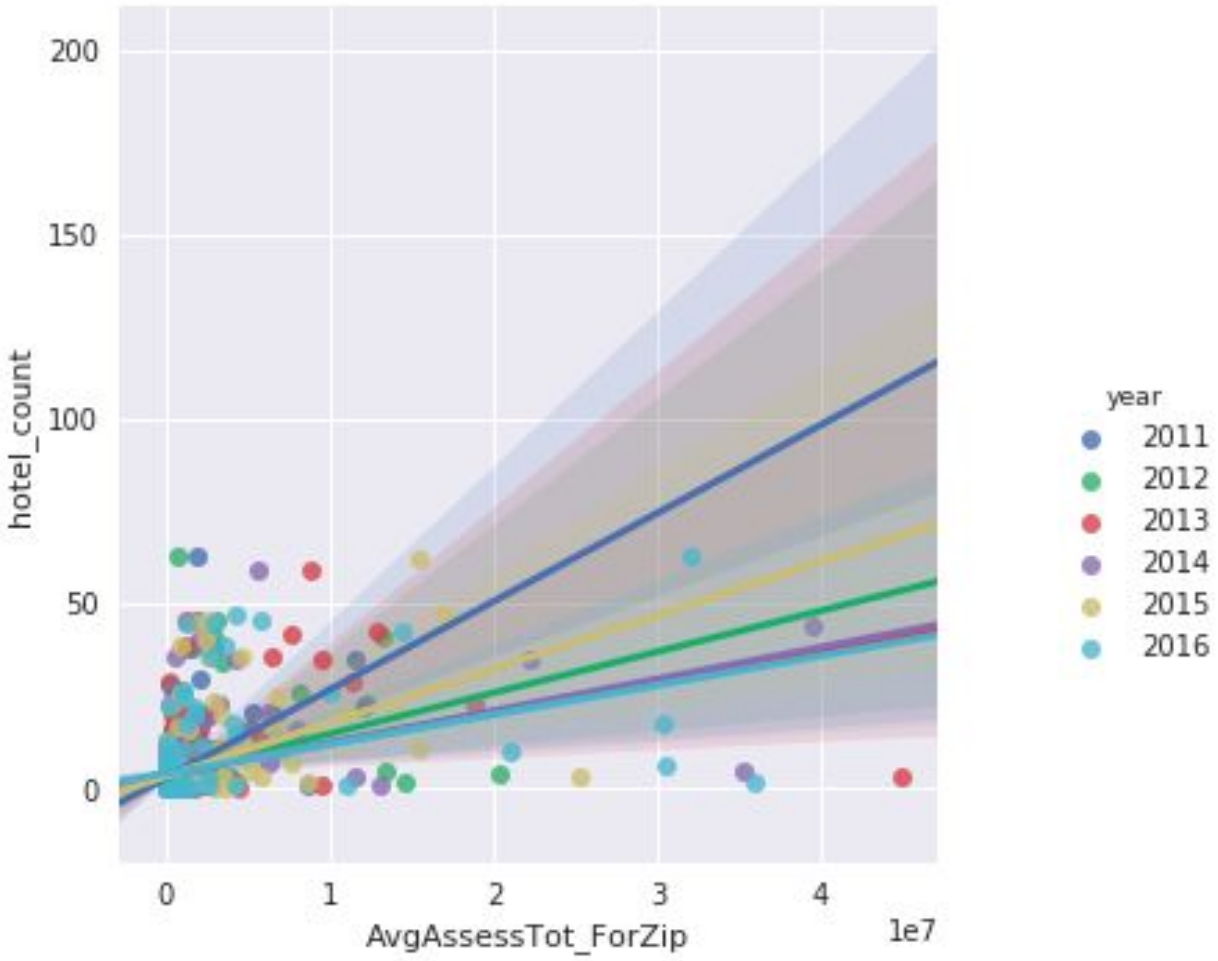
Appendix A:

Hotel Count in Relation to Covariates: Total Population, Total Assessed Value, and Shopping Count

Hotel Count in Relation to Covariates: <u>Total Population</u>

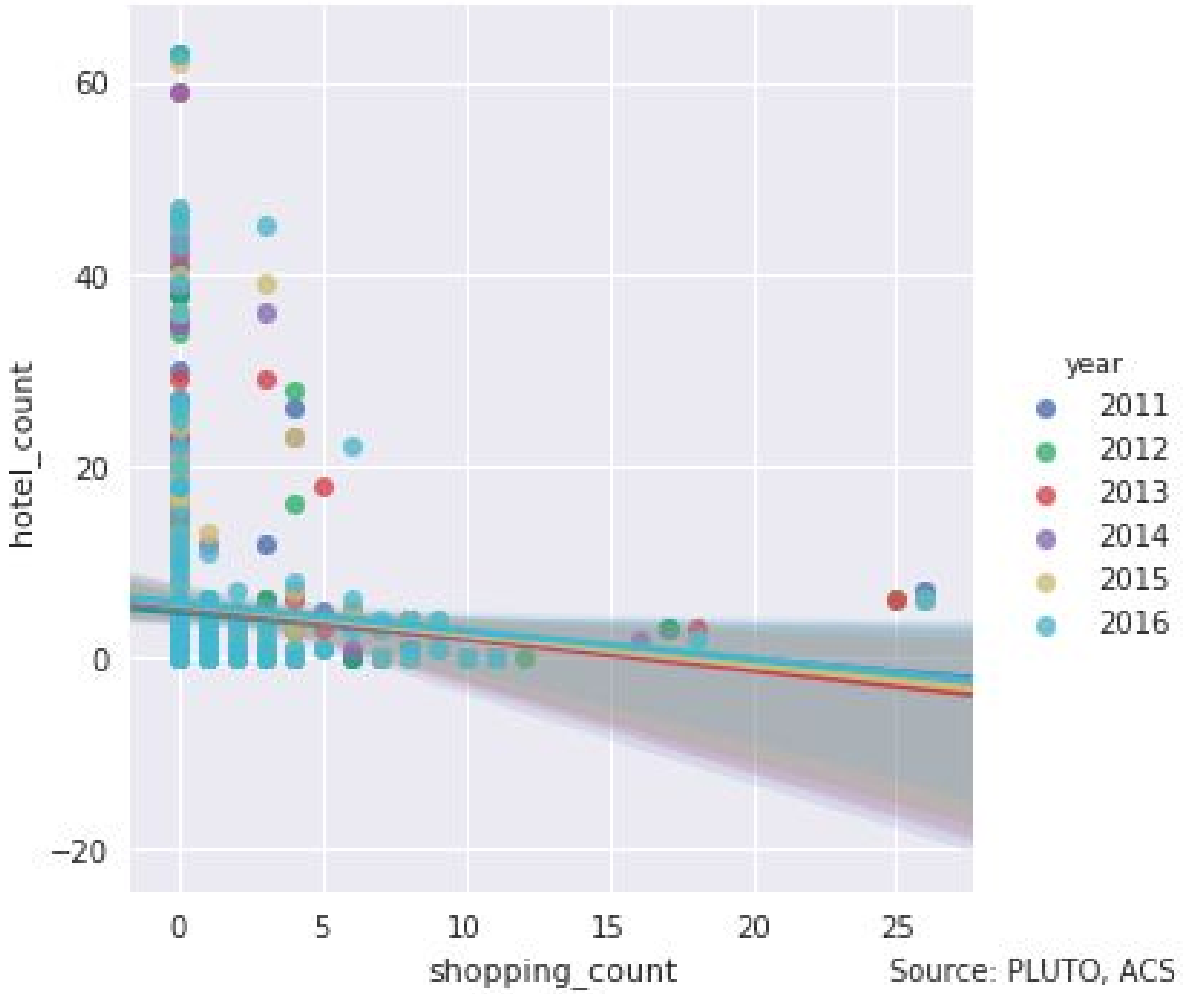


Hotel Count in Relation to Covariates: Total Assessed Value



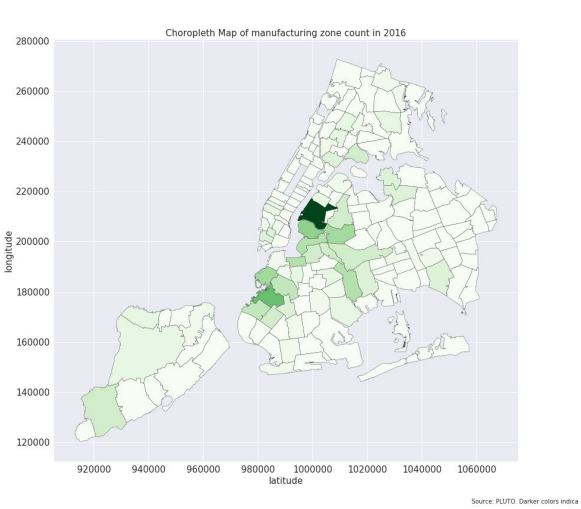
Source: PLUTO, ACS

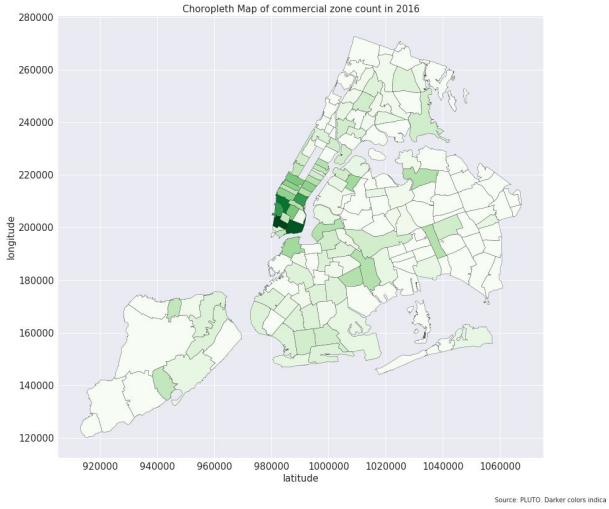
Hotel Count in Relation to Covariates: <u>Shopping Count</u>

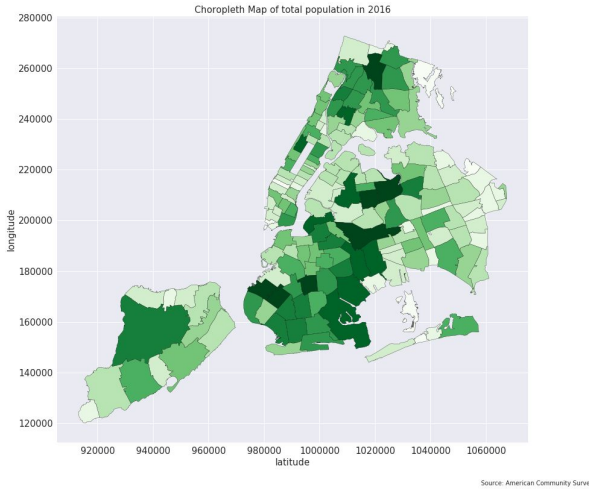


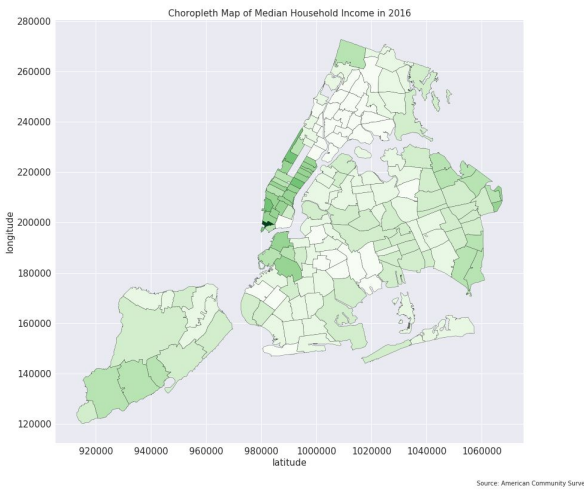
Appendix B: Choropleth Maps of Covariates Distributed on New York City Zipcode Level.

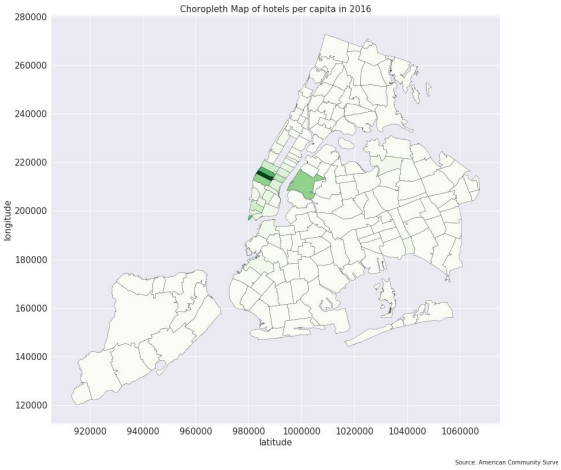
Darker Colors Indicate Larger Concentration of Data

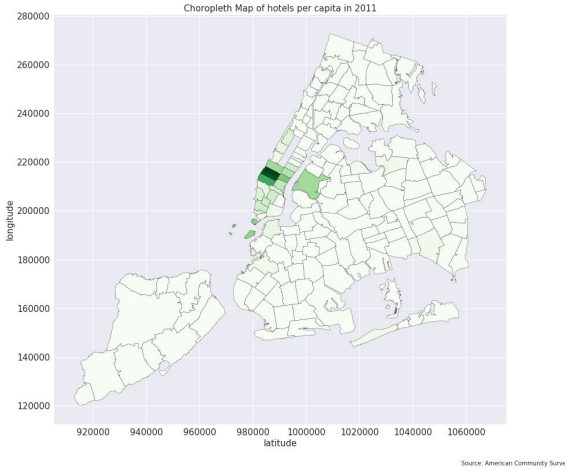


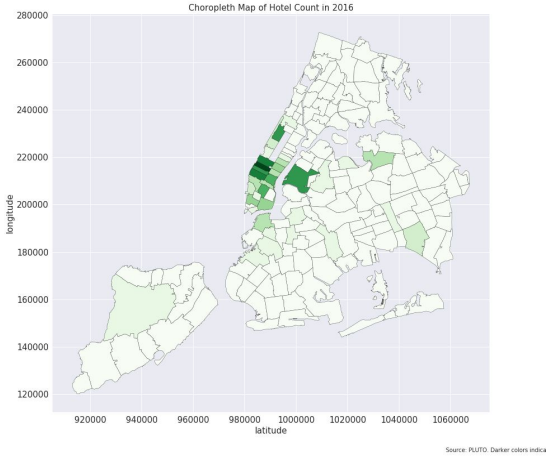


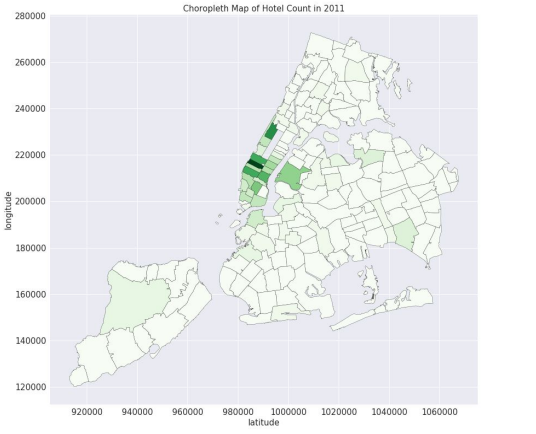


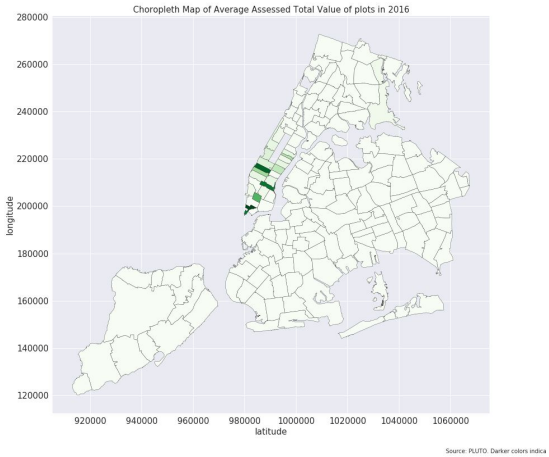


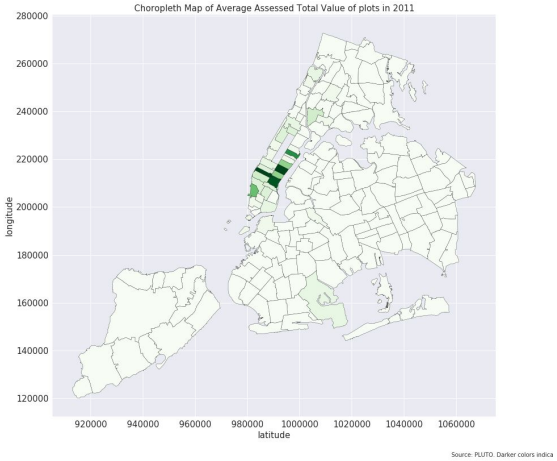

Choropleth Map of hotels per capita in 2016
longitude
latitude
280000
260000
240000
220000
200000
180000
160000
140000
120000
920000
940000
960000
980000
1000000
1020000
1040000
1060000
Source: American Community Surve


Choropleth Map of hotels per capita in 2011
longitude
latitude
280000
260000
240000
220000
200000
180000
160000
140000
120000
920000
940000
960000
980000
1000000
1020000
1040000
1060000
Source: American Community Surve


Choropleth Map of Hotel Count in 2016
longitude
latitude
280000
260000
240000
220000
200000
180000
160000
140000
120000
920000
940000
960000
980000
1000000
1020000
1040000
1060000
Source: PLUTO. Darker colors indica


Choropleth Map of Hotel Count in 2011
longitude
latitude
280000
260000
240000
220000
200000
180000
160000
140000
120000
920000
940000
960000
980000
1000000
1020000
1040000
1060000
Source: PLUTO. Darker colors indica


Choropleth Map of Average Assessed Total Value of plots in 2016
longitude
latitude
280000
260000
240000
220000
200000
180000
160000
140000
120000
920000
940000
960000
980000
1000000
1020000
1040000
1060000
Source: PLUTO. Darker colors indica


Choropleth Map of Average Assessed Total Value of plots in 2011
longitude
latitude
280000
260000
240000
220000
200000
180000
160000
140000
120000
920000
940000
960000
980000
1000000
1020000
1040000
1060000
Source: PLUTO. Darker colors indica

References：

1. Ashworth, G., 1989. Urban tourism: an imbalance in attention. Prog. Tour. Recreat. Hosp. Manag. 1, 33–54.
2. Ashworth, G., Tunbridge, J., 1990. The Tourist-Historic City. Belhaven, London.
3. Barros, C.P., 2005. Measuring efficiency in the hotel sector. Ann. Tour. Res. 32 (2), 456–477.
4. Clow, K.E., Garretson, J.A., & Kurtz, D.L. (1994). An Exploratory Study into the Purchase Decision Process Used by Leisure Travellers in Hotel Selection. Journal of Hospitality and Leisure Marketing, 4, 53-71.
5. Dev, Chekitan S., John D. Buschman, and John T. Bowen. "Hospitality marketing: A retrospective analysis (1960-2010) and predictions (2010-2020)." Cornell Hospitality Quarterly 51.4 (2010): 459-469.
6. Dolnicar, Sara, and The Otter. "Which hotel attributes matter? A review of previous and a framework for future research." (2003).
7. Edwards, D., Griffin, T., Hayllar, B., 2008. Urban tourism research: developing an agenda. Ann. Tour. Res. 35 (4), 1031–1052.
8. Issahaku, A., Francis, E.A., 2013. Dimensions of hotel location in the Kumasi Metropolis, Ghana. Tour. Manag. Perspect. 8, 1–8.
9. Lew, A., & McKercher, B. (2006). Modeling tourist movements: A local destination analysis. *Annals of tourism research*, *33*(2), 403-423.
10. Li, Mimi, et al. "A spatial–temporal analysis of hotels in urban tourism destination." International Journal of Hospitality Management 45 (2015): 34-43.
11. New York City Department of City Planning CENTRAL OFFICE (2017). NYC Hotel Market Analysis Existing Conditions and 10-Year Outlook. New York. NY.
12. Pearce, D.G., 1995. Tourism Today: A Geographical Analysis, 2nd ed. Longman, Har-low.
13. Pearce, D.G., 1998. Tourism development in Paris: public intervention. Ann. Tour. Res. 25 (2), 457–476.

14. Peiró-Signes, A., Segarra-Oña, M. D. V., Miret-Pastor, L., & Verma, R. (2015). The effect of tourism clusters on US hotel performance. *Cornell Hospitality Quarterly*, *56*(2), 155-167.
15. Plitt, Amy, 2017. NYC Tourism to hit record numbers in 2017. Retrieved from https://ny.curbed.com/2017/11/20/16678672/new-york-tourism-2017-nyc-and-company
16. Porter, R., Tarrant, M.A., 2001. A case study of environmental justice and federal tourism sites in southern Appalachia: a GIS application. J. Travel Res. 40 (1), 27–40.
17. Ruggero, S., Rodolfo, B., 2014. Structural social capital and hotel performance: is there a link. IJHM 37, 99–110.
18. Shoval, N., McKercher, B., Ng, E., & Birenboim, A. (2011). Hotel location and tourist activity in cities. *Annals of Tourism Research*, *38*(4), 1594-1612.
19. Yadegaridehkordi, Elaheh, et al. "Predicting determinants of otel success and development using Structural Equation Modelling (SEM)-ANFIS method." Tourism Management 66 (2018): 364-386.
20. Yang, Y., Wong, K.K., Wang, T., 2012. How do hotels choose their location? Evidence from hotels in Beijing. IJHM 31 (3), 675–685.
21. Yang, Yang, Hao Luo, and Rob Law. "Theoretical, empirical, and operational models in hotel location research." International Journal of Hospitality Management 36 (2014): 209-220.