Project Portfolio

William Dean

Updated 05/01/2018

Contents

Humor Study

1. Introduction

1.1 Overview

Our client, Dr. Oppliger, a professor at Boston University's School of Communications, administered a questionnaire (using Qualtrics), completed by approximately 700 undergraduate students at Boston University and East Carolina University. The questionnaire consists of 35 jokes and is split into two sections: 32 jokes compose "Study 1" and 3 jokes compose "Study 2." Four different versions of the questionnaire exist where the sex of the joke teller and joke target have been manipulated in each of the 4 versions. The study participants are, among other things, asked to rate each joke on a 1 (not funny) to 5 (funny) scale. For the Study 2 jokes, the students are also asked to give each joke an offensive score (1 - I strongly disagree that this joke is offensive to 5 - I strongly agree that this joke is offensive). Dr. Oppliger is interested in investigating the differences between male and female appreciation of humor. Exploratory data analysis led to a more specific research question regarding the Study 2 jokes: Does the gender of the joke teller influence the offensiveness ratings, and is this influence stronger for female respondents compared to male respondents? While a rigorous test (linear mixed effects model) did not show the difference in offensiveness ratings described above to be statistically significant at an alpha level of .05, other interesting relationships are brought to light. For example, the participants gender, political party affiliation, attitude towards sex, and attitudes towards women, are significant predictors of offensiveness rating. Furthermore, the direction of the association between offensiveness ratings and the gender of the joke teller does suggest that, for this particular set of jokes, male joke tellers are perceived as more offensive.

That is, do humor scores differ between the two genders of survey participants? Are humor scores dependent on gender of the joke teller or the gender of the joke target? Does offensiveness rating depend on gender of the joke teller?

1.2 Outline

The outline of this report is as follows. Firstly, we display the conclusions of our exploratory data analysis and state a specific research question of interest. Next, we describe the methods employed to answer our question via a statistical model. Lastly, we interpret our model and discuss the conclusions.

2. Exploratory Data Analysis (EDA)

For the Study 1 jokes, we analyze the humor response of four jokes where the gender of the joke teller and the gender of the joke target were manipulated. Similarly, we analyze the humor response of 28 jokes where only the gender of the joke teller was manipulated across surveys. For the Study 2 jokes, we analyze both the humor and offense response manipulating the gender of the joke teller.

Study 1: The humor responses over the four possible gender of teller/target combinations do not appear to differ largely by participants' gender. While we see a slight change in females' responses for the "hit" joke, this effect does not appear to be consistent across the jokes.

Furthermore, for the remaining 28 jokes in Study 1, the plot in Appendix 6.2 displays that, for each joke, there is little difference in humor ratings for female versus male joke teller.

Study 2: However, we find the Study 2 data suggest that the gender of the joke teller has an effect on the offensive scores for female participants. We see women tend to find the jokes less offensive when delivered by a female rather than a male.

It appears that for the same set of jokes, male respondents react indifferently with the gender of a teller, but female respondents tend to think the joke more offensive when it is told by a male. Thus, we hypothesize that gender of teller influences offensive scores for female respondents and conduct a regression analysis.

2.1 Research Question

Does gender of joke teller influence the offensive scores differently for male and female survey participants?

3. Methods

3.1 Variable Selection

To predict an offensive score we have 12 potentially informative variables: political score, social score, political party, feminine sex score, masculine sex score, positive attitude towards women, negative attitude towards women, attitude towards sex, funniness score, age, joke teller's gender, and respondent's gender. As intuition suggests that some of these predictor variables contain redundant information, we first check the existence of multicollinearity. That is, we examine if any of the predictor variables are highly linearly related.

For instance, A respondent with social score of 1 is more likely to have a political score of 1 as well, and a respondent with a social score of 2 is more likely to have a political score of 2, etc. The result shows a strong correlation between political and social scores. A second analysis, examining variance inflation factors, tells a similar story for the feminine and masculine sex score variables (see Appendix 6.3 for a thorough explanation). Thus, as a multiple regression model with collinear predictors may not give valid results, we decide to remove political score and masculine sex score from the remainder of our analysis.

3.2 Mixed Effects Model

Because each respondent is rating multiple jokes, and because each joke has its own overall or innate funniness and offensiveness rating, we choose to include respondent ID and joke ID as random effects in our model.

Our final model has 9 predictor variables including a random intercept for the joke, a random intercept for each survey participant, and a fixed effect interaction term between gender of the joke teller and gender of survey participant, which speaks directly to our research question. That is, our model accounts for the variability of an individual's response, the variability in response to each particular joke, and the effect of each possible gender combination (female teller and female participant, female teller and male participant, etc).

The continuous variables were centered and scaled (said differently, the variables were standardized). This allows us to more easily compare effect size across the model. Also note that the reference level in our model is male participant, male joke teller, and PolParty1 (political party, Republican). In other words, an average aged Republican male with average social, feminine sex, positive attitude towards women, negative attitude towards women, attitude towards sex and funniness scores would have a predicted offensive score of 3.13 (the intercept of the model) when the joke is told by a male.

4. Model

In our model, the numeric variables (social score (`Socially`), feminine sex score (`SexFem`), positive attitude towards women score (`PosATW`), negative attitude towards women score (`NegATW`), attitude towards sex score (`AttSex`), funniness score (`Score`) and age (`Age`)) if the variable increase by 1 standard deviation, our model would predict the offensive score to be driven up/down by the estimated coefficient.

For example, the age variable has a mean of 19.8 and a standard deviation of 2. Holding all the other variables constant, a 21.8 year old male (one standard deviation above the mean) will have an estimated offensive score 0.0166 lower than the 19.8 year old male.

For categorical variables (political party (`factor(PolParty)`), joke teller's gender (`Order2FemaleTeller`) and respondent's gender (`GenderFemale`)), the offensive score will change by the estimated coefficient in comparison to the reference level. For example, looking at the political party category and holding the remaining variables constant, a Democrat has a predicted offensive score 0.26 higher than a Republican.

Note the interaction term `Order2FemaleTeller : GenderFemale` included in the model. This represents the effect of the joke teller's gender on female respondents compared to male respondents. Taking it piece by piece: the `GenderFemale` coefficient tells us that women tend to have higher offensive ratings than males. For all participants, a female teller tends to result in lower offensive ratings (given by the coefficient for `Order2FemaleTeller`). However, if the participant is female AND the teller is female, this effect is stronger, as the estimated coefficient for `Order2FemaleTeller : GenderFemale` is negative.

While interpreting the coefficients is valuable, we must also determine if their values are significantly different from 0 in order to have confidence that the relationships described above are indeed present.

Tests of model fit were conducted to verify the appropriateness of its use.

We use our model to check the hypothesis that respondents' offensive ratings depend on gender of the joke teller, and that this effect is more prevalent for women. Note that the confidence intervals for the `Order2FemaleTeller` and `Order2FemaleTeller : GenderFemale` coefficients include zero, indicating that this effect is not significant at the 5% level. Furthermore, the offensive score prediction intervals for female survey participants given a male joke teller is (2.67, 3.93). Given a female joke teller, the offensive score prediction interval is (2.51, 3.79). We see a lot of overlap in this prediction interval; again, from a rigorous statistical point of view, the difference is not significant at 5% confidence level. However, the negative coefficients for `Order2FemaleTeller` and the interaction term `Order2FemaleTeller : GenderFemale` are informative. They suggest that participants, both male and female, are less offended by these jokes when they are told by a female and that this effect is stronger for female participants than males.

5. Discussion

Our model does give further explanations of offensive scores. For example, the following effects are significant at the alpha = .05 level:

- Women find jokes more offensive than men.
- Political party is a significant predictor of offensiveness (Democrats are the most offended).
- A high score on the 'positive attitude toward women' scale leads to the participant finding the jokes more offensive. This is a reasonable conclusion, as the subject matter of the offensive jokes could be classified as gender issues, and a participant that regards women highly would tend to find the jokes inappropriate.
- A high score on the 'attitude towards sex' scale leads to the participant finding the jokes less offensive. Again, this is consistent with the notion that people who are open-minded in regards to sex will be less offended by the sexual jokes.

City of Boston Autonomous Vehicles

1. Introduction

1.1 Project Background

The advent of autonomous vehicles has seemed like a foregone conclusion in recent years, especially with the successes seen at the DARPA autonomous vehicle challenges and the rollout of Tesla's semi-autonomous autopilot to their model S. An autonomous vehicle as described by Jo K et al. is "a self-driving vehicle that has the capability to perceive the surrounding environment and navigate itself without human intervention." 1 Furthermore, they say that in order for this to happen, "complex autonomous driving algorithms, including perception, localization, planning, and control, are required with many heterogeneous sensors, actuators, and computers."1 As autonomous vehicles become more of a reality, challenges are posed to cities that are trying to adopt the use of autonomous vehicles. Questions such as what regulations must be put in place to protect the safety of the public and what changes to existing infrastructure must be answered. While there are many obstacles to making autonomous vehicles a fully embraced part of our daily lives, the opportunities that they afford are also plentiful.

One of the most important impacts of autonomous vehicles is their ability to improve road safety. Human drivers have the propensity to get distracted i.e. texting, cellphone use, conversations. Autonomous vehicles do not have such an issue. Furthermore, the onboard computers and technology have the ability to gather and process more information faster than their human counterparts. A fleet of autonomous vehicles can play same role as Uber and Lyft without the burden of having to find safe and reliable human drivers. In recent news, we hear about problems that Uber and Lyft have been having with their driver screening processes. Many people don't feel safe getting in a car with a stranger. Autonomous vehicles would have no such issue. As for the contributions that Uber and Lyft give that a fleet of autonomous vehicles could take over, it could alleviate some of the burden on public transportation, and reduce the amount of parking space available in the city. Finally, since most autonomous vehicles are electric or hybrid gas/electric vehicles, they will reduce the usage of fossil fuels.

As with any new technology, the positive aspects are accompanied with negatives. Since autonomous vehicles are completely reliant on the performance of their computer systems, there is a concern for the security of these computer systems. Cybersecurity is currently one of the most important fields, and as technology progresses and becomes more integrated into our daily lives, it will become even more important. Also, there will be economic consequences. Autonomous vehicles will reduce the need for drivers, so there will be a drop in employment in that sector. There will also be a hit to the number of people who use traditional public transportation.

The impacts of autonomous vehicles, both positive and negative, will shape the future of our cities and so we must be proactive in understanding what is about to come.

1.2 Research Objectives

Our team was asked how autonomous vehicles will impact and influence the city of Boston and what the city can do to prepare for autonomous vehicles. Specifically, we were asked what data would be needed in order for the city to make decisions and overall be better prepared for the advent of autonomous vehicles. With such a broad objective given to us, we started off exploring the datasets that the city had already collected. With this collection of data sets, we were then able to formulate more specific questions that could be answered. Unfortunately with all the data sets that we found, there were issues that prevented us from doing any meaningful analysis. We were then given data for school buses in Boston and asked to assess the efficiency of the routes taken by the buses when compared to routes that an autonomous vehicle would take.

A measure for bus route efficiency was developed as well as a web based app to easily navigate through routes on different dates.

2. Data Description

2.1 Data Overview

Of all the datasets found on the Analyze Boston website, the traffic flow dataset seemed to be the most relevant one to the discussion of autonomous vehicles since it could be used to map the traffic patterns of the city on a daily basis. This contains information about various intersections throughout Boston. It counts the number of motor vehicles, pedestrians, and cyclists passing through the particular intersection in 15 minute increments for a 12 hour period. Understanding traffic patterns would be useful for planning how we use autonomous vehicles. Autonomous vehicles could be programmed to find alternate routes avoiding areas of high congestion, thus alleviating the congestion. Unfortunately, the data set is not sufficiently wide enough to get a good sense of the traffic patterns throughout the whole city.

The Massachusetts Department of Transportation has a dataset on motor vehicle crashes in the state of Massachusetts. The data is based on reports submitted by state and local police, other police departments, and prior to 2011 the vehicle operators. According the website though, "Some of the information in these reports has been aggregated and some may have been incorrectly or incompletely reported to us. Therefore, MassDOT makes no representation as to the accuracy or completeness of the crash records or the data collected from them." This becomes an issue when trying to make any kind of inference based on this data and thus it was not useful in our pursuits.

With these setbacks in the datasets, we then looked at how autonomous vehicles could help the sick and elderly by transporting them to hospitals and medical appointments. The Boston Region Metropolitan Planning Organization website contains a dataset that provides information on transportation services available to seniors in Boston. However the dataset does not contain detailed information on how often these services are used and what routes they take.

Finally, we had datasets that gave the locations of schools and hospitals around Boston. This would have been useful if the traffic flow data for those areas were available. We could have looked at the traffic congestion for these areas but unfortunately the data was not available.

This led to the final datasets that were given to us. We were given the path data for school bus B296 for a date range from November 8th to November 17th. This dataset gave the latitude and longitude position of the bus which allowed us to map the routes the bus took during that date range. Furthermore, we were given an idle spot data set which gave the latitude and longitude position for where the bus was idle for a given period of time. These idle spots could have been bus stops, stop sign stops, or heavy traffic stops. What matters though is that they gave us clearly defined legs of a route which we were then able to compare to routes suggested by Google Maps which would probably be the route an autonomous vehicle would take. With these two datasets, we were able to proceed to the next phase of our project.

3. Analysis

3.1 Method Description

Using the B296 Path data and the B296 Idle Spots data, we were able to map the route taken by the bus on any day within the date range of November 8th through November 17th. Furthermore, we were able to map the route suggested by Google Maps which would be a good estimation for the route an autonomous vehicle would take. We then calculated two types of bus efficiency. The leg efficiency of the bus route was calculated by dividing the distance the bus route traveled on a certain leg by the distance of the route suggested by Google Maps. The other efficiency calculated was the cumulative efficiency for the bus. This was calculated by taking the total distance the bus traveled, from the beginning of the route to the leg specified, and dividing that by the distance suggested by Google Maps. With these efficiencies, we can see how many times further a school bus would travel when compared to an autonomous vehicle.

3.2 Limitation of Data and Method - Future Steps

Looking through our results, we saw a few problems:

There were missing idle stops, so no route estimation was calculated from Google Maps indicated an untracked loop. This shows an issue with the idle spots data. The bus took this path, but the device that records the idle spots was never triggered during this path.

Another issue is with the path data. There are big gaps in the times that were recorded. This caused us to not be able to accurately calculate the time duration for each leg and so we could not create a time efficiency measure.

There was another issue with the way data had to be obtained from Google Maps. The functions used to get the Google Maps data were returning the same location for some of the idle spots that were close together, thus giving a distance of 0.

These issues would have to be addressed for any sort of exploration of the data to happen with this app.

4. Conclusion

Scaling up the application would allow production of daily efficiency scores per route, which in turn would facilitate both individual and aggregate-level analysis against various factors. In order for a scaling up of the application to occur, a few things would have to be taken into consideration. The above mentioned issues with the idle spots and path data would have to be addressed. A great deal of data processing was done in order for us to upload the data into the application. Finally, the functions used to gain the information from Google Maps have limits on daily usage. In order to avoid such limits, an API key would have to be obtained or a partnership with Google or Uber would give the required data.

Improving the quality of the data in all of the data sets mentioned in this report would allow further exploration as to how autonomous vehicles would impact the city. As mentioned above, the traffic flow data can be very valuable to understanding the traffic patterns of the city and how autonomous vehicles could make a positive impact. Our recommendation is that this dataset is expanded to include a better sampling of the intersections in Boston, and have flow data for multiple days. Furthermore, some useful data that is lacking is information on where people are going in the city. Since autonomous vehicles would be acting in a similar manner to Uber and Lyft, having their flow data and or origin/destination data would be extremely useful. If a partnership between the city and Uber or Lyft were to happen, they would probably insist on a loss of spatial resolution due to privacy concerns. An assessment of the impact on this loss of spatial resolution would have to be conducted during discussions with these 3rd party companies.

5. Appendix and Reference

- K. Jo, J. Kim, D. Kim, C. Jang and M. Sunwoo, "Development of Autonomous Car—Part I: Distributed System Architecture and Development Process," in IEEE Transactions on Industrial Electronics, vol. 61, no. 12, pp. 7131-7140, Dec. 2014.
- MassDOT Crash Portal, http://services.massdot.state.ma.us/crashportal/
- Analyze Boston, https://data.boston.gov/

Beantown - A Neighborhood Comparison of Boston & Chicago

1. Overview

1.1 Motivation

I am currently living in Boston, MA, but plan on relocating next year to Chicago. I have visited a few times but, from the large size in comparison to Boston, I have found it difficult to grasp which areas in Chicago I would enjoy let alone want to live. In general, in my recent travels, I have found difficulty exploring bigger cities. When spending a few days in a certain city, I often end up spending a lot of time traveling within the city, not having enough time to connect with the people and the cultures.

1.2 Goal

The **goal of this project** is to use data from areas and neighborhoods of Boston to see how communities compare to others within it and to Chicago.

When looking each neighborhood, there a few things that I am interested in measuring:

- What are the people like? How are the people that live in the area and what do people who visit think of the area and their time there.
- What are the living situations like? What kind of houses are in the area.
- What is the culture like? What are these communities known for and what are people talking about?

The question I want to answer with this project is: **What neighborhoods in Chicago are similar to the neighborhood I currently live in Boston?** That is, if I were to relocate to Chicago from Boston, what areas would I like based on my current location.

1.3 Data

I wanted data that would reflect all these measure. Here are a few sources I found.

Airbnb

The Airbnb Datasets provides insight into many cities in the US and around the world. Not only are there reviews written by people who use the service, but also the listings have information about the neighborhood, the house, and the host.

Using this information will provide information for the people who live in the neighborhood, details of the neighborhoods from the hosts, and the experiences of the people who visit.

Twitter

Twitter data gives insight to what is going on within a neighborhood, and what people are talking about. This accomplished trying to capture the neighborhood culture and potentially what people are like.

1.4 Methods

Doing the analysis, I mainly used a bag of words approach to the text analysis. For a neighborhood's set of reviews, I collected all the defining words for a given neighborhood as well as important sets of two words for a neighborhood.

When comparing two or more neighborhoods to each other, I used latent Dirichlet allocation (LDA) to group into a set of k topics.

This approach seemed appropriate since it will group neighborhoods with neighborhoods whose texts are similar and are talking about the same topics.

2. Calibration: The Neighborhoods of Boston

2.1 People from each Neighborhood

2.1.1 Hosts

Each Airbnb Host provides a description of them self and their household. Using the methods stated above, I grouped terms from each neighborhood. This process appeared to catch terms associated with people and aspects of their lives.

In one example, this method caught the word **accountant downtown**. In context, the self description was: "Lorin works as an **accountant downtown** and takes the expressway bus which is a quick 20 minute ride."

Using this method to grab characteristics of the people in each neighborhood, certain areas are grouped together. For example, the Fenway area is grouped with Longwood area. These areas are located around many hospitals in Boston. Other grouping is the North End and Downtown. These areas are close to many financial jobs and have many people that would be further in a career. This does appear to separate Boston into areas that have similar host.

2.1.2 People's Reviews

The contents of what people are saying into their reviews seems to depend on what neighborhood they are staying. Again, many neighborhoods far from the heart of Boston look to be grouped together, and as the neighborhoods get closer, there appear to stay grouped together. South Boston and the Waterfront are grouped together, and the Fenway area is grouped with Back Bay. The city appear to be separated into reasonable sections based on shops and attractions throughout the city.

2.2 Boston Neighborhoods

2.2.1 Neighborhood Descriptions

Every Airbnb host provides a description of the neighborhood where their property. Using this, we can see which neighborhood descriptions are similar. We see that some of the adjacent neighborhoods in Boston are grouped together. For example, Downtown and Seaport are grouped together. Other than that, many of the neighborhood are described differently enough from each other that they are not grouped together.

2.2.2 Housing

Using each Airbnb listing's detailed description of the property, we gain information about the type of homes in each location are where housing is similar. One interesting grouping is Back Bay with Fenway. This areas is filled with brownstone housing, having most streets with the same style of house.

2.2.3 Culture

Looking at tweets associated with each neighborhood either through direct mention or in a hashtag, we can pick up on what people in each area are talking about. The tweets were sample during the week of November, 30th.

3. Calibration: Comparison of Boston and Chicago

After looking at Boston, there is interpretability within each section above and does appear to give some insight into the similar neighborhoods in Boston. Let's consider how another Boston's neighborhoods compare to the ones in Chicago and see if we can extend this onto another city.

3.1 Tweets

This data will help understand which neighborhoods in Boston are talking about similar topics as the neighborhoods in Chicago. Using LDA with 25 topics for all 25 Boston neighborhoods, we can check if any of of the Boston neighborhoods match with neighborhoods in Chicago and which neighborhoods are not grouped together.

To have some sense of comparison, below are the areas in Chicago that are similar to Downtown Boston based of the twitter data. Interestingly enough, the projected similar locations in Chicago are in fact close to the downtown area of Chicago as well. This worked out well and suggests that this method of comparison is catching similarities between the cities.

4. Moving to Chicago

It appears this method does have some grounds in comparing not only neighborhoods in a certain city, but also has some ability to catch where similar places are in different cities as well.

Using this method to see which areas in Chicago are similar to Allston, MA, there are some interesting results. Although I am not too familiar with Chicago, there is one area that I am aware of. One of the marked areas in Little Italy and the West Loop. In this area, there's the University of Illinois at Chicago along with its students, many shops and restaurants. Similarly, Allston is close to Boston University and has many of its students live in the area during and after college. There are many strips in Allston that provide many food options and is a big destination for eating.

5. Conclusion

The neighborhoods in Boston have a interesting relationship with each other and are similar in as many ways as they are different. There are neighborhoods in Boston that provide many of the same attractions, but differ other aspects like in housing and residents. Using the LDA topic model, these similarities and differences did show with in the clusters which had interpretability that went along with characteristics of the areas in Boston. Extending this model onto Chicago, we saw that it was able to match the downtown area of Boston to areas around Downtown in Chicago. Knowing that this method has some sense of calibration, this process was able to be used to find areas in Chicago that are similar to Allston. Allston was mapped to a college area in Chicago which is quite reassuring. I am unaware of other areas that were marked similar to Allston, but they are on my radar for my next trip to Chicago.

5.1 Next Steps

Even though I was able to find out what locations are similar to Allston, I think this is just one approach to the question. Continuing on, I would like to see what other data I could use to expand on this. If there is data that provides a fine-grain look into the neighborhoods, I may be able separate each neighborhood into further classifications.

5.2 Limitations with the Data

The Boston Airbnb Data set had a lot more information about each listing than the listings from Chicago. When doing the analysis, I was limited to using only the twitter data and the housing listing for the comparison because that was all that was available. If I can get the full listings from the Chicago listings, it would interesting to see how results change.

5.3 What I Learned

With this project, this was the first time where I used only text data to answer questions. I got very familiar with the packages in R for text mining, mainly `tidytext`, `tm`, and `topicmodels`.

Analysis of Boston Rent Prices

1. Introduction

Boston is known for its high housing prices. However, these costs vary greatly upon the property location within the city. Throughout Boston, different places provide different sets of amenities as well as the proximity to various places of interest that make certain areas more desirable to live than others. Also, desirability of certain housing styles throughout the city can impact the price of rentals. Overall, when looking at rental prices in Boston, there is a delicate balance between the access the location provides and its affordability. The aim of this project is to investigate the the direct impact amenities have on the price to rent a room in Boston.

2. Data

2.1 Data Collection

The website PadMapper provides information about rent through the city. Along with this rent information, the number of rooms being rented was provided. This gives users the ability to look at the price of rent per rooms. Google's Places API provides information regarding shopping, dinner, grocery, entertainment, transportation options, nightlife, and what other amenities that are available in an area of a city. Information regarding parks and outdoor resources can be found on Boston's city website.

2.2 Data Description

After aggregating the data about the listings from the sources and extracting characteristics, each room listings had the following attributes:

Variables to capture the proximity of local amenities. Namely, the closest distance in miles to all of the following: T stop, bar, grocery store, coffee shop, restaurant, historical site, landmark, bike path, dog park, and park. These variables describe the local transportation, food, leisure, and entertainment and give insight to how accessible they are for that area.

Some variables reflecting where the listing is in the city. These include an indicator if the listing is within a half mile of a sports arena, another indicator if downtown, and one if the listing is within a half mile of the Boston Common. Similarly, the neighborhood of the city is attached to each listing and the distance in miles to the Boston Common.

There are two variables to capture the characteristics of the listings and other listings within the area. They are the average number of rooms of the closest 10 other apartments which tell what living situations in the area is like as well as an indicator if the listing is a studio apartment.

With each listings, the average traffic count of the closest 3 intersections was recorded in order to describe how busy the location is.

The variable of interest to estimate is the rent, however the log(Rent) for the listing provides a better comparison with the right skewness of rent prices. Figure 1 shows the location of the data points throughout the city of Boston and the log(Rent) price of that data point.

3. A Need for GLM

Even if you are not familiar with Boston, it is probably apparent that the rent prices are going vary, not only throughout the city, but throughout the neighborhoods in the city. It's also likely to believe how much they vary will depend upon where they are and what amenities the area provides.

For instance, the average rent can change between neighborhoods as well as how much that rent varies also depends upon the neighborhood. Areas like Allston have many types of living situations ranging from family homes to new luxury apartments which increase the variability of a room's price throughout the neighborhood. On the other hand, some areas provide fairly uniform living arrangements. For example, a downtown neighborhood like Chinatown has very similar apartment/loft style living or a very residential area like Roslindale may only have family homes which would affect the variability of the rent prices. Although a crude example, the dispersion of the data is likely to also depend upon the covariates of interest.

That being said, instead of just modeling the rent on the covariates, modeling both the mean and dispersion will likely capture the variation of the data. Using a GLM to jointly model the mean and dispersion will allow for the model such a scenario. That is, after modeling the mean of the log(Rent) with a linear model and the dispersion of data with a gamma model, a new model will take in account both the covariate's effect on the mean and dispersion and will better reflect the data.

4. EDA

Investigating the relationship the covariates have with the outcome, many appear to have the same relationship with the rent regardless of the location of the listing. For instance, the average number of rooms of closest listings increases on the log scale, the rent per room also decreases on the log scale. Not only does it decreases, but it generally decreases for all neighborhoods. This appears to be the case for these two variables as well as other relationships. Neighborhood appears to have no effect on the trend of the relationship. For many other relationships, the location does not change the effect of the covariate.

In comparison, the effect that some of the covariates appear to vary depending on the neighborhood. The relationship the log of distance to closest grocery store has with the rent changes. The rent on the log scale in Charlestown appears to decrease linearly as the closest to a grocery store increases, however in Jamaica Plain the relationship between these two variables appears a little different and moves in the opposite direction. The appendix shows a similar neighborhood effect with the variable for closeness to bicycle paths. There appears to exist interactions between some of the variables and the neighborhood of the listings.

Many of the variables appear to be more linearly related with the response variable when taking the log of the variable. That is, the marginally the data is very right skewed, having many amenities close, however a few listings that have a sparsity of amenities. From here, all distance variables are used on the log scale to account for this characteristic.

With all of this into consideration, it seems apparent to include an interaction term between the log distance to grocery store and neighborhood as well as the log distance to a bike path and the neighborhood. Many of the other variables suggest less of an impact of interaction with neighborhood so they'll be modeled without an interaction term with the neighborhood.

5. Model

5.1 Mean

With the response of log(Rent) being a continuous variable, we are able to model the mean of the data with a linear model.

This linear model has variables with no interaction term with the neighborhood of the listings. Namely, the log of the distance to the Boston Common, the log distance to the closest: T stop, luxurious apartment building, dog park, park, historical site, landmark, coffee shoe, restaurant, and bar all do not have an interaction term with neighborhood. Also the log of the number of rooms does not have an interaction.

In order to capture the effect of the neighborhood, an interaction term is included between the neighborhood with log distance to grocery store as well as the log distance to closest bike path.

Two other indicators are present in this model. They are indicators for if the listing is a studio and if it is within a mile of a sports arena.

After fitting the this model, we are able to then model the residuals and capture the missing variability from this model.

5.2 Dispersion

After fitting just the linear model for the mean, it is apparent that the model does not capture all the variability of the data.

Modeling the squared residuals from the mean model with a gamma GLM give insight to the dispersion's dependency on the covariates. A log link is used to relate the mean of the dispersion to the parameters.

The predicted dispersion from this model can be used weights to refit the linear regression.

5.3 Final Model

Refitting the first linear model with new weights, this model now takes into account the varying dispersion because of our covariates. The linear models use the same covariates but only differ by the weighting of the variances.

5.4 Model Comparison

The first model fit does not model the dispersion even though it is likely to have varying dispersion based off our covariates. Without modeling this dispersion, our AIC is -4057. In comparison, weighting our linear regression with a modeled dispersion, the model has an AIC of -5919.

Using a lower AIC to choose between models, it suggest that modeling both the mean and dispersion is provides a better fit for the data.

Looking at the residuals of the first model in comparison to the second model suggests the first model does not explain the variation of the data as well as the second model as the fitted values are larger.

The second model appears to capture the variation of the data better than the first model. The regression with the modeled dispersion has an R squared of 0.894 which is an improvement from the first regression's R squared of 0.85.

These few signs suggest that there has been an improvement in the fit of the data after modeling the dispersion, which better reflects our data.

6. Discussion

6.1 Effects of Covariates

Many of the covariates effects appear to follow intuition. For instance, the final model suggests a significantly negative effect of increasing log distance from closest T stop with a 95% confidence interval for the parameter is between -0.022 and -0.011. This suggests that as you get further away from the closest T stop, the price of a room will generally decrease. This makes sense with the T transport being used many people through the city and is clearly favorable to be near city transportation. Similarly, the model suggests that as the average number of rooms in the area increases on the log scale, the expected rent will decrease. This suggests that areas where there are living situations offer rooms will have generally have lower rent per bedroom prices.

The model suggests also some insights that may not be so apparent as well. For instance, as a listing gets farther away from its closest restaurant, there is a statistically significant decrease in its rent compared to

other listings. Being closer to a restaurant will, in general, increase the value of the room being rented. On the contrary, the closer a listing is to a bar, the lower on average the rent will be per room. The data suggests that living closely to a bar will generally have a lower rent than a comparable listing further away from a bar.

Consider the effect proximity a coffee shops has on the rent, our model suggest that there is no significant effect. The 95% confidence interval for the log distance to the closest coffee shop has a lower bound of -0.006 and upper bound 0.006. This model suggest that effects of a close coffee shop is insignificant after controlling for all other variables.

It is to note that the effect of the log distance to the closest coffee shop in the first model has a 95% confidence interval lower bound of 0.009 and upper bound 0.025. This is contrary to what the second model says; however, our second model reflects the data better so that is used for inference.

One effect that seems contrary to intuition is that the model suggests having a studio room will decrease the rent on average. The model's 95% confidence interval for this parameter is statistically negative, which seems rather odd that this effect would be the case. However, looking at the data suggests that non-studio rooms with comparable average number of rooms as a studio do appear to have lower rents, making some sense of this negative parameter estimate.

6.2 Challenges

While modeling this problem, there is a balance between capturing the relationships from the data while keeping a model that does not over simplify the situation. Most notably, being able to capture the effects of each listings attribute without having the added complexity that comes along with the interaction with the neighborhood of the listing. As we saw in the EDA, there potentially exists an interaction effect covariates. In order to keep a model that is rather interpretable and parsimonious many of these may be overlooked to provide that simplicity.

The data had no attributes which described the interior of home which could lead to further explanation of the variability of a room's rental price. For example, year of home or condition may greatly reflect in the rental price and could improve the fit of our model.

7. Conclusion

Figuring out what factors influence the price of a room rental throughout Boston could be explained with a linear model. However, it also seemed apparent that variability in the price also was influenced because of these factors as well. Because of that, it seemed appropriate to use a gamma GLM to model the dispersion of the data as well. After doing so, we saw improvements in our linear model ability to explain variability of the log price of a room rental through a few measures of goodness of fit.

After improving our model fit, we turned to inference from the model where we saw many conclusion that meet intuition as well as some insights about the effects of surrounding amenities and home features.

Jointly modeling the mean and dispersion of our data allows for a better fit model, leading to better inference about our parameters, and provides a understanding behind the amenities that drive room prices in Boston.

Boston Public School Shuttle Allocation

1. Introduction

Boston Public Schools (BPS) allow students from any district to attend any other school throughout the city but, when first implementing this policy, the efficiency and equability of the students was not taken in consideration. During 2017, around $2,000 was spent per student on transportation in BPS in comparison to the $350 national average. Currently, there is no rational behind the allocation of the shuttle transportation option so, with clear room for improvement, BPS was curious to how the procedure of shuttle placements could be more methodical.

2. Objective

Simulated locations of BPS students were provided however, without the true locations of students, we were unable to directly give insights on shuttle allocation. With this limitation of the data in mind, we aimed to give BPS the tools to analyze a school's current allocation and understand how other schools' shuttles compare.

3. Deliverable

We developed a web app which had two features:

Investigate transit saving shuttle locations for a school's student body. In doing so, our app gives a map of the place of top locations that would decrease the transit times of students the greatest. The current shuttle cost is presented in comparison to judge how effective the current allocation is.

Our second feature compares how different schools would gain from beginning to use shuttles. With this, we show the transit time that could be saved by a school if a shuttle was allocated to them in comparison to another school.

4. References

- Emanuel, Gabrielle. (2017, May 14). *High Costs And Empty Seats: Why Boston Public Schools Spends Millions On Bus Transportation*. Retrieved from https://news.wgbh.org

Impact of Feed on Clownfish Reproduction

1. Introduction

Based on the concept of life history theory, life history traits are predicted to be negatively correlated but are found to be positively correlated in clown fish. Our client Tina, a PhD candidate from BU's school of Biology, hypothesizes this may be due to varying environmental factors, interested in how feeding clown fish effects a multitude of measures associated with their reproduction and the relationships between those measurements.

To investigate her research question, she conducted an experiment and collected the data. In this experiment, there are 120 groups of clown fish in total. She randomly divided these clown fish anemones into 2 subgroups based on whether they were fed or not. The clown fish in control groups were not fed and the remaining 60 groups were fed. She measured clown fish's female and male growth, number of clutches laid, number of eggs laid per clutch, number of eggs hatched per clutch, and took videos of parental care. In the structure of data frame , each anemone is categorized as receiving the treatment or not. Within each anemone measurements are taken on the female (largest) and male (second largest) fish. Measurements on the number of egg clutches produced, number of eggs per clutch, and number of hatched eggs per clutch were also collected. The anemones were found at 10 different reefs (an average of 6 anemones per reef).

2. Research Questions

1. What is the impact of treatment on a number of dependent variables?
2. What is the correlation structure between the dependent variables(does it change given treatment)?

3. Regression Models

In order to understand the impact the treatment has on the dependent variables, we considered univariate relationship between the feed variable and all other response. We preferred individual univariate models over a single multivariate model to enable us to handle each response variable appropriately. That is, we wanted to correctly model each response variable individually in order to capture the variation from the feed variable.

We also preferred univariate models because each response variable had a different hierarchical structure where the measurements took place. For instance, there is only growth per Modeling the response one by one with their hierarchical structure allows us to capture any relationship we would expect between each response measurement.

After fitting appropriate regression for the fish tending times, growths, number of eggs laid, and hatch probability, we found the feeding of the fish only has a significant positive impact on the tending times of clown fish. That is, both fed male and female fish have a higher tending time than their non-fed counterparts.

4. Correlations

In order to investigate a change in the correlation structure of the dependent variables, we computed bootstrapped 95% confidence intervals for the difference all variable pairs between the two groups. Namely, the difference in correlation between fed group of fish and control group of fish.

Of the 11 pairs of dependent variables, only the correlation between male and female growth of the fed group is suggested to be different from the same correlation of the control group.

5. Summary

In order to understand the impact that feeding clown fish has on variables associated with reproduction, we fit individual mixed effect model to account for the hierarchical structure of the data. Our results suggest that only the tending times of fed fish are significantly different than fish that are not fed.

Investigating the change in the correlation structure between the control and fed group, needed us to bootstrap 95% for the difference between the correlations of the two groups. Of all pairs of dependent variables, the difference was only significant between the correlation of the growths of the two genders of fish.

Similarity of Boston Neighborhoods

Abstract

Boston's neighborhoods have distinguishable characteristics among each other, however pockets within these neighborhoods exude similar attributes and amenities. There are many benefits of knowing the similarities between different areas, and allows residents and visitors to find areas which fit one's desired needs. Using a graph to relate areas of the city, similar areas can be grouped together to show what areas in Boston possess commonalities in what they have to offer.

Introduction

Boston's neighborhood each have something unique to offer. There are, however, many aspects throughout the neighborhoods that are common among one another. Home rental service, Airbnb.com, attempts to find such commonalities by allowing users to choose desired characteristics for a neighborhood and shows comparable neighborhoods in the city. For instance, a user can desire an area with good transportation and nightlife options, and all neighborhoods which satisfy those characteristics are shown.

This is a powerful tool since many areas can differ within a single neighborhood despite being characterized by many similar attributes. For instance, one side of a neighborhood may provide access to city transportation, whereas the other side may not be as accessible. This lack of granularity limits Airbnb from accurately finding a wide range of commonalities.

Finding areas that are similar and satisfy all desired characteristics would improve the satisfaction of people who would now be more likely to get what they wanted.

Using a graph to represent different areas of the city and their edges to relate areas to each other ensures areas that are similar in a certain characteristic will be closer in the graph. With graphs representing all desired characteristics for an area, these graphs can be aggregated together in order to relate all areas based on specified characteristics. Having one graph relating all areas to each other with certain specification in mind will clearly show areas that are heavily connected to reflect the similarities of those specifications.

Data Description

Airbnb has a dataset of Boston rental listings which can be used as proxies for areas throughout the city. That is, each listing is used to assess how much that relative area satisfies some degree of particular characteristic.

Data surrounding particular characteristics were aggregated from a few sources. Boston data hub provides information surrounding the city services. For instance, T stop locations are provided on the Boston city website. Other amenity locations could be accessed through the Google Places API. Searching a particular keyword with this service gives us the locations of all places in the area that relate to that keyword. Fore example, searching *bars* with this API gives information surrounding all bars in the city.

The Boston Airbnb listings show locations where there are households in the city. For each one of these listings, there is the degree that a certain characteristic is satisfies. For example, how many transportation options are in the area or if the listings is close to a restaurant strip and so forth.

Methods

There were 3 main steps needed to be considered when approaching this problem: how to create graphs that exude different characteristics, how to combine the graphs which contain preferable characteristics, and how to cluster to find the similar listings.

Formation of Graphs

The first step is creating graphs that will characterize a particular attribute of an area and how each area relates to others throughout the city.

Nodes

The Airbnb listings are going to be the nodes because in finding similar houses throughout the city will also likely indicate similar areas. We are able to use the listings to assess a small area of the city for commonalities.

Edges

Since we are looking for a graph that relate one area of the city to another, it makes sense to relate one Airbnb listing to another. In connecting listings, we are finding houses which a similar to some degree. However, the similarity of two listings can be difficult to express which had me try a few options.

Firstly, I tried to connect link listings that were similar in proximity to places. That is, if two listings were within a close distance from an amenity, say a bar, then there would exist an edge between the two. This process fails to quantify the difference between listings that are in areas with a busy bar strips versus a single standing bar. Looking at just proximity does not take in account how many places the listing is close to.

In order to account for that, instead of looking at the closest amenity to the listings, I considered the number of a certain amenity that fall into a walking distance of that listings. In this method, if the difference in the number for two listings was small, then the listings would have an edge between them. This approach was a little better in connecting the listings, however when considering different types of amenities, the size of the differences changes and requires a lot of parameter tuning for results.

After trying these two methods, it seems obvious that the two most important characteristics in the similarity of two listings are both the proximity to the amenities as well as density of the amenities. Defining similarity like this is very intuitive since it would distinguish areas with, say, dense restaurant strips from a corner restaurant if they were both within walking distance.

To first account for the density, the amenities are clusters prior to considering the listings which tells high covered areas from areas with more sparsity of that amenity. The amenities were clustered using k-means distance with each other and the number of clusters was determined by the elbow of the curve of the total within sum of squares with number of clusters. Figure 5 in the appendix shows the clusters for bars throughout the city. This shows areas which have highly dense bars as well as smaller strips.

After the amenities are clustered, all amenities within walking distance of a listings are selected. This takes into account the proximity to each amenity. To account for the density of the amenities, the proportion of dense amenities to low density amenities is compared to other listings. Each listings will have a vector indicating the proportion of varying density of amenities and the dot product between two listings' vectors is used as a measure to how similar those listings are.

All of the listings have a value between zero and one, which can be viewed as weights of an adjacency matrix. This adjacency matrix can be transformed into a graph which will relate listings throughout the city to other based on a specific characteristic.

Graph Aggregation

After finding the similarity between two listings for multiple characteristics, we have multiple adjacency matrices corresponding to how much listings connect for different reasons. In order to combine the different graphs, I first considered using unions and intersections, however the weights of each adjacency matrix gets disregarded in the process, which lead me toward using weighted sums to combine the adjacency. That is, each adjacency matrix is multiplied by a weight which is between zero and one with all of the weights summing to one.

The resulting adjacency matrix also relates each listing to one another, but will be influenced by the individual edge weights from the other adjacency matrices.

The benefit of using weighted sums of the adjacency matrices allows priority of one graph over another. That is, if one of the weights is larger than another, the influence that that graph's edge has on the resulting graph will show more than that of the other matrices.

Clustering

With one adjacency weighted matrix taking into account all desired preferences, nodes which are more heavily connected will be areas that are more similar based upon those characteristics.

Since the matrix is a weighted matrix, it seems obvious to use a random walk clustering algorithm. I also tried fast greedy to improve the speed of the clustering, but random walk clustering allows smaller clusters to form which is beneficial for the point of this project. Namely, if there is a small area that should be separated from the rest, it should appear like that in the clusters.

Using random walking clustering leads to clusters within the graph that are similar based upon desired characteristics.

Results

Using these methods, we are able to find areas throughout the city which are similar to some standard.

For instance, consider the similarity of listings based upon T transportation. Ideally, areas throughout the city which have a lot of transportation options would be clustered together and places without easy assess to the MBTA will also be connected to each other.

Instead of clustering on the graph relating listing because of the T, consider of similarity of listings based on just bar locations throughout the city. In this scenario, the resulting clusters of the listings differ from those using the listings' similarity in bars.

Comparing areas based on one characteristic, T access and bars, results in two different adjacency matrices. However, if the adjacency matrices for both the T and bars are summed together each with equal weighting, this results in an adjacency matrix that values both T and bars the same. This adjacency matrix graph can be clustered as well.

The weights the two graphs were equal making the influence of each graph the same. Instead of weighting each of the two graph equally, a larger emphasis on the connections between listings that are similar in bars. Since we put more emphasis on the bar similarities, clusters from the bar graph will be more prominent. Weighting the adjacency matrices leads to a different clustering results.

Discussion

Conclusion

Using this method leads to clustering of areas based upon certain characteristics.

Aggregating the particular graphs allow for combinations of the individual characteristics to find similar listings based upon multiple specifications.

Upcoming

Currently when making the similarity matrix for a particular trait, each listing will have a probability vector of being close to a particular density type of amenities and to calculate the edge weight for each listing, the

dot product of the two vectors is used. This is probably not the best measure for similarity. Namely, cosine similarity would already be better. Other measures of similarities may be interesting to explore in the future to compare the clusters while using different similarity functions.

Currently, it is easy to interpret clusters of single graphs but, as the number of aggregated graphs increases, the clusters become harder to understand the similarities and differences. When looking at the individual graphs, we can make a sense of meaning from the clusters and then make some sense of the aggregated clusters. However, for the future, finding a way to see why two clusters are the same or different may be interesting and insightful.