DC Area Rent Analysis

DNSC 6211: Programming for Analytics

Dmitry Chudinovskikh
Patrick Steeves
Daniel Swart
Szu-Ying Yang
Xin Yuan

Abstract

Given the fact that rents in Washington, D.C. area have risen by 4 percent and it may always be the largest monthly expense of D.C. commenters, our group wanted to know what the significant factors are and how they could predict rent in the big D.C. area. To answer these questions, we used web scraping to obtain main data from Rental websites apartments.com and complementary Google location data for metro stations as well as IRS data for zip code. For the next few steps, we will use our analytic tools and sources to analyze the correlation between the dependent variable rent and those independent variables, produce data visualizations and develop an R-Shiny application aimed at interactively providing answers to the questions. We believe that our insights will help D.C. Rentals get a better understanding of their customers with providing the general picture of the larger concerns of D.C. renters specifically and that the project could be expanded to provide even more value for specific customer positioning or even public policy setting.

Contents

1 Introduction

Compared to last year, rents in the nations capital have risen by 4 percent and Foggy Bottom becomes the most expensive neighborhood for D.C. rentals. According to a new report by Apartment List, one-bedroom rentals in Washington, D.C. now cost a median $2,220 per month, while two-bedrooms are now asking a median 3,050 per month. Except for living in neighborhoods in D.C., commenters tend to choose cities nearby to settle for lower rent. As students studying in GWU, we are interested in finding out what the significant factors are and how they could predict rent in the big D.C. area, whether a given factor plays the leading role or has limited impacts. On the other hand, the descriptive and predictive analysis on rent would also give us the general picture of the larger concerns of D.C. renters specifically.

2 Background

Considering that the Washington DC metro area has one of the highest rental rates in the country, according to many new sources to include Business Insider, our group wanted to investigate the driving factors of the rental rates in the area. Apartments.com is a widely used internet services for finding apartments for rent that provides relevant data for each rental listing. We decided to use this as our main data source. After closer inspection of the data, it was found that a lot of the data was qualitative rather than quantitative. In order to explore more regression type analysis, we decided that additional data sources would be necessary. Intuitively we decided that area incomes, proximity to metro stations and local education systems are related to rental rates. We then decided to pull data from tax information, metro station locations and possibly SAT scores (if request is approved by the College Board). With these data sets, we ended up with a broader spectrum of predictors to analyze to determine which are the strongest in determining local rental rates.

3 Method

The major question we were trying to find the answer to was what were the major contributors to the amount of rent in different areas. In addition, we were considering possible dependencies between predictors such as median household income by zip code, distance to the nearest metro station and so on. The major reason for that was to find out whether there was multicollinearity between independent variables that could possibly affect our analysis. Not to mention that these findings will pose certain value on their own. At this point of our progress we are yet to find out which questions we will not be able to respond in a satisfactory manner. We can assume that some dependencies between predictors and/or rent are not going to be significant for our overall analysis without oversampling and, thereby, can be neglected. However, only further analysis will reveal which questions cannot have satisfactory answers.

4 Organization

So far we have gone through the data scraping part, in which the code was primarily generated by Patrick and was discussed and modified by the group. Following that, Dan will be in charge of the data exploration as well as the mapping part. Dmitry will be primarily working on Shiny while Szu and Cora focusing on building and interpreting the Linear Regression model. To present our results, we will deliver a joint contribution on recordings, making presentations as well as writing the report in a LaTex form.

4.1 Workflow

The first part of the workflow is to retrieve data from sources like Apartmenet.com, Google location, and IRS Tax database, where we utilize technique that involves web scraping and data cleaning. After that, we will process the data and generate statistical figures in order to perform initial data explorations by looking at the correlation and distribution of each variable. Complementary to this process, we will create some visualizations of showing plots by using ggplot, and matplotlib. In addition, we will also perform R Shiny to demonstrate an interactive way to show the results. We will then fit our data to the map and build up a regression model that predicts the dependent variable price corresponding to the set of independent variables.

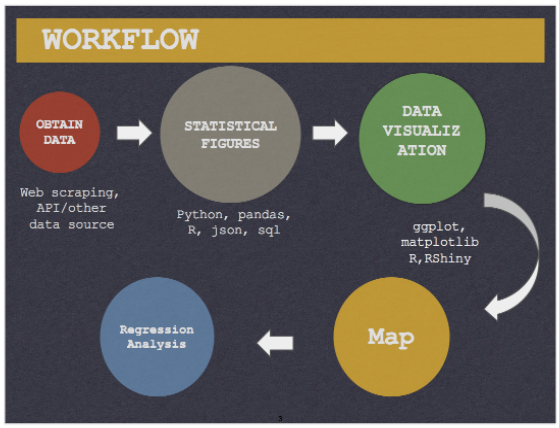

Figure 1: The project workflow

4.2 Project structure

We used three different online data sources for this project. All three sources were used to gather data on apartment rent levels in the greater DC area, and factor variables to predict it. First, we scraped apartment data from apartments.com for the cities of Washington (DC), Arlington (VA), Alexandria (VA), Bethesda (MD), and Silver Spring (MD). We collected the rent (dependent variable) and numerous independent variables such as number of bedrooms, Zip code, number and type of amenities, style of offering (for example studio or apartment), etc. We also obtained Google location data for all DC metro stops. This maps data was used to examine the distance between apartments and the nearest metro stop, which was added as a potential predictor for rent. Finally, we obtained income data from the IRS website. Using this data, we were able to determine the median income data for the zip codes scraped from apartments.com, and try to use this median income to predict rent.

4.3 Figures and Tables

List your tables and figures and explain why you chose to use them. Explain how these tables and / or figures contribute to your "story." Limit this to 250 words.

5 Discussion

This section requires you to discuss your experience. Describe the value of your project. What are two main "selling points" of your project. Limit this to 150 words.

5.1 Learnings

Discuss some of your "better moments" in this projects - the ones you enjoyed. Also describe. what you learned in this project. Limit to 150 words

5.2 Challenges

Discuss some of your "difficult moments" in completing this project. You may want to write about things you wanted to do but could not complete and why. Limit to 150 words.

6 Bullets and numbered lists (FYI, delete in your report)

This is up to you. If you want to add another section. This section explains how to make lists. In you final report you should delete this part.

6.1 Bulleted and Numbered Lists

LaTeX is very good at providing clean lists. Examples are shown below.

- Bulleted items come out properly indented and spaced, every time.
  - Sub-bullets are a virtual no-brainer: just nest another `itemize` block.
  - Note how the bullet character automatically changes too.
- Just keep on adding `\item`s...
- ...until you're done.

Numbered lists are almost identical, except that you specify `enumerate` instead of `itemize`. List items are specified in exactly the same way (thus making it easy to change list types).

1. A list item
2. Another list item
3. A list item with multiple nested lists
   - Nested lists can be of mixed types.
   - That's a lot of power and flexibility for the price of learning a handful of directives.
     (a) Like nested bullet lists, nested numbered lists also "intelligently" change their numbering schemes.
     (b) Meanwhile, all *you* have to write is `\item`. LaTeX does the rest.

4. Back to your regularly scheduled list item

BTW, this is a great site to generate tables in Latex and learn how to do it in Latex – `http://www.tablesgenerator.com/`

7 Conclusion

Wrap up your paper with an executive summary of the paper itself, reiterating its subject and its major points. Limit this to 150 words.