Assignment: ASSIGNMENT 8.3 Final Project Step 1

Name: Rapuru, Supraja

Date: 2021-08-01

Project: Analysis of how AirBnB rentals prices affects the nearby housing rental prices in Chicago

Introduction

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in that locale. It currently covers more than 100,000 cities and 220 countries worldwide. It largely does not own dwellings or real estate of its own; instead, it collects fees by acting as a broker between those with dwellings to rent and those looking to book lodging.

The company has been criticized for a direct correlation between increases in the number of its listings and increases in nearby rent prices and creating nuisances for those living near leased properties.

The problem here I am addressing is how the the prices of Chicago AirBnB rentals affect the prices of the nearby neighborhood rent prices.

Data science algorithm will help here to predict the prices of Chicago AirBnB rentals and also help to understand the correlation between the prices of Chicago AirBnB rentals and neighborhood rent prices.

Research questions

- What are the Airbnb rental prices for different areas in Chicago?
- What is the correlation between the Airbnb rental prices and Chicago neighborhood rental prices?
- What are the average rental prices by the neighborhood?
- What are the average rental prices for Airbnb by the neighborhood?
- What type of houses are most rented on Airbnb?
- What is the monthly rent from the Airbnb properties?
- What are the rental property options by neighborhood?
- How much profit does Airbnb make monthly?

Approach

Approach involves analyzing data to discover correlations, patterns and create machine learning model to predict how AirBnB rentals prices affects the nearby housing rental prices in Chicago based of various factors i.e. neighborhood, zip code, Airbnb prices, number of reviews, housing rental area, housing rental units etc.

- The approach is to start with finding the most important predictors for the regression model.
- Once the predictors are decided then I will look into the R2 , Adjusted R2 statistics, p-value.
- I will then calculate the betas for the predictors in the regression model. It will tell me how the 1 standard deviation change in predictor will impact dependent (response) variable.
- I will then calculate confidence intervals which indicate that the estimates how the model are likely to be representative of the true population values.
- I will then perform an analysis of variance on all models to compare performance of different models.
- I will then calculate standardized residuals, the leverage, cooks distance, and covariance rations
- At last I will check if the regression model unbiased and then will select the unbiased model for the prediction of the Airbnb prices

How your approach addresses (fully or partially) the problem.

Approach focus on to give enough data inputs to be able to address the problem completely. The approach will help to predict direct correlation between increases in the number of its listings and increases in nearby rent prices. It will help uncover various data patterns to answer multiple research questions. It will help understand cause and effect relationship between Airbnb prices and nearby housing rental prices. It also intends to develop a model to predict Airbnb prices based on given variables.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

1) AibBnb listing dataAs of October 2020, the dataset has 226030 rows and 17 columns of Airbnb listings in the U.S. The dataset includes NaNs, and data is of mixed types.

```
> airbnb_df <- read.csv("AB_US_2020.csv")
> str(airbnb_df)
'data.frame':   226030 obs. of  17 variables:
 $ i..id                          : int  38585 80905 108061 155305 160594 209068 213006 246315 259576 295496 ...
 $ name                           : chr  "Charming Victorian home - twin beds + breakfast" "French Chic Loft" "Walk to stores/parks/downtown. Fenced
 $ host_id                        : int  165529 427027 320564 746673 769252 1029919 1098412 1292070 1362726 1501882 ...
 $ host_name                      : chr  "Evelyne" "Celeste" "Lisa" "BonPaul" ...
 $ neighbourhood_group            : chr  "" "" "" "" ...
 $ neighbourhood                  : chr  "28804" "28801" "28801" "28806" ...
 $ latitude                       : num  35.7 35.6 35.6 35.6 35.6 ...
 $ longitude                      : num  -82.6 -82.6 -82.6 -82.6 -82.5 ...
 $ room_type                      : chr  "Private room" "Entire home/apt" "Entire home/apt" "Entire home/apt" ...
 $ price                          : int  60 470 75 90 125 134 48 65 71 50 ...
 $ minimum_nights                 : int  1 1 30 1 30 7 1 3 28 90 ...
 $ number_of_reviews              : int  138 114 89 267 58 54 137 57 537 31 ...
 $ last_review                    : chr  "16/02/20" "07/09/20" "30/11/19" "22/09/20" ...
 $ reviews_per_month              : num  1.14 1.03 0.81 2.39 0.52 0.49 1.35 0.53 5.01 0.29 ...
 $ calculated_host_listings_count: int  1 11 2 5 1 1 1 2 1 1 ...
 $ availability_365               : int  0 288 298 0 0 294 0 106 207 339 ...
 $ city                           : chr  "Asheville" "Asheville" "Asheville" "Asheville" ...
```

Data Source:
http://insideairbnb.com/get-the-data.html

Affordable rental housing data

The rental housing developments listed below are among the thousands of affordable units that are supported by City of Chicago programs to maintain affordability in local neighborhoods.
The dataset has 488 rows and 14 columns

```
> housing_df <- read.csv("Affordable_Rental_Housing_Developments.csv")
> str(housing_df)
'data.frame':   488 obs. of  14 variables:
 $ Community.Area.Name  : chr  "Edgewater" "Roseland" "Humboldt Park" "Grand Boulevard" ...
 $ Community.Area.Number: int  77 49 23 38 42 36 36 8 24 18 ...
 $ Property.Type        : chr  "Multifamily" "Senior" "Multifamily" "Multifamily" ...
 $ Property.Name        : chr  "Winthrop Apts." "Victory Center of Roseland" "Nelson Mandela Apts." "Cornerstone Apts." ...
 $ Address              : chr  "6214 N. Winthrop Ave." "10450 S. Michigan Ave." "3114 W. Franklin Blvd." "611 E. 50th St." ...
 $ Zip.Code             : int  60660 60628 60624 60615 60637 60653 60653 60610 60642 60707 ...
 $ Phone.Number         : chr  "773-477-7070" "773-468-6400" "773-227-6332" "312-577-5555" ...
 $ Management.Company   : chr  "Hunter Properties" "Pathway Senior Living" "Bickerdike Apts." "The Community Builders Inc." ...
 $ Units                : int  108 81 6 8 33 148 76 7 3 3 ...
 $ X.Coordinate         : num  1167689 1178829 1155445 1181237 1182661 ...
 $ Y.Coordinate         : num  1941496 1835494 1903206 1871959 1864419 ...
 $ Latitude             : num  42 41.7 41.9 41.8 41.8 ...
 $ Longitude            : num  -87.7 -87.6 -87.7 -87.6 -87.6 ...
 $ Location             : chr  "(41.9950154575665, -87.6585160357341)" "(41.7038907515241, -87.6207711552983)" "(41.8902013034932, -87.7045868112095)" "
```

Data Source:
https://data.cityofchicago.org/Community-Economic-Development/Affordable-Rental-Housing-Developments/s6ha-ppgi

Average rent Chicago neighborhood

This dataset contains 181 rows and 2 columns of average housing rental details for Chicago neighborhood.

```
> avg_rental_df <- read.csv("Avg_Rental_prices_chicago.csv")
> str(avg_rental_df)
'data.frame':   70 obs. of  2 variables:
 $ Neighboourhood: chr  "Near North Side" "Lakeview" "West Town" "Loop" ...
 $ Average.Rent  : chr  "$2,200 " "$1,395 " "$1,600 " "$2,350 " ...
>
```

Data Source:
https://www.zumper.com/rent-research/chicago-il

Required Packages

Packages for data transformation
1. dplyr
2. purr

Packages to Regression diagnostics
1. QuantPsyc - To get standard regression coefficients
2. car - Use durbinWatsonTest() to test the assumption of independent error
3. lm.test - Use dwtest() to test the assumption of independent error

Package for interactive plotting, model fitting, and stats about data
Rcmdr

Packages for data visualization and visual evaluation
1. ggplot2 - Useful to plot various charts to evaluate assumptions of linear regression
2. qqplotr - Useful to plot various charts to evaluate assumptions of linear regression

Plots and Table Needs
Histogram – To check normal distribution (a bell-shaped curve).
Scatterplot (Residual vs Fitted) - Access linearity of data
QQ plot of residuals - Access normality of residuals
Density plot

Questions for future steps

1) What are the other datasets (like crime data or school data) available that can impact the analysis?
2) Can we use different model for the predictions?
3) How can we check the quality of available data for the analysis?