Assignment: ASSIGNMENT 8.3 Final Project Step 1

Name: Anjale, Jiteshwar

Date: 2021-05-15

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

Introduction

Airbnb, Inc operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. The platform is accessible via website and mobile app. Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking.

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

1) What are the Airbnb rental prices for different areas in Chicago?
2) What is the correlation between the Airbnb rental prices and Chicago neighborhood rent prices?
3) What are the average rent prices by the neighborhood?
4) What are the average rent prices for Airbnb by the neighborhood?
5) What type of houses are most rented on Airbnb?
6) What is the monthly rent from the Airbnb properties?
7) What are the rental property options by neighborhood?
8) How much profit does Airbnb made 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, 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. I 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. I 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 data-
   As 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.

AibBnb-listing-data.csv

```
> str(airbnb_df)
spec_tbl_df [226,030 x 17] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ id                            : num [1:226030] 38585 80905 108061 155305 160594 ...
 $ name                          : chr [1:226030] "Charming Victorian home - twin beds + breakfast" "French Chic Loft" "Wal
k to stores/parks/downtown. Fenced yard/Pets OK" "Cottage! BonPaul + Sharky's Hostel" ...
 $ host_id                       : num [1:226030] 165529 427027 320564 746673 769252 ...
 $ host_name                     : chr [1:226030] "Evelyne" "Celeste" "Lisa" "BonPaul" ...
 $ neighbourhood_group           : logi [1:226030] NA NA NA NA NA NA ...
 $ neighbourhood                 : num [1:226030] 28804 28801 28801 28806 28801 ...
 $ latitude                      : num [1:226030] 35.7 35.6 35.6 35.6 35.6 ...
 $ longitude                     : num [1:226030] -82.6 -82.6 -82.6 -82.6 -82.5 ...
 $ room_type                     : chr [1:226030] "Private room" "Entire home/apt" "Entire home/apt" "Entire home/apt" ...
 $ price                         : num [1:226030] 60 470 75 90 125 134 48 65 71 50 ...
 $ minimum_nights                : num [1:226030] 1 1 30 1 30 7 1 3 28 90 ...
 $ number_of_reviews             : num [1:226030] 138 114 89 267 58 54 137 57 537 31 ...
 $ last_review                   : chr [1:226030] "16/02/20" "7/9/2020" "30/11/19" "22/09/20" ...
 $ reviews_per_month             : num [1:226030] 1.14 1.03 0.81 2.39 0.52 0.49 1.35 0.53 5.01 0.29 ...
 $ calculated_host_listings_count: num [1:226030] 1 11 2 5 1 1 1 2 1 1 ...
 $ availability_365              : num [1:226030] 0 288 298 0 0 294 0 106 207 339 ...
 $ city                          : chr [1:226030] "Asheville" "Asheville" "Asheville" "Asheville" ...
```

2) 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 429 rows and 19 columns.

affordable-rental-housing-developments.cs

```
> str(housing_df)
spec_tbl_df [428 x 18] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Community Area Name       : chr [1:428] "Portage Park" "West Englewood" "Englewood" "Washington Park" ...
 $ Community Area Number     : num [1:428] 15 67 68 40 23 38 3 43 5 28 ...
 $ Property Type             : chr [1:428] "ARO" "Multifamily" "Multifamily" "Senior HUD 202" ...
 $ Property Name             : chr [1:428] "4812-15 W. Montrose Apts." "New West Englewood Homes" "Antioch Homes II" "St. E
dmund's Corners" ...
 $ Address                   : chr [1:428] "4812-15 W. Montrose Ave." "2109 W. 63rd St." "301 W. Marquette Road" "5556 S. M
ichigan Ave." ...
 $ Zip Code                  : num [1:428] 60641 60636 60621 60637 60624 ...
 $ Phone Number              : chr [1:428] "630-694-6968" "773-434-4929" "773-994-4546" "773-667-7583" ...
 $ Management Company        : chr [1:428] "@properties" "Interfaith Housing Corp." "Universal Management Service, Inc." "S
t. Edmund's Redevelopment Corp." ...
 $ Units                     : num [1:428] 2 12 69 53 6 60 60 136 4 84 ...
 $ X Coordinate              : num [1:428] NA NA 1175445 1178070 1155238 ...
 $ Y Coordinate              : num [1:428] NA NA 1860492 1867952 1903559 ...
 $ Latitude                  : num [1:428] NA NA 41.8 41.8 41.9 ...
 $ Longitude                 : num [1:428] NA NA -87.6 -87.6 -87.7 ...
 $ Historical Wards 2003-2015: num [1:428] NA NA 31 53 41 1 37 32 NA 48 ...
 $ Wards                     : num [1:428] NA NA 32 4 46 10 18 33 NA 46 ...
 $ Community Areas           : num [1:428] NA NA 66 7 24 4 31 39 NA 29 ...
 $ Zip Codes                 : num [1:428] NA NA 21559 22260 21184 ...
 $ Census Tracts             : num [1:428] NA NA 479 403 177 165 633 381 NA 89 ...
```

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

Average_rent_Chicag
o_neighbourhood.xls›

```
> str(avg_rent_df)
tibble [180 x 2] (S3: tbl_df/tbl/data.frame)
 $ Neighborhood: chr [1:180] "Dearborn Park" "Printer's Row" "Streeterville" "Near East Side Chicago" ...
 $ Average Rent: num [1:180] 2419 2419 2410 2316 2308 ...
```

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. lmtest - 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 datesets(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?