Analysis of AirBnB rentals prices affect on near by housing in Chicago

Ghanta, Madhavi

2023-05-14

Case Overview

Problem Statement

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. 1. The approach is to start with finding the most important predictors for the regression model. 2. Once the predictors are decided then I will look into the R2 , Adjusted R2 statistics, p-value. 3. 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. 4. I will then calculate confidence intervals which indicate that the estimates how the model are likely to be representative of the true population values. 5. I will then perform an analysis of variance on all models to compare performance of different models. 6. I will then calculate standardized residuals, the leverage, cooks distance, and covariance rations 7. 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.

Packages

```r
# loading the required packages.

## Load the readxl package
library(readxl)
## Load the plyr package
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
## Load the plyr package
library(plyr)
```

```
## -------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## -------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
## Load the tidyverse package
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.1     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x plyr::arrange()   masks dplyr::arrange()
## x purrr::compact()  masks plyr::compact()
## x plyr::count()     masks dplyr::count()
## x plyr::desc()      masks dplyr::desc()
## x plyr::failwith()  masks dplyr::failwith()
## x dplyr::filter()   masks stats::filter()
## x plyr::id()        masks dplyr::id()
## x dplyr::lag()      masks stats::lag()
## x plyr::mutate()    masks dplyr::mutate()
## x plyr::rename()    masks dplyr::rename()
## x plyr::summarise() masks dplyr::summarise()
## x plyr::summarize() masks dplyr::summarize()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
```

Data Description

Data is imported from online Airbnb , Data from City of Chicago and Zumper.com repository.

```
airbnb_chicago_df = fread("http://data.insideairbnb.com/united-states/il/chicago/2023-03-19/data/listing
housing_df = fread("https://data.cityofchicago.org/Community-Economic-Development/Affordable-Rental-Hous
```

Names
The dataset present in:
* listings.csv.gz will be addressed as **airbnb_chicago_df** dataset
* Data set in Affordable rental housing data will be addressed as **housing_df** dataset * Data of Average rent Chicago neighborhood will be addressed as **avg_rent_df** dataset

These names will be frequently used in this report

Before diving deep into the data, we begin by having a glimpse of the data

library(readxl) library(dplyr) library(tidyverse) library(ggplot2) #library(purrr) #install.packages("ggpubr") #library(ggpubr)

#install.packages("caTools") #library(caTools)

Set the working directory to the root of your DSC 520 directory

setwd("C:/Users/mghan/Documents/dsc520/FinalProject")

#Above data set contains information across US cities #Filtering the data based on city==Chicago as we are focusing on Chicago library(readr) airbnb_chicago_df <- readr::read_csv('airbnb-listings.csv') str(airbnb_chicago_df) #AibBnb listing data , 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.

Load the Affordable rental housing dataset

housing_df=read.csv("Affordable_Rental_Housing_Developments.csv") glimpse(housing_df) #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

Load the Average rent Chicago neighborhood dataset

avg_rent_df <- read_excel("Average_rent_Chicago_neighbourhood.xlsx") glimpse(avg_rent_df)

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

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?

#Merge the airbnb df with rental housing df based on neighbourhood final_1_df <- left_join(airbnb_chicago_df,housing_df,
) glimpse(final_1_df) head(final_1_df)

#Merge the above df with Average rent df based on neighbourhood final_2_df <- inner_join(x=final_1_df,y=avg_rent_df,by
) glimpse(final_2_df)

#By looking at the data we can say that #Airbnb data # 1. Variable id is just an identifier and we can ignore it. # 2. We can factor the field room.type - Private room,Entire home/apt,Hotel # room, Shared room # 3. We can drop the host.id and host.name,neighbourhood.group,name fields # from the dataset # 4. We can drop fields like last.review,number.of.reviews, # reviews.per.month,calculated.host.listings.count

#Average rent Chicago neighborhood data # 5. We can drop Property Name,Phone Number,Management Company,Units,Zip Codes # from the # dataset

#Average rent Chicago neighborhood data # 6. rename the Average Rent to Average_Rent

Apply above transformation to the dataframe

final_df <- subset(final_2_df, select = c("neighbourhood_cleansed", "latitude", "longitude", "room_type", "price","minimum_nights", "availability_365", "property_type", "Zip.Code","X.Coordinate", "Y.Coordinate", "Latitude","Longitude", "Average Rent")) glimpse(final_df)

#Rename Average Rent to Average_Rent colnames(final_df)[14] <- "Average_Rent"

Checking the summary of data set to gauge the value range of each numerical

variable

summary(final_df)

7. Range of values prices are varies from 0 to 10000.

# It looks like there are outliers in the field.

8. Range of values minimum_nights varies from 1 to 365.

# It looks like there are outliers in the field.

9. Range of values for availability_365 varies from 0 to 365.

10. Range of values for Average_Rent varies from 675 to 2350.

#Calculate the 30 days price for airbnb property. final_df$airbnb_30_days_price = final_df$price * 30 summary(final_df)

#Check missing values apply(final_df, 2, function(x) any(is.na(x)))

#It looks like there are some missing values for #X.Coordinate ,Y.Coordinate, Latitude, Longitude, Zip.Code

2.What does the final data set look like?

glimpse(final_df)

3. Questions for future steps.

a) Need to learn how to visualize more than two variables.

b) Need to learn application of variable scaling and techniques.

c) Need to learn how lm() function takes care of variable scaling.

d) Need to learn correlation between different variables.

4.What information is not self-evident?

To uncover new information in the data that is not self-evident -

1. visualize data to uncover patterns and trends

2. correlation among variables

3. Check data distribution of variables

4. detect outliers and influencial cases

5.What are different ways you could look at this data?

Checking relation between airbnb_30_days_price and Average_Rent using

```
ggplot() library(ggplot2) ggplot(data = final_df, aes(x = airbnb_30_days_price, y = Average_Rent)) +
geom_point() + geom_smooth(fill=NA)
```

Checking relation between neighbourhood_cleansed and Average_Rent using

```
ggplot() library(ggplot2) ggplot(data = final_df, aes(y = neighbourhood_cleansed, x = Average_Rent)) +
geom_point() + geom_smooth(fill=NA)
```

Checking relation between neighbourhood_cleansed & airbnb_30_days_pric using

```
ggplot() library(ggplot2) ggplot(data = final_df, aes(y = neighbourhood_cleansed, x = airbnb_30_days_price))
+ geom_point() + geom_smooth(fill=NA)
```

#We can see that there is relationship between neighbourhood and prices # Checking if data distribution of numeric variables is normal # combining pipe operator between dplyr transformation and ggplot final_df %>% select(airbnb_30_days_price, Zip.Code, Average_Rent) %>% gather() %>% ggplot(., aes(sample = value)) + stat_qq() + facet_wrap(vars(key), scales ='free_y')

#None of the variables looks normally distributed ggplot(data = final_df, aes(x = neighbourhood_cleansed , y = airbnb_30_days_price)) + geom_boxplot() + ylab("airbnb_30_days_price")

# We can see that there are so many outliers for many neighbourhoods

thus data is not normally distributed

ggplot(data = final_df, aes(x = room_type , y = airbnb_30_days_price)) + geom_boxplot() + ylab("30 days price")

# We can see that there are so many outliers for room_type

thus data is not normally distributed

ggplot(data = final_df, aes(x = property_type , y = Average_Rent)) + geom_boxplot() + ylab("Property Type")

# We can see that there are so many outliers for Property_Type

thus data is not normally distributed

6.How do you plan to slice and dice the data?

unique(final_df[c("Zip.Code")])

unique(final_df[c("neighbourhood_cleansed")])

#I think need to slice the datasets by zip codes or neighbourhood to analyze # the data in more granular level

7.How could you summarize your data to answer key questions?

library("ggpubr")

ggscatter(final_df, x = "airbnb_30_days_price", y = "Average_Rent", add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson", xlab = "Miles/(US) gallon", ylab = "Weight (1000 lbs)")

#a) What are the Airbnb rental prices for different areas in Chicago? ggplot(data=final_df,aes(y=neighbourhood_cleansed)) + geom_histogram(stat = "count")

ggplot(aes(y=neighbourhood_cleansed,x=airbnb_30_days_price),data=final_df)+ geom_point()

From graph it looks like “West town” have major number of airbnb properties

Also the prices of “West town” properties are high for airbnb rental.

b) What is the correlation between the Airbnb rental prices and Chicago

neighborhood rent prices?

cor(final_df$airbnb_30_days_price, final_df$Average_Rent)

It is evident from the plots that there is positive correlation between

airbnb prices and average rent

c)What are the average rent prices by the neighborhood?

ggplot(aes(y=neighbourhood_cleansed,x=mean(final_df$Average_Rent)),data=final_df)+ geom_point()

#The average rent price is ~1600 per month # d)What are the average rent prices for Airbnb by the neighborhood?

ggplot(aes(y=neighbourhood_cleansed,x=mean(airbnb_30_days_price)),data=final_df)+ geom_point()

#The average airbnb price is ~ 5400 per month

e) What type of houses are most rented on Airbnb?

ggplot(data=final_df,aes(y=room_type)) + geom_histogram(stat =“count”)

#It looks like Entire home/apt are most rented on Airbnb

f)What is the monthly rent from the Airbnb properties?

df1 <-final_df%>%select(neighbourhood_cleansed, airbnb_30_days_price, Average_Rent) df1 %>% group_by(neighbourhood_cleansed) %>% summarize(mean_airbnb_30_days_price = mean(airbnb_30_days_price))

#Airbnb monrthly average rent is 5462.622

9)Do you plan on incorporating any machine learning techniques to answer

your research questions? Explain.

performing multiple linear regression

splitting the data into training and test set

```
library(caTools) mymodel_1 <-lm(airbnb_30_days_price ~ neighbourhood_cleansed,data = final_df) summary(mymodel_1)
```

```
mymodel_2 <-lm(airbnb_30_days_price ~ Zip.Code,data = final_df) summary(mymodel_2)
```

Questions for future steps?

1. I would like to plot the airbnb properties on map

2. I think I need to look for more data to determine the correlation and to

predict prices accurately