Capital One Data challenge

Alex Kiiru

January 22, 2021

Executive Summary

The Airbnb and Zillow datases provides us with a fantastic source of data to better understand New York's rental landscape. With over 50k listings registered in the last 9 years, New York has proven to be one of Airbnb's fastest growing cities. Between 2010 and 2017, as more and more people adopted the use of the internet as a service provider, the number of listings doubled every year. From our analysis, downtown Manhattan and the adjacent parts of Brooklyn had the highest concentration of listings by far. Staten Island and Bronx were the slow adopters but the data showed an increasing trend.

Problem Statement

You are consulting for a real estate company that has a niche in purchasing properties to rent out short-term as part of their business model specifically within New York City. The real estate company has already concluded that two bedroom properties are the most profitable; however, they do not know which zip codes are the best to invest in.

Objective

The task given is to analyse the datasets of Airbnb and Zillow to answer the following question:

- Which zip codes are the most profitable to invest in within New York City for two bed room properties

Assumptions

- The investor will pay for the property in cash (i.e. no mortgage/interest rate will need to be accounted for).
- The time value of money discount rate is 0% (i.e. $1 today is worth the same 100 years from now).
- All properties and all square feet within each locale can be assumed to be homogeneous (i.e. a 1000 square foot property in a locale such as Bronx or Manhattan generates twice the revenue and costs twice as much as any other 500 square foot property within that same locale.)
- The Occupancy Rate has been derived on the basis of certain conditions relating to Neighborhood and Overall Ratings (will be further discussed in the Analysis)
- Imputation has been done on the basis of the MICE package in R. The package creates multiple imputations (replacement values) for multivariate missing data.
- We now know that Covid-19 is a factor that would be have to be accounted for in order to make accurate predictions for investors in 2021, but this factor is omitted from this analysis because it is based only on data from 1996-2017.

Key Insights

- Renters in the neighbourhoods of Staten Island and Queens have limited choices because there are fewer properties listed there, but these are desirable neighborhoods for investors because they cost less, are highly popular rentals, and return higher rental prices.
- Properties in Manhattan and Brookly have seen major property price increase in the last 5-10 years and thus could have high resale value. Property prices in Staten Island and Queens have been stagnant for over a decade.

Results

Using different approaches discussed in the model building segment of this report, the best zipcodes to invest in are:

1. Optimal Revenue Model: 11231 ,11217 ,11215, 10036,10025 ,10003 ,10011 ,10128
2. Break Even Model: 11434,10304,10309, 11234
3. Mean Return Rate Model: 11434, 11234, 11304

These zipcodes derive higher daily rental value, provide more choices, and are popular among current consumers in terms of number of reviews and occupancy.

Future steps & Recommedations

- Expand this analysis to multiple cities with compare patterns and trends between different cities. From the insights that have been derived, I would also like to build predictive models using different features from the dataset.
- Merge this data with Covid-19 data to see how the pandemic would affect the demand and supply of airbnb rentals in the next few years in New York.
- Introduce seasonality and weather data to understand trends in occupancy rates throughout the year and create a model to predict occupancy rate based on this insight.
- Use NLP models to analyze the qualitative part of Airbnab data. We can account for the trends in reviews and get sentimental insights from word cloud that drive demand and occupancy rates.

Metadata Created

- Annual_Return_rate - revenue generated by property per year dived by predicted price of the property
- mean_Return_Rate - mean revenue genarated by property per year
- predicted_price - predicted cost price from the zillow data set. In the data set only the price as of february 2021 2019 is taken and is in dollar amount
- occupancy_rate - Percentage occupancy of the airbnb listing. It is represented as intervals.
- breakeven_period - Time it takes for the property to return it's cost price. This is also known as breakeven period and it is taken in the form of years
- Annual_return - revenue generated by property per year

Package Loading

Required Packages

- Tidyverse (dplyr, ggplot2..) - Data Read, Manipulation and visualisation
- Data Explorer - EDA & Generate Reports
- Caret - Pre Processing, Feature Selection
- Plotly - Interactive Visualization
- KableExtra - Styling Data Tables within Markdown
- MatrixStats - for manipulating operating on rows and columns of matrices.
- Choropleth packages -for spatial data
- Ggmap- for spatial data
- Other minor packages described in config file.

Loading Data into R Environment

Revenue(Airbnb_data) and Cost(Zillow) Datasets are loaded into R-Enviroment for Exploratory data analysis. If new markets or cities have to be explored at a later stage, new files must be loaded into the same directory and the code will automatically scale to account for it.

Loading Airbnb_data Data

```
Airbnb_data <- read.csv("C:/Users/kiiru/Desktop/Cpaital_one _data_challege/listings.csv")
```

Loading Zillow Data

```
Zillow_data <- read.csv("C:/Users/kiiru/Desktop/Cpaital_one _data_challege/Zip_Zhvi_2bedroom_2021.csv",check.names= FALSE,fi
leEncoding = "UTF-8-BOM")
```

Dimension and Shape Analysis of Zillow and Airbnb_data data

Checking Dimension of both data sets (Number of Rows, Number of Columns)

```
## [1] 48895   106
```

```
## [1] 8946  262
```

Data Structure

We check the data structuctures and variable names in each data set.

Initial Data Preprocessing

Data Preparation Zillow data set

In the Zillow dataset below, most of the columns represent median prices for 2-bedroom homes between 1996 to 2017 with one month spread.

RegionID	RegionName	City	State	Metro	CountyName	SizeRank	1996-04	1996-05	1996-06	1996-07	1996-08	199

RegionID	RegionName	City	State	Metro	CountyName	SizeRank	1996-04	1996-05	1996-06	1996-07	1996-08	199
61639	10025	New York	NY	New York	New York	1	NA	NA	NA	NA	NA	N
84654	60657	Chicago	IL	Chicago	Cook	2	167700	166400	166700	167200	166900	1669(
61637	10023	New York	NY	New York	New York	3	NA	NA	NA	NA	NA	N

Quality Check of Zillow data

Missing value Check

Median Price for early years (1996-2013) has many Nulls as shown in the table below. This is not consistent across all region names. Steps are taken in the following section to filter columns with higher percentage of Nulls/NA.

	total_missing_values
RegionID	0
RegionName	0
City	0
State	0
Metro	0

Duplicates values check

We can see that our data does not have duplicates.

0 rows | 1-10 of 262 columns

```
## [1] TRUE
```

```
## [1] TRUE
```

Negative and Zero valued columns check

There are no negative or zero Values in any of the non-character columns(Int/Numeric) as shown in the following table.

	zillow_neg_and_zero
RegionID	0
RegionName	0
SizeRank	0
1996-04	0
1996-05	0
1996-06	0
1996-07	0

Filtering Data to get Desired Region (New york)

To reduce the number of columns and only focus on our desired region, the Zillow data is filtered to only show New York region. In the future, if a different city is desired, the name of the city can be changed to the desired city.

zipcode	1996-04	1996-05	1996-06	1996-07	1996-08	1996-09	1996-10	1996-11	1996-12	1997-01	1997-02	1997-03	1997-04	1997-05	1997-06	1997-07
10025	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10023	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10128	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10011	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10003	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11201	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Predicting prices for houses untill 2021

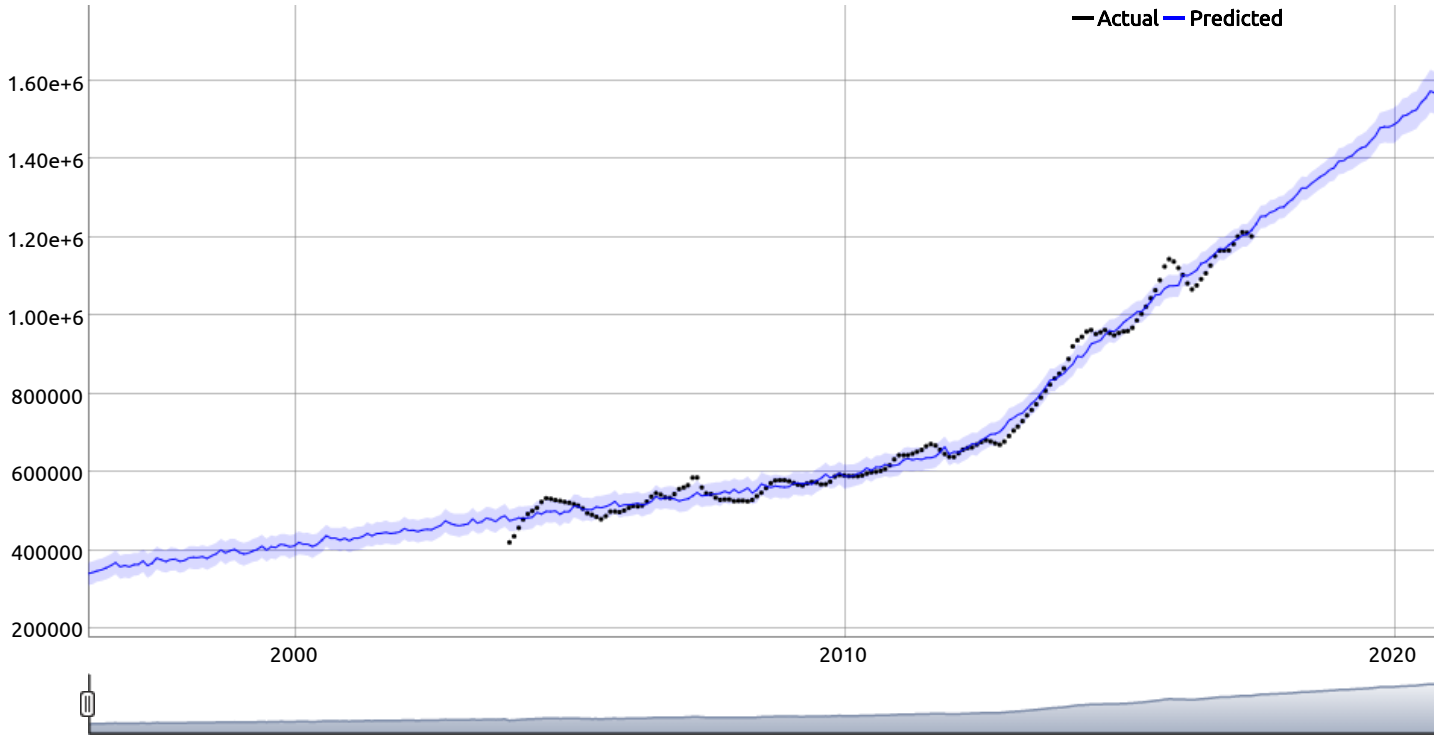
Used Facebook open source Prophet package to identify and extrapolate the trends in house prices in New York.

	zipcode <int>	1996-04 <int>	1996-05 <int>	1996-06 <int>	1996-07 <int>	1996-08 <int>	1996-09 <int>	1996-10 <int>	1996-11 <int>
1	10025	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
2	10023	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
3	10128	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
4	10011	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
5	10003	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
6	11201	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*

6 rows | 1-10 of 258 columns

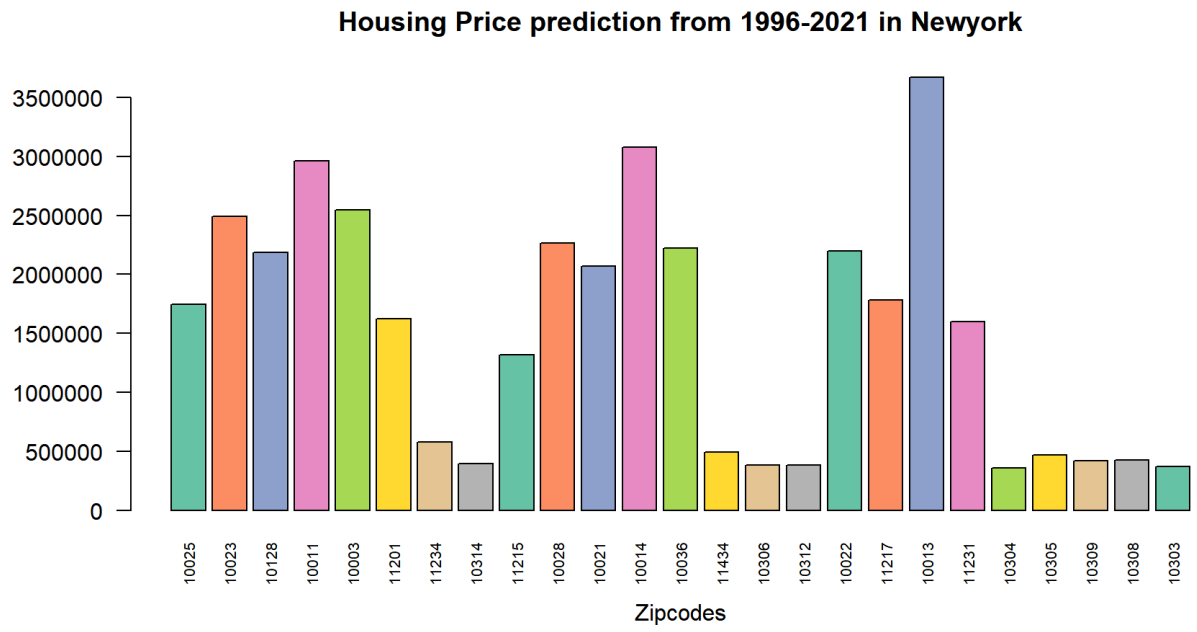
Visualizing predicted price

Randomly picked a zip code (11231) to demonstrate the results generated by the model. The model predicts an increase in housing prices to around $1.6 million by 2021.



Visualizing Housing Price prediction from 1996-2021 in New York

This graph shows the predicted price for February 2021, for each of the zip codes.



Filtering final data for merging

	zipcode	2017-06	predicted_price
20	11231	1202900	1600232.0
21	10304	328300	360413.1
22	10305	425100	469420.8
23	10309	390500	416740.8
24	10308	409500	424353.4
25	10303	327700	369223.6

Data Preparation Airbnb_data data set

Revenue data contains a mix of information including details about the properties, like address, zip code, bedrooms, bathrooms, information about host, daily/weekly and monthly price details for stays.

id	listing_url	scrape_id	last_scraped	name	summary	space
2539	https://www.airbnb.com/rooms/2539 (https://www.airbnb.com/rooms/2539)	20190708031610	2019-07-09	Clean & quiet apt home by the park	Renovated apt home in elevator building.	Spacious, renovated, block to F train, 25 mi

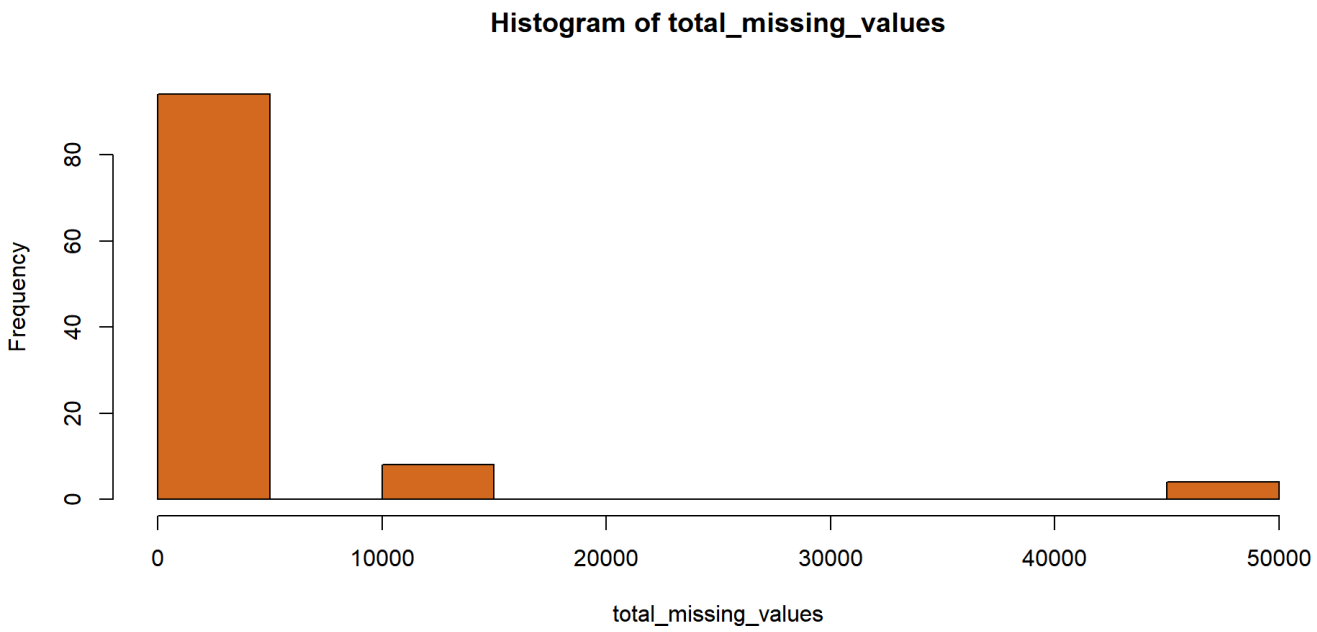
Quality Check of Airbnb_data

Missing value Check

The data Airbnb also had null values. To preserve all the information, we will impute or drop the rows and columns containing null values later on in this project, while conducting exploratory analysis that makes use of these features.

	total_missing_values
id	0
listing_url	0
scrape_id	0
last_scraped	0
name	0

We construct a Histogram plot to analyze the missing values for the variables that we use in our exploratory analysis



Duplicates values check

There were no duplicates found in the Airbnb data.

0 rows | 1-9 of 106 columns

Negative and Zero valued columns check

There were no negative or zero valued columns found in the Airbnb data.

	Airbnb_data_neg_Zero_count
id	0
scrape_id	0
thumbnail_url	0
medium_url	0
xl_picture_url	0
host_id	0

	Airbnb_data_neg_Zero_count
host_listings_count	901

Checking columns with no data

Missing value analysis

	missing.values
thumbnail_url	48895
medium_url	48895
xl_picture_url	48895
square_feet	48487
review_scores_location	11082
review_scores_value	11080
review_scores_checkin	11078
review_scores_accuracy	11060
review_scores_communication	11055
review_scores_cleanliness	11043
review_scores_rating	11022
reviews_per_month	10052
bathrooms	56
beds	40
bedrooms	22
host_listings_count	21
host_total_listings_count	21

```
## [1] 0
```

Zip code data integrity

48,372 zip codes had the correct length, whereas 523 zip codes had incorrect lengths.

zipcodes
with
correct
length

zipcode
48372

zipcodes
with
incorrect
length

zipcode
523

Filtering final data for merging

id <int>	host_id <int>	street <chr>	neighbourhood <chr>	neighbourhood_cleansed <chr>
2539	2787	Brooklyn , NY, United States	Brooklyn	Kensington
2595	2845	New York, NY, United States	Manhattan	Midtown
3647	4632	New York, NY, United States	Harlem	Harlem
3831	4869	Brooklyn, NY, United States	Brooklyn	Clinton Hill
5022	7192	New York, NY, United States	East Harlem	East Harlem
5099	7322	New York, NY, United States	Midtown East	Murray Hill

6 rows | 1-5 of 31 columns

```
## [1] 48895    31
```

Cleaning columns with dollar and comma sign

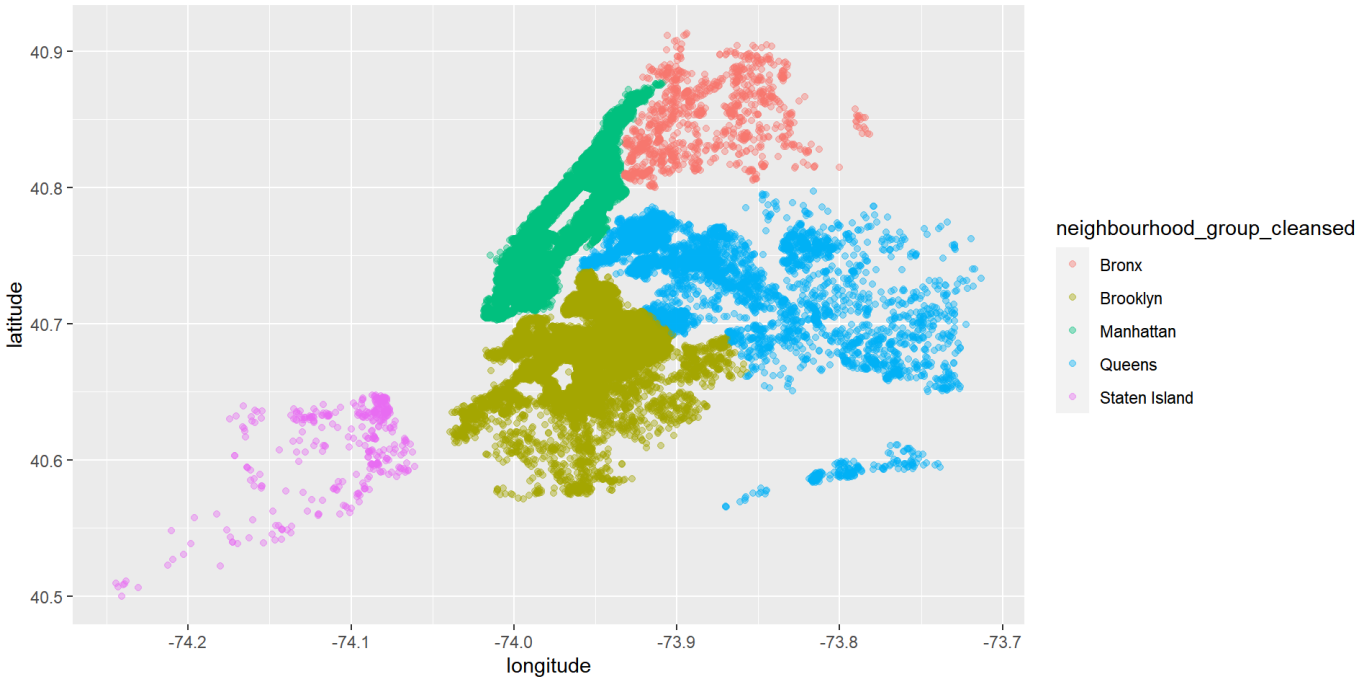
I modified colums with comma and dollar sign values and converted them to numeric format.

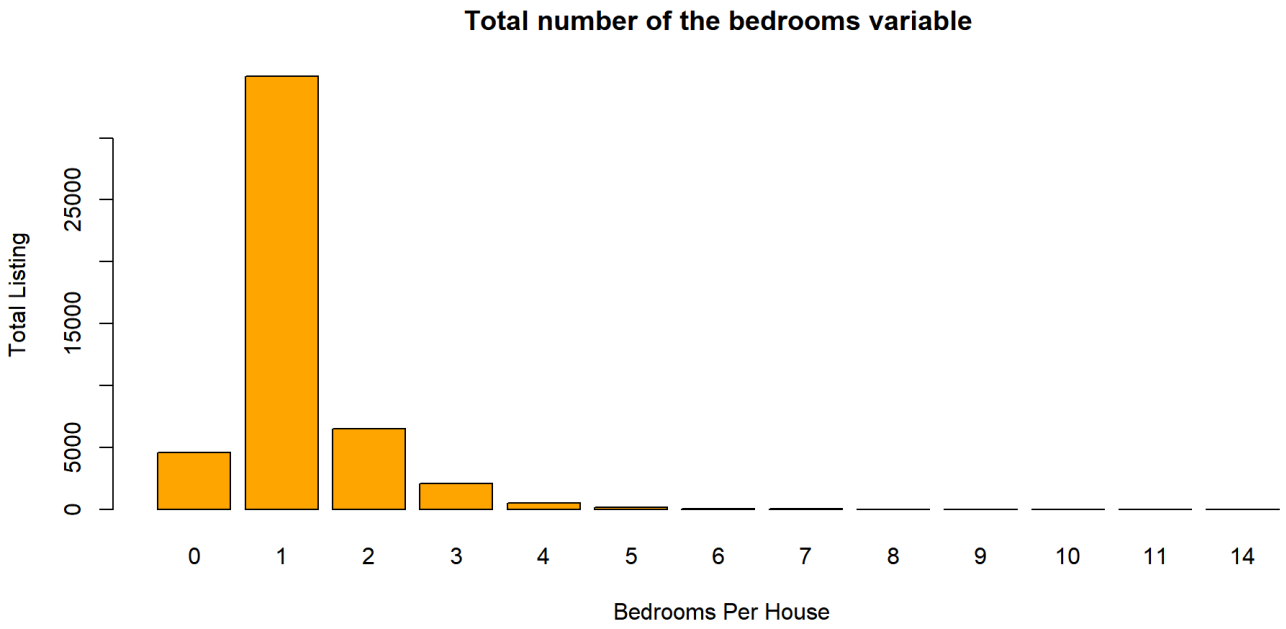
id <int>	host_id <int>	street <chr>	neighbourhood <chr>	neighbourhood_cleansed <chr>
2539	2787	Brooklyn , NY, United States	Brooklyn	Kensington
2595	2845	New York, NY, United States	Manhattan	Midtown
3647	4632	New York, NY, United States	Harlem	Harlem
3831	4869	Brooklyn, NY, United States	Brooklyn	Clinton Hill
5022	7192	New York, NY, United States	East Harlem	East Harlem
5099	7322	New York, NY, United States	Midtown East	Murray Hill

6 rows | 1-5 of 31 columns

Analyzing Longitude and Latitude Data

I mapped the point data of the original Airbnb dataset, grouped by several variables, including neighbourhood_group_cleansed , neighborhood, room type, and price per night.





Final Data Preprocessing

Data Join of Airbnb and Zillow data sets

	zipcode <chr>	id <int>	host_id <int>	street <chr>	neighbourhood <chr>
21	10003	9905142	15869370	New York, NY, United States	Manhattan
23	10003	14759888	6049738	New York, NY, United States	East Village
25	10003	30014439	35422741	New York, NY, United States	Manhattan
38	10003	34642119	4516135	New York, NY, United States	East Village
39	10003	20028846	142240474	New York, NY, United States	Flatiron District
49	10003	21783251	158725307	New York, NY, United States	Noho

6 rows | 1-6 of 34 columns

Checking columns with no data

```
colSums(is.na(final))
```

```
##                     zipcode                          id
##                           0                           0
##                     host_id                      street
##                           0                           0
##               neighbourhood      neighbourhood_cleansed
##                           0                           0
## neighbourhood_group_cleansed                        city
##                           0                           0
##                      market               smart_location
##                           0                           0
##                    latitude                   longitude
##                           0                           0
##                accommodates                   bathrooms
##                           0                           3
##                    bedrooms                        beds
##                           0                           0
##                       price                 weekly_price
##                           0                        1351
##                monthly_price             security_deposit
##                        1398                         420
##                 cleaning_fee       review_scores_location
##                         212                         389
##          review_scores_value              guests_included
##                         389                           0
##                extra_people              availability_30
##                           0                           0
##              availability_60              availability_90
##                           0                           0
##             availability_365          cancellation_policy
##                           0                           0
##                    room_type                      2017-06
##                           0                           0
##              predicted_price
##                           0
```

Missing Value Treatment Using Mice Package

```
#Converting Character Fields/Columns to Factors using 'mutate_if' function
na.model<-final%>%
  mutate_if(is.character, as.factor)%>%
  select(c(id, host_id, street, neighbourhood, neighbourhood_cleansed, neighbourhood_group_cleansed,   city,   zipcode,    m
arket, smart_location, latitude,   longitude, accommodates, bathrooms, bedrooms, beds, price, weekly_price, monthly_price, s
ecurity_deposit,  cleaning_fee, review_scores_location,review_scores_value,  guests_included, extra_people, availability_30,
availability_60, availability_90, availability_365, cancellation_policy,predicted_price ))

set.seed(123)
miceMod <- mice(na.model, method='cart',m=1,maxit=1)  # perform mice imputation, based on CART/Decision Tree Algorithm. Numb
er of iterations here is one
```

```
##
##  iter imp variable
##   1   1  bathrooms  weekly_price  monthly_price  security_deposit  cleaning_fee  review_scores_location  review_scores_va
lue
```

```
final_clean <- complete(miceMod) #New Data is generated with imputed values for missing observations

#Checking for missing values
anyNA(final_clean)
```

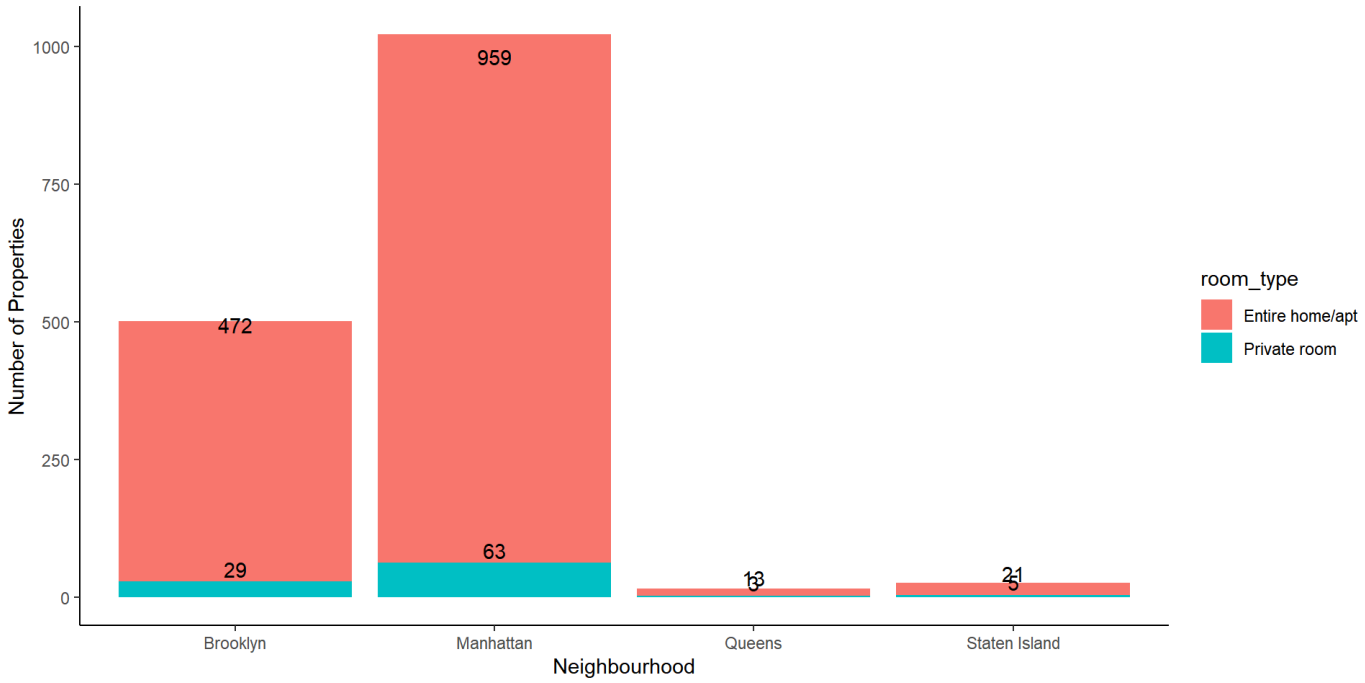
```
## [1] FALSE
```

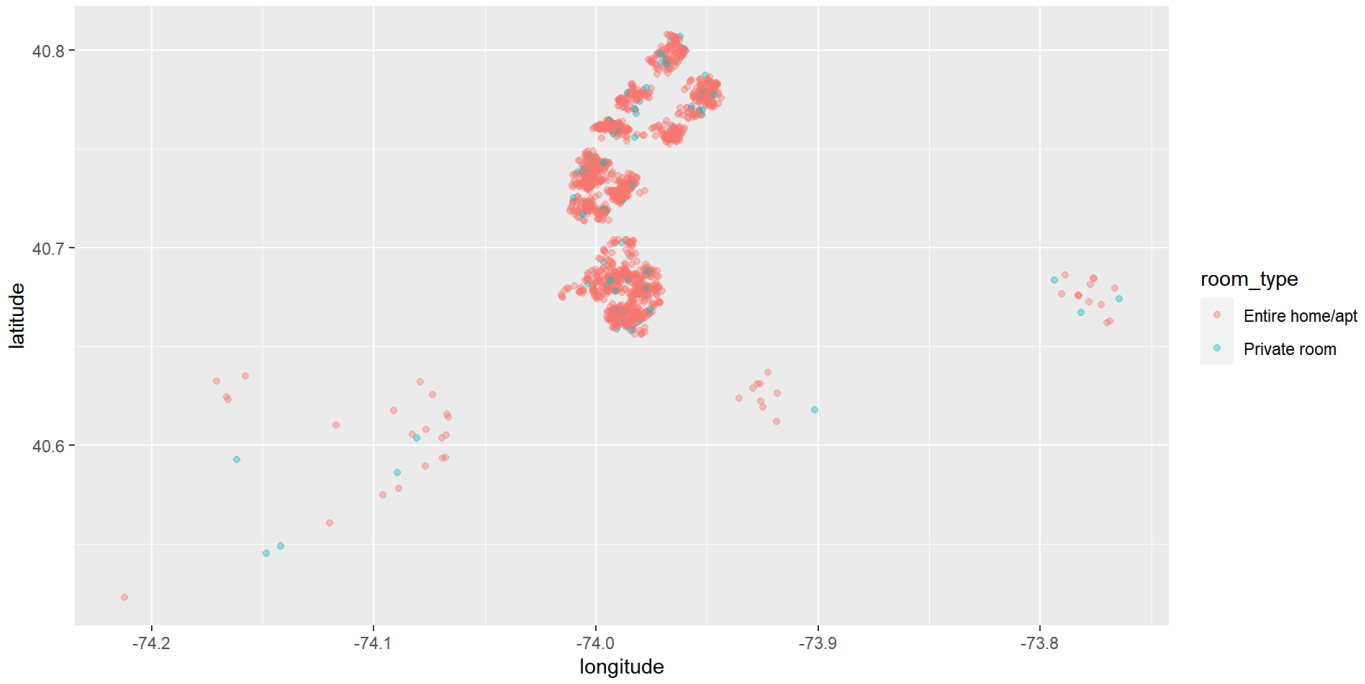
Exploratory Data Analaysis

Univariate Data Analysis

Visualizing the number of properties listed by room type

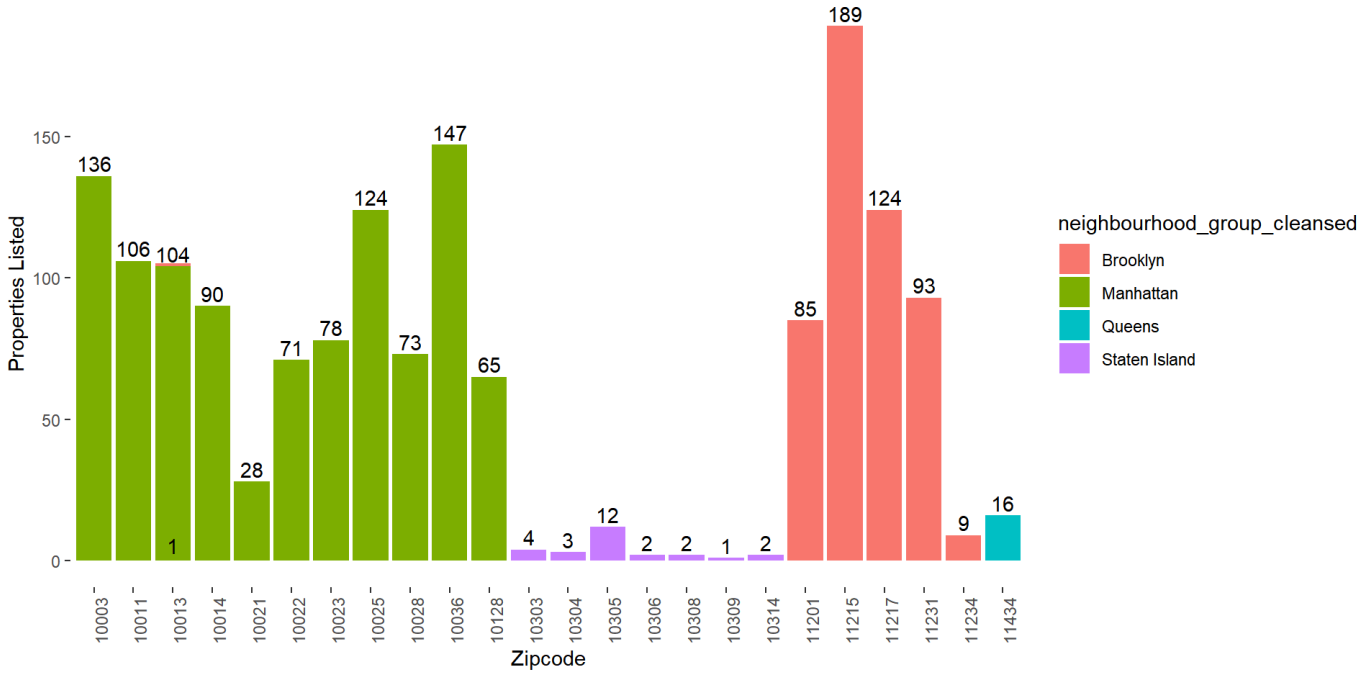
Manhattan has the most listings, followed by Brooklyn. Most of the listings are either “Entire home/apt” or “Private room”.





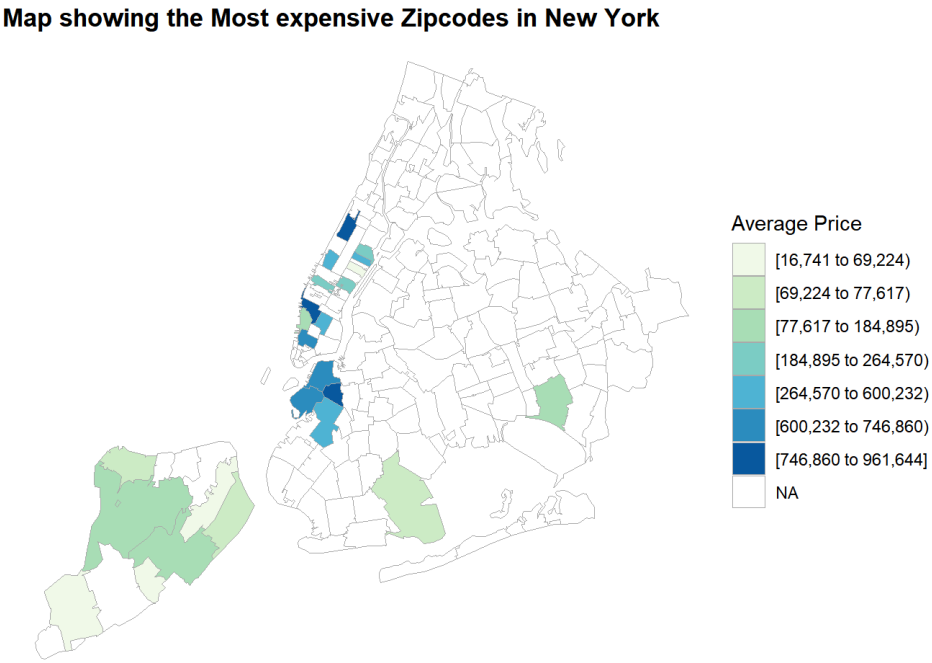
Visualizing Number of Property Listing by Zipcode

Zipcodes 10003, 10011,10013,10014,10025,10036,11201,11215,11217, and 11231 have the highest number of properties listed.



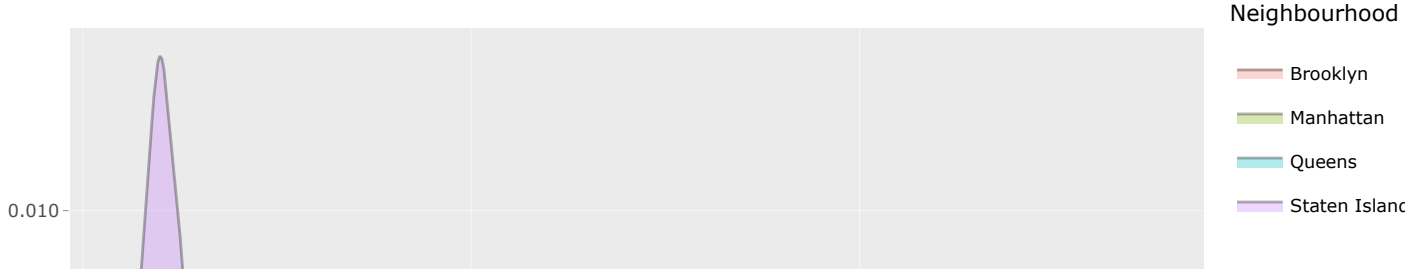
Which Neigbourhood is the most Expensive

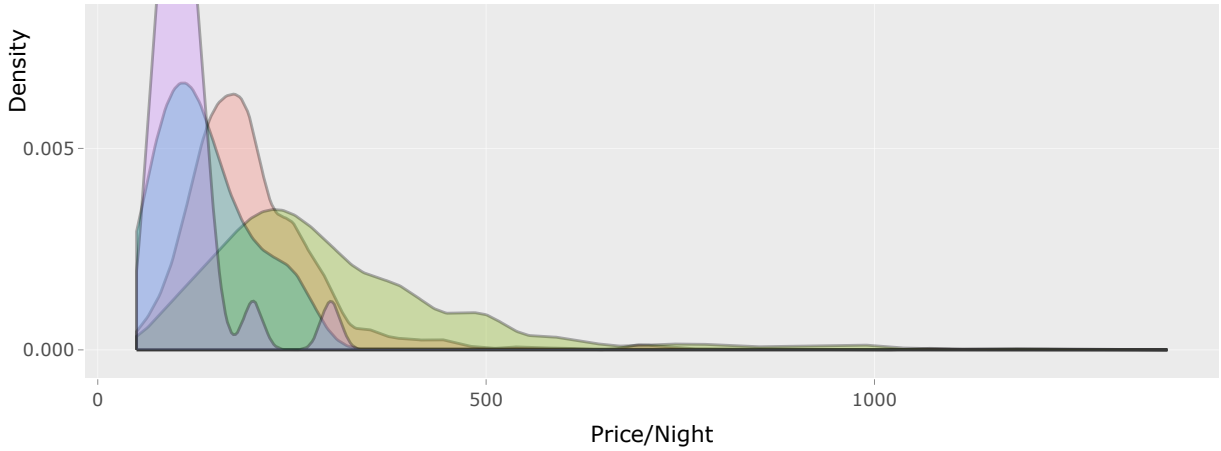
Manhattan and Brooklyn neighbourhoods are the most expensive.



Visualizing Density of Property Listing vs Price

The spread of prce per neighbourhood. + Manhattan has the widest price spread. + Brooklyn has almost a perfect bell curve, which is a characterstic of normal real-world behaviour. + Staten Island and Queens have narrower distribution due to small sample size.

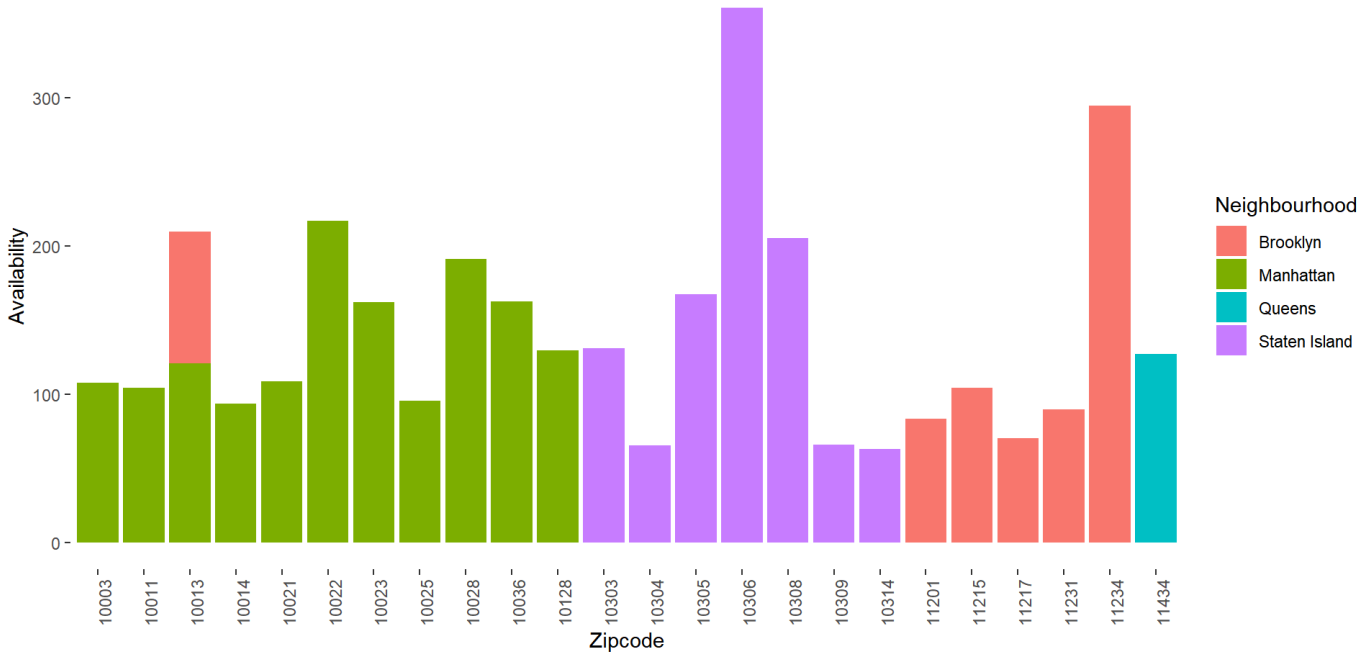




Bivariate Data Analysis

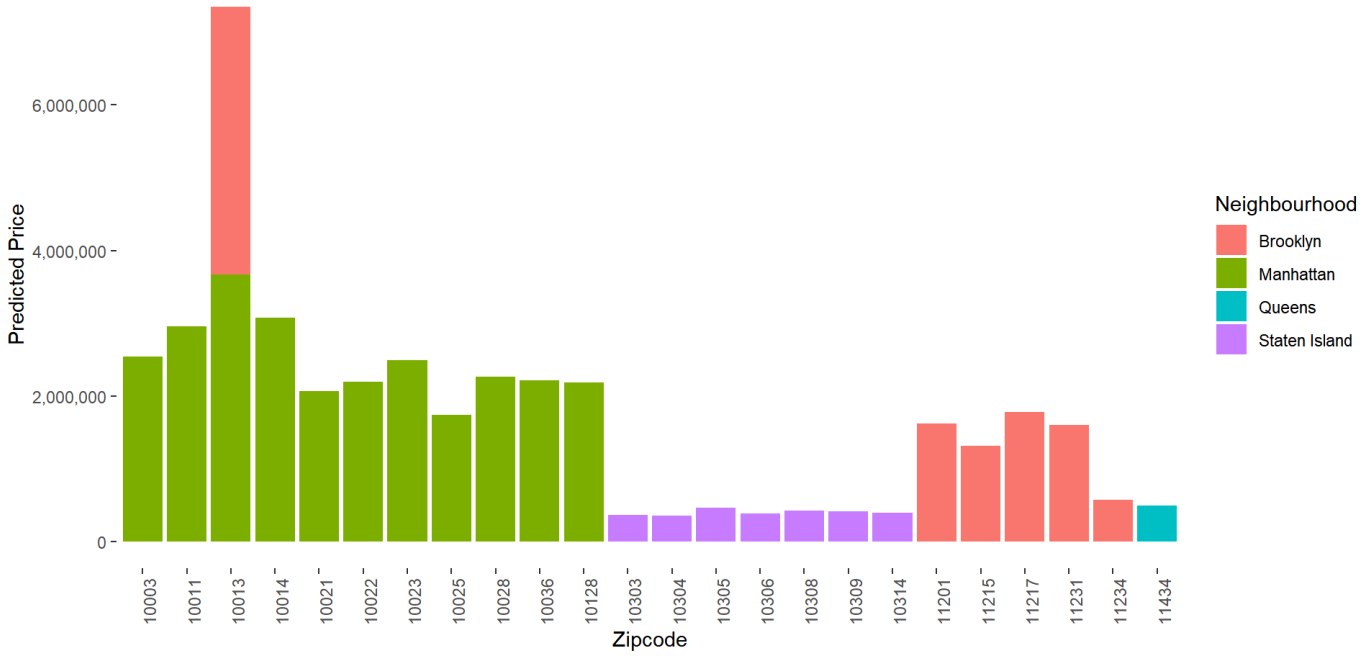
Visualizing Property Availiability In Each Zipcode

```
final %>% group_by(neighbourhood_group_cleansed,zipcode) %>% summarise_all(funs(mean)) %>% ggplot(aes(x =zipcode,y= availabi
lity_365, fill = neighbourhood_group_cleansed )) + geom_bar(stat = "identity") + scale_y_continuous(labels = scales::comma)
 +  scale_colour_brewer(palette = "Pastel2") + labs( y = "Availability", x = "Zipcode") + theme_bw() + theme(plot.background
= element_blank(),panel.grid.major = element_blank(),panel.grid.minor = element_blank(),panel.border = element_blank(),axis.
text.x = element_text(angle = 90, hjust = 1)) + guides(fill = guide_legend(title = "Neighbourhood"))
```



Visualizing Property Predicted in Each Zipcode for February 2021

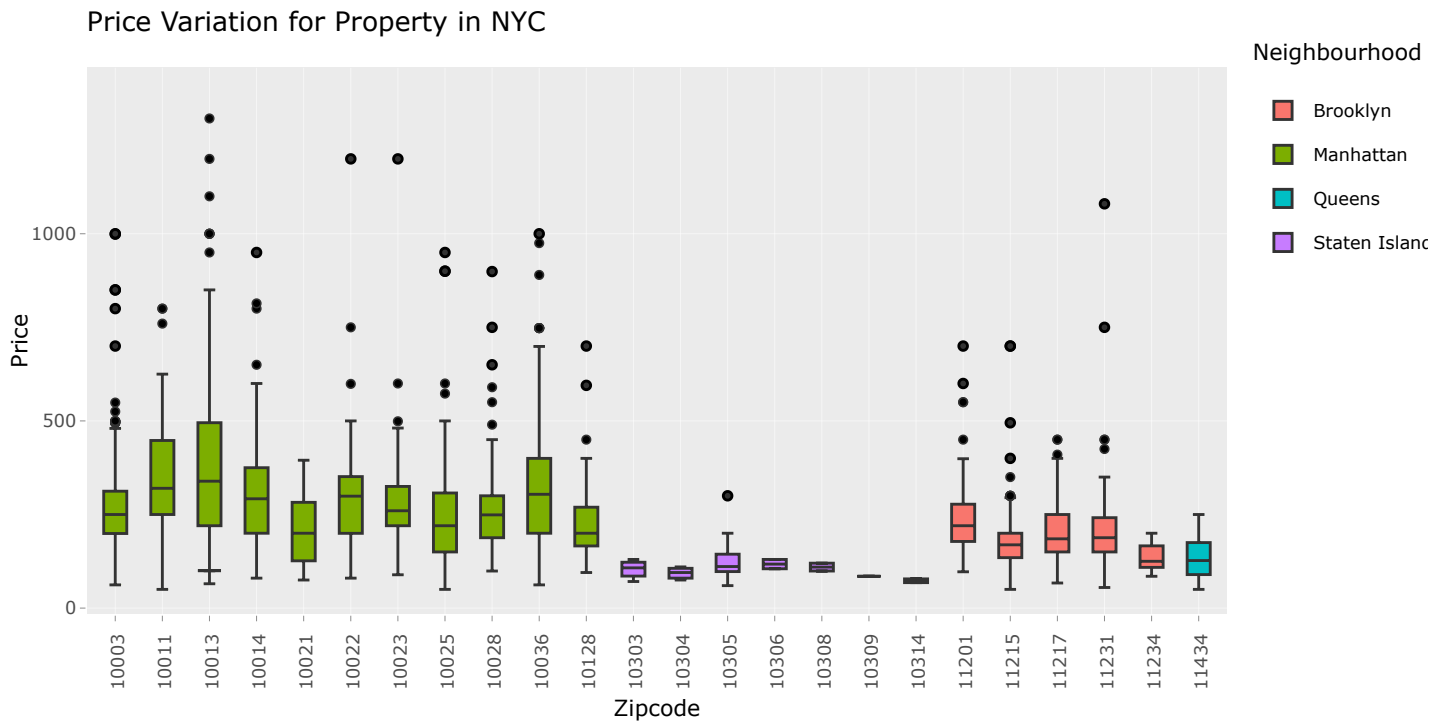
```
final %>% group_by(neighbourhood_group_cleansed,zipcode) %>% summarise_all(funs(mean)) %>% ggplot(aes(x =zipcode,y= predicte
d_price, fill = neighbourhood_group_cleansed )) + geom_bar(stat = "identity") + scale_y_continuous(labels = scales::comma) +
scale_colour_brewer(palette = "Pastel2") + labs( y = "Predicted Price", x = "Zipcode") + theme_bw() + theme(plot.background
 = element_blank(),panel.grid.major = element_blank(),panel.grid.minor = element_blank(),panel.border = element_blank(),axi
s.text.x = element_text(angle = 90, hjust = 1)) + guides(fill = guide_legend(title = "Neighbourhood"))
```



Visualizing Price Variation for Property by Zipcode

The top 10 zipcodes which fetch highest price per night are: 10011,10013,10014,10022,10023,10028,10036,11201,11217, and 11231.

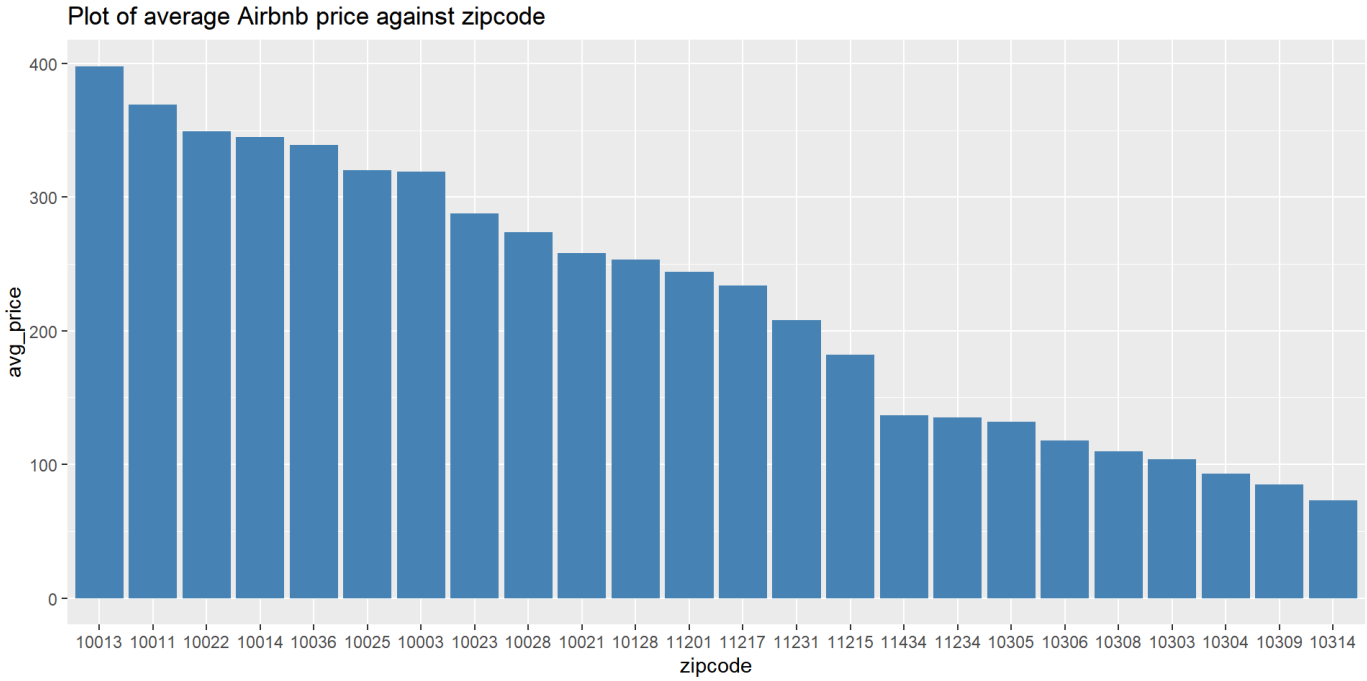
```
g<-ggplot(final, aes(x = zipcode,y = price, fill = neighbourhood_group_cleansed)) + geom_boxplot() + scale_y_continuous(limi
ts = quantile(final$price, c(0, 0.99))) + labs(x = "Zipcode", y = "Price") + theme(axis.text.x = element_text(angle = 90, hj
ust = 1))  +ggtitle("Price Variation for Property in NYC")+guides(fill = guide_legend(title = "Neighbourhood"))
ggplotly(g)
```



Visualizing Average Price for Property by Zipcode

The average price decreases from 380 to 70.

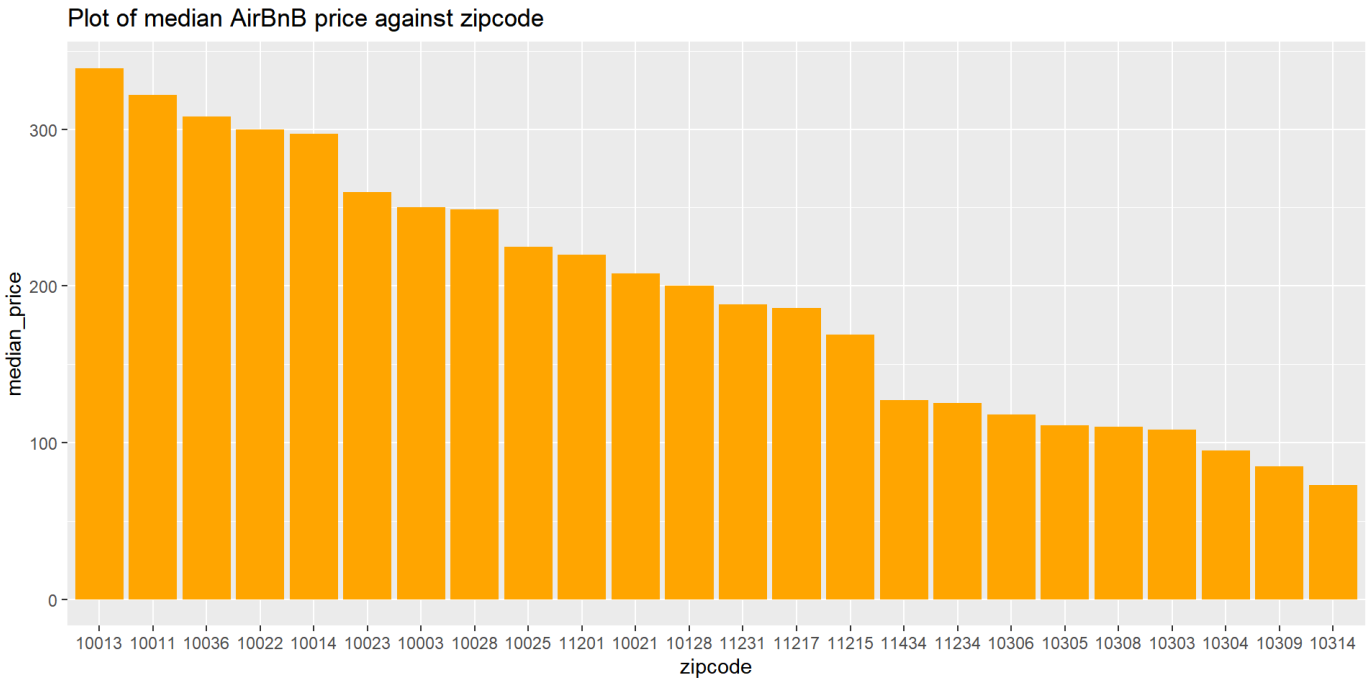
```
final %>% select(zipcode,price) %>% filter(zipcode>0)%>% group_by(zipcode)%>% summarise(avg_price=mean(price,na.rm = T),coun
t=n()) %>%arrange(desc(avg_price),count)%>%mutate_if(is.numeric,round,digits=0)%>% top_n(n = 25)%>% ggplot(.,mapping = aes(r
eorder(zipcode,-avg_price),avg_price))+geom_bar(stat = "identity",fill='steelblue')+ggtitle("Plot of average Airbnb price ag
ainst zipcode")+xlab('zipcode')
```



Visualizing Median Price for Property by Zipcode

Most prices are between $100 and $300. The median price is decreasing from $350 to $70.

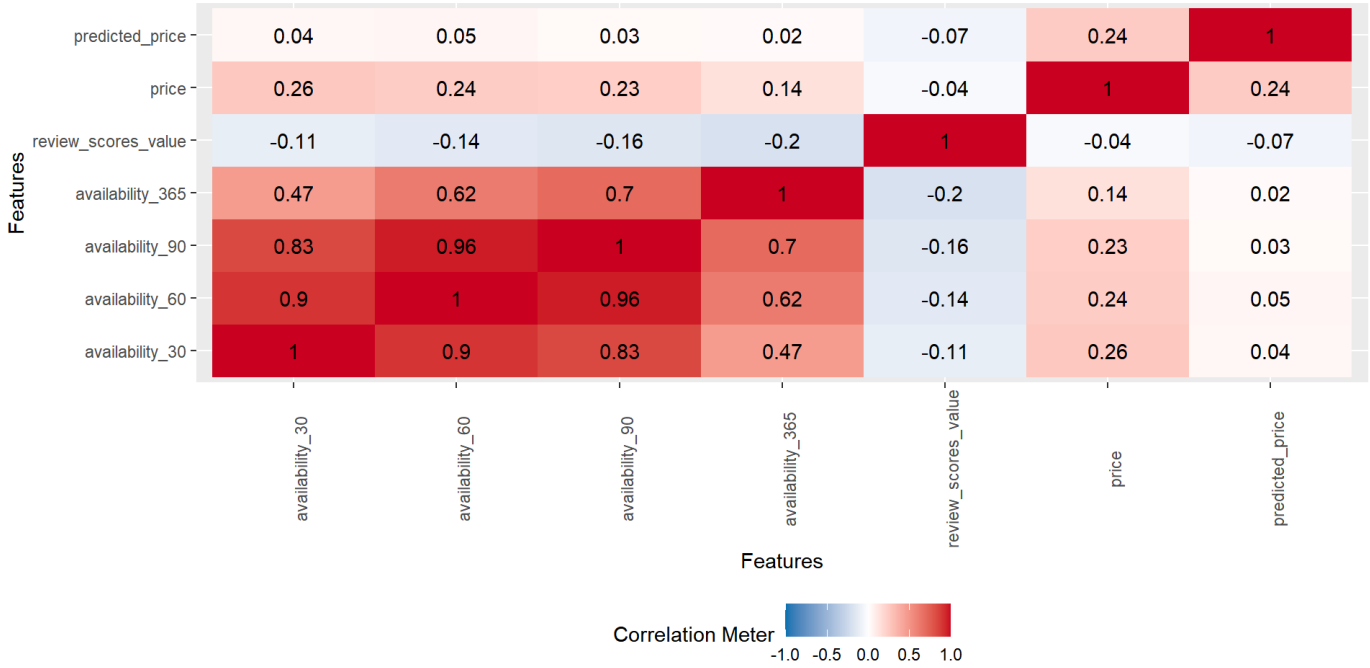
```
final %>% select(zipcode,price) %>% filter(zipcode>0)%>% group_by(zipcode)%>% summarise(median_price=median(price,na.rm = T
),count=n()) %>%arrange(desc(median_price),desc(count))%>%mutate_if(is.numeric,round,digits=0)%>% top_n(n = 25)%>% ggplot(.,
mapping = aes(reorder(zipcode,-median_price),median_price))+geom_bar(stat = "identity",fill="orange")+ggtitle("Plot of media
n AirBnB price against zipcode")+xlab("zipcode")
```



From our observation above, the averages prices are higher than median prices because of outliers.

Correlation Plot of Price, Reviews, Availability, and Predicted Price

```
corCols <- c("availability_30","availability_60","availability_90","availability_365","review_scores_value","price","predicted_price")
plot_correlation(final_clean[corCols])
```



Model Bulding and Evaluation

The model that we are going to cosnider while looking for a suitable propety to invest will be based on three approaches.

Investing Approaches

1. Optimal Revenue Model: In this approach, I am assuming that property will be sold after few years. Both revenue and real price, which is determine by the real estate appreciation cost of the property, is being considered. The discount factor has been assumed to be 0%.
2. Break Even Model: In this approach, we consider that property is going to be held indefinitely, and so the focus will only be on the revenues generated and the initial outlay. Discount factor will again be 0%. Our approach here is to calculate the number of years it will take for the property to break even.
3. Mean Return Rate Model: In this approach, We are assuming that the expected return from a Zip will be the average of the revenue generated by all the listings in that Zip. We will then compare this average value of revenue with the property prices given to get mean annual return rate.

Calculating Annual Returns and Annual Rate of Returns

```
#Occupancy Rates based on Rating and Location conditions
#Generally good Neighbourhoods and well Rated accomodations will have better Occupancy Rates
final_clean<-final_clean%>%
  mutate(OccupancyRate=ifelse(review_scores_value >=9 &
                                (neighbourhood_group_cleansed=="Manhattan"|
                                   neighbourhood_group_cleansed=="Brooklyn"),0.75,
                              ifelse(review_scores_value >=9&
                                neighbourhood_group_cleansed=="Queens",0.70,ifelse((review_scores_value >=8  &review_scores_value<9)&
                                   (neighbourhood_group_cleansed=="Manhattan"|
                                      neighbourhood_group_cleansed=="Brooklyn"),0.65,0.55))))
```

```
## [1] 1330
```

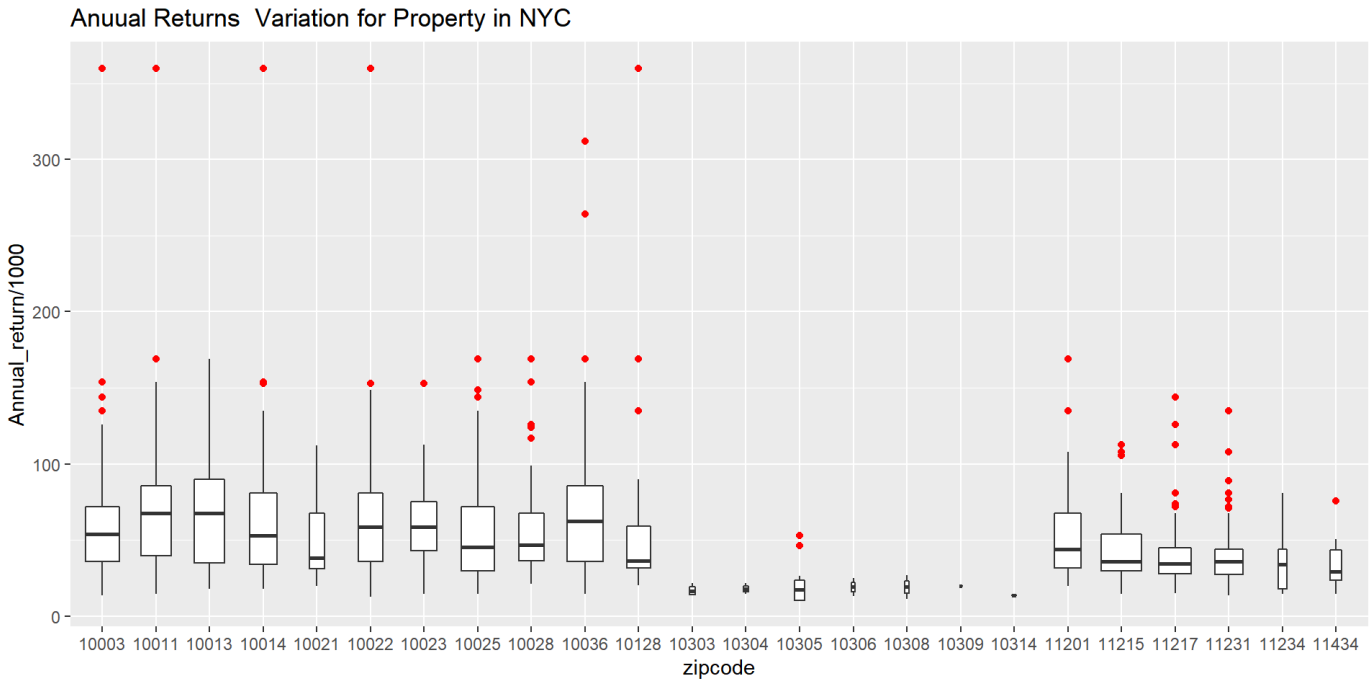
```
final_clean<-final_clean%>%
  mutate(Annual_return=ifelse(is.na(monthly_price),365*price*OccupancyRate,
                              12*monthly_price*OccupancyRate))%>%
  mutate(Annual_Return_Rate=Annual_return/predicted_price)
```

id	host_id	street	neighbourhood	neighbourhood_cleansed	neighbourhood_group_cleansed	city	zipcode
9905142	15869370	New York, NY, United States	Manhattan	East Village	Manhattan	New York	10003
14759888	6049738	New York, NY, United States	East Village	East Village	Manhattan	New York	10003
30014439	35422741	New York, NY, United States	Manhattan	East Village	Manhattan	New York	10003

```
## [1] 1565   34
```

Visualizing Annual Return Variation for Property in NYC

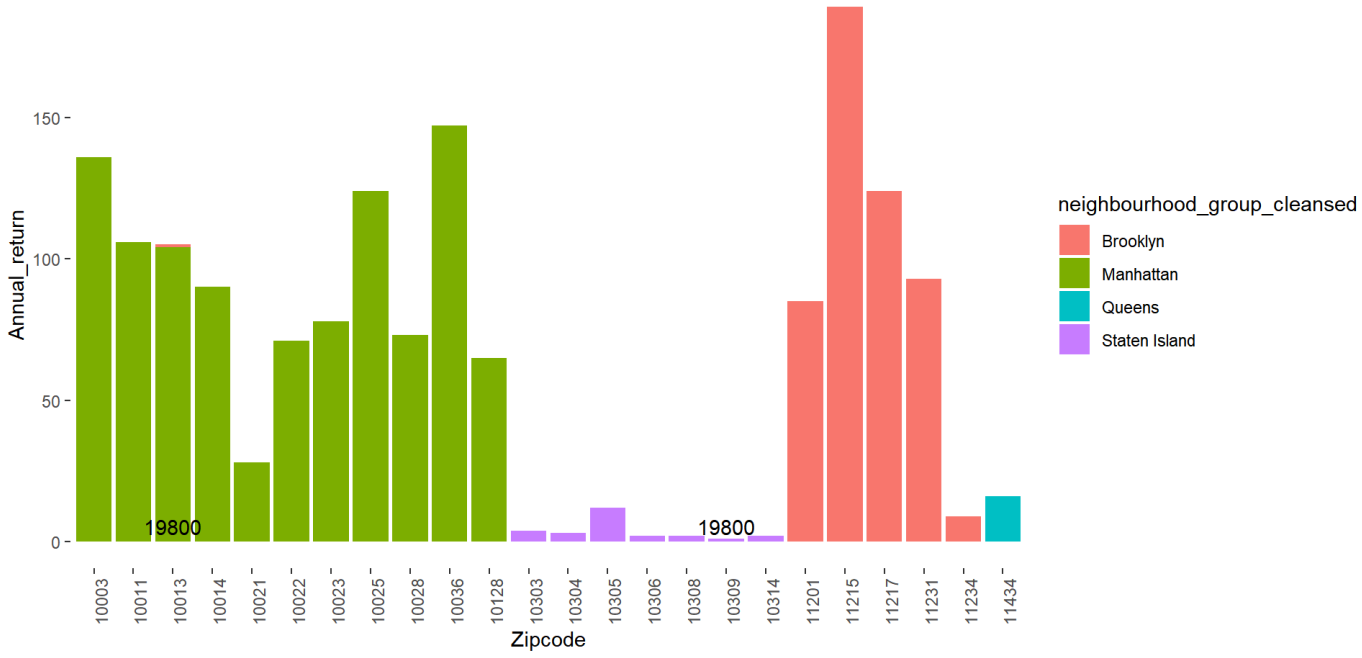
```
# boxplot to show variations in price within a zipcode
ggplot(data=final_clean,mapping = aes(zipcode,Annual_return/1000))+
geom_boxplot(outlier.colour = "Red",varwidth = TRUE)+ggtitle("Anuual Returns  Variation for Property in NYC")
```



Visualizing Annual Return for Property in each ZipCode

From the two conclusions above and the third table, we can also infer that the number of properties in Brooklyn and Manhattan is pretty high compared to the others. Staten Island and Bronx have less than 1000 properties

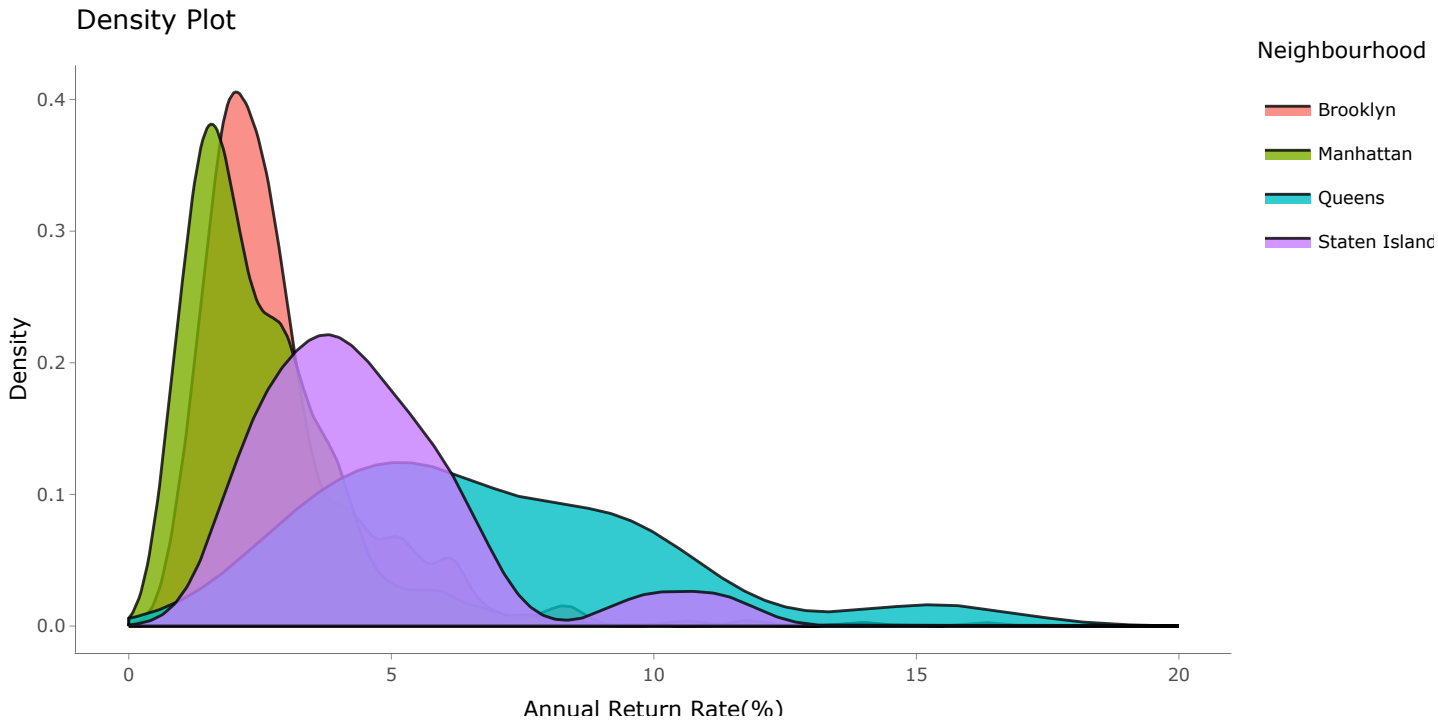
```
ggplot(final_clean, aes(zipcode, fill = neighbourhood_group_cleansed)) + geom_bar() + labs(x = "Zipcode", y = "Annual_retur
n") + geom_text(stat='count', aes(label=Annual_return), vjust= -0.3) + theme(axis.text.x = element_text(angle = 90, hjust =
1))  + theme_bw() + theme(plot.background = element_blank(),panel.grid.major = element_blank(),panel.grid.minor = element_bl
ank(),panel.border = element_blank(),axis.text.x = element_text(angle = 90, hjust = 1))
```



Visualizing Annual Return for Property in each Neigbourhood

The Neighbourhood Queens has very high return on investments, between 10% and 20% for many zip codes. 11109 is the zip code with the highest return rate, 19%.

```
## usings plptly to create interactive plot
theme_set(theme_classic())
g <- ggplot(data=final_clean, aes(Annual_Return_Rate*100))
g.g<-g + geom_density(aes(fill=factor(neighbourhood_group_cleansed)), alpha=0.8) +
  labs(title="Density Plot",
       subtitle="Annual Return Rate Grouped by Neighbourhood",
       caption="Source: Airbnb Data",
       x="Annual Return Rate(%)",
       y="Density",
       fill="Neighbourhood ")+scale_x_continuous(limits = c(0, 20))
p <- ggplotly(g.g)
p
```



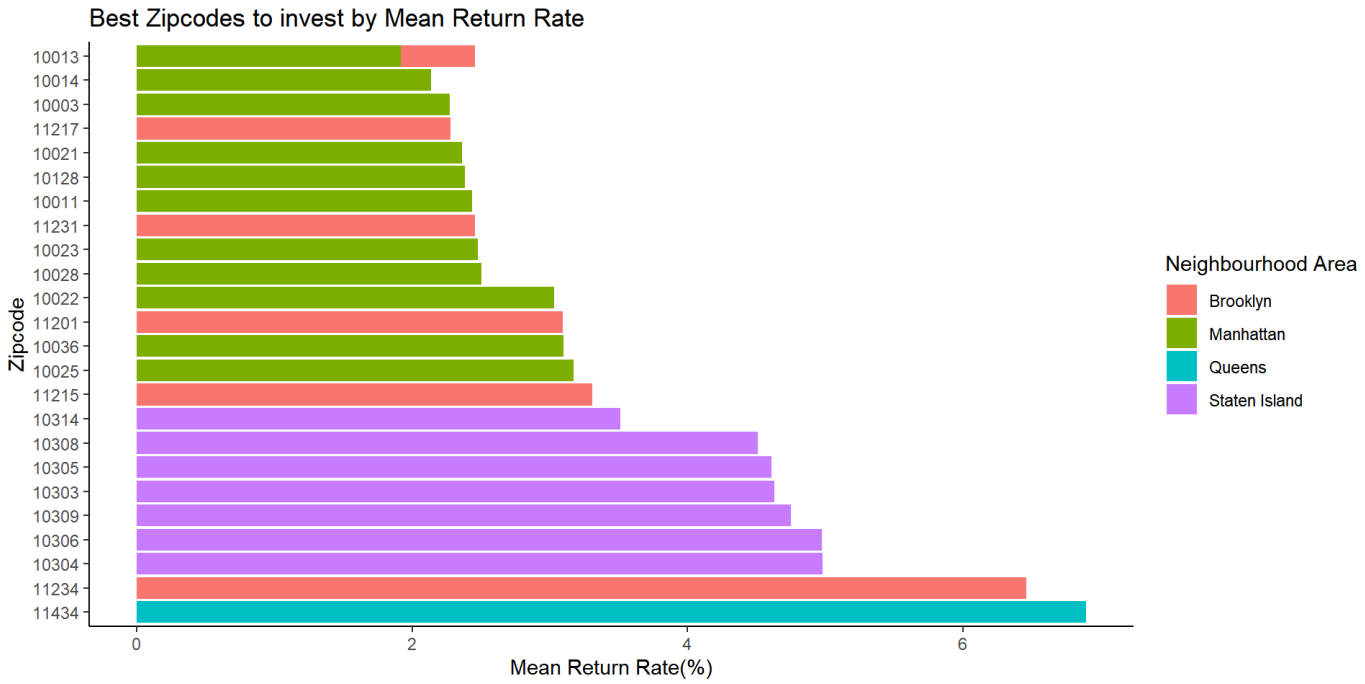
Mean Return Rate Model:Best zipcodes to Invest by Mean Return Rate

Mean property price in Manhattan and Brooklyn exceeds the mean property price in other neighborhoods. This is quite predictable, as these are busy cities, and there is heavy competition among real estate agents. Also the presence of corporate offices and buildings in that area boosts the property prices.

```
final_clean4<-final_clean%>%
  filter(!is.na(zipcode))%>%
  select(zipcode,neighbourhood_group_cleansed,Annual_Return_Rate )%>%
  group_by(zipcode,neighbourhood_group_cleansed)%>%
  summarise(Mean_Return_Rate=mean(Annual_Return_Rate,na.rm=T))%>%
  arrange(Mean_Return_Rate)

final_clean4<-final_clean4%>%
  select(zipcode,neighbourhood_group_cleansed,Mean_Return_Rate)%>%
  head(25)

ggplot(data=final_clean4,aes(x=reorder(zipcode,-Mean_Return_Rate),y=Mean_Return_Rate*100,fill=neighbourhood_group_cleansed))
+
  geom_bar(stat="identity")+
  #scale_fill_manual(values=c("red", "yellow","dark green","blue","grey"))+
  coord_flip()+ggtitle("Best Zipcodes to invest by Mean Return Rate")+xlab("Zipcode") +ylab("Mean Return Rate(%)") + labs(fi
ll = "Neighbourhood Area")
```

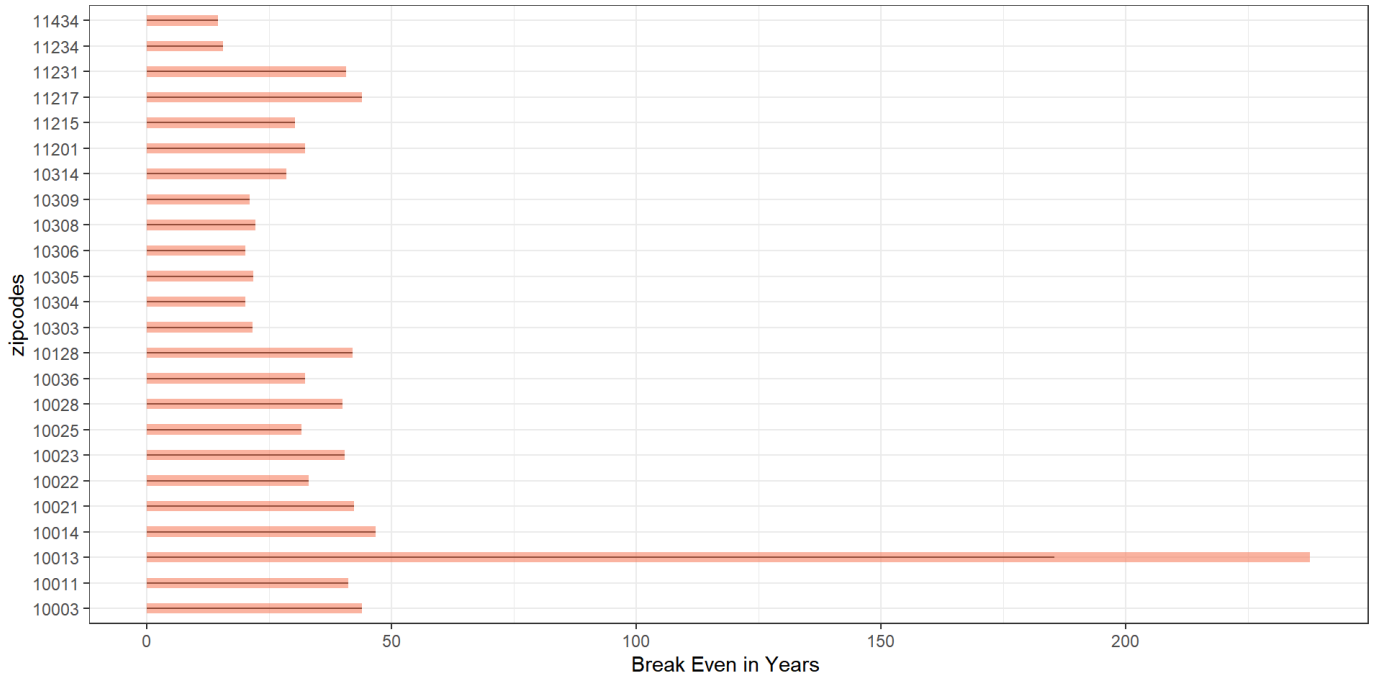


Break Even Model: Calculating The Break Even Period

Visulaizing the Break Even Period in Years

The Break even period is calculated based on the mean predicted price of each zipcode. From the chart below,we can see that median amount of time taken by a zipcode to breakeven is between 30 and 50 years, not taking into account property appretion price.

```
final_clean5 %>%
  arrange(break_even) %>%    # First sort by val. This sort the dataframe but NOT the factor levels
  mutate(zipcodes=factor(zipcode)) %>%   # This trick update the factor levels
  ggplot( aes(x=zipcodes, y=break_even)) +
    geom_segment( aes(xend=zipcodes, yend=0)) +
    geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
    coord_flip() +
    theme_bw() +
    ylab("Break Even in Years ")
```



Conclusion

Using different approaches discussed in our model building segment, the best zipcodes to invest in are:

1. Optimal Revenue Model: 11231 ,11217 ,11215, 10036,10025 ,10003 ,10011 ,10128
2. Break Even Model: 11434,10304,10309, 11234
3. Mean Return Rate Model: 11434, 11234, 11304

Future Steps

- Expand this analysis to multiple cities with compare patterns and trends between different cities. From the insights that have been derived, I would also like to build predictive models using different features from the dataset.
- Merge this data with Covid-19 data to see how the pandemic would affect the demand and supply of airbnb rentals in the next few years in New York.
- Introduce seasonality and weather data to understand trends in occupancy rates throughout the year and create a model to predict occupancy rate based on this insight.
- Use NLP models to analyze the qualitative part of Airbnab data. We can account for the trends in reviews and get sentimental insights from word cloud that drive demand and occupancy rates.

References

1. Exploratory Data Analysis of NYC Airbnb Listing Info and Demographics https://xukeren.rbind.io/post/2019/12/31/exploratory-data-analysis-of-nyc-airbnb-listing-info-and-demographics/ (https://xukeren.rbind.io/post/2019/12/31/exploratory-data-analysis-of-nyc-airbnb-listing-info-and-demographics/)
2. Create a choropleth of US Zip Codes https://arilamstein.com/documentation/choroplethrZip/reference/zip_choropleth.html#examples (https://arilamstein.com/documentation/choroplethrZip/reference/zip_choropleth.html#examples)

3. Analysis of Airbnb Data in NYC 2019 https://web.stanford.edu/~kjytay/courses/stats32-aut2019/Session%208/Airbnb_analysis.html (https://web.stanford.edu/~kjytay/courses/stats32-aut2019/Session%208/Airbnb_analysis.html)