Airbnb Data Analysis - Chicago

Data Mining and Predictive Analysis

Project Title : Explanatory and Predictive Data Analysis: Airbnb Chicago

"We, the undersigned, certify that the report submitted is our original work; all authors participated in the work substantively; all authors have seen and approved the report as submitted; the text, images, illustrations, and other items included in the manuscript do not carry any infringement/plagiarism issue upon any existing copyrighted materials."

Member No.	Team Member
Team Member 1	Ajay Iyer
Team Member 2	Kavya Purushothaman
Team Member 3	Mihir Mohite
Team Member 4	Rohan Chaudhari
Team Member 5	Shruthi Nair

Executive Summary

This project analyzes the Airbnb data to understand the factors affecting high booking rates of the properties in Chicago to guide a potential investor in choosing a property with a high booking rate. The analysis also helps the existing property owners to make necessary modifications to their properties to get a higher booking rate.

One of the key findings of our exploratory analysis is the importance of location. It shows the North and West regions have the highest booking rate and the Southern region has the lowest booking rate. We also found that the ease of access to public transportation had a considerable impact on the booking rate. Moreover, an interesting observation was that property being near the coast has no significant effect on the booking rate, even though Chicago is a coastal city. We also found properties that were described as being more family-friendly and quiet surroundings had better chances of getting high booking rates. We narrowed down on these findings by first doing exploratory data analysis and then by testing the accuracy of our predictions after incorporating each of these factors. Importance was given to avoid misclassifying a property having a low booking rate as the one having a high booking rate. The model parameters were adjusted such that it would lead to the lowest number of misclassifications.

Research Questions

Our primary focus was to help a potential investor choose a property with a high booking rate. To shortlist the exact questions, we started with market analysis.

1. What factors should a potential investor consider before buying a property in Chicago that he/she/they wish to list on Airbnb?
- Is location an important factor? Which areas within Chicago have the highest booking rates?
- Does connectivity to major hubs/accessibility transport impact booking rates of listings?
- Are there any underlying trends in guest behavior? What kind of crowd usually stays at an Airbnb in Chicago?
- How has the Chicago Airbnb market been growing over time? Are there any specific trends?
- Are there any specific property types that are more popular and likely to have high booking rates?
- How should a property be priced? Is there a lot of variation in the prices between regions and how to decide the ideal price of a property?

2. Are there any amenities that have an impact on the high booking rate?
- Which are some of the basic amenities which guests expect to have, irrespective of the price?
- Are there any amenities that could be game-changers?

Methodology

Market Research:

Some key findings from market research that helped us narrow down our research questions have been listed below.

- Location - Chicago has been divided into well-defined neighborhoods, which form communities, which are further grouped into broader sides. The Central side including neighborhoods such as Lincoln Park and a few others have most of the important tourist spots. We expect these areas to have high booking rates.
- Crime - Chicago has had high crime rates over the years. With a violent crime rate of 1006.1 per 100000 people in 2018, we were curious about the impact of this on the booking rate of listings.
- Connectivity - Chicago ranks 6th among all US states when it comes to public transportation. We expect properties that are nearby public transport to have higher booking rates. We would also like to explore whether guests still prefer having private transport facilities.
- Regulations - Chicago has very strict laws and regulations when it comes to shared rentals. We would be exploring the impact of this on booking rates - whether guests consider this before booking a property.
- Longer stays - We found that these above regulations were implemented because most owners listed properties for a minimum stay of 30 days, thus competing with other small businesses, endangering residential neighborhoods, etc. We would like to explore the relationship between the minimum stay period and the high booking rate.

- Tourists - Being a coastal region, we expect most guests to be tourists. Thus, we would be exploring the impact of proximity to beaches/lakes as well as whether too restrictive house rules (no smoking/partying, etc) have an impact on booking rates.
- Multi-listings - We found that most property owners have multiple listings in Chicago. We would like to analyze whether owning multiple properties has an impact on response time as well as booking rate.

Data processing:

Data Cleaning:

- Standardization - Removed symbols and converted data to relevant data types.
- Missing Values - Handled N/A values by mean imputation & zero imputation based on variable types.
- Mean Imputations: We replaced NAs in the following variables by mean: *Price,availability_30, availability_90, availability_365, accommodates*. The reason we replaced the missing values with mean in this case is that there weren't a large number of NA values, hence, this wouldn't take a toll on the variability of the data. Apart from that, it also makes sense as using the median would just skew the whole data and hence affect the analysis we do further.
- Zero Imputations: We replaced NAs in the following variables by 0: *Extra_people, cleaning_fee*. The reason we replaced these with zero is that, there would be value in here if something extra would be charged from the customers, if not, then it obviously is zero. Hence, logically, if a value is not available, zero is the perfect option.

Data Tranformation

```
airChicago$security_deposit<- as.numeric(gsub("\\$", "",
airChicago$security_deposit))

## Warning: NAs introduced by coercion

airChicago$price <- as.numeric(gsub("\\$", "", airChicago$price))

## Warning: NAs introduced by coercion

airChicago$cleaning_fee <- as.numeric(gsub("[$]", "",
airChicago$cleaning_fee))

## Warning: NAs introduced by coercion

airChicago$extra_people <- as.numeric(gsub("\\$", "",
airChicago$extra_people))
airChicago$bathrooms <- as.numeric(airChicago$bathrooms)
airChicago$beds <- as.numeric(airChicago$beds)
airChicago$accomodates <- as.numeric(airChicago$accommodates)
airChicago$availability_30 <- as.numeric(airChicago$availability_30)
airChicago$availability_60 <- as.numeric(airChicago$availability_60)
airChicago$availability_90 <- as.numeric(airChicago$availability_90)
```

```
airChicago$availability_365 <- as.numeric(airChicago$availability_365)
airChicago$guests_included <- as.numeric(airChicago$guests_included)
airChicago$host_response_rate <- as.numeric(gsub("\\%", '',
airChicago$host_response_rate))
```

Data Imputation

```
airChicago$review_scores_rating[is.na(airChicago$review_scores_rating)] <- 95
airChicago$review_scores_value[is.na(airChicago$review_scores_value)] <- 10
airChicago$security_deposit[is.na(airChicago$security_deposit)] <- 0
airChicago$price[is.na(airChicago$price)] <- 142
airChicago$beds[is.na(airChicago$beds)] <- 2
airChicago$bedrooms[is.na(airChicago$bedrooms)] <- 2
airChicago$host_response_rate[is.na(airChicago$host_response_rate)] <- 97
airChicago$host_response_time[is.na(airChicago$host_response_time)] <-
'within an hour'
airChicago$extra_people[is.na(airChicago$extra_people)] <- 0
airChicago$cleaning_fee[is.na(airChicago$cleaning_fee)] <- 0
airChicago$review_scores_rating[is.na(airChicago$review_scores_rating)] <- 0
airChicago$accommodates[is.na(airChicago$accommodates)] <- 4
airChicago$availability_30[is.na(airChicago$availability_30)] <- 16
airChicago$availability_90[is.na(airChicago$availability_90)] <- 57
airChicago$availability_365[is.na(airChicago$availability_365)] <- 177
```

Exploratory Analysis:

```
airChicago_transit1 <- airChicago1 %>%
  filter(high_booking_rate == 1)
```

We created a wordcloud to show the various **modes of transportation** the airbnbs having a high booking rate.

```
wordcloud(d5$word, d5$freq, random.order = FALSE, rot.per = 0.3, scale = c(4,
.5), max.words = 50, colors = brewer.pal(8, "Dark2"))
```

Observations:

- Presence of **"bus", "station", "line", "cta", "red/brown" in "Transit"** for high booking rate listings.
- According to Chicago Tribune, Chicago's public transit ranks 6th among all US cities.

```
airChicago_hbr1 <- airChicago1 %>%
  filter(high_booking_rate == 1)
```

We created a wordcloud to show the various **amenities** in the airbnbs having a high booking rate.

```
wordcloud(d1$word, d1$freq, random.order = FALSE, rot.per = 0.3, scale =
c(1.5, .5), max.words = 50, colors = brewer.pal(8, "Dark2"))
```

Observations:

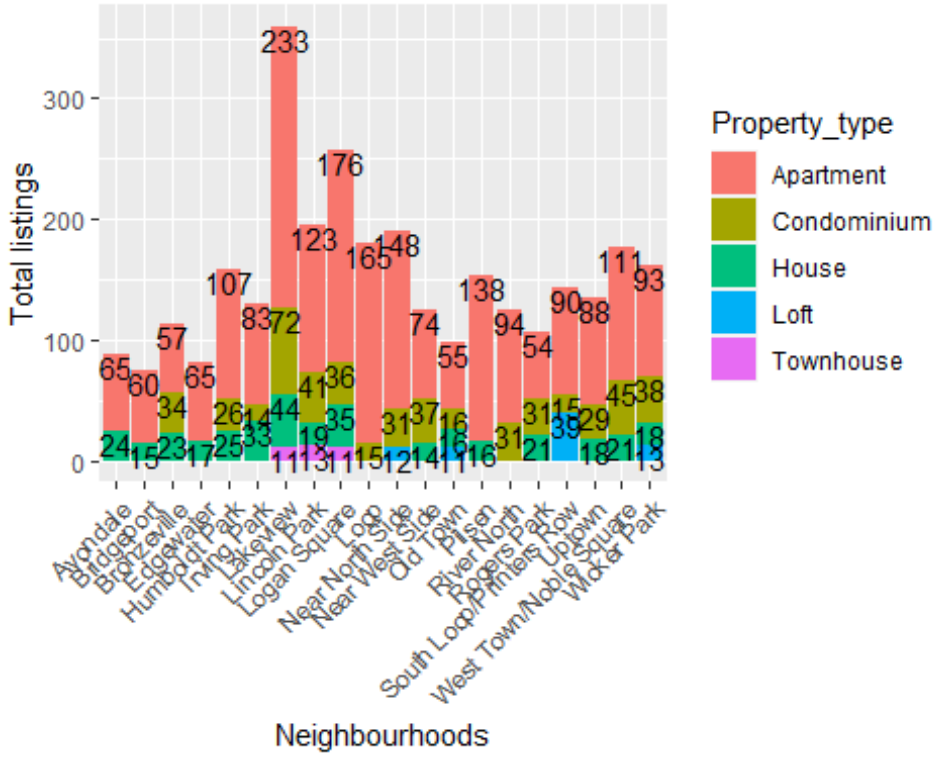
- Must-have amenities - **AC, Heater, Parking, Workspace, WiFi, Internet, Cooking facilities, Long-term stays, Security, No stairs/steps**

```
airChicago1$amenities <- gsub(" ", "_", airChicago$amenities)
airChicago1$amenities <- gsub("[^[:alnum:]_]"," ",airChicago1$amenities)
```

This graph shows the **property types** among the airbnbs having the highest booking rate.

```
plot_property_type <- ggplot(table3, aes(Neighbourhoods, Count_of_property,
fill = Property_type)) +
  geom_bar(stat = "identity") + geom_text(aes(label = Count_of_property),
position = "stack", vjust = 1) + xlab("Neighbourhoods") + ylab("Total
listings") +
  theme(axis.text.x = element_text(angle = 45, size = 9, hjust = 1))

plot_property_type
```

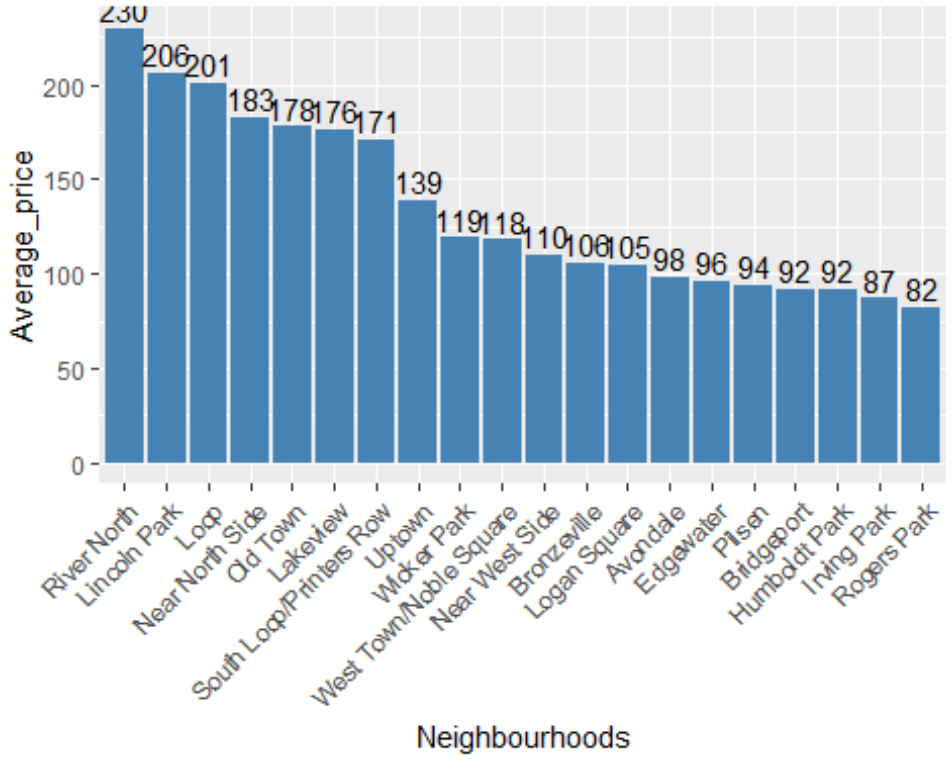


Apartments, Condominiums and **Houses** are the most common property types with high booking rates in Chicago.

This graph shows us the average price of a listing with high booking rate within the particular neighbourhood.

```
plot_average_price <- ggplot(average_price, aes(reorder(neighbourhood,
desc(Average_price)), Average_price)) + geom_bar(stat = "identity", fill =
"steelblue") +
  geom_text(aes(label = Average_price), position = position_dodge(1), vjust =
-0.3) + xlab("Neighbourhoods") + ylab("Average_price") +
  theme(axis.text.x = element_text(angle = 45, size = 9, hjust = 1))

plot_average_price
```



We created these two maps for chicago, in order to see **how does the house size impact the high booking rate**.

```
ggplotly(ggmap(chicago_stamen) +
  geom_point(data = airChicago %>%filter(high_booking_rate==1),
             mapping = aes(x = longitude,
                           y = latitude, color=bedrooms)))
```

```
## PhantomJS not found. You can install it with webshot::install_phantomjs().
If it is installed, please make sure the phantomjs executable can be found
via the PATH variable.
```

```
ggplotly(ggmap(chicago_stamen) +
  geom_point(data = airChicago%>%filter(high_booking_rate==0),
             mapping = aes(x = longitude,
                           y = latitude, color=bedrooms)))
```

Observation:

- Houses with **more bedrooms** have a higher booking rate.
- House placement, i.e **near the coast or not** does not really have an impact.
- **South** has a lower booking rate.

Feature Extraction:

We could think about some 'would-be' useful information still lacking such as the actual location of the property, the impact of having good transport facilities near the property,

why do people generally visit Chicago, and what are their accommodation preferences. These features were latent in the variables- amenities and transit. On the other hand, although the neighborhood gave a sense of the property's location, we were looking for segregated groups of the nearby locations, that would help us identify location trends more clearly.

Based on market analysis, we decided to create/extract some new columns pertaining to our research questions:

Location:

Sides - There are 87 distinct neighborhoods in the dataset. Since we didn't find any major variations based on these and found that neighborhoods are further divided into sides from market analysis, we decided to group these into broader, well-defined regions known as sides (Source - https://en.wikipedia.org/wiki/Community_areas_in_Chicago). Thus, a new column - 'sides" was added.

```
#Neighbourhoods based on sides
side <- read_csv("side.csv")

## Parsed with column specification:
## cols(
##   neighbourhood = col_character(),
##   side = col_character()
## )

airChicago <- left_join(x=airChicago,y=side)

## Joining, by = c("neighbourhood", "side")
```

Beach Proximity - To analyze the impact of proximity to beaches, we created a new column called "beach" which has '1' for location involving "beach", "lake" and '0' for listings without these.

```
airChicago <- airChicago %>%
  mutate(beach = (str_detect(description,"beach")))

airChicago$beach[is.na(airChicago$beach)] <- FALSE
```

Transportation:

To check the importance of transportation accessibility, we created a new column - "good_trans" - which has value 1 for listings that have used "bus", "line", "station", "subway" in the "Transit" column description.

```
airChicago <- airChicago %>%
  mutate(good_trans = (str_detect(transit,"bus")))

airChicago$good_trans[is.na(airChicago$good_trans)] <- FALSE

airChicago <- airChicago %>%
```

```
  mutate(good_transline = (str_detect(transit,"line")))

airChicago$good_transline[is.na(airChicago$good_transline)] <- FALSE
```

Regulation:

Columns associated with regulations - "license required". However, it is true for all listings and hence we would not be including this in the model. To analyze the impact of regulations, we created a new column - "reg" which has 1 for host_about that includes "license", "registered" and '0' for listings without this.

```
airChicago <- airChicago %>%
  mutate(reg = (str_detect(description,c("license","registered"))))

## Warning in stri_detect_regex(string, pattern, negate = negate, opts_regex
=
## opts(pattern)): longer object length is not a multiple of shorter object
length

airChicago$reg[is.na(airChicago$reg)] <- FALSE
```

Length of Stay:

New column - "long_stay" was created. Properties with minimum nights as 30 or more would have 1 and the other 0.

```
airChicago <- airChicago %>%
  mutate(long_stay = ifelse(minimum_nights>=30,1,0))

airChicago$long_stay[is.na(airChicago$long_stay)] <- FALSE
```

Multiple Listings:

New column "multi_listings" with host listings count > 1 as "1" and others as "0".

```
airChicago <- airChicago %>%
  mutate(multiple_listings = ifelse(host_listings_count>1,1,0))

airChicago$multiple_listings[is.na(airChicago$multiple_listings)] <- FALSE
```

Amenities:

Also to check which amenities are more significant to cause higher booking rates, we created new columns for the properties having certain amenities-"Family/kid-friendly", "Self-check-in", "Refrigerator", "Coffee Maker" which have values '1' if the respective amenities are present, otherwise '0'.

```
airChicago <- airChicago %>%
  mutate(quiet = (str_detect(description,"quiet")))

airChicago$quiet[is.na(airChicago$quiet)] <- FALSE
```

```r
airChicago <- airChicago %>%
  mutate(amen_family = (str_detect(amenities,"Family/kid friendly")))

airChicago <- airChicago %>%
  mutate(amen_check = (str_detect(amenities,"Self check-in")))

airChicago <- airChicago %>%
  mutate(amen_ref = (str_detect(amenities,"Refrigerator")))

airChicago <- airChicago %>%
  mutate(amen_cof = (str_detect(amenities,"Coffee")))

head(airChicago)
```

```
## # A tibble: 6 x 79
##       id high_booking_ra~ access accommodates amenities availability_30
##    <dbl> <fct>            <chr>         <dbl> <fct>                <dbl>
## 1 1.05e6 0                The w~            4 "{TV,Wif~                0
## 2 1.02e6 1                Their~            2 "{Intern~               28
## 3 1.05e6 0                <NA>              1 "{TV,Wif~               29
## 4 1.07e6 0                <NA>              2 "{TV,Wif~                0
## 5 1.12e6 0                All a~            4 "{TV,Wif~                0
## 6 1.15e6 0                <NA>              4 "{TV,\"C~                0
## # ... with 73 more variables: availability_365 <dbl>, availability_60
<dbl>,
##   availability_90 <dbl>, bathrooms <dbl>, bed_type <fct>, bedrooms
<dbl>,
##   beds <dbl>, cancellation_policy <fct>, city <chr>, cleaning_fee <dbl>,
##   description <chr>, extra_people <dbl>, guests_included <dbl>,
##   host_about <chr>, host_acceptance_rate <chr>, host_has_profile_pic
<fct>,
##   host_identity_verified <fct>, host_is_superhost <fct>,
##   host_listings_count <dbl>, host_location <chr>, host_neighbourhood
<chr>,
##   host_response_rate <dbl>, host_response_time <fct>, host_since <chr>,
##   host_verifications <chr>, house_rules <chr>, instant_bookable <fct>,
##   interaction <chr>, is_business_travel_ready <fct>, is_location_exact
<fct>,
##   latitude <dbl>, longitude <dbl>, market <chr>, maximum_nights <dbl>,
##   minimum_nights <dbl>, monthly_price <chr>, neighborhood_overview
<chr>,
##   neighbourhood <fct>, notes <chr>, price <dbl>, property_type <fct>,
##   require_guest_phone_verification <fct>,
##   require_guest_profile_picture <fct>, requires_license <fct>,
##   review_scores_accuracy <dbl>, review_scores_checkin <dbl>,
##   review_scores_cleanliness <dbl>, review_scores_communication <dbl>,
##   review_scores_location <dbl>, review_scores_rating <dbl>,
##   review_scores_value <fct>, room_type <fct>, security_deposit <dbl>,
##   space <chr>, square_feet <dbl>, state <chr>, transit <chr>,
##   weekly_price <chr>, zipcode <chr>, `{randomControl}` <dbl>, side
```

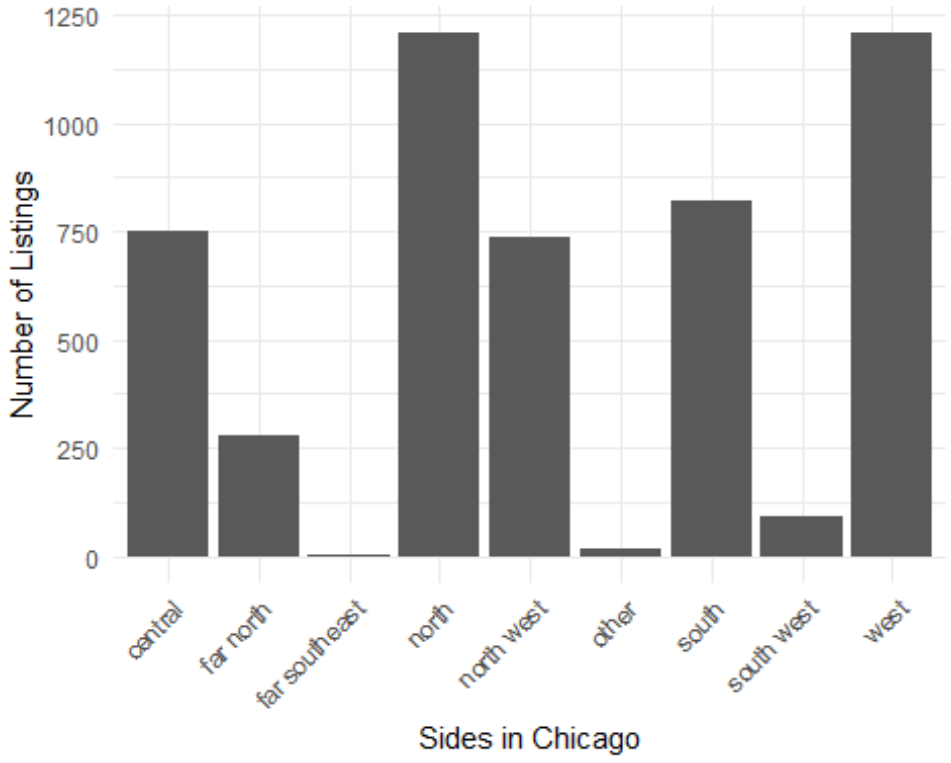
```
<fct>,
## #   accomodates <dbl>, beach <lgl>, good_trans <lgl>, good_transline
<lgl>,
## #   reg <lgl>, long_stay <dbl>, multiple_listings <dbl>, quiet <lgl>,
## #   amen_family <lgl>, amen_check <lgl>, amen_ref <lgl>, amen_cof <lgl>
```

This is the final dataframe we recieved after we added the columns that we thought were essential.

Research Questions

Impact of Location on Booking Rate

```
g <- ggplot(data=airChicago, aes(x=side))
g + geom_bar(aes()) + xlab("Sides in Chicago") + ylab("Number of Listings") +
theme(axis.text.x = element_text(angle=45,size=9, hjust=1))
```




```
g <- ggplot(data=airChicago, aes(x=side))
g + geom_bar(aes(fill = high_booking_rate), position="fill") + xlab("Sides in
Chicago") + ylab("Proportion of High Booking Rate") + theme(axis.text.x =
element_text(angle=45,size=9, hjust=1))
```

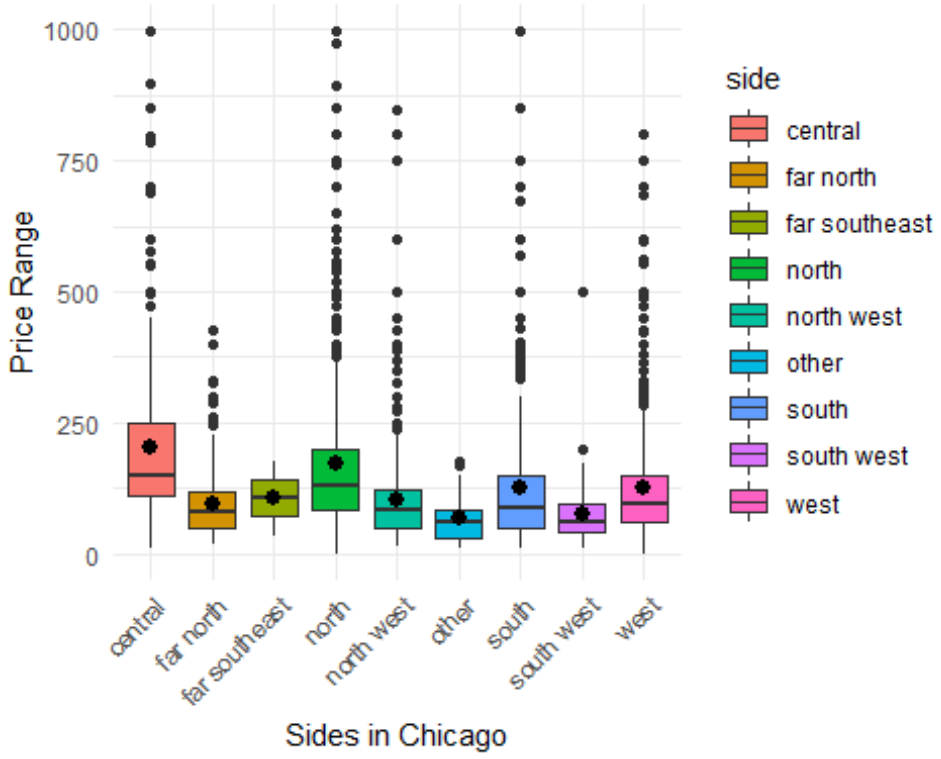
Observations:

- The **South region** (Including south, far southeast & southwest) has low listings count as well as a proportion of high booking rate.
- The **central region** has high listings count but a low proportion of high booking rate.
- The **North & West** have high listings count as well as a proportion of high booking rate.

```
g <- ggplot(airChicago, aes(x = side, y = price))

# Notched box plot with mean points
g + geom_boxplot(notch = FALSE, aes(fill=side))+ xlab("Sides in Chicago") +
ylab("Price Range") + theme(axis.text.x = element_text(angle=45,size=9,
hjust=1)) +
  stat_summary(fun.y = mean, geom = "point",
               size = 2.5, color = "black")

## Warning: `fun.y` is deprecated. Use `fun` instead.
```



Observations:

- There appears to be a **correlation between low listings count & booking rate in the south region** with higher crime rates in the corresponding region.
- Central region's low booking rate despite higher listings count could be explained by the **higher price range**.
- Potential hosts could benefit best by **focussing on the North & West sides**.

Impact of Regulations on New Listings

```
```{r, include=FALSE}

# total_hosts <- airChicago %>%

#   dplyr::select("host_since", "host_listings_count", "host_location",
"host_neighbourhood")

# total_hosts <- unique(total_hosts)

# total_hosts <- total_hosts %>%

#   group_by(host_since) %>%

#   tally() %>%

#   arrange(desc(n))

#

# total_hosts$host_since <- yearmonth(total_hosts$host_since)
```
```

```
# total_hosts <- aggregate(total_hosts["n"], by = total_hosts["host_since"],
sum)

#

# total_hosts <- as_tsibble(total_hosts, key = host_since, index = n)

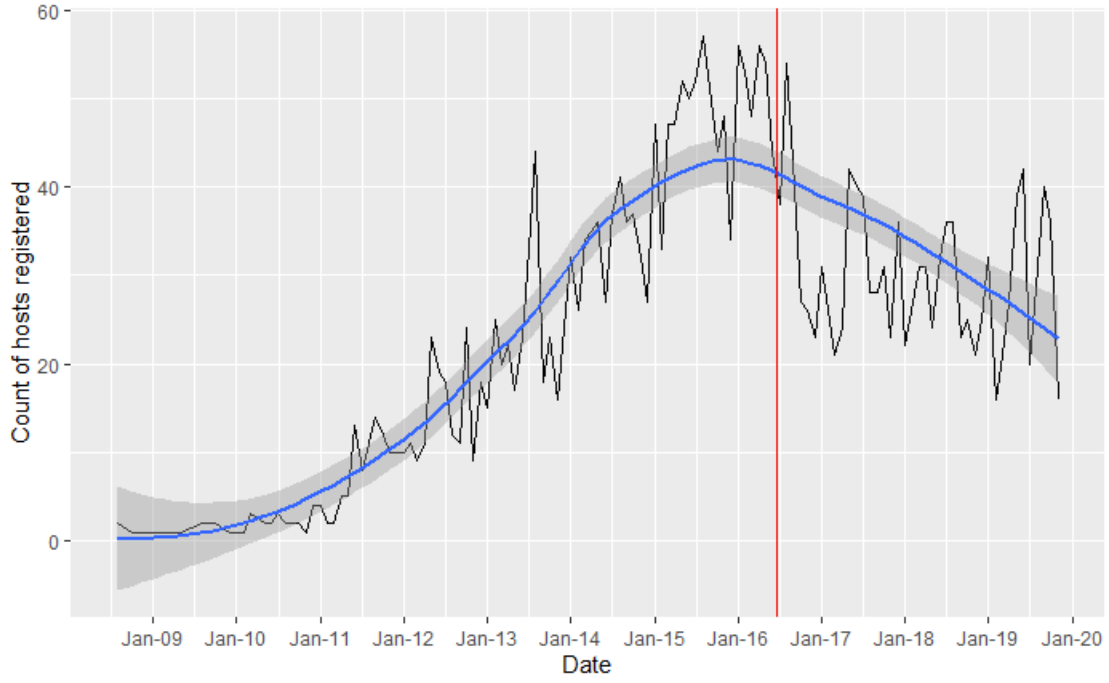
# total_hosts
```

```
plot_total_hosts <- total_hosts %>%
  ggplot(aes(x = host_since, y = n)) + geom_line(color = "black") +
scale_x_date(date_labels = "%b-%y", date_breaks = "12 month") +
  xlab("Date") + ylab("Count of hosts registered") + geom_smooth() +
geom_vline(xintercept = as.Date("2016-06-22"), col = "red")

## Error in eval(lhs, parent, parent): object 'total_hosts' not found

plot_total_hosts

## Error in eval(expr, envir, enclos): object 'plot_total_hosts' not found
```



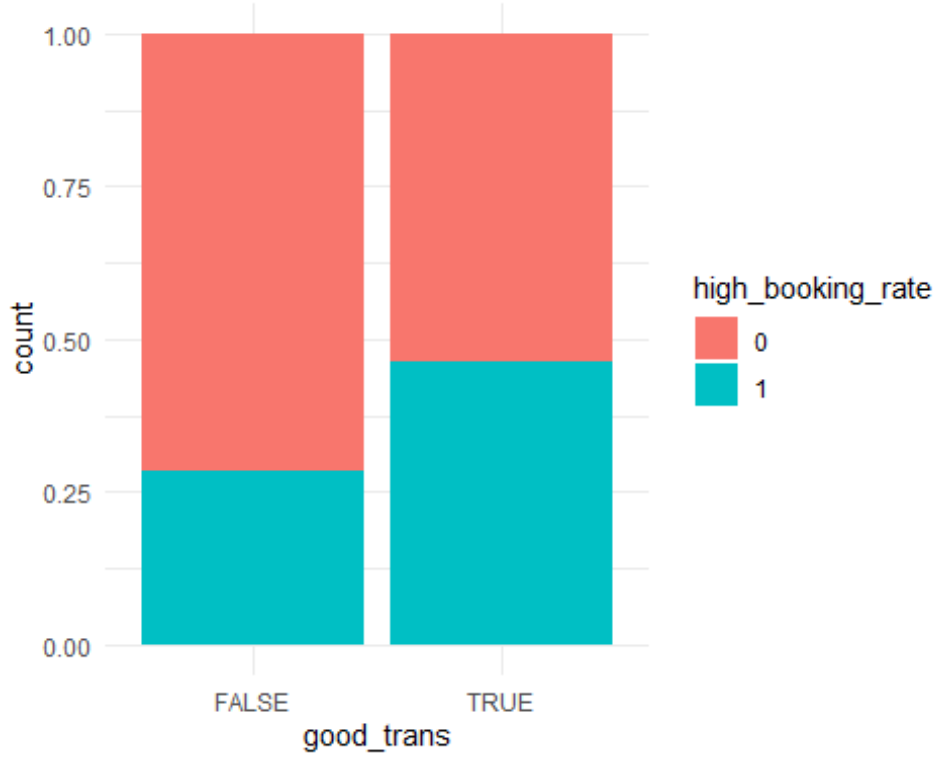
Observations:

- Shared Housing Ordinance issued by Chicago City Council on June 22, 2016, imposed restrictions on shared/vacation rental listing owners to better regulate & to impose taxes. The negative impact of this is evident by the drop in the number of new Airbnb host registrations.

- 7.6% of all housing units in Chicago are prohibited from hosting on Airbnb. Potential hosts should be wary of all Chicago regulations and requirements beforehand.

Impact of Transportation on Booking Rate

```
g <- ggplot(data=airChicago, aes(x=good_trans))
g + geom_bar(aes(fill=high_booking_rate), position="fill")
```



Observations:

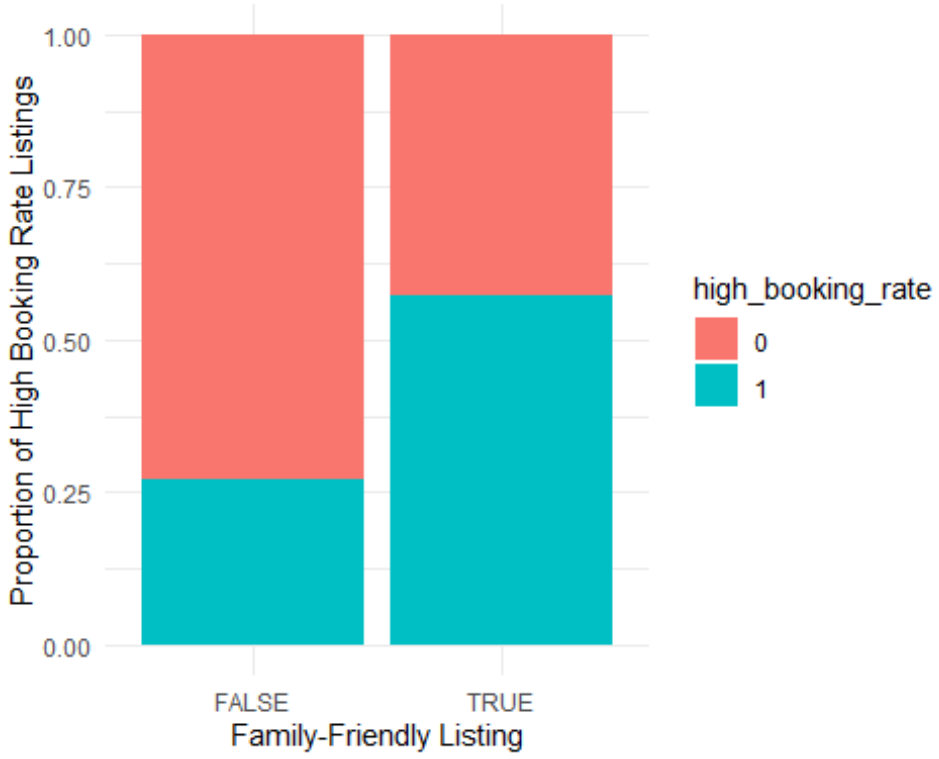
- Presence of **"bus", "station", "line", "cta", "red/brown" in "Transit"** for high booking rate listings.
- According to Chicago Tribune, Chicago's public transit ranks 6th among all US cities.

Inference:

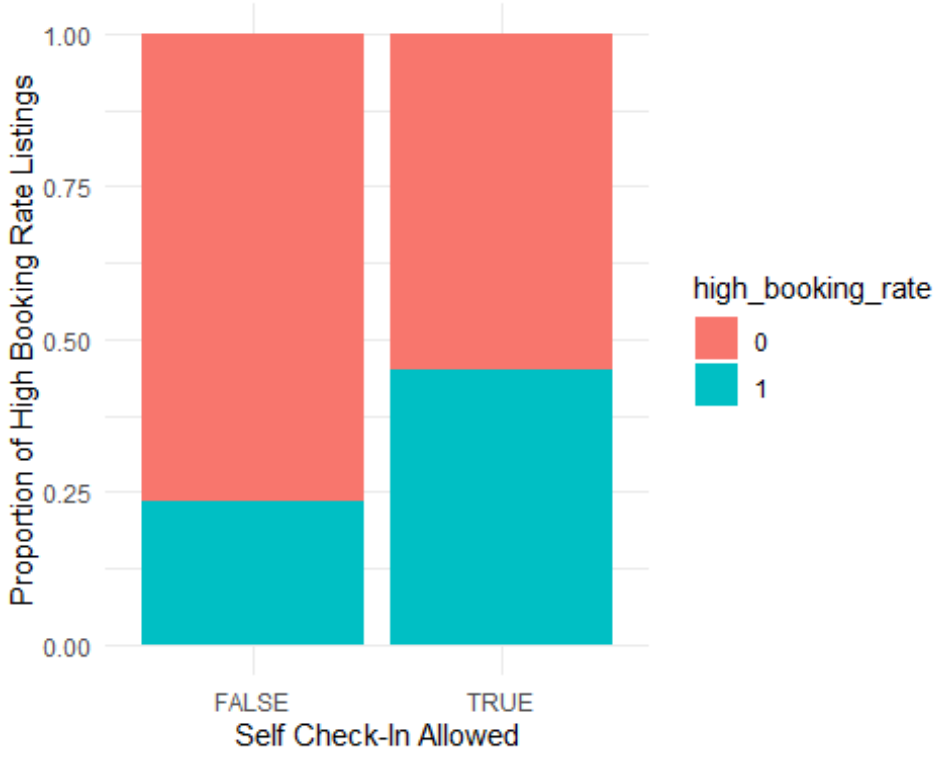
This suggests listings that have **better public transport accessibility** have higher chances of having high booking rates.

Impact of Amenities on Booking Rate

```
g1 <- ggplot(data=airChicago, aes(x=amen_family))
g1 + geom_bar(aes(fill=high_booking_rate), position="fill") + xlab("Family-
Friendly Listing") + ylab("Proportion of High Booking Rate Listings")
```




```
g1 <- ggplot(data=airChicago, aes(x=amen_check))
g1 + geom_bar(aes(fill=high_booking_rate), position="fill") + xlab("Self
Check-In Allowed") + ylab("Proportion of High Booking Rate Listings")
```



Observations:

- Must-have amenities - **AC, Heater, Parking, Workspace, WiFi, Internet, Cooking facilities, Long-term stays, Security, No stairs/steps**
- Differentiating Factors - Family-friendliness, Self check-in option

Inferences:

Importance of **family-friendliness, security etc. indicates guests prefer quieter & safer neighbourhoods**. This could also mean guests are predominantly families.

Model:

During the first part of the project (Kaggle competition), we focussed primarily on identifying variables that had a huge impact on the booking rate, irrespective of a specific market. We chose Linear Discriminant Analysis as the statistical model for this reason. We used LDA because we wanted to remove the dependent variables by changing the dataset to a lower-dimensional space so that we were able to retain most of the data. In addition to that, since we only used limited features in our model for the Kaggle competition, it was necessary for us to LDA over logistic regression. This is because logistic regression tends to be more unstable when used with a small number of features.

For the second part, we started off with the same variables as we felt these were more generic and would have an impact on the booking rate, irrespective of the market. Shortlisted variables have been listed below:

Variables Based on Kaggle Model:

1. Describing the characteristics of the property- *Property price, number of bedrooms, beds, and baths, number of people it accommodates, instant_bookable, is_location_exact*

2.About the host themselves- *host response time, host response rate, cancellation policy, if a cleaning fee charged, review_scores_rating, host_response_time, host_is_superhost, host_identity_verified, cancellation_policy, price, accommodates, availability_30, availability_90, availability_365, requires_license, guests_included, host_listings_count*

```
set.seed(333)

airChiTrain <- airChicago %>% sample_frac(.70)
airChiTest <- setdiff(airChicago, airChiTrain)

# Converting the dependent variable into a factor
airChiTrain <- airChiTrain %>%
               mutate(high_booking_rate = as.factor(high_booking_rate))
airChiTest <- airChiTest %>%
               mutate(high_booking_rate = as.factor(high_booking_rate))

fitrf <- train((high_booking_rate) ~
     cleaning_fee
     +review_scores_rating
```

```
     +security_deposit
     +host_response_rate
     +host_response_time
     +host_is_superhost
     +host_identity_verified
     +host_listings_count
     +host_has_profile_pic
     +cancellation_policy
     +price
     +accommodates
     +availability_30
     +availability_365
     +instant_bookable
     +is_location_exact
     +requires_license
     +guests_included
     +room_type
     +property_type
     +maximum_nights
     +minimum_nights
     +latitude
     +longitude
     +bathrooms
     +bed_type
     +beds
     +bedrooms
     #+side
     #+amen_family
     #+amen_check
     #+amen_cof
     #+amen_ref
     #+good_transline
     #+good_transbus
     , family='binomial', data=airChiTrain, method='rf')

resultsLog <-fitrf%>%
  predict(airChiTest,type='raw') %>%
  bind_cols(airChiTest, predictedClass = .)
 #as.factor(high_booking_rate))
#Creating the confusion matrix
resultsLog %>%
  xtabs(~predictedClass+high_booking_rate, .) %>%
  confusionMatrix(positive = '1')

## Confusion Matrix and Statistics
##
##                high_booking_rate
## predictedClass    0    1
##              0 3200  181
##              1  110 1626
```

```
##
##                Accuracy : 0.943
##                  95% CI : (0.936, 0.949)
##     No Information Rate : 0.647
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.874
##
##  Mcnemar's Test P-Value : 4.07e-05
##
##             Sensitivity : 0.900
##             Specificity : 0.967
##          Pos Pred Value : 0.937
##          Neg Pred Value : 0.946
##              Prevalence : 0.353
##          Detection Rate : 0.318
##    Detection Prevalence : 0.339
##       Balanced Accuracy : 0.933
##
##        'Positive' Class : 1
##
```

The new variables created such as the *'side', 'amen_family', 'amen_check', 'good_transline', 'good_transbus'* had a positive impact on the accuracy.

Whereas the variables created such as the *'good_transsubway' and 'good_transstation'* did not have a positive impact on the accuracy which is a bit surprising. Maybe the buses are preferred over subways.

```
# Converting the dependent variable into a factor
airChiTrain <- airChiTrain %>%
               mutate(high_booking_rate = as.factor(high_booking_rate))
airChiTest <- airChiTest %>%
               mutate(high_booking_rate = as.factor(high_booking_rate))

fitrf <- train((high_booking_rate) ~
     cleaning_fee
     +review_scores_rating
     +security_deposit
     +host_response_rate
     +host_response_time
     +host_is_superhost
     +host_identity_verified
     +host_listings_count
     +host_has_profile_pic
     +cancellation_policy
     +price
     +accommodates
     +availability_30
     +availability_365
```

```
    +instant_bookable
    +is_location_exact
    +requires_license
    +guests_included
    +room_type
    +property_type
    +maximum_nights
    +minimum_nights
    +latitude
    +longitude
    +bathrooms
    +bed_type
    +beds
    +bedrooms
    +side
    +amen_family
    +amen_check
    +amen_cof
    +amen_ref
    +good_trans
    +good_transline
    , family='binomial', data=airChiTrain, method='rf')


resultsLog <-fitrf%>%
  predict(airChiTest,type='raw') %>%
  bind_cols(airChiTest, predictedClass = .)
 #as.factor(high_booking_rate))
#Creating the confusion matrix
resultsLog %>%
  xtabs(~predictedClass+high_booking_rate, .) %>%
  confusionMatrix(positive = '1')

## Confusion Matrix and Statistics
##
##                high_booking_rate
## predictedClass    0    1
##              0 3205  165
##              1  105 1642
##
##                Accuracy : 0.947
##                  95% CI : (0.941, 0.953)
##     No Information Rate : 0.647
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.884
##
##  Mcnemar's Test P-Value : 0.00033
##
##             Sensitivity : 0.909
```

```
##             Specificity : 0.968
##          Pos Pred Value : 0.940
##          Neg Pred Value : 0.951
##              Prevalence : 0.353
##          Detection Rate : 0.321
##    Detection Prevalence : 0.341
##       Balanced Accuracy : 0.938
##
##        'Positive' Class : 1
##
```

Final Accuracy we achieved: Final Spoecificity we achieved:

Threshold Selection:

Putting ourselves in the shoes of the property owners, we thought that which misclassification error would cost us more. So, the prediction that the booking rate is high i.e 1 and it actually being 0 is most costly which is False Positive Type 2 error. In order to improve our model to be stricter in labeling properties to have a high booking rate, we had to increase our cut-off. The cutoff value of 0.65 gave us the highest specificity. (We tested this using our glm model)

```
# Converting the dependent variable into a factor

airChiTrain <- airChiTrain %>%
               mutate(high_booking_rate = as.factor(high_booking_rate))
airChiTest <- airChiTest %>%
               mutate(high_booking_rate = as.factor(high_booking_rate))
resultsLog <-
  glm((high_booking_rate) ~
     cleaning_fee
     +review_scores_rating
     +security_deposit
     +host_response_rate
     +host_response_time
     +host_is_superhost
     +host_identity_verified
     +host_listings_count
     +host_has_profile_pic
     +cancellation_policy
     +price
     +accommodates
     +availability_30
     +availability_365
     +instant_bookable
     +is_location_exact
     +requires_license
     +guests_included
     +room_type
     +maximum_nights
```

```
    +minimum_nights
    +latitude
    +longitude
    +bathrooms
    +bed_type
    +beds
    +bedrooms
    +side
    +amen_family
    +amen_check
    +good_trans
    +good_transline, family='binomial', data=airChiTrain)%>%
  predict(airChiTest,type='response') %>%
  bind_cols(airChiTest, predictedProb = .)%>%
  mutate(predictedClass=as.factor(ifelse(predictedProb >0.65, 1, 0)))
 #as.factor(high_booking_rate))
#Creating the confusion matrix
resultsLog %>%
  xtabs(~predictedClass+high_booking_rate, .) %>%
  confusionMatrix(positive = '1')

## Confusion Matrix and Statistics
##
##               high_booking_rate
## predictedClass    0    1
##              0 3128 1113
##              1  182  694
##
##                Accuracy : 0.747
##                  95% CI : (0.735, 0.759)
##     No Information Rate : 0.647
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.373
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.384
##             Specificity : 0.945
##          Pos Pred Value : 0.792
##          Neg Pred Value : 0.738
##              Prevalence : 0.353
##          Detection Rate : 0.136
##    Detection Prevalence : 0.171
##       Balanced Accuracy : 0.665
##
##        'Positive' Class : 1
##
```

Results and Findings

The following table highlights the odds ratio & percentage increase in booking rate of the variables that we believe have a positive effect on our model and support our hypothesis-

Variable.	Odds Ratio	Percentage
side_central	1.32	32%
side_north	1.42	42%
side_west	1.55	55%
family_friendly	3.55	255%
Self Check in	1.99	99%
Refridgerator	2.19	119%
Trans bus	1.47	47%
Trans Line	1.31	31%

As shown above, we see that listings in the north have a 42% higher chance of getting a high booking rate as compared to listings in south. (Reference variable used for side is south), listings in west have 55% higher chances and central have 32% higher chances as compared to south. Listings that are connected to bus & metro have 47% and 31% higher chances of having a high booking rate. Listings that are family friendly, have self check-in and a refrigerator have extremely high chances of having a high booking rate as compared to listings that do not have these amenities, as indicated by the above table.

Summary

Location, crime rate, and accessibility to public transportation are some key parameters that have a huge impact on the booking rate of listings. We concluded this through market analysis, exploratory data analysis, and adding these variables to model led to an increase in accuracy, which reinstates our analyses. Buses are the most preferred means of transport in Chicago, which explains the relative increase in accuracy upon the inclusion of buses rather than subways or lines. We expected house rules including restricting partying, smoking, need to be quiet after 10 pm, etc to have a negative impact on booking rate. However, this was disproved as listings that have the “quiet after 10 pm” rule have higher booking rates. Restrictions on partying/smoking have no significant impact. Combining this with the finding that the presence of a “family-friendly” tag has a huge positive impact on booking rates and including these variables in the model led to an increase in the accuracy, we can conclude that the guests are primarily families and prefer quieter neighborhoods. Furthermore, being closer to the coast does not seem to have an impact on the booking rate.

Based on the data and analysis, we could not find significant conclusions on the impact of regulations on the booking rate. We, however, found that there has been a huge impact on the number of hosts registering on the platform. This further suggests two things-

1. Potential investors need to carefully **check for all regulations before buying a property**
2. **Lower competition in a highly growing market**, which could be an advantage for a potential buyer.

Recommendations for potential investors

Before Investing:

- Focusing on North & West sides could lead to a higher booking rate. * If a buyer wants to focus on income and not booking rate, or are planning on renting the property only intermittently, focussing on the central region would be best as just a few bookings would lead to higher income due to higher price range. It is best to avoid the South region due to higher crime rates.
- Focusing on connectivity to public transport would lead to higher bookings.
- Make sure the neighborhood has a low crime rate.
- Look at all regulations carefully. Make sure property is not listed on the prohibited properties list.
- Focus on family-friendly neighborhoods.

After Investing:

- Allowing self-check-in can have a huge impact on the booking rate of the property.
- Making the property family and kid-friendly, adding rules that restrict noise, partying, etc. can have a positive impact on the booking rate.
- Having a refrigerator and coffee maker can increase the high booking rate.

Conclusion:

Key findings/recommendations for investors

- Listings in North and West regions appear to perform well in terms of booking rate, so purchasing properties in those areas would be advisable.
- Properties with transportation facilities nearby such as bus stops, train stations, etc. are much more attractive for potential customers.
- Having a family friendly property with an easy check-in policy at listings with a quieter neighbourhood and specifically having amenities such as refrigerator and coffee-maker helps with achieving a higher booking rate.