Team5 Report

Josh Liu, Imran Haluk Senlik, Billy Xing, Xiaomian Wu

10/7/2017

“All models are wrong, but some are useful”- George E. P. Box

“An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.” – John Tukey

“On two occasions I have been asked [by members of Parliament], ‘Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?’ I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.” –Charles Babbage

Topic Question

Airbnb, a lodging company where anyone can make a few dollars by renting our their spare room, has created competition for hotels. It has helped lower the barriers of entry for lodging and created a market of itself. Although historically it has struggled with breaking into the high end market as there are very few, if any, Airbnb locations that can compete with such luxuries as swimming pool, massage and spa services of an expensive hotel, it has been a beacon of lodging services for people looking for an inexpensive resort.

How much should an Airbnb host charge for their property– particularly, can we predict future prices based off of historical trends, and which factors are statistically significant in determining the price of a listing?

The model assumes that the average host is looking to maximize their revenue while still remaining competitive with the market. We believe the following characteristics will be significant in predicting the price of a listing:

Value of the Property - Number of Venues that are walking distance away - The demographics of the area, specifically the median household income - A combination of the Bed Type and number of people the listing Accommodates - Whether the day is a weekday/weekend

Non-Technical Executive Summary

To answer the question of the optimal price of a listing, we created a model that predicts the price based off of the traits of the listing. We created a web application that provides Airbnb hosts with an optimal price based off of the characteristics of their listing and our model.

Technical Executive Summary

Introduction and Methods

Our group developed an interactive Web App for property owners to help them make informed decisions about the optimal price they can charge. In the Web App, homeowners can get an estimate of the value of their rental property, by simply providing a few pieces of information, such as date of rental, number of guests they can accommodate, number of bathrooms, number of bedrooms, cancellation policy, etc. With this tool, property owners can get an estimate of the value of one night’s stay in their property, calculated based on information from the past. Click here to access our Airbnb host toolkit for pricing listings: **Airbnb Property Owners Toolkits (https://smartbilly.shinyapps.io/citadel_datathon/)**

To create the model, ANOVA tests were run on the different variables to determine the statistically significant ones. Pearson's correlation tests were run on the categorical data to evaluate the likelihood of differences between the sets as our model was based off of unpaired data with a large sample space.

The following factors were found to be significant at the Alpha=5%:

- Accommodates
- Bathrooms
- Bedrooms
- Instant Book-ability
- Review Scores Rating
- Room Type : {Private, Shared}
- Cancellation policy

The null hypothesis that there is no difference between the customers that use Airbnb during the weekdays Monday-Thursday and weekend Friday-Sunday proved to be not false. Property type was also insignificant in predicting prices of listings. We found that listings that do not provide a "Real Bed" are popular among the customers looking for a cheaper place to stay. Median household income and day of the week are significant factor however they were not built into our model due to time restrictions.

We expected that the price of listings would increase as the number of venues that are walking distance away increased. This hypothesis is based on the assumption that the number of venues is a strong indicative of the density of a location which is positively correlated to the demand for a listing.

The size of the data was small enough that we could run our models on the full data without incurring a high computational time cost.

Through the use of the R package "GGMap" we were able to to render a visualization of the number of venues listed in the "venues" dataset. Next, using the latitude and longitude we were able to plot the density of venues in each neighborhood. This is shown in the maps below (Figure A.1).

The heat map was then merged with the listings dataset by rounding to consider the popularity of a certain area based off of the number of attractions within walking distance (0.5 miles) of the listing property. The listings was used to create a heat map of the prices based off of the location (latitude and longitude, Figure B) which was then compared with the heat map of the number of venues.

The New York City, DC, and Chicago area heat maps show a high correlation between the number of venues and price of the listing. However, in Denver we found a wider distribution when it comes to predicting the price of a listing based off of the number of venues in its neighborhood. We believe that a possible reason for the range of correlation is the popularity of medium to high sized businesses and the banking industry in the different cities which creates a market not only for venues but also for hosts.

Figure A.2 is a vector of days of bar charts of the number of properties available at different prices. It suggested signs of price gouging between the weekdays and weekends so we ran t-tests for significancs. Our model shows that property rental price on AirBnB is significantly correlated with the day of week. Nor surprisingly, price is the highest on Friday night and Saturday night, and drops immediately on Sunday, staying on a low level until the next weekend. We can conclude that most AirBnb users are travelers for pleasure, rather than business travelers. The number of available properties shows that is very little difference in the number of available listings on the weekends and weekdays (Figure C).

The Pearson Correlation coefficient between the different types of ratings proved to be too high (Figure D), and the following variables were subtracted from the model: Review Scores Communication, Review Scores Cleanliness, and Review Scores CheckIn.

Discussion

Our application provides Airbnb hosts with an accurate value of their listing which would help them incur the highest possible revenue by maximizing their listing price while still remaining competitive.

We considered adjusting the model for weekdays and for weekends, expecting that weekday prices to be higher than weekends. Traveling during the week tends to be business related whereas weekends tend to be more for pleasure. Airfare companies have capitalized on this by increasing their prices for trips that are within the week. The number of available listings and the price of these listings change marginally between weekday and weekend however there was a significant different in price between certain days.

Our model is limited by our assumptions. Holidays is a factor we have not considered. Revenue is not necessarily the single most important factor to an Airbnb host, perhaps some hosts prefer to have a higher profit margin and fewer guests.

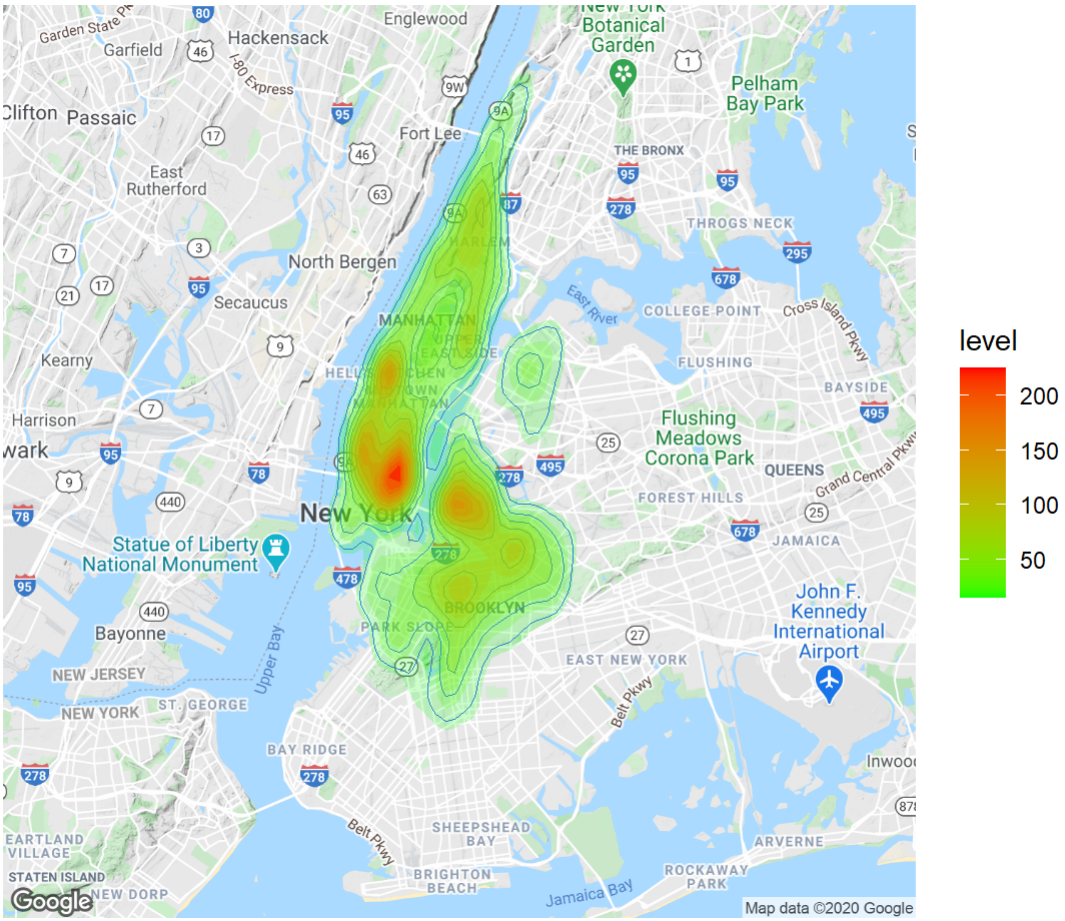
Beyond

As the demand for Airbnb increased over time, hosts were provided with a bigger basket of customers. This has lead to multiple reports of Airbnb hosts canceling reservations based off of the race of the customer. Airbnb has fined a number of hosts up to $5000(1). If we had information on demographics of the guests we could study whether hosts have a preference to popular American names based off of the acceptance/cancellation rates of the names of the guests.

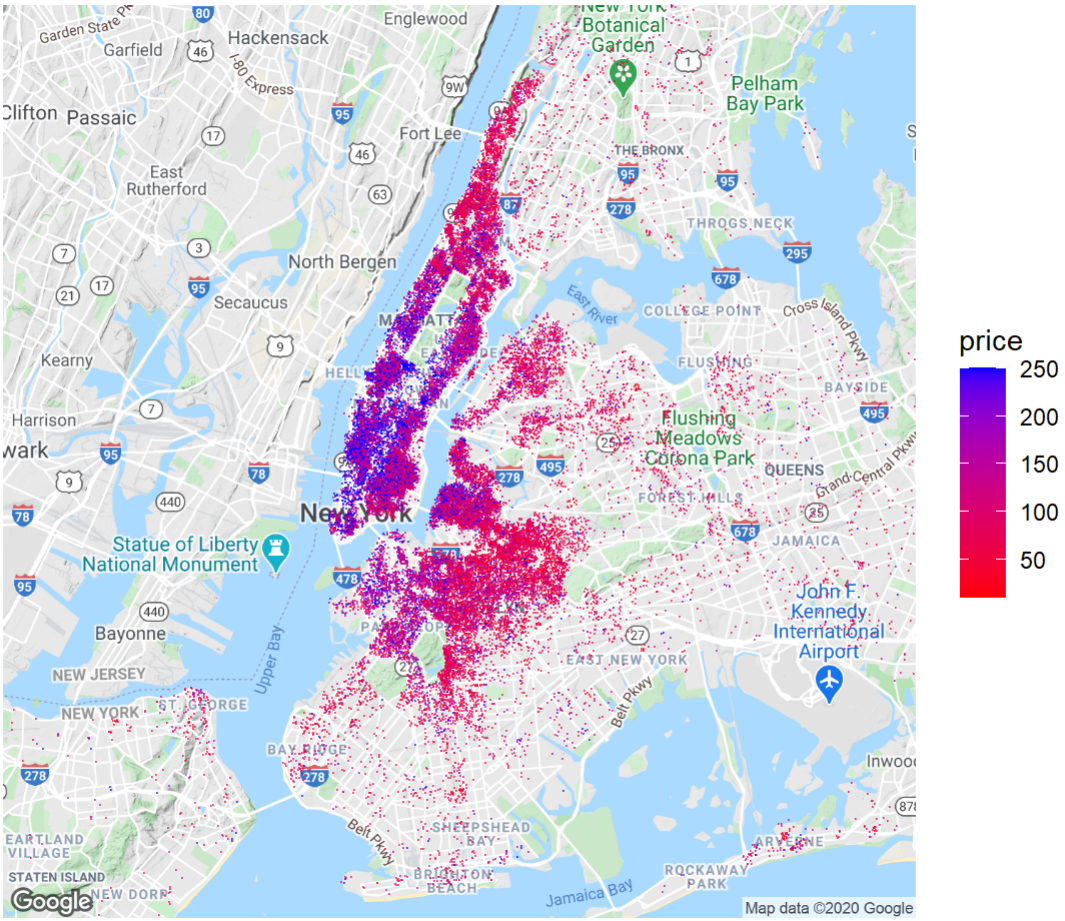
Figure A

NYC: The map shows that properties are usually more expensive in populous area, such as Manhattan, Flushing, Hudson River bank, and price decreases as the location moves further from the center area.

NYC Density Heatmap

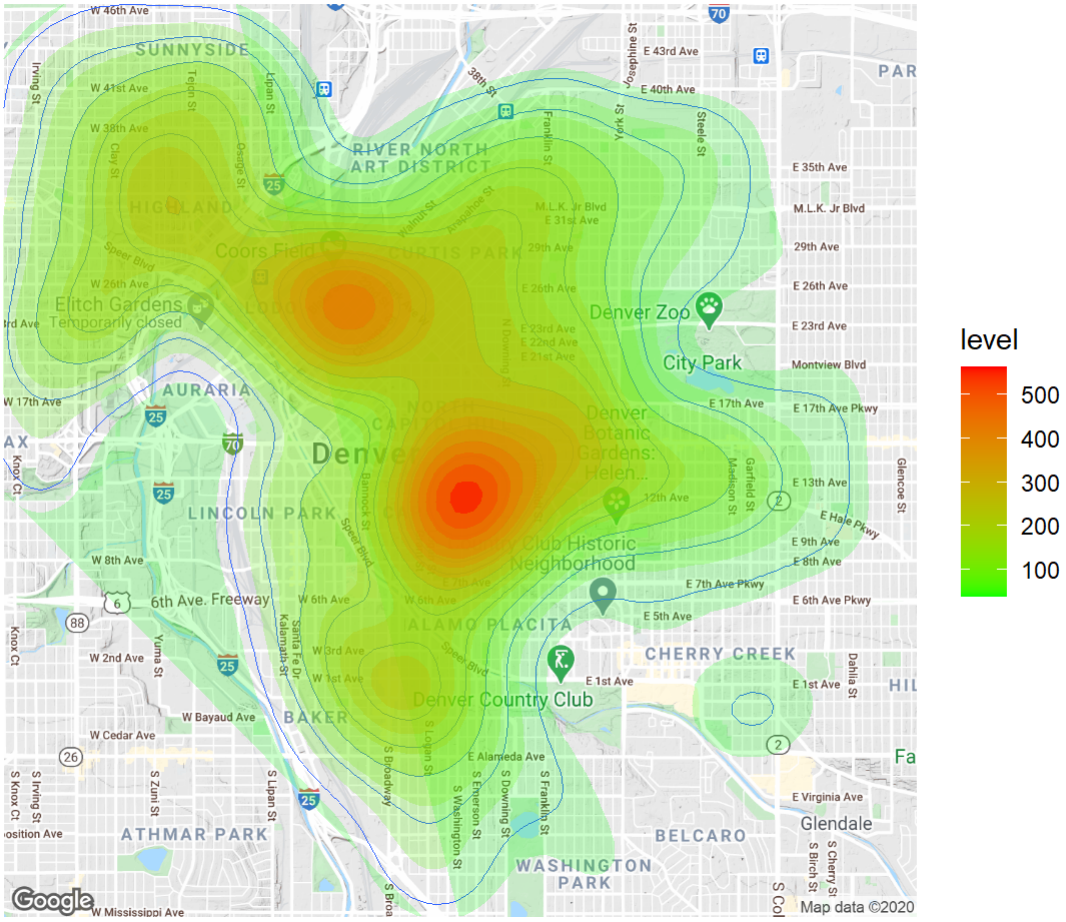


NYC Price Plot

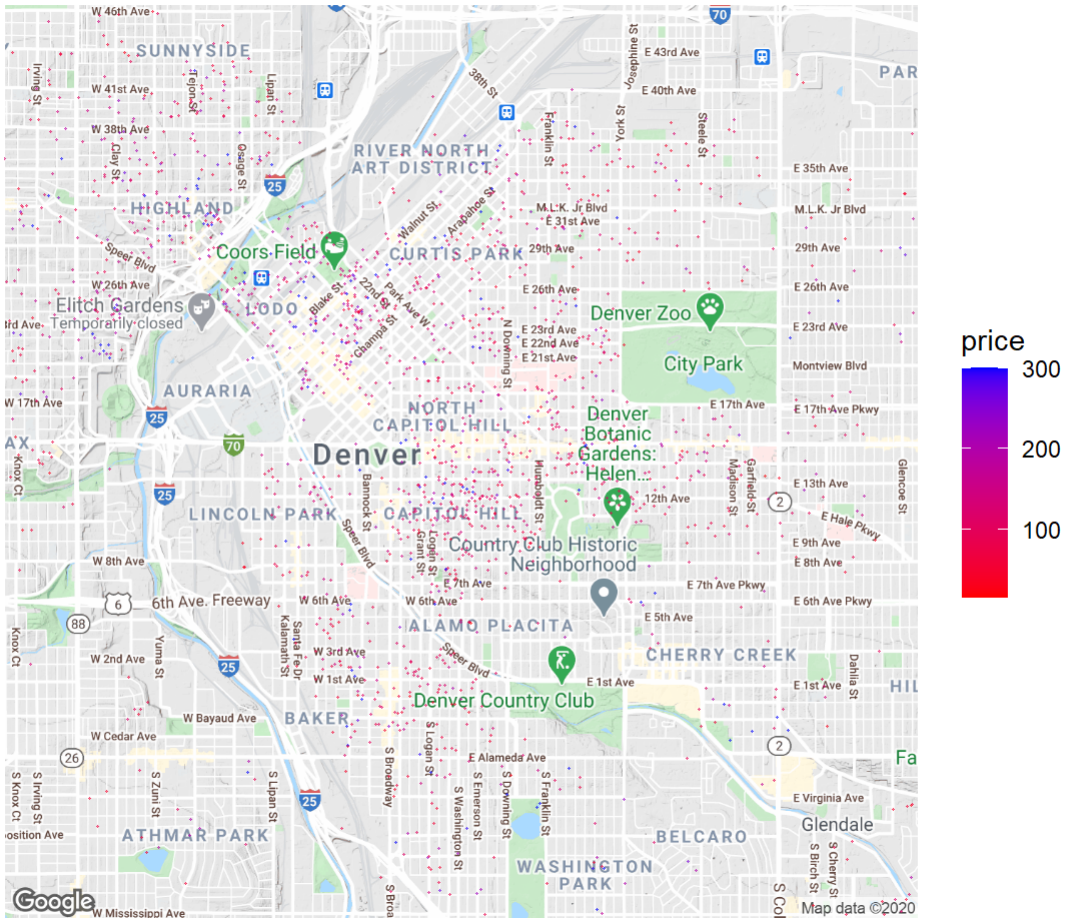


Denver:The map shows that high-priced properties are congregated in two residential areas. The contour map indicates that homeowners are able to leverage their property better, most probably because tourists prefer to stay in a convenient area, in terms of restaurants, tourist attractions, and public transportation.

Denver Density Heatmap

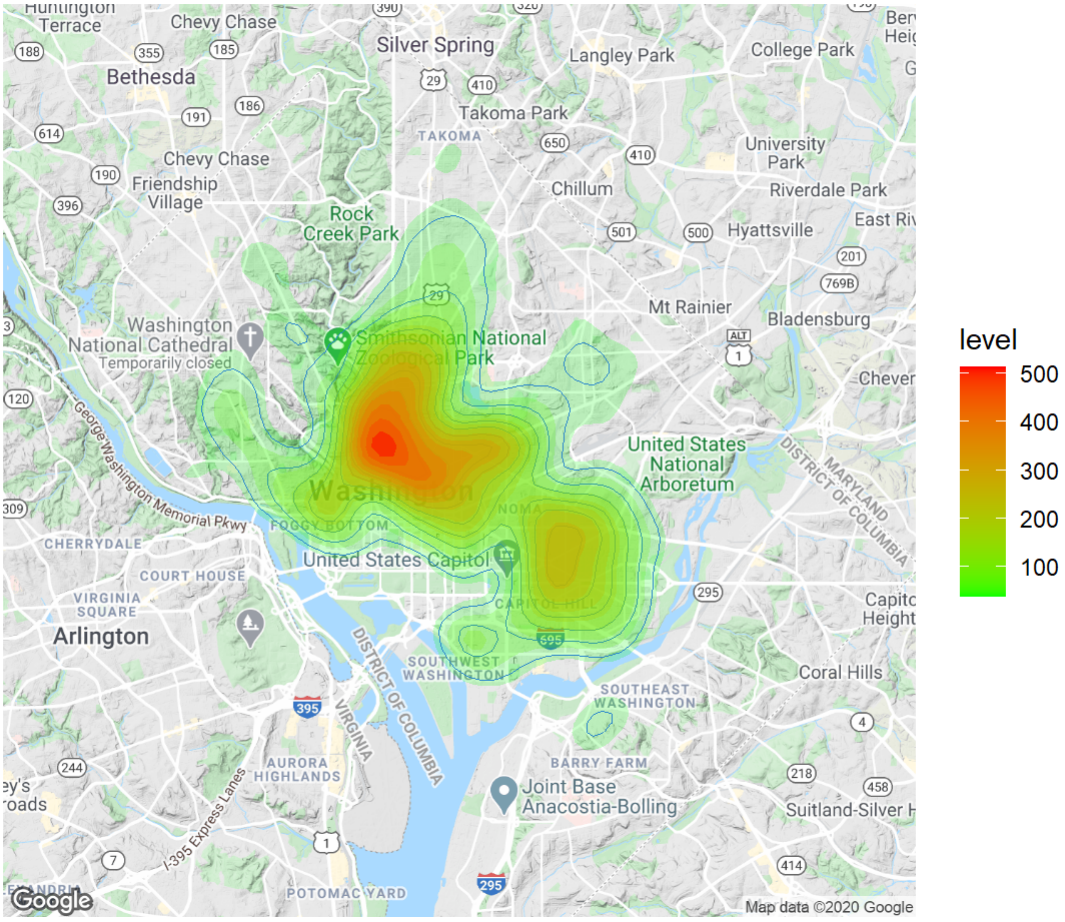


Denver Price Plot

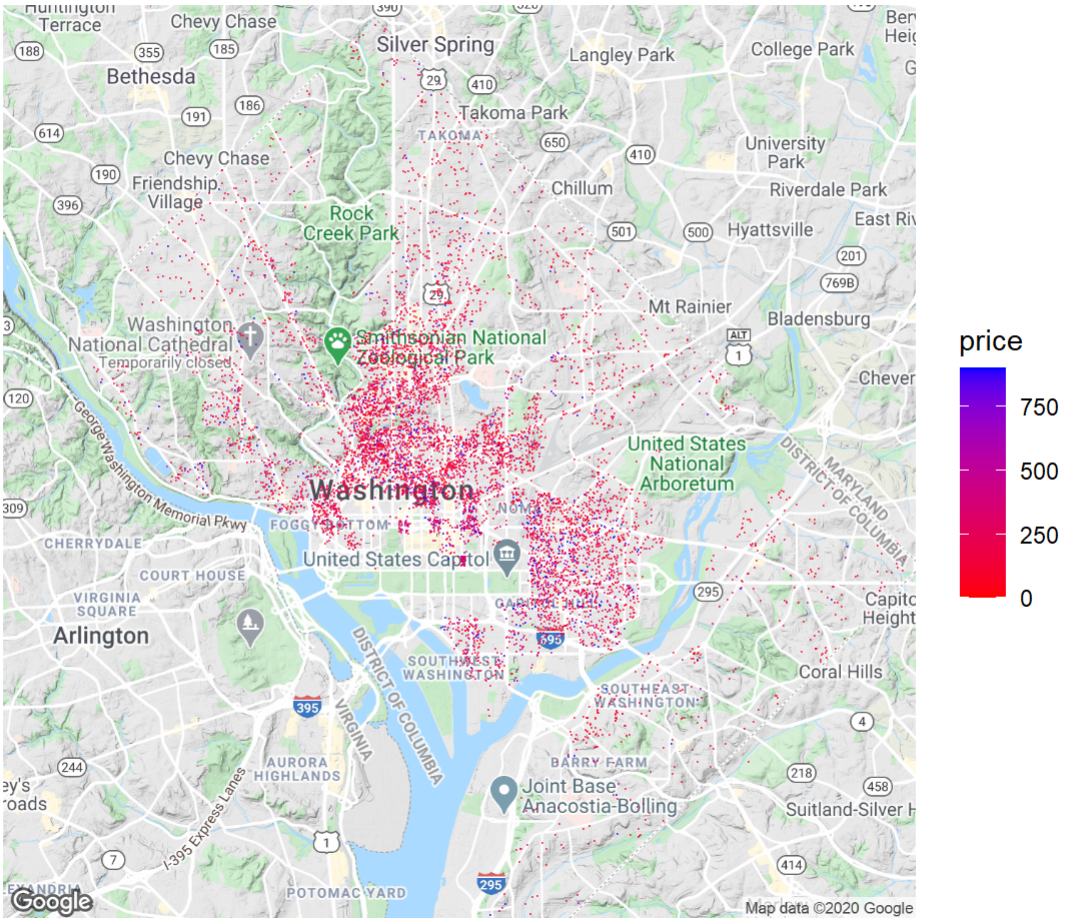


DC:The map of Washington, D.C shows that property rental prices are the highest in downtown DC, mostly in Arlington. It is probably because most tourist attractions, businesses, and government agencies are located in this area. It is very similar to the distribution of rental properties in Denver.

DC Density Heatmap

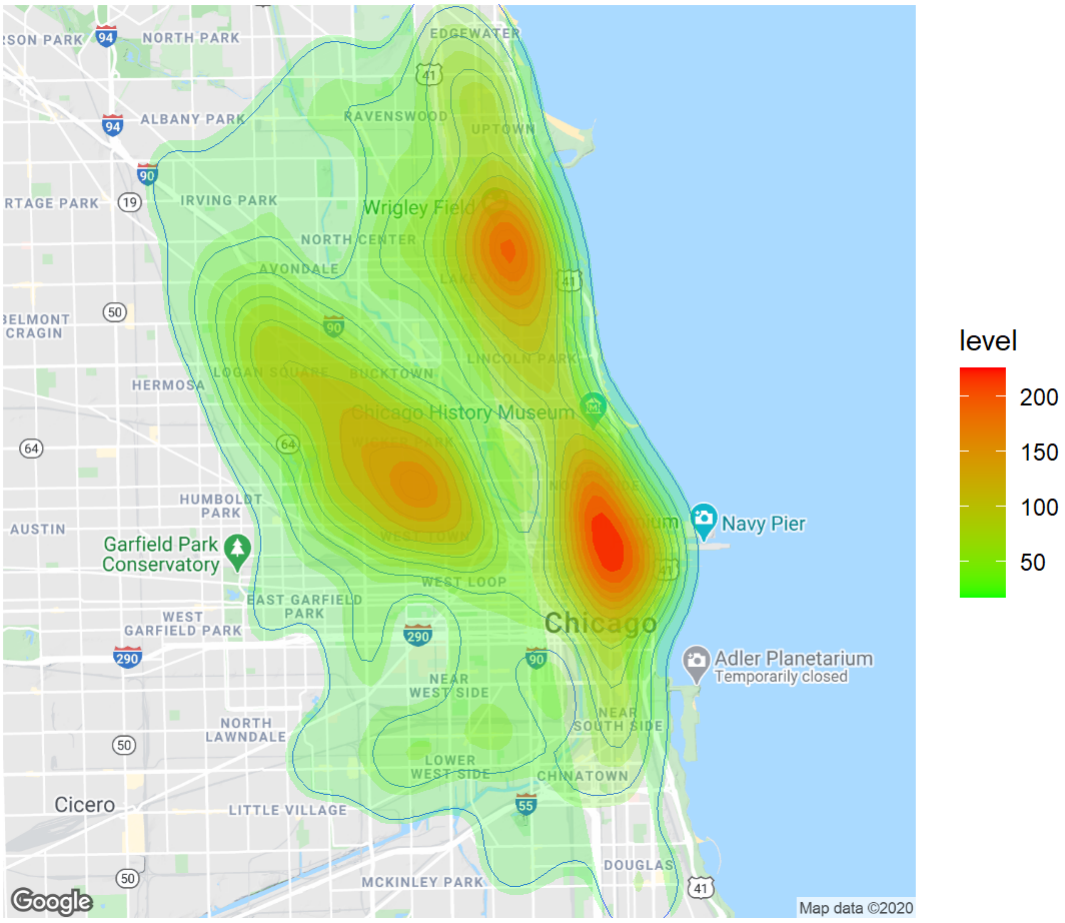


DC Price Plot

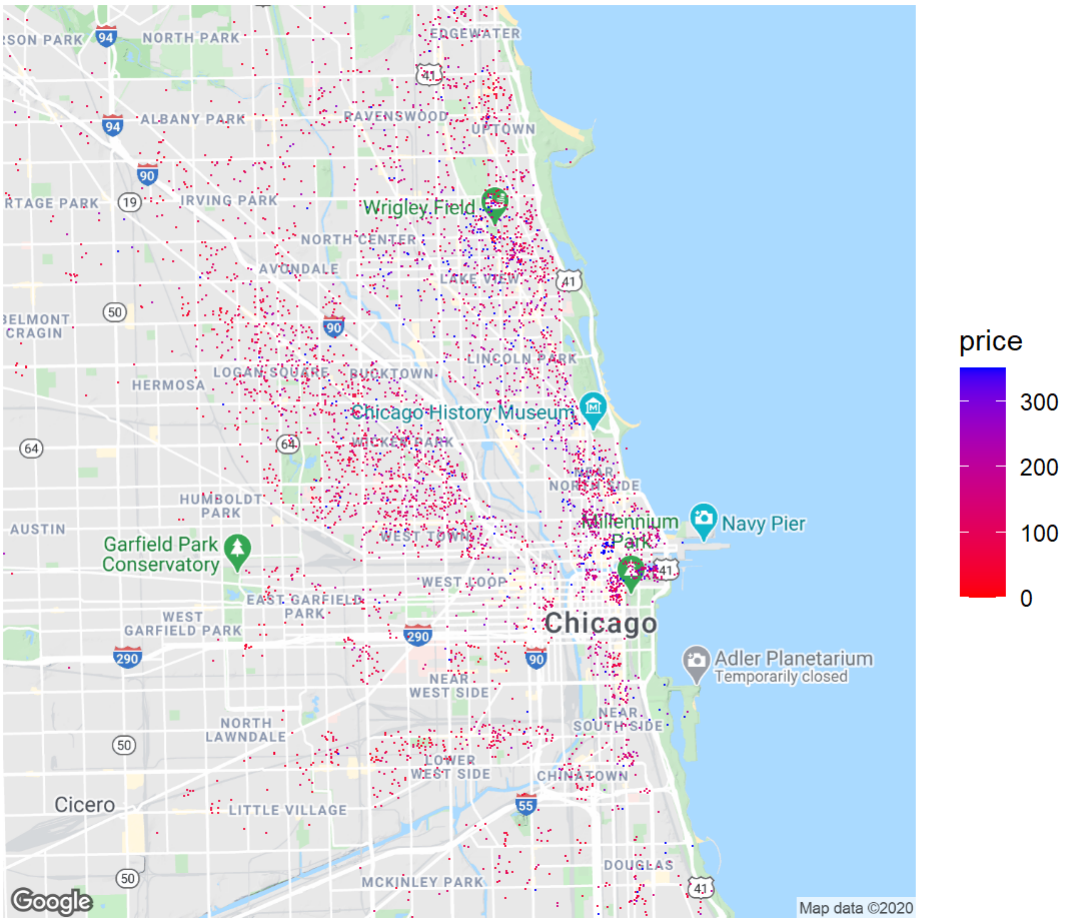


Chicago:Different from all other cities, in Chicago, properties are more expensive along the Michigan Lake bank, with downtown area having the most pricey accommodation.

Chicago Density Heatmap

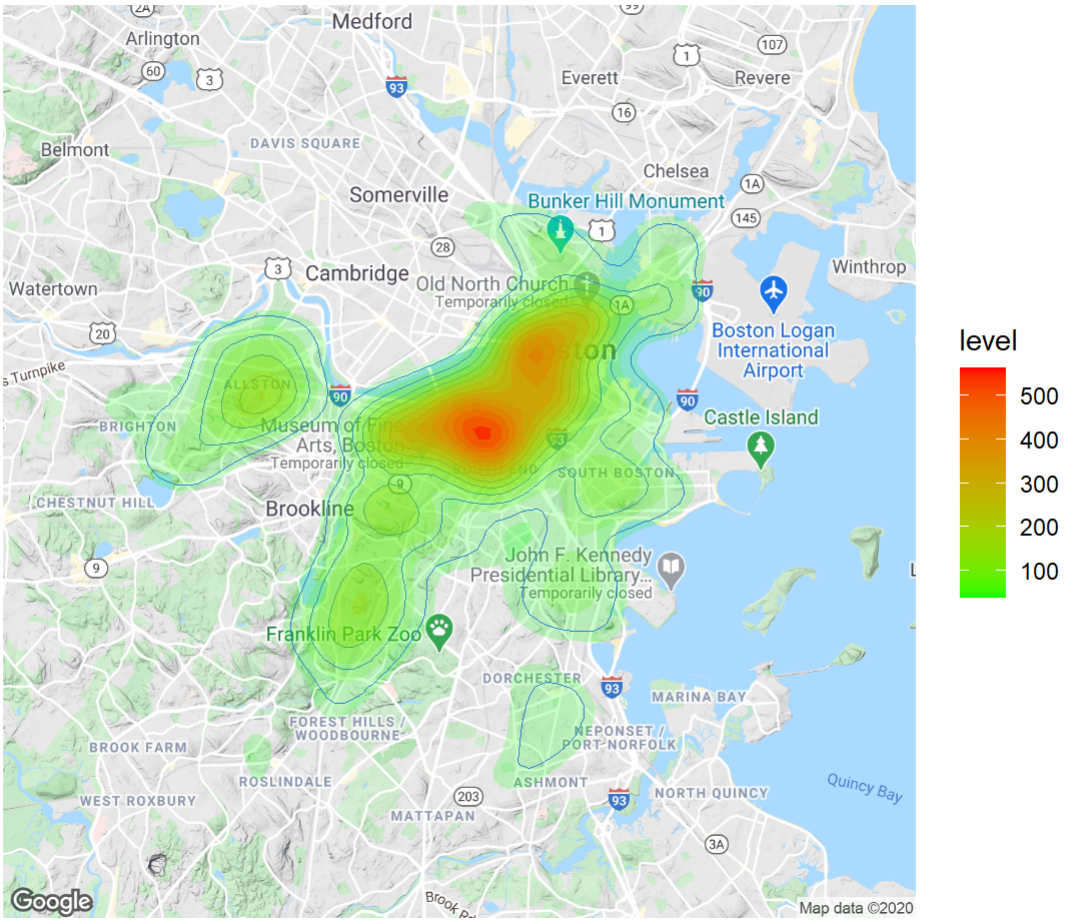


Chicago Price Plot



Boston:Boston’s most expensive properties are located in downtown, as well. Different from other cities, its contour map shows two peaks are very close together.

Boston Density Heatmap



Boston Price Plot

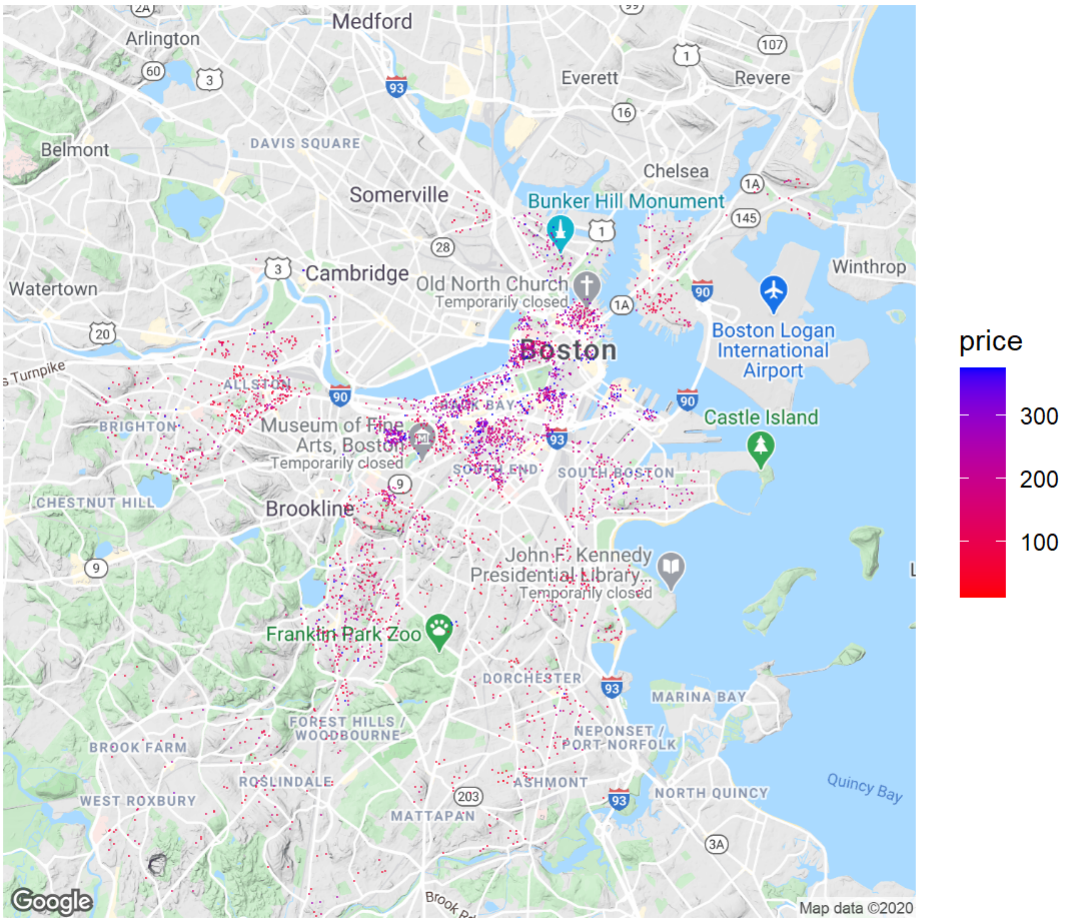
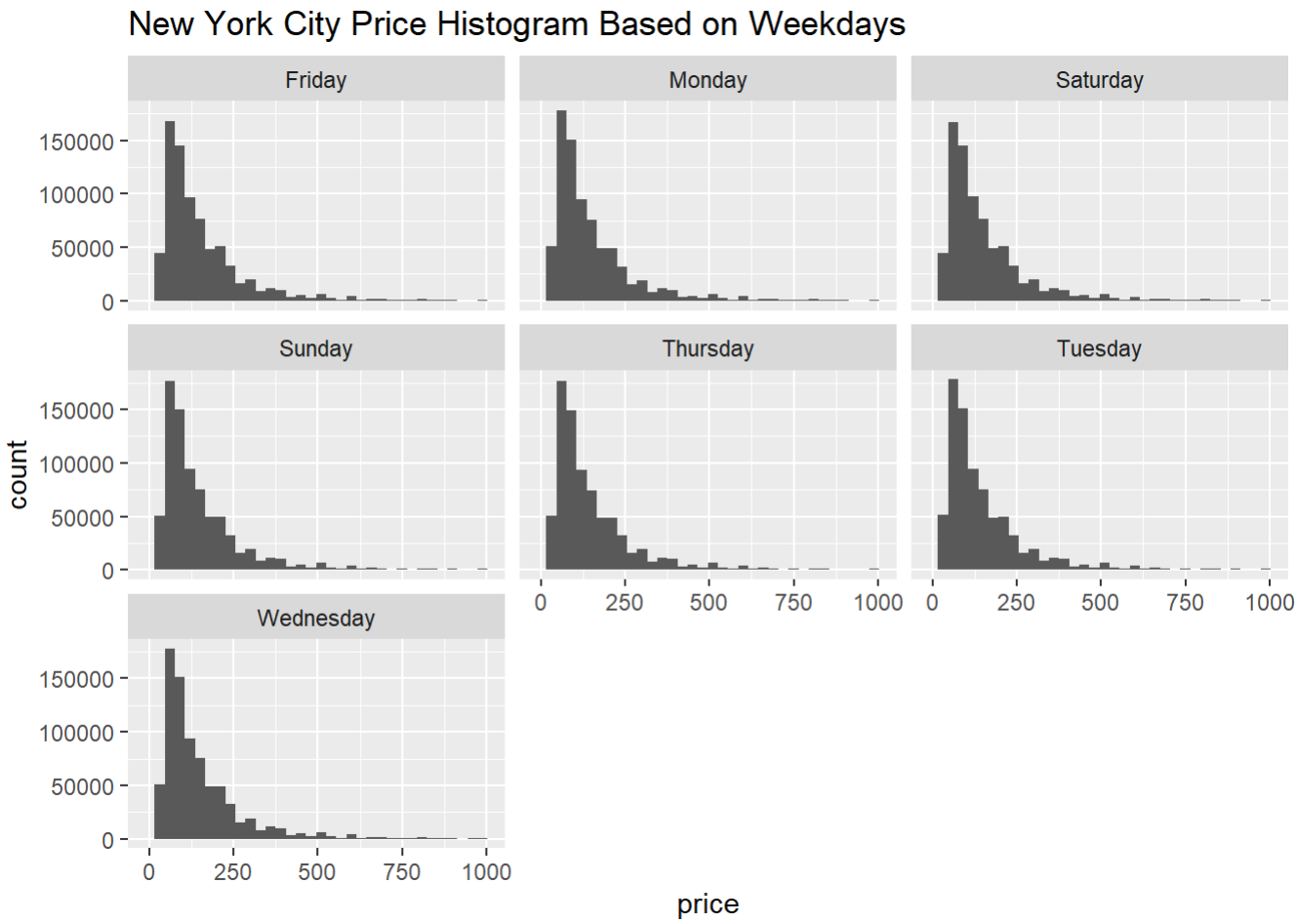
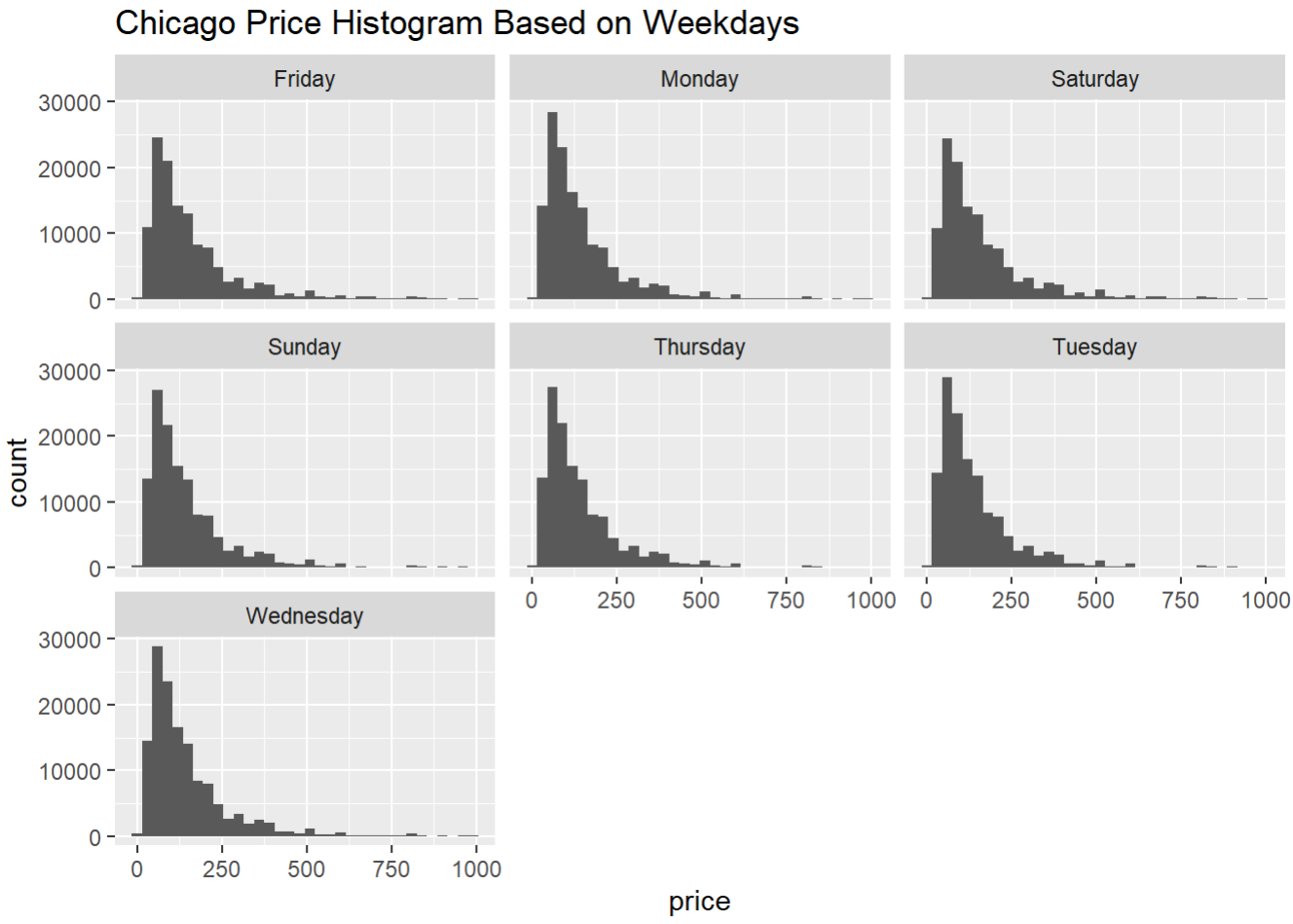


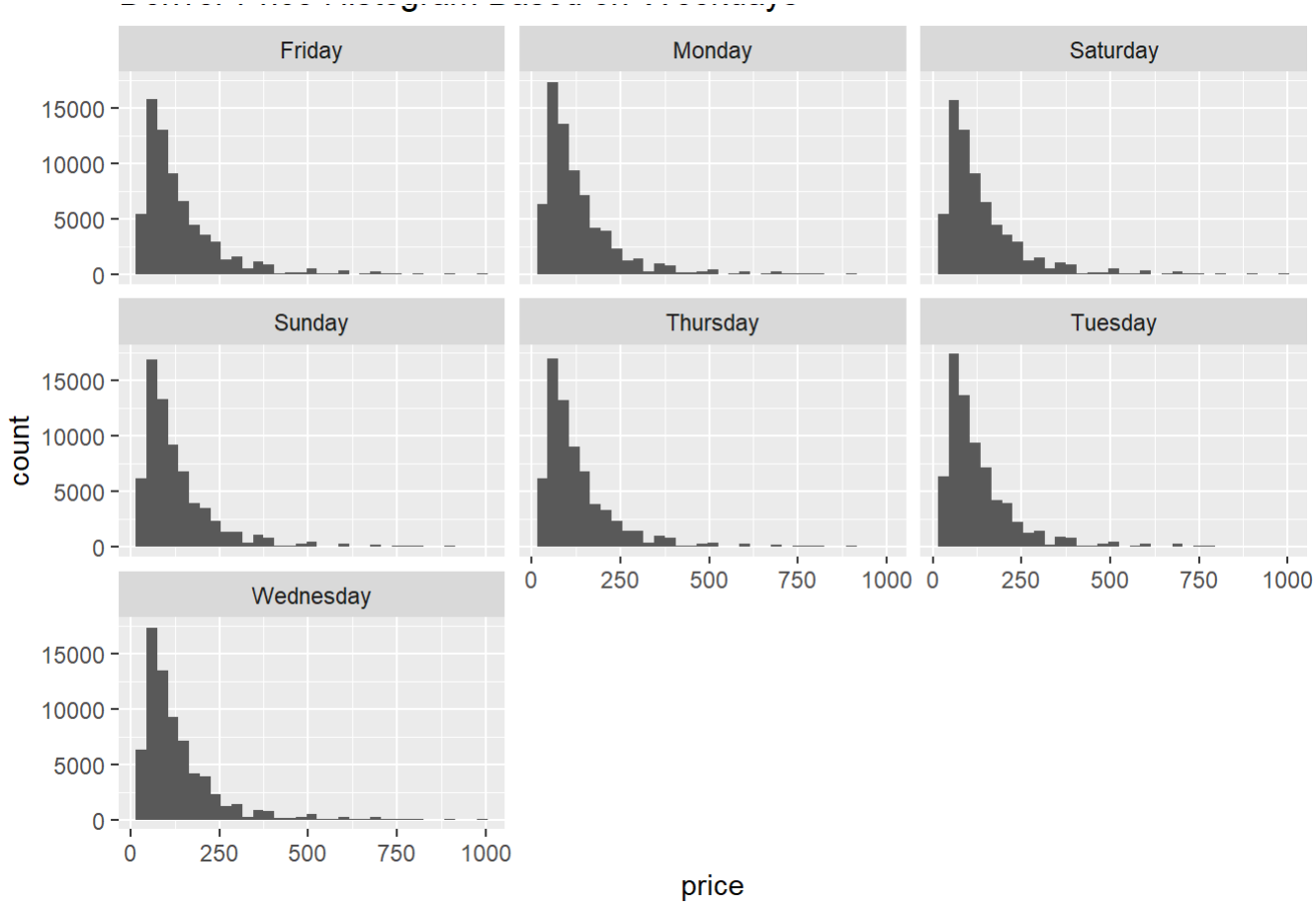
Figure B

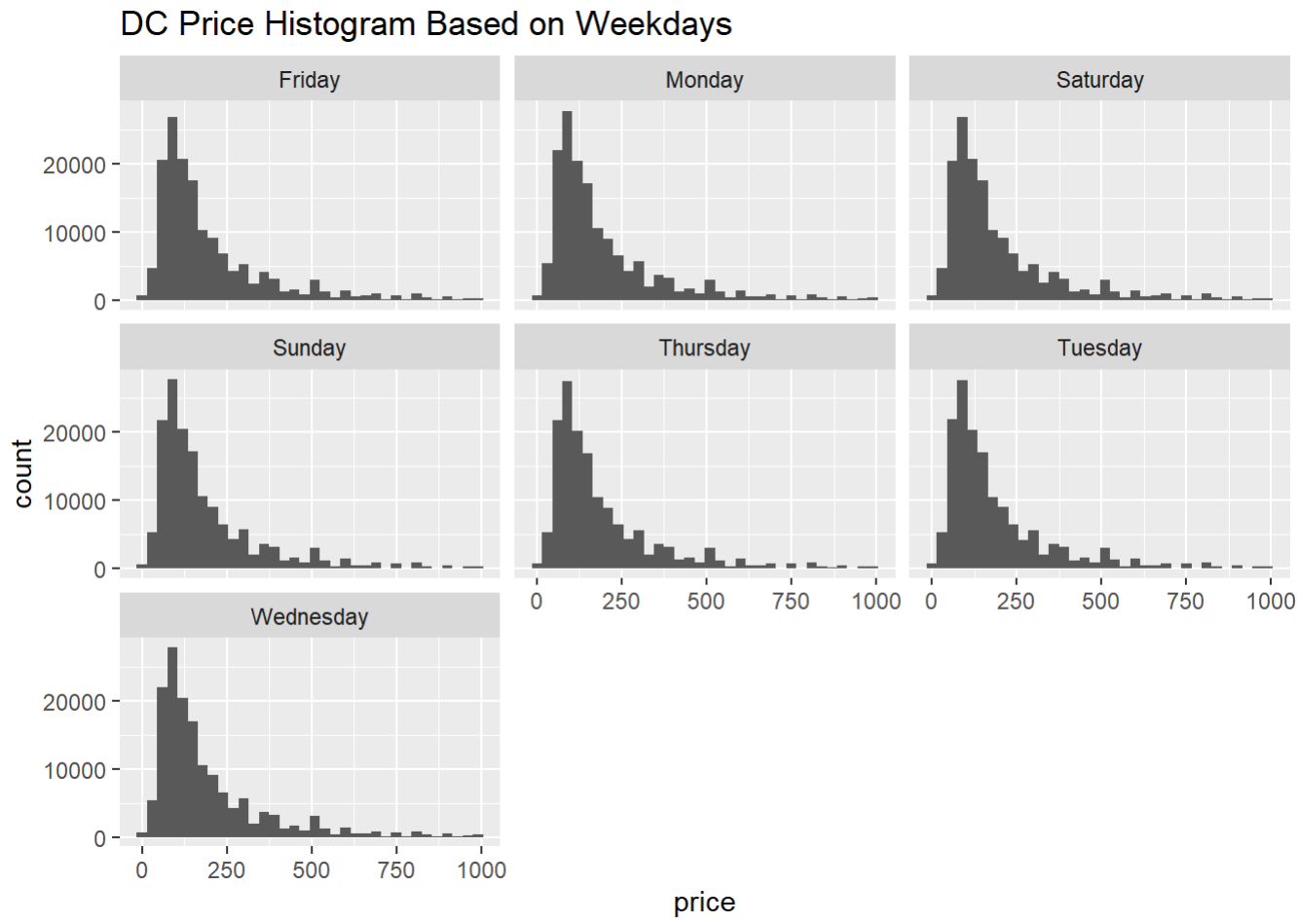
```
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
```

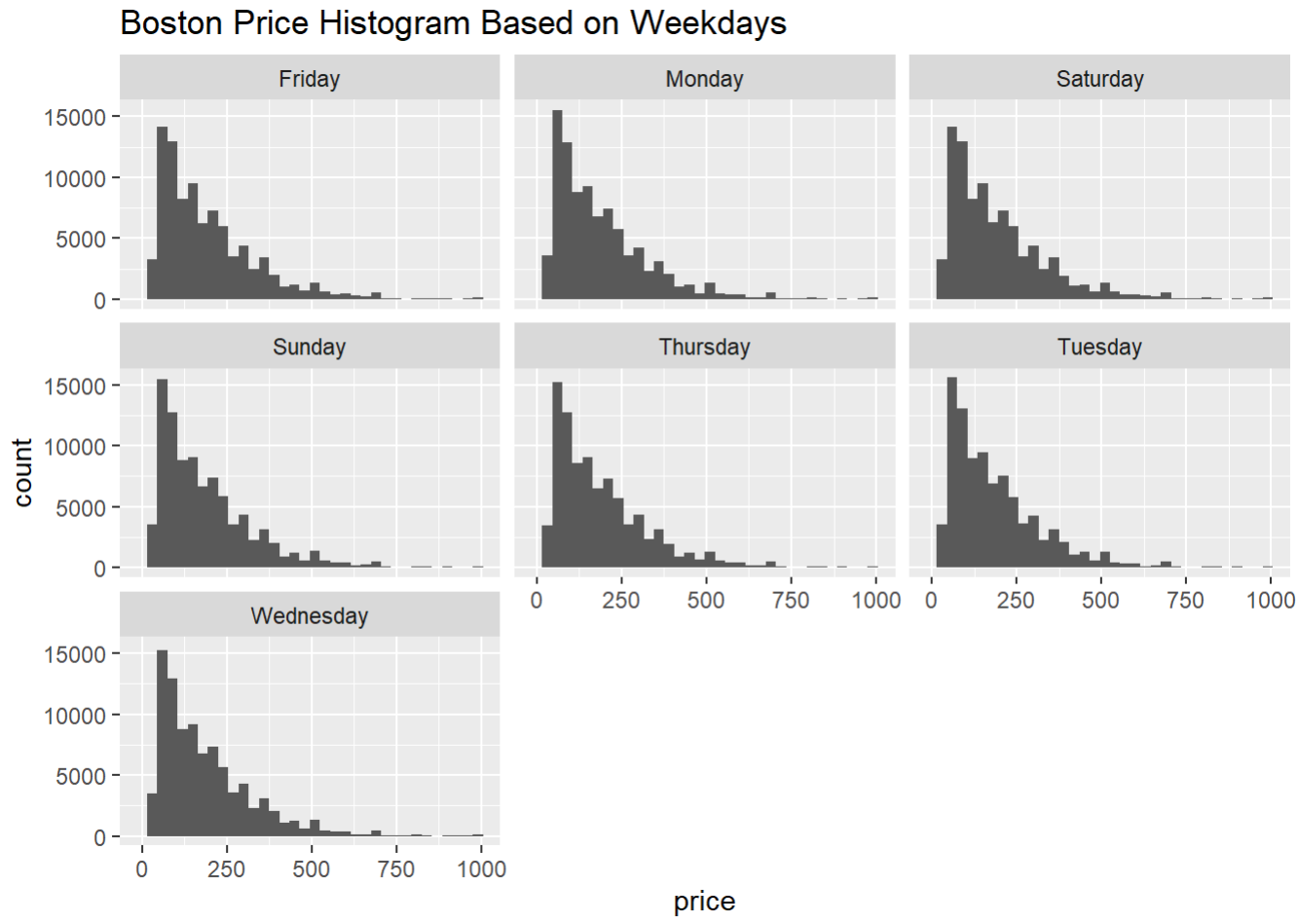




Denver Price Histogram Based on Weekdays


Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
count
price
0
5000
10000
15000
0
250
500
750
1000


DC Price Histogram Based on Weekdays
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
count
price
0
10000
20000
0
250
500
750
1000


Boston Price Histogram Based on Weekdays
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
count
price
0
5000
10000
15000
0
250
500
750
1000

####

Figure C

```
## # A tibble: 7 x 2
##   weekday   `number of available properties`
##   <fct>                                <int>
## 1 Friday                              773357
## 2 Monday                              786396
## 3 Saturday                            774023
## 4 Sunday                              784725
## 5 Thursday                            776854
## 6 Tuesday                             784600
## 7 Wednesday                           784273
```

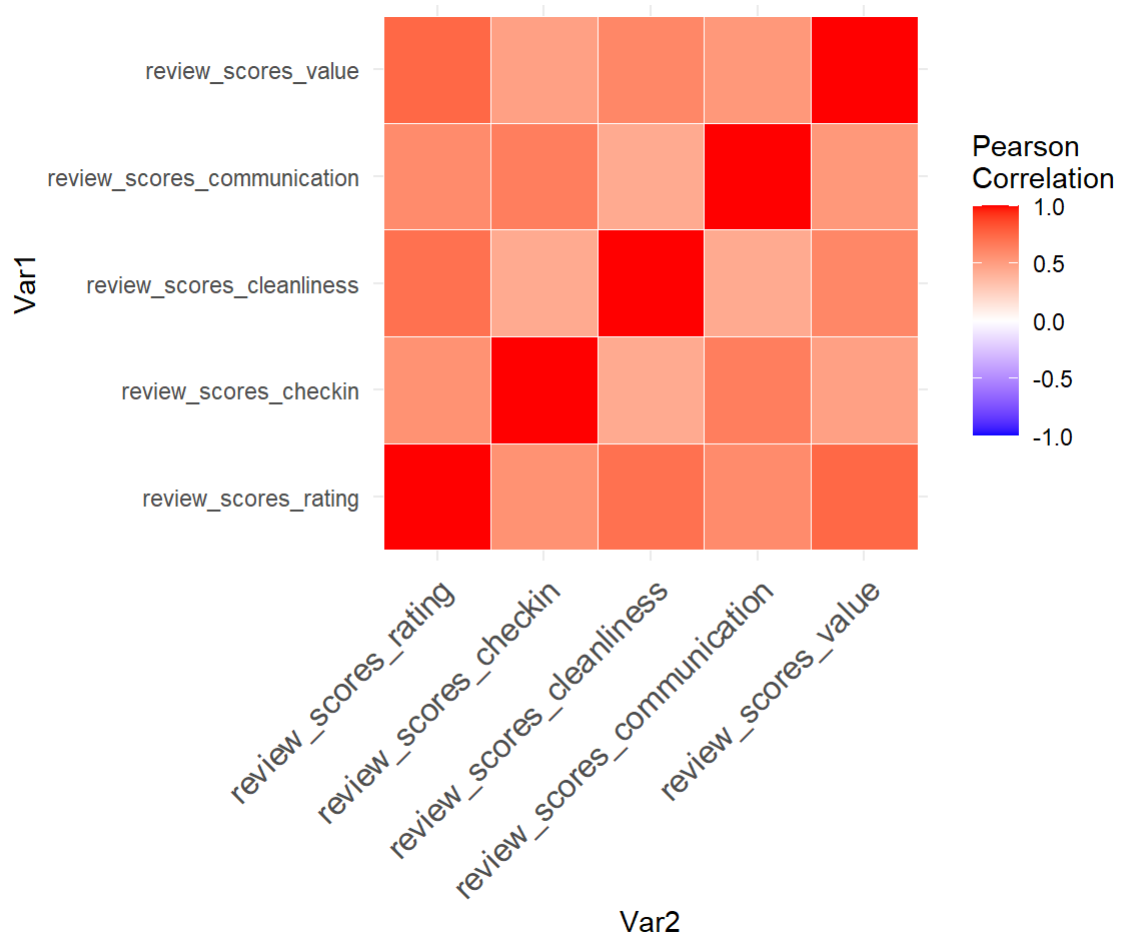
```
## # A tibble: 7 x 2
##   weekday   `number of available properties`
##   <fct>                                <int>
## 1 Friday                              125338
## 2 Monday                              135466
## 3 Saturday                            124458
## 4 Sunday                              130624
## 5 Thursday                            130895
## 6 Tuesday                             136698
## 7 Wednesday                           137272
```

```
## # A tibble: 7 x 2
##   weekday   `number of available properties`
##   <fct>                                <int>
## 1 Friday                               68678
## 2 Monday                               70711
## 3 Saturday                             68298
## 4 Sunday                               69574
## 5 Thursday                             69350
## 6 Tuesday                              71229
## 7 Wednesday                            70665
```

```
## # A tibble: 7 x 2
##   weekday   `number of available properties`
##   <fct>                                <int>
## 1 Friday                              156608
## 2 Monday                              158213
## 3 Saturday                            156649
## 4 Sunday                              158298
## 5 Thursday                            156794
## 6 Tuesday                             157268
## 7 Wednesday                           158101
```

```
## # A tibble: 7 x 2
##   weekday   `number of available properties`
##   <fct>                                <int>
## 1 Friday                               90998
## 2 Monday                               92413
## 3 Saturday                             91068
## 4 Sunday                               92208
## 5 Thursday                             91202
## 6 Tuesday                              93332
## 7 Wednesday                            91816
```

Figure D



"Prediction is very difficult, especially about the future." – Niels Bohr

"All generalizations are false, including this one." –Mark Twain

"In God we trust. All others must bring data." –W. Edwards Deming

Links

1. https://www.theguardian.com/technology/2017/jul/13/airbnb-california-racist-comment-penalty-asian-american (https://www.theguardian.com/technology/2017/jul/13/airbnb-california-racist-comment-penalty-asian-american)