Airbnb Price Analysis STAT232 Final Report

Matthew Barclay

Riley Baumgarten

Megan Dang

Arvind Kamboh

2024-03-20

Introduction

This report delves into an analysis of Airbnb with the aim of understanding the factors that impact pricing. Airbnb, a leading platform in the hospitality industry, offers customers the opportunity to book, list, or discover places to stay worldwide. Understanding the data behind Airbnb can provide key insights into customer preferences, market trends, and business opportunities. The dataset used in this analysis is from Kaggle and comprises information gathered from AirBnB listings. It contains various attributes related to listings, such as location, property type, price, availability, and reviews. This data offers a comprehensive view of the Airbnb ecosystem and can be leveraged to uncover patterns and trends within the platform. Through the analysis of the Airbnb data, we aim to address several key business questions:

- Which property types are in the highest demand among travelers?
- How do prices vary across different locations and property types?

By answering these questions, Airbnb property owners can make informed decisions regarding pricing strategies, and property management to optimize their performance on the Airbnb platform.

Data Source: https://www.kaggle.com/datasets/lovishbansal123/airbnb-data/data (https://www.kaggle.com/datasets/lovishbansal123/airbnb-data/data)

Exploratory Data Analysis

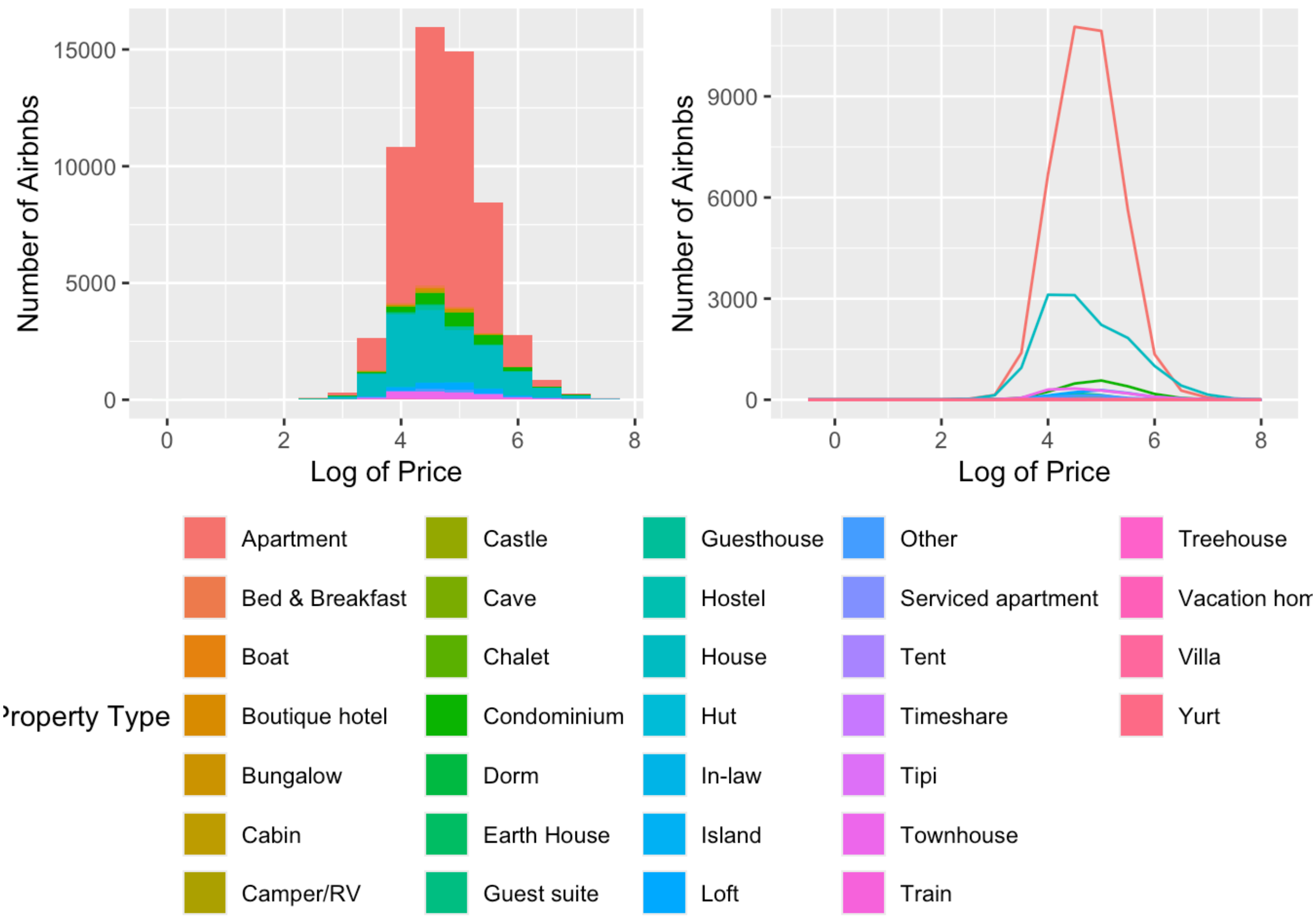
Data Cleaning

The data had a usable format for all variables except *amenities* which was a comma-separated list within each row. We converted this list to encoded rows for each unique value in the amenities lists. This resulted in about 120 new columns in our data frame. To simplify the data, we combined similar amenities into broad categories. This reduced the number of amenities columns to 28, with each column representing the percentage of amenities present in the listing for each listing. This transformation changes the interpretation of the predictors from binary categorical to a continuous scale where adding additional amenities in the predictor group will increase or decrease the log price by that coefficient times the percent of amenities added in the group.

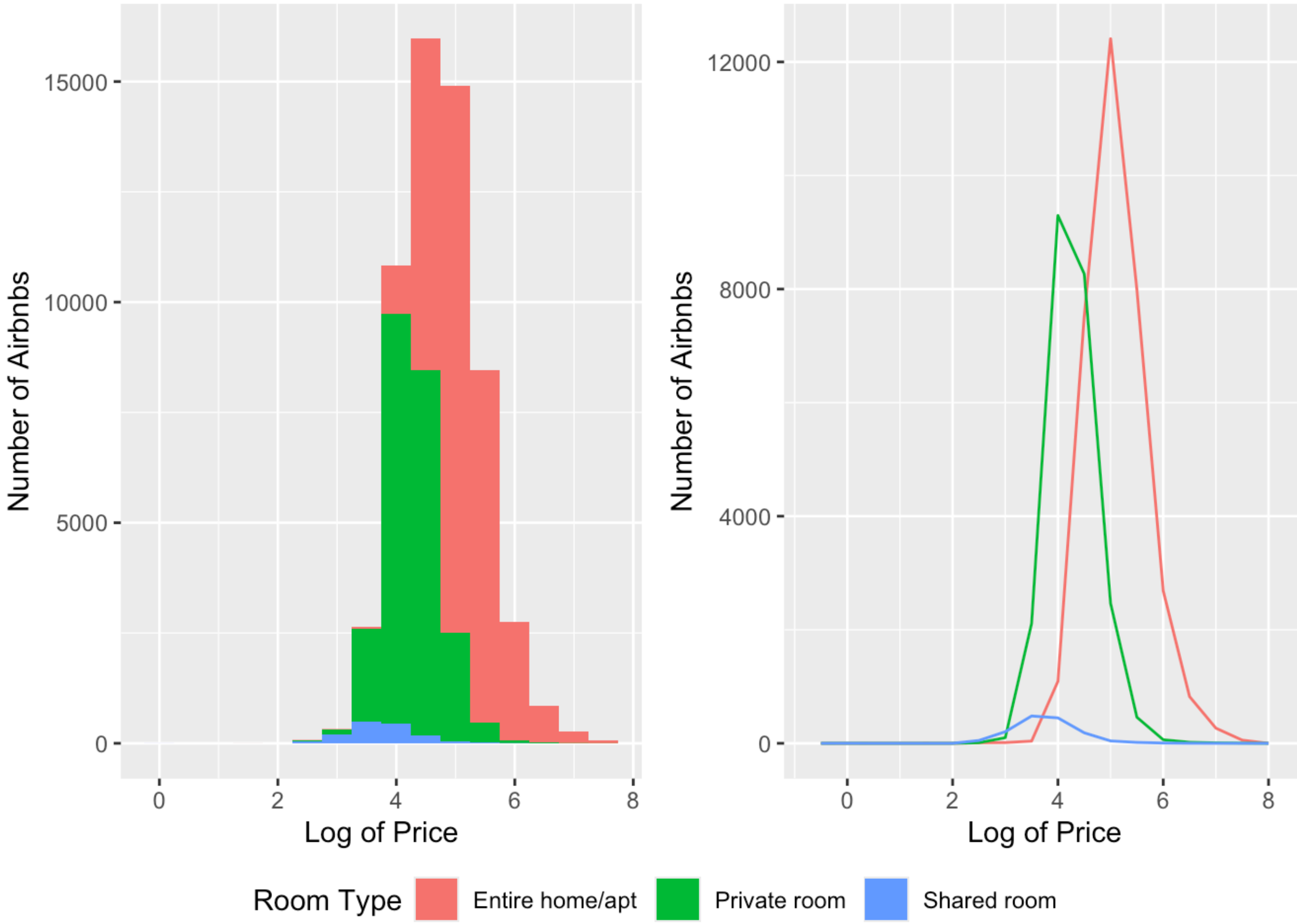
Research Question

> What are the factors influencing Airbnb nightly prices?

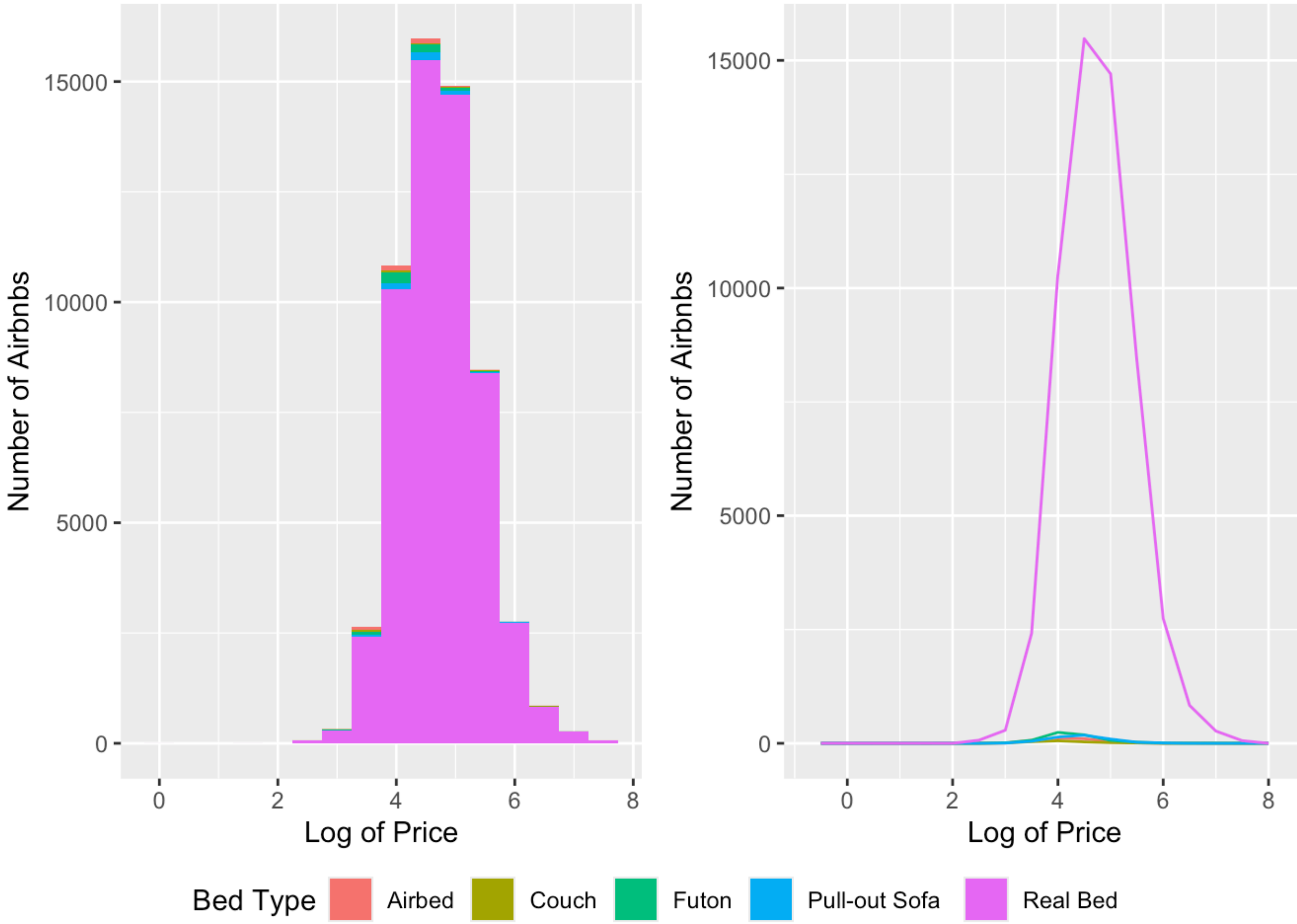
Data Visualization



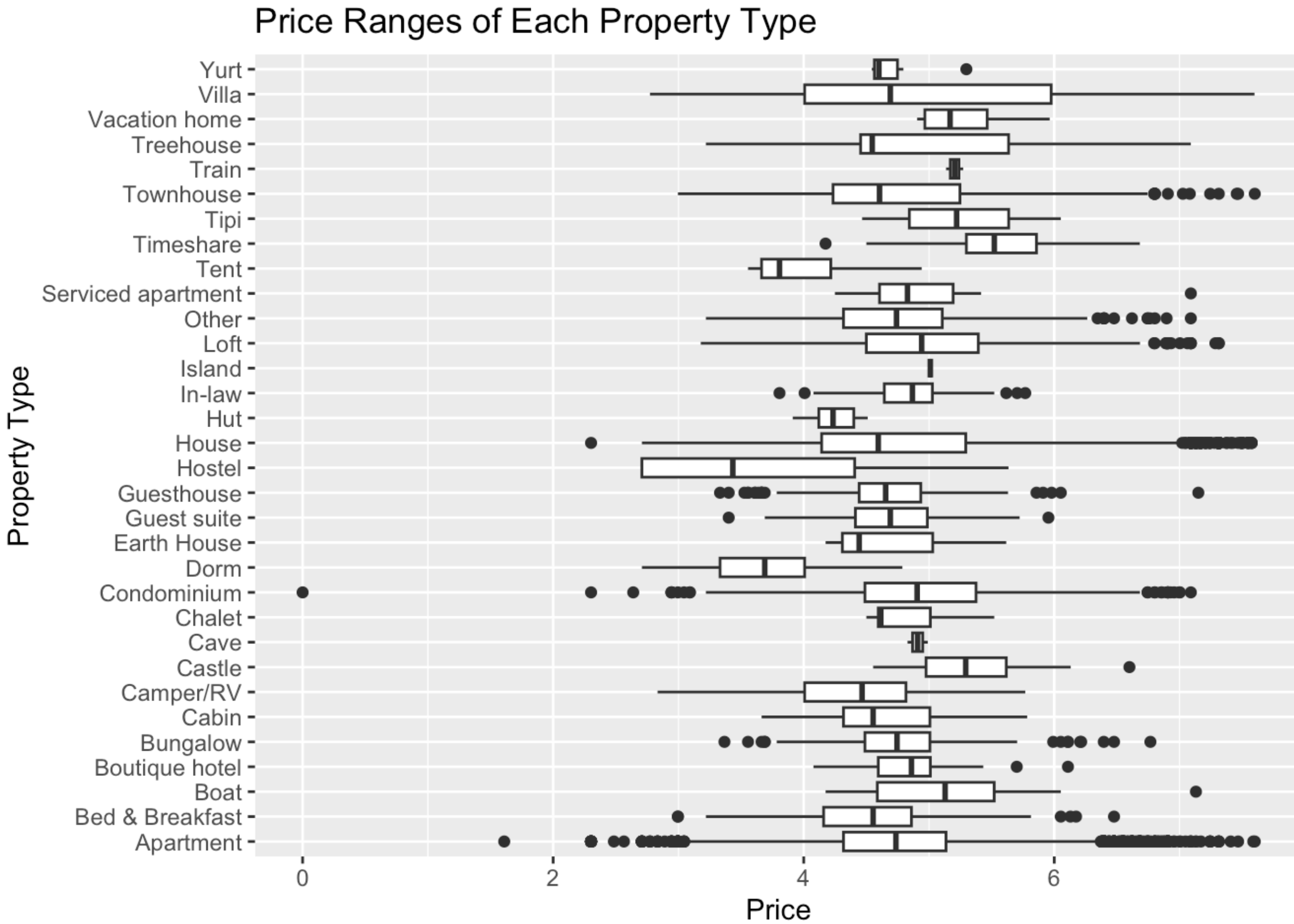
These plots show the distribution of property types against price. We can see that the most popular type of property is apartments. The distribution of property type by price does not appear to have any distinct patterns with all property types having similar price distributions.



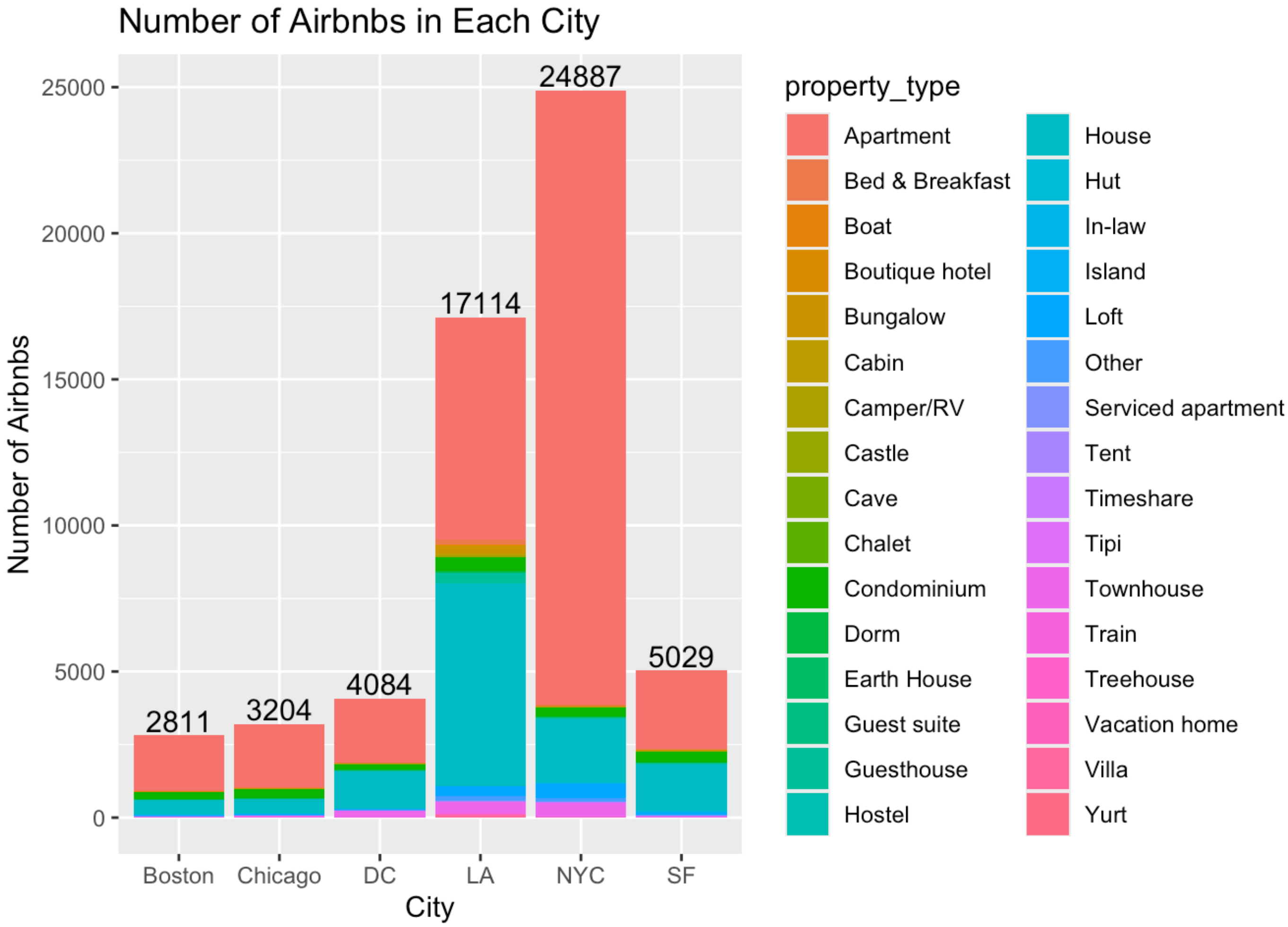
These plots show the distribution of room type across prices. It shows that entire apartments/homes are most prevalent and command a higher nightly rate, followed by private rooms and shared rooms being the least common and cheapest.



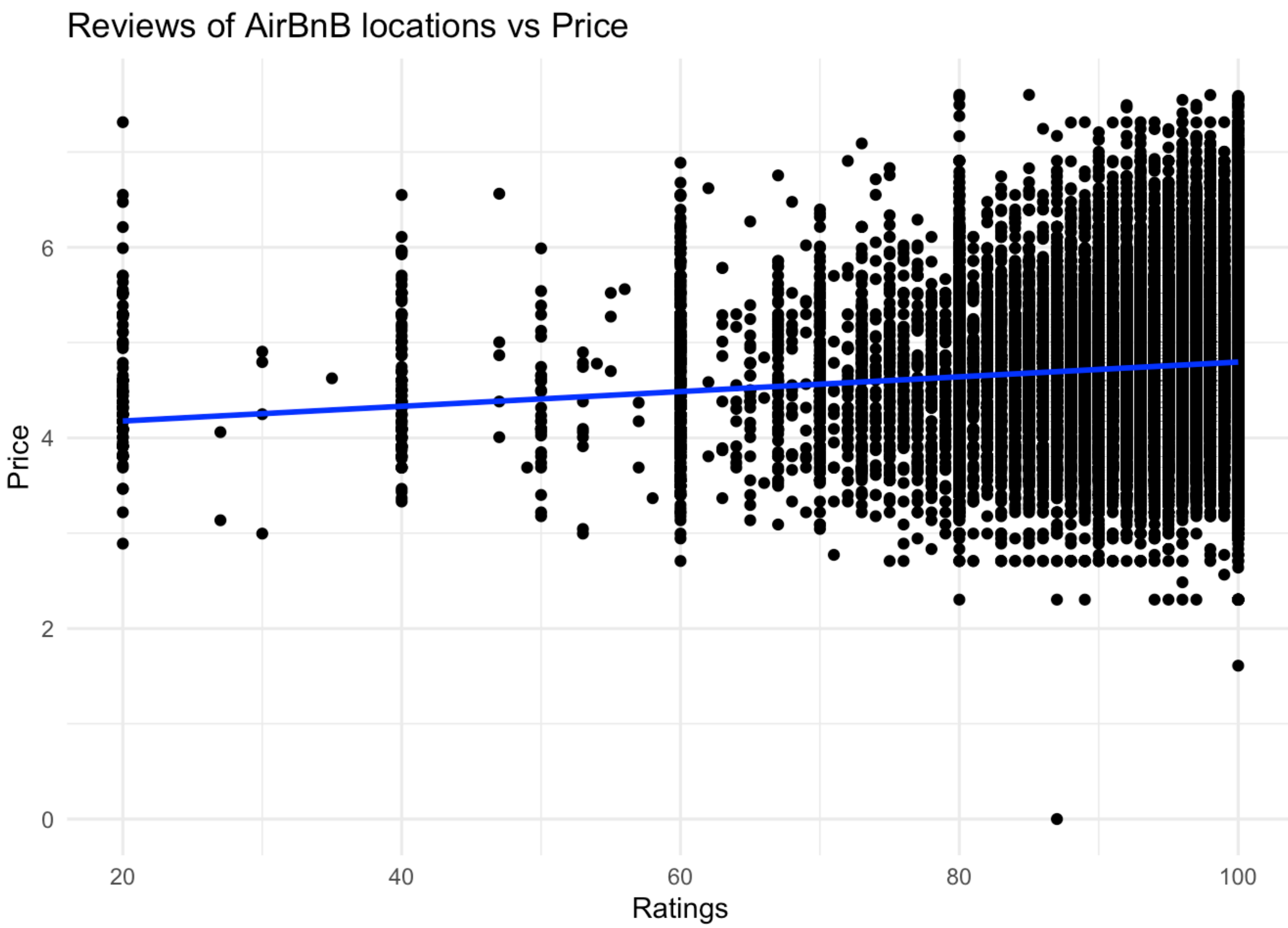
These plots show that real beds are most common in Airbnbs. There does not appear to be a relationship between the bed type and price which was unexpected, as we believed that sleeping on couches and airbeds would drive down the nightly prices.



This boxplot shows the range of price for each property type. While most property types have similar distributions, notable exceptions are tents, hostels, and dorms which have much lower prices in general. With these property types being the least luxurious and often shared with many other guests, the lower pricing is reasonable.

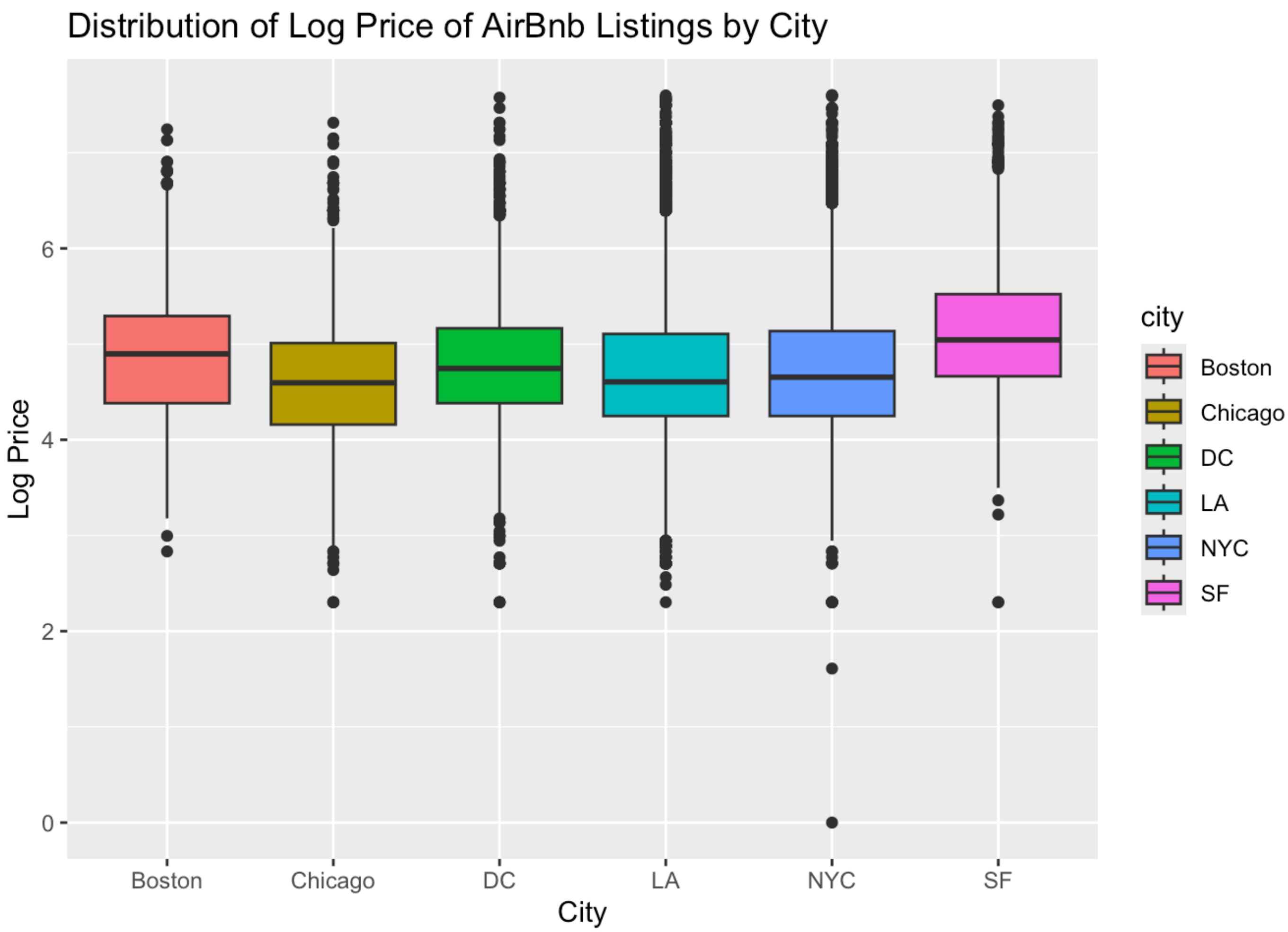


This bar plot shows the number of Airbnbs in each city. NYC has the most Airbnbs, with LA closely following behind.

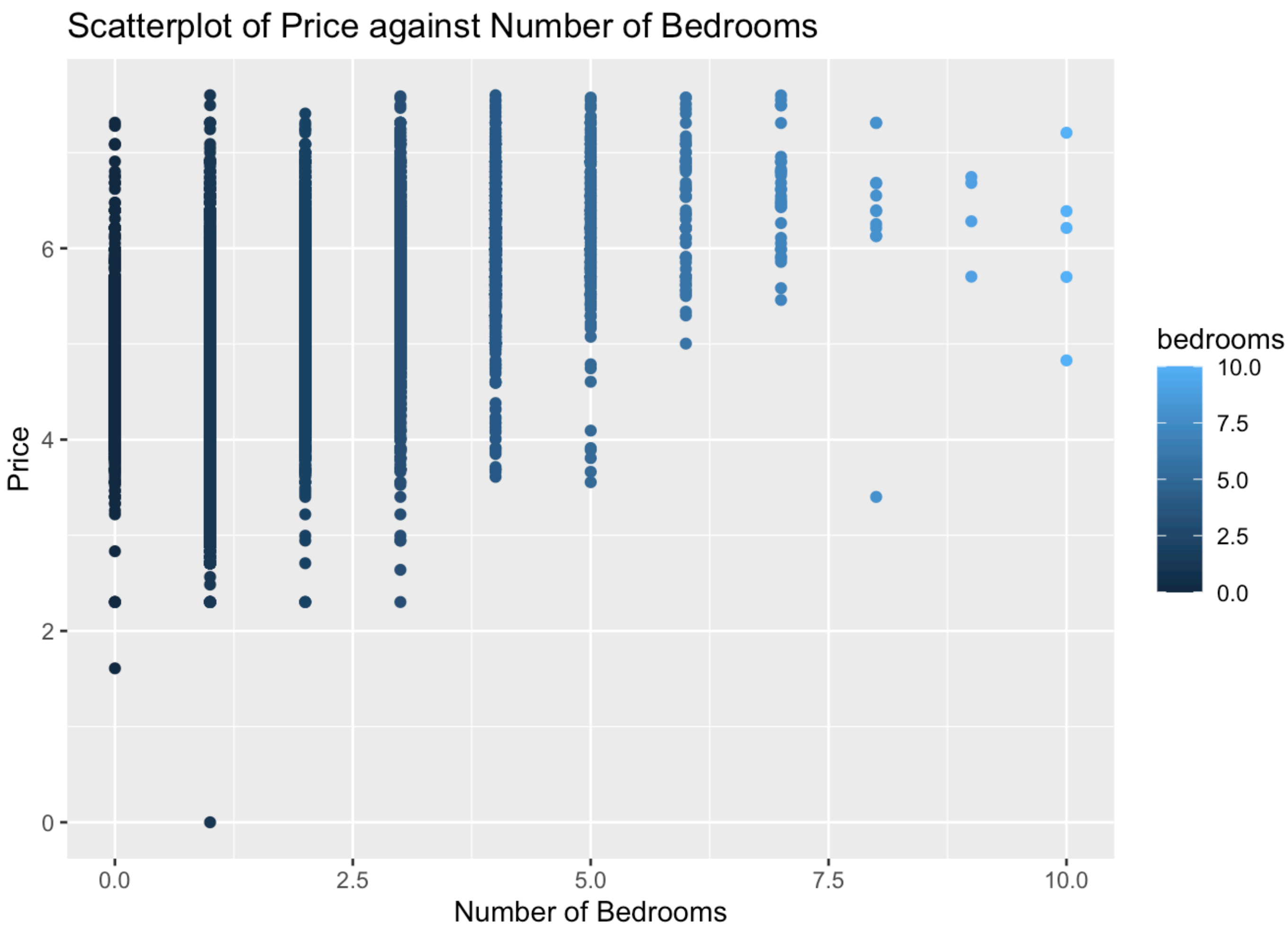


This scatter plot shows the relationship between the ratings the AirBnB locations received and how it relates to price. From the plot we can see a linear relationship and that the lower-rated locations have lower prices. Price steadily increases as the reviews get higher.

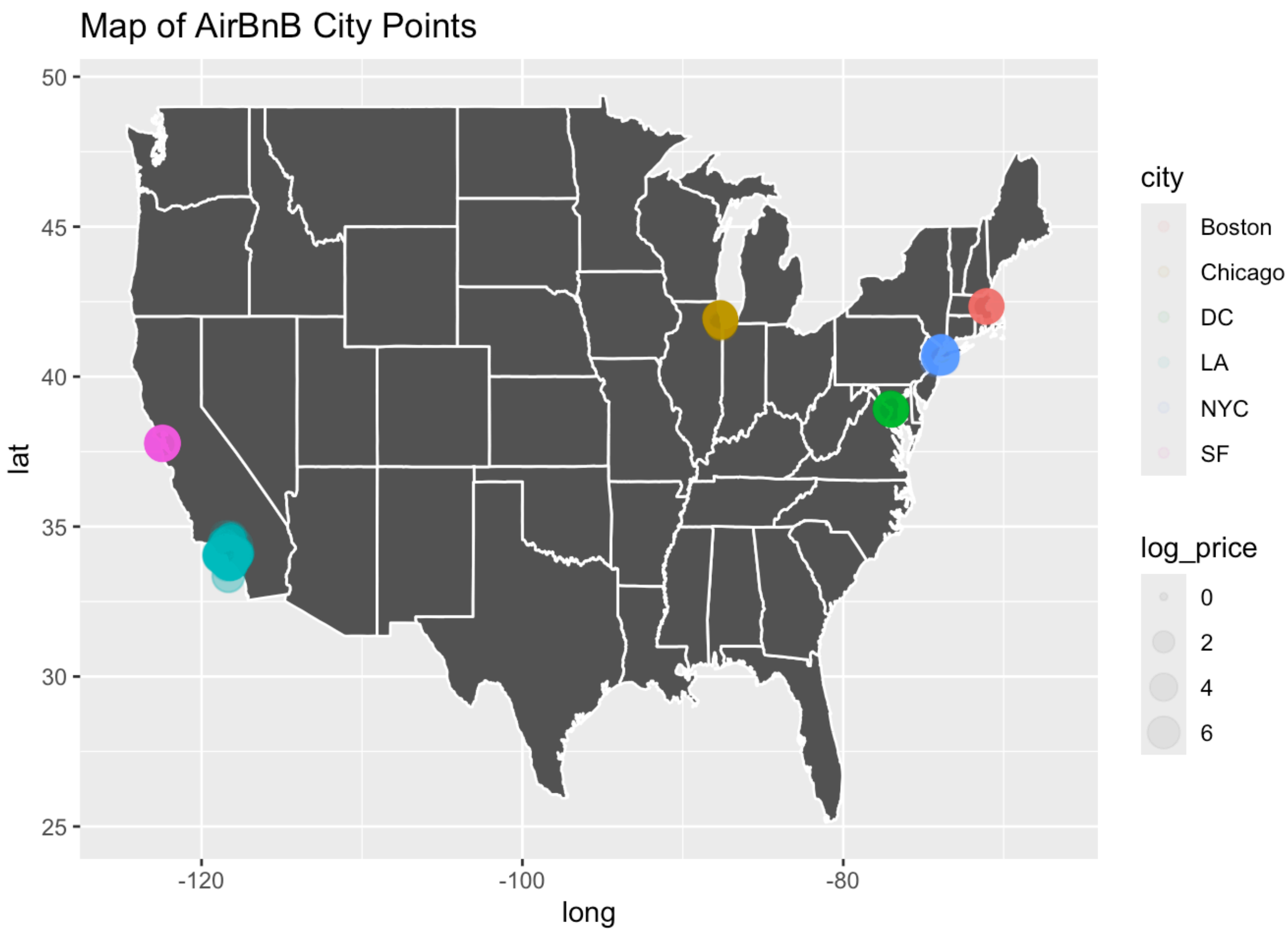

Mean Log Price of AirBnb Listings by City
Mean Log Price
0
1
2
3
4
5
Boston
Chicago
DC
LA
NYC
SF
City
city
Boston
Chicago
DC
LA
NYC
SF



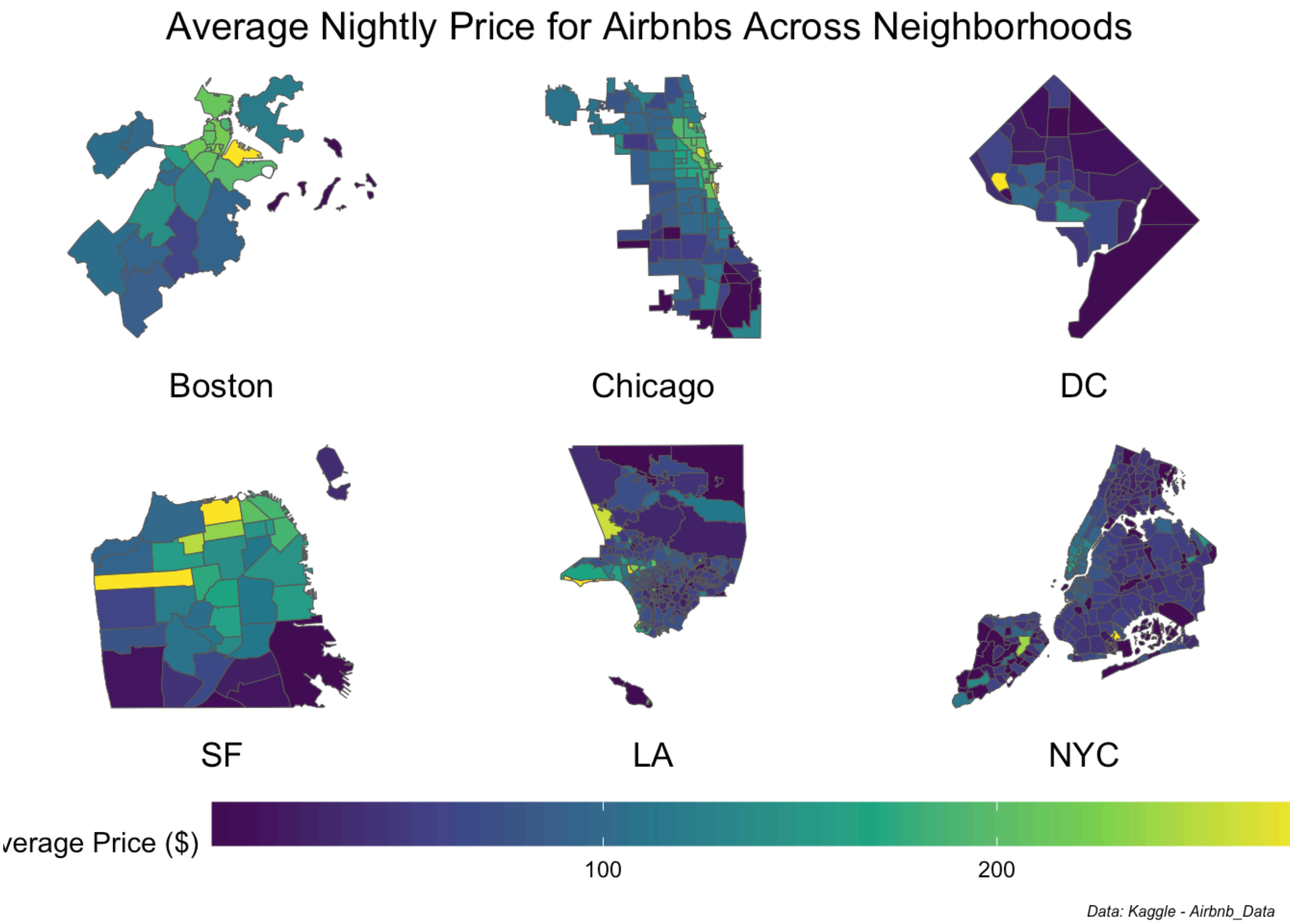
This boxplot shows the relationship between the AirBnB locations based on city and how it relates to price. Though the average price is similar across the cities, San Francisco has the highest price while Chicago has the lowest, which aligns with higher cost of living in coastal cities.

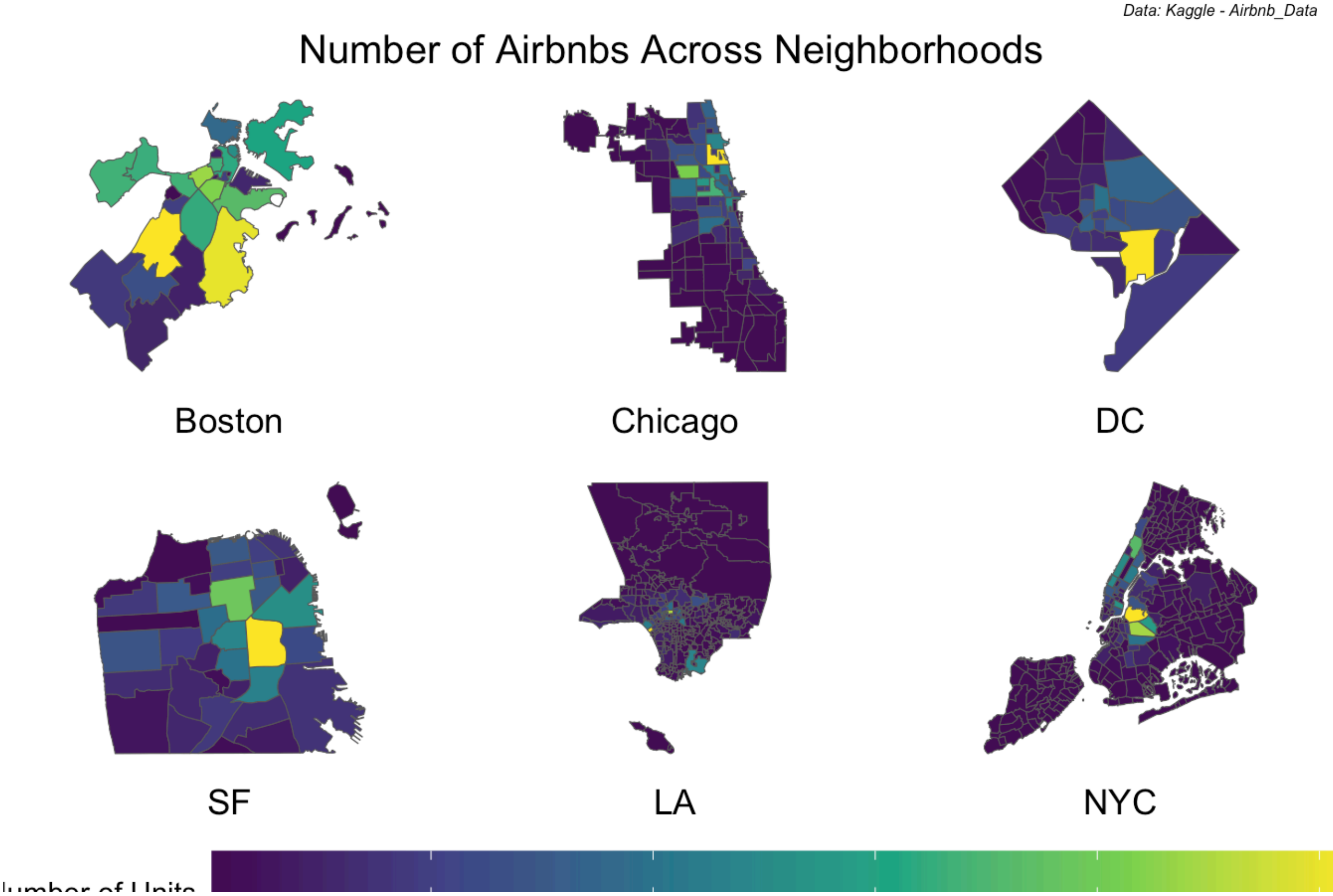


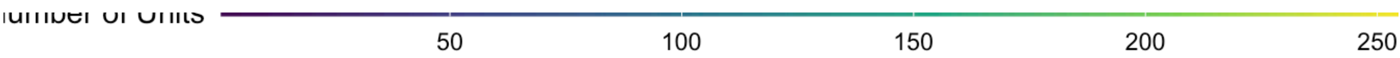
This scatterplot shows the relationship between the number of bedrooms in an AirBnB location and the price. There appears to be diminishing returns on how high the price will be increased as the number of bedrooms increases as there is a plateau after about 7 bedrooms.



This map shows the points across the US where the airbnb locations are based on the dataset. The size of the points are adjusted by the price of the airbnb locations.


Average Nightly Price for Airbnbs Across Neighborhoods
Boston
Chicago
DC
SF
LA
NYC
verage Price ($)
100
200
Data: Kaggle - Airbnb_Data


Number of Airbnbs Across Neighborhoods
Boston
Chicago
DC
SF
LA
NYC

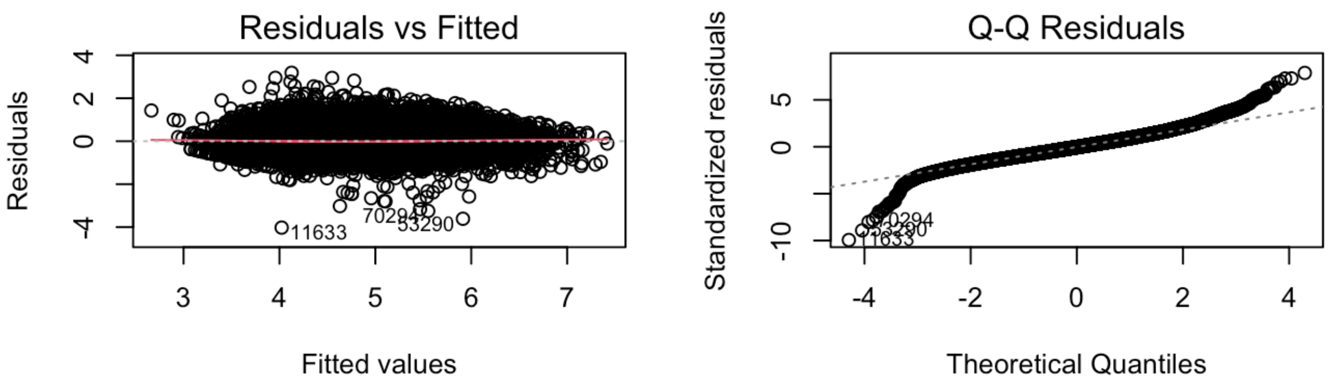


Data: Kaggle - Airbnb_Data

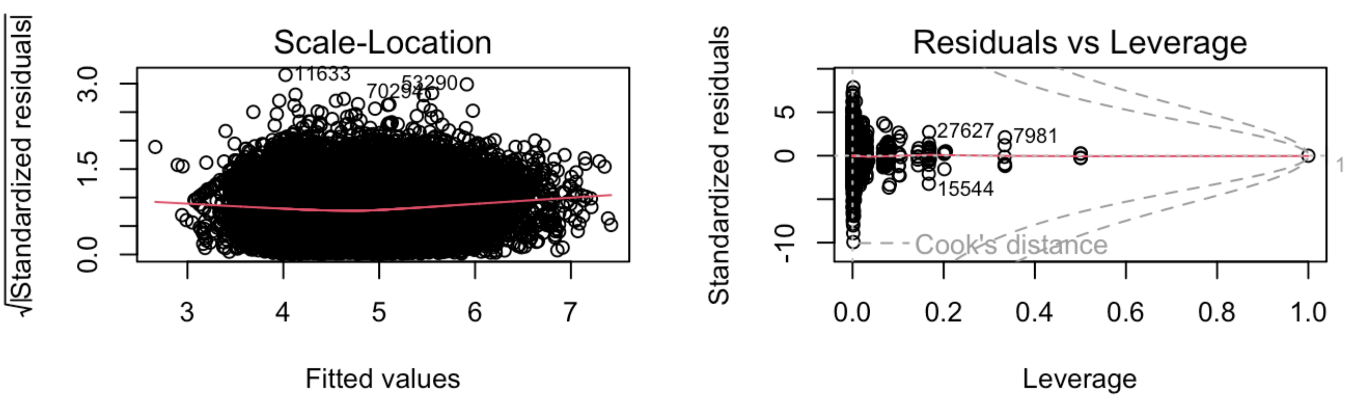
While nightly rates differ between cities, they also differ within the city. While we will not use this neighborhood-level data in our modeling, it is useful to visualize where the most Airbnbs are located and where the average prices are distributed. In each city, the majority of Airbnbs appear to be located in just one neighborhood but the average prices in that neighborhood are lower. The highest average prices appear either along the beach or in the city center, where we would expect to see the higher prices to be located.

Models

Multiple Linear Regression

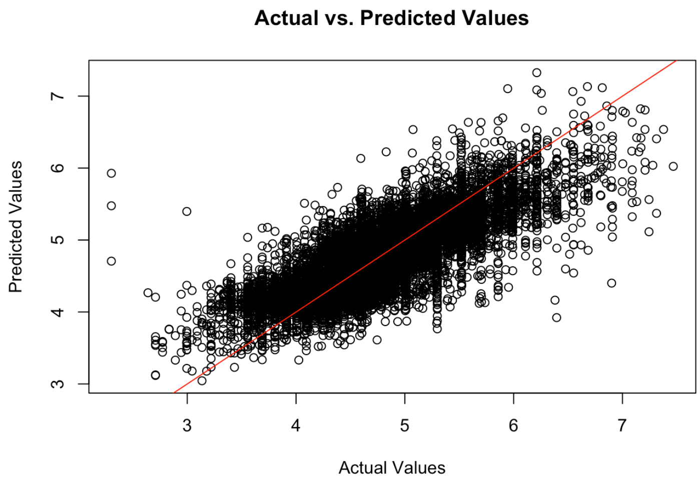


Here



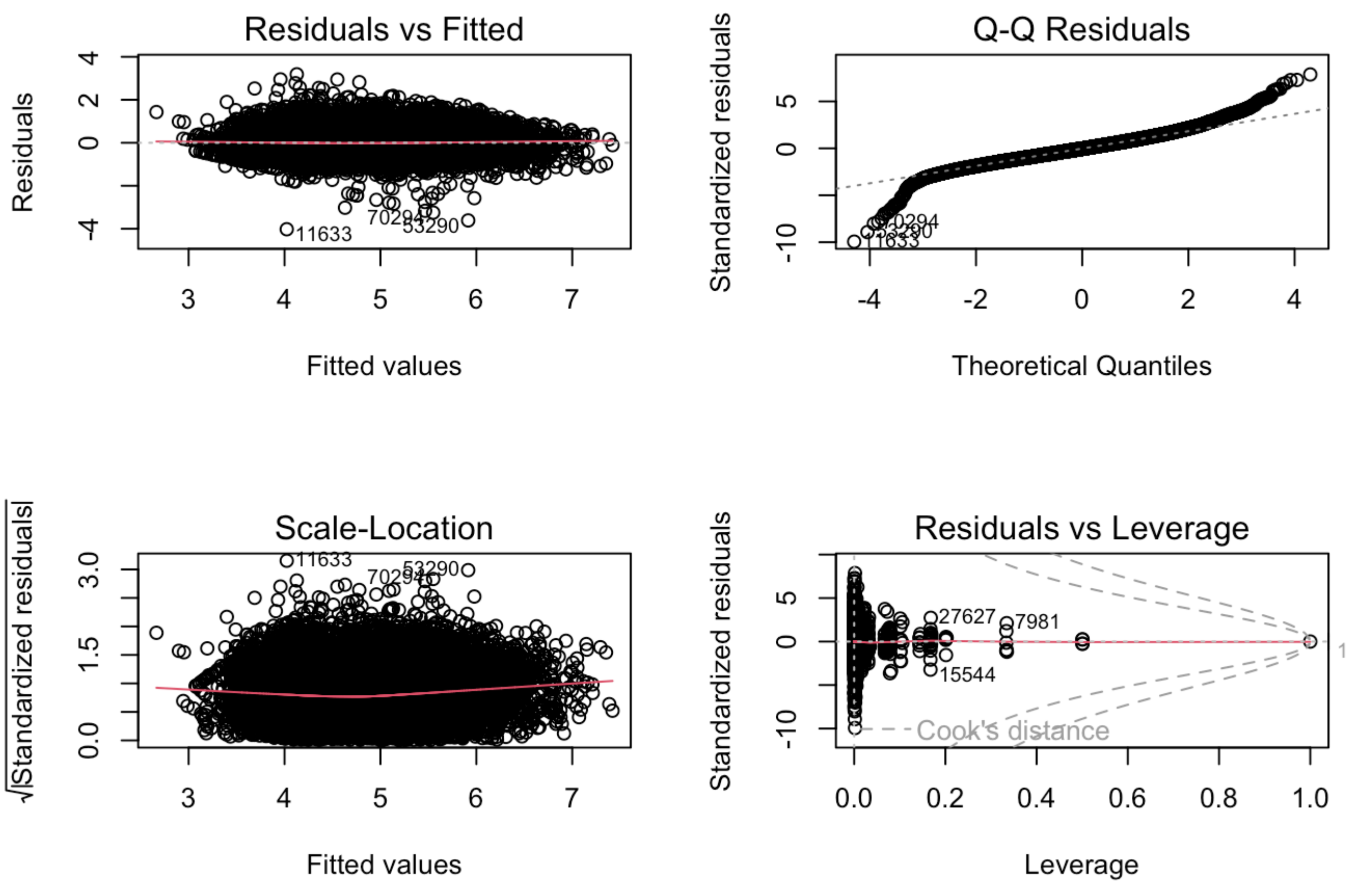
is a simple linear regression model, we excluded zip code, longitude and latitude because of their inherent unimportance.

Cross-Validated MLR



We then created a train and test set for multilinear regression with a 80 train 20 test partitioning. In this model, none of the variables were removed and we see a very similar R^2 from the simple model which is a good sign we are not overfitting.

MLR- Removed insignificant predictors



	MLR	Cross-Validated MLR	Reduced MLR

(Intercept)	3.909 *** (0.000)	3.909 *** (0.000)	3.910 *** (0.000)
property_typeBed & Breakfast	0.120 *** (0.000)	0.098 *** (0.000)	0.120 *** (0.000)
property_typeBoat	0.346 *** (0.000)	0.256 *** (0.000)	0.346 *** (0.000)
property_typeBoutique hotel	0.162 ** (0.006)	0.195 ** (0.005)	0.162 ** (0.006)
property_typeBungalow	0.034 (0.145)	0.057 * (0.031)	0.034 (0.145)
property_typeCabin	-0.007 (0.892)	-0.006 (0.919)	-0.007 (0.886)
property_typeCamper/RV	-0.125 * (0.012)	-0.121 * (0.028)	-0.126 * (0.012)
property_typeCastle	0.421 *** (0.000)	0.258 * (0.044)	0.422 *** (0.000)
property_typeCave	0.407 (0.156)	0.421 (0.142)	0.407 (0.156)
property_typeChalet	0.148 (0.414)	0.034 (0.884)	0.150 (0.407)
property_typeCondominium	0.069 *** (0.000)	0.074 *** (0.000)	0.069 *** (0.000)
property_typeDorm	-0.378 *** (0.000)	-0.383 *** (0.000)	-0.378 *** (0.000)
property_typeEarth House	0.131 (0.575)	0.529 (0.192)	0.130 (0.577)
property_typeGuest suite	-0.030 (0.462)	0.021 (0.636)	-0.033 (0.417)
property_typeGuesthouse	0.000 (0.984)	0.010 (0.651)	0.001 (0.975)
property_typeHostel	-0.580 *** (0.000)	-0.524 *** (0.000)	-0.580 *** (0.000)
property_typeHouse	-0.010 * (0.038)	-0.009 (0.113)	-0.011 * (0.034)
property_typeHut	-0.303 * (0.048)	-0.447 (0.056)	-0.303 * (0.048)
property_typeIn-law	-0.115 * (0.026)	-0.078 (0.188)	-0.117 * (0.023)
property_typeIsland	0.894 * (0.027)	0.899 * (0.027)	0.894 * (0.027)
property_typeLoft	0.154 *** (0.000)	0.153 *** (0.000)	0.154 *** (0.000)
property_typeOther	0.054 ** (0.007)	0.054 * (0.019)	0.054 ** (0.007)
property_typeServiced apartment	0.139 (0.185)	0.030 (0.795)	0.137 (0.190)
property_typeTent	-0.308 ** (0.005)	-0.285 * (0.015)	-0.309 ** (0.004)
property_typeTimeshare	0.415 *** (0.000)	0.411 *** (0.000)	0.415 *** (0.000)
property_typeTipi	0.747 ** (0.001)	0.949 *** (0.001)	0.747 ** (0.001)

property_typeTownhouse	0.004 (0.743)	-0.003 (0.846)	0.004 (0.746)
property_typeTrain	0.701 * (0.014)	0.711 * (0.013)	0.700 * (0.015)
property_typeTreehouse	0.210 (0.205)	0.207 (0.210)	0.209 (0.207)
property_typeVacation home	0.333 * (0.044)	0.321 (0.113)	0.331 * (0.046)
property_typeVilla	0.039 (0.282)	-0.016 (0.713)	0.039 (0.283)
property_typeYurt	0.228 (0.168)	0.236 (0.244)	0.229 (0.167)
room_typePrivate room	-0.576 *** (0.000)	-0.577 *** (0.000)	-0.576 *** (0.000)
room_typeShared room	-1.039 *** (0.000)	-1.044 *** (0.000)	-1.039 *** (0.000)
accommodates	0.070 *** (0.000)	0.069 *** (0.000)	0.071 *** (0.000)
bathrooms	0.123 *** (0.000)	0.124 *** (0.000)	0.123 *** (0.000)
bed_typeCouch	0.056 (0.156)	0.015 (0.736)	0.056 (0.156)
bed_typeFuton	-0.002 (0.936)	-0.019 (0.533)	-0.002 (0.935)
bed_typePull-out Sofa	0.047 (0.099)	0.041 (0.201)	0.047 (0.098)
bed_typeReal Bed	0.051 * (0.022)	0.047 (0.061)	0.051 * (0.022)
cancellation_policymoderate	0.011 * (0.027)	0.007 (0.189)	0.011 * (0.022)
cancellation_policystrict	0.038 *** (0.000)	0.037 *** (0.000)	0.038 *** (0.000)
cancellation_policysuper_strict_30	0.254 *** (0.000)	0.215 *** (0.000)	0.254 *** (0.000)
cancellation_policysuper_strict_60	0.702 *** (0.000)	0.987 *** (0.000)	0.704 *** (0.000)
cleaning_feeTrue	0.001 (0.813)	0.002 (0.754)	
cityChicago	-0.343 *** (0.000)	-0.349 *** (0.000)	-0.343 *** (0.000)
cityDC	-0.135 *** (0.000)	-0.137 *** (0.000)	-0.135 *** (0.000)
cityLA	-0.176 *** (0.000)	-0.176 *** (0.000)	-0.176 *** (0.000)
cityNYC	-0.019 * (0.023)	-0.022 * (0.018)	-0.019 * (0.024)
citySF	0.273 *** (0.000)	0.267 *** (0.000)	0.273 *** (0.000)
instant_bookablet	-0.042 *** (0.000)	-0.042 *** (0.000)	-0.042 *** (0.000)
number_of_reviews	-0.000 *** (0.000)	-0.000 *** (0.000)	-0.000 *** (0.000)

review_scores_rating	0.005 *** (0.000)	0.005 *** (0.000)	0.005 *** (0.000)
bedrooms	0.145 *** (0.000)	0.151 *** (0.000)	0.145 *** (0.000)
beds	-0.037 *** (0.000)	-0.037 *** (0.000)	-0.037 *** (0.000)
Airconditioning	0.024 *** (0.000)	0.019 *** (0.000)	0.024 *** (0.000)
Heating	0.015 * (0.034)	0.018 * (0.021)	0.015 * (0.032)
Essentials	-0.021 *** (0.000)	-0.017 ** (0.009)	-0.021 *** (0.000)
Internet	-0.056 *** (0.000)	-0.067 *** (0.000)	-0.056 *** (0.000)
Pool	0.027 *** (0.001)	0.033 *** (0.000)	0.027 *** (0.001)
Gym	0.080 *** (0.000)	0.075 *** (0.000)	0.080 *** (0.000)
Hottub	0.020 ** (0.006)	0.019 * (0.018)	0.020 ** (0.006)
Suitableforevents	0.065 *** (0.000)	0.064 *** (0.000)	0.065 *** (0.000)
Smokingallowed	-0.031 *** (0.000)	-0.030 ** (0.001)	-0.031 *** (0.000)
Indoorfireplace	0.125 *** (0.000)	0.122 *** (0.000)	0.125 *** (0.000)
Flat	-0.105 *** (0.000)	-0.123 *** (0.000)	-0.105 *** (0.000)
Longtermstaysallowed	-0.058 *** (0.000)	-0.065 *** (0.000)	-0.056 *** (0.000)
Cleaningbeforecheckout	0.039 (0.171)	0.077 * (0.017)	
Airpurifier	-0.321 ** (0.005)	-0.386 ** (0.003)	-0.322 ** (0.004)
DigitalEntertainment	0.237 *** (0.000)	0.239 *** (0.000)	0.237 *** (0.000)
Services	0.187 *** (0.000)	0.190 *** (0.000)	0.187 *** (0.000)
Pets	0.002 (0.853)	0.014 (0.244)	
Children	0.017 (0.584)	0.045 (0.206)	
Laundry	0.065 *** (0.000)	0.069 *** (0.000)	0.065 *** (0.000)
Bathroom	0.168 *** (0.000)	0.164 *** (0.000)	0.169 *** (0.000)
FullKitchen	0.038 * (0.049)	0.050 * (0.024)	0.033 (0.085)
ExtraKitchen	-0.075 ** (0.003)	-0.084 ** (0.003)	-0.080 ** (0.001)
Bedroom	-0.061 ** (0.010)	-0.054 * (0.040)	-0.062 ** (0.009)

Accessibility	0.611 *** (0.000)	0.648 *** (0.000)	0.613 *** (0.000)
LocationFeatures	-0.096 (0.064)	-0.116 * (0.047)	
Parking	-0.229 *** (0.000)	-0.244 *** (0.000)	-0.229 *** (0.000)
Emergency	-0.012 * (0.046)	-0.017 * (0.012)	-0.012 * (0.046)
Safety	-0.237 *** (0.000)	-0.243 *** (0.000)	-0.237 *** (0.000)
Privacy	-0.097 *** (0.000)	-0.097 *** (0.000)	-0.098 *** (0.000)
N	57129	45702	57129
R2	0.634	0.635	0.634
logLik	-29397.731	-23505.581	-29400.518
AIC	58965.462	47181.162	58961.035

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
## The percentage of prediction error is 8.528524 %
```

After conducting a multiple linear regression (MLR) analysis, we observed that the variables zipcode, longitude, latitude, children, extrakitchen, parking, safety, accessibility, locationfeatures, laundry, pets and bathroom were not statistically significant for our model. Despite this, our initial model yielded a satisfactory adjusted R-squared value of 0.6332, indicating a good fit.

To optimize our model, we excluded these insignificant factors and re-ran the linear regression analysis. The resulting model achieved an adjusted R-squared value of 0.634, slightly higher than the previous model. Additionally, upon examining the residual plot, we observed a close-to-zero line, indicating that our model has a decent fit and captures the variability in the data well.

While removing non-significant variables led to a minor decrease in the adjusted R-squared value, the residual plot supports the adequacy of our model's fit, suggesting that our revised model remains a reasonable choice for predicting the response variable. Additionally, we calculated the RSE (0.4) and percentage of prediction error which was 8.53%.

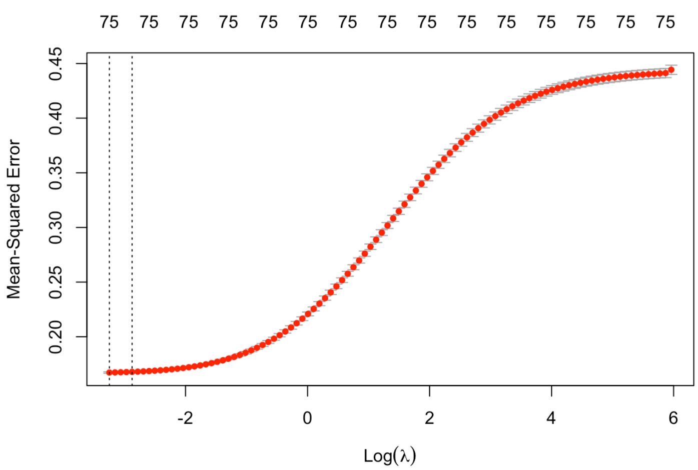
AIC

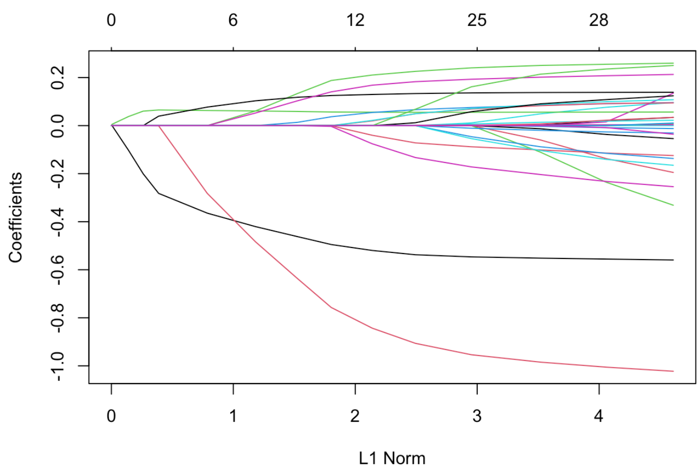
Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
			5.7e+04	9.36e+03	-1.03e+05
- Pets	1	0.00562	5.7e+04	9.36e+03	-1.03e+05

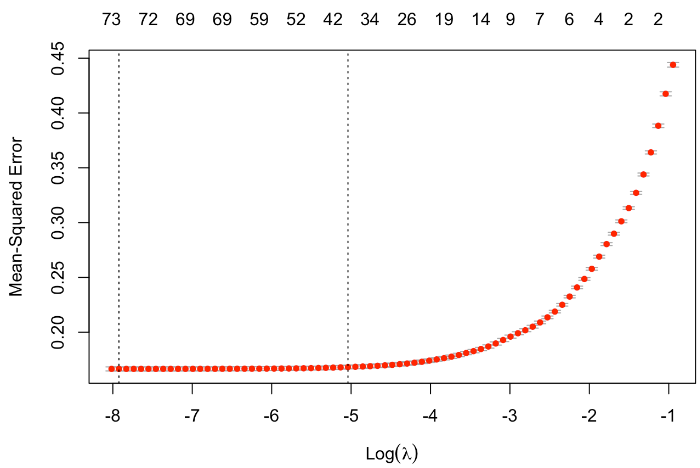
- cleaning_fee	1	0.00911	5.7e+04	9.36e+03	-1.03e+05
- Children	1	0.051	5.7e+04	9.36e+03	-1.03e+05
- Cleaningbeforecheckout	1	0.309	5.7e+04	9.36e+03	-1.03e+05

After conducting a backward step to confirm the removal of appropriate variables, we identified that “Pets,” “cleaning_fee,” “Children,” and “Cleaningbeforecheckout” were to be removed from the model. Upon analyzing the AIC values, we observed that all these variables contributed negatively, indicating a reduction in the rate of information loss. This implies that excluding these variables leads to a more efficient model with improved information retention.

Lasso Regression








```
## [1] "Best Lamda:  0.0388379108211007"
```

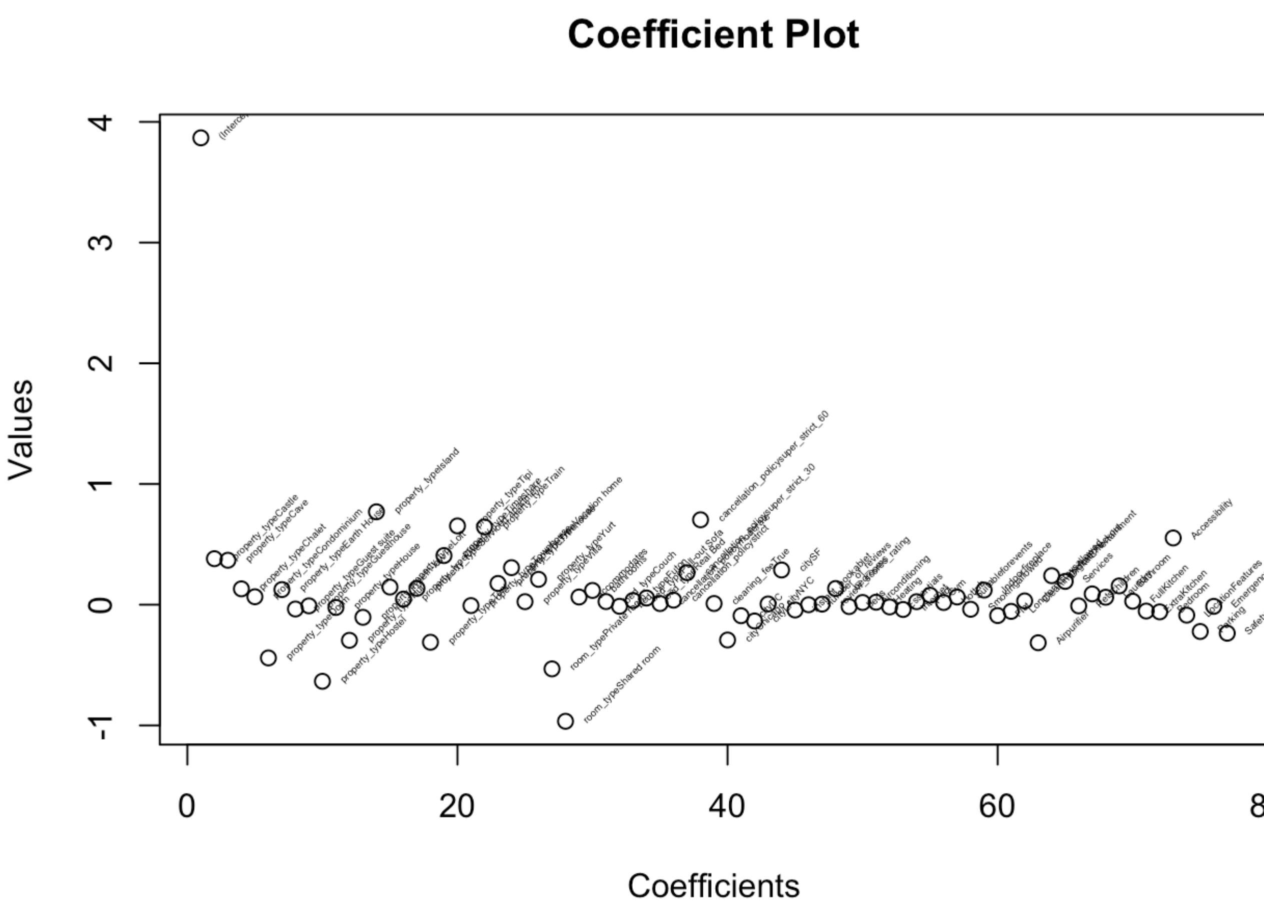
```
## [1] "RMSE:  0.405586452423197"
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.96575 -0.03967  0.02484  0.08733  0.13395  3.86766
```

```
##                     (Intercept)              property_typeCastle
##                    3.8676605619                     0.3811330334
##               property_typeCave              property_typeChalet
##                    0.3681345114                     0.1313511400
##        property_typeCondominium                property_typeDorm
##                    0.0658055516                    -0.4410423891
##        property_typeEarth House         property_typeGuest suite
##                    0.1251606330                    -0.0358308161
##         property_typeGuesthouse              property_typeHostel
##                   -0.0097185401                    -0.6346478942
##              property_typeHouse                 property_typeHut
##                   -0.0226491871                    -0.2951523847
##             property_typeIn-law              property_typeIsland
```

```
##                       -0.1044905627                        0.7695216228
##                    property_typeLoft                  property_typeOther
##                        0.1460650239                        0.0466547286
##      property_typeServiced apartment                   property_typeTent
##                        0.1355721273                       -0.3111389255
##              property_typeTimeshare                    property_typeTipi
##                        0.4070778674                        0.6526483114
##              property_typeTownhouse                  property_typeTrain
##                       -0.0075101157                        0.6445839878
##              property_typeTreehouse          property_typeVacation home
##                        0.1760083023                        0.3067548606
##                  property_typeVilla                   property_typeYurt
##                        0.0250656652                        0.2100122822
##                room_typePrivate room                room_typeShared room
##                       -0.5317018451                       -0.9657521795
##                         accommodates                           bathrooms
##                        0.0637195489                        0.1180552963
##                        bed_typeCouch                       bed_typeFuton
##                        0.0258124352                       -0.0133226523
##                bed_typePull-out Sofa                    bed_typeReal Bed
##                        0.0350885973                        0.0553916148
##         cancellation_policymoderate           cancellation_policystrict
##                        0.0094269855                        0.0376642392
## cancellation_policysuper_strict_30 cancellation_policysuper_strict_60
##                        0.2652259420                        0.7035367018
##                     cleaning_feeTrue                         cityChicago
##                        0.0098431983                       -0.2924707148
##                               cityDC                              cityLA
##                       -0.0912341491                       -0.1348706300
##                              cityNYC                              citySF
##                        0.0059819110                        0.2875273908
##                    instant_bookablet                   number_of_reviews
##                       -0.0438592736                       -0.0003190128
##                 review_scores_rating                            bedrooms
##                        0.0048503792                        0.1339484665
##                                 beds                     Airconditioning
##                       -0.0156914197                        0.0183471047
##                              Heating                          Essentials
##                        0.0206326792                       -0.0186290482
##                             Internet                                Pool
##                       -0.0396657917                        0.0248427345
##                                  Gym                              Hottub
##                        0.0782032518                        0.0182131895
##                    Suitableforevents                      Smokingallowed
##                        0.0638701520                       -0.0377876913
##                      Indoorfireplace                                Flat
##                        0.1209247439                       -0.0900927565
##                 Longtermstaysallowed               Cleaningbeforecheckout
##                       -0.0551710801                        0.0314002713
##                          Airpurifier                DigitalEntertainment
##                       -0.3152362886                        0.2401436993
##                             Services                                Pets
```

```
##                 0.1929948326                     -0.0098897243
##                     Children                           Laundry
##                 0.0894161221                      0.0616286507
##                     Bathroom                       FullKitchen
##                 0.1556138062                      0.0287343627
##                  ExtraKitchen                          Bedroom
##                -0.0525200097                     -0.0593394394
##                Accessibility                  LocationFeatures
##                 0.5528465746                     -0.0867216766
##                      Parking                         Emergency
##                -0.2219569329                     -0.0138167561
##                       Safety
##                -0.2363021596
```



Lasso selected an optimal lambda value of 0.0388379108211007, indicating the level of regularization applied to the model. The RMSE of approximately 0.41 indicates that the model's predictions closely align with the actual values on average, which is a positive sign of the model's effectiveness. Some variables like “Accessibility,” “DigitalEntertainment,” “Pets,” and “Children” have significant coefficients, indicating a strong influence on the target variable. Conversely, variables such as “Cleaningbeforecheckout,” “Smokingallowed,” and “Flat” have coefficients close to zero, suggesting a minimal impact on the target variable according to Lasso regularization. The model's combination of a low RMSE and informative coefficients signifies that the Lasso-regularized model is performing well, capturing meaningful patterns in the data while mitigating overfitting.

Conclusion

> Which property types are in the highest demand among travelers?

Amongst travelers, apartments are in the highest demand among travelers across all the cities, with houses being second in demand.

> How do prices vary across different locations and property types?

Across the different cities, we notice that SF has the highest mean log price but on average, all the cities have around the same mean log price.

The multiple linear regression (MLR) model with cross-validation shows an intercept of 3.911 with a high statistical significance ($p < 0.001$). Among the property types, Bed & Breakfast, Boat, Boutique hotel, Castle, Condominium, Island, Loft, Timeshare, Tipi, and Train are statistically significant predictors of prices with positive coefficient values, while Dorm, Camper/RV, House, Hut, and Hostel had negative coefficients. Island, Tipi, Train, Castle, and boat had the largest coefficients out of all property types, which indicates that more rare property types like these can command a higher nightly price by offering a unique experience to guests.

Room types such as Private and Shared also have significant negative coefficients, indicating these factors reduce the price. The model achieves an R-squared value of around 0.63, indicating a moderate level of predictive power.

Recommendations

In order to command the highest nightly prices, Airbnb owners should prioritize unique properties that offer whole property privacy. The properties should have a large amount of bedrooms and be able to accommodate many guests in order to charge as much as possible per night. They should offer amenities such as pools and hottubs, digital entertainment, and accessibility features. Finally, the property should be in San Francisco and have high review ratings. Utilizing these features, Airbnb owners will be able to charge the highest nightly rate according to our models.

Appendix

Cleaned Columns:

DigitalEntertainment: TV,CableTV, Gameconsole

Internet: WirelessInternet, Internet, Pocketwifi, Ethernetconnection, Laptopfriendlyworkspace

Services: Doorman,24hourcheckin,SelfCheckIn,Hostgreetsyou,Buzzerwirelessintercom,DoormanEntry,Luggagedropoffallowed

Pets: Dogs,Cats,Otherpets,Petsallowed,Petsliveonthisproperty

Children: Childrensbooksandtoys,Childrensdinnerware,Changingtable,Babymonitor,Babybath, Highchair, Familykidfriendly,Babysitterrecommendations,PacknPlaytravelcrib, Outletcovers,Windowguards,Tablecornerguards,Crib,Fireplaceguards,Stairgates

Laundry: WasherDryer, Washer, Dryer, Iron, Hangers

Bathroom: Hotwater,Bodysoap,Handsoap,Bathtowel,Handorpapertowel,Toiletpaper,Bathtub,Shampoo,Hairdryer

FullKitchen: Kitchen, Cookingbasics, Oven, Stove, Dishwasher, Dishesandsilverware, Refrigerator

ExtraKitchen:Coffeemaker, Microwave, Hotwaterkettle

Bedroom: Firmmattress,Firmmatress,Bedlinens,Extrapillowsandblankets,Roomdarkeningshades,Roomdarkeningshades

Accessibility: Wheelchairaccessible, Stepfreeaccess, Wideclearancetobed, Accessibleheightbed, Widedoorway, Accessibleheighttoilet, Wideentryway, Widehallwayclearance, smoothpathwaytofrontdoor, Welllitpathtoentrance, Wideclearancetoshowerandtoilet, Rollinshowerwithchair, Disabledparkingspot, Bathtubwithshowerchair, Fixedgrabbarsforshowertoilet, Wideclearancetoshowertoilet, Rollinshowerwithchair, Pathtoentrancelitatnight, Flatsmoothpathwaytofrontdoor, Grabrailsforshowerandtoilet, Wideclearancetoshowertoilet,Handheldshowerhead, Elevator, Groundflooraccess, Elevatorinbuilding

LocationFeatures: Waterfront,Beachfront,Lakeaccess,SkiinSkiout,Beachessentials,Gardenorbackyard,BBQgrill,Patioorbalcony

Parking: Paidparkingoffpremises,Freeparkingonstreet,Freeparkingonpremises,EVcharger

Emergency: Smokedetector, Fireextinguisher, Carbonmonoxidedetector, Firstaidkit

Safety: Safetycard, Lockonbedroomdoor, Lockbox,Keypad, Smartlock

Privacy: Privatelivingroom,Privateentrance,Privatebathroom