Project report on "Price prediction of Airbnb Accomodation"

Submitted towards partial fulfilment of the criteria for award of PGPDSE by Great Lakes Institute of Management

Submitted By
Group No. 4 [Batch: PGPDSE – FT Chennai Jul21]

Group Members

1. **Mohanraj Kandhasamy**
2. **Rajendra Karthick Sharan**
3. **Yogeshwar Mohan**
4. **Dineshkumar Anbalagan**
5. **Pydipati Praneeth Kumar**

Research Supervisor
Mr. Mupidi Srikar

GREAT LAKES
EXECUTIVE LEARNING
Great Lakes Institute of Management

ABSTRACT

The project will be based on the retail domain. In this project we would like to analyze the factors that affect the pricing of Airbnb accommodations and to help hosts to give the best price to the customers and also make good profit in this post pandemic resurgence. We would like to do this with the help of a Machine learning algorithm to find hidden information and relations. The Target variable is a numeric variable hence this is a regression problem. So we have built different models and also applied ensemble techniques to the get the best result that can be achieved at the moment. To do this we have used scikit learn and pycaret libraries for Machine Learning processes. Using the Pycaret Library we were able to create and compare various Models like Random Forest, Linear Regressor, Decision tree, Gradient Boost etc. The Best model that we arrived at is the Light Gradient Boosting Machine. After tuning this algorithim we were able to observe Mean R2 is 0.677, The RMSE score is 77.53$ with a SD of 1.64$ and the MAPE is 0.344 with SD 0.0044.

ACKNOWLEDGEMENT

Any endeavour in a specific field requires the guidance and support of many people for successful completion. The sense of achievement on completing anything remains incomplete if the people who were instrumental in its execution are not properly acknowledged. We would like to take this opportunity to verbalize our deepest sense of indebtedness to our project mentor, Mr. Mupiddi Srikar, who was a constant pillar of support and continually provided us with valuable insights to improve upon our project and make it a success. Further, we would like to thank our parents for encouraging us and providing us a platform wherein we got an opportunity to design our own project.

Date:

Place:

CERTIFICATE OF COMPLETION

I hereby certify that the project titled 'Price prediction of Airbnb Accommodation' for case resolution was undertaken and completed under my supervision by <<>> of Post Graduate Program in Data Science and Engineering (PGP – DSE)

<< Name and signature of the mentor>>

Date:

Place:

TABLE OF CONTENTS

LIST OF TABLES AND GRAPHS

EXECUTIVE SUMMARY

The Objective of our project was to predict the price of the Airbnb Accommodation with respect to several factors that have been given in the open-source dataset. The data used included city, country, state, no of bedrooms, Type of room, Type of property, Review score, amenities etc. The first base line model we formed was a Random forest Regressor for which we got a Adj-R2 of 0.62. After getting the base line model we did some amount of feature engineering and have used different Machine Learning algorithms to get the best result. The we were able to get the best result for Light Gradient Boosting Machine. We got an R2 value of 0.66 and a RMSE value of 78. Then we further tuned the model to get an RMSE of 77.

CHAPTER 1

1.1 Problem Statement:

The title of the project Price prediction of Airbnb Accommodation. Since 2008, Airbnb has helped guests and hosts to travel in a more unique, personalized way. The company went from a single air mattress for rent to global cooperation valued at more than 30 billion dollars all thanks to its energetic founder- Brian Chesky. 2020 was supposed to be the golden year for Airbnb as it would go public and issue the worlds' most sought after stocks. Tragically, Coronavirus happened. The travel sector was gutted by the pandemic. Airbnb now faces burning cash, angry hosts and an uncertain future as 2000 employees could potentially be discharged from their positions and the billion dollars debts with a high-interest rate that is being built to refund their customers. With the help of this project, we would like to help predict the prices of the Airbnb Accommodation based on several factors and help the hosts to quote a suitable price that will satisfy the travelers and to ensure that the hosts with healthy profits, which in-turn will help Airbnb get back on its feet.

Data set Source: https://public.opendatasoft.com/explore/dataset/airbnb-listings/table/?disjunctive.host_verifications&disjunctive.amenities&disjunctive.features

1.2 Data Description

id	integer	Airbnb's unique identifier for the listing
scrape_id	bigint	Inside Airbnb "Scrape" this was part of
last_scraped	datetime	UTC. The date and time this listing was "scraped".
name	text	Name of the listing

description	text	Detailed description of the listing
neighborhood_overview	text	Host's description of the neighbourhood
picture_url	text	URL to the Airbnb hosted regular sized image for the listing
host_id	integer	Airbnb's unique identifier for the host/user
host_url	text	The Airbnb page for the host
host_name	text	Name of the host. Usually just the first name(s).
host_since	date	The date the host/user was created. For hosts that are Airbnb guests this could be the date they registered as a guest.
host_location	text	The host's self-reported location
host_about	text	Description about the host
host_response_time		
host_response_rate		
host_acceptance_rate		That rate at which a host accepts booking requests.
host_is_superhost	boolean [t=true; f=false]	
host_thumbnail_url	text	
host_picture_url	text	
host_neighbourhood	text	
host_listings_count	text	The number of listings the host has (per Airbnb calculations)
host_total_listings_count	text	The number of listings the host has (per Airbnb calculations)
host_verifications		
host_has_profile_pic	boolean [t=true; f=false]	

host_identity_verified	boolean [t=true; f=false]	
neighbourhood	text	
neighbourhood_cleansed	text	The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.
neighbourhood_group_cleansed	text	The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.
latitude	numeric	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
longitude	numeric	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
property_type	text	Self selected property type. Hotels and Bed and Breakfasts are described as such by their hosts in this field
room_type	text	[Entire home/apt|Private room|Shared room|Hotel]
accommodates	integer	The maximum capacity of the listing
bathrooms	numeric	The number of bathrooms in the listing
bathrooms_text	string	The number of bathrooms in the listing.
bedrooms	integer	The number of bedrooms
beds	integer	The number of bed(s)
amenities	json	
price	currency	daily price in local currency

minimum_nights	integer	minimum number of night stay for the listing (calendar rules may be different)
maximum_nights	integer	maximum number of night stay for the listing (calendar rules may be different)
minimum_minimum_nights	integer	the smallest minimum_night value from the calender (looking 365 nights in the future)
maximum_minimum_nights	integer	the largest minimum_night value from the calender (looking 365 nights in the future)
minimum_maximum_nights	integer	the smallest maximum_night value from the calender (looking 365 nights in the future)
maximum_maximum_nights	integer	the largest maximum_night value from the calender (looking 365 nights in the future)
minimum_nights_avg_ntm	numeric	the average minimum_night value from the calender (looking 365 nights in the future)
maximum_nights_avg_ntm	numeric	the average maximum_night value from the calender (looking 365 nights in the future)
calendar_updated	date	
has_availability	boolean	[t=true; f=false]
availability_30	integer	avaliability_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
availability_60	integer	avaliability_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.

availability_90	integer	avaliability_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
availability_365	integer	avaliability_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
calendar_last_scraped	date	
number_of_reviews	integer	The number of reviews the listing has
number_of_reviews_ltm	integer	The number of reviews the listing has (in the last 12 months)
number_of_reviews_l30d	integer	The number of reviews the listing has (in the last 30 days)
first_review	date	The date of the first/oldest review
last_review	date	The date of the last/newest review
review_scores_rating		
review_scores_accuracy		
review_scores_cleanliness		
review_scores_checkin		
review_scores_communication		
review_scores_location		
review_scores_value		
license	text	The licence/permit/registration number

instant_bookable	boolean	[t=true; f=false]. Whether the guest can automatically book the listing without the host requiring to accept their booking request. An indicator of a commercial listing.
calculated_host_listings_count	integer	The number of listings the host has in the current scrape, in the city/region geography.
calculated_host_listings_count_entire_homes	integer	The number of Entire home/apt listings the host has in the current scrape, in the city/region geography
calculated_host_listings_count_private_rooms	integer	The number of Private room listings the host has in the current scrape, in the city/region geography
calculated_host_listings_count_shared_rooms	integer	The number of Shared room listings the host has in the current scrape, in the city/region geography
reviews_per_month	numeric	The number of reviews the listing has over the lifetime of the listing

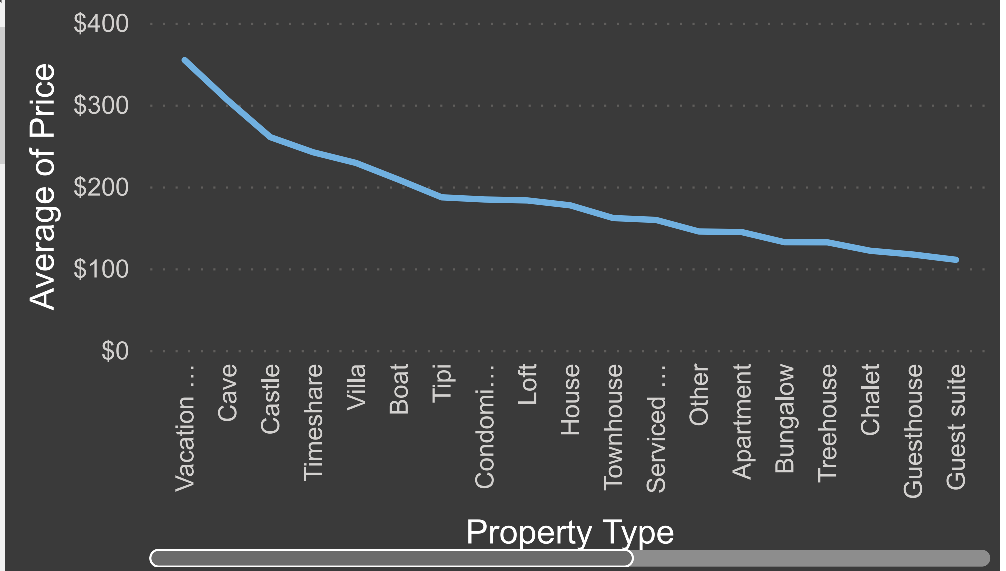
CHAPTER 2: EXPLORATRY DATA ANALYSIS

(ONE PARAGRAGHP ABOUT THE TOOLS USED TO GENERATE THE GRAPHS AND HOW WE DID IT)

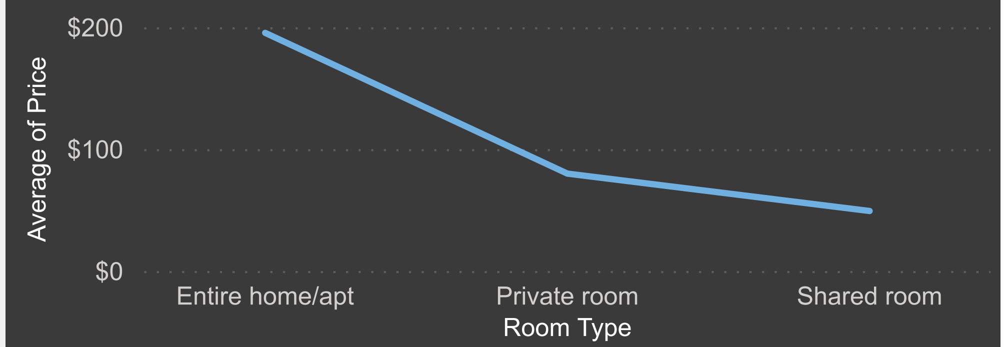
(The graphs are illustrated below :)

City	Neighbourhood	Property Type	Room Type	Bedrooms	Bed Type	Beds	Bathrooms	Average of Price
Alhambra	Alhambra	House	Entire home/apt	0	Real Bed	1	1.00	$8
Alhambra	Alhambra	House	Entire home/apt	1	Real Bed	1	1.00	$71
Alhambra	Alhambra	House	Entire home/apt	2	Real Bed	2	1.00	$11
Alhambra	Alhambra	House	Entire home/apt	2	Real Bed	3	1.00	$10
Alhambra	Alhambra	House	Entire home/apt	2	Real Bed	4	1.00	$13
Alhambra	Alhambra	House	Entire home/apt	2	Real Bed	2	1.50	$18
Alhambra	Alhambra	House	Entire home/apt	2	Real Bed	3	2.00	$11
Alhambra	Alhambra	House	Entire home/apt	3	Real Bed	3	3.00	$29
Alhambra	Alhambra	House	Private room	1	Real Bed	1	0.00	$3
Alhambra	Alhambra	House	Private room	1	Futon	1	1.00	$4
Alhambra	Alhambra	House	Private room	1	Real Bed	1	1.00	$45
Alhambra	Alhambra	House	Private room	1	Real Bed	2	1.00	$44.666666666666
Alhambra	Alhambra	House	Private room	1	Real Bed	2	2.00	$3
Alhambra	Monterey Park	House	Private room	1	Real Bed	1	2.50	$4
Allston	Allston-Brighton	House	Private room	1	Real Bed	1	2.00	$7
Altadena	Altadena	House	Entire home/apt	0	Real Bed	1	1.00	$8
Altadena	Altadena	House	Entire home/apt	0	Real Bed	2	1.00	$159
Altadena	Altadena	House	Entire home/apt	1	Futon	1	1.00	$5
Altadena	Altadena	House	Entire home/apt	1	Real Bed	1	1.00	$4
Altadena	Altadena	House	Entire home/apt	1	Real Bed	2	1.00	$9
Altadena	Altadena	House	Entire home/apt	1	Pull-out Sofa	3	1.00	$8
Altadena	Altadena	House	Entire home/apt	2	Real Bed	2	1.00	$144.333333333333
Altadena	Altadena	House	Entire home/apt	2	Real Bed	3	1.00	$147
Altadena	Altadena	House	Entire home/apt	1	Real Bed	2	1.50	$8
Altadena	Altadena	House	Entire home/apt	1	Real Bed	2	2.00	$9
Altadena	Altadena	House	Entire home/apt	2	Real Bed	3	2.00	$162
Altadena	Altadena	House	Entire home/apt	3	Real Bed	3	2.00	$242
Altadena	Altadena	House	Entire home/apt	3	Real Bed	4	2.00	$34
Altadena	Altadena	House	Entire home/apt	3	Real Bed	5	2.00	$50
Altadena	Altadena	House	Entire home/apt	3	Real Bed	6	3.00	$30
Altadena	Altadena	House	Entire home/apt	4	Real Bed	6	3.00	$37
Altadena	Altadena	House	Entire home/apt	5	Real Bed	7	3.00	$47
Altadena	Altadena	House	Entire home/apt	4	Real Bed	4	3.50	$60
Altadena	Altadena	House	Entire home/apt	4	Real Bed	6	3.50	$72
Altadena	Altadena	House	Entire home/apt	5	Real Bed	6	3.50	$38
Altadena	Altadena	House	Entire home/apt	4	Real Bed	10	4.00	$43
Altadena	Altadena	House	Entire home/apt	6	Real Bed	8	4.00	$49
Total								**$176.964110929853**

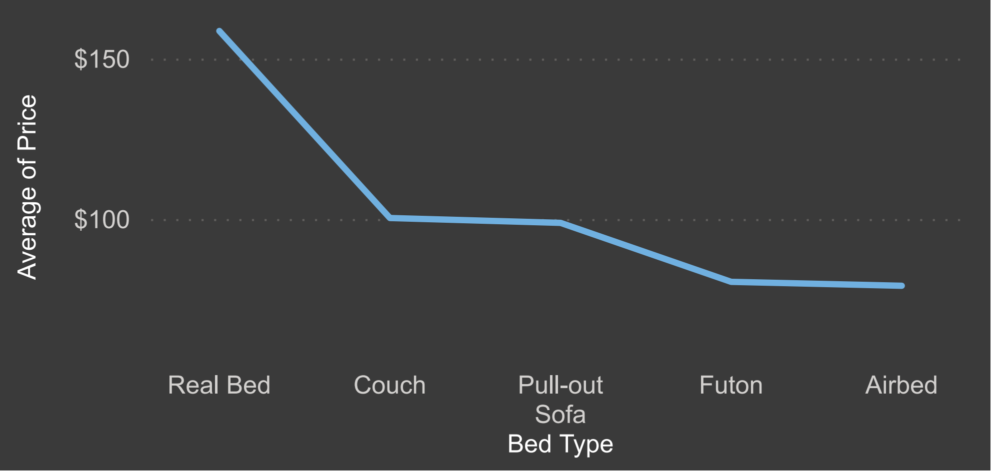
Average of Price by Property Type



Average of Price by Room Type

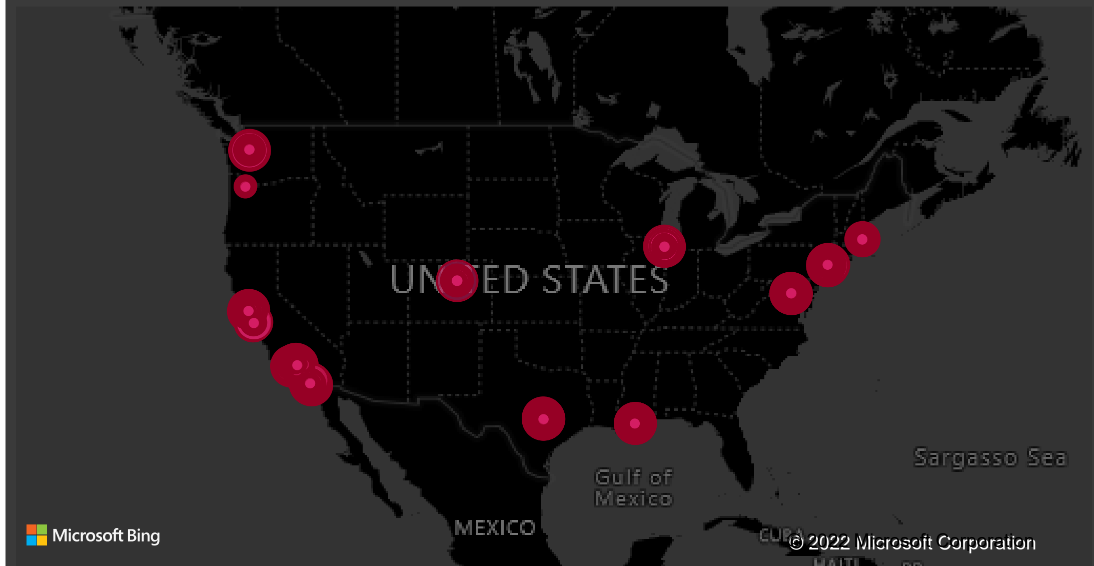


Average of Price by Bed Type



AIRBNB DEEP DIVE PRICE ANALYSIS

PRICES BASED ON LOCATIONS



State	City	Neighbourhood	Property Type	Average of Pric
CA	Alhambra	Alhambra	Apartment	$97.7
CA	Alhambra	Alhambra	Bungalow	$60.0
CA	Alhambra	Alhambra	Condominium	$119.6
CA	Alhambra	Alhambra	Guesthouse	$80.0
CA	Alhambra	Alhambra	House	$77.8
CA	Alhambra	Alhambra	Loft	$80.0
CA	Alhambra	Alhambra	Other	$54.0
CA	Alhambra	Alhambra	Townhouse	$59.1
CA	Alhambra	Downtown	Apartment	$250.0
CA	Alhambra	Monterey Park	House	$40.0
CA	Altadena	Altadena	Apartment	$112.0
CA	Altadena	Altadena	Bed & Breakfast	$119.0
Total				**$156.8**

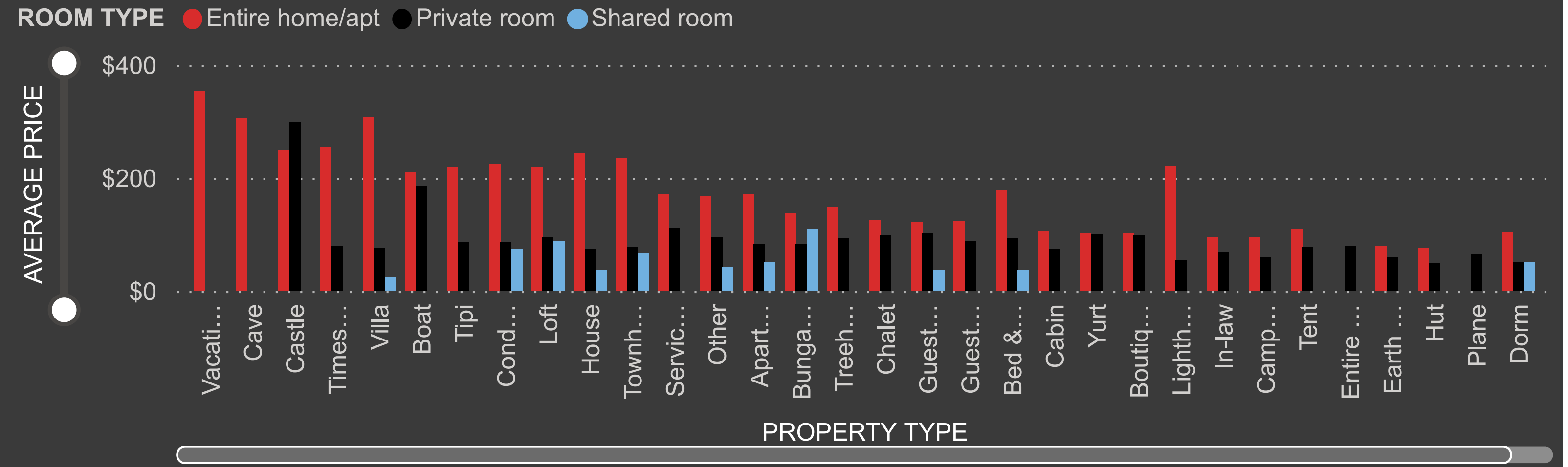
STATE	CITY	NEIGHBOURHOOD	PROPERTY T...
All	All	All	All

REVIEW RATING

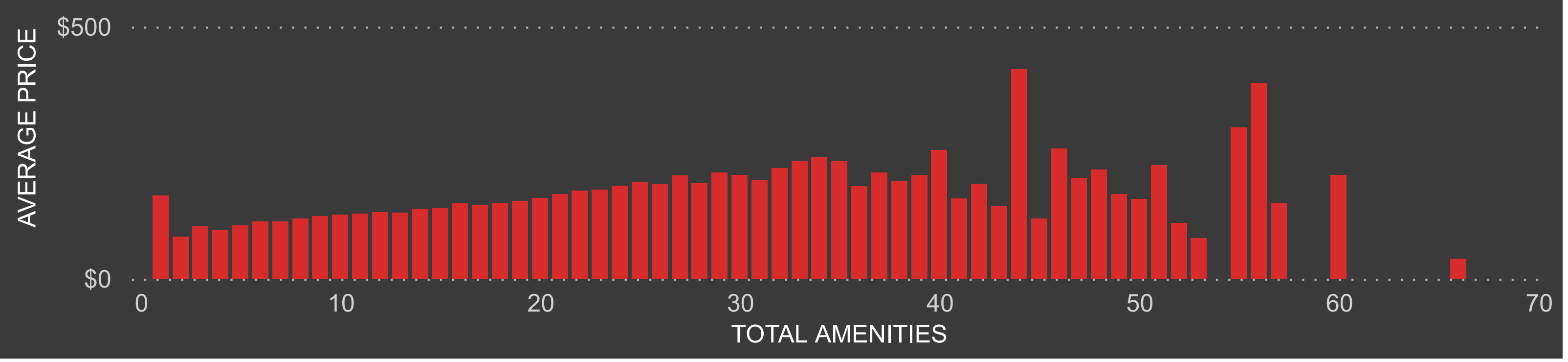
$156.81

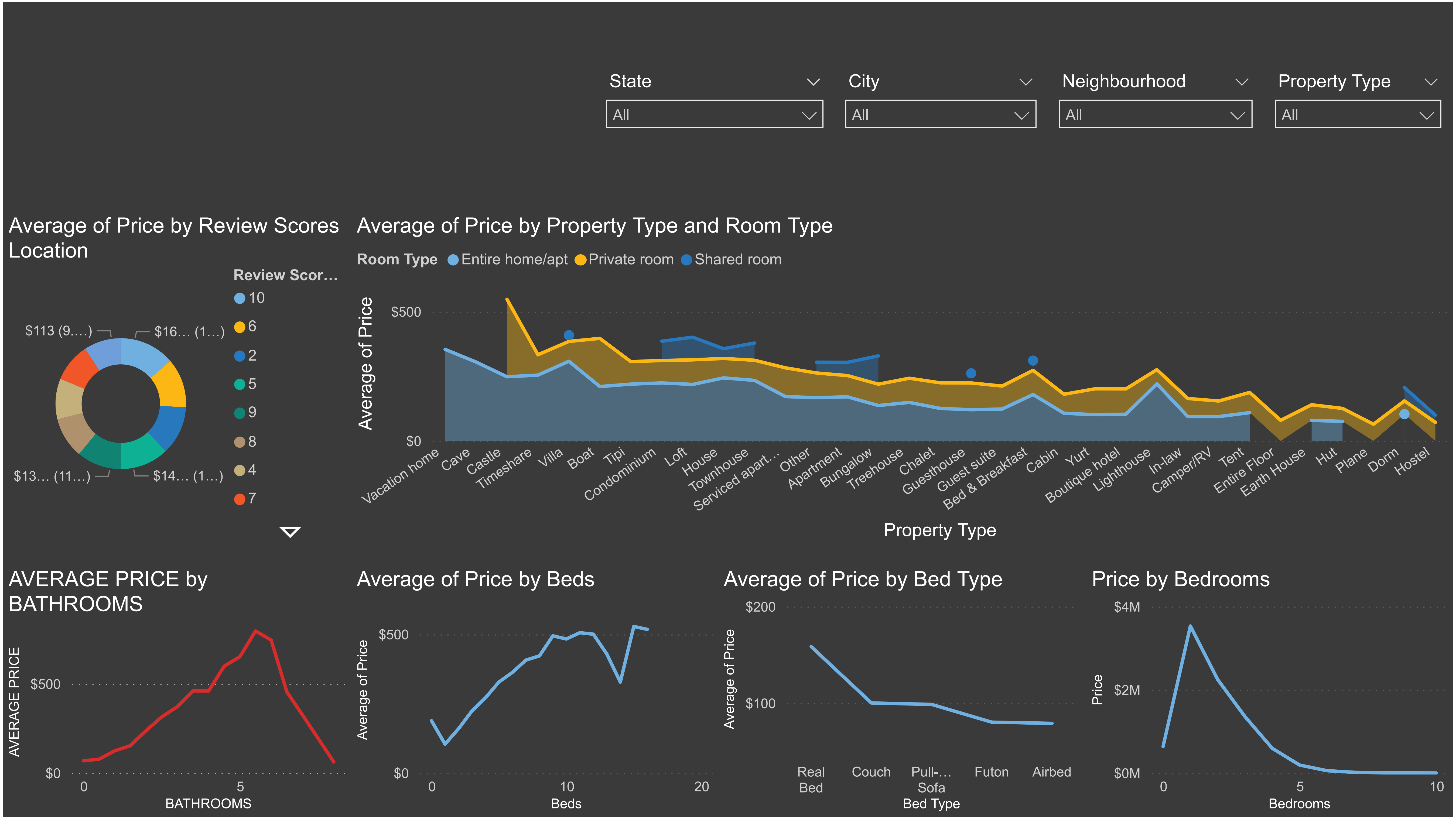
AVERAGE PRICE

AVERAGE PRICE BY PROPERTY AND ROOM TYPE

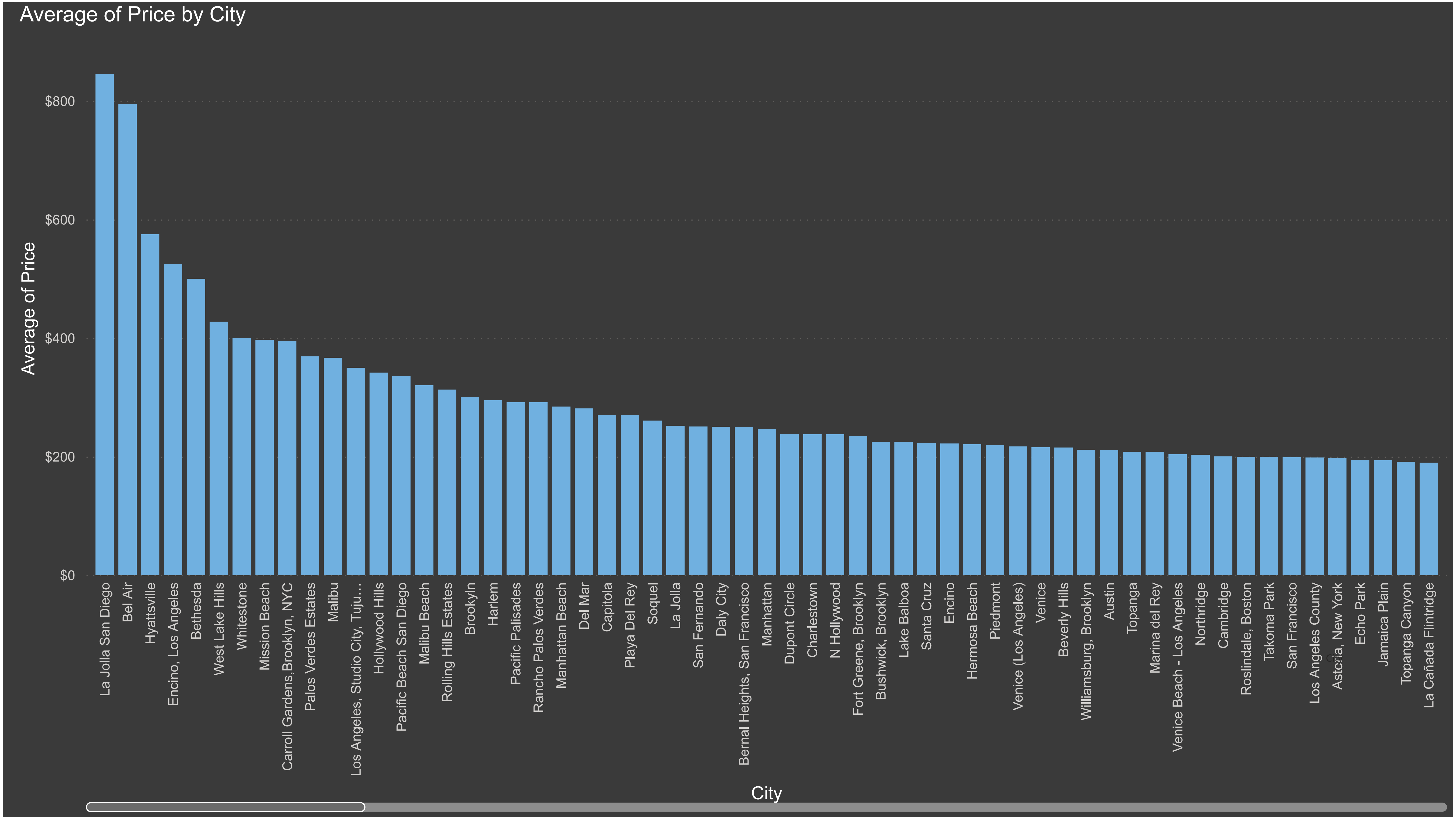


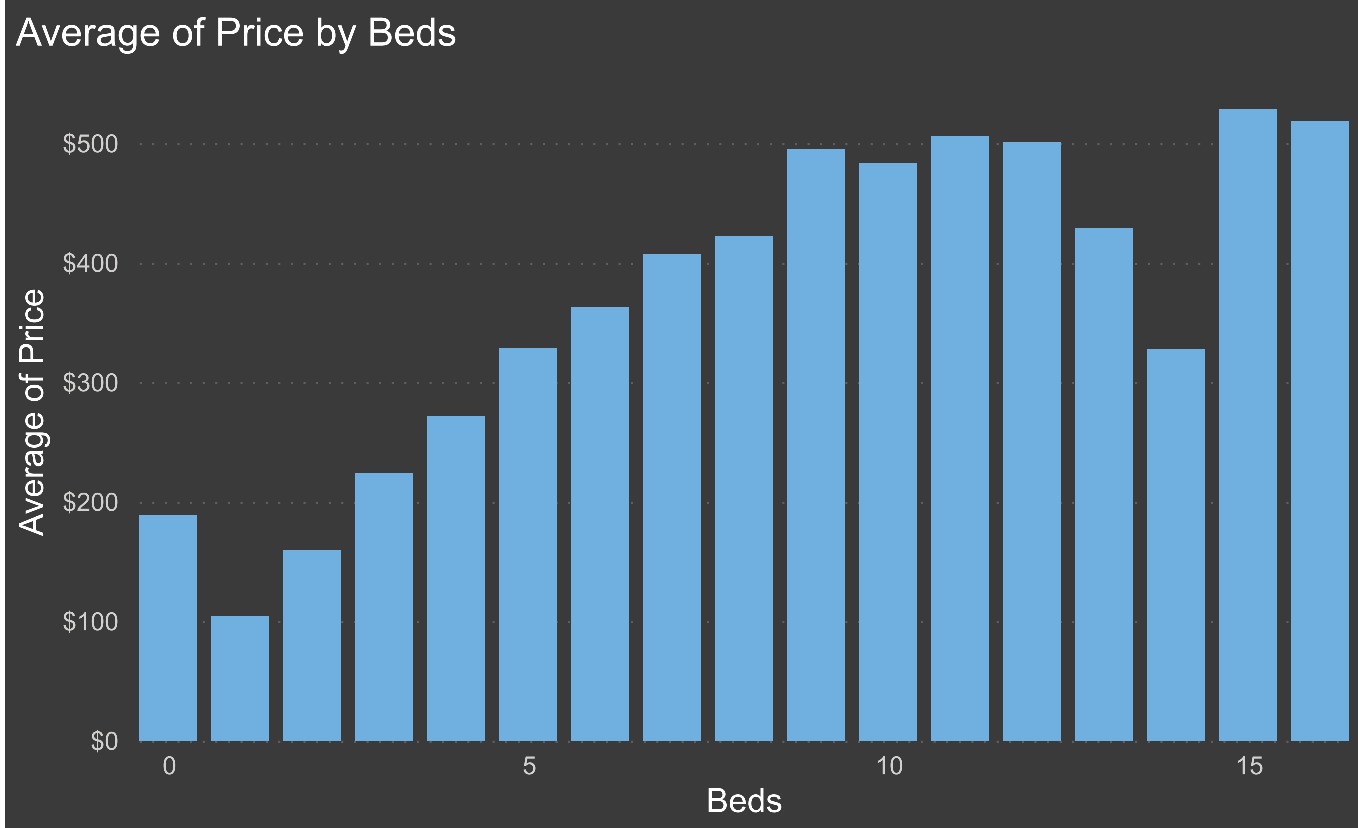
AVERAGE PRICE BASED ON AENITIES



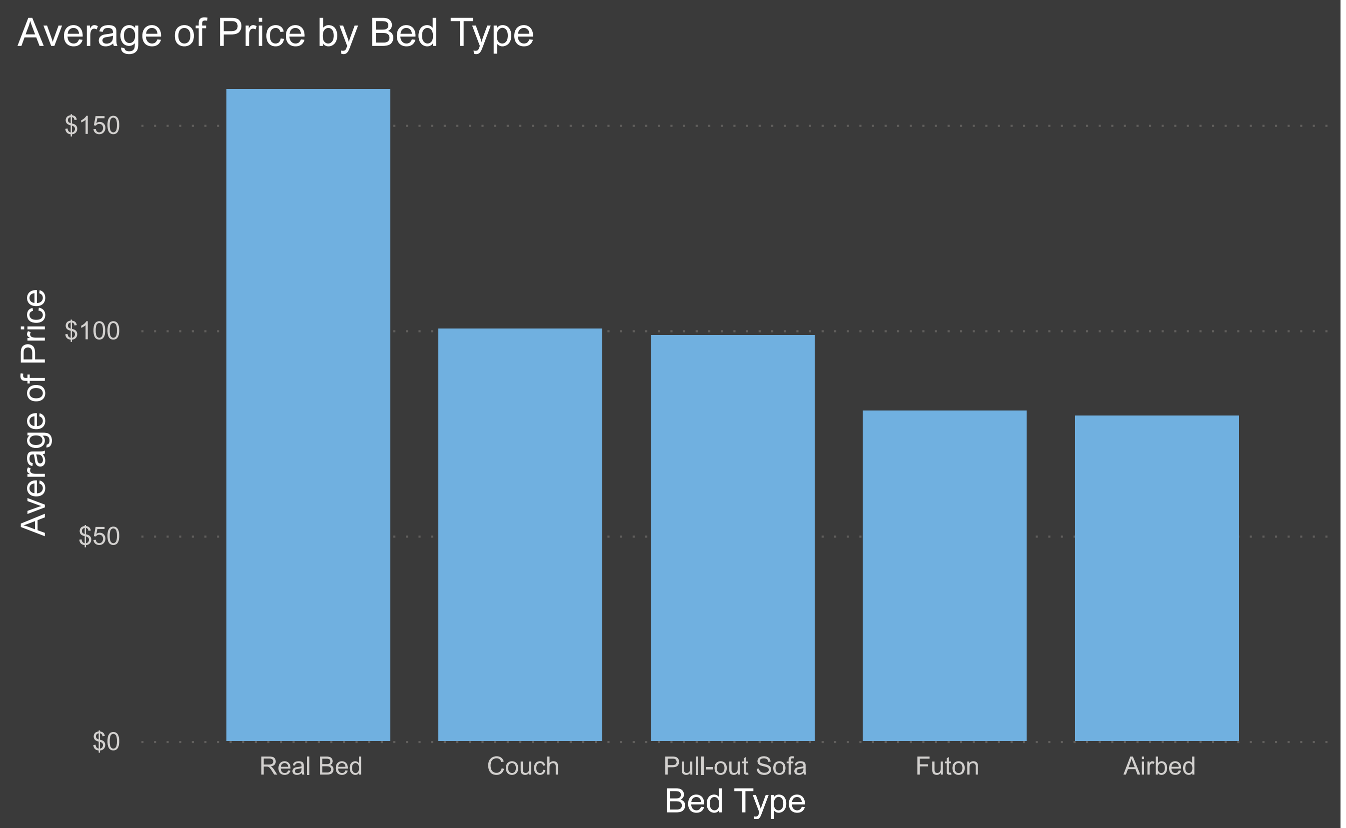


State
All
City
All
Neighbourhood
All
Property Type
All
Average of Price by Review Scores Location
Review Scor…
10
6
2
5
9
8
4
7
$113 (9.…)
$16… (1…)
$13… (11…)
$14… (1…)
Average of Price by Property Type and Room Type
Room Type Entire home/apt Private room Shared room
Average of Price
$500
$0
Vacation home
Cave
Castle
Timeshare
Villa
Boat
Tipi
Condominium
Loft
House
Townhouse
Serviced apart…
Other
Apartment
Bungalow
Treehouse
Chalet
Guesthouse
Guest suite
Bed & Breakfast
Cabin
Yurt
Boutique hotel
Lighthouse
In-law
Camper/RV
Tent
Entire Floor
Earth House
Hut
Plane
Dorm
Hostel
Property Type
AVERAGE PRICE by BATHROOMS
AVERAGE PRICE
$500
$0
0
5
BATHROOMS
Average of Price by Beds
Average of Price
$500
$0
0
10
20
Beds
Average of Price by Bed Type
Average of Price
$200
$100
Real Bed
Couch
Pull-… Sofa
Futon
Airbed
Bed Type
Price by Bedrooms
Price
$4M
$2M
$0M
0
5
10
Bedrooms

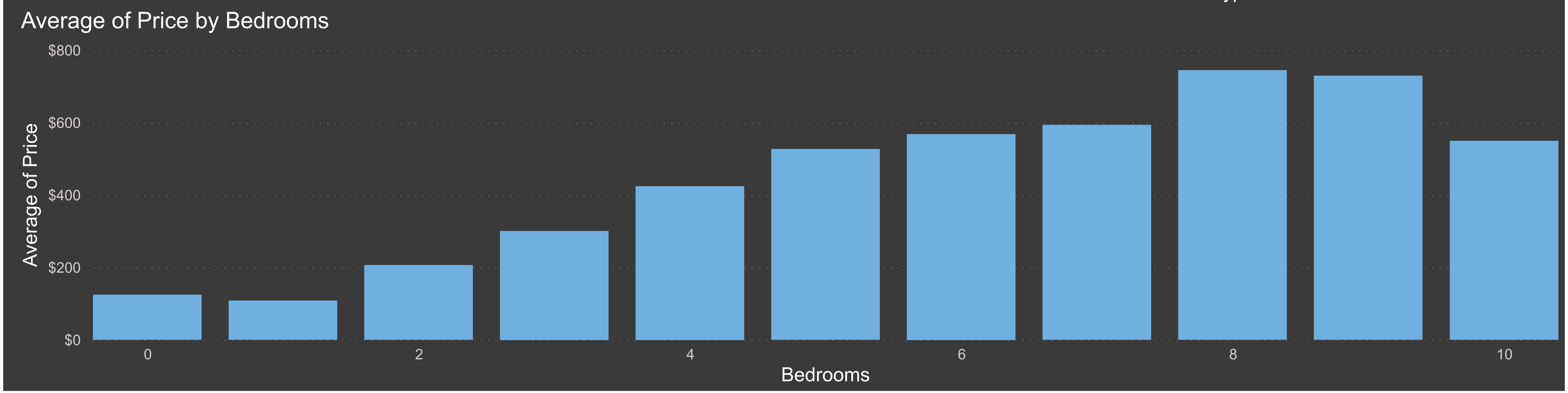

Average of Price by City
Average of Price
$800
$600
$400
$200
$0
La Jolla San Diego
Bel Air
Hyattsville
Encino, Los Angeles
Bethesda
West Lake Hills
Whitestone
Mission Beach
Carroll Gardens,Brooklyn, NYC
Palos Verdes Estates
Malibu
Los Angeles, Studio City, Tuju...
Hollywood Hills
Pacific Beach San Diego
Malibu Beach
Rolling Hills Estates
Brookyln
Harlem
Pacific Palisades
Rancho Palos Verdes
Manhattan Beach
Del Mar
Capitola
Playa Del Rey
Soquel
La Jolla
San Fernando
Daly City
Bernal Heights, San Francisco
Manhattan
Dupont Circle
Charlestown
N Hollywood
Fort Greene, Brooklyn
Bushwick, Brooklyn
Lake Balboa
Santa Cruz
Encino
Hermosa Beach
Piedmont
Venice (Los Angeles)
Venice
Beverly Hills
Williamsburg, Brooklyn
Austin
Topanga
Marina del Rey
Venice Beach - Los Angeles
Northridge
Cambridge
Roslindale, Boston
Takoma Park
San Francisco
Los Angeles County
Astoria, New York
Echo Park
Jamaica Plain
Topanga Canyon
La Cañada Flintridge
City



Average of Price by Beds
Average of Price
$500
$400
$300
$200
$100
$0
0
5
10
15
Beds

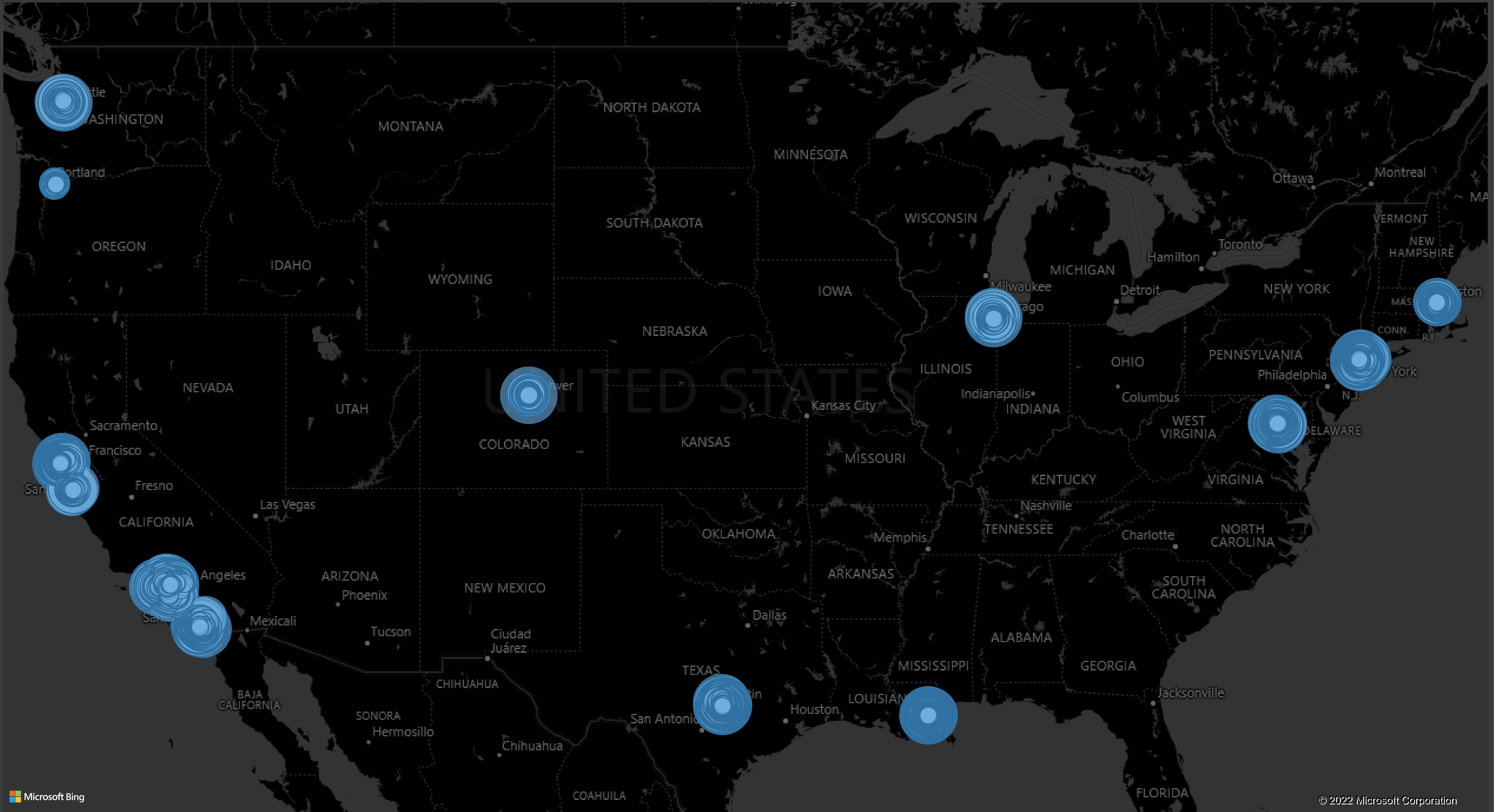


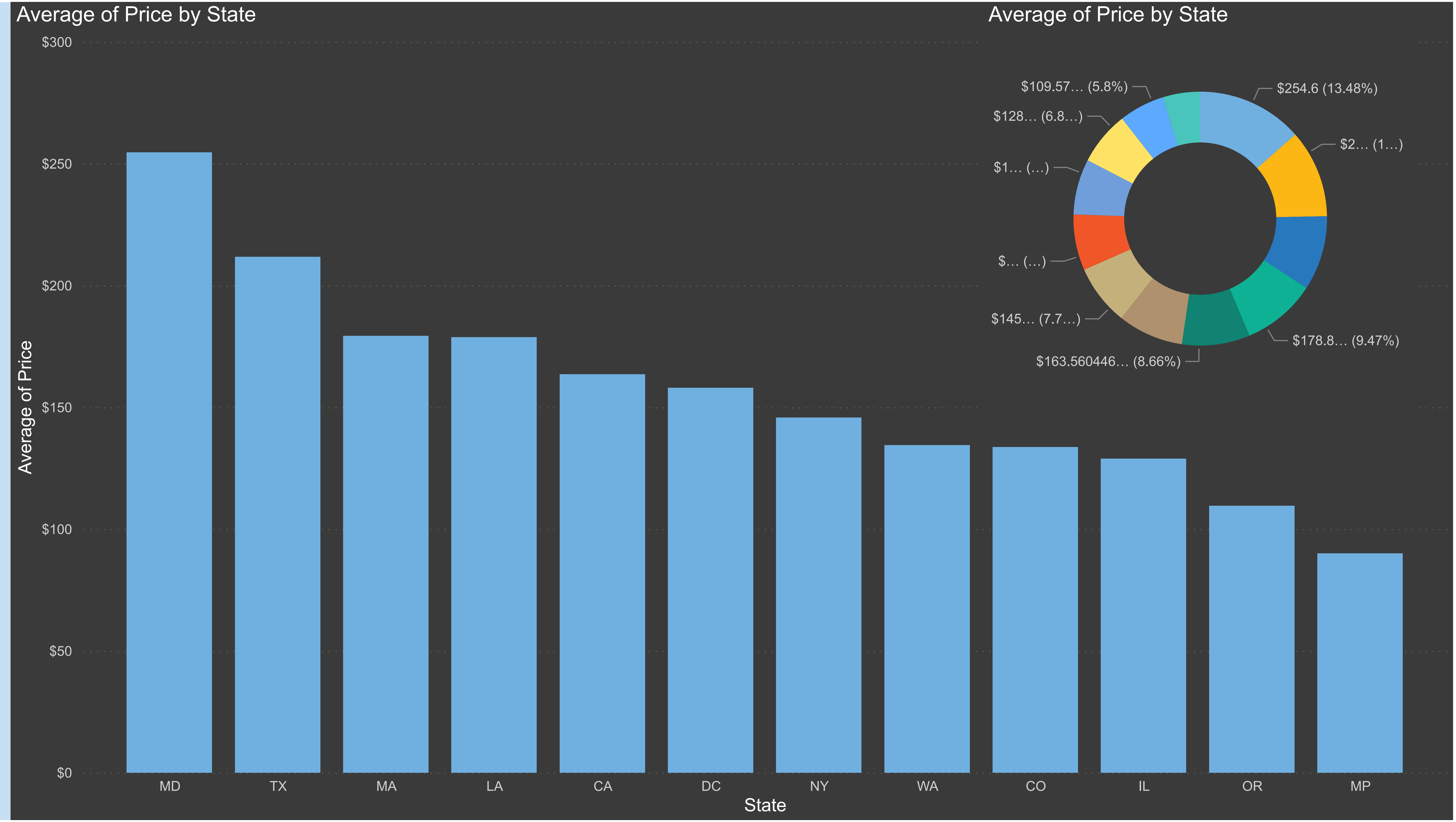
Average of Price by Bed Type
Average of Price
$150
$100
$50
$0
Real Bed
Couch
Pull-out Sofa
Futon
Airbed
Bed Type



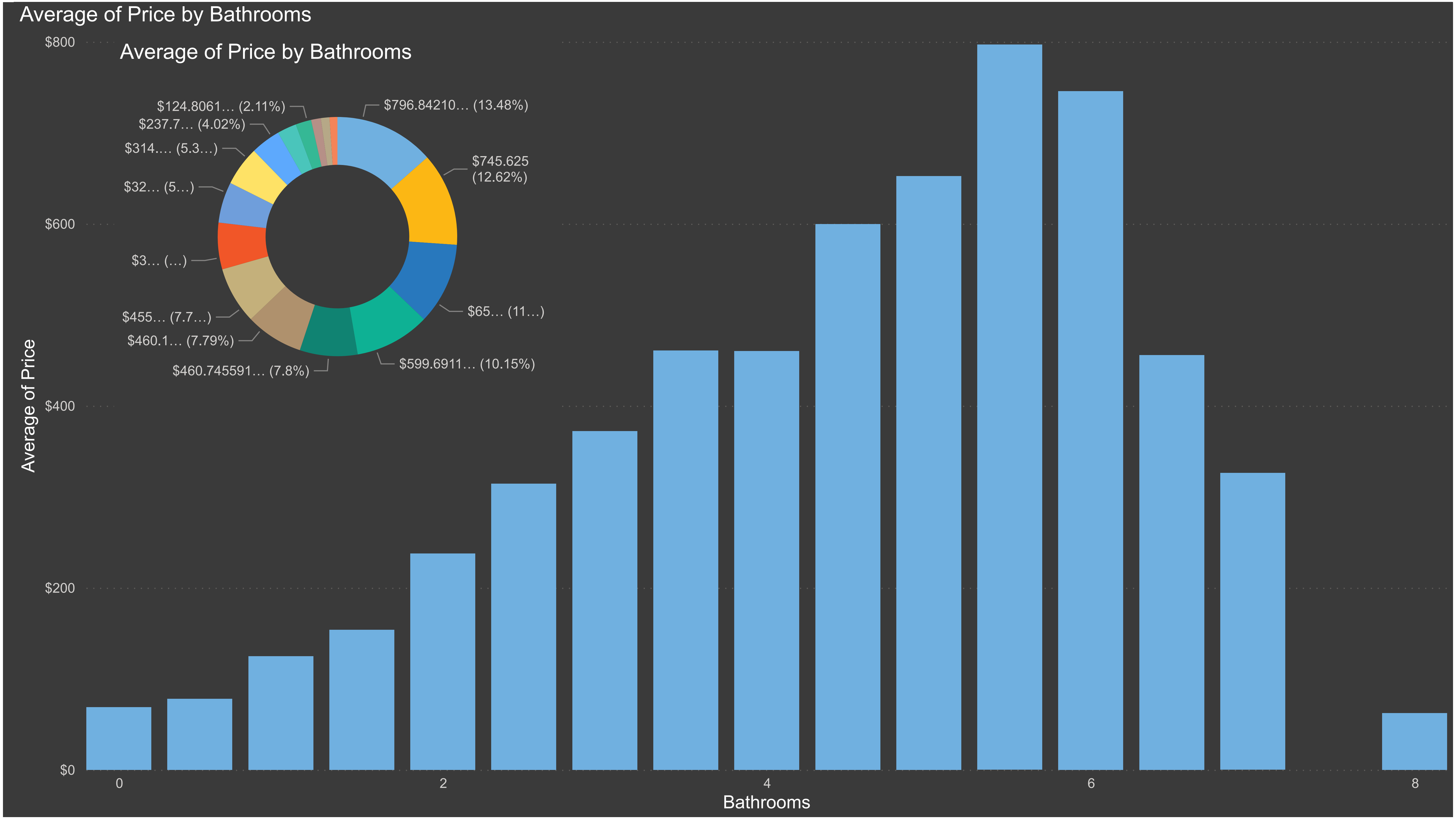
Average of Price by Bedrooms
Average of Price
$800
$600
$400
$200
$0
0
2
4
6
8
10
Bedrooms

Average of Price by Latitude and Longitude

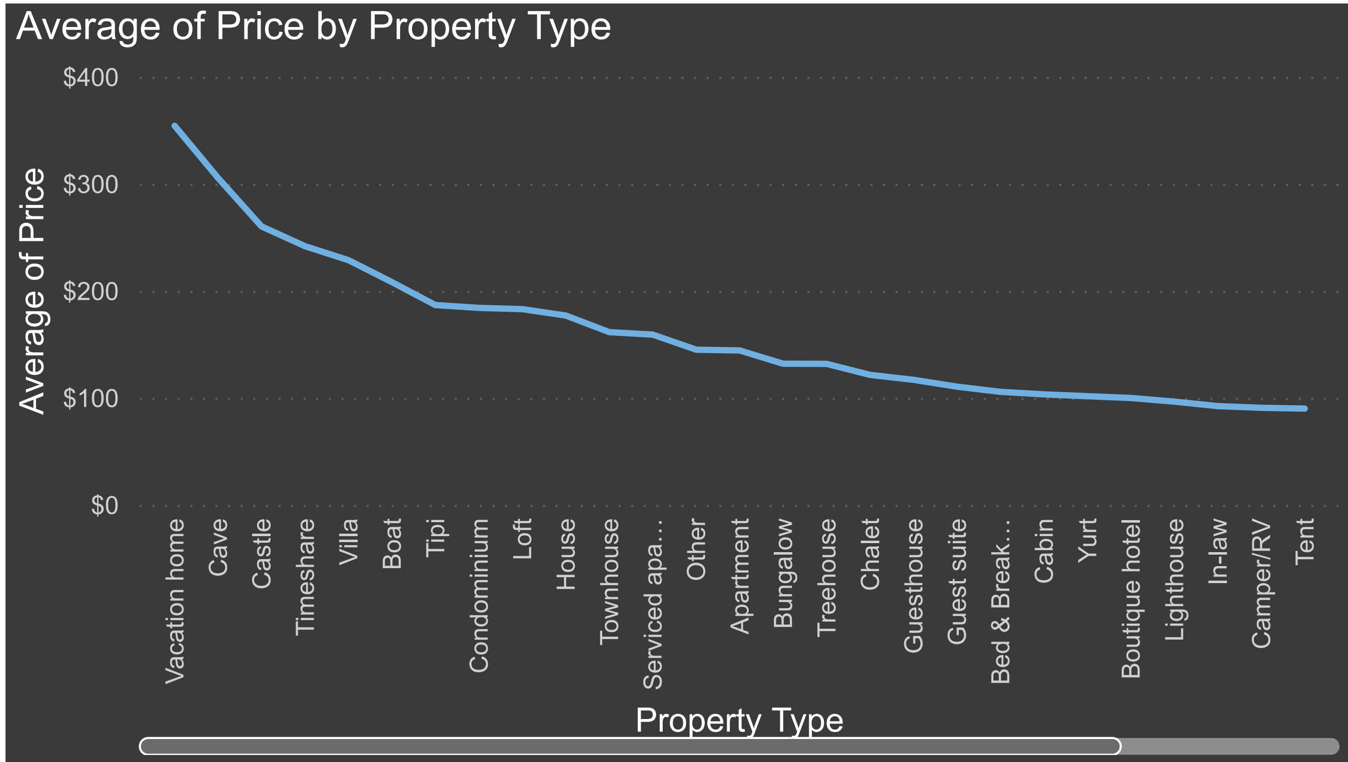


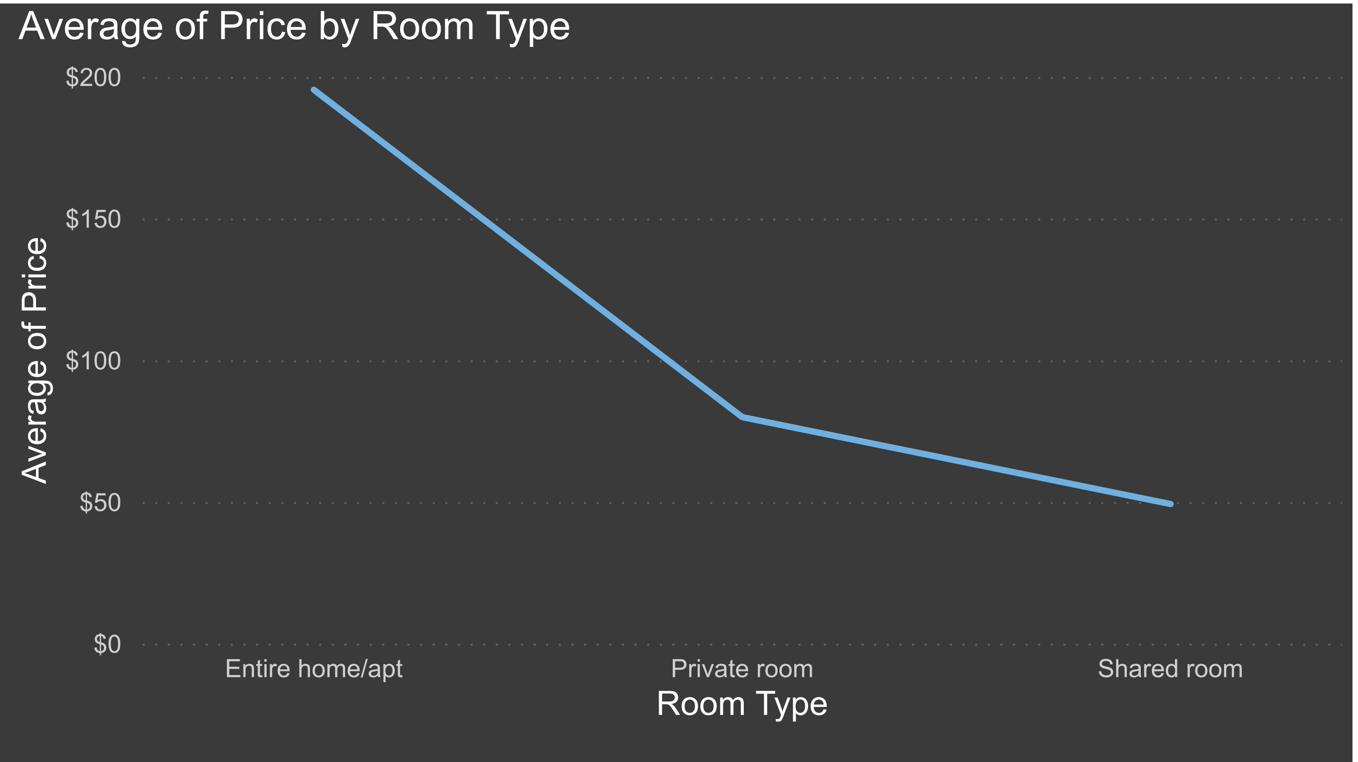


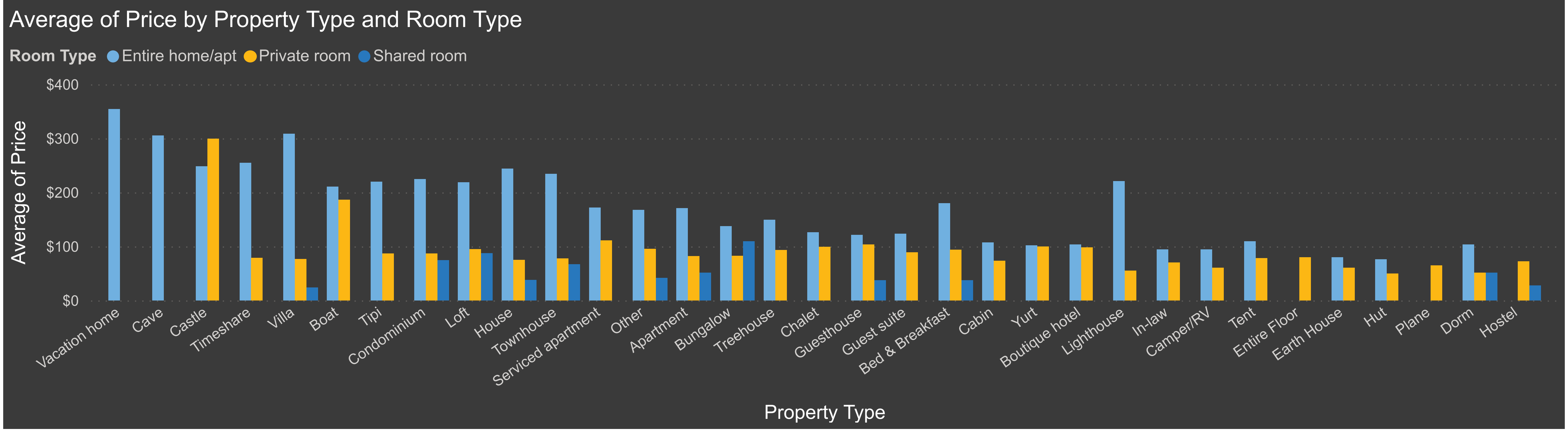
Average of Price by State
$300
$250
$200
$150
$100
$50
$0
Average of Price
MD
TX
MA
LA
CA
DC
NY
WA
CO
IL
OR
MP
State
Average of Price by State
$109.57… (5.8%)
$254.6 (13.48%)
$128… (6.8…)
$2… (1…)
$1… (…)
$… (…)
$145… (7.7…)
$178.8… (9.47%)
$163.560446… (8.66%)

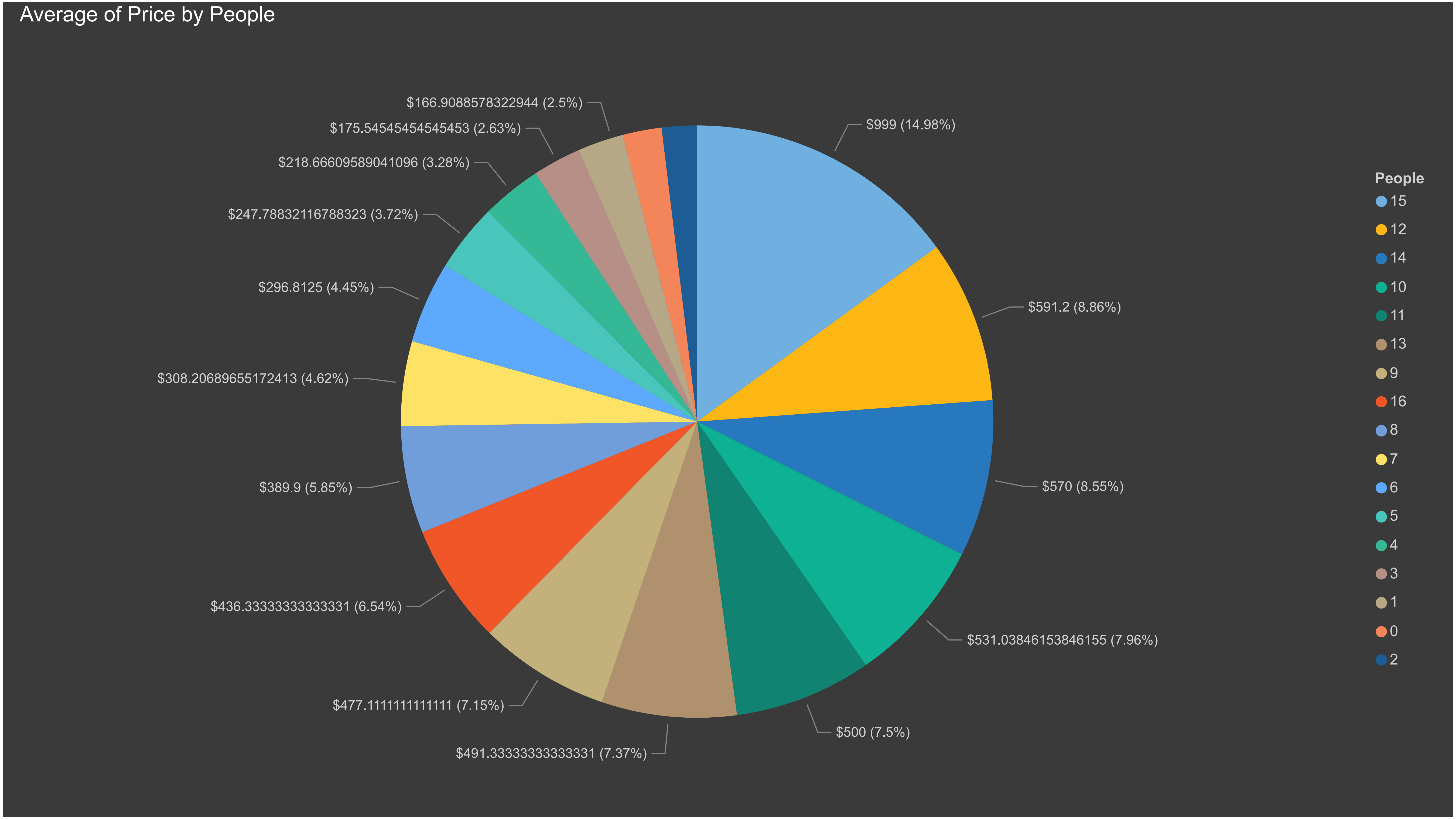


Average of Price by Bathrooms
Average of Price by Bathrooms
$124.8061… (2.11%)
$796.84210… (13.48%)
$237.7… (4.02%)
$314…. (5.3…)
$745.625
(12.62%)
$32… (5…)
$3… (…)
$65… (11…)
$455… (7.7…)
$460.1… (7.79%)
$599.6911… (10.15%)
$460.745591… (7.8%)
$800
$600
$400
$200
$0
Average of Price
0
2
4
6
8
Bathrooms

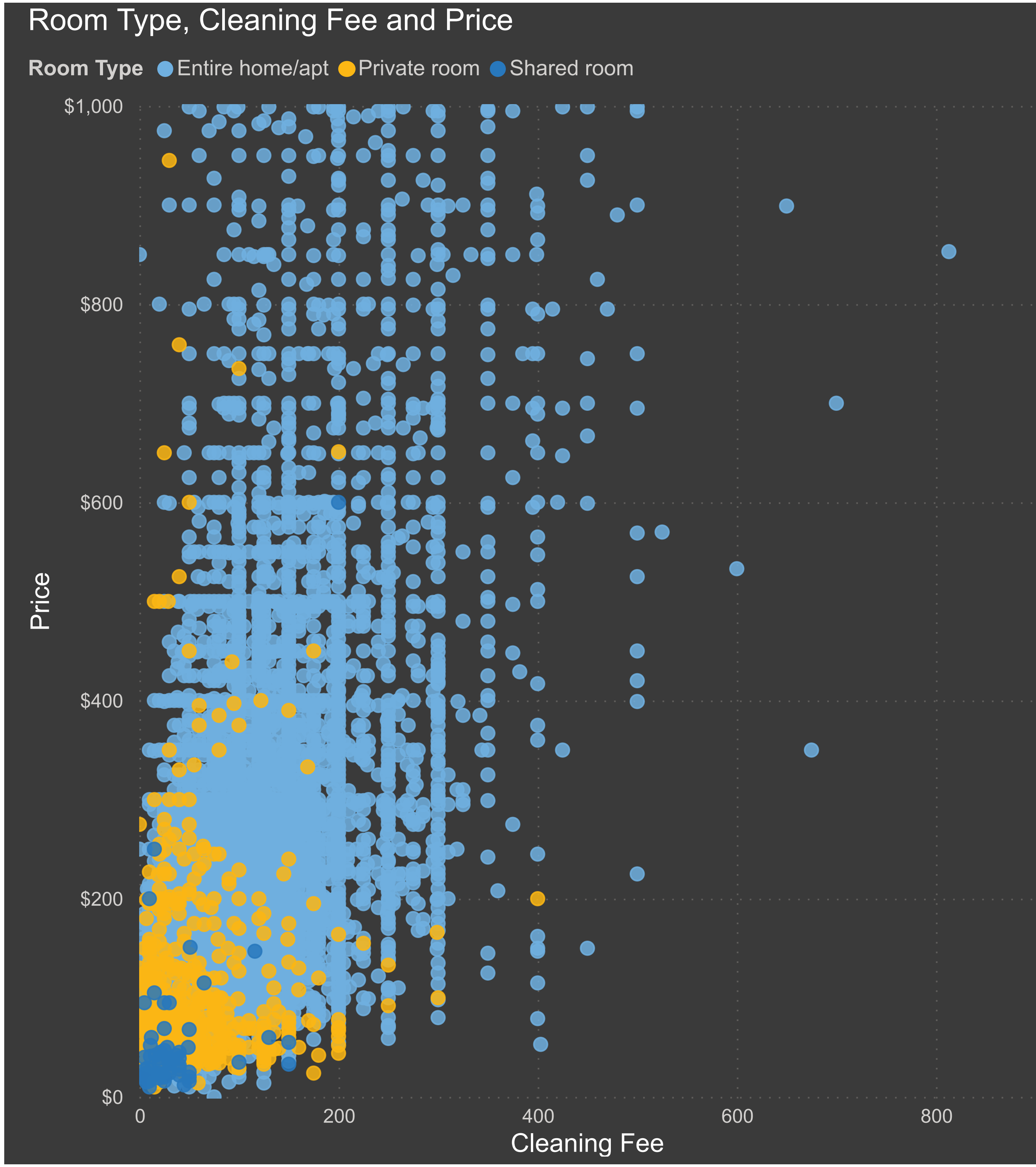

Average of Price by Property Type
$400
$300
$200
$100
$0
Average of Price
Vacation home
Cave
Castle
Timeshare
Villa
Boat
Tipi
Condominium
Loft
House
Townhouse
Serviced apa…
Other
Apartment
Bungalow
Treehouse
Chalet
Guesthouse
Guest suite
Bed & Break…
Cabin
Yurt
Boutique hotel
Lighthouse
In-law
Camper/RV
Tent
Property Type

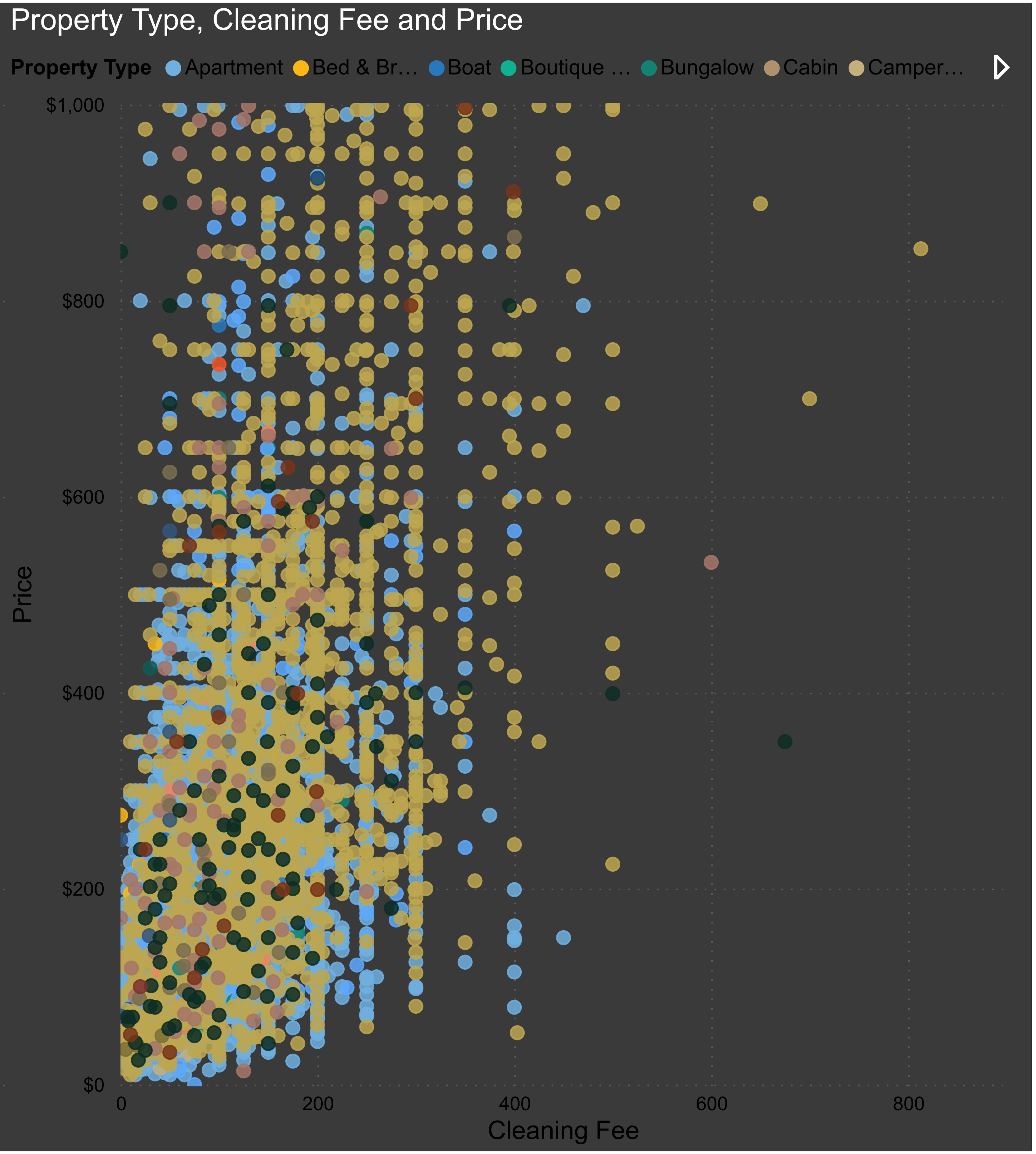

Average of Price by Room Type
$200
$150
$100
$50
$0
Average of Price
Entire home/apt
Private room
Shared room
Room Type


Average of Price by Property Type and Room Type
Room Type
Entire home/apt
Private room
Shared room
$400
$300
$200
$100
$0
Average of Price
Vacation home
Cave
Castle
Timeshare
Villa
Boat
Tipi
Condominium
Loft
House
Townhouse
Serviced apartment
Other
Apartment
Bungalow
Treehouse
Chalet
Guesthouse
Guest suite
Bed & Breakfast
Cabin
Yurt
Boutique hotel
Lighthouse
In-law
Camper/RV
Tent
Entire Floor
Earth House
Hut
Plane
Dorm
Hostel
Property Type



Average of Price by People
$999 (14.98%)
$591.2 (8.86%)
$570 (8.55%)
$531.03846153846155 (7.96%)
$500 (7.5%)
$491.33333333333331 (7.37%)
$477.1111111111111 (7.15%)
$436.33333333333331 (6.54%)
$389.9 (5.85%)
$308.20689655172413 (4.62%)
$296.8125 (4.45%)
$247.78832116788323 (3.72%)
$218.66609589041096 (3.28%)
$175.54545454545453 (2.63%)
$166.9088578322944 (2.5%)
People
15
12
14
10
11
13
9
16
8
7
6
5
4
3
1
0
2


Room Type, Cleaning Fee and Price
Room Type
Entire home/apt
Private room
Shared room
$1,000
$800
$600
$400
$200
$0
Price
0
200
400
600
800
Cleaning Fee


Property Type, Cleaning Fee and Price
Property Type
Apartment
Bed & Br...
Boat
Boutique ...
Bungalow
Cabin
Camper...
$1,000
$800
$600
$400
$200
$0
Price
0
200
400
600
800
Cleaning Fee

CHAPTERS 3: STEP-BY-STEP WALK THROUGH OF THE SOLUTION

After the exploratory data analysis & preparation of final data, the next steps we took towards solving our problem were as follows:

i. Splitting the data into training & validation datasets, i.e. training dataset : the actual dataset that we use to train the model, model sees and learns from this data; validation dataset : the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

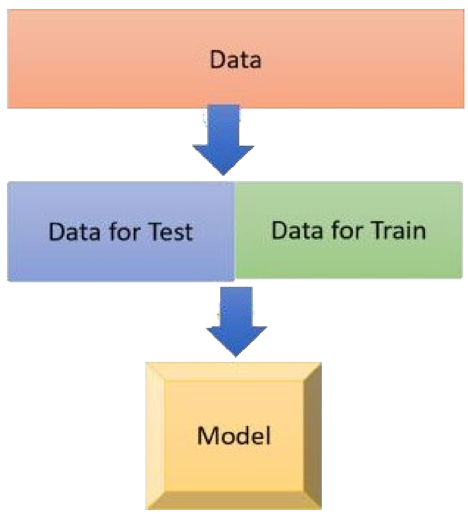


Figure 25: Splitting of Data

ii. Building the base line regression model taking into consideration all the variables. The base line model used was Random Forest. Random Forest is a type of Non-linear model which has many decision trees to learn the data.

iii. After building the base line model, & doing further analysis on the model summary, we removed variables one by one by taking into consideration of the features used to learn the data. The final base line model was then ready after doing all the above mentioned steps, but we wanted to know the features which are most important for predicting the target variable.

iv. We have an ensemble technique which provides us with the relative importance of the variables i.e. Gradient Boosting. Feature importance is used commonly & tells us that which variable is contributing the most to a model & is critical to interpreting the results. The technique resulted in providing some visualization & insights from which we were able to recognize the important variables & tried removing the insignificant variables, taking into consideration the relative importance values.

Feature Engineerning

i. The data has a column named 'Host Since', this column present to us since which date the current host has taken charge of the property. Since its data type is date and time we were not able to get any significant inferences by analysing it. Hence we subtracted the date with the current date and arrive at a new column named 'host since days' Which has the of days for which they have been the host and we have changed it to numeric variable to get better insights from it.
ii. The data had a Column 'Amenities', columns had the name of amenities as comma separated, we engineer this column to get better interpretation. We took the count of amenities as no of amenities and added as a new column 'Total_Amenities' and dropped the Amenities column.
iii. The data had a Column 'Features, columns had the name of features as comma separated, we engineer this column to get better interpretation. We took the count of features as no of amenities and added as a new column 'Total_Features' and dropped the features column.

Selecting the best Model

i. After building the baseline model we have built another model with Gradient boosting algorithm. This had a R2 value of 0.65 and RMSE Score of 83.93. Keeping this as the second base line model we got the important features.

ii. In this we thought that there would be high Cardinality problem in the 'City' column as the no of unique levels was 724. And then we tried dropping the city column and got a R2 score of 0.65 and RMSE score of 83.7. The R2 and the RMSE score were the same for both the models hence we considered to drop the Column after doing a few more tests.

iii. The Third model we built was the XGBoost model which gave us a better performance. The R2 value of the model is 0.66 and RMSE of 81.4. After looking at the model we also planned to build different models to check for the best models. To do this we used a third-party library called Pycaret with this we were able to build the several models and compare them.

iv. The Model with the best metrics is the Light Gradient Boosting Machine. This model had the R2 score of 0.668, the RMSE Score of 78.68 and MAPE of 0.351. Hence we have selected it as our final model and proceeded with Hyper-parameter tuning.d

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	47.5500	6194.7346	78.6887	0.6680	0.3936	0.3513	3.8040
rf	Random Forest Regressor	48.8057	6531.5000	80.8004	0.6500	0.4002	0.3615	99.9540
et	Extra Trees Regressor	49.1462	6658.8478	81.5876	0.6432	0.4016	0.3651	117.5080
gbr	Gradient Boosting Regressor	50.1439	6889.2691	82.9864	0.6308	0.4155	0.3763	25.3020
br	Bayesian Ridge	56.1198	8415.0590	91.7144	0.5492	0.4868	0.4406	1.5670
ridge	Ridge Regression	56.1517	8417.6931	91.7286	0.5490	0.4878	0.4410	0.4500
lar	Least Angle Regression	56.1563	8418.4023	91.7325	0.5490	0.4879	0.4411	0.1420
lasso	Lasso Regression	55.8609	8442.5378	91.8640	0.5477	0.4773	0.4365	1.5110
en	Elastic Net	56.0510	8600.2812	92.7179	0.5393	0.4698	0.4351	1.4270
omp	Orthogonal Matching Pursuit	56.3890	8648.7291	92.9802	0.5366	0.4811	0.4412	0.1470
lr	Linear Regression	59.3074	9685.6521	98.3648	0.4811	0.4984	0.4653	8.1310
ada	AdaBoost Regressor	86.9859	13318.3008	115.1946	0.2859	0.6605	0.8578	13.4370
dt	Decision Tree Regressor	68.3502	13547.7070	116.3731	0.2735	0.5488	0.4814	1.4130
huber	Huber Regressor	71.6273	15839.2039	124.0355	0.1522	0.6038	0.4908	9.5800
knn	K Neighbors Regressor	85.6969	17414.2678	131.9413	0.0670	0.7024	0.7417	17.7580
llar	Lasso Least Angle Regression	91.2955	18666.8130	136.6068	-0.0001	0.7468	0.8802	0.1600
dummy	Dummy Regressor	91.2955	18666.8127	136.6068	-0.0001	0.7468	0.8802	0.0810
par	Passive Aggressive Regressor	1955.0254	59454847646.5381	147647.0177	-3130663.1440	0.8289	21.8571	0.4980

COMPARE R2 AND ADJUSTED_R2

	Model	R2	Adjusted_R2
lightgbm	Light Gradient Boosting Machine	0.6680	0.6674
rf	Random Forest Regressor	0.6500	0.6494
et	Extra Trees Regressor	0.6432	0.6425
gbr	Gradient Boosting Regressor	0.6308	0.6301
br	Bayesian Ridge	0.5492	0.5484
ridge	Ridge Regression	0.5490	0.5482
lar	Least Angle Regression	0.5490	0.5482
lasso	Lasso Regression	0.5477	0.5469
en	Elastic Net	0.5393	0.5385
omp	Orthogonal Matching Pursuit	0.5366	0.5357
lr	Linear Regression	0.4811	0.4801
ada	AdaBoost Regressor	0.2859	0.2846
dt	Decision Tree Regressor	0.2735	0.2722
huber	Huber Regressor	0.1522	0.1506
knn	K Neighbors Regressor	0.0670	0.0653

Hyper-Parameter Tuning

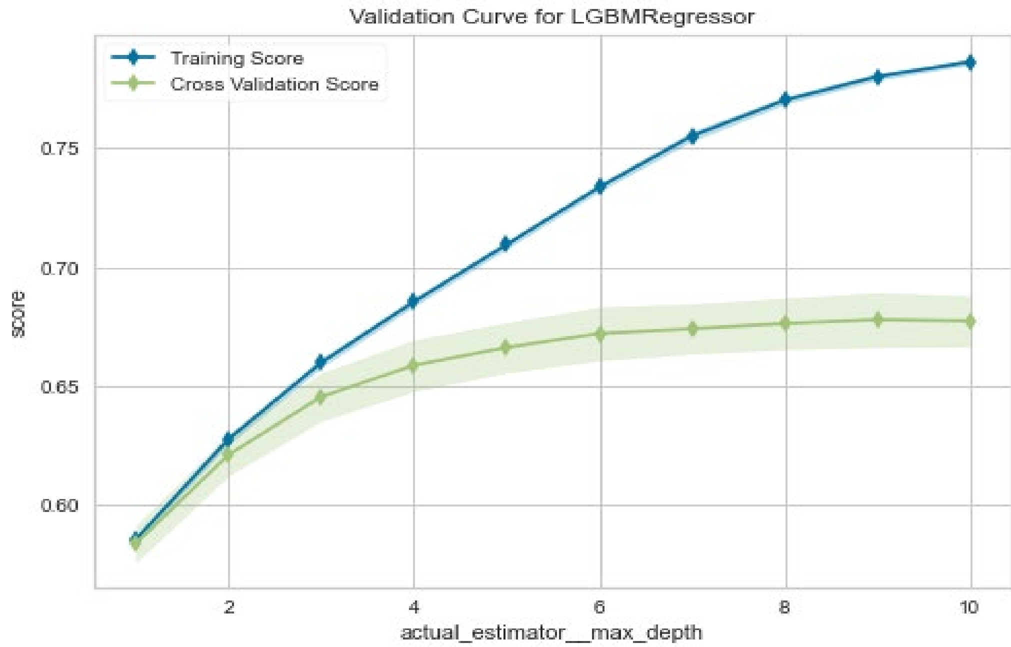
Parameters : bagging_fraction=0.6, bagging_freq=2, boosting_type='gbdt',class_weight=None, colsample_bytree=1.0, feature_fraction=0.4, importance_type='split', learning_rate=0.1, max_depth=-1, min_child_samples=41, min_child_weight=0.001, min_split_gain=0.9, n_estimators=260, n_jobs=-1, num_leaves=70, objective=None, random_state=123, reg_alpha=2, reg_lambda=3, silent='warn', subsample=1.0, subsample_for_bin=200000, subsample_freq=0

	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	46.3715	5873.4891	76.6387	0.6800	0.3911	0.3435
1	46.9955	5962.3763	77.2164	0.6871	0.3914	0.3454
2	48.0260	6385.7296	79.9108	0.6683	0.3953	0.3445
3	46.3177	5787.9146	76.0783	0.6765	0.3802	0.3365
4	46.2583	5897.1425	76.7929	0.6728	0.3878	0.3421
5	47.7254	6211.0249	78.8101	0.6863	0.3943	0.3526
6	47.3480	6209.5841	78.8009	0.6639	0.3861	0.3413
7	46.1371	5604.3148	74.8620	0.6842	0.3911	0.3456
8	48.0818	6398.0752	79.9880	0.6624	0.4015	0.3513
9	46.2216	5821.6848	76.3000	0.6952	0.3881	0.3434
Mean	46.9483	6015.1336	77.5398	0.6777	0.3907	0.3446
SD	0.7486	255.8831	1.6473	0.0103	0.0055	0.0044

The Above figure shows the Results after Hyper-parameter tuning. The Observed Mean R2 is 0.677 and the Adj R2 is 0.6771. The RMSE score is 77.53$ with a SD of 1.64$ and the MAPE is 0.344 with SD 0.0044. The no of Cross-validation iterations done is 10 with the above given Hyper-parameters. This is the best performing model we have gotten so far.

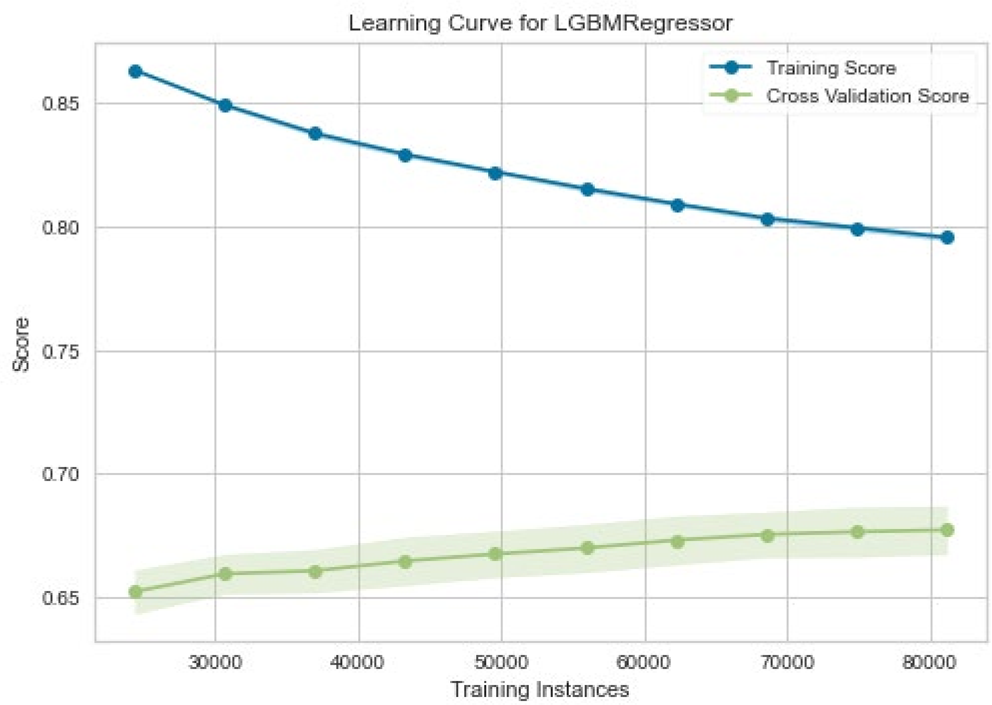
CHAPTER 4 - EVALUATION

Validation Curve:

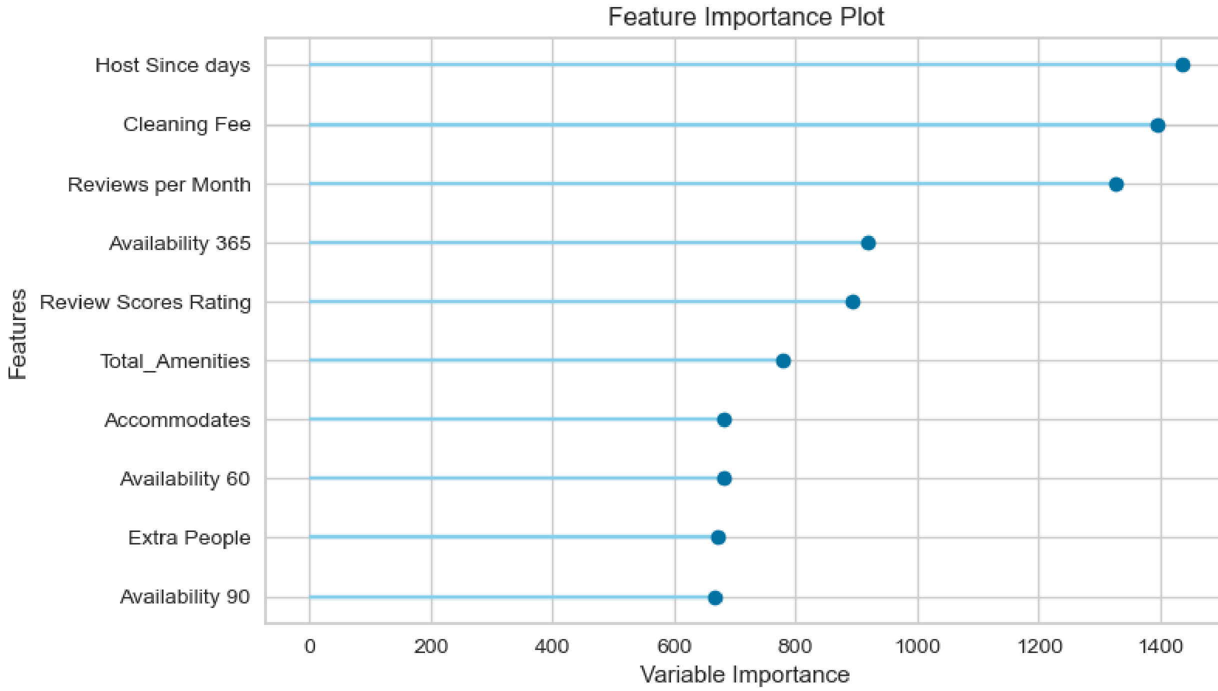


We can observe that the Training Score is around 0.78 and the Cross Validation score curve is maturing around 0.675.

Learning Curve:

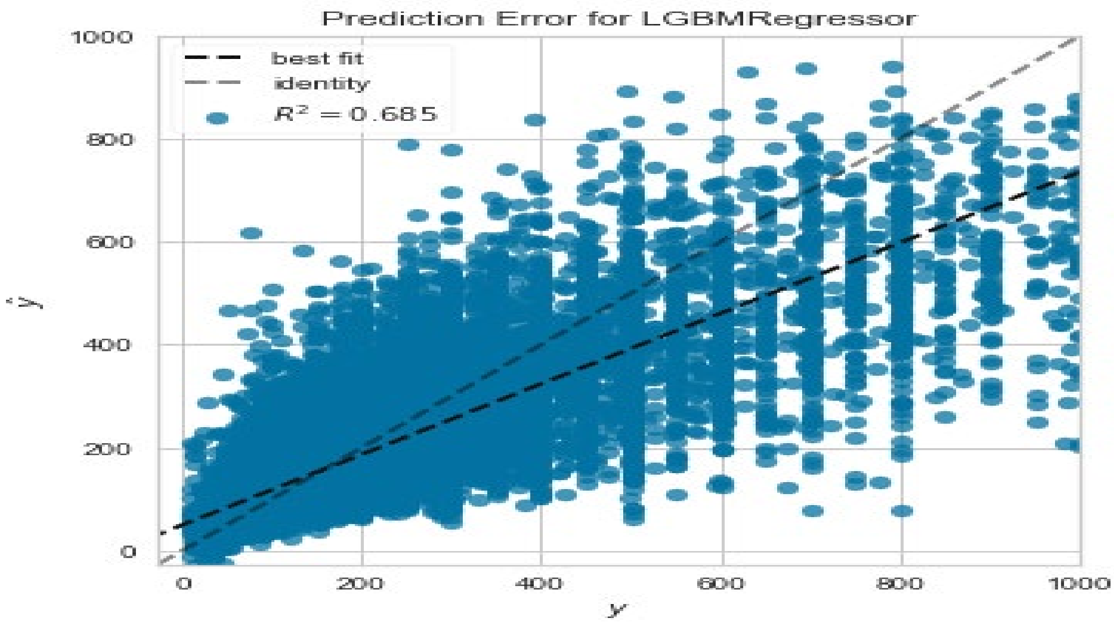


Feature Importance Plot:

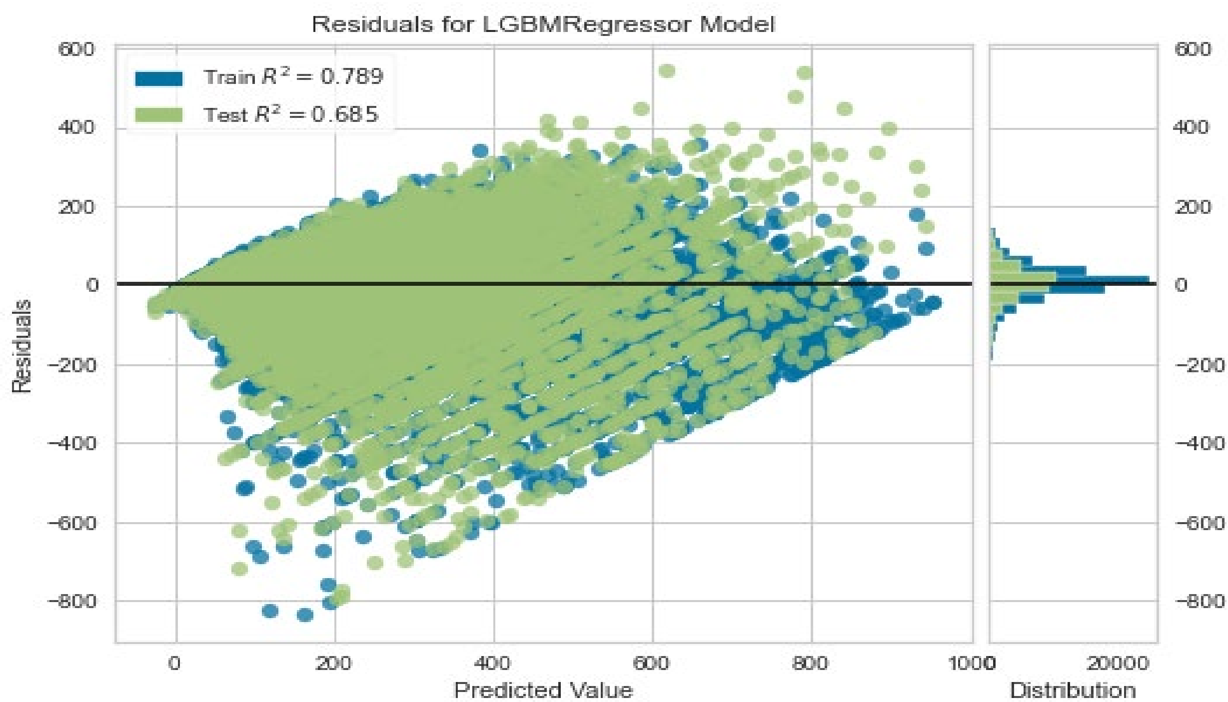


The most important features are 'Host Importance days', 'Cleaning Fee', 'Reviews per Month', 'Availability 365', 'Review Scores Rating', 'Total Amenities', 'Accommodates', 'Availability 60', Extra People' and 'Availability 90'

Prediction Error Plot:



Residual Plot:



The report should be prepared on A4 size paper, typed in New Times Roman. The font size of Chapter Heading should be 14 bold, subheading – 12 bold and further text in 12