

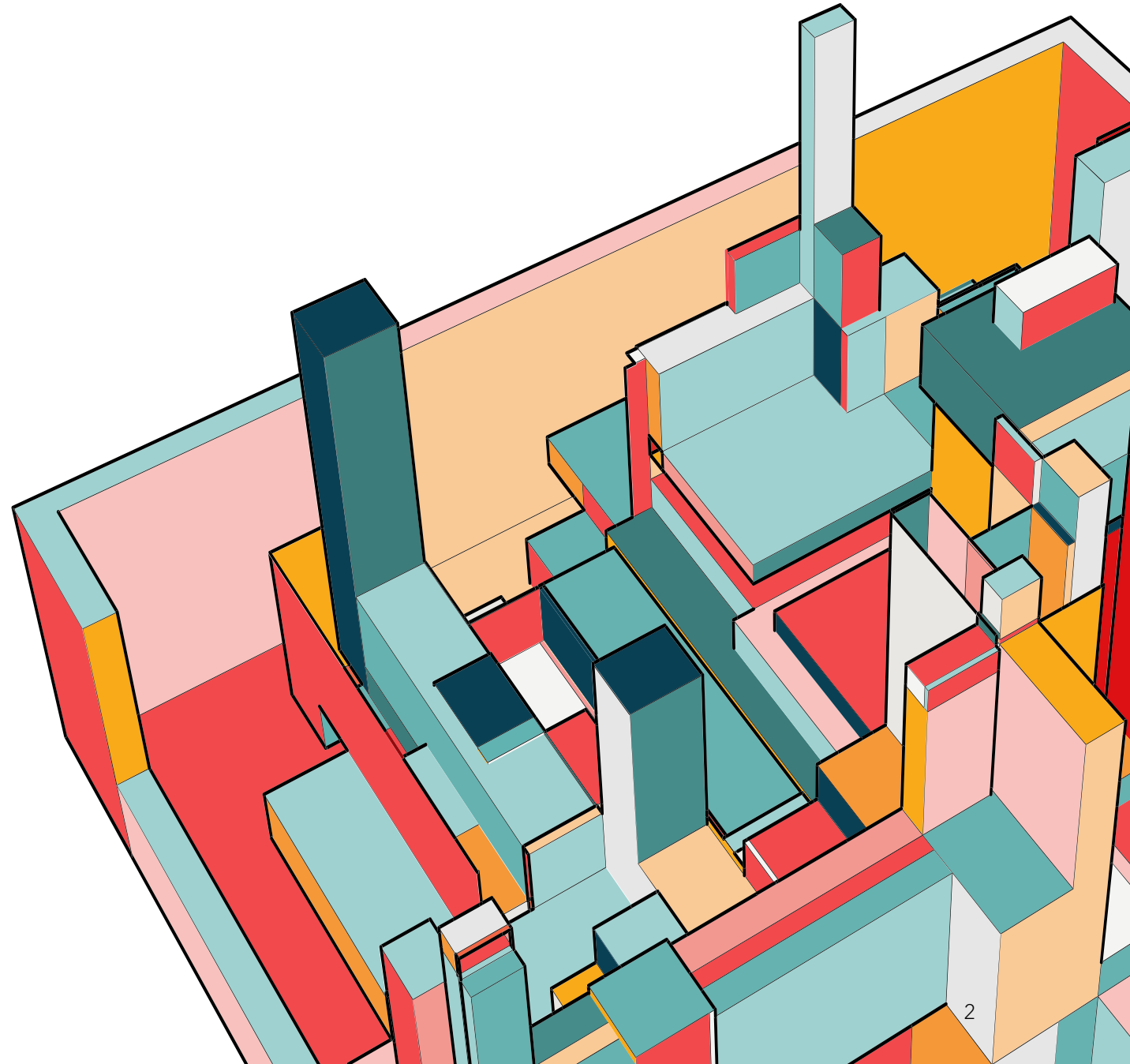


UNLOCKING AFFORDABLE AIRBNB STAYS WORLDWIDE: A DATA ANALYSIS

NIDHISH ABHIJIT GORE

OUR AIM

- Determine the Definition of "Affordable" Airbnb
- Compare Airbnb Costs by Country
- Assess Amenities within Budget Constraints
- Categorize Types of Airbnb Accommodations



OVERVIEW OF DATASET

1. AMENITIES

Internet, TV, Heating, Air-conditioning, Washer, Kitchen, Smoke detectors, Bathroom amenities, Pet friendly.

2. HOST FEATURES

Has profile pic, Is super host, Identity verified, Location exact, Instant bookable, Required guest profile picture, Require guest phone verification, Requires license.

3. LOCATIONS

Street, Neighbourhood, City, State, Zip code, Country code, Country, Latitude, Longitude.

4. PROPERTY DETAILS

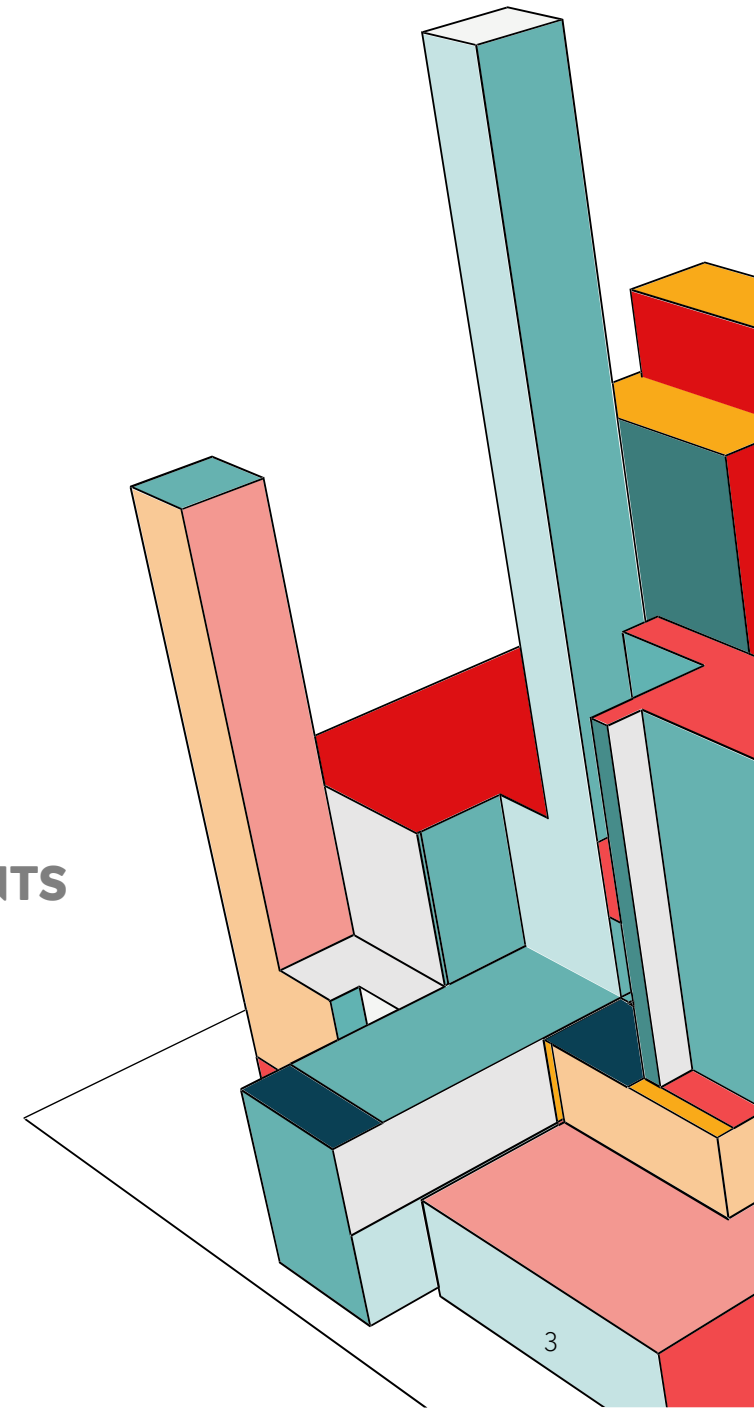
Property type, Room type, Accommodates, Bathrooms, Bedrooms, Beds, Bed types, Minimum nights, Maximum nights, Cancellation policy.

5. PRICE AND AVAILABILITY

Price, Security deposit, Cleaning fee, Availability 30,60,90 and 365.

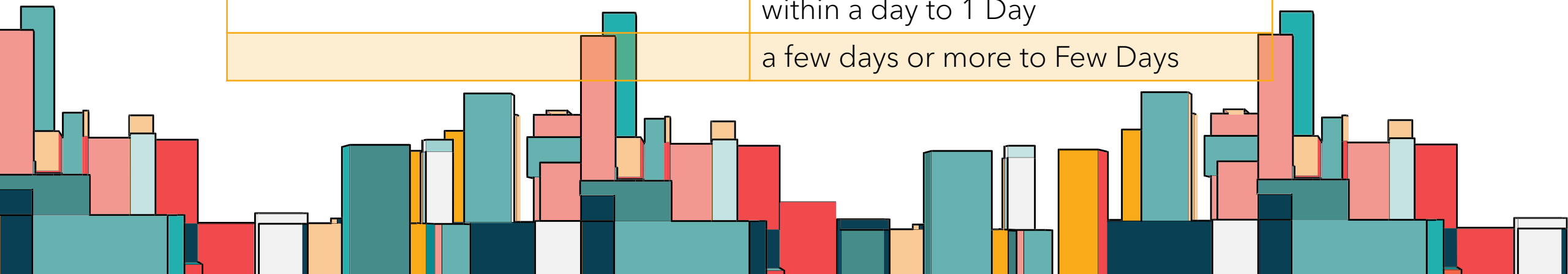
6. REVIEWS AND LISTING COUNTS

Number of reviews, Last review, First review, Review scores, Reviews per month, Host listing counts.



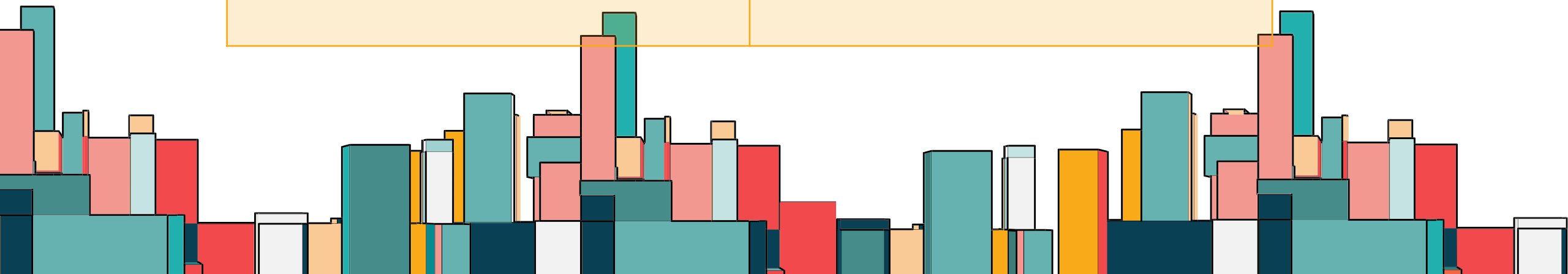
OVERVIEW OF CLEANING DONE TO DATASET

Variable	Changes
Amenities	Unique columns with different amenities (TV, Internet, etc)
Host Features	Divided host features in individual columns (Host Since, Superhost, etc)
Experienced Offered	Changed NA to none & replaced 1 with none
Host Response Time	Changed within an hour to 1 Hour
	within a few hours to Few Hours
	within a day to 1 Day
	a few days or more to Few Days



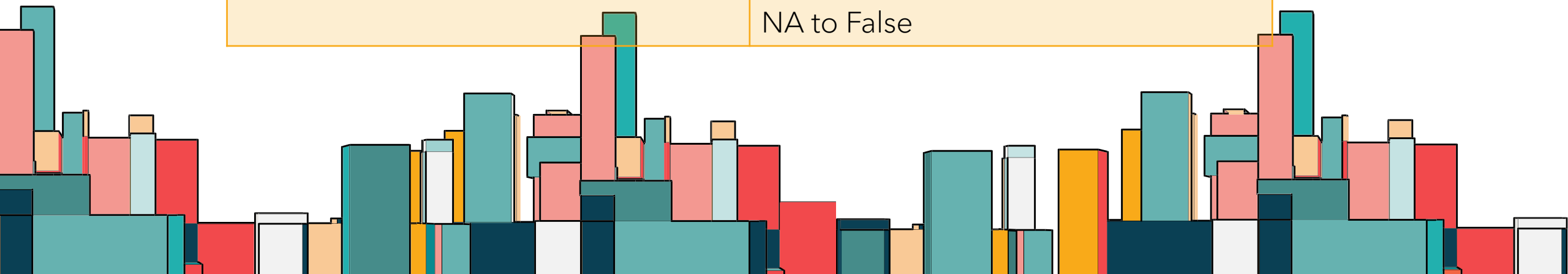
OVERVIEW OF CLEANING DONE TO DATASET

Host Response Time	Filled NA to 100
Host Acceptance Rate	Replaced Apartment to 100
Country	Checked corresponding lat lon and filled 0 to United States
Neighbourhood Cleansed	Filled Na to 0
Price	Dropped the NA rows
Country	Replaced NA with Italy after checking corresponding city state.



OVERVIEW OF CLEANING DONE TO DATASET

Cancellation Policy	Strict_new to Strict
	Moderate_new to moderate
	Flexible_new to flexible
	Super_strict_60_new to Super_strict_60
	Super_strict_30_new to Super_strict_30
Has Availability	Changed t to True
	0.21 to False
	NA to False



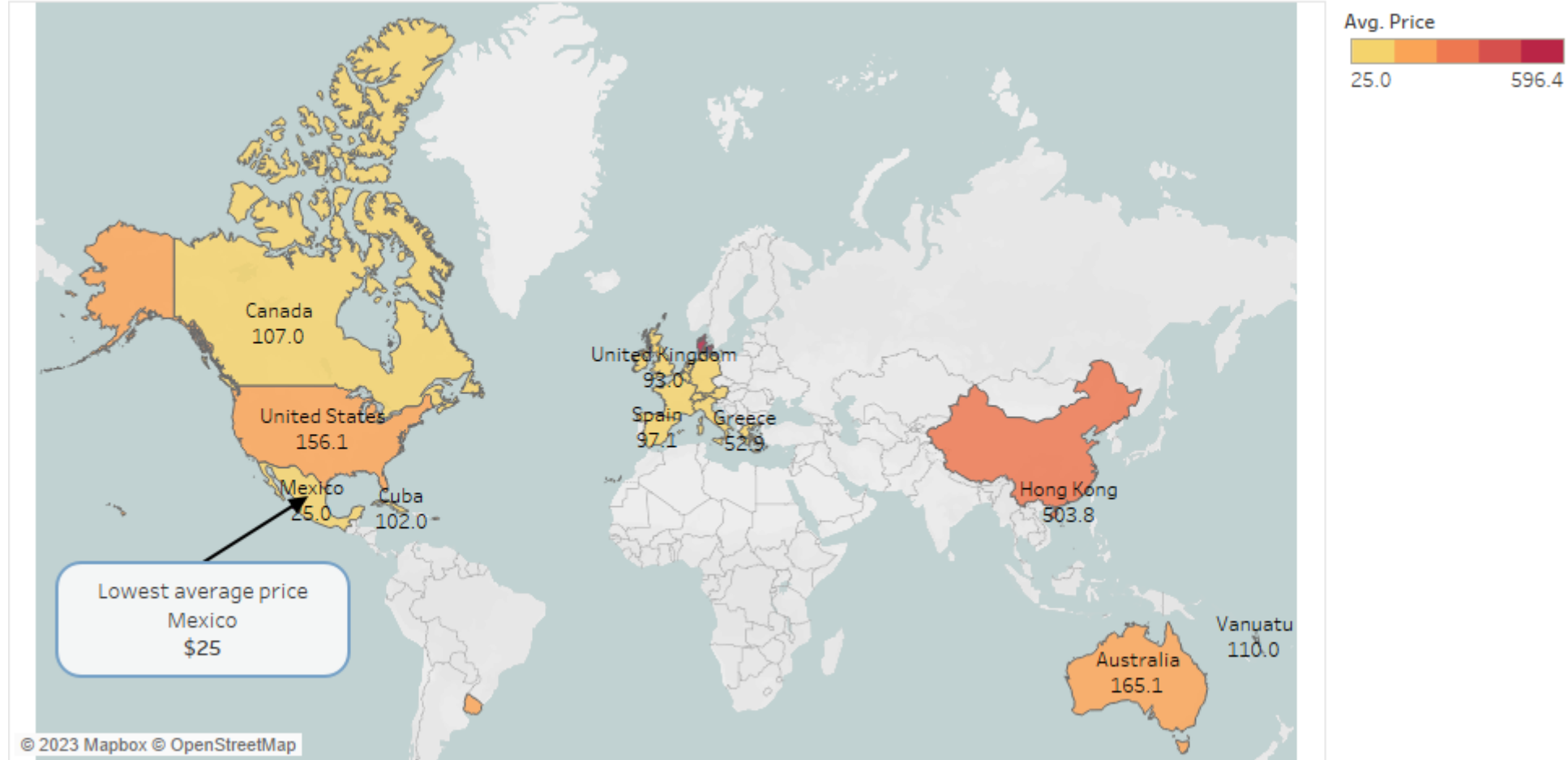


DATA AFTER CLEANING

- So, after cleaning the dataset the shape of the it has changed from (494954, 89) to (486996, 75).
- We dropped few columns and added few new columns.
- We have 75 attributes after adding all the features, amenities, dropping off the column and dropping the null values of the price column

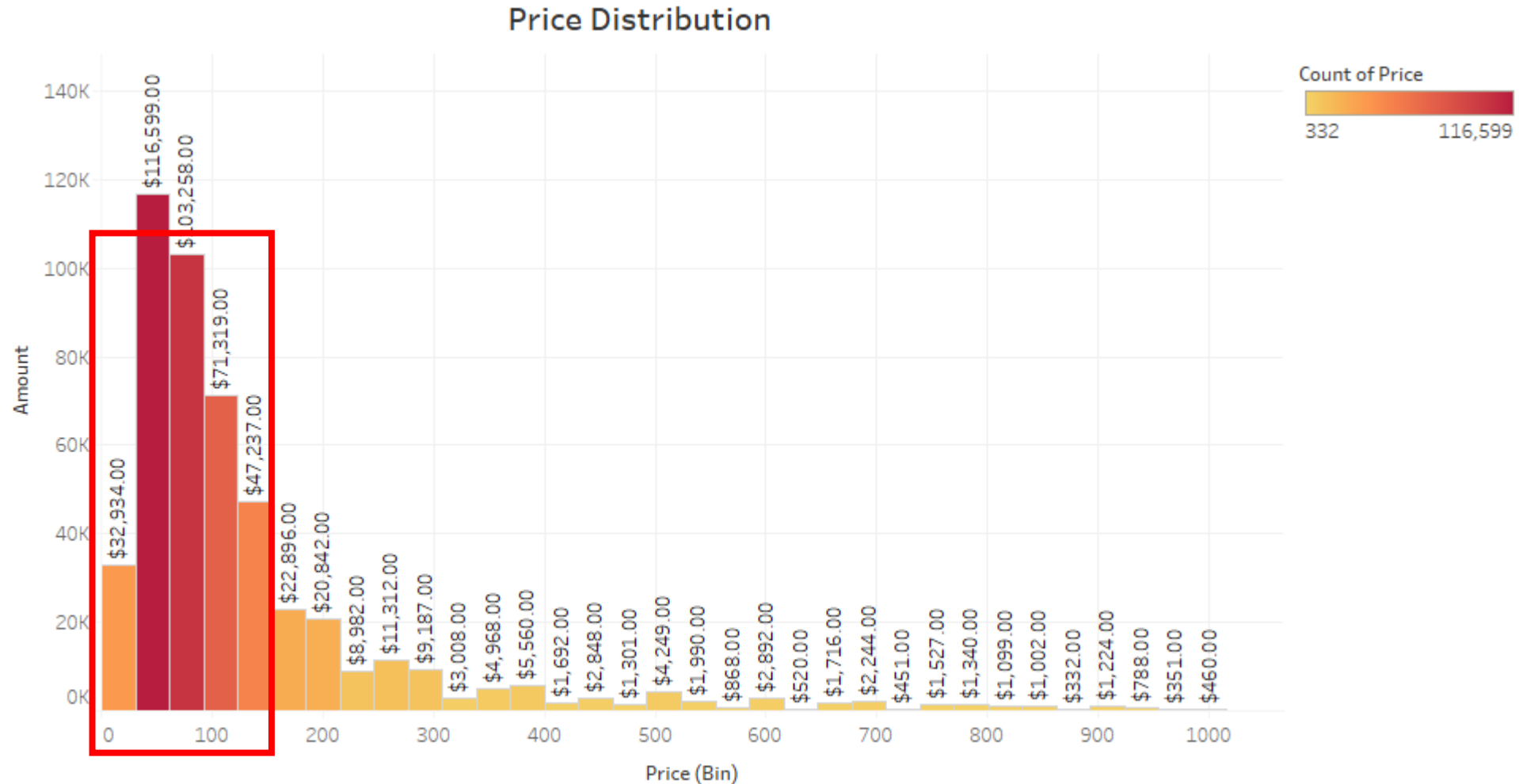
LET'S CHECK THE PRICES!

Lowest Average Price

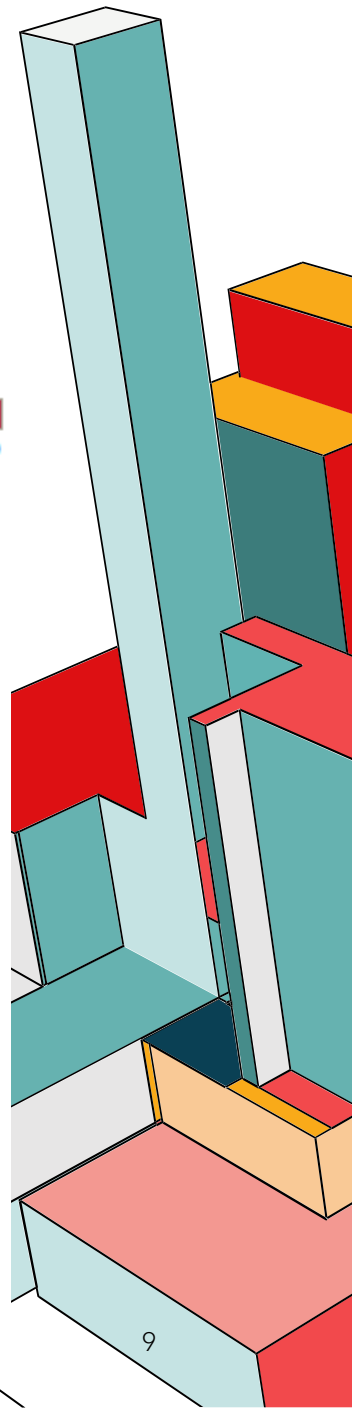


Map based on Longitude (generated) and Latitude (generated). Color shows average of Price. The marks are labeled by Country and average of Price. Details are shown for Country. The view is filtered on Country and average of Price. The Country filter keeps 22 of 22 members. The average of Price filter ranges from 25.0 to 596.4.

WHAT'S THE PRICE DISTRIBUTION?

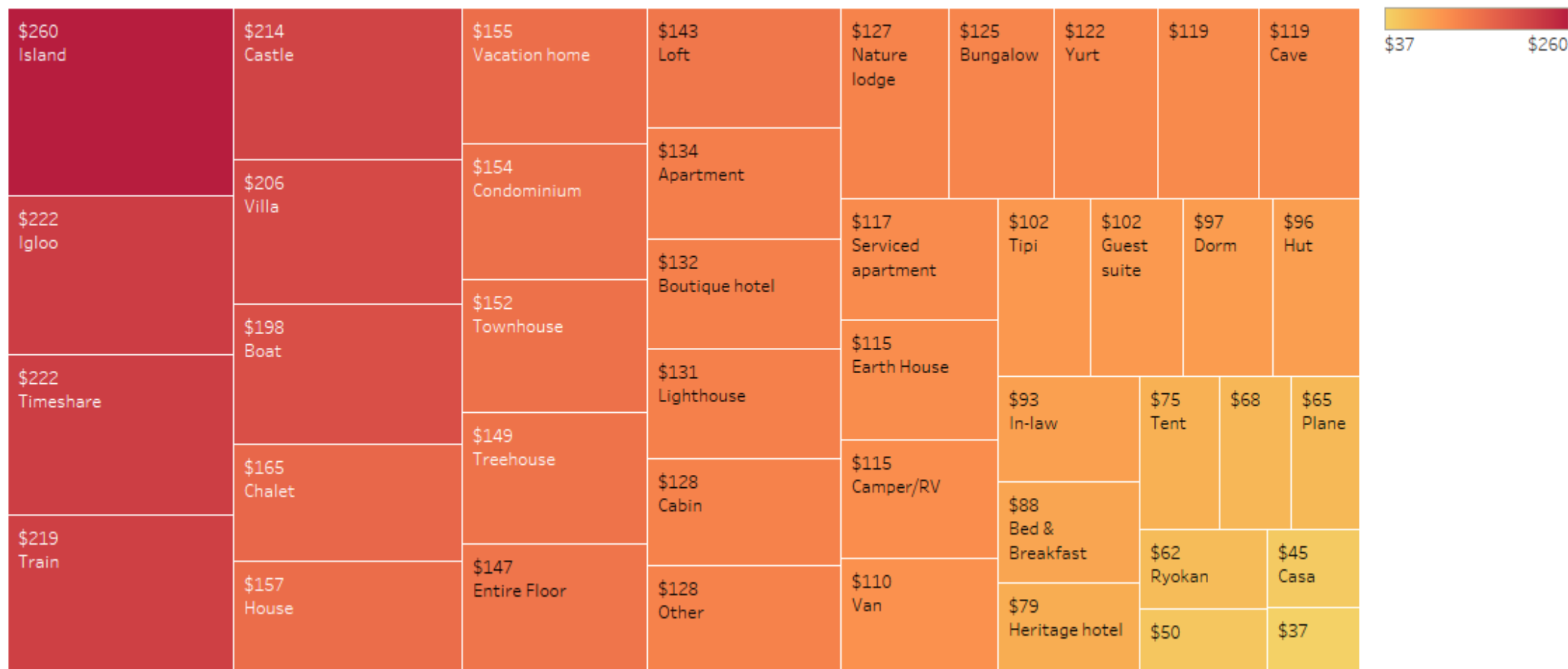


The trend of count of Price for Price (bin). Color shows count of Price. The marks are labeled by count of Price.



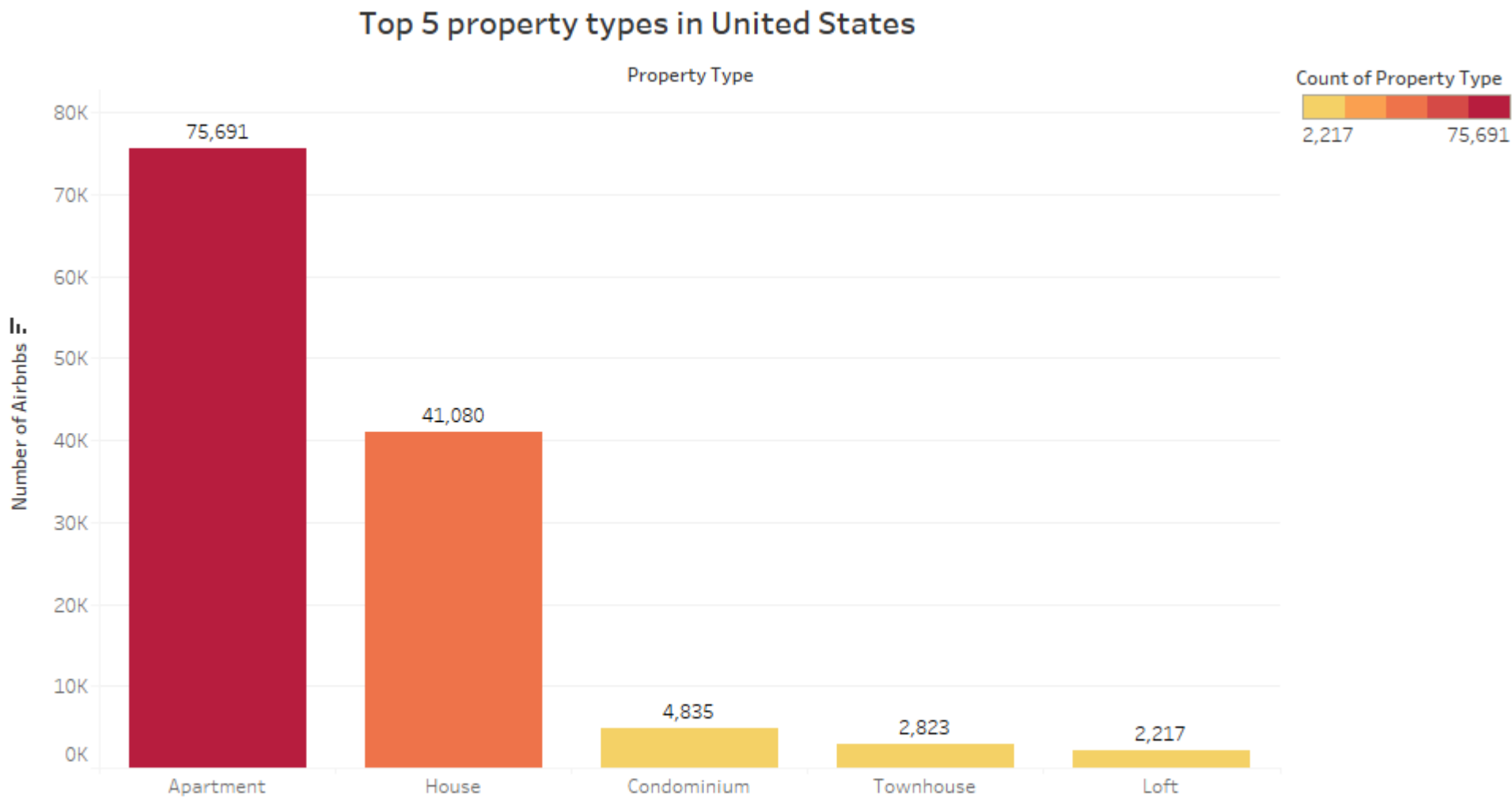
LET'S SEE THE TYPES OF PROPERTY!

Price Per property type



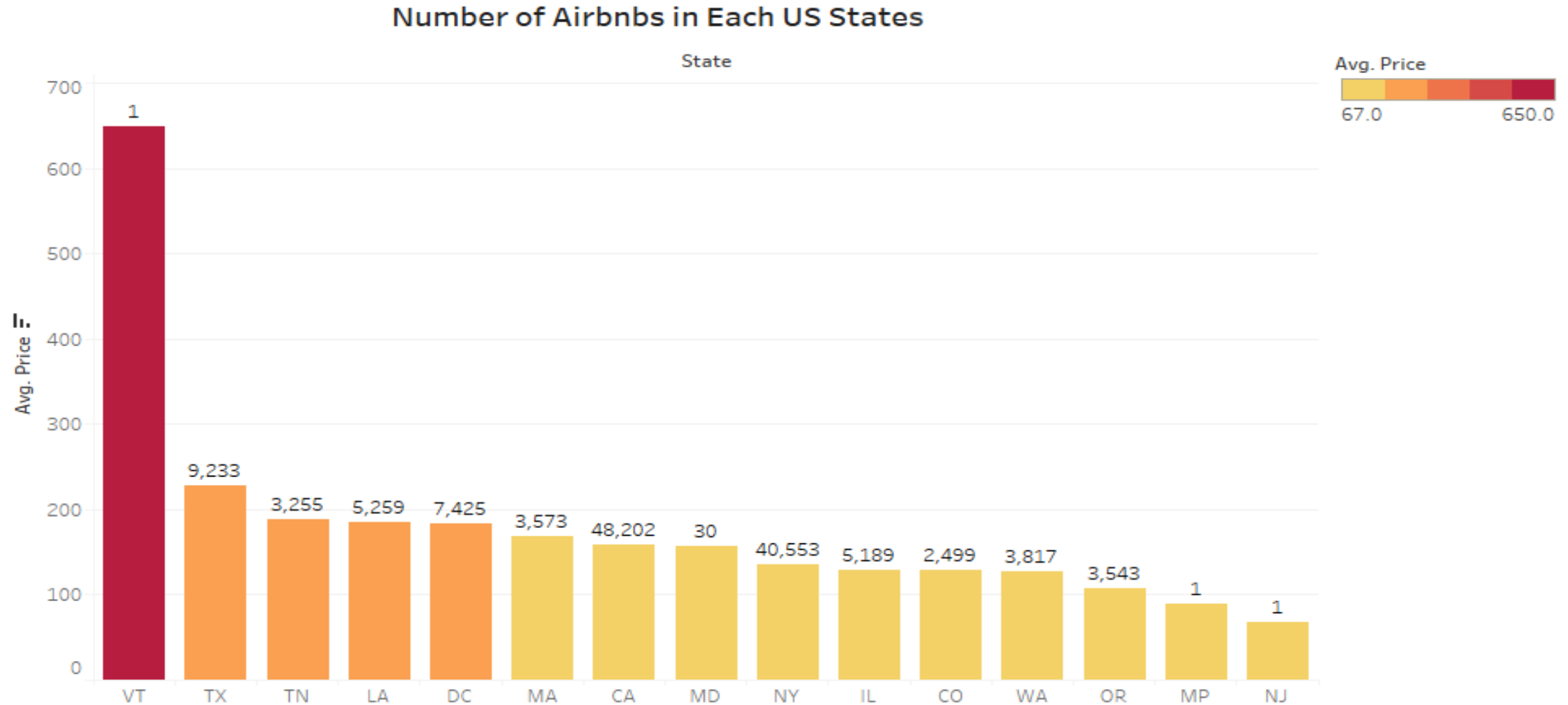
Average of Price and Property Type. Color shows average of Price. Size shows average of Price. The marks are labeled by average of Price and Property Type. The view is filtered on Property Type, which excludes Car.

LET'S SEE THE AVERAGE PRICE!



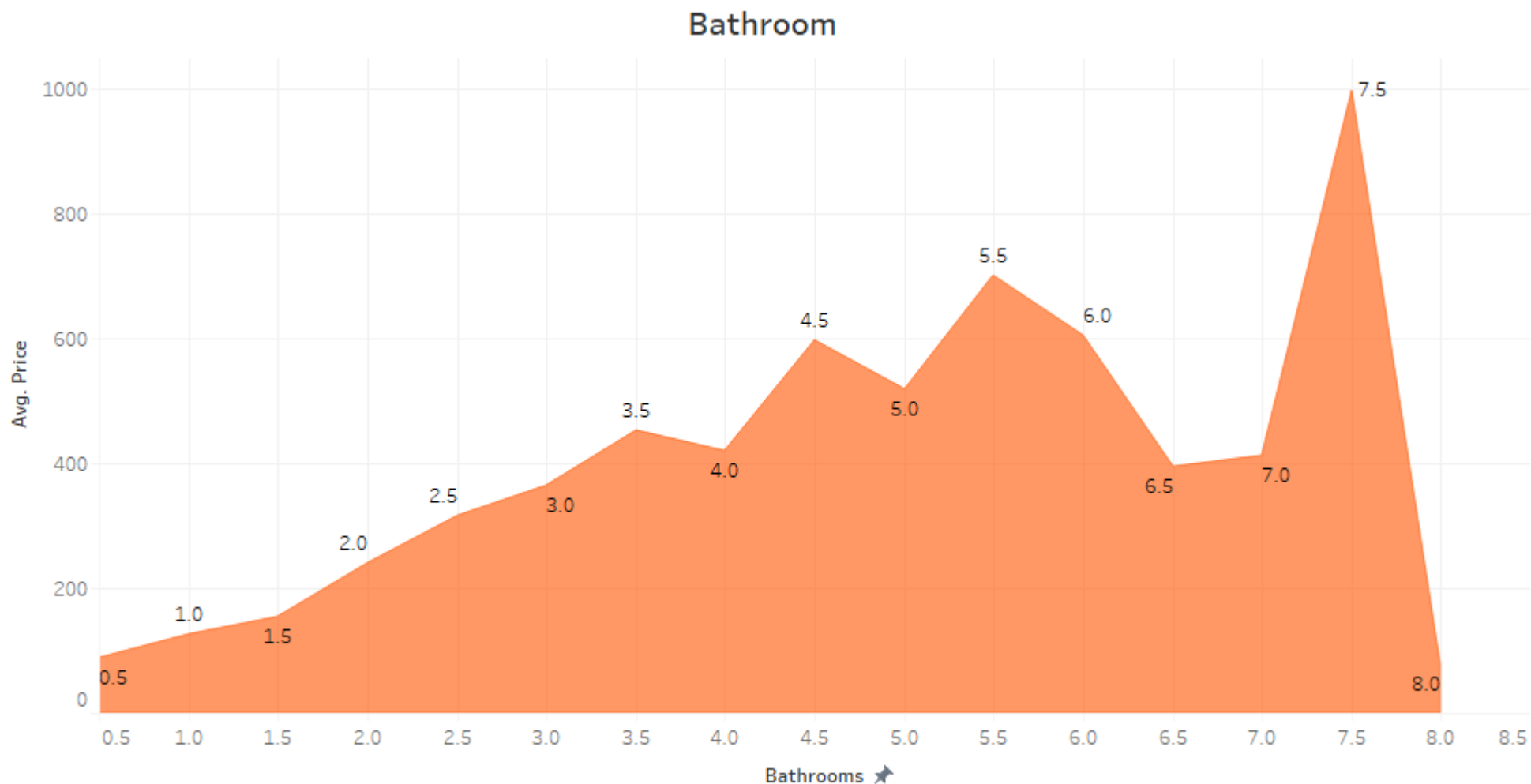
Count of Property Type for each Property Type. Color shows count of Property Type. The marks are labeled by count of Property Type. The data is filtered on Country, which keeps United States. The view is filtered on Property Type, which keeps 10 of 44 members.

AVERAGE AIRBNB PRICES IN US STATES



Average of Price for each State. Color shows average of Price. The marks are labeled by count of State. The data is filtered on Country, which keeps United States. The view is filtered on State, which excludes 70.00.

LET'S CHECK THE BEDROOMS AND BATHROOMS



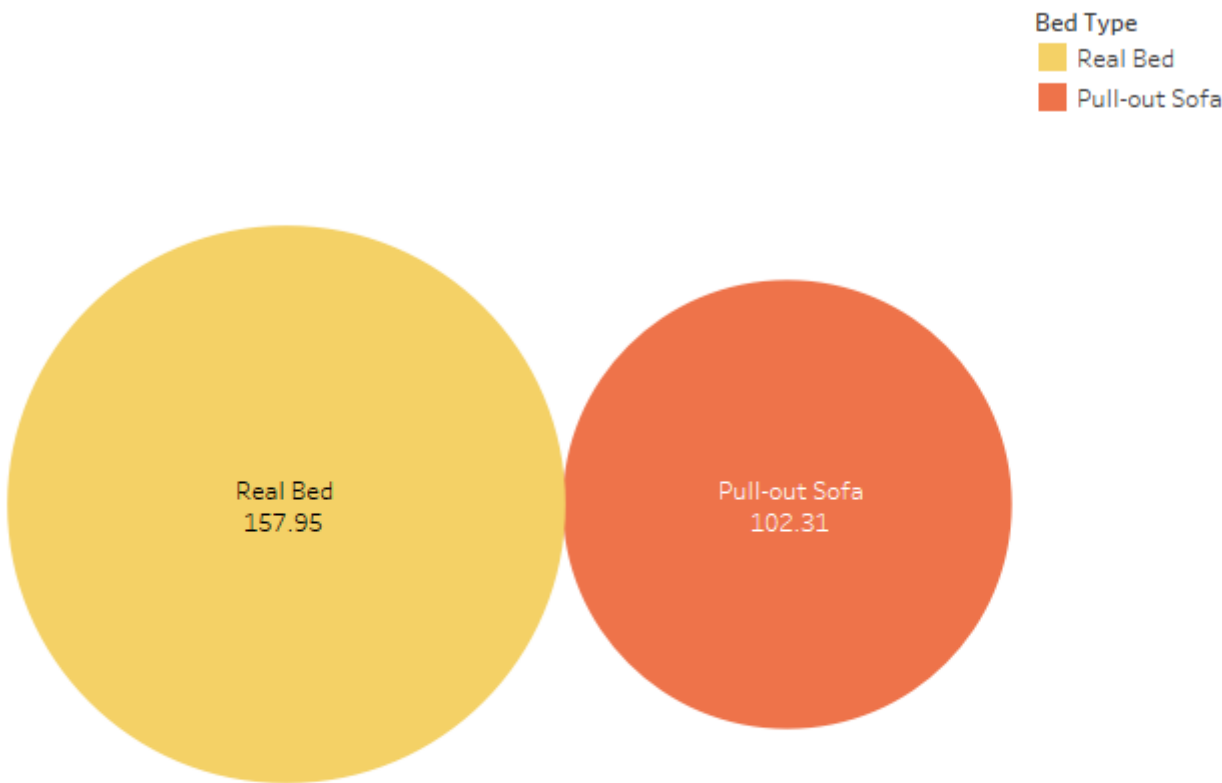
The plot of average of Price for Bathrooms. The marks are labeled by Bathrooms. The data is filtered on Country and Bathrooms. The Country filter keeps United States. The Bathrooms filter excludes 0.

BED TYPES?

Bed Type Distribution

Bed Type	
Real Bed	128,715
Pull-out Sofa	1,037
Couch	480
Airbed	894
Futon	1,456

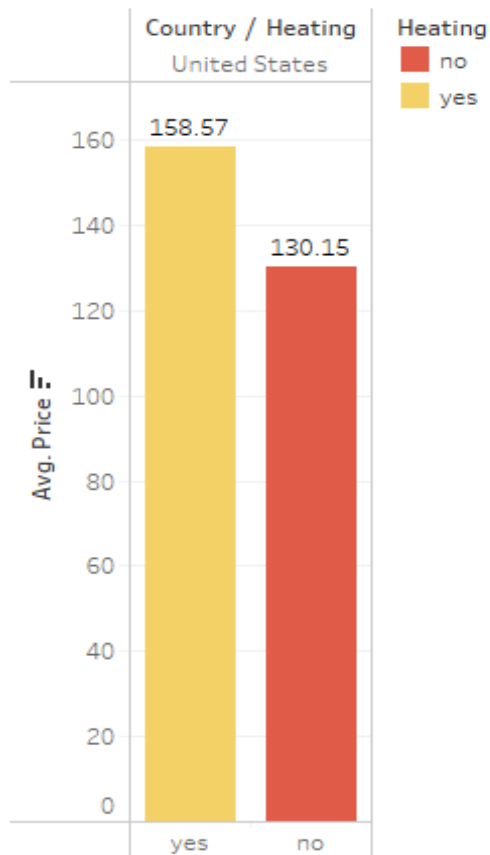
Bed types with price filter



Bed Type and average of Price. Color shows details about Bed Type. Size shows average of Price. The marks are labeled by Bed Type and average of Price. The data is filtered on Country, which keeps United States. The view is filtered on average of Price, which ranges from 100.00 to 160.00.

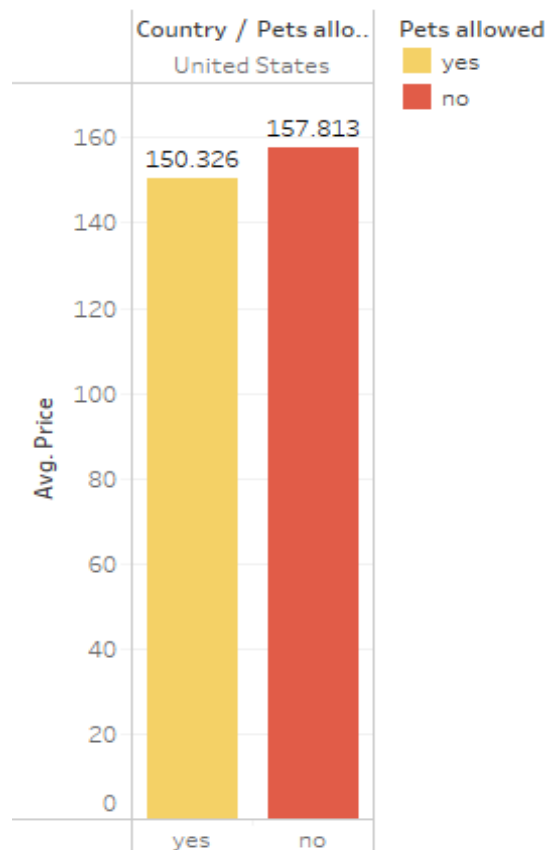
THE TRUTH ABOUT AIRBNB AMENITIES

Heating



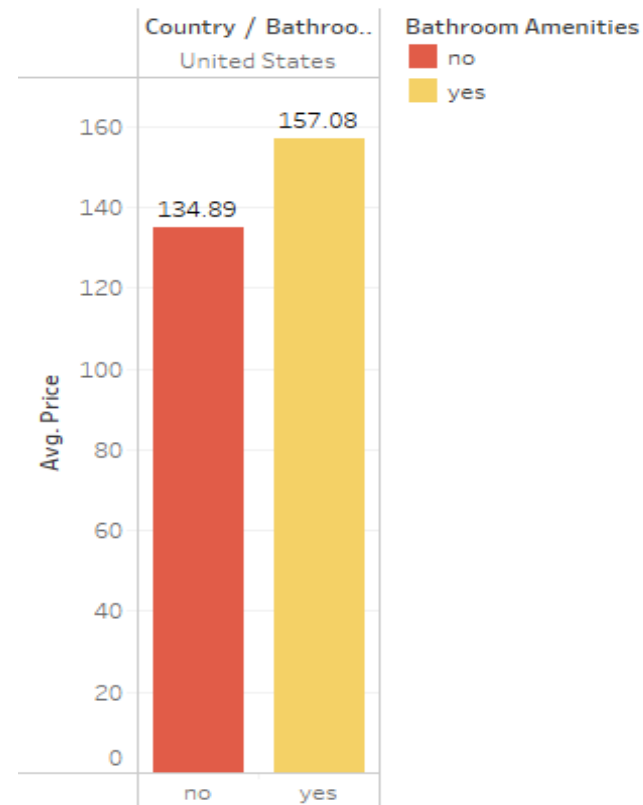
Average of Price for each Heating broken down by Country. Color shows details about Heating. The marks are labeled by average of Price. The view is filtered on Country, which keeps United States.

Man's Best Friend



Average of Price for each Pets allowed broken down by Country. Color shows details about Pets allowed. The marks are labeled by average of Price. The view is filtered on Country, which keeps United States.

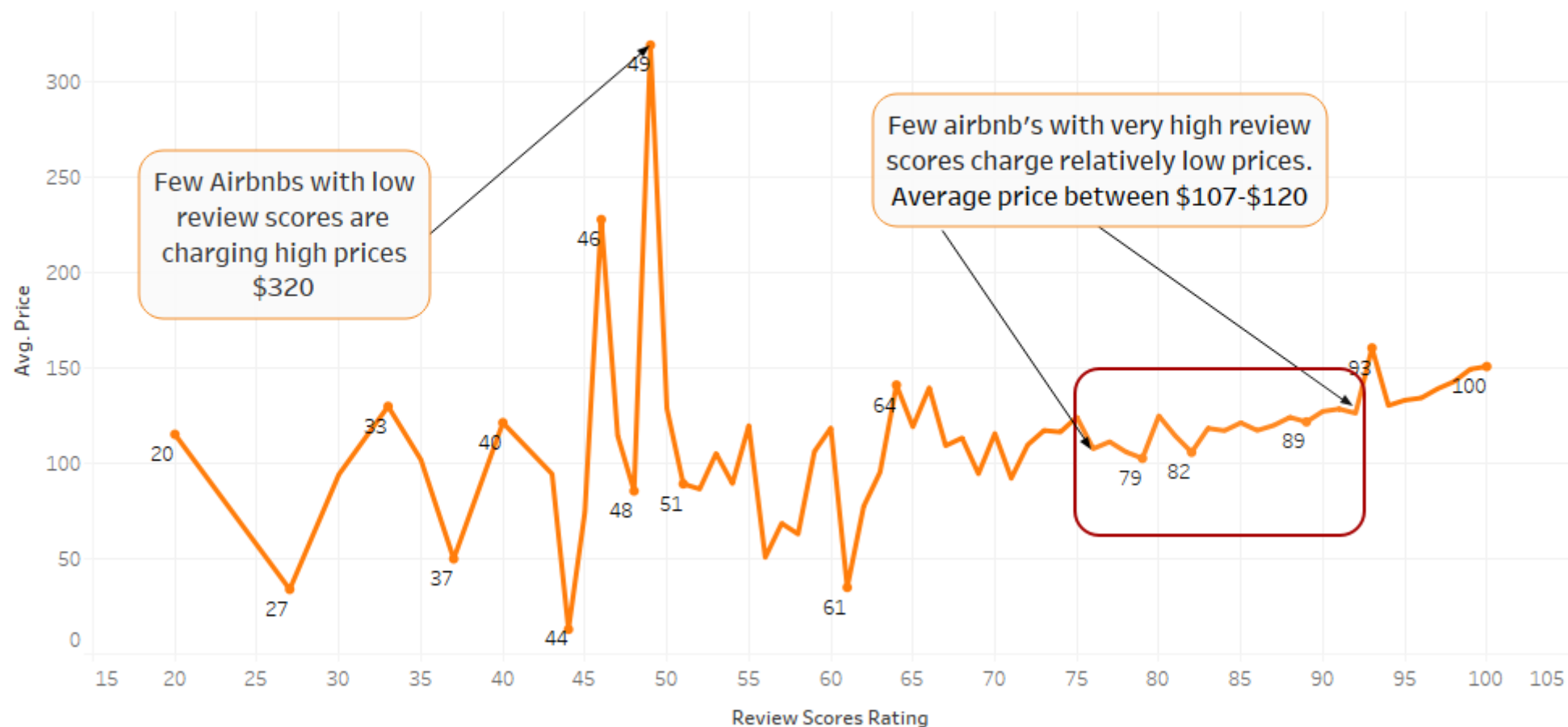
Bathroom Amenities



Average of Price for each Bathroom Amenities broken down by Country. Color shows details about Bathroom Amenities. The marks are labeled by average of Price. The view is filtered on Country, which keeps United States.

SOME REVIEW STATS!

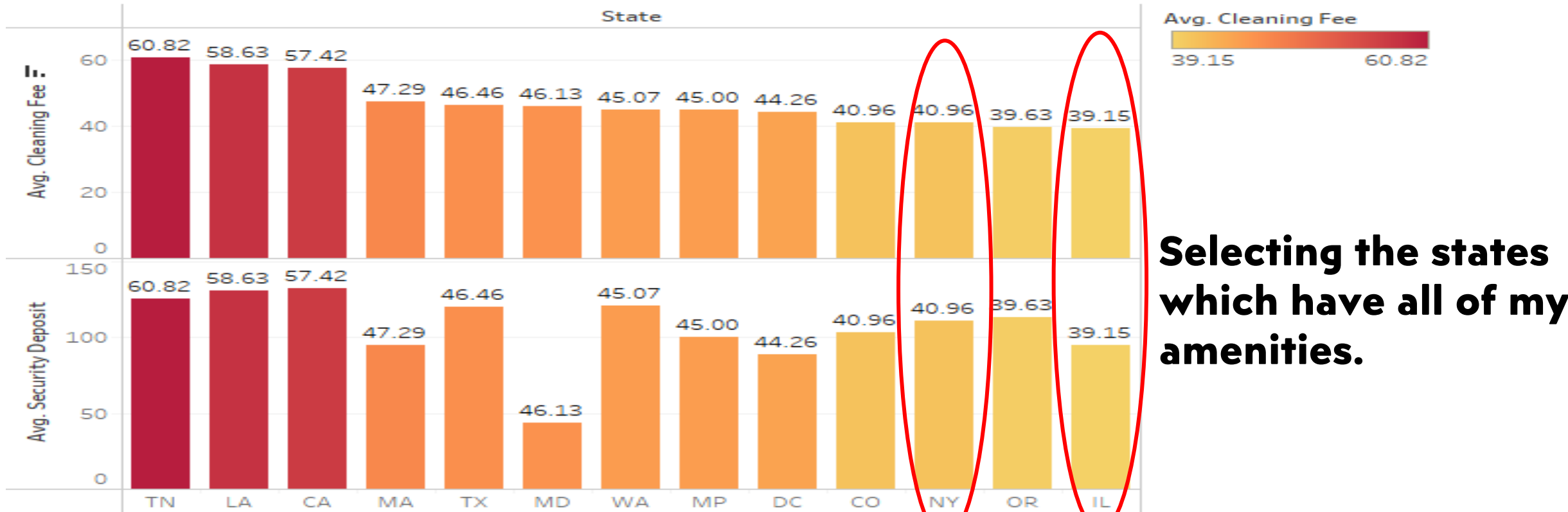
How Review Scores Affect Prices



The trend of average of Price for Review Scores Rating. The marks are labeled by Review Scores Rating. The data is filtered on Country, which keeps 8 of 22 members.

AVERAGE AIRBNB PRICE BY MUST-HAVE AMENITIES

State wise average cleaning fee and security fee



Average of Cleaning Fee and average of Security Deposit for each State. Color shows average of Cleaning Fee. The marks are labeled by average of Cleaning Fee. The data is filtered on Country, which keeps United States. The view is filtered on State, which excludes 70.00, NJ and VT.

Selecting the states which have all of my amenities.

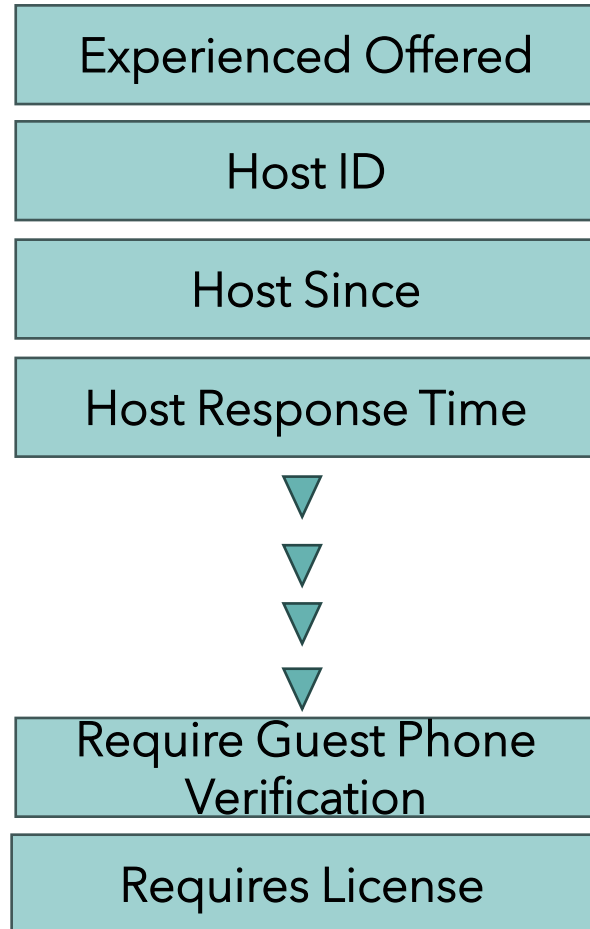


ANALYSIS PART -2 PRICE PREDICTION MODEL

NIDHISH ABHIJIT GORE

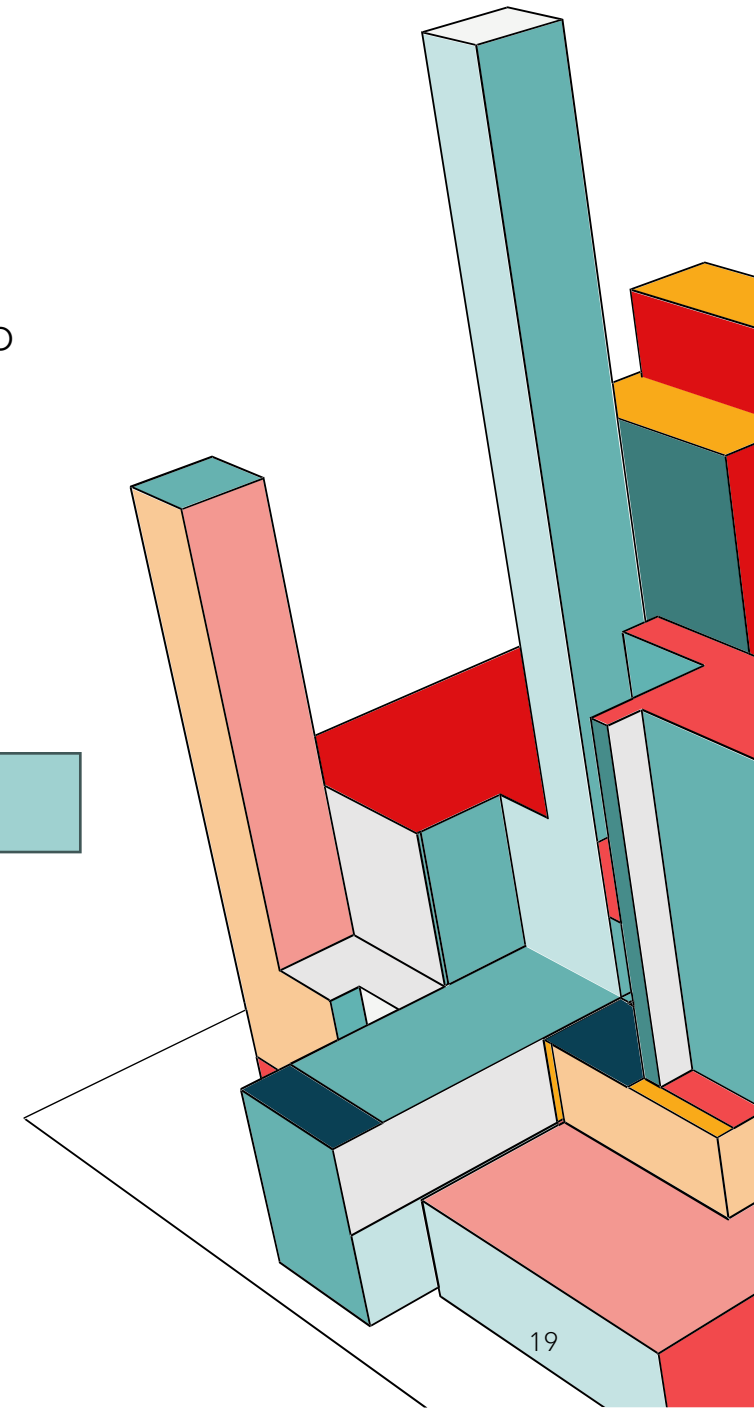
OVERVIEW OF DATASET

Information related to Airbnb listings
Input Variables



After filtering the data according to
US as country, we have -
It's (132582 , 74)

Output Variable



DETAILS OF OUR DATASET

Attributes: 72 Independent and 1
Dependent

Data Type:
Object - 39
Integer - 27
Float - 5
DateTime - 3
Boolean - 1

Host Details

Host ID, Host Since, Host Response Time,
Host Response Rate, Host Total Listings Count

Address Information

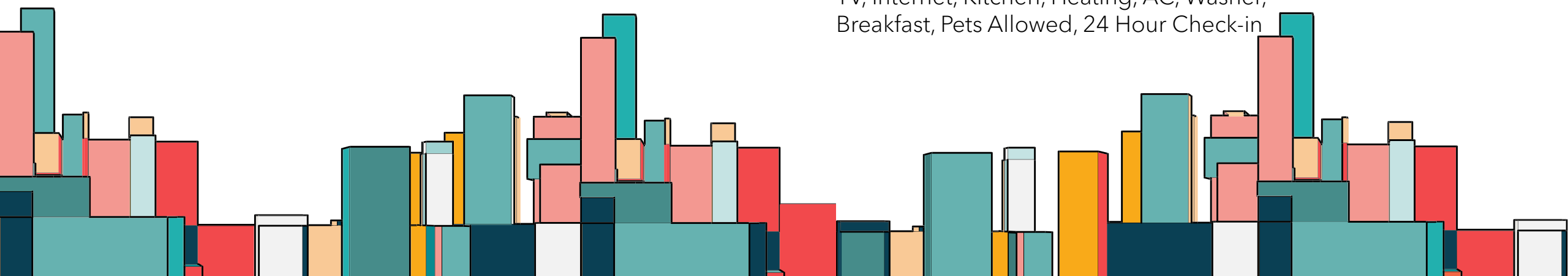
Street, Neighbourhood, Neighbourhood Cleansed,
City, State, Zipcode, Country, Latitude, Longitude

Property Details

Property Type, Room Type, Accommodate, Bathrooms,
Bedrooms, Beds, Bed Type

Amenities

TV, Internet, Kitchen, Heating, AC, Washer,
Breakfast, Pets Allowed, 24 Hour Check-in

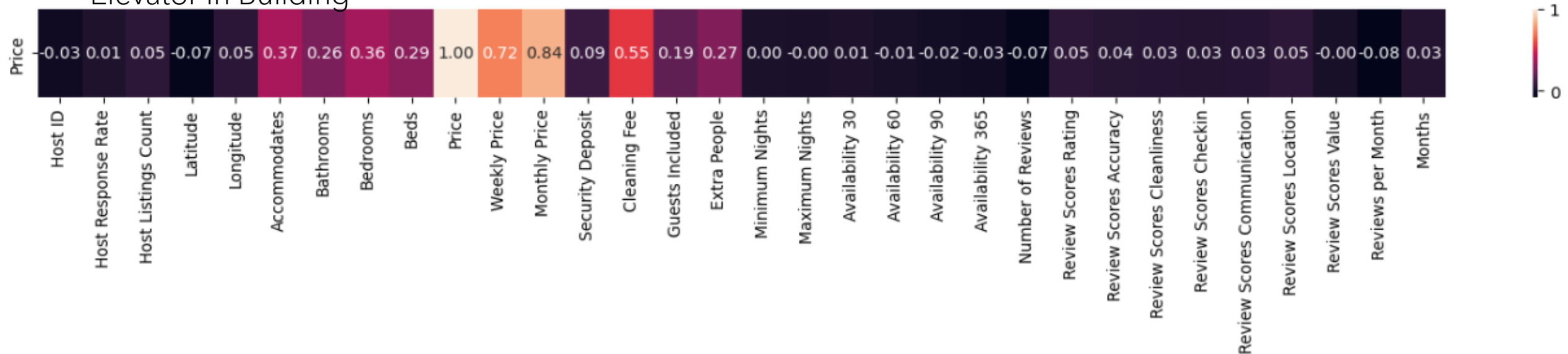


DROPPING COLUMNS

- Host Since
- Host ID
- First Review
- Last Review
- Experience Offered
- Street
- Neighbourhood
- Neighbourhood Cleansed
- Country
- Carnbon Monoxide detector
- Elevator In Building
- Instant Bookable
- Require Guest Profile Picture
- Require Guest Phone Verification
- Requires License
- Host Profile Pic

Splitting Dependent and Independent Variable

Attributes: 57 Independent and 1 Dependent



ONE HOT ENCODING

- Categorical variables present in the dataset like - City, State, Bed Type, TV, Internet
- Total Categorical variable - 26(Object)
- For us to run regression models, we need to convert this variables into numbers.
- We convert each categorical value into a new categorical column and assign a binary value of 1
- A example from our dataset -

Original

Bed Type
Entire home/apt
Private room
Shared Room

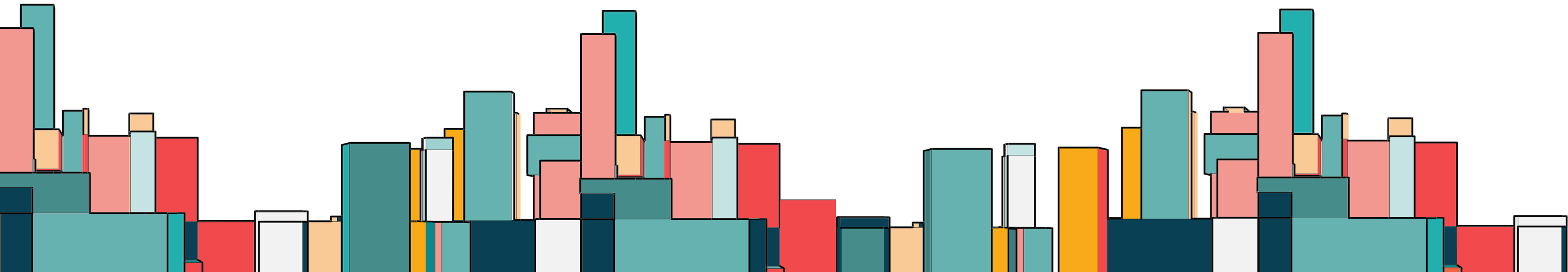


One Hot Encoded

Bed Type	Entire home/apt	Private room	Shared room
Entire home/apt	1	0	0
Private room	0	1	0
Shared Room	0	0	1

Dataset after One Hot Encoding

Shape : (132582, 1129)

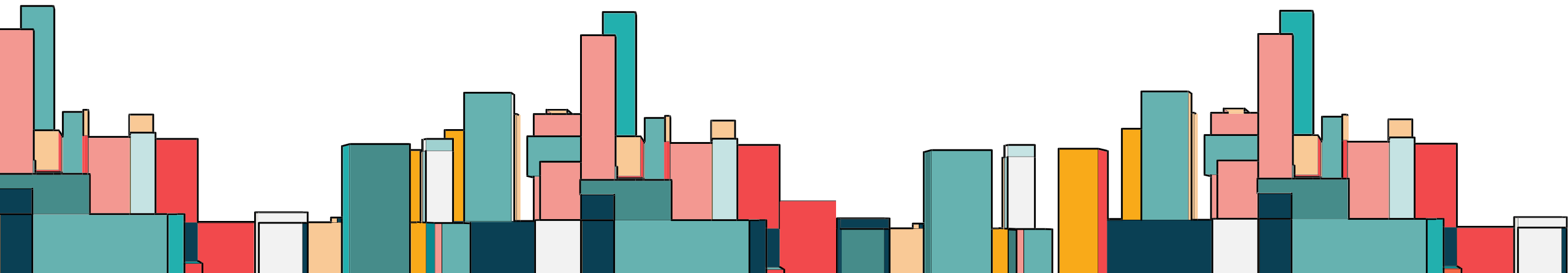


DATA SPLITTING

- Splitting the dataset into two different dataset
- Training Dataset - 75%
- Validation Dataset - 25%
- Shape of training data set after one hot encoding and data splitting
(99436, 1129)
- Shape of validation data set after one hot encoding and data splitting
(33146, 1129)

SCALLING – STANDARD

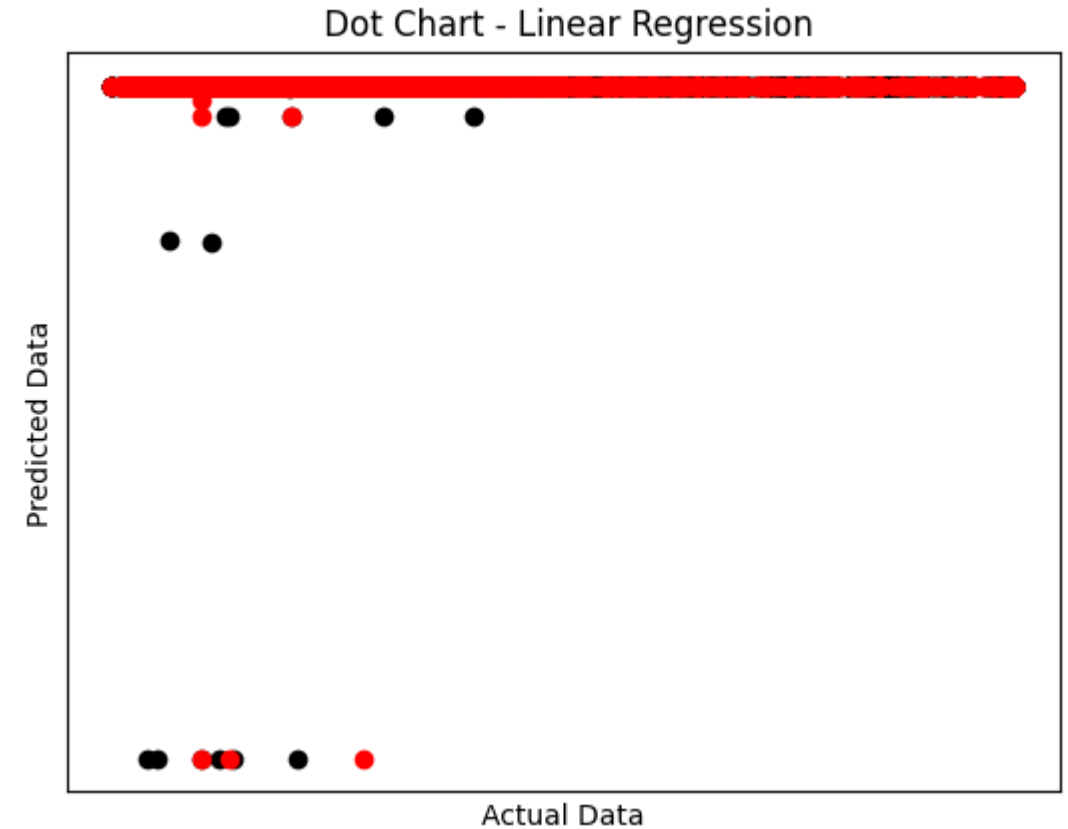
- Standardization preserves meaningful differences between values, which is important in context of price prediction
- Standardization is less affected by outliers compared to normalization
- Standardization is less prone to data leakages.



PRINCIPAL COMPONENT ANALYSIS (PCA)

- PCA used for feature extraction and dimensionality reduction.
- Very low coefficient of determination for Training as well as validation data

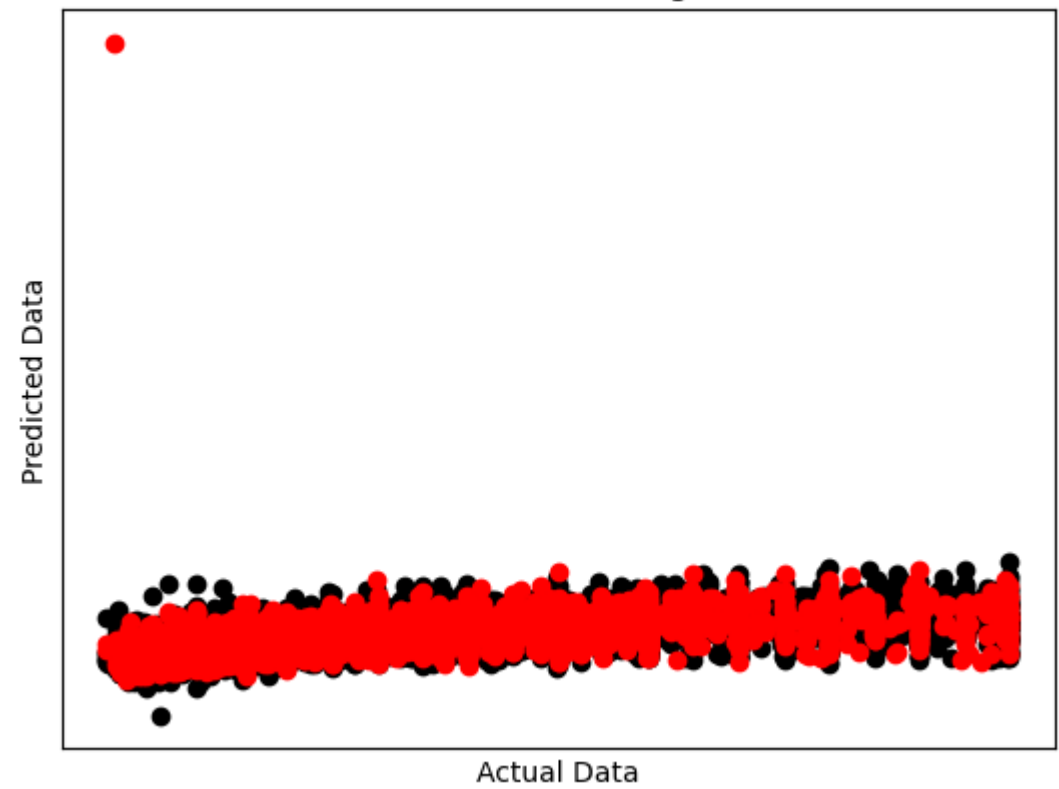
Training data - Mean squared error: 18833.54
Validation data - Mean squared error: 18345.74
Training data Coefficient of determination: 0.00
Validation data Coefficient of determination: -0.00



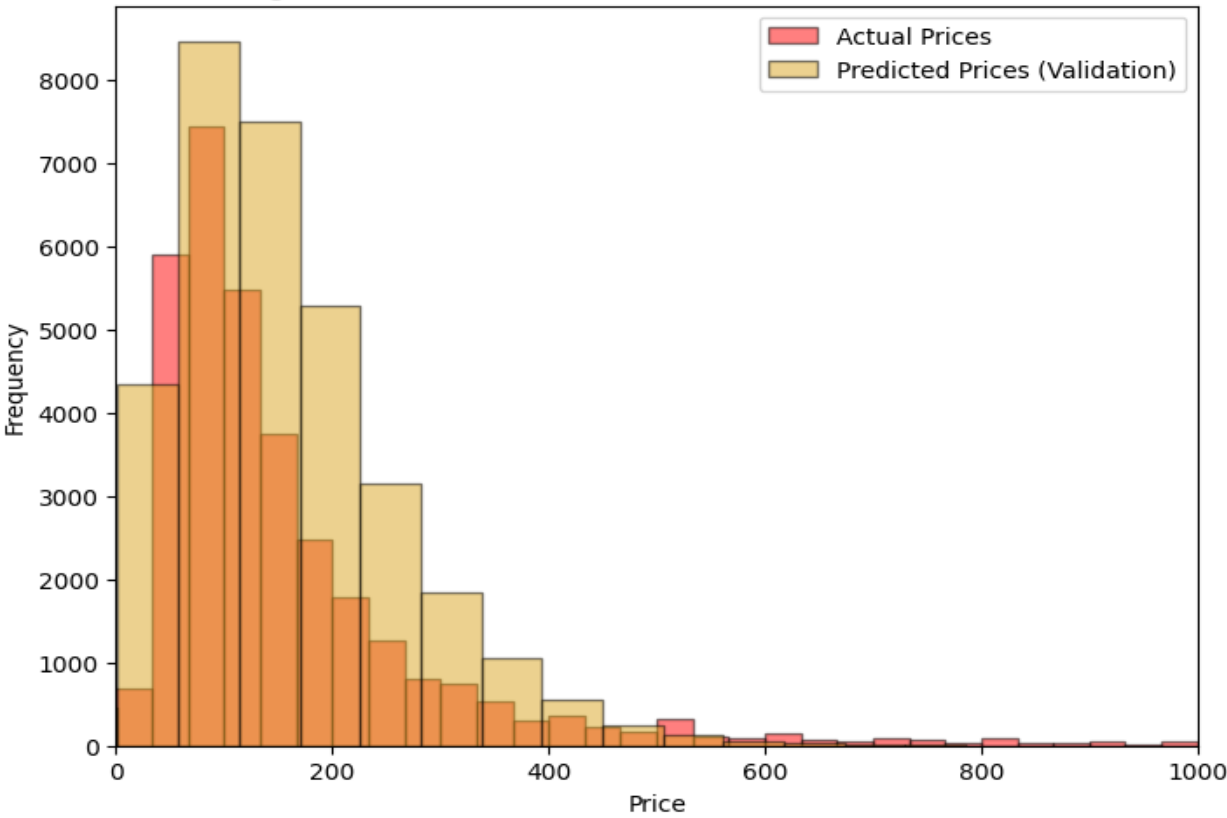
REGRESSION ANALYSIS

Training data -Mean squared error: 8138.14
Validation data - Mean squared error: 8869.18
Training data Coefficient of determination: 0.57
Validation data Coefficient of determination: 0.52
Adjusted R2 for Training data: 0.56
Adjusted R2 for Validation data: 0.50

Dot Chart - Linear Regression



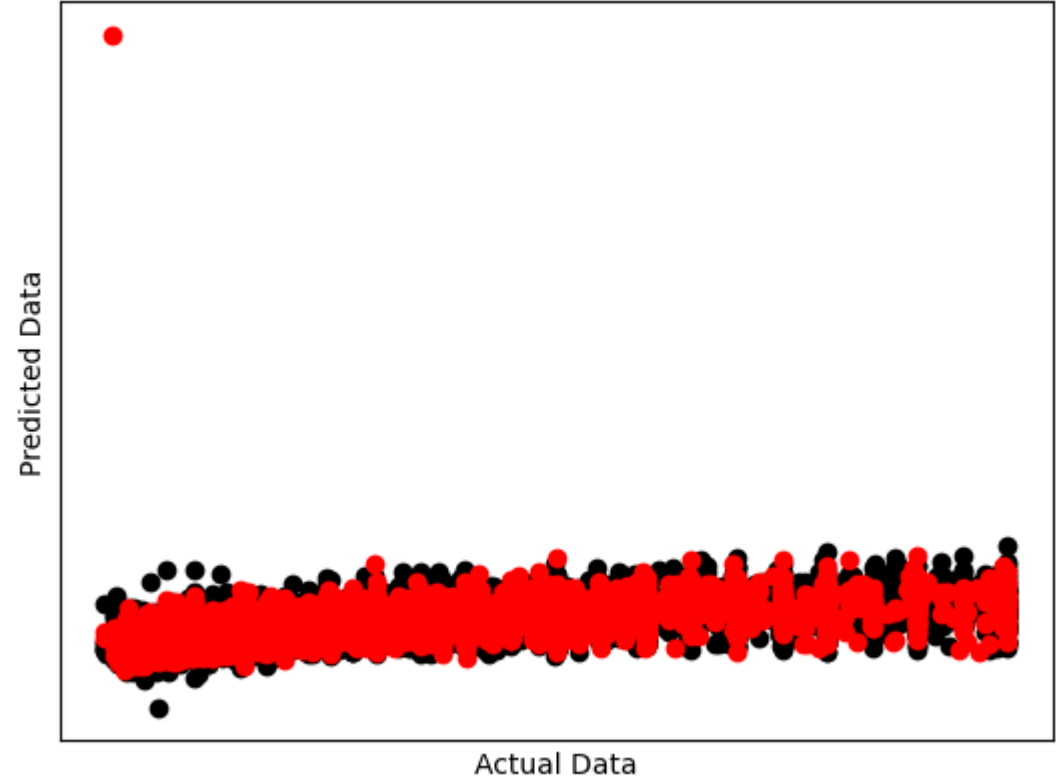
Histogram of Actual Prices vs. Predicted Prices (Validation Data)



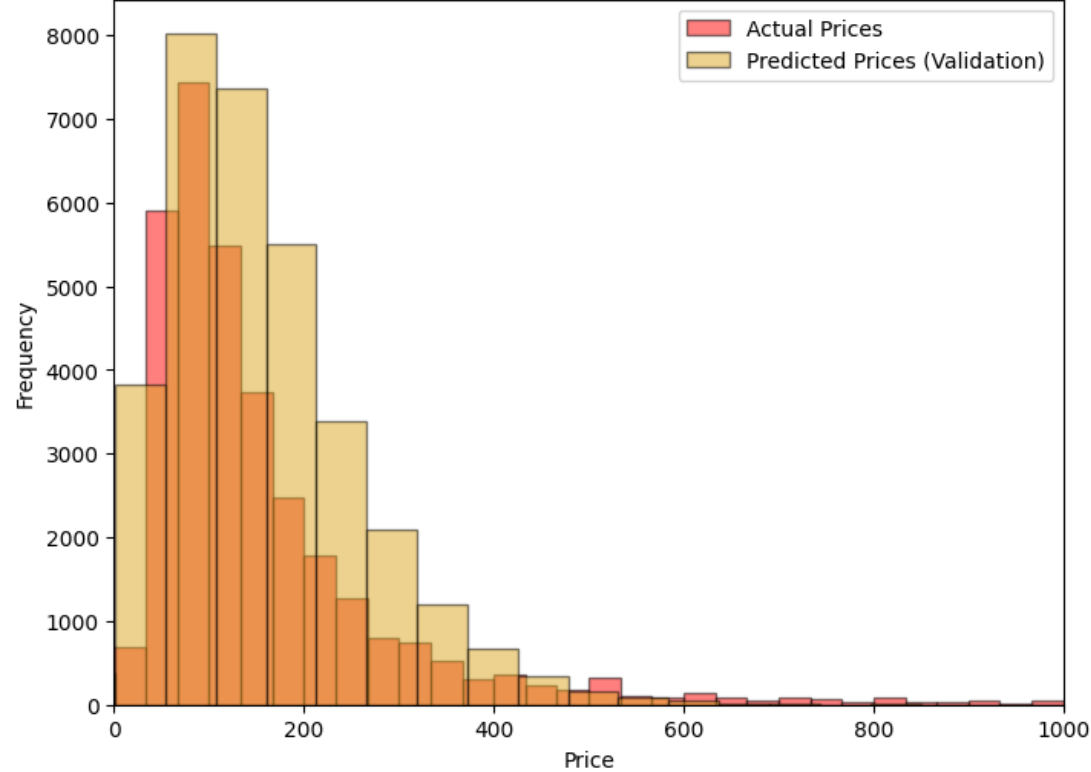
REGRESSION ANALYSIS

Training data -Mean squared error: 8163.59
Validation data - Mean squared error: 8762.34
Training data R2 Score: 0.57
Validation data R2 Score: 0.52
Adjusted R2 for Training data: 0.56
Adjusted R2 for Validation data: 0.51

Dot Chart for Ridge Regression



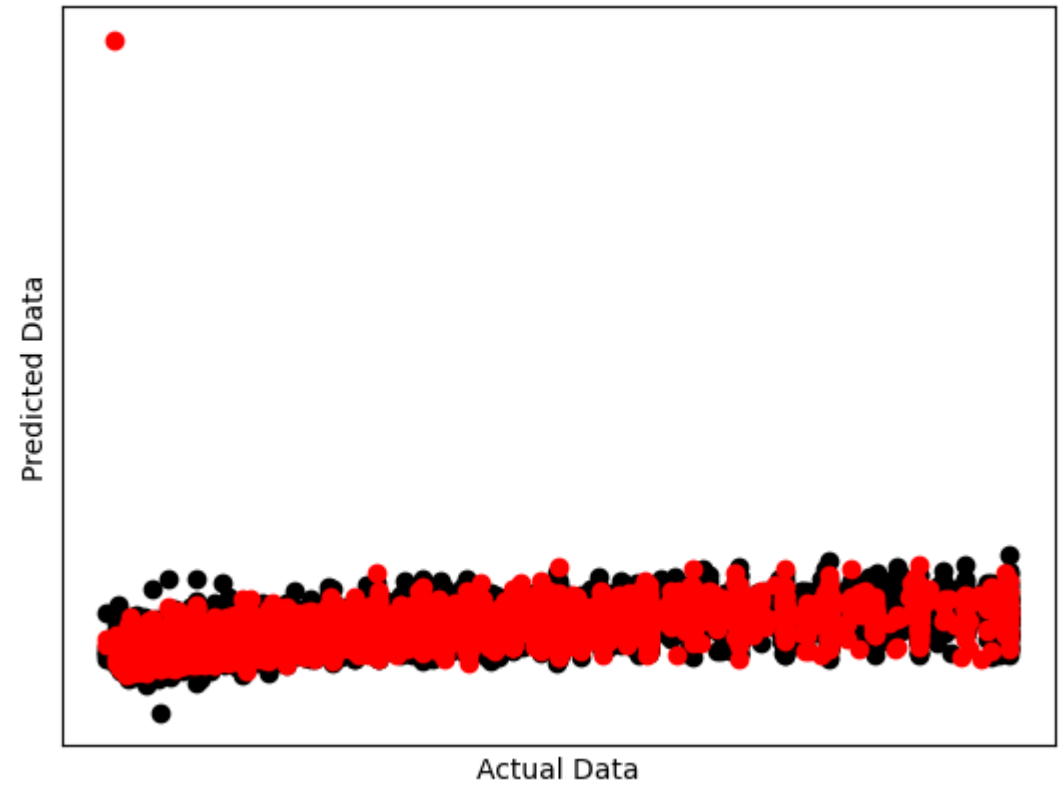
Histogram of Actual Prices vs. Predicted Prices (Validation Data)



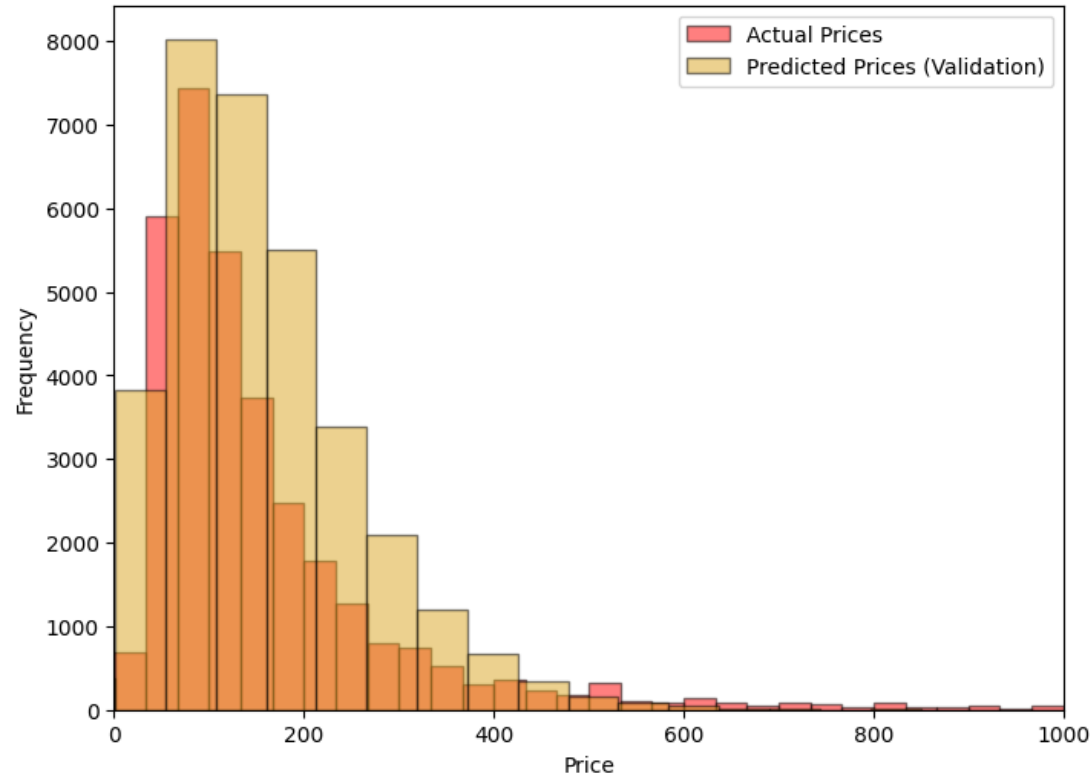
REGRESSION ANALYSIS

Training data -Mean squared error: 8147.30
Validation data - Mean squared error: 8812.41
Training data R2 Score: 0.57
Validation data R2 Score: 0.52
Adjusted R2 for Training data: 0.56
Adjusted R2 for Validation data: 0.50

Dot Chart for Lasso Regression



Histogram of Actual Prices vs. Predicted Prices (Validation Data)



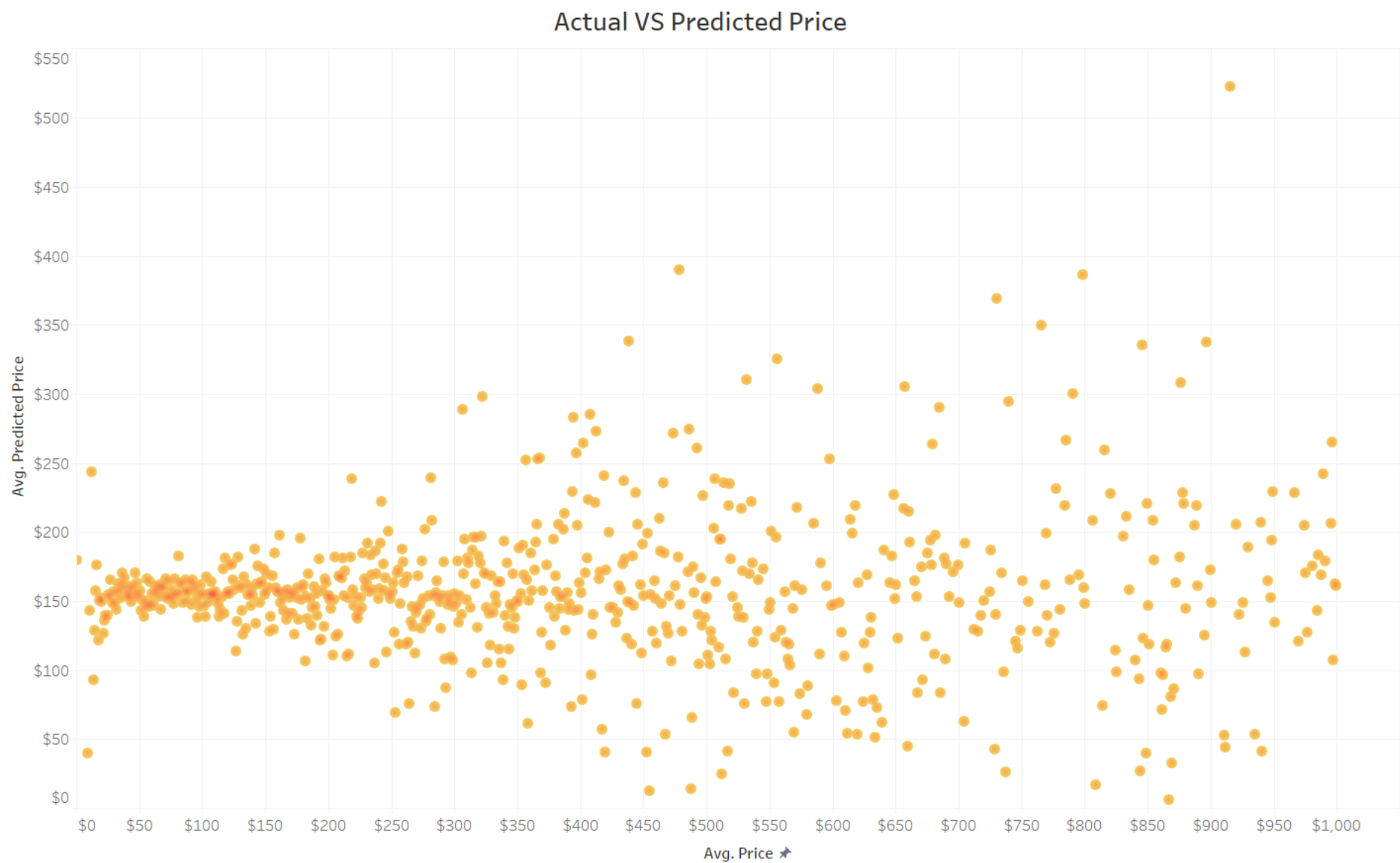
REGRESSION ANALYSIS

Models	MSE (Training)	MSE (Validation)	R2 (Training)	R2 (Validation)	Adj. R2 (Training)	Adj. R2 (Validation)
Linear Regression	8183.14	8869.18	0.57	0.52	0.56	0.50
Ridge Regression	8163.59	8762.34	0.57	0.52	0.56	0.51
Lasso Regression	8147.30	8812.41	0.57	0.52	0.56	0.50

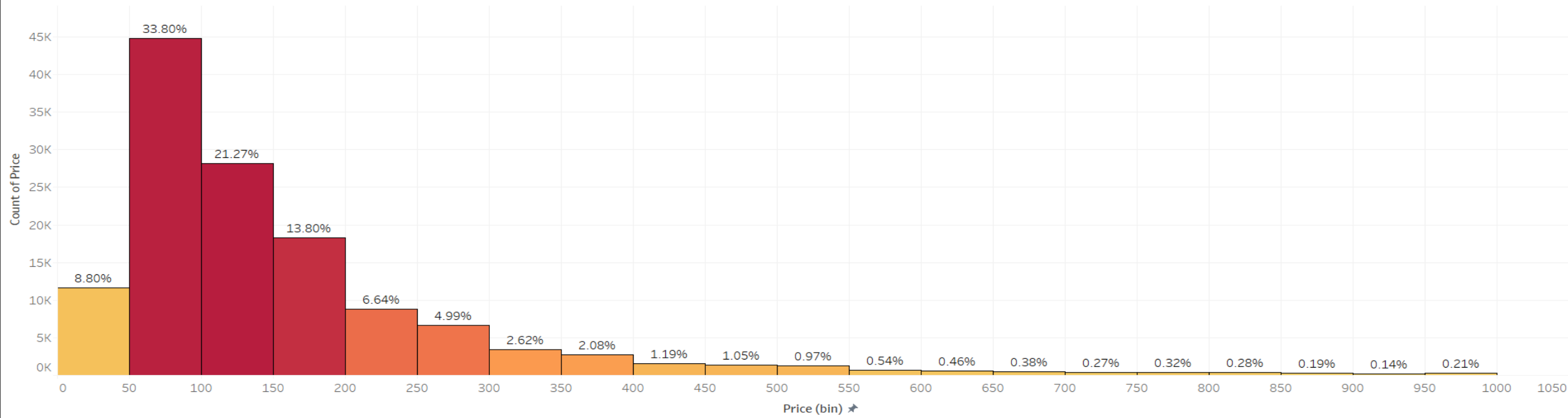
Ridge regression is the best model, so proceeding with it for our price prediction

- Deals with Multicollinearity
- Flexible in handling large features
- Prevents Overfitting
- It gives a consistent output.

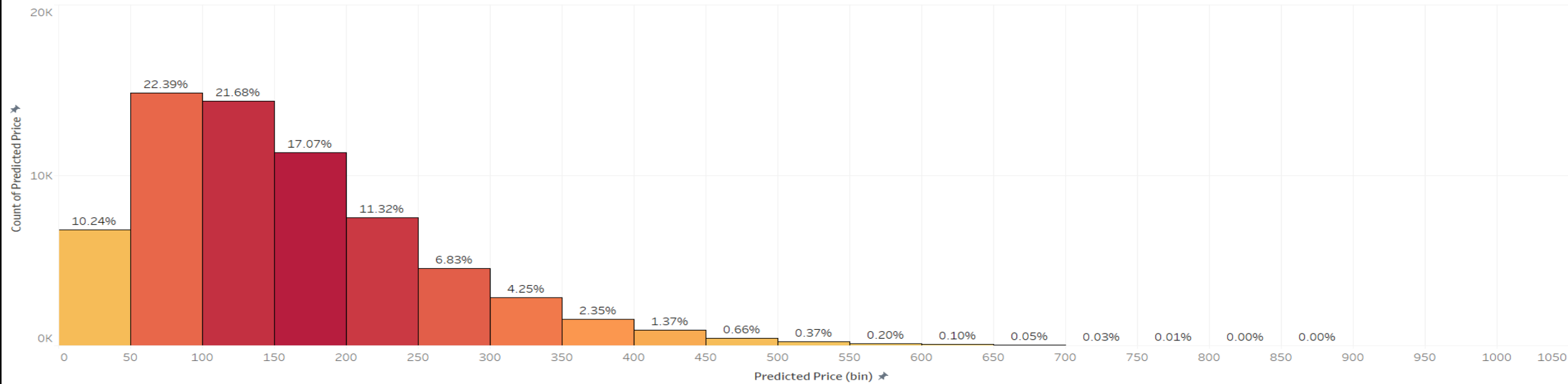
COMPARING ACTUAL PRICE WITH PREDICTED PRICE



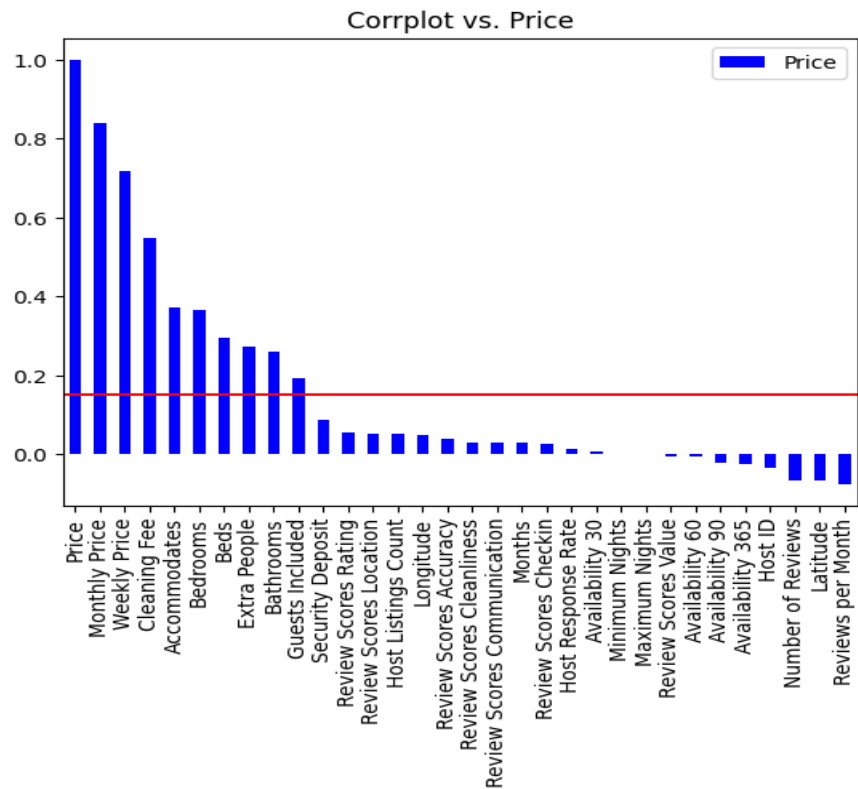
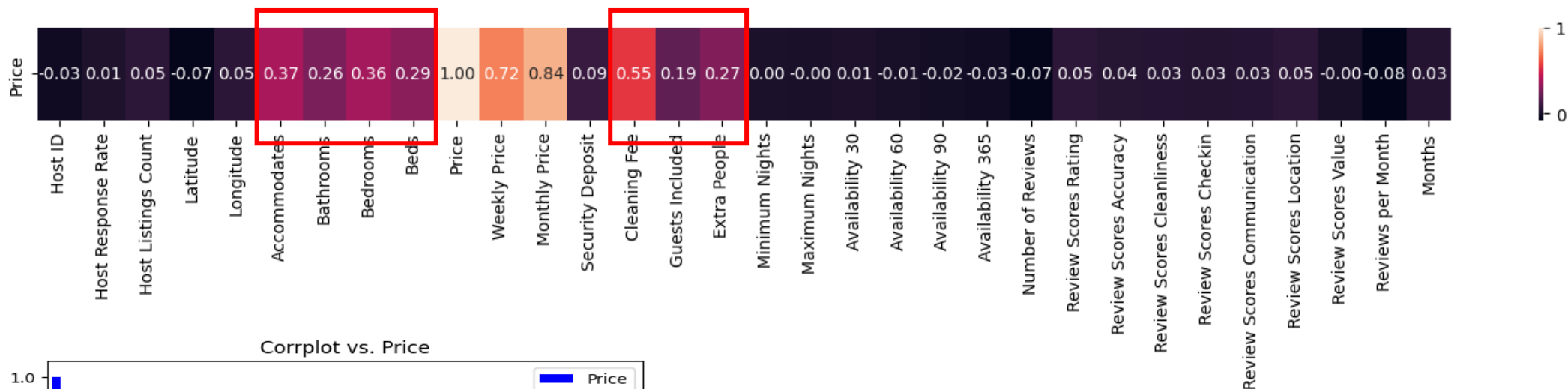
Actual Price



Predicted Price



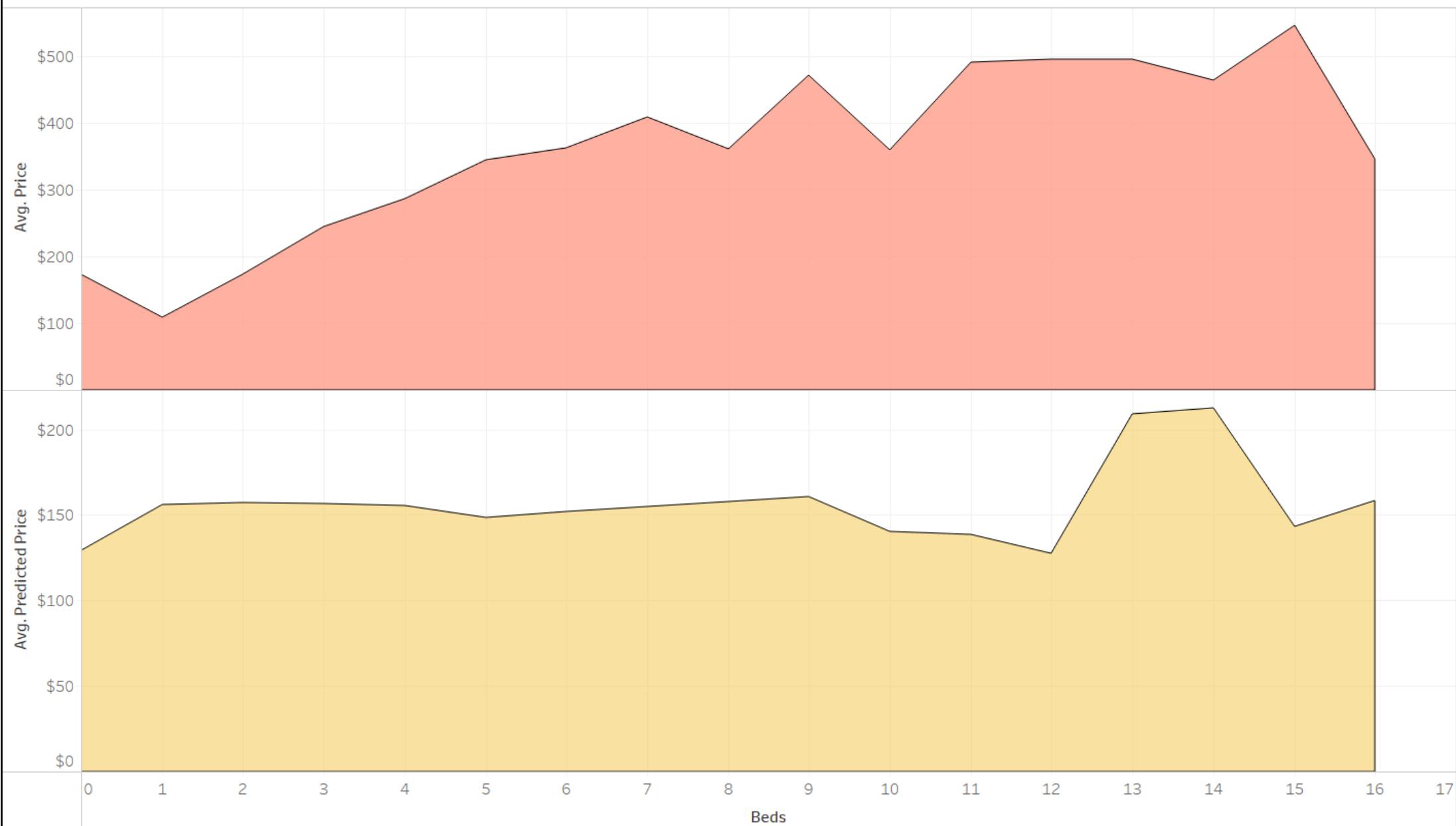
CHECKING CO-RELATION



Top Variables

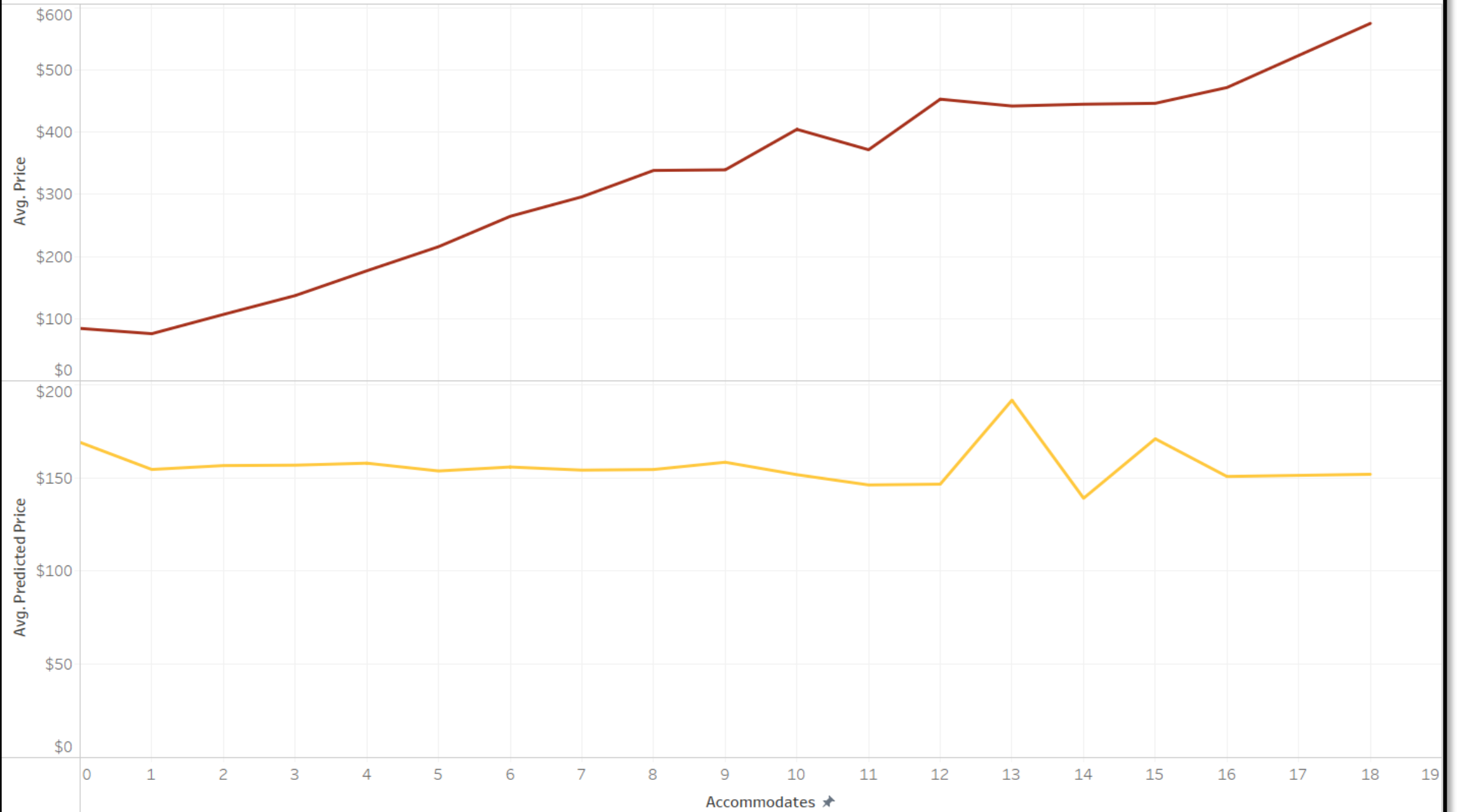
- Accommodates
- Bathrooms
- Bedrooms
- Beds
- Extra People
- Cleaning Fee

Beds: Actual VS Predicted Price



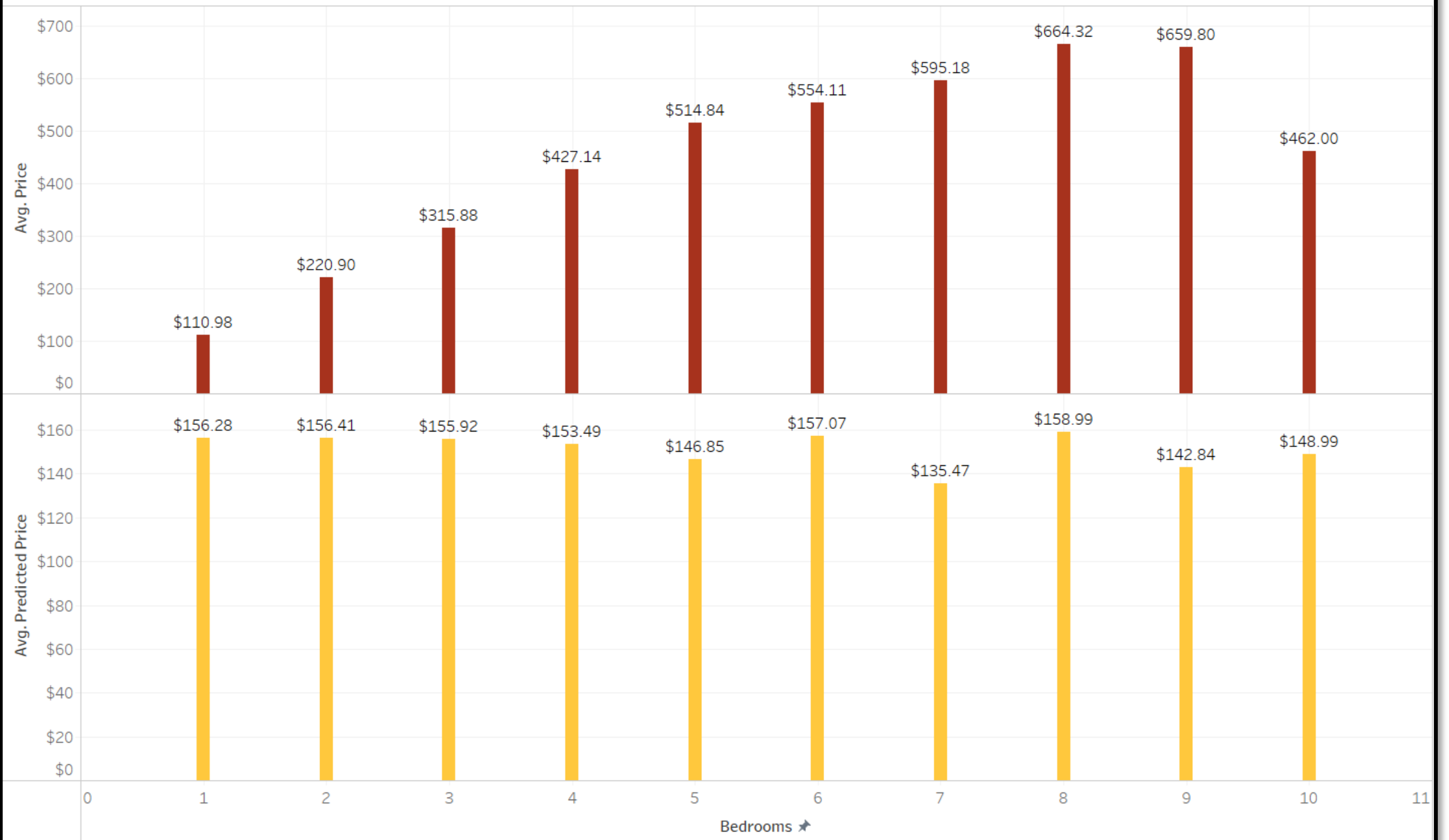
The plots of average of Price and average of Predicted Price for Beds. For pane Average of Predicted Price: Color shows average of Predicted Price. For pane Average of Price: Color shows average of Price.

Accommodates: Actual VS Predicted Price

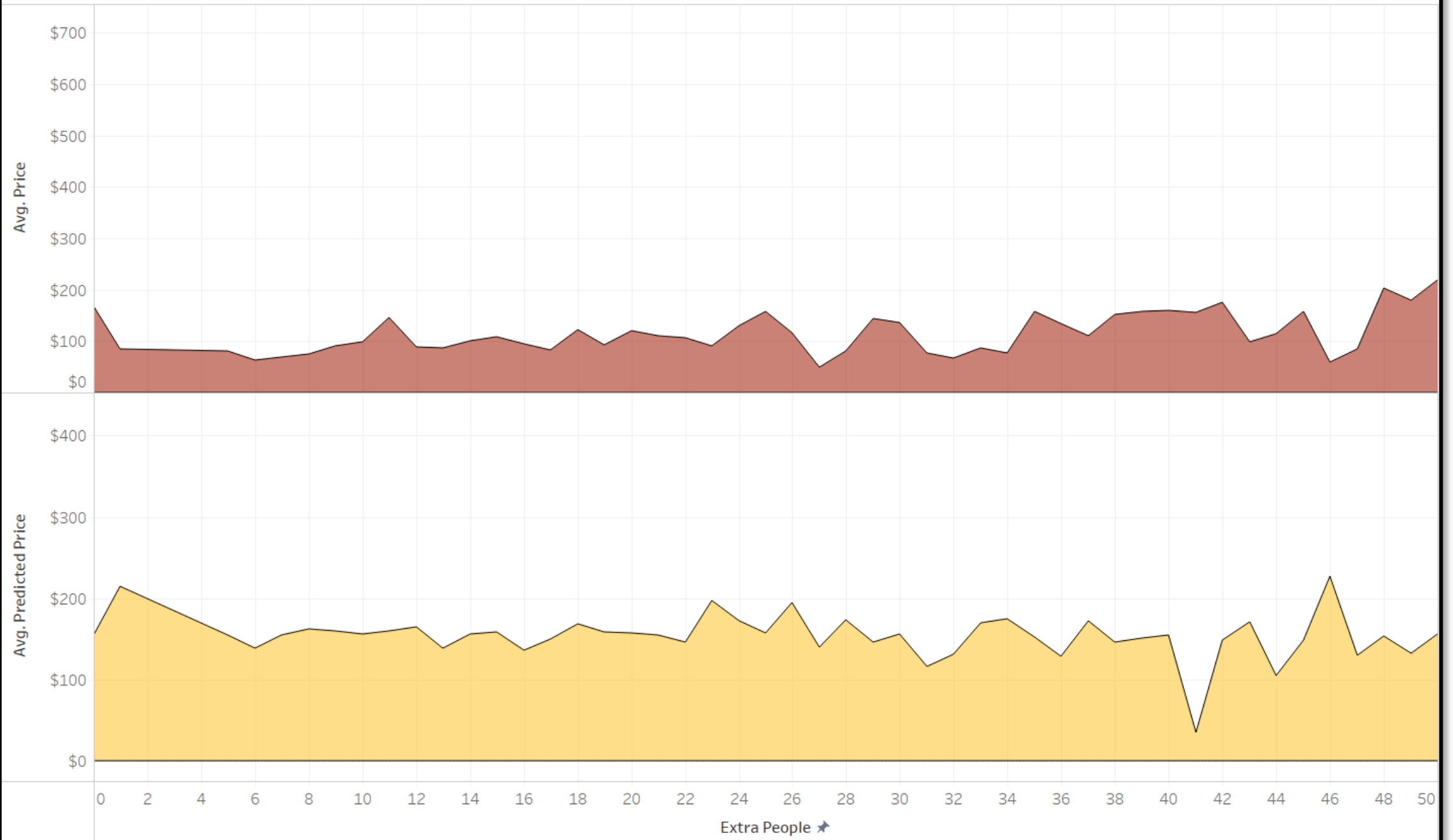


The trends of average of Price and average of Predicted Price for Accommodates. The data is filtered on Accommodates, which excludes 192.

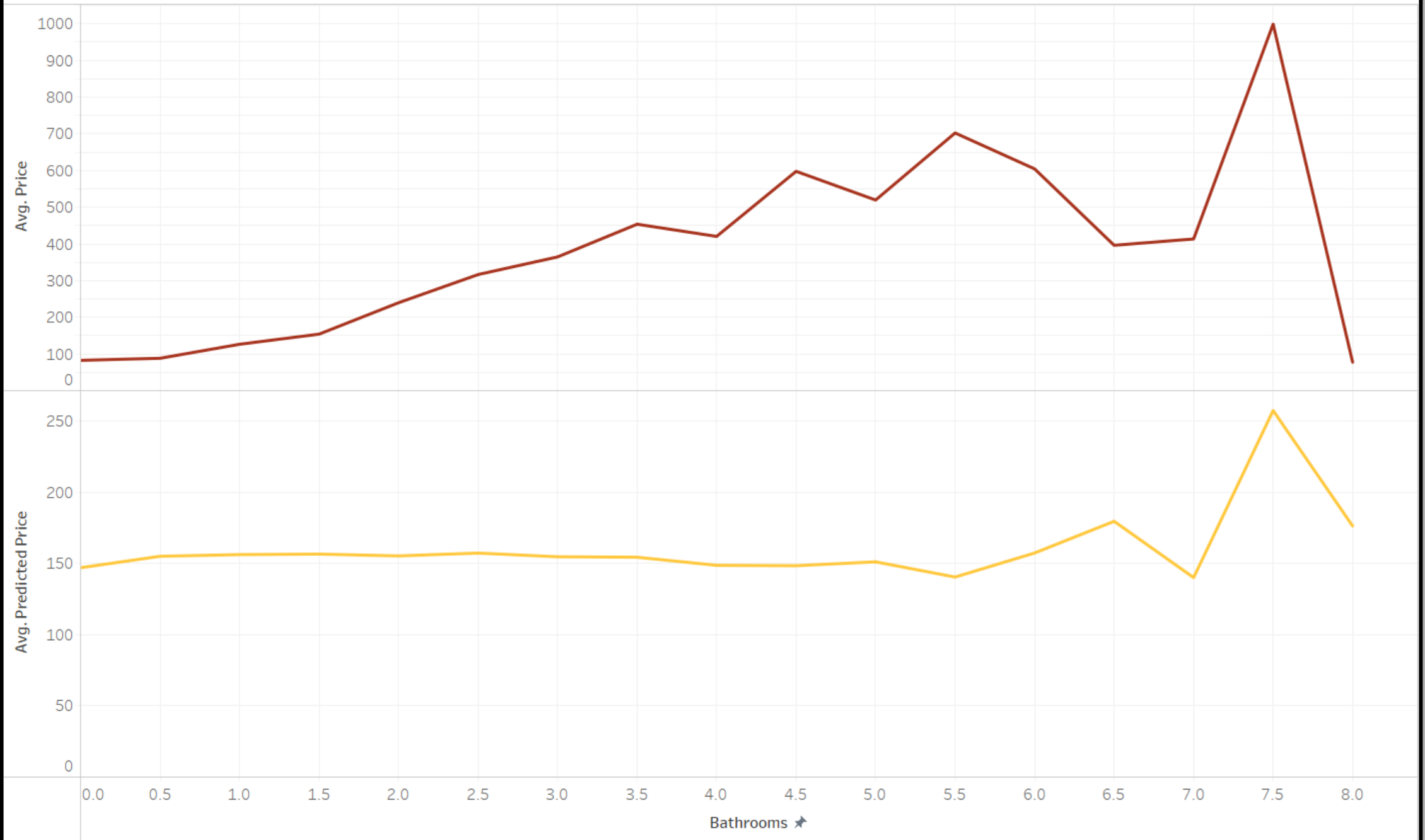
Bedrooms: Actual VS Predicted Price



Extra People: Actual VS Predicted Price



Bathrooms: Actual VS Predicted Price





THANK YOU