Analyzing Seattle Airbnb Dataset for Price Prediction

KIRTI CHAUDHARI

Goals

Analyze	Problem Statement: Analyze various trends in Airbnb bookings in Seatttle and identify factors affecting the price of property based on various characteristics
Mine	Mine dataset to understand busiest times of the year, how prices spike, general trend of Airbnb listings and Airbnb visitors etc.
Apply	Apply Machine learning on the dataset to predict rental pricings or factors which affect listing price

Data Acquisition

- Source : http://insideairbnb.com/get-the-data.html
- Listing.csv: Listings, including full descriptions and average review score.
- *Reviews.csv*: Reviews, including unique id for each reviewer and detailed comments.
- *Calendar.csv*: Calendar, including listing id and the price and availability for that day.

Seattle, Washington, United States

See Seattle data visually here.

Date Compiled	Country/City	File Name	Description
17 June, 2020	Seattle	listings.csv.gz	Detailed Listings data for Seattle
17 June, 2020	Seattle	calendar.csv.gz	Detailed Calendar Data for listings in Seattle
17 June, 2020	Seattle	reviews.csv.gz	Detailed Review Data for listings in Seattle
17 June, 2020	Seattle	listings.csv	Summary information and metrics for listings in Seattle (good for visualisations).
17 June, 2020	Seattle	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
N/A	Seattle	neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
N/A	Seattle	neighbourhoods.geojson	GeoJSON file of neighbourhoods of the city.

show archived data

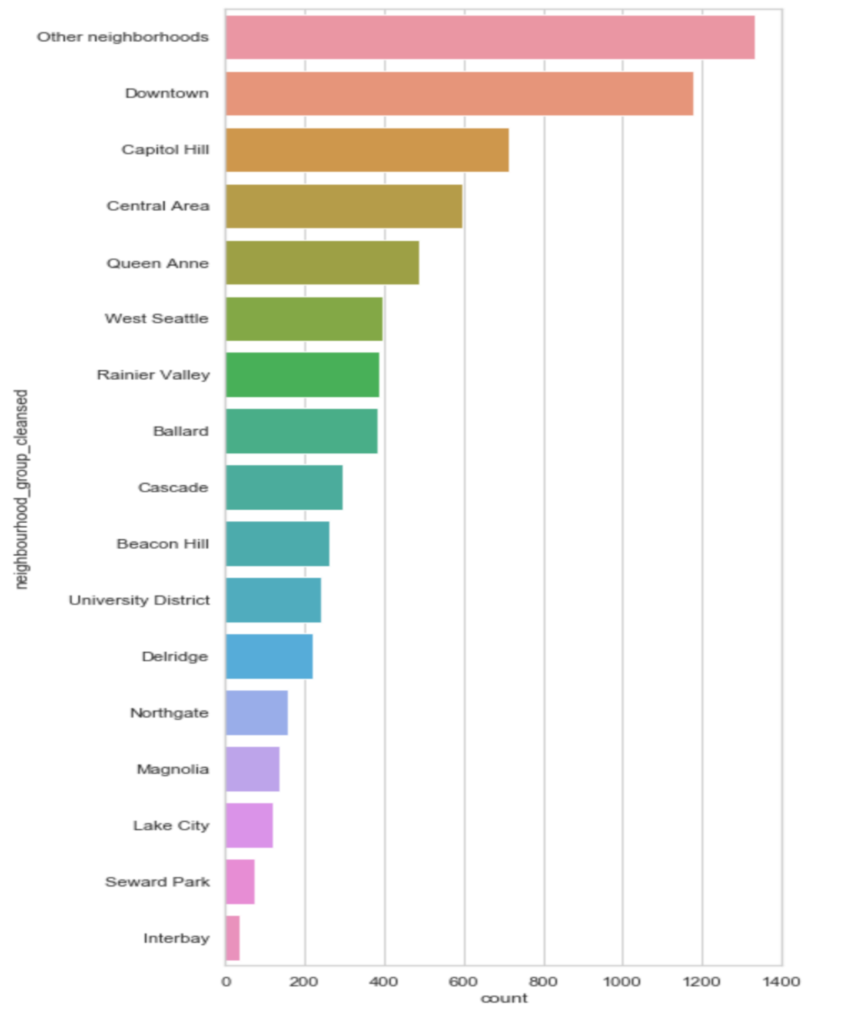
Data Preparation and Cleaning

- Check empty/null values
- Price columns
  - remove chars such as ‘$’, ‘,’
  - Replace nan with 0
- Correct Datatypes for columns
  - Price columns – string to int
  - Date columns – change datatype to datetime

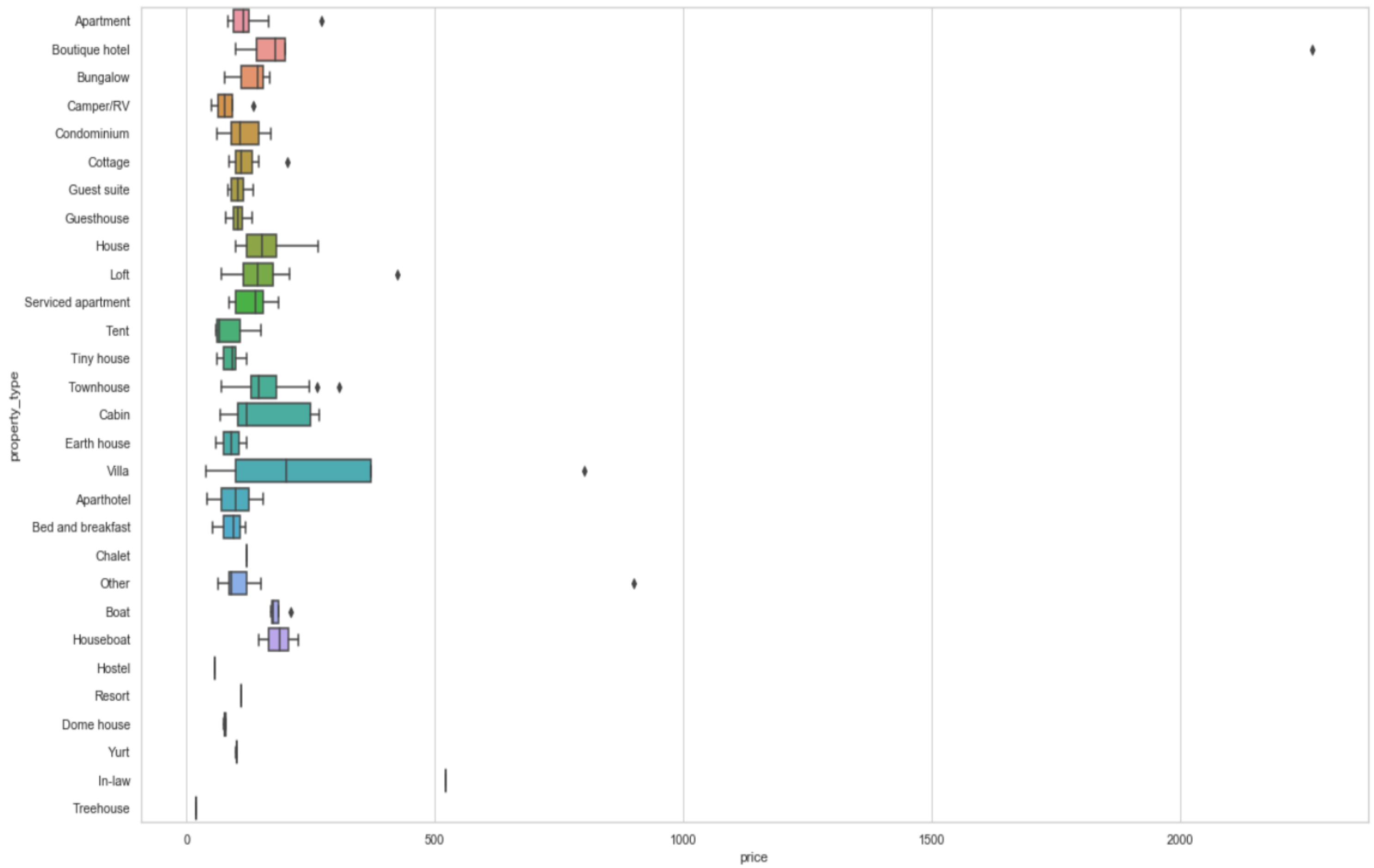
Data Wrangling

- Replace values in Columns
  - availability – replace t and f with 1 and 0.
- Create New columns –
  - Availability – no of days available.
- Join Datasets – listings and calendars

Data Visualization

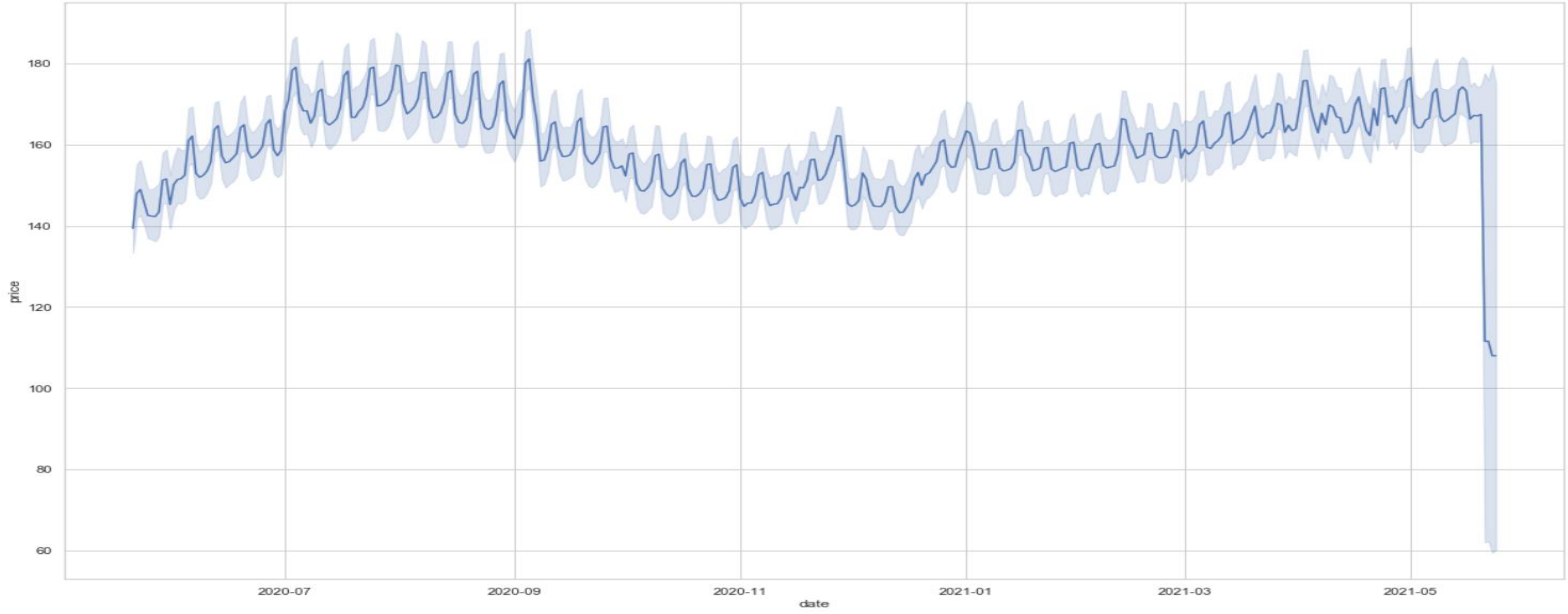


Number of Properties per Neighborhood



Price distribution Per Property Type

Data Visualization

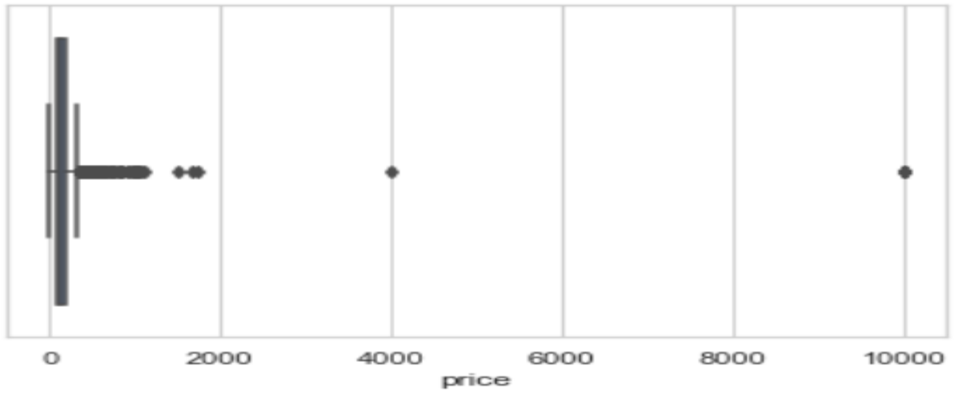


Price Trend per Calendar Date

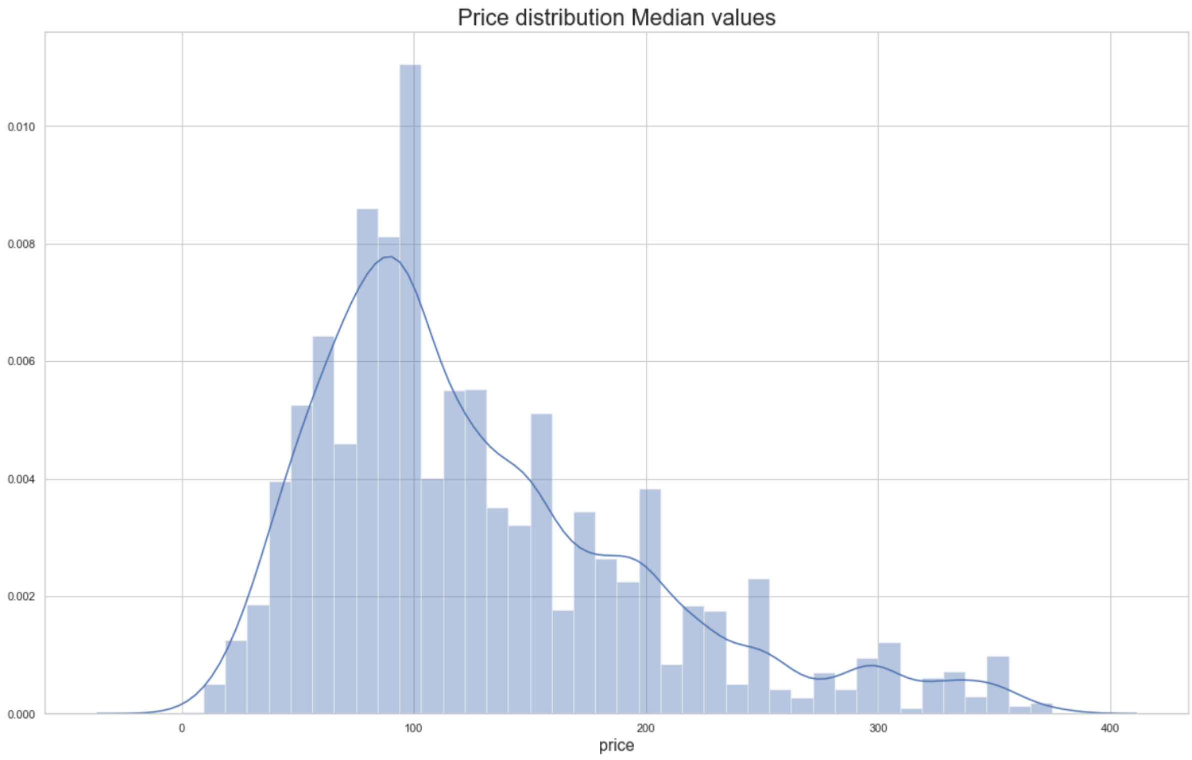
Data Analysis

- Primary Goal : Determine if there is any strong correlation between pairs of independent variables or between an independent and dependent variables (price)
- Remove outliers from price distribution
  - Calculate First and third quartile ranges of dataset
  - Find lower and upper bounds of dataset – 2 times IQR (Inter Quartile Range)
  - Anything above upper bounds can be removed from dataset

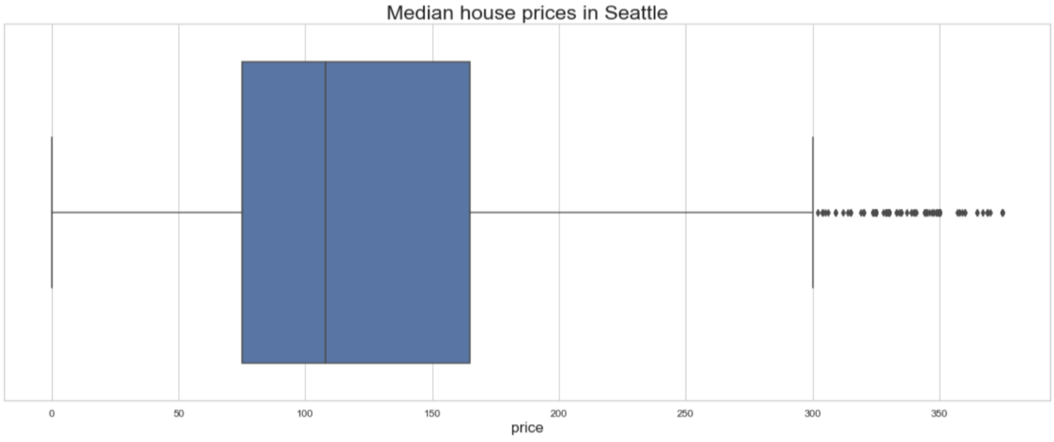
Data Analysis



Box Plot- Before removing Outliers

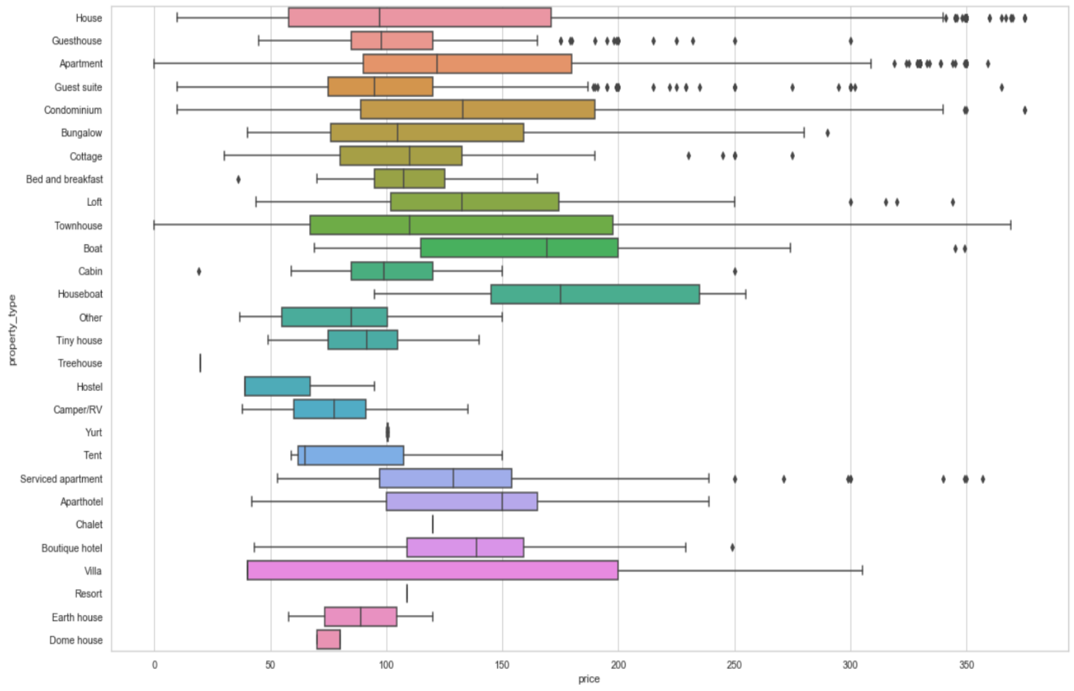


Price distribution Median Values

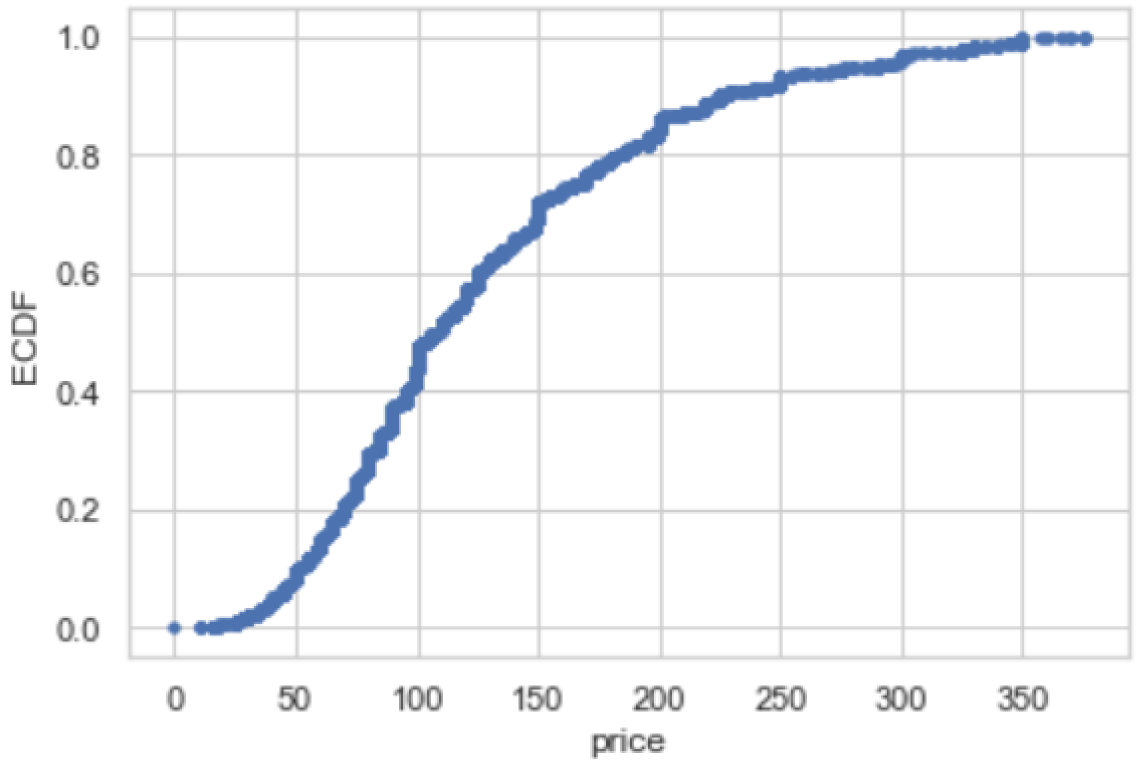


Box Plot - After removing Outliers

Data Visualization

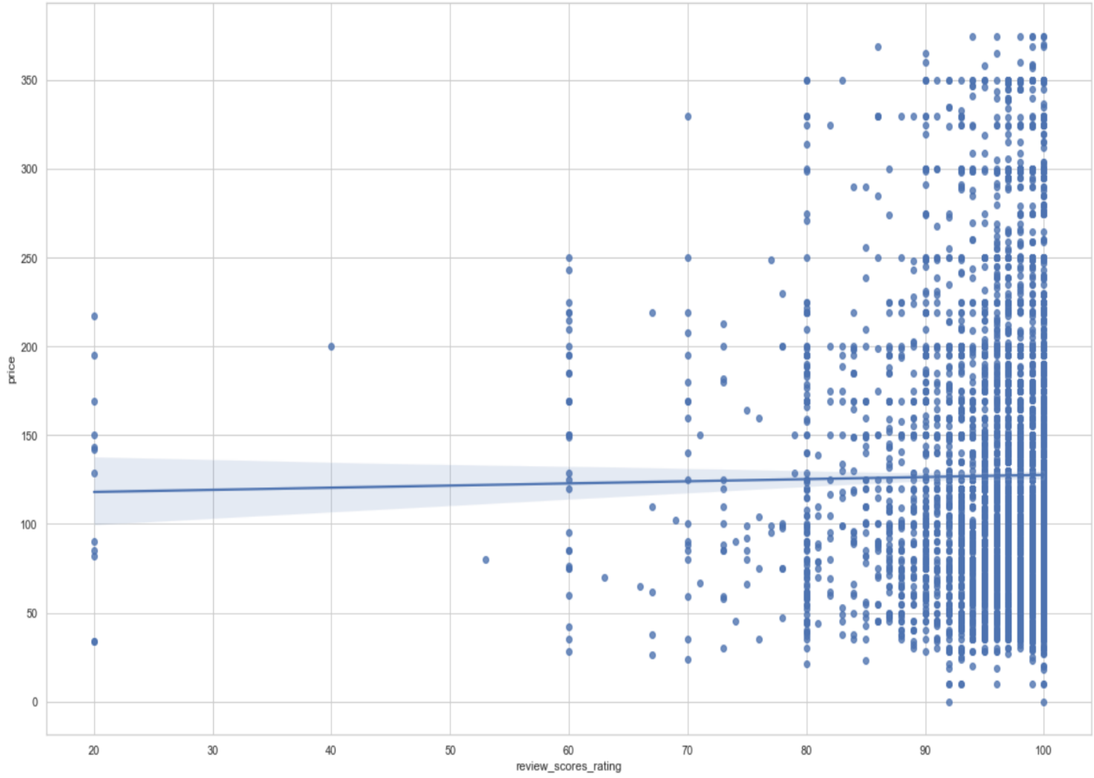


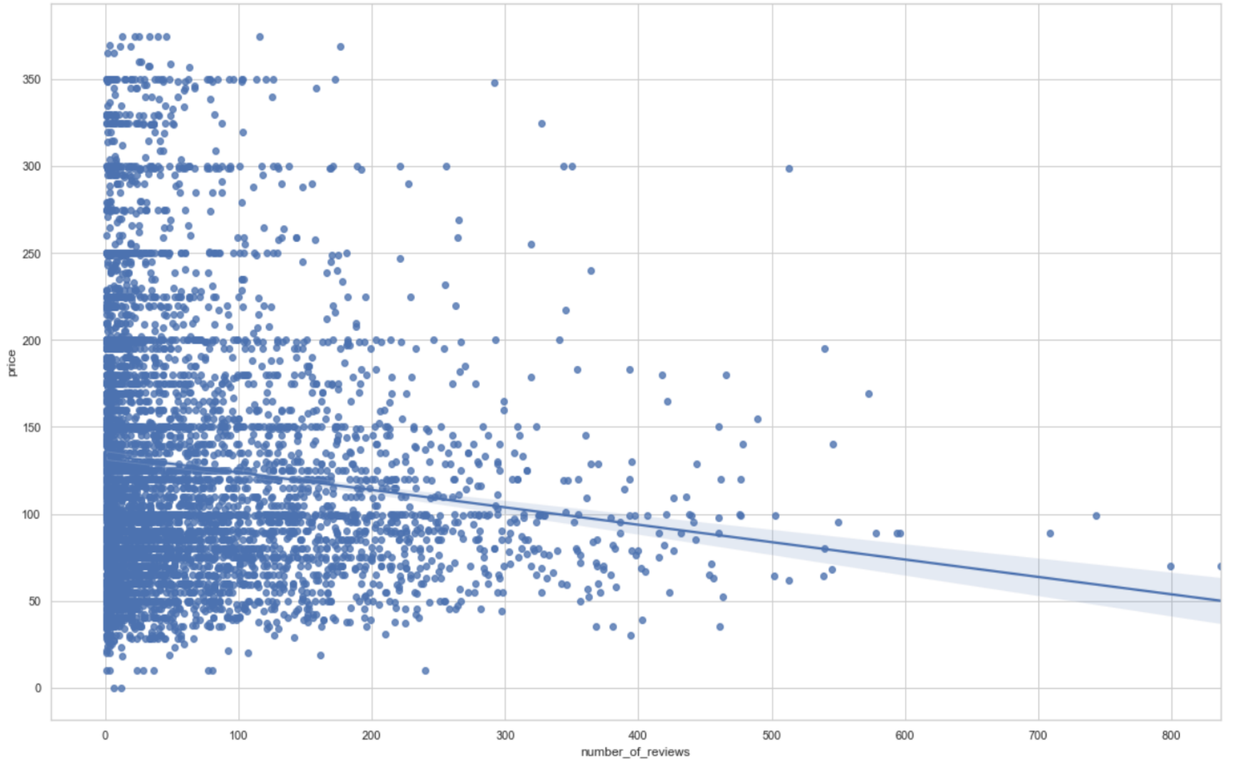
Price distribution Per Property type



ECDF plot for price column

Data Visualization



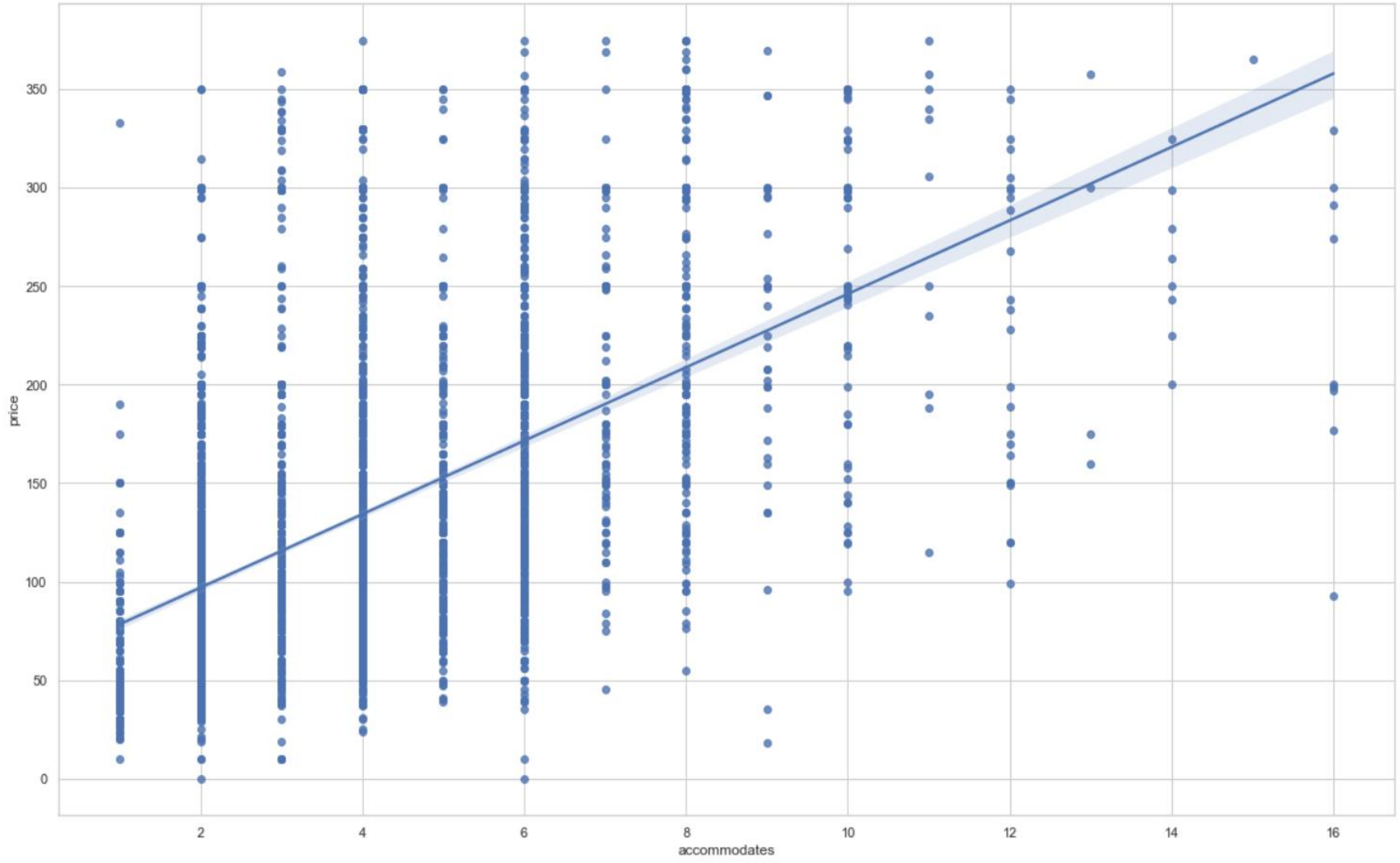


Review Score Rating vs Price

Number of Review vs Price

Data Visualization

- Strong linear correlation between Number of accommodates and Price
- Listing price directly affected by no of accommodates allowed in the property



Number of Accommodates vs Price

Data Modelling

- Select Specific columns in dataset to reduce dimensionality
- Drop Nan rows
- Perform one-hot encoding for categorical columns
- Split dataset into – 70% training and 30 % testing
- Random Forest Regressor to train and fit the model

```
#train RF regressor model
forest = RandomForestRegressor(n_estimators=100,
                               criterion='mse',
                               random_state=42,
                               n_jobs=-1,verbose=1)
forest.fit(X_train, y_train.squeeze())
```

```
[Parallel(n_jobs=-1)]: Using backend ThreadingBackend with 16 concurrent workers.
[Parallel(n_jobs=-1)]: Done  18 tasks      | elapsed:    0.2s
[Parallel(n_jobs=-1)]: Done 100 out of 100 | elapsed:    0.6s finished
```

```
RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                      max_depth=None, max_features='auto', max_leaf_nodes=None,
                      max_samples=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      n_estimators=100, n_jobs=-1, oob_score=False,
                      random_state=42, verbose=1, warm_start=False)
```

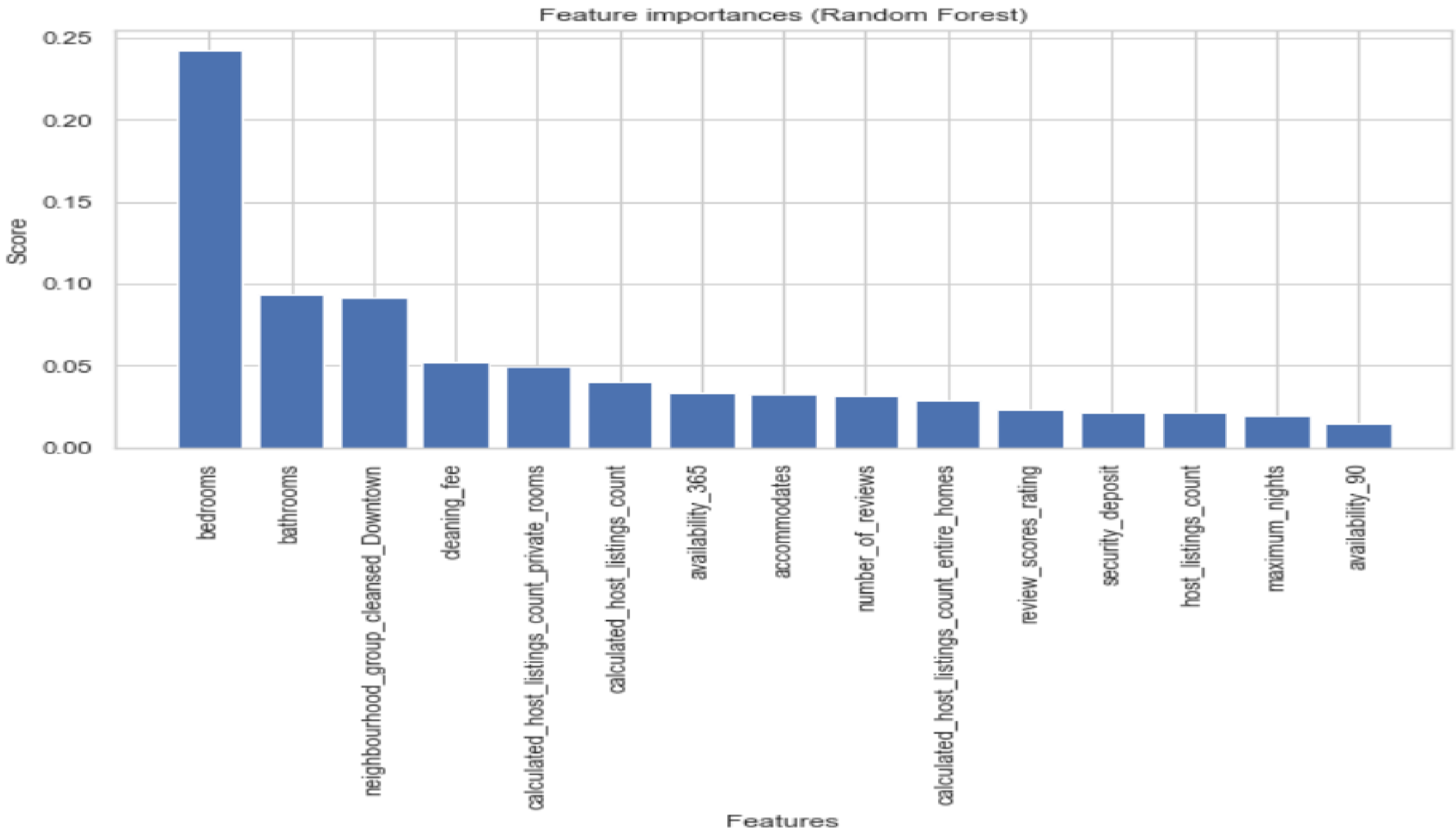
```
#calculate scores for the model
y_train_preds = forest.predict(X_train)
y_test_preds = forest.predict(X_test)

print(f'Random Forest MSE train: %.3f, test: %.3f' % (
        mean_squared_error(y_train, y_train_preds),
        mean_squared_error(y_test, y_test_preds)))
print('Random Forest R^2 train: %.3f, test: %.3f' % (
        r2_score(y_train, y_train_preds),
        r2_score(y_test, y_test_preds)))
```

```
[Parallel(n_jobs=16)]: Using backend ThreadingBackend with 16 concurrent workers.
[Parallel(n_jobs=16)]: Done  18 tasks      | elapsed:    0.0s
[Parallel(n_jobs=16)]: Done 100 out of 100 | elapsed:    0.0s finished
[Parallel(n_jobs=16)]: Using backend ThreadingBackend with 16 concurrent workers.
[Parallel(n_jobs=16)]: Done  18 tasks      | elapsed:    0.0s
[Parallel(n_jobs=16)]: Done 100 out of 100 | elapsed:    0.0s finished
```

```
Random Forest MSE train: 265.215, test: 1879.684
Random Forest R^2 train: 0.949, test: 0.632
```

Feature Importance Matrix



Conclusion

- Data acquired from Airbnb for Seattle listings
- Data cleaning, wrangling steps were performed on the dataset
- Removed outliers in target variable (price) and visualized data in various properties.
- Trained and fit the model using Random Forest regressor.
- From the feature importance matrix- no of bedroom was seen as important feature affecting the prices.
- From the charts, we also saw strong correlation between no of accommodates against the price.

References

- Project Github:

https://github.com/kirti-chaudhari/SpringBoard_DataScience_Career

- Jupyter Notebook :

https://github.com/kirti-chaudhari/SpringBoard_DataScience_Career/blob/master/CapstoneProject/AirbnbProject.ipynb