AML Project
AirBnB New York City

SUBMITTED BY: RADHIKA RAJEEVAN & SUNIT NAIR

INSTRUCTOR: DR. FARID ALIZADEH

Project Scope and Objective

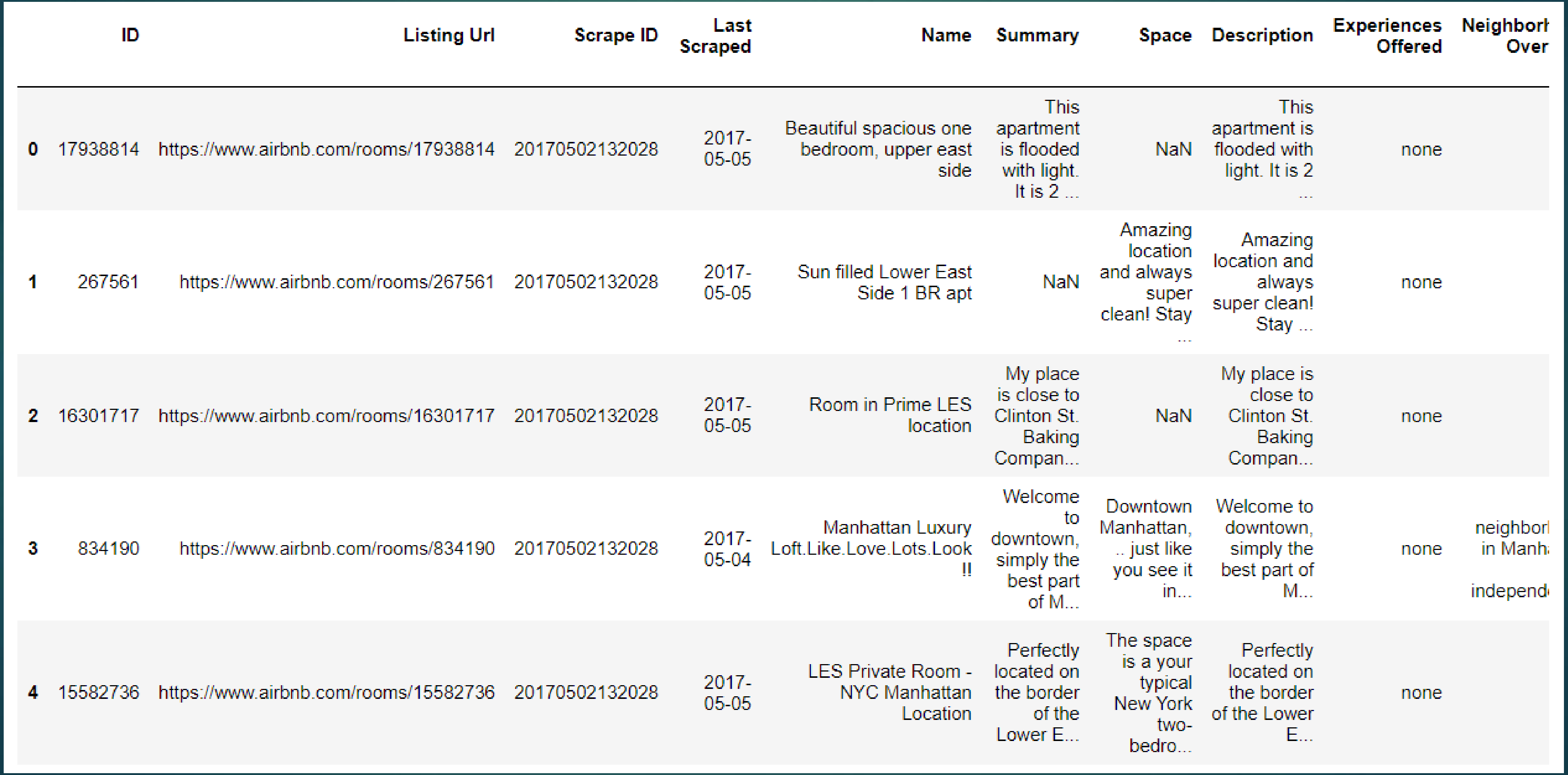
- Analyze AirBnB data to extract possible/relevant features.
- Use techniques learned as part of course and additional methods to create regression models to predict price of listing.
- Clean and extract features from original data file (Python).
- Analyze and plot relevant graphs to understand data (Python).
- Load clean data set (CSV) to database (MySQL).
- Read data in R through database connection (MySQL).
- Derive new features from existing features.
- Run regression models to predict price against relevant features.

The Dataset

- Original data set: 494,954 records
- USA data set: 134,545
- New York data set: 19,528
- Columns in data set: 89
- Data set after cleaning: 19,273
- Columns after cleaning and extraction: 130
- Column types
  - Identification: 1
  - Numerical: 23
  - Categorical: 106

Python: Data Cleaning

Preview of Raw data : 19528 observations and 89 columns



	ID	Listing Url	Scrape ID	Last Scraped	Name	Summary	Space	Description	Experiences Offered	Neighborl Over
0	17938814	https://www.airbnb.com/rooms/17938814	20170502132028	2017-05-05	Beautiful spacious one bedroom, upper east side	This apartment is flooded with light. It is 2 ...	NaN	This apartment is flooded with light. It is 2 ...	none	
1	267561	https://www.airbnb.com/rooms/267561	20170502132028	2017-05-05	Sun filled Lower East Side 1 BR apt	NaN	Amazing location and always super clean! Stay ...	Amazing location and always super clean! Stay ...	none	
2	16301717	https://www.airbnb.com/rooms/16301717	20170502132028	2017-05-05	Room in Prime LES location	My place is close to Clinton St. Baking Compan...	NaN	My place is close to Clinton St. Baking Compan...	none	
3	834190	https://www.airbnb.com/rooms/834190	20170502132028	2017-05-04	Manhattan Luxury Loft.Like.Love.Lots.Look !!	Welcome to downtown, simply the best part of M...	Downtown Manhattan, .. just like you see it in...	Welcome to downtown, simply the best part of M...	none	neighborl in Manh independ
4	15582736	https://www.airbnb.com/rooms/15582736	20170502132028	2017-05-05	LES Private Room - NYC Manhattan Location	Perfectly located on the border of the Lower E...	The space is a your typical New York two-bedro...	Perfectly located on the border of the Lower E...	none	

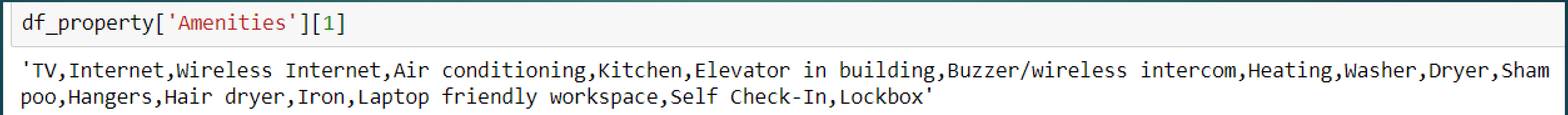
Python: Data Cleaning Tasks

1. Treating null values : Replace with 0 or delete

- Host Response Time : NAs were replaced with 1 hour
- Host Response Time : hours and days were converted to hours.
- Ratings : NAs were replaced with 0
- Neighbourhood : NAs were deleted due to lack of information.

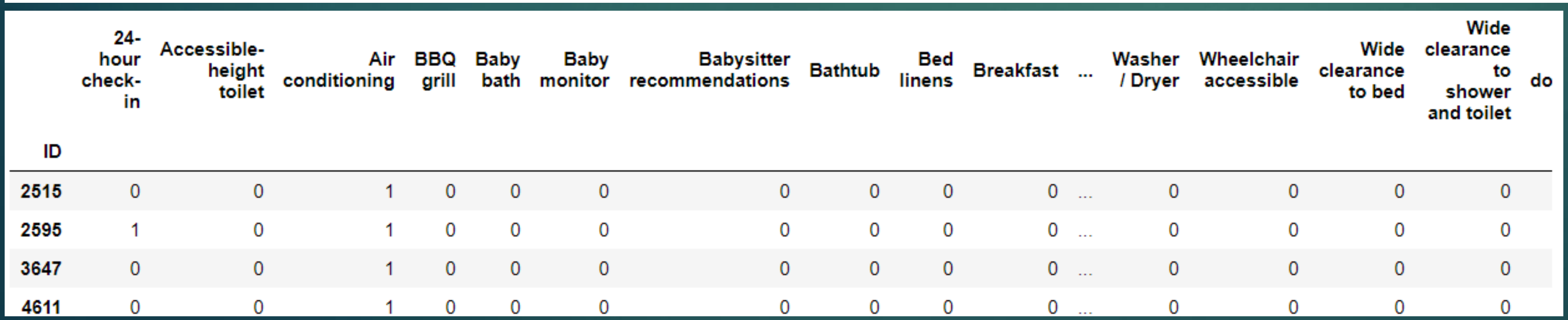
2. Creating Dummy Variables : comma separated format to binary columns

- Amenities
- Review Features



```
df_property['Amenities'][1]
```

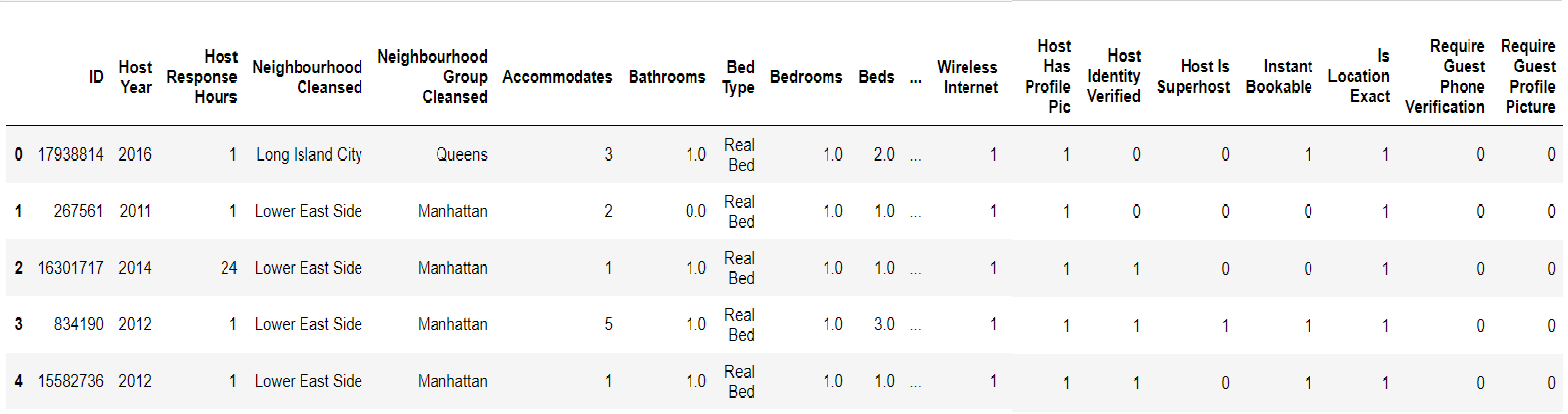
```
'TV,Internet,Wireless Internet,Air conditioning,Kitchen,Elevator in building,Buzzer/wireless intercom,Heating,Washer,Dryer,Sham
poo,Hangers,Hair dryer,Iron,Laptop friendly workspace,Self Check-In,Lockbox'
```



	24-hour check-in	Accessible-height toilet	Air conditioning	BBQ grill	Baby bath	Baby monitor	Babysitter recommendations	Bathtub	Bed linens	Breakfast	...	Washer / Dryer	Wheelchair accessible	Wide clearance to bed	Wide clearance to shower and toilet	do
ID																
2515	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	
2595	1	0	1	0	0	0	0	0	0	0	...	0	0	0	0	
3647	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	
4611	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	

Python: Data Cleaning Tasks

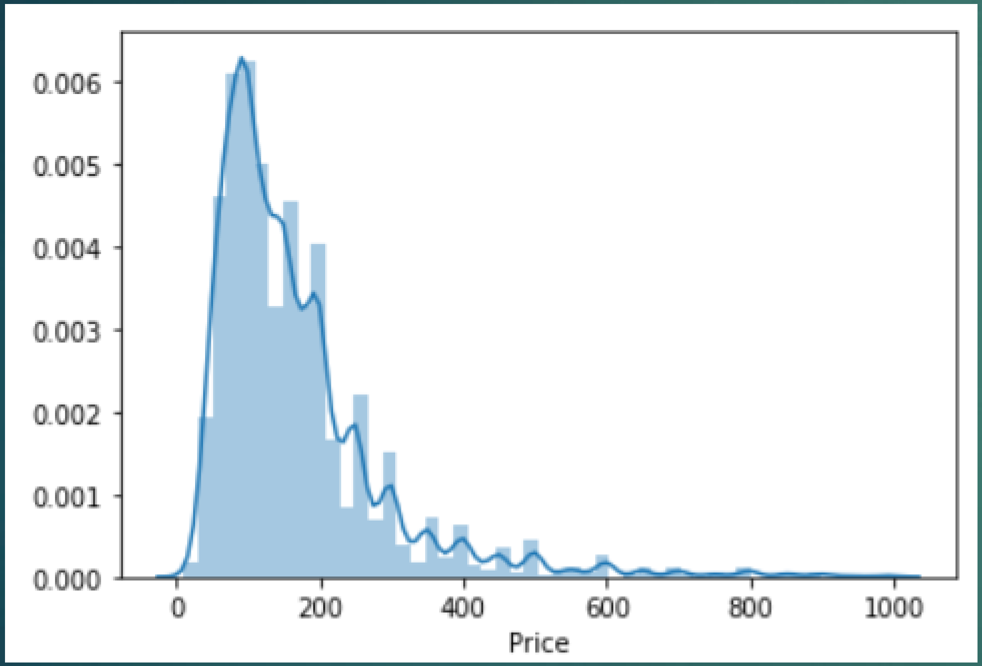
Preview of Cleaned Data : 19273 observations and 130 columns

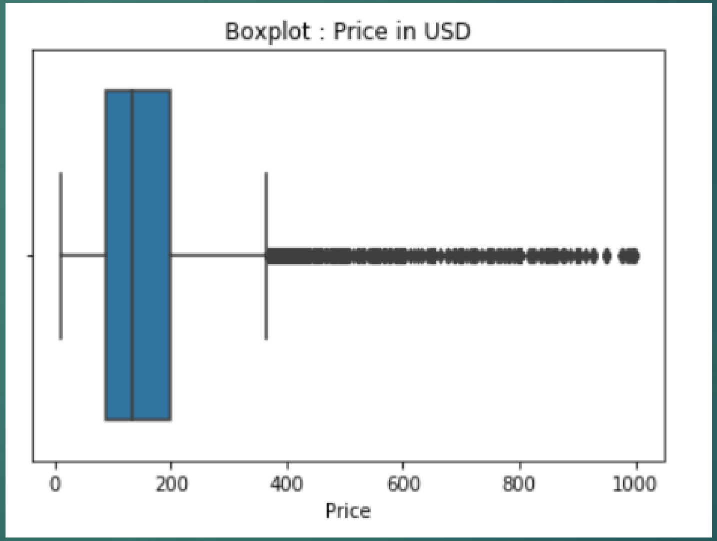


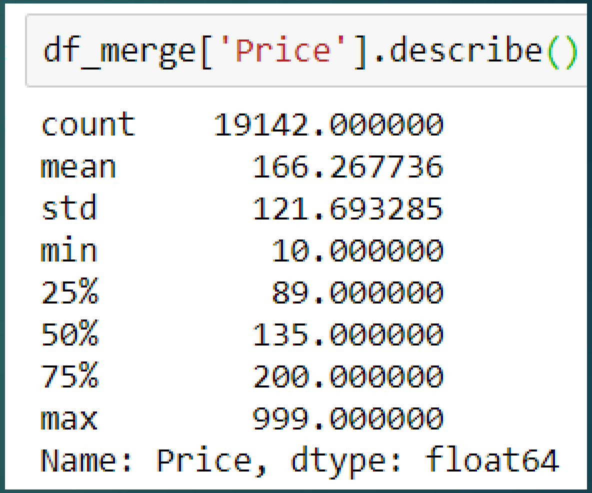
	ID	Host Year	Host Response Hours	Neighbourhood Cleansed	Neighbourhood Group Cleansed	Accommodates	Bathrooms	Bed Type	Bedrooms	Beds	...	Wireless Internet	Host Has Profile Pic	Host Identity Verified	Host Is Superhost	Instant Bookable	Is Location Exact	Require Guest Phone Verification	Require Guest Profile Picture
0	17938814	2016	1	Long Island City	Queens	3	1.0	Real Bed	1.0	2.0	...	1	1	0	0	1	1	0	0
1	267561	2011	1	Lower East Side	Manhattan	2	0.0	Real Bed	1.0	1.0	...	1	1	0	0	0	1	0	0
2	16301717	2014	24	Lower East Side	Manhattan	1	1.0	Real Bed	1.0	1.0	...	1	1	1	0	0	1	0	0
3	834190	2012	1	Lower East Side	Manhattan	5	1.0	Real Bed	1.0	3.0	...	1	1	1	1	1	1	0	0
4	15582736	2012	1	Lower East Side	Manhattan	1	1.0	Real Bed	1.0	1.0	...	1	1	1	0	1	1	0	0

Python: Data Analysis

Price : Univariate Distribution




Boxplot : Price in USD



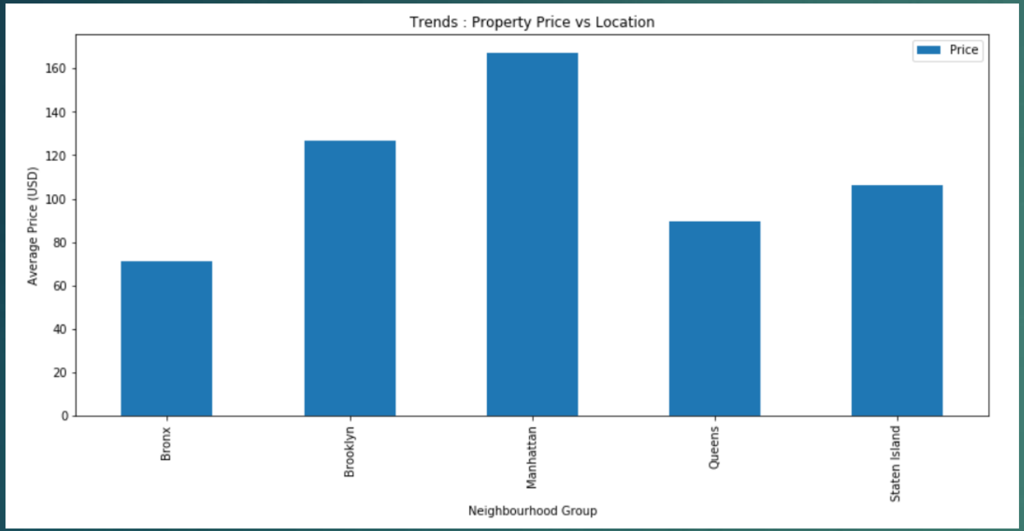
```
df_merge['Price'].describe()
```

```
count    19142.000000
mean       166.267736
std        121.693285
min         10.000000
25%         89.000000
50%        135.000000
75%        200.000000
max        999.000000
Name: Price, dtype: float64
```

- The price distribution is skewed to the right.
- 50% of the properties are priced between $90 - $200

Python: Data Analysis

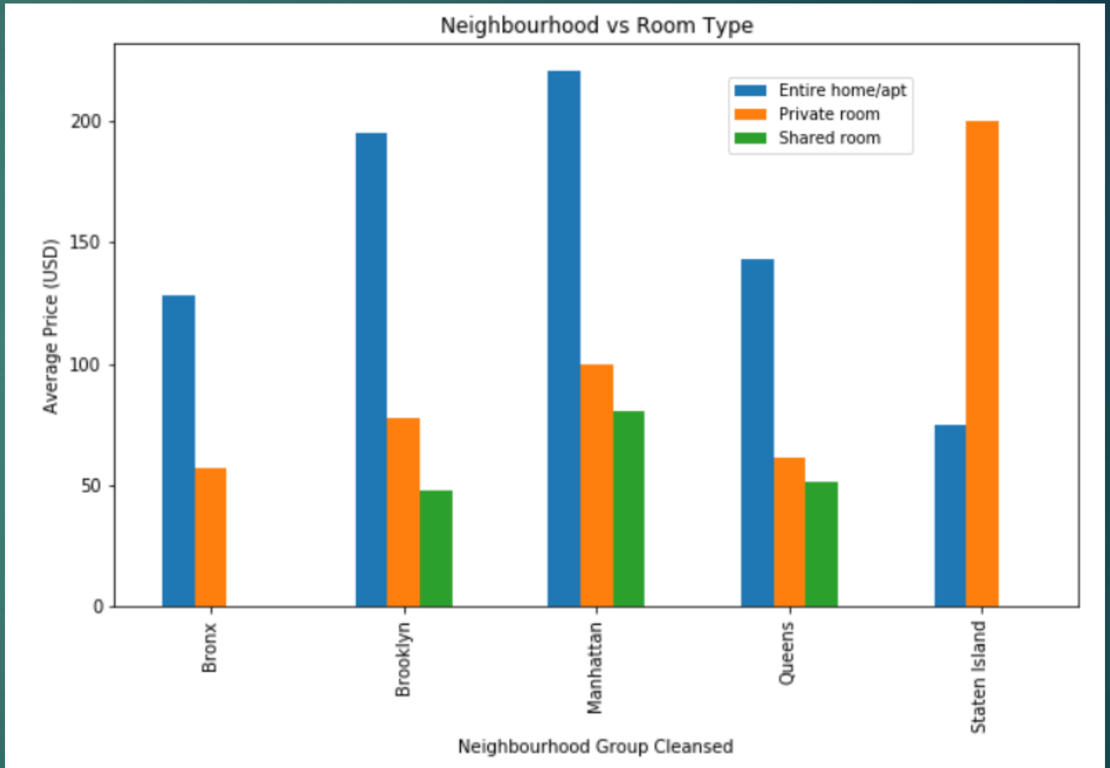
Property Price vs Neighbourhood


Trends : Property Price vs Location

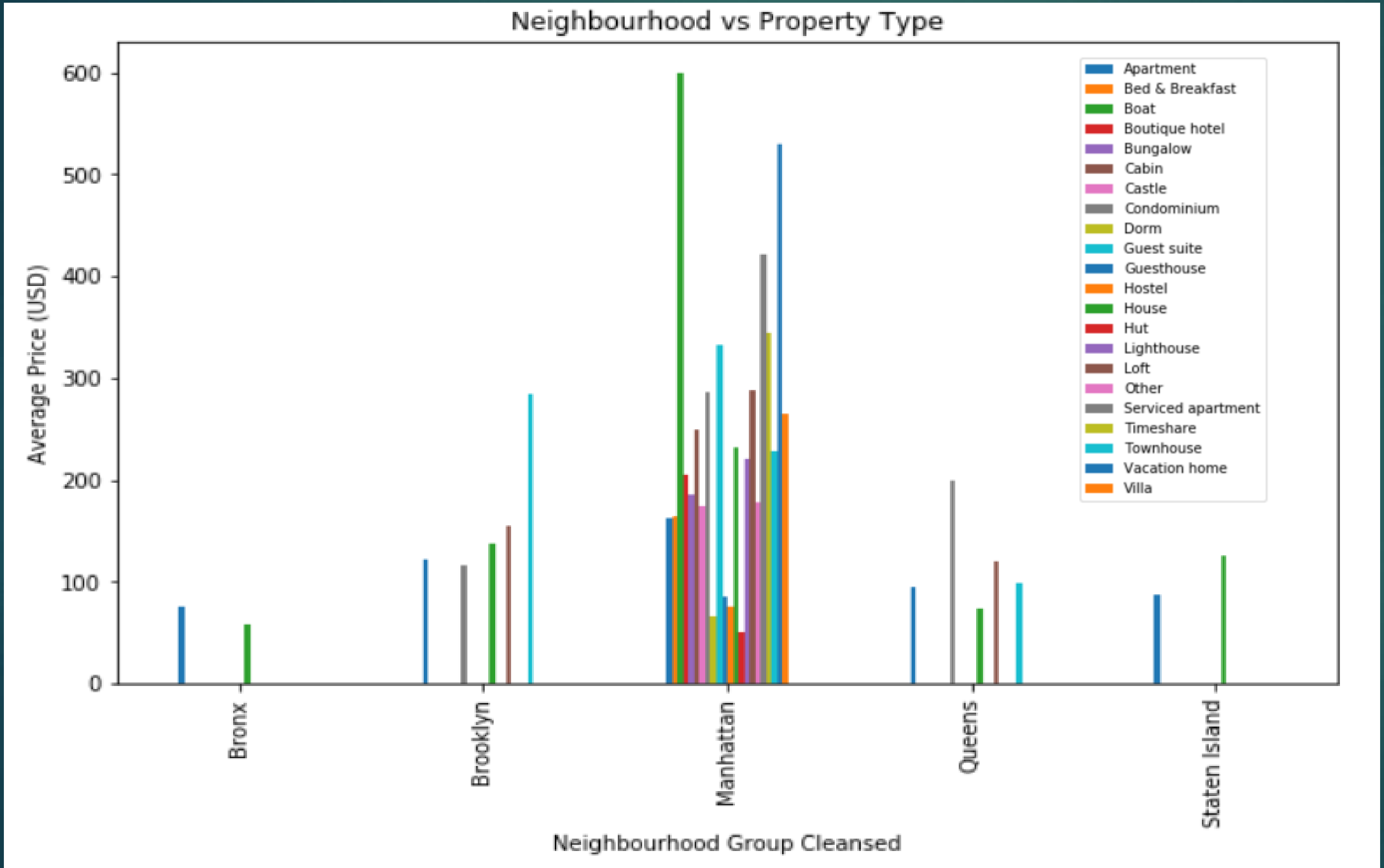
- Manhattan & Brooklyn have more expensive property listings.

- Average Price $170 and $125 respectively.

Property Price vs Room Type


Neighbourhood vs Room Type

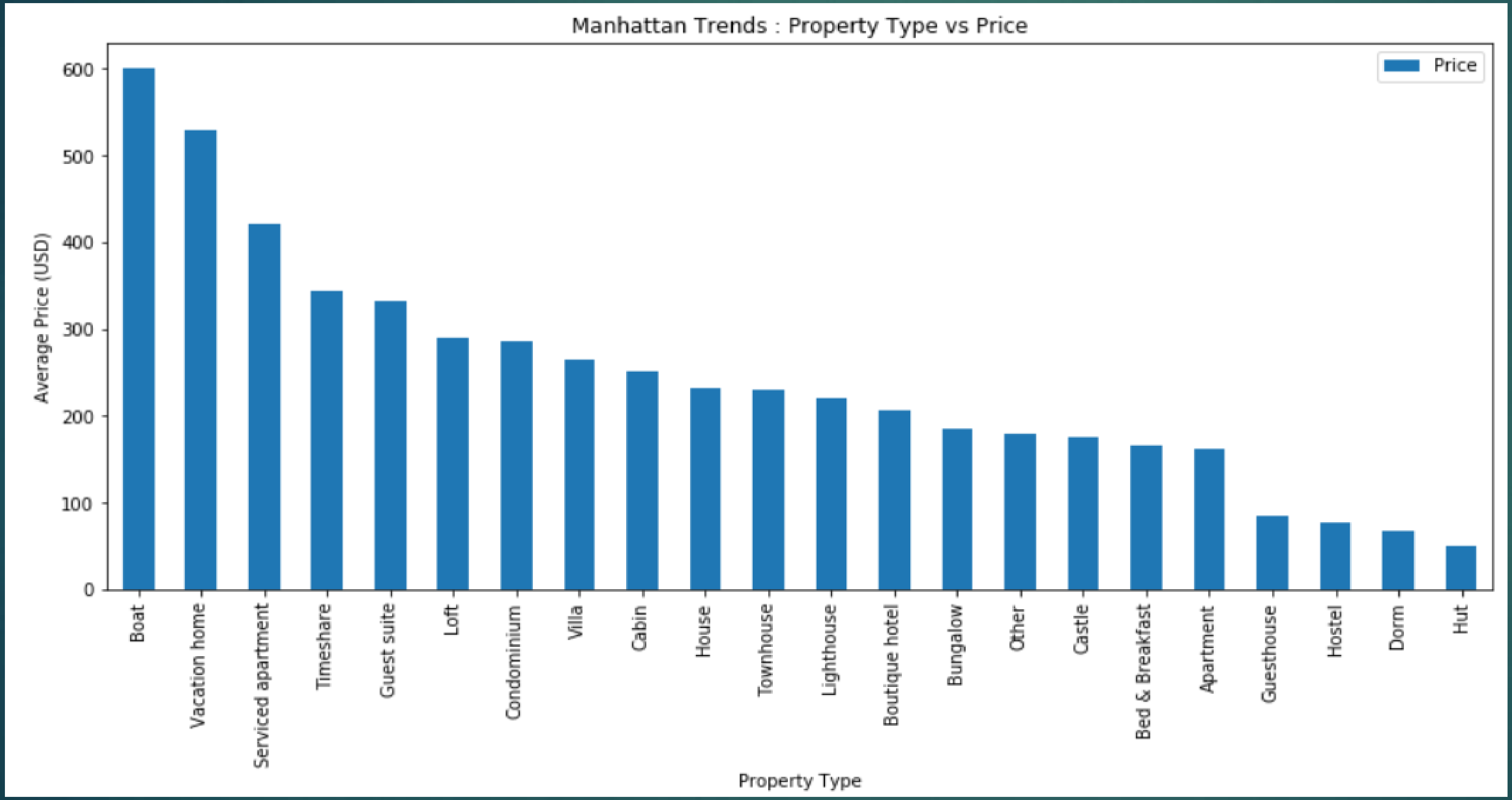
Average Property Price in Neighbourhood vs Property Type



- Price for property types difference with change in Neighbourhood groups.
- Rent a boat in Manhattan at $600 or Live in a Hut for $50
- Boats are cheaper in Bronx, Brooklyn, Queens, Staten Island

Python: Data Analysis

Price vs Property Type

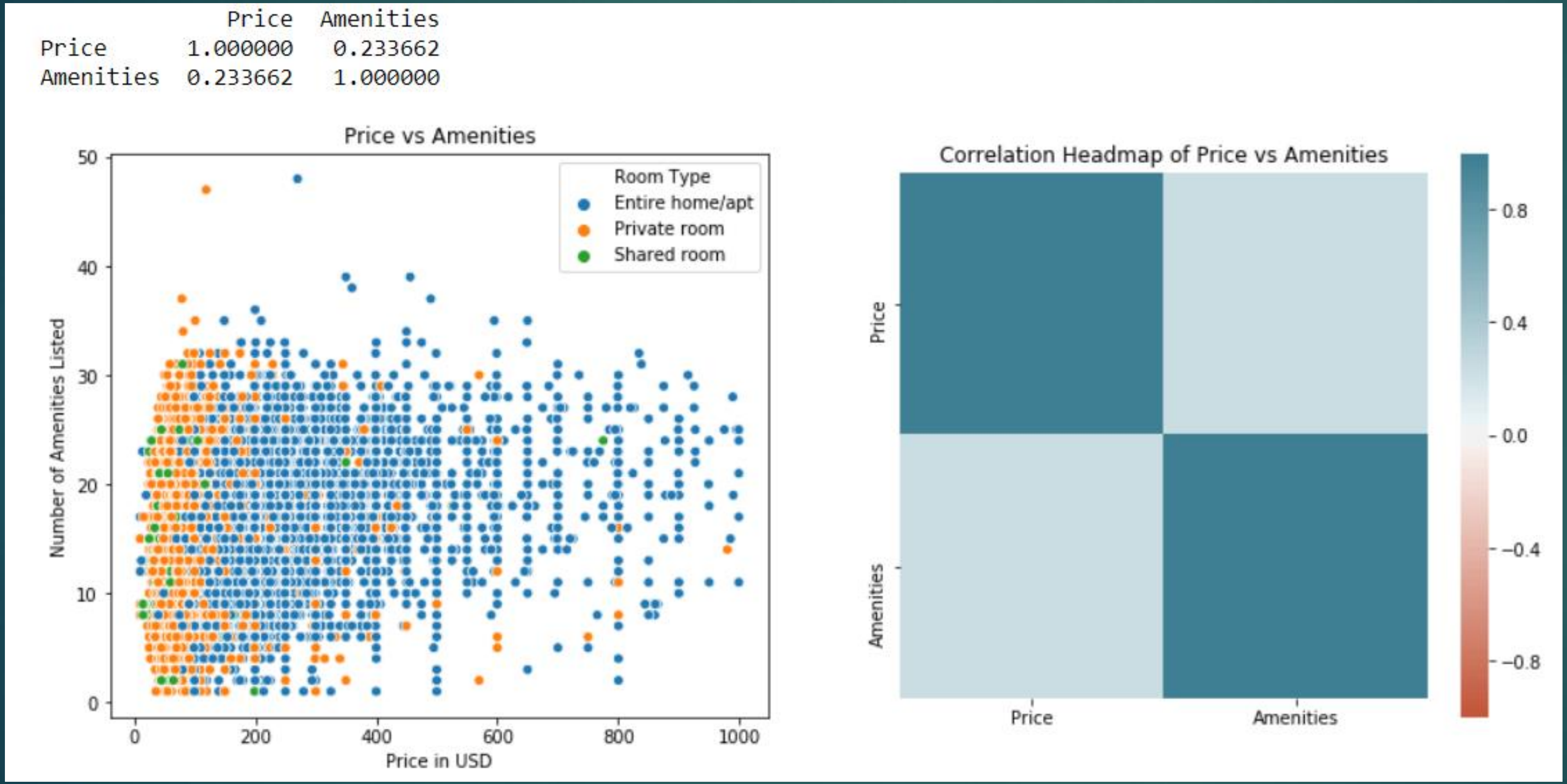


- Boats and Vacation homes most expensive
- Lower price options are Hostels, Dorm or Huts
- Budget friendly options are Cabins or Houses

Decreasing order of average Price based on Property type in Manhattan

Python: Data Analysis

Price vs Amenities

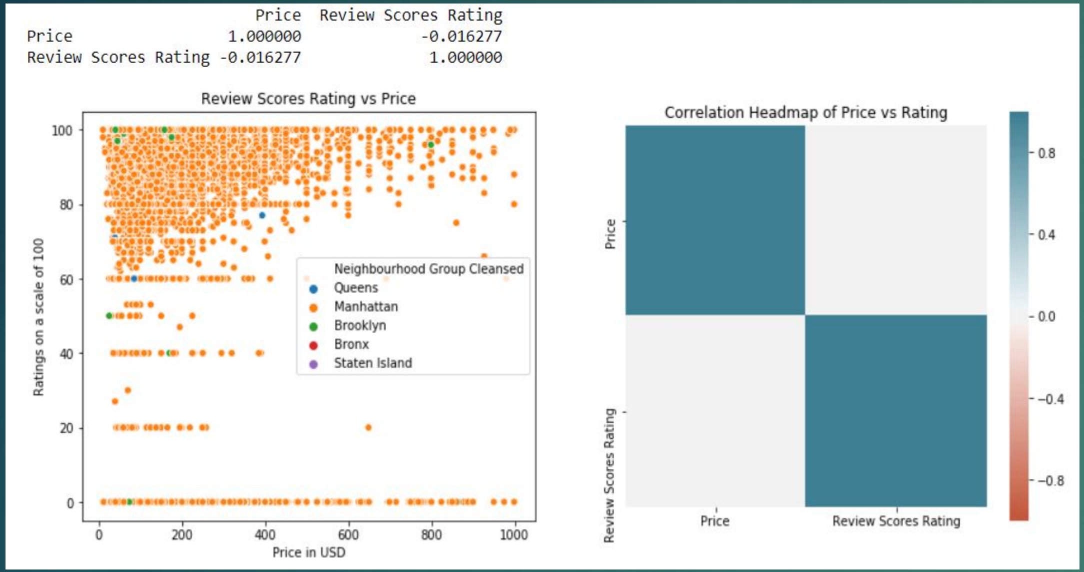


```
                Price  Amenities
Price        1.000000   0.233662
Amenities    0.233662   1.000000
```

- Price variations are significant according to the Room Type.
- Private Rooms are priced mostly < $200.
- Price is higher for an Entire Apartment.

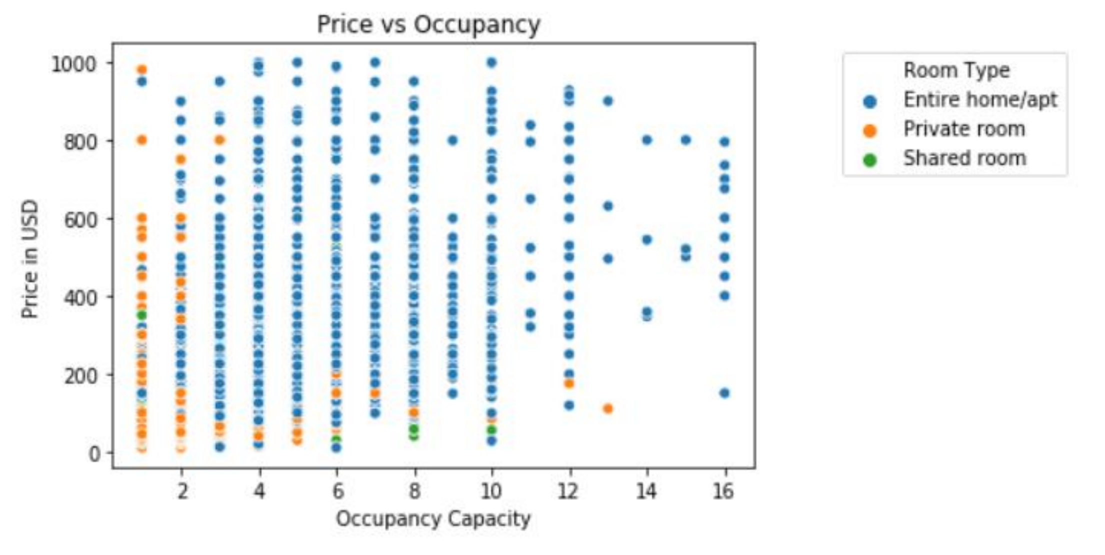
Python: Data Analysis

Price vs Rating



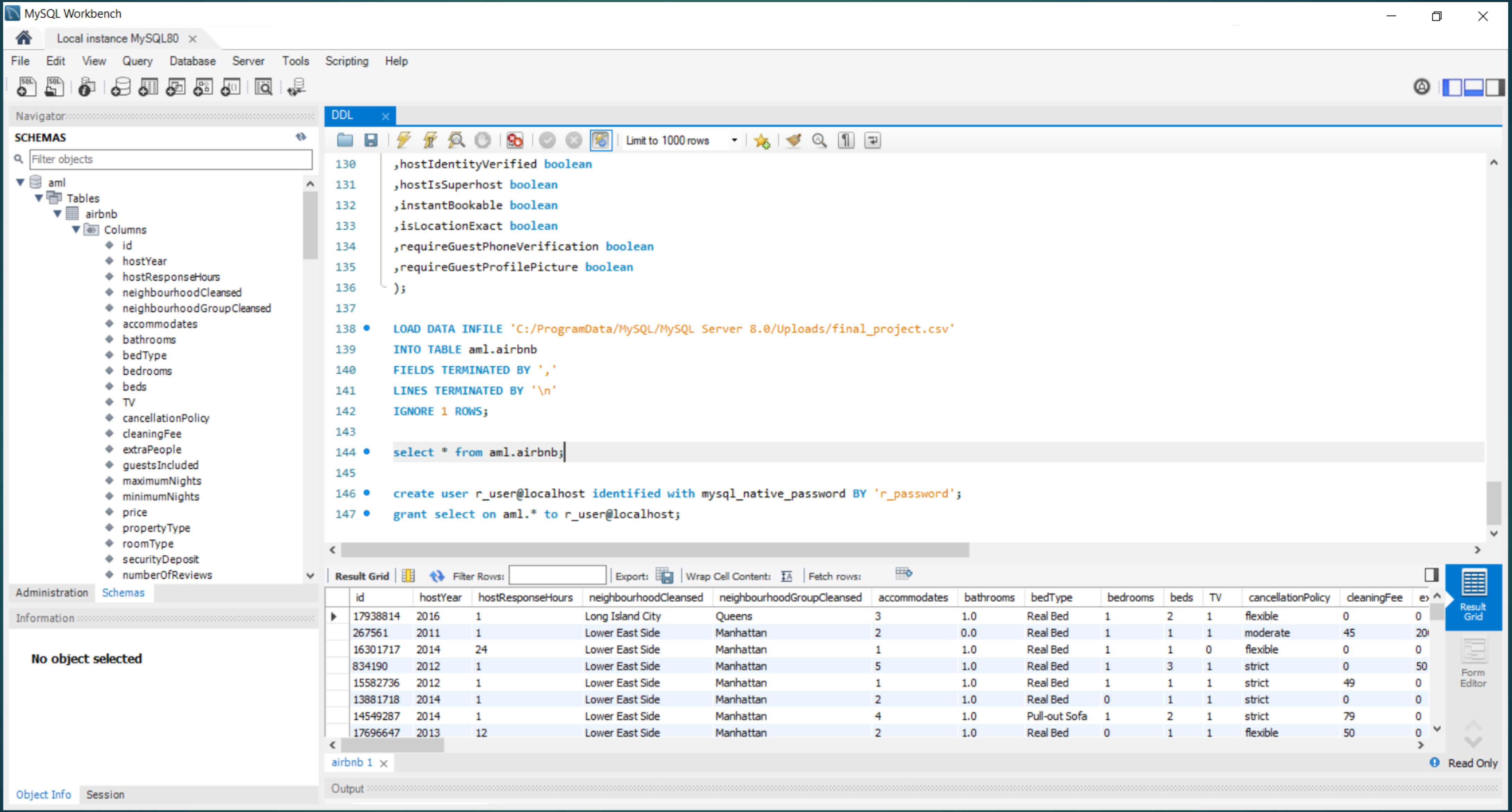
```
                        Price  Review Scores Rating
Price                1.000000             -0.016277
Review Scores Rating -0.016277             1.000000
```

Price vs Occupancy



- Budget friendly properties have higher ratings.
- Properties with 80% and above rates are concentrated over an average price of $200
- This pattern also could be because there are fewer in the extremely high price bracket.

Database

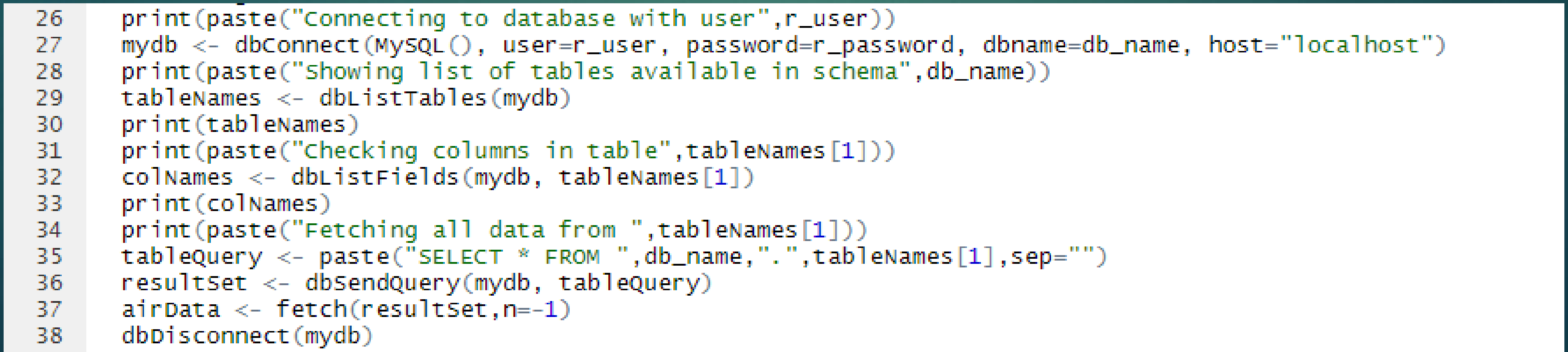


```
,hostIdentityVerified boolean
,hostIsSuperhost boolean
,instantBookable boolean
,isLocationExact boolean
,requireGuestPhoneVerification boolean
,requireGuestProfilePicture boolean
);

LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/final_project.csv'
INTO TABLE aml.airbnb
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;

select * from aml.airbnb;

create user r_user@localhost identified with mysql_native_password BY 'r_password';
grant select on aml.* to r_user@localhost;
```

id	hostYear	hostResponseHours	neighbourhoodCleansed	neighbourhoodGroupCleansed	accommodates	bathrooms	bedType	bedrooms	beds	TV	cancellationPolicy	cleaningFee	ex
17938814	2016	1	Long Island City	Queens	3	1.0	Real Bed	1	2	1	flexible	0	0
267561	2011	1	Lower East Side	Manhattan	2	0.0	Real Bed	1	1	1	moderate	45	20
16301717	2014	24	Lower East Side	Manhattan	1	1.0	Real Bed	1	1	0	flexible	0	0
834190	2012	1	Lower East Side	Manhattan	5	1.0	Real Bed	1	3	1	strict	0	50
15582736	2012	1	Lower East Side	Manhattan	1	1.0	Real Bed	1	1	1	strict	49	0
13881718	2014	1	Lower East Side	Manhattan	2	1.0	Real Bed	0	1	1	strict	0	0
14549287	2014	1	Lower East Side	Manhattan	4	1.0	Pull-out Sofa	1	2	1	strict	79	0
17696647	2013	12	Lower East Side	Manhattan	2	1.0	Real Bed	0	1	1	flexible	50	0

R: Connection to Database

- Library RMySQL is used to connect to mySQL database.
- All DDL and DML commands are runnable through connection (provided user has corresponding privileges).
- Query results can be read into variables similar to reading CSV files.
- Numerical data types are converted to decimal automatically (with warnings).



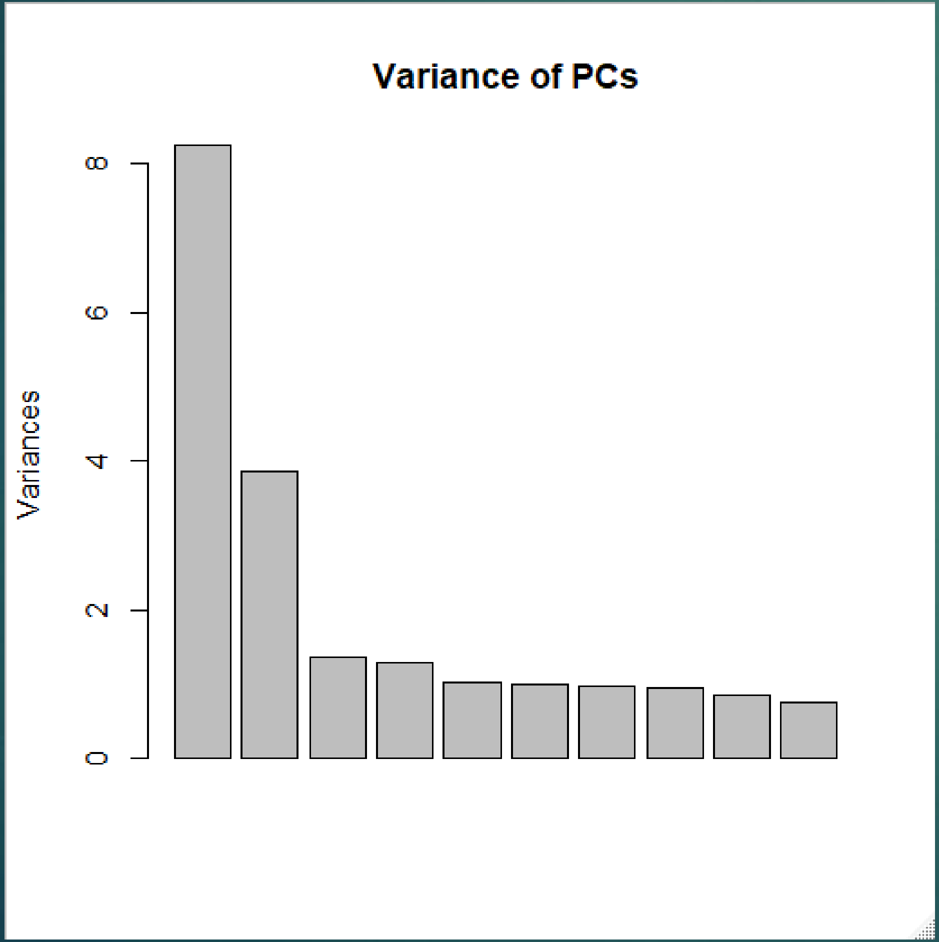
```r
print(paste("Connecting to database with user",r_user))
mydb <- dbConnect(MySQL(), user=r_user, password=r_password, dbname=db_name, host="localhost")
print(paste("Showing list of tables available in schema",db_name))
tableNames <- dbListTables(mydb)
print(tableNames)
print(paste("Checking columns in table",tableNames[1]))
colNames <- dbListFields(mydb, tableNames[1])
print(colNames)
print(paste("Fetching all data from ",tableNames[1]))
tableQuery <- paste("SELECT * FROM ",db_name,".",tableNames[1],sep="")
resultSet <- dbSendQuery(mydb, tableQuery)
airData <- fetch(resultSet,n=-1)
dbDisconnect(mydb)
```

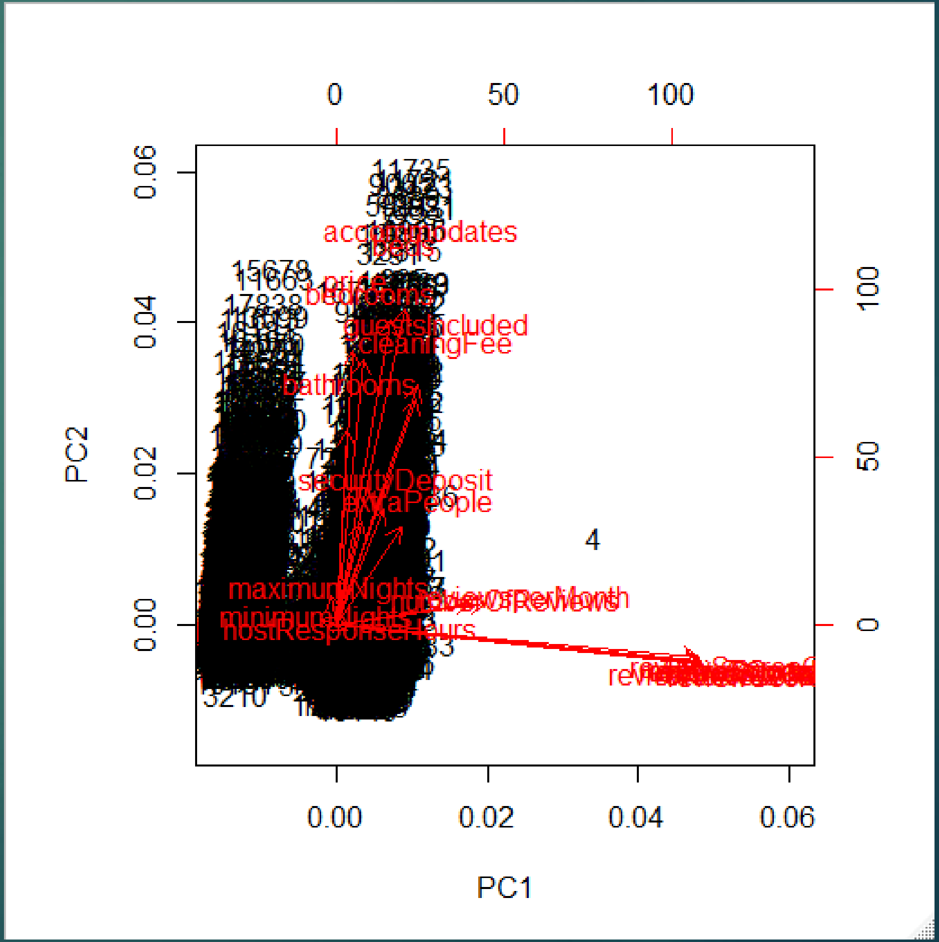
R Code: Data manipulation

- Converting Boolean data types to factor (0/1) features.
- Derive yearsAsHost from hostYear.
- Drop columns used to derive features.
- Club infrequent categorical levels into 'Other' to prevent errors during prediction on validation set.
- Remove records with price = 0 (to be used purely for prediction).
- Number of records in prediction set: 131
- Create training and test data sets from remaining valid data.
- Remove features as required at various stages.

R Code: Data analysis (PCA)

- Analysis of numerical features.
- Features with similar factor loadings were removed.





R Code: Regression Models

- Linear Regression.
- K-fold cross validation for linear regression after PCA.
- K-fold cross validation for linear regression after removal of features without significance in previous regression model.
- Regression Tree.
- Cross validation based pruning for tree.
- Ridge regression.
- Lasso regression.
- XGBoost.

R Code: Model Performance

- Linear regression provided similar error rate and R-squared for all levels (with all features, after PCA-based removal, and p-value based removal).
- Regression Tree provided comparable results and pruning resulted in the same tree.
- Ridge and Lasso regression provided comparable results with notable difference only in weights of features.
- XGBoost provided highest R-squared and smallest RMSE.
- Cross validation based estimation of parameters required long time (~35-40 minutes per run x approximately 10 runs).
- XGBoost yielded almost +5% increase in R-squared value for same data.

R Code: Prediction output to CSV

- XGBoost model run on prediction set.
- Output written to CSV file.
- All models run with log(price).
- XGBoost prediction differed on average by ~$28 for price prediction vs log(price) prediction.

Conclusion

- Despite large amount of features, explanatory power of features limited.
- Pricing of AirBnB listing in NYC possibly subjective and/or dependent on other features.
- Relationship between price and features may be non-linear.

Thank you.