Problem Statement 1

BRAZIL HOUSE RENT PREDICTION

PROBLEM STATEMENT:

Explore the given Brazil house rent data set using EDA techniques visualize the results and build a suitable model to predict the house rent.

OBJECTIVE:

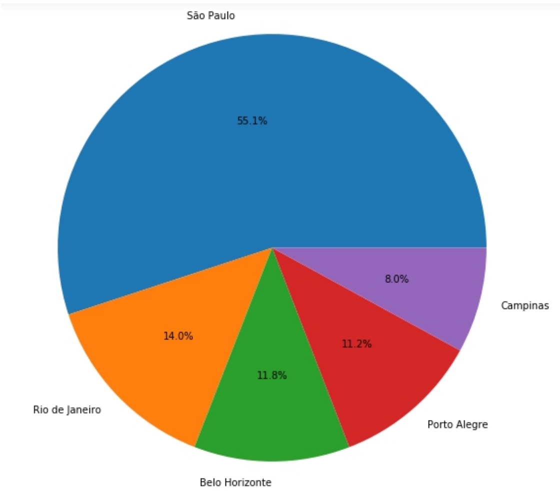
- Exploratory Data Analysis
- Data Pre-processing
- Feature Selection
- Model Building
- Validation

BACKGROUND:

The given dataset is based on classification where to predict the Brazil House Rent for new data. The dataset consists of 10692 rows and 13 columns.

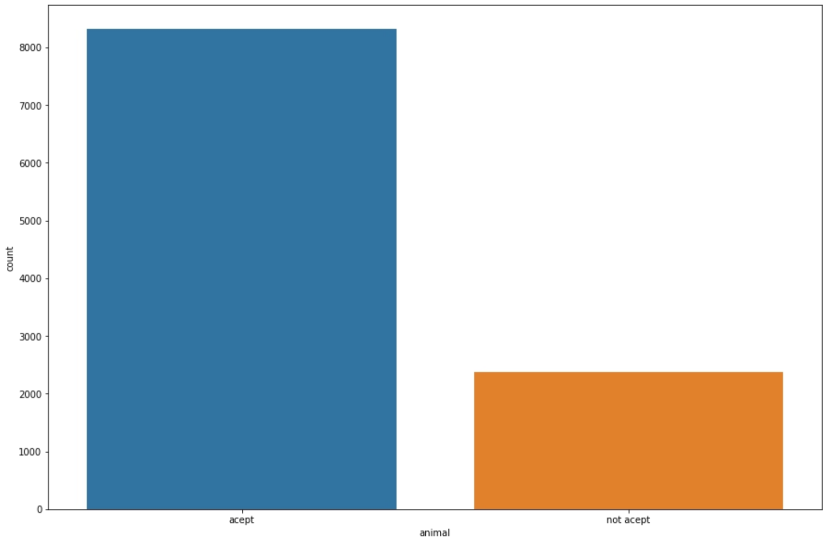
EXPLORATORY DATA ANALYSIS:

1. How is the distribution of each city?



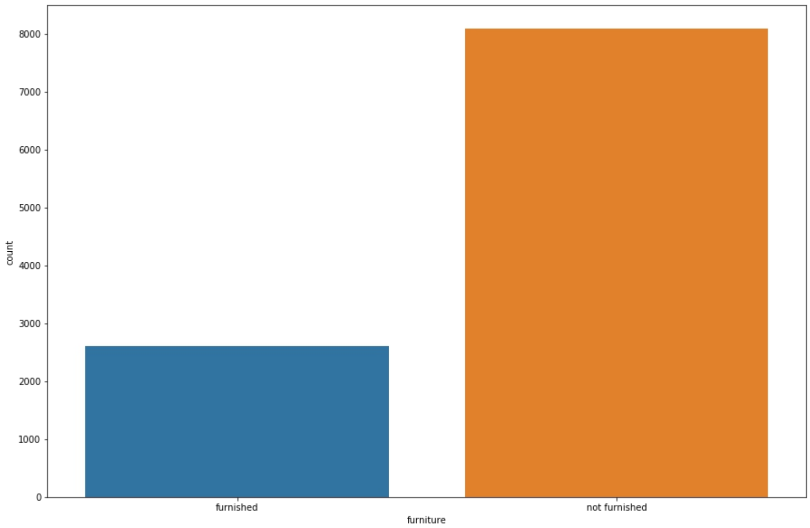
From the chart, São Paulo is the city with more houses

2. How many house owners accept animals in the home?



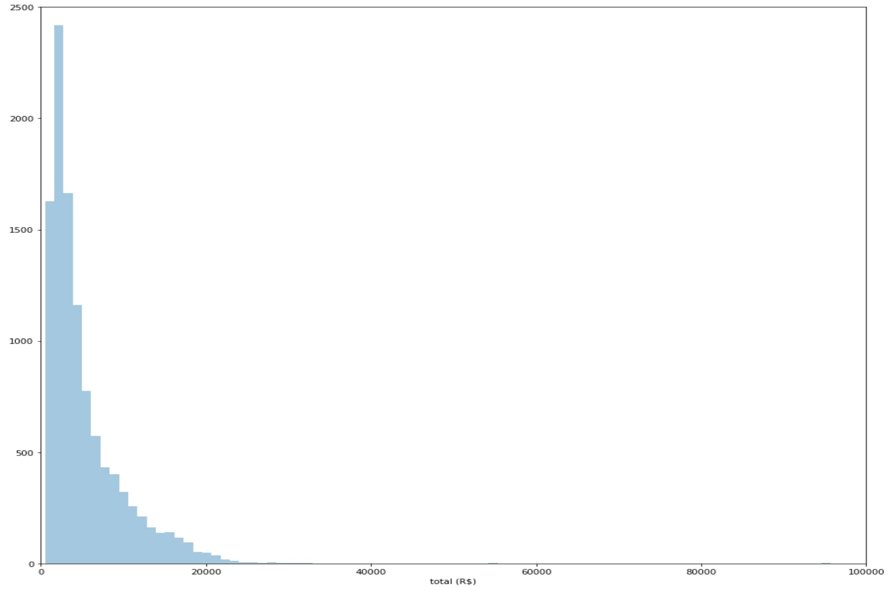
From the chart, most houses accept pet animals.

3. How many houses are furnished?



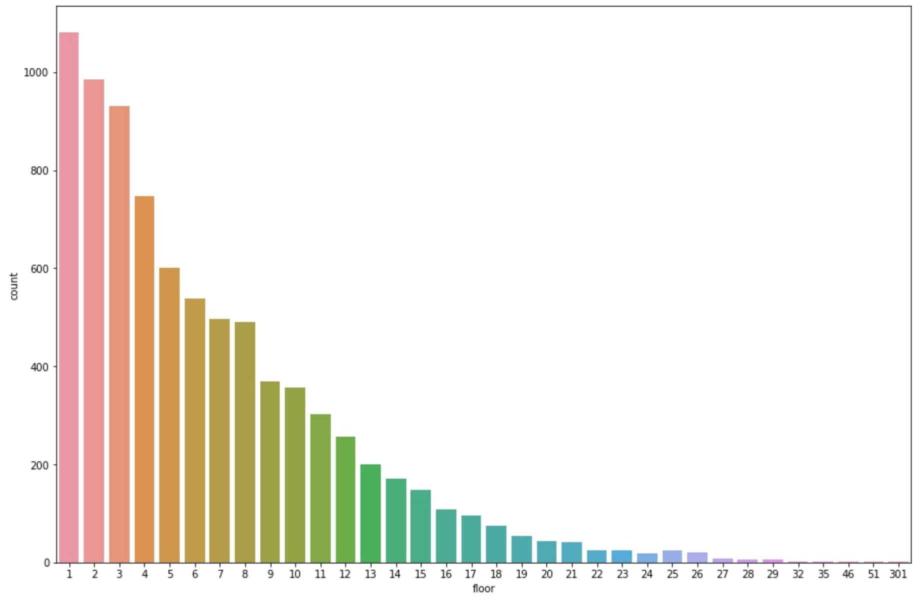
From the chart, most houses are not furnished.

4. Where is the accumulation point of total price?

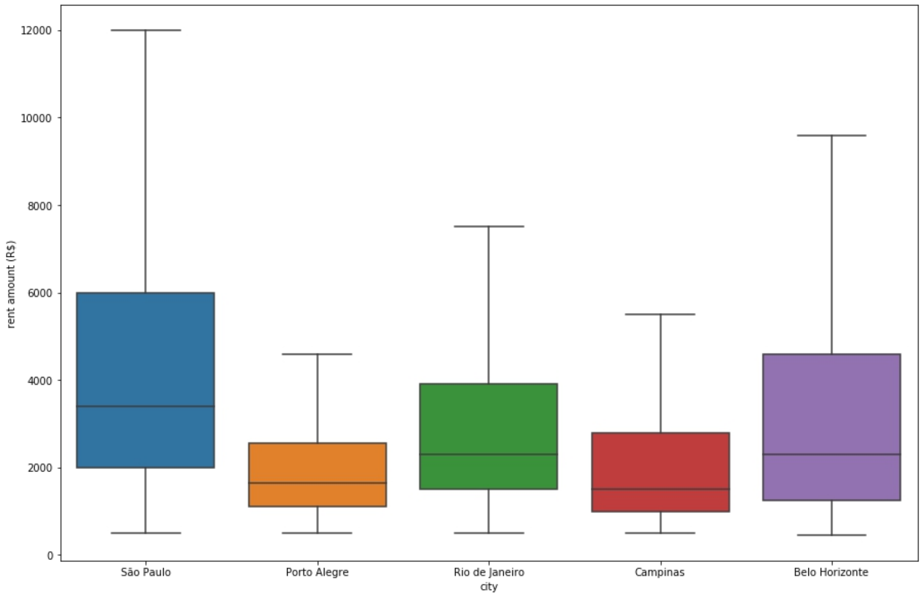


The accumulation point is between 2000 and 3000.

5. How is the distribution of floors?

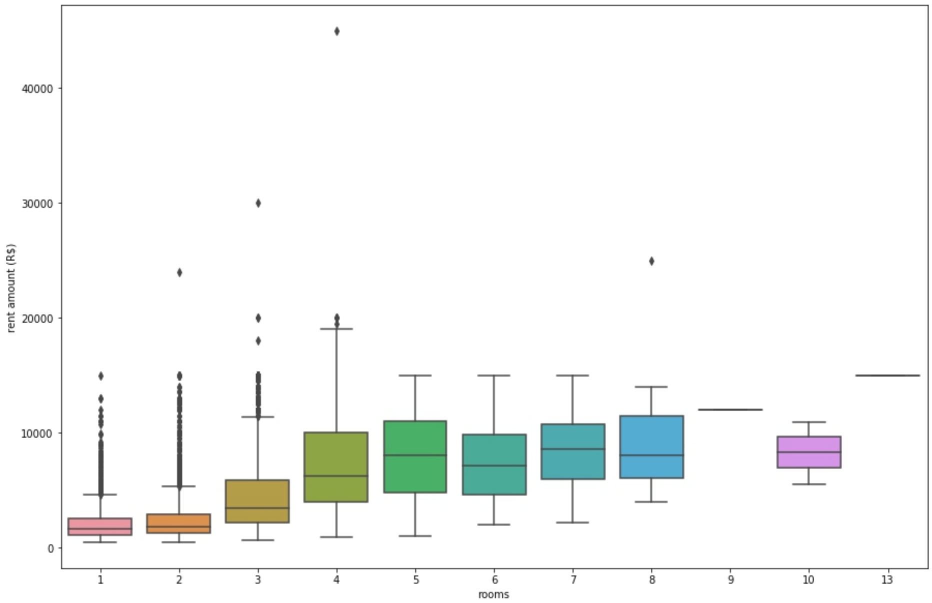


6. Which city has the most expensive rent prices?



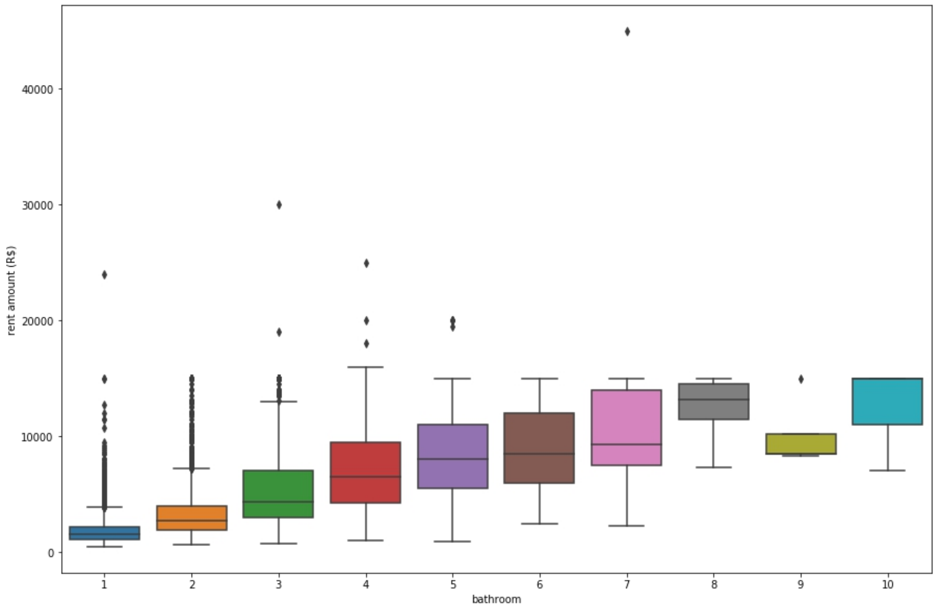
It seems like Sao Paulo has the most expensive rent prices.

7. Which floor is the most expensive?



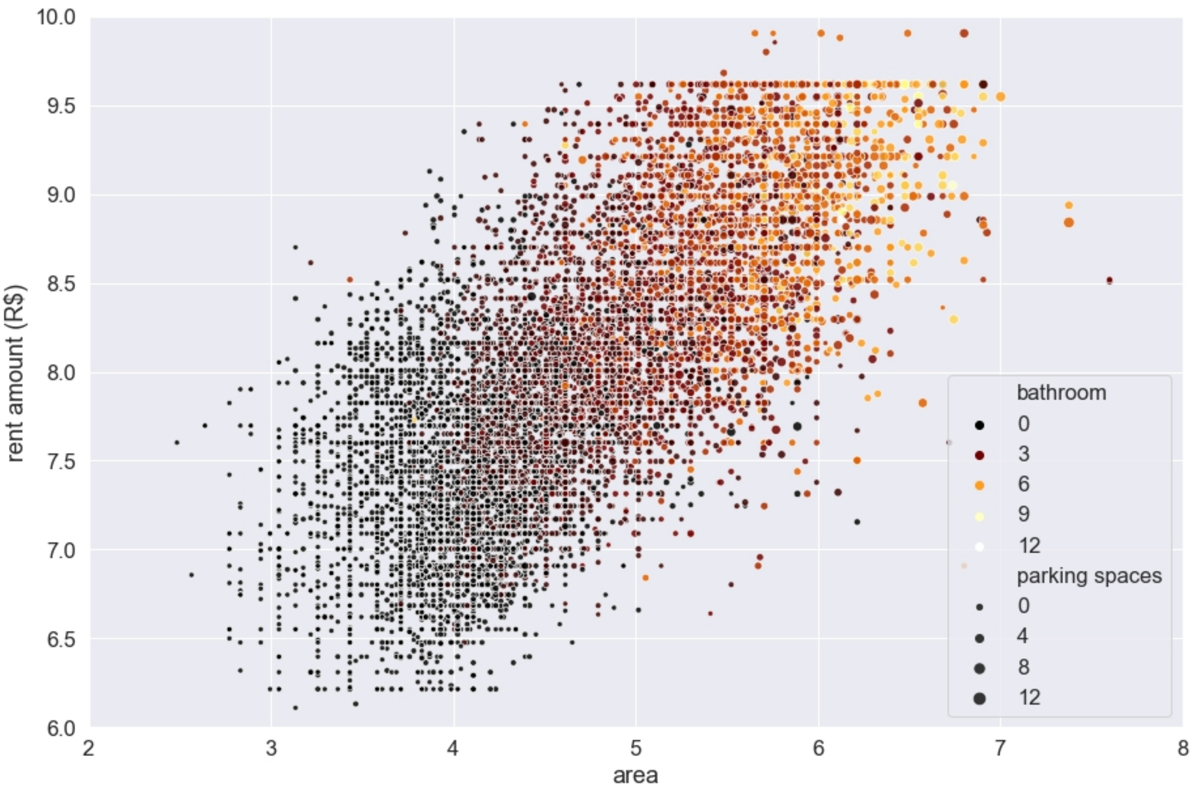
From the graph,the floors 5-8 are almost expensive. The answer could be either 5th or 7th floor.

8. Does the number of bathrooms affect the rent amount?



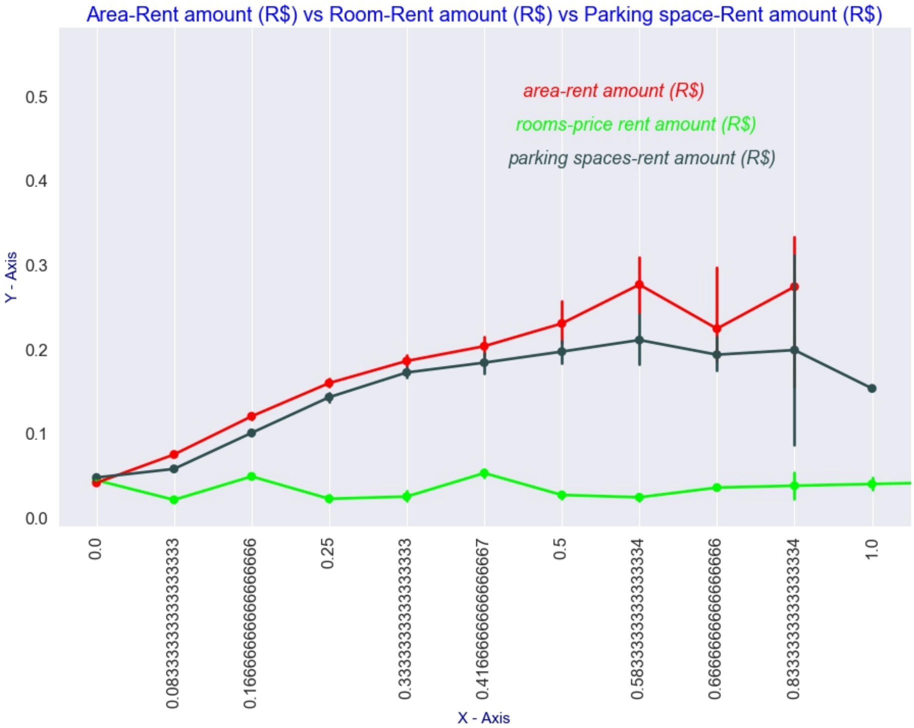
Yes, as the number of bathrooms in a house increases, the rent also increases.

9. How strong is the correlation between area,number of bathroom and rent amount?



All have positive correlation with the rent.

10. Which feature is correlated the most with rent amount: Area? Number of rooms? Parking Spaces?



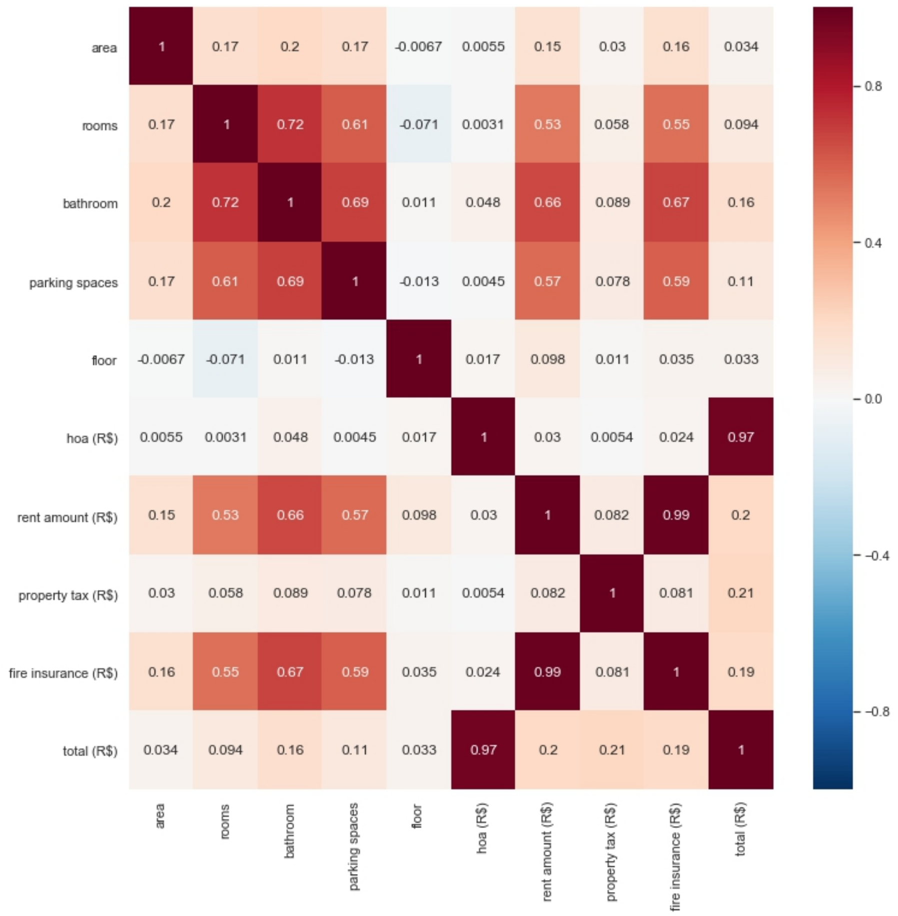
Area is correlated the most with rent.

PREDICTIVE ANALYSIS:

- **DATA PRE-PROCESSING**

1. **Cleansing the Data**
   The floor variable has an unwanted symbol '-' and it is removed.
2. **Dealing with outliers**
   To treat the outliers, the interquartile range is used and performed this analysis in every city.
3. **Data Wrangling**
   We used a labelencoder for furniture because it only has two values.
   For the cities we have used OneHotEncoder and dropped the first column to avoid the dummy variable trap.

- **FEATURE SELECTION**



We have used the columns that have more correlation with the variable that we want to predict.

- **MODEL BUILDING**

  I have used several models and analyzed the best among them.
  These are the models:

- Linear Regression
- Ridge Regression
- Decision Tree
- Random Forest
- Support Vector Regression (SVR)
- KNearestNeighbours (KNN)
- Lasso Regression
- GridSearch to find the best parameters on Lasso and Ridge

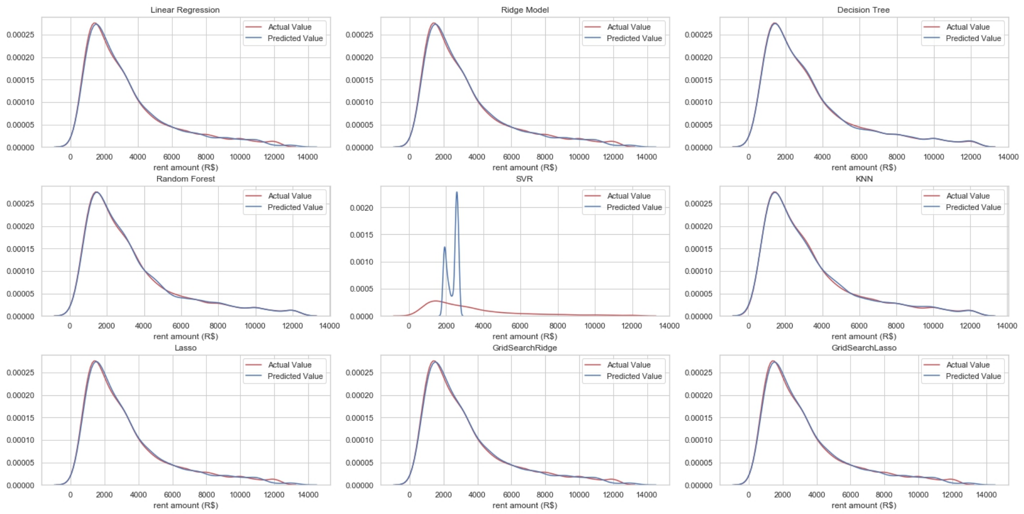
- **VALIDATION:**

  For validation MAE, RMSE and R2 score is used.

```
Linear Regression
MAE: 248.99289449586416
RMSE: 372.6152499362306
R2: 0.978435563565699
****************************************
Ridge Model
MAE: 248.98912238422588
RMSE: 372.61549419180517
R2: 0.9784355352939869
****************************************
Decision Tree
MAE: 141.2116552152166
RMSE: 346.8138229906298
R2: 0.9813185896505354
****************************************
Random Forest
MAE: 141.08075583369595
RMSE: 295.98079124251603
R2: 0.9863935786362991
****************************************
SVR
MAE: 1551.9569900522486
RMSE: 2569.299940825056
R2: -0.025289420627535364
****************************************
KNN
MAE: 160.46795856999665
RMSE: 315.0206757145174
R2: 0.9845867231272063
****************************************
Lasso
MAE: 247.54779770272793
RMSE: 372.82011833939237
R2: 0.9784118442672647
****************************************
GridSearchRidge
MAE: 248.95527069031468
RMSE: 372.6177481766283
R2: 0.9784352744024033
****************************************
GridSearchLasso
MAE: 248.99274499244535
RMSE: 372.6152612820053
R2: 0.9784355622524664
****************************************
```

ANALYSIS OF THE RESULTS:

Visualization of the plot's for each regressor



	model	MAE	RMSE	R2
0	Random Forest	141.080756	295.980791	0.986394
1	Random Forest	141.080756	295.980791	0.986394
2	Random Forest	141.080756	295.980791	0.986394
3	Random Forest	141.080756	295.980791	0.986394
4	KNN	160.467959	315.020676	0.984587
5	KNN	160.467959	315.020676	0.984587
6	KNN	160.467959	315.020676	0.984587
7	KNN	160.467959	315.020676	0.984587
8	Decision Tree	140.383545	344.691739	0.981547

RandomForest it's our best performer in all three metrics