# Melbourne House Price Prediction

Abdullah All Mamun
M.Sc in Data Science
Beuth University of Applied Science
ID: 902258

# Short Introduction

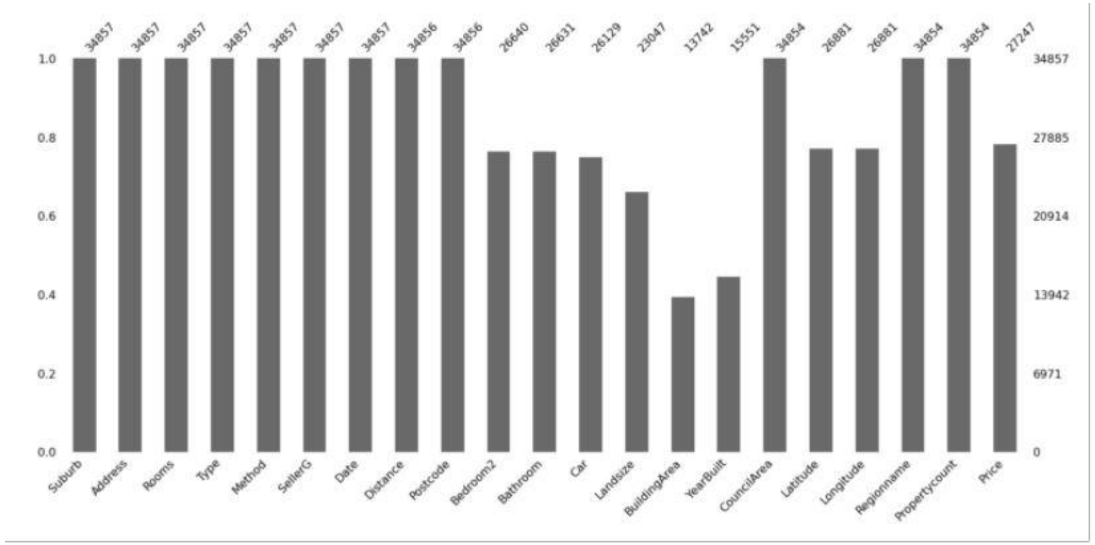
- Dataset
- Data Preprocessing
- Data Visualisation
- Feature selection
- Predective modeling
- Hyperparameter Tuning
- Performance Metrics

# Dataset

- Kaggle Dataset
- 34,857 records with 21 attributes
- Attributes information:
- SellerG: Real Estate Agent
- Date: Date sold
- Distance: Distance from CBD in Kilometres
- Regionname: General Region (West, North West, North, North east …etc)
- Propertycount: Number of properties that exist in the suburb.
- Bedroom2 : Scraped # of Bedrooms (from different source)
- Bathroom: Number of Bathrooms
- Car: Number of carspots
- Landsize: Land Size in Metres

- Suburb: Suburb
- Address: Address
- Rooms: Number of rooms
- Price: Price in Australian dollars
- Method:
- Type:
- BuildingArea: Building Size in Metres
- YearBuilt: Year the house was built
- CouncilArea: Governing council for the area
- Lattitude: Self explanitory
- Longtitude: Self explanitory

# Data Preprocessing



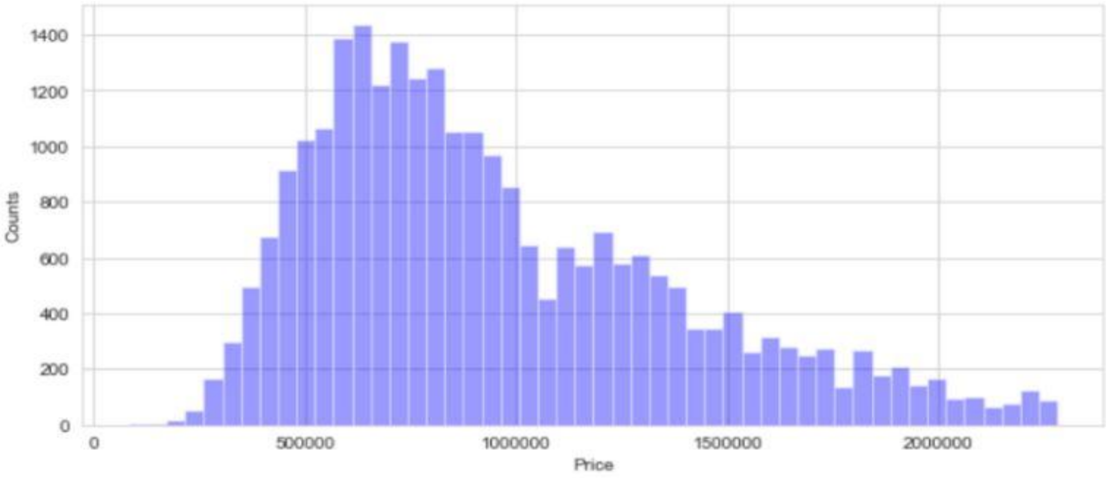
❑ **Missing values Handling**

- Missing values of price are dropped
- Missing values of  Bathroom and car are filled  using medians
- Features having missing  values are dropped

# Outliers Handling

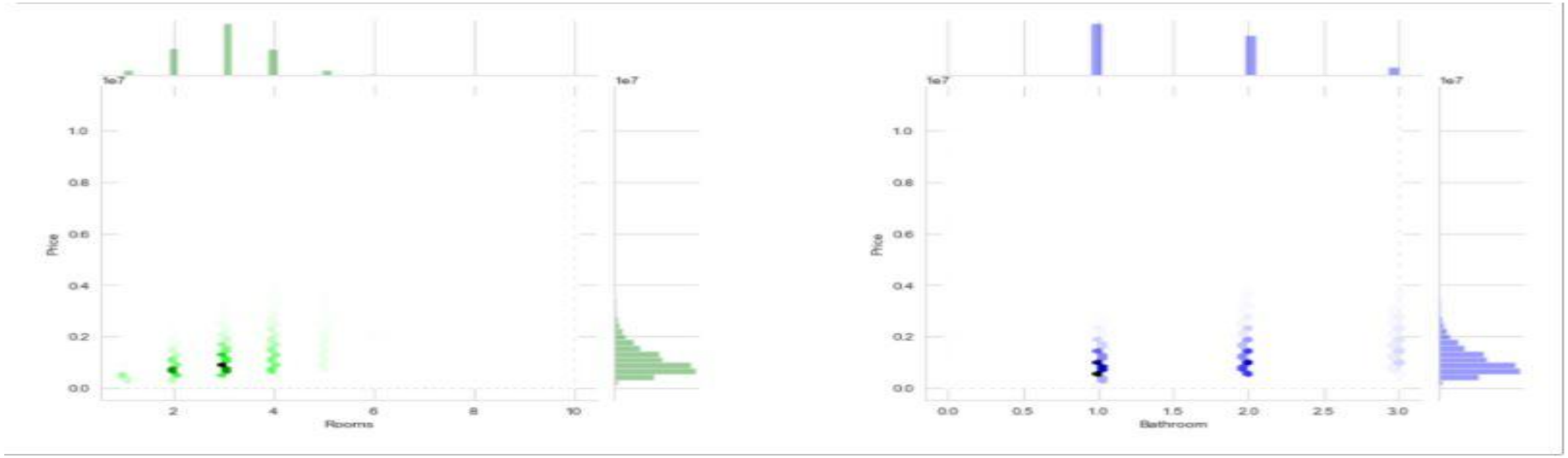
- Using Interquartile Range
- Drop if data points falls above the 3rd quartile and below the 1st quartile

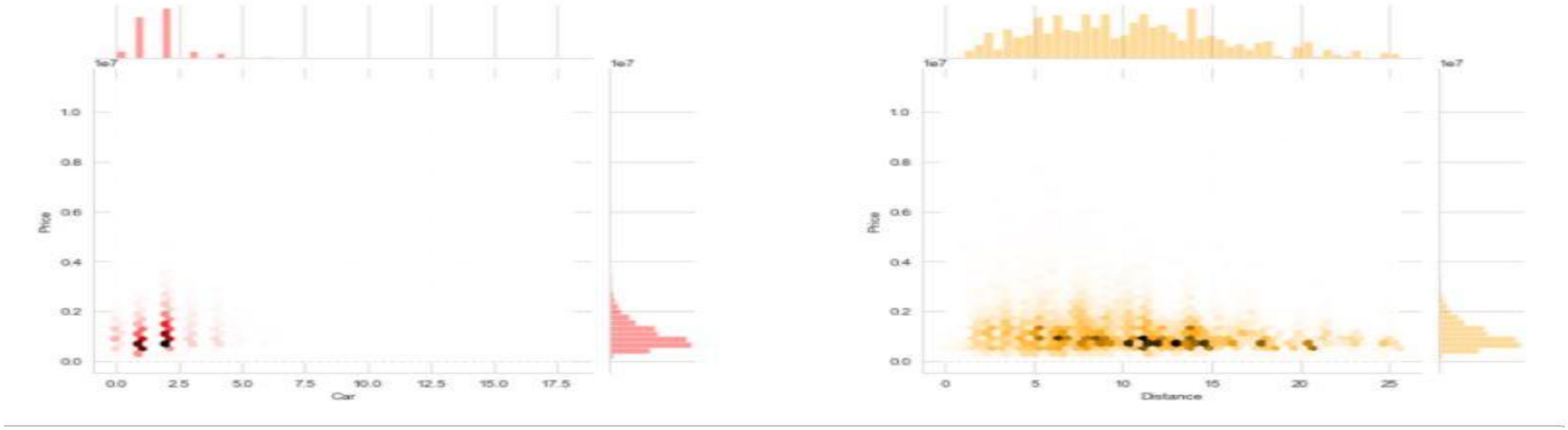
- Price $635000 and $1295000
- Rooms – 2 rooms and 4 rooms
- Distance-6.4 kilometers and 14 kilometers
- Bathroom- 1 and 2 rooms
- Car- 1 and 2 spots



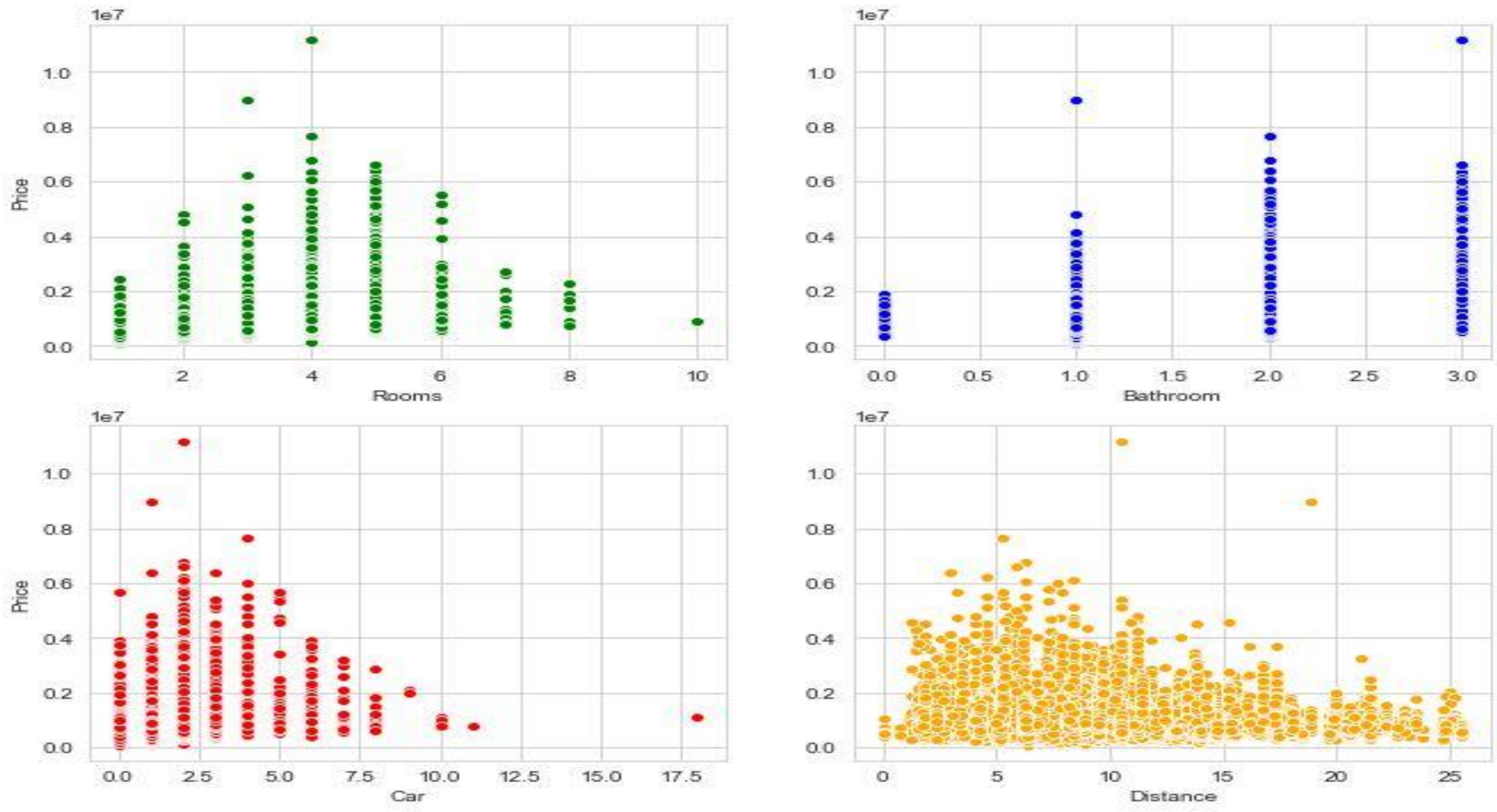
The Distribution of Price After Removing Outliers

# Exploratory Data Analysis(EDA)

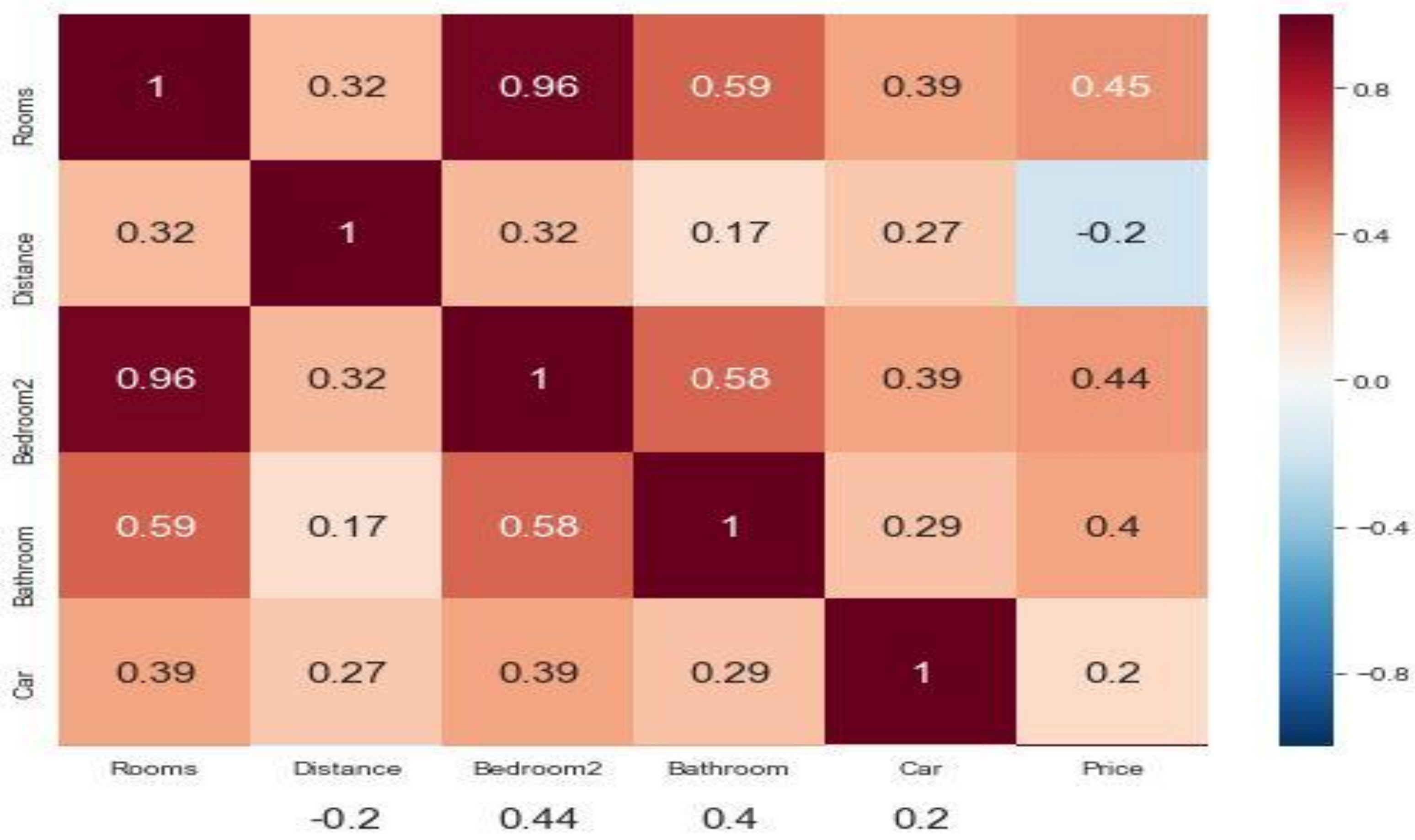




- ❑ Rooms Vs Price
- ❑ Bathroom Vs Price
- ❑ Car Vs Price
- ❑ Distance  Vs Price

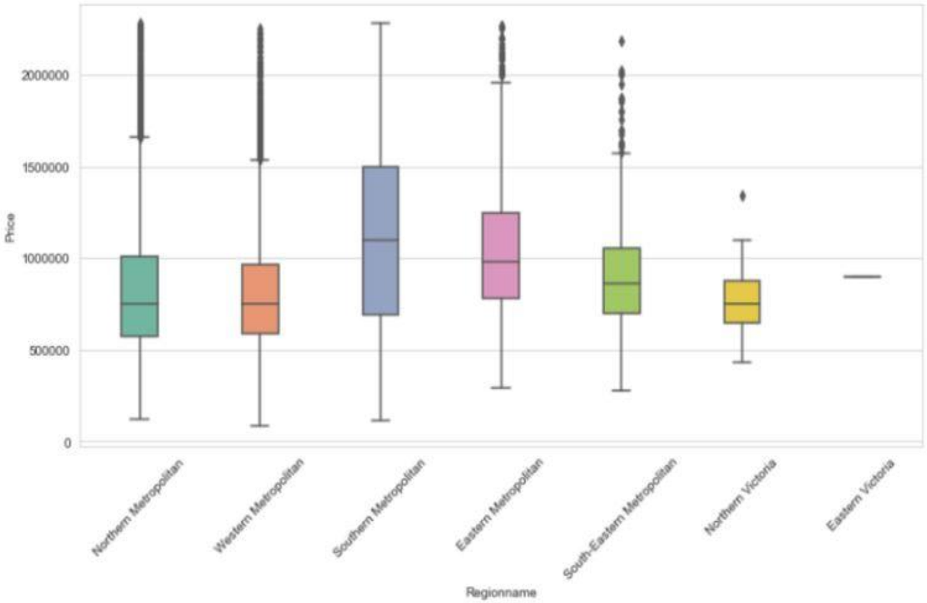


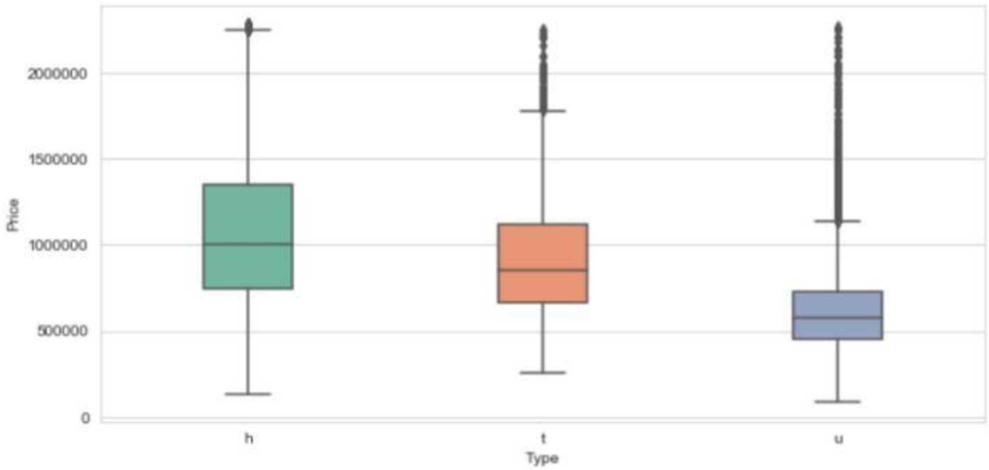
- Heatmap


Rooms
Distance
Bedroom2
Bathroom
Car
1
0.32
0.96
0.59
0.39
0.45
0.32
1
0.32
0.17
0.27
-0.2
0.96
0.32
1
0.58
0.39
0.44
0.59
0.17
0.58
1
0.29
0.4
0.39
0.27
0.39
0.29
1
0.2
Rooms
Distance
Bedroom2
Bathroom
Car
Price
-0.2
0.44
0.4
0.2
0.8
0.4
0.0
−0.4
−0.8

# Categorical Features

- Regionname and Type





H=House, cottage; t=townhouse; u=unit, duplex

# Predictive Modeling

- Machine learning model

□ Linear Regression

□ Ridge Regression

□ K-Nearest Neighbors

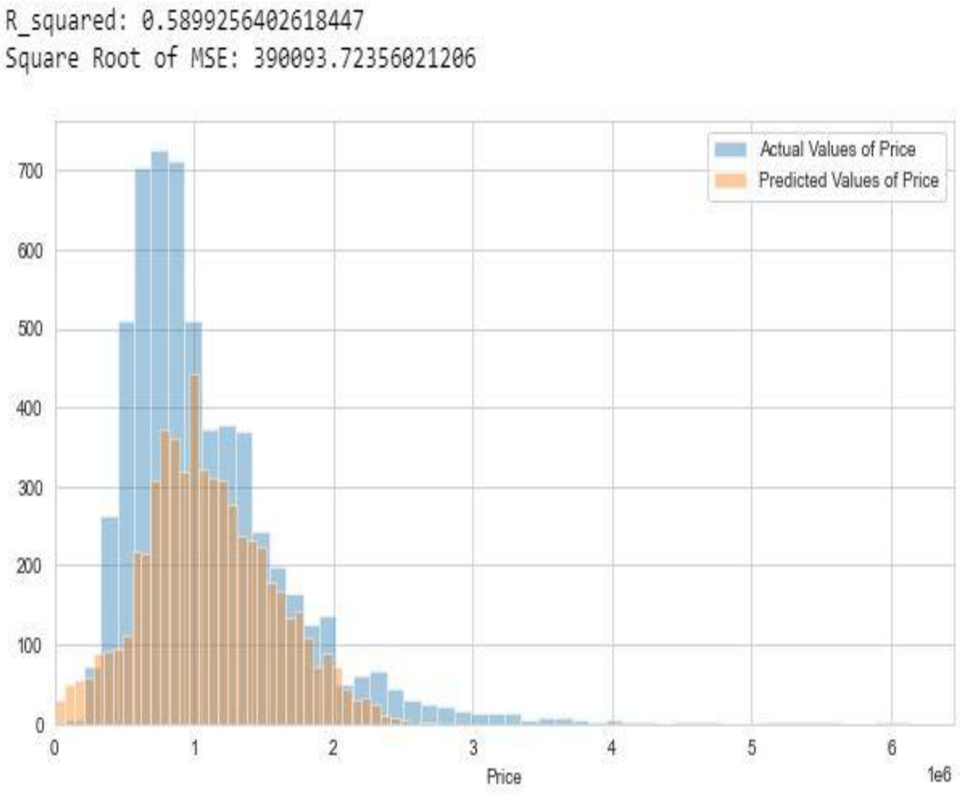
□ Decision Tree

- Performance Metrics

□ Coeffiecient of Determination

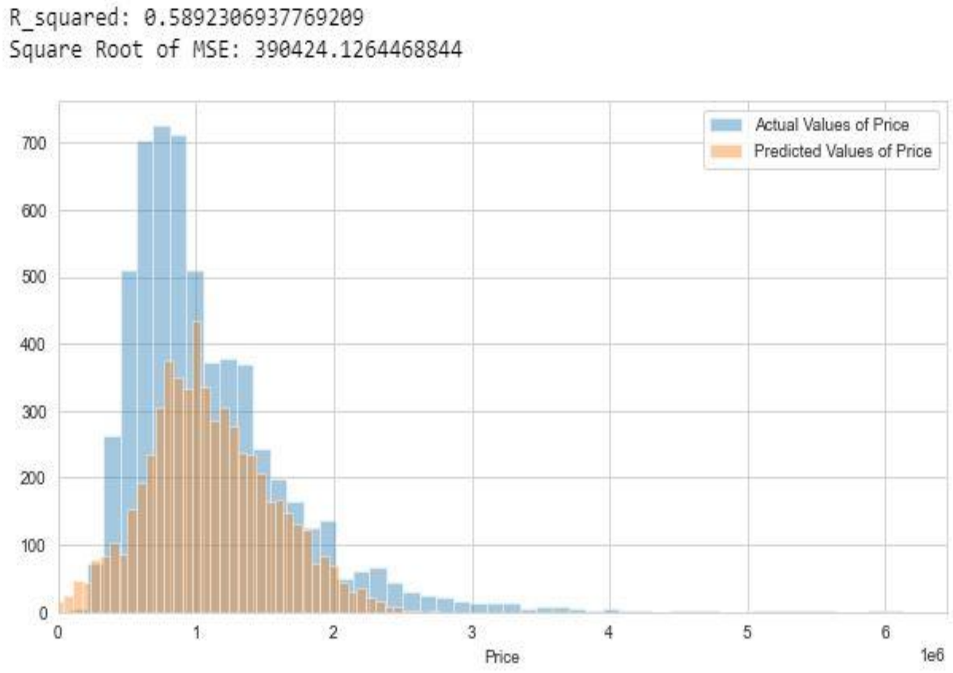
□ MSE(Mean Squared Error)

# Linear and ridge regression



```
R_squared: 0.5899256402618447
Square Root of MSE: 390093.72356021206
```

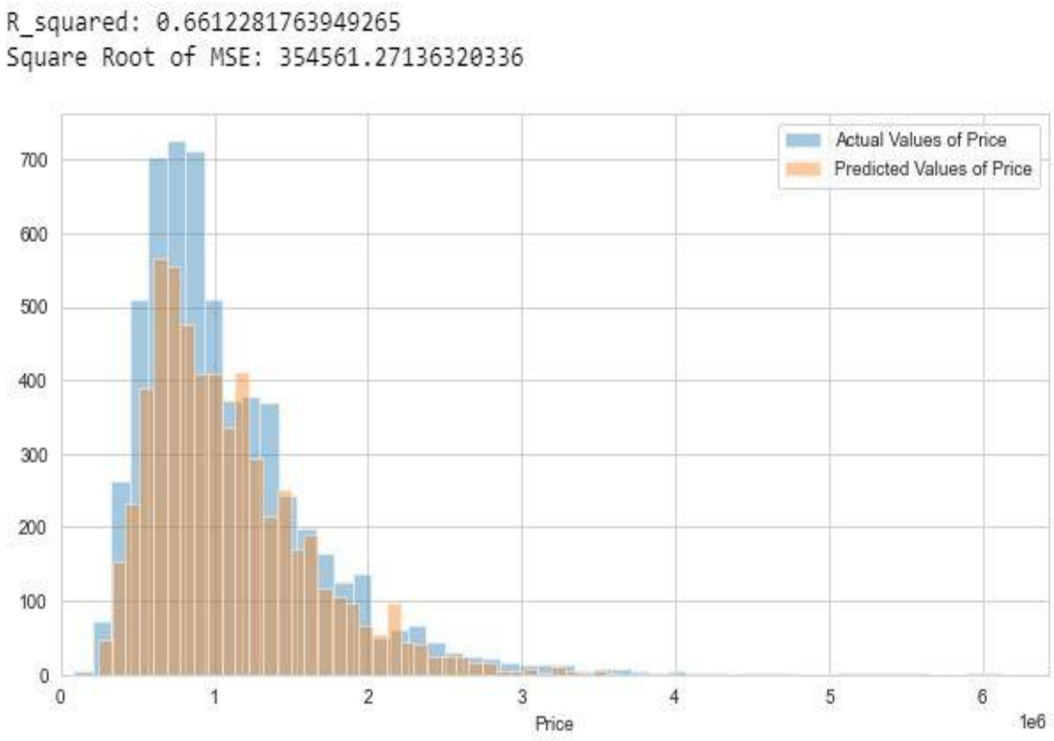
Linear regression



```
R_squared: 0.5892306937769209
Square Root of MSE: 390424.1264468844
```

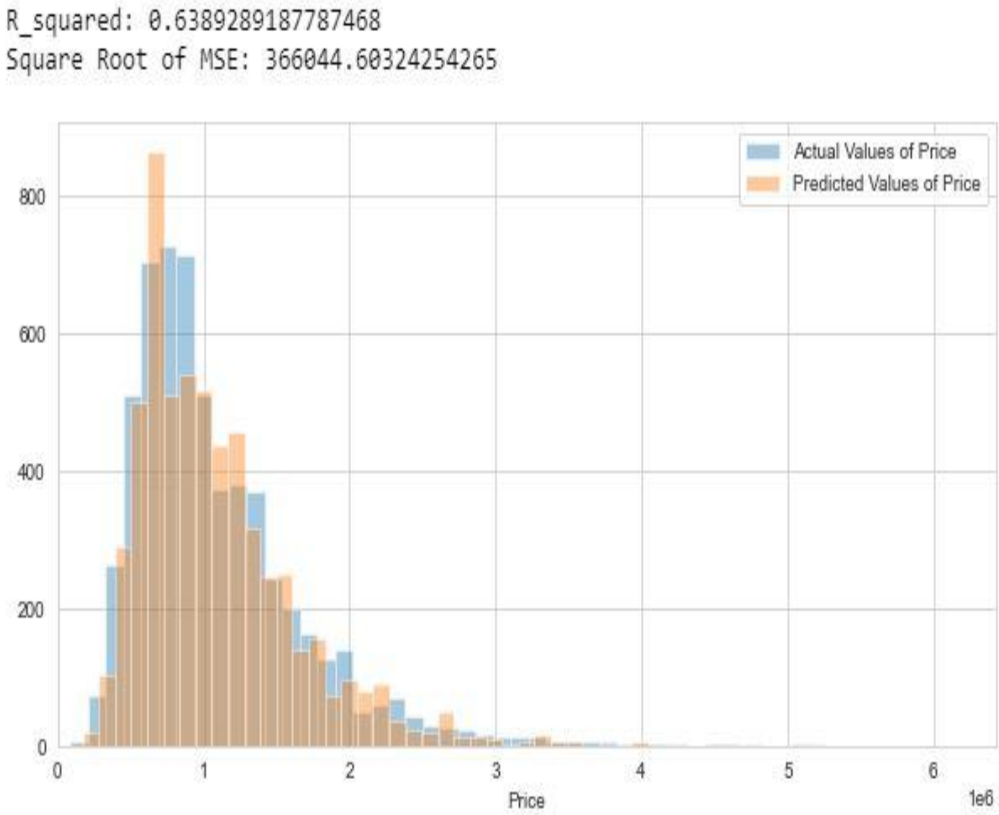
Ridge regression

# KNN and Decision TRee



```
R_squared: 0.6612281763949265
Square Root of MSE: 354561.27136320336
```

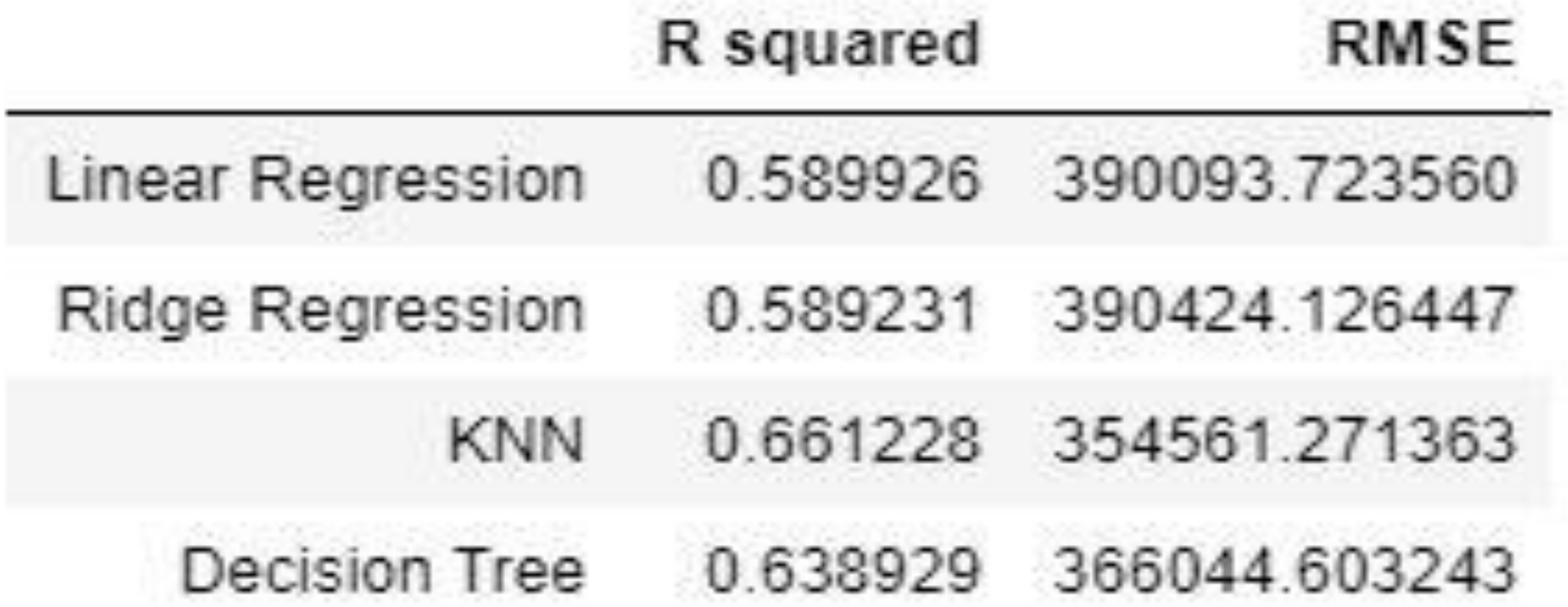
Knn



```
R_squared: 0.6389289187787468
Square Root of MSE: 366044.60324254265
```

Decision Tree

# Performance Summary



| | R squared | RMSE |
|---|---|---|
| Linear Regression | 0.589926 | 390093.723560 |
| Ridge Regression | 0.589231 | 390424.126447 |
| KNN | 0.661228 | 354561.271363 |
| Decision Tree | 0.638929 | 366044.603243 |

# Cross Validation and Grid Search

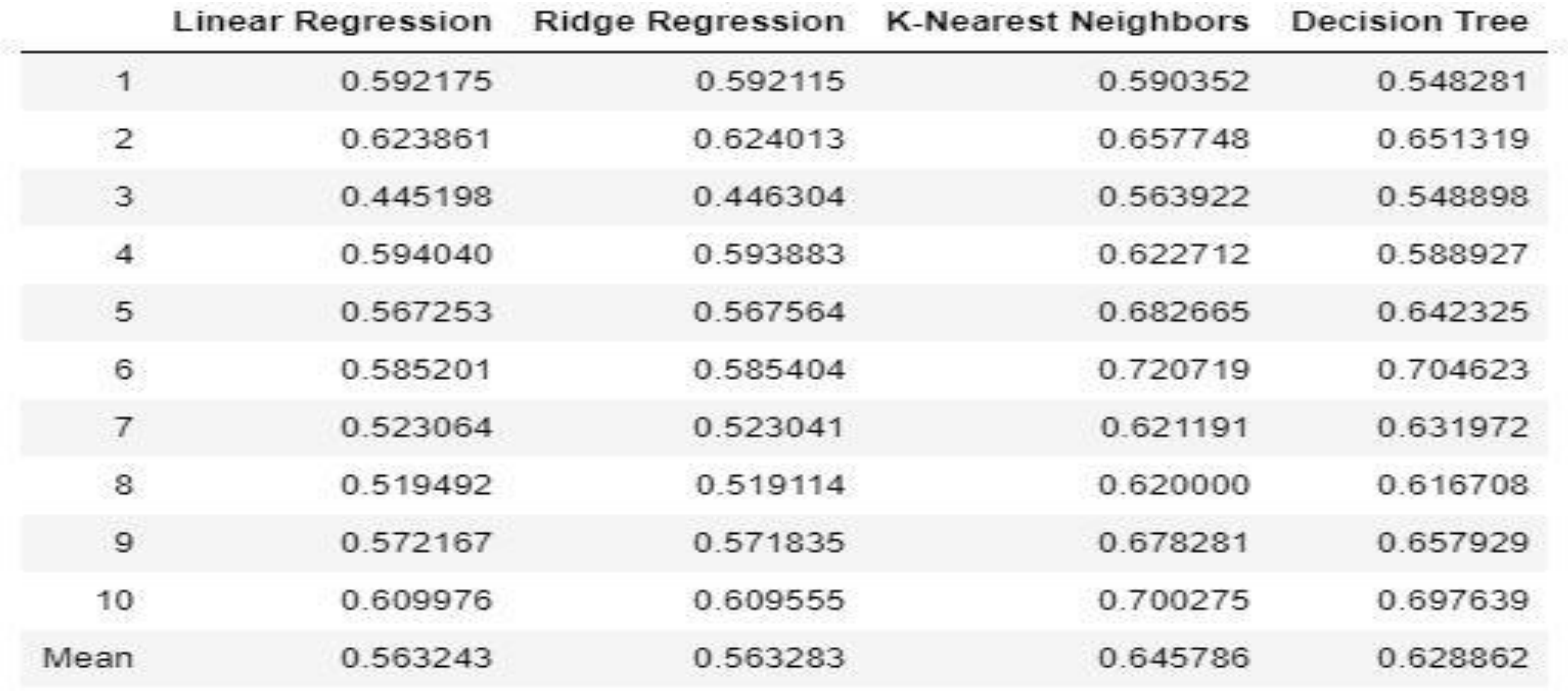
**Cross validation**

- Re-sampling procedure
- Data splits into k-folds
- Fit a model using (k-1) folds and validate the model using  the remaining fold
- Find the average of the score

**Grid Search**

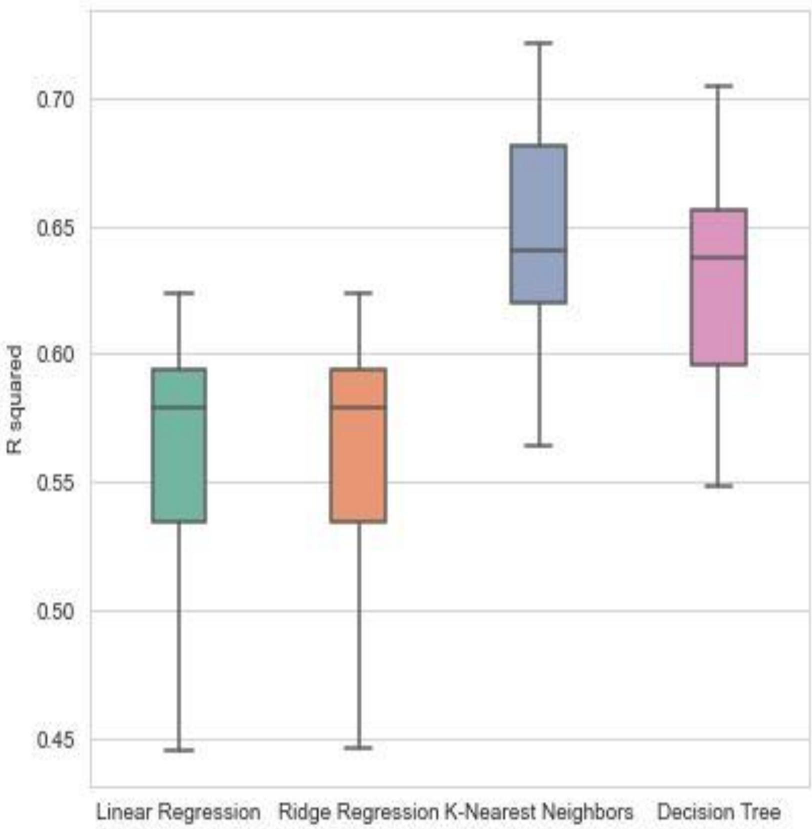
- Hyper parameters tunning to find the optimal values of the parameters for a model

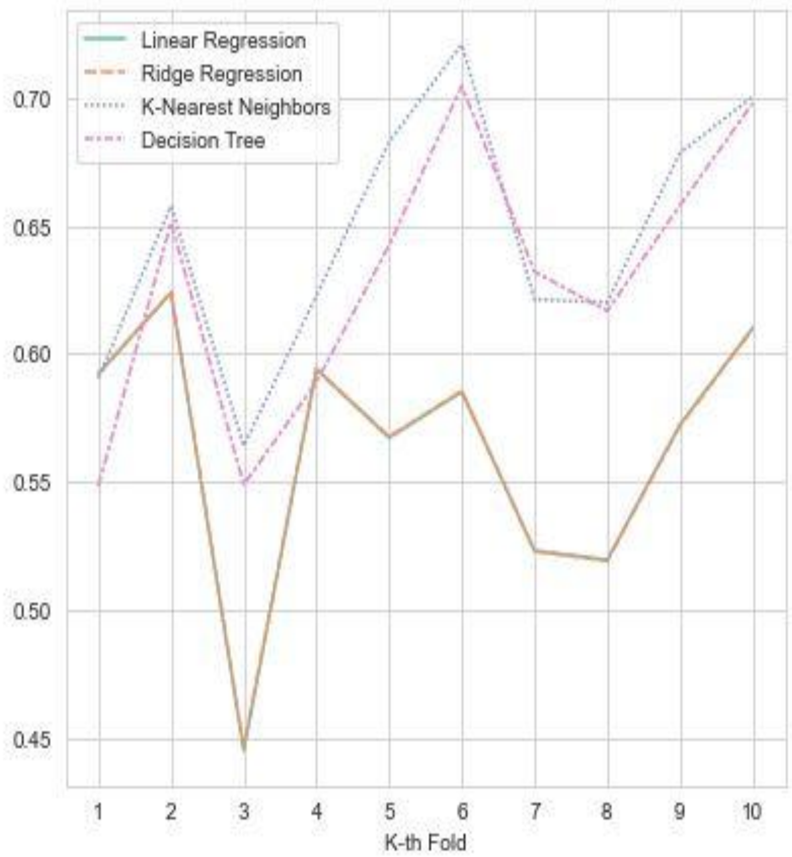
# Cross validation summary after parameter tuning



| | Linear Regression | Ridge Regression | K-Nearest Neighbors | Decision Tree |
|---|---|---|---|---|
| 1 | 0.592175 | 0.592115 | 0.590352 | 0.548281 |
| 2 | 0.623861 | 0.624013 | 0.657748 | 0.651319 |
| 3 | 0.445198 | 0.446304 | 0.563922 | 0.548898 |
| 4 | 0.594040 | 0.593883 | 0.622712 | 0.588927 |
| 5 | 0.567253 | 0.567564 | 0.682665 | 0.642325 |
| 6 | 0.585201 | 0.585404 | 0.720719 | 0.704623 |
| 7 | 0.523064 | 0.523041 | 0.621191 | 0.631972 |
| 8 | 0.519492 | 0.519114 | 0.620000 | 0.616708 |
| 9 | 0.572167 | 0.571835 | 0.678281 | 0.657929 |
| 10 | 0.609976 | 0.609555 | 0.700275 | 0.697639 |
| Mean | 0.563243 | 0.563283 | 0.645786 | 0.628862 |

# Boxplot and line plot for the perfomance





# The End