UNIVERSITY OF CALGARY

STAT 641/543 Group Project : GROUP 5

By Ethan Scott, [redacted]

UNIVERSITY OF CALGARY

Pricing wine batches from new wine vendor of Hotel X, using classification

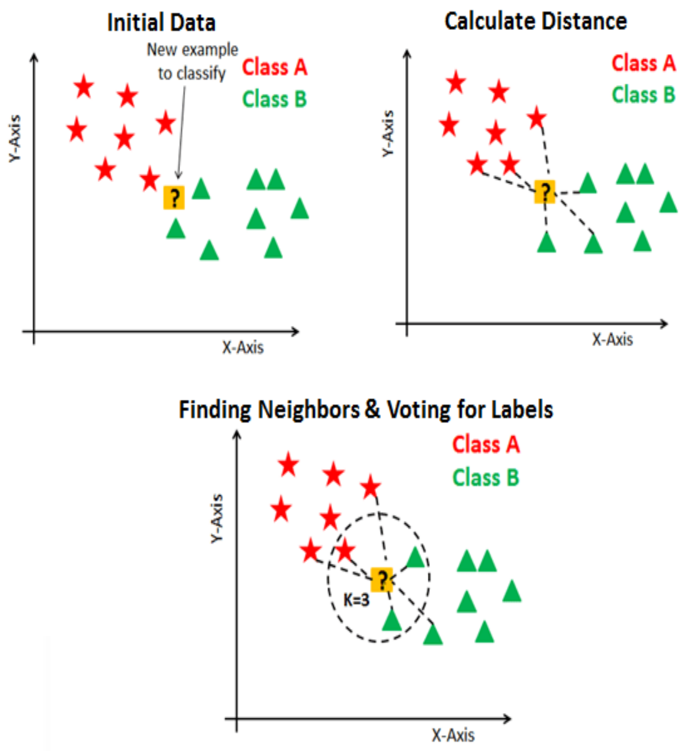
What is classification?

- It’s a predictive modeling method, where a class label (qualitative category) is predicted for a given input data X
- A classification model will use the training dataset and will determine how to best map input data into specific class labels or predict probability of class membership

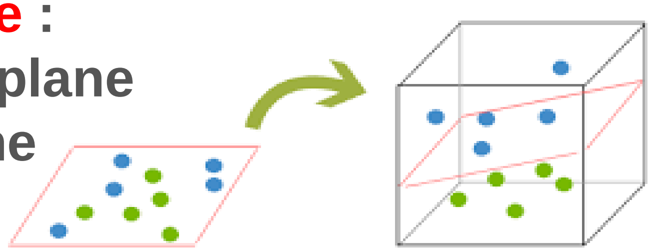
- Binary , Multi-Class , Multi-Label & Imbalanced classification are the main 4 categories of classification algorithms discussed in Machine Learning
- k-Nearest Neighbors, Decision Trees, Naive Bayes, Random Forest, Gradient Boosting & Support Vector Machines are some commonly used classification algorithms
- Some popular diagnostic for evaluating predicted class or class probabilities are confusion matric, Precision, Recall, F-Measure & ROC curve

- A few real-world applications of classification algorithms are, pattern recognition, fraud detection, credit scoring, anomaly detection

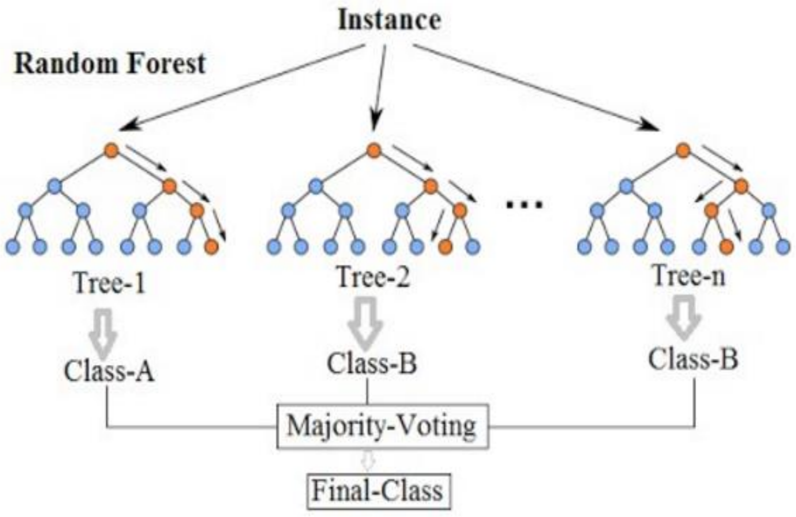
Overview : Classification methods



1. Support Vector Machine : Determine the best hyperplane which linearly separate the classes



2. Random Forest Model : Determine the best class suited, through a series of decision trees and majority voting system



3. K-NN Model : An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors

Overview : Evaluating classification methods

		Real Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

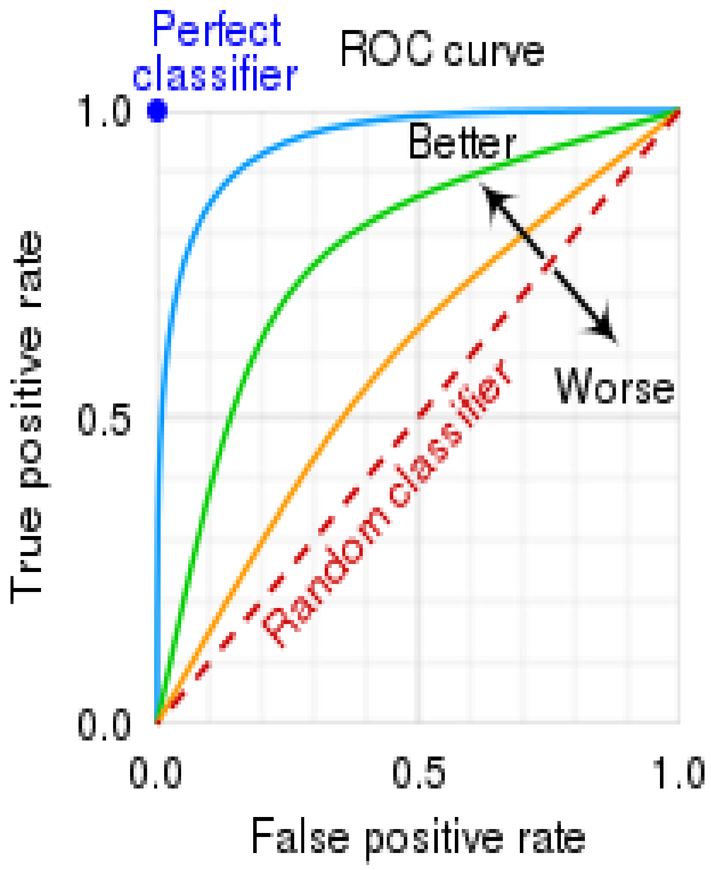
Accuracy : The fraction of correct classifications

Precision : fraction of relevant classes among the retrieved classes

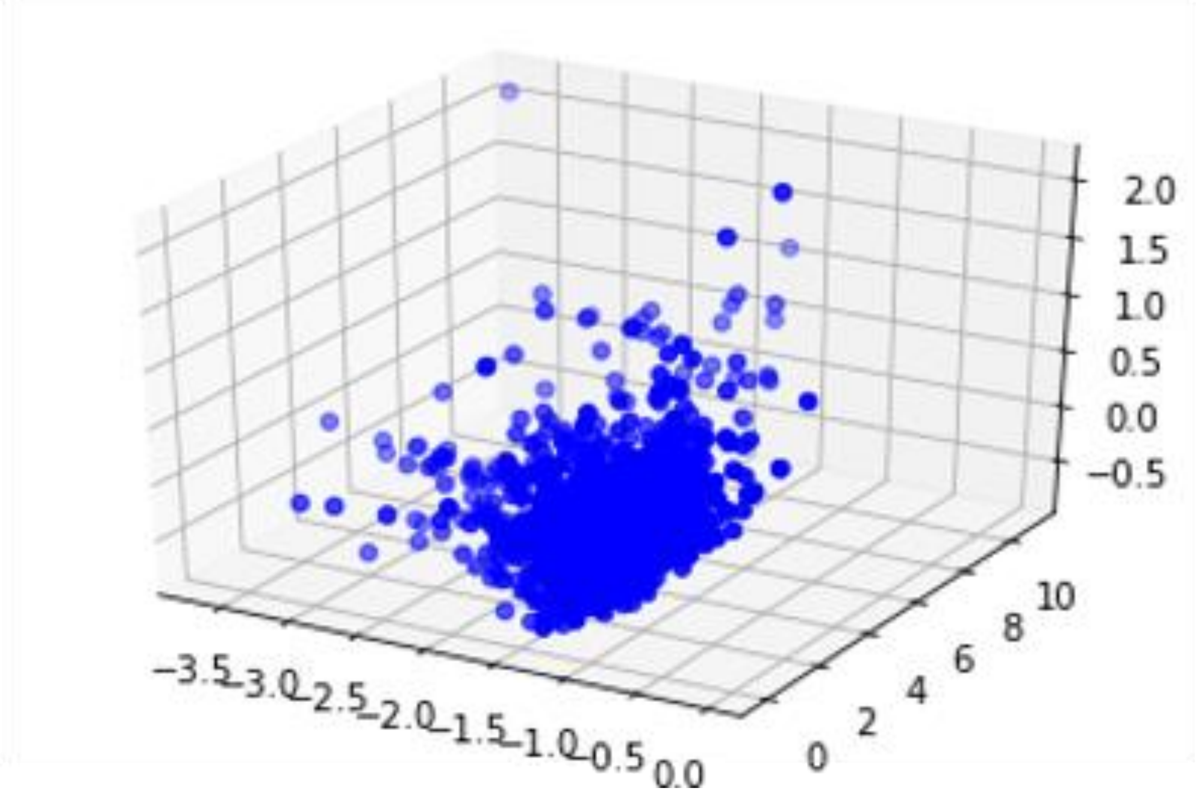
Recall: the fraction of the correctly classified classes, out of retrieved classes

F-measure : Harmonic mean of precision and recall

ROC Curve

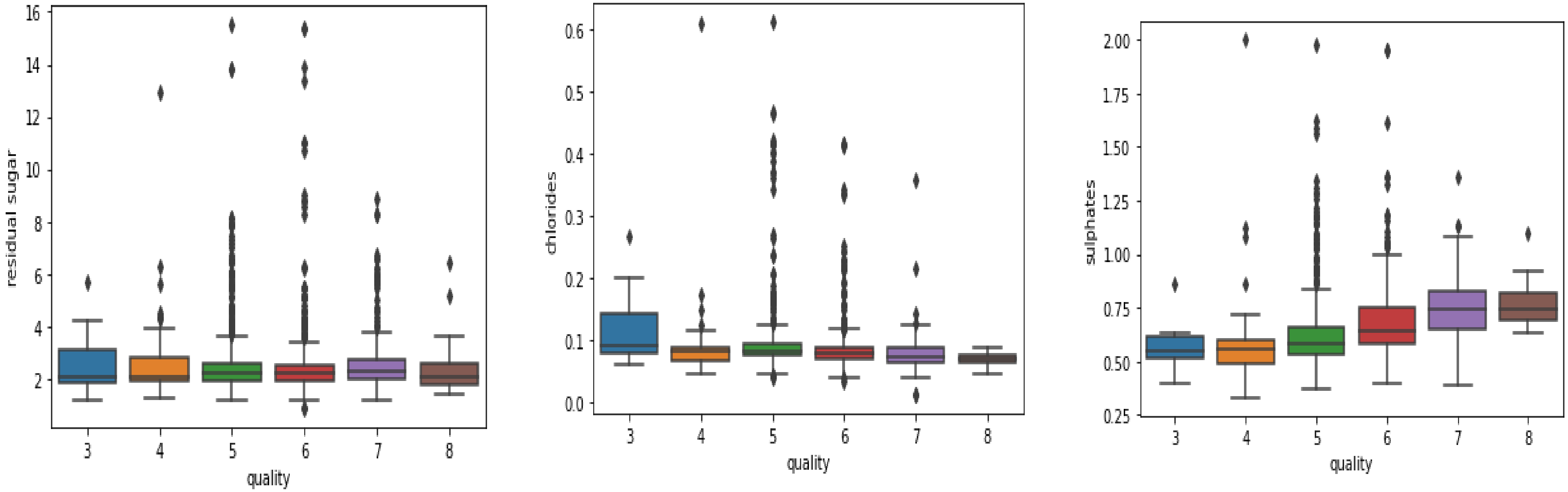


Overview of Data & the goal of analysis



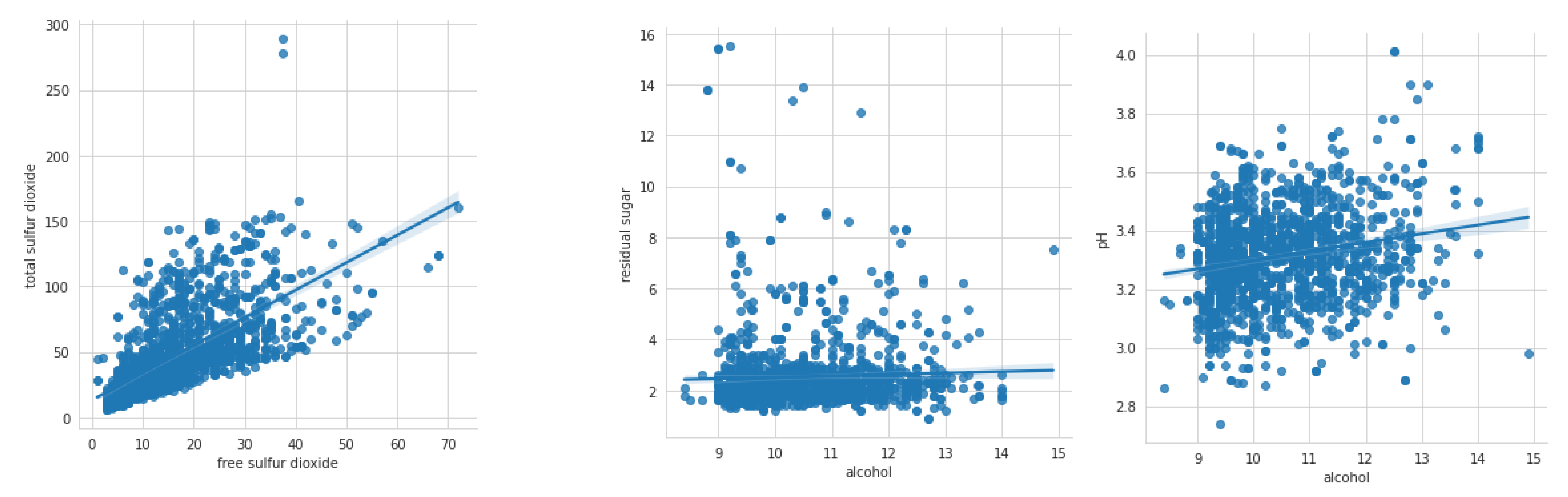
- Data Set : Wine Quality Data Set
- No of data : 1599
- No of features/predictors : 11 quantitative variables
- Response Variable : Quality of the wine (Class 3, 4, 5, 6, 7, 8)

The Hotel X, has recently changed their wine vendor and the management wants to price each batch of wine bottles delivered to them based on their own wine quality control process. For example, the wine batch which falls into class 8 (best quality) will be priced at $ 1000 per bottle and so on



Key Takeaways from the Descriptive Analysis

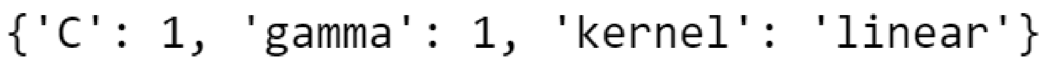
- When grouped by the quality of wine, the data set seems to have significant number of both mild and extreme outliers with respective to several features such as, residual sugar content, sulphate content, no of calories etc.



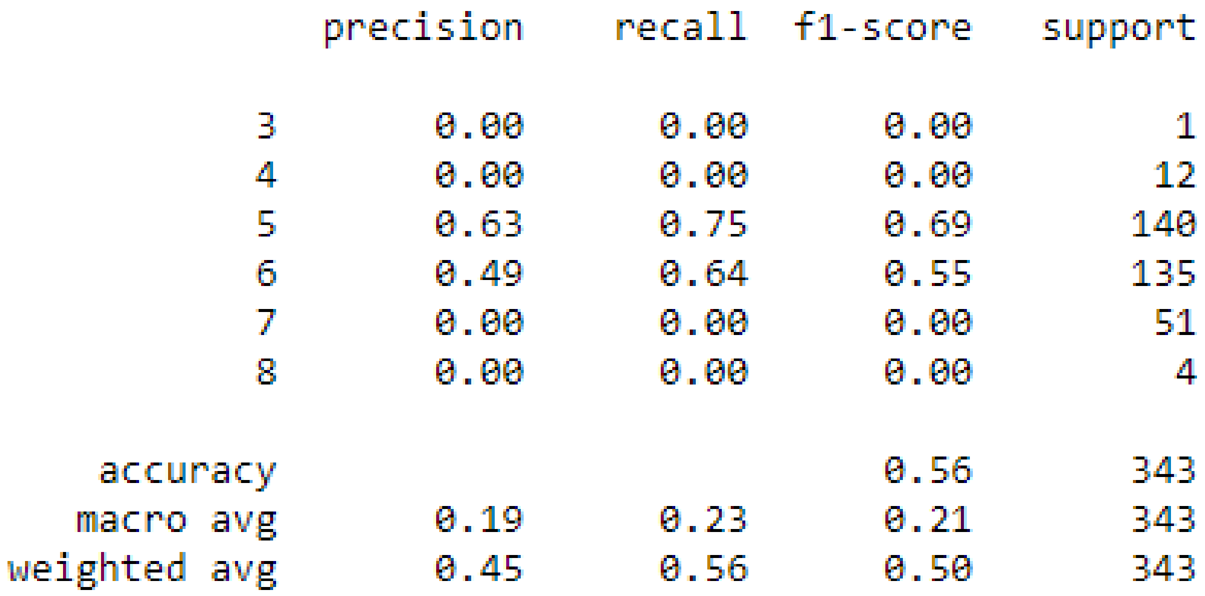
Key Takeaways from the Descriptive Analysis cont.

- The predictors total sulfur dioxide content and free sulfur content seems to have a noticeable correlation (r=0.6) as expected
- Remaining predictors are not significantly correlated

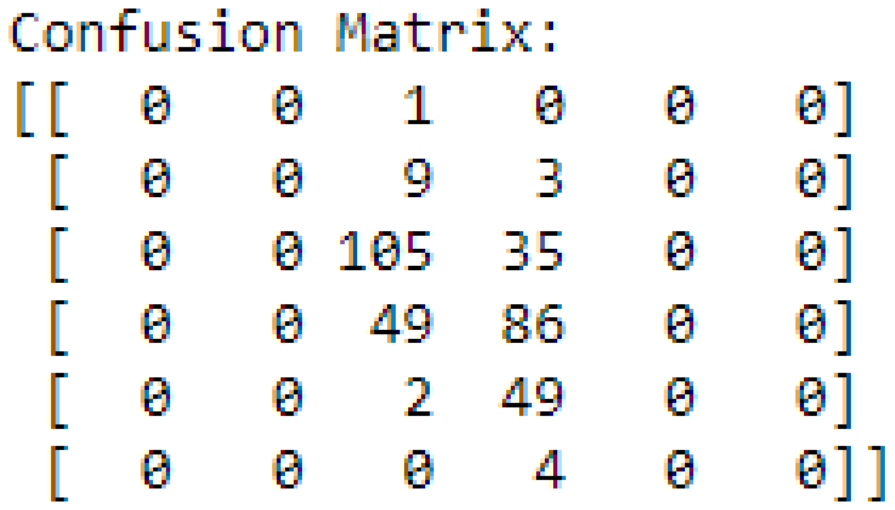
SVM Model

Best Hyperparameters: 

```
{'C': 1, 'gamma': 1, 'kernel': 'linear'}
```

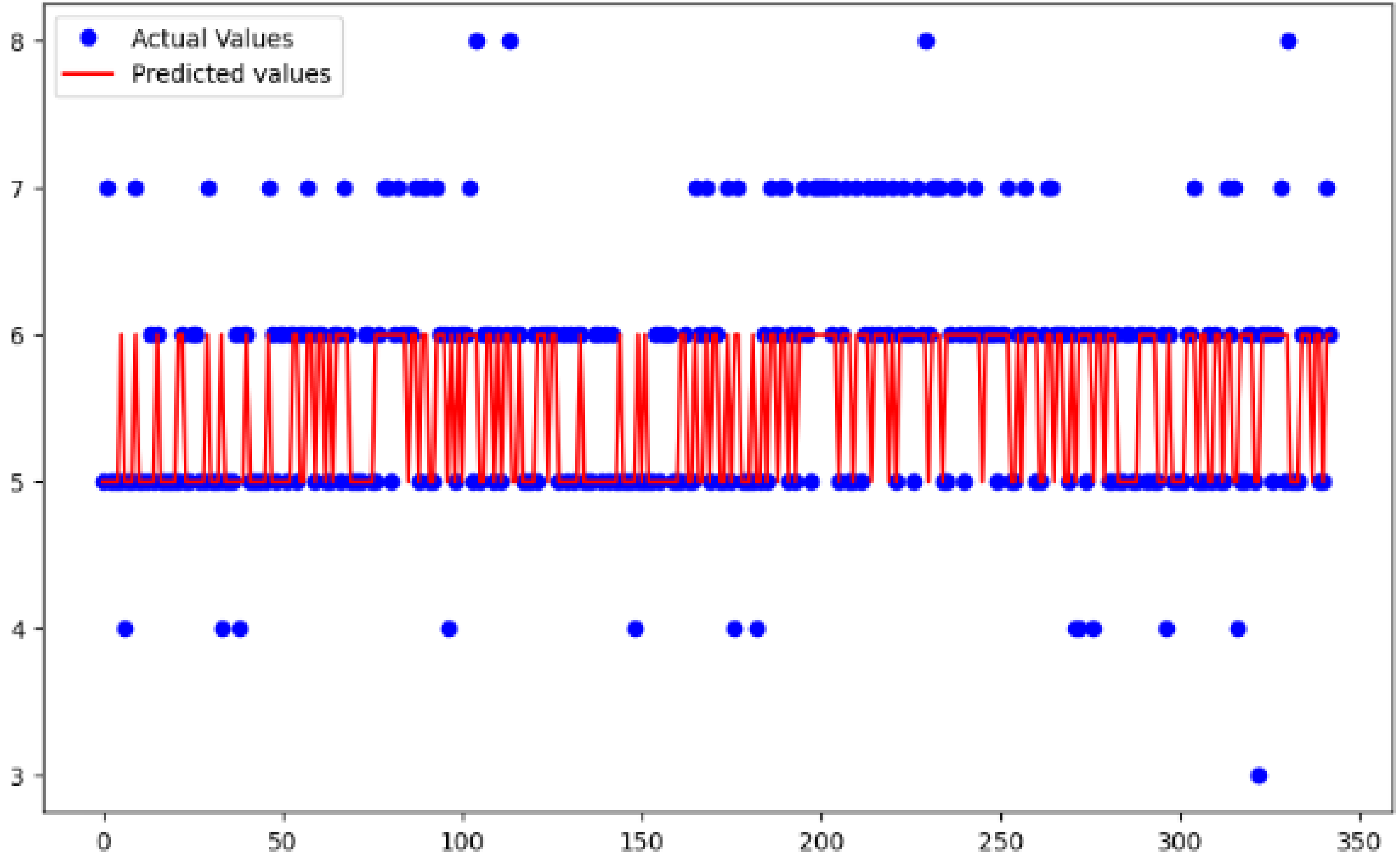


```
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.00      0.00      0.00        12
           5       0.63      0.75      0.69       140
           6       0.49      0.64      0.55       135
           7       0.00      0.00      0.00        51
           8       0.00      0.00      0.00         4

    accuracy                           0.56       343
   macro avg       0.19      0.23      0.21       343
weighted avg       0.45      0.56      0.50       343
```

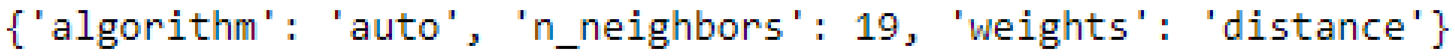


```
Confusion Matrix:
[[  0   0   1   0   0   0]
 [  0   0   9   3   0   0]
 [  0   0 105  35   0   0]
 [  0   0  49  86   0   0]
 [  0   0   2  49   0   0]
 [  0   0   0   4   0   0]]
```

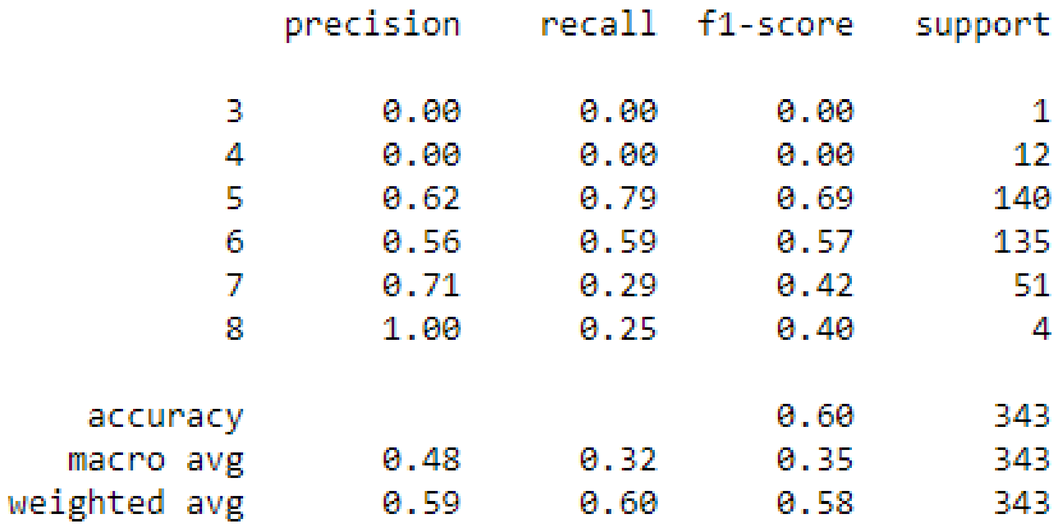
SVM Predicted Vs Actual Values



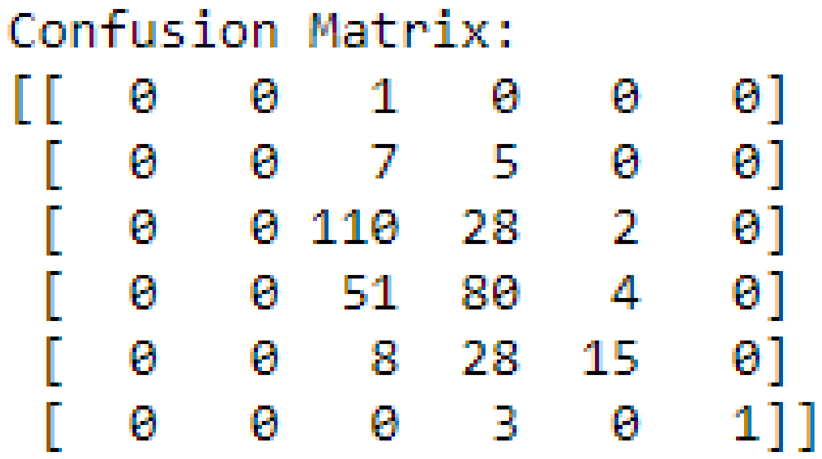
KNN Model

Best Hyperparameters: 

```
{'algorithm': 'auto', 'n_neighbors': 19, 'weights': 'distance'}
```

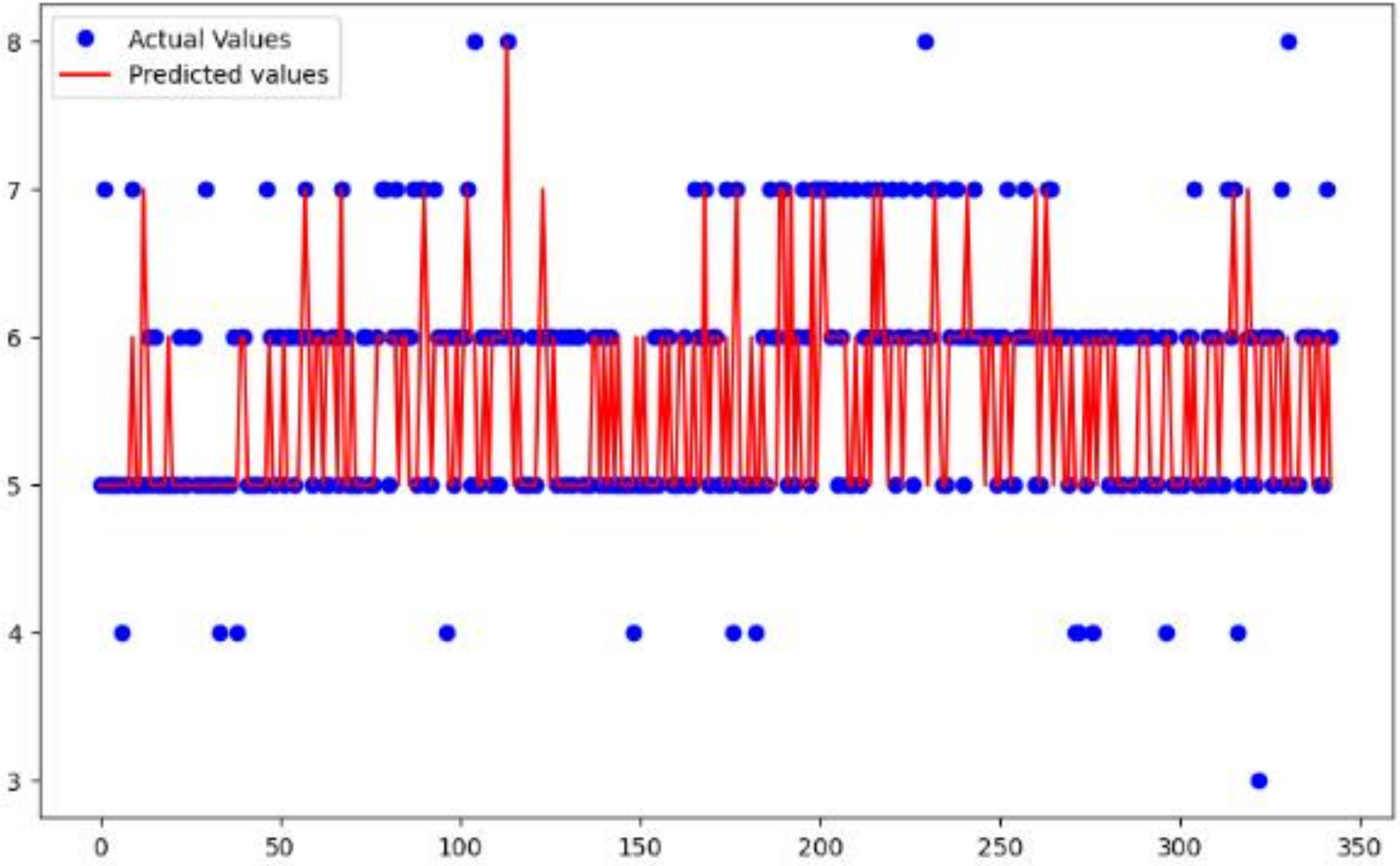


```
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.00      0.00      0.00        12
           5       0.62      0.79      0.69       140
           6       0.56      0.59      0.57       135
           7       0.71      0.29      0.42        51
           8       1.00      0.25      0.40         4

    accuracy                           0.60       343
   macro avg       0.48      0.32      0.35       343
weighted avg       0.59      0.60      0.58       343
```



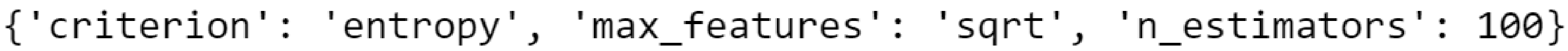
```
Confusion Matrix:
[[  0   0   1   0   0   0]
 [  0   0   7   5   0   0]
 [  0   0 110  28   2   0]
 [  0   0  51  80   4   0]
 [  0   0   8  28  15   0]
 [  0   0   0   3   0   1]]
```

KNN Predicted Vs Actual Values

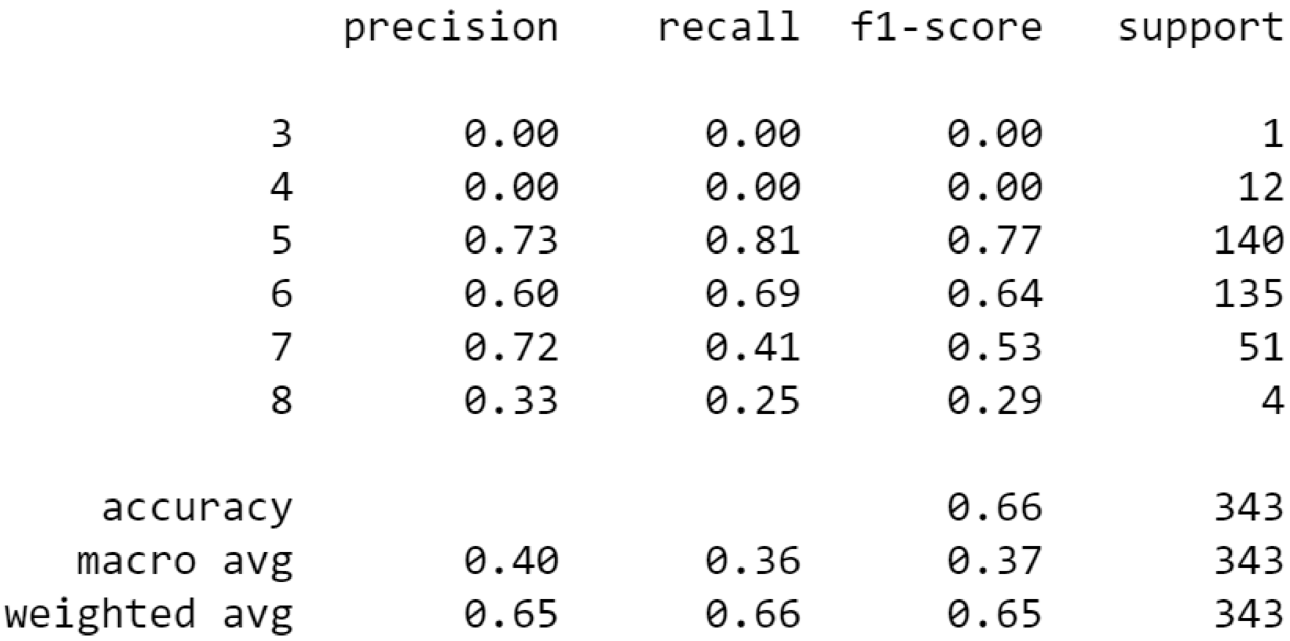


Random Forest Model

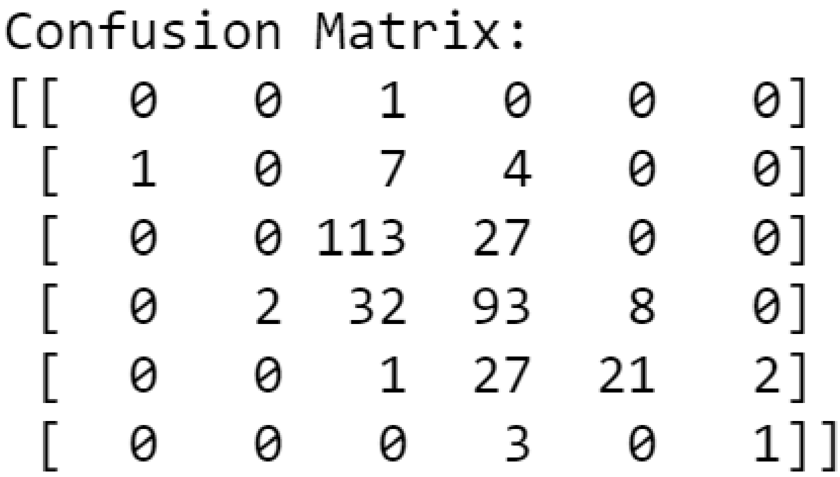
Best Hyperparameters:



```
{'criterion': 'entropy', 'max_features': 'sqrt', 'n_estimators': 100}
```

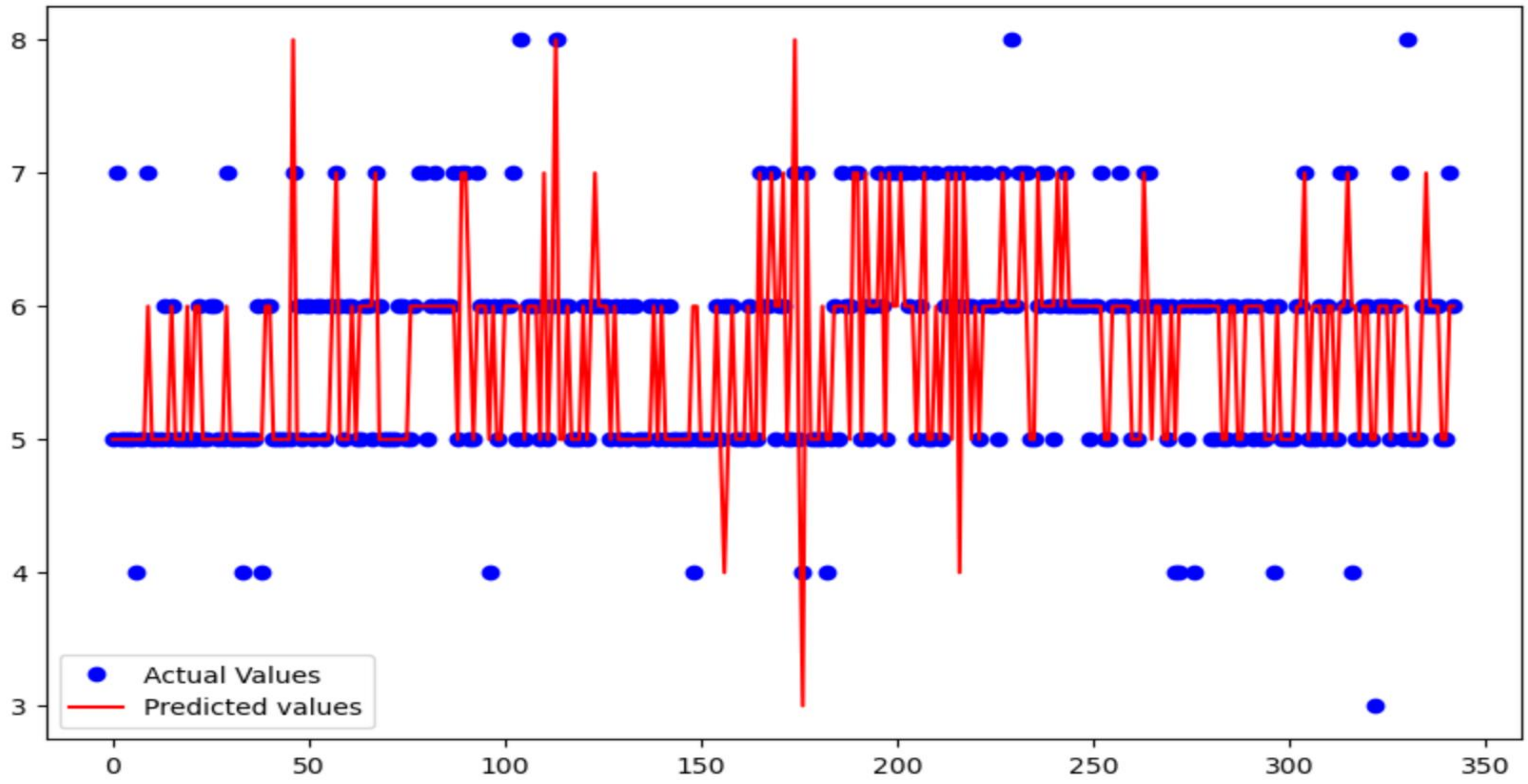


```
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.00      0.00      0.00        12
           5       0.73      0.81      0.77       140
           6       0.60      0.69      0.64       135
           7       0.72      0.41      0.53        51
           8       0.33      0.25      0.29         4

    accuracy                           0.66       343
   macro avg       0.40      0.36      0.37       343
weighted avg       0.65      0.66      0.65       343
```



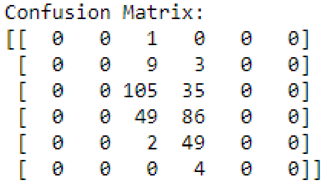
```
Confusion Matrix:
[[  0   0   1   0   0   0]
 [  1   0   7   4   0   0]
 [  0   0 113  27   0   0]
 [  0   2  32  93   8   0]
 [  0   0   1  27  21   2]
 [  0   0   0   3   0   1]]
```

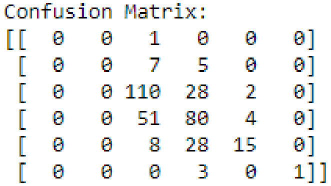
Random Forest Predicted Vs Actual Values

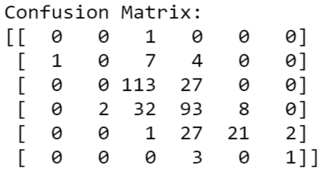


Summary of the model fitting

Classification Method	Predicted no of classes	Prediction accuracy	Confusion matrix	F- measure (Weighted Average)	Precision (Weighted Average)	Recall (Weighted Average)

SVM	**2**	**0.56**	
			`Confusion Matrix:`
			`[[  0   0   1   0   0   0]`
			`[  0   0   9   3   0   0]`
			`[  0   0 105  35   0   0]`
			`[  0   0  49  86   0   0]`
			`[  0   0   2  49   0   0]`
			`[  0   0   0   4   0   0]]`
				0.50	**0.45**	**0.56**

KNN	**4**	**0.60**	
			`Confusion Matrix:`
			`[[  0   0   1   0   0   0]`
			`[  0   0   7   5   0   0]`
			`[  0   0 110  28   2   0]`
			`[  0   0  51  80   4   0]`
			`[  0   0   8  28  15   0]`
			`[  0   0   0   3   0   1]]`
				0.58	**0.59**	**0.60**

Random Forest	**6**	**0.66**	
			`Confusion Matrix:`
			`[[  0   0   1   0   0   0]`
			`[  1   0   7   4   0   0]`
			`[  0   0 113  27   0   0]`
			`[  0   2  32  93   8   0]`
			`[  0   0   1  27  21   2]`
			`[  0   0   0   3   0   1]]`
				0.65	**0.65**	**0.66**

Conclusion

- The Random Forest Model outperformed the other models in all methods of evaluation and is the best model to classify the wine quality

UNIVERSITY OF CALGARY

The Maximum Allowable Mortgage Loan for the Prospective Customer in Melbourne, using Ridge Regression, Lasso and Random Forest

Introduction to the question

- Prior approach when lending a mortgage loan is, to request employment information with last six months' pay stubs.
- Due to the economic crisis default risk increased
- Implemented a new plan to estimate the house price according

  debtor’s requirements.
- Maximum allowable mortgage plan is 70% of the predicted house price

Our objectives are

To answer the previous question according to the below criterias

- Identify the important features effecting the house price using an exploratory data analysis.

- Predict the house prices using statistical learning techniques

About the dataset

- Data source: **www.kaggle.com**
- **34,857 Observations**
- **21 variables**

Categorical - 14

Quantitative – 7

Descriptive analysis

- Response variable : Price
- Predictor variables :

Location Based

1. Type
2. Address
3. Suburb
4. Post code
5. Property Count
6. Distance
7. Council Area
8. Region Name
9. Latitude
10. Longitude

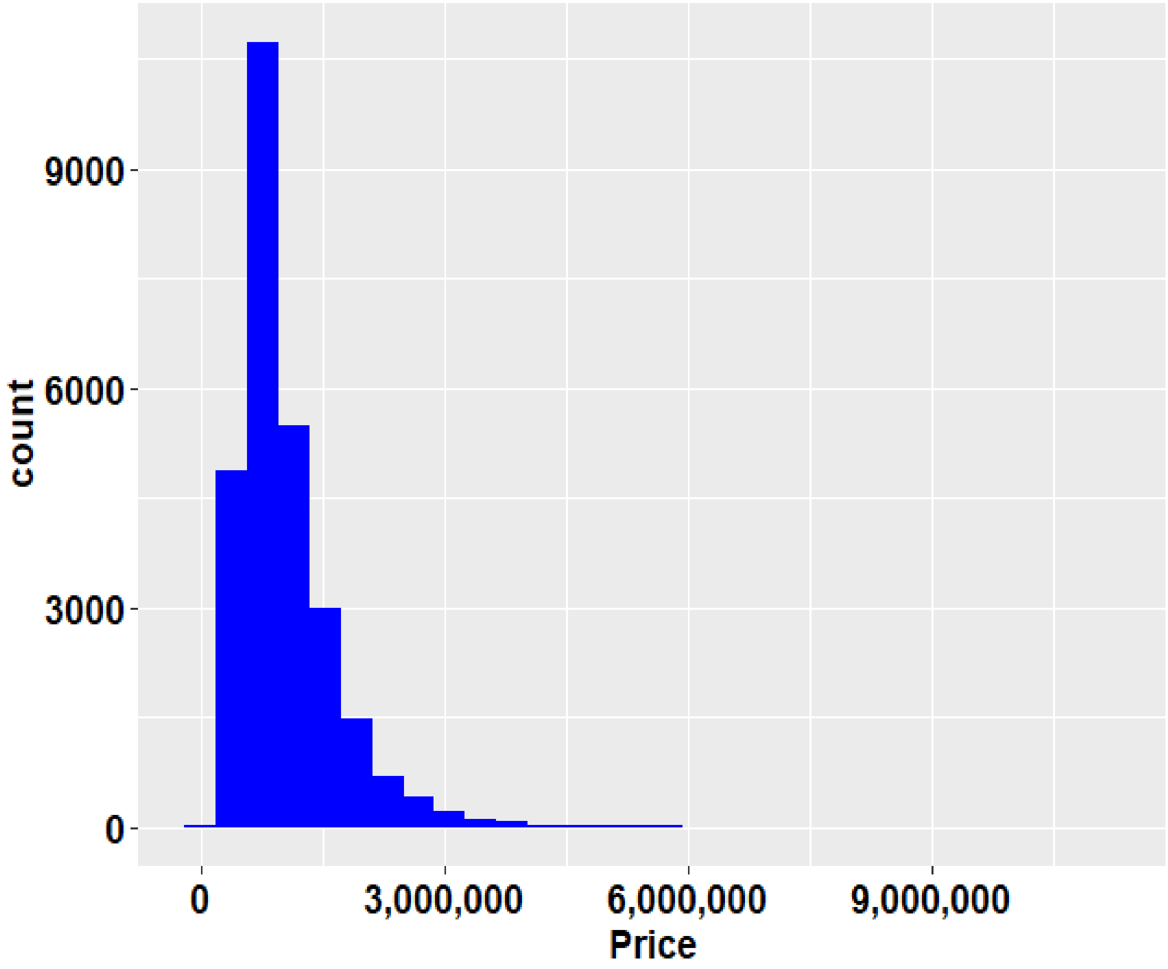
House Related

1. Rooms
2. Bedroom2
3. Bathroom
4. Car
5. Building Area
6. Landsize
7. Year Built

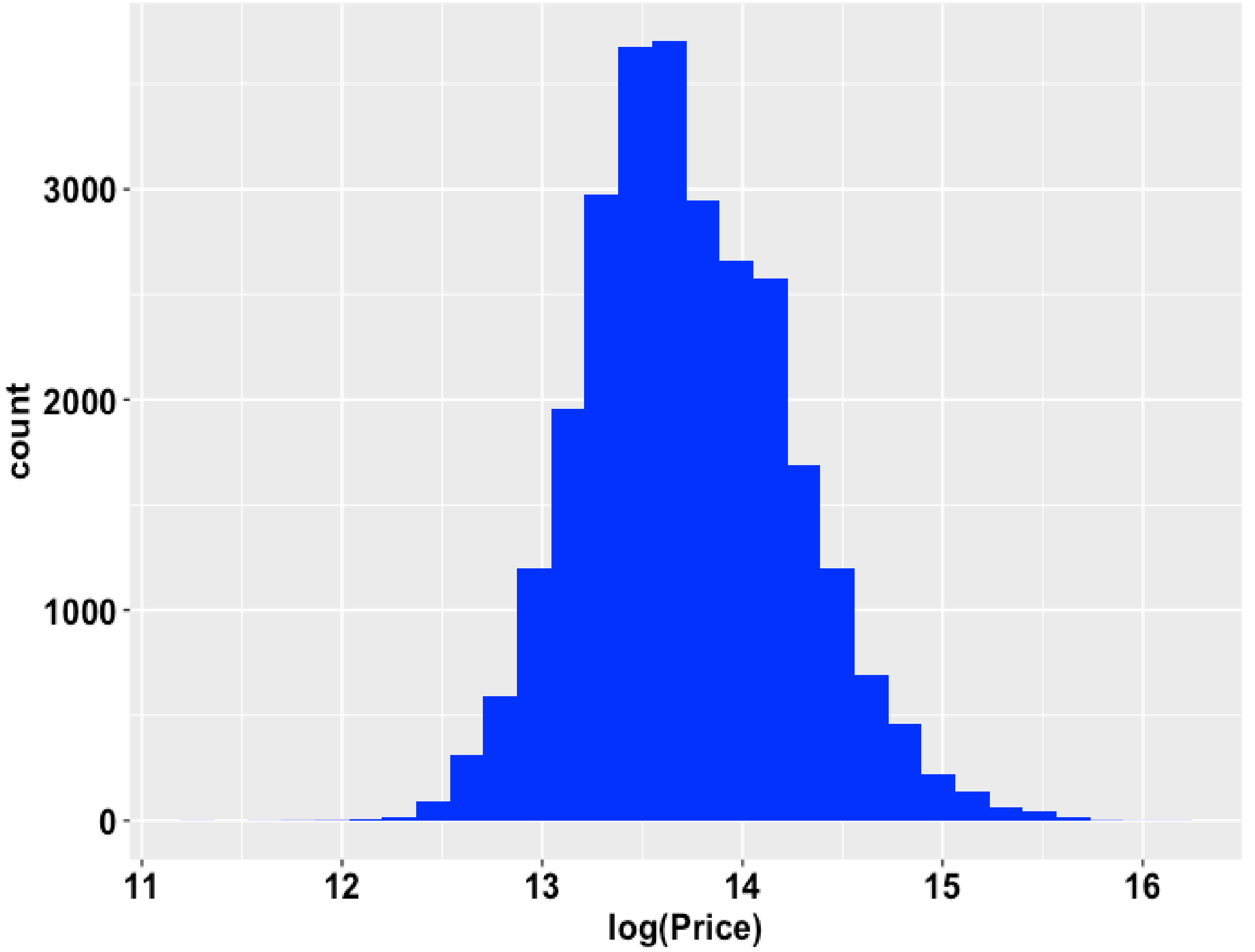
Seller Related

1. SellerG
2. Method
3. Date

Distribution of response variable

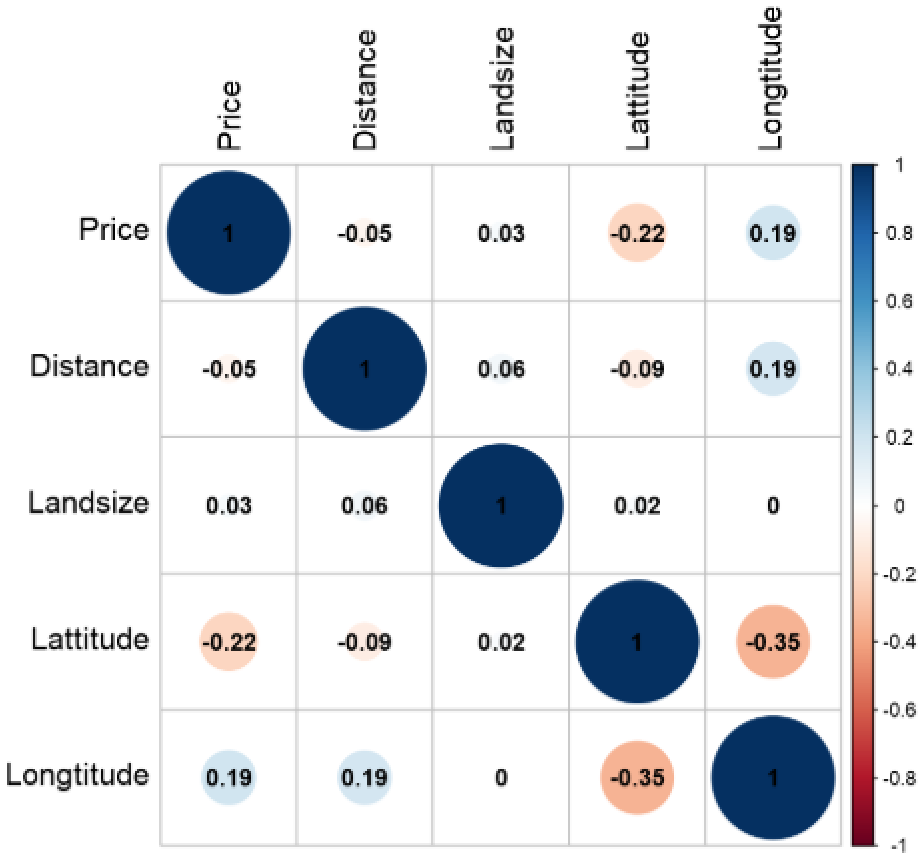


Distribution of the response – “Price”



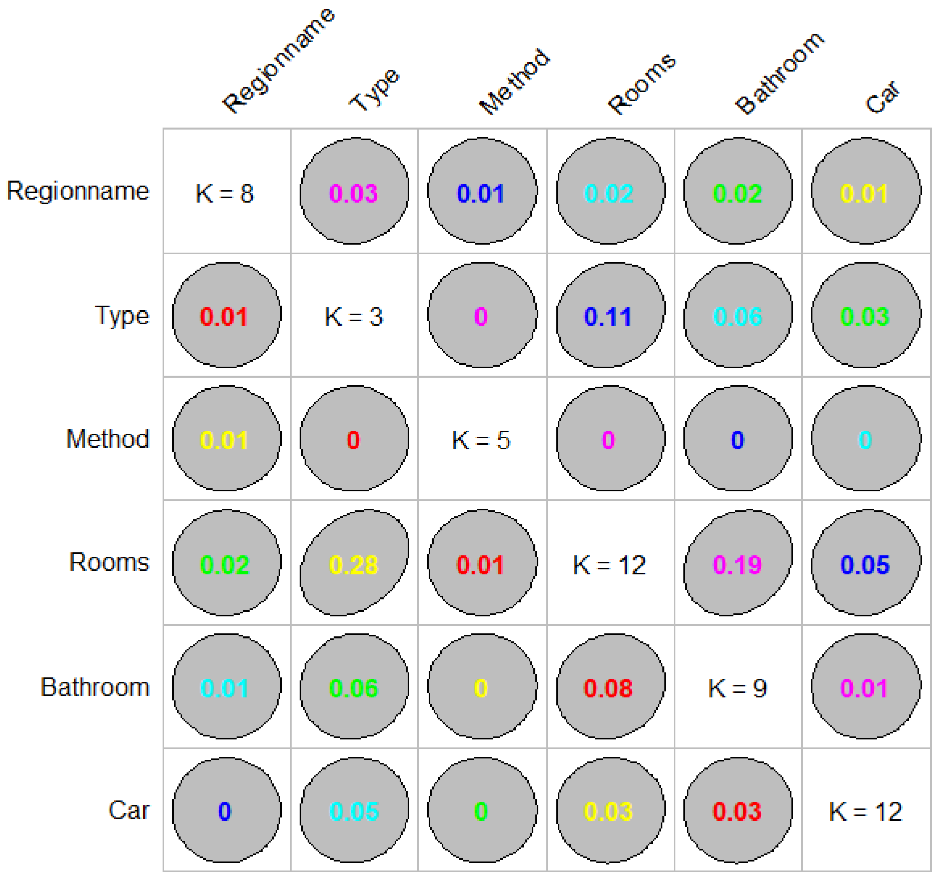
Distribution of the response – “log(Price)”

Correlation Plots



	Price	Distance	Landsize	Lattitude	Longtitude
Price	1	-0.05	0.03	-0.22	0.19
Distance	-0.05	1	0.06	-0.09	0.19
Landsize	0.03	0.06	1	0.02	0
Lattitude	-0.22	-0.09	0.02	1	-0.35
Longtitude	0.19	0.19	0	-0.35	1

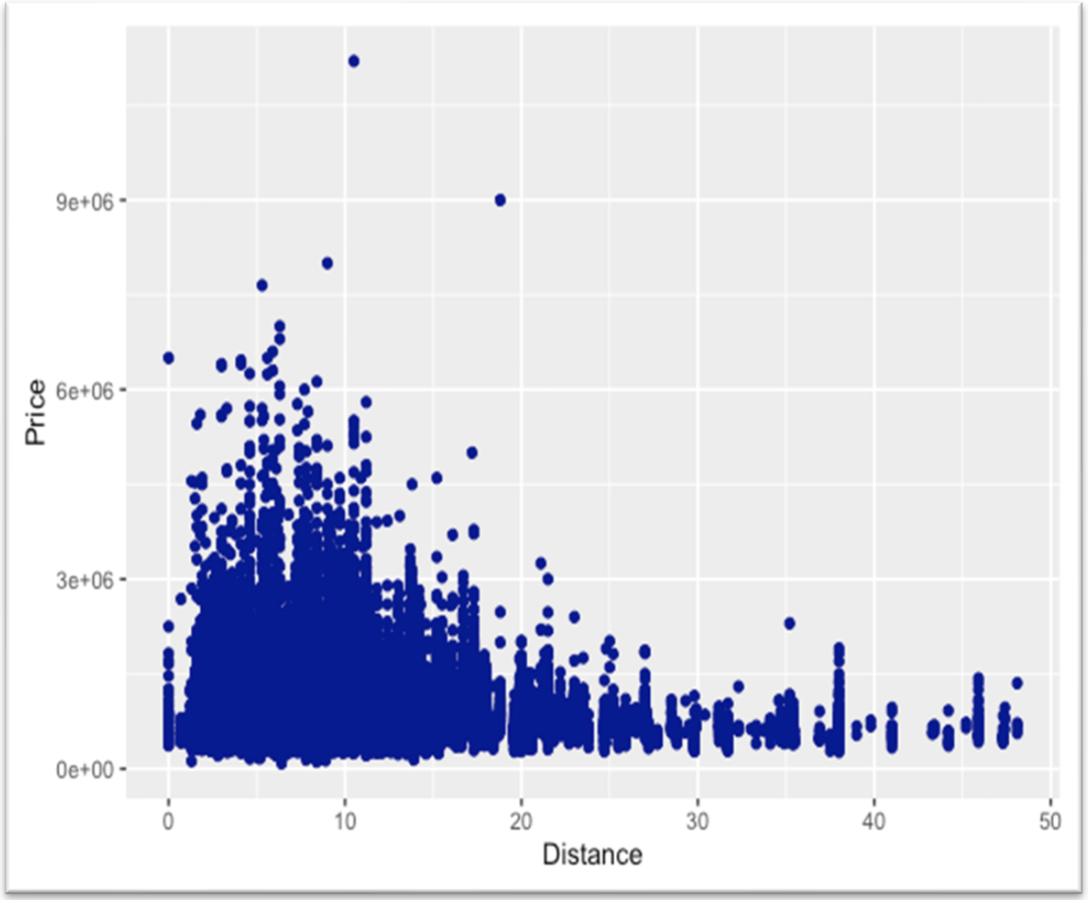
Pearson's correlation plot



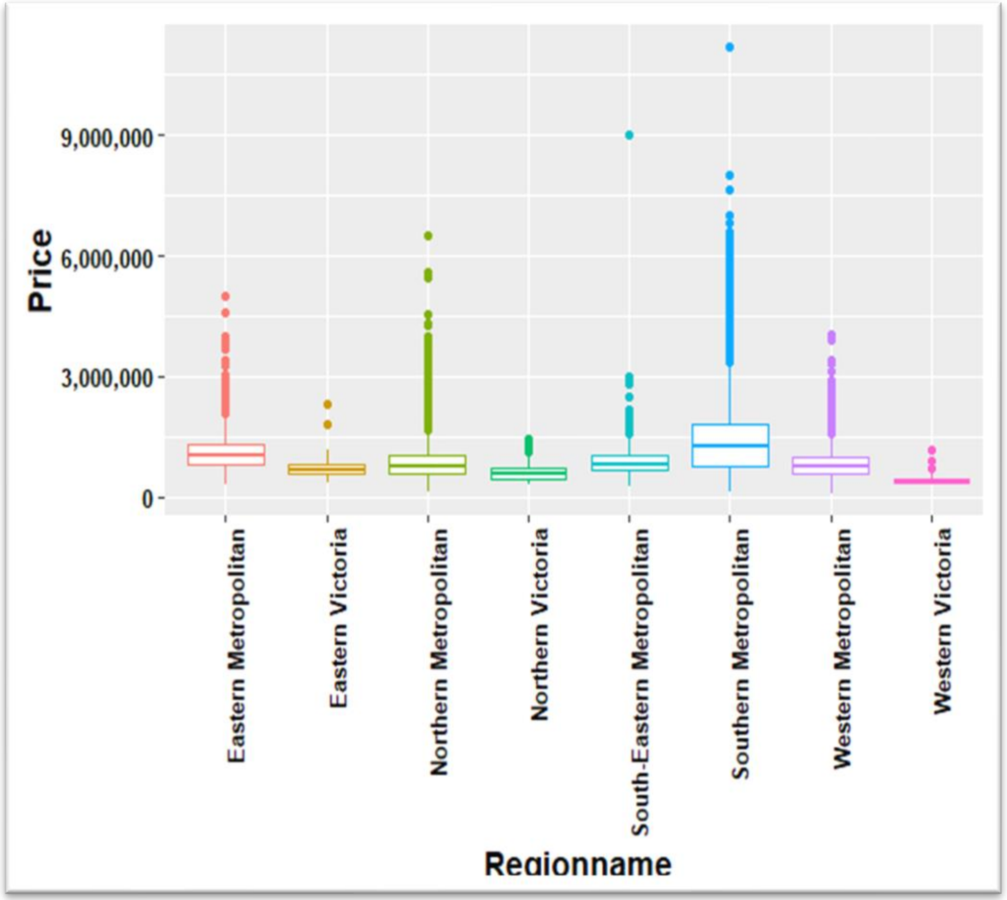
	Regionname	Type	Method	Rooms	Bathroom	Car
Regionname	K = 8	0.03	0.01	0.02	0.02	0.01
Type	0.01	K = 3	0	0.11	0.06	0.03
Method	0.01	0	K = 5	0	0	0
Rooms	0.02	0.28	0.01	K = 12	0.19	0.05
Bathroom	0.01	0.06	0	0.08	K = 9	0.01
Car	0	0.05	0	0.03	0.03	K = 12

Goodman – Kruskal plot

Distribution of important explanatory variables vs price

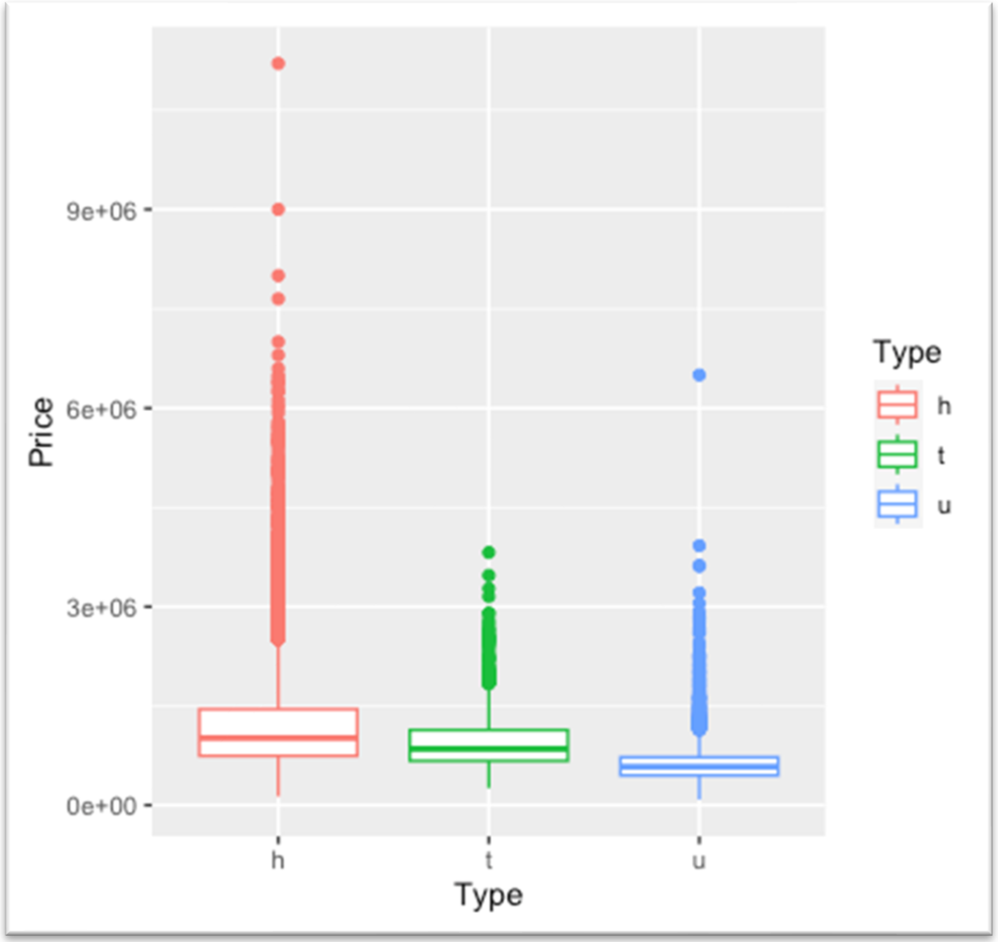


Distance vs Price

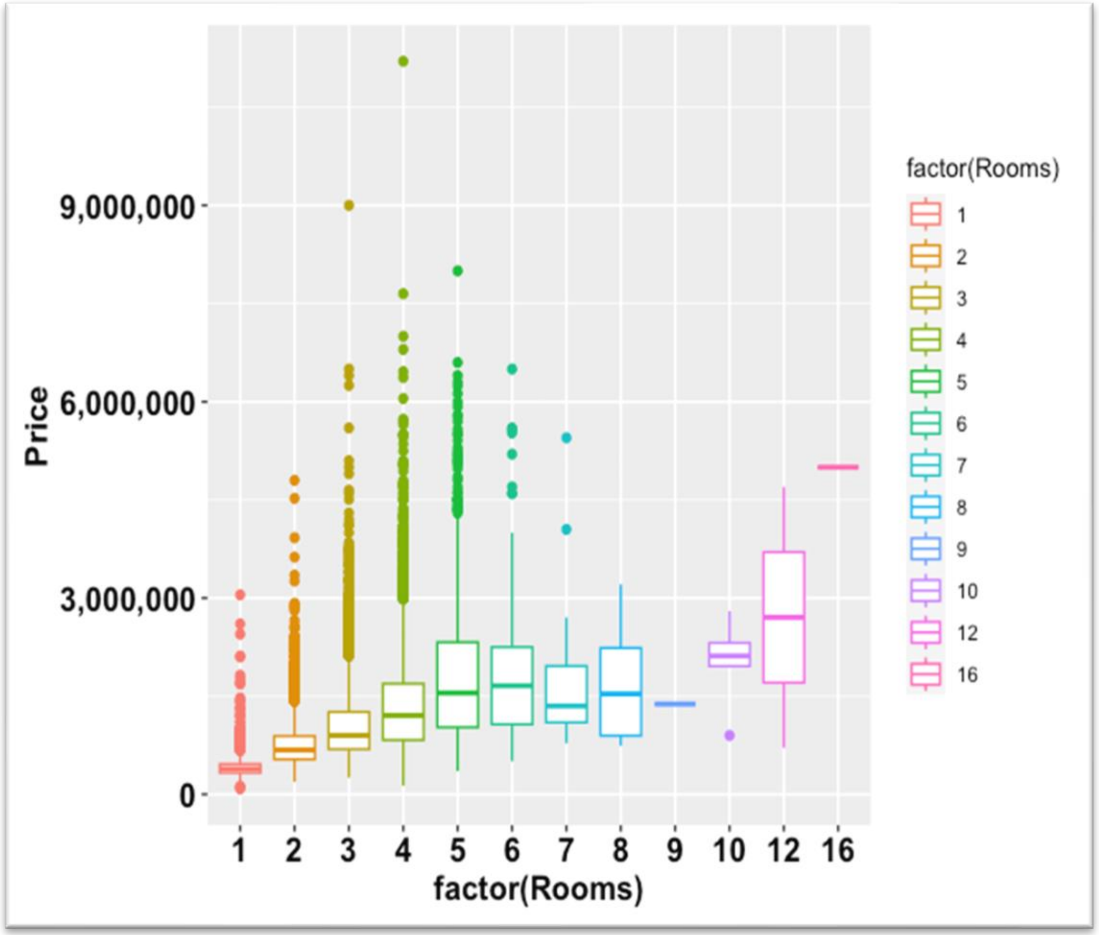


Region name vs price

Distribution of important explanatory variables vs price

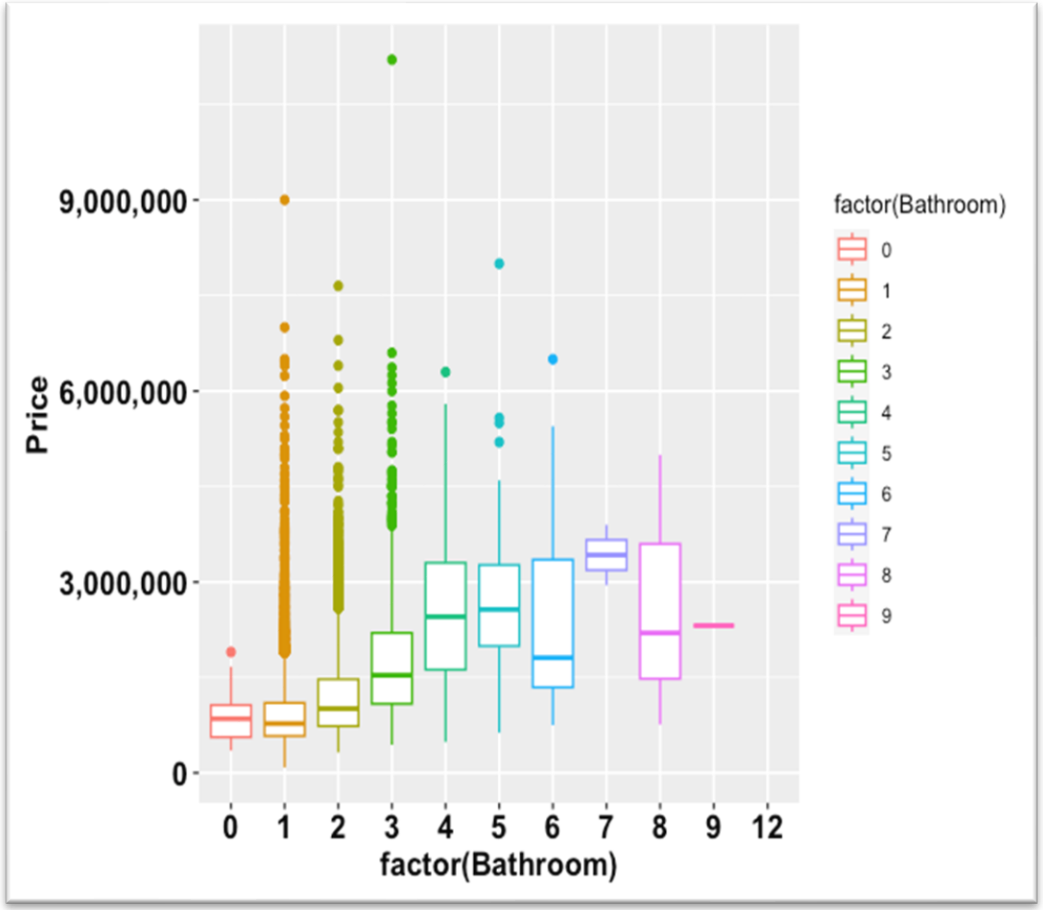


Type vs Price

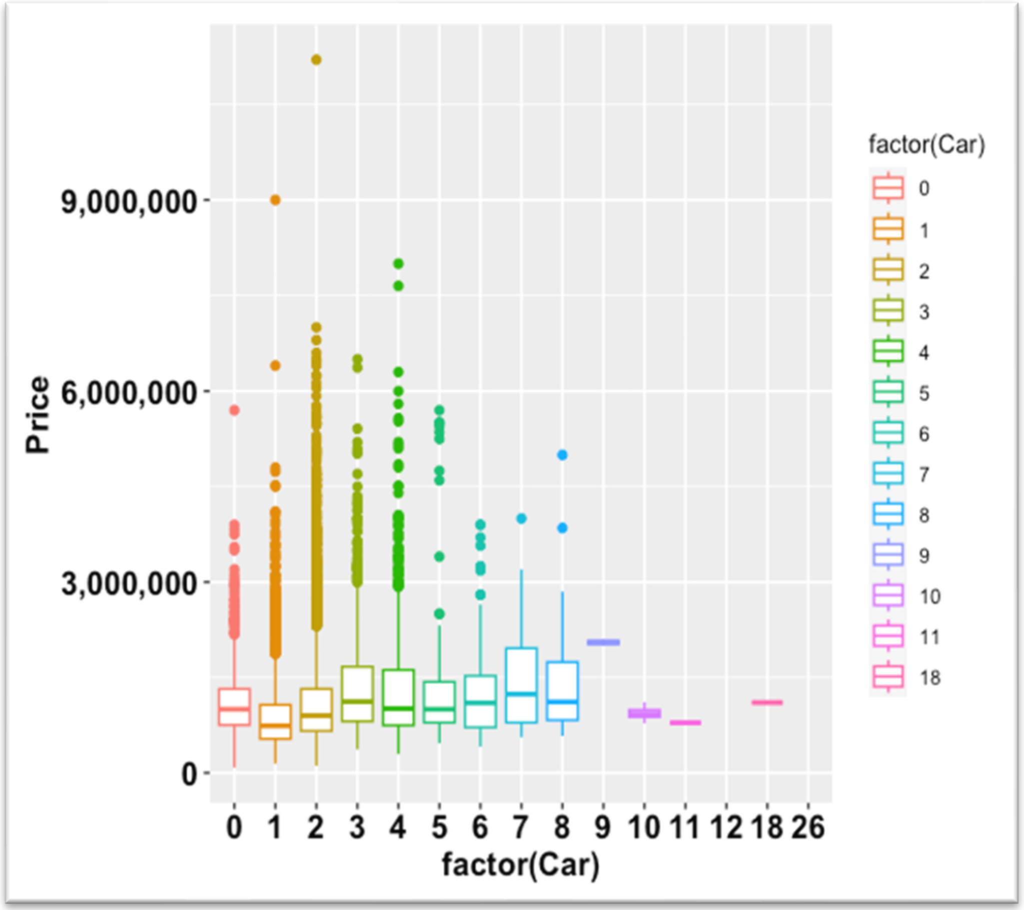


Number of Rooms vs Price

Distribution of important explanatory variables vs price

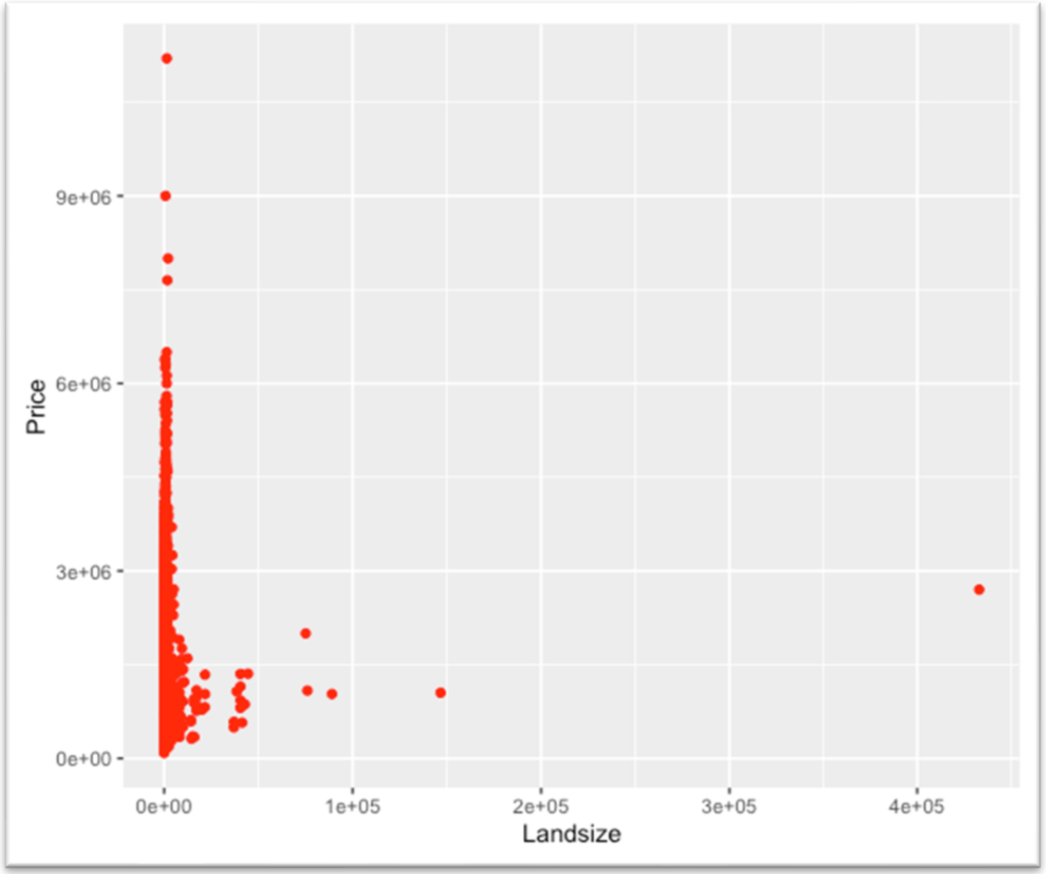


Number of Bathrooms vs Price

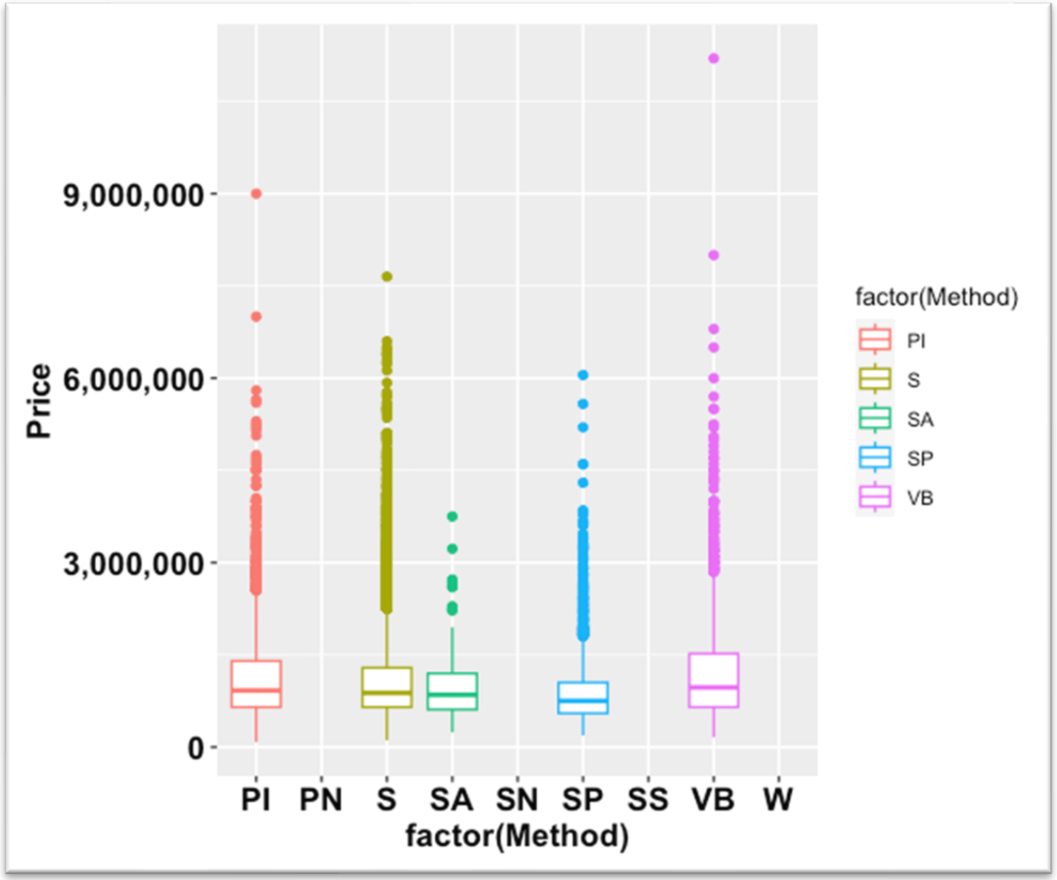


Number of car spots vs Price

Distribution of important explanatory variables vs price



Landsize vs Price



Method vs Price

Summary of the Descriptive analysis

UNIVERSITY OF CALGARY

- **Following variables are chosen for the Advanced analysis**

  - Type
  - Distance
  - Region Name
  - Rooms
  - Bathroom
  - Car
  - Land Size
  - Method

- **Due to the multicollinearity present among the explanatory variables, following statistical methods have used to develop the model**

- Ridge
- Lasso
- Random Forest

Advanced Analysis

(1) Data cleaning process

- Since “Price” is the response variable, records that are having missing values in the price column were removed. **(7,610 records)**
- Log e value of the Price variable was considered.
- There was one duplicate record - removed
- Distance, Bathroom, and Car variables had missing values – imputed with the mode.
- Landsize variable imputed using MICE package.

✓**Finally ended up with 8 predictor variables and 27,246 observations**

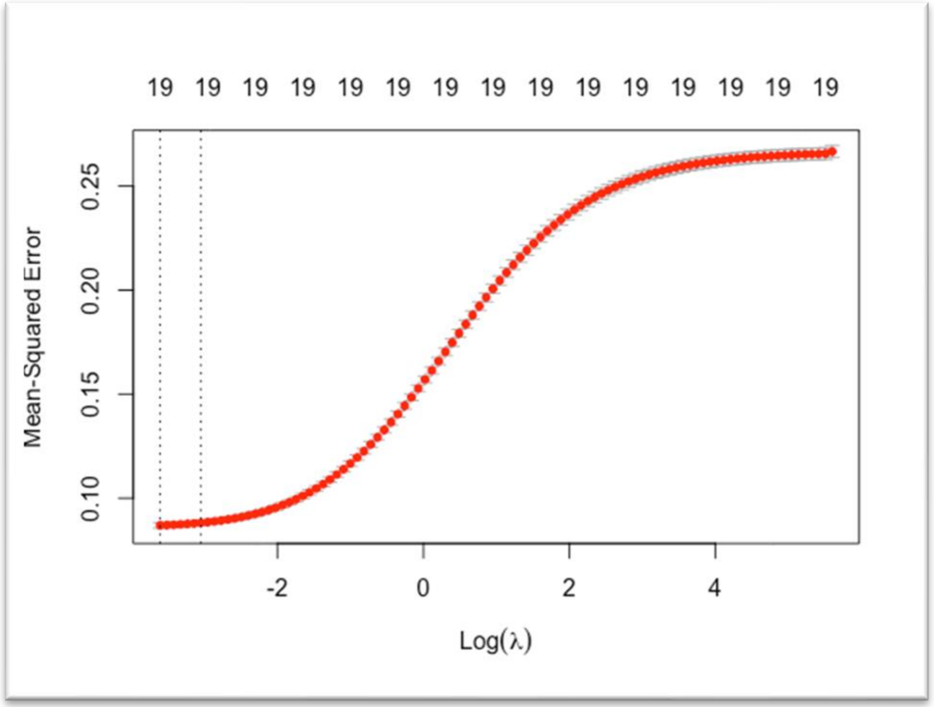
Advanced Analysis continued....

(2) Introduction to model fitting

- **Ridge Regression**
  - A method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated.
- **Lasso Regression**
  - Regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.
- **Random Forest**
  - A random forest, selects observations and specific variables to build multiple decision trees from the input and then averages the results.

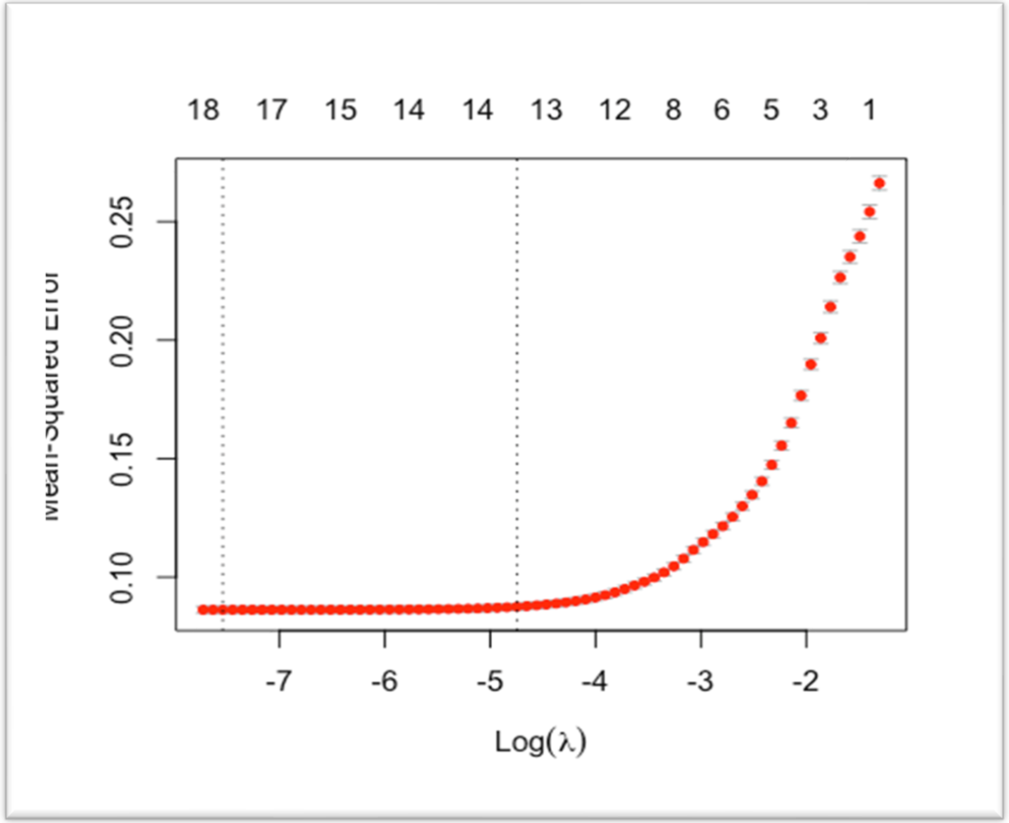
Results of Advanced analysis

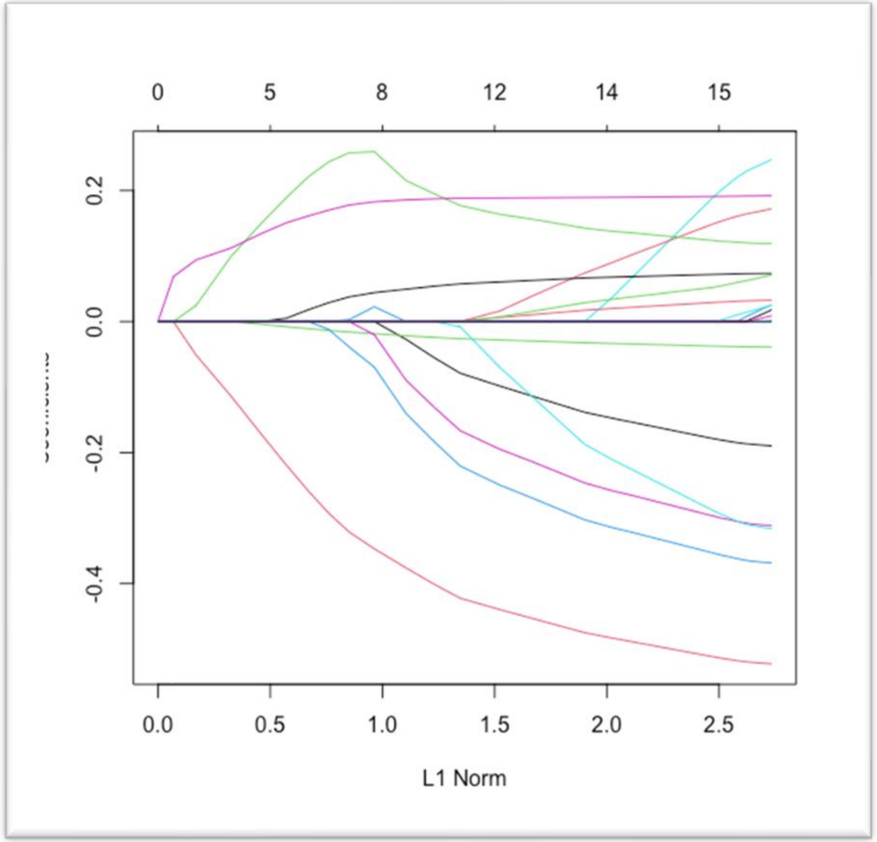
1.Ridge Regression



Best Lambda	Test MSE
0.02713538	0.088802527

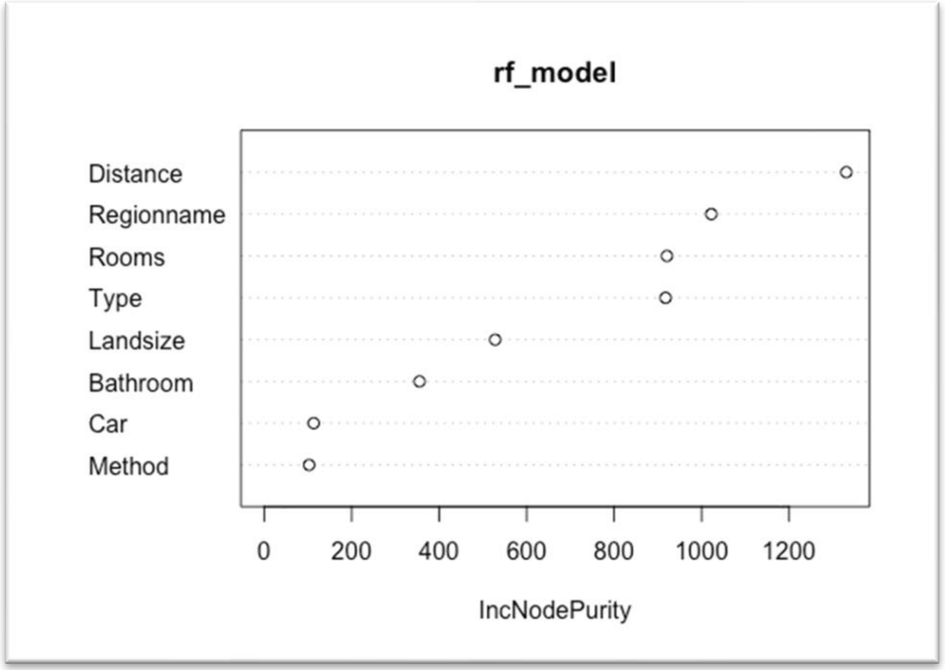
2.Lasso model





Best Lambda	Test MSE
0.00148221	0.08747227

3. Random Forest



Test MSE
0.1309824

Summary

Model	Test MSE
Ridge	0.088802527
Lasso	0.08747227
Random Forest	0.1309824

Best model

Final Lasso model

Variable	Co-efficient
Intercept	13.61600
Type(t)	-0.1868281
Type(u)	-0.5147224
Method(S)	0.05931903
Method(SA)	0.0145875
Method(SP)	0.01066759
Method(VB)	-
Distance	-0.03789760
Region(EV)	0.2111078
Region(NM)	-0.2989450
Region(NV)	-
Region(SEM)	0.1586974
Region(SM)	0.1264151
Region(WM)	-0.3568105
Region(WV)	-0.3200066
Region(EM)	-
Landsize	0.00000334
Rooms	0.1934605
Bathroom	0.07375295
Car	0.02961863

Conclusion

There are 8 variables that are most associated

- Type
- Distance
- Region Name
- Rooms
- Bathroom
- Car
- Land Size
- Method

Lasso model outperformed the other models and identified as the best model when predicting the house prices

UNIVERSITY OF CALGARY

Thank You