EXPLORE || DATA SCIENCE ACADEMY

EDSA APPLE PRICES CHALLENGE

TEAM_TS4_DSFT

PROBLEM STATEMENT

THE DATA SCIENCE PROCESS

Step 1: Data Collection

Durban Fresh Produce Market

Step 2: Data Cleaning

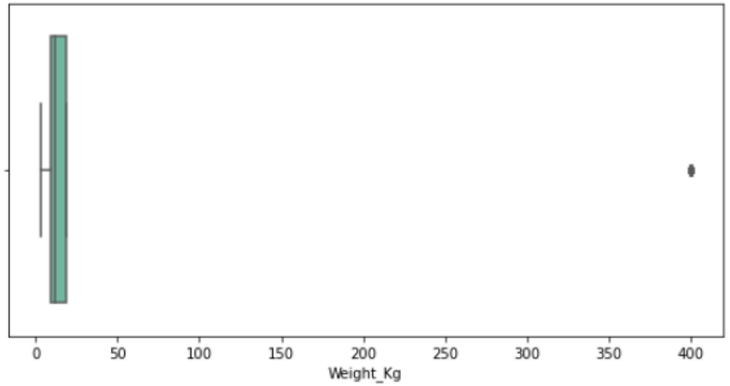
- No NAN Values
- No infinity Values

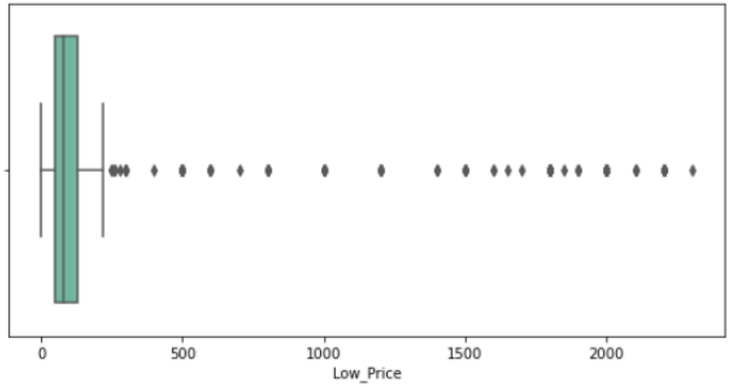
Step 3: Exploratory Data Analysis

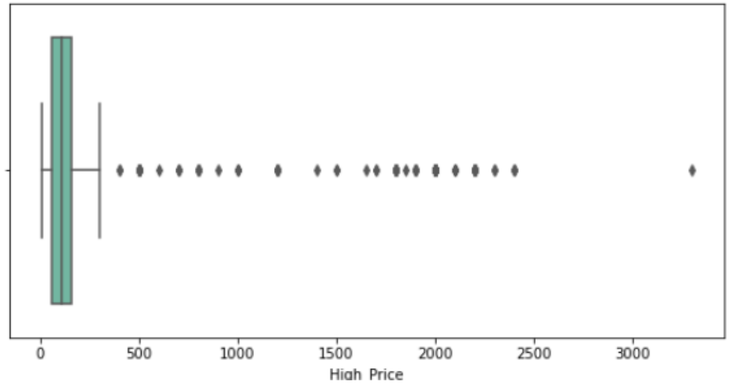
- Univariate Analysis
- Multivariate Analysis

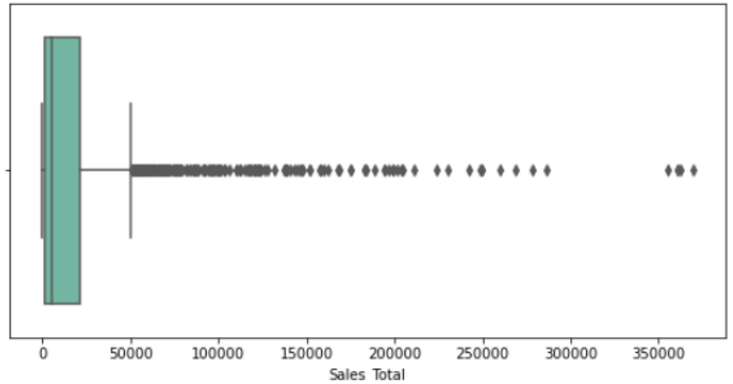
Univariate analysis

Summarised Visuals

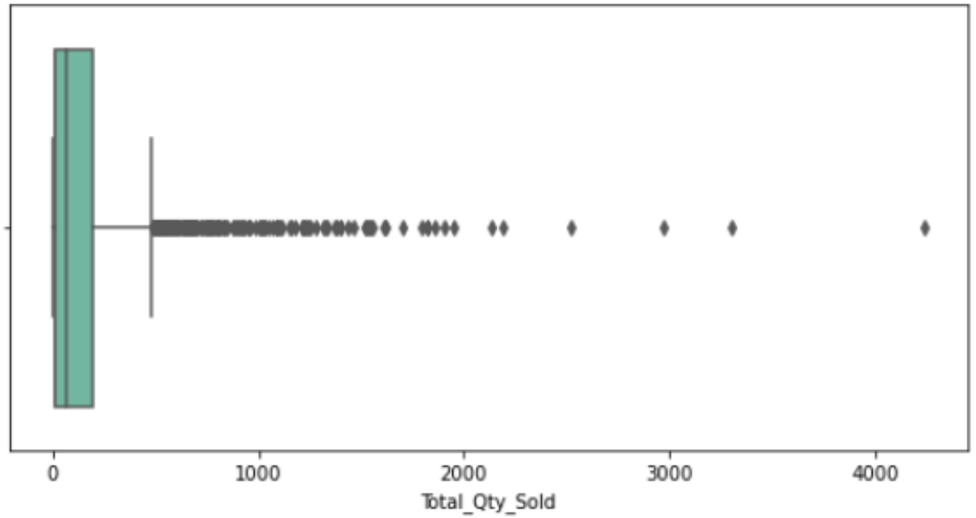


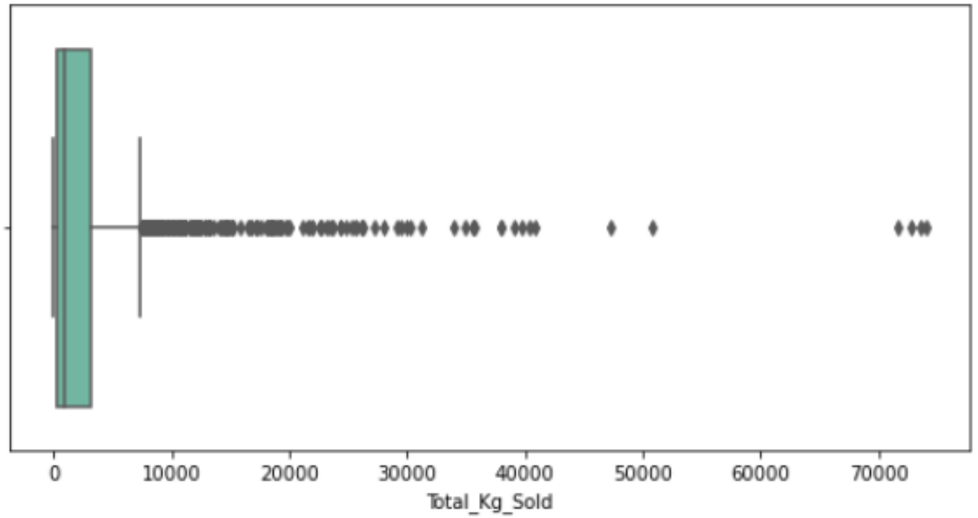


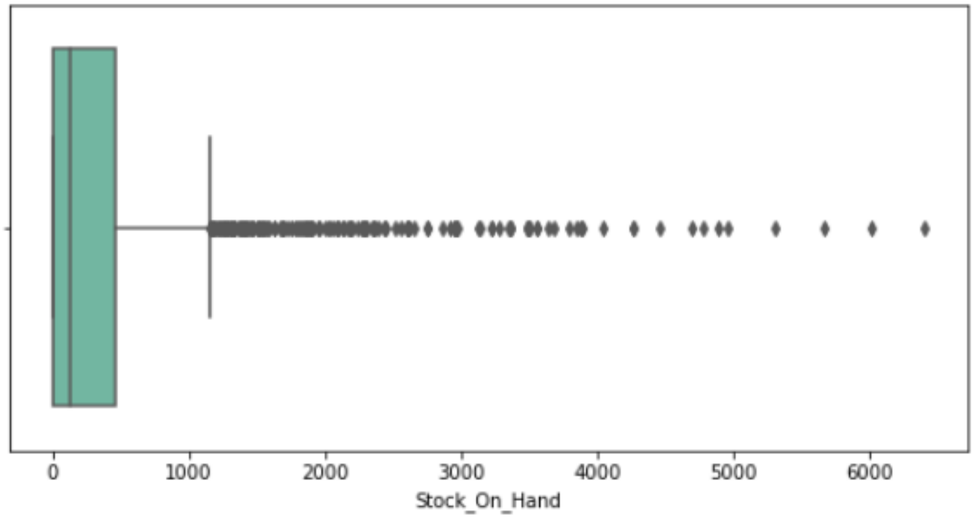


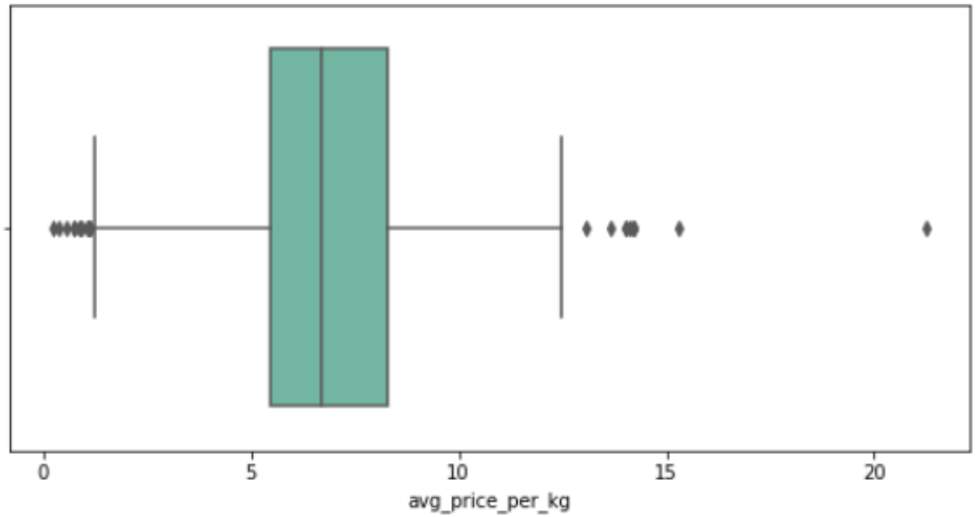


Summarised Visuals

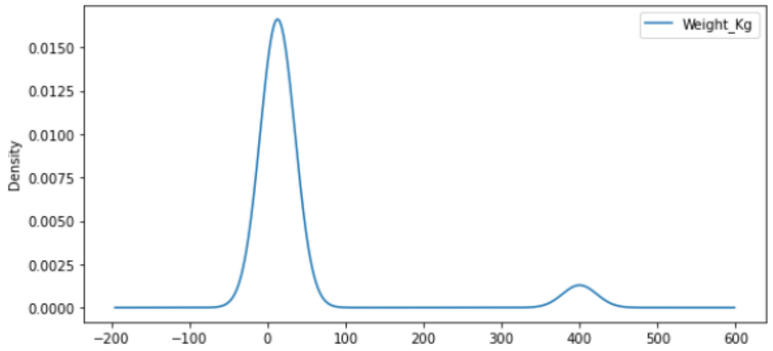


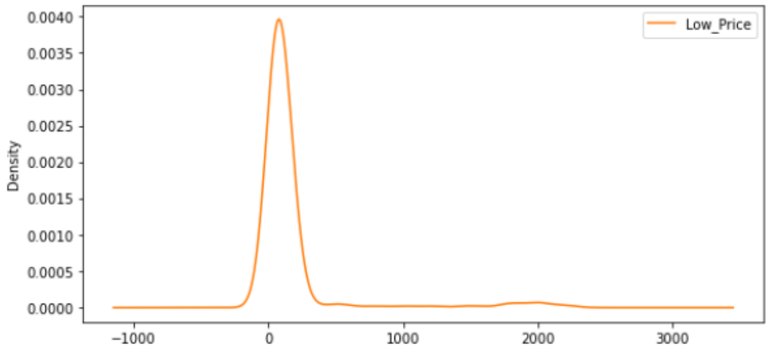


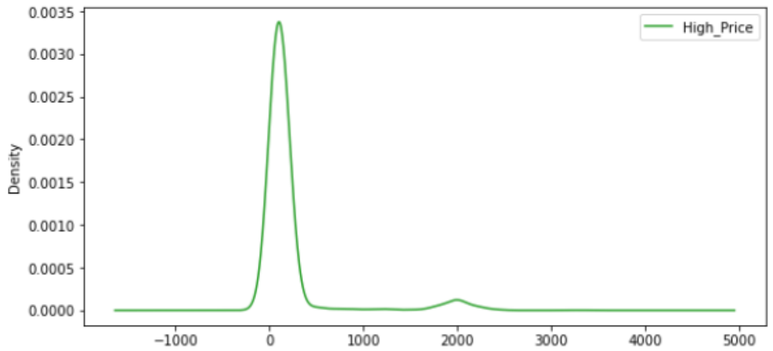


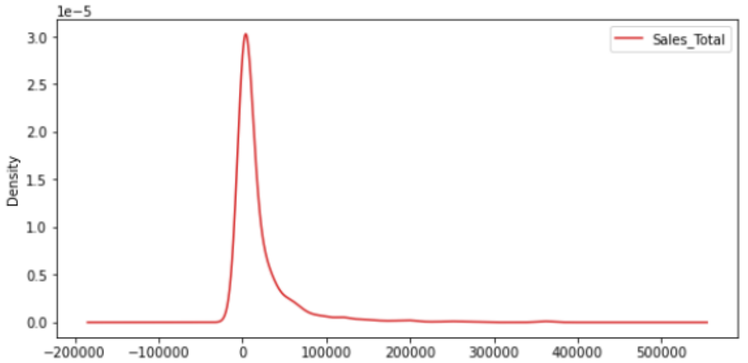


Skewness and Kurtosis

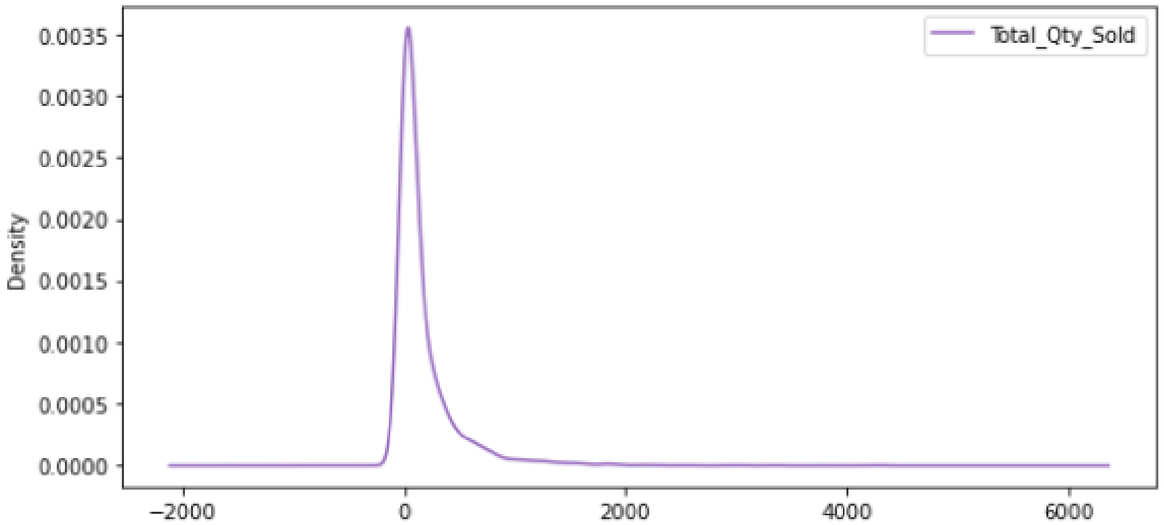


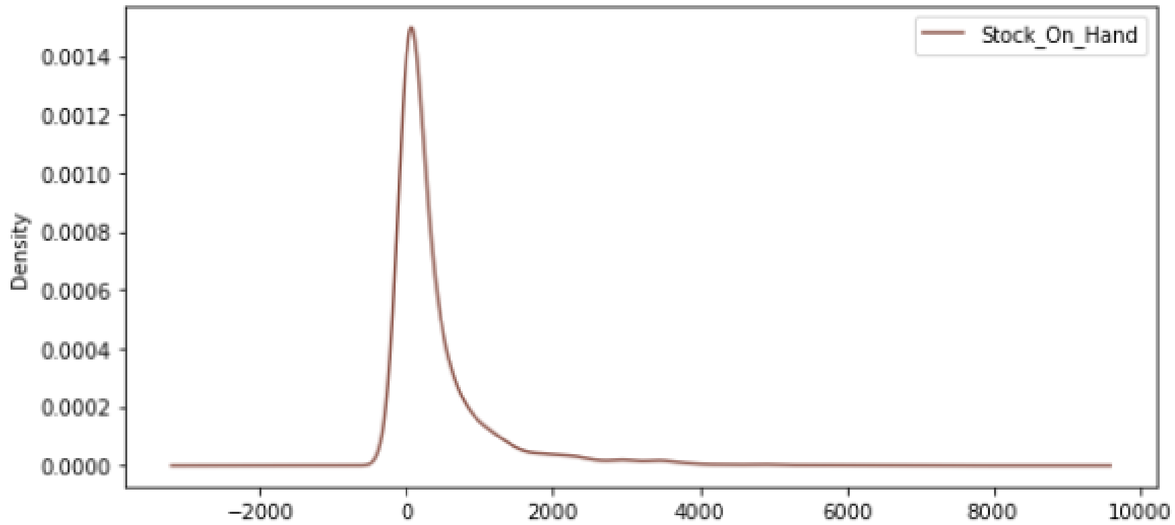


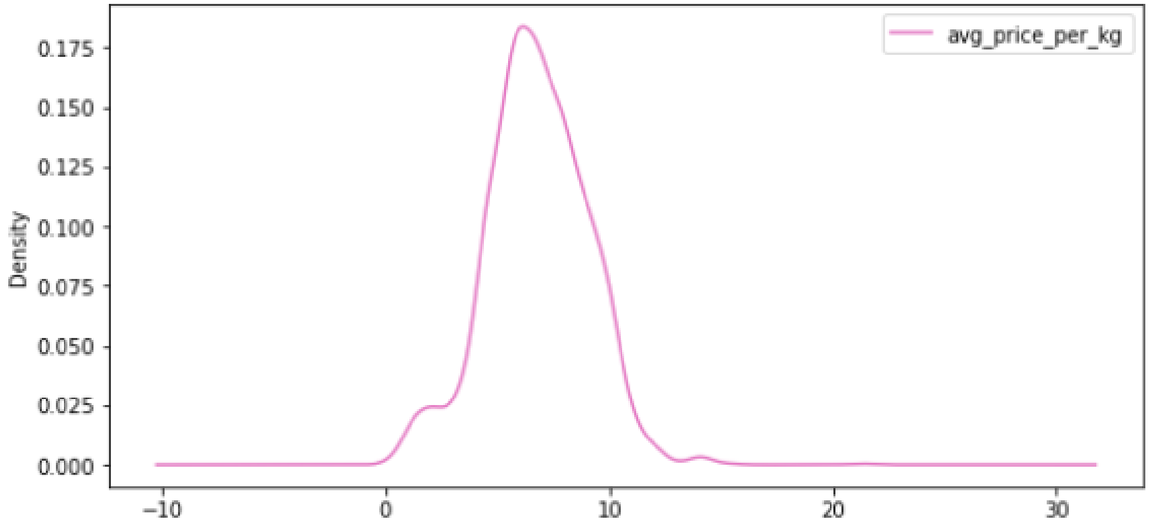




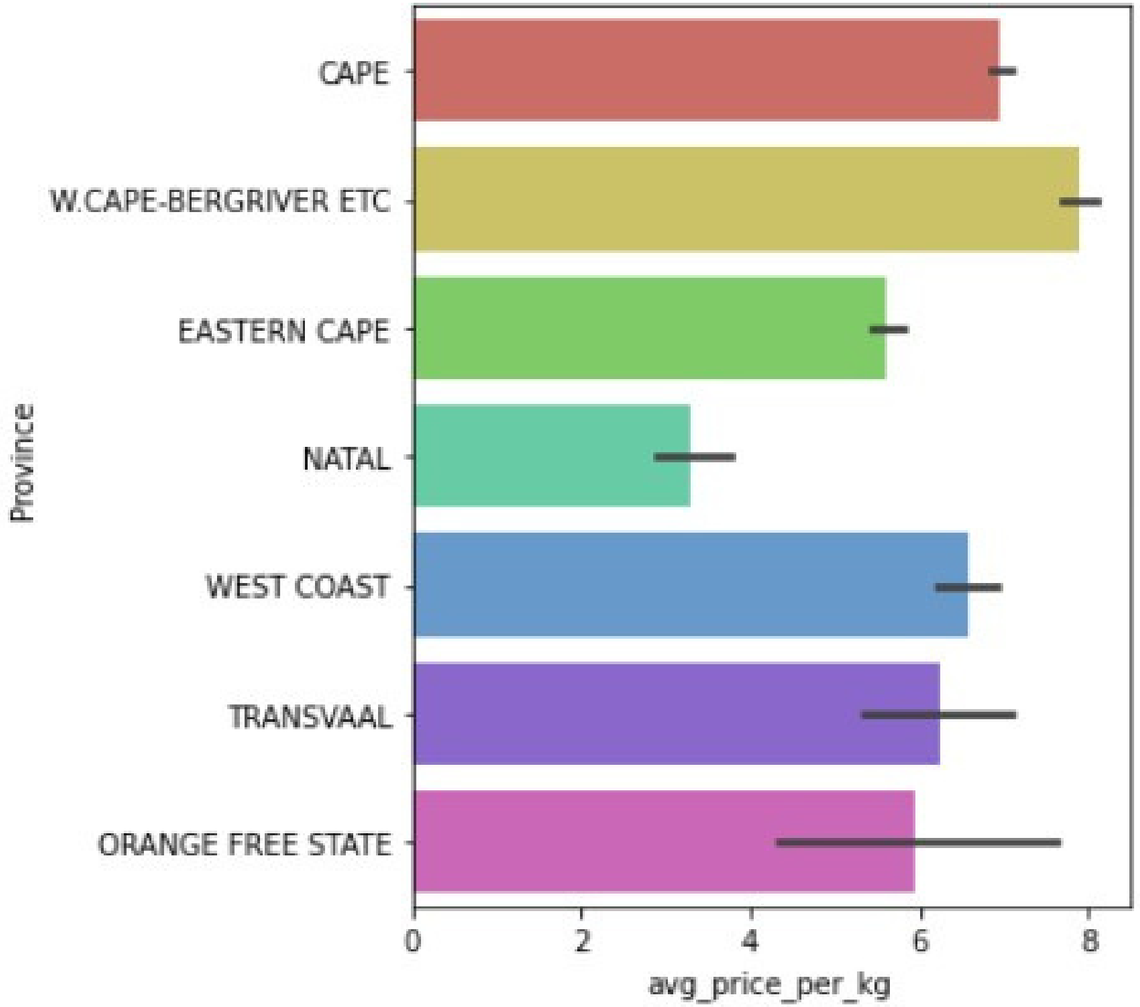
Skewness and Kurtosis

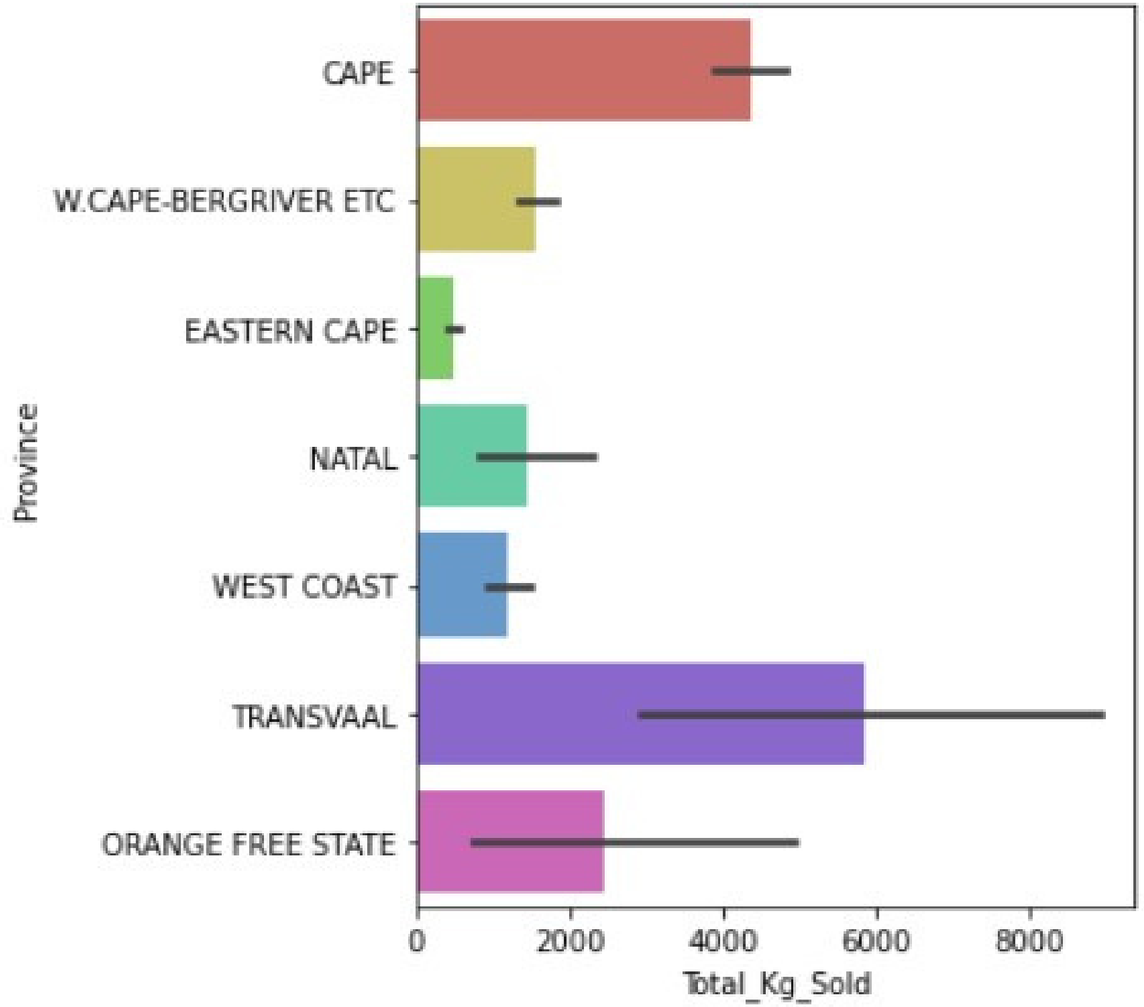




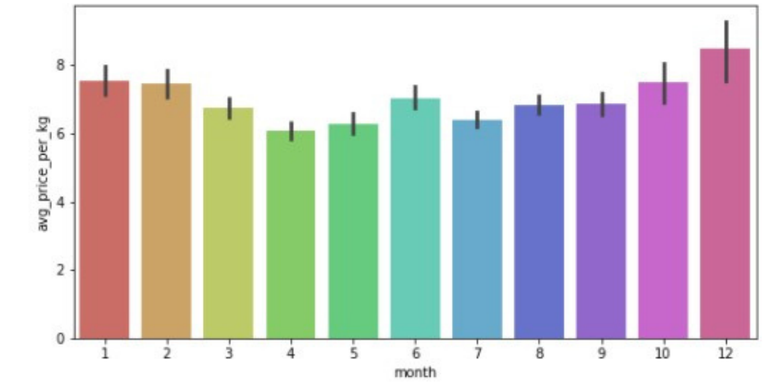


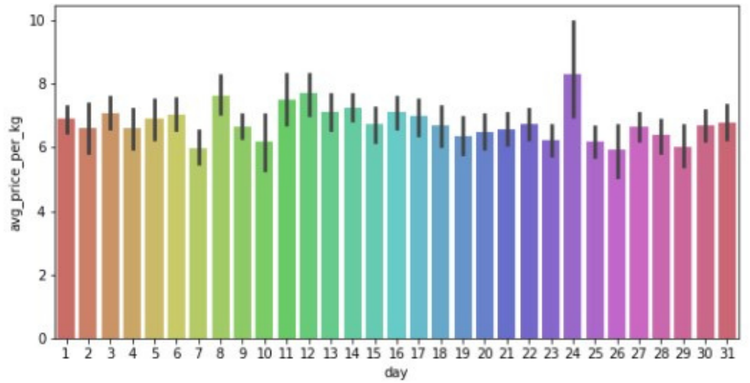
Multivariate Analysis

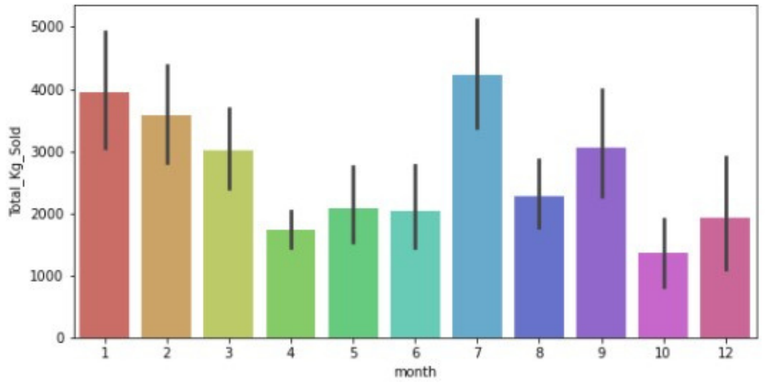


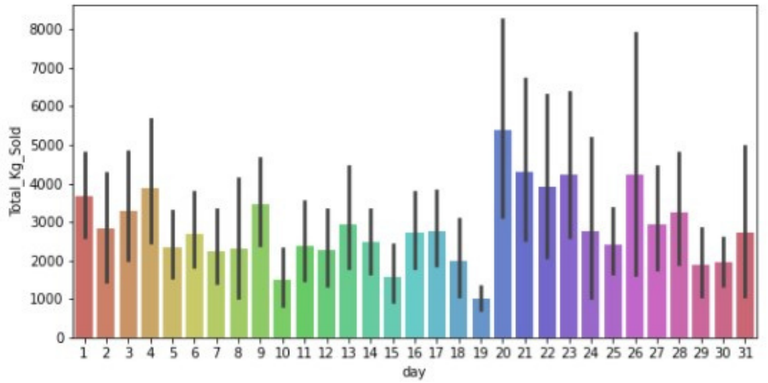


Relationship between months and days

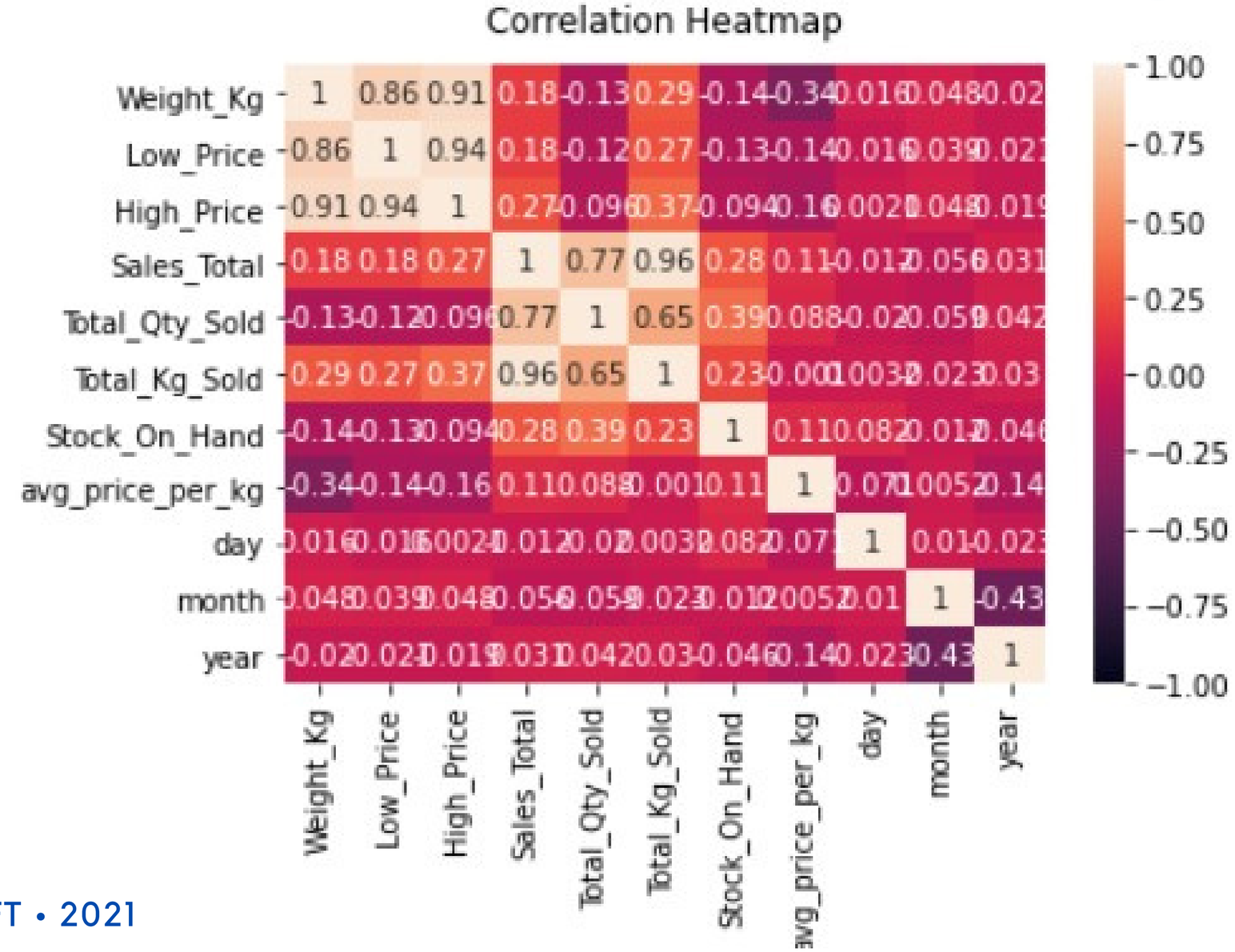




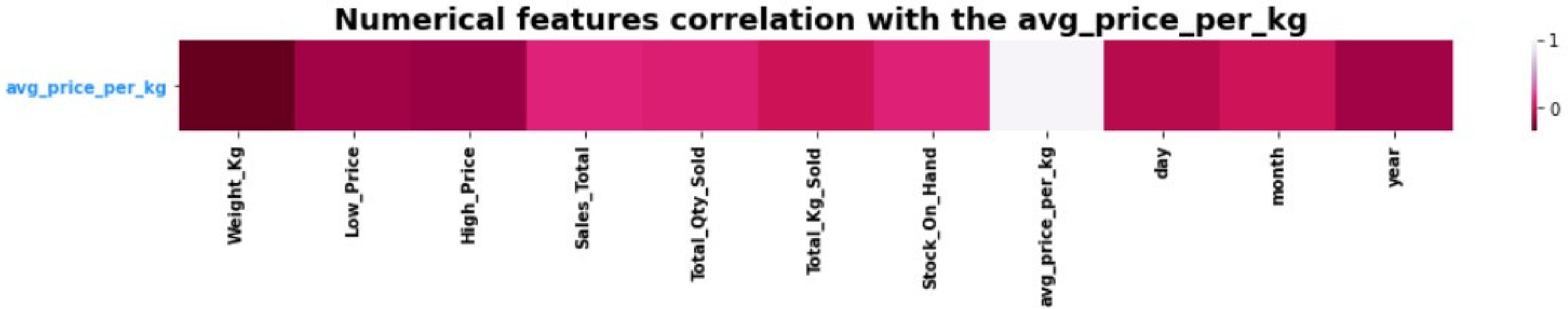




Let's visualize the correlations using a heat map



Correlations between all our features and our target variable



Dummy Variabe Encoding

- In regression analysis, a dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical feature.
- Dummy variable encoding has our dataset transformed from 15 columns to 36

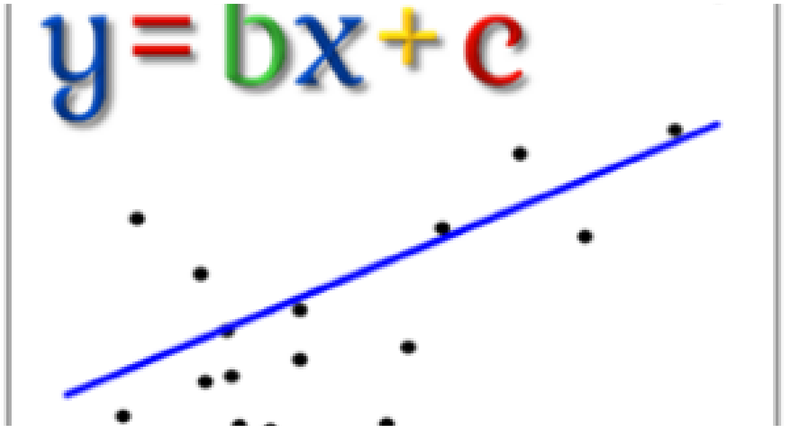
Variable Selection by Variance Thresholds

- Variable selected as a result of variable selection by variance thresholds
- A threshold of 3%

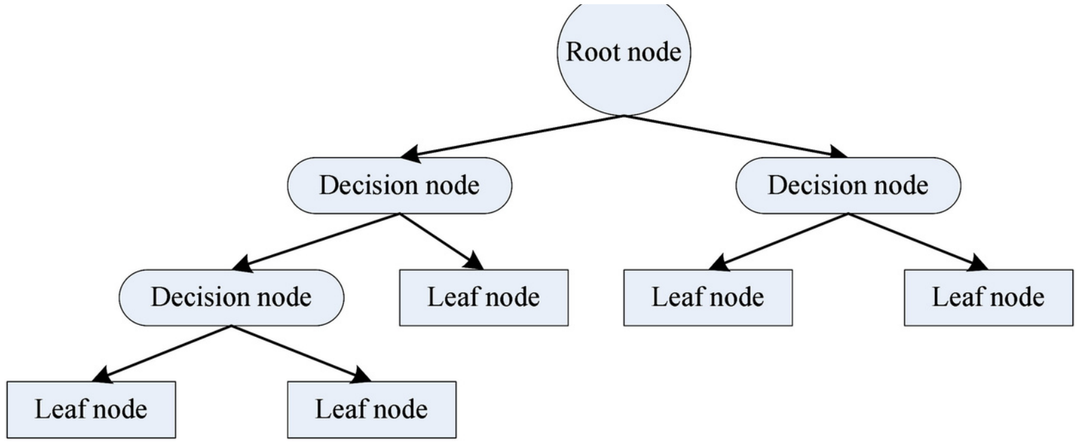
- The result was a decrease of variables, 36 to 15

The Models

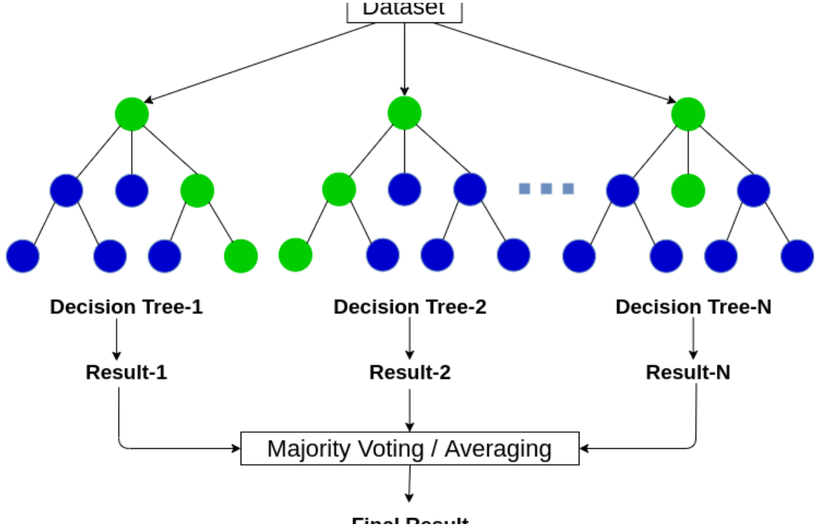
LINEAR REGRESSION



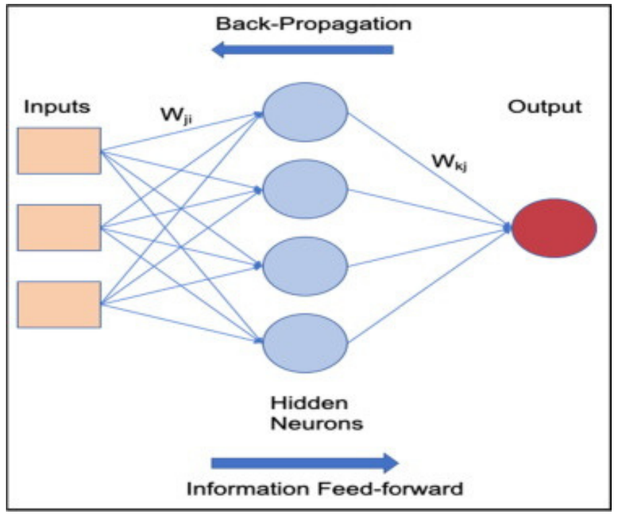
BUILDING THE DECISION TREE REGRESSOR MODEL



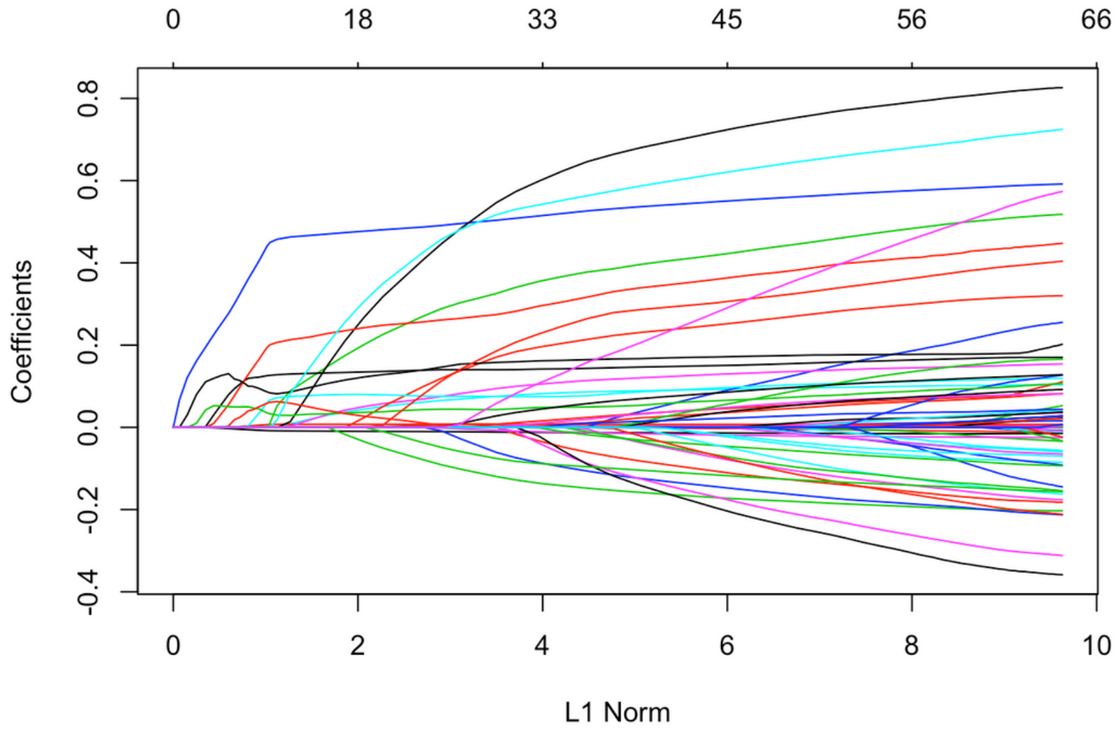
BUILDING THE RANDOM FOREST MODEL



BUILDING THE XGBOOST MODEL



BUILDING A LASSO REGRESSION MODEL



The RMSE Criterion for Model Selection

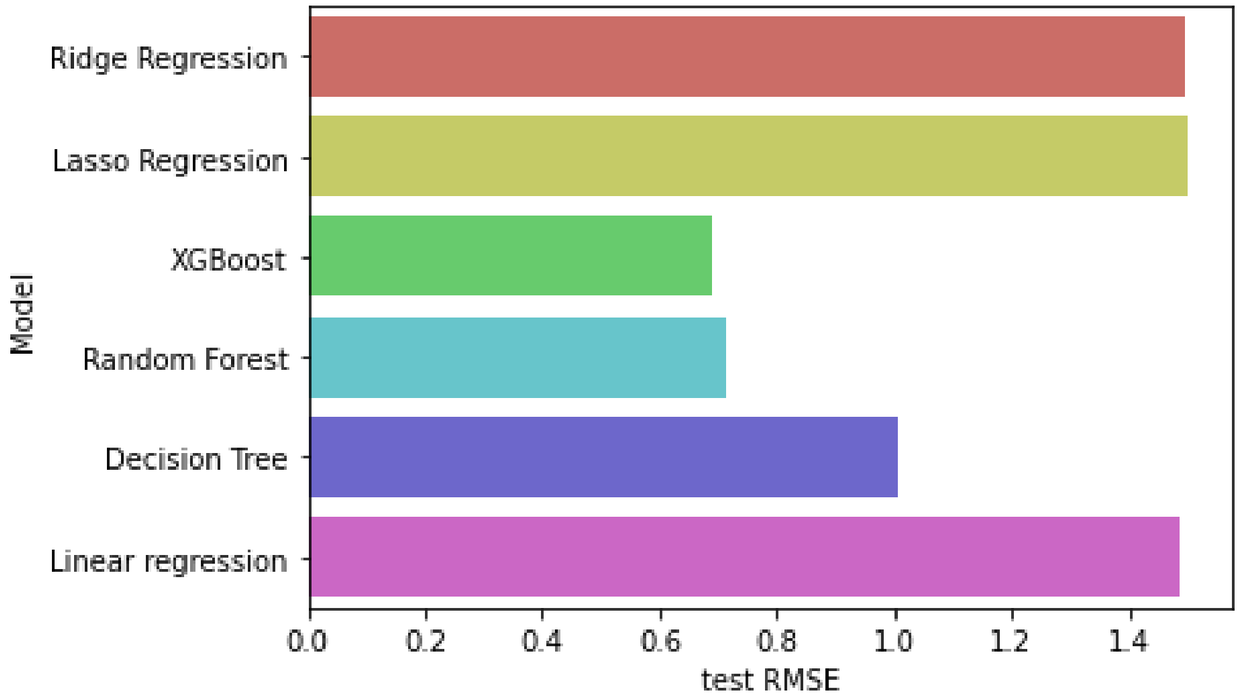
- How RMSE and R-squared help with model selection

- The RMSE results of the candidate models

- Ridge Regression : 1.4949277683692117
- Lasso Regression : 1.5024482644764312
- XGBoost: 0.6911297404630196
- Random Forest : 0.7119468540737508
- Decision Tree : 1.003683521955450 7
- Linear regression : 1.487728622763673 2

Optimal Regression Model Selection

As expected, the lasso and the ridge models fell short as compared to the XGBoost, Random Forest even the Linear regression slightly.



The Optimal Model

- The optimal model is XGBOOST
- This model is Extreme Gradient Boosting
- Key advantages of XGBOOST

Conclusion

- We started with a dataset of 64376 and 15 columns.
- RMSE was used as a criterion to for selecting the optimal model for use in our prediction of average prices per kilogram
- The Optimal mode we find is XGBoost model.

MEET OUR TEAM MEMBERS

Thato Bogopane

Namhla Sokapase

Sibongile Maluleka

Mahlomola Mothogoane

Katlego Mathole


Any
questions