INTRODUCTION TO DATA SCIENCE EXPLORATORY DATA ANALYSIS ASSIGNMENT

- Tejas Srinivasan (PES1201800110)

- Joseph Dominic (PES1201800328)

DATASET DESCRIPTION-

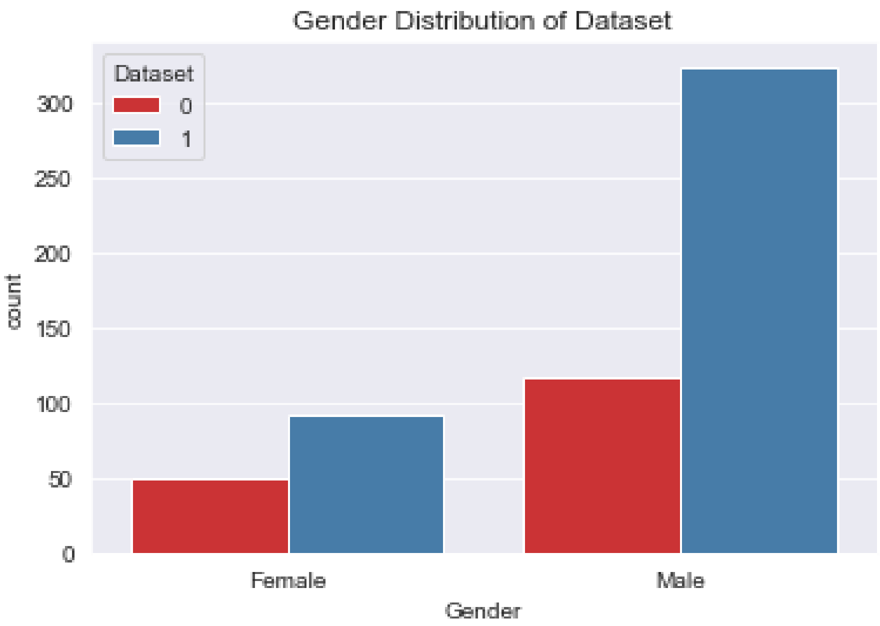
- This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India.
- The “Dataset” column is a class label used to divide groups into liver patient (liver disease) or not (no disease). This data set contains 441 male patient records and 142 female patient records.
- Any patient whose age exceeded 89 is listed as being of age “90”.
- There are 583 rows and 11 columns, out of which two are categorical and rest are numerical.

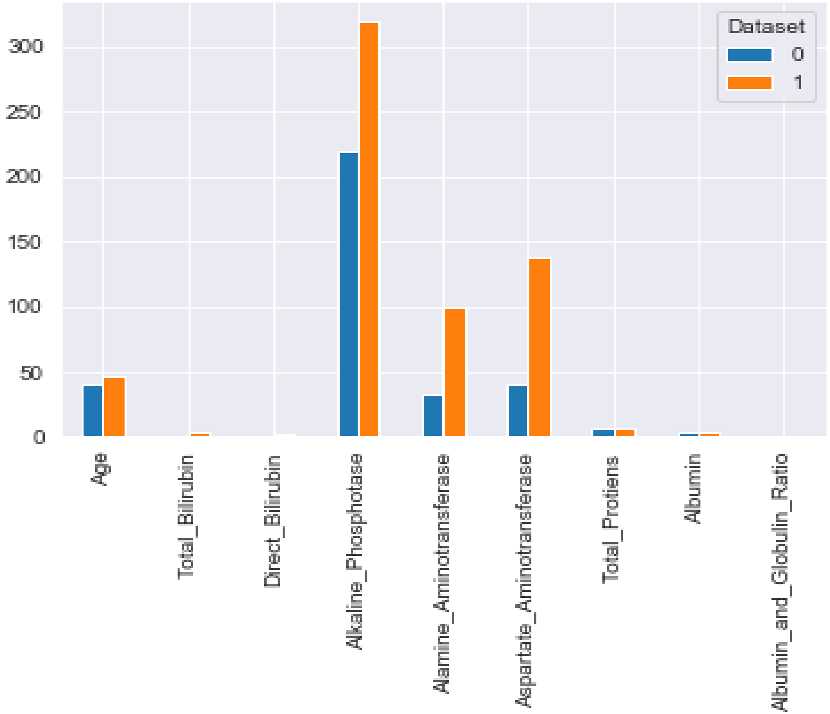
DATA CLEANING -

- Only 4 values were missing from the Albumin_and_Globulin_Ratio column.
- The column did not have enough NA/missing values to drop so the missing values were replaced instead.
- These values were filled using interpolation.
- Since there is no accepted medical definition of an "outlier", the entire range of values in the dataset was considered.
- The categorical values (2) in the "Dataset" column were changed to 0s for easier understanding and visualization of the column.

VISUALIZATIONS -

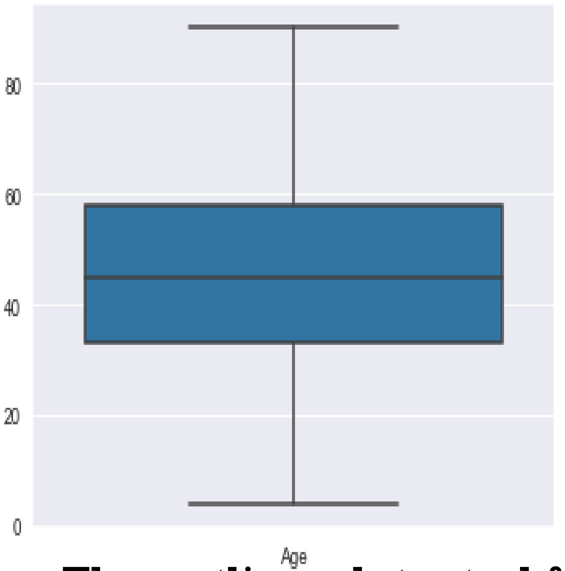
Bar plots -



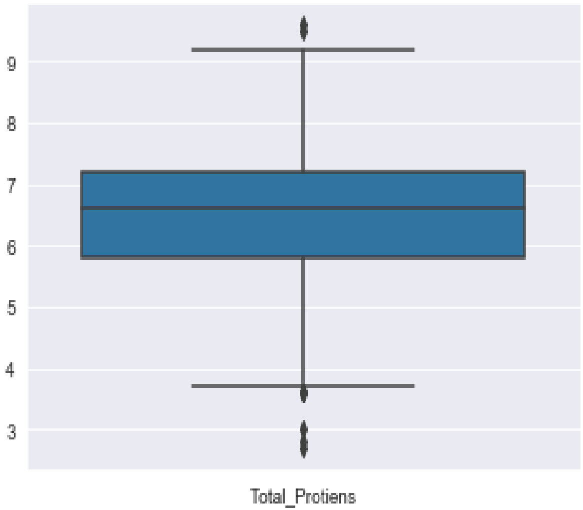


BOXPLOTS -

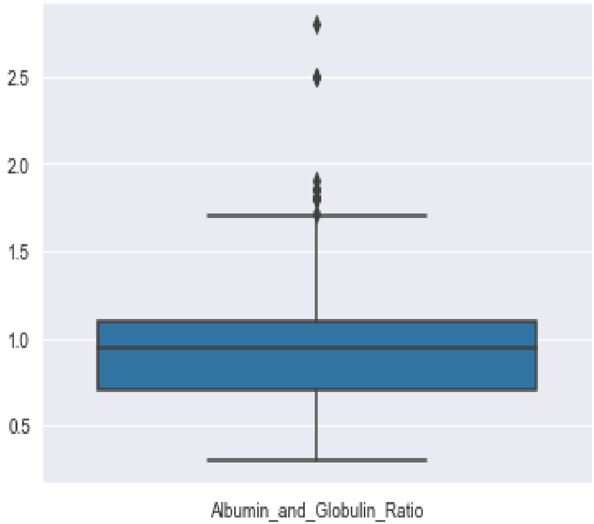
Median value for Age is about 45 years



Median Total_Proteins value is about 6.7

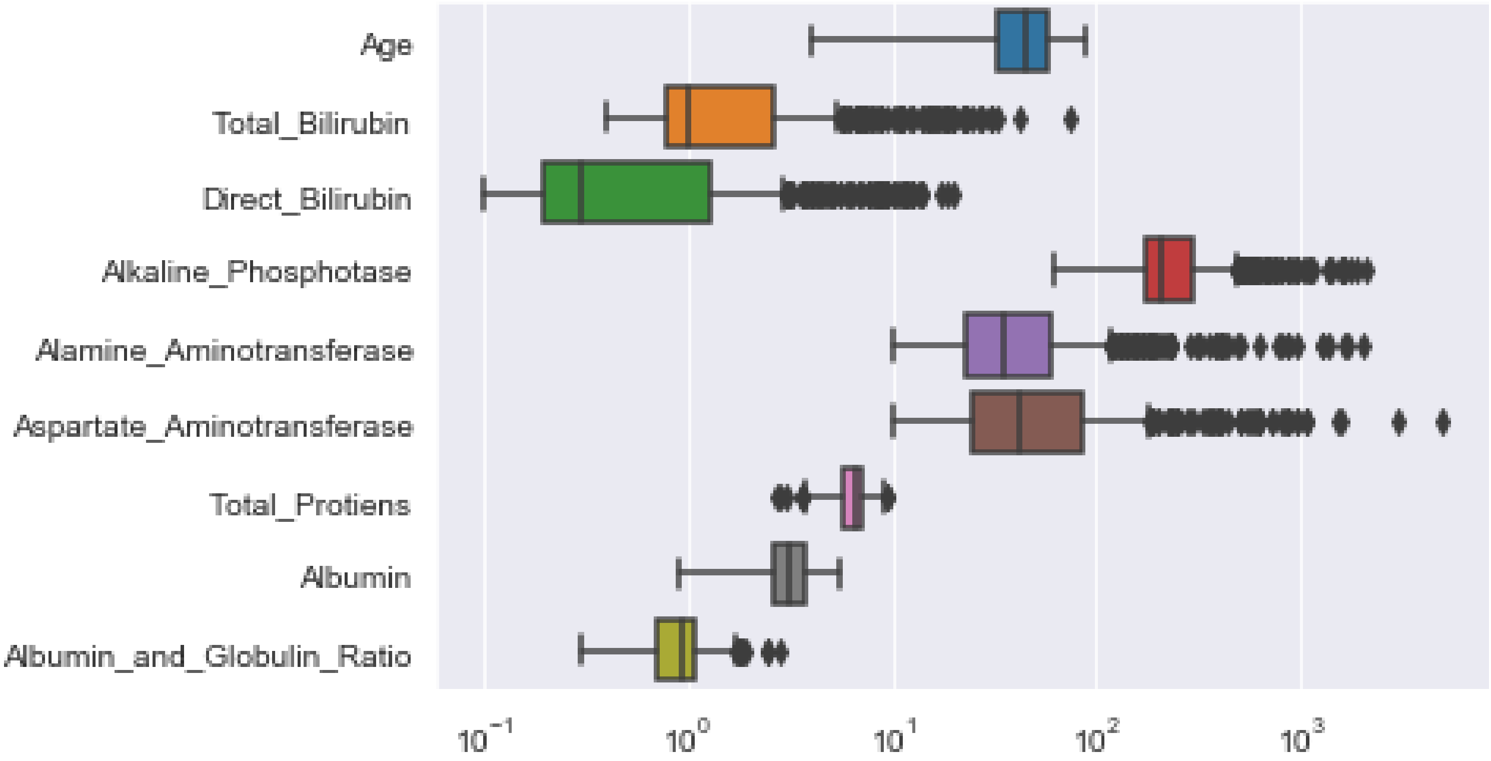


Median AGR is about 9.9

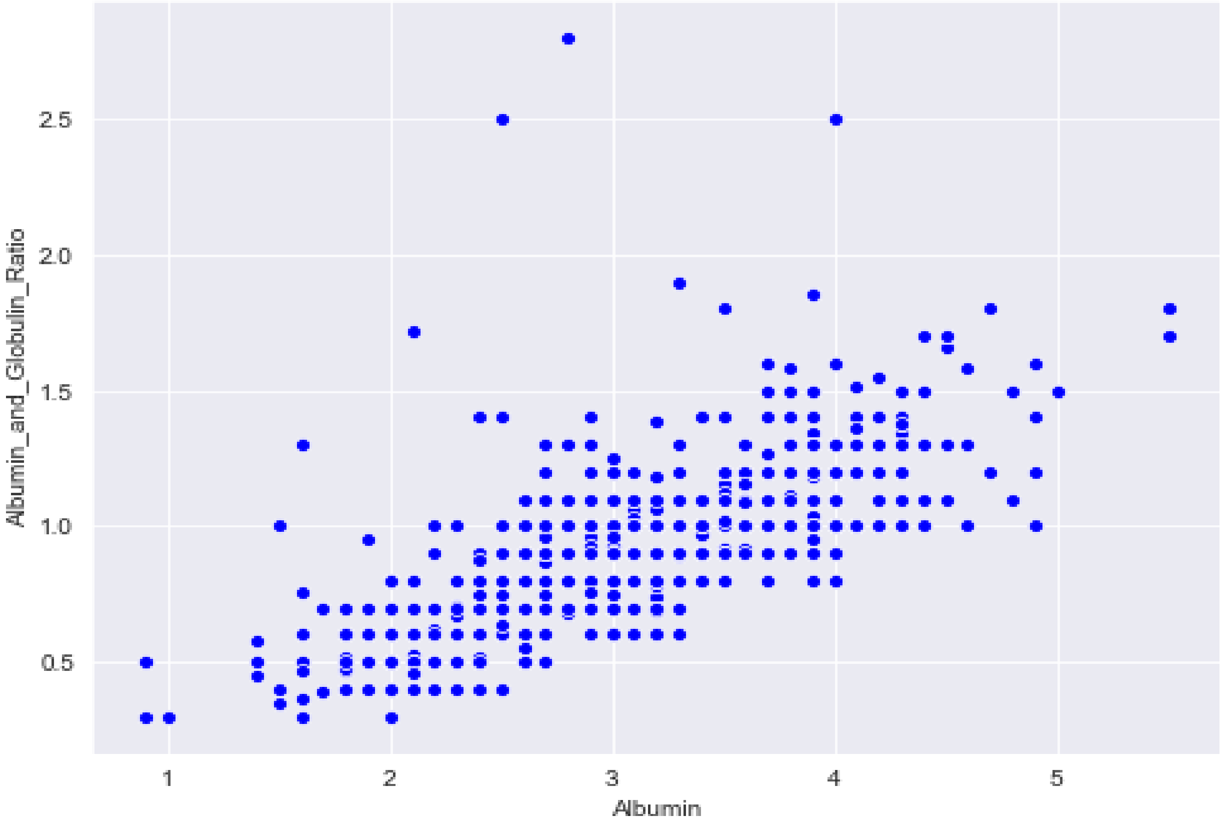


The outliers detected for Total Proteins and AGR were ignored.

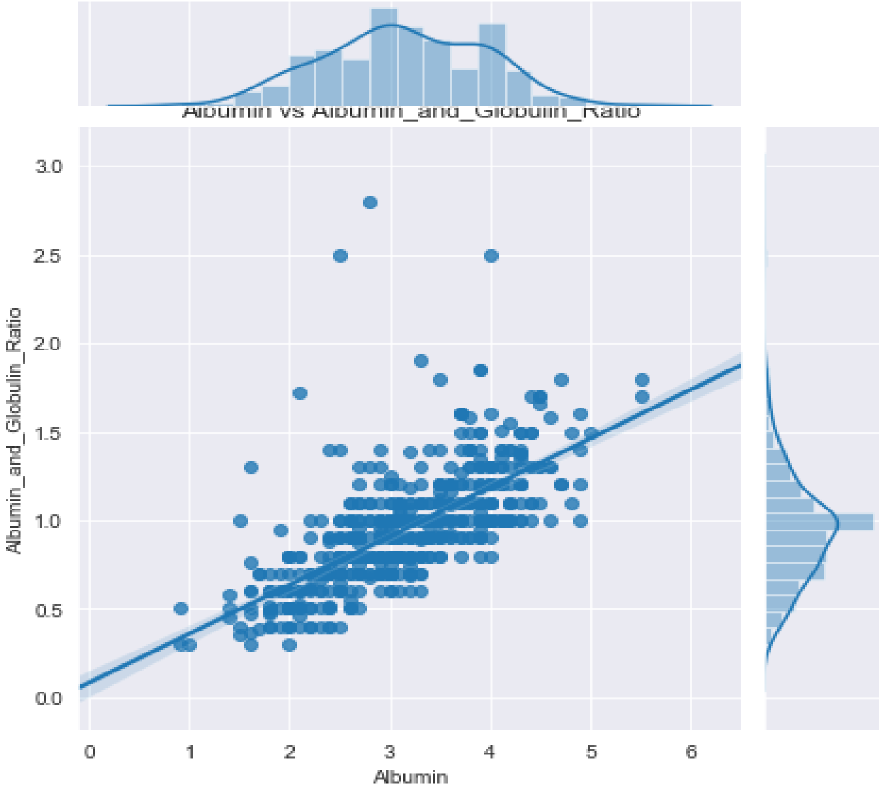
BOXPLOT WITH ALL THE ATTRIBUTES -

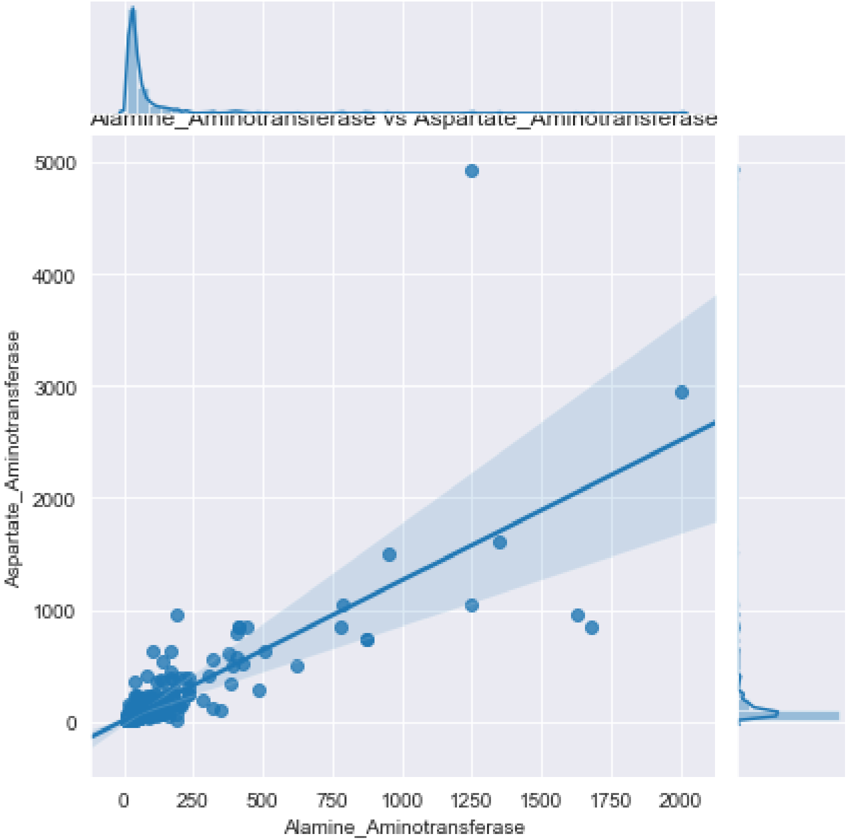


SCATTER PLOTS -

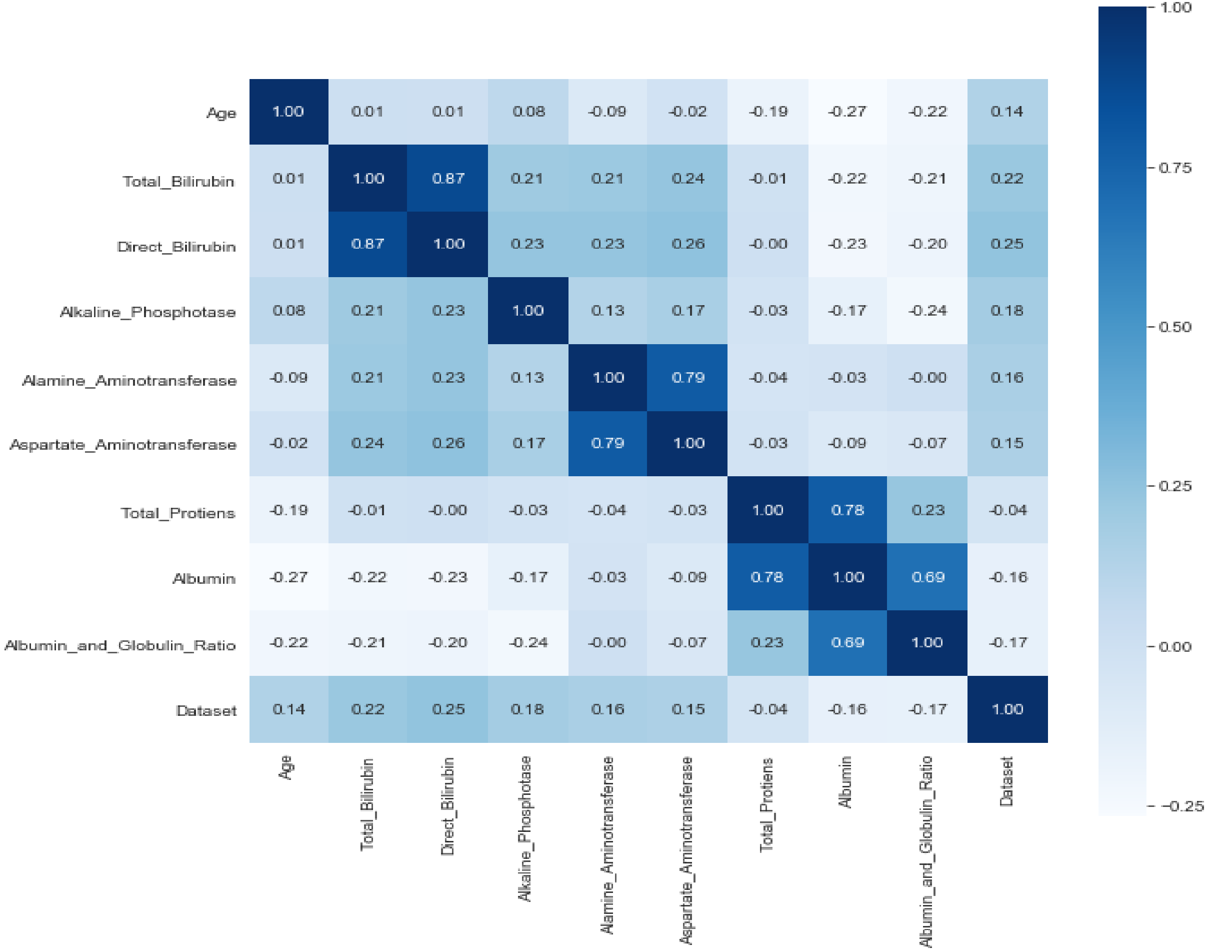


JOINTPLOTS TO CHECK FOR LINEARITY IN CORRELATION BETWEEN DIFFERENT VARIANTS OF THE SAME PROTEIN -





CORRELATIONS BETWEEN DIFFERENT COLUMNS

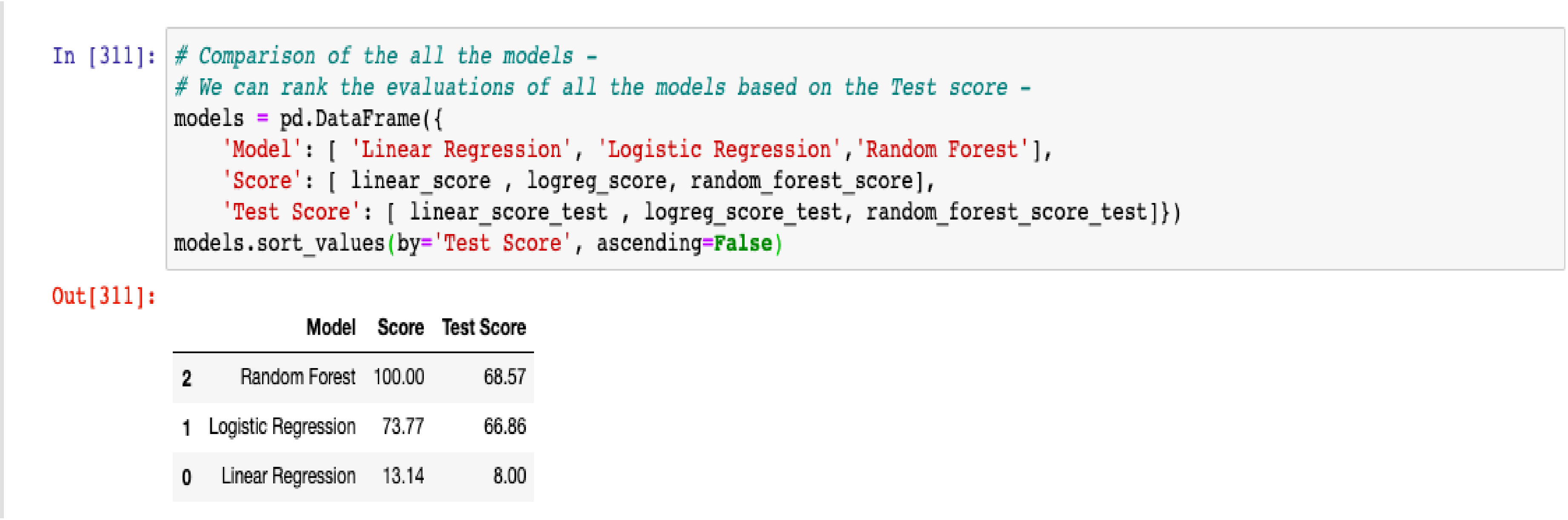


	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Dataset
Age	1.00	0.01	0.01	0.08	-0.09	-0.02	-0.19	-0.27	-0.22	0.14
Total_Bilirubin	0.01	1.00	0.87	0.21	0.21	0.24	-0.01	-0.22	-0.21	0.22
Direct_Bilirubin	0.01	0.87	1.00	0.23	0.23	0.26	-0.00	-0.23	-0.20	0.25
Alkaline_Phosphotase	0.08	0.21	0.23	1.00	0.13	0.17	-0.03	-0.17	-0.24	0.18
Alamine_Aminotransferase	-0.09	0.21	0.23	0.13	1.00	0.79	-0.04	-0.03	-0.00	0.16
Aspartate_Aminotransferase	-0.02	0.24	0.26	0.17	0.79	1.00	-0.03	-0.09	-0.07	0.15
Total_Protiens	-0.19	-0.01	-0.00	-0.03	-0.04	-0.03	1.00	0.78	0.23	-0.04
Albumin	-0.27	-0.22	-0.23	-0.17	-0.03	-0.09	0.78	1.00	0.69	-0.16
Albumin_and_Globulin_Ratio	-0.22	-0.21	-0.20	-0.24	-0.00	-0.07	0.23	0.69	1.00	-0.17
Dataset	0.14	0.22	0.25	0.18	0.16	0.15	-0.04	-0.16	-0.17	1.00

PREDICTIONS MADE:

Linear Regression	Logistic Regression	Random Forest
Values are linearly plotted	Values are plotted for binary classes using a sigmoid function	Values are plotted using a RF classifier
Score = 13.14	Score = 73.77	Score =100

COMPARISON OF ALL THE TEST AND TRAIN SCORES OF ALL APPROACHES -



In [311]:

```
# Comparison of the all the models -
# We can rank the evaluations of all the models based on the Test score -
models = pd.DataFrame({
    'Model': [ 'Linear Regression', 'Logistic Regression','Random Forest'],
    'Score': [ linear_score , logreg_score, random_forest_score],
    'Test Score': [ linear_score_test , logreg_score_test, random_forest_score_test]})
models.sort_values(by='Test Score', ascending=False)
```

Out[311]:

	Model	Score	Test Score
2	Random Forest	100.00	68.57
1	Logistic Regression	73.77	66.86
0	Linear Regression	13.14	8.00

HYPOTHESIS TESTING -

- The Hypothesis we tested was: Men above 45 are more susceptible to liver disease than Women above 45.
- H0 : Proportion of affected men above 45 - Proportion of affected women above 45 <=0
- H1 : Proportion of affected men above 45 - Proportion of affected women above 45 > 0
- Used the Chi – Square Test.
- Rejected H0. Implying Men above 45 are more susceptible than women above 45.

CONCLUSIONS DRAWN -

- The different variants of the proteins of the same type (Eg. Aspartate and Alanine Transferase and Total and Direct Bilirubin are linearly correlated - as shown by the jointplots).
- The skewness of the data towards men (in plots such as Gender vs Total_Bilirubin and Gender vs Albumin) is due to a higher number of men in the dataset, as shown in the barplot.
- The 'Albumin_and_Globulin_Ratio' column has the highest correlation (about 0.64) with the 'Dataset' column. The 'Age' column has the lowest correlation with 'Dataset' , (0.013) and can be dropped if need be.
- Prediction of Liver Disease has been been performed using Linear Regression, Logistic Regression and Random Forest and it was found that Random Forest gave the best accuracy since it takes a model subset of the features instead of all of them.
- From the hypothesis test, it has been concluded that Men above 45 are more susceptible to liver disease than Women above 45.

Thank you!