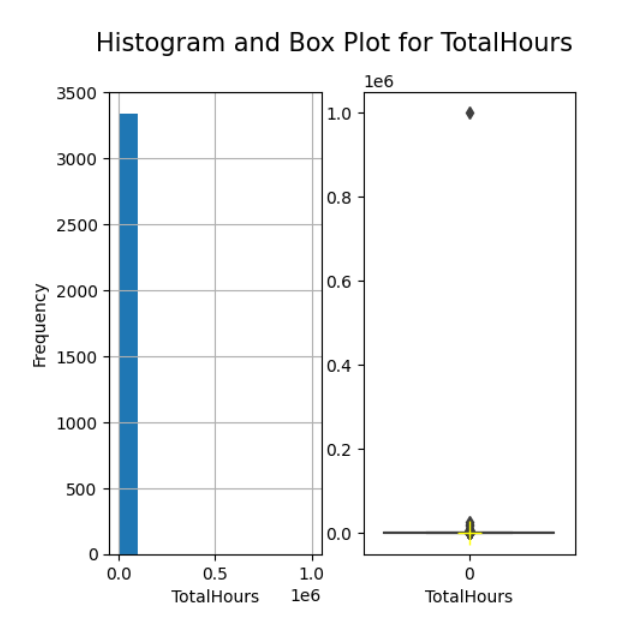
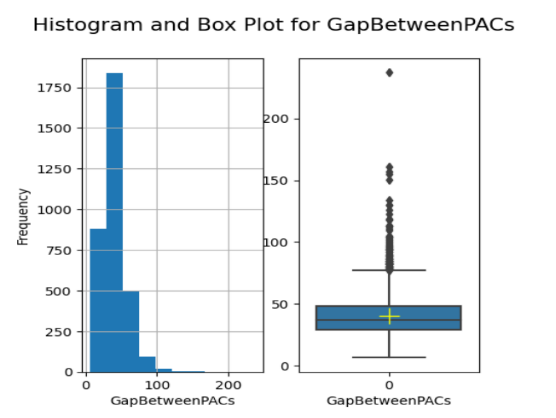
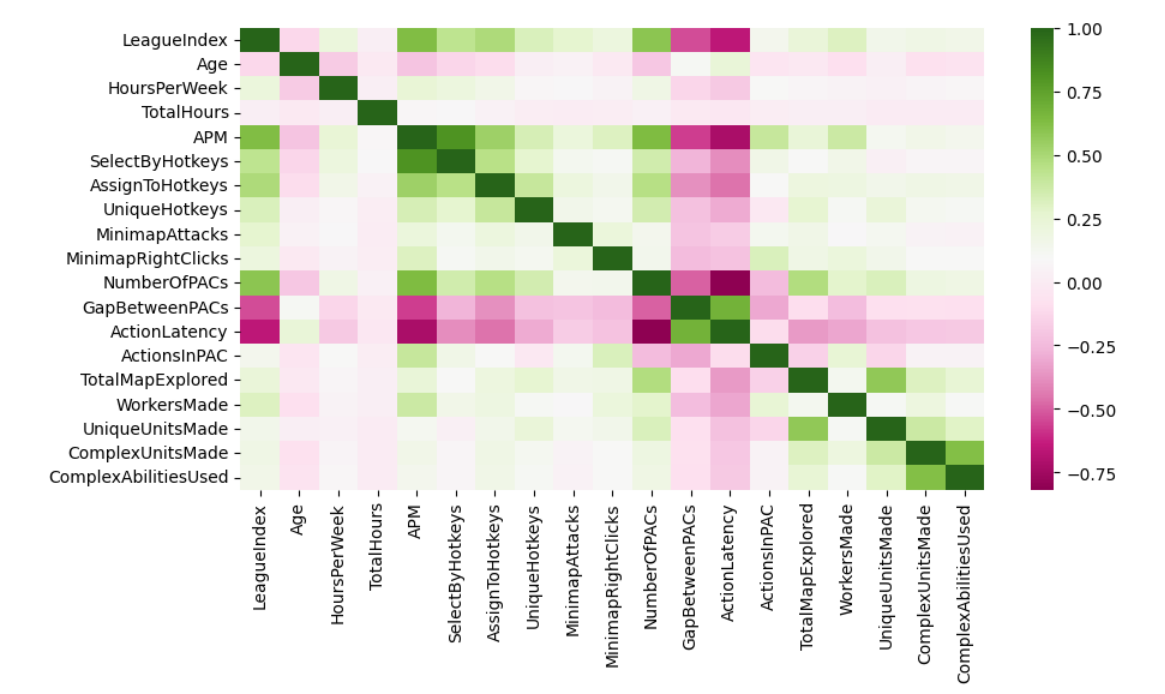
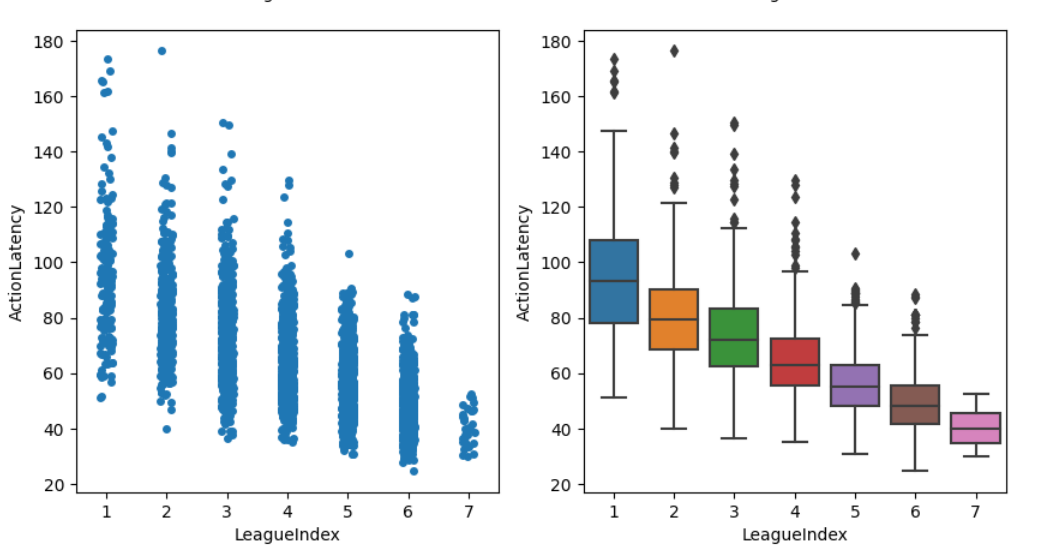
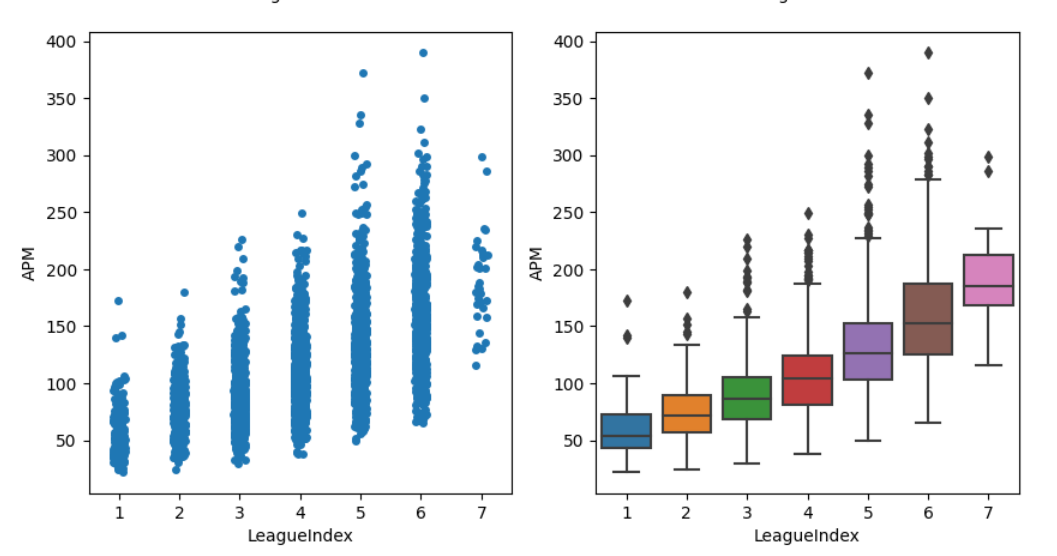
SUMMARY

* **Data characteristics:** The Dataset consists of 3395 rows and 20 columns.

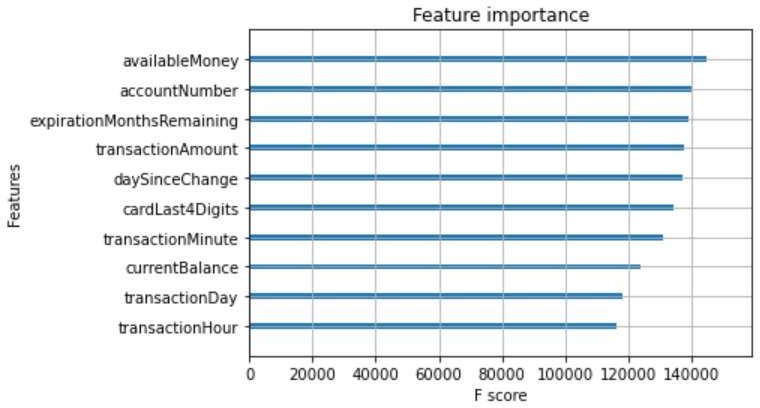
#### Data Exploration (Key Takeaways):

* 1. **Missing Value Analysis:** The Dataset had few columns (Age, HoursPerWeek, TotalHours) which had an object data type in spite of being quantitative. On further analysis, it was found that a few of the rows had the value "?". These values can be dropped or imputed. They have been dropped in this workflow. They can also be imputed by using a Regression Model to predict the missing Age, HoursPerWeek and TotalHours. This led to all the LeagueIndex having 8 to be dropped.
  2. Check for Imbalanced Data :
  3. **Outlier Detection:** The Data had many columns which had very high values for some rows. Hence Outlier Detection was done using Interquartile Range to remove the outliers from the data.

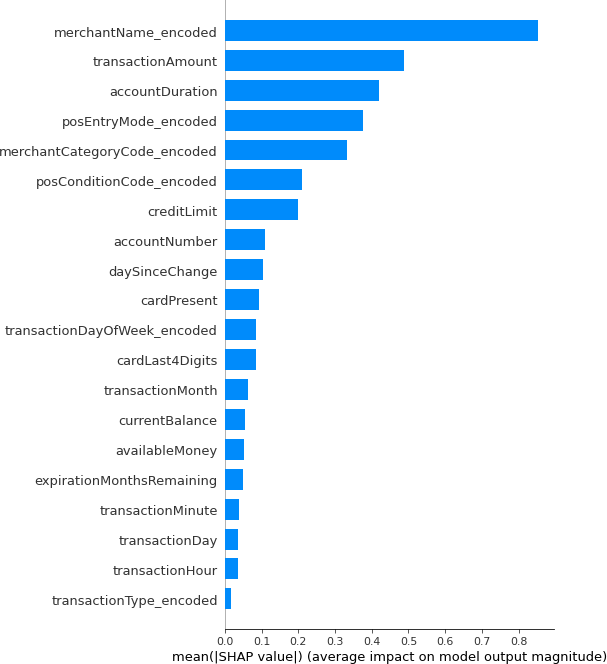
* 1. **Correlation:** The % of fraudulent transactions is higher(**2.8%**) when the cardCVV and enteredCVV does not match, as compared to complete dataset where the % of fraudulent transactions is **1.57%. 39.7%** of all fraudulent transactions are when the merchantCategoryCode is **online\_retail**. Also, out of all fraudulent transactions, the **card is not present** in **72%** of those fraudulent transactions
  2. Bivariate Analysis : Average transaction amount of fraudulent transactions is higher **($ 225)** as compared to non-fraudulent transactions **($ 135)**. However, there are very high value transactions above $ 1600 which are not fraudulent.



* 1. Modelling :
  2. Used XGBoost Feature importance to gain an insight on which variables played a major role in the construction of trees in XGBoost while training the model.



* 1. Used shapley values to understand the feature which have the largest contribution to the model prediction. This can be done on a global level as well as on a local level where we can drill down on a particular row to see which variables helped in making the decision. Shapley values has only been run on a small subset of Test Data because of high calculation time.



## Steps to run the code

* There is only one jupyter notebook that needs to be run which has markdowns explaining the steps