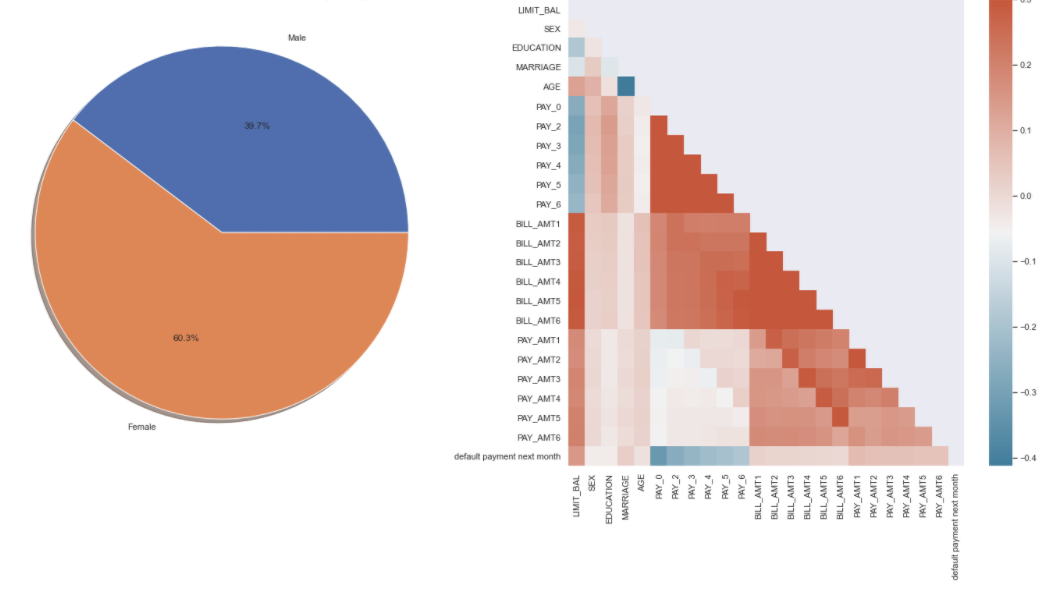
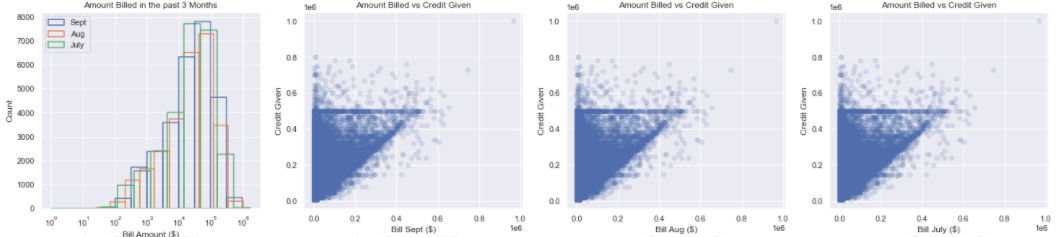
# Lessons Learned

Potential Business Value

* Close to 2/3 of the data comes from female customers, this can give a useful demographic to target future adds. However, for data analytics purposes, this means that the possibility exists that any model created from this data might have a bias towards females. This must be taken into account when deploying the final model.
* A correlation matrix formed using all the features in the data frame shows a high correlation between all the monthly bill payments and also the payment status of the customer. This suggests the model will prioritize these features over others. Nevertheless, it is not possible to determine a clear group of features essential for data modeling, this is why machine learning will be essential for model building.
* Scatter plots and histograms reveal a slight patter in the data of monthly bills and the limit credit score. The histograms suggest a normal distribution on a log scale for the first three months of bill data. Furthermore, there seems to be some anomaly in the bill amount when dealing with a credit score value of around $500,000. The Bill amount plateaus as if it hit an artificial ceiling. This might be important to address before doing any machine learning modeling.

Main lessons

* While modeling is a powerful tool for Data Analytics, it is essential to do an Exploratory Data Analysis before trying to model anything. This will give the team a better idea of what features might be the most important and which ones are irrelevant. Furthermore, any potential mistakes, outliers or anomalies in the data could be recognized early in the machine learning process, saving time, and computing power.
* In terms of powerful tools for EDA, pandas profiling tool is a great way to start and look at an overview of the entire data. Interesting relationships and findings can be further explored subsequently.
* Seaborn has a plethora of data visualizations which are far superior to what matplotlib library can do in the same amount of time. It automizes much of the coding and allows for faster data manipulation with less code.
* Shapash as data modeling tool allows for easy and quick visualization of the data model. It uses a webapp to create interactive graphs which can then be downloaded and used offline. I look forward to keep reading about it and hopefully incorporate it in the final report.

Recommendations

* The team must focus on detecting any outliers that might negatively impact the data modeling, specifically the plateau in the monthly bill amount at $500,000 credit.
* It is possible that many of the features are not important to the model and therefore might be best to identify them and remove them.
* Lastly, it is important to take into account the percentage of males and females recorded in the data, as long as gender is considered an essential feature for the machine learning algorithm.

Most of the hard work will be done by the machine learning algorithms, Random Forest Classifier and Regression. Refining these models and attempt to get accurate predictions will be the task of the data analytics team.