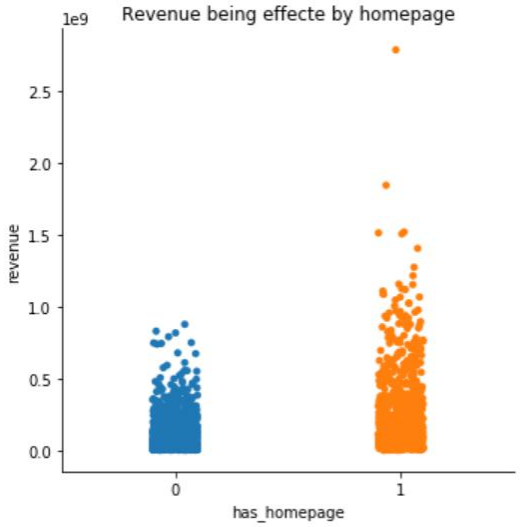
Report for Assignment 3

Data Cleaning :

For muti-variate linear regression, one of the most important questions is the process of choosing the features. I used the approach of data visualization via scatter plot to decide whether a particular feature is effecting the revenue or not.

Following is one example of a scatter plot of few successful and unsuccessful feature plots.

Figure 1 :



The presence of the homepage is effecting the revenue, hence we decided to include that in the feature list.

We used similar approach for the rest of the features and narrowed down on the following list of features to consider.

cast,crew,genres,homepage,production_companies, spoken_languages,tagline,release_date

Now the next issue at hand was that the data in the features was in the form of json.
The solution that was observed was to utilize json.loads and then read the data.

The next issue at hand was that the dataset for cast and crew is huge in terms of numbers and how to utilize it for data modelling.

Data Transformation :

Approach Followed	Observation
Sci-Kit One-Hot-Encoder	This approach failed because there were a lot of data inputs and the time limit of the assignment was not being met.
Using pandas.dummies	Pandas dummies would create new columns similar to one hot encoding but again there were a lot of data points
Frequency approach	Here we used the frequency of the singular data point checking and then encoding the ones that are frequently present in the data.

	This approach was useful and fast as well.
Fetching Day from date	Here we changed the date to date-time object, Then extracted day from it.
Normalizing the data for budget and revenue	Since the values of budget and revenue are huge hence we normalized the data set so that the values are to scale.
Failed Scenario - Combined Trained & Validation Data	I had concated training and validation data then after combined cleaning, split the data set to get back the same Train and Valid Data. This gave a PCC of 0.71 but not too sure whether the approach was right hence was dropped.
Failed Scenario - Difference between size of training data frame and validation data frame	When using One-Hot-Encoding then, we observed dimension mismatch for the train data frame and validation data frame. Since the size was huge hence iteration was not useful, and neither was re-index or any other approach.
Adjusting Revenue for Inflation	We adjusted the revenue for inflation and then fit the data model,but for some reason that also did nor prove useful.
Significance of PCC	It took me sometime to understand the significance of PCC

Part 1 : Regression Analysis

Question : To predict the revenue of the given movie based on selective features.

Choice of Regression utilized : ***Linear***

I utilized the linear regression to get the values, of PCC and MSR.
Also looked into RandomForestRegressor and other regression techniques like lasso etc.

At the end the best possible regressor were Linear & RandomForestRegressor.

Part 2 : Classification

In part 2 I relied on the same concepts of data cleaning and transformation.

The following table illustrates the approach that was followed :

Approach	*Observation*
KNN Classifier	*It gave the accuracy of .60*
GradientBoostClassifier	*This increased the accuracy and now it is 0.72*
KNeighboursClassifier	*Using this also improved the accuracy of the model.*

At the end the best possible one was GradientBoostClassifier with accuracy of 0.72