WEEKLY SUMMARY REPORTS

Week 1

Objective of this week is to find the relationship between loyalty points, age, remuneration, and spending scores. Reviews CSV file was loaded into a dataframe, data is checked for dimensions, null values. Language & platform columns were dropped from dataframe and assigned to new one. Column names were changed for easy access. Modified dataframe is saved as a new csv file.

Ordinary least squares (OLS) as a way to estimate a linear regression model and fit a linear equation to observed data. Statsmodels package is used to fit a simple linear regression model.

Linear Regression is calculated between spending vs loyalty, remuneration vs loyalty and age vs loyalty. Plots show a linearity between remuneration vs loyalty, spending vs loyalty but there is no linearity between age vs loyalty.

From the linear regression plots, we could figure out there is a close relation between remuneration, loyalty and spending score. Age and loyalty have no relations. Loyalty points is high when there is a less remuneration or more remuneration. Loyalty points increase when there is an increase in spending score.

Week 2

Objective of this week is to Use *k*-means clustering to identify the optimal number of clusters and then apply and plot the data using the created segments. Modified dataframe is loaded and , only remuneration and spending score columns are used for doing cluster analysis.

We'll begin with a simple scatterplot. So, we have our bivariate setting here, namely remuneration and spending score. A scatterplot is called using these settings. Next, we create a pair plot and see the relations between the settings, we could see a clear clustering between these settings. Perform k-means clustering. Elbow method is used to find the clustering, plotting the result showed clusters as 5 as ideal number. Silhouetted method was also used to confirm if the predicted number is correct, it also showed that with 5 cluster has sil score which is closer to 1.

Pair plots were created for cluster 5 & 6. With pair plot we could see clear distinction with cluster count of 5, though all clusters are not evenly distributed, most of them are.

In case of 6 cluster, we could see some overlapping. If we could see that spending score of average remuneration is constant, whereas one with low or high remunerations have varying spending scores.

This insight provides us that we can group the customers based on remunerations, and we can to targetted sales for the customers based on the clusters they belong.

Week 3

Objective of this week is this week is to find common words which form the positive and negative reviews using NLP.

Modified CSV file is used for this objective, the file is loaded as dataframe, data wrangling is done by removing null values, duplicates. Only review and summary are extracted in to a new data frame, Further cleaning of data is done by converting them into lowercase and removing punctuation marks. Duplicates are removed from both reviews and summary. WordCloud is created for both reviews and summary. Tokenization was performed for summary and reviews using nltk_tokenize libnrary. Frequency distribution is calculated and found there were many stop words. Stopwords were removed and worldcloud was plotted again. We could see more words related to the games like fun, game etc. Polarity was performed on each tokenized summary and review, to neutral, negative, positive. Histogram plot were created for summary and review, based on the analysis typically there is more neutral trends which is aligned towards positive.

Week 4

Objective for this week was to gather insights from scatterplots, histograms, and boxplots about the data set using R. necessary library was imported, tutle_sales.csv file is loaded into a dataframe. Descriptive statistics of the dataframe is done using View(), glimpse().

Scatter plots, Box plots and Histogram visualizations were plotted between NA_Sales vs Platform, EU_Sales Vs Platform, Global_Sales Vs Platform. Product was not considered because a product could use multiple platforms and have many generations, platform is one which can help in providing information on which platform is yielding to more sales. Histogram doesn't provide any insights when Platform is considered. Scatterplot and Box plot gives insights that there is some outliers for certain platform. Outliers when examined, we could not figure out if

there was mistake on sales data or specific sales were more at that instance due to other factors.

Week 5

Objective for the week is to explore normality of the data set. Loaded turtle_sales.csv into a dataframe, min, max, mean sales for all region are calculated. Aggreagation was doperfome by grouping it by product. Visualization was created for the data frames for EU_Sales, NA_Sales, Global_Sales. Normality is created using qqnorm, linear line drawn in this case we're comparing the shape of the data to the shape of a normal distribution. So, if there is a good correspondence between the two variables, if there were a perfect fit between the two then these points would lie exactly on a straight line. Shapiro-Wilk test is performed for hypothesis tests. The p-value is far less than 0.05 so we can conclude its not normal distribution. Skewness is calculated suggests positive skewneees. The kurtosis is a measure of this tailedness of the distribution. And as our benchmark for kurtosis, we use the normal distribution. So here you need to know that the normal has a kurtosis equal to three. The kurtosis is heigh value, which proves to be heavy tailed. Correlation is calculated between the sales, and it looks like its strong positive. Visualization is created using ggplot2 to show the relation between sales and product.

Week 6

Objective is to find relation between sales using single linear and multiple regression models and predict the values. Sales columns are filtered from the dataframe used before and single and multiple regression models is created. Sample values are generated and prediction is done using predict() with a confidence interval. The predicted values when compared with the observed values are close and falled between the range. One observation is the higher sales values are far away from the observed but the lower sales values are closer.