ASSIGNMENT ACTIVITY FOR COURSE THREE

Background/context of the business

This presentation is for Turtle Games, a game manufacturer and retailer. They manufacture and sell their own products, along with sourcing and selling products manufactured by other companies. Their product range includes books, board games, video games and toys. They have a global customer base and have a business objective of improving overall sales performance by utilising customer trends. I will be using two data set provided (turtle_review.csv and turtle_sales.csv).

In particular, the Game company wants to understand:

- How customers accumulate loyalty points.
- How useful are remuneration and spending scores data.
- Can social data be used in marketing campaigns.
- What is the impact on sales per product.
- The reliability of the data
- If there is any possible relationship(s) in sales between North America, Europe, and global sales.

And I will be exploring python and R as a tool to import, clean and manipulate the dataset provided. Techniques such as Linear regression, clustering with K-means and NLP to provide solution to some of the questions in making prediction, finding relationship, correlation and to obtain some descriptive statistics and present different graphical representations.

Analytical approach

For python, I lunch jupyter notebook, and upload the turtle_review.csv provided. After I opened a new notebook: Import the python libraries and packages.

import numpy as np, import pandas as pd, import matplotlib.pyplot as plt, import seaborn as sns, import statsmodels.api as sm, from statsmodels.formula.api import ols, import nltk, import os, import matplotlib.pyplot as plt, from wordcloud import WordCloud, from nltk.tokenize import word_tokenize, from nltk.probability import FreqDist, from nltk.corpus import stopwords, from textblob import TextBlob, from scipy.stats import norm All needed to perform the analysis on jupyter notebook.

I explored the data, using functions such info () and describe () to determine the number of rows and columns, missing values, metadata and descriptive statistics. Also, Data cleaning and manipulation were done.

In the process of the analysis, there is a need for new dataframe, this is done by creating subsets. This is well detailed in my ipynb files.

For R, I opened R studio and import tidyverse library which is package needed for exploring, cleaning the dataset and to plot graphs. After I import the data set (turtle_review.csv).

I import library(dplyr) for data manipulation.

Visualisation and insights

The chart I used was scatterplot, very good for numerical variable, histogram, boxplot these are very good when analysing normal distribution able to see if the data is normally skewed or positive/negatively skewed. Also, Shapiro-Wilk test was used to determine normality. Performing relationship among the variable, linear regression and multilinear regression was performed. Also, in performing clustering, number of K was determined using Elbow and Silhoutte methods to determine optimal number of clusters. Furthermore, performing NLP analysis, here dataset was pre-process, tokenization was performed, frequency distribution, polarization and sentiment analysis was performed.

Patterns and predictions

Question one answer

45% of the total variability of **loyalty point**, is explained by the variability of **spending score**. The P-value < (0.05), the set of variables

of the regression model are significant. if the spending score increases by 1 unit, the loyalty point will change by 33.0617 units.

38% of the total variability of **loyalty point**, is explained by the variability of **renumeration**. The P-value < (0.05), the set of variables of the regression model are significant. if the **renumeration** changes by 1 unit, the **loyalty point** will change by 34.1878 units.

2% of the total variability of **loyalty point**, is explained by the variability of **age**.The P-value (0.06) < (0.05), the variables of the regression model are not significant, if the **age** changes by 1 unit, the **loyalty point** will change by -4.0128 units.

Question two answer

Scatterplot and pairplot were used to determine if any correlation and possible clusters between **renumeration** and **spending score**. There is a correlation between the renumeration and spending scores of customers. looking at the result obtained when K = 5, 6 give the best result (groups) and distribution is better. The chart can be seen on the ipynb.

Question three answer

The horizontal bar chart help to reveal 15 most commonly used words online in ranking order. This plot shows that most review responses express a positive sentiment direction. Also, most customers are

neutral in their summary response, though there were few positive positive sentiment responses.

Question four answer

Scatterplot was created with a Y-variables (NA_Sales, EU_Sales and Global_Sales), the result shows same pattern and shape across the chart. Also, the result revealed high sales outlier. The histogram revealed the output shows the variables are highly skewed to the right. Boxplot revealed that the output of the three sales variables is skewed to the right and high sales outliers.

Question five answer

The histogram revealed the three sales plots are highly skewed to the right.

From the Q-Q plot, the data is pretty far from the straight line, it is not a perfect good fit for the data. Shapiro-Wilk test on all the sales data revealed evidence to reject the assumption of normality as the P-value < 0.05.

Furthermore, the distribution of the data set was evaluated with the skewness and kurtosis tests. The output indicated a kurtosis greater than 3, suggesting a heavy tail with extreme outliers. The skewness test indicated that the data is highly skewed to the right. There is a strong positive correlation relationship among the three variables (NA_Sales_sum, EU_Sale_sum, Global_Sales_sum).

Question six answer

From the multiple linear regression model, it can be deduced that North America sales and Europe sales is statistically significant as the Pvalue (0.0000000000000002) < 0.05. This means an increase in America sale and Europe sales will lead to a corresponding increase in Global sales. Furthermore, 93.47% of the variance in Global Sales can be explained by the predictors (North American and Europe sales). This model a very good fit. Moreover, to verify the accuracy, I compare the output of the predict () function with the actual dataset. The result revealed a relatively close output.