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 (e.g. customer reviews) be used in marketing campaigns.
- What is the impact on sales per product.
- The reliability of the data (e.g. normal distribution, Skewness, Kurtosis).
- 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. The DataFrame consists of 2000 rows and eleven columns, though two unnecessary columns were dropped, now left with nine. The metadata indicates that both columns have 300 non-null values, which means there is no missing data. Some data cleaning and manipution were perform in the course of the activities.

Analytical approach

I will be using python and R,

For python, I lunch jupyter notebook, and upload the turtle_review.csv provided. After I opened a new notebook to commence work: 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 these are needed to

perform the analysis on jupyter notebook. After I load the csv file and give a name.

I explored the data, get some info (), shape () and describe () to have a broader view about the dataset. Data cleaning and manipulation were done.

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

For R, I opened R studio and import tidyverse library which is package needed for exploring, cleaning the dataset and to plot graps. 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 is 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 preprocess,

tokenization was performed, frequency distribution, polarization and sentiment analysis was performed.

Patterns and predictions

From the analysis, the P-value is less than (0.05), therefore the is a significant relationship between independent variables(x) and dependent variable(y). From the plot, though line of best fit is not perfect. However, it is a positive upward trend pattern

This plot histogram of review sentiment score polarity expresses a positive sentiment direction.

The histogram of summary sentiment score polarity shows that most customers are neutral, however, the plot still expresses a positive sentiment direction.

The data points are not a perfect line of fit as it does not align perfectly with the line of good fit, it is too far.

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 also observation of outliers which really affect the distribution.

The correlation between NA_Sale_sum and Global_Sales_sum is 0.92. A strong positive correlation.

The correlation between EU_Sale_sum and Global_Sales_sum is 0.85. A strong positive correlation.

The correlation between NA_Sale_sum and EU_Sales_sum is 0.62. A positive correlation.