CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

---

Module Title:	Data Visualisation and Communication
Assessment Title:	CA2
Lecturer Name:	Marina Iantorno
Student Full Name:	Afia Serwaa
Student Number:	Sba23069
Assessment Due Date:	24/03/2024
Date of Submission:	24/03/2024

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

In this assessment I will discuss how an e-commerce platform called Dr. Axe concentrates on lifestyle and wellness products. Dr. Axe is a chiropractic doctor who focuses on nutrition, fitness , healthy recipes, and trending news on his platform. The platform is grouped into Health, Nutrition, Beauty, essential oils and fitness. This assessment will focus on how the company goal to understand its current market dynamics, customer demographics, product performance and reginal sales patters. I will discuss the missing values, outliers in the dataset. A descriptive statistic will be carried out to perform the needed Visualizations be demonstrated on the age, Customer life Value, and preference. Different Categories in the dataset will be visualized. For instance, the Session Duration Minutes a customer browses on the platform, the demographic Regions, the subscription status, and the product category. Lastly, I will use geo-visualization to identify the strong and week market trends in the United States Regions.

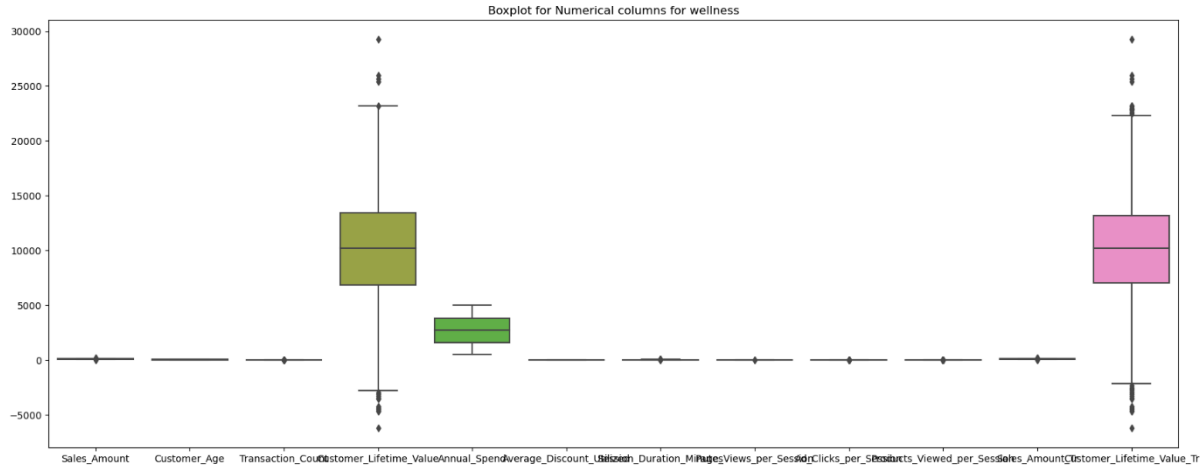
Assess the quality of the dataset. Identify and address any missing values, outliers, or inconsistencies.

First, to be able to evaluate the data quality. The relevant libraries were imported. These libraries can be found in Jupiter notebook. I imported this library import warnings to prevent warnings from disordering the output in the code. The dataset was load into the library. With the help of imported libraries. The dataset was named ‘dt_wellnesses. The dataset was described,(dt-wellness. Describe(include='all') this code will calculate the mean, mode, and the standard deviation of the variables in the dataset. dt-wellness was loaded. This will load the data frame columns. dt_wellness.info() will help me know the number of objects, float, and integer in the data frame.

The null values were checked with the help of this code dt_wellness.isnull().sum(). This code will print out the missing values in each column of the dataset. Sales Amount and Customer Lifetime was having missing values. Sales Amount is having (99) missing vales and Customer_Lifetime_value is having 100 missing data. The data was cleaned .

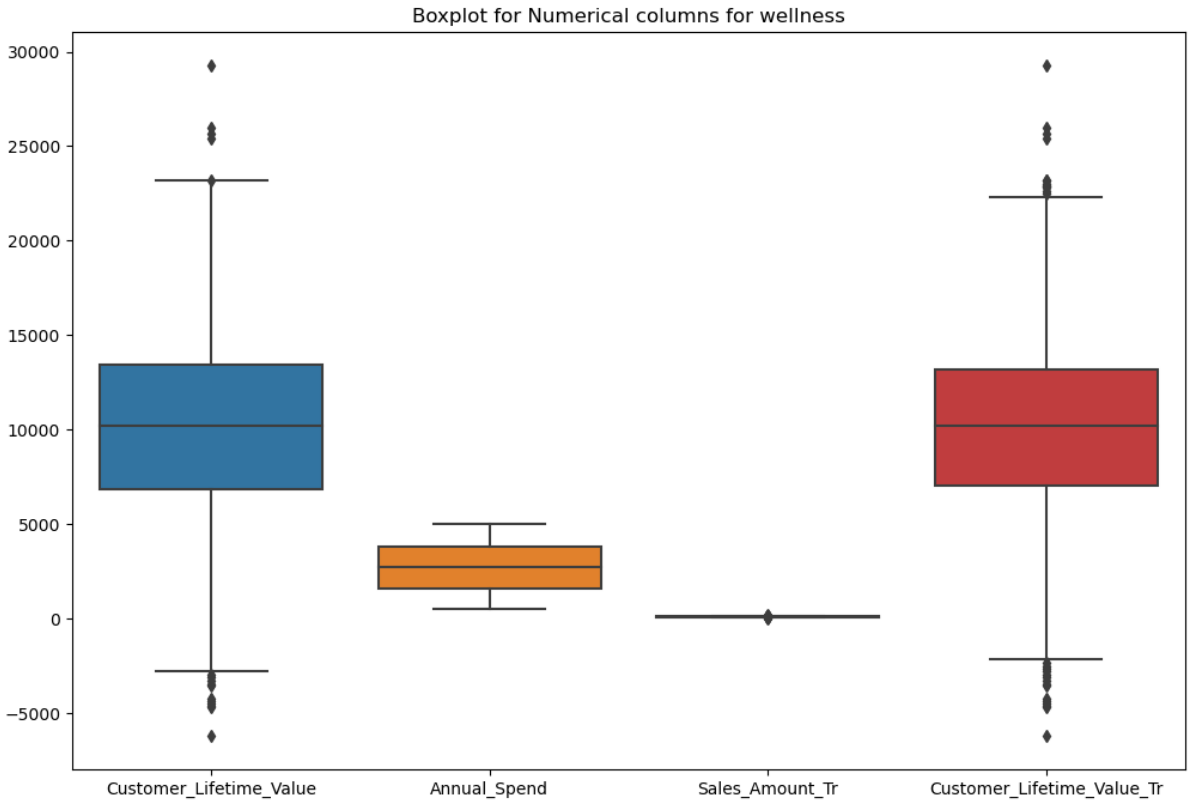
The mean method was used to treat the missing data for sales- amount and Customer_lifetime_value. I used the mean method because sales-amount and customer_life_value are numerical data, and the mean has a lower effect on the variability and the data after cleaning. The mean method is the simplest method to clean the data. The mean method can help to keep the original dataset , since the missing values is still there after analysing the treated data.

A new numerical code was created to be able to visualise all the numerical data. As seen in the Jupiter Notebook. Name for numerical data was “num_columns”. Box plot was created to be able to identify outlier through visualisation. Firstly, all the numerical data was visualised with boxplot. As seen below.



Few column names were selected to be able to see the visualization properly. That's is

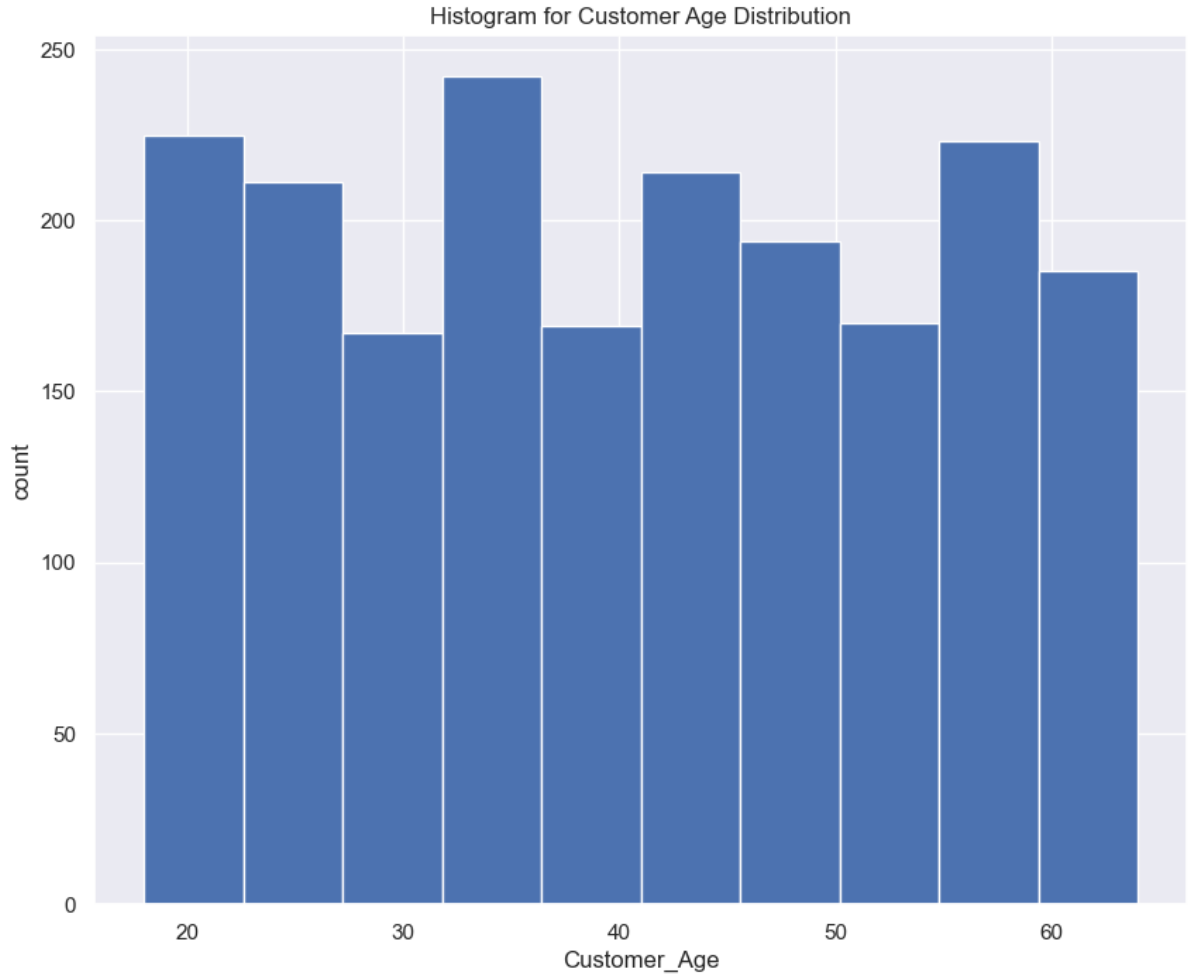
Customer Lifetime Value, Annual Spend, Sales Amount Tr, 'Customer_Lifetime_Value_Tr' was visualized using box plot. As code seen in Jupyter Notebook.



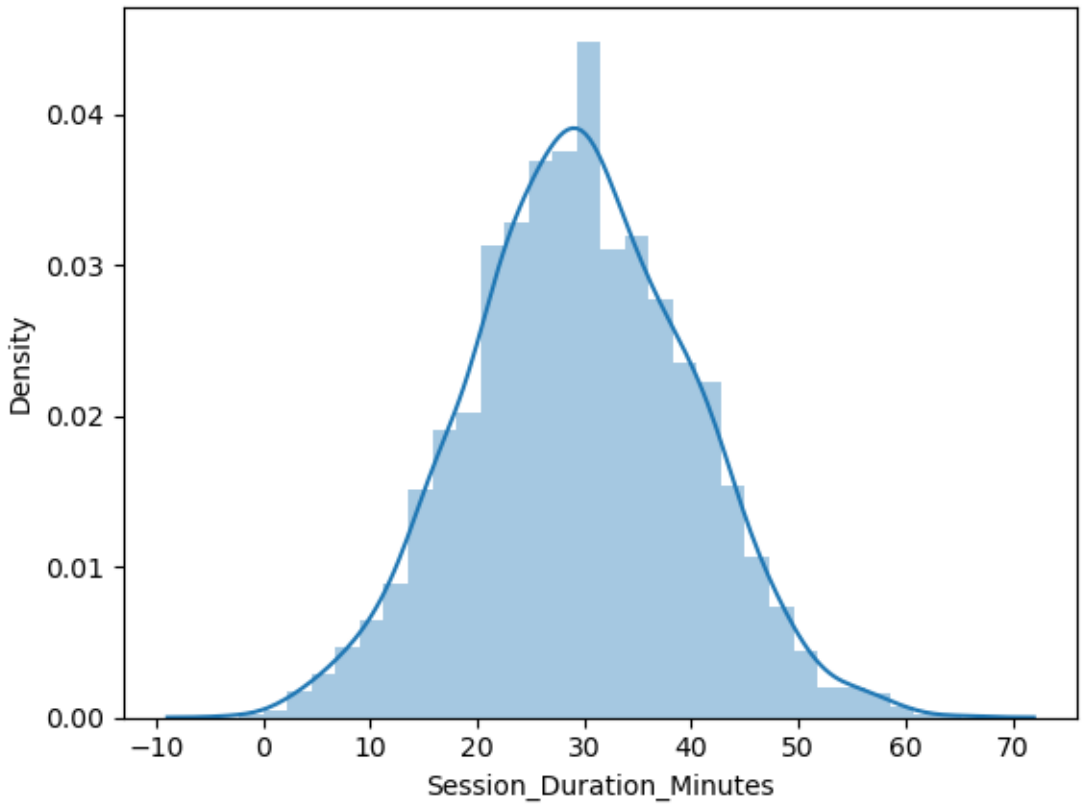
As seen in the above graph before Customer lifetime value was treated there was outliers of greater than 22000. Same as when Customer lifetime Value was treated. The outlier was above 22000. Annual spending has no outliers. A code was created to identify outliers by code.

Conduct descriptive statistical analysis to understand the current scenario and perform the necessary visualisations. This includes, but not limited to, demographic distribution of the customer base, age, gender, preferences, etc.

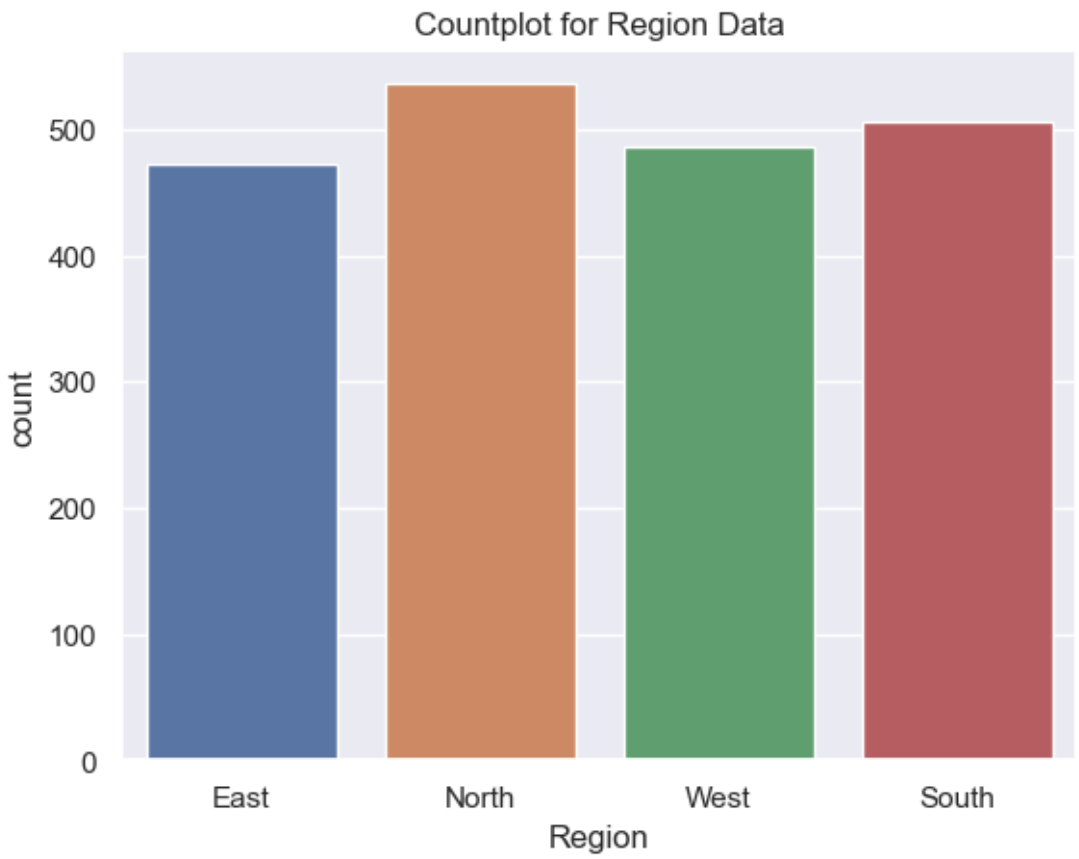
Firstly, a descriptive statistic was performed on customer-Age. To be able to identify the average age who uses the platform. As seen in the diagram above histogram was used to be able to identify the popular age which is age 35years. Followed by age 55years. The least years is age 30 and 40 as showed in the graph.



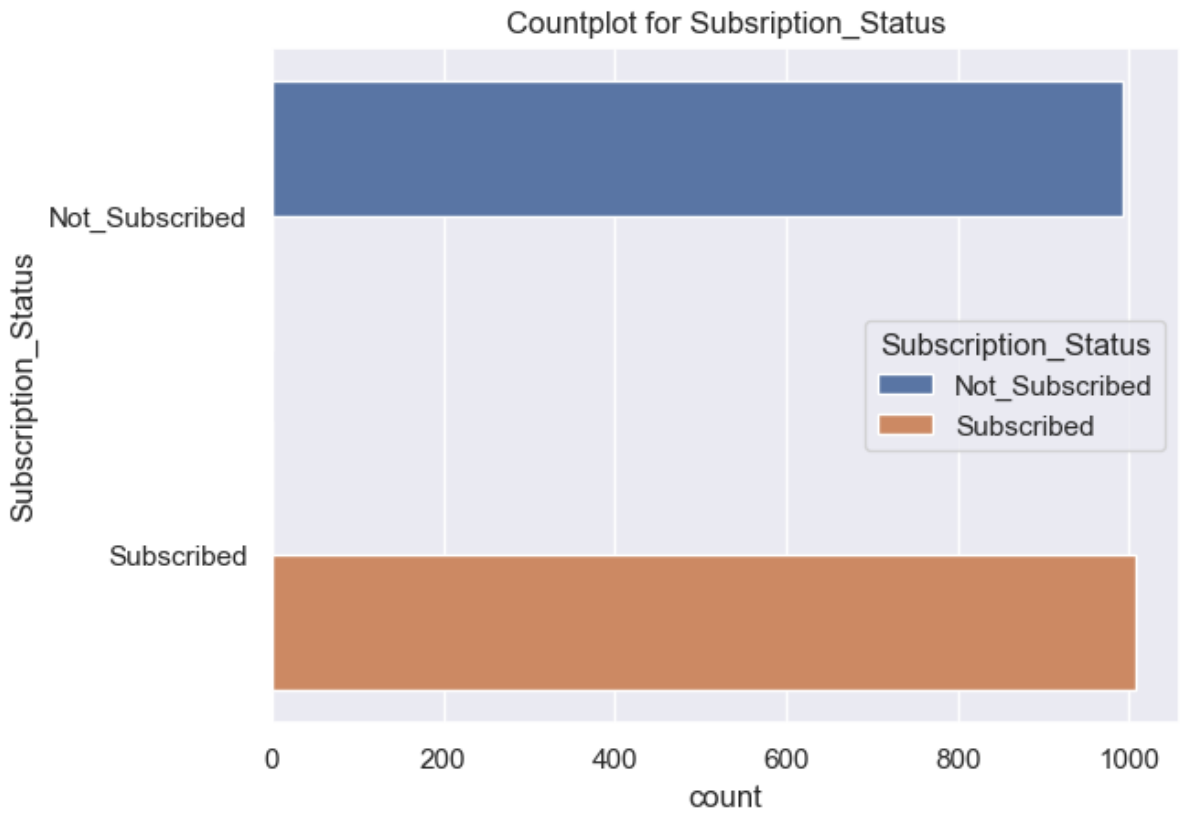
Statistics of Session Duration in minutes was described in the dataset. when the describe code was used as seen in Jupyter notebook the average minutes spend on the platform was 29.5minutes. see below a visualized graph in distplot. On the above Graph, on average customers spent 30 minutes on the platform.

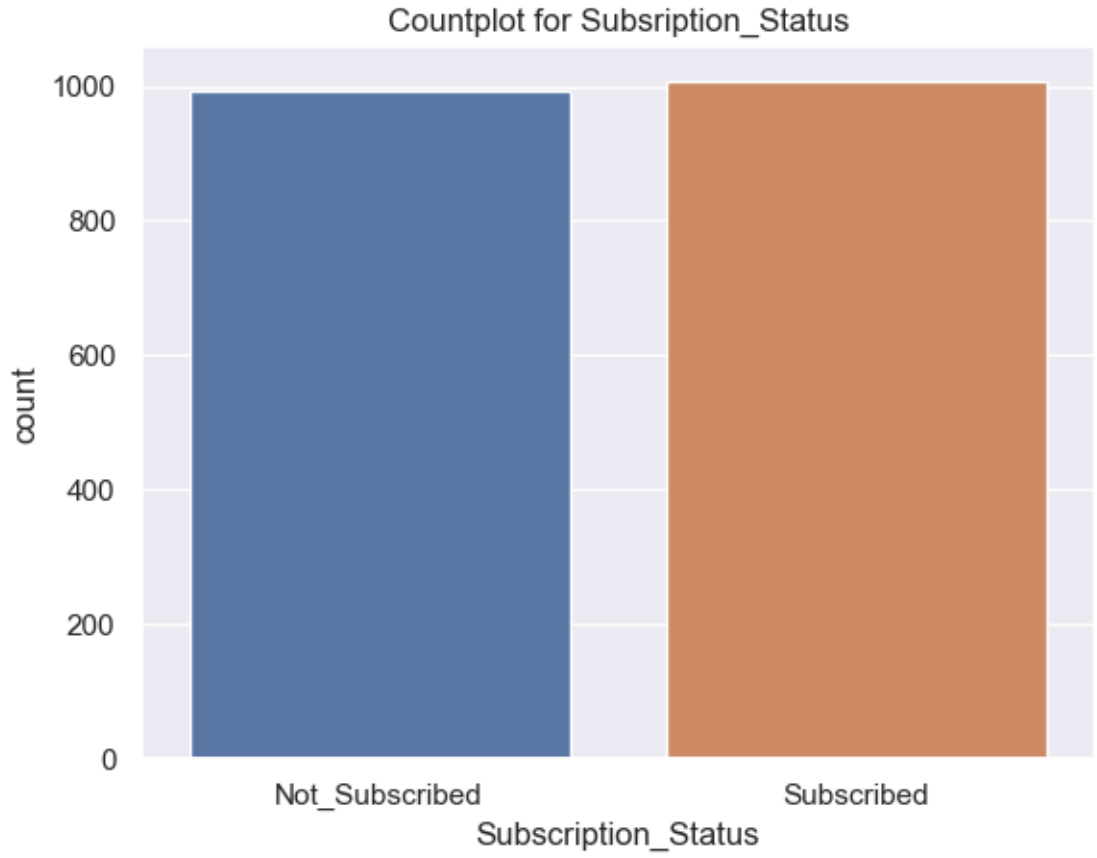


Statistics for Region was analysis. Dt wellness Tr .describe(). Was used to describe the data. North region was the most Region a customer purchases an item. The north Region has the highest purchase. Followed by the South Region, and West and East. See below the graph.

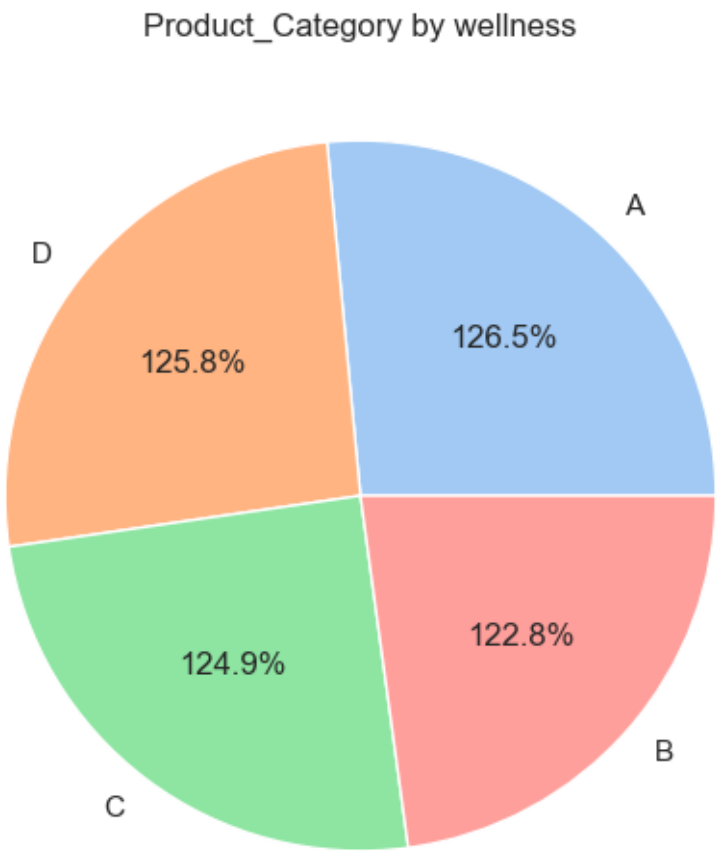


Subscription Status of the customer was checked through visualization of count plot. The subscription status was selected to be able to identify the trends weather a customer has subscribe to the platform or not. As shown in the graph the number of subscribed customers is the highest as compared to the not subscribed customers.


Countplot for Subsription_Status
Not_Subscribed
Subscribed
Subscription_Status
Subscription_Status
Not_Subscribed
Subscribed
0
200
400
600
800
1000
count

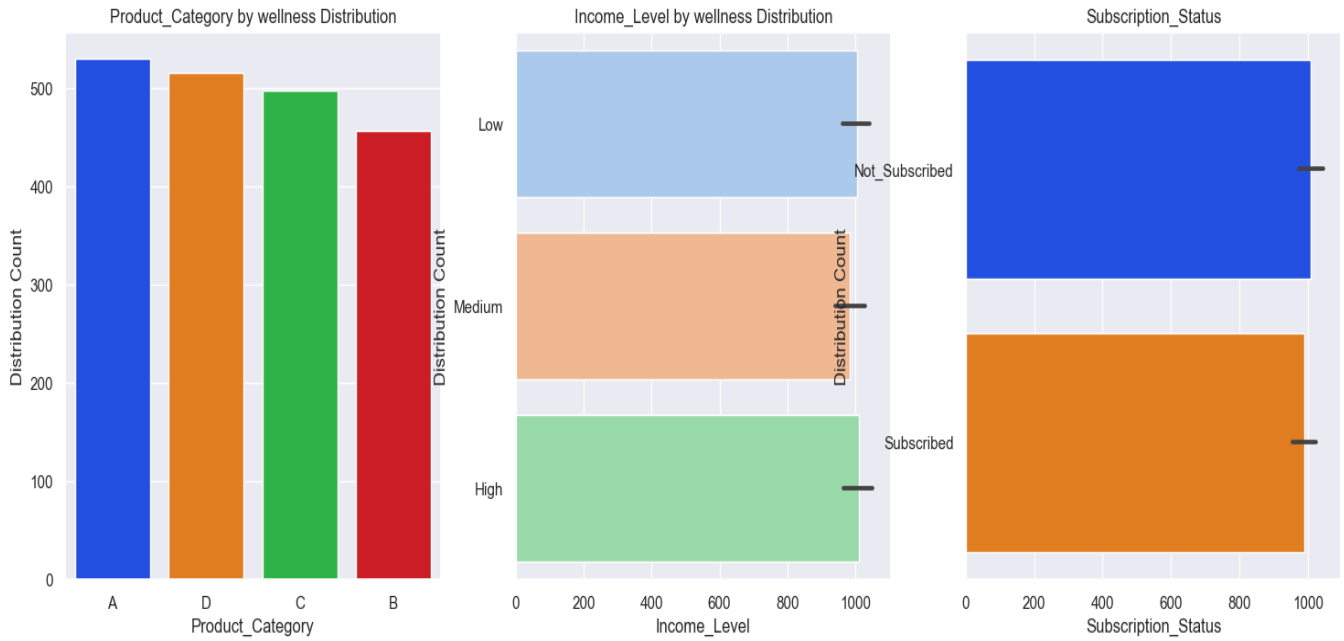

Countplot for Subsription_Status
1000
800
600
400
200
0
count
Not_Subscribed
Subscribed
Subscription_Status

Product Category by wellness was described and analysed by the use of pie chart. The products is divided in various types, that is type A,B,C,D. Category A has the highest data purchased of 126.5%. Followed by D, which has 125.8%, C has 124.9% and laslty Category B has the lowest as dipicted in the graph.

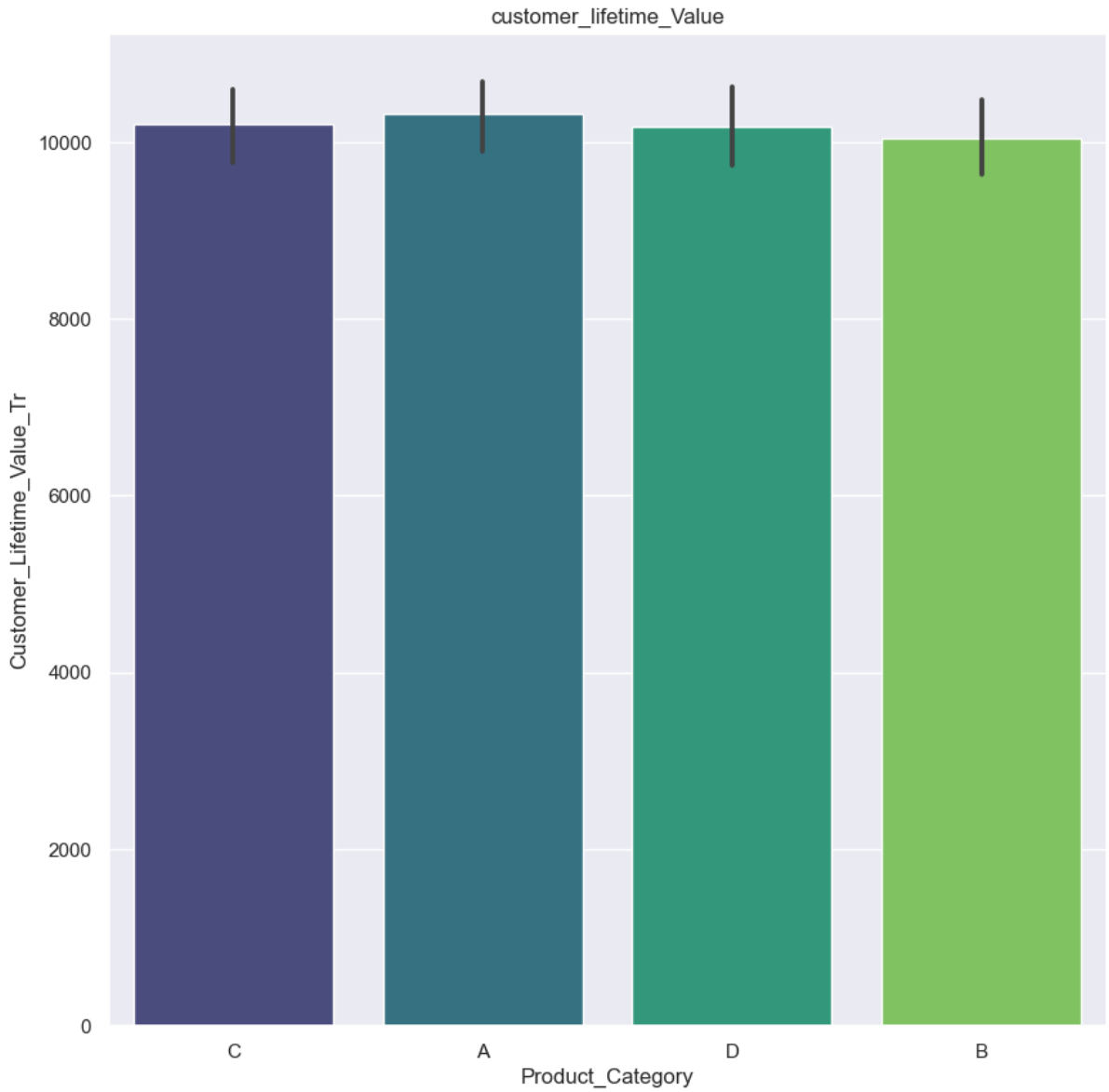


Use visualizations to compare the performance of different product categories. Highlight top□performing products and categories, as well as those underperforming.

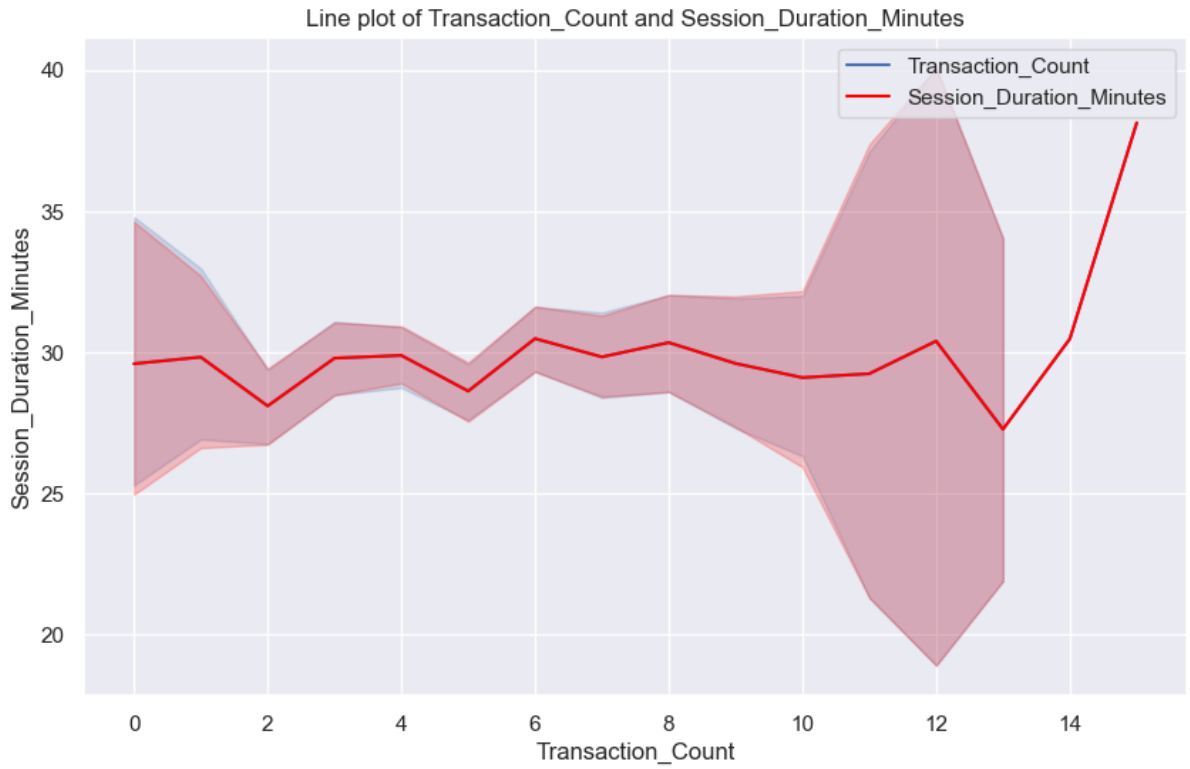
Product category, Income level ,and subscription status was described and visualised in a single axis. To identify how these categories is distributed. Looking at the graph there is high low-income customers on the platform. On the subscription status, the non-subscription status is higher than the subscription status. The category products B has the lowest under performance in the product category.



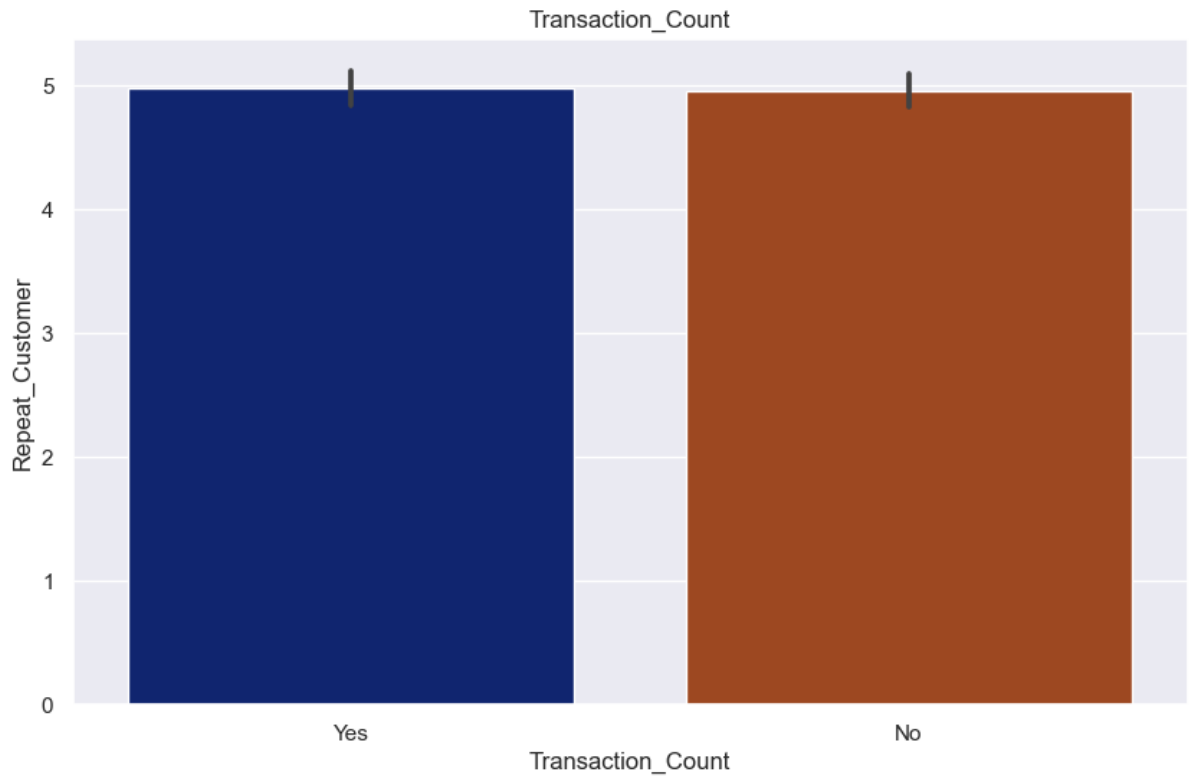
Customer life time and product category was compared and analysised. The customer lifetime has a highest profit of 11000 and this was able to achive because the popular product category is A . The graph shows the relationship between customer life time value and product category.



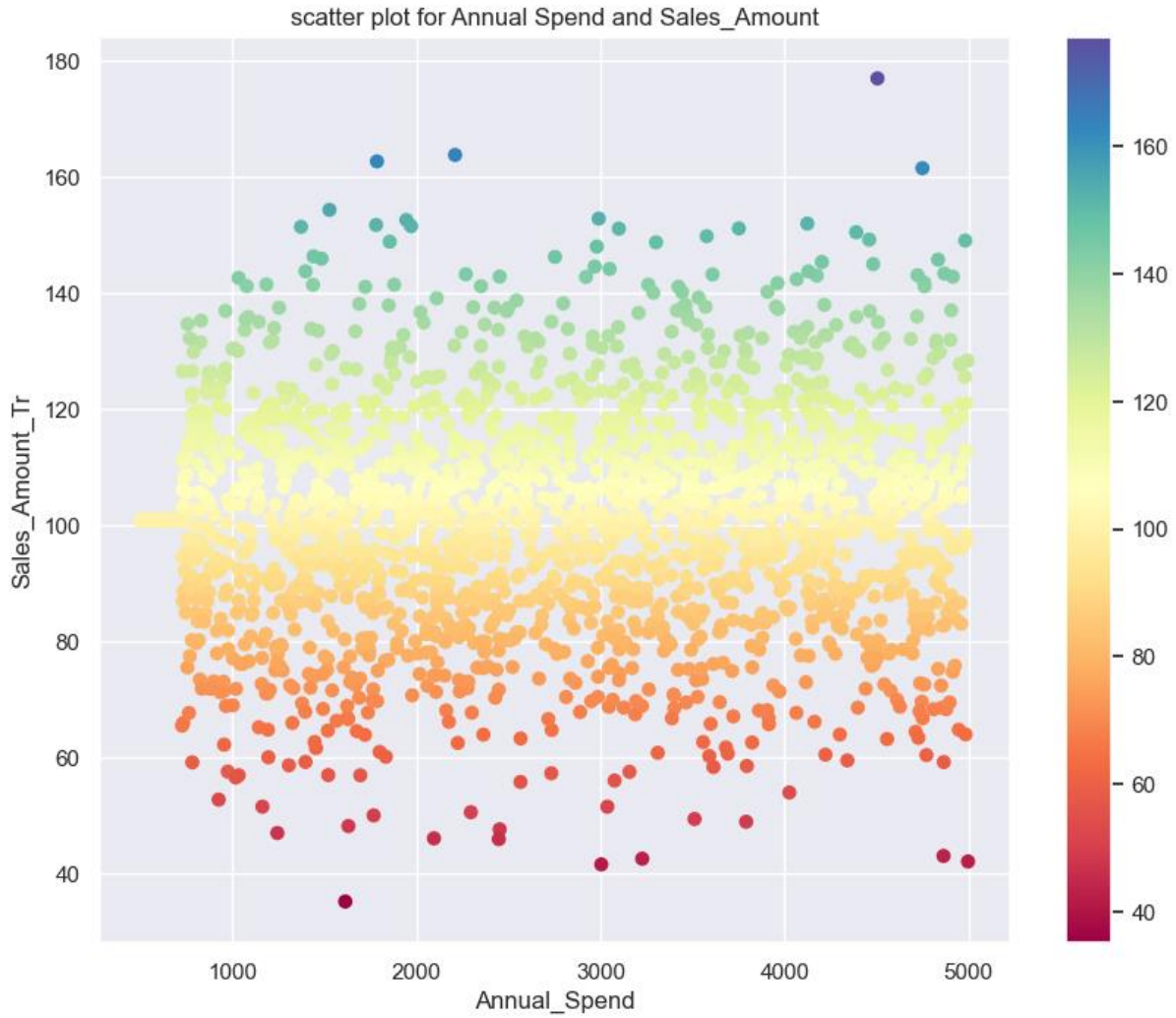
A line plot was created to compare the transaction count and session duration minutes. Transaction count has 14 counts as the highest. And session duration of 40 mins as the highest.



Transaction count and a Repeat customer was created to compare how much a customer purchase an item on the platform. 'No' has the highest repeated customer as found in descriptive statistics done by using a code as seen in the Jupiter Notebook.

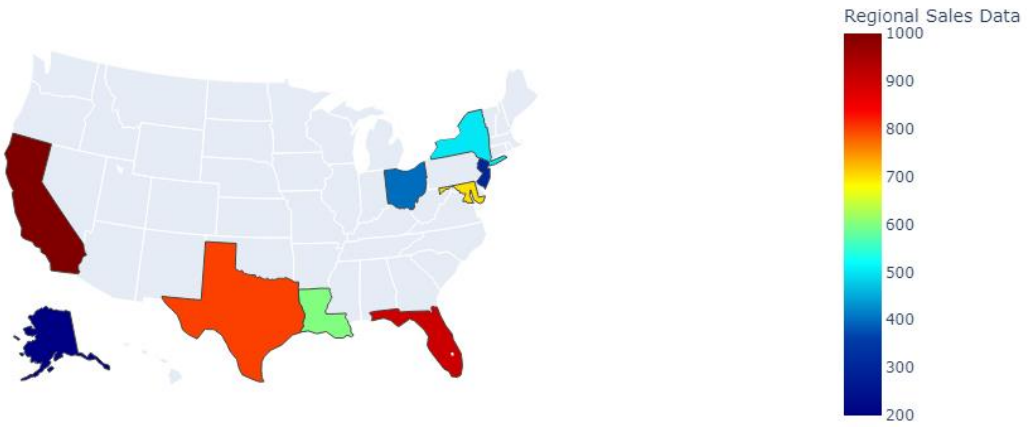


Scatter plot was created for Annual pend and sales Amount as seen below. Based on the graph below and the statistics performed in Jupyter Notebook . The samallest amount in sales Amount was 40 as showed in the graph below . The minimum spend annually was 500. In this Graph there was no correlation between Sales Amount and Annual spend as seen in the graph . the graph scattered evenly.



Analyse regional sales data to identify strong and weak markets. You can use geo-visualizations at this point.

Visualisation geo-visualisation was created using choropleth maps in various regions in the United States, that is the Alaska, 'New Jersey, Ohio, New York, Louisiana, Maryland, Texas, Florida, California. The code can be found in Jupiter Notebook. The geo visualisation is an interactive graph as seen in Jupyter notebook. California has the highest market sales, followed by Florida and Texas. As seen in the picture Alaska has the lowest market sales, New Jersey has the second lowest in market sales.



By comparing these regions, the Dr. Axe platform will be able to see the various Regions underperforming. This will help to be able to implements any further marketing strategies. Such as giving promotions, discounts on the weak Regions. To level up the sales amount. The platform must be continuous be observed, to be able to detect the weak and product Category been purchased the most on the website. Regions with strong sales, the company must continue the market strategies to be able to grow further.

REFERENCE:

Baker, Dan. "I've Built a Public World Atlas with 2,500 Datasets to Explore." *Medium*, 24 Mar. 2024, towardsdatascience.com/ive-built-a-public-world-atlas-with-2-500-datasets-to-explore-8b9ae799e345. Accessed 24 Mar. 2024.

plotly graphs. "Mapbox County Choropleth." *Plotly.com*, 24 Mar. 2024, plotly.com/python/mapbox-county-choropleth/. Accessed 24 Mar. 2024.

Axe, Josh. "Dr. Axe | Health and Fitness News, Recipes, Natural Remedies." *Dr. Axe*, 24 Mar. 2024, draxe.com/. Accessed 24 Mar. 2024.

Brownlee, Jason. "4 Automatic Outlier Detection Algorithms in Python." *Machine Learning Mastery*, 24 Mar. 2024, machinelearningmastery.com/model-based-outlier-detection-and-removal-in-python/. Accessed 24 Mar. 2024.