**Name: Aqsa Younus**

**ERP: 18571**

**Dataset: Online Shoppers Purchasing Intention**

<https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset>

**Domain Knowledge**

According to Wikipedia, web analytics is the measurement, collection, analysis, and reporting of web data to understand and optimize web usage. Web analytics is not just a process for measuring web traffic but can be used as a tool for business and market research and assess and improve website effectiveness.

Similarly, this data is to keep track of how effective the website is in persuading people to buy products from this ‌business website.

Keeping this domain knowledge in mind we will more forward with the analysis

**Content**

The dataset comprises 10 numerical and 8 categorical attributes.

The 'Revenue' attribute can be used as the class label.

**Problem Statement**

* To explore web analytics and business insights via websites
* To understand how persuasive the website is for enabling users to shop
* To illustrate key insights and user flow of the dataset of this specific website
* To highlight the aspects of e-commerce websites and relating it to users' shopping interest

**KPIs Terminology Definitions**

*(important for those who are not familiar with the web analytic terms)*

**Product Related:** represents the number of different ‌pages visited by the visitor in that session.

**Bounce Rate:** defined as the percentage of visitors that leave a webpage without taking an action.

**Exit Rates:** metric referring to the number of times visitors have left a site from a single page.

**Page Values:** the average value for a page that a user visited before landing on the goal page or completing an Ecommerce transaction.

**Special Day:** A feature shows the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be completed with transactions. The value of this attribute is determined by considering the dynamics of e-commerce, such as the duration between the order date and delivery date.

**Informational Duration:** total time spent on informational pages

**Possible Dashboard Audience**

* The Business who owns this particular website

**Data Transformation/Wrangling**

Because the data is large. I did extensive data wrangling through Python because it makes it much easier to understand the data and its nuances. I added the ipython notebook. The in-depth details are written on the markdown (objectives/assumptions/reasoning). Below, I summarise what I have done to transform the data.

* Divided the wrangling into multiple steps
* Checked for missing values
* Created msno matrix as well
* Changed data type of columns where needed
* Looked into data description of potential KPIs
* Created a heat map to check for correlation in the columns
* Deleted columns that barely had any correlation
* Checked for unique values and count of potential dimensions to eliminate incorrect inputs
* Created histograms and box plots for anomaly detection
* Since there is only one csv file, no joining/union needed to be done on tableau.
* The data does not have any row removal requirements as the percentage of values that are 0 is low and the inputs that are 0 are of significance.

*(further details in the ipython notebook)*

**Potential Dimensions and KPIs**

*“The KPI acronym stands for key performance indicator—it’s a metric that measures how projects, individuals, departments or businesses perform in terms of strategic goals and objectives. KPIs are a way for stakeholders to see if they’re making progress or if the business is on track. A dimension is a structure that categorises facts and measures in order to enable users to answer business questions.”*

One thing to note is during the data wrangling, I dropped a few columns that were not significant. However, that does not mean that all the columns that I have kept would be definitely considered as KPIs and Dimensions. In the end, it depends on the business what they are intending to see as KPIs.

**Dimensions**

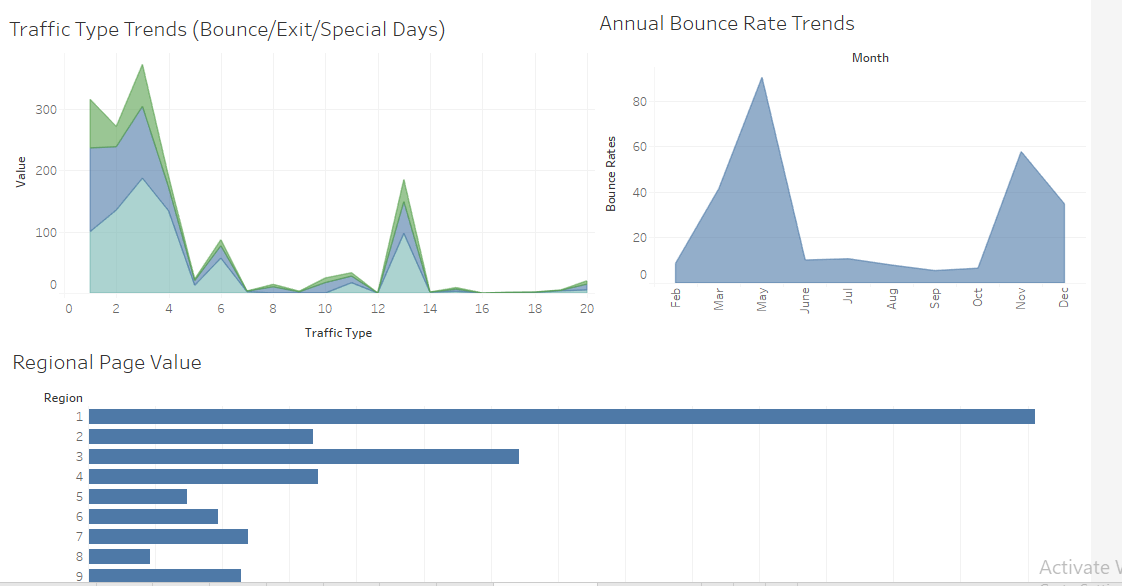
* Month
* Browser
* Region
* Traffic Type
* Visitor Type
* Revenue

**KPIs**

* Bounce Rates
* Exit Rates
* Page Values
* Special Day
* Informational Duration
* Product Related

**Charts and Explanation**

*(because time was limited, I did not get the time to work on the aesthetic of the charts. I usually prefer a black background with white axis and bold colours. Did not have the opportunity to do that this time)*



**Traffic Type Trends (Bounce/Exit/Special Days)**

The insight it provides:

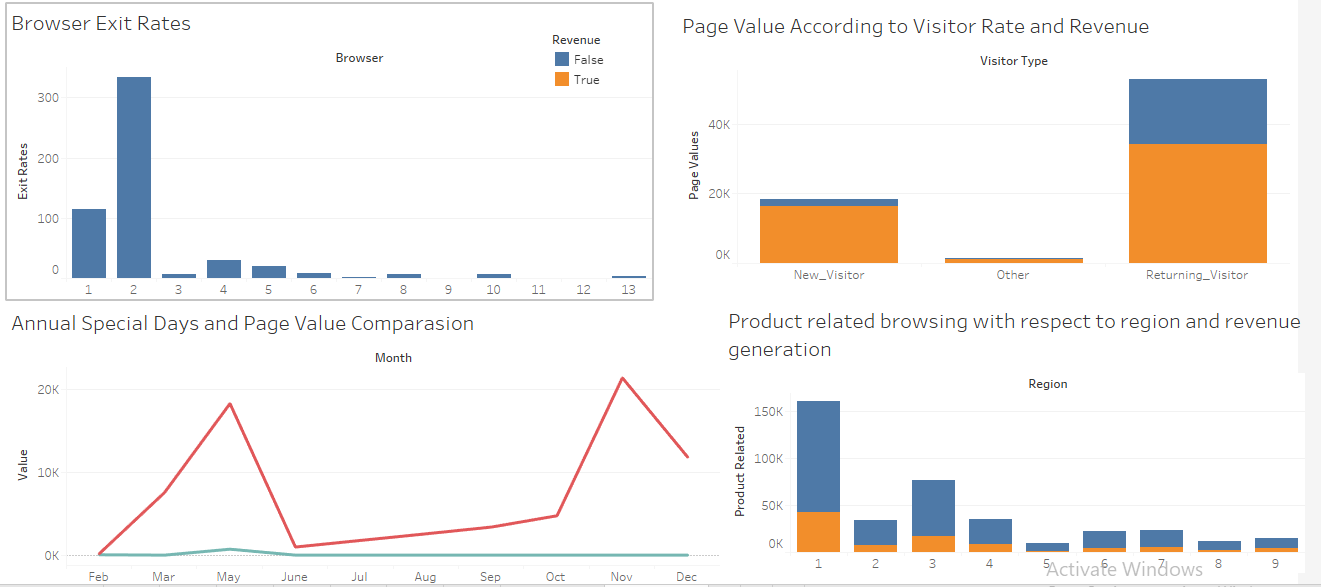
* Comparison of bounce/exit/special days according to the type of traffic it was divided into
* Able to understand user traffic trends with respect to their movement within the ecommerce website
* Able to understand the significance of special days with respect to traffic type whether a specific traffic type shows more inclination towards special days transactions or not.

**Annual Bounce Rate Trends**

* The lower the bounce rate, the better it is for the website, we are able to understand the trend of bounce rates
* Can further delve into understanding reasons behind the peaks in bounce rates

**Regional Page Value**

* Understanding the importance of page value according to the region.
* Helpful Determining which regions require improvement in order to boost page value.



**Browser Exit Rates**

The insight it provides:

* Significance of browser and exit rates
* Trend shows one specific browser has the most exit rates
* Able to understand the root cause behind the odd trend
* Essentially asks the question “*Is our page effective in gathering potential customers on a specific browser?”*

**Annual Special Days and Page Value Comparison**

The insight it provides:

* Tries to show whether special days has any form of impact on page value
* Gives annual insights on the trend of the page value comparison and special days
* Gives the business an insight on their own data collection *“why are there no special days input within the data after June?”*

**Page Value According to Visitor Type and Revenue**

* We can understand the importance of page value according to visitor type its significance in revenue generation
* Shows how new visitor are also contributing half as much as previous customers
* Can drill down more into subcategories of “other” customers as they seem to also generate page value and revenue for the business.

**Product Related Browsing with Respect to Region and Revenue Generation**

* Shows link between product related browsing, region and revenue generation
* Gives insights on window shoppers and how they contribute to revenue generation
* Also illustrates which region has the most revenue generation and also which region do they belong to the most.

**Potential Analysis**

* What was the bounce rate trend throughout the year?
  + *The bounce rate shows two peaks May and Nov, where the bounce rates were above 90 and 50 respectively. Other than that, it showed a fairly linear trend throughout the year.*
* Which region gathered the most page value? Which one had the least?
  + *Region one had the most values whereas region 8 has gathered the least page value*
* What browser has the most exit rates?
  + *We can see that there is no set trend within the browser and exit rates. However, browser 2 has a very significant exit rate compared to the rest.*

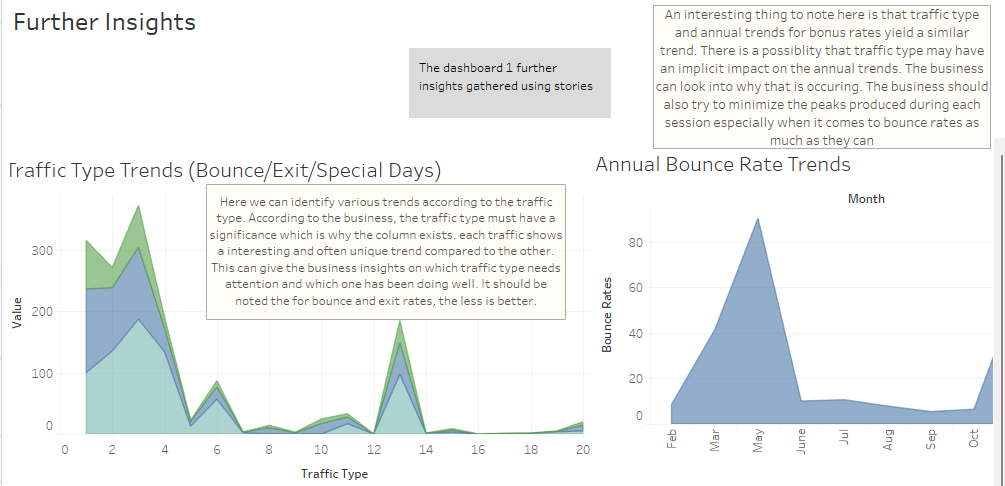
* Explain the trend of Special days in comparison with page value throughout the year
  + *Consistent to the correlation on our heat maps, the special days and page values do not have a very significant relation. The page value has a strikingly interesting trend, similar to that of bounce rates. Whereas the special days fall flat right after May.*
* Which visitor type generates the most page value?
  + *Returning visitors generates the most page value. However, new customers also show a significant contribution to page value. These page values are marked at revenue generation in the data set as well.*
* Which traffic type has the most exit rates? Which one has the least?
  + *Traffic three has interestingly the highest exit/bounce and special days associated with it. Whereas traffic type 17 has the least.*
* Region that has the most product related browsing with respect to revenue generation
  + *Product related browsing is seen in region one the most, however the product related browsing in comparison has least contribution in revenue generation. Many people who do product related browsing are probably online window shoppers.*

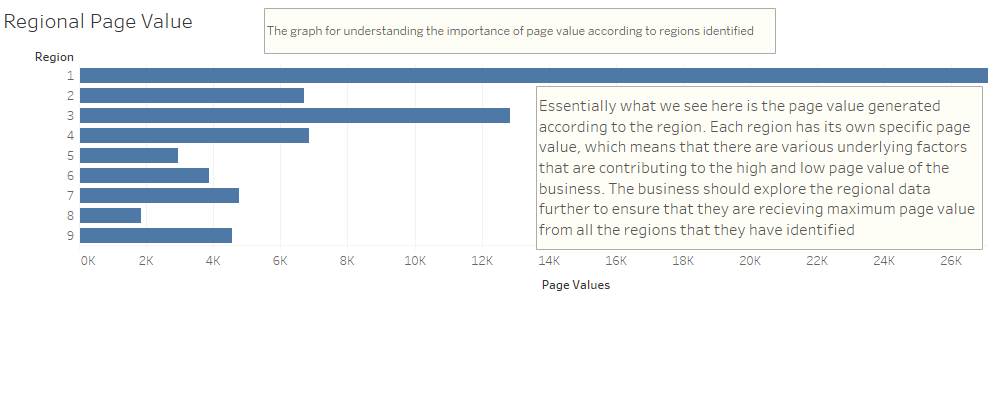
**Further Insights**

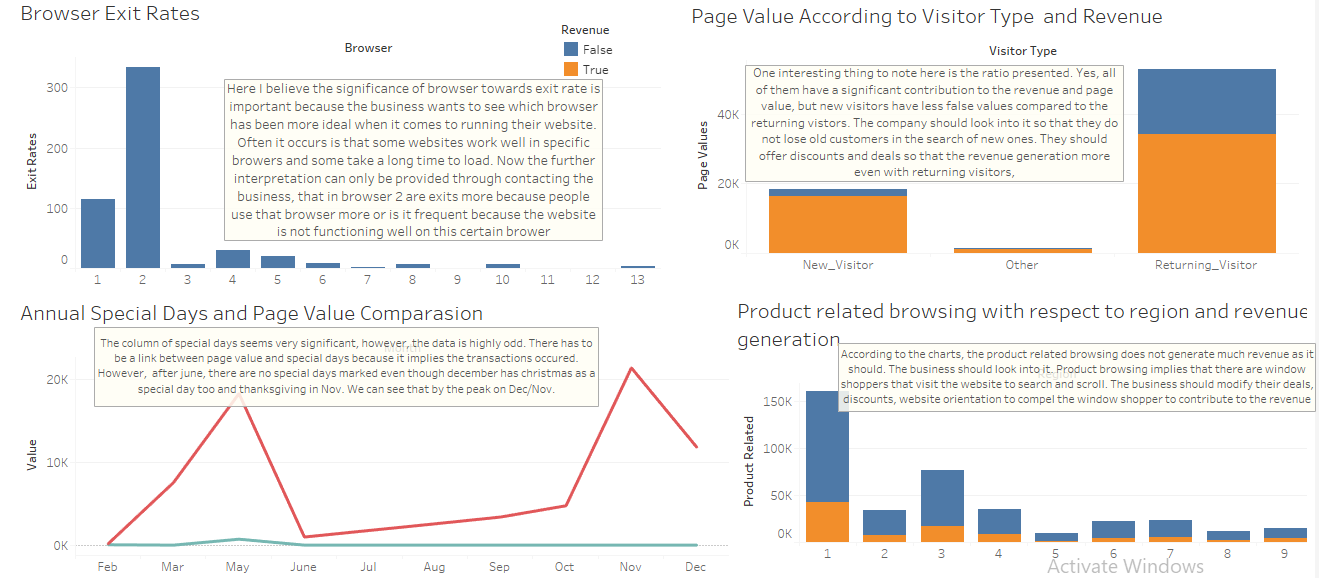
* According to the heat map generated through python, it showed odd correlation between the columns. It was almost as if the dataset was specifically prepared for machine learning. I found that out after I had cleaned and wrangled the data. These correlations in a way, may help machine learning algorithms, but greatly hinders the work of a data analyst.
* A lot of the work is based on research and assumptions since a lot of categorical data is numbered rather than with their names, therefore concealing the appropriate use of these columns in data analytics. This would require continuous back and forth discussions with the business to ensure what these numbers imply.
* The business, if they have added a browser as their column, surely they want to understand the significance of the browser on the consumer. They could look into why a specific browser has the most exit rather, whether the other browsers have it low because people do not use it as much. They can also look into the relation between their website and the browser their customers use.
* Similarly, they can also look into the bounce rates and try to convert it into conversion rates. They can look into why there are 2 peaks in the bounce rates and try to bring them lower in the upcoming year.
* They should also look into why there is barely any input on special days. It seems like a significant column as it would tell the transactions that occurred close to the special days so that they are able to design their marketing campaigns accordingly.
* There is also a very interesting link between bounce, exit and page value. The peaks are very similar. The business can look into why despite having high peaks in bounce and exit rates, the company’s page value has the peaks for page value as well.
* The regional data is essential as it provides insight on which region is doing well in terms of gathering potential customers. The business can further provide data on regions which can be broken down and understood as to which region needs strategic changes in marketing or not.

The best part about the dataset is that all visualisations are interlinked with two aspects: marketing and human psyche. Any change in marketing strategies can impact the trends of these graphs. Similarly, the independent variable, the human psyche, can cause shifts in the graphs as well. Some people surf the web because of various emotional states. The goal of the business would be to ensure that they gain maximum attention from the user.

*(referring to the stories in the book is ideal because the screenshots have bad pixel quality)*

**





**Final Thought/Opinions**

**About the Data:**

The data was quite interesting. The business using these insights can dive deeper into the psyche of the consumer to address their behaviour. They can look at the indirect links that caused the user to exit pages or bounce off a page and think about bringing about a change on their website and optimise it accordingly. The only issue with the data is that complex graphs could not be created because of correlations which was not ideal according to the data analysis standards. Furthermore, the data is originally compiled for machine learning purposes. If it was for data analytics, then the revenue would have been a float number and would be used as a KPI, which would have given detailed insights on the link between revenue and web analytics.

And as aforementioned, the dataset is created mainly for the purpose of machine learning which is why many of the dimensions are marked with numbers rather than by their categorical names. As a data analyst, I would prefer having the dimensions such as region, traffic type, e.t.c according to their name as it would help me provide more valuable insights to the data.

**About the Final:**

The final made me appreciate the importance of time. I realised how much time I take to carefully go through each dataset provided by the instructor and do a thorough research on the set before selecting it and creating aesthetically pleasing dashboards. Because of the little time I had, I felt like my command and understanding towards the data was not as much as it should have been. However, I tried my best to remain consistent with the detailed analysis I have provided throughout the semester for each dataset. Data analytics is truly time intensive work, and it needs a lot of research and dedication to ensure that you are able to explain to the businesses what analysis you are able to produce.

**Reference(s):**

<https://en.wikipedia.org/wiki/Web_analytics>

<https://backlinko.com/hub/seo/bounce-rate>

[https://support.google.com/analytics/answer/2695658?hl=en#:~:text=Page%20Value%20is%20the%20average,more%20to%20your%20site's%20revenue](https://support.google.com/analytics/answer/2695658?hl=en" \l ":~:text=Page Value is the average,more to your site's revenue).