**R-01254238 Data Engineer/Analyst - Assessment**

**Analyses and Assumptions**

I will include all the file (required and analysis) in a zip file

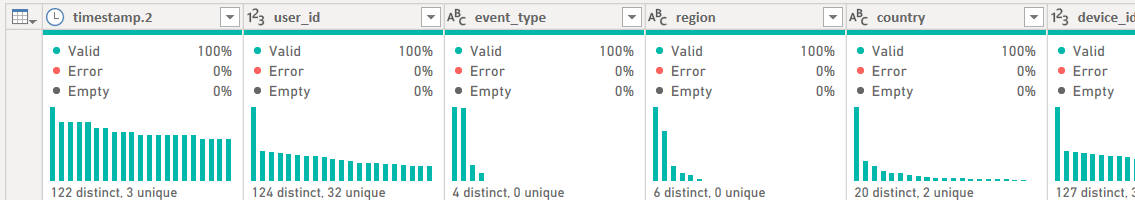
Sai Siva Prabhu Teja Kondapalli

**OBSERVATIONS:**

* Data Ranges - Total user – 3077, records with missing country – 115 , date – 1, time – (12AM -12 PM), Regions – 6 , Referred – 4 (without NA)
* To begin with I have noticed that there are empties or blanks in the columns/dimensions – Country and Referrer
  + Country - only the users under region ‘EMEA’ have blanks for few of the users.
  + Referrer – 59% of the users have blanks in referrer (this could mean that they did use any referrer, or the data is not captured)
* Time stamp column has date and time for each record but if we clearly notice the date same for all the records, which indicated that this information is collected for single day and records spanning from 12 AM – 12 PM. In this case we don’t have much of use with the date in time stamp column. But if we have different or dates over the week or month or year, it could help use find trends with respect to dates.

**Assumption & Implementations:**

* As soon as I saw the data in the excel, I checked for blanks and found two columns with blanks (country, referrer). As almost 60% of referrer is blank I thought it might be useless and I thought I should remove it and if possible replace or fill the missing countries, a habit from working on data science projects where I fill the missing values with mean or mode most of the time if there are very few % of missing values.
  + Both of my initial assumptions were wrong, country is a categorical variable which represents demographic, and I can’t just replace missing values like I can do with the numerical values. I should not do that because it could lead to wrong metrics (false information).
    - But there is one thing that I have found , there is this specific user - 2442 where I has a country along with region and device id, but 7 of his record doesn’t contain country, so in sql I kinda added the country for those 7 records using update statement. I tried to verify similarly for other user where country is missing but I couldn’t find any useful information based on region and device id, so I decided no to go head with the replacement in others case.
    - Query to find user with country **select distinct(USER\_ID) from my\_data where country is null**
  + After careful analysis this referrer could turn into a useful metric to see how many of them had referred from different sources, so that I could help company to identify if they wanna improve referrers from a specific categories. For the visualization I assumed that blanks in referred as missing data (data that is not collected). so, under this assumption I didn’t include that in my visualization by filtering it out.
* I removed the duplicates before calculating the users at each stage of the funnel because I am assumed the I need to find only the users count (how many of them are there) but not how many times they done it. So, to filter duplicated by using event type and user id (set them to be one for each pair and not more than that).
* I made an assumption that a user should or must have all the four stage to be counted as a successful conversion rate , which I was doing some analysis using SQL. But that’s a wrong assumption because there are users with out first step , but have rest of the steps (example user 13). This lead to wrong summarization in my initial analysis of the grouping and aggregation. But once I found out the reason by working with python and sql and power bi , I was able to figure it out the use case or flow cases. This helped me understand that I am in correct direction of analysis.
* I used SSMS – import flat file option to import csv file into sql database table and used for writing queries.
* My formula assumption for calculating the conversion rates are as follows:
  + **home\_to\_order\_rate = (order\_page\_visits / home\_page\_visits) \* 100**
  + **order\_to\_checkout\_rate = (checkout\_page\_visits / order\_page\_visits) \* 100**
  + **checkout\_to\_success\_rate = (successful\_checkouts / checkout\_page\_visits) \* 100**
* For data model I went to flat schema, which means all the information in one table, we have only 7 columns of data, it doesn’t make sense to apply other schema. I feel its not necessary to split and forcefully form the schema which it is not helpful.
  + When ever we are doing analytics , its better to de-normalize the data, which help with efficient data reads, if we have large dataset. While analytics it should be read efficient focused. But for application databases (transaction loads) we need write and read optimized and de duplication – that why we go for normalization for transactional load.
  + If we have huge amounts of columns and splitting actually helps then I would prefer star schema rather then snowflake. The reason behind it is that snowflake splits the dimension into sub dimensions which lead to analyst doing lot of join for a simple query. Whereas star schema will reduce the joins, and it is more close to denormalization.
  + I have included sample schema diagram, along with code to generate the diagram – which can be done using - https://dbdiagram.io/d
* **POWER BI**
  + I did ETL and during transformation, I converted the timestamp to datetime data type and I split the date and time , cause it has only one date it is not help full , but time is helpful so , I saved and changed it type to time.

****

**A screenshot of a computer

Description automatically generated**

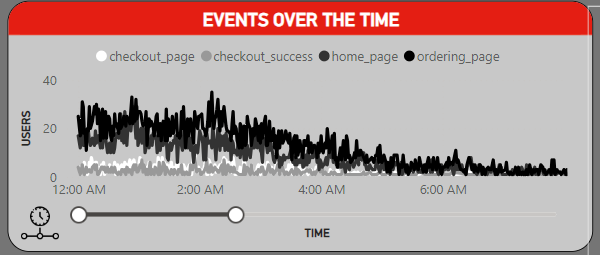
* + Did some data profiling and distribution check to see the quality of the data. For most of the part it looks fine to me except for blank in referred, which I have replace with NA , for better separation and identification.
  + I have also verified all the data type if there are appropriate or not.
  + The reason to choose black, red and while is because these colors are the ThermoFisher Scientific branding colors.
  + Lets discuss about the visualizations:

**A screenshot of a computer

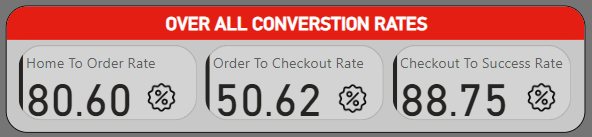
Description automatically generated**

**Note: This visualization is interaction, and other visualizations interaction will effect this visualization.**

* + The reason I choose time over the date and dumped the date in power bi while transformation is because the data has only one date and I don’t provide valuable information, but if we have range of date spanning over the week or month or year , we can do different trend analysis respectively.
  + But here time give us useful information like how are event happening over the time period of 24 hours, you can see that there is high amount of activity before 6 AM and after 6 PM, and in between the is very less activity. With this we can also see these are peak hours of interaction happening.



* + The slider under the visualization allows you to adjust the hours to see the specific activity. In the sample above image you can see activity between 12 AM-6 AM, and you can allow click on the event type on the top so that you can only see specific event one at a time.



* + Note: The values are fixed and will not be updated while interacting with other visualizations. This is because this calculation is don’t as a whole using measure but not for specific country or region. Because going detail will effect the number of use being used in the calculation, which will result in % reaching beyond 100 which shouldn’t happen. In order to prevent that I restricted it to not interact. In data science we have similar issue and the term we use is Simpson paradox – grouping and ungrouping of data reveals different versions of truth.
  + For the conversion rate I use card (new) visualization which help me to display the conversion rate over the stage all at once simultaneously, with using 3 different card (which I need to use if I have used old version of card visualization)
  + You can see only 80.60% of the user are moving from **home to order** out of 80.60% only 50.62% are **going to checkout page** and out this 50.62% only 88.75% are **going to successful page.** This is what the above visualization convey or I wanted it to convey.

A screenshot of a graph

Description automatically generated

* + Note: This visualization is interaction, and other visualizations interaction will effect this visualization.
  + I used funnel chart to display the number of user at each funnel stage, I could use bar chart or pie chart for this funnel chart fits better for the information that I am displayed as these represents stage and this give better information at a glance and shows information as stages.

* + You can select specific event and it filters out reset of the visualizations accordingly if you want to look in details with respect to event.

A screenshot of a user's information

Description automatically generated

* + Note: This visualization is interaction, and other visualizations interaction will effect this visualization.
  + I used bar graph here so that I can clearly display which region has more user base compared to other. Like the funnel and time visualizations, this bar graph is interactive, if you want to look at metrics based on a specific region you can do that by selecting the region in the bar graph.

A map of the world with black circles

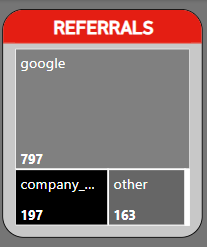
Description automatically generated

* + Note: This visualization is interaction, and other visualizations interaction will effect this visualization.
  + I used map here so that I can clearly display which countries and the user base compared to other in a glace. Like the funnel and time visualizations, this viz is interactive, if you want to look at metrics based on a specific country you can do that by selecting the country in the map and also you can hover and get the user count.

A gray and white pie chart with a red and white text

Description automatically generated

* + I used pie char to display the % user at each stage of conversion in over all user 100%(3077 user). This is different from the “OVER ALL CONVERSION RATE” visualization. Because that will deal with % with in the previous stage (i.e., how much % of the home to order are going to order to checkout). The “% users per stage” show how much percentage of total user that is (3077 users) in a particular stage.
  + I know that there aren’t labels for the visualization for this, which makes use to take an extra step that is to hover over the visualization to get details above stage. This is because the labels doesn’t space to display them at a comfortable font size.



* + I used the treemap to show which referral are more compared to other, how many user each category is occupying. This helps company to focus on any particular category that they want to improve.