Data Analysis of Conversation chats

The report holds information of the analysis of the chat conversation between Vivid money agent and its customer. The data is present in excel file and holds raw data of customers and their activities on the website. The aim of this analysis is to provide insight into the chat conversations data held between customers and the agents and the steps taken to organize the data for analysis and visualization. Further the analysis report gives the idea of the customers behavior and the ways it can be utilized to make the customer experience better in the future.

Below are the key steps taken to perform analysis.

Data exploration:

The aim of data exploration is to get general idea of the data and forming hypotheses about how data can answer technical and business goals. I began with observing the attributes that look promising for analysis, learning the characteristics of data, finding missing values, relationships among the attributes and steps that can be taken to organize the data for further analysis.

The excel file holds two sheets, the first sheet holds the data of conversation records between the customer and agent and the second sheet holds the behavior of the customers for eg; the interaction between the customer and the application.

I have used MS SQL Server and Google spread sheets to explore the data. The tools help me understand the relationship among the attributes, columns with unique values ,missing values present, duplicate values ,format of the data, number of records per attribute etc. The glossary in the excel sheet provides the definition of the attributes.

Few of the steps taken to explore the data are present in the SQL file please refer the file name: ‘**data_exploration_queries**’

	Sheet 1- Conversation	Sheet 2- Customer_Monthly
Total number of records:	40,315	39535
Attributes with Missing values	CLOSED_AT, RATING_NUM, LANGUAGE_CD, FRT_DU_IN_MIN	PLAN_NM
Attributes Data type	float,object	float,object
Attributes Duplicate records	CONVERSATION_HK, CUSTOMER_HK	CUSTOMER_HK

The key for conversation and customers holds duplicate values which does not allow to mark them as primary keys for the table when the data is uploaded into the database. Few attributes have missing values which can affect the future analysis steps. Further the data type of all the attributes are similar which can make data manipulation difficult. Hence we need to take care of all these factors in the next process i.e. data cleaning.

Data Preparation:

In this step we will be taking care of the following things i.e. accessing the data, cleaning, formatting, combining, transforming and finally analyze the data.

Data cleaning:

After the first analysis of the raw data I understood the characteristics of the data the next step is to prepare the data for analysis. I found that the data had many missing values and duplicates in the attributes. In order to work on the data for analysis I need to proceed with cleaning of the data.

I have used Jupyter Notebook to clean and organize the data. The raw data file is present in .xlsx excel file format. In order to work individually on the data sheets I have downloaded the excel sheets in to CSV format and used Python library Pandas data frames and Numpy for data manipulation and cleaning.

I have mentioned the steps taken to clean the data in the python file please refer file:

Below table holds the information of the data after cleaning process.

	Sheet 1- Conversation_edit	Sheet 2- Customer_Monthly_edit
Number of records	39535	26766
Duplicate values removed	780	12769
Unique attribute	CONVERSATION_HK (primary key)	CUSTOMER_HK + DAY_DT (composite key)

There were several duplicate records in the key attribute for customer and conversation, I found that the duplicate key for conversation table holds same values for the entire record and I removed the replicated record and kept one copy in the data. Similarly for customers there were multiple duplicates for the month of November, December and January, since the last date time for every month is similar I carefully deleted the rest of the occurrences which held the exact value as per my assumption.

The data type of each attribute was set to its original dtype to ensure no difficulties during the manipulation phase. The values missing in the attributes are being neglected as they hold less significance when it comes to analysis and we can assign placeholders.

Data Transformation:

Now that we have cleaned the raw data according to standards we can begin with preparing the data for transformation. This step is needed to organize the attributes so that they are ready for analysis and combining the data for finding the relationship among them.

I have used MS SQL Server and Power BI to carry out the steps necessary for data transformation. After cleaning the data in python I extracted the dataframes back to csv files. The csv file is ready to be transformed and the relationship among the tables are defined while ingesting the data into SQL Server.

The data loaded into the database is structured, the attribute ‘CONVERSATION_HK’ is set as the primary key for table conversations and the attributes ‘CUSTOMER_HK’ and ‘DAY_DT’ are set as composite key. We can now join both the tables based on the relationship among them. The next step is to connect the database and the Power BI tool to perform visualization. Also we can begin with the analysis in SQL server and finding insights.

Data Analysis:

I have used Python and SQL to analyze the data.
Please refer python and sql query file:

cohort_analysis.ipynb
analysis_queries.sql

Data Visualization:

I have used Power BI for visual representation of data by creating dashboards, please refer Power BI file: ‘**Analysis Dashboard.pbix’**