Data Glacier

Project: Data Collection Pipeline (Data Acquisition to Story Telling)

Week 9: Deliverables

Name: Salih Eren Yüzbaşıoğlu

Email: algosalih@gmail.com

Group Name: Initial Group

Country: Turkey

Specialization: Data Analyst

Batch Code: LISUM20

Date: 19 May 2023

Submitted to: Data Glacier

Reshaped by: Ahmet Metin Zengin

Table of Contents:

1. Problem description

XYZ company is a global organization that provides various products and services to customers worldwide. As part of their efforts to enhance customer experience and satisfaction, the company collects customer feedback and preferences through online surveys conducted using Google Forms and Survey Monkey.

However, managing and analyzing the collected survey data poses a significant challenge for the company. The data received from these surveys often contain inconsistencies, such as duplicate entries, junk data, or missing values, which can affect the accuracy and reliability of the analysis.

To address this challenge, XYZ company aims to develop a data pipeline that efficiently collects, cleans, and processes the survey data. The pipeline will perform data validation and quality checks to ensure the data's integrity and eliminate any inconsistencies. One critical aspect of the pipeline is to identify and eliminate duplicate entries based on customers' unique email IDs.

Additionally, the company wants to visualize the collected data in a dashboard to gain valuable insights into customer preferences, satisfaction levels, and demographics. The dashboard will provide visual representations and interactive tools to explore the data and derive actionable insights for the company's decision-making processes.

By implementing this data pipeline and dashboard, XYZ company aims to streamline the data collection and analysis process, ensuring the availability of clean and reliable data for making informed business decisions. The pipeline's ability to identify and handle data issues will improve data accuracy and allow the company to gain valuable insights into customer behavior, preferences, and satisfaction levels, ultimately leading to enhanced customer satisfaction and improved business performance.

2. GitHub Repo Link:

https://github.com/AhmetMetinZengin/Data_Glacier_Virtual_Internship/tree/master/Week_09

3. Data Cleansing and Transformation

During our data understanding, we recognized the importance of data cleaning in ensuring the accuracy and reliability of our results. To this end, we performed a thorough data cleaning process to address any potential outliers, missing data, and duplicate columns in the dataset.

Firstly, we utilized the wisconsizing technique to identify and handle any potential outliers in the dataset professionally. This technique involves analyzing the distribution of the data and determining if any values fall outside the range of typical values. By addressing these outliers, we minimized any potential bias or errors in our analysis, which ultimately led to more accurate and reliable results.

Additionally, we identified duplicate columns in the merged file and, specifically, the email columns, which we removed to avoid redundancy in the data. This step ensured that our analysis was based on unique and relevant variables, which ultimately improved the quality of the data and minimized any potential errors or biases.

Furthermore, we deleted any missing or NA values in the dataset to ensure that the data was complete and accurate. This step is crucial in data analysis, as missing data can significantly impact the results of our analysis and lead to biased or inaccurate conclusions. By removing these missing values, we were able to ensure that our analysis was based on complete and reliable data, which ultimately led to more accurate and meaningful insights.