Data Glacier

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

Week 8: Deliverables

Name: Patryk Potocki

Email: ejotpl91@gmail.com

Group Name: Initial Group

Country: Poland

Specialization: Data Analyst

Batch Code: LISUM20

Date: 19 May 2023

Submitted to: Data Glacier

Table of Contents:

1. Project Plan

Week	Date	Plan
Week 7	19 May 2023	Data acquisition. (Generate data)
Week 8	26 May 2023	Collecting all Datasets into master data using script. Understand dataset insights.
Week 9	02 June 2023	Clean the data and perform dedup check.
Week 10	09 June 2023	Visualize the data into Dashboard.
Week 11	16 June 2023	Create a batch which will at specified time and dump the data into master file
Week 12	23 June 2023	Document the challenges encountered during this implementation.
Week 13	30 June 2023	Final Project Report and Code

2. Problem Statement

XYZ company is collecting the data customer using google forms and they have floated n number of forms on the web.

The company wants to create a pipeline which will collect all the data of these google forms and visualize the data in the dashboard.

The dataset needs to be clean and if there is any data issue present in the data then it should be treated by this pipeline (duplicate data or junk data. Dedup check should be performed on the email id of the customer.

3. Data Understanding

The source of the data used in this analysis is company data. It includes personal information about individuals such as their age, gender, marital status, education level, number of children, and number of languages spoken. The data also includes a satisfaction rate variable, which represents the level of satisfaction of the individuals. The data was collected from various sources and merged into a single file to provide a comprehensive view of the customers. By analyzing this data, we can gain insights into the characteristics and preferences of the customers, which can help companies make informed decisions about their products and services.

4. Types of Data

Based on the merged data, we can identify the following types of data:

1.Categorical Data: Gender, Marital Status, Education Level, and Languages Spoken are categorical variables. They have a limited number of possible values and often represent characteristics or attributes of the customers.

2.Numerical Data: Age, Number of Children, Number of Languages Spoken, and Satisfaction Rate are numerical variables. They have a range of values and often represent quantitative measurements of the customers.

5. Data Insights

After analyzing the data, we found that the majority of the variables are skewed positively, indicating that the data is more concentrated towards the higher end of the scale. However, the Satisfaction Rate variable is skewed negatively, indicating that the data is more concentrated towards the lower end of the scale. We did not find any outliers or missing

data in the dataset, indicating that the data is of good quality. However, we did find duplicate columns in the merged file, specifically the email columns, which we will remove to avoid redundancy in the data. Overall, the data is in relatively good condition, and by addressing the issues found during our analysis, we ensured that our results are accurate and reliable.

6. Data Cleaning

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.