Project name: Data Analyst:: Data Collection Pipeline (Data Acquistion to Storytelling) - Data Collection Pipeline (Data Acquisition to Story telling)

Debika Mukherjee (Solo project)
Email: debika.isms@gmail.com
Country: United Kingdom
University: Cardiff metropolitan university
Specialization: Data Analyst

Problem description:

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

Company wants to create a pipeline which will collect all the data of these google forms/survey monkey and visualize the data in the dashboard.

Company wants clean data 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

Business understanding :

The XYZ company is currently leveraging Google Forms and SurveyMonkey to collect valuable customer data through various web-based forms. These forms have been distributed across the internet to gather insights and information from their customer base. In order to streamline their data management processes and gain actionable insights, the company is looking to establish a comprehensive data pipeline and visualization dashboard.

The primary objectives of this initiative are as follows:

- Centralized Data Collection: The company aims to aggregate data from the multiple Google Forms and SurveyMonkey sources into a unified data repository. This consolidation will enable a holistic view of customer interactions and responses across different forms.

- Data Visualization: To effectively interpret and analyze the collected data, the company plans to visualize the information on a dashboard. This dashboard will provide interactive and intuitive graphical representations of key metrics, trends, and insights derived from the data.

- Data Cleaning and Quality Assurance: Ensuring the reliability and accuracy of the collected data is of paramount importance. The pipeline should include mechanisms to identify and address potential issues within the data. Specifically, it will detect and handle duplicate records and erroneous entries, ensuring that only high-quality data is used for analysis.

- Deduplication Check: One significant aspect of data quality is identifying and removing duplicate records. To achieve this, the pipeline will focus on performing deduplication checks based on the email IDs of the customers. This will enhance the accuracy of customer-related insights and prevent redundant data from skewing the analysis.

Overall, the successful implementation of this data pipeline and visualization dashboard will empower XYZ company to make well-informed business decisions based on reliable, clean, and meaningful customer data. By addressing data issues and presenting insights in a user-friendly manner, the

company aims to enhance its strategic planning, customer engagement, and overall operational efficiency.

Project lifecycle and deadline:

Project Planning and Research: [19/08/2023] - [26/08/2023]
Data Collection and Preprocessing: [26/08/2023] - [02/09/2023]
Model Development: [02/09/2023] - [09/09/2023]
Testing and Validation: [09/09/2023] - [16/09/2023]
Documentation and Reporting: [16/09/2023] - [23/09/2023]
Final Presentation and Submission: [23/09/2023] - [30/09/2023]

Data intake report:

I will create survey by using google forms to collect the data on product usage and customer experience as well as feedback on marketing and communication based on demographics. Currenly I am in a process to collect the data. I will use qualitative as well as quantitative method to collect the data. I will provide the whole data intake report on 26th September when I will submit the planning and research.

Github repository:

debikaisms/Final-project: Data Collection Pipeline (Data Acquistion to Storytelling) Data Collection Pipeline (Data Acquisition to Story telling) (github.com)