Data Analyst: Data Collection Pipeline (Data Acquisition to Storytelling) - Group Project

Data Collection Pipeline (Data Acquisition to Storytelling)

Problem Statement:

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

Task:

1. **Perform data acquisition using more than 2 google forms/Survey monkey:**
   -Created 2 google forms manually simulating Company XYZ's product & customer service feedback
   -Each form has 10 manually inputted responses
   -Linked each form to a Google Sheet and then exported to csv

2. Collect all the forms data into a master form using script.

-Concatenated to combine both files in Jupyter Notebook

-Master form created successfully.

3. Clean the data and perform dedup check

-Standardized email field (lowercase)

-Dropped duplicates

-NaN values are expected since each form collected different fields, they were retained to preserve the structure

-A new column named "Form Source" was created to label each record's origin

4. Visualize the data into dashboard.

Visualization can be based on the number of positive responses received location-wise or countrywide

5. Create a batch which will run at specific time and dump the data into master file (or data lake)

-Created a batch pipeline script in Spyder 6

-Set specific time through Windows Task Scheduler

5. Document the challenges encountered during this implementation:

-Struggled on figuring out whether to create forms manually or to find forms online but ended up manually creating 2 google forms as a "simulation" for XYZ essentially.

-Ran into small hiccups with coding but solved issues

Week 7 Deliverables

- Submit a pdf document which should contain the following details:
- **Team members' details :** Data Collector, Katrina Kirby, katkirby9217@gmail.com, United States, Virginia Commonwealth University, Specialization (Data Analyst)
- **Problem description:** Read above at top of page
- **Business understanding:** Company XYZ needs assistance with managing their data. This company is seeking services in collecting customer data using Google Forms and SurveyMonkey. The goal is to consolidate those responses from multiple forms and make into one singular, clean dataset. I need to create a data collection pipeline that cleans, deduplicates by email, stores, and visualizes the data in a dashboard. If the company is dealing with unorganized, inconsistent, and scattered customer data, this can affect strategic business decision-making and how they are interpreting the data. I can help centralize and clean form data, which can enable

better analysis of customer satisfaction trends, geographic response distribution, and data readability.

- **Project lifecycle along with deadline:**

Week	Task	Status	Tools/Deliverables
Week 7	**Data acquisition, cleaning, deduplication, visualization, and pipeline documentation**	**In progress**	**Google forms, Python (pandas), GitHub, PDF write-up**
Week 8	**Data understanding and profiling, identify, NA/outliers/skew**	**Upcoming**	**Pandas, summary stats, exploratory notes, GitHub update**
Week 9	**Data cleansing & transformation (2+ methods)**	**Upcoming**	**Jupyter Notebook, regex, mean/median/model imputation techniques**

Week 10	**EDA and insights reporting**	**Upcoming**	**Seaborn, matplotlib, summary of trends, GitHub notebook**
Week 11	**EDA presentation for business and technical recommendations**	**Upcoming**	**Powerpoint or Google Slides, business/technical slide split**
Week 12	**Model selection and dashboard creation for Data Analyst track**	**Upcoming**	**ML models (Linear, Ensemble, Boosting). Plotly/Tableau/Power BI dashboard**
Week 13	**Final report, polished code, GitHub link, solo solution presentation**	**Upcoming**	**Final GitHub repo, PowerPoint, PDF report**

- Data Intake report – Submitted separately via GitHub
- GitHub Repo link-