Week 8 Deliverables

Katrina Kirby

LISUM43

4/25/2025

Team member's details : Data Collector, Katrina Kirby, katkirby9217@gmail.com, United States, Virginia Commonwealth University, Specialization (Data Analyst)

Problem description: XYZ company is collecting the data customer using google forms/survey monkey and they have floated in 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

Data understanding: Two datasets were collected from different Google Forms:

1. **Product Survey Data** – Captures customer feedback on purchased products.
2. **Service Survey Data** – Captures customer support experience and resolution quality.

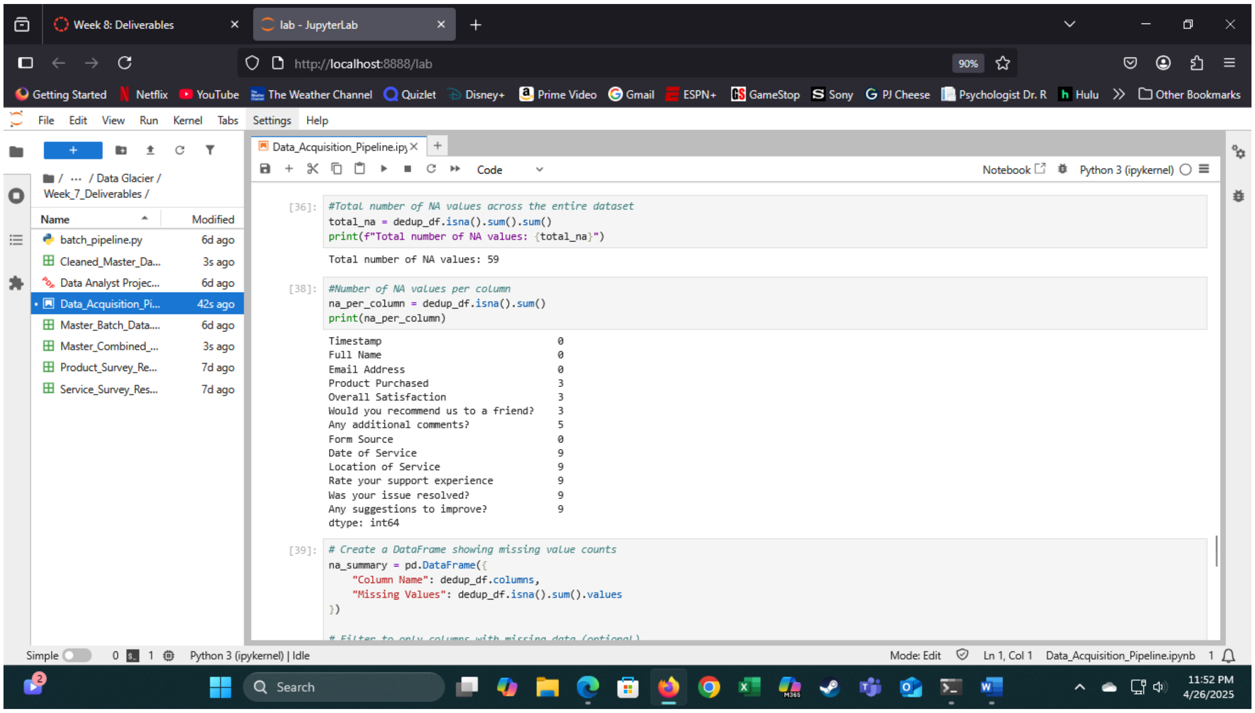
Each dataset contains user inputs such as name, email, satisfaction rating, comments, and timestamps. Some key fields from the Product Survey data that were collected are Timestamp, Full Name, Email Address, Product Purchased, Overall Satisfaction, and Recommendation. Key fields collected from the Service Survey Data include Timestamp. Full Name, Email Address, Date of Service, Location of Service, Support Rating, and Issue Resolved. I collected different kinds of data types such as text, categorical, integers, and dates. The primary identifier of the data is the email address and some customers might submit both reviews separately which means that deduplication based on email addresses is important.

What type of data you have got for analysis:

I have survey responses that are structured neatly into fields. Each row represents a single customer feedback entry. The data is organized into a tabular format and easily readable. I started with direct customer inputs via Google Forms and each row captures feedback at the individual customer level. My collection was web-based surveys containing a variety of data categories. I used categorical variables to help with grouping, segmentation, and visual plots. Numerical variables were used for averages, distributions, and rating analytics. The date and time variables were useful for time-based trend and scheduling analytics.

What are the problems in the data (number of NA values, outliers , skewed etc):

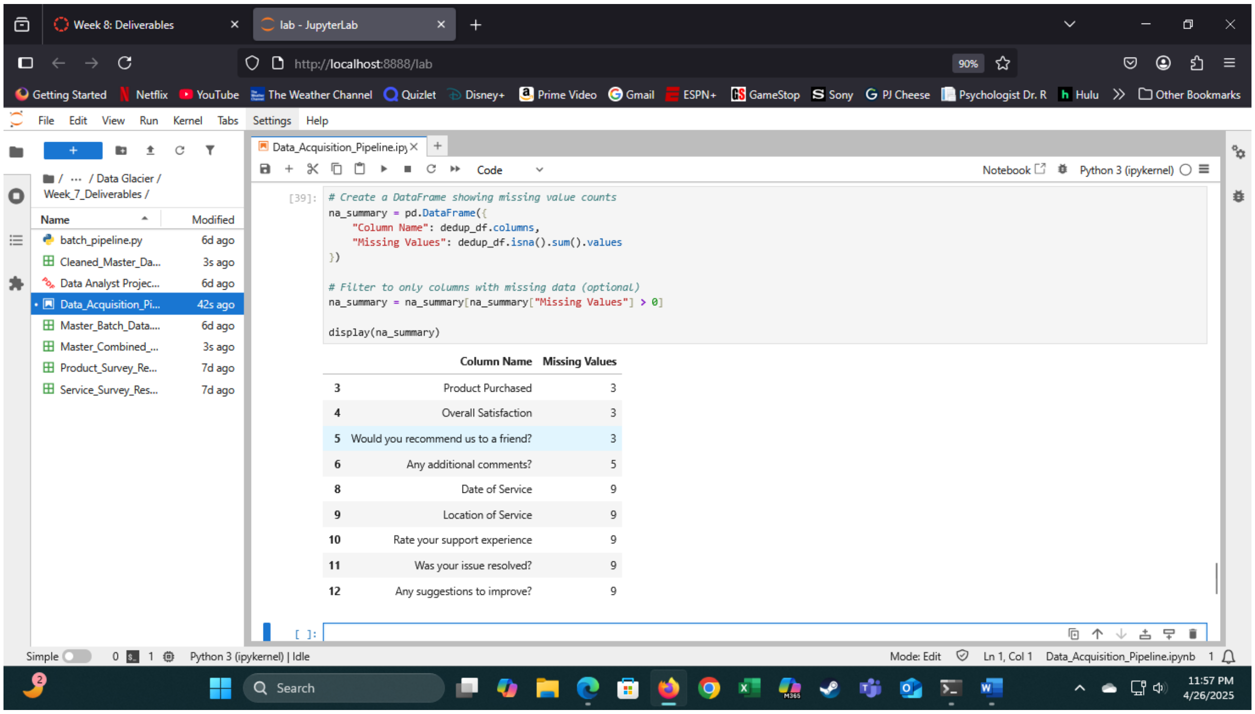
After merging the two datasets into one master form, I found NaN values in columns. There is a total of 59 missing values.


Week 8: Deliverables
lab - JupyterLab
http://localhost:8888/lab
File Edit View Run Kernel Tabs Settings Help
/ … / Data Glacier / Week_7_Deliverables /
Name Modified
batch_pipeline.py 6d ago
Cleaned_Master_Da... 3s ago
Data Analyst Projec... 6d ago
Data_Acquisition_Pi... 42s ago
Master_Batch_Data.... 6d ago
Master_Combined_... 3s ago
Product_Survey_Re... 7d ago
Service_Survey_Res... 7d ago
Data_Acquisition_Pipeline.ipynb
Code
Notebook
Python 3 (ipykernel)
Simple
Python 3 (ipykernel) | Idle
Mode: Edit
Ln 1, Col 1
Data_Acquisition_Pipeline.ipynb
Search
11:52 PM
4/26/2025

```
[36]: #Total number of NA values across the entire dataset
      total_na = dedup_df.isna().sum().sum()
      print(f"Total number of NA values: {total_na}")

      Total number of NA values: 59

[38]: #Number of NA values per column
      na_per_column = dedup_df.isna().sum()
      print(na_per_column)

      Timestamp                          0
      Full Name                          0
      Email Address                      0
      Product Purchased                  3
      Overall Satisfaction               3
      Would you recommend us to a friend?    3
      Any additional comments?           5
      Form Source                        0
      Date of Service                    9
      Location of Service                9
      Rate your support experience       9
      Was your issue resolved?           9
      Any suggestions to improve?        9
      dtype: int64

[39]: # Create a DataFrame showing missing value counts
      na_summary = pd.DataFrame({
          "Column Name": dedup_df.columns,
          "Missing Values": dedup_df.isna().sum().values
      })

      # Filter to only columns with missing data (optional)
```

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

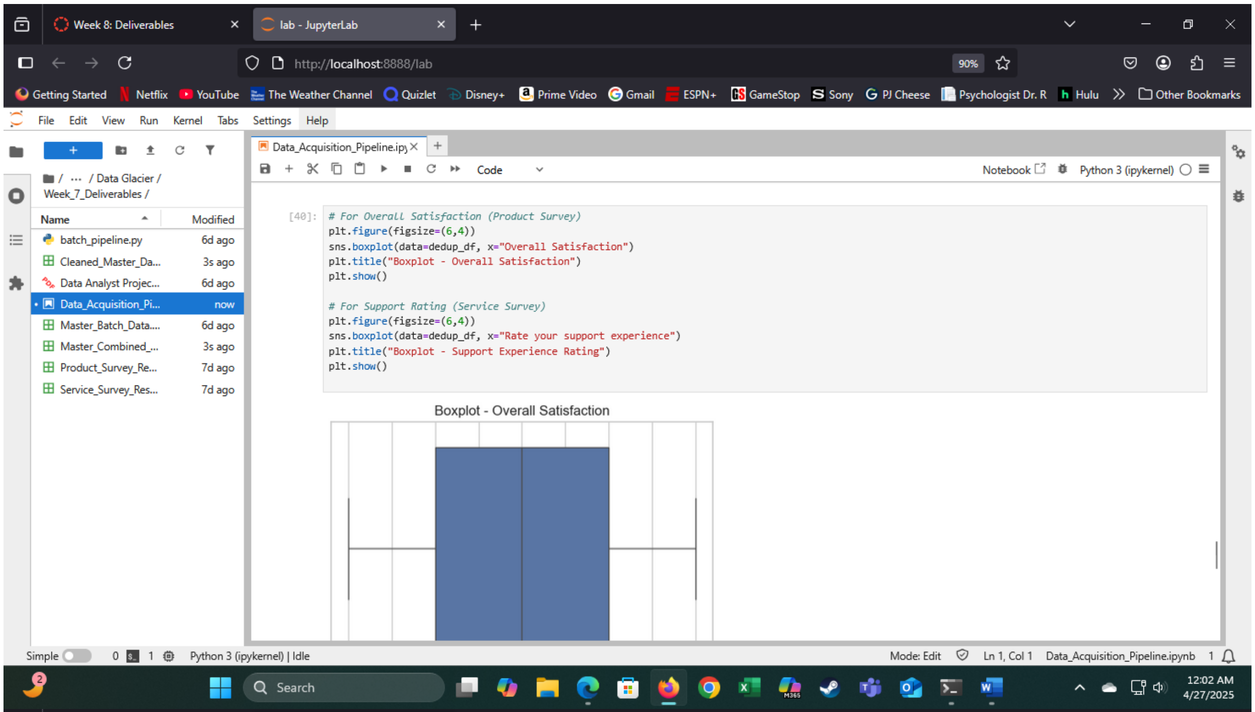
Since the two forms collected different sets of information, the merged data naturally contained NaN values in fields. To maintain the data integrity, I filled the missing values with the placeholder "Not Applicable" to clearly indicate the absence of certain responses. I performed analysis to identify missing or non-applicable fields.



```
[39]: # Create a DataFrame showing missing value counts
na_summary = pd.DataFrame({
    "Column Name": dedup_df.columns,
    "Missing Values": dedup_df.isna().sum().values
})

# Filter to only columns with missing data (optional)
na_summary = na_summary[na_summary["Missing Values"] > 0]

display(na_summary)
```

	Column Name	Missing Values
3	Product Purchased	3
4	Overall Satisfaction	3
5	Would you recommend us to a friend?	3
6	Any additional comments?	5
8	Date of Service	9
9	Location of Service	9
10	Rate your support experience	9
11	Was your issue resolved?	9
12	Any suggestions to improve?	9

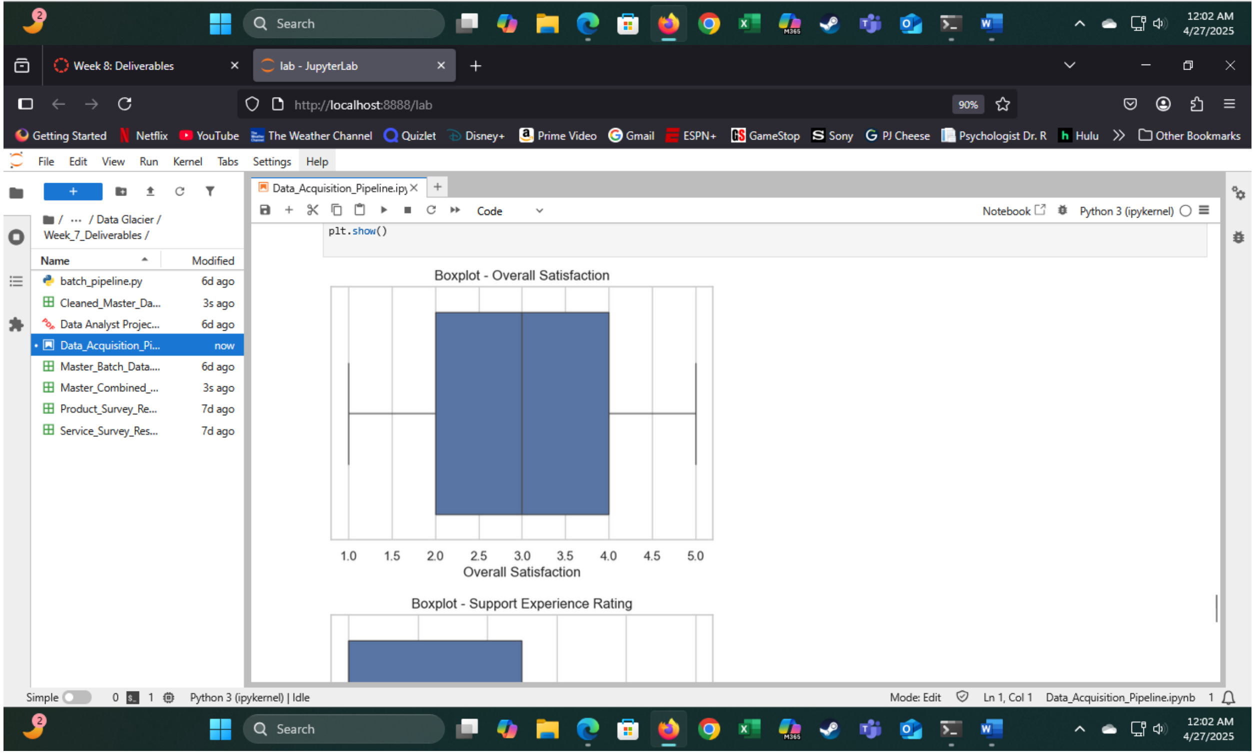
I conducted outlier analysis using a quick boxplot visual and no significant outliers were shown. This was expected seeing as it was constrained within a 1-5 rating scale.



Name	Modified
batch_pipeline.py	6d ago
Cleaned_Master_Da...	3s ago
Data Analyst Projec...	6d ago
Data_Acquisition_Pi...	now
Master_Batch_Data...	6d ago
Master_Combined_...	3s ago
Product_Survey_Re...	7d ago
Service_Survey_Res...	7d ago

[40]:

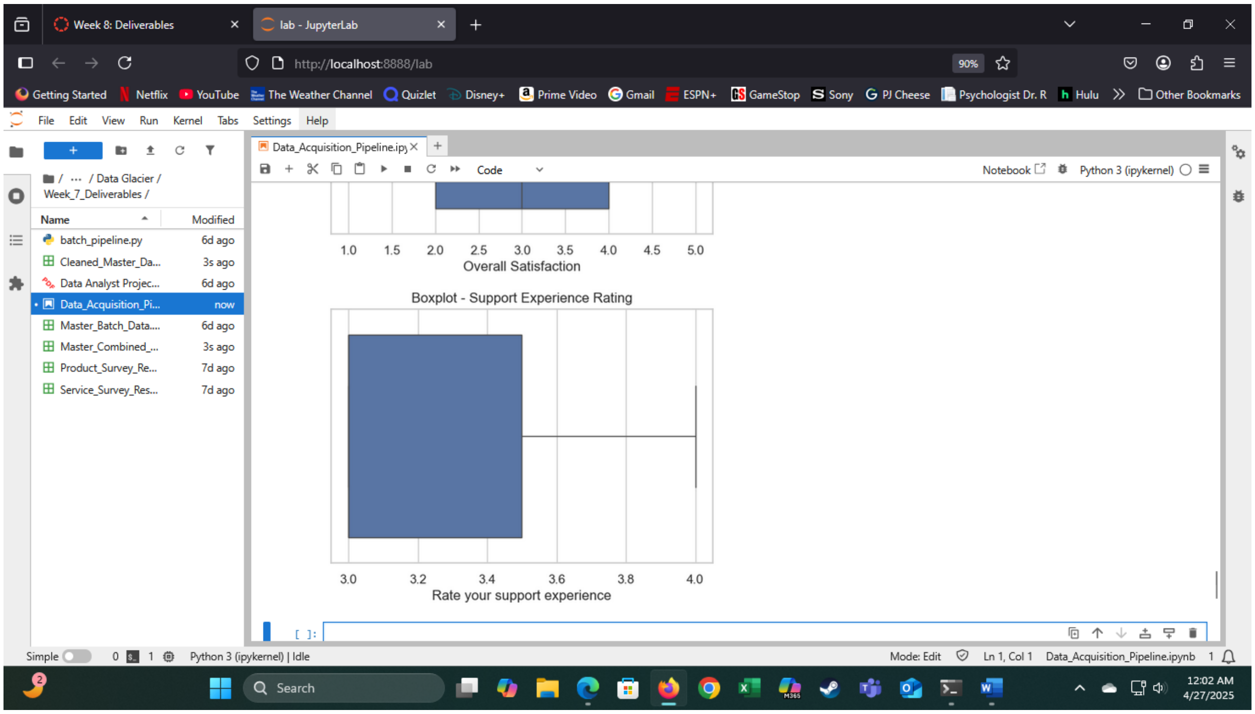
```
# For Overall Satisfaction (Product Survey)
plt.figure(figsize=(6,4))
sns.boxplot(data=dedup_df, x="Overall Satisfaction")
plt.title("Boxplot - Overall Satisfaction")
plt.show()

# For Support Rating (Service Survey)
plt.figure(figsize=(6,4))
sns.boxplot(data=dedup_df, x="Rate your support experience")
plt.title("Boxplot - Support Experience Rating")
plt.show()
```



Week 8: Deliverables
lab - JupyterLab
http://localhost:8888/lab
90%
File
Edit
View
Run
Kernel
Tabs
Settings
Help
Data_Acquisition_Pipeline.ipy
Code
Notebook
Python 3 (ipykernel)
/ … / Data Glacier / Week_7_Deliverables /
Boxplot - Overall Satisfaction
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Overall Satisfaction
Boxplot - Support Experience Rating
Simple
Python 3 (ipykernel) | Idle
Mode: Edit
Ln 1, Col 1
Data_Acquisition_Pipeline.ipynb
Search
12:02 AM
4/27/2025

Name	Modified
batch_pipeline.py	6d ago
Cleaned_Master_Da...	3s ago
Data Analyst Projec...	6d ago
Data_Acquisition_Pi...	now
Master_Batch_Data...	6d ago
Master_Combined_...	3s ago
Product_Survey_Re...	7d ago
Service_Survey_Res...	7d ago

```
plt.show()
```



Name	Modified
batch_pipeline.py	6d ago
Cleaned_Master_Da...	3s ago
Data Analyst Projec...	6d ago
Data_Acquisition_Pi...	now
Master_Batch_Data....	6d ago
Master_Combined_...	3s ago
Product_Survey_Re...	7d ago
Service_Survey_Res...	7d ago

GitHub Repo link: https://github.com/kirbykt/Week-7-Data-Analyst-Project

*Repository is named week 7 but will contain all files for this project. Week 8 PDF will be inside the repository.