Web Analytics and Business Intelligence

Data Cleaning using Tableau

Steps Performed in Data Cleaning:

1. Connect to the Data Source

- **Goal**: Ensure the data is loaded and visible in Tableau's Data Source tab.
- **Action**: Open Tableau and connect to your data source.
- **Steps**:
  - From the start page, click **"Connect"** and select your data source type (Excel, CSV, database, etc.).
  - Locate and select the data file or connect to the database using appropriate credentials.

2. Explore and Understand Your Data

1. **Action**: Examine your dataset to identify issues like data types, null values, duplicates, or inconsistent formats.
2. **Steps**:
   - Navigate to the **Data Source** tab.
   - Check field types (string, date, number) and ensure appropriate classification.
   - Branch ID to Integer
   - Duration to Float or Decimal
   - Time to Date/Time
   - Worker ID to Integer
   - Review a sample of rows to spot missing or inconsistent data.

3. Perform Basic Cleaning

- **Action**: Resolve common data quality issues.
- **Steps**:
  - Remove ‘o’ from Branch ID using calculated field Branch ID 2, and hide Branch ID.
  - Rename Fields: Rename F1 column to Record ID.
  - Handle Null Values: Replace nulls by using calculated fields like for both numerical values and text.
  - Use a calculated field to replace nulls for numerical columns with “0”, for the Response columns
  - Text with the right text information for the Query columns.

- ❖ Use Regex, clean Query7 to create new column with only alphabet.
- ❖ Fix Query9 by filling the column with the correct text "Will you recommend this factory to a friend or family member?", using calculated field.

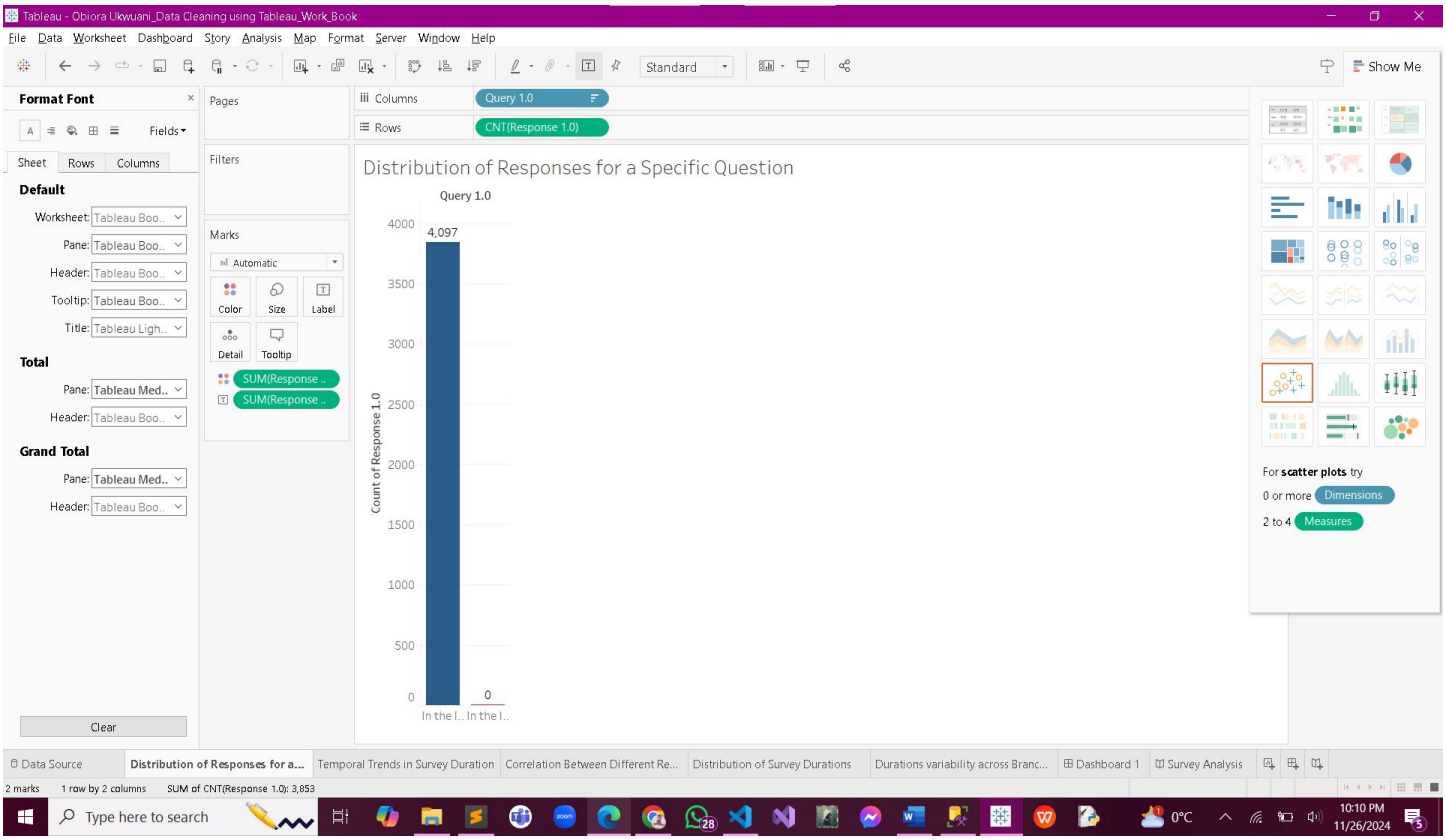
Visualizations

1. Bar Chart: Distribution of Responses for a Specific Question

Purpose: Visualizes the frequency of responses for a given question to identify the most common answers.

Steps:

- ➢ Open Tableau and connect to your CSV file.
- ➢ Drag the Query 1.0 to **Columns**.
- ➢ Drag the corresponding Response 1.0 to **Rows**.
- ➢ Use **COUNT** as the aggregation method for responses.
- ➢ Sort the bars in descending order and optionally add labels for clarity.

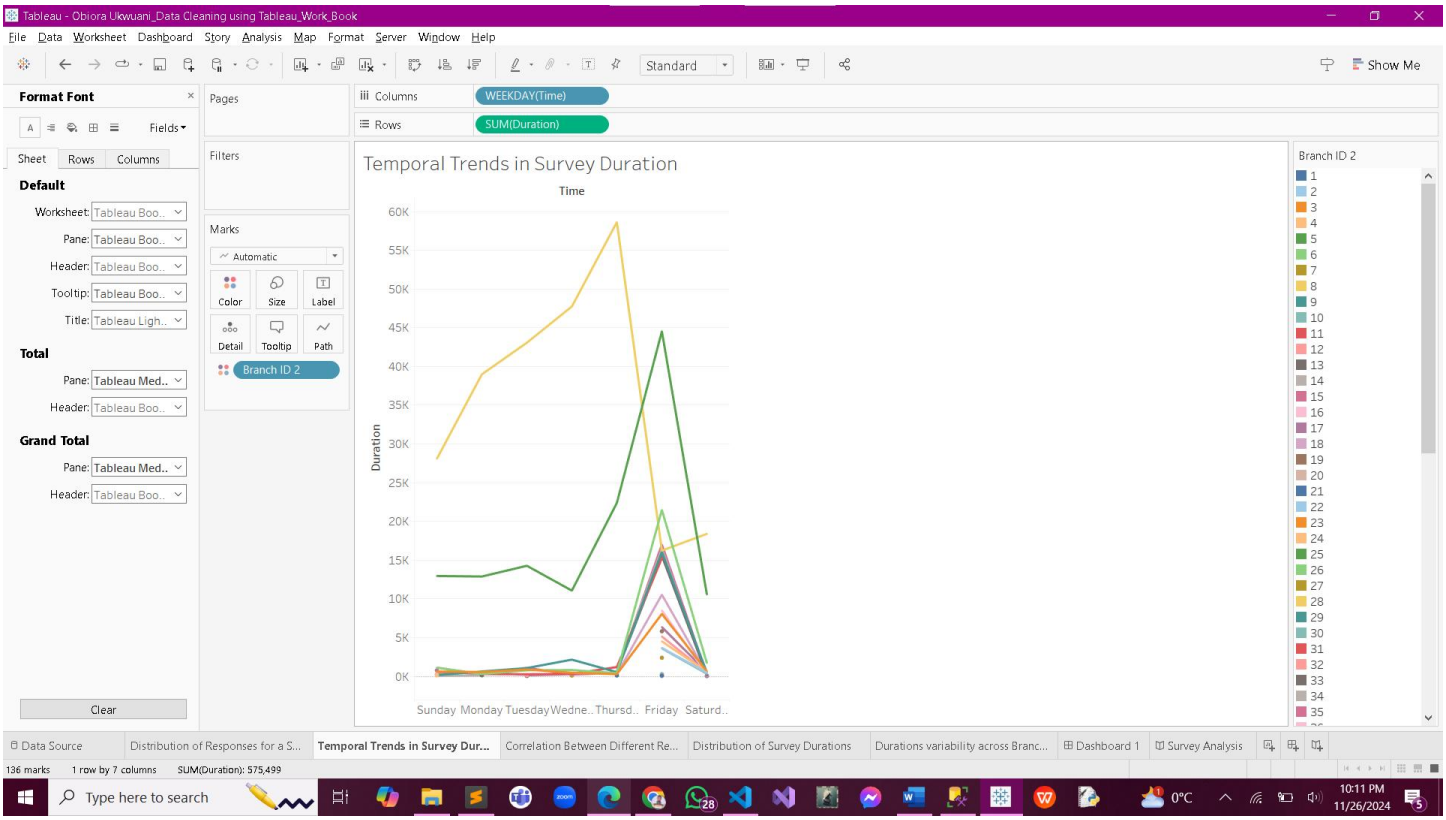


2. Line Chart: Temporal Trends in Survey Duration

Purpose: Shows how survey durations change over time.

Steps:

- Drag Time to **Columns** and convert it to a continuous field if needed.
- Drag Duration to **Rows**.
- Add Branch ID 2 to **Color** to differentiate trends by branches.
- Right click on the Time and set the axis to weekly to suit the granularity.



3. Scatter Plots: Correlation Between Different Response Categories

Purpose: Highlights the relationship between multiple response categories.

Steps:

- **Drag the Response Fields to Columns and Rows**:
  - Drag one response field Response 2.0 to **Columns**.
  - Drag another response field Response 3.0 to **Rows**.
- **Add a Measure to Color**:
  - Drag the Worker ID field to **Color** on the **Marks Card**.
  - Tableau automatically applies the **COUNT** aggregation, counting how many times each combination of Response 1.0 and Response 2.0 occurs.
- **Change the Mark Type to Square**:
  - On the **Marks Card**, select **Square** to create a heatmap-style visualization.
- **Customize the Color Gradient**:
  - Click the **Color** button on the **Marks Card,** Select **Edit Colors** to apply a gradient scale from light to dark or blue to red.
  - Use a diverging color palette for better emphasis on extremes.
- **Optional: Add Labels**:
  - Drag the Record ID and Worker ID to **Label** to apply count.
  - Adjust the label format for clarity.

Explanation

- Each square represents a combination of responses from Response 1.0 and Response 2.0.
- The color intensity reflects the frequency of that combination, making it easy to spot popular or rare correlations between responses.

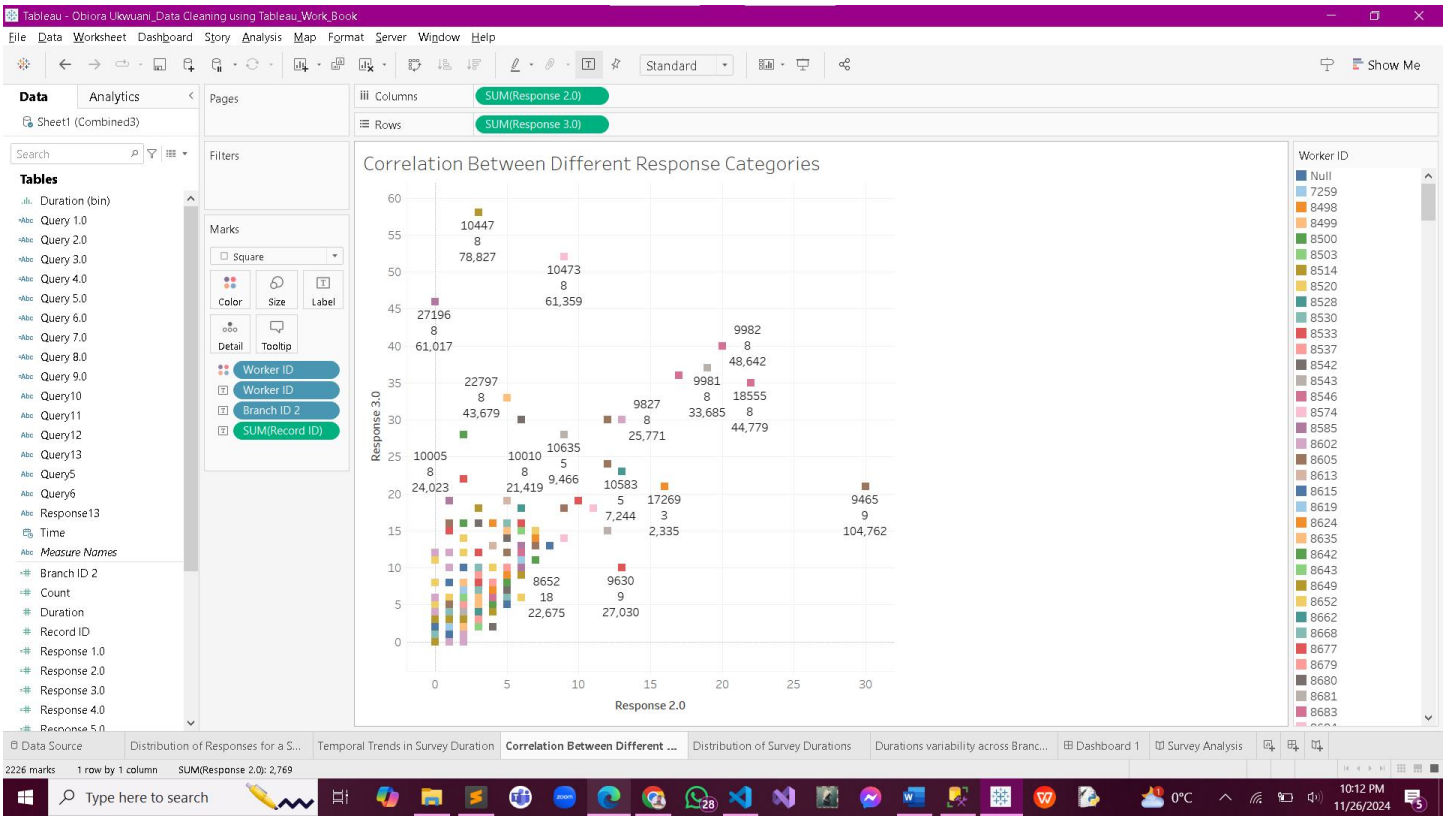


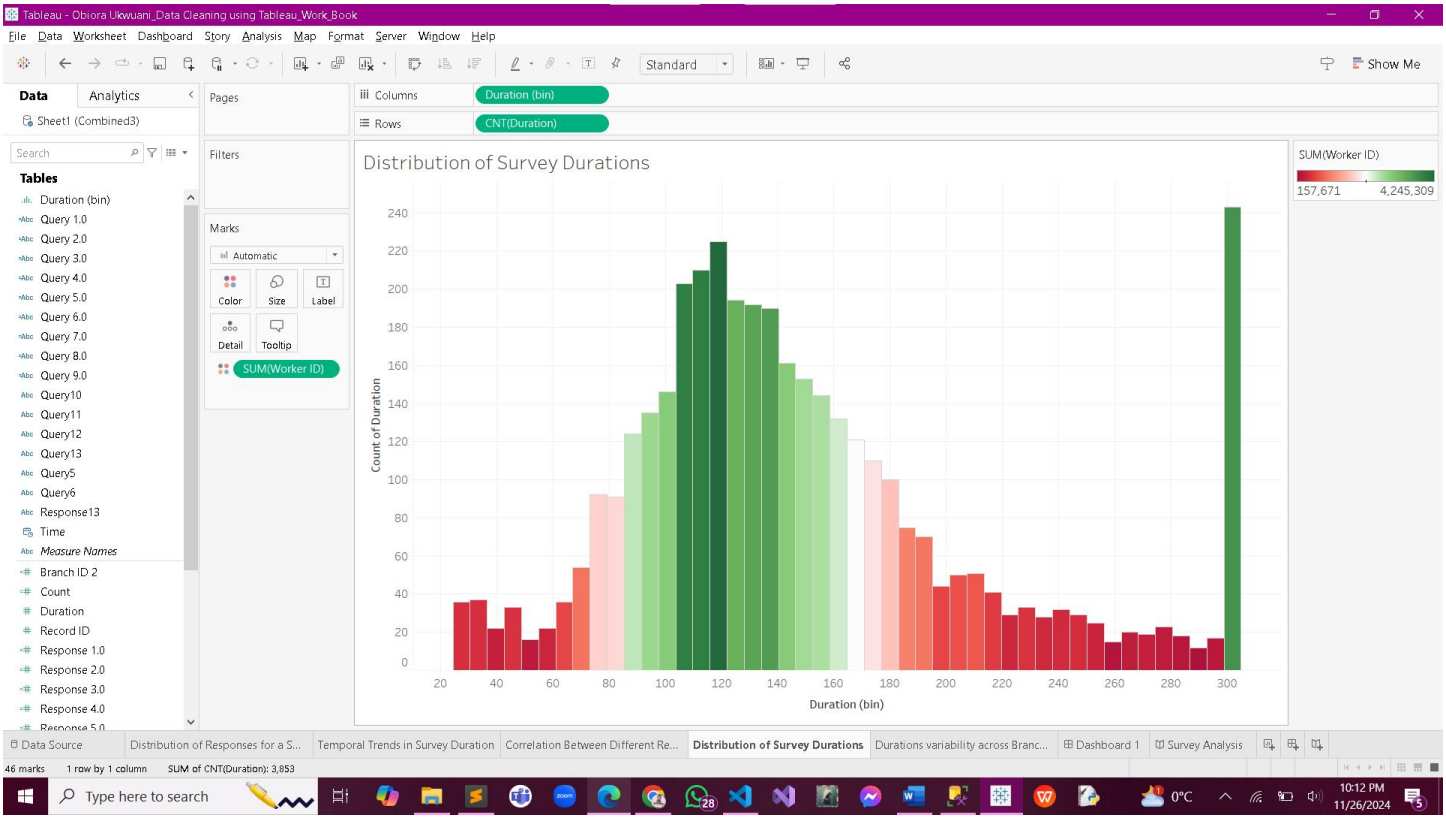
Tableau - Obiora Ukwuani_Data Cleaning using Tableau_Work_Book
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
Sheet1 (Combined3)
Search
Tables
Duration (bin)
Query 1.0
Query 2.0
Query 3.0
Query 4.0
Query 5.0
Query 6.0
Query 7.0
Query 8.0
Query 9.0
Query10
Query11
Query12
Query13
Query5
Query6
Response13
Time
Measure Names
Branch ID 2
Count
Duration
Record ID
Response 1.0
Response 2.0
Response 3.0
Response 4.0
Response 5.0
Pages
Filters
Marks
Square
Color
Size
Label
Detail
Tooltip
Worker ID
Worker ID
Branch ID 2
SUM(Record ID)
Columns
SUM(Response 2.0)
Rows
SUM(Response 3.0)
Correlation Between Different Response Categories
Response 3.0
Response 2.0
10447 8 78,827
10473 8 61,359
27196 8 61,017
9982 8 48,642
22797 8 43,679
9981 8 33,685
18555 8 44,779
9827 8 25,771
10005 8 24,023
10010 8 21,419
10635 5 9,466
10583 5 7,244
17269 3 2,335
9465 9 104,762
8652 18 22,675
9630 9 27,030
Worker ID
Null
7259
8498
8499
8500
8503
8514
8520
8528
8530
8533
8537
8542
8543
8546
8574
8585
8602
8605
8613
8615
8619
8624
8635
8642
8643
8649
8652
8662
8668
8677
8679
8680
8681
8683
Data Source
Distribution of Responses for a S...
Temporal Trends in Survey Duration
Correlation Between Different ...
Distribution of Survey Durations
Durations variability across Branc...
Dashboard 1
Survey Analysis
2226 marks
1 row by 1 column
SUM(Response 2.0): 2,769
Type here to search
0°C
10:12 PM
11/26/2024

4. Histogram: Distribution of Survey Durations

Purpose: Examines the frequency of different survey durations to identify outliers or common ranges.

Steps:

- Drag Duration to **Columns**.
- Select the histogram option from the **Show Me** menu.
- Adjust bin sizes as needed for better granularity.



5. Box-and-Whisker Plot: Variability in Responses Across Branches

Purpose: Visualizes the spread and outliers of responses across branches.

- **Understand Duration Variability**:
  - Identify how survey durations differ across branches and response categories.
  - Spot branches with high variability in duration, which might indicate operational inconsistencies.
- **Compare Responses Across Branches**:
  - Examine how response patterns (e.g., satisfaction levels captured by Response 5.0) relate to survey duration within each branch.
  - Identify trends, such as longer durations correlating with specific response levels.
- **Detect Outliers**:
  - Highlight outliers where survey durations are significantly higher or lower than typical.

  This visualization helps decision-makers pinpoint branches with unusual survey behavior and understand relationships between response patterns and survey time. Let me know if you'd like further clarification

Steps:

1. Drag Branch ID 2 and Response 5.0 to **Rows**.
2. Drag Duration to Column
3. Select **Box Plot (interquartile range (IQR)** from the **Analytics Pane**, drag and release on the chart.
4. Customize the box plot with colors or labels.

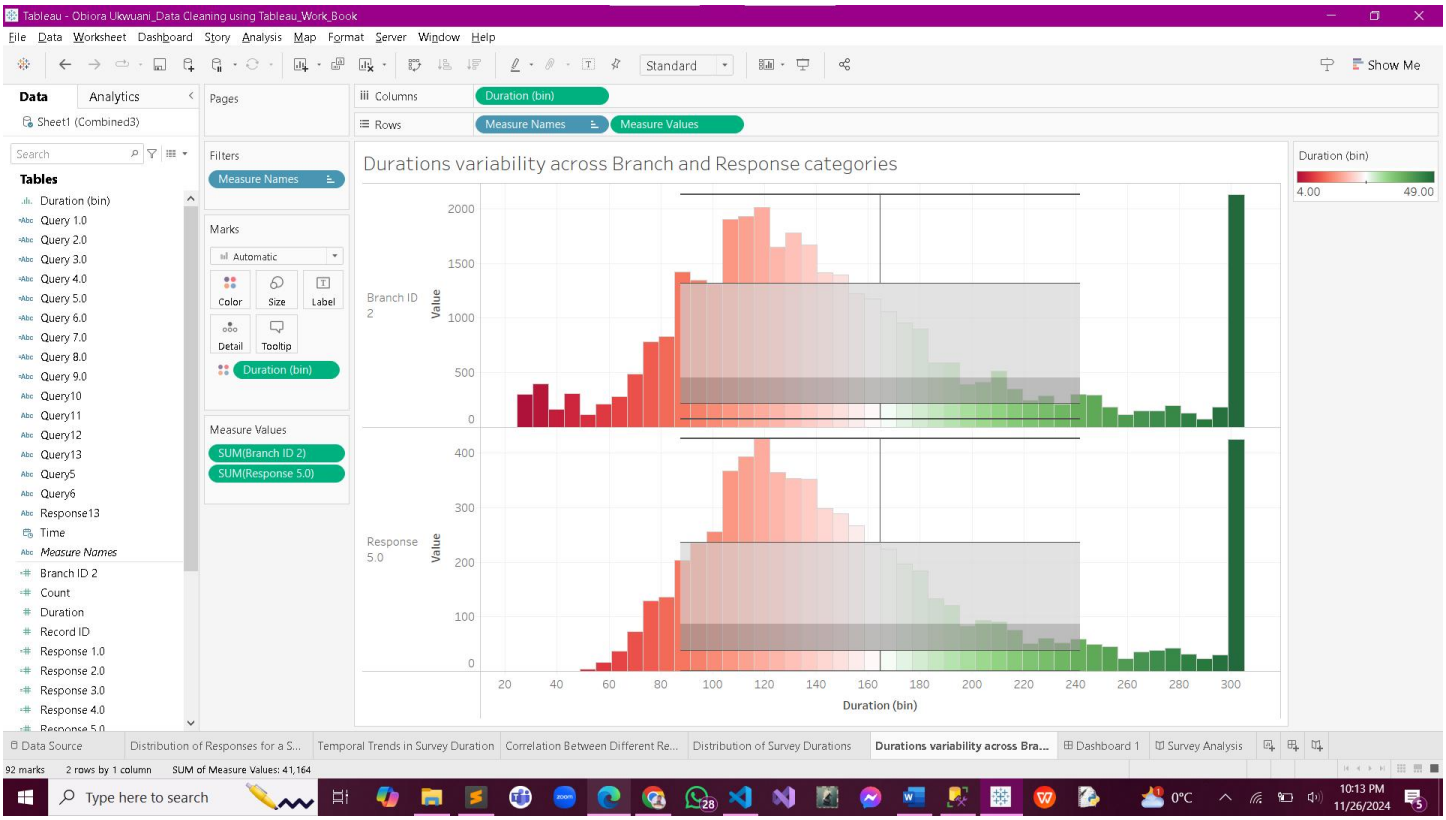


Tableau - Obiora Ukwuani_Data Cleaning using Tableau_Work_Book
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Data Analytics
Pages
Filters
Measure Names
Marks
Automatic
Color Size Label
Detail Tooltip
Duration (bin)
Measure Values
SUM(Branch ID 2)
SUM(Response 5.0)
Columns
Duration (bin)
Rows
Measure Names
Measure Values
Durations variability across Branch and Response categories
Branch ID 2
Response 5.0
Value
Duration (bin)
Duration (bin)
4.00
49.00
Data Source
Distribution of Responses for a S...
Temporal Trends in Survey Duration
Correlation Between Different Re...
Distribution of Survey Durations
Durations variability across Bra...
Dashboard 1
Survey Analysis
92 marks
2 rows by 1 column
SUM of Measure Values: 41,164
Type here to search
10:13 PM
11/26/2024

Dashboard:

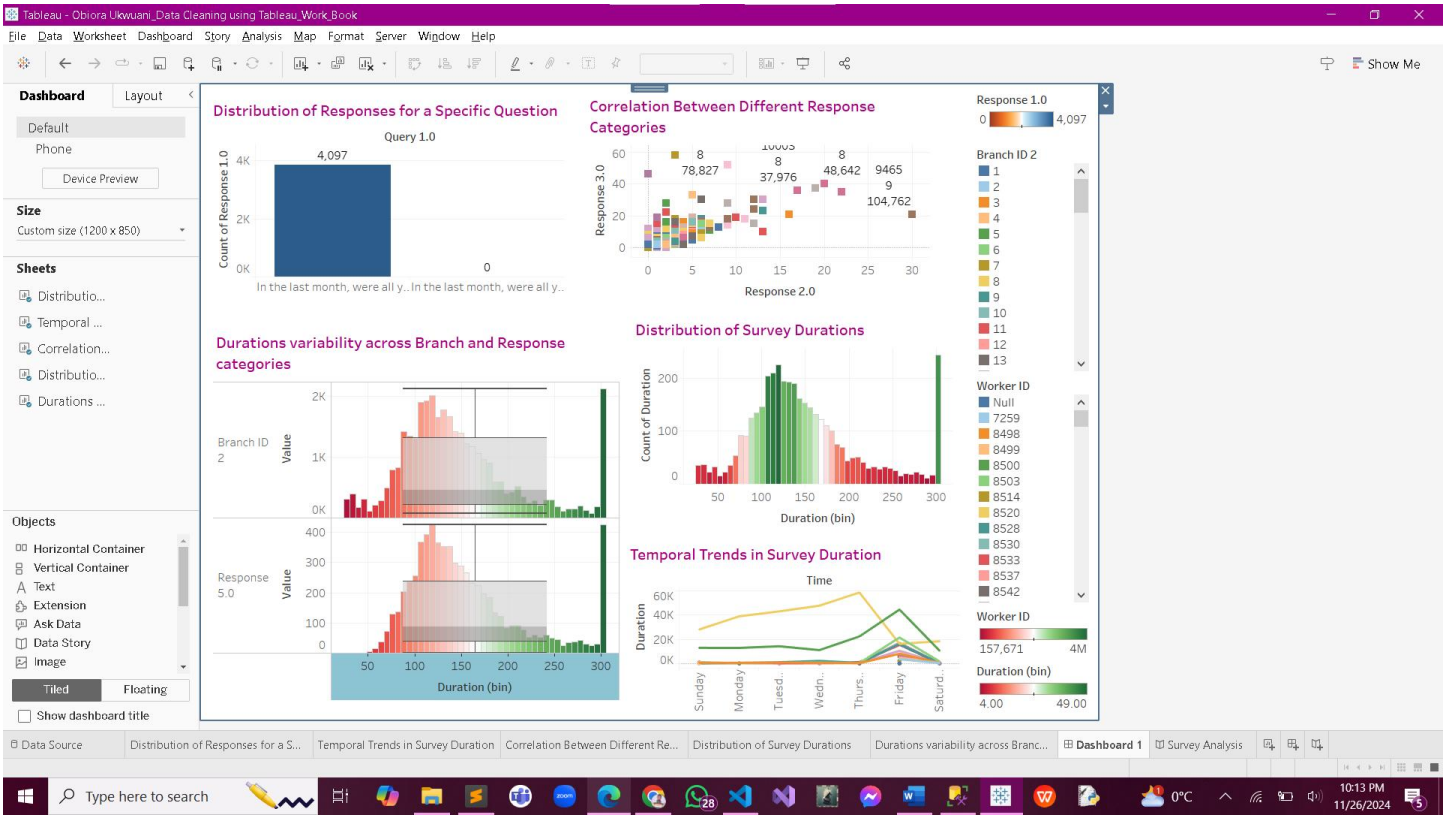


Tableau - Obiora Ukwuani_Data Cleaning using Tableau_Work_Book
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Dashboard
Layout
Default
Phone
Device Preview
Size
Custom size (1200 x 850)
Sheets
Distributio...
Temporal ...
Correlation...
Distributio...
Durations ...
Objects
Horizontal Container
Vertical Container
Text
Extension
Ask Data
Data Story
Image
Tiled
Floating
Show dashboard title
Distribution of Responses for a Specific Question
Query 1.0
Count of Response 1.0
4,097
0
In the last month, were all y... In the last month, were all y...
Correlation Between Different Response Categories
Response 3.0
Response 2.0
78,827
37,976
48,642
9465
104,762
Durations variability across Branch and Response categories
Branch ID 2
Response 5.0
Value
Duration (bin)
Distribution of Survey Durations
Count of Duration
Duration (bin)
Temporal Trends in Survey Duration
Time
Duration
Sunday
Monday
Tuesd...
Wedn...
Thurs...
Friday
Saturd...
Response 1.0
0
4,097
Branch ID 2
Worker ID
Null
7259
8498
8499
8500
8503
8514
8520
8528
8530
8533
8537
8542
Worker ID
157,671
4M
Duration (bin)
4.00
49.00
Data Source
Distribution of Responses for a S...
Temporal Trends in Survey Duration
Correlation Between Different Re...
Distribution of Survey Durations
Durations variability across Branc...
Dashboard 1
Survey Analysis
Type here to search
10:13 PM
11/26/2024

Survey Insights Dashboard:

This presents a comprehensive analysis of survey responses to identify trends, patterns, and relationships in the data. The visualizations focus on the distribution of responses, temporal changes, branch-level insights, and textual feedback, offering a multifaceted understanding of survey results.