**Data Quality Assessment and Cleaning Plan for Cases Dataset**

**Dataset Overview:**

The dataset encompasses various fields associated with cases, including case ID, type, location coordinates, closure details, and categorical attributes such as department, neighborhood, and category. Each field has undergone evaluation for missing values, uniqueness, and the range of values.

**Issues Identified:**

**DAYS TO CLOSE:**

Problem: 1.7% missing values.

Plan: Consider imputing missing values using a suitable method, such as mean or median, since it represents a numeric variable.

**CLOSED DATE:**

Problem: 0.8% missing values, and date format inconsistency (mixed formats: mm/dd/yyyy).

Plan: Impute missing values using an appropriate strategy. Standardize date format to a consistent format (e.g., yyyy-mm-dd) for consistency and ease of analysis.

**POLICE DISTRICT:**

Problem: 2.3% missing values.

Plan: Explore options for imputing missing values, considering the nature of the data. Potential methods include using the most frequent value or predictive modeling if feasible.

**CATEGORY2:**

Problem: 64.1% missing values.

Plan: Assess the impact of the missing values on analysis. If significant, explore methods for imputation, considering the data distribution. If the impact is negligible, it may be reasonable to exclude this variable from certain analyses.

**NEIGHBORHOOD:**

Problem: 2.9% missing values.

Plan: Similar to the police district, explore imputation methods based on the nature of the data. Imputing with the most frequent value or using a predictive model might be considered.

**ZIP CODE:**

Problem: 0.1% missing values and inconsistency (mixed numeric and string formats).

Plan: Convert all ZIP codes to a consistent numeric or string format. Impute missing values using a suitable strategy based on the distribution of the data.

**CATEGORY3:**

Problem: 89.9% missing values.

Plan: Assess the importance of this variable for analysis. If critical, explore methods for imputation. If not, consider excluding it from certain analyses.

**COUNTY:**

Problem: 4.3% missing values.

Plan: Assess the impact of missing values on analysis. If significant, explore imputation methods. If the impact is negligible, consider excluding it from certain analyses.

**General:**

Problem: Some fields suggest combining certain value levels due to a small number of counts.

Plan: Investigate the specific values and counts to determine if combining is necessary for better analysis. This might involve grouping less frequent values into broader categories.

**Timestamp Fields (e.g., CREATION TIME, CREATION DATE):**

Problem: These fields may require conversion to a consistent timestamp format for analysis.

Plan: Convert these fields to a standard timestamp format for uniformity and ease of analysis.

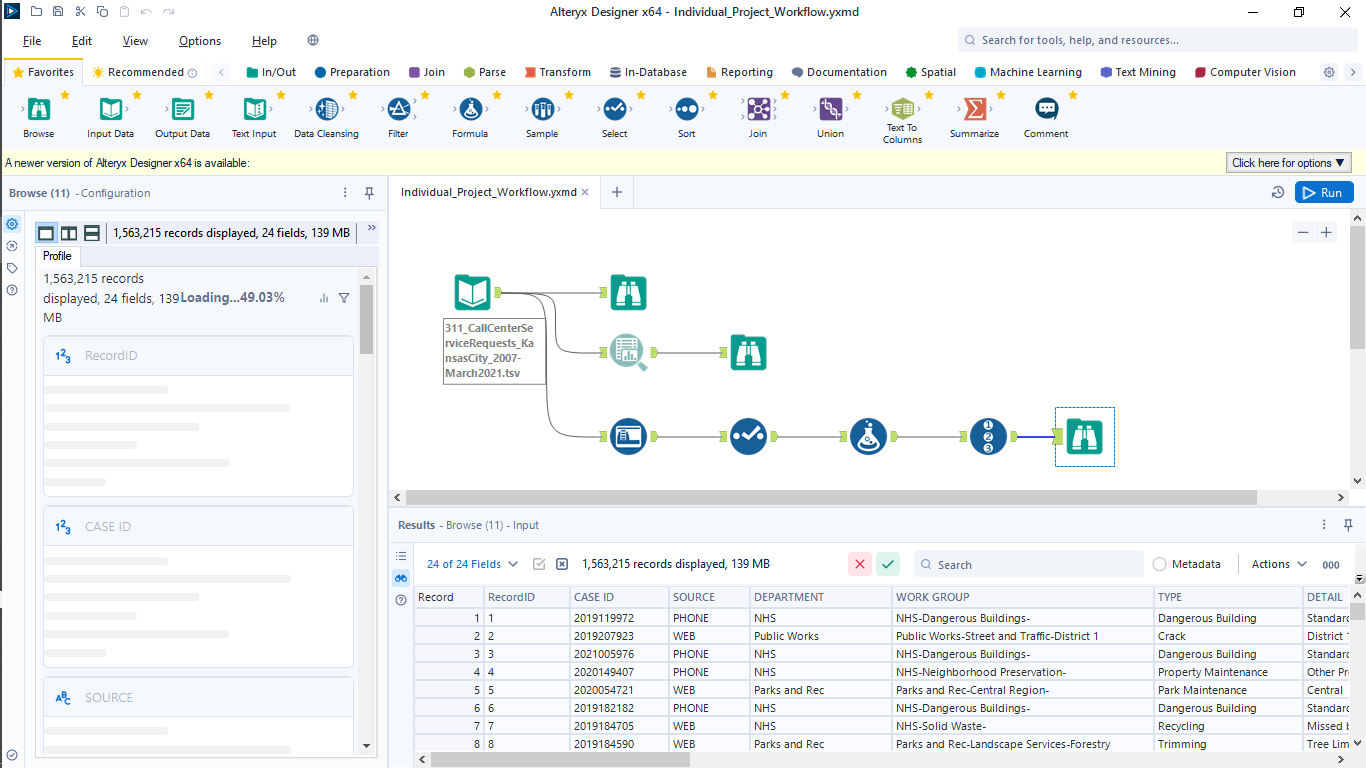
**Conclusion:**

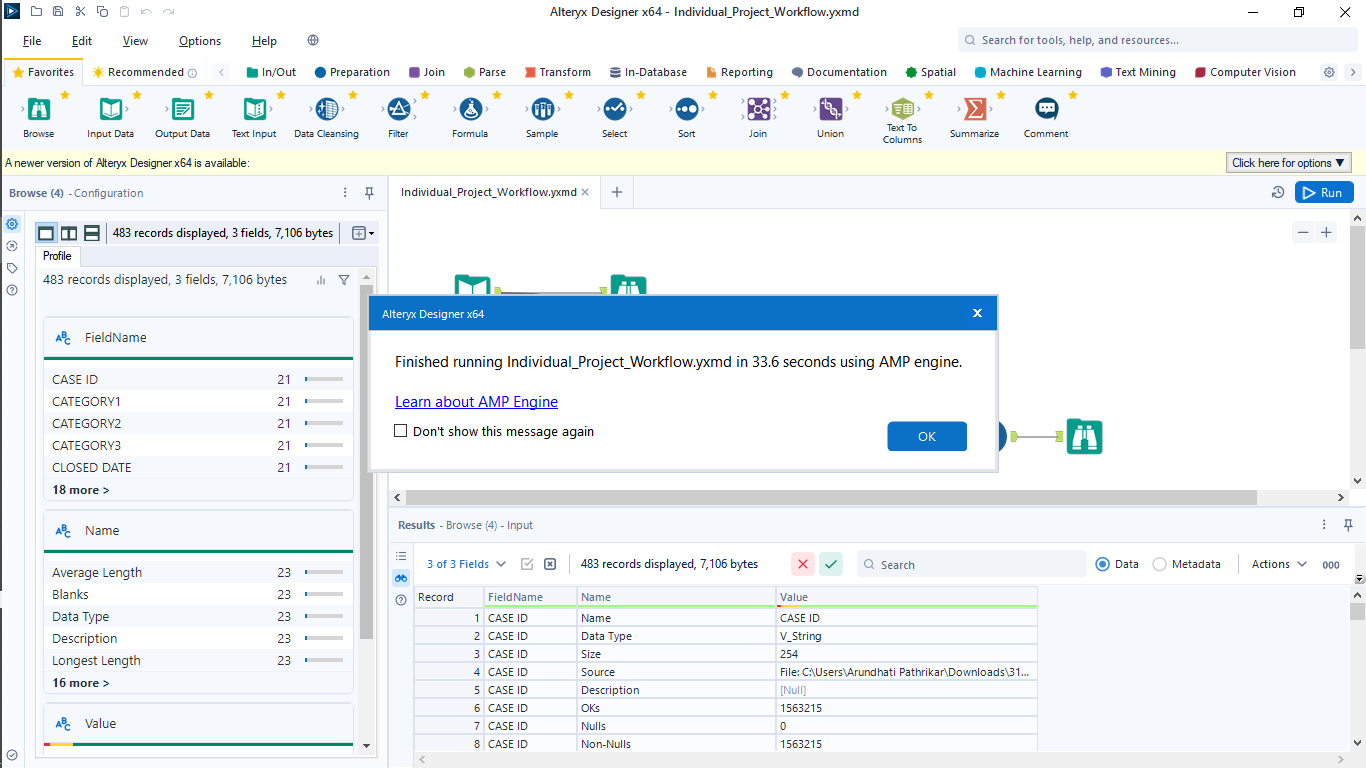
The dataset displays several data quality issues, encompassing missing values, inconsistent data types, outliers, and inaccuracies. Implementing data cleaning techniques will enhance the reliability and usability of the dataset for analysis and decision-making purposes.

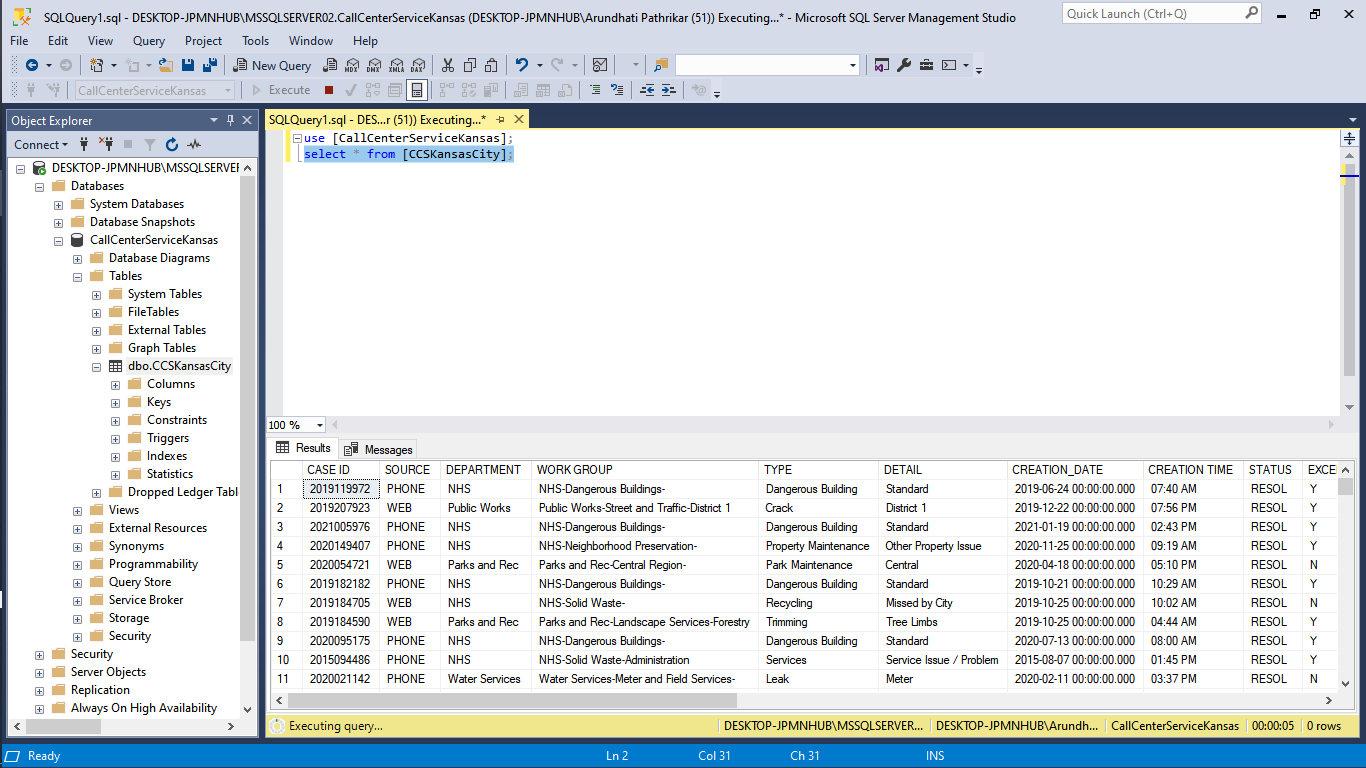
SQL Query :

1. SELECT YEAR(CREATION\_DATE) AS 'YEAR', COUNT([CASE ID]) AS 'SERVICE REQUESTS' FROM [CallCenterServiceKansas] GROUP BY YEAR(CREATION\_DATE) HAVING YEAR (CREATION\_DATE ) IN (2018, 2019, 2020, 2021);
2. SELECT MONTH(CREATION\_DATE) AS 'MONTH', YEAR(CREATION\_DATE) AS 'YEAR', COUNT([CASE ID]) AS 'SERVICE REQUESTS' FROM [CallCenterServiceKansas] GROUP BY YEAR(CREATION\_DATE), MONTH(CREATION\_DATE) HAVING YEAR (CREATION\_DATE ) IN (2018, 2019, 2020, 2021) ORDER BY YEAR(CREATION\_DATE), MONTH(CREATION\_DATE);
3. SELECT SOURCE, COUNT([CASE ID]) AS RequestCount FROM [CallCenterServiceKansas] GROUP BY Source ORDER BY RequestCount DESC;
4. SELECT DEPARTMENT, COUNT([CASE ID]) AS RequestCount FROM[CallCenterServiceKansas] GROUP BY DEPARTMENT ORDER BY RequestCount DESC;
5. SELECT TOP 10 [ZIP CODE], COUNT([CASE ID]) AS RequestCount FROM [CallCenterServiceKansas] GROUP BY [ZIP CODE] ORDER BY RequestCount DESC;
6. SELECT DEPARTMENT, COUNT([CASE ID]) AS Workload FROM CallCenterServiceKansas GROUP BY DEPARTMENT ORDER BY Workload DESC;
7. SELECT [WORK GROUP], COUNT([CASE ID]) AS Workload FROM CallCenterServiceKansas GROUP BY [WORK GROUP] ORDER BY Workload DESC;
8. SELECT DEPARTMENT, AVG([DAYS TO CLOSE]) FROM CallCenterServiceKansas GROUP BY DEPARTMENT ORDER BY DEPARTMENT ASC;
9. SELECT YEAR([CREATION\_DATE]) AS Year, STATUS, COUNT([CASE ID]) AS StatusCount FROM CallCenterServiceKansas WHERE YEAR(CREATION\_DATE) BETWEEN 2018 AND 2021 GROUP BY Year(CREATION\_DATE), STATUS ORDER BY YEAR ASC;
10. SELECT CATEGORY1, AVG([DAYS TO CLOSE]) AS AvgDaysToClose FROM CallCenterServiceKansas WHERE CATEGORY1 != 'Data Not Available' GROUP BY CATEGORY1 ORDER BY AvgDaysToClose DESC;
11. SELECT TOP 10 CATEGORY1, AVG([DAYS TO CLOSE]) AS AvgDaysToClose FROM CallCenterServiceKansas WHERE CATEGORY1 != 'Data Not Available' GROUP BY CATEGORY1 ORDER BY AvgDaysToClose DESC;
12. SELECT DEPARTMENT, COUNT([CASE ID]) AS Workload, AVG([DAYS TO CLOSE]) AS AvgEfficiency FROM CallCenterServiceKansas GROUP BY DEPARTMENT ORDER BY DEPARTMENT;

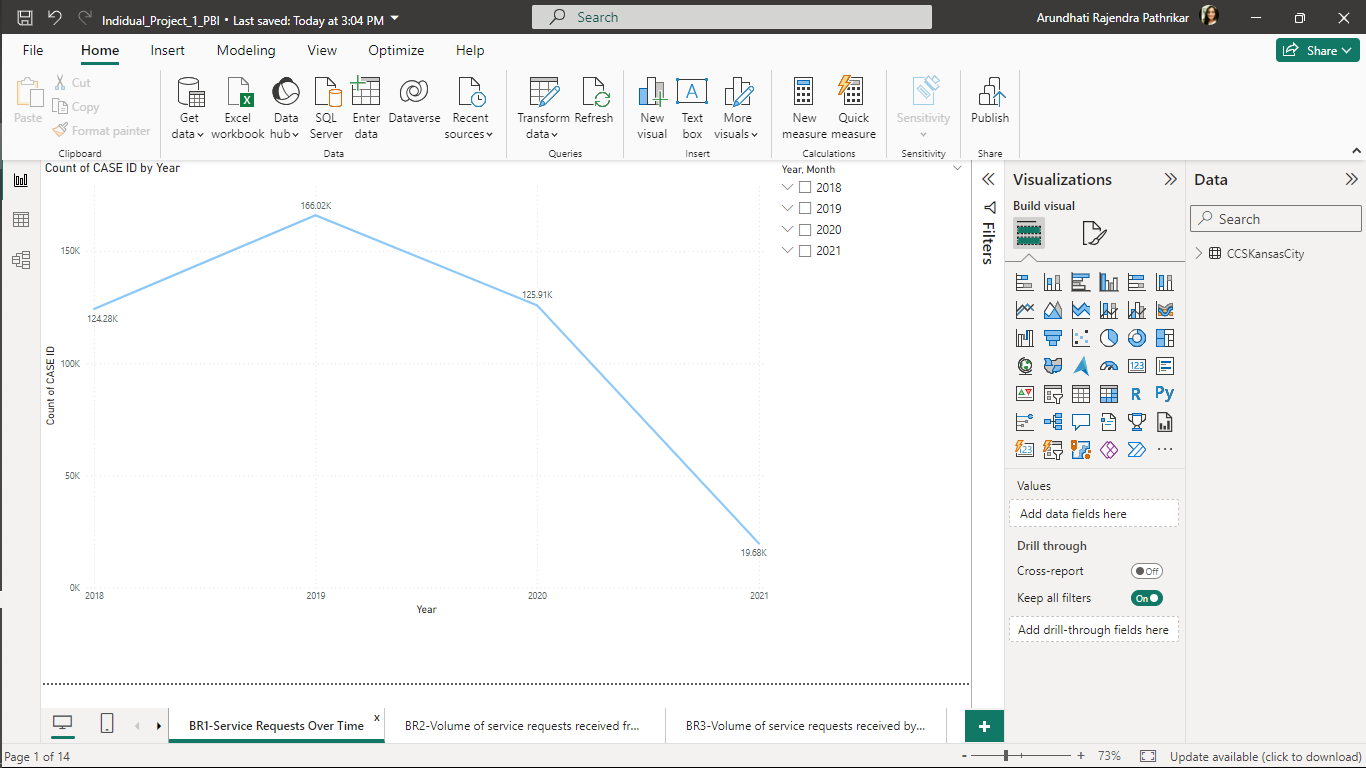
Alteryx:



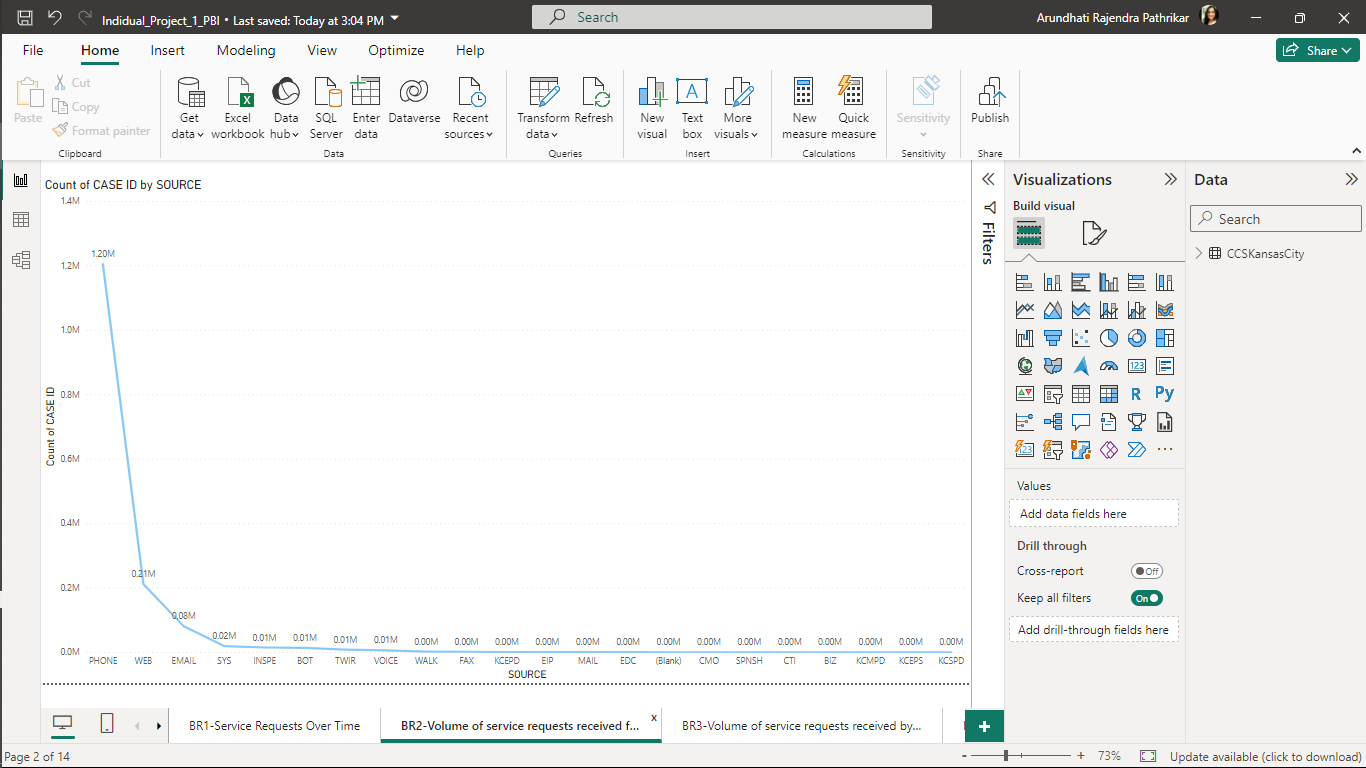


PowerBI screenshots:

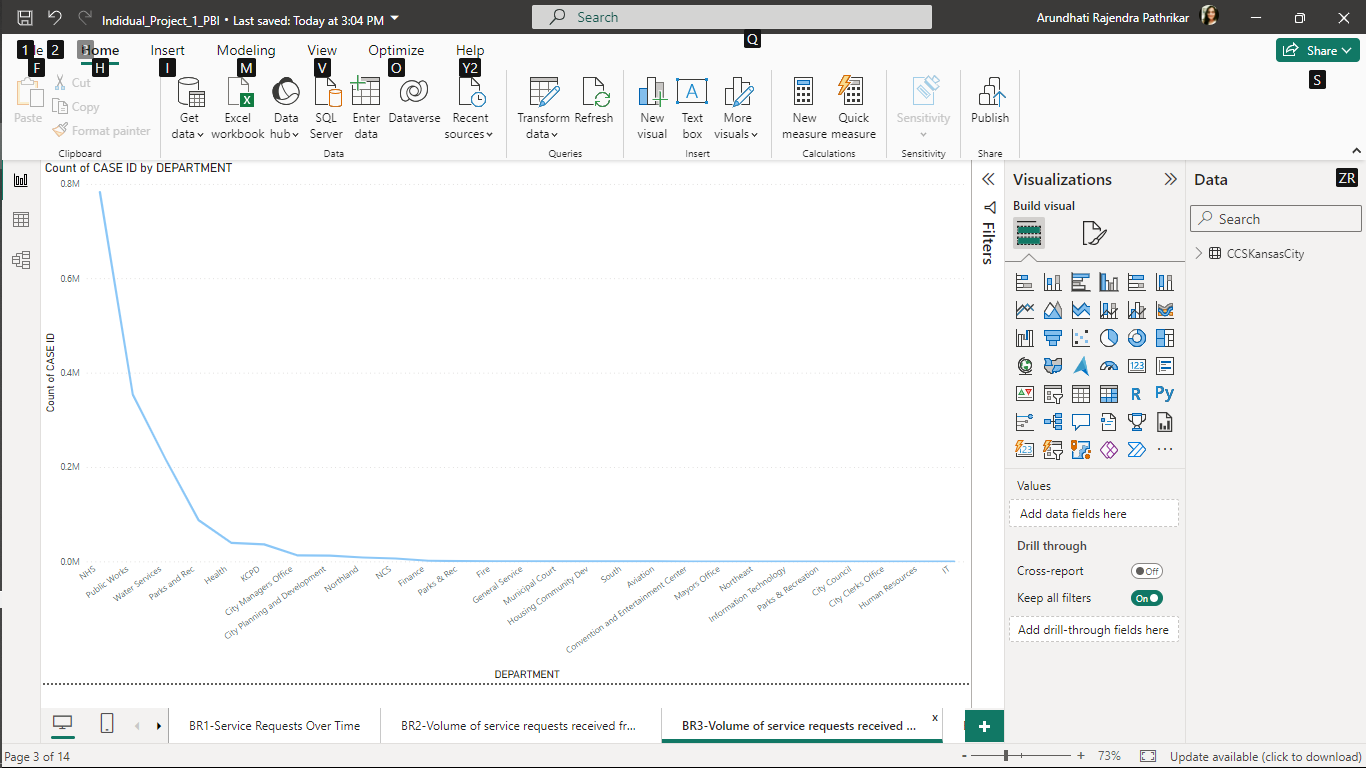
BR1



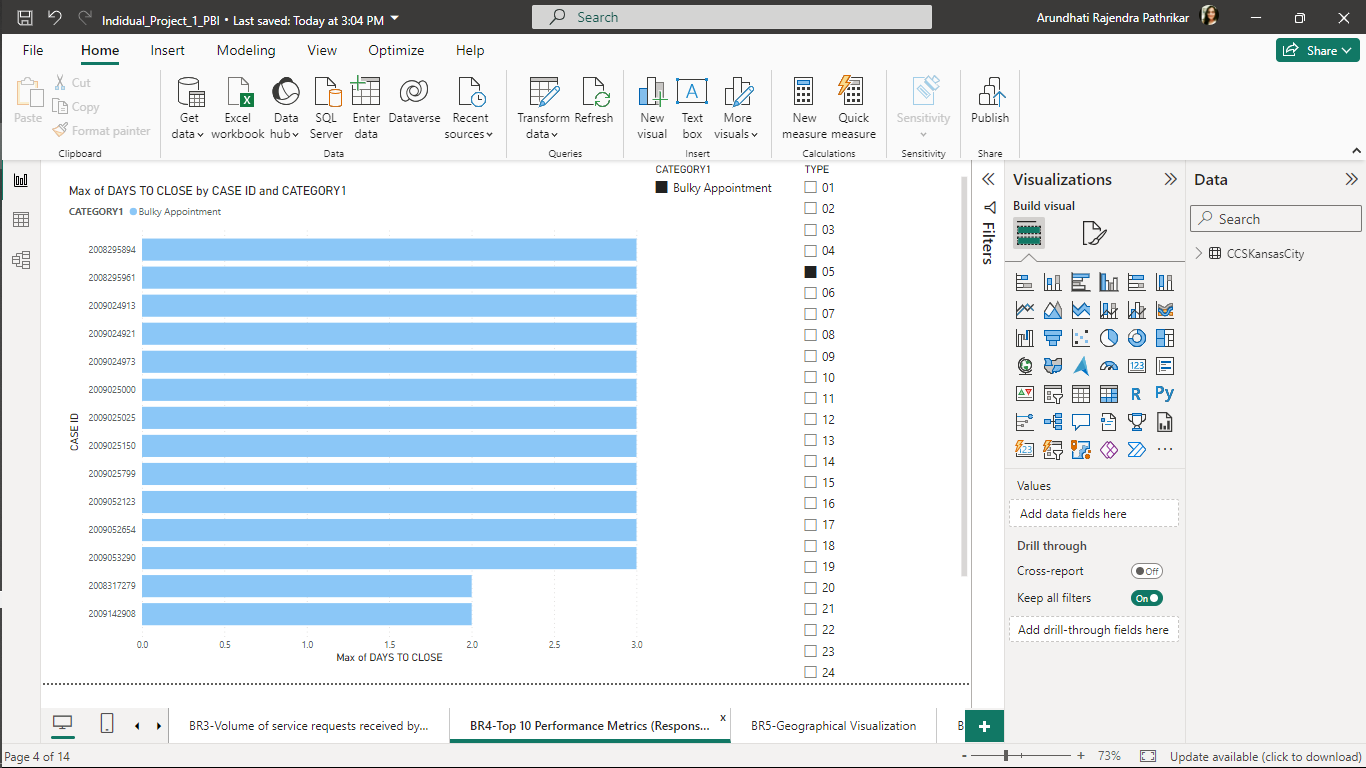
BR2



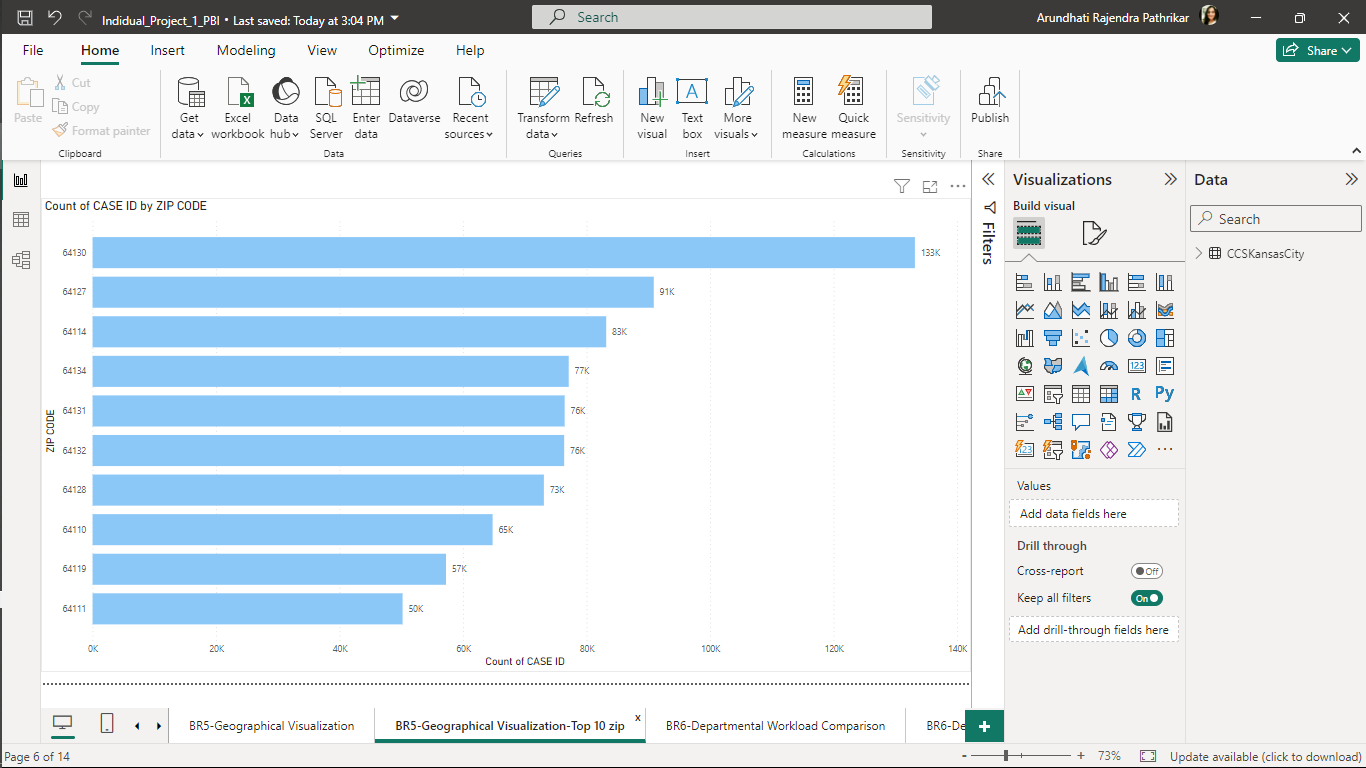
BR3

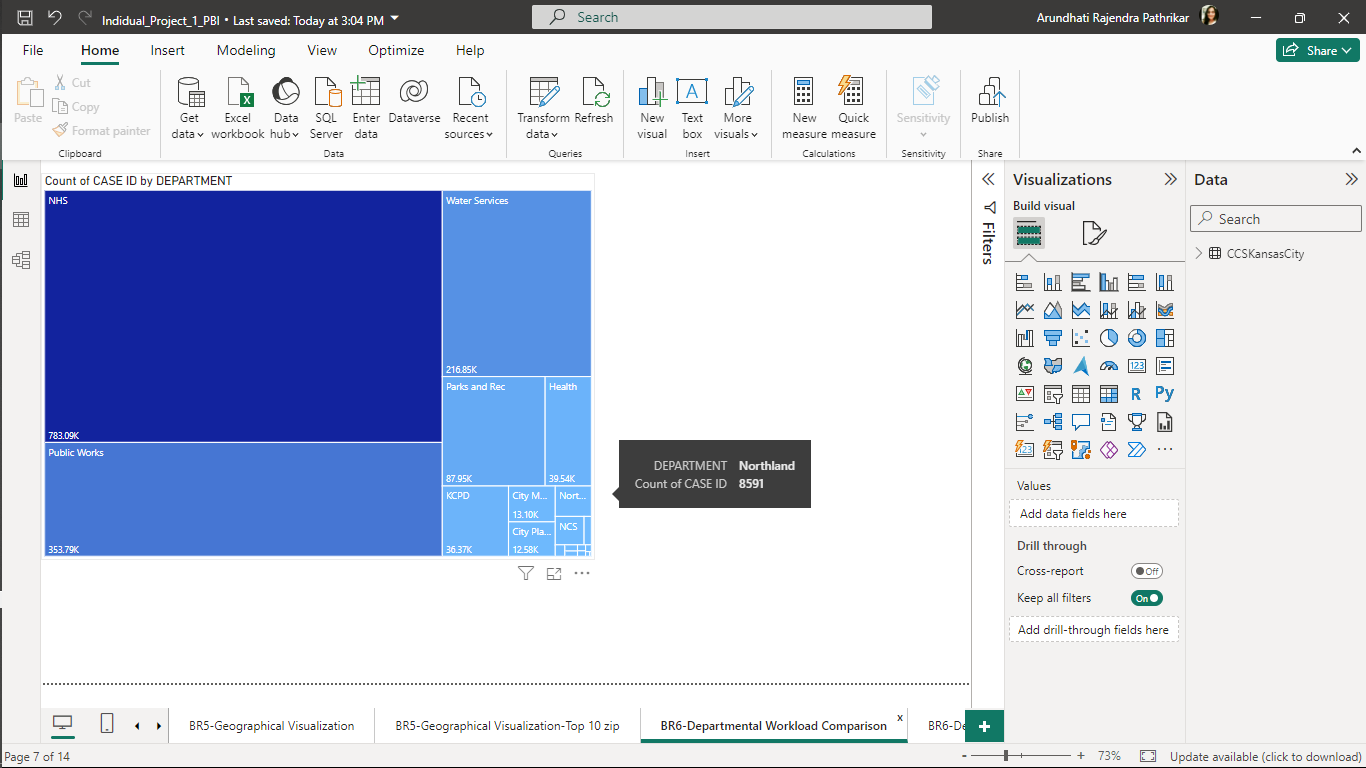


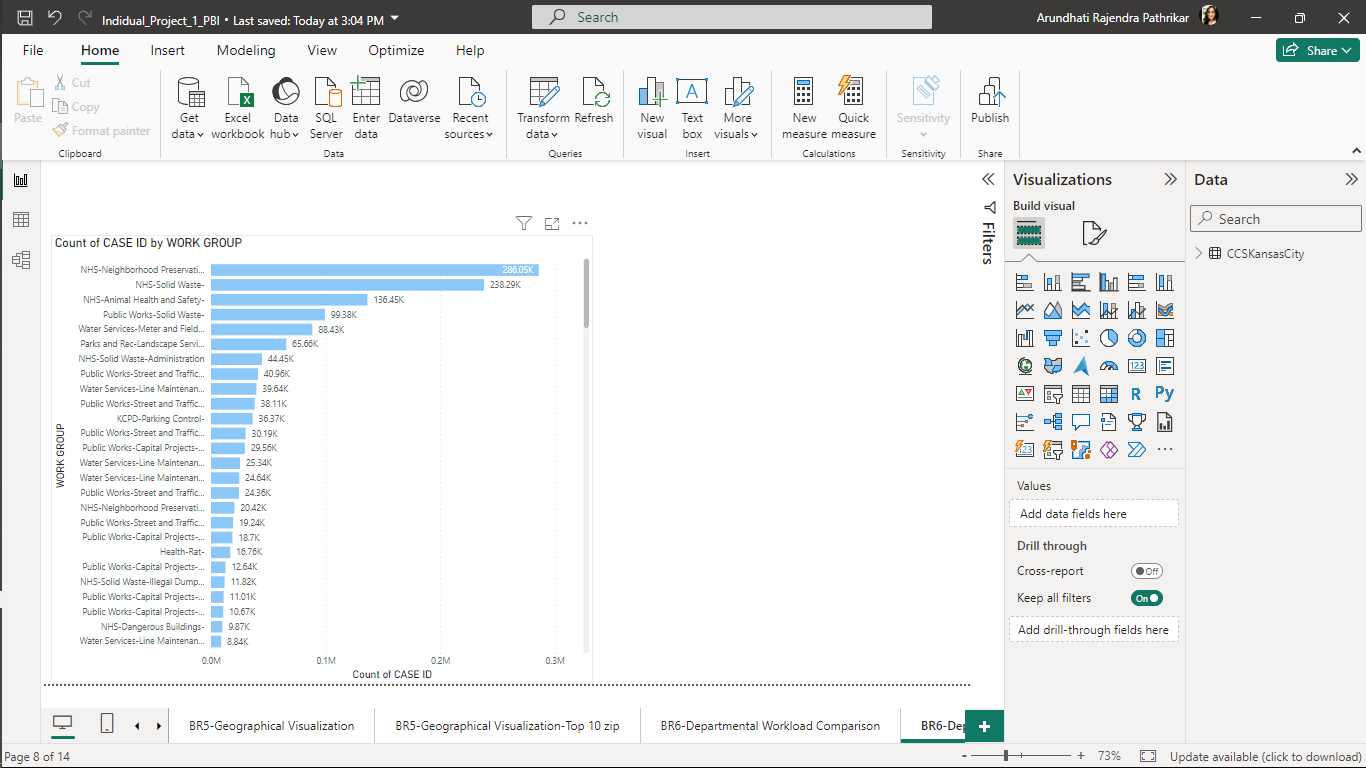
BR4

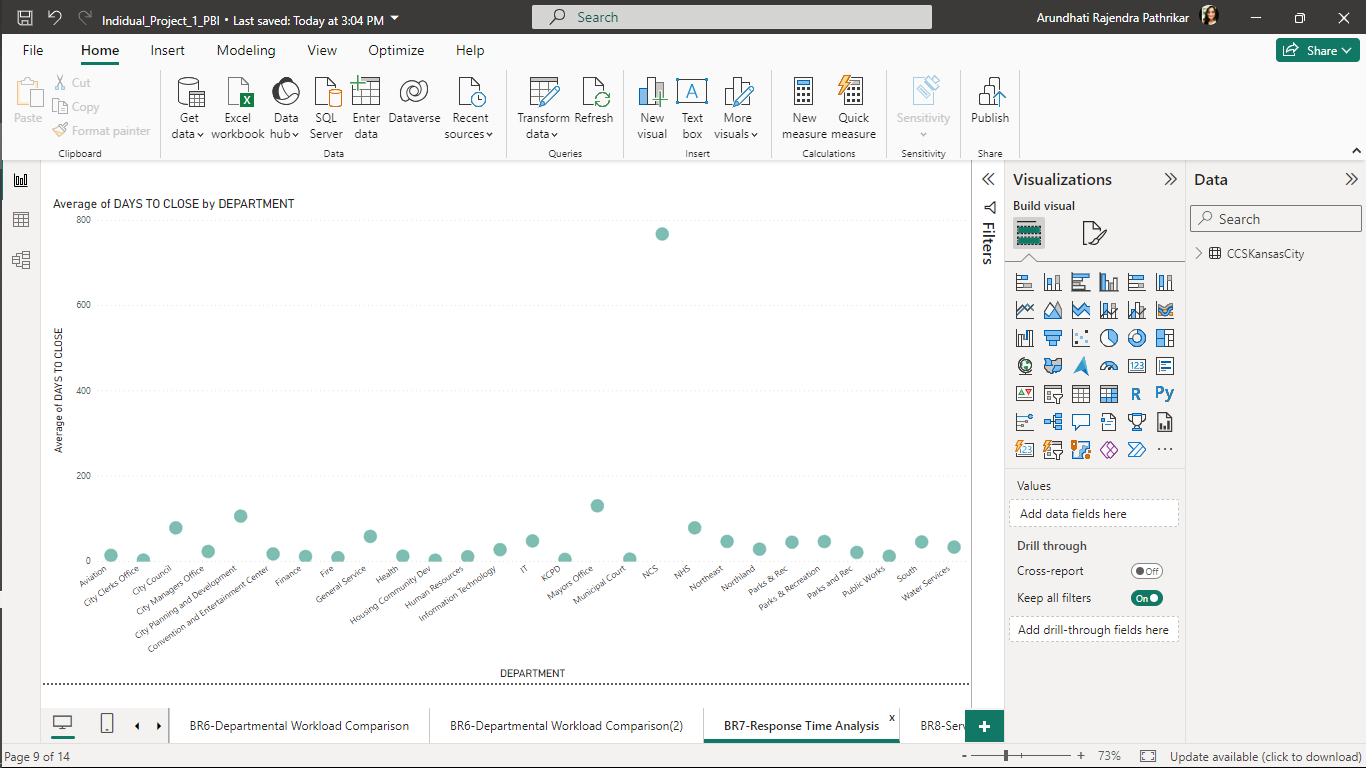


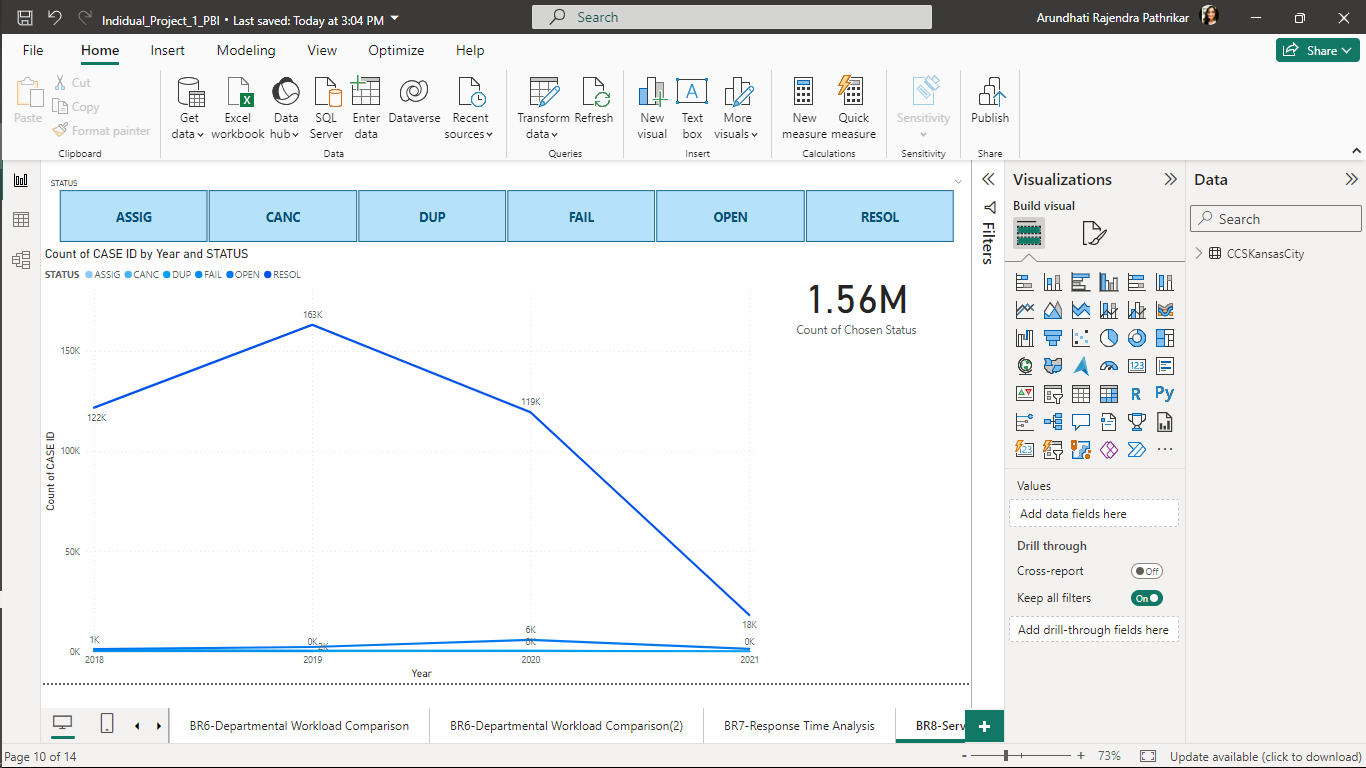


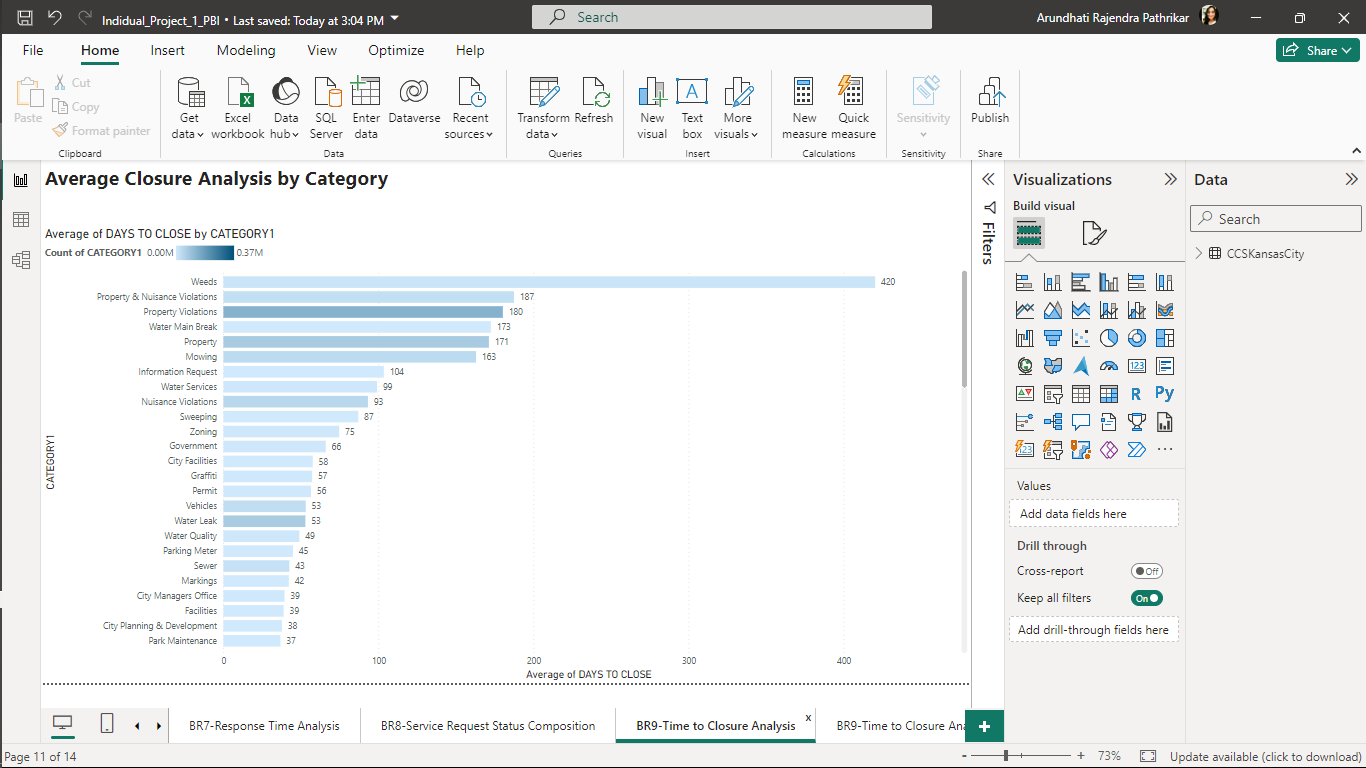


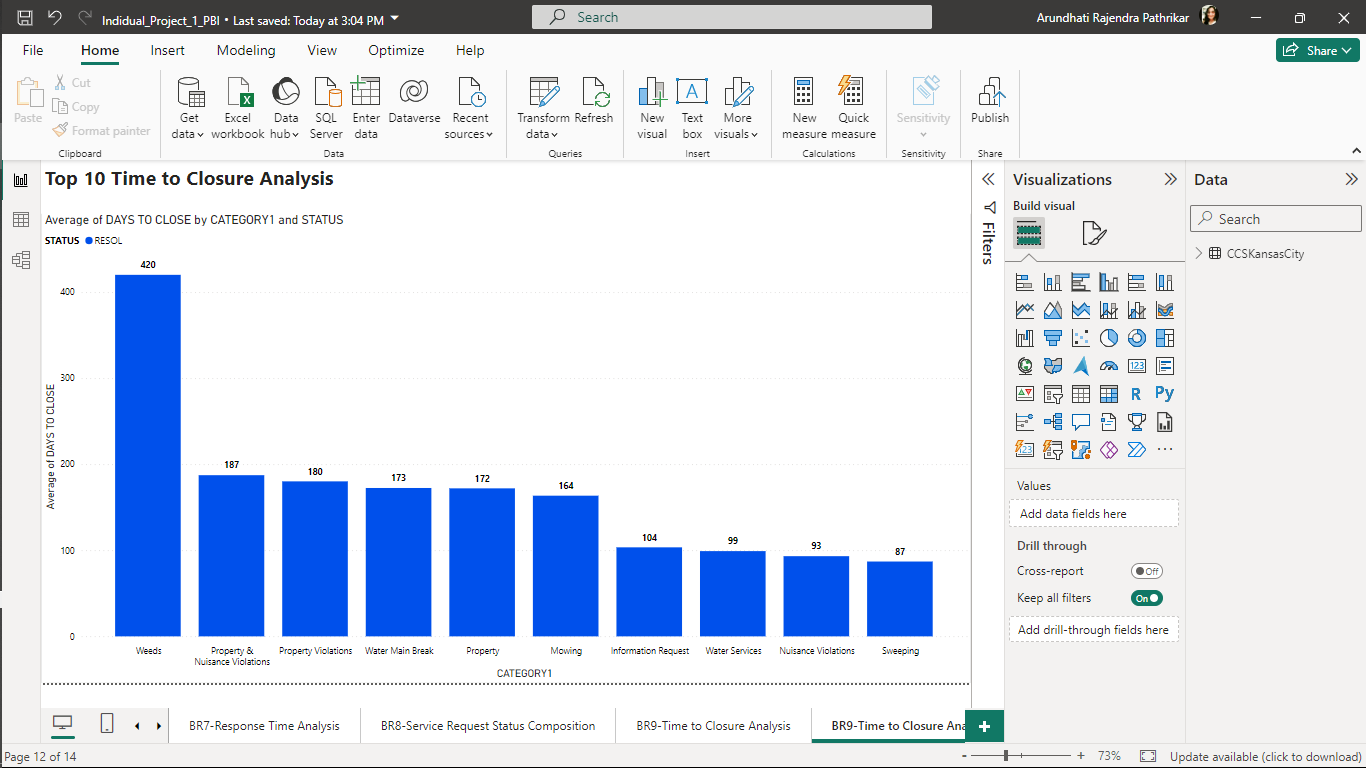


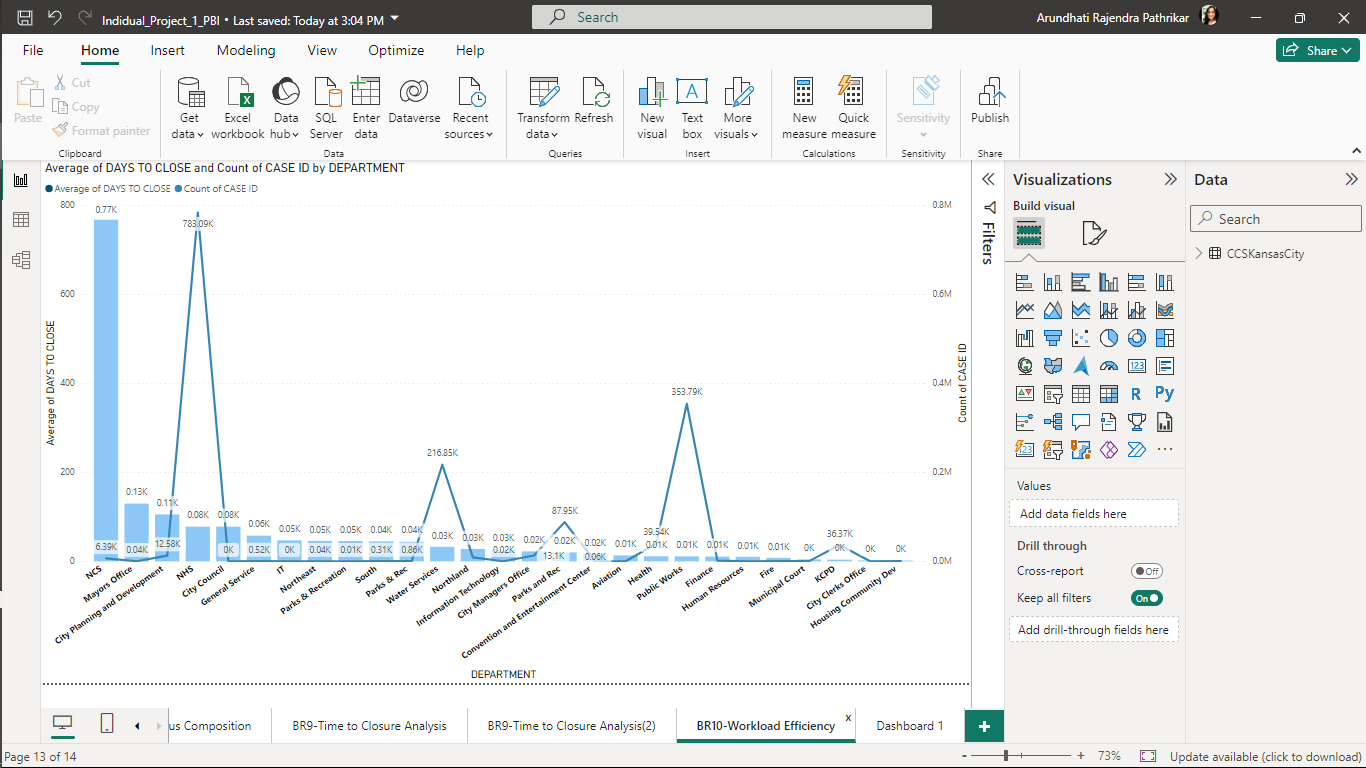


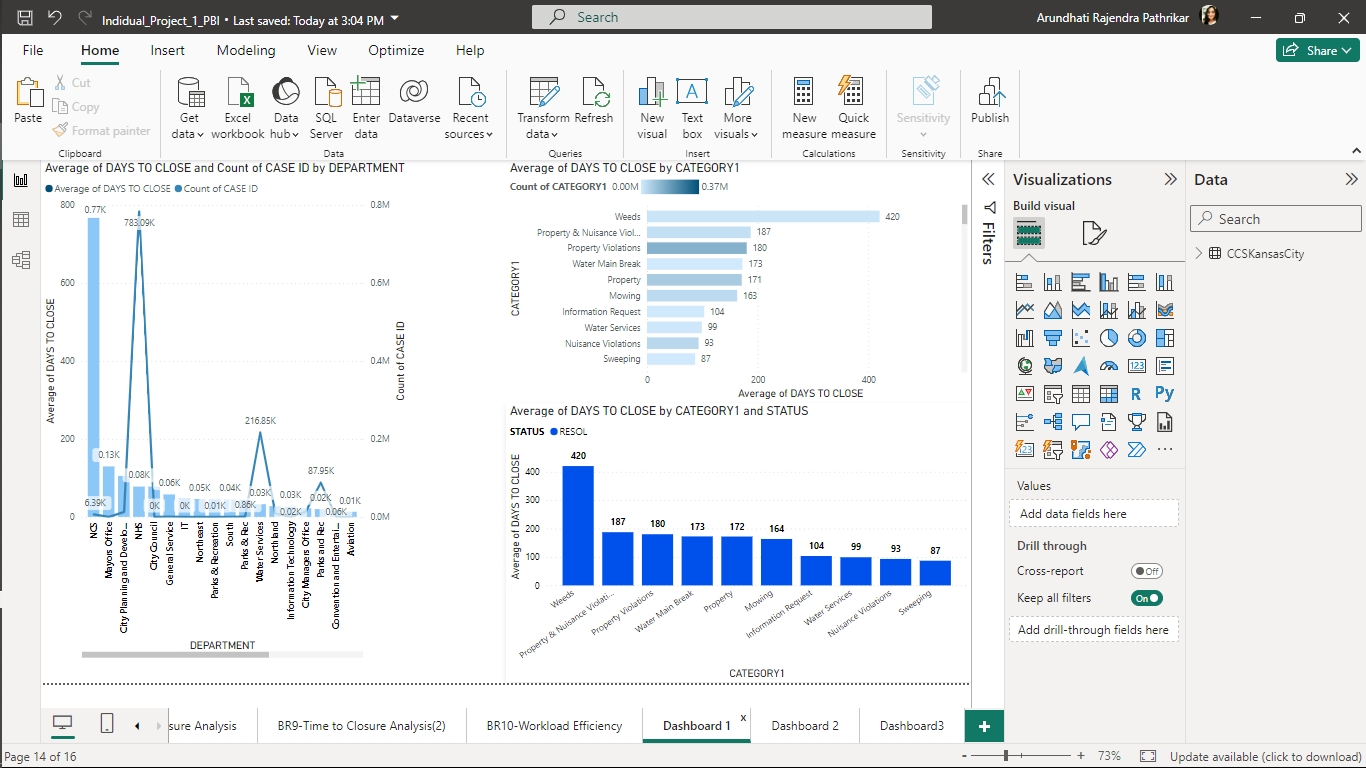


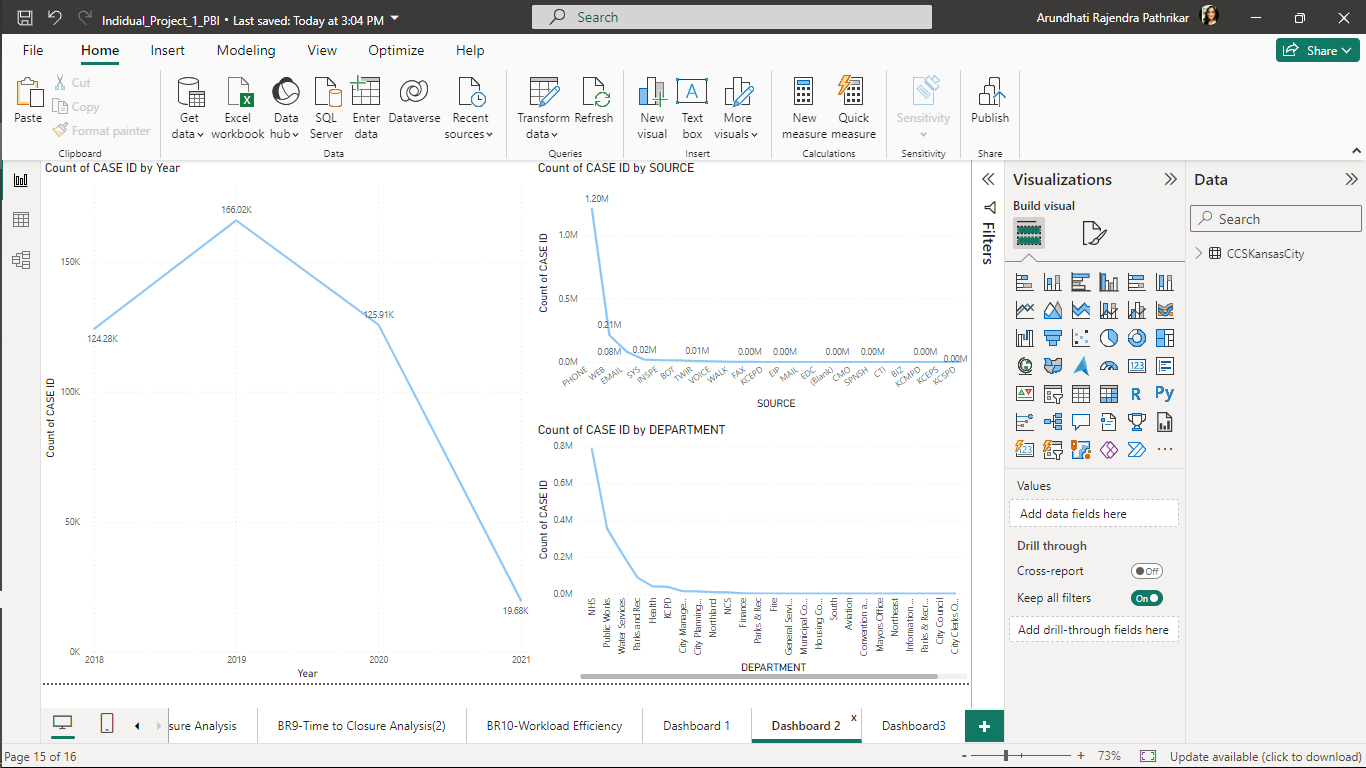


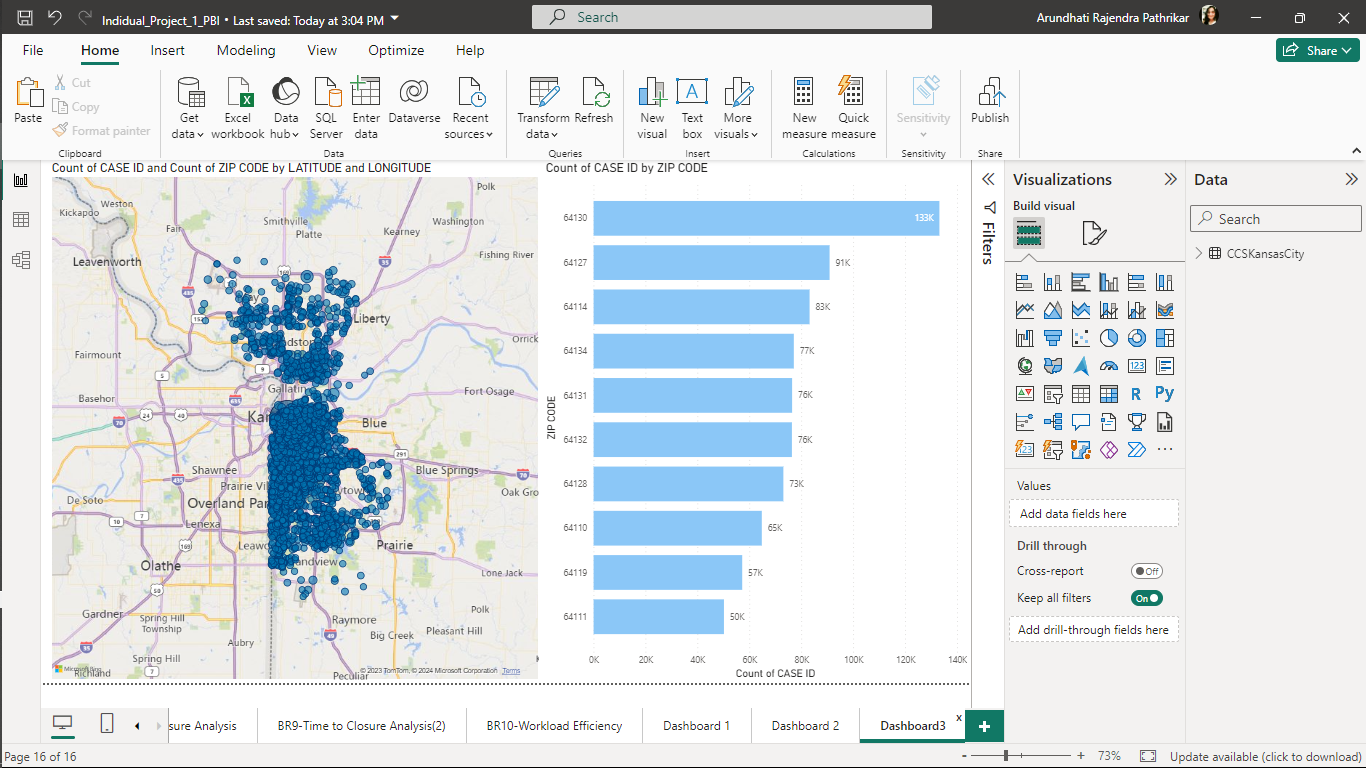












TABLEAU

