**Here is the excel project (just click on the link to download it):**

<https://docs.google.com/spreadsheets/d/1TagEgL7AwRL_VS5jaLnYhyrQC19vJg_9/edit?usp=sharing&ouid=109480930024828513670&rtpof=true&sd=true>

**First, we began by cleaning the data:**

* **Column Selection for Cleaning:** We selected specific columns to clean based on the questions we were addressing:
  + For Question 1, "What are the most common reasons for delays and breakdowns?" we selected the bus\_company and How\_Long\_Delayed columns.
  + For Question 3, "Is there a correlation between specific days of the week and the frequency of breakdowns or delays?" we selected the Occurred\_On column.
  + Columns marked in orange indicate those that required cleaning, while those in green represent the cleaned data.
  + We utilized the WEEKDAY function to analyze the Occurred\_On date, resulting in numbers ranging from 1 to 6, representing days from Sunday (1) to Friday (6). Notably, Saturday was absent, suggesting that buses don't operate on this day.
* **Data Cleaning:**
  + We observed inconsistencies in the bus\_company names, such as "SMART PIC" and "SMART PIC INC," and standardized them to a single name.
  + The How\_Long\_Delayed column was divided into two distinct columns: one for the short delay time estimate and another for the long delay time estimate.
* **Addressing the Common Delay Question:**
  + We took the reason and breakdown\_or\_running\_late columns and created a pivot table. We placed reason in the rows, counted occurrences, and placed breakdown\_or\_running\_late in the columns. Afterward, we generated a chart using the column data.
* **Analyzing Delay Times by Bus Company and Borough:**
  + A pivot table was created, placing the cleaned bus\_company column in the rows area.
  + We added the short and long delay time estimates to the values area and calculated their average.
  + A bar chart was then generated, which highlighted the need to further filter the data due to its messiness.
  + To refine the analysis, we focused on the top 10 averages of short delay time estimates.
* **Exploring Correlations Between Weekdays and Breakdown/Delay Frequency:**
  + A pivot table was created, where we initially cleaned the data by assigning numbers (1 to 7) to the Days\_of\_week columns. However, we later realized that using day names instead of numbers would provide clearer representation in a line chart.
  + The first pivot table included Day\_Of\_Week in the rows, the count of breakdowns\_or\_running\_late in the values, and breakdowns\_or\_running\_late in the columns, resulting in a line chart.
    - The column label selected was "running late."
  + The second pivot table included the same structure but focused on breakdown as the selected column label.