12/20/2024

Trend Analysis of Transit Crime Complaints on Daily Subway Ridership

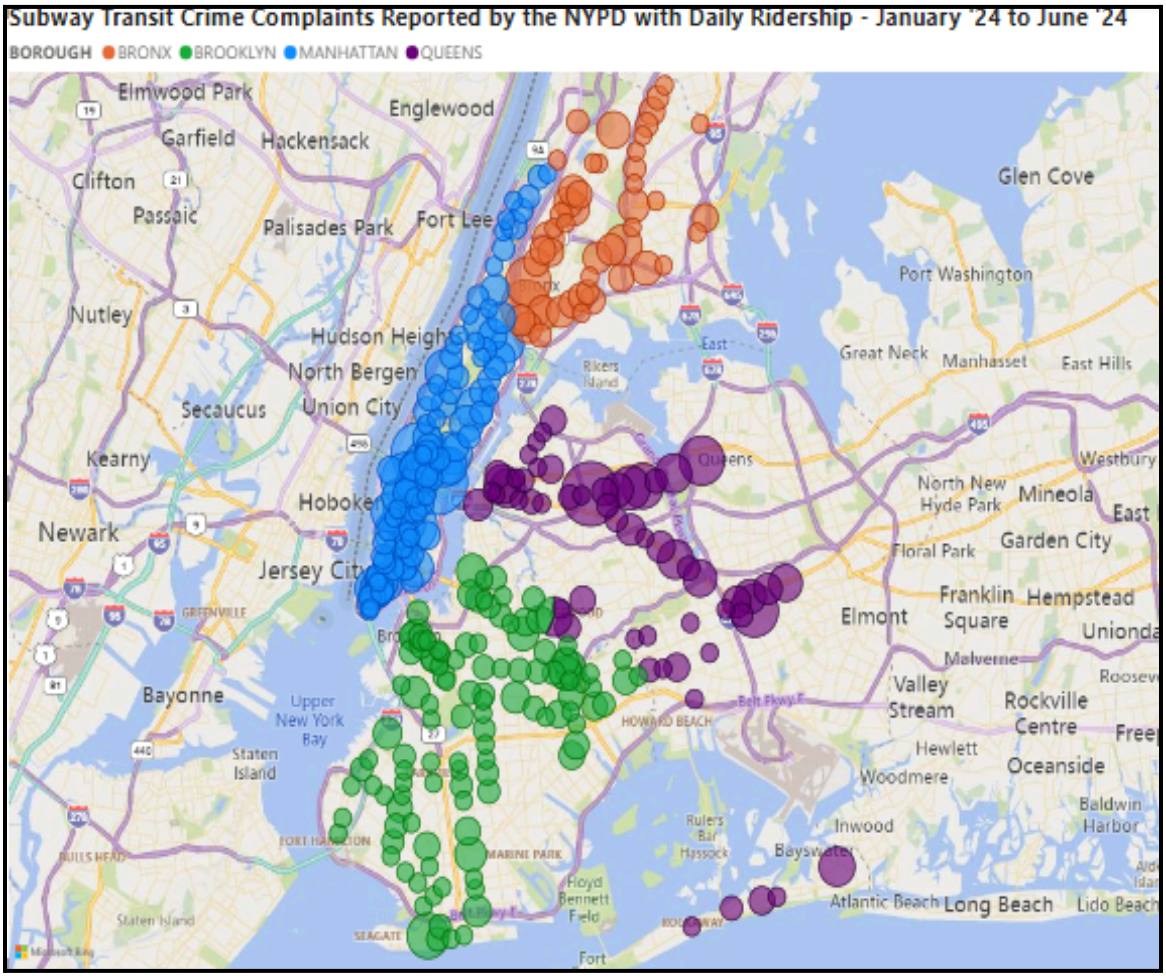
Aniruddha "Partha" Parthasarathy
College Aide (Graduate) - Fall 2024
Bus Technology, MTA New York City Transit

Spread of Subway Station Locations Across NYC

- **Manhattan** has a dense cluster of stations along major north-south routes.
- **Brooklyn** and **Queens** have broader coverage but with sparser clusters.
- **The Bronx** stations are concentrated primarily in the southern half.
- Subway lines are lesser toward the border of **Queens**.

Summary of Exploratory Data Analysis (EDA):

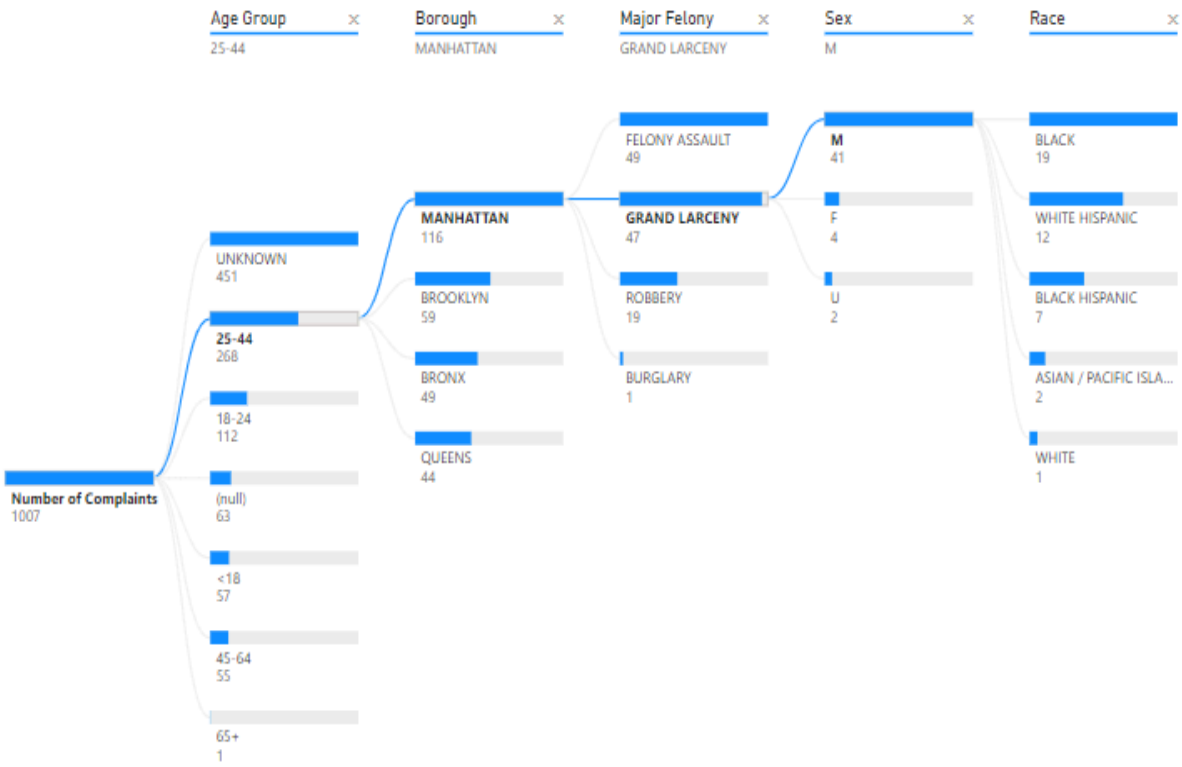
- **Total # Days in the Time Period**: 182 *(01/01/24 to 6/30/24)*
- **Total # Felony Complaints:** 2,254
- **Total # Felony Types:** 16
- **Total # Major Felony Complaints:** 1,007
- **Types of Major Felonies:** Grand Larceny, Felony Assault, Robbery, Burglary, Rape



Major Felony Complaints - Dense Cluster of Complaints Around Midtown Manhattan

- **# Days with Major Felony Complaints**: 182
- **Average # Daily Complaints** (All Major Felonies Combined): 5.53
- **Maximum # Major Felony Complaints in a Day**: 15 (occurred on May 4)

- **Most Common Type of Major Felony:** Grand Larceny with 541 complaints (24% of total) and a daily average of 2.97 complaints.



Decomposition Tree of Major Felony Complaints

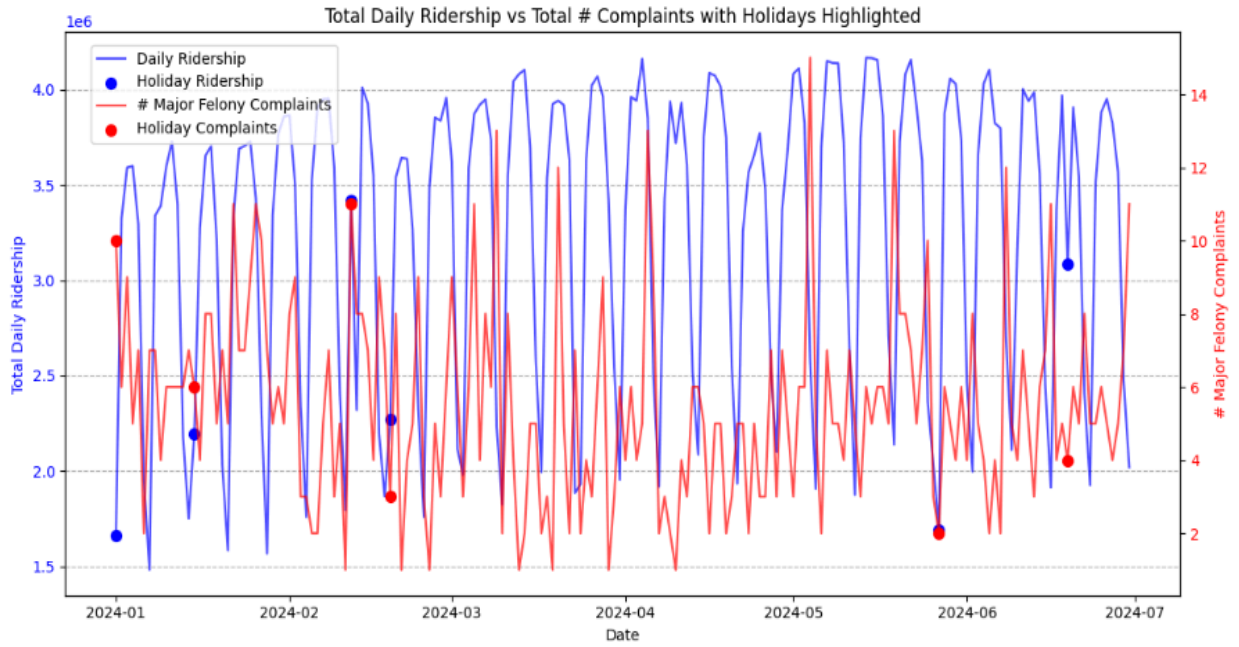
Daily Ridership Statistics on Complaint Dates:

Per Complaint Station, for Major Felonies:

- **Total Number of Complaint Stations:** 256
- **Average Daily Ridership:** 10,784.23
- **Standard Deviation:** 3,412.59
- **Median Ridership**: 5,598.0
- **Minimum Ridership:** 10.0
- **25% of Stations had ridership ≤:** 9,452.82
- **75% of Stations had ridership ≤:** 12,352.86
- **Maximum Ridership:** 158,893.0

NYC FISA Payroll

2024 List of Holidays	
Monday, January 1	New Year's Day
Monday, January 15	Martin Luther Kind, Jr.Day
Monday, February 12	Lincoln's Birthday (Floating Holiday)*
Monday, February 19	President's Day
Monday, May 27	Memorial Day
Wednesday, June 19	Juneteenth



The above plot confirms a cyclic pattern of peak ridership during weekdays and dips on weekends.

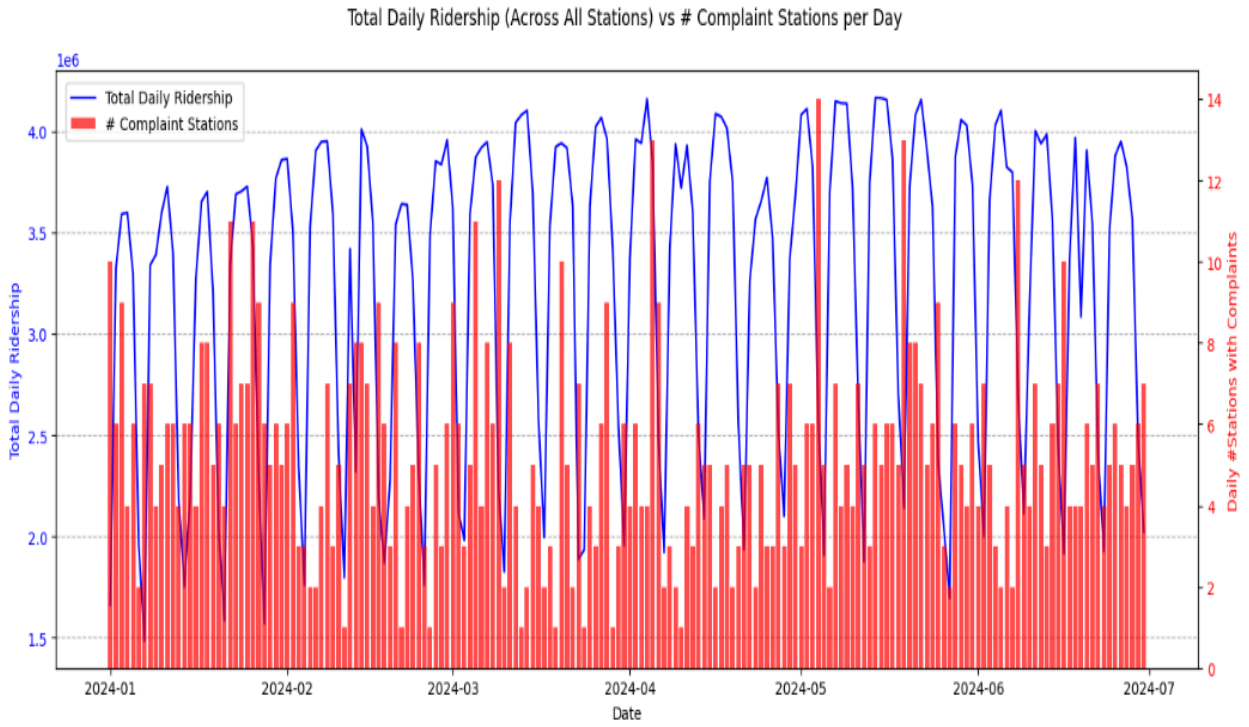
Daily subway ridership significantly decreases on holidays, but it picks up sharply afterwards, while the immediate trend in number of complaints remains variable. In general, **complaints show sporadic peaks but have been lower on all holidays** (except NYD and Lincoln's birthday)

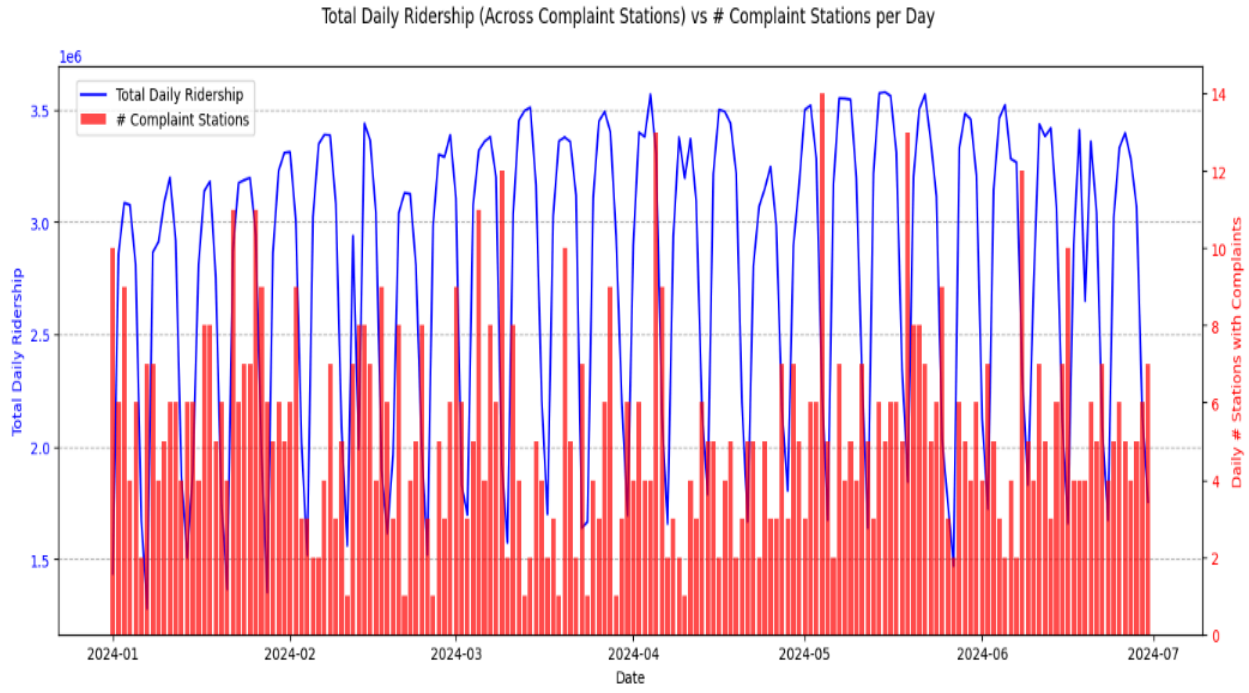
Type of Day	Average Daily Ridership	Total # Complaints	Average # Complaints per Day
Holiday	14,550.08	36	6.0
Weekday (non-Holiday)	27,992.62	649	5.23
Weekend	18,318.98	322	6.19

Holidays and Weekends:

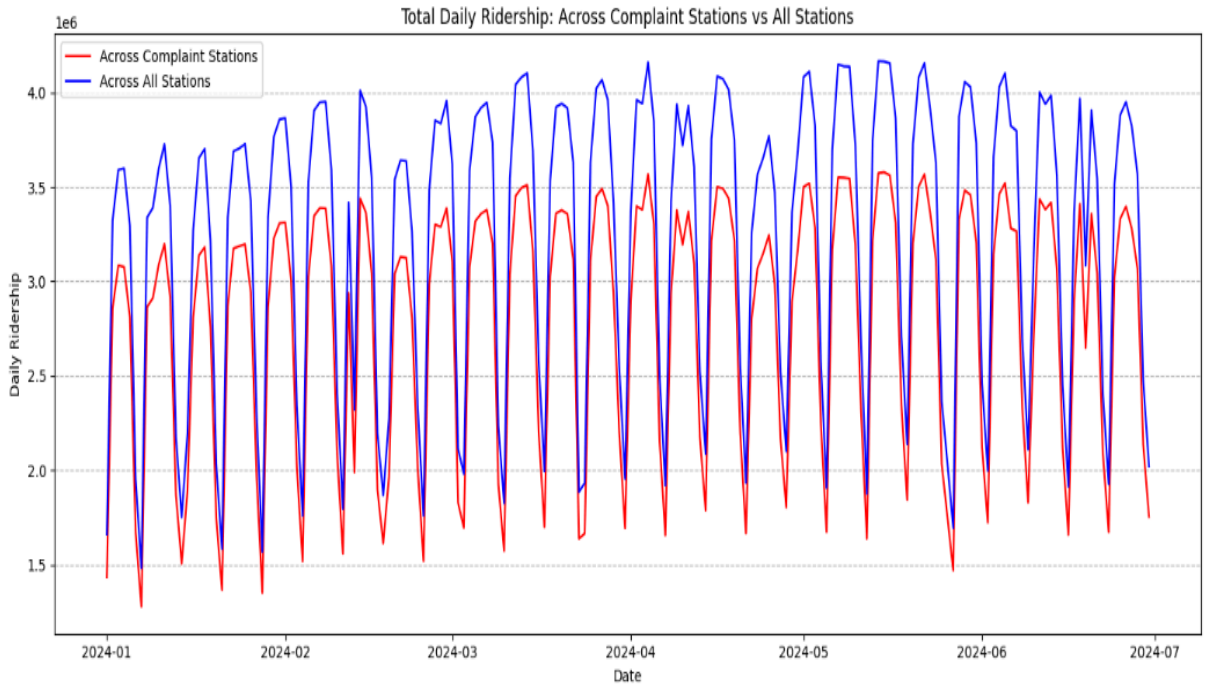
- Both holidays and weekends show elevated complaint averages, suggesting these days may attract more crimes.
- This could be linked to higher social activity, events, and tourism.

Non-Holiday Weekdays:
Weekdays without holidays have the lowest average complaints, likely reflecting regular travel patterns.

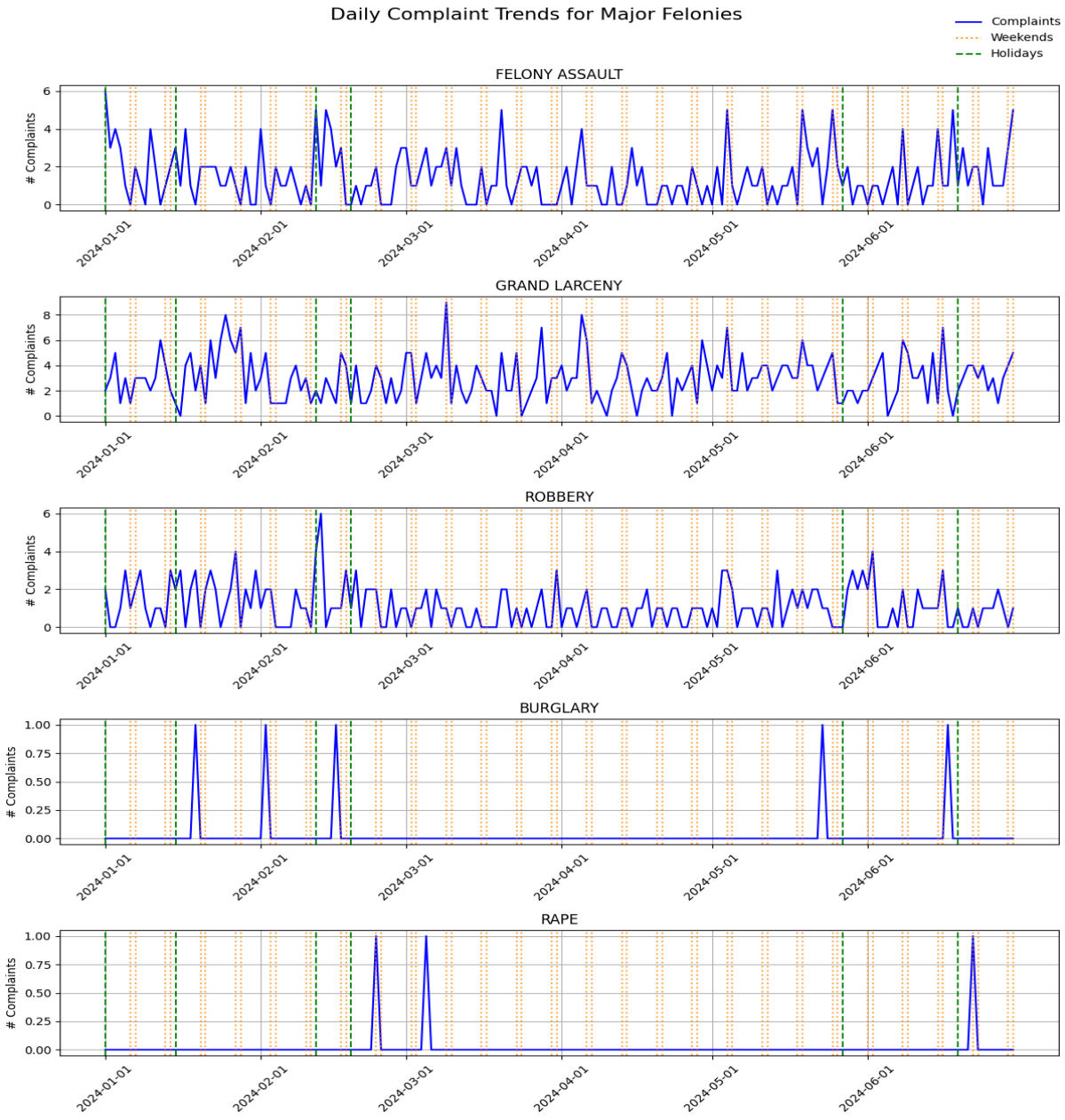




More stations report complaints on days with high ridership peaks.



Ridership across complaint stations closely mirrors the trend across all stations. Complaint stations contribute significantly to overall ridership patterns. On average, **85.9% of the total daily ridership in the system comes from stations with major felony complaints**. Since felony station ridership accounts for a large share of it, the two plots above will look very similar.



- **Felony Assault and Grand Larceny** are the most frequent offenses, with **persistent daily fluctuations but have no consistent weekend/holiday correlation**.

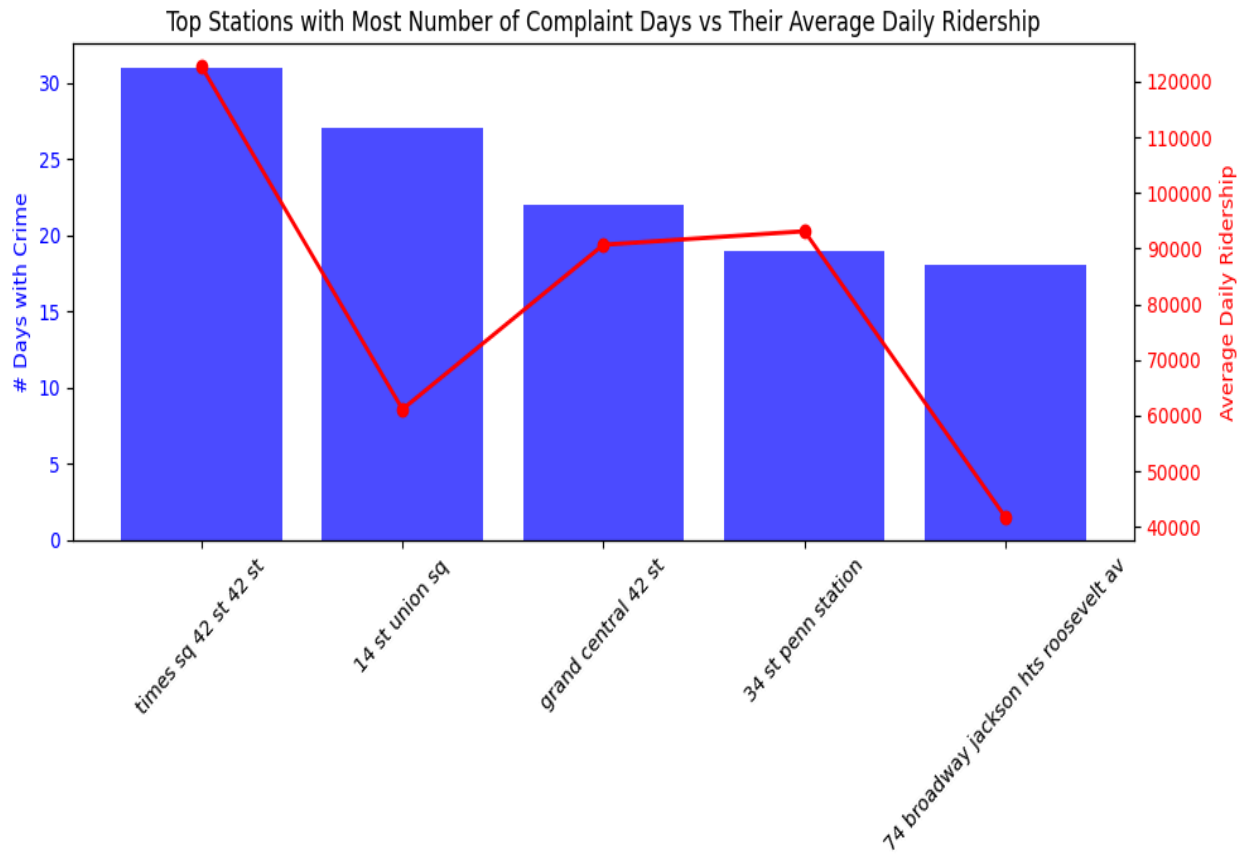
- **Robbery** complaints are less frequent but show occasional spikes, **mostly midweek**.

- Burglary and Rape occur rarely with isolated incidents with **Burglary having more presence near holidays**.

- Overall, there is **no clear drop or rise in major complaints during holidays and weekends**, indicating consistent complaint behavior.

Top Stations - stations with at least ~10% of Days with Complaints (18.2 out of 182 days):

- **Times Sq - 42 St:** 31 days
- **14 St - Union Sq:** 27 days
- **Grand Central - 42 St:** 22 days
- **34 St - Penn Station:** 19 days
- **74 Broadway - Jackson Heights Roosevelt Av:** 18 days



Higher daily ridership usually correlates with more complaint days except for **14 St-Union Sq** and **74 Broadway - Jackson Heights Roosevelt Av** which means that factors apart from felony complaints may be affecting ridership at these particular stations, such as socio-economic environment and design vulnerabilities (like poor lighting, entrance/exit congestion, lack of patrol presence).

The table below shows the distribution of top major felonies across top complaint stations:

Station Name	Grand Larceny (# Complaints & # Days)	Felony Assault (# Complaints & # Days)	Robbery (# Complaints & # Days)	Total (# Complaints & # Days)
Times Sq-42 St / 42 St	15 (15)	10 (8)	7 (7)	32 (30)
14 St-Union Sq	21 (20)	4 (3)	4 (4)	29 (27)
Grand Central 42 St	23 (21)	0 (0)	2 (2)	25 (23)
34 St - Penn Station	13 (13)	5 (5)	2 (2)	20 (20)
74 Broadway - Jackson Heights Roosevelt Av	19 (16)	3 (3)	0 (0)	22 (19)

Conclusions:

- While overall ridership remains resilient to crimes like Grand Larceny, more **severe crimes like Felony Assault may trigger sharper ridership declines**.
- Higher daily ridership usually pairs with more felony complaints. But **complaint frequency, along with other factors, influence ridership at** stations like **14 St-Union Sq** and **74 Broadway - Jackson Heights Roosevelt Av**.

Further Analysis Recommendations:

- **Time-of-Day Analysis**: Break down felony trends by different times to identify high-risk periods.
- **Demographic Insights**: Correlate age group, gender, and offender/victim profiles with felony types to develop targeted awareness programs.
- **Anomaly Analysis**: Investigate outliers like **14 St-Union Sq** and **74 Broadway** backed by statistical analysis to identify structural, social, or external factors' impact on transit crime.
- **Impact of Events**: Assess subway crime patterns during large city events (e.g., parades, high profile sports events, major holidays) to determine how special activities influence complaint trends.

- **Predictive Modeling**: Use more historical ridership and complaint data to better forecast high-risk crime days and allocate patrolling personnel resources proactively.

Post-internship Analysis

Anomaly Stations: 14 St-Union Sq and 74 Broadway - Jackson Heights Roosevelt Av

Non-anomaly Stations: Times Sq - 42 St, Grand Central - 42 St and 34 St - Penn Station

Investigation of Anomaly Stations:

To statistically investigate the trend of low ridership at high complaint frequency for anomaly stations, hypothesis testing was carried out.

But first, Z-scores for # days with complaints and average daily ridership were calculated for top complaint stations listed earlier. Calculating Z-scores is a method of standardizing values. They represent deviations from the mean value.

Formula: $Z = (Value - Mean) / Standard\ Deviation$

Station	# Days with Complaints	Z-score	Average Daily Ridership	Z-score	Ridership to # Complaints Ratio	Z-score Ratio
Times Sq-42 St / 42 St	31	1.54	122,757	1.31	3,959.90	0.38
14 St-Union Sq	27	0.73	61,151	-0.66	2,264.85	-1.07
Grand Central 42 St	22	-0.28	90,702	0.28	4,122.82	0.52
34 St - Penn Station	19	-0.89	93,136	0.36	4,901.89	1.19
74 Broadway - Jackson Heights Roosevelt Av	18	-1.10	41,746	-1.28	2,319.22	-1.02

Breaking Down Calculation of Z-scores:

- **Z-score for # Days with Complaints:**
  - Measures how each station's # Days with Complaints compares to the average across all top stations.
  - Mean = 23.4 days
  - Standard Deviation = 5.5 days

  Example:

  For **14 St-Union Sq**:

  ➢ # Days with complaints = 27
  ➢ Z-score = (27−23.4) / 5.5 = 0.73
  ➢ This implies that *the station had complaint days 0.73 standard deviation units above the mean value of 23.4 days*.

- **Z-score (Average Daily Ridership):**
  - Measures how each station's average daily ridership compares to the mean of average daily ridership across all top stations.
  - Mean = 81,898 riders
  - Standard Deviation = 31,489 riders

  Example:

  For **14 St-Union Sq**:

  ➢ Average daily ridership = 61,151
  ➢ Z-score = (61,151−81,898) / 31,489 = −0.66
  ➢ This implies that *the station had average daily ridership that is 0.66 standard deviation units below the mean value of 81,898 riders per day*.

3. Z-score (Ridership to # Complaints Ratio)

- Measures how much each station's *average daily ridership per complaint day* deviates from the mean value across all top stations.
- Mean = 3,513.74 riders per complaint day
- Standard Deviation = 1,170.86 riders per complaint day

Example:

For **14 St-Union Sq**:

➢ Ratio value = 2,264.85 riders per complaint day
➢ Z-score = (2,264.85−3,513.74) / 1,170.86 = −1.07

➢ This implies that t*he station had 1.07 standard deviation units fewer riders per complaint day than the mean value of 3,513.74 riders per complaint day*.

Hypothesis Testing:

A t-test is used to estimate the true difference between means of two groups of data using the ratio of the difference in group means over the pooled standard error. Here the independent variable is station group (anomaly or non-anomaly) and dependent variable is either # complaints days or average daily ridership.

Before a t-test, an *F-test for variance is used to determine the type of t-test to be used*. Here, the *null hypothesis* - a statement that there is no significant difference or effect is - Variances of the two groups (anomaly vs non-anomaly top stations) are equal.

F-test results:

- **For # days with complaints:**

  F-statistic: **1.0385**

  F-critical (one-tail): **18.5128**

  P-value (one-tail): **0.4154**

  P-value (two-tail): **0.8308**

  F < F-critical (one-tail) and p-values (one and two-tail) were > 0.05 significance level. So *null hypothesis cannot be rejected* and an equal variance independent t-test is needed.

- **For Average Daily Ridership:**

  F-statistic: **1.6915**

  F-critical (one-tail): **199.5**

  P-value (one-tail): **0.4777**

  P-value (two-tail): **0.9553**

  F < F-critical (one-tail) and p-values (one and two-tail) were > 0.05 significance level. So *null hypothesis cannot be rejected* and an equal variance independent t-test is needed.

Therefore, the F-tests determined that an equal variance independent t-test is needed for both cases.

Null hypothesis: Means of the two groups (anomaly vs non-anomaly top stations) are equal.

T-test results:

- **For # days with complaints:**

t-statistic: -0.2614
p-value (two-tail): 0.8107
p-value (one-tail): 0.4053

The t-value is almost zero barely having an effect on the mean difference. And since both p-values are > 0.05, the test *fails to reject the null hypothesis*. This means that there is **no significant difference between the mean # complaint days between anomaly and non-anomaly top stations. The anomaly here is likely due to random chance**.

- **For Average Daily Ridership**:

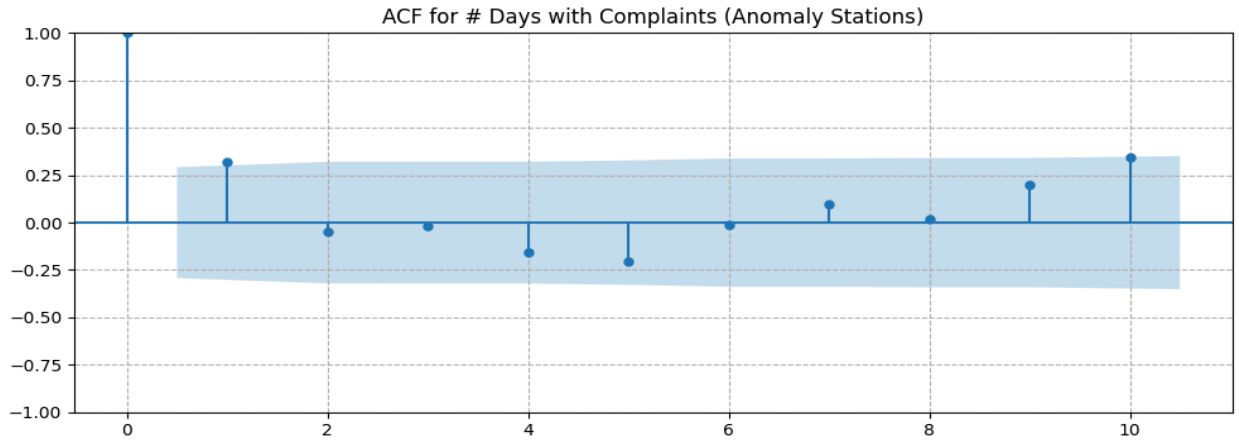
  t-statistic: -3.352
  p-value (two-tail): 0.044
  p-value (one-tail): 0.022

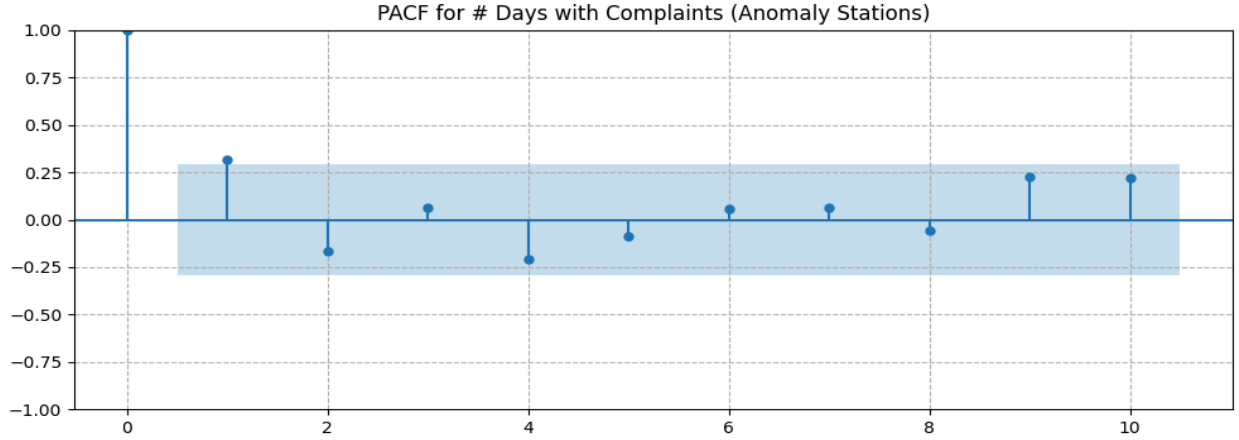
  The t-value deviates enough from 0 to have an effect on the mean difference. And with both p-values < 0.05, it implies that the test *rejects the null hypothesis*. This suggests that **there is a statistically significant difference in the means of average daily ridership between anomaly and non-anomaly top stations**.

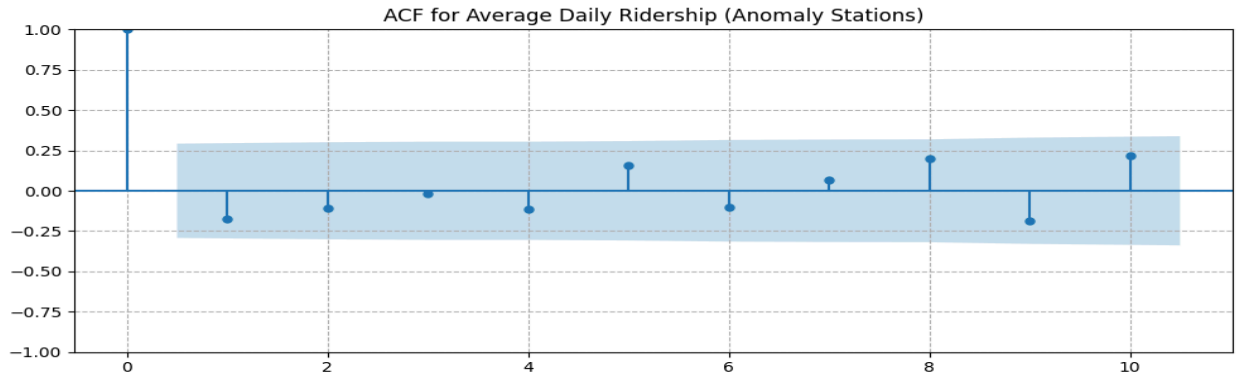
So, while complaint patterns are consistent across all top stations, average daily ridership at the two anomaly stations deviates notably from the norm.

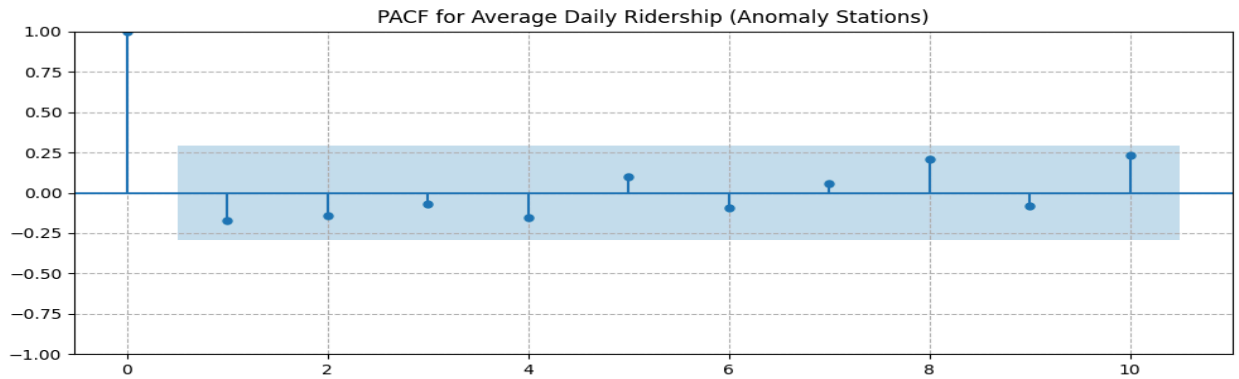
Given the nature of our data, **short-term predictions such as anomaly detection can be performed**. But to decide what kind of model to use for it, we have to look at auto-correlation and cross-correlation between # days with complaints and average daily ridership. If there is no significant cross-correlation between the two, we will use a static regression model. Otherwise, we will use a time-series/lagged regression model.

Interpretation of cross-correlation depends on the fact that there is zero to minimal autocorrelation (self-dependency on past values). The results of auto-correlation analysis are shown below:


ACF for # Days with Complaints (Anomaly Stations)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
2
4
6
8
10


PACF for # Days with Complaints (Anomaly Stations)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
2
4
6
8
10


ACF for Average Daily Ridership (Anomaly Stations)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
2
4
6
8
10


PACF for Average Daily Ridership (Anomaly Stations)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
2
4
6
8
10

- **For # Days with Complaints:**
    - ACF values drop off quickly, showing weak autocorrelation and suggesting that patterns for # complaint days are not strongly dependent on past days.
    - PACF values have minimal spikes showing no significant direct lagged effects and suggesting # complaint days are largely independent from day-to-day.
- **For Average Daily Ridership:**
    - ACF values also drop off quickly, suggesting that patterns for ridership are not strongly dependent on past days.
    - PACF values also have minimal spikes suggesting no strong influence from past ridership values.

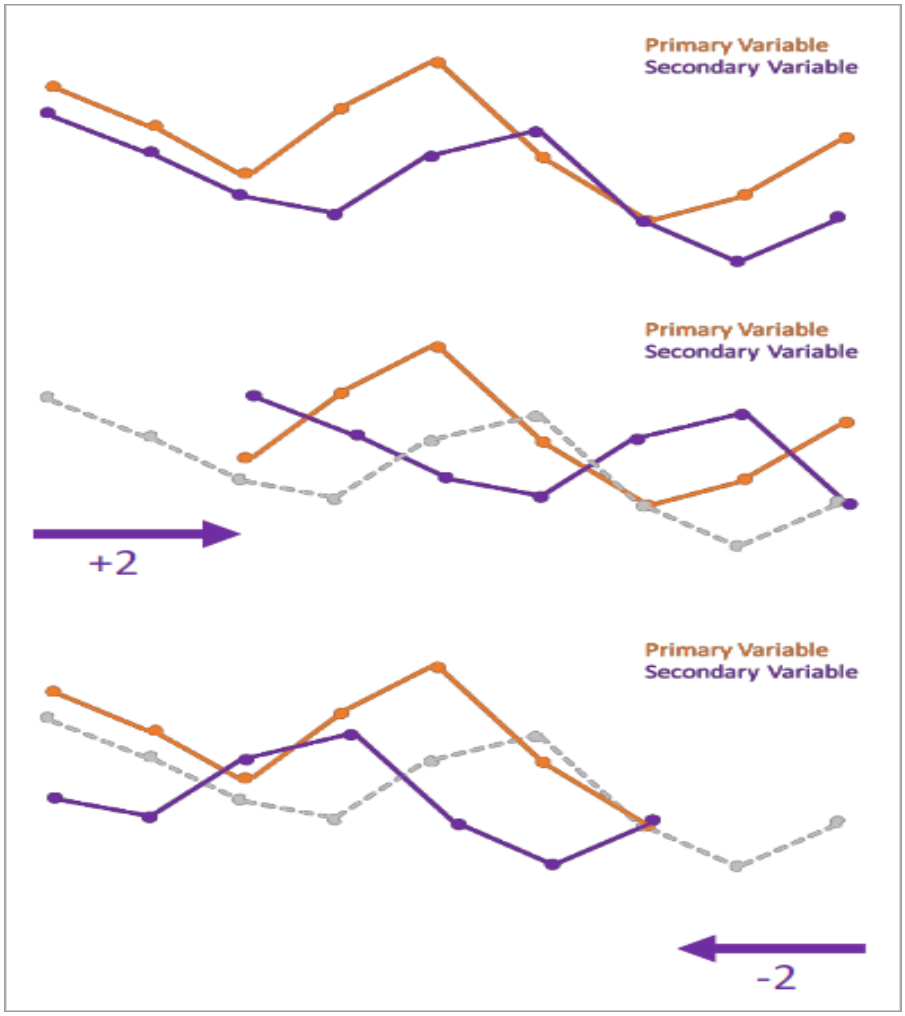
So no strong past-value dependency exists for both # days with complaints and average daily ridership at the anomaly stations. In other words*, there is no significant auto-correlation in both cases.*

We can now look at cross correlation safely. Cross correlation values measure the strength of the linear relationship between two time series: when one time series increases, does the other tend to increase, decrease, or not change?

Because there are often delayed effects between two time series (for example, the delay between an increase in the number of predators in an ecosystem and changes in the population of the prey), cross correlation values are always calculated with respect to a time lag. The time lag is a shift of the secondary variable relative to the first, and a new cross correlation value is calculated for the new corresponding pairs of values between the two time series.

Cross-correlation function is given by: $R_{xy}(k) \triangleq \sum_{n} x(n)y(n+k)$ where k is the lag.

- *If k is positive*, then changes in y(n) occur before changes in x(n) implying y(n) leads x(n) by k values.
- *If k is negative*, then changes in x(n) occur before changes in y(n) implying x(n) leads y(n) by k values.
- *If k = 0*, then x(n) and y(n) have no delayed effects on each other.



I wanted to focus on looking for a possible causal effect of # days with complaint days on average daily ridership for the anomaly stations.

Let x(n) be # complaint days and y(n) be average daily ridership. Lag values are measured in days here.

For a causal relationship to exist between them, the maximum correlation has to be closer to 1.0 at a particular negative lag.

Results after Checking for Cross-correlation at Different Lags:

Lag Range (Days)	Best Lag (Days)	Max Correlation
0	0	-0.15
-1 to +1	1	0.17

-2 to +2	2	0.18
-3 to +3	2	0.18

The correlation did not improve further at extended negative lags (-4 to -5).

The best correlation was observed at a positive lag of 2 which implies that changes in average daily ridership occurs 2 days before changes in # complaint days. Irrespective of its practicality, this is not my intended direction of focus. It also shows a very weak relationship ($R_{xy}(k)$ = 0.18).

This suggests that there is **no strong evidence for # days with complaints having any significant effect on the average daily ridership** at anomaly stations.

Anomaly Detection:

We can still develop a baseline model to flag anomaly values for # days with complaints and average daily ridership for those stations. This can help in pattern recognition to take proactive security measures.

Since our data is stable and shows no extreme outliers, we can use statistical thresholding to detect anomalies. Here, we consider # daily complaints instead, along with average daily ridership for thresholds which are calculated using the formula:

$$Threshold = Mean \pm (2 * Standard\ Deviation)$$

Those thresholds were chosen because in a normal distribution, ~95% of the data lies within that range. And even though our data does not have perfect normal distribution, this method still works well for data with stable patterns and no extreme outliers.

Note:

1. With only 2 official holidays in the time period considered, the data is not sufficient to establish reliable thresholds specific to holidays. So we ignore them for the flagging model.
2. The minimum value of the lower threshold for # daily complaints is zero since this count can never be negative.

Threshold and Anomaly Detection Results:

1. For 14 St-Union Sq:

- **# Daily Complaints**:
  - Mean: 1.08
  - Standard Deviation: 0.28
  - Upper Threshold: 1.63 complaints

    - Lower Threshold: 0.53 complaints

Out of 25 (non-holiday) days with complaints, two were flagged as anomalies for # daily complaints.

- **Average Daily Ridership**:
    - Mean: 59,356.80 riders/day
    - Standard Deviation: 12,115.97
    - Upper Threshold: 83,588.73 riders/day
    - Lower Threshold: 35,124.87 riders/day

Out of 25 (non-holiday) days with complaints, one was flagged as an anomaly for average daily ridership.

2. For 74 Broadway - Jackson Heights Roosevelt Av:

- **# Daily Complaints**:
    - Mean: 1.22
    - Standard Deviation: 0.55
    - Upper Threshold: 2.32 complaints
    - Lower Threshold: 0.13 complaints

Out of 18 (non-holiday) days with complaints, one was flagged as an anomaly for # daily complaints.

- **Average Daily Ridership**:
    - Mean: 41,799.17 riders/day
    - Standard Deviation: 6,819.74
    - Upper Threshold: 55,438.65 riders/day
    - Lower Threshold: 28,159.68 riders/day

Out of 18 (non-holiday) days with complaints, two were flagged as anomalies for average daily ridership.

These thresholds, calculated from current complaint and ridership data, can be used as a baseline for detecting anomalies in the future. **Any future data can be plugged into them and flagged accordingly**!

Conclusions:

- From January 1st to June 30th of 2024, any anomalies in # daily complaints at **14 St-Union Sq** and **74 Broadway - Jackson Heights Roosevelt Av** stations compared to other top complaint stations were likely due to random chance. However, **the anomalies in average daily ridership were not random**.

- There was **no significant effect of # daily complaints on the average daily ridership** at these stations, suggesting that deviations in ridership may have been caused by other factors.

- That said, since our data is stable and free from extreme outliers, I was able to apply statistical thresholding to develop a baseline anomaly detection model.

- **5.56% and 11.12% of days with complaints were flagged as anomalies in average daily ridership** at the two stations, respectively. This model allows for plugging in new data to identify similar anomalies.
