**Delays and Demographics: A Data-Driven Look at NYC Transit**

**Why this data:**

Commuting in New York City often comes with its share of challenges, and many residents can recall at least one frustrating experience navigating the city via the Metropolitan Transportation Authority (MTA). In an effort to improve transparency and public engagement, the MTA has recently released a wide range of datasets related to ridership, service delays, and major incidents. This project aims to perform a comprehensive statistical and time series analysis of this data to better understand the patterns of disruption—specifically, which demographics and subway lines are most affected by delays and major incidents.

**Datasets:**

**MTA Subway Hourly Ridership** - Source : <https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-2020-2024/wujg-7c2s/about_data>

This dataset provides subway, SIR, and Roosevelt Island Tramway ridership, compiled at the hourly level, by station complex and fare payment method, and consolidated fare class.

**Limitations:** Due to the automated collection method from the actual MTA data servers, there are virtually no limitations to the data collected. Additionally, because of the vast amount of data available, we will only be looking at 2023-2024 for this analysis.

**Ethical Considerations:** All of the data has been made publicly available by the MTA with no limitations on its usage. Additionally, there is no identifying information that could possibly infringe on anyone’s privacy.

**MTA Subway Trains Delayed** – Source: <https://data.ny.gov/Transportation/MTA-Subway-Trains-Delayed-2020-2024/wx2t-qtaz/about_data>

This dataset reflects the number and percentage of subway trains delayed per weekday and weekend based on scheduled service, broken down into specific delay categories. This data is calculated separately for weekdays and weekends.

**MTA Subway Major Incidents** – Source: <https://data.ny.gov/Transportation/MTA-Subway-Major-Incidents-2020-2024/j6d2-s8m2/about_data>

Major Incidents are unplanned incidents that delay 50 or more trains. Such events cause the most disruption to customers. For each month and subway line, the dataset includes the division of the route (A Division or B Division), the day type (weekday or weekend), the delay category, and the count of major incidents.

**Limitations:** The assigning of delays and major incidents, to incident causes, depends upon the subjective judgement of the dispatcher. Often, a single subway line can be affected by multiple incidents at the same time, making it difficult to neatly assign a specific delayed train to a single incident. Furthermore, incidents can sometimes have multiple causes, making it difficult to neatly categorize incidents, and the delays assigned to them, by cause. Thus, analysts looking at delay data should look at delay categories as the MTA’s “best guess” of what the delay cause was, nothing more. Additionally, because of the vast amount of data available, we will only be looking at 2023-2024 for this analysis.

**Ethical Considerations:** All of the data has been made publicly available by the MTA with no limitations on its usage. Additionally, there is no identifying information that could possibly infringe on anyone’s privacy.

**MTA Subway Customer Journey-Focused Metrics** – Source: <https://data.ny.gov/Transportation/MTA-Subway-Customer-Journey-Focused-Metrics-2020-2/4apg-4kt9/about_data>

The Metropolitan Transportation Authority (MTA) operates the largest transit network in North America, serving parts of New York and Connecticut. To better reflect the rider experience, the MTA introduced Customer Journey-Focused Metrics as part of the Subway Action Plan, moving beyond traditional operational metrics like on-time performance. These new metrics include Additional Platform Time (APT), Additional Train Time (ATT), and Additional Journey Time (AJT), which measure delays experienced by passengers compared to scheduled service. Journey Time captures the full duration of a trip, while Customer Journey Time Performance (CJTP) measures the percentage of trips completed within 5 minutes of the scheduled time. These metrics offer a more accurate view of how service impacts passengers’ daily commutes.

**Ethical Considerations:** All of the data has been made publicly available by the MTA with no limitations on its usage. Additionally, there is no identifying information that could possibly infringe on anyone’s privacy.

**Limitations:** The data is broken down monthly at the line level. APT, ATT, AJT, CJTP, and Journey Time are measured on weekdays between 6 AM and 11 PM, with the peak period being between 7 AM and 10 AM and 4 PM and 7 PM. Off-peak hours are the times of day outside of those hours. These metrics are sensitive to changes in service scheduled in the timetable (service frequency and run times) since the service scheduled is an input in their calculation, and since APT, ATT, and AJT measure the additional time riders spend waiting for a train, onboard a train, or both, relative to the timetable, and CJTP measures the percentage of riders with AJT less than 5 minutes.

**U.S Decennial Census 2020-** Source : <https://popfactfinder.planning.nyc.gov/explorer/cities/NYC?compareTo=1>

This data set provides demographic data for NYC broken down on a borough level. These demographics include total population, population for different age groups, old age / child dependency ratios, and race/ethnicity numbers.

**Limitations:** While the Census is known for being extremely thorough and providing accurate counts for demographics. This data will be slightly outdated for this project since the data is from 2020 while our other data sets cover the years 2023-2024. But this is the most accurate demographic data available for NYC.

**Ethical Considerations:**

All of the data has been made publicly available by the US Census Bureau and NYC Population Fact Finder with no limitations on its usage. Additionally, there is no identifying information that could possibly infringe on anyone’s privacy.

**Labor Statistics for the New York City Region-** Source: <https://dol.ny.gov/labor-statistics-new-york-city-region>

This is provided by the NYS Department of Labor and provides Labor Force numbers for NYC broken down by borough and month/year. This includes total labor force population, how many from the force are employed/unemployed, and unemployment rate.

**Ethical Considerations:**

All of the data has been made publicly available by NYS Department of Labor with no limitations on its usage. Additionally, there is no identifying information that could possibly infringe on anyone’s privacy.

**ACS 1 Year Demographic Data for NYC 2023-** Source: <https://www.nyc.gov/content/planning/pages/resources/datasets/american-community-survey>

The American Community Survey (ACS) is the most comprehensive nationwide survey available. From its annual releases, users can examine the city’s detailed demographic, social, economic, and housing characteristics.

**Ethical Considerations:**

All of the data has been made publicly available by the ACS and the city of New York with no limitations on its usage. Additionally, there is no identifying information that could possibly infringe on anyone’s privacy.

**Limitations:**

It is also important to note that ACS estimates are derived from a survey and are subject to sampling error. Therefore, all ACS estimates are published with Margins of Error (MOEs) at a 90% confidence interval.

**Data Cleaning Summary:**

Thankfully due to the efforts of the MTA, US Census Bureau, and the NYS Department of Labor none of the data sets contained any missing or duplicate values. However some wrangling will be necessary to compile all datasets into one for exploration.

**Data Wrangling:**

New columns were created in all datasets that converted the time related columns to datetime format in python.

* **MTA Subway Hourly Ridership –** 
  + Transit\_timestamp – converted to datetime, then separated into columns date/time/month for the sake of merging with other datasets
  + Using a pivot table – Each fare-class category turned into its own column giving the total count for that fare-class aggregated for a given month and borough
    - List of new columns -
      * MetroCard – Fair Fare •
      * MetroCard – Full Fare •
      * MetroCard – Other •
      * MetroCard – Senior & Disability •
      * MetroCard – Students •
      * MetroCard – Unlimited 30-Day •
      * MetroCard – Unlimited 7-Day •
      * OMNY – Full Fare •
      * OMNY – Other •
      * OMNY – Seniors & Disabilities •
* **Labor Statistics for the New York City Region**
  + In excel a new column MonthDT was created by merging the Year and Month columns –
    - This new column MonthDT was reformatted in python to match ridership month column
  + Labor and Ridership were then merged into one dataset on the respective month/monthDT and borough columns.
* **ACS 1 Year Demographic Data for NYC 2023**
  + Dropped Columns
    - All margin of error, percent, and percent margin of error columns were dropped as they will not be needed for this project
  + Dropped Rows
    - **Sex ratio (males per 100 females),** **18 years and over, Male, Female, Sex ratio (males per 100 females), 65 years and over, Male, Female, Sex ratio (males per 100 females) One Race , Two or more races, Aztec, Blackfeet Tribe of the Blackfeet Indian Reservation of Montana, Maya, Native Village of Barrow Inupiat Traditional Govt., Navajo Nation, Nome Eskimo Community, Other American Indian and Alaska Native, Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian, Chamorro, Native Hawaian, Samoan, Other Native Hawaiian and Other Pacific Islander** **White and Black or African American, White and American Indian and Alaska Native, White and Asian, White and Some Other Race, Black or African American and American Indian and Alaska Native, Black or African American and Some Other Race, Race alone or in combination with one or more other races, Total population, White, Black or African American, American Indian and Alaska Native, Asian, Native Hawaiian and Other Pacific Islander, Some Other Race, Mexican, Puerto Rican, Cuban, Other Hispanic or Latino, Not Hispanic or Latino, Two or more races**

**Data Profile:**

**MTA Subway Hourly Ridership**

|  |  |  |
| --- | --- | --- |
| **Data Label** | **Data Type** | **Data Description** |
| transit\_timestamp | DATE  object | Timestamp payment took place in local time. All transactions here are rounded down to the nearest hour. For example, a swipe that took place at 1:37pm will be reported as having taken place at 1pm. |
| transit\_mode | TEXT  object | Distinguishes between the subway, Staten Island Railway, and the Roosevelt Island Tram |
| station\_complex\_id | ALPHANUMERIC  object | A unique identifier for station complexes |
| station\_complex | TEXT  object | The subway complex where an entry swipe or tap took place. Large subway complexes, such as Times Square and Fulton Center, may contain multiple subway lines. The subway complex name includes the routes that stop at the complex in parenthesis, such as Zerega Av (6). |
| borough | TEXT  object | Represents one of the boroughs of New York City serviced by the subway system (Bronx, Brooklyn, Manhattan, Queens). |
| payment\_method | TEXT  object | Specifies whether the payment method used to enter was from OMNY or MetroCard. |
| fare\_class\_category | TEXT  object | The class of fare payment used for the trip. The consolidated categories are: • MetroCard – Fair Fare • MetroCard – Full Fare • MetroCard – Other • MetroCard – Senior & Disability • MetroCard – Students • MetroCard – Unlimited 30-Day • MetroCard – Unlimited 7-Day • OMNY – Full Fare • OMNY – Other • OMNY – Seniors & Disabilities |
| ridership | NUMERIC  int64 | Total number of riders that entered a subway complex via OMNY or MetroCard at the specific hour and for that specific fare type |
| transfers | NUMERIC  int64 | Number of individuals who entered a subway complex via a free bus-to-subway, or free out-of-network transfer. This represents a subset of total ridership, meaning that these transfers are already included in the preceding ridership column. Transfers that take place within a subway complex (e.g., individuals transferring from the 2 to the 4 train within Atlantic Avenue) are not captured here. |
| latitude | DECIMAL  float64 | Latitude for the specified subway complex |
| longitude | DECIMAL  float 64 | Longitude for the specified subway complex |
| georeference | ALPHANUMBERIC  object | Longitude and Latitude for the specified subway complex |
| datetime\_timestamp | DATE  datetime64 | A copy of the timestamp column saved in datetime format |
| date | DATE  datetime64 | Column containing just the date in datetime format |
| time | DATE  datetime64 | Column containing just the time in datetime format |

**MTA Subway Trains Delayed**

|  |  |  |
| --- | --- | --- |
| **Data Label** | **Data Type** | **Data Description** |
| month | DATE  object | The month in which subway trains delayed is being calculated (yyyy-mm-dd). |
| division | TEXT  object | The A Division (numbered subway lines and S 42nd) and B Division (lettered subway lines). |
| line | TEXT  object | Each subway line (1, 2, 3, 4, 5, 6, 7, A, C, E, B, D, F, M, G, JZ, L, N, Q, R, W, S 42nd, S Rock, S Fkln). |
| day\_type | NUMERIC  int64 | Represents weekday as 1 and weekend as 2. |
| reporting\_category | TEXT  object | The six categories that delays are reported under are Crew Availability, External Factors, Infrastructure & Equipment, Operating Conditions, Planned ROW Work, and Police & Medical). |
| subcategory | TEXT  object | The subcategories are Crew Availability, External Debris on Roadbed, Inclement Weather, External Agency or Utility, Propulsion, Door-Related, Other - CE, Braking, Service Delivery, Signal Modernization Capital Project, Rail and Roadbed, Other - Sig, Other Infrastructure, Train Brake Activation - Cause Unknown, "Fire, Smoke, Debris", Other Internal Disruptions, Service Management, Other Operating Environment, Work Equipment, Capital Work - Other Planned ROW, Subway Maintenance, Insufficient Supplement Schedule, Other Planned ROW Work, Persons on Roadbed, "Public Conduct, Crime, Police Response", and Sick/Injured Customer |
| delays | NUMERIC  int64 | The number of train delays per category and time period. |
| datetime\_month | DATE  datetime64 | Copy of the month column but saved in datetime format |

**MTA Subway Major Incidents**

|  |  |  |
| --- | --- | --- |
| **Data Label** | **Data Type** | **Data Description** |
| month | DATE  object | The month in which the Major Incidents are being calculated (yyyy-mm-dd). |
| division | TEXT  object | The A Division (numbered subway lines) and B Division (lettered subway lines). |
| line | TEXT  object | Each subway line (1, 2, 3, 4, 5, 6, 7, A, C, E, B, D, F, M, G, JZ, L, N, Q, R, W, S 42nd, S Rock, S Fkln). |
| day\_type | NUMERIC  int64 | Represents weekday as 1, weekend as 2. |
| category | TEXT  object | The six categories that fall under the definition of a Major Incident: Track, Signals, Persons on Trackbed/Police/Medical, Stations and Structure, Subway Car, and Other. |
| count | NUMERIC  int64 | The number of major incidents that occurred per month and per subway line. |
| datetime\_month | DATE  datetime64 | Copy of the month column but saved in datetime format |

|  |  |  |
| --- | --- | --- |
| **Data Label** | **Data Type** | **Data Description** |
| month | DATE | The month in which the metrics are being calculated (yyyy-mm-dd). |
| division | TEXT | The A Division (numbered subway lines and S 42nd) and B Division (lettered subway lines). |
| line | TEXT | Each subway line (1, 2, 3, 4, 5, 6, 7, A, C, E, B, D, F, M, G, JZ, L, N, Q, R, W, S 42nd, S Rock, S Fkln). |
| period | TEXT | Represents both the peak and off-peak service periods. |
| num\_passengers | NUMERIC | Total number of estimated passengers reported each month and on each line. |
| additional platform time | NUMERIC | The average estimated additional time in minutes (above scheduled time) customers wait for their train, reported each month and on each line. |
| additional train time | NUMERIC | The average estimated additional time in minutes (above scheduled time) customers spend onboard a train, reported each month and on each line. |
| total\_apt | NUMERIC | The total number of estimated additional time in minutes (above scheduled time) customers wait for their train, reported each month and on each line. |
| total\_att | NUMERIC | The total number of average additional time in minutes (above scheduled time) customers spend onboard a train, reported each month and on each line |
| over\_five\_mins | NUMERIC | The estimated total number of customers whose journeys are not completed within 5 minutes of the scheduled time, reported each month and on each line. |
| over\_five\_mins\_perc | PERCENT | The estimated percentage of customers whose journeys are not completed within 5 minutes of the scheduled time, reported each month and on each line. |
| customer journey time performance | PERCENT | The estimated percentage of customers whose journeys are completed within 5 minutes of the scheduled time, reported each month and on each line. |

**Labor Statistics for the New York City Region**

|  |  |  |
| --- | --- | --- |
| **Data Label** | **Data Type** | **Data Description** |
| Borough | TEXT  object | Represents one of the boroughs of New York City (Bronx, Brooklyn, Manhattan, Queens, Staten Island). |
| Year | NUMERIC  int64 | The year in which the labor statistics were calculated |
| Month | NUMERIC  object | The month in which the labor statistics were calculated |
| MonthDT | NUMERIC  int64 | The month and year in which the labor statistics were calculated |
| Labor Force | NUMERIC  object | The total size of the labor force |
| Employed | NUMERIC  object | The number of people from the labor force that are employed |
| Unemployed | NUMERIC  object | The number of people from the labor force that are unemployed |
| Unemployment Rate | NUMERIC  float64 | The percentage of people in the labor force that are unemployed |

**US Census Bureau Demographics**

|  |  |  |
| --- | --- | --- |
| **Data Label** | **Data Type** | **Data Description** |
| Year | NUMERIC  int64 | The year in which the demographics were calculated |
| |  | | --- | | Borough | | TEXT  object | Represents one of the boroughs of New York City (Bronx, Brooklyn, Manhattan, Queens, Staten Island). |
| Total Population | NUMERIC  int64 | Total Population |
| Male | NUMERIC  int64 | The number of males in the population |
| Female | NUMERIC  int64 | The number of females in the population |
| Under 5 years | NUMERIC  int64 | The number of people under the age of 5 in the population |
| 5 to 9 years | NUMERIC  int64 | The number of people aged 5 to 9 years |
| 10 to 14 years | NUMERIC  int64 | The number of people aged 10 to 14 years |
| 15 to 19 years | NUMERIC  int64 | The number of people aged 15 to 19 years |
| 20 to 24 years | NUMERIC  int64 | The number of people aged 20 to 24 years |
| 25 to 29 years | NUMERIC  int64 | The number of people aged 25 to 29 years |
| 30 to 34 years | NUMERIC  int64 | The number of people aged 30 to 34 years |
| 35 to 39 years | NUMERIC  int64 | The number of people aged 35 to 39 years |
| 40 to 44 years | NUMERIC  int64 | The number of people aged 40 to 44 years |
| 45 to 49 years | NUMERIC  int64 | The number of people aged 45 to 49 years |
| 50 to 54 years | NUMERIC  int64 | The number of people aged 50 to 54 years |
| 55 to 59 years | NUMERIC  int64 | The number of people aged 55 to 59 years |
| 60 to 64 years | NUMERIC  int64 | The number of people aged 60 to 64 years |
| 65 to 69 years | NUMERIC  int64 | The number of people aged 65 to 69 years |
| 70 to 74 years | NUMERIC  int64 | The number of people aged 70 to 74 years |
| 75 to 79 years | NUMERIC  int64 | The number of people aged 75 to 79 years |
| 80 to 84 years | NUMERIC  int64 | The number of people aged 80 to 84 years |
| 85 years and over | NUMERIC  int64 | The number of people aged 85 years or over |
| Median age (years) | NUMERIC  float64 | The median age in years for the given borough |
| Under 18 years | NUMERIC  int64 | The number of people aged under 18 years |
| 65 years and over | NUMERIC  int64 | The number of people aged 65 years or over |
| Age dependency ratio | NUMERIC  float64 | Derived by dividing the combined under 18 and over 65 populations by the 18-64 population and multiplied by 100 |
| Old-age dependency ratio | NUMERIC  float64 | Derived by dividing the over 65 populations by the 18-64 population and multiplied by 100 |
| Child dependency ratio | NUMERIC  float64 | Derived by dividing the combined under 18 population by the 18-64 population and multiplied by 100 |
| Hispanic | NUMERIC  int64 | The number of people who identify as Hispanic |
| White non-Hispanic | NUMERIC  int64 | The number of people who identify as White non-hispanic |
| Black non-Hispanic | NUMERIC  int64 | The number of people who identify as Black non-hispanic |
| Asian non-Hispanic | NUMERIC  int64 | The number of people who identify as Asian non-hispanic |
| Some other race non-Hispanic | NUMERIC  int64 | The number of people who identify as some other race non-hispanic |
| Non-Hispanic of two or more races | NUMERIC  int64 | The number of people who identify as non-hispanic of two or more races |

**Defining Questions:**

**1. Unemployment Impact**

**Hypothesis**: Boroughs with higher unemployment rates will show lower subway ridership.

**2. Labor Force Size and Ridership**

**Hypothesis**: Boroughs with larger labor force populations will exhibit higher average subway ridership.

**3. Age Group Influence**

**Hypothesis**: Boroughs with a higher percentage of 18–34-year-olds will have higher subway ridership. Specifically with student issued metrocards

**4. Race and Transit Dependency**

**Hypothesis**: Boroughs with higher percentages of non-white residents have higher subway ridership, reflecting greater public transit dependency.

**5. Hourly Patterns by Borough**

**Hypothesis**: Boroughs with higher youth populations will show later ridership peaks compared to boroughs with older populations.

**6. Geographic Correlation Between Unemployment and Ridership**

**Hypothesis**: Boroughs with persistently high unemployment rates will show more **flattened ridership patterns** across the day (less of a sharp morning/evening commute peak), reflecting fewer regular commuters.

**7. Spatial Clustering of High-Ridership Stations**

**Question**: Are high-ridership stations spatially clustered within specific boroughs, and how does that relate to population density or employment hubs?

**8. Spatial Distribution of Transit Dependency**

**Hypothesis**: Boroughs located farther from central business districts (like Manhattan) show higher **per capita ridership**, indicating greater **transit dependency**