NYC Subway: Interlined, or Deinterlined? (COMP3125 Individual Project)

*Note: Do not used sub-title

Matthew Santorsa
dept. name of organization

Abstract*—This electronic document is a "live" template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. **CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract. (*provide a short abstract*)

Keywords—Public Transit, New York City, MTA (provide 3-5 keywords)

I. Introduction (*Heading 1*)

The New York City subway is an interesting beast, created back in the early 20th century to transport commuters throughout the boroughs, to help clear congestion within the city. Several lines were created, primarily running on shared tracks through downtown Manhattan. This is commonly known as interlining, and it has its problems, such as delaying trains if one breaks down within a station, or if the track along the line fails. A single delay on one train line can fan out to delay all the other lines. This brings up the question, how does interlining affect the NYC subway? Does it significantly change the average margins of time that it takes for a train to take its route, end to end? Was interlining the right way for the Metropolitan Transportation Authority (MTA) to go?

II. Datasets

A. *Source of dataset*

My dataset originates from the MTA itself, on the NY State Govt website. This makes it credible, as it is maintained by the state government, and the MTA itself. They are produced from statistics generated from subway metrics, and originally were produced at the start of 2019, as that is the timeframe of available data from when I began this project.

B. *Dataset Characteristics*

The dataset's format is a CSV file, and it contains the following columns: Month, Scheduled Day Type, Time Period, Line, Direction, Stop Path ID, Average Actual Runtime, 25th Percentile Runtime, 50th Percentile Runtime, 75th Percentile Runtime, Actual Trains, Distance, Average Speed, Average Scheduled Runtime, Scheduled Trains, Origin Station ID, Destination Station ID, Origin Station Name, Destination Station Name, and Number of Stops. When cleaning the Data, I removed all of the columns except for Month, Line, the percentile columns, average runtime, and the destination and origin station names. After I separated the lines, I ended up dropping the destination and origin station names as well, to help clean up the data further. Finally, I created a datetime column, and a time_numeric column to allow for the proper handling of yearly trends over the 6 years since the dataset started. For the dataset size, it is roughly 112 thousand rows of data, pre cleaning.

III. Methodology

In this part, you should give an introduction of the methods/model. First, what's the method/model. What's the assumption of this method/model. What's the advantage/disadvantage of this method/model. Why did you choose it. What Python module or function do you apply to apply this method/model. Any optional input/extra work did you adjust to make the results better. If you have multiple methods, feel free to use subsection A., B. to separate them.

A. *Average Lines*

To find the fastest line, I utilized the average runtime column as my main data type. This is because it is an aggregate of the entire months' data, and is not a quartile like the other 3 available runtimes. The only downside is that I am taking an average of an average, which may be less accurate than if I had the original runtimes, but the MTA does not provide this data to the public. Utilizing Pandas, I was able to complete my data cleaning, and then eventually arrive at my results.

B. *Linear Regression*

For my graphs, I utilized matplotlib, as well as sklearn linear model for my linear regression, on top of seaborn to make my graphs look more readable. Of course, Pandas was still in use for the datasets, as it is the backbone of my project.

C. *A Note about Methodology*

It should be noted that for my model I was utilizing two dataframes for non interlined, shuttle and no shuttle. This is because if I left the shuttle lines in, the fastest line would always be the shuttles, due to their short length. Therefore in the end, shuttles end up getting eliminated from the running, except for their own special category as fastest of the shuttles.

IV. Results

In this section, present your findings using an appropriate method, such as equations, numerical summaries, or visualizations like charts and graphs. Clearly explain all results and provide guidance on how to interpret them. If any unexpected results arise, discuss possible reasons or contributing factors. To improve clarity and organization, consider using subsections (e.g., A, B) to separate different aspects of your results

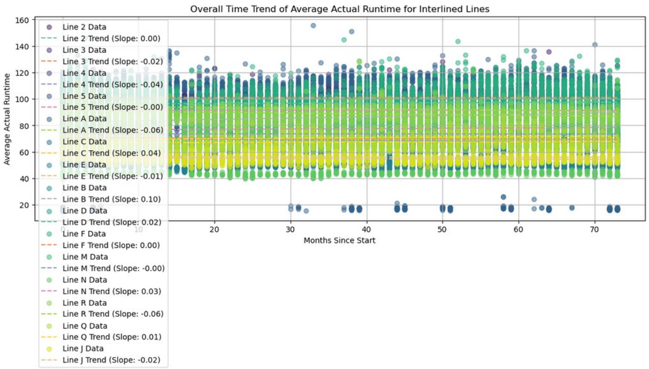
For my results, there are 2 main categories, with 1 subcategory for the deinterlined category. This is shuttle and non shuttle, with the other primary category being interlined lines. These have no shuttles, as shuttles exclusively run along their own tracks.

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

A. *Fastest Lines*

1) Interlined

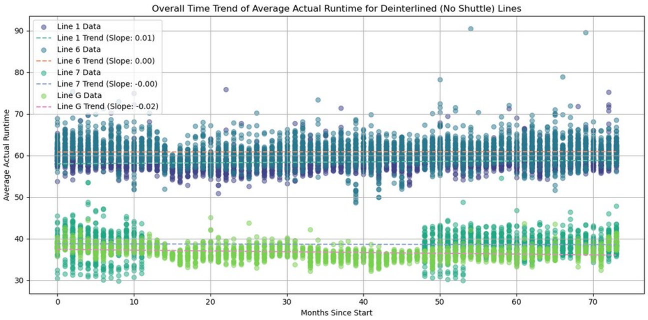
From running my code, I can determine that the fastest interlined line is the J, with an average runtime of ~55 minutes. This can be seen in the chart below:



As such, we can determine that the fastest line is the J, among all the available interlined lines.

2) Deinterlined

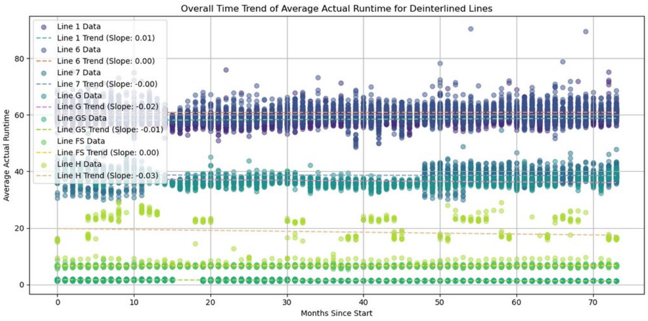
Looking back at the code results, I can determine that the fastest deinterlined line is the G, with an average runtime of ~37 minutes. This is reflected in the following chart:



As such, the fastest line can be determined to be the G, which is fairly consistent as well, as seen in the chart.

3) Shuttles

Finally, the shuttle lines. Once again looking at the results my code provided, I can determine that the 42nd St Shuttle (the GS in the code) is the fastest shuttle line. This is most likely due to it being only 2 stops, causing it to be extremely short. This is seen in the chart below, where it is compared to the other interlined lines as well, proving that it is the fastest shuttle line on the NYC subway, at roughly 2 minutes total runtime.



B. *Interlined vs Deinterlined Speeds*

Looking at the graphs from before, we can clearly tell that the interlined lines take a significantly longer time to reach their designated termini, with the linear regression lines depicting this the best. This of course is to be expected, as running more trains through a single line will decrease the amount of overall trains that each branch can handle.

C. *Will the fastest lines remain the fastest?*

1) Interlined

Looking back to the chart from before, we can determine that the J will remain the fastest line, as its regression line is still the lowest of all of them, and is trending down. It is expected to drop, according to the regression, to a ~53 minute runtime, as well.

2) Deinterlined

According to the chart, and the results from the code, the G will remain the fastest line. This will also result in a 1 minute overall decrease in end to end time according to the regression line, reuslting in a ~36 minute travel time.

3) Shuttles

Finally, according to the chart with shuttles, we can tell that the 42nd street shuttle is to remain the fastest, losing about 30 seconds from its average runtime.

V. Discussion

My project's primary shortfall is the inability to properly compare the interlined sections vs their branches, and rather only exclusively compares interlined and deinterlined lines. This is due to time primarily, but also due to the lack of proper data from the MTA. Due to everything being averages, it makes it harder to work with the data, and there is no way to obtain data between certain stations, only between designated termini stations and certain stations used for early turnbacks, where trains terminate early to run in the other direction for schedule changes, maintenance, or due to rush hour scheduling. Without the data for the stations where there aren't turnbacks, I am unable to properly determine how much of a slowdown each indivdual interline is causing. The only real fix to this is to petition the MTA to release the data publicly, or dig through the currently released data for a way to calculate these values.

VI. Conclusion

In this part, you should summarize your project. What important results did you find for your topic and what's the effect of this result on the real-world?

Overall, I found that deinterlined lines end up being faster than interlined ones, which shows that the MTA should attempt to work towards deinterlining more of their lines over time, as it would help to improve subway speeds across the system, and as it stands, line speeds will remain roughly the same without any changes to the system. Deinterlining seems to be the key to improving the New York City subway, and therefore it should be accomplished in the coming decades.

Acknowledgment *(Heading 5)*

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

Use the IEEE format for the citation. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence:

"Reference [3] was the first ..." Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

[1] [1] M. T. Authority, "MTA subway end-to-end running times: Beginning 2019: State of New York," MTA Subway End-to-End Running Times: Beginning 2019 | State of New York, https://data.ny.gov/Transportation/MTA-Subway-End-to-End-Running-Times-Beginning-2019/sp9g-mzjh/about_data (accessed Apr. 9, 2025).