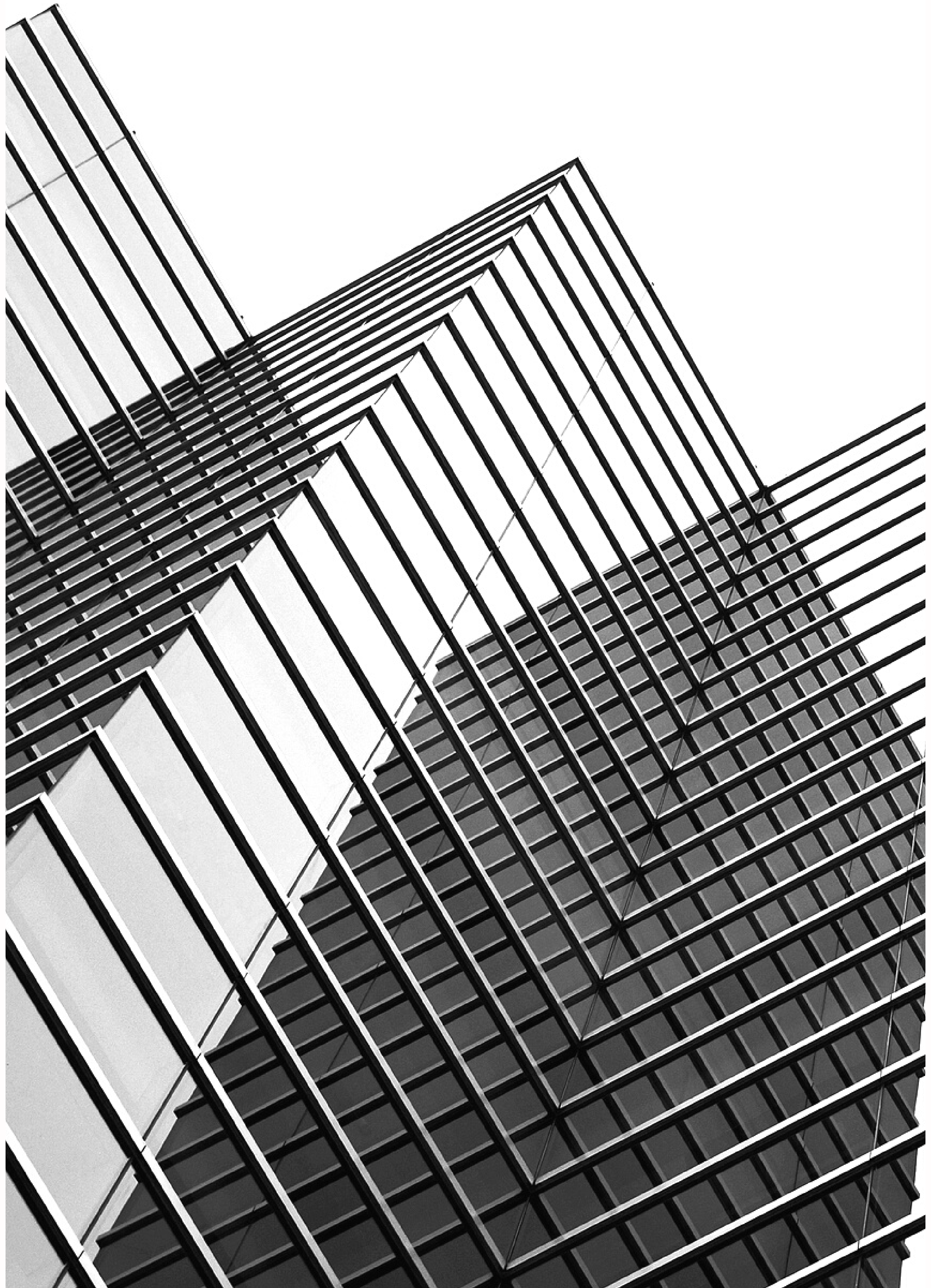
FORECASTING SUBWAY USE POST COVID

New York City transit use dropped by 90 percent in March of 2020.

Extreme changes make it hard to plan resource and budgeting

How can we better plan for these expected changes?

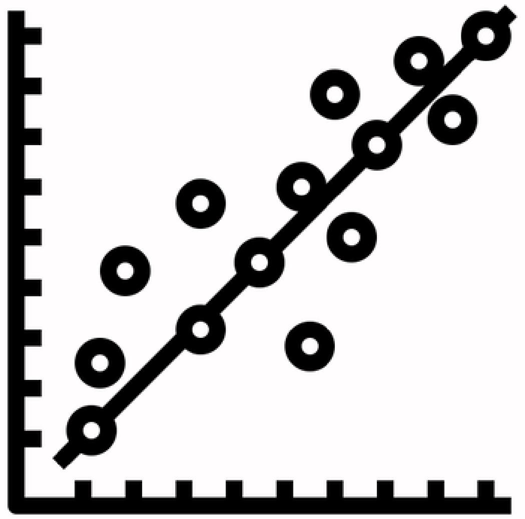
PRESENTED BY PATRICK BROWN



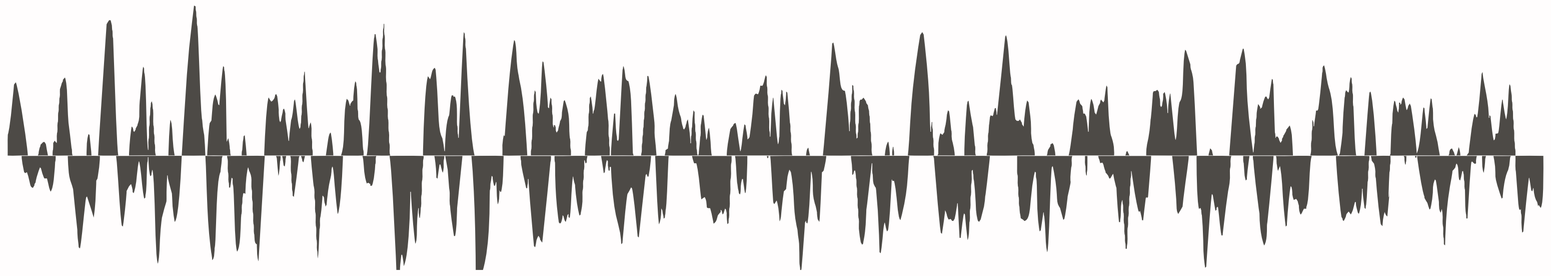
METHODS

Algorithms

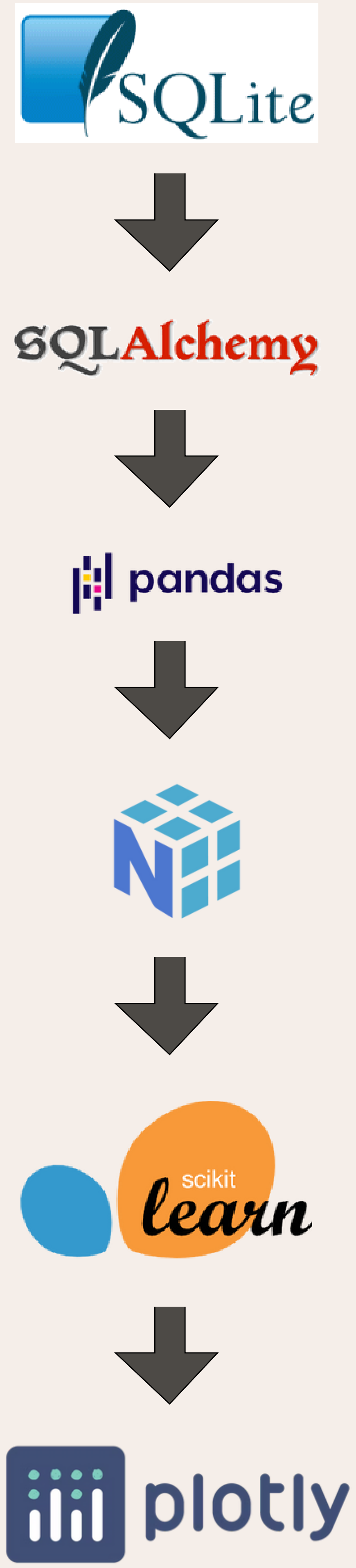
- Linear Regression
  - ○ Ordinary Least Squares



- Fast Fourier Transform



TOOLS



SQLite

SQLAlchemy

pandas

scikit learn

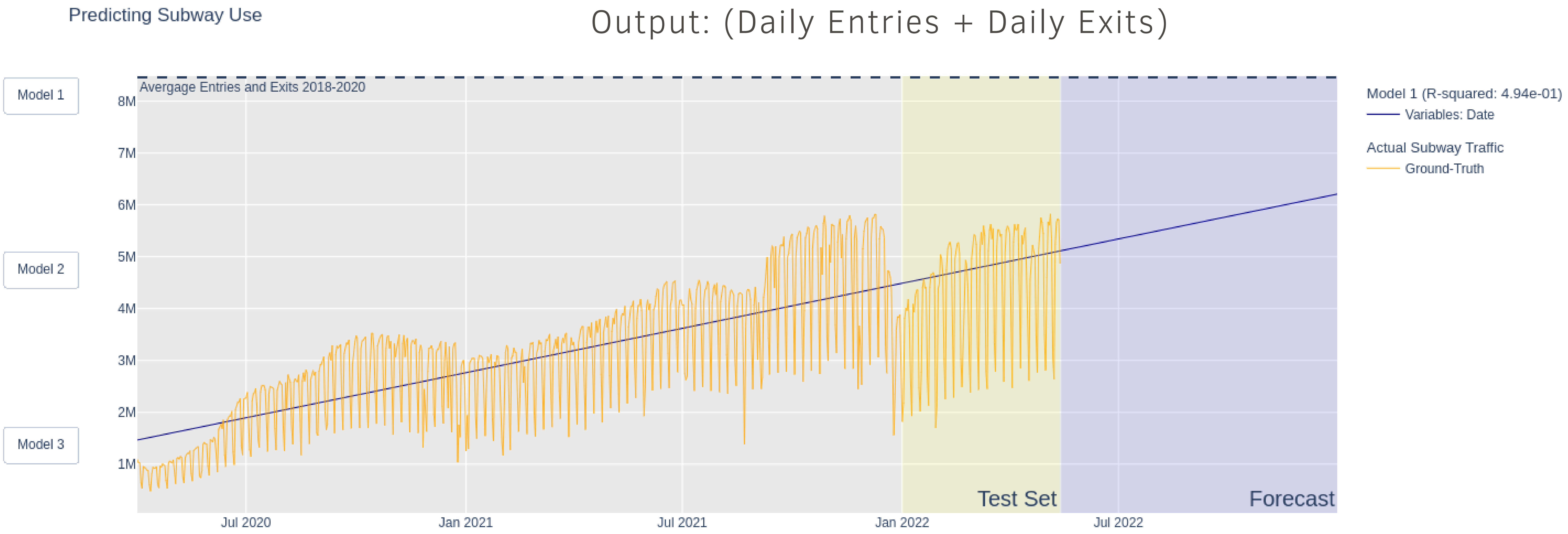
plotly

BUILDING OUR MODEL

STEP 1

Input: Days Post SaH Order

Output: (Daily Entries + Daily Exits)

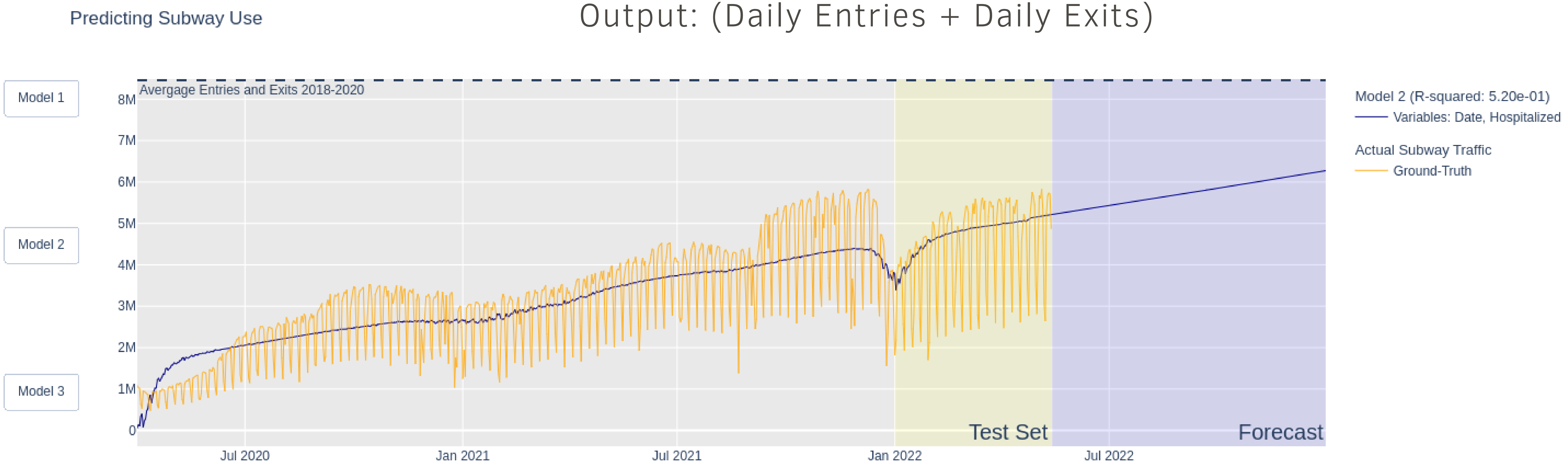


BUILDING OUR MODEL

STEP 2

Input: Days Post SaH Order, COVID Hospitalizations

Output: (Daily Entries + Daily Exits)

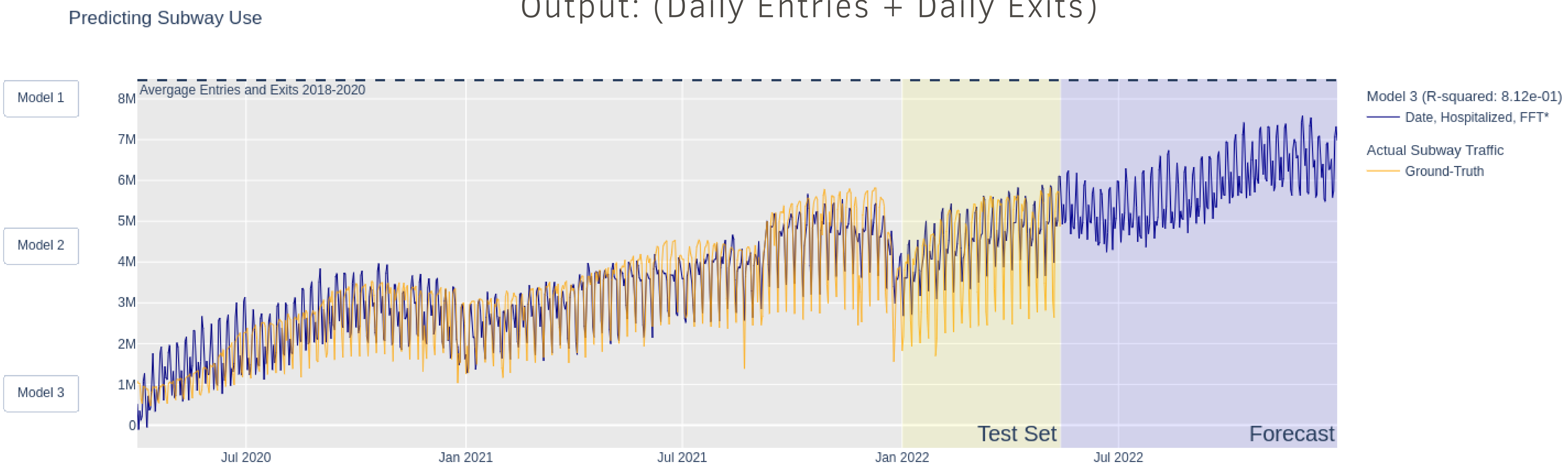


BUILDING OUR MODEL

STEP 3

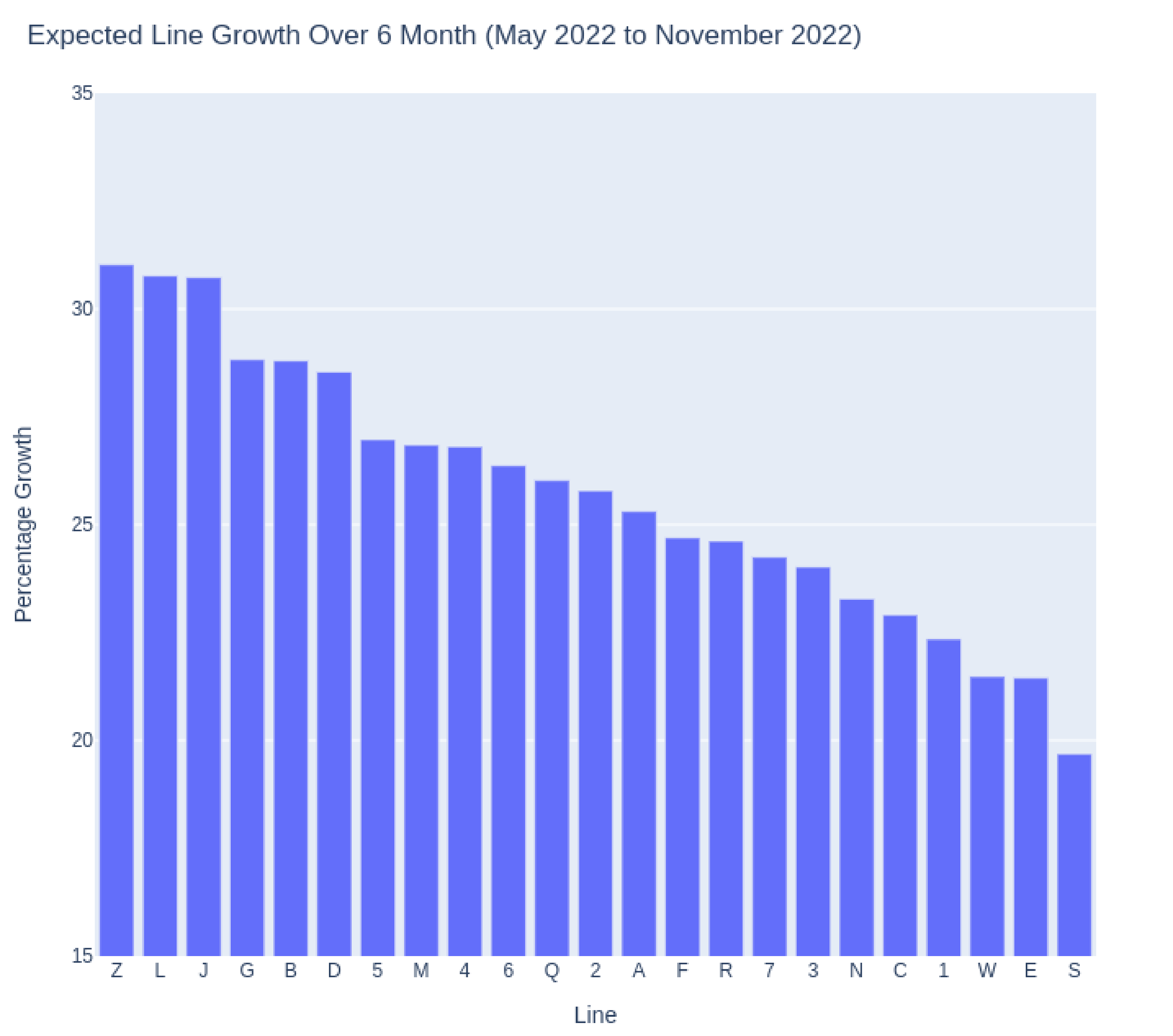
Input: Days Post SaH Order, COVID Hospitalizations, FFT

Output: (Daily Entries + Daily Exits)



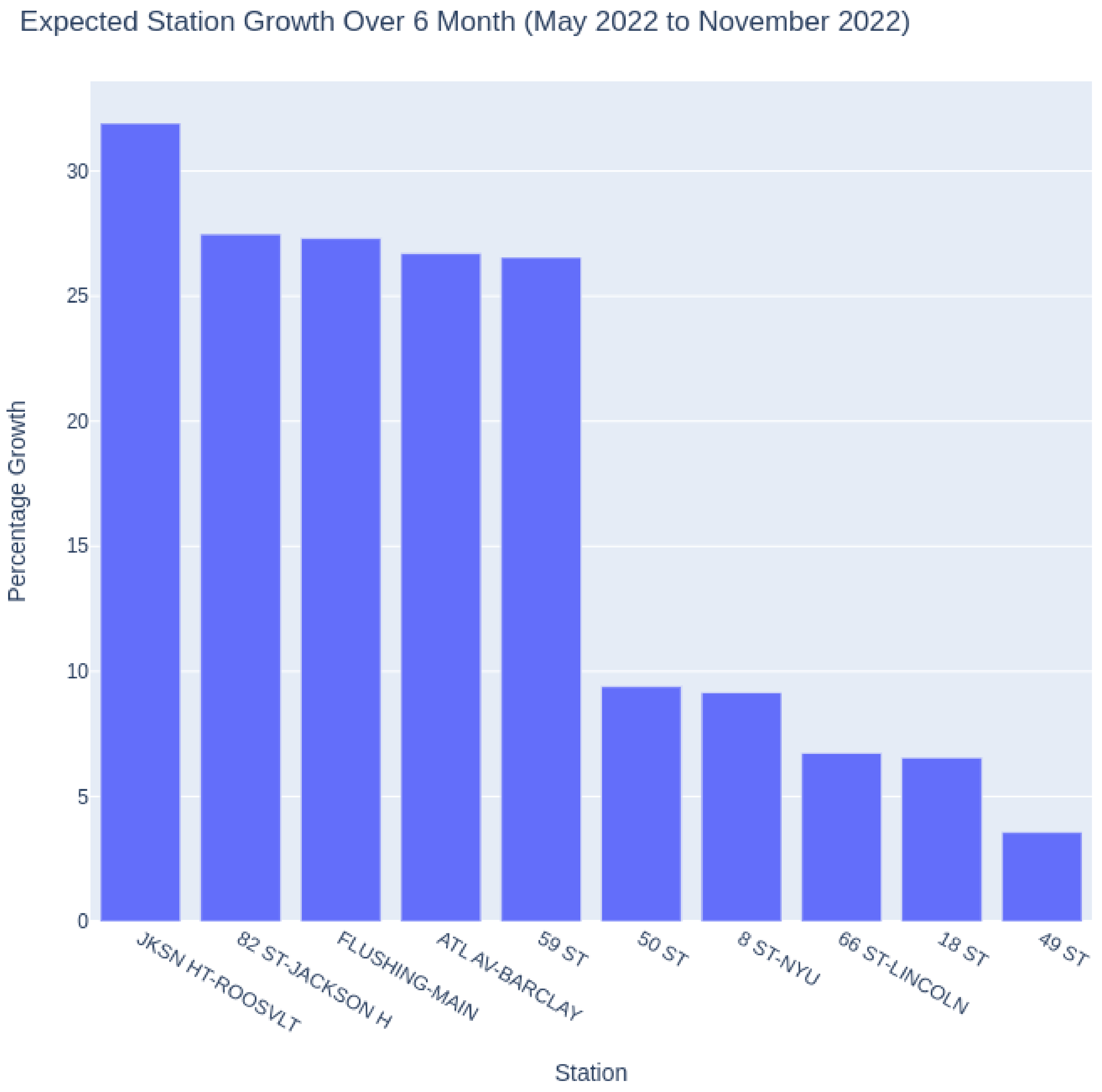
FORECASTING LINE GROWTH

- Nearly all lines are expected to grow 20%-30% over the next 6 months
- W, E, S Lines show the smallest growth (<22%)
- Z, L, J Lines are predicted to grow the most (>30%)



FORECASTING STATION GROWTH

- Analysis included 369 Stations
- Stations have a wider range of expected growth (<5% - >30%)
- Stations predictions are less confident R^2 range 0.1-0.85
- Stations shown are expected to have more than 2000 daily entries and exits in 6 months



CONCLUSIONS

MTA can expect continued positive growth not reaching pre-SaH levels until at least 2023.

Growth will not be uniform for stations and lines. Resource and budgeting can take advantage.

Z, L, J Lines will require the most proportional budgeting increase to accommodate growth

Avoid extra investment in low growth stations such as 18th St and 49th st.

FUTURE WORK

Explore

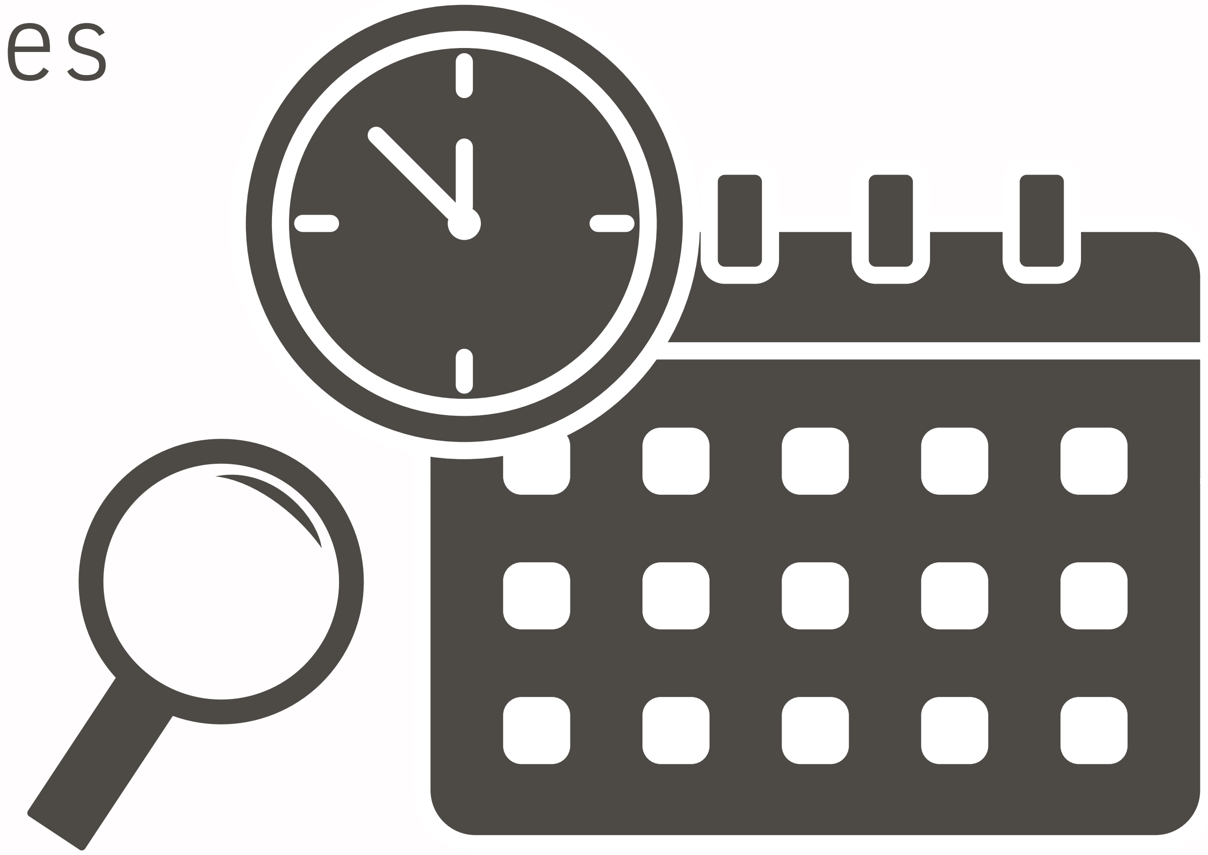
- Lag effects of hospitalization n days before/after

Data

- Expand Time Frame to include years prior to 2020
- Include more features
  - SaH Order Active (Binary)
  - Polynomial Features (Feature Interactions)

Modeling

- Test more algorithms such as RANSAC, MLP
- Utilize non-negative least squares



APPENDIX

Feel free to ask me any questions on inspiration or implementation!

CONTENTS

-Inspiration

-Data

-Features

-Algorithms

-Performance

-Additional Insights

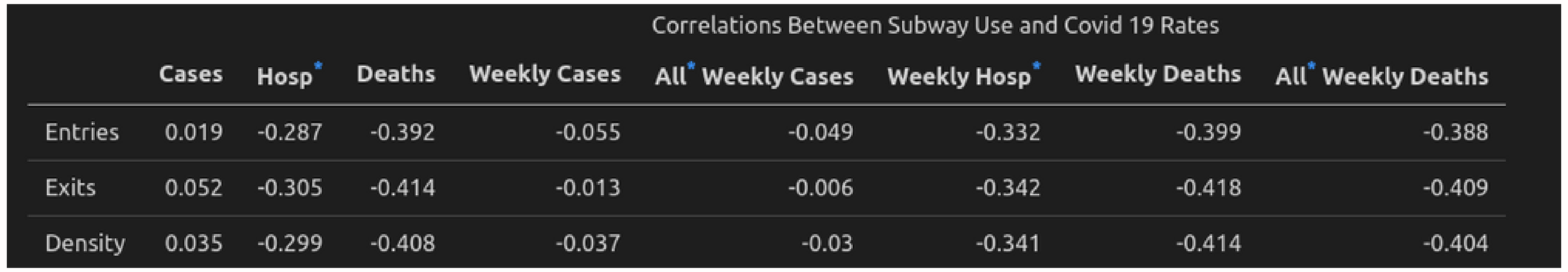
-Limitations

Contact

Email: patrick.ty.brown@gmail.com
Linkedin: https://www.linkedin.com/in/patrick-ty-brown/

INSPIRATION

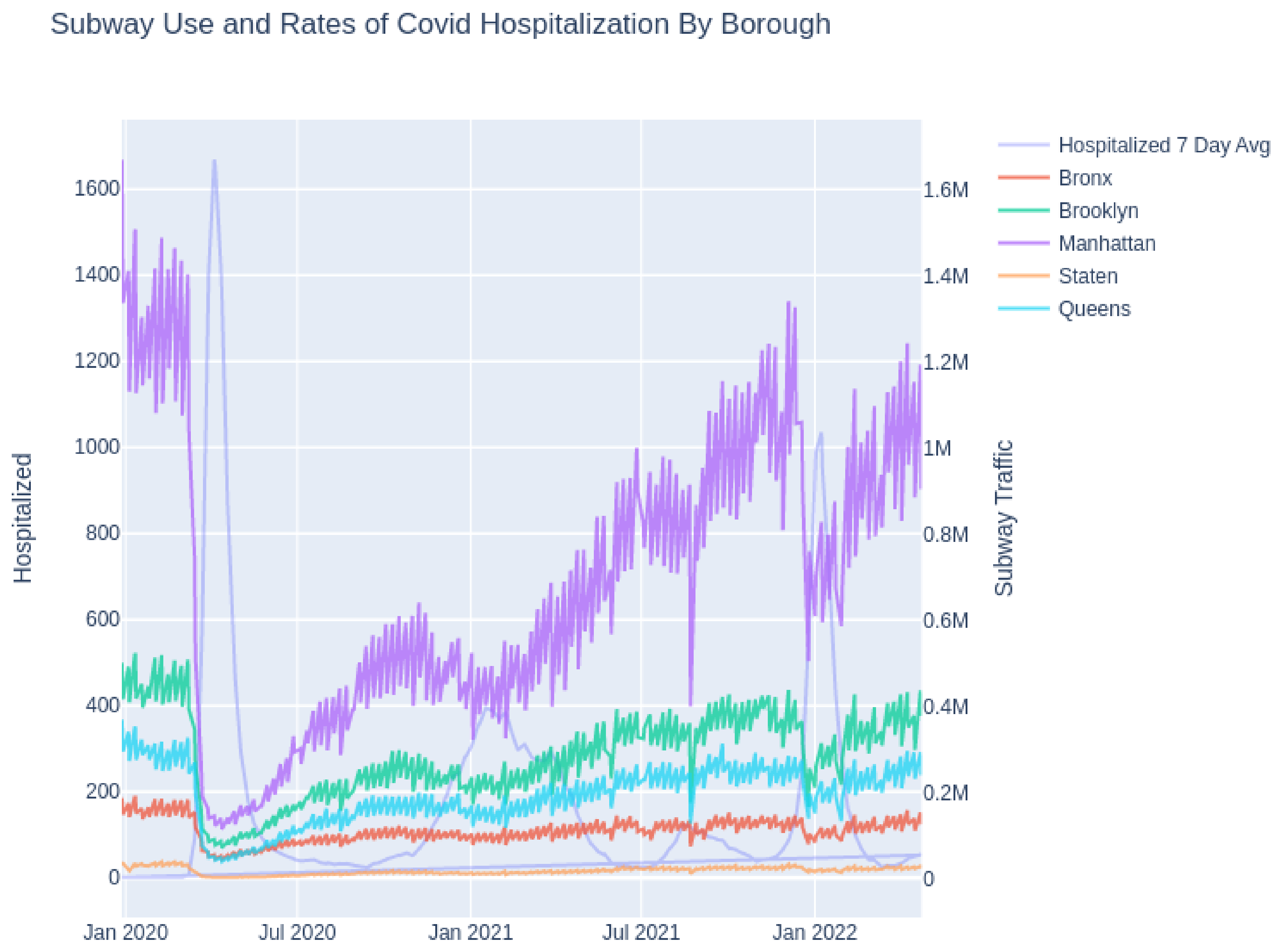
Initial exploration show large correlations between variables



Correlations Between Subway Use and Covid 19 Rates

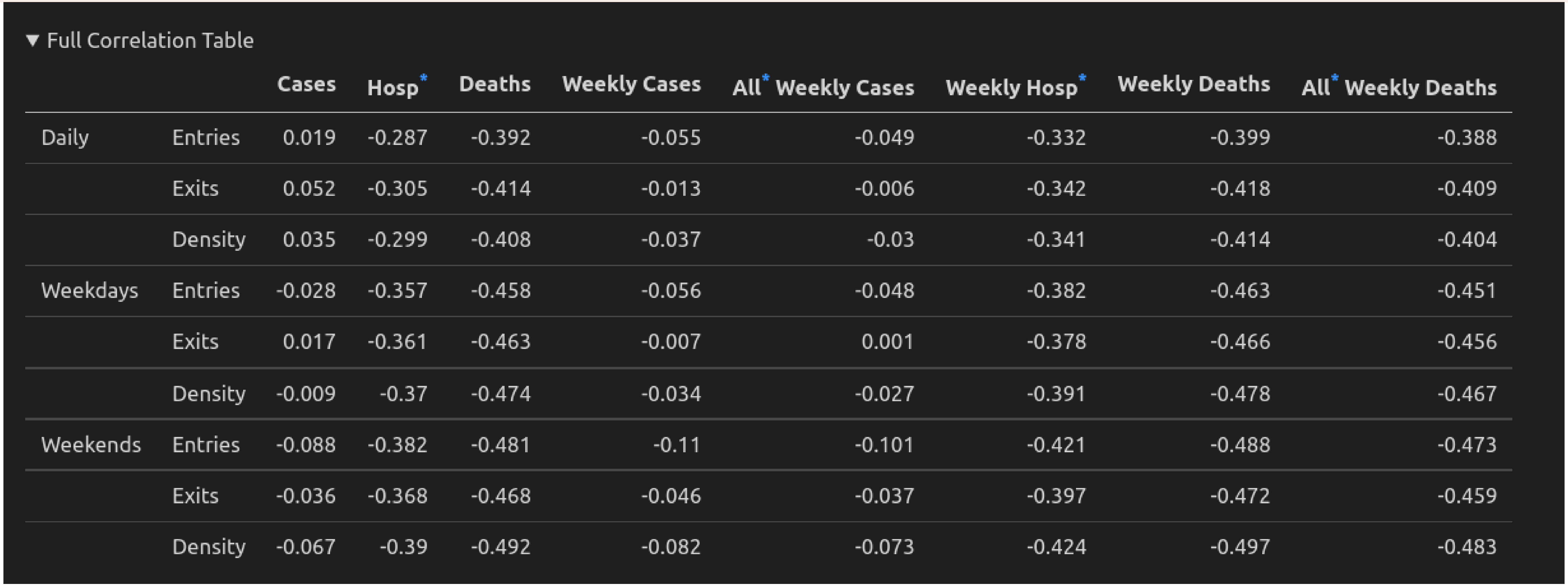
	Cases	Hosp*	Deaths	Weekly Cases	All* Weekly Cases	Weekly Hosp*	Weekly Deaths	All* Weekly Deaths
Entries	0.019	-0.287	-0.392	-0.055	-0.049	-0.332	-0.399	-0.388
Exits	0.052	-0.305	-0.414	-0.013	-0.006	-0.342	-0.418	-0.409
Density	0.035	-0.299	-0.408	-0.037	-0.03	-0.341	-0.414	-0.404

Further Exploration showed discrepancies between borough and subway use (controlled for number of stations). Could this insight be used to inform actions with Lines and Stations?



Subway Use and Rates of Covid Hospitalization By Borough

INSPIRATION CONTINUED...



▼ Full Correlation Table

		Cases	Hosp*	Deaths	Weekly Cases	All* Weekly Cases	Weekly Hosp*	Weekly Deaths	All* Weekly Deaths
Daily	Entries	0.019	-0.287	-0.392	-0.055	-0.049	-0.332	-0.399	-0.388
	Exits	0.052	-0.305	-0.414	-0.013	-0.006	-0.342	-0.418	-0.409
	Density	0.035	-0.299	-0.408	-0.037	-0.03	-0.341	-0.414	-0.404
Weekdays	Entries	-0.028	-0.357	-0.458	-0.056	-0.048	-0.382	-0.463	-0.451
	Exits	0.017	-0.361	-0.463	-0.007	0.001	-0.378	-0.466	-0.456
	Density	-0.009	-0.37	-0.474	-0.034	-0.027	-0.391	-0.478	-0.467
Weekends	Entries	-0.088	-0.382	-0.481	-0.11	-0.101	-0.421	-0.488	-0.473
	Exits	-0.036	-0.368	-0.468	-0.046	-0.037	-0.397	-0.472	-0.459
	Density	-0.067	-0.39	-0.492	-0.082	-0.073	-0.424	-0.497	-0.483

DATA

MTA Turnstile Data

- http://web.mta.info/developers/turnstile.html

Covid 19 Rates for NYC

- https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3

Station Zip Codes

- Shared by Dave Salorio on the EDA Slack channel April 12

FEATURES

Raw Features

- Covid Hospitalization
- Date
- Station Name
- Station Lines
- Station Borough
- Station Zip Code

Engineered Features

- Subway use (Subway Daily Entries + Daily Exits)
- Line Use (Station Subway use / # of Lines at station)
- Days Post NYC Stay at home order (Date – April 1, 2020)
- Fourier Transformation (Effect of Periodicity)

ALGORITHMS

Linear Regression

- Takes an input and fits a line to an output.
- When multiple inputs are used each gain independent slopes.

Singular Input

$$y = \beta_0 + \beta_1 X_1$$

Multiple Inputs

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$$

Fast Fourier Transform

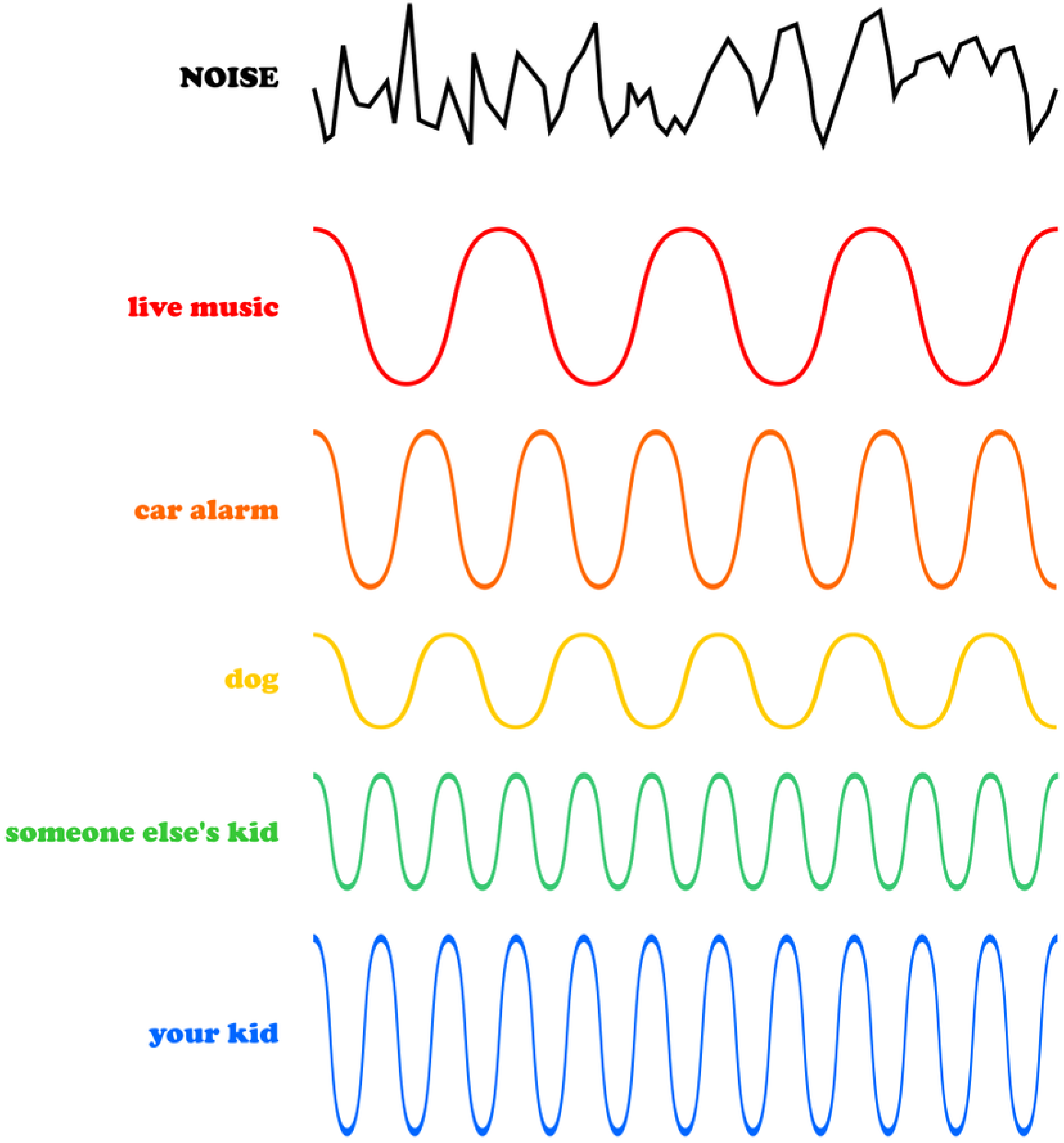
- Powerful tool for audio, image and forecasting domains.
- Takes a frequency (residuals of initial regression) breaks them down to component sine waves.

ALGORITHMS CONTINUED...

Fast Fourier Transform

- Each wave can represent a different periodic effect.

- Adding the waves together we reconstruct the original "noise."



MODEL PERFORMANCE

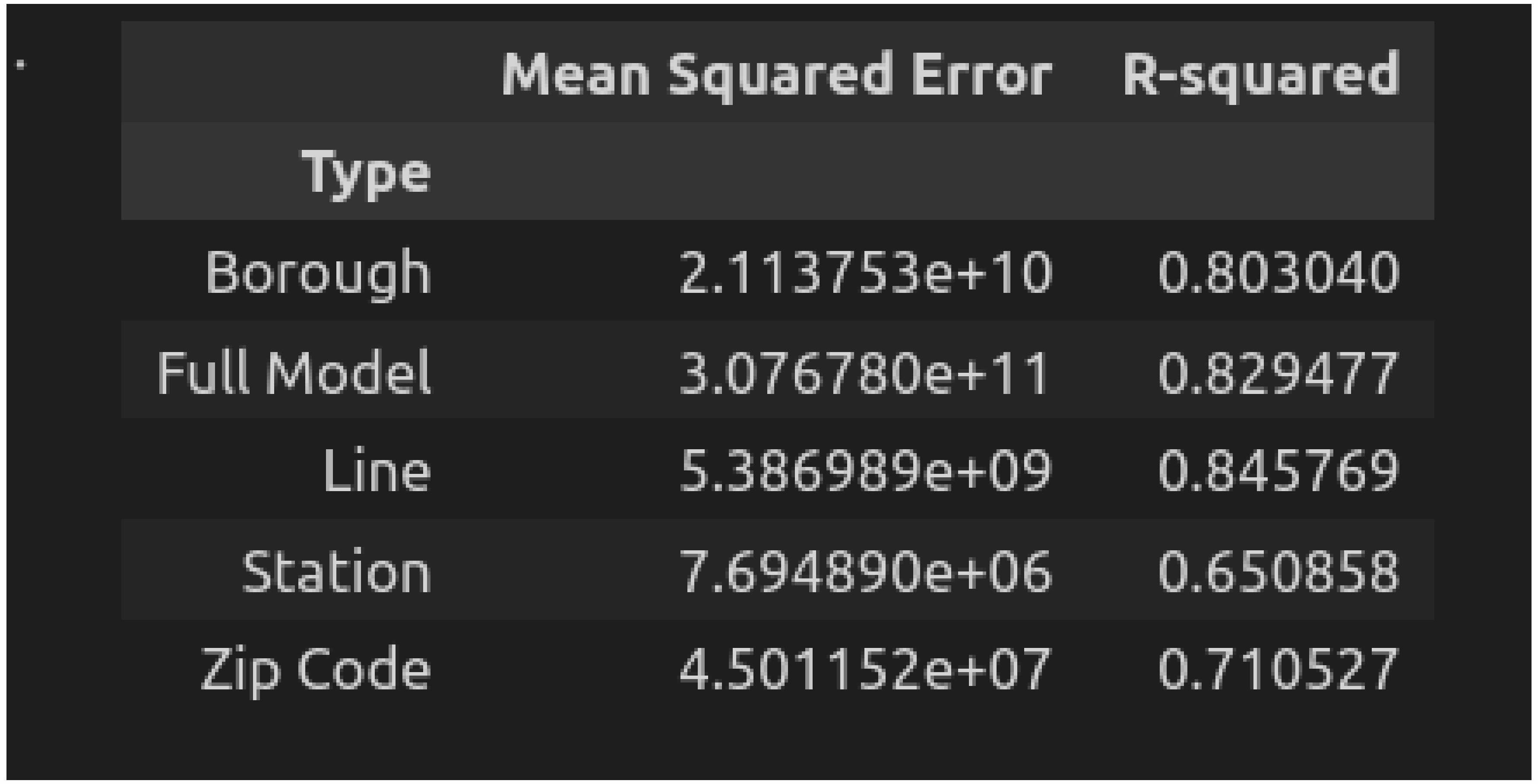
Models were developed for each level of prediction.

- Stations: 369
- Zip Codes: 125
- Lines: 23
- Boroughs: 5
- City as whole: 1

Data for 2022 was held out to better gauge predictive viability.

Models were fit using an 80:20 split on data before 2022. Scores on the test split are presented.

Average Model Performance

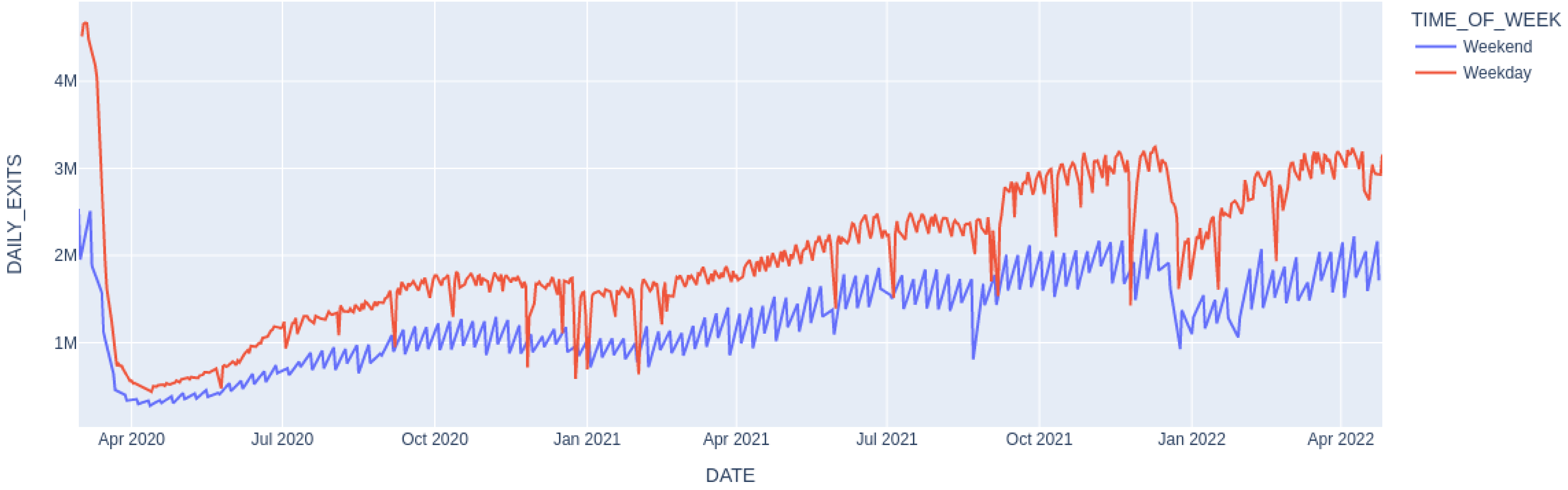


	Mean Squared Error	R-squared
Type		
Borough	2.113753e+10	0.803040
Full Model	3.076780e+11	0.829477
Line	5.386989e+09	0.845769
Station	7.694890e+06	0.650858
Zip Code	4.501152e+07	0.710527

ADDITIONAL INSIGHTS

There is a large day of the week effect on subway use. Use is much higher on Weekdays than weekends.

We hypothesize this is due to commuting for work.



LIMITATIONS

There are a few limitations in this analysis.

- Only data after the stay at home order was used. For the scope of the project it was unnecessary to include.
- Select station models showed extremely poor performance R^2 < .05. These were included in aggregates of performance to avoid misrepresentation. Further investigation will be needed to understand the reasoning.
- Linear regression is ultimately linear and may not capture the full complexity of transit use in NYC compared to more advanced models.