Time series forecasting for Yellow Taxi in New York Inner Urban Boroughs

Ngoc Duy Tran
Student ID: 1156724
Github repo

August 20, 2023

1 Introduction

Conventional yellow taxi in New York City is currently facing challenges due to the thriving of technologically-driven ride-hailing services like Uber, Grab, and Lyft. Therefore, there exists a substantial demand for taxi companies such as TLC to strategically re-allocate their taxi drivers to specific highly-demanded areas at different times during the day to maximize profit.

1.1 Research focus

This research focuses on hourly prediction for highly-demanded taxi boroughs to help the target audience **TLC** answer a million-dollar question: *Given the data in a borough from the previous $w = 96$ hours, what are the predictions for yellow taxi demand in the next $k = 8$ hours?* This research focuses on **Yellow Taxis** due to its flexibility, in which they have no restrictions in picking up, including Manhattan, the core borough of this study.

1.2 Dataset

The major dataset is **TLC Taxi Trip Record Data**, capturing taxi ridership features namely pick-up and drop-off locations [1]. Additionally, this research confines the timeframe from 2022 February to 2023 February, ending up with a dataset of 43,172,888 instances and 19 features. Data before February 2022 was also excluded, attributed by a major surge in COVID-19 cases and fatalities, preventing this data from capturing the genuine underlying pattern in taxi ridership demand [2].

Two external datasets integrated in modelling process are **MTA Subway Hourly Ridership** [3] and **Global Hourly Integrated Surface Database** [4]. The former dataset presents subway ridership estimates in hourly interval in various boroughs starting from February 2022 with 5,423,857 records and 11 features. Furthermore, the latter dataset incorporates hourly records measured from JFK International Airport from 2022 (13,344 records and 100 features) and 2023 (8475 records and 95 features). More information about these 3 datasets and preprocessing steps will be introduced in Section 2.

2 Preprocessing

2.1 Landing layer to Raw layer

After all datasets and supplementary datasets (taxi zone lookup and shape files) are downloaded into the landing layer, the three primary datasets undergo data parsing to ensure consistent data types.

2.2 Raw layer to Curated Layer

2.2.1 TLC Trip Record Data

Having inspected the dataset, the dataset comes with a few different problems requiring further filtering and analysis.

- **Filtering records not in the defined timeframe:** Dataset shape: (43172300,19)
- **Filtering trips with short distance:** Records where trip distance falls below 0.4 miles (500 meters) are excluded as a discerning passengers would not typically opt for a taxi when the travel distance is below 500 meters. (41811443, 19)
- **Filtering trips with short commuting time 60 seconds:** It would be unreasonable to finish the whole trip under a minute. An extra column is created to find the difference (41753931, 20)
- **Filtering trips with long commuting time 4 hours:** The two furthest locations in New York City is Floral Park and The Conference House, which results in only 45 miles and nearly 2 hours if taken by cars according to Google Maps. The threshold is set at 4 hours in case there is a return. (41704217, 20)
- **Filtering trips with long distance:** Records where trip distance exceeds 100 miles are excluded, which doubles the distance from Floral Park to Conference House. (41702718, 20)
- **Filtering trips with fare amount exceeds $300:** With current choice of maximum trip distance of 100 miles and average price per mile of $3. (41504885, 20)
- **Filtering trips with pickup locations within New York City:** For visualization, all trips with pick-up location ID outside the range [2-263] are removed. An extra column indicating borough is added. Newark Airport is also excluded because it is not the target of the study, inner-city boroughs (40984627, 21)

2.2.2 Hourly Integrated Surface Database

The 2022 and 2023 weather datasets undergo the process of data wrangling and feature selection.

- **Feature selection:** `date`, `temperature` and `wind`, are preserved. (13344,3), (8475,3)
- **Outlier filtration:** This involves removing outliers for 2 features (denoted as 9999 according to data dictionary) and instances with unqualified quality to avoid any miscalculations. (12967,3), (8235,3)
- **Merging:** Having been aggregated hourly, both 2022 and 2023 weather datasets are joined with an extra feature created `hour` for joining. (`year_month_date`, `hour`, `avg_tmp_observation`, and `avg_speed`) (8760,4), (5528,4)
- **Timeline filtering:** Both datasets are refined to the specific timeframe ranging from 01/02/2022 to 28/02/2023, (8016,4), (1416,4), resulting in a final weather dataset with 9432 instances and 4 features in total

2.2.3 MTA Subway Hourly Ridership

The MTA Subway hourly Ridership undergoes a similar process of data wrangling

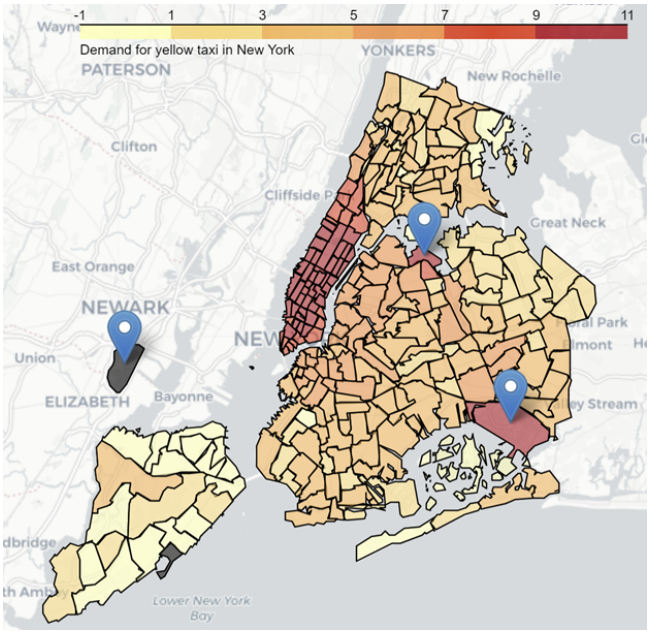
- **Feature selection:** 3 relevant features, namely `transit_timestamp`, `borough` and `ridership`, are preserved. (5423857, 3)
- **Timeline filtering:** The dataset is refined to the pre-chosen timeframe. (3913645, 3)
- **Outlier filtration:** The maximum hourly `ridership` is 999, which is found in Manhattan and queens boroughs. However, this number is within the expected range as both boroughs often experience ridership around 900 during rush hours. (3913645, 3)
- **Aggregation and Borough filtering:** The data is then aggregated based on `borough` and `transit_timestamp`. 4 separate datasets are subsequently stored for 4 unique boroughs: Queens, Manhattan, Brooklyn, and Bronx (9432,3)

For TLC Trip Record Data, 4 inner-city boroughs are aggregated and saved separately, before joining with MTA Subway and Integrated Surface Database on the timestamp to produce 4 other datasets ready for modelling (9432 instances, 5 features).

3 Geospatial Visualisation and Feature Analysis

For **Section 3 Visualization and Feature Analysis**, this research covers highly-demanded boroughs with a focus on Manhattan; while **Modelling Section 4** only covers Manhattan time series forecasting due to computational limitations. Newark Liberty International Airport is not within the 5 boroughs; therefore, will not be covered in the analysis.

3.1 Geospatial visualization



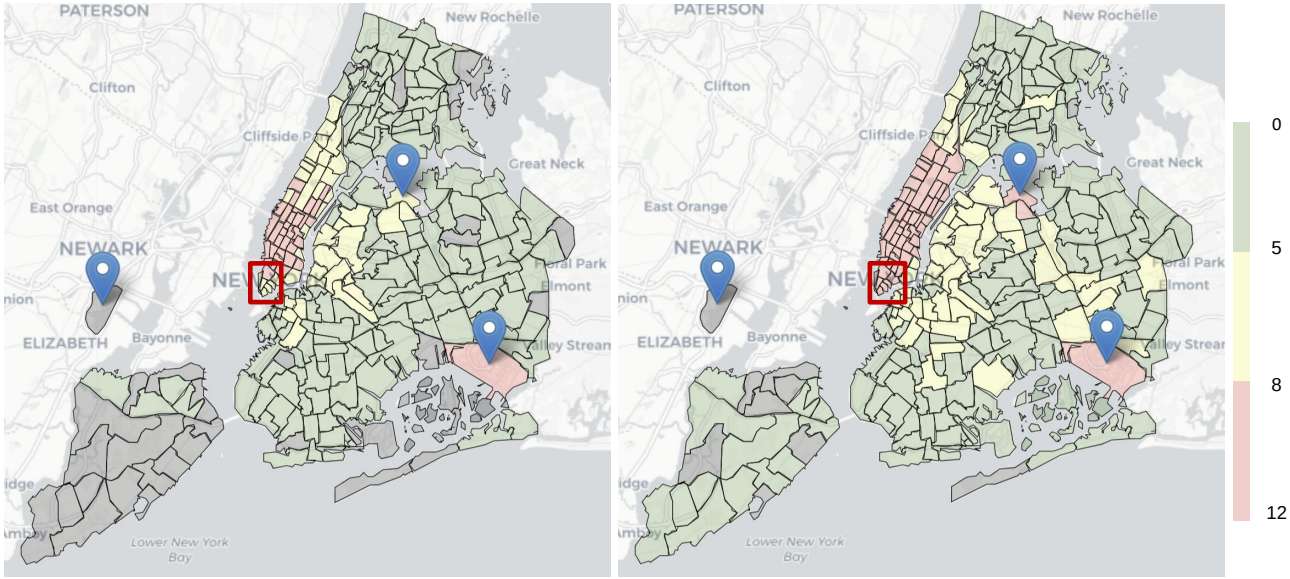
From Figure 1, Manhattan, JFK International Airport, and Laguardia Airport stand out as regions with enormous demand for yellow taxis; thereby, attracting an influx of yellow taxi drivers. Therefore, it would be reasonable for TLC to distribute their yellow taxi drivers in proximity to these locations because demand for yellow taxi in areas adjacent to these 3 major hotspots is still high.

On the other hand, Staten Island observes the lowest demand for yellow taxis among the 5 boroughs, which is reasoned by several factors, including higher prevalence of other taxi brands namely Clove Lake Cars and small population (493,194 in Staten Island vs 1,628,706 in Manhattan).

Figure 1: Log Hourly demand for Yellow Taxi from February 2022 - February 2023

3.2 Geospatial visualization across different times during the day

Figure 2a at 2AM presents findings partially aligning with Figure 1. Though JFK remains highly demanded, it reveals that Laguardia Airport's demand diminishes and Manhattan's demand has now condensed into the downtown and midtown area. These insights could act as TLC's strategy to navigate their yellow taxi drivers to JFK and particularly midtown Manhattan where the profit remains significant, nearly 10,000 rides for a zone at 2 AM. However, not all parts of lower Manhattan are equally busy, illustrated by the red box in Figure 2a. This will be shortly be explained in the next paragraph and followed up in Section 6 Recommendations. Additionally, demand is little outside lower Manhattan and surrounding neighbours, especially in Staten Island, with many zones with no trips recorded (grey).



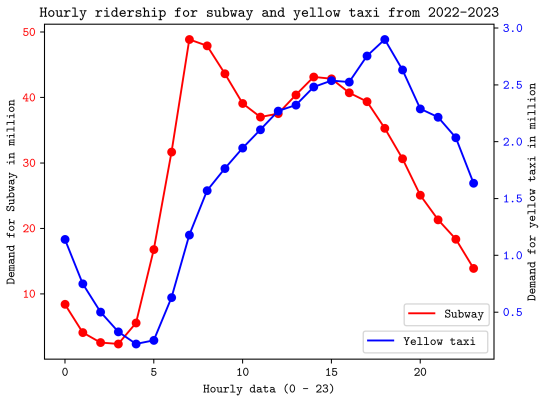
(a) t = 02:00:00 (b) t = 14:00:00

Figure 2: Log Hourly demand for Yellow Taxi across different hours

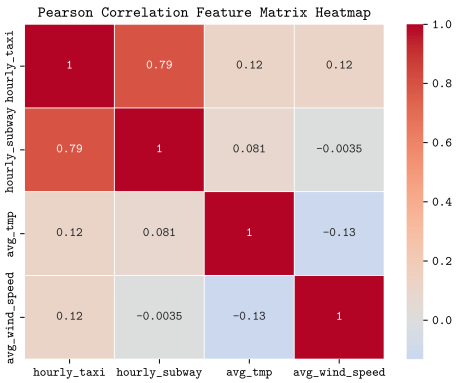
Figure 2b shows while there are discrepancies in 2AM findings, the geospatial demand findings at 14:00 is consistent with results in section 3.1. An interesting observation is that the red box is heavily demanded 12 hours later. This observation is because red-box region comprises of the Financial District experiencing a major surge in working commuting activities. This is further supported by the region's hosting of various tourist attractions, including Statue of Liberty cruises, attracting many taxi drivers around to transport tourists. Therefore, TLC should encourage their yellow taxi drivers towards lower Manhattan during this period to accommodate the elevated demand, while allocating green taxis to other boroughs to achieve maximum profit.

3.3 Feature analysis

Firstly, from Figure 3a, there is a positive correlation between `hourly subway ridership` and `hourly yellow taxi demand` - response variable. Despite certain differences in the mid-day pattern, both features exhibit similar pattern in the early and late parts. They both show a decreasing trend in demand from 6P.M until 5 A.M in the following day, followed by a spike due to substantial demand for means of transportation during work commutes. This finding is further supported by Figure 9b, revealing a Pearson correlation of 0.8 between these 2 features for the Manhattan borough. These 2 evidence signals the significance of `hourly subway ridership` as a powerful predictor.



(a) Hourly ridership for subway and yellow taxi



(b) Pearson Correlation for Manhattan borough

Figure 3: Feature analysis for weather and subway features

Despite low correlation between `temperature`, `wind_speed` and response variable `count`, these features are retained during feature selection. Real-world cases as illustrated by Tan(2022) and Pfosi(2022), showcased that during extreme weather events, the demand for cab was substantially reduced, signalling weather variables are highly powerful for modelling [5, 6].

Regarding **feature distribution**, while the boxplots for `Hourly taxi ridership` and `Hourly subway ridership` exhibit left-skewness, that for hourly temperature is a relatively right-skewed distribution, within a reasonable range from -15 to 35 degrees Celcius. The box plot also reveals outliers in hourly wind speed, with speed exceeding 12.5 meters/second. The author decides not to remove these outliers as they have passed all quality control checks from NCEI data source with quality code of 5. Given the city's unusual weather events such as storms, these outliers are considered reasonable.

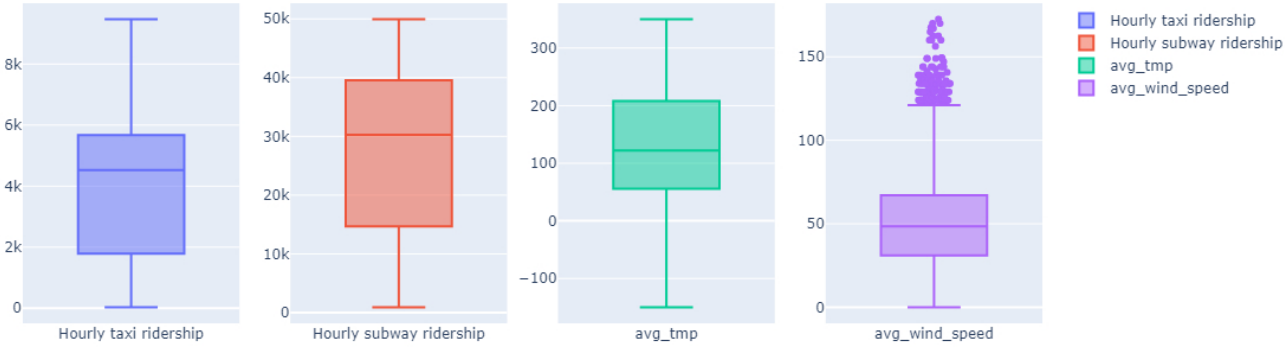


Figure 4: Feature Distribution Boxplots

Regarding missing data for modelling, only 1 out of 9432 instances is found missing in Manhattan borough, which is 0.01% of the dataset. Therefore, linear interpolation based on preceding and subsequent timestamps is implemented for imputation [7].

4 Modelling, an end-to-end example with Manhattan

As mentioned in Section 3, this study will exclusively conduct modelling in Manhattan, the busiest borough where yellow taxi has an unique advantage. Manhattan data (9432 instances, 5 features) is

transformed via window sliding method [8], involving iterative traversal and collection of time-series features and their corresponding responses.

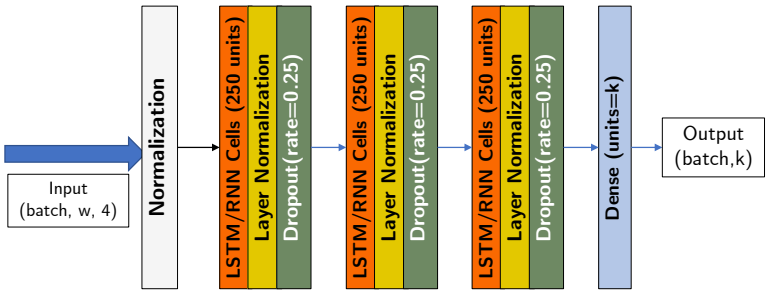
4.1 Methodology overview

Based on a time series $\{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_w}\}$ where $\boldsymbol{x_i}$ is a vector of n features at time i, a time series forecasting algorithm needs to learn the input and returns a function that maps input to predicted values for response variable y from time $w+1$ to $w+k$ for some pre-chosen $k \geq 1$. For example, for w is 96, and k is 4, which means based on data from the previous 96 hours, we are going to predict yellow taxi demand in the next 4 hours.

$$y_{w+i} = f_i(\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_w}) + \epsilon_i \text{ for } 1 \leq i \leq k \tag{1}$$

where y_{w+i} is the response variable value for time $w+i$, w is the window size, k is the future length we need to predict, f_i is the prediction function at time $w+i$ learned from data by deep learning methods, and ϵ_i is the random error - $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

The dataset is then partitioned approximately in ratio: 70% for training (February 2022 - October 2022), 15% for validation (November 2022 - December 2022), and 15% for testing (January 2023 - February 2023)



The data is then fed into 2 Deep Learning Networks, as illustrated in the left. The network utilizes Seq2Seq approach, implementing LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Units) as its components layers. Both LSTM and GRU are known for their exceptional capabilities to process sequential data, thanks to their brilliant architecture using gates [9, 10].

Figure 5: Deep Learning Architectures

5 Error analysis

For this experiment, we will be examining both deep learning architectures in various prediction steps (different future steps) based on training loss of **Mean Squared Error**, and evaluation metrics of **RMSE** . The overall results are summarized in the table below:

Future prediction (k)	LSTM	GRU
1 hour	750.02	**686.46**
4 hours	934.55	**776.28**
16 hours	991.13	**924.81**

Figure 6: Performance (RMSE) at various k for w = 96

Window size (w)	LSTM	GRU
1 day	954.09	**871.37**
2 days	936.43	**922.47**
8 days	**950.24**	986.12

Figure 7: Performance (RMSE) at various w for k = 8

Overall, among the 2 models, GRU networks outperforms LSTM in terms of model's performance (smaller error bolded) because GRU tends to be more powerful than LSTM in small datasets [10].

More importantly, as the future prediction steps in Table 6 increase, the errors observed in both models also increase, which is consistent with the intuition that the more uncertain about the future, the greater error models are likely to make [8].

Another key conclusion from Table 7 is that the greater the window size w, the worse performance GRU network attains, which is reasoned by conditional on more recent data, the response variable and older predictors are weakly related. Therefore, using a large window size w will introduce a higher dimension of covariates with few useful information, resulting in a decline in model's performances [8].

However, when the look-back window is large, LSTM becomes the more powerful model, which is explained by LSTM's superior capability in capturing the long-range high-complexity sequences. Therefore, depending on different purposes whether long or short forecasting for instance, TLC should be cautious in their choice of model and other hyperparameters for the best of their interests.

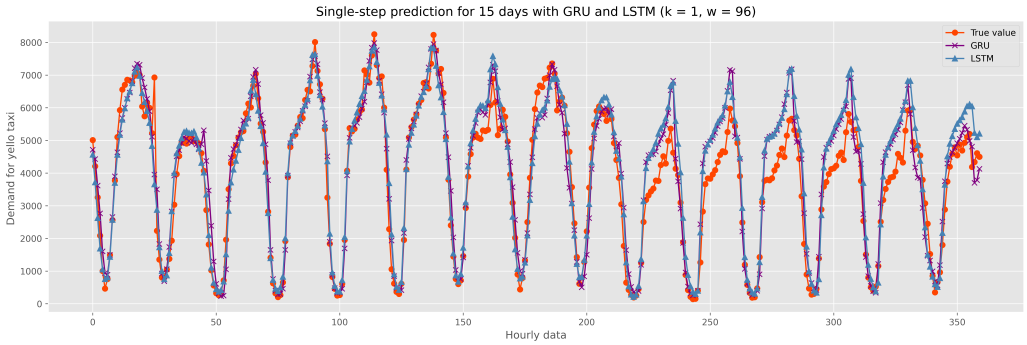


Figure 8: Single-step prediction for 15 days with GRU and LSTM (k=1, w=96)

Figure 8 shows both architectures have achieved strong performance when the predicted values closely match the ground truth; however, GRU is the more powerful model, as illustrated in Table 6 and Table 7. In addition, the highest demand for yellow taxi is during the noon when there is a peak during this period every 24 hours, which implies that TLC should encourage their yellow taxi drivers to be around Manhattan due to peak demands.

However, from 150^{th} hour (January 6, 2023), both models over-estimated the true demand, which could be several factors including seasonal factor, which could be improved by introducing a complex seasonal decomposition block to account for the seasonal factor. However, the models' performance are strong enough illustrated by competitve results and errors.

6 Recommendations

Figure 2b suggests that during daytime hours, TLC should suggest yellow taxi drivers target at high-demanded areas such as JFK, Laguardia Airport and Manhattan. However during the nights and early morning, only JFK and a region of Manhattan should be targeted. Highly-demanded regions comprise of a variety of entertainment establishments such as restaurants and night clubs. Therefore, within Manhattan, TLC should propose taxi drivers with some regions namely Diamond District) illustrated in Figure 9a, hosting numerous clubs and bars; which is the potential market as intoxicated customers usually opt for a private means of transport.

Figure 8 has illustrated how precise both models could be although they are foundational and basic. Hence, TLC should consider investing towards hiring machine learning engineers to apply transfer learning on SOTA pretrained models such as Temporal Fusion Transformer and Autoformers or design

(a) Diamond District - potential market during the nights

(b) DoorDash App - an example interface navigating their shippers towards busy zones

Figure 9: Recommendations for TLC

a novel architecture. Given data is the core element of deep learning, their competitive advantage in terms of massive data collection facilitates generating highly accurate predictions.

With a highly powerful model in hand, TLC could conduct training process at a smaller scale based on zone location ID Madison Square for instance, instead of the entire borough. Building an app accessing these crucial model predictions will act as a reference/assistant to help yellow taxi drivers understand about substantially-demanded zones, from which they could approach. This plan is certainly achievable as a similar case DoorDash, a food shipping business, has already developed an app to help their shippers earn money more efficiently as in Figure 9b.

7 Conclusion

This research focuses on a real-world application of time series forecasting on the demand for yellow taxi in inner urban areas of New York City, with an end-to-end example on Manhattan. With the help integration of external datasets such as Hourly Integrated Surface Database and MTA Subway Dataset, this has created a simple yet highly effective in forecasting the demand.

Different experiments with various hyper-parameters have been conducted to help TLC understand about the models output and its potential. Therefore, this author highly recommends TLC to apply this method and take recommended actions to create a more balanced play field with Uber and Lyft.

References

[1] NYC Taxi and Limousine Commission. *TLC Trip Record Data*. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08-07.

[2] The New York Times. *Track Covid-19 in New York*. https://www.nytimes.com/interactive/2023/us/new-york-covid-cases.html. Accessed: 2022-08-07.

[3] New York State. *MTA Subway Hourly Ridership: Beginning February 2022*. https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-Beginning-February-202/wujg-7c2s. Accessed: 2022-08-07.

[4] National Centers for Environmental Information. *Integrated Surface Dataset (Global)*. https://www.ncei.noaa.gov/access/search/data-search/global-hourly. Accessed: 2022-08-21.

[5] Ming Hui Tan. *Predicting Hourly Demand for Taxi Rides in NYC*. https://canvas.lms.unimelb.edu.au/courses/158133/files/15975429?module_item_id=4966415. Accessed: 2022-08-07.

[6] Nick Pfosi. *Major Nor'easter blankets U.S. East Coast with snow, heavy winds*. https://www.reuters.com/world/us/us-east-coast-prepares-heavy-snow-plunging-temperatures-blizzard-hits-2022-01-29/. Accessed: 2022-08-07.

[7] Pandas. *pandas.series.interpolate*. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.interpolate.html. Accessed: 2022-08-07.

[8] Duy N. Tran. "Multivariate Time Series Forecasting with N-beats Architecture". University of Melbourne, 2023. URL: https://ms.unimelb.edu.au/__data/assets/pdf_file/0009/4524228/Duy_Ngoc_Tran_-_Multivariate_time_series_forecasting_with_N-beats_architecture1.pdf.

[9] Haşim Sak, Andrew Senior, and François Beaufays. "Long Short-term Memory Based Recurrent Neural Network Architectures For Large Vocabulary Speech Recognition". In: *arXiv preprint arXiv:1402.1128* (2014). URL: https://arxiv.org/abs/1402.1128.

[10] Junyoung Chung et al. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *arXiv preprint arXiv:1412.3555* (2014). URL: https://arxiv.org/abs/1412.3555.

The author did use ChatGPT as a tool to achieve Latex and coding visualizations. This author also reuses code chunks and previous findings which were written by this author 6 months ago - Ngoc Duy Tran, conducted under Unimelb Summer Vacation Scholarship Program 2023 [8]

In case I understand this wrong, this link is the last commit I have made.