TTC Subway Delay Prediction*

Junwei Chen

January 24, 2024

With a focus on its importance in everyday urban life, this project studied and explored Toronto's subway delay issue. In order to advance scientific knowledge of urban transportation dynamics, this project develop predictive model for delay estimation, find patterns in the data, and determine the factors that contribute to delays. The paper presents the dataset sourced from open source opentorontodata, undertakes an exploratory data analysis to identify temporal patterns, and ends with the presentation of the predictive model.

1 Introduction

The Toronto subway system plays a crucial role in the daily lives of residents and commuters. Understanding the factors contributing to delays is essential for improving the overall transit and commuting experience. In this paper, we studied and took a deep dive into the delay data, examining trends, identifying potential causes, and ultimately creating a predictive model for estimating delay durations.

The paper is organized as follows: first, we present the dataset and its context; next, we perform an Exploratory Data Analysis to understand the delay patterns; finally, we introduce our predictive model and discuss its implications for the transit system.

2 Data

TTC Subway & SRT Train Service Delay Data sourced from the Toronto Open Data (data catalog), providing a comprehensive record of subway delays over a particular period. This dataset is invaluable for assessing the reliability of the subway system and identifying potential areas for improvement. Information gathered between January 1, 2023, and December 31, 2023.

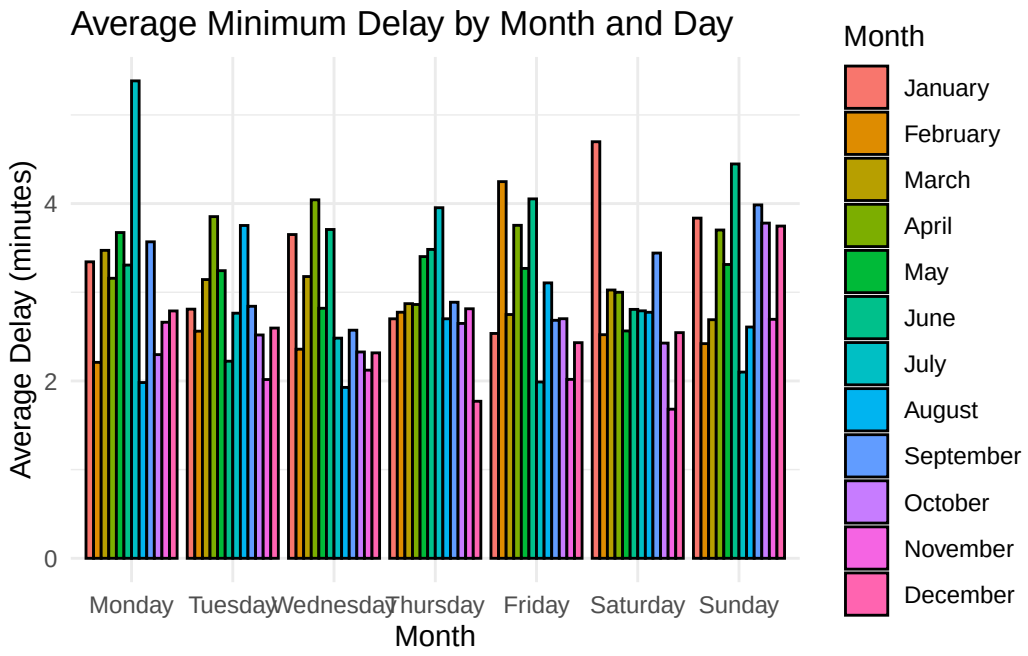
*Code and data are available at: Github Repository

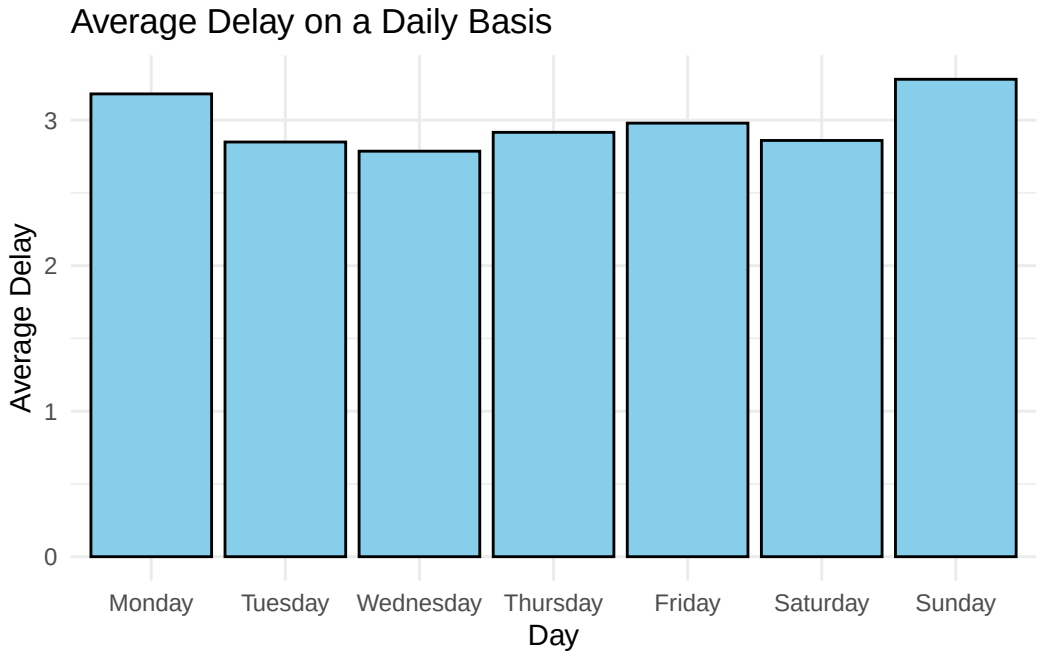
Date	Time	Day	Station	Code	Min Delay	Min Gap
2023-01-01	02:22:00	Sunday	MUSEUM STATION	MUPAA	3	9
2023-01-01	02:30:00	Sunday	KIPLING STATION	MUIS	0	0
2023-01-01	02:33:00	Sunday	WARDEN STATION	SUO	0	0
2023-01-01	03:17:00	Sunday	KEELE STATION	MUIS	0	0
2023-01-01	07:16:00	Sunday	BATHURST STATION	MUIS	0	0

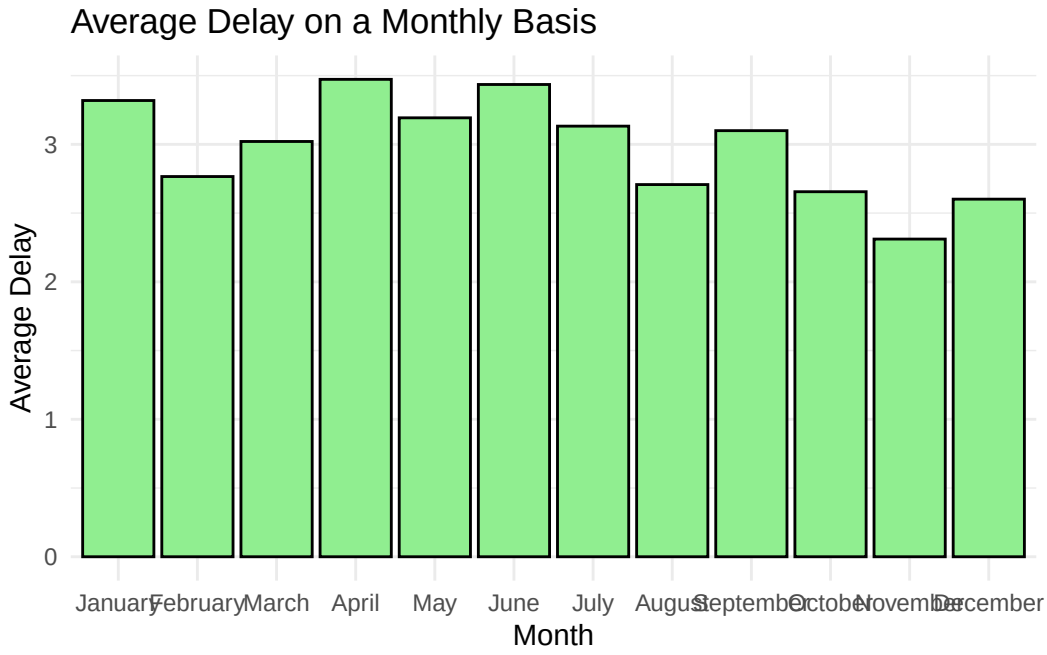
3 Ethical and Statistical Considerations

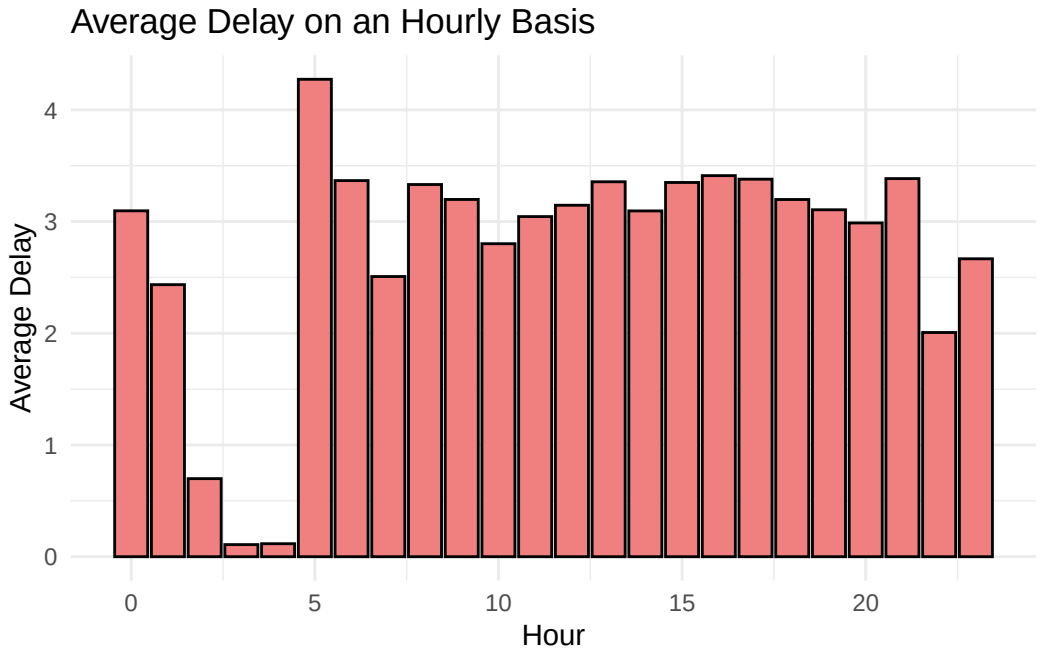
Before performing advanced analysis, it is crucial to acknowledge the ethical and statistical dimensions of the data. This paper consider issues related to data privacy, the representativeness of the sample, and the potential impact on various demographics. Statistical methods used in the analysis are carefully chosen to ensure the robustness of our findings.

4 Exploratory Data Analysis




Average Delay on a Daily Basis
Average Delay
0
1
2
3
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day


Average Delay on a Monthly Basis
Average Delay
0
1
2
3
January
February
March
April
May
June
July
August
September
October
November
December
Month



In 2023, the biggest delay happens at 5 a.m., with an average delay of 3 minutes. It's interesting to consider that the train may leave the station at five in the morning and not leave until every passenger has boarded. It's also noteworthy that, although Sundays are supposed to be free days, there is a high average weekly train delay.

5 Predictive Model

Min Delay	Min Gap	Min Delay Pred
0	0	0.4488083
0	0	2.3179212
0	0	2.3179212
0	0	2.3179212
7	0	2.3179212
0	0	1.2311845
9	11	4.8893226
5	9	4.1318643
0	0	1.2311845
0	0	1.2311845

To enhance our understanding of subway delays, we develop a predictive model using machine learning techniques, Random Forest. Details of the model and its performance can be

found om github repository. By isolating December data as test set, the model can achieve Mean Absolute Error of 1.222971 minutes and Root Mean Squared Error (RMSE) of 1.909541 minutes.

6 References

References to R and relevant R packages (Wickham 2016; Xie 2023; Iannone et al. 2022) are included in the bibliography. Moreover, we acknowledge the Toronto Open Data initiative as the source of the subway delay dataset.