

Analyzing Toronto Islands Ferry Tickets Sales and Redemption Count Data

Yun Chu

September 23, 2024

The paper firstly discussed the background of the dataset, then presented and analyzed the cleaned dataset. In the result section, the relationship between the sales and redemption data, the distribution of redemption count, and the trend of redemption count with respect to time is analyzed. The discussion section comes up with some suggestions to Toronto Islands ferries based on the findings and discuss the limilation of the analysis.

Table of contents

1	Introduction	1
2	Data	2
2.1	Raw Data	2
2.2	Cleaned Data	2
2.3	Summary of Data	3
3	Result	4
3.1	Investigating the Relationship between the Sales and Redemption Data	4
3.2	Investigating the Frequency of Redemption Count Data	5
3.3	Investigating the Monthly Redemption Data	6
4	Discussion	6
4.1	Suggestions to adjusting price based on season	6
4.2	Measures government can take on to cope with high-volume season	7
4.3	Monthly Schedules based on historical data	7
4.4	Weaknesses and next steps	7
	Appendix	8

A Additional data details	8
References	9

1 Introduction

Ferries to Toronto Island Park is a service that operates all year round. Ferries carry passenger from Jack Layton Ferry Terminal to Center Island, Hanlan's Point and Ward's Island or from the three islands to the terminal (City of Toronto (n.d.)).

They are popular places to visit for tourists and people who live in Toronto. Most people take the ferries to and from Toronto Islands while others take water taxis. By analyzing the sales and redemption count data of the ferry to Toronto Islands, we can better understand visitor trends to help with maximizing resource allocation and analyze revenue data.

Through identifying peak seasons for ferry usage, it would help understand at what times are the islands most popular. This would inform details about resource allocation and operational planning information for governments. For example, the information could be used to design the departure and returning schedule of the ferry and the number of staff on duty in each season. By examining sales data over time, revenue trends can be identified and thus help assess the financial health of the ferry service and guide pricing strategies.

This paper reveals that the sales and redemption data are highly correlated and is highly likely to have a linear relationship. It also finds that in summer months, the number of redemption is much higher compared with winter months. Moreover, the 95% of redemption is below 16,076 per day.

The remainder of this paper is structured as follows. Section 2 discusses about the features of raw data and the cleaned data, as well as some summary of the cleaned data. Section 3 discusses about the details of investigating the relationship between Sales and Redemption Data through graphs, linear regression and correlation. Besides, the section also presents and analyses the distribution of the redemption count data based on frequency and month. Section 4 elaborates further on the implication and application of the analysis in Section 3. It also discusses about the weakness of the analysis and the ways to improve.

2 Data

2.1 Raw Data

The data used in this paper is taken from Toronto Open Data and is read into the paper using the opendatatoronto library (Gelfand 2022). The raw data contains the timestamp data per 15 minutes and the sales and redemption count data for each 15-minute time collected from

online and POS-kiosk sales(City of Toronto (2024)).The it contains data from July 11st, 2023 to September 20, 2024. The data is updated hourly.

The dataset is downloaded, cleaned and analysed using R (R Core Team (2023)). The following packages are also utilized:

- tidyverse (Wickham (2021))
- lubridate (Grolemund and Wickham (2011))
- dplyr (Wickham et al. (2023))
- knitr (Xie (2015))
- scales (Wickham (2018))

2.2 Cleaned Data

A sample of the cleaned data (Table 1) and a line plot (Figure 1) showcasing every sales and redemption count data by date are shown below.

Table 1: Sample of Cleaned Sales and Redemption Data

Date	Sales	Redemption
2023-07-11	124	44
2023-07-12	8866	9322
2023-07-13	6651	5465
2023-07-14	12817	11960
2023-07-15	12296	14028
2023-07-16	15684	17008

From the line graph, we can see that the sales and redemption data is highly overlapped and that the values are much bigger in summer compared with winter.

2.3 Summary of Data

To have a better understanding of the data, the mean sales and redemption counts are calculated in Table 2.

Table 2: Mean Sales and Redemption Count

Mean Sales	Mean Redemption
4746.505	4785.342

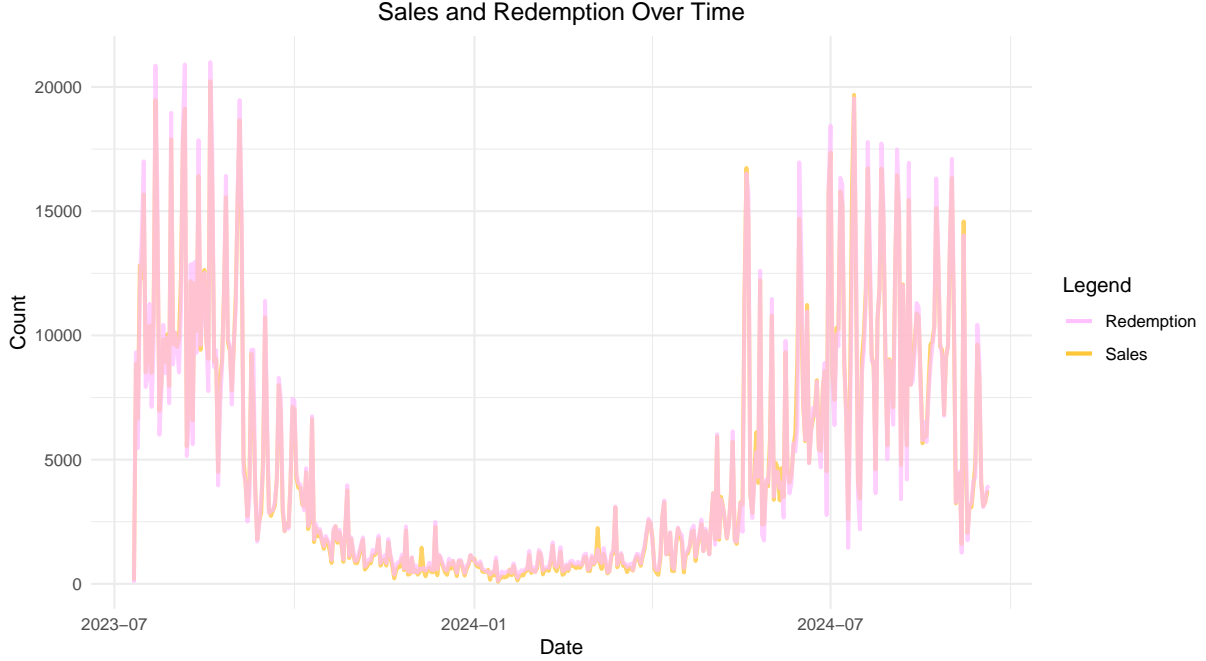


Figure 1: line plot of sales and redemption data with respect to time

The mean sales is very close to mean redemption. From both Figure 1 and Table 2, the sales and redemption data look very alike. The analysis of the relationship between sales and redemption is done in the next section.

3 Result

3.1 Investigating the Relationship between the Sales and Redemption Data

To investigate the relationship between sales and redemption data directly, we first plot Figure 2 where it describe the relationship between the variables visually.

It is obvious to tell that they have a linear relationship. To further investigate, we do a linear regression with redemption count as the response variable and sales count as the predictor variable. As Table 3 shows, the R-squared statistic is 0.9882 and the correlation between redemption count and sales count is 0.994. These values indicates that the redemption count and sales count are indeed highly correlated.

In this paper we'll be mainly focusing on analyzing the redemption count data since it is the direct reflection of the visitor data and using it to analyze revenue is accurate enough since the two variables are higly overlapped and correlated.

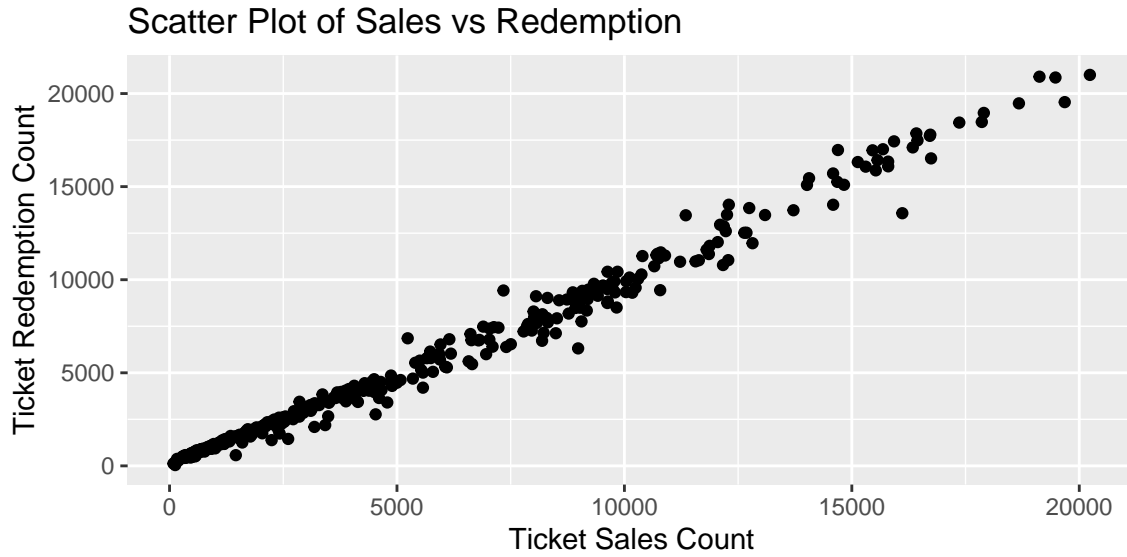


Figure 2: Relationship between Sales and Redemption

Table 3: R-squared and Correlation Statistic

Statistics	Value
R-Squared	0.9882393
Correlation	0.9941023

3.2 Investigating the Frequency of Redemption Count Data

Figure 3 shows that for most of the days, the redemption count is below 10,000 times, while the rest of days has redemption count up to 20234 times in one day. For our purpose of using the information to help with resource allocation, monthly data would be sufficient enough so that we can create monthly schedule for staff. Creating daily schedules would be not as feasible and efficient as monthly schedules.

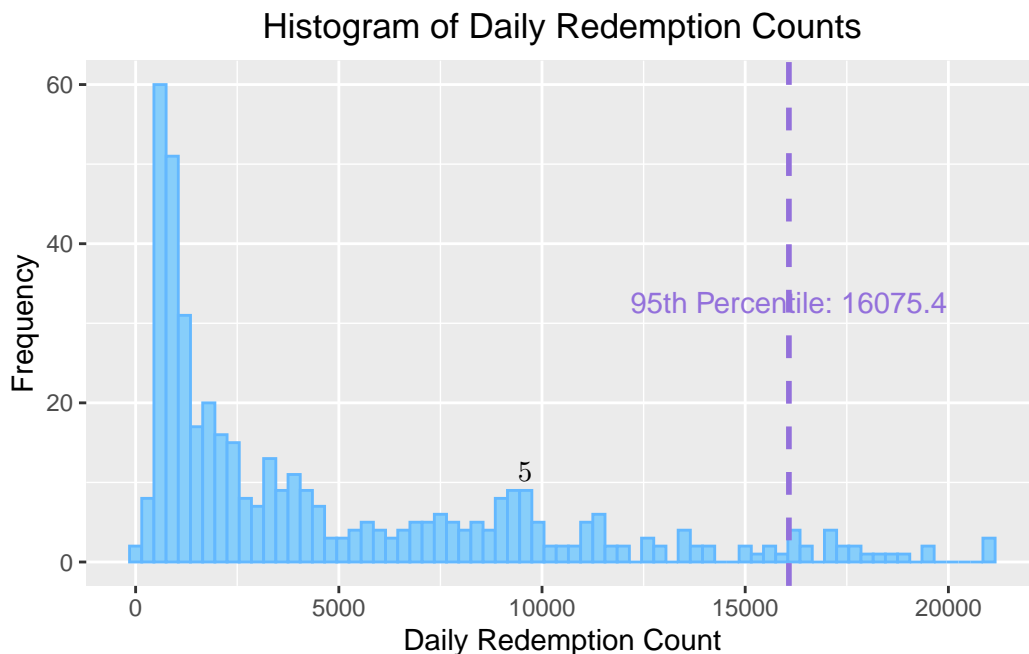


Figure 3: Histogram of Daily Redemption count data

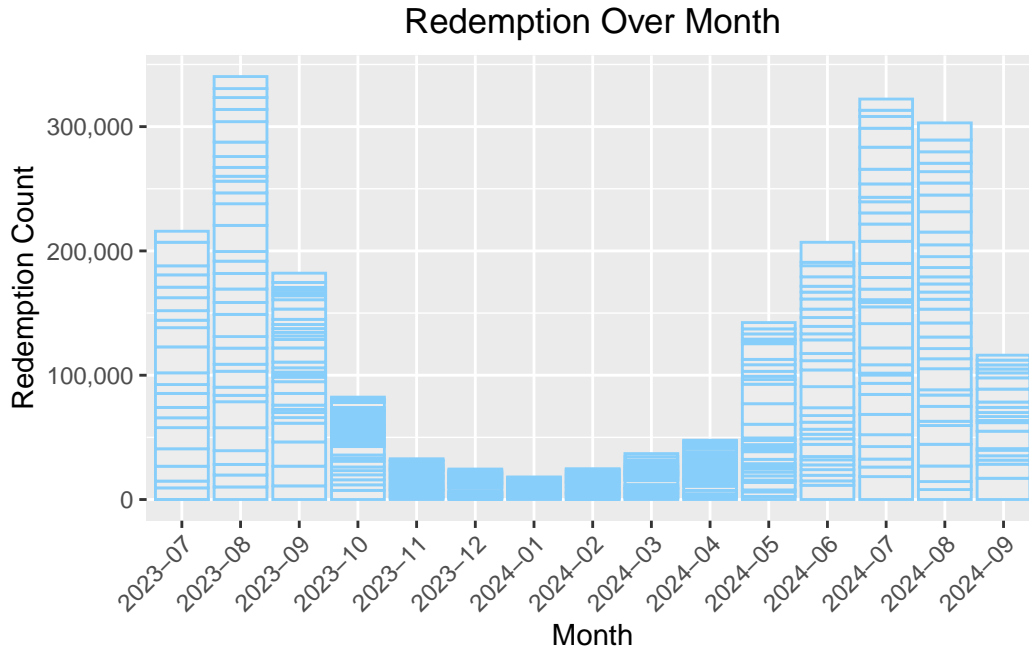


Figure 4: Redemption Count of Each Month

3.3 Investigating the Monthly Redemption Data

From July to September 2023 and May to September 2024, the monthly redemption count exceeds 100,000 each month. From October 2023 to April 2024, the redemption count is less than 100,000 each month. It would be reasonable to design two schedules for the ferries and assign different number of staff to be on duty based on this threshold.

4 Discussion

4.1 Suggestions to adjusting price based on season

From the high correlation and linear relationship between the sales and redemption data, one possible reason might be that the ferry price is flat for all years. If the price is set to be lower in off-peak season, that might help increase revenue in winter while the loss of revenue in summer might not be high.

4.2 Measures government can take on to cope with high-volume season

Based on Figure 3, one measure that Toronto Islands ferries could adapt is that during the peak seasons, integrate the staff schedules to be prepared for large visitor volumes. For example,

having some janitors to be on-call in case there is a large amount of people using the ferries service and there's not enough staff to clean the garbage. This also leads to the next discussion point, which is the on-call schedule should be set based on peak season and off-peak season. If it's winter months, preparing the number of staff to cope with more than 20,000 visitors per day is just waste of resources.

4.3 Monthly Schedules based on historical data

from Figure 4, a noticeable difference is the height of the bars between May-September and the rest of the months in a year. Right now, the Toronto Island ferries, already adapt different schedules for summer and winter months to allocate resources efficiently. Taking a closer look at the graph, July and August have much higher redemption count compared with the rest of summer months. Further splitting the schedules might be helpful to better meet visitors' need and avoid wasting resources.

4.4 Weaknesses and next steps

By examining the linear relationship of redemption and sales data in Section 3.1, the analysis is done using redemption data for simplicity and the representation of the actual visitor volume of the ferries. If we're want to get a better sense of anything related to revenue, analysis with the sales data would be more accurate.

Appendix

A Additional data details

References

- City of Toronto. 2024. “Toronto Island Ferry Ticket Counts.” <https://open.toronto.ca/dataset/toronto-island-ferry-ticket-counts/>.
- . n.d. *Ferries to Toronto Island Park*. <https://www.toronto.ca/explore-enjoy/parks-recreation/places-spaces/beaches-gardens-attractions/toronto-island-park/ferries-to-toronto-island-park/>.
- Grolemund, Garrett, and Hadley Wickham. 2011. *Lubridate: Make Dealing with Dates a Little Easier*. <https://CRAN.R-project.org/package=lubridate>.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Wickham, Hadley. 2018. *Scales: Scale Functions for Visualization*. <https://CRAN.R-project.org/package=scales>.
- . 2021. *Tidyverse: Easily Install and Load the 'Tidyverse'*. <https://CRAN.R-project.org/package=tidyverse>.
- Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.
- Xie, Yihui. 2015. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. <https://CRAN.R-project.org/package=knitr>.