Marriage License Statistics Analysis*

Jerry Xia

2024-09-19

This report analyzes Marriage License Statistics in Toronto using data from Open Data Toronto. We simulate data using the Poisson distribution with lambda = 10, clean the raw dataset, and provide insights with visualizations. A discussion of key results and limitations follows.

1 Introduction

This paper presents an analysis of Marriage License Statistics in Toronto, combining simulated data with actual datasets retrieved from the Open Data Toronto portal. Our analysis focuses on the trends in marriage licenses issued over time, providing insight into potential influencing factors. This document also explores limitations and future directions.

2 Data

We obtained the raw data on Marriage License Statistics in Toronto from Open Data Toronto (Gelfand (2022)).

Below is a visualization of the simulated data.

The graph in Figure 1 demonstrates the fluctuating counts of marriage licenses in the simulated dataset. These counts are generated using the Poisson distribution, which is ideal for modeling count data with an average rate.

*Code and data are available at: https://github.com/Jerryx2020/starter_folder

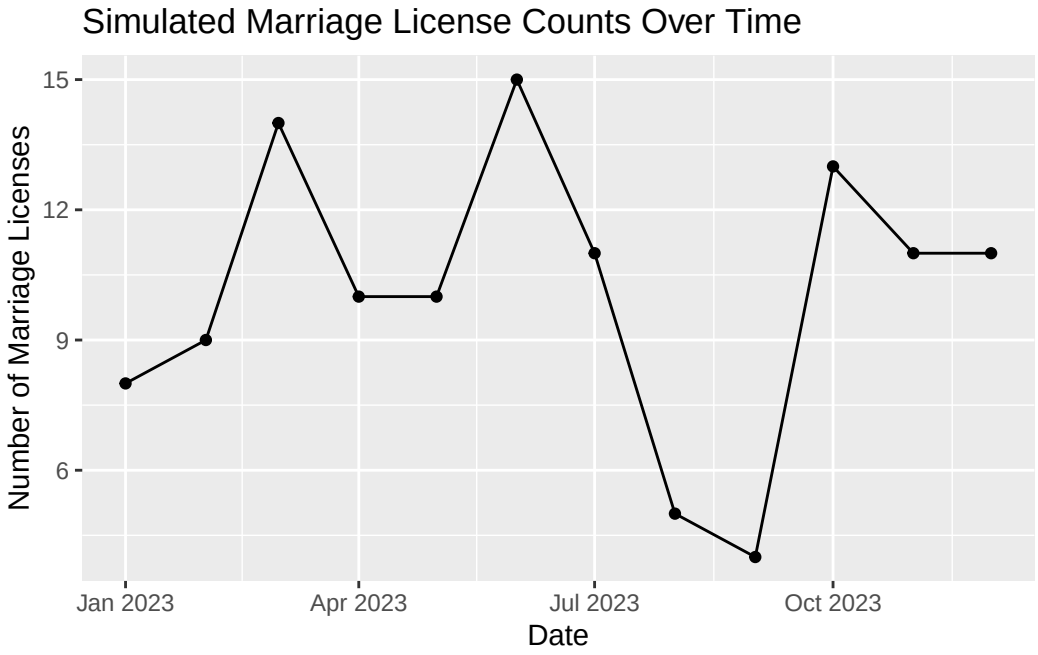

Figure 1: Simulated Marriage License Data

3 Discussion

3.1 Observations from the Simulated Data

The graph shows that the number of marriage licenses fluctuates monthly, reflecting randomness as expected in real-world events. The average count aligns with the Poisson distribution's expected value of 10.

3.2 Future Directions

While this simulated data provides a baseline for analysis, integrating actual marriage license data from Open Data Toronto allows for deeper insights. Key future steps include comparing trends across years and investigating potential external factors influencing license issuance.

3.3 Weaknesses and Next Steps

The simulated data, while useful for testing, does not capture seasonal or external trends that may exist in real data. Future work should focus on cleaning and analyzing the real dataset to identify patterns and correlations.

Appendix

3.1 Data Simulation Code

```
# Simulating marriage license data using Poisson distribution
set.seed(123)
dates <- seq(ymd('2023-01-01'), by = "month", length.out = 12)
marriage_licenses <- rpois(12, lambda = 10)

data_simulated <- tibble(date = dates, marriage_licenses = marriage_licenses)
data_simulated
```

```
# A tibble: 12 x 2
   date       marriage_licenses
   <date>                 <int>
 1 2023-01-01                 8
 2 2023-02-01                 9
 3 2023-03-01                14
 4 2023-04-01                10
 5 2023-05-01                10
 6 2023-06-01                15
 7 2023-07-01                11
 8 2023-08-01                 5
 9 2023-09-01                 4
10 2023-10-01                13
11 2023-11-01                11
12 2023-12-01                11
```

4 References

- Wickham et al. (2019), R Core Team. (2021). R: A Language and Environment for Statistical Computing.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.