Data Science Report
Group 2: Foundations of Data Science

University of Waterloo

December 10, 2023

Collaborators:

Anuj, Singh, a794sing
Adrian, Cortes, a4cortes
Abhishek, Patel, a564pate
Abd Alrhman, Alloush, aaallous
Aaiman, Aamir, a5aamir

Table of Contents

Abstract

The comprehensive data science report investigates the relationship between fire response times and socioeconomic status across Toronto's wards. Utilizing datasets from the City of Toronto's Open Data Portal, the study aimed to discern if a significant difference in response times exists between higher-income and lower-income wards. The analysis involved preprocessing a substantial dataset, integrating external income data, and conducting a detailed statistical examination. Despite initial hypotheses suggesting wealthier wards would experience shorter response times, the findings reveal a weak correlation between ward income and response time, indicating that income alone does not significantly dictate the efficiency of emergency response services. The choropleth map analysis further elucidated the spatial distribution of response times, highlighting a concentration of longer response times in central wards and suggesting the influence of urban dynamics. The report concludes that socioeconomic status is not a standalone predictor of response times, recommending that future research incorporate broader urban planning variables to uncover deeper insights into the factors affecting emergency response efficiency.

Objectives

Goals:

The goal of this report is to investigate and analyze the temporal relationship between fire alarm initiation time and arrival time of fire response units for fire incidents in many different wards of Toronto, with a specific focus on assessing whether there is a significant difference in response times between wealthier and lower-income wards. The study aims to contribute valuable insights into the factors influencing emergency response efficiency, providing a basis for potential improvements in service delivery.

Questions to be answered:

1. Data Collection Considerations
    - What is the dataset's source and timeframe for fire incidents in Toronto?
    - How is socioeconomic status measured for different wards in Toronto?
    - What variables are available in the dataset, and how are they defined?
2. Temporal Analysis:
    - How is the time delta between alarm time and arrival time calculated?
    - Are there any temporal patterns or trends in the overall dataset?
3. Socioeconomic Factors:
    - How are the socioeconomic levels determined for each ward?

- What other socioeconomic variables might influence emergency response times?
- Are there any additional factors (e.g., population density) that may need consideration?

4. Statistical Analysis:

- Are there any null values that limit the outcome of the analysis?
- Are there potential confounding variables that need to be controlled for in the analysis?
- Can we identify specific wards or groups of wards that exhibit notable differences in response times?
- Are there outliers or exceptional cases that require special consideration?
- How do response times vary across different income levels within the wealthier and lower-income wards?

5. Findings and Outcomes:

- How can the findings contribute to policy recommendations or improvements in service delivery?
- Are there any ethical considerations or potential biases in the analysis that need to be addressed?
- How can the study be expanded or improved in future investigations?

Hypothesis:

We posit that there exists a statistically significant difference in the temporal interval between the recorded alarm initiation time and the actual arrival time of emergency response services for fire incidents in Toronto. Specifically, we anticipate that this time delta is shorter in wards characterized by higher socioeconomic status compared to those in wards with lower income levels.

Approach:

1. Identify Key Variables:
    a. Examine 43 columns in the dataset, focusing on essential variables:
    b. ***TFS_Alarm_Time*** and ***TFS_Arrival_Time*** for proper time format alignment.
    c. ***Response_time***, calculated by subtracting alarm time from arrival time.
    d. ***Incident_Ward***, providing ward identification.
2. Data Simplification:
    a. Condense the dataset (29419 rows × 44 columns) to include only the identified key variables.
3. Organize Data:
    a. Arrange the table in ascending order, omitting rows with missing values.
    b. Convert time format to a more readable representation (e.g., minutes).
    c. Introduce a new column for rounded-up average response times per ward, organized in descending order.
4. Add Additional Data Points:
    a. Supplement the dataset with household income information by ward, sourced externally.

b. Extract and integrate the average household income for each ward.

5. Compare Key Data Points:
    a. Analyze the relationship between average response times and household incomes to validate the hypothesis that higher-income wards exhibit shorter response times than lower-income counterparts.

Data Preparation

Data Source:

This data has been sourced from The City of Toronto's Open Data Portal. This is an open-source delivery tool to bring updated and accessible datasets to any users for free. This specific data set includes only fire incidents as defined by the Ontario Fire Marshal (OFM) up to December 31, 2021. Sourced from: https://open.toronto.ca/dataset/fire-incidents/ (The City of Toronto and Wu)

Data Quality:

The City of Toronto's Open Data Portal provides a Data Quality score using their scorecard metrics. A chart is provided below on the breakdown and additional notes regarding the data quality. Overall, the quality is very high with few to no null values.

This dataset also provides more detail than the basic incidents dataset provides for only fire Incidents to which the Toronto Fire Service(TFS) responds to. The format is similar to the reporting data sent by TFS to the OFM.

Metric	**Feedback from Source**	**Project Notes**
Freshness	Last Refreshed on 2023-11-12	The data is current from the present month
Metadata	87% score - The following metadata field(s) are empty: ○ information_url	Only one metadata field is missing
Accessibility	75% - not accessible from searchable words	
Completeness	100% - No Issues	For this analysis, we require an additional data set on ward names and household income.
Usability	No issues	

Data Challenges:

- **Data Collection:**
    - The initial challenge concerning data collection centered on obtaining accurate and current income figures for the wards. Statistics Canada conducts a census every five years, with the most recent one completed in 2021. However, the City of Toronto website exclusively provides the 2016 census data for each ward. Consequently, our options were constrained, compelling us to rely on this 2016 dataset as the most recent information available for the average household income statistics of each ward.
    - Another challenge emerged due to the City of Toronto's utilization of varying ward profiles over time. Presently, a 25-ward model is in effect; however, before 2018, and during the collection of the 2016 census data, the city employed a 44-ward model. To align with the timeframe of the census data collection and preemptively circumvent the unavailability of data under the 25-ward model, we chose to use the 44-ward model for our analysis.
- **Handling Ward "0" Concerns:** A secondary yet significant concern arose from the presence of reports assigned to "Ward 0" in the raw fire incidents data. There is no mention of Ward 0 on the city's website, which prompted us to contact city representatives for clarification directly. We received a response indicating that Ward 0 designates incidents where the TFS apparatus responded to events outside of Toronto for mutual aid. Recognizing that these incidents fall beyond the boundaries of known wards, we decided to omit them from the dataset for data accuracy and coherence.
- Based on the decision above, only fire incidents up till 2018 are included in the dataset. Since wards 26-44 may be underreported concerning the larger dataset due to the downsizing of wards.
- After finalizing the ward model, our next task was to gather the average household income stats for each ward. The initial challenge was figuring out how to extract data from PDFs. To tackle this, we used the PDFReader method from PyPDF2 to read the text in each ward's PDF and extract the necessary information, which, in this case, was the average household income.
- Our group aimed to automate most of the data gathering and cleaning steps, allowing the Jupyter notebook to run smoothly cell by cell and produce the desired results without requiring manual pre-processing from the user. Initially, we achieved this for all steps, except one: automatically loading the names and numbers of each ward into a dataframe without typing or copying them manually.
    - We tried scraping this information from the City of Toronto's website, similar to previous steps. However, the webpage had JavaScript, which BeautifulSoup alone couldn't handle. So, we shifted our approach and used Selenium. We ran a headless browser to load the target URL with the table, then used Pandas' read_html method to load the table into a dataframe. This dataframe was then merged with our existing one to consolidate all the necessary features.

Data Analysis

As mentioned on page 4 of this report, the approach for this data set involved thorough data preprocessing to support the analysis. The following sections walk through these steps.

Data Preprocessing

1. Data Acquisition:

The code begins by downloading the "Fire Incidents Data" CSV file from the Toronto website using a specified URL. The **`requests`** library is used to fetch the data, and the file is saved in the current working directory.

2. Data Loading and Cleaning:

The dataset is loaded into a Pandas DataFrame, and a specific data type is set for the 'Incident_Number' column to address a warning about mixed types. The 'TFS_Alarm_Time' and 'TFS_Arrival_Time' columns are converted to datetime format, and a new column 'response_time' is created to represent the time difference between alarm and arrival times. Rows with 'Incident_Ward' equal to 0 are removed, and the dataset is further subsetted to the relevant columns and date range (2011-2017).

3. Data Transformation:

The 'response_time' column, represented as a timedelta, is converted to a float ('minutes') to facilitate further analysis. The mean response time for each ward is calculated and stored in a new DataFrame called **`fires_avg`**.

4. Downloading and Extracting Information from Ward PDFs:

PDFs containing ward information are downloaded, and average household income is extracted from each PDF. The extracted income values are added to the **`fires_avg`** DataFrame.

5. Data Integration:

The ward information is obtained from a second dataset, the "Ward List," using Selenium. Both datasets (**`fires_avg`** and **`ward_list`**) are then merged based on the ward number.

6. Data Export:

The final merged dataset is exported to a CSV file for further analysis.

7. Exploratory Data Analysis:

Finally, the code explores the data by sorting it based on the "Average response time (minutes)" column in ascending order.

8. Additional Information:

The code blocks also include some commented-out code for installing packages (beautifulsoup4 and pypdf2) and downloading additional PDFs to support the additional data set on household incomes by ward.

Data Outputs

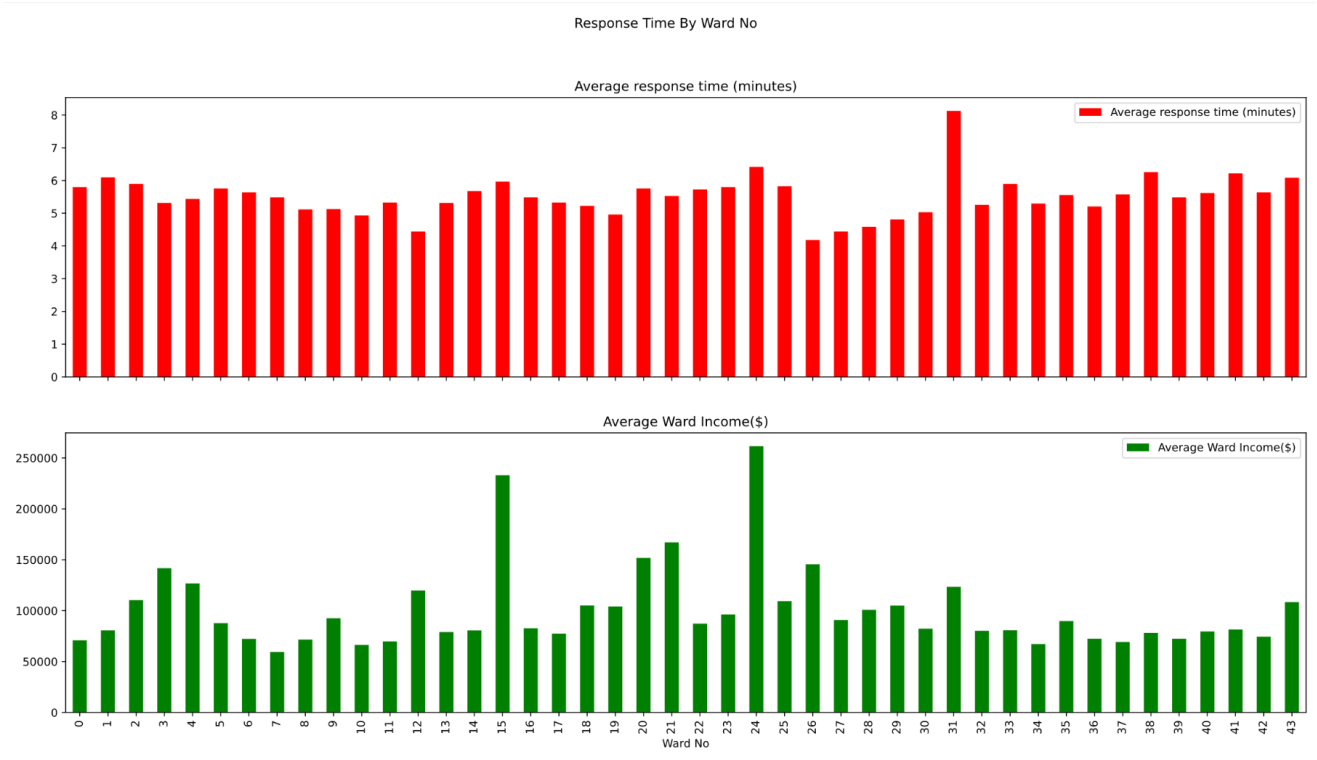
The key table to analyze lists all 44 wards, sorted by average response times and annual household incomes.

ward_no	ward_name	Average response time (minutes)	Average Ward Income($)
1	Etobicoke North	5.8	70939
2	Etobicoke North	6.1	80685
3	Etobicoke Centre	5.9	110440
4	Etobicoke Centre	5.32	141783
5	Etobicoke-Lakeshore	5.44	126802
6	Etobicoke-Lakeshore	5.76	87794
7	York West	5.64	72300
8	York West	5.49	59421
9	York Centre	5.12	71681
10	York Centre	5.13	92587
11	York South-Weston	4.93	66447
12	York South-Weston	5.33	69783
13	Parkdale-High Park	4.44	119822
14	Parkdale-High Park	5.32	79073
15	Eglinton-Lawrence	5.68	80644
16	Eglinton-Lawrence	5.97	233103
17	Davenport	5.49	82761
18	Davenport	5.33	77510
19	Trinity-Spadina	5.23	105144
20	Trinity-Spadina	4.96	104119
21	St. Paul's	5.76	151848
22	St. Paul's	5.53	167159
23	Willowdale	5.73	87263
24	Willowdale	5.8	96321
25	Don Valley West	6.42	261595
26	Don Valley West	5.83	109375
27	Toronto Centre-Rosedale	4.18	145614
28	Toronto Centre-Rosedale	4.44	90911
29	Toronto-Danforth	4.58	100840
30	Toronto-Danforth	4.81	105059
31	Beaches-East York	5.03	82373
32	Beaches-East York	8.13	123511
33	Don Valley East	5.26	80276

34	Don Valley East	5.9	80779
35	Scarborough Southwest	5.3	67193
36	Scarborough Southwest	5.56	89840
37	Scarborough Centre	5.21	72415
38	Scarborough Centre	5.58	69301
39	Scarborough Agincourt	6.26	78138
40	Scarborough Agincourt	5.49	72405
41	Scarborough-Rouge River	5.62	79646
42	Scarborough-Rouge River	6.22	81529
43	Scarborough East	5.64	74537
44	Scarborough East	6.09	108436

Data Visualization:

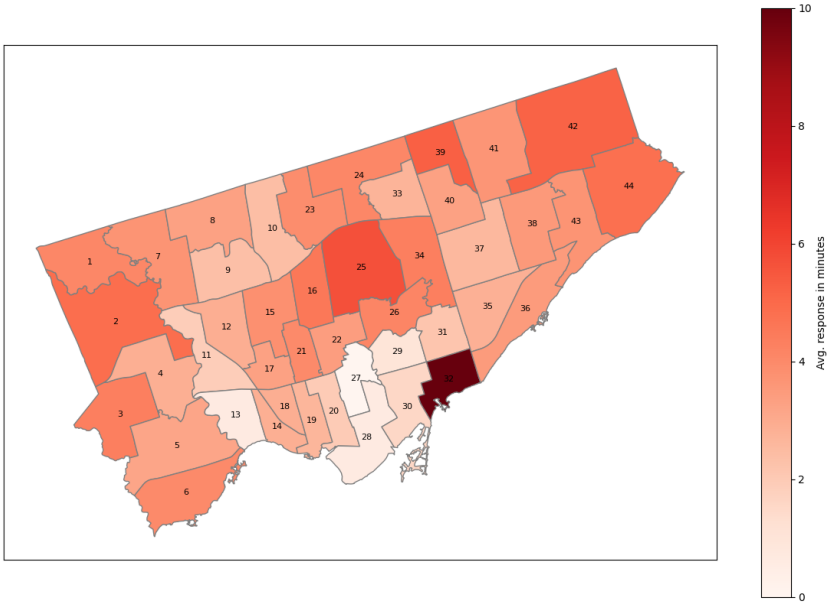
1. The average response time across all wards is approximately 5.52 minutes.
- The average income per ward varies significantly, with a mean of around $100,209 and a standard deviation of about $40,920, indicating substantial variation in income levels across wards.
- The histogram below shows the two variables as a visual aid.



2. The choropleth map provided depicts the average response times in minutes for Toronto's wards. The colour intensity corresponds to the length of the response

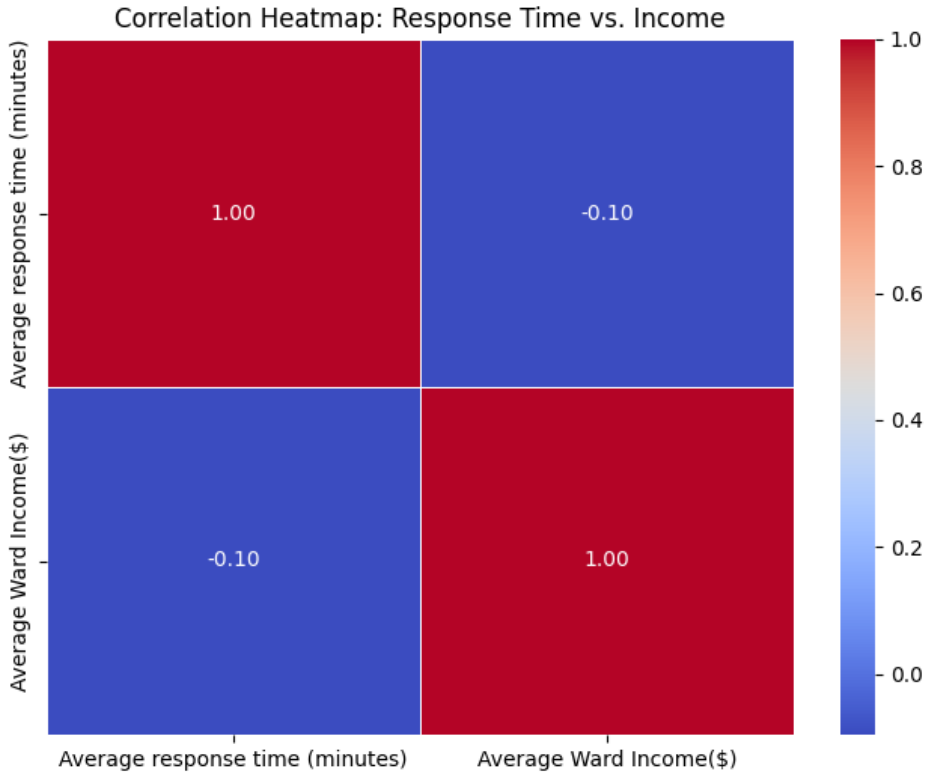
time, with darker shades indicating longer response times and lighter shades representing shorter response times. From the map, we can observe:

- **Spatial Distribution:**
    a. There is a ward in the central region with a notably darker shade, which suggests it has a longer average response time compared to surrounding areas.
    b. The central to southern regions of the map have wards with varying response times, as indicated by the mix of colour intensities.
- **Potential Urban vs. Rural Divide:**
    c. If the darker central area corresponds to a more urbanized part of Toronto, this could suggest longer response times possibly due to traffic congestion or higher incident rates. Conversely, the lighter shades might correspond to less urbanized or suburban areas where response times are potentially shorter.
- **Concentration of Longer Response Times:**
    d. There appears to be a concentration of longer response times in the central wards, whereas the outskirts are lighter, indicating shorter response times.
- **Outliers:**
    e. There are a couple of wards that stand out due to their significantly darker colour. This could be due to specific local factors affecting response times, such as ward size, road network complexity, or the location of fire stations.

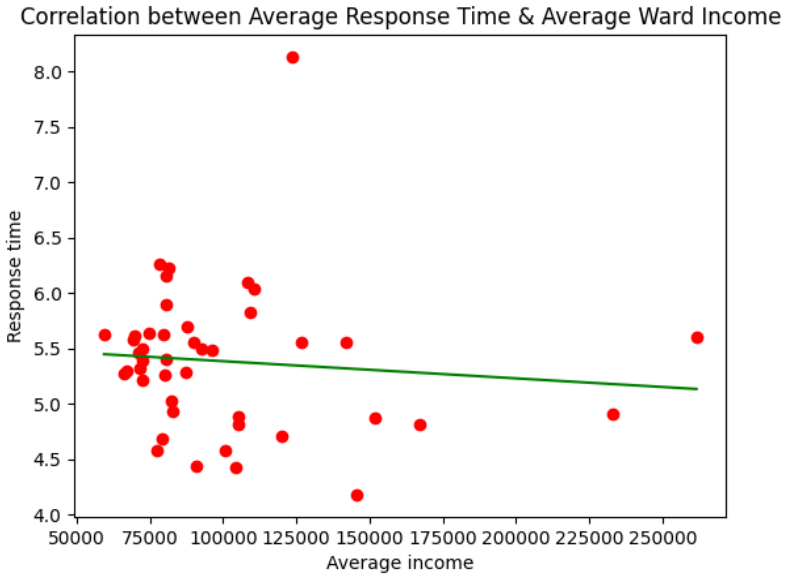


3. The correlation coefficient between 'Average response time (minutes)' and 'Average Ward Income($)' is approximately -0.10, as indicated in the heatmap below.

- This value suggests a weak negative correlation, meaning that there is a slight tendency for wards with higher incomes to have lower response times, but this relationship is not strong.
- The correlation matrix also includes the correlation of each variable with itself, which is always 1.



4. The scatterplot below illustrates the relationship between average response time (in minutes) and average ward income. From the visualization, the following observations can be made:



- **Trend:** There appears to be a negative trend indicated by the line of best fit in the plot. This suggests that as the average income increases, average response times decrease. However, the distribution of the points suggests a very weak correlation, as the points are fairly dispersed and not closely clustered around the trend line.
- **Patterns:** The data points are spread across a wide range of income levels. There is a concentration of data points around the lower to mid response time values, indicating that for most wards, the response time falls within a narrower range.
- **Outliers:** There are a few data points well above the general cluster that could be considered outliers. One in particular is an incident where an emergency fire service responded after 11 hours, which could also have been an input error. These points represent wards with significantly higher average incomes compared to others, and they do not follow the general trend of the majority of the data. They may indicate wards with particularly high property values or higher socioeconomic status.
- **Distribution:** 56% (25/44)of the wards have average response times between 5 to 6 minutes, and incomes ranging broadly from around $50,000 to $125,000. There are fewer wards with very high average incomes, and these do not show a clear trend in terms of response times.
- In summary, while there is a weak negative trend, the spread of the data points and the presence of outliers with high average incomes suggest that the relationship between response time and income is not straightforward and may be influenced by other factors not captured in this two-dimensional analysis. Additional data and a more in-depth statistical analysis would be helpful to understand the underlying factors contributing to this relationship.

Conclusion

The analysis of fire department response times and household income across wards in Toronto reveals a nuanced relationship. Despite initial hypotheses suggesting a potential correlation between income levels and response efficiency, the data indicates otherwise. The weak negative correlation observed through heatmap and scatterplot analyses is not sufficient to substantiate a definitive link between higher-income wards and quicker fire emergency responses. The conclusion underscores the complexity of factors influencing response times, suggesting that socioeconomic status alone is not a determinant.

The choropleth map underscores the geographic variation in response times across the city, with central wards showing a mix of response times, including some outliers with significantly longer times. It's important to note that while there is a general trend observed, the correlation is not strong, which implies that factors other than income might be influencing response times to a greater extent. The outliers observed in the scatterplot, particularly at higher income levels, suggest that specific local characteristics or external factors could be impacting response times.

These findings highlight the importance of considering a range of socio-economic and urban planning factors when analyzing emergency response data. Future studies could benefit from incorporating additional variables such as population density, traffic patterns, and the location of emergency services to better understand the dynamics at play.

Overall, the review of this dataset disproves our hypothesis. **There is no statistical relationship between socioeconomic status and fire services' response times.**

Works Cited

The City of Toronto, and Kevin Ku. "About Fire Incidents." *The City of Toronto*, November 2023, https://open.toronto.ca/dataset/fire-incidents/%20. Accessed 19 November 2023.