PDF Report Detailing Method

Chathurangi Godahewa Gamage, Dury Kim, Renata King, Shantikrishna Panicker

2024-09-28

Methods

1. Data Sources and Loading

The data used for this project comes from multiple sources:

- Spark Telecommunications Data: This dataset (sp_data.csv.gz) contains population counts for different SA2 codes over time, which serves as a proxy for population density.
- Vodafone Telecommunications Data: Similar to the Spark data, this dataset (vf_data.parquet) contains population counts for various SA2 codes, but in Parquet format.
- SA2 Concordance Files: The concordance data (sa2_2023.csv) provides mappings from SA2 codes to region names, which allows us to enrich the telecommunications data with geographical context.
- Urban/Rural Indicators: Two files (urban_rural_to_indicator_2023.csv and urban_rural_to_sa2_concord_2023.csv) provide urban/rural classifications, which help distinguish between different types of areas (e.g., urban, rural, small settlements).

The data was loaded using read_csv() for CSV files and read_parquet() for Parquet files. For some datasets, the first few rows were skipped to avoid metadata and focus only on the relevant data.

2. Data Cleaning

Spark Data The Spark data required the following cleaning steps:

- Timestamp Conversion: The ts column, representing the timestamp, was converted from character to datetime format using the as_datetime() function.
- Handling Missing Values: The cnt column, representing the population count, had some missing values. These were replaced with 0 using replace_na(), assuming that missing data implies no recorded population.

Vodafone Data The Vodafone data underwent similar cleaning:

- Timestamp Conversion: The dt column was converted to a proper datetime format.
- Handling Missing Values: The devices column, representing population count, had missing values replaced with 0.

3. Merging with SA2 Concordance Data

After cleaning the Spark and Vodafone datasets, the next step was to enrich them with geographical information from the SA2 Concordance file.

- Spark Data: The cleaned Spark data was merged with the SA2 concordance data on the SA2 code column. This added region names to each SA2 code.
- Vodafone Data: Similarly, the Vodafone data was merged with the same concordance file using the left_join() function in R.

To ensure consistency during the merge, the SA2 code column in both datasets was converted to the same data type (numeric or character as required).

4. Merging with Urban/Rural Indicators

The urban/rural indicator datasets were also cleaned before merging:

- Splitting Columns: Certain columns, such as UR2018 V1.0.0, contained multiple values separated by commas. These were split into separate columns using separate(), providing the urban/rural classification and settlement type.
- Converting Data Types: The UR2023_code and UR2018_code columns were converted to character to ensure compatibility during the join process.

Once the urban/rural indicators were cleaned, they were merged with the already merged Spark and Vodafone datasets using the SA2 code.

5. Aggregating Population Data

The cleaned and merged Spark and Vodafone data were aggregated to compute total population counts for each SA2 code at each timestamp:

- Spark Data Aggregation: The Spark data was grouped by SA2 code and timestamp, and the total population count for each group was computed.
- Vodafone Data Aggregation: The Vodafone data was aggregated similarly, grouped by SA2 code and timestamp.

After both datasets were aggregated, they were combined using full_join(), allowing us to include all available data from both Spark and Vodafone.

6. Handling Missing Data After Merge

Once the data from Spark and Vodafone was combined, some SA2 codes had missing values from one of the sources. To handle this:

- Replacing Missing Values: Missing counts from either Spark or Vodafone were replaced with 0 to ensure completeness.
- Total Population Count: A new column was created that summed the Spark and Vodafone counts to produce the total population count for each SA2 code at each timestamp.

7. Final Output

The final dataset contains the following columns:

- SA2 Code: The Statistical Area Level 2 (SA2) code for each region.
- Timestamp: The datetime for each observation, converted to NZST.
- Total Population Count: The combined population count from both Spark and Vodafone datasets.
- Region Names and Urban/Rural Classification: Added geographical context, including the name of each SA2 area and whether it is classified as urban or rural.

Finally, the data was exported as a gzipped CSV file using the write_csv() function, producing the file final_combined_data_with_regions.csv.gz.

Charts

#134800 auckland university area code #326600 christchurch central area code #251400 wellington central area code

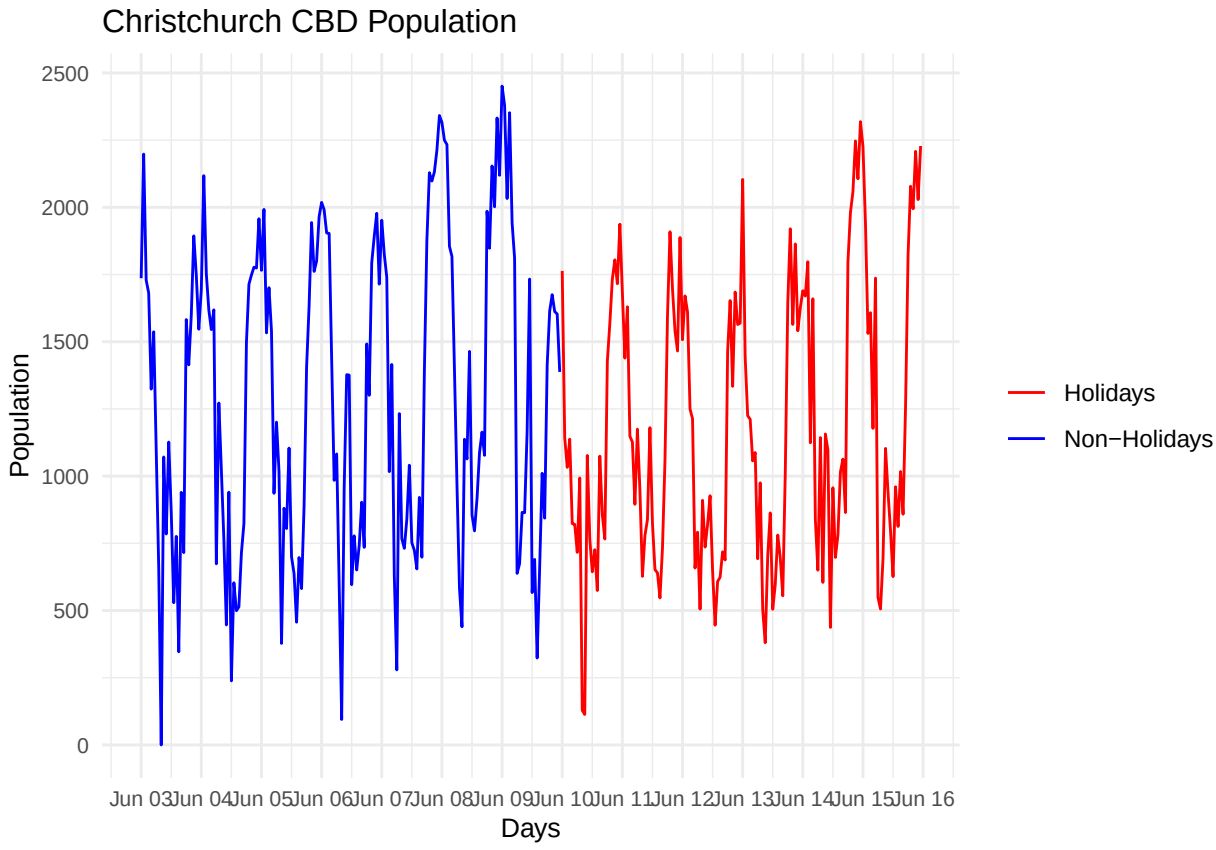
#start holidays 2024-06-10 #end holidays 2024-6-16

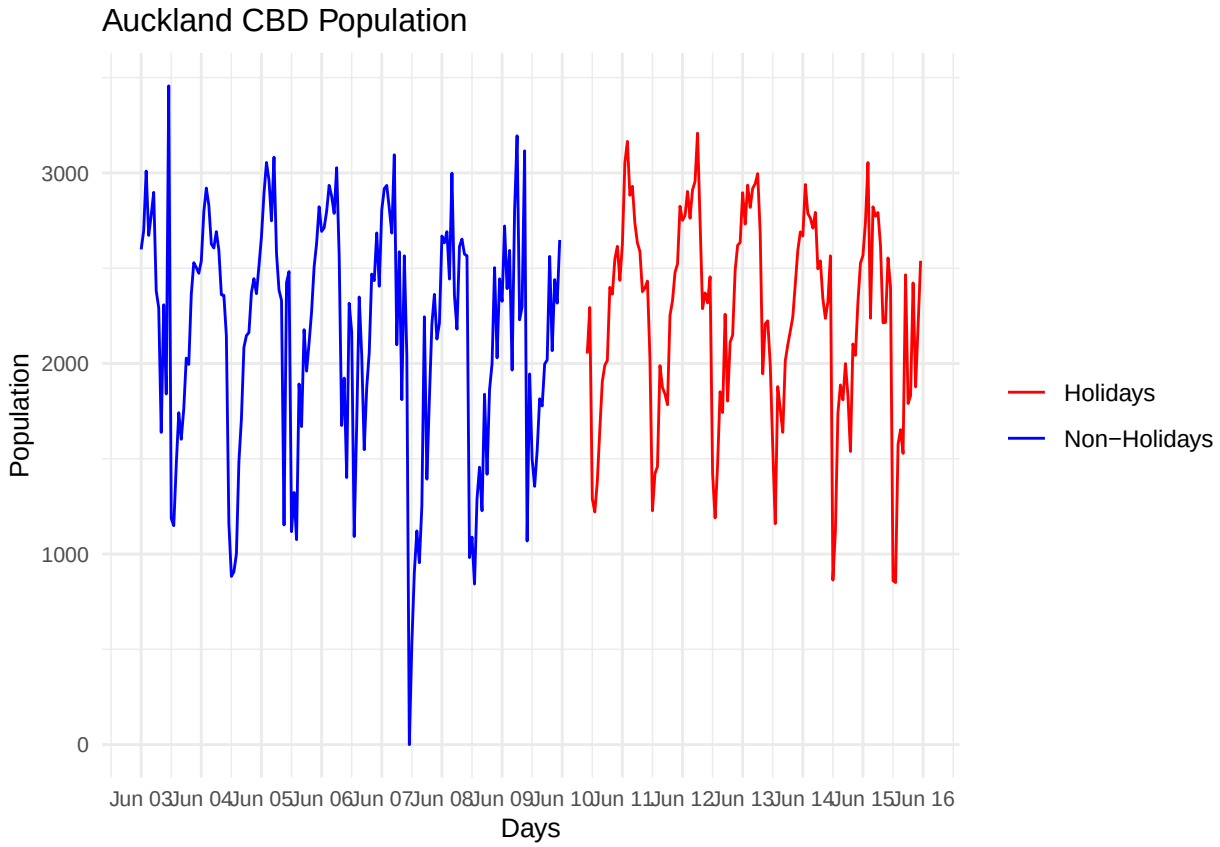
```
# Load the gzipped CSV file
final_cleaned_data <- read_csv("final_cleaned_data.csv.gz")
```

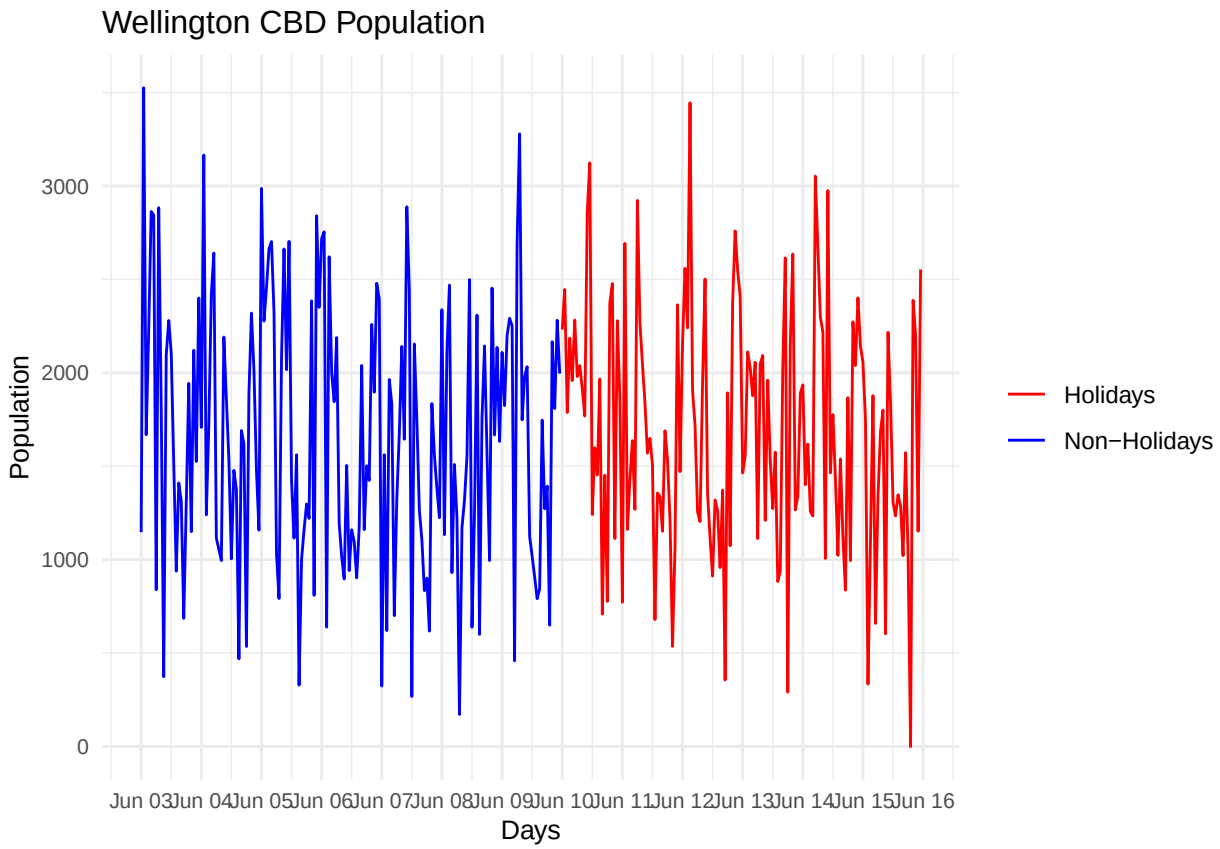
```
## Rows: 945840 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (6): sa2, count.x, count.y, total_count, UR2023_code, UR2018_code
## dttm (1): ts
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Preview the first few rows to check that the data was loaded correctly
head(final_cleaned_data)
```

```
## # A tibble: 6 x 7
##      sa2 ts                  count.x count.y total_count UR2023_code UR2018_code
##    <dbl> <dttm>                <dbl>   <dbl>       <dbl>        <dbl>        <dbl>
## 1 100100 2024-06-02 12:00:00    793.    340.       1133.         1001           21
## 2 100100 2024-06-02 12:00:00    793.    340.       1133.         1013           22
## 3 100100 2024-06-02 13:00:00    742.    318.       1059.         1001           21
## 4 100100 2024-06-02 13:00:00    742.    318.       1059.         1013           22
## 5 100100 2024-06-02 14:00:00   1233.    528.       1761.         1001           21
## 6 100100 2024-06-02 14:00:00   1233.    528.       1761.         1013           22
```




Auckland CBD Population
Population
3000
2000
1000
0
Jun 03
Jun 04
Jun 05
Jun 06
Jun 07
Jun 08
Jun 09
Jun 10
Jun 11
Jun 12
Jun 13
Jun 14
Jun 15
Jun 16
Days
Holidays
Non−Holidays


Wellington CBD Population
Population
3000
2000
1000
0
Jun 03
Jun 04
Jun 05
Jun 06
Jun 07
Jun 08
Jun 09
Jun 10
Jun 11
Jun 12
Jun 13
Jun 14
Jun 15
Jun 16
Days
Holidays
Non−Holidays

Limitations of Data

1. Incomplete or Missing Data

- While the Spark, Vodafone and SA2 population estimates provide substantial information, they do have limitations regarding data completeness. Missing values in key columns ("cnt" from Spark and "devices" from Vodafone) can affect the reliability of the analysis as a whole.
- For instance, if certain time intervals lack data this can skew trends and give inaccurate conclusions.

2. Temporal Resolution

- The data collected is based on specific time stamps, which might not capture the differences of population behavior or device usage patterns accurately. Having hourly aggregation intervals can overlook important fluctuations in usage, mainly during peak traffic times.

3. Data Source Reliability

- The datasets come from different sources, each with their own methodologies for data collection and processing. For example, the spark datas focused on population counts while Vodafone reflects device usage. Differences in data collection standard can lead to inconsistencies and complicate cross-data comparisons.

4. Unknown Bias:

- Vodafone data could unknowingly underrepresent areas with low mobile coverage or populations that are less likely to use mobile devices, like elderly or economically disadvantaged groups.
- Similarly Spark data might be influenced by seasonal fluctuations such as public holidays.

5. Data Mismatch

- Trying to merge data from different sources involves assumptions that the concordance is accurate and complete. But any inaccuracies or missing mappings between different data sets can skew population estimates in certain regions or time periods.
- When splitting and renaming columns in data sets there is a risk that some data may not align perfectly, leading to potentially incorrectly categorized rows.

6. Geographical Scope

- This project is purely focused on CBDs of Auckland, Wellington and Christchurch. But these data sets include data for much larger geographical locations.
- Filtering and aggregating the data to the CBD level introduces potential for misclassification or exclusion of relevant areas.

7. Uncertainty in GS Mapping

- The use of SA2-level population data introduces a limitation in precision. Population densities may vary significantly within SA2 regions. Especially in CBD areas with high population turnover. SA2 boundaries might not capture this fine grained line effectively.

Future Improvements

- Enhanced Day-Specific Recommendations: Expand the analysis to incorporate suggestions related to particular days of the week, such as weekdays compared to weekends. Identify which days have the lowest population densities, making them ideal for roadwork activities. This strategy directly addresses the key question of the best days for daytime roadwork, offering practical recommendations that can assist Fulton Hogan in reducing disruptions.

- Geospatial Visualization: Incorporate visual tools like heatmaps or geographic maps to illustrate variations in population density across various regions and time periods. The visualization of geospatial data provides valuable insights into regional patterns and demographic shifts, making it easier for non-technical stakeholders to quickly understand the information.
- Detailed Limitation Section: Include a detailed section that addresses the limitations of telecommunications data. Explore possible biases, inaccuracies, or gaps in the data, including variations in network coverage and challenges related to anonymization. A thorough analysis of data limitations promotes transparency and enables future users of the dataset or analysis to gain a clearer understanding and context for the findings.