Methods

Main Pipeline

00_load.R

Load in all required libraries for main pipeline.

Read in raw data.

01_clean.R

Run previous step in pipeline.

Reformat date time data so both Spark and Vodafone match.
Add the variable "data_from" to ensure merging is done correctly.
Remove duplicate entries.
Add 12 hours to each row of Spark data (explained in time_shift_justification.R).

Clean variable names for each data frame.

02_merge.R

Run previous step in pipeline.

Combine telecommunications and location data.
Combine further into a cleaned_data set. Reformat date time data (issues with formats when testing).

Created clean dataset:

Give all variables proper names.
Combine device counts for Spark and Vodafone data. Remove NA values.
Use formula (from population_linear_model) to predict people_count from device_count.

Commented section for including only observations taken during the day.
7AM to 6PM chosen as data was taken during daylight savings time.

03_analysis.R

Run previous step in pipeline.

Create new variable, which is the current device count minus the device count of the previous hour.
This models the change in the amount of people between hours.

04_vizualize.R

Run previous step in pipeline.

Add new variable to each row that shows the day of the week, and which hour the row was recorded in.
Aggregate data by day/hour.
Plot the absolute difference in the amount of people for each day/hour.

Additional Scripts

export_data.R

Export clean data set to be given to Fulton Hogan data scientists.

population_liner_model.R

Get population estimates for each area (from pop_estimates) and tidy variable names.
Combine to match population estimates with the name of area.
Add new variable which takes the device count in each area.
5AM on Tuesday, 11th of June chosen to model device counts, as Tuesday was shown to have the lowest change in device count, and 5AM chosen for similar reasons.
Fit linear model to predict population using device count.
Intercept of zero chosen, as a negative population doesn't make sense.
Summary and plot of data.

required_packages.R

Installs required packages to run scripts.

time_shift_justification.R

In the raw data, the Spark data starts starts 12 hours earlier than the Vodafone data.
This presents multiple issues:
The first and final days only having 12 hours worth of device counts.
The data starting and ending in the afternoon instead of midnight.
The data between mobile carriers being out of sync.

We decided mutating the Spark data was the best approach, as the Vodafone data started at midnight, and encapsulated exactly two weeks.
To determine whether we would remove the first and last 12 hours or shift every observation foward by 12 hours, we created these plots.

As the plots show, when shifting time forwards, the relative counts for both carriers matched up perfectly, and when removing rows they were out of sync.
This is how we justified moving on by shifting the time for each row of the spark data forwards by 12 hours.

auckland_cbd.R

Runs the main pipeline, but filters data to only include areas within the Auckland CBD.

christchurch_cbd.R

Runs the main pipeline, but filters data to only include areas within the Christchurch CBD.

wellingotn_cbd.R

Runs the main pipeline, but filters data to only include areas within the Wellington CBD.

total_cbd.R

Runs the main pipeline, but filters data to only include areas within the CBDs for each city.

Limitations

With this data, we have a device count for each area each hour, and we modeled movement by taking the difference in device count, with a lower device count meaning people were leaving, and a higher device count meaning people were coming in. With this data, we could not account for people both leaving and coming into an area in the same hour. For example, is 4,000 devices were leaving Christchurch Central, and 4,100 devices were coming into Christchurch Central in the same hour, we would only see a change of +100 devices, even though there were 8,000 other devices moving between the area.

Due to this, areas which experience similar levels of people both entering and leaving would be under counted. For our analysis, we had to assume this affected all CBD areas equally, and affected each day equally, so the results could still produce useful insight.

Another limitation when using device counts is cell towers going down. When inspecting the data we observed areas with a device count of 0, when the previous and next hour had a device count greater than 0. This is likely due to cell towers malfunctioning, or reporting the device count wrong. This disproportionately affected CBD areas, as they tended to have higher device counts leading to higher difference counts when the device count was not recorded properly.

This is why we decided to remove observations with device counts of zero in CBD areas. This is a limitation, as we do not know exactly what caused the device count to be miscounted, which could affect the patterns we saw in the plots. For example, if cell towers were down for maintenance, and they were always scheduled for a Tuesday, we would see significantly lower changes for Tuesdays. Luckily we only had two weeks worth of data, but this effect would be magnified for longer periods of time, so a different approach should be considered for larger data sets.

Future Improvements