Where are the Crimes?

STATS 765 - Milestone 2

Sujay Anjankar - 2433337734

Contents

1 Goal 3

2 Data Source 3

3 Data Processing 3
- 3.1 Missing / Redacted Values . . . 3
- 3.2 Date Range . . . 3
- 3.3 Reducing Dimensionality . . . 3
- 3.4 Additional Processing . . . 4

4 Data Exploration 4
- 4.1 Victimization Data . . . 4
  - 4.1.1 Distribution of Committed Offenses . . . 4
  - 4.1.2 Criminal Activity Hotspots . . . 5
- 4.2 Census . . . 5
  - 4.2.1 Victimization Rate / 1000 Households for Large SA2s . . . 5

5 Analytical Plan 5

6 Appendix 6
- 6.1 Data Loading . . . 6
- 6.2 Data Processing . . . 6
  - 6.2.1 Single Value Columns . . . 6
  - 6.2.2 Restricting to 2023 . . . 7
  - 6.2.3 Purge Extra Columns . . . 7
  - 6.2.4 Prepare Final Dataframe . . . 8
- 6.3 Visualization: Victimization Data . . . 9
  - 6.3.1 Distribution of Committed Offenses . . . 9
  - 6.3.2 Criminal Activity Hotspots . . . 9
- 6.4 Visualization: Census - Victimization Correlation . . . 10
  - 6.4.1 Overall Crime Rate / 1000 Households in Large Areas . . . 11
  - 6.4.2 Median Household Income vs Total Victimizations by Neighborhood . . . 12
  - 6.4.3 Vehicle Ownership (Multiple) vs Total Victimizations . . . 13
  - 6.4.4 Communication Access vs Total Crimes by Neighborhood . . . 14
  - 6.4.5 Weekly Rental vs Total Victimizations . . . 15
  - 6.4.6 Subdivision vs Family Type . . . 17

1 Goal

This project aims to establish a relationship between the types and rates of crimes based on the features of neighborhoods from the census data. This report specifically aims at doing so for different characteristics of households.

2 Data Source

The analysis is performed using the following three data sources.

Dataset	Detail
Census	Household attributes (income, rent, # of vehicles)
Meshblock	Mapping table for meshblock to SA2 lookup
Victimisations	Victimisation data with meshblock and timestamp

These datasets are at the meshblock level,[1] but we have aggregated them up to statistical area - 2 level.[2] Statistical Area 2 usually have a population of 2000 - 4000, which is why the analysis was performed at this level. Meshblocks offer a much finer granularity.

The data loading is described in Appendix 5.1.

3 Data Processing

The two major parts of the data processing process are outlined here.

3.1 Missing / Redacted Values

According to StatsNZ, the data made available by them may contain these two special values: `-997` means "Data not collected" and `-999` means "Suppressed due to confidentiality".

The values are essentially `NA`, so when the files are loaded in Appendix 5.1, they are treated as `NA`.

3.2 Date Range

The Census dataset contains household and dwelling data for the years 2013, 2018 and 2023. The Victimizations dataset spans from 2021 to 2025. The only overlap here is 2023, so it was decided to limit the scope of analysis to the year 2023 to ensure completeness.

Victimizations data is a long table, so only records for the year 2023 were retained, and the rest were dropped. But the Census data is a wide table, with columns for a particular feature for every year. The columns look like this, for example, `Rental - 2023 - Weekly Rent`, `Income - 2018 - Median Household Income`. Due to this, columns *not* containing the year `2023` in their name were dropped from the dataframe. This is demonstrated in Appendix 5.2.2.

3.3 Reducing Dimensionality

Dropping the 2023 columns from the Census dataframe already reduces the dimensionality for that dataset significantly, almost takes it down to a third of what it was originally.

The next step for reducing dimensionality was determining columns with low predictive power. A simplified definition for such columns in this context was considering columns which only have one values. Such columns were identified on all datasets and then dropped. This is demonstrated in Appendix 5.2.1.

Lastly, looking at the reduced column list of all the dataframes, columns which were not pertinent to the analysis were identified. The purpose of this analysis is to establish a relationship between household attributes to the regions where victimizations occurred. For instance, the size of the area in square kilometers, the day/time of

[1] https://en.wikipedia.org/wiki/Meshblock

[2] https://en.wikipedia.org/wiki/List_of_statistical_areas_in_New_Zealand

the occurrence are not related to the location of the victimization, so they were dropped from the dataframe. A significant such reduction is seen in the Meshblock dataset, since other than the meshblock code, and the code and name of the statistical area 2, all other columns were irrelevant and were dropped.

3.4 Additional Processing

For increasing readability and ease of execution, the following activities were also performed in Appendix 5.2.3. 1. Census: The column names were excessively long, they were shortened by regular expression based renaming. 2. All datasets: The columns containing meshblock ID, and statistical area 2 code & name were renamed to `meshblock`, `sa2_code` and `sa2_name` for simplifying joins.

Final dimensions of the dataframes after all this processing are as described (changed dimensions in **bold**) here.

Dataset	Original Dimensions	Reduced Dimensions
Census	2395 × 269	2395 × **88**
Meshblock	57539 × 55	57539 × **3**
Victimizations	1133581 × 17	**302968** × **9**

Finally these dataframes were joined in Appendix 5.2.4, resulting in a dataframe with dimensions $\boxed{\mathbf{204912 \times 97}}$.

4 Data Exploration

Exploring this data is separated into two parts, exploring the victimizations data independently, and then attempting to identify Census variables that may potentially explain victimization numbers.

4.1 Victimization Data

4.1.1 Distribution of Committed Offenses

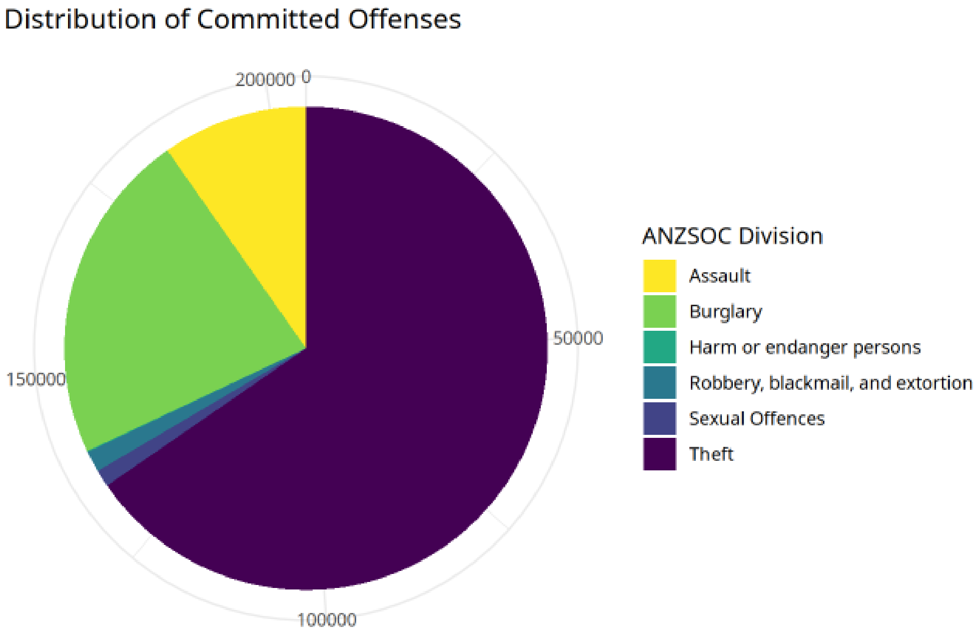


Figure 1: Distribution of Committed Offenses

4.1.2 Criminal Activity Hotspots

Statistical Area 2 Name	Number of Victimizations
Palmerston North Central	1939
Rotorua Central	1791
Hamilton Central	1612
Te Rapa South	1560
Hutt Central North	1342
Napier Central	1240
Sylvia Park	1138
Whanganui Central	1116
Tauranga Central	1104
Hastings Central	1072

4.2 Census

4.2.1 Victimization Rate / 1000 Households for Large SA2s

SA2 Name	Victimizations	Households	CR / 1000 HH
Te Rapa South	1909	168480	11.330722
Napier Central	1403	204600	6.857283
Rotorua Central	2278	338499	6.729710
Hastings Central	1170	189744	6.166203
Gisborne Central	1096	216216	5.069005
New Lynn Central	846	183183	4.618333
Nelson Central-Trafalgar	883	202272	4.365409
Masterton Central	737	184266	3.999653
Frankton Junction	699	196098	3.564544
Whanganui Central	1204	348192	3.457862

5 Analytical Plan

6 Appendix

6.1 Data Loading

```r
df_victimizations <- read_csv(
    paste(path, "/victimizations-data.csv", sep = ""),
    na = c("", "NA"), show_col_types = FALSE
)

df_census <- read_csv(
    paste(path, "/2023_Census_totals_by_topic_for_households_by_SA2.csv", sep = ""),
    # Read the special values as NA for simplicity.
    na = c("", "NA", -999, -997), show_col_types = FALSE
)

df_meshblock <- read_csv(
    paste(path, "/meshblock-higher-geographies-2023-generalized.csv", sep = ""),
    na = c("", "NA"), show_col_types = FALSE
)
dim(df_census)
```

```
## [1] 2395  269
```

```r
dim(df_meshblock)
```

```
## [1] 57539    55
```

```r
dim(df_victimizations)
```

```
## [1] 1133581      17
```

6.2 Data Processing

6.2.1 Single Value Columns

```r
dfs <- list(
    victimizations = df_victimizations,
    census = df_census,
    meshblock = df_meshblock
)

single_value_column_names <- lapply(dfs, function(df) {
    df |>
        summarize(across(everything(), ~ n_distinct(.))) |>
        select(where(~ .x == 1)) |>
        names()
})

df_victimizations <- df_victimizations |>
    select(-all_of(single_value_column_names$victimizations))

df_census <- df_census |>
    select(-all_of(single_value_column_names$census))

df_meshblock <- df_meshblock |>
    select(-all_of(single_value_column_names$meshblock))
```

```r
dim(df_census)
```

```
## [1] 2395  261
```

```
dim(df_meshblock)
```

```
## [1] 57539    55
```

```
dim(df_victimizations)
```

```
## [1] 1133581      15
```

6.2.2 Restricting to 2023

```
df_victimizations |>
    distinct(year = str_sub(`Year Month`, start = -4L, end = -1L))
```

```
## # A tibble: 5 x 1
##   year
##   <chr>
## 1 2021
## 2 2022
## 3 2023
## 4 2024
## 5 2025
```

```
df_victimizations <- df_victimizations |>
    filter(str_sub(`Year Month`, start = -4L, end = -1L) == "2023")

df_census <- df_census |>
    select(!matches("2013|2018"))
```

6.2.3 Purge Extra Columns

```
df_victimizations <- df_victimizations |>
    select(-c(
        # Contain the same values as `Year Month`
        `Year Month (copy 2)`, `Month Year`,
        # `Not relevant to analysis
        `Occurrence Day Of Week`, `Occurrence Hour Of Day`, `Territorial Authority`,
        `Location Type`
    )) |>
    rename(meshblock = Meshblock)

df_meshblock <- df_meshblock |>
    # This dataset is to just join the two datasets, so drop any excess columns.
    select(c(MB2023_V1_00, SA22023_V1_00, SA22023_V1_00_NAME)) |>
    # Converted the string meshblock to a numeric value before joining.
    mutate(meshblock = as.numeric(MB2023_V1_00)) |>
    # Rename it to meshblock for ease of joining.
    rename(
        sa2_code = SA22023_V1_00,
        sa2_name = SA22023_V1_00_NAME
    ) |>
    select(-MB2023_V1_00)

df_census <- df_census |>
    # Rename overly long columns.
    rename_with(
        ~ .x |>
```

```
            str_replace(
                "Subject pop: Households in rented occupied private dwellings, Year: ",
                "Rentals - "
            ) |>
            str_replace(
                "Subject pop: Households in occupied private dwellings, Year: ",
                "Households - "
            ) |>
            str_replace(", Measure: Count, Var1:", " -") |>
            str_replace(", Measure: Median, Var1: ", " - Median: ") |>
            str_replace(", Measure: Mean, Var1: ", " - Median: ") |>
            str_replace(" paid by household", "") |>
            str_replace("Total household income", "Income") |>
            str_replace("Sector of landlord", "Landlord") |>
            str_replace("Tenure of household", "Tenure") |>
            str_replace("Number of usual residents in household", "# of Residents") |>
            str_replace("Household crowding index", "Crowding Index") |>
            str_replace("Household composition", "Composition") |>
            str_replace("Access to telecommunication systems", "Telecom Access") |>
            str_replace("Number of motor vehicles", "Vehicle Count") |>
            str_replace(
                "Count of households in occupied private dwellings",
                "Household Count"
            )
    ) |>
    # These columns are not required to the analysis.
    select(-c(
        OBJECTID,
        `Statistical area 2 (SA2) 2023 name no macrons`,
        `Area square kilometres`,
        `Land area square kilometres`,
        `Shape__Area`,
        `Shape__Length`
    )) |>
    rename(
        sa2_code = `Statistical area 2 (SA2) 2023 code`,
        sa2_name = `Statistical area 2 (SA2) 2023 name`
    ) |>
    # Since it is all 2023 data, remove the year from column names.
    rename_with(~ .x |> str_replace(" - 2023", ""))
```

6.2.4 Prepare Final Dataframe

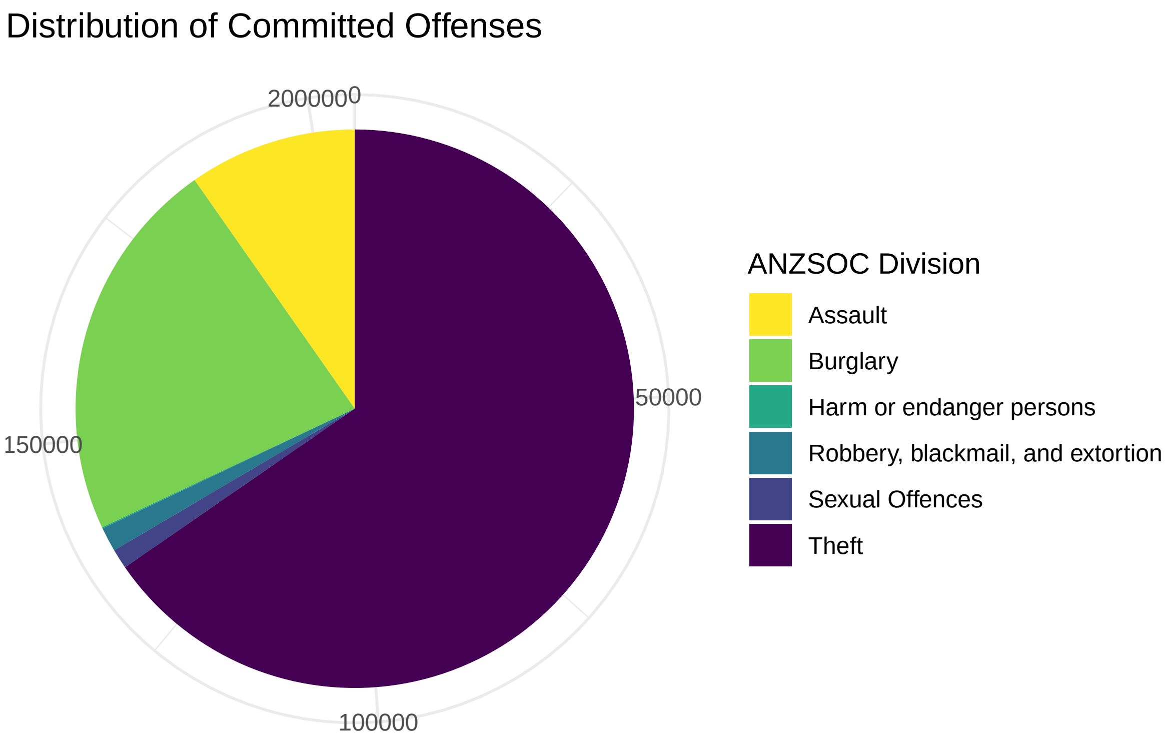
```
df_merge <- inner_join(
    df_census,
    inner_join(
        df_victimizations, df_meshblock,
        by = "meshblock"
    ),
    by = c("sa2_code", "sa2_name")
)
dim(df_merge)
```

```
## [1] 204912     97
```

6.3 Visualization: Victimization Data

6.3.1 Distribution of Committed Offenses

```
# Most committed offenses
df_merge |>
    group_by(`ANZSOC Division`) |>
    summarize(count_by_subdivision = n(), .groups = "drop") |>
    ggplot(aes(x = "", y = count_by_subdivision, fill = `ANZSOC Division`)) +
    geom_bar(stat = "identity", width = 1) +
    scale_fill_viridis_d(direction = -1) +
    coord_polar("y", start = 0) +
    labs(
        title = "Distribution of Committed Offenses",
        x = "",
        y = ""
    ) +
    theme_minimal()
```



6.3.2 Criminal Activity Hotspots

```
# Areas with a high prevalence of crime
knitr::kable(
    df_merge |>
        group_by(sa2_name) |>
        summarize(count_by_sa2 = n(), .groups = "drop") |>
        slice_max(n = 10, order_by = count_by_sa2),
    col.names = c("Statistical Area 2 Name", "Number of Victimizations")
)
```

Statistical Area 2 Name	Number of Victimizations
Palmerston North Central	1939
Rotorua Central	1791
Hamilton Central	1612
Te Rapa South	1560
Hutt Central North	1342
Napier Central	1240
Sylvia Park	1138
Whanganui Central	1116
Tauranga Central	1104
Hastings Central	1072

6.4 Visualization: Census - Victimization Correlation

6.4.1 Overall Crime Rate / 1000 Households in Large Areas

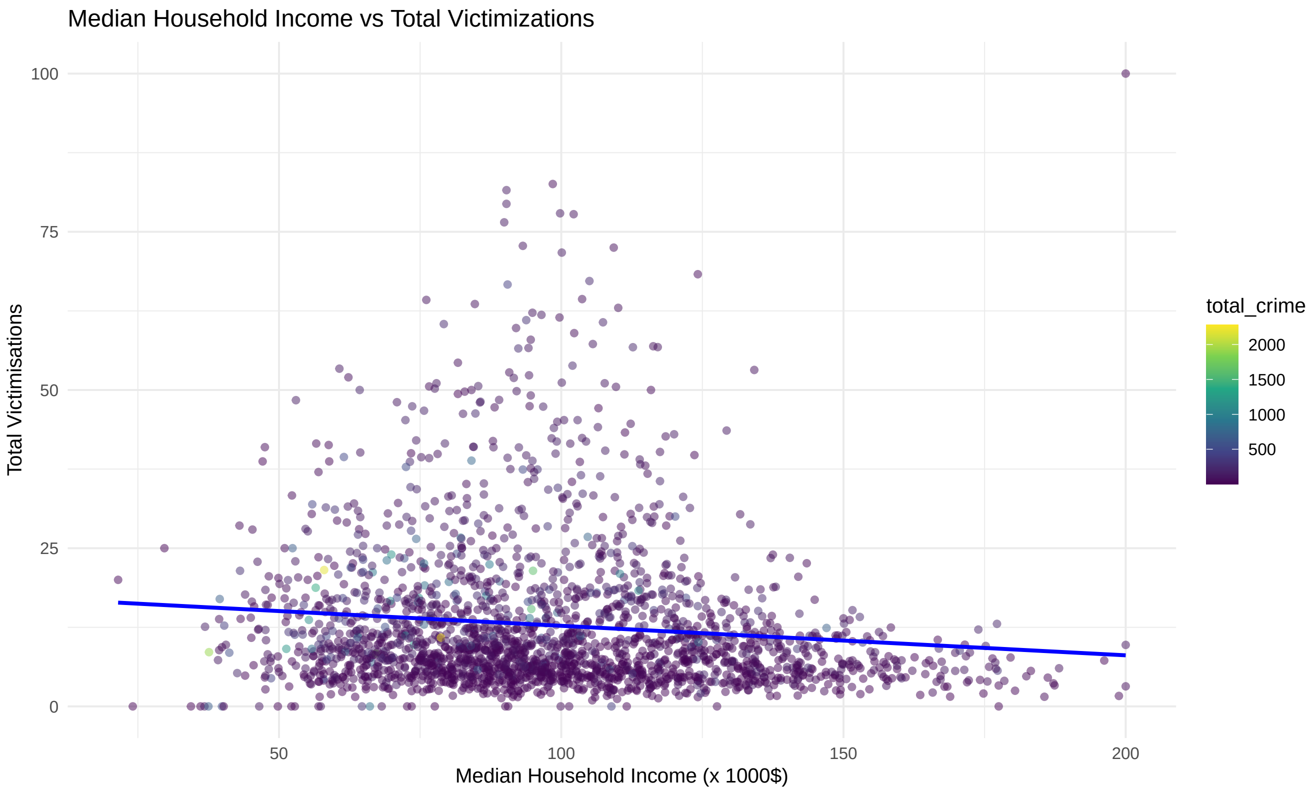
```r
total_number_of_households <- sum(df_merge$`Households - Household Count (Total)`)
# Consider a 0.1% of the total household count as the threshold to identify large SA2s.
threshold <- total_number_of_households * 0.001

knitr::kable(
    df_merge |>
        # Filter all NAs
        filter(
            !is.na(Victimisations),
            !is.na(`Households - Household Count (Total)`)
        ) |>
        group_by(sa2_code, sa2_name) |>
        summarize(
            total_victimisations = sum(Victimisations),
            households = sum(`Households - Household Count (Total)`),
            .groups = "drop"
        ) |>
        filter(households > threshold) |>
        mutate(
            # Crime rate per 1,000 households
            crime_rate_per_1000hh = (total_victimisations / households) * 1000
        ) |>
        select(-sa2_code) |>
        arrange(desc(crime_rate_per_1000hh)) |>
        head(n = 10),
    col.names = c("SA2 Name", "Victimizations", "Households", "CR / 1000 HH")
)
```

SA2 Name	Victimizations	Households	CR / 1000 HH
Te Rapa South	1909	168480	11.330722
Napier Central	1403	204600	6.857283
Rotorua Central	2278	338499	6.729710
Hastings Central	1170	189744	6.166203
Gisborne Central	1096	216216	5.069005
New Lynn Central	846	183183	4.618333
Nelson Central-Trafalgar	883	202272	4.365409
Masterton Central	737	184266	3.999653
Frankton Junction	699	196098	3.564544
Whanganui Central	1204	348192	3.457862

6.4.2 Median Household Income vs Total Victimizations by Neighborhood

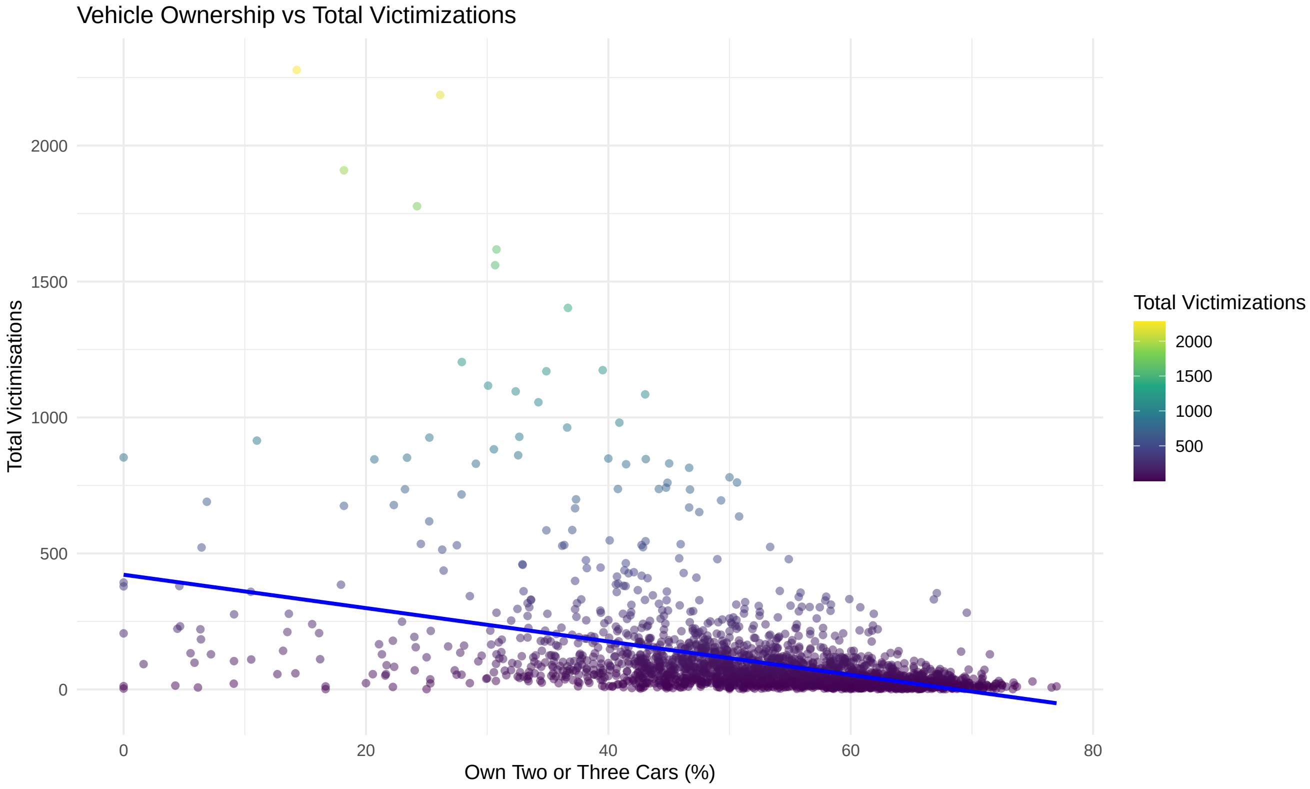
Observation: Number of victimizations is observed to trend downwards with an increase in median household income.

```r
df_merge |>
    group_by(sa2_name) |>
    summarize(
        median_income = median(`Households - Median: Income (Median ($))`, na.rm = TRUE),
        crowding_rate = (
            sum(`Households - Crowding Index (Crowded)`, na.rm = TRUE) +
                sum(`Households - Crowding Index (Two or more bedrooms needed (severely crowded))`, na.rm
                sum(`Households - Crowding Index (One bedroom needed (crowded))`, na.rm = TRUE)
        ) * 100 / sum(`Households - Crowding Index (Total stated)`, na.rm = TRUE),
        total_crime = sum(Victimisations, na.rm = TRUE)
    ) |>
    ggplot(aes(
        x = median_income / 1000,
        y = crowding_rate,
        color = total_crime
    )) +
    scale_color_viridis_c() +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE, color = "blue") +
    labs(
        title = "Median Household Income vs Total Victimizations",
        x = "Median Household Income (x 1000$)",
        y = "Total Victimisations",
    ) +
    theme_minimal()
```



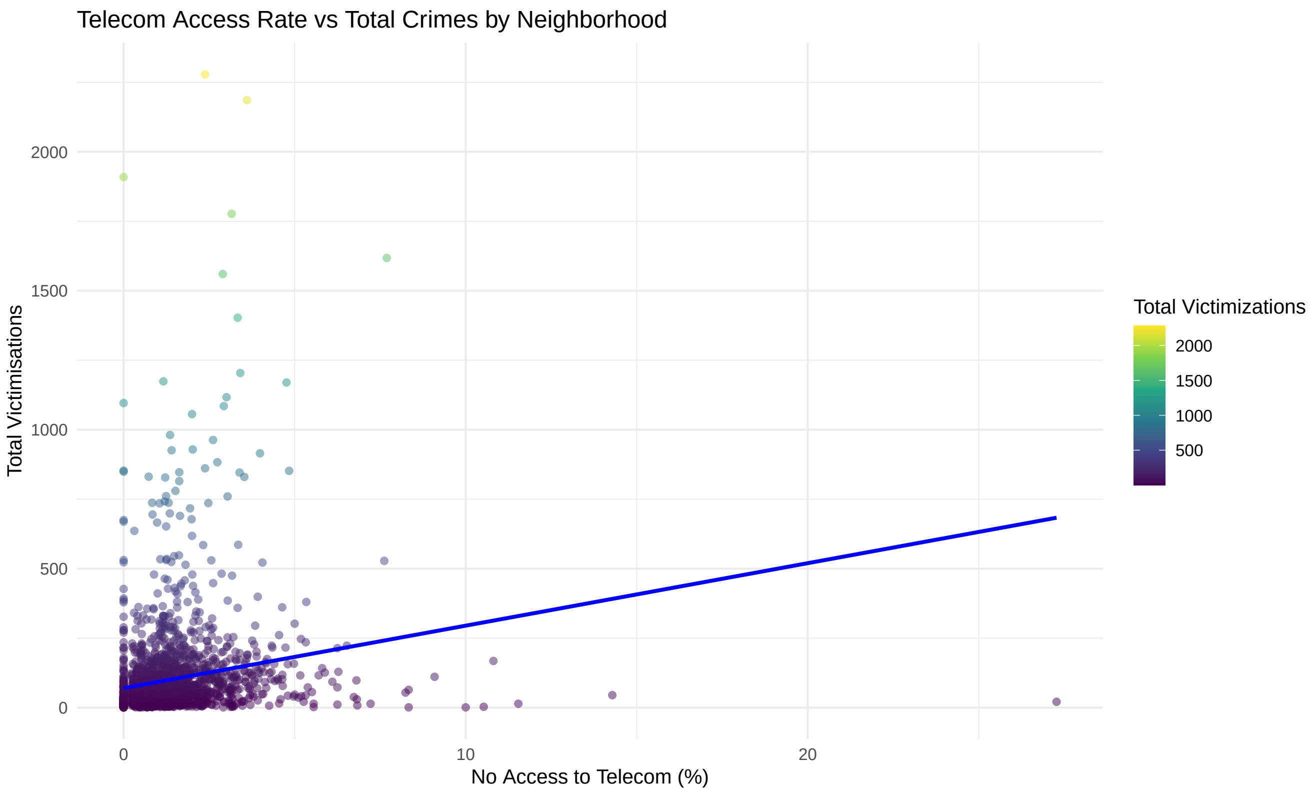
6.4.3 Vehicle Ownership (Multiple) vs Total Victimizations

```r
df_merge |>
    group_by(sa2_name) |>
    summarize(
        vehicle_count = ((
            sum(`Households - Vehicle Count (Three motor vehicles)`, na.rm = TRUE)
            + sum(`Households - Vehicle Count (Two motor vehicles)`, na.rm = TRUE)
        ) * 100)
        / sum(`Households - Vehicle Count (Total stated)`, na.rm = TRUE),
        total_crime = sum(Victimisations, na.rm = TRUE)
    ) |>
    ggplot(aes(
        x = vehicle_count,
        y = total_crime,
        color = total_crime
    )) +
    scale_color_viridis_c(name = "Total Victimizations") +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE, color = "blue") +
    labs(
        title = "Vehicle Ownership vs Total Victimizations",
        x = "Own Two or Three Cars (%)",
        y = "Total Victimisations",
    ) +
    theme_minimal()
```



6.4.4 Communication Access vs Total Crimes by Neighborhood

```
df_merge |>
    group_by(sa2_name) |>
    summarize(
        `No Access to Telecom (%)` = sum(`Households - Telecom Access (No access to telecommunication syst
            na.rm = TRUE
        ) /
            sum(`Households - Telecom Access (Total stated)`, na.rm = TRUE) * 100,
        total_crime = sum(Victimisations, na.rm = TRUE)
    ) |>
    ggplot(aes(x = `No Access to Telecom (%)`, y = total_crime, color = total_crime)) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", color = "blue", se = FALSE) +
    scale_color_viridis_c(name = "Total Victimizations") +
    labs(
        title = "Telecom Access Rate vs Total Crimes by Neighborhood",
        x = "No Access to Telecom (%)",
        y = "Total Victimisations",
    ) +
    theme_minimal()
```



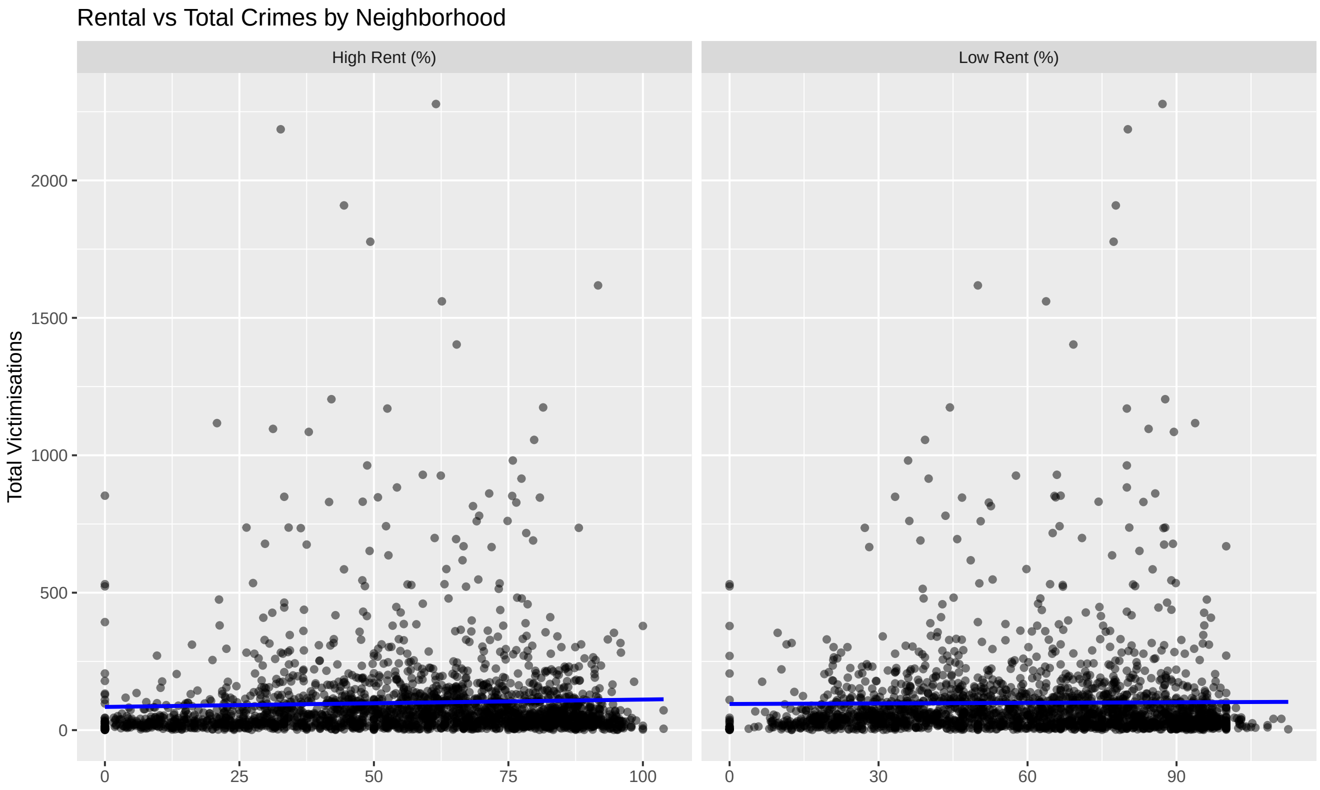
6.4.5 Weekly Rental vs Total Victimizations

```r
df_merge |>
    group_by(sa2_name) |>
    summarize(
        `Low Rent (%)` = (
            sum(`Rentals - Weekly rent (Under $200)`, na.rm = TRUE) +
                sum(`Rentals - Weekly rent ($200 - $299)`, na.rm = TRUE) +
                sum(`Rentals - Weekly rent ($300 - $399)`, na.rm = TRUE) +
                sum(`Rentals - Weekly rent ($400 - $499)`, na.rm = TRUE)
        ) * 100 /
            sum(`Rentals - Weekly rent (Total stated)`, na.rm = TRUE),
        `High Rent (%)` = (
            sum(`Rentals - Weekly rent ($400 - $499)`, na.rm = TRUE) +
                sum(`Rentals - Weekly rent ($500 - $599)`, na.rm = TRUE) +
                sum(`Rentals - Weekly rent ($600 - $699)`, na.rm = TRUE) +
                sum(`Rentals - Weekly rent ($700 - $799)`, na.rm = TRUE) +
                sum(`Rentals - Weekly rent ($800 and over)`, na.rm = TRUE)
        ) * 100 /
            sum(`Rentals - Weekly rent (Total stated)`, na.rm = TRUE),
        total_crime = sum(Victimisations, na.rm = TRUE)
    ) |>
    pivot_longer(
        cols = c(`High Rent (%)`, `Low Rent (%)`),
        names_to = "attribute",
        values_to = "value"
    ) |>
    ggplot(aes(x = value, y = total_crime)) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE, color = "blue") +
    facet_wrap(~attribute, scales = "free_x") +
    labs(
        title = "Rental vs Total Crimes by Neighborhood",
        x = NULL,
        y = "Total Victimisations"
    )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

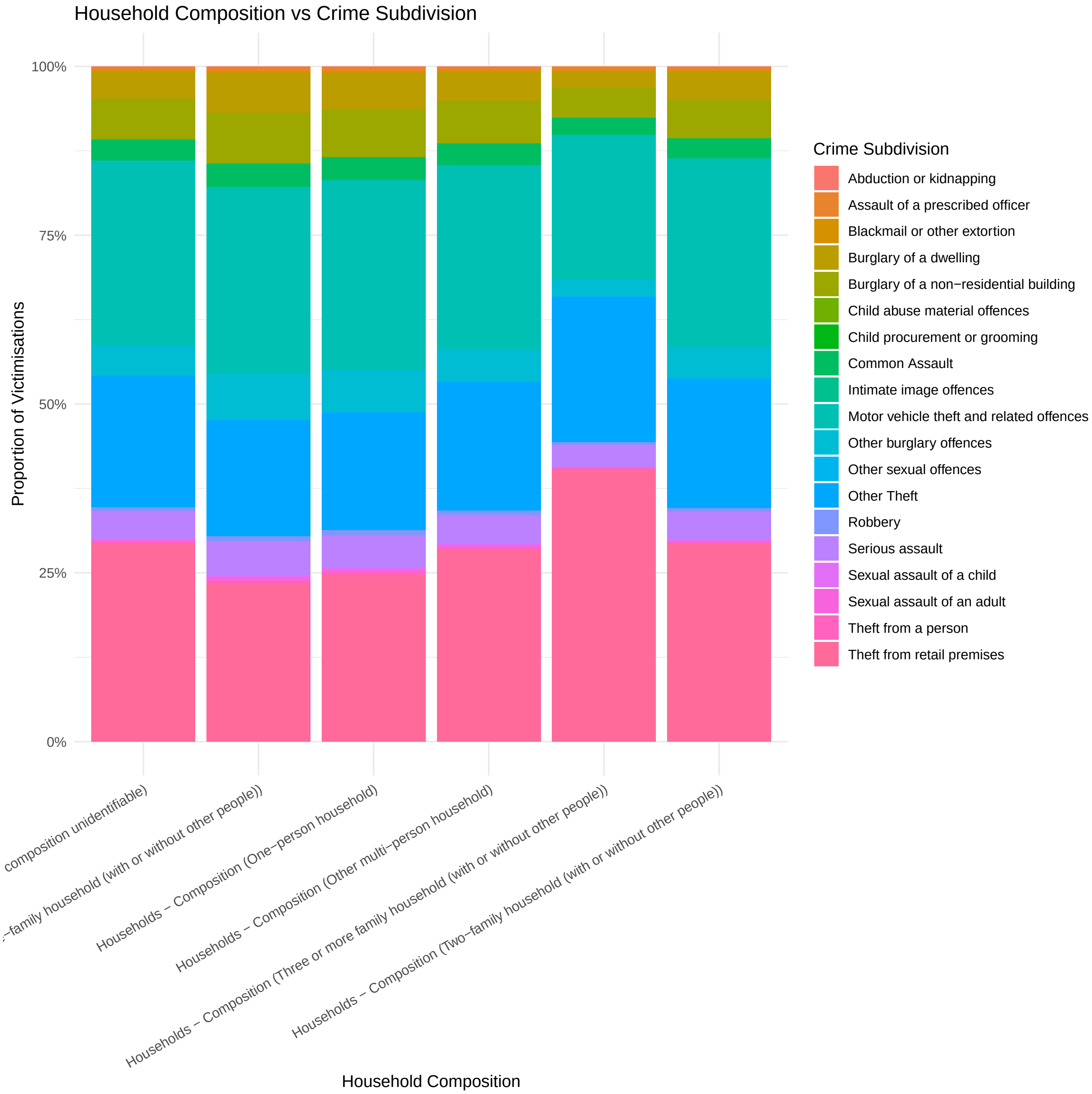
```
## Warning: Removed 80 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
```


Rental vs Total Crimes by Neighborhood
High Rent (%)
Low Rent (%)
Total Victimisations
2000
1500
1000
500
0
0
25
50
75
100
0
30
60
90

6.4.6 Subdivision vs Family Type

```
df_victimizations |>
    select(c(`ANZSOC Subdivision`, meshblock, Victimisations)) |>
    # Roll up at the meshblock level.
    group_by(meshblock, `ANZSOC Subdivision`) |>
    summarize(total_victimizations = sum(Victimisations), .groups = "drop") |>
    inner_join(
        df_meshblock |> select(meshblock, sa2_code, sa2_name),
        by = "meshblock"
    ) |>
    # Get rid of meshblock, prepare to roll up to SA2.
    select(-meshblock) |>
    # Roll up to SA2.
    group_by(sa2_code, sa2_name, `ANZSOC Subdivision`) |>
    summarize(total_victimizations = sum(total_victimizations), .groups = "drop") |>
    inner_join(
        df_census |>
            select(
                sa2_code,
                sa2_name,
                starts_with("Households - Composition") & !contains("Total")
            ),
        by = c("sa2_code", "sa2_name")
    ) |>
    # Pivot table longer, no individual column for family composition type.
    pivot_longer(
        cols = starts_with("Households - Composition"),
        names_to = "Household_Composition",
        values_to = "Household_Count"
    ) |>
    # Get rid of SA2, the table is now purely family composition & victimization subdivision.
    select(-c(sa2_code, sa2_name)) |>
    # Dedupe, because for every SA2, the stats would repeat for different crime subdivisions.
    distinct_all() |>
    # Now plot.
    ggplot(aes(x = Household_Composition, y = total_victimizations, fill = `ANZSOC Subdivision`)) +
    geom_col(position = "fill") +
    # Change sums to percentages.
    scale_y_continuous(labels = scales::percent) +
    labs(
        title = "Household Composition vs Crime Subdivision",
        x = "Household Composition",
        y = "Proportion of Victimisations",
        fill = "Crime Subdivision"
    ) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 30, hjust = 1))
```


Household Composition vs Crime Subdivision
Proportion of Victimisations
100%
75%
50%
25%
0%
composition unidentifiable)
–family household (with or without other people))
Households – Composition (One–person household)
Households – Composition (Other multi–person household)
Households – Composition (Three or more family household (with or without other people))
Households – Composition (Two–family household (with or without other people))
Household Composition
Crime Subdivision
Abduction or kidnapping
Assault of a prescribed officer
Blackmail or other extortion
Burglary of a dwelling
Burglary of a non–residential building
Child abuse material offences
Child procurement or grooming
Common Assault
Intimate image offences
Motor vehicle theft and related offences
Other burglary offences
Other sexual offences
Other Theft
Robbery
Serious assault
Sexual assault of a child
Sexual assault of an adult
Theft from a person
Theft from retail premises