Homelessness

Exploring Homelessness in Australia

Jason Wang s4134626

Last updated: 18 October, 2024

Introduction

- Homelessness is an increasing issue in Australia.The Australian Bureau of Statistics (ABS) defines homelessness as where a person lacks access to suitable accommodation alternatives.

- A person is considered homeless if:

  - *Inadequate dwelling:They live in places such as makeshift dwellings or shelters, or tents.*
  - *No security of tenure:They have no secure living arrangement. Stay is short term.*
  - *Severely overcrowded: Living arrangement is crowded with inadequate facilities.*

- Understanding characteristics of homelessness is essential for targeted interventions.These characteristics stem from housing affordability, simply being too high.Additionally, unemployment, family breakdown, and mental health issues contribute to the growing trend. Homelessness is increasing despite ongoing efforts from government and non-government organisations.

- This investigation will be prove beneficial because it:

  - *Records changes over time to identify growing risks and emerging vulnerable groups.*
  - *Recognises needed resource allocation by understanding where homelessness is most prevalent.*

- With this in mind, this study will analyze homelessness trends from 2006 to 2021, providing insights into which operational groups and regions have been most affected.The findings aim to guide policy decisions for housing security and homelessness prevention.

Problem Statement

- The investigation aims to explore:
  - *How has the rate of homelessness changed from 2006 to 2021?*
  - *What are the most affected operational groups?*
- This study uses data from the ABS 2021 census to determine homelessness by category, across time periods and groups.

Data scraping + cleaning

- Retreived data from the ABS. Collected from the census (every 5 years).
- Scrape then tidy data to perform data analysis better without errors.
- Remove unwanted rows, rename columns for better clarity.

```
        # Data scraping
        homelessness.url <- "https://www.abs.gov.au/statistics/people/housing/estimating-homelessness-
census/2021/20490do001_2021.xlsx"
        homelessness <- read.xlsx(homelessness.url, sheet = "Table_1.1", startRow = 5)

        # Remove unnecessary rows at the bottom
        homelessness <- homelessness[1:(nrow(homelessness) - 14), ]

        # Remove unwanted rows
        homelessness_clean <- homelessness %>%
          slice(-c(1, 2, 3, 11, 21, 24, 27, 31))

        # Rename columns
        colnames(homelessness_clean) <- c(
          "Category", "2006_Count", "2006_Percentage_proportion", "2006_Rate_per_10k",
          "2011_Count", "2011_Percentage_proportion", "2011_Rate_per_10k",
          "2016_Count", "2016_Percentage_proportion", "2016_Rate_per_10k",
          "2021_Count", "2021_Percentage_proportion", "2021_Rate_per_10k"
        )
```

Data explanation.

- Variables:
  - *Category (Factor): Grouping variables to categorise similar types of homelessness.*
  - *Year (Categorical):The census year for data.*
  - *Count (Numeric):The number of homeless people in each category.*
  - *Percentage (Numeric): Percentage of the total homeless population within this category.*
  - *Rate (Numeric): Rate of homelessness per 10,000 people.*
- Levels of the factor 'Category':
  - *Living conditions: People living in improvised dwellings or tents, supported accommodation for homeless, staying temporarily with others, boarding houses, temporary lodgings and crowded dwellings.*
  - *Age: Spanning from under 12 to 75 and over*
  - *Sex: Male or Female*
  - *Indigenous Status: Indigenous or non-indigenous*
  - *State or Territory:Which state or territory.*

Data explanation cont. + pivot longer.

- Scale of numeric variables:
  - *Count: Number representing the amount of homeless individuals.*
  - *Percentage: Percentage value between 0 and 100 indicating the proportion of homeless people.*
  - *Rate: Rate per 10,000 population.*

```
          # Convert to long format
          homelessness_tidy <- homelessness_clean %>% pivot_longer( cols = -Category,  names_to = c("Year", ".value"),names_sep =
"_" )


          head(homelessness_tidy, n = 3)
```

Category <chr>	Year <chr>	Count <chr>	Percentage <chr>	Rate <chr>
People living in improvised dwellings, tents, or sleeping out	2006	7252	8	3.7
People living in improvised dwellings, tents, or sleeping out	2011	6810	7	3.2
People living in improvised dwellings, tents, or sleeping out	2016	8200	7	3.5

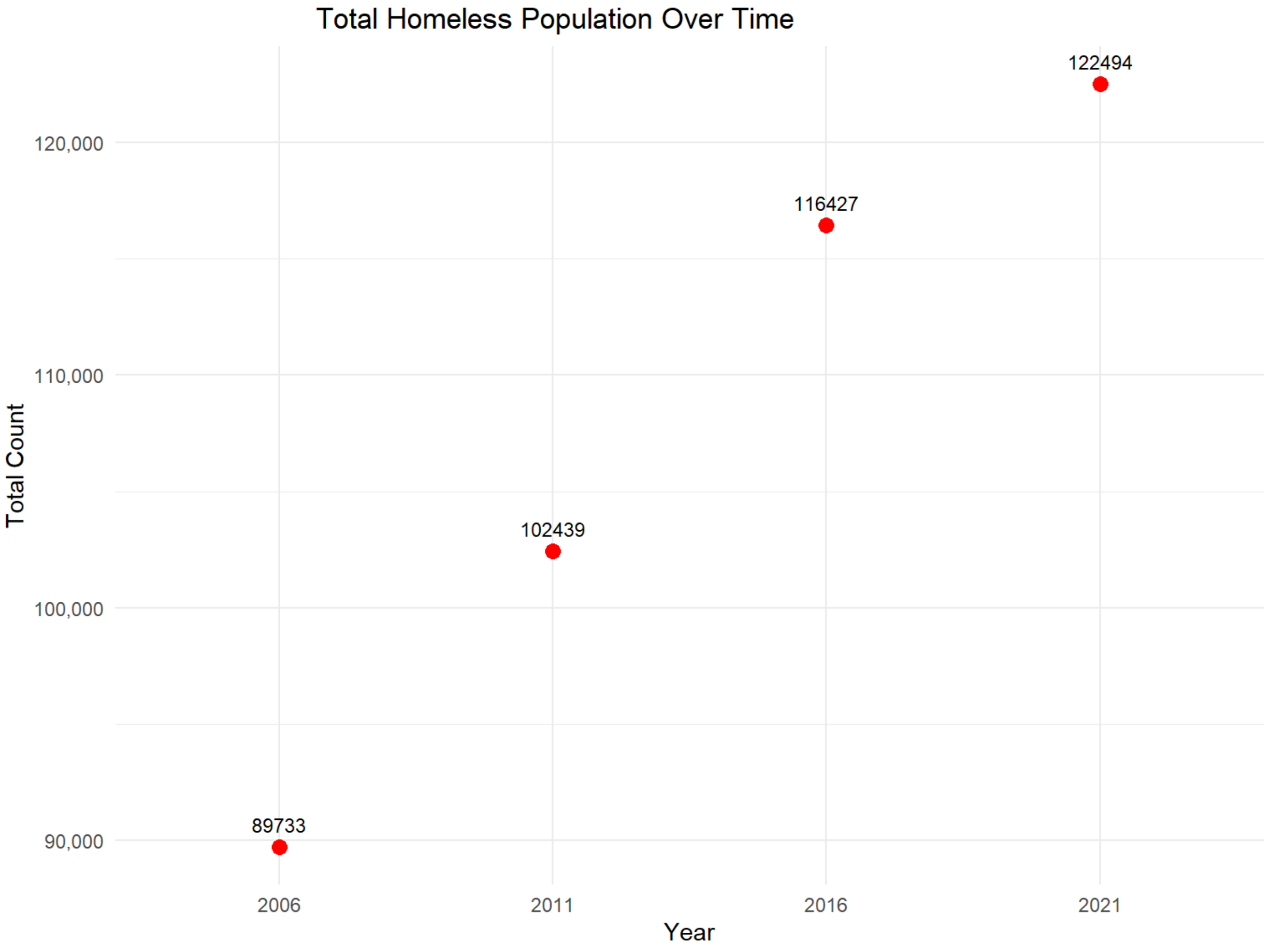
3 rows

Descriptive Statistics

- Category (Factor): Represents different living conditions, age groups, sex, Indigenous status, and state/territory.
- Year (Categorical): Depicts the census year.
- Count (Numeric): The total number in the respective category.
- Percentage (Numeric): The proportion of the total homeless population in each category.
- Rate (Numeric): Rate of homelessness per 10,000 people in the general population.

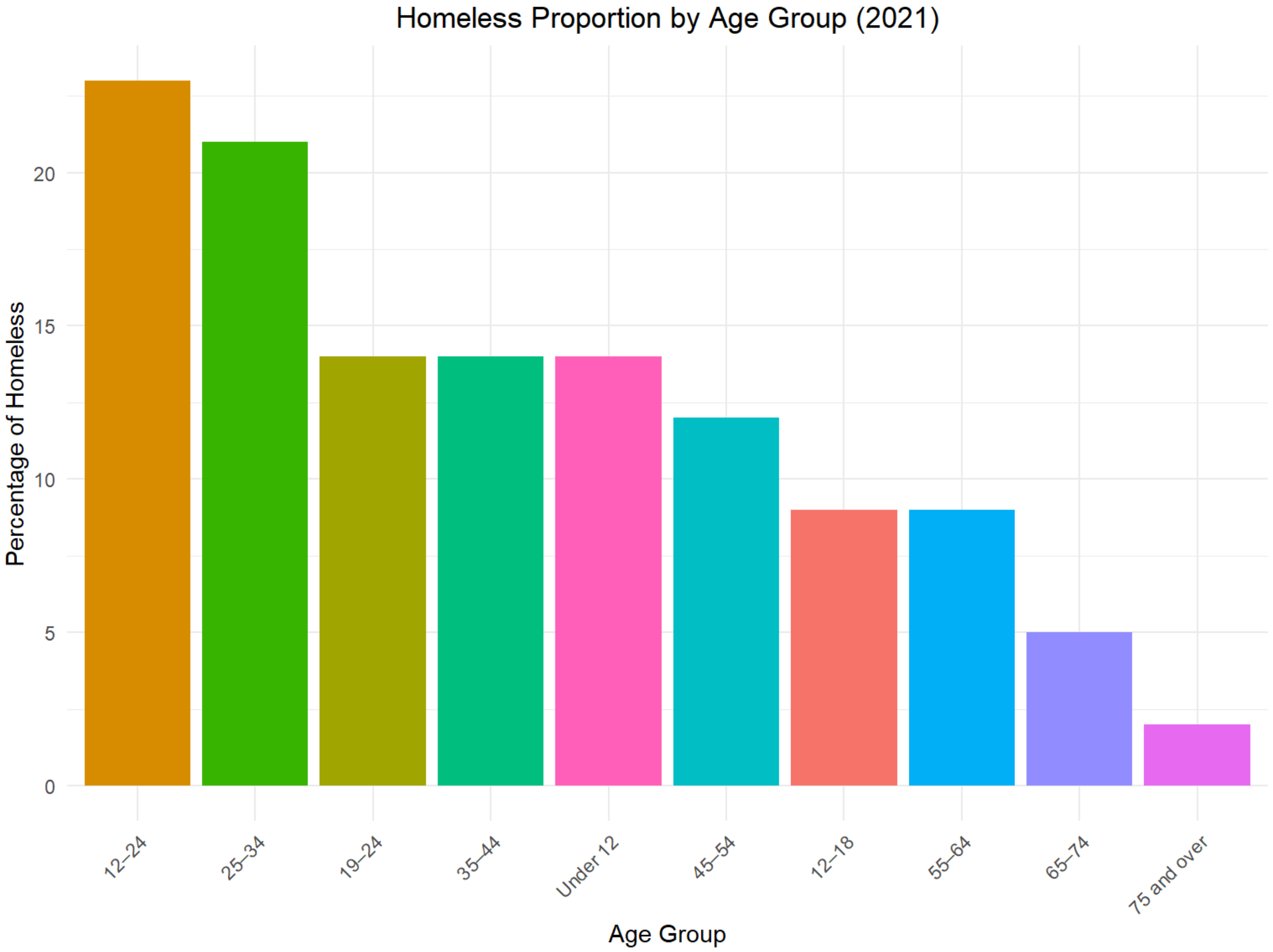
Visualisation - Total homeless

```
        total_homeless <- homelessness_tidy %>% filter(Category == "All homeless persons") %>% mutate(Count = as.numeric(Count))

        ggplot(total_homeless, aes(x = Year, y = Count)) +
          geom_line(size = 1.1, color = "blue") + geom_point(size = 3, color = "red") + geom_text(aes(label = Count), vjust = -1,
size = 3) +
          scale_y_continuous(labels = scales::comma) + labs(title = "Total Homeless Population Over Time", x = "Year", y = "Total
Count") +
          theme_minimal() + theme(plot.title = element_text(hjust = 0.3))
```



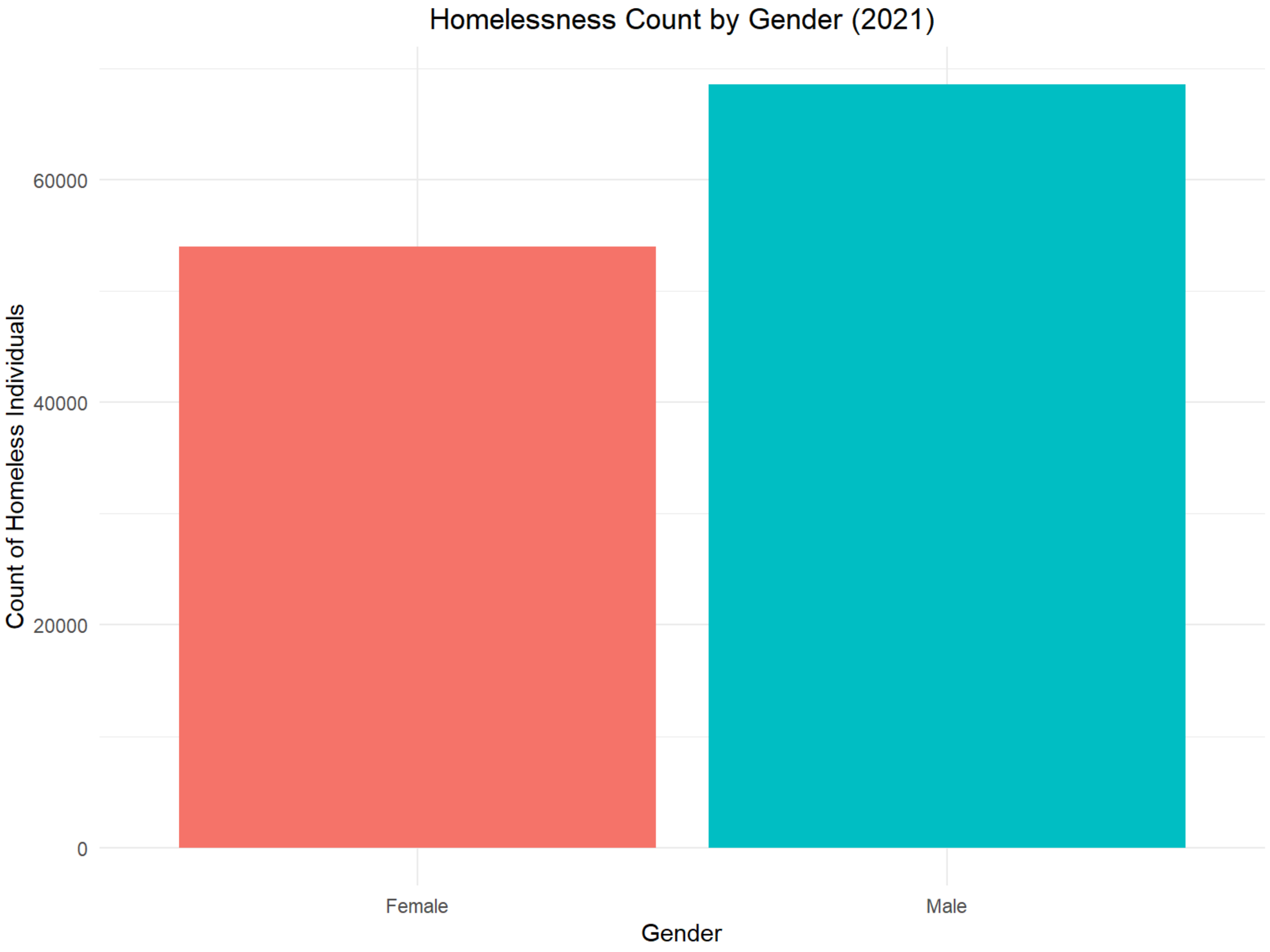
Visualisation - Homeless proportion by age group

```
        age_data_2021 <- homelessness_tidy %>% filter(Year == 2021 & grepl("Under|\\d{2}-\\d{2}|75 and over", Category)) %>%
mutate(Percentage = as.numeric(Percentage))

        ggplot(age_data_2021, aes(x = reorder(Category, -Percentage), y = Percentage, fill = Category)) +
        geom_col() + labs(title = "Homeless Proportion by Age Group (2021)", x = "Age Group", y = "Percentage of Homeless") +
theme_minimal() +
        theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 45, hjust = 1), legend.position =
"none")
```



Visualisation - Homelessness count by gender

```
        gender_data_2021 <- homelessness_tidy %>% filter(Year == 2021 & Category %in% c("Male", "Female")) %>% mutate(Count =
as.numeric(Count))

        ggplot(gender_data_2021, aes(x = Category, y = Count, fill = Category)) +geom_col() +
        labs(title = "Homelessness Count by Gender (2021)", x = "Gender", y = "Count of Homeless Individuals") +
        theme_minimal() + theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
```



Hypothesis Testing

- It would like to be determined whether homelessness rate has changed significantly between 2006 and 2021.
  - *Null Hypothesis (H_0):There is no significant difference in the average homelessness rate between 2006 and 2021.*
  - *Alternative Hypothesis (H_1):There is a significant difference in the average homelessness rate between 2006 and 2021.*
- Assumptions for t test include:
  - *Independence:The two samples (2006 and 2021) must be independent.*
  - *Normality:The data for each year should follow a normal distribution.*
  - *Equal Variances:The variances between the two samples should be equal or similar.*

```
        data_test <- homelessness_tidy %>% filter(Year %in% c(2006, 2021)) %>% select(Year, Rate) %>% mutate(Rate =
as.numeric(Rate))

        rate_2006 <- data_test %>% filter(Year == 2006) %>% pull(Rate)
        rate_2021 <- data_test %>% filter(Year == 2021) %>% pull(Rate)

        t_test_result <- t.test(rate_2006, rate_2021, var.equal = TRUE)
```

Hypthesis Testing Cont.

```
print(t_test_result)
```

```
## 
##  Two Sample t-test
## 
## data:  rate_2006 and rate_2021
## t = 0.28452, df = 62, p-value = 0.777
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -58.71346  78.20096
## sample estimates:
## mean of x mean of y 
##  75.24375  65.50000
```

- t = 0.285
  - *This indicates how many standard errors the difference in means is from 0.*
- df = 62
  - *This is used to calculate p-value.*
- p-value = 0.777
  - *Since p-value > 0.05, we fail to reject the null hypothesis.*
  - *This means there is no statistically significant difference in the homelessness rates between 2006 and 2021.*
- 95% confidence interval = Range: (-58.71, 78.20)
  - *We are 95% confident that the true difference in homelessness rates between 2006 and 2021 lies within this range.*
  - *Interval contains 0. Difference between the two means could be 0.*

Hypthesis Testing Cont.

```
print(shapiro.test(rate_2006))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rate_2006
## W = 0.37315, p-value = 1.382e-10
```

```
print(shapiro.test(rate_2021))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rate_2021
## W = 0.47226, p-value = 1.307e-09
```

- Since p-value is <0.05 for both, we reject null hypothesis that the data follows a normal distribution.

- It can be concluded that:

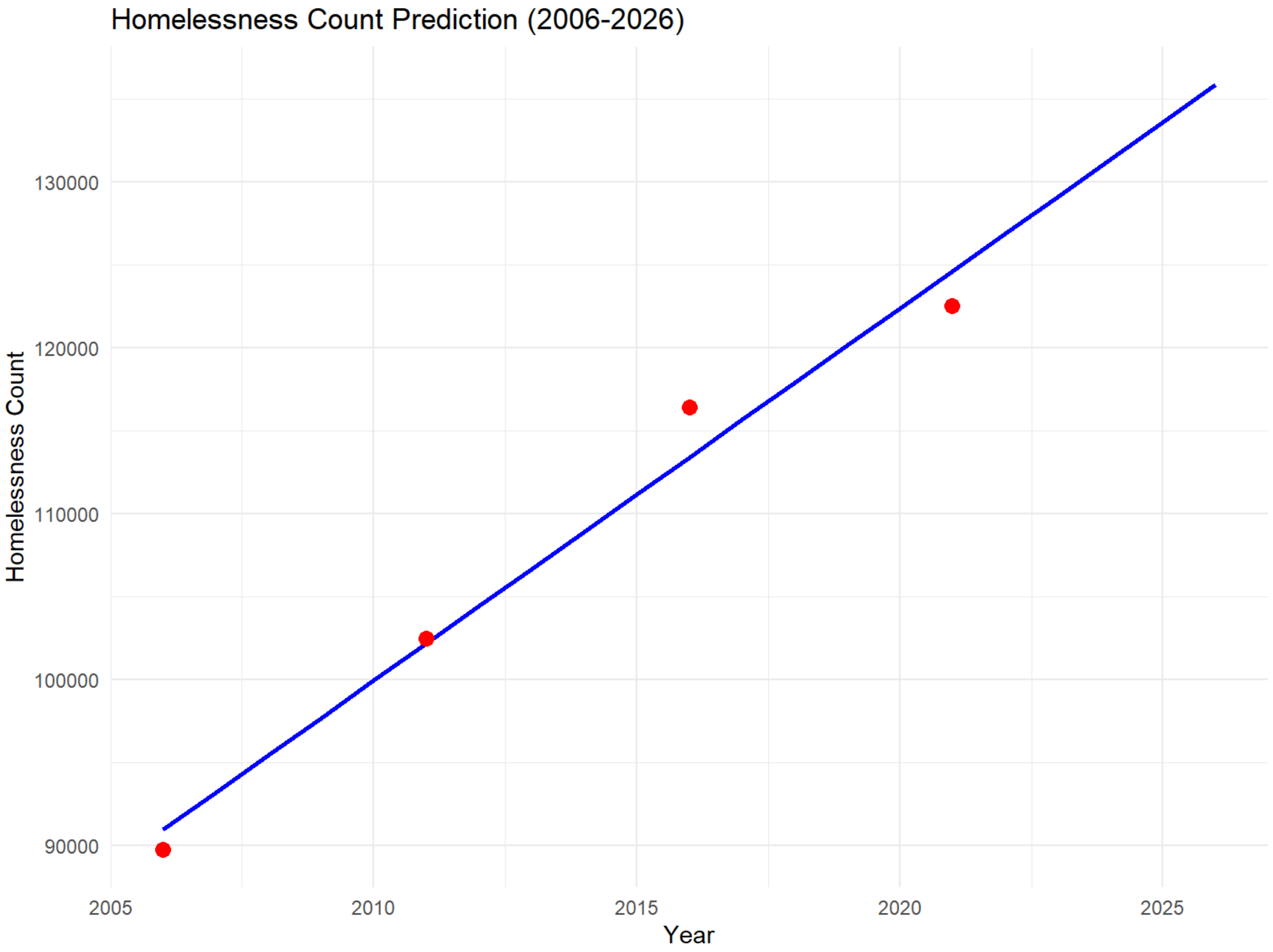
  - *The difference in homelessness rates between 2006 and 2021 is not statistically significant due to the p value.*
  - *The difference between 2006 and 2021 is likely due to random variation because the CI contains 0.*
  - *We fail to reject null hypothesis meaning there is no strong evidence suggesting that homelessness rates have changed significantly.*

Regression analysis

- Null Hypothesis (H_0):
    - *There is no significant linear relationship between year and homelessness count.*
- Alternative Hypothesis (H_1):
    - *There is a significant linear relationship between year and homelessness count.*
- Linearity: The relationship between year and homelessness count should be linear.
- Independence: The observations are independent.
- Homoscedasticity: The residuals (errors) should have constant variance.
- Normality of Residuals: The residuals should follow a normal distribution.

Regression analysis cont.

```
        original_data <- data.frame(Year = c(2006, 2011, 2016, 2021), Count = c(89733, 102439, 116427, 122494))
        model <- lm(Count ~ Year, data = original_data)
        complete_years <- data.frame(Year = 2006:2026)
        complete_years$Predicted_Count <- predict(model, newdata = complete_years)

        ggplot(complete_years, aes(x = Year, y = Predicted_Count)) + geom_line(color = "blue", size = 1) + geom_point(data =
original_data, aes(x = Year, y = Count),
        color = "red", size = 3) + labs(title = "Homelessness Count Prediction (2006-2026)", x = "Year", y = "Homelessness
Count") + theme_minimal()
```



Regression analysis cont.

```
summary(model)
```

```
##
## Call:
## lm(formula = Count ~ Year, data = original_data)
##
## Residuals:
##       1       2       3       4
## -1199.6   279.3  3040.2 -2119.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4413380     497362  -8.874   0.0125 *
## Year            2245        247   9.090   0.0119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2762 on 2 degrees of freedom
## Multiple R-squared:  0.9764, Adjusted R-squared:  0.9646
## F-statistic: 82.63 on 1 and 2 DF,  p-value: 0.01189
```

- Intercept = -4,413,380
- Slope (Year) = 2,245
- p-value for intercept = 0.0125
- p-value for year = 0.0119

Regression analysis cont.

- Multiple R-squared = 0.9764
  - *97.64% of the variation in homelessness count is explained by the year*
- Adjusted R-squared = 0.9646
- F-statistic: 82.63 with p-value = 0.01189
  - *p-value < 0.05, the model is statistically significant*
- It can be concluded that:
  - *Year is a significant predictor of homelessness count. It increases 2,245 people per year on average.*
  - *The high R-squared (97.64%) shows that the year explains homelessness well.*
  - *It is statistically significant due to p-value.*

Discussion

- Major findings:
  - *From hypothesis testing, the p-value from the t test returned 0.777, showing no significant difference in the homelessness rates between 2006 and 2021.The 95% CI had an interval of (-58.71, 78.20). It includes 0, indiciating that observed difference can be due to random variation.*
  - *From regression analysis, the p-value from the model showed a statiscally significant positive trend.This means that homelessness count increased over time.The model estimated that for every year on average, homeless people increases by 2,245. R squared value of 97.64% indicates the year explains most of the variability in homelessness count.*
- Strengths:
  - *Regression r squared value of 97.64% suggests strong relationship between year and homelessness count.This positive trend in homelessness provides valuable insight for intervention.*
- Limitations:
  - *T test compared only the years 2006 and 2021. Regression model only used year as a predictor. So there aren't many variables, let alone observation count of 2006, 2011, 2016 and 2021.Thus, a big limiter of this analysis are the lack of observation counts and variables used.*
- Proposed Directions for Future Investigations:
  - *Collect more frequent data. Other factors such as economic conditions, house prices and policies and unemployment rates may impact homelessness and can account for this data.*

Conclusion

- The main finding of this investigation is that homelessness in Australia has been increasing over time. An average increase of 2,245 extra homeless people per year is quite staggering. The year is a signficant predictor for homelessness count. The model used also fits well with the data, as it had an R squared of 97.64%.

- Relating back to the statistical question, it can be seen that the growing risk are that the numbers of homeless people are increasing at a large rate. Most vulnerable groups are the younger groups, especially the younger generation. Ages 0-34 represent almost 50% of the homeless population, which would make their futures miserable, let alone experiencing life as non-existent. Proactive intervention is needed, because the increase in homelessness is scary. We must address the underlying causes to mitigate this upward trend.

References

- “https://www.abs.gov.au/statistics/people/housing/estimating-homelessness-census/2021/20490do001_2021.xlsx”