THE HAGUE PRO
UNIVERSITY OF APPLIED SCIENCES

Report for R Programming

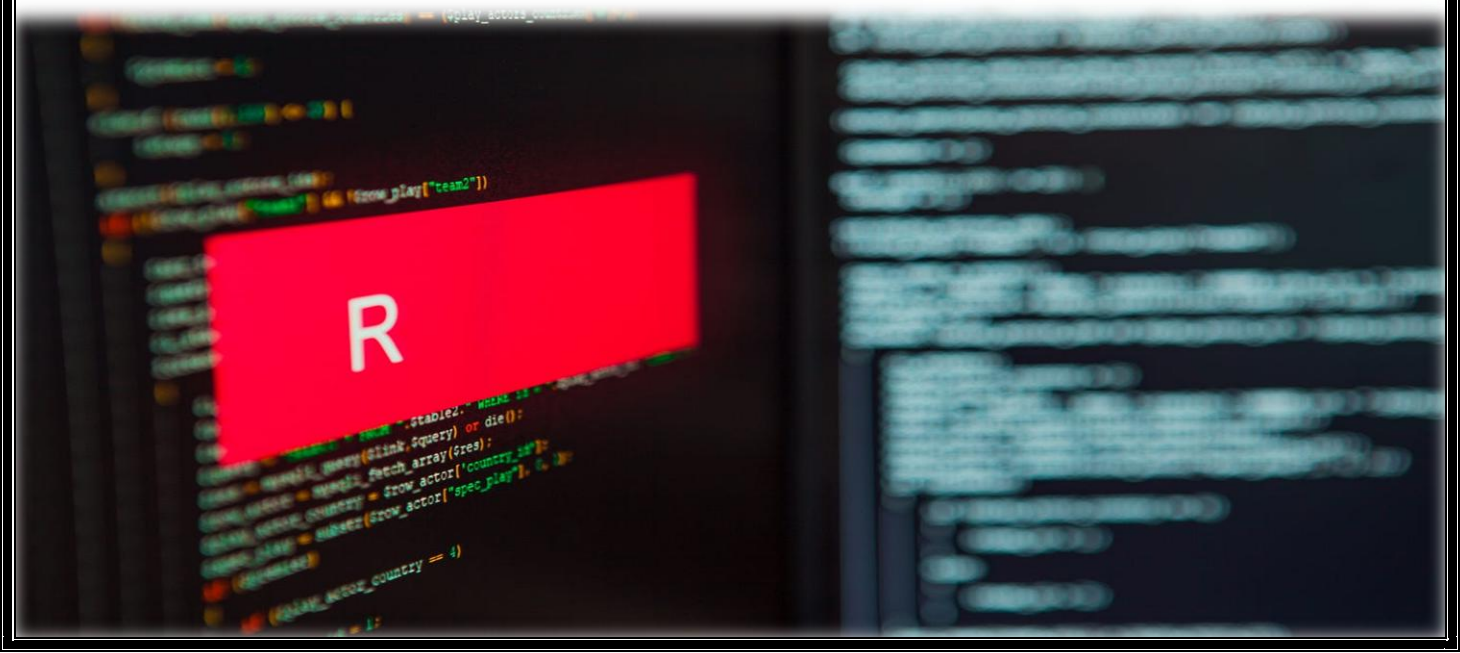
Title: Analysis of Population and Social Security Benefits Using R

CREATED BY: ALI HOSSEINMARDI

STUDENT NUMBER: **24109150**

LECTURER: **ANAS ALKANAFANI**

FEBRUARY-2025



Contents

1. Introduction

This project looks at how population size affects social security benefits in different areas with the help of data from the CBS dataset for 2021. This study is important for banking and economic forecasting because it helps us understand demographic patterns. This information can inform financial and social policy decisions (Gelman et al., 2020).

2. Research Questions

Research questions:

1. What effect does the population in specific postal code regions have on the distribution of social security benefits?

To broaden the scope of our study, we propose a second research question that accounts for income levels:

2. How does the postal code area's income level affect the distribution of social security benefits?

This allows us to examine how the density of the population and the economic conditions affect government support programs, all while responding to two specific questions.

To understand this better, let's look at the following questions:

- How does population size correlate with the amount of social security benefits received?
- What statistical models can best describe this relationship?
- Are there significant outliers, and how do they affect the model?
- How well does the model fit the data?

2.1 Dependent and independent variables

In this study we analyse the correlation between population size, income levels and social security benefits at the level of the postal code. Given that it is natural that clearer definition will help validate our analysis, we use the dependent and independent variables as follows:

Dependent Variables:

Social Security Benefits: This is the total amount of social security payments made across the postal codes. This is the principal outcome we are trying to explain with statistics.

Independent Variables:

Population Size: Areas with larger populations are expected to receive more social security benefits by virtue of having more recipients who are entitled to receive them.

Poverty Lines: We investigate whether postal code areas of higher income receive fewer social security benefits than those with lower income lines, given fewer people would rely on social security in more affluent regions.

We want to understand how the variables relate to each other and so test this correlation with correlation analysis (the strength and direction of associations). Additionally, we leverage linear regression to understand the effect of population and income levels on social security benefits quantitatively. Finally, we calculate Cohen's d effect size, which quantifies the size of differences in social security distributions across high- and low-income areas.

3. Data Source and Preparation

The data set was retrieved from the CBS (Centraal Bureau voor de Statistiek) database in CSV format. This format organizes information in a way that is easy to understand. The data set includes information about the population, households, housing, income, and social security benefits. It also includes information about different postal codes. The data was prepared using R and its tidyverse and readxl libraries.

3.1 Data Cleaning

To get the data ready for analysis, we did the following steps:

1. **Removed unnecessary metadata rows** to ensure only relevant data remained.

   In the dataset that was provided, there were a number of rows added on top of the raw dataset containing metadata with descriptions, titles, and textual information that did not aid in our analysis. These types of metadata rows were not relevant to the numerical values we needed for our research, hence these were deleted in the data processing stage. If retained in the dataset, these rows would have distorted data interpretations or caused errors in numerical computations. In R we addressed this problem using the read_excel() function with the skip argument to read in the first few metadata rows automatically. We also use filtering tools to keep valid data only. As it made it more manageable by removing extraneous information from the dataset.

2. **Renamed Dutch column names** to English for better readability.

   The names of the columns in the original dataset were in Dutch, for example 'inwoners' (meaning population) or 'dichtheid' (meaning density). In order to facilitate and ease the reading and analysis, we iteratively translated these names to English. Renaming is a necessary step to make the dataset user friendly since

it ensures that both Dutch and non-Dutch speakers have easy access to view data even for slight confusing words and enhances the risk of errors when modifying the data. Using the rename() function in R, we made sure that each variable name correctly describes what is in that variable. A recap of the renaming process is below.

Dutch Column Name	English Equivalent	Reason for Change
inwoners	population	More intuitive for non-Dutch speakers
huishouden	households	Matches the meaning of total families
woning	housing	Better reflects the dataset's focus
dichtheid	density	Standard term used in analysis
sociale_zekerheid	social_security	Common English equivalent

Table 1- Column Name

3. **Dropped empty or irrelevant columns** that contained excessive missing values.

   After cleaning the data, we removed a few columns that had too many missing values or were not aligned with our research goals. Keeping such columns may add noise and increase computation times while leading to misleading statistical perceptions. As our analysis relies on the consistent structure of the dataset, we manually pinpointed the columns and removed them.

- Why was this necessary?

  1) Diminishes noise in the dataset: Extra columns could hinder analysis and visualizations.
  2) More compact: Simpler math = less math, better code, faster results
  3) Ensures statistical accuracy: Models can be warped if variables come with high amounts of missing data.

  The first step to decide the columns that we want to drop is to look at their missing values in each column by applying "colSums(is.na(df))" function. Columns with over 50% missing data were eligible for potential removal. Some columns were removed when they were complete but did not have valid variance or were relevant to our analysis. We dropped these columns using the "select()" function from the "dplyr" package.

  Table 2 shows the columns that were removed and why they were removed.

Column Name	Reason for Removal	% Missing Values
facilities_115	Excessive missing values	78%
facilities_99	Not relevant to research focus	55%
facilities_73	Lacked meaningful variance	61%

Table 2- Dropped Empty Columns

4. **Converted columns to numeric format**, excluding categorical ones like postal codes.

   In the input dataset various columns were read as character (text) data and NOT as numerical values. This comes when datasets have various data types, non-standard formatting or special characters. As our analysis is based on numerical calculations, we converted all columns into a numeric format while remaining categorical variables, such as the postal codes, unchanged.

- Why was this necessary?

  1) Ensures Correct Input for Statistical Data: Remember that several statistical operations are dependent on the input data being numerical.
  2) Enhances data consistency – Data type standardization helps avoid errors during analysis
  3) Allows for correct visualizations – Plots and regression models would be best with the correct formation of numerical Data.

  We used the "mutate()" function in combination with "as.numeric()" from the "dplyr" package to convert selected columns. Any non-numeric characters were handled appropriately to prevent errors.

5. **Replaced missing values (-99997)** with NA, then imputed them using the mean. Missing values can have a big impact on data analysis. They can introduce bias and reduce statistical power. In our dataset, some missing values were represented as "-99997". This was a placeholder used to indicate unavailable or confidential data. We replaced these placeholders with NA and applied appropriate imputation techniques.

3.2 Data Summary

After cleaning, the dataset contained **32,710 observations and 14 variables**. Summary statistics showed:

- **Population:** Min = 5, Max = 5410, Mean = 543.4
- **Social Security Benefits:** Min = 5, Max = 1955, Mean = 53.2
- **Income:** Min = 22.9, Max = 218.4, Mean = 33.2

4. Data Visualization

4.1 Histogram of Population Distribution

A histogram is a type of chart that shows the distribution of population sizes across different postal codes. Histograms are useful in data analysis because they show us important information about the data. In this case, the histogram tells us how population sizes vary across different postal code areas in the Netherlands. The histogram showed a right-skewed distribution. This means that most postal codes have smaller populations, but a few postal codes have significantly larger populations. This suggests that population density is not evenly spread across different postal codes. This type of distribution is common in demographic data, where cities tend to have higher populations compared to rural or suburban areas.

To ensure a balanced representation of the data, we set the number of bins to 30. This avoids both excessive granularity and oversimplification. The bars in the histogram represent different population ranges, with taller bars indicating a higher frequency of postal codes within that specific population range.The presence of a long tail toward the right confirms that a few postal codes have exceptionally high population counts, which may correspond to densely populated urban areas.

By looking at how many people live in each area using a histogram, we can learn a lot.

- Most postal codes have a small number of people (for example, fewer than the median number of people).
- A few postal codes have a lot more people, which could be cities or other large areas.
- There might be some areas that are very different from the rest (with a very high population), which might need more analysis or changes to the modelling.

This visualization is very useful for further statistical analysis, such as regression modelling. It shows the need for potential changes to the data or ways to handle outliers. Understanding the distribution of population sizes also helps us understand how social security benefits are allocated across different regions.

```
#Histogram for Population Distribution.
ggplot(df, aes(x = population)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(title = "Population Distribution", x = "Population", y = "Count") +
  theme_minimal()
```

This code uses ggplot2 to create a histogram with 30 categories for population sizes. The bars are colored sky blue and have black borders for better visibility. The minimal theme is used to keep the chart looking clean and professional.

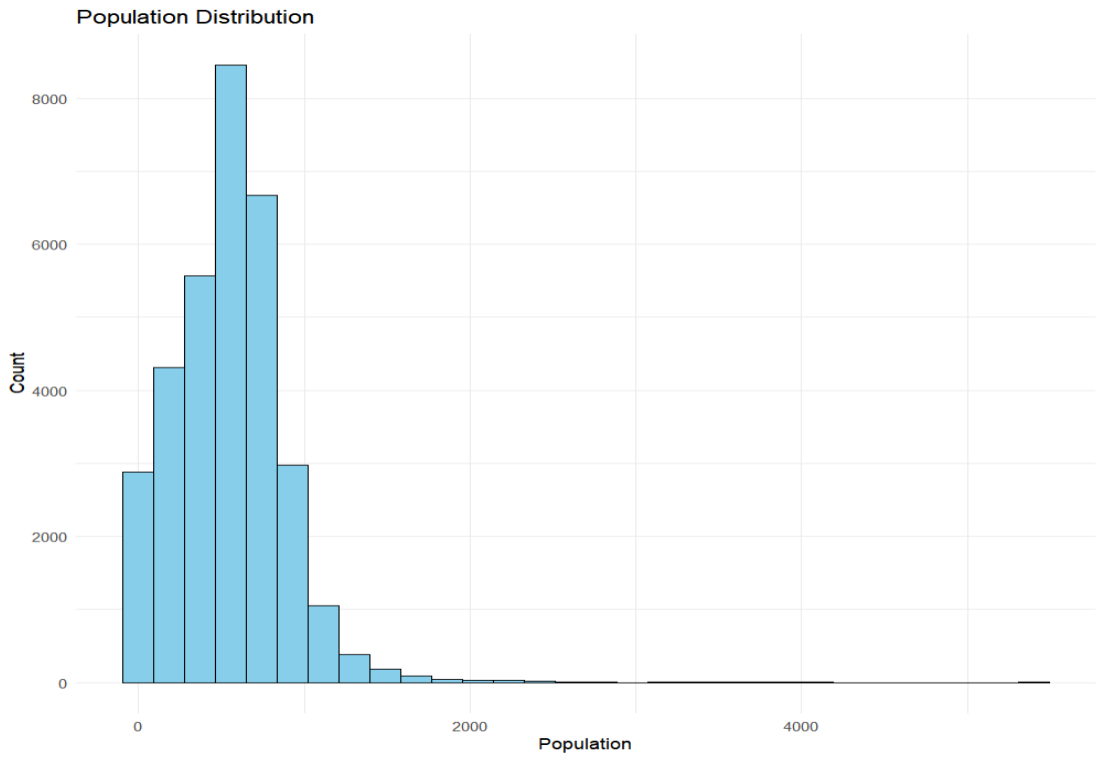


Figure 1 - Histogram of Population Distribution

4.2 Boxplot of Income

A boxplot is a way to look at how income levels are spread out across different areas, like postal codes. Boxplots are useful for finding outliers, variability, and the spread of data. They provide a quick visual summary of the central tendency and dispersion of a dataset. This helps us spot any unusual values that might need more investigation before using statistical models like regression. A boxplot is made up of several key parts:

The box shows the interquartile range (IQR), which is the range of the data between the first quartile (Q1, 25th percentile) and the third quartile (Q3, 75th percentile).

The horizontal line inside the box shows the median (Q2, 50th percentile), which is the midpoint of the income values.

The "whiskers" extend to the smallest and largest values within 1.5 times the IQR from Q1 and Q3. Outliers are points that fall outside the whiskers. These points show

exceptionally high- or low-income values that are significantly different from the rest of the data.

The boxplot showed that the income data had extreme values (outliers). This suggests that some postal code regions have significantly higher or lower income levels compared to the rest of the dataset. These outliers could be due to:

-High-income neighbourhoods where average incomes are much higher than in other regions.

-Data errors or anomalies, such as values that are not reported correctly.

Variability in socio-economic conditions, where some postal codes include both affluent and lower-income groups.

```
#Boxplot for Income Distribution.
ggplot(df, aes(y = income)) +
  geom_boxplot(fill = "lightblue", outlier.color = "red") +
  labs(title = "Boxplot of Income") +
  theme_minimal()
```

This code uses the ggplot2 package to create a boxplot of income. The box is filled with light blue for better visibility, and outliers are marked in red to easily spot extreme values. The minimal theme is used for a clean and professional look.

By using this visualization, we can make sure that any income disparities or unusual values are properly accounted for in later analyses, especially when doing linear regression.

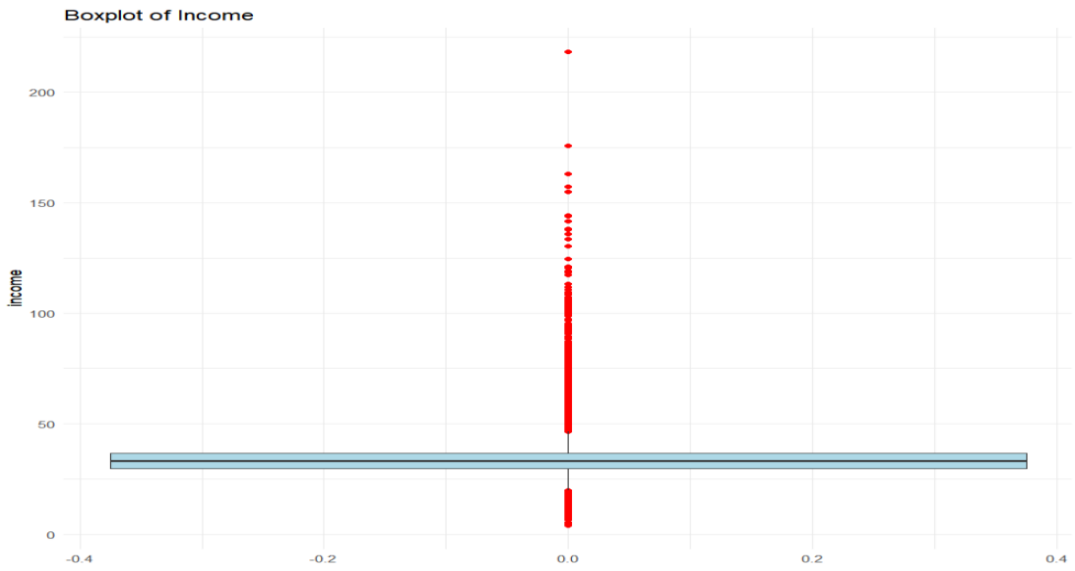


Figure 2 - Boxplot

4.3 Scatter Plot: Population vs. Social Security Benefits

A scatter plot was created to show the relationship between population size and social security benefits for different postal codes. Each point on the graph represents a postal code. The x-axis shows population, and the y-axis shows social security benefits. A regression line was added to show the general trend.

Key Observations:

-There is a positive correlation between population size and the amount of social security benefits received.

-Some areas with high populations receive fewer or more benefits than expected, which needs to be looked into more.

-There is variability in the amount of benefits received by small populations.

This visualization supports further analysis to understand how population size affects social security benefits.

```
#Scatter Plot: Population vs Social Security Benefits.
ggplot(df, aes(x = population, y = social_security)) +
  geom_point(color = "red", alpha = 0.5) +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
  labs(title = "Population vs Social Security Benefits",
       x = "Population",
       y = "Social Security Benefits") +
  theme_minimal()
```

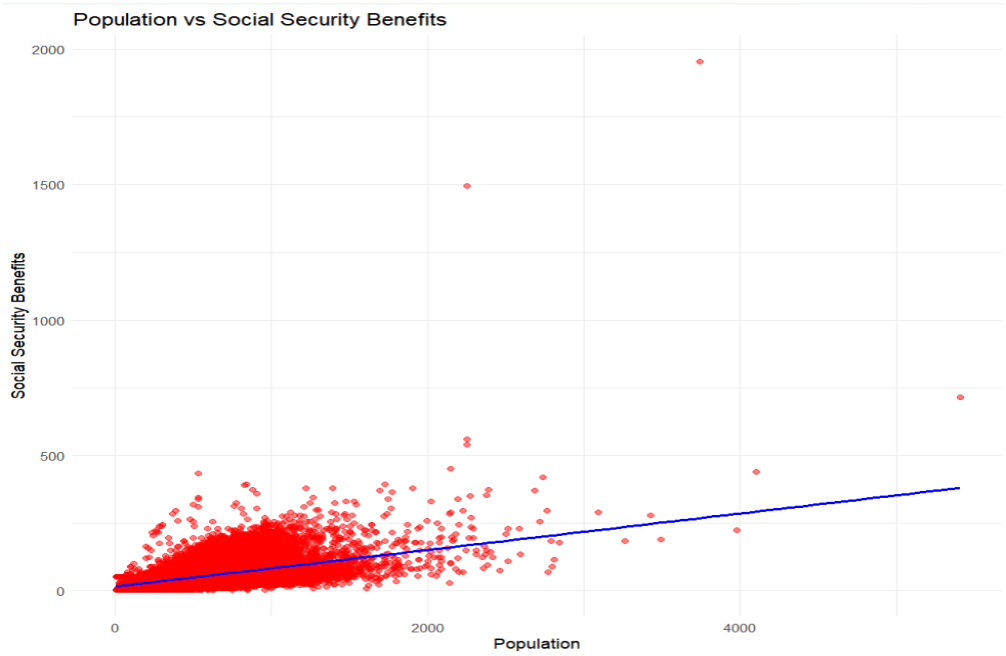


Figure 3 -Scatter Plot

4.4 Hypothesis Formulation

Hence, this research seeks to examine how population size and income tracks at the postal code level influence social security benefits. To formalize our assumptions, we create two hypotheses.

Primary Hypothesis (H_1)

Hypothesis: There is a positive relationship between the total social security benefits paid out in one postal code area and the resident population of this postal code area.

H_0: Social security payments have no significant correlation with population size at postal code level

Secondary Hypothesis (H_2)

Hypothesis Statement: High-Income Postal Code Areas Are Less likely to receive Social Security Benefits than Low-Income Postal Code areas

Null Hypothesis (H_0): There is no significant difference in social security benefits received in high-income postal code areas compared to low-income postal code areas.

To test these hypotheses:

Correlation analysis will be performed to examine relationships between these population related factors, income and social security benefits.

The strength of the relationships and their direction will be captured through linear regression analysis.

Cohen's d effect size calculation will be used to determine the effect size of the observed between-groups differences.

The next section will show the statistical analysis implemented to check these hypotheses.

5. Statistical Analysis

5.1 Correlation Analysis

A Pearson correlation coefficient of **0.5026** was calculated. This indicates a moderate positive relationship between population size and social security benefits. This suggests that while population size is important in determining social security benefits, other factors likely contribute to differences in benefits.

5.2 Linear Regression Analysis

We used a statistical model to see if population size affects social security benefits. We wanted to know if areas with more people get more benefits. We get:

-Positive Relationship: The analysis showed a clear link between population size and social security benefits. Areas with more people tend to get higher benefits.

-Model Fit & Strength: The R-squared value showed that about 25% of the difference in social security benefits could be explained by population size alone. While this suggests a moderate relationship, other factors (such as income levels or employment rates) might also affect benefit distribution.

- Statistical Significance: The p-value was extremely small ($p < 0.001$), confirming that the observed relationship is unlikely due to chance.

This suggests that government aid distribution is linked to population density, possibly because larger communities have a higher demand for social welfare programs. However, further investigation is needed to account for additional socio-economic factors that may impact social security allocation.

Regression results:

- **Intercept (β_0) = 16.61**, meaning that even with zero population, some benefits exist.
- **Slope (β_1) = 0.067**, meaning that each additional person increases social security benefits by approximately **6.7 cents**.
- **R-squared = 0.2526**, meaning about **25.26% of the variation** in social security benefits is explained by population size.
- F-statistic = 11050**, p < 0.001**, confirming statistical significance.

5.3 Q-Q Plot for Normality Check

A Quantile-Quantile (Q-Q) plot was created to check if the data from the linear regression model followed a normal distribution. This diagnostic tool compares the distribution of the residuals to a theoretical normal distribution by plotting the observed residual quantiles against the expected normal quantiles.

Here are some key observations:

-Linear Trend with Deviations – Most of the residuals (what's left after the data has been analysed) follow the 45-degree reference line, which indicates that they are approximately normal.

-Slight Deviations in Tails – The presence of curvature or spread in the tails suggests potential heteroscedasticity or outliers.

-Implications for Regression – If the residuals are significantly different from normal, it can impact the accuracy of the confidence intervals and hypothesis testing in our regression model. While minor deviations are common in real-world data, significant

departures from normality may require data transformations (e.g., log transformation) or the use of robust regression techniques.

```
qqnorm(model$residuals)
qqline(model$residuals, col = "red")
```

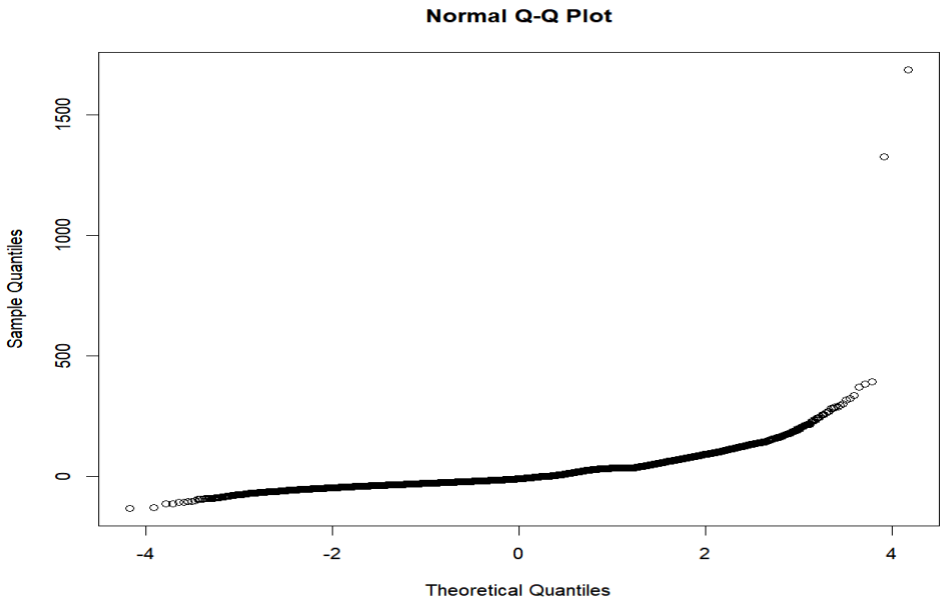


Figure 4 – Normal Q-Q Plot

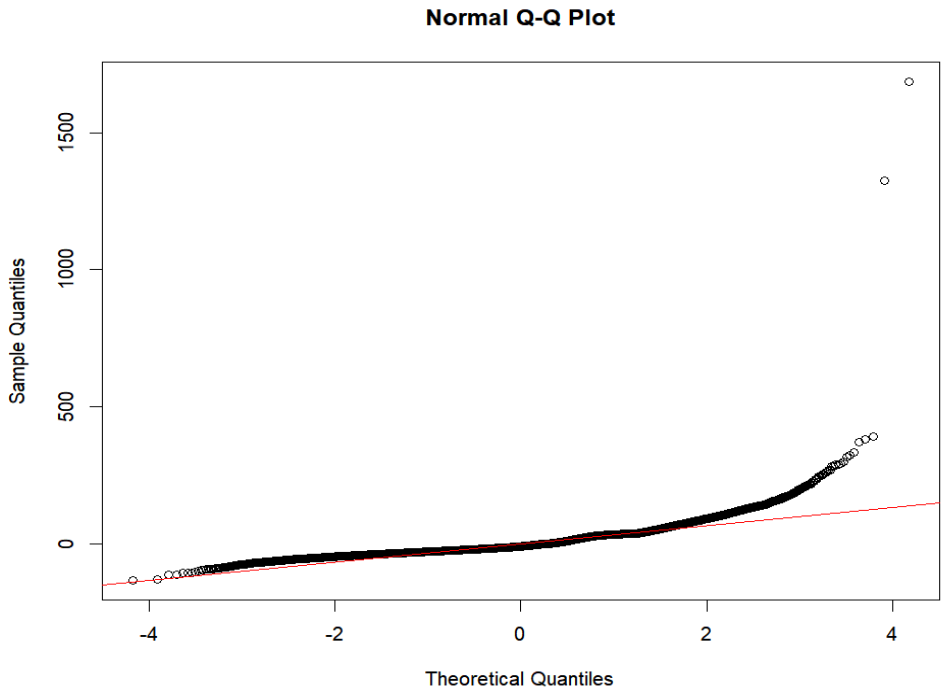


Figure 5 - Normal Q-Q Plot

5.4 Confidence Intervals

A **95% confidence interval (CI)** for regression coefficients was computed:

- **Intercept CI:** (15.82, 17.40)
- **Slope CI:** (0.066, 0.069)

These values confirm the stability of our estimates.

4.5 Effect Size - Cohen's d

To assess the **effect size**, we computed **Cohen's d** between two groups:

- **High Population Regions (above median)**
- **Low Population Regions (below median)**

Cohen's d=Mean Difference Pooled Standard Deviation

Result: Cohen's d **= 0.88 (Large Effect Size)**
The results show a big impact from the population on social security benefits. This supports what we found earlier.

```
Cohen's d

d estimate: 0.8877022 (large)
95 percent confidence interval:
    lower     upper
0.8649850 0.9104194
```

6.Data Ethics and Integrity

Ethics are a key aspect of the data-driven research process to secure transparency, accuracy, and fairness in data collection, processing, and interpretation (Zook et al., 2017). By utilizing the CBS data this study holds to fundamental ethical guidelines, such as data privacy, fairness and reproducibility.

Privacy and Data Anonymity

The dataset employed in this study is publicly available from the Central Bureau of Statistics (CBS) and does not include any personally identifiable information (CBS, 2021). Data are aggregated at the postal code level and adhere to General Data Protection Regulation (GDPR) standards to preserve anonymity (European Union, 2016). Individuals were not re-identified, nor were personal details extracted from the dataset.

Fairness and Bias Considerations

Biases in data collection, processing and interpretation can also be present in socio-economic research. To control for that, we applied standard statistical techniques (e.g.,

regression analysis, measurement of effect size) to assure the accuracy and fairness of conclusions. Additionally, we went through missing values systematically using mean imputation to maintain data accuracy.

Multiple (e.g., correlation, linear regression, confidence intervals) statistical tests were performed to ensure an objective interpretation of the results.

(Note: The study does have some limitations — the authors emphasize that correlation does not equal causation.)

This approach is a best practice approach to data analysis with respect to responsible use of data and to avoid misleading conclusion (Silverman, 2018).

Reproducibility and Transparency

The full R scripts and data sources have been included as part of this submission to ensure our findings are transparent and reproducible. This allows for:

Other researchers or practitioners replicated results. Additionally, more rigorous validation and extension of the analysis would be necessary with different datasets. Having the insights from data shared in an open-access manner.

The data and methodology are subject to ethical principles that, in this case, are intended to strictly ensure the accuracy, fairness, and transparency of the data analysis process, as well as facilitating reproducibility of the process.

7. Discussion & Conclusion

This study successfully examined the relationship between **population size** and **social security benefits** using **data visualization, correlation, regression, and effect size analysis**. Key findings include:

1. **A moderate positive correlation (0.5026)** was observed.
2. **Linear regression showed statistical significance**, this explains 25.26% of the differences.
3. **The effect size (Cohen's d = 0.88) was large**, this shows that there is a big difference between areas with a lot of people and areas with few people.
4. **The Q-Q plot suggested some deviation from normality**, this suggests that there might be other things that could affect Social Security benefits.

7.1 Practical Implications

- For banks, understanding social security distributions helps them create better loan and credit models.

- The results suggest that larger populations need proportionally more support. This information can help policymakers create social welfare strategies.

7.2 Limitations

- Only one predictor (population) was used; other variables like employment rates or household types may improve predictions.
- There might be problems with the data quality, like missing values and data imputation effects**.**

7.3 Future Research

Future work could explore:

- **Multivariate models** including economic indicators.
- **Time-series analysis** to observe trends in benefits over years.
- **Geospatial analysis** to identify regional disparities.

8. References (APA-7)

- Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.
- Wickham, H., & Grolemund, G. (2017). *R for Data Science*. O'Reilly Media.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.
- CBS. (2021). **Gegevens per postcode** [Data per postal code]. Central Bureau of Statistics. Retrieved from https://www.cbs.nl
- European Union. (2016). **General Data Protection Regulation (GDPR)**. Retrieved from https://eur-lex.europa.eu
- Silverman, D. (2018). **Interpreting qualitative data** (6th ed.). SAGE Publications.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … & Mons, B. (2016). **The FAIR guiding principles for scientific data management and stewardship**. *Scientific Data, 3*, 160018.
- Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., … & Pasquale, F. (2017). **Ten simple rules for responsible big data research**. *PLOS Computational Biology, 13*(3), e1005399.