Population and Income Data in Rio Grande do Sul (2010)
Exploratory Data Analysis (EDA)

Note: used RStudio to code the data, code will be attached at the very bottom of the document

Nature of the Dataset

The dataset used for this analysis comes from the 2010 Brazilian Census, which provides demographic and economic information about Brazilian municipalities. The specific subset of data selected for this study focuses on Rio Grande do Sul (RS), in Brazil.

Key Features of the Dataset:

- code_muni: Unique identification code for municipalities
- abbrev_state: State abbreviation
- V6526: Income in all jobs, measured in the number of minimum wages
- V0010: Sample weight for statistical adjustments

This dataset is a combination of structured numerical and categorical data, allowing for geographic, economic, and statistical analysis.

Data

- Timeframe: The dataset represents data collected in the 2010 Census
- Geographical Coverage: The dataset contains data from all municipalities in Brazil, but the analysis is limited to Rio Grande do Sul (RS)
- Entities Represented: Municipal-level economic data is included, meaning the unit of analysis is each municipality

Data Cleaning

The glimpse(jobs) function provides an overview of the dataset, confirming the presence of key variables and their data types. The dataset is structured in a tidy format, making it ready for processing. To ensure data quality, I checked for missing values using colSums(is.na(jobs)). The presence of NA values can distort analysis, so missing income values (V6526) are handled using na.rm = TRUE when computing summary statistics. Since the dataset includes all Brazilian municipalities, I filtered it to only include data from RS.

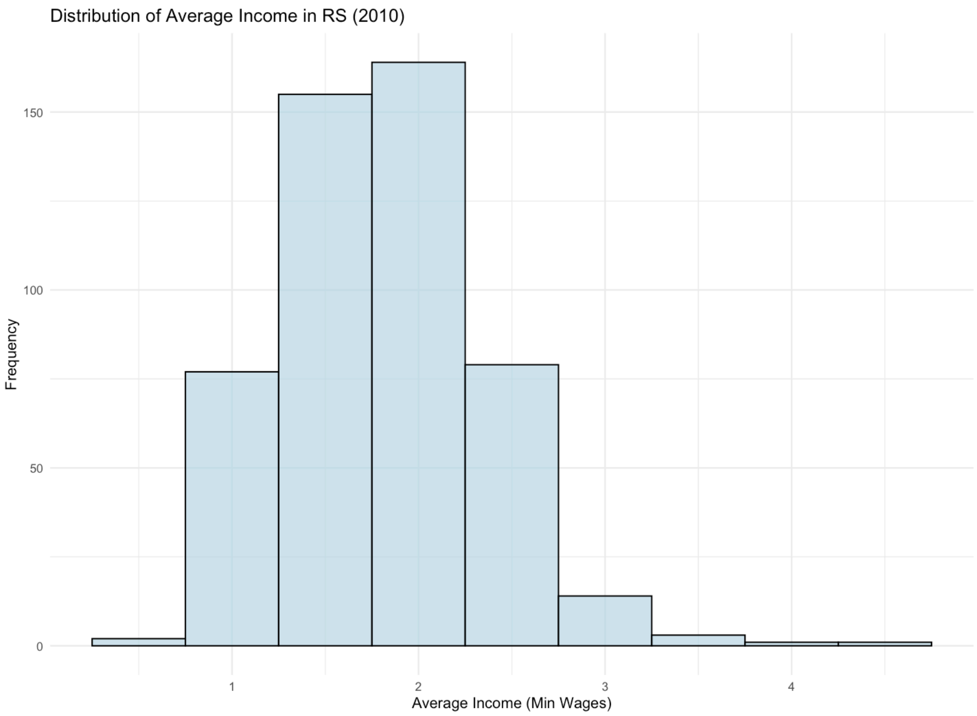
Key Findings from the EDA

Using summary(muni_income$avg_income), I generated key insights about income levels in RS:

- Mean Income: The average income in municipalities varies across RS
- Median Income: Since income distributions tend to be skewed, the median offers a better representation of central tendency
- Standard Deviation: A high standard deviation suggests significant income disparity across different municipalities

Visualization and Insights

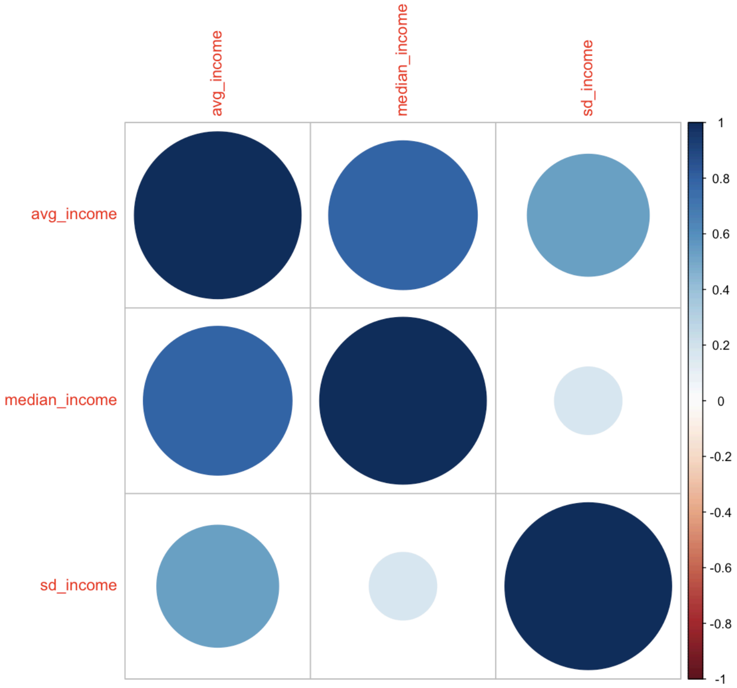
Figure A.



Income Distribution Observations: *(figure A.)*

- The histogram shows that most municipalities have low average incomes, with a right-skewed distribution.
- A few municipalities exhibit higher income levels, possibly indicating economic hubs like Porto Alegre.

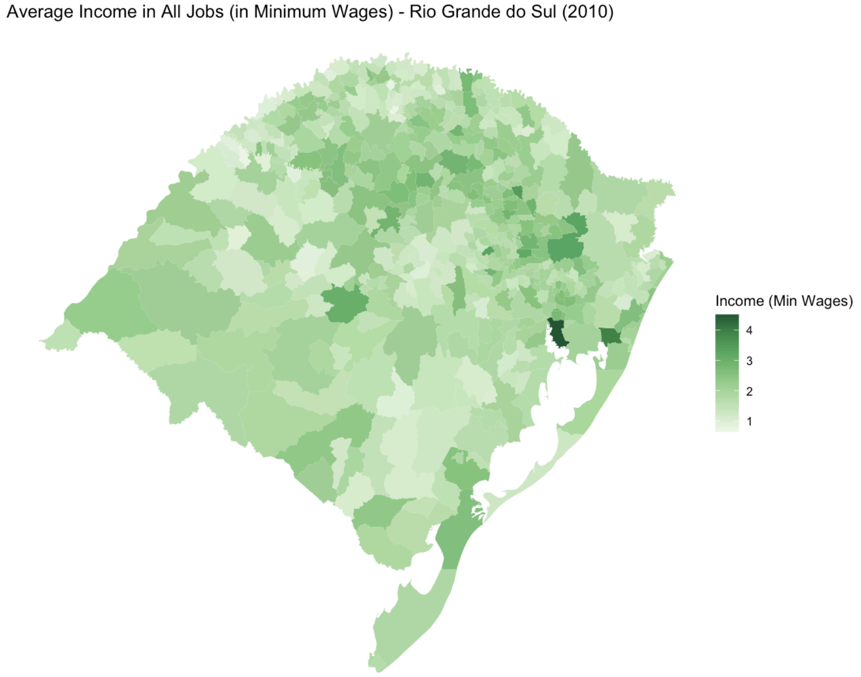
Figure B.



Correlation Analysis Takeaways: *(figure B.)*

- Strong correlation between median and mean income, indicating that in most municipalities, income levels are fairly consistent.
- High standard deviation in some areas suggests economic inequality, meaning some municipalities have a few individuals with very high incomes, pulling the average up.

Figure C.


Average Income in All Jobs (in Minimum Wages) - Rio Grande do Sul (2010)
Income (Min Wages)
4
3
2
1

Geospatial Analysis/Choropleth Map Key Observations: *(figure C.)*

- Urban municipalities tend to have higher incomes, with cities like Porto Alegre showing significantly higher values.
- Rural municipalities have lower income levels, highlighting economic disparities.
- Clusters of higher-income municipalities may correspond to industrial zones or areas with higher economic activity.

Challenges in the Data

Challenges of the data included missing values, some records had missing income data (V6526). While ignoring the missing values (na.rm = TRUE) avoids distortion, further exploration is still needed. There is a high standard deviation which suggests significant inequality in economic disparity. Further study can compare income with education or employment data. There was a skewed distribution which means a few municipalities significantly exceeded the average, using log transformations may help interpret the income disparities more effectively.

Further Exploration

Advanced statistical analysis conducting regression analysis to understand factors influencing income and comparing income distributions over multiple census years. Additional data sources on education and employment data to find the factors contributing to income differences. And, a time series analysis to explore how income distribution has evolved over time.

Summary

Income distribution is highly skewed, with some municipalities earning significantly more than others. Higher-income municipalities tend to be urban centers, while rural areas exhibit lower wages. Correlation analysis suggests that municipalities with high mean income also have high median income, but there is a variation due to inequality. Geospatial analysis confirms regional disparities, emphasizing urban-rural income gaps. Data quality issues include missing values and high variance, requiring careful statistical treatment.

```r
### CODE

library(librarian)
# using librarian to install and load packages
librarian::shelf(censobr, geobr, sf, tidyverse, dplyr, ggplot2, arrow, viridis,
corrplot)

# downloading the population dataset
jobs <- read_population(year = 2010,
                        columns = c(
                          'code_muni', # municipalities
                          'abbrev_state', # state abbreviation
                          'V6526', # income in all jobs in number of minimum wages
                          'V0010' # sample weight
                        ),
                        add_labels = 'pt', # using 'pt' for Portuguese labels
                        showProgress = FALSE,
                        as_data_frame = TRUE,
                        cache = TRUE
)

# checking the structure of the dataset
dplyr::glimpse(jobs)

# checking for missing values
colSums(is.na(jobs))

# filtering dataset to Rio Grande do Sul
muni_income <- jobs %>%
  filter(abbrev_state == 'RS') %>%
  group_by(code_muni) %>%
  summarise(
    avg_income = mean(V6526, na.rm = TRUE),
    median_income = median(V6526, na.rm = TRUE),
    sd_income = sd(V6526, na.rm = TRUE),
    count = n()
  )

# display summary statistics
summary(muni_income$avg_income)

# plotting histogram of income distribution
ggplot(muni_income, aes(x = avg_income)) +
  geom_histogram(binwidth = 0.5, fill = "lightgreen", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Average Income in RS (2010)",
       x = "Average Income (Min Wages)",
       y = "Frequency") +
  theme_minimal()

# checking correlations between variables
cor_matrix <- cor(muni_income %>% select(avg_income, median_income, sd_income), use =
"complete.obs")

# visualizing correlation matrix
corrplot(cor_matrix, method = "circle")

# getting municipality boundaries
municipalities <- read_municipality(code_muni = "RS", year = 2010)

# merging geospatial and jobs income data
rsincome <- municipalities %>%
  left_join(muni_income, by = c("code_muni" = "code_muni"))

# plotting the results in a choropleth map
ggplot() +
  geom_sf(data = rsincome, aes(fill = avg_income), color = NA) +
  scale_fill_distiller(palette = "Greens", direction = 1) +
  theme_void() +
  labs(
    title = "Average Income in All Jobs (in Minimum Wages) - Rio Grande do Sul (2010)",
    fill = "Income (Min Wages)"
  )
```