STAT 847: Analysis Assignment 3

DUE: Saturday April 8, 2024 by 11:59pm Eastern

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark. Organization and comprehensibility is part of a full solution. Consequently, points will be deducted for solutions that are not organized and incomprehensible. Furthermore, if you submit your assignment to Crowdmark, but you do so incorrectly in any way (e.g., you upload your Question 2 solution in the Question 1 box), you will receive a 5% deduction (i.e., 5% of the assignment's point total will be deducted from your point total).

There are a total of 40 marks.

1. (2 points) Use the Stats Canada API (that is, the `cancensus` package) to get a list of all the vectors/variables publicly available from the 2021 census (`CA21`). Report the number of such variables.

```r
list_census_datasets()

ca21 <- list_census_vectors("CA21")

num_vectors <- length(ca21$vector)
print(num_vectors)
```

There are 7709 unique variable vectors.

2. (2 points) Use the list of all variables to find the vectors for "`Total Employment income (%)`" and "`Total Market income (%)`", not split by sex.

```
TEI_data <- find_census_vectors("Employment income (%)", dataset = "CA21", type="Total", query_type = "s

TMI_data <- find_census_vectors("Market income (%)", dataset = "CA21", type = "total", query_type = "ser
```

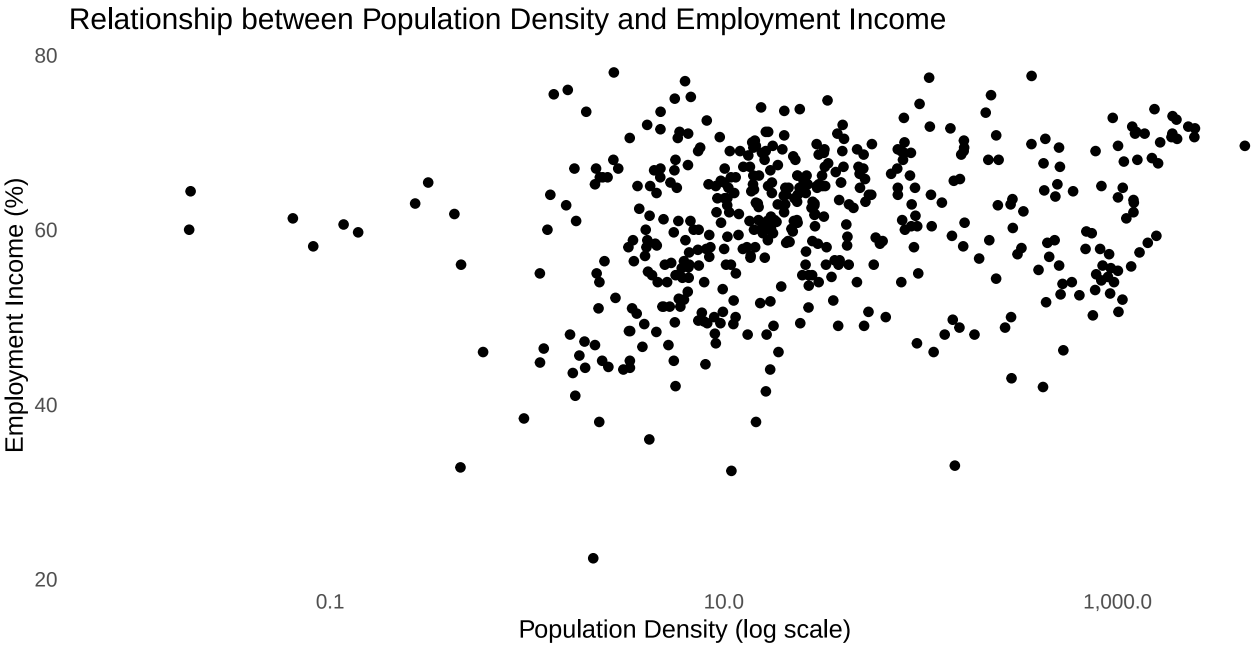
Therefore the vectors for "`Total Employment income (%)`" and "`Total Market income (%)`", not split by sex are v_CA21_650 and v_CA21_647 respectively.

3. (6 points) Get the census data for all of the Census Subdivisions of Ontario for "`Total Employment income (%)`" and "`Total Market income (%)`". Include your code and report the dimensions of the dataset you get.

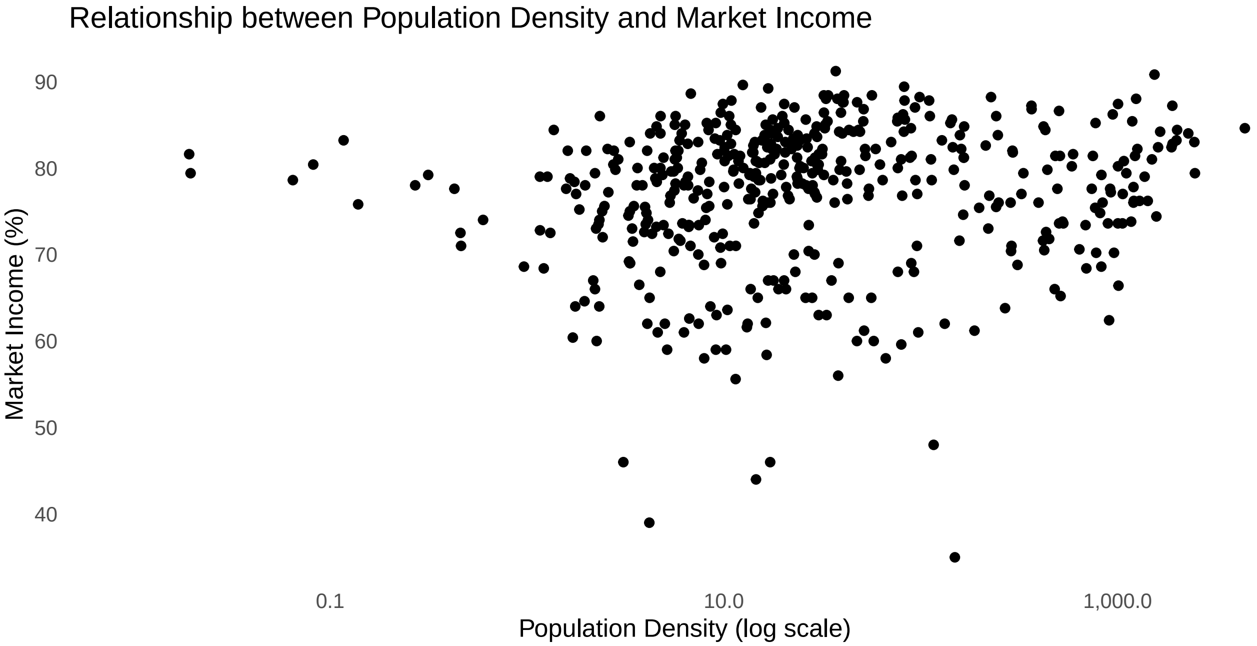
```
Q3_census_data <- get_census(dataset='CA21', regions=list(PR="35"), vectors = c("v_CA21_650", "v_CA21_64

Q3_census_data

dim_Q3 <- dim(Q3_census_data)
```

The dimensions of the census data for all of the Census Subdivisions of Ontario for "`Total Employment income (%)`" and "`Total Market income (%)`" are 577, 12

4. (6 points) Make a scatterplot in ggplot in which the x axis is the population density (population divided by area (sq km)), and the y axis is employment income (%). Use an appropriate title and labels. You may have to rename some variables with `names(census_data)` to remove spaces. Use a log-scale for density (hint: `scale_x_continuous`).



5. (2 points) Make another scatterplot in ggplot in which the x axis is the population density (population divided by area (sq km)), and the y axis is market income (%). Use an appropriate title and labels. Use a log-scale for density (hint: `scale_x_continuous`).



6. (4 points) Consider that market income is employment income plus self-employment income from market activity (e.g., sales of farm goods). Use this information and the two previous graphs to explain the income differences between high-density and low-density sub divisions of Ontario.

Lower population density areas tend to exhibit lower employment income due to reliance on self-employment, notably in agriculture. This trend is evident in the first graph, where we see lower employment income percentage in the lower population density areas.

However, the second graph unveils that low-density areas compensate with higher percentages of market income, indicating supplementary income from self-employment activities like agricultural sales. Notably, low market income percentages aren't observed until a population density of approximately 9 on the log scale, suggesting a balance between employment and market income even in low-density regions.

These insights underscore the intricate dynamics between population density, employment, and market income, highlighting the role of self-employment in mitigating income disparities across Ontario's subdivisions

7. (2 points) In the eurodata package, use the browseDataList function to find the code name for the dataset (not the table) “Volume of passenger transport relative to GDP”. Report the code name.

```r
library(eurodata)

browseDataList()
```

After using the browseDatalist function and finding the “Volume of passenger transport relative to GDP” dataset, I found that the code name is ‘tran_hv_pstra’.

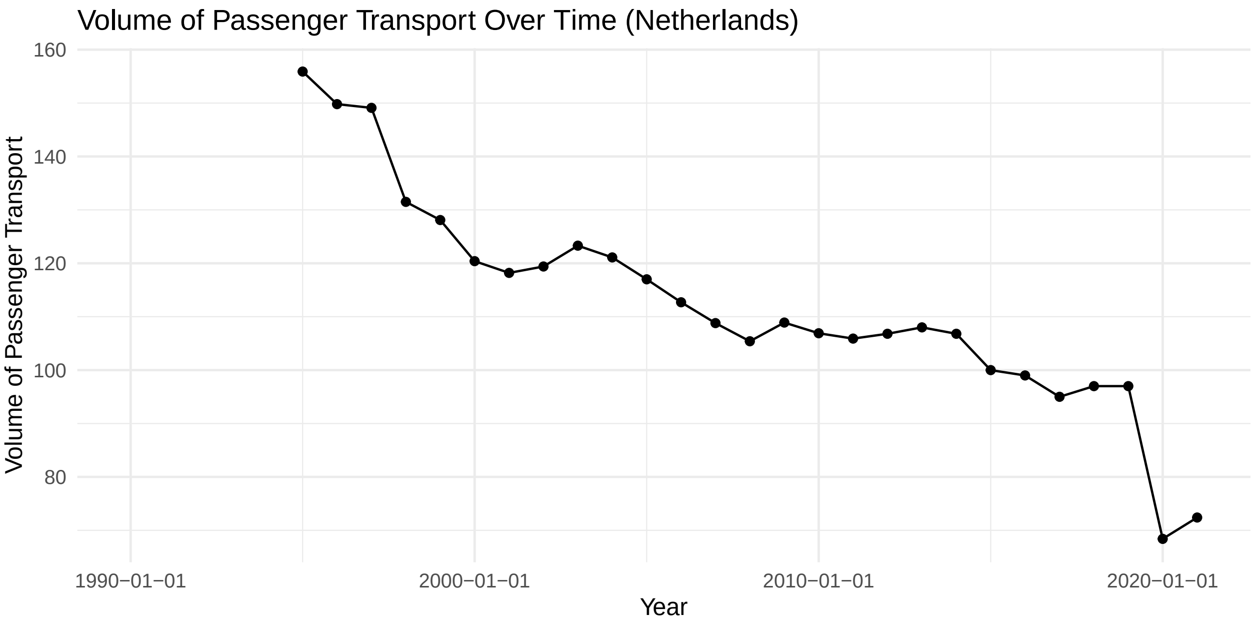
8. (2 points) Use the `eurodata` package to import and save the “Volume of passenger transport relative to GDP” dataset. Report the dimensions of the data.

```
library(eurodata)

Q8_dta <- importData('tran_hv_pstra')

dim_Q8 <- dim(Q8_dta)
```

The dimensions of the “Volume of passenger transport relative to GDP” dataset” are 1184, 6

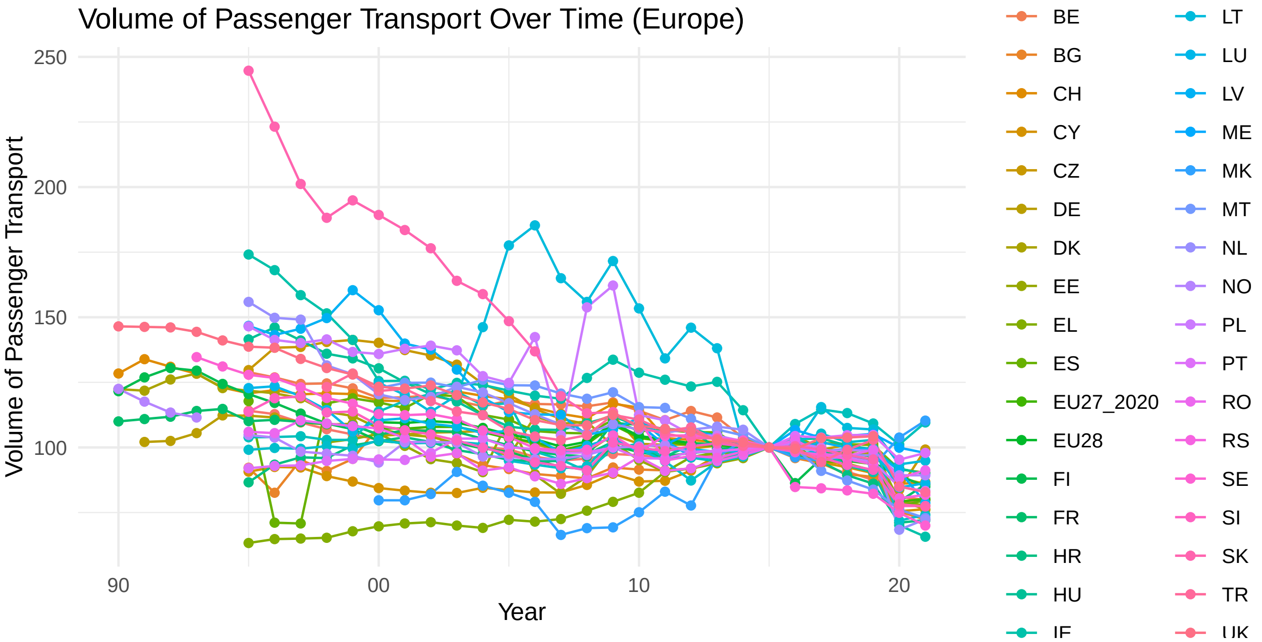
9. (5 points) Use ggplot and plot a line plot of the `value_` (that is, the volume of passenger transport) over `TIME_PERIOD`, using only the `NL` geography (that is, only use the Netherlands). Interpret `TIME_PERIOD` as a date to make the x-axis cleaner. Use an appropriate title and labels.

```r
library(hrbrthemes)

netherlands_data <- Q8_dta %>%
  filter(geo == "NL")

netherlands_data$TIME_PERIOD <- as.Date(paste(netherlands_data$TIME_PERIOD, "-01-01", sep = ""))

ggplot(netherlands_data, aes(x = TIME_PERIOD, y = value_)) +
  geom_line() +
  geom_point() +
  labs(title = "Volume of Passenger Transport Over Time (Netherlands)",
       x = "Year",
       y = "Volume of Passenger Transport") +
  scale_x_date(date_labels = "%Y-%m-%d") +
  theme_minimal()
```



10. (3 points) Use ggplot and plot a set of lines of `value_` over `TIME_PERIOD`, where each `geo` is a distinct line. Use only the last two digits of year.

```r
library(ggplot2)

Q8_dta$TIME_PERIOD <- as.Date(paste(Q8_dta$TIME_PERIOD, "-01-01", sep = ""))

ggplot(Q8_dta, aes(x = TIME_PERIOD, y = value_, color = geo)) +
  geom_line() +
  geom_point() +
  labs(title = "Volume of Passenger Transport Over Time (Europe)",
       x = "Year",
       y = "Volume of Passenger Transport") +
  scale_x_date(date_labels = "%y") +
  theme_minimal()
```



11. (4 points) Comment on the prevailing trends found in the previous two plots.

In the plot “Volume of Passenger Transport Over Time (Netherlands),” a significant downward trend in the volume of passenger transport in the Netherlands is evident. Notably, there are prominent decreases in volume observed from 1997 to 1998 and from 2019 to 2020. Research into the years 1997 and 1998 reveals notable events such as the 1998 Dutch general election and the 1998 FIFA World Cup. One possibility is that dissatisfaction with election outcomes or the allure of the World Cup led to a portion of the population either leaving the country, thereby reducing participation in passenger transport, or temporarily staying in neighboring countries, particularly given the proximity of the Netherlands to France. The decrease observed in 2019 to 2020 is due to the lock down from the COVID-19 Pandemic.

In the plot “Volume of Passenger Transport Over Time (Europe),” a considerable variance in passenger transport volume is observed during the early years (1990-2007). The plot indicates that the volume of passenger transport across all countries converges around 2015, suggesting a baseline for comparison. Notable observations include Spain (ES) exhibiting a low volume of passenger transport in 1995 and 1996, Italy showing a strong upward trend from 2002 to 2006, and the UK having the highest volume of passenger transportation from 1994 to 2004, followed by a significant decreasing trend. Additionally, similar to the previous plot, a decrease in passenger transportation volume across all countries is observed due to the COVID-19 pandemic.