Final Project STAT206A

RAMEEN JALIL, 251710245

2024-06-16

Loading Packages

Reading data sets

Each dataset was imported into Excel as separate sheets. This ensures that the data from each source is correctly represented within the same workbook.

```
exceldata<-read_excel("/Users/rafeiamunir/Final Project Spring/Counties Data Set 1.xlsx")
View(exceldata)

exceldata2<-read_excel("/Users/rafeiamunir/Final Project Spring/Counties Data Set 2.xlsx")
View(exceldata2)

exceldata3<-read_excel("/Users/rafeiamunir/Final Project Spring/Counties Data Set 3.xlsx")
View(exceldata3)
```

Combining the data

bind_rows is a function from dplyr that combines multiple data frames (or tibbles) by rows.

```
library(dplyr)
combined_data <- bind_rows(exceldata,exceldata2,exceldata3)
```

Checking for missing values

A check for missing values was performed using the **colSums function**. This helped in identifying any missing data that needs to be addressed in subsequent analyses.

```
missing_values <- colSums(is.na(combined_data))
print(missing_values)
```

```
##             county              state        pop.density                pop
##               1014               1014               1014                  0
##         pop.change            age6574              age75              crime
##               1019                  0                  3               1023
##            college             income               farm           democrat
##                  0                  4               1019               1045
##         republican              white              black            turnout
##                 27                  0                  0                  3
##         pop change              CRIME               FARM          Democrate
##               2122               2122               2122               2123
##             COUNTY              STATE population density
##               2127               2127               2127
```

Missing values can either be removed, replaced, or kept as it is. I don't think any missing values should be removed from this data set, as that will cause omission of the entire information of a county, making it *impossible to analyse those areas, making our analysis incomplete and problematic.*

Imputing Missing Values with Mean

The three datasets were successfully combined into a single data frame, and the percentage of missing values for each variable was calculated.

```
combined_data <- combined_data %>%
  mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

str(combined_data)
```

```
## tibble [3,141 x 23] (S3: tbl_df/tbl/data.frame)
##  $ county            : chr [1:3141] "Autauga" "Baldwin" "Barbour" "Bibb" ...
##  $ state             : chr [1:3141] "AL" "AL" "AL" "AL" ...
##  $ pop.density       : num [1:3141] 61 67 29 28 62 18 28 191 62 36 ...
##  $ pop               : num [1:3141] 34222 98280 25417 16576 39248 ...
##  $ pop.change        : num [1:3141] 11.9 35.4 2 9.2 10.6 ...
##  $ age6574           : num [1:3141] 5.7 9.2 8.2 6.7 7.4 ...
##  $ age75             : num [1:3141] 4.1 6 6.4 6 5.6 ...
##  $ crime             : num [1:3141] 4996 3329 3192 0 2052 ...
##  $ college           : num [1:3141] 14.5 16.8 11.8 4.7 7 ...
##  $ income            : num [1:3141] 32240 30199 23838 23714 26323 ...
##  $ farm              : num [1:3141] 1.8 1.7 2.4 0.9 4.7 ...
##  $ democrat          : num [1:3141] 30.9 26.2 46.4 43.2 32.9 ...
##  $ republican        : num [1:3141] 55.9 56.5 42.9 46.5 53.8 ...
##  $ white             : num [1:3141] 79.3 86 55.5 78.7 97.8 ...
##  $ black             : num [1:3141] 20 12.86 44.04 20.98 1.33 ...
##  $ turnout           : num [1:3141] 45.5 47.3 41 40.5 42.1 ...
##  $ pop change        : num [1:3141] 3.74 3.74 3.74 3.74 3.74 ...
##  $ CRIME             : num [1:3141] 2873 2873 2873 2873 2873 ...
##  $ FARM              : num [1:3141] 7.52 7.52 7.52 7.52 7.52 ...
##  $ Democrate         : num [1:3141] 38.8 38.8 38.8 38.8 38.8 ...
##  $ COUNTY            : chr [1:3141] NA NA NA NA ...
##  $ STATE             : chr [1:3141] NA NA NA NA ...
##  $ population density: num [1:3141] 201 201 201 201 201 ...
```

```
missing_values_after_imputation <- colSums(is.na(combined_data))
print(missing_values_after_imputation)
```

```
##             county              state        pop.density                pop
##               1014               1014                  0                  0
##         pop.change            age6574              age75              crime
##                  0                  0                  0                  0
##            college             income               farm           democrat
##                  0                  0                  0                  0
##         republican              white              black            turnout
##                  0                  0                  0                  0
##         pop change              CRIME               FARM          Democrate
##                  0                  0                  0                  0
##             COUNTY              STATE population density
##               2127               2127                  0
```

Identifying Outliers

Outliers were identified using the IQR method.

```
identify_outliers <- function(data) {
  Q1 <- quantile(data, 0.25, na.rm = TRUE)
  Q3 <- quantile(data, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  return(data < lower_bound | data > upper_bound)
}

outliers <- combined_data %>%
  select(where(is.numeric)) %>%
  summarise_all(identify_outliers)
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## i The deprecated feature was likely used in the dplyr package.
##   Please report the issue at <https://github.com/tidyverse/dplyr/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
outlier_counts <- colSums(outliers)
print(outlier_counts)
```

```
##        pop.density                pop         pop.change            age6574
##                122                379                431                 84
##              age75              crime            college             income
##                 82                285                202                131
##               farm           democrat         republican              white
##                356                442                 35                148
##              black            turnout         pop change              CRIME
##                412                 65               1019               1019
##               FARM          Democrate population density
##               1019               1018                104
```

Outliers too in this case, are necessary to be kept because they can provide valuable insight and removing them may cause us to have an incomplete analysis. Moreover, if we trim certain data, it will not be as useful, as the outliers in this case are very large, as we can see from the boxplots. By trimming in one variable for e.g., we will remove data of a whole county and a thorough analysis will not be possible.

Handling Outliers with Capping

I will cap the outliers in the numeric columns to the 1.5 IQR range limits.

```
cap_outliers <- function(data) {
  Q1 <- quantile(data, 0.25, na.rm = TRUE)
  Q3 <- quantile(data, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  data <- ifelse(data < lower_bound, lower_bound, data)
  data <- ifelse(data > upper_bound, upper_bound, data)
  return(data)
```

```r
}

combined_data <- combined_data %>%
  mutate(across(where(is.numeric), cap_outliers))
```

Capping outliers helps maintain the accuracy and reliability of statistical analyses and models by reducing the impact of extreme values. Outliers can distort measures like the mean and standard deviation, skewing results and potentially misleading interpretations. By capping these values within a defined range (1.5 times the interquartile range), I can ensure that our data analysis remains robust and that visualizations accurately reflect underlying patterns without being unduly influenced by extreme deviations.

Calculating the five value summary

I have calculated the five-value summary (minimum, maximum, mean, median, and standard deviation) for all continuous variables in the combined dataset.

```r
five_value_summary <- function(data) {
  summary_stats <- data %>%
    select(where(is.numeric)) %>%
    summarise_all(list(
      Min = ~ min(., na.rm = TRUE),
      Max = ~ max(., na.rm = TRUE),
      Mean = ~ mean(., na.rm = TRUE),
      Median = ~ median(., na.rm = TRUE),
      SD = ~ sd(., na.rm = TRUE)
    ))

  summary_stats_long <- summary_stats %>%
    pivot_longer(everything(), names_to = c("Variable", "Statistic"), names_sep = "_") %>%
    pivot_wider(names_from = "Statistic", values_from = value)

  return(summary_stats_long)
}

summary_table <- five_value_summary(combined_data)
print(summary_table)
```

```
## # A tibble: 19 x 6
##    Variable                Min       Max     Mean   Median       SD
##    <chr>                 <dbl>     <dbl>    <dbl>    <dbl>    <dbl>
##  1 pop.density             0        542.     140.       93      132.
##  2 pop                    52     121384.   39168.   22085    39134.
##  3 pop.change            -16.5        23.2      5.42     7.83    10.4
##  4 age6574                 3.00       13.4      8.25     8.20     2.08
##  5 age75                   0.650      12.3      6.55     6.30     2.32
##  6 crime                   0        5563     2898.    3071.    1465.
##  7 college                 0          25.2     13.1     11.8      5.28
##  8 income              12022.     43530.   28214.   27361     6371.
##  9 farm                    0          10.9      5.18     5.91     3.14
## 10 democrat               26.5        52.5     39.8     40.2      6.89
## 11 republican             16.7        62.7     39.8     39.3      8.45
## 12 white                  53.5       100       87.8     94.1     13.8
## 13 black                   0          24.8      6.49     1.50     8.90
## 14 turnout                23.9        64.2     43.9     44.1      7.53
## 15 pop change              3.74        3.74     3.74     3.74     0
```

```
## 16 CRIME              2873.      2873.    2873.    2873.      0
## 17 FARM                  7.52       7.52     7.52     7.52    0
## 18 Democrate            38.8       38.8     38.8     38.8     0
## 19 population density    0        339.     163.     201.      81.8
```