Project 1: Wrangling, Exploration, Visualization

SDS322E

Data Wrangling, Exploration, Visualization

Filina Nurcahya-Tjoa UTEID : fnt226

Introduction Paragraph or two introducing your data sets and variables, why they are interesting to you, etc.

My first data set contains the minimum wages for 54 different states throughout a period lasting from 1968 to 2020. My second data set shows COVID-19 statistics for 51 different states throughout 2020. It has many variables such as the number of deaths, people tested, etc. I am interested in this data because I am curious to see how COVID-19 has affected us all differently based on the location we reside in. For my project, I'm going to be investigating the relationship between the minimum wage of each state and how its residents were affected by COVID-19. From this, I am hoping to find out how the financial situation of a location influences the survivability of COVID-19.

From my two datasets, I'm going to be focusing on a total of five different variables. Namely, the number of tests, number of deaths, number of infections, the population size, and the minimum wage of each state. I feel like these variables would best give me a snapshot of the information I am looking for. I also indexed out the minimum wages data set so that it only had information for 2020 (which is when a majority of the duration of the pandemic occurred). I expect to find an inversely proportionate relationship between the numbers of death/infection and the minimum wage rate. I predict this outcome as people would have more money to spend on health care and have better access to resources.

```
# Importing the First Dataset and Indexing Desired
# Variables.
MinimumWageDataSet <- read.csv("Minimum Wage Data (1).csv")
MinimumWageDataSet <- MinimumWageDataSet[, c(1, 2, 7)]

# Importing the First Dataset and Indexing Desired
# Variables.
COVID19Dataset <- read.csv("COVID19_state.csv")
COVID19Dataset <- COVID19Dataset[, c(1:5)]

# Indexing Out the Desired Rows with the Desired Year.
MinimumWageDataSet <- MinimumWageDataSet[MinimumWageDataSet$Year ==
    2020, ]
```

Tidying: Reshaping If your data sets are tidy already, demonstrate that you can reshape data with pivot wider/longer here (e.g., untidy and then retidy). Alternatively, it may be easier to wait until the wrangling section so you can reshape your summary statistics. Note here if you are going to do this.

```
library(tidyverse)
library(stringr)
```

```r
# Joining Datasets based on ID.
combineddataset <- inner_join(COVID19Dataset, MinimumWageDataSet,
    by = "State")

# Making the data set longer.
untidydataset <- combineddataset %>%
    pivot_longer(2:7, names_to = "Catergories", values_to = "Numbers")

untidydataset %>%
    glimpse()
```

```
## Rows: 306
## Columns: 3
## $ State       <chr> "Alaska", "Alaska", "Alaska", "Alaska", "Alaska", "Alaska"~
## $ Catergories <chr> "Tested", "Infected", "Deaths", "Population", "Year", "Eff~
## $ Numbers     <dbl> 620170.00, 17057.00, 84.00, 734002.00, 2020.00, 10.19, 135~
```

```r
# Making the data set wider (back to original data set).
retidieddataset <- untidydataset %>%
    pivot_wider(names_from = Catergories, values_from = Numbers)

retidieddataset %>%
    glimpse()
```

```
## Rows: 51
## Columns: 7
## $ State                 <chr> "Alaska", "Alabama", "Arkansas", "Arizona", "Ca~
## $ Tested                <dbl> 620170, 1356420, 1363429, 1792602, 18912501, 20~
## $ Infected              <dbl> 17057, 194892, 113641, 248139, 930628, 109910, ~
## $ Deaths                <dbl> 84, 2973, 1985, 5982, 17672, 2105, 4627, 647, 7~
## $ Population            <dbl> 734002, 4908621, 3038999, 7378494, 39937489, 58~
## $ Year                  <dbl> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020,~
## $ Effective.Minimum.Wage <dbl> 10.19, 7.25, 10.00, 12.00, 13.00, 12.00, 11.00,~
```

```r
# Unique IDs in Each Dataset.
COVID19Dataset %>%
    select(State) %>%
    summarise_all(n_distinct)
```

Joining/Merging

```
##   State
## 1    51
```

```r
MinimumWageDataSet %>%
    select(State) %>%
    summarise_all(n_distinct)
```

```
##   State
## 1    54
```

```r
# Joining Datasets based on ID.
combineddataset <- inner_join(COVID19Dataset, MinimumWageDataSet,
    by = "State")

inner_join(COVID19Dataset, MinimumWageDataSet, by = "State") %>%
    summarise(count = n())
```

```
##   count
## 1    51
```

```r
# Seeing which were dropped from the Dataset and seeing
# which IDs appeared in one but not the other.
anti_join(MinimumWageDataSet, COVID19Dataset, by = "State")
```

```
##   Year               State Effective.Minimum.Wage
## 1 2020                Guam                 8.25
## 2 2020         Puerto Rico                 7.25
## 3 2020 U.S. Virgin Islands                10.50
```

```r
anti_join(MinimumWageDataSet, COVID19Dataset, by = "State") %>%
    summarise(count = n())
```

```
##   count
## 1     3
```

```r
anti_join(COVID19Dataset, MinimumWageDataSet, by = "State")
```

```
## [1] State      Tested     Infected   Deaths     Population
## <0 rows> (or 0-length row.names)
```

```r
anti_join(COVID19Dataset, MinimumWageDataSet, by = "State") %>%
    summarise(count = n())
```

```
##   count
## 1     0
```

```r
# Use all six core dplyr functions.  mutate() to turn
# population and testing Rates into a categorical
# variables.
combineddataset2 <- combineddataset %>%
    mutate(Testing_Rate = Tested/Population) %>%
    rename(Minimum_Wage = "Effective.Minimum.Wage")

combineddataset2 <- combineddataset2 %>%
    mutate(Population_Level = case_when(Population >= median(Population) ~
        "More than Median", Population < median(Population) ~
        "Less than Median")) %>%
```

```
    mutate(Testing_Level = case_when(Testing_Rate >= 0.5 ~ "More than Half",
        Testing_Rate < 0.5 ~ "Less than Half"))

combineddataset2 <- combineddataset2 %>%
    mutate(Death_Rate = Deaths/Population)

# filter() to find average minimum wage in states that have
# more than more than 50% of population tested.
combineddataset2 %>%
    filter(Testing_Level == "More than Half") %>%
    summarise(Average_Minimum_Wage = mean(Minimum_Wage))
```

Wrangling

```
##   Average_Minimum_Wage
## 1             9.873529
```

```
# arrange() to find the top three states with the highest
# population sizes.
combineddataset2 %>%
    arrange(desc(Population)) %>%
    slice(1:3)
```

```
##        State   Tested Infected Deaths Population Year Minimum_Wage Testing_Rate
## 1 California 18912501   930628  17672   39937489 2020        13.00    0.4735526
## 2      Texas  8291703   909257  18097   29472295 2020         7.25    0.2813389
## 3    Florida 10126764   801371  17043   21992985 2020         8.56    0.4604543
##   Population_Level  Testing_Level   Death_Rate
## 1 More than Median Less than Half 0.0004424915
## 2 More than Median Less than Half 0.0006140343
## 3 More than Median Less than Half 0.0007749289
```

```
# select() and summarize() to find mean of only population
# and number of people tested.
combineddataset2 %>%
    select(Population, Tested) %>%
    summarise_all(mean)
```

```
##   Population  Tested
## 1    6496451 2904946
```

```
# group_by() to find count of each state in each population
# level.
combineddataset2 %>%
    group_by(Population_Level) %>%
    summarise(n())
```

```
## # A tibble: 2 x 2
##   Population_Level `n()`
##   <chr>            <int>
## 1 Less than Median    25
## 2 More than Median    26
```

```r
# Using a Stringr function (and Regex) to find number of
# states with a minimum wage of more than two digits.
combineddataset2 %>%
    mutate(Minimum10andUp = str_count(Minimum_Wage, "^[0-9]{2}")) %>%
    group_by(Minimum10andUp) %>%
    summarise(n())
```

```
## # A tibble: 2 x 2
##   Minimum10andUp `n()`
##            <int> <int>
## 1              0    34
## 2              1    17
```

```r
# Create summary statistics of all numeric data. (Using 5
# Unique Functions inside Summarize) Mean of all numeric
# variables.
combineddataset2 %>%
    select(c(2:8)) %>%
    summarise_all(mean)
```

```
##    Tested Infected   Deaths Population Year Minimum_Wage Testing_Rate
## 1 2904946 179626.7 4357.745    6496451 2020     9.103137    0.4740247
```

```r
# IQR of all numeric variables.
combineddataset2 %>%
    select(c(2:8)) %>%
    summarise_all(IQR)
```

```
##    Tested Infected Deaths Population Year Minimum_Wage Testing_Rate
## 1 2581944 161537.5   4333    5785682    0         3.48    0.2385959
```

```r
# Standard deviation of all numeric data.
combineddataset2 %>%
    select(c(2:8)) %>%
    summarise_all(sd)
```

```
##    Tested Infected   Deaths Population Year Minimum_Wage Testing_Rate
## 1 3590449 208077.9 5637.548    7450657    0     2.028774    0.2058035
```

```r
# Median of all numeric data.
combineddataset2 %>%
    select(c(2:8)) %>%
    summarise_all(median)
```

```
##    Tested Infected Deaths Population Year Minimum_Wage Testing_Rate
## 1 1792602   120865   2113    4499692 2020          8.7    0.4468277
```

```r
# Max of all numeric data.
combineddataset2 %>%
    select(c(2:8)) %>%
    summarise_all(max)
```

```
##     Tested Infected Deaths Population Year Minimum_Wage Testing_Rate
## 1 18912501   930628  25838   39937489 2020           14      1.11427
```

```
# Count Number of distinct values in each column.
combineddataset2 %>%
    summarise_all(n_distinct)
```

```
##   State Tested Infected Deaths Population Year Minimum_Wage Testing_Rate
## 1    51     51       51     50         51    1           24           51
##   Population_Level Testing_Level Death_Rate
## 1                2             2         51
```

```
# Using two functions after grouping by a catergorical
# variable.  Average Testing Rate by Population Level.
combineddataset2 %>%
    group_by(Population_Level) %>%
    select(Testing_Rate) %>%
    summarise_all(mean)
```

```
## # A tibble: 2 x 2
##   Population_Level Testing_Rate
##   <chr>                   <dbl>
## 1 Less than Median        0.497
## 2 More than Median        0.452
```

```
# Average Minimum Wage by Testing Level.
combineddataset2 %>%
    group_by(Testing_Level) %>%
    select(Minimum_Wage) %>%
    summarise_all(mean)
```

```
## # A tibble: 2 x 2
##   Testing_Level  Minimum_Wage
##   <chr>                 <dbl>
## 1 Less than Half         8.72
## 2 More than Half         9.87
```

```
# Count of states with high/low testing rates and
# population level. (Summarizing after Grouping by Two
# Variables)
combineddataset2 %>%
    group_by(Testing_Level, Population_Level) %>%
    summarise(n())
```

```
## # A tibble: 4 x 3
## # Groups:   Testing_Level [2]
##   Testing_Level  Population_Level `n()`
##   <chr>          <chr>            <int>
## 1 Less than Half Less than Median    17
## 2 Less than Half More than Median    17
## 3 More than Half Less than Median     8
## 4 More than Half More than Median     9
```

```r
# Compute a Summary Statistic with a User-Defined Function.
proportion <- function(x) (x/sum(x))
combineddataset2 %>%
    summarise(ProportionPop = proportion(Population)) %>%
    head()
```

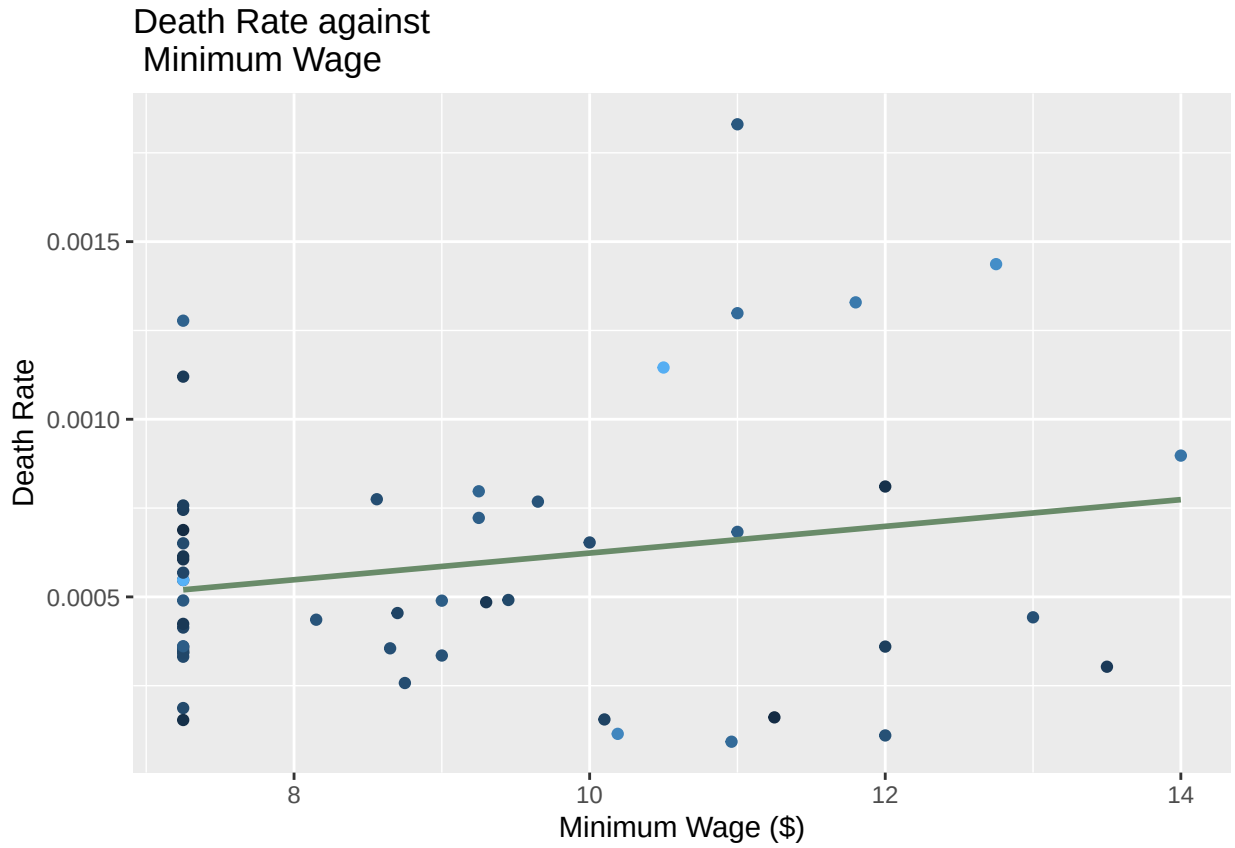
```
##   ProportionPop
## 1   0.002215394
## 2   0.014815393
## 3   0.009172426
## 4   0.022270061
## 5   0.120540899
## 6   0.017643196
```

To summarize the numerical variables, I used the summarize function to find the mean, median, IQR, standard deviation, and the max to which I found a lot of variables that don't really mean anything to be because there is no threshold of significance or context. To make the data for applicable to my project, I mutated the data to have a variable that calculates the death rate of each state. By finding the median of the state's population and using 50% as a testing rate threshold, I divided the data into two new columns that divided the numerical data into categorical variables. I found the proportion of the state's population against the total population of the country by using a user generated function into the summarize function. By using the six care dpylr functions, I was able to investigate the data set.

I found several interesting pieces of information when grouping the data and summarizing it by group. Firstly, there seems to be only 17 states with a minimum wage of over $10 (which came out to be only 1/3 of states). Secondly, states with a lower population seem to have a slightly higher testing rate than states with higher population rates. Thirdly, the states with a higher minimum wage seem to have higher testing levels than those with lower minimum wage. Finally, there are only 9 states with more than a 50% testing rate and a higher population level. All these findings contradicted what I previously have thought.

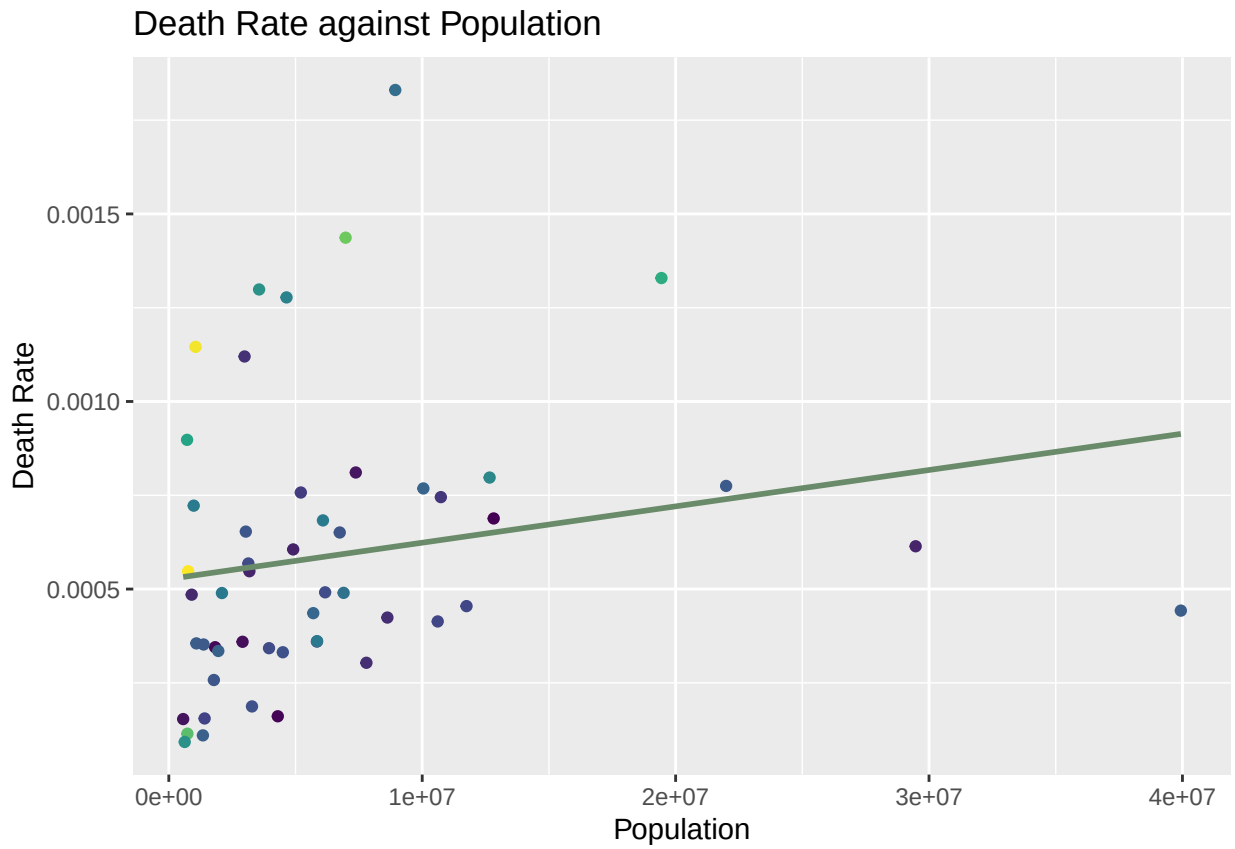
```r
combineddataset2 %>%
    ggplot(aes(Minimum_Wage, Death_Rate, col = Testing_Rate)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE, col = "darkseagreen4") +
    xlab("Minimum Wage ($)") + ylab("Death Rate") + theme(legend.position = "none") +
    ggtitle("Death Rate against \n Minimum Wage")
```

Visualizing



This plot shows a positive, linear relationship between the minimum wage($) and the death rate. This means that the death rate increases with the minimum wage. This was not the relationship I was expecting as I thought higher minimum wages means more access to health care. However, this relationship may be because the higher cost of living in those areas mean that people have less disposable income. This relationship is to be expected.

```
combineddataset2 %>%
    ggplot(aes(Population, Death_Rate, col = Testing_Rate)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE, col = "darkseagreen4") +
    xlab("Population") + ylab("Death Rate") + theme(legend.position = "none") +
    ggtitle("Death Rate against Population") + scale_color_viridis_c()
```



This plot shows a positive, linear trend between the population size and the death rate of a state. This means that the death rate increases with population size. This may be because the lack of supply for health care supplies and services vs the demand. The hospitals in the more populated areas may be more overwhelmed by the demand than those in less populated areas. This may lead to people dying from there not being help available to them. The virus may be more susceptible to being spread in more densely populated areas. This relationship is to be expected.

```
combineddataset2 %>%
    ggplot(aes(Testing_Level, Death_Rate, fill = Testing_Level)) +
    geom_boxplot() + xlab("Testing Level") + ylab("Death Rate") +
    theme(legend.position = "none") + ggtitle("Death Rate against Testing Level")
```

This plot shows that the death rate increases with testing rate with states with more than half the population being testing having a higher death rate than states with less than half the population tested. This relationship is rather unexpected as you would think that people who got tested would be taking more precautions against the virus. This relationship may be because states with higher death rates are pushing harder against testing leading to testing being more available to the public.

Concluding Remarks If any!