COVID-19 and other Risk Factors on Infant Mortality in Texas

Part 1: Introduction

This project examines (1) how infant mortality rates have changed from 2016 - 2021 in Texas, and (2) risk factors for infant mortality on the county level.

There is no statistical testing involved and therefore, all findings in this report should be taken lightly. The purpose of this project is more so to practice data wrangling and visualization in R.

DATASETS:

The ‘deaths’ dataset includes the average infant mortality rate in all Texas counties by year. The ‘income’ dataset includes the average income in all Texas counties by year. The ‘smoking’ dataset includes the average smoking rate in all Texas counties by year.

All datasets are obtained from countyhealthrankings.org.

IMPORTANCE:

Infant mortality is a tragedy that requires immediate attention. While risk factors for infant mortality have been studied within human subjects, few studies look at counties as their unit of analysis. A broader perspective may offer some interesting insights, and strengthen pre-existing knowledge on risk factors for infant mortality.

PREDICTIONS:

1. Infant mortality will slightly decrease from 2016 - 2019 due to urbanization and developments in healthcare, then increase starting in 2020 due to the COVID-19 pandemic.

2. Counties that have a higher average income should have lower rates of infant mortality than counties with lower average income. Additionally, counties that have higher rates of smoking should have higher rates of infant mortality than counties with lower rates of smoking.

First, we’ll install tidyverse. I’ve also installed 24 datasets that will be informative to answering our questions.

```
# Install necessary packages:
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.2     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.2     ✓ forcats 0.5.1
```

```
## ── Conflicts ──────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(stringr)

# Install data. We will add the year to each dataset as well to stay organized:

# INFANT DEATHS:
deaths2016 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2016_deaths.csv') %
>%
  mutate(year = 2016)
deaths2017 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2017_deaths.csv') %
>%
  mutate(year = 2017)
deaths2018 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2018_deaths.csv') %
>%
  mutate(year = 2018)
deaths2019 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2019_deaths.csv') %
>%
  mutate(year = 2019)
deaths2020 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2020_deaths.csv') %
>%
  mutate(year = 2020)
deaths2021 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2021_deaths.csv') %
>%
  mutate(year = 2021)

# AVERAGE INCOME:
income2016 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2016income.csv') %>%
  mutate(year = 2016)
income2017 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2017income.csv') %>%
  mutate(year = 2017)
income2018 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2018income.csv') %>%
  mutate(year = 2018)
income2019 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2019income.csv') %>%
  mutate(year = 2019)
income2020 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2020income.csv') %>%
  mutate(year = 2020)
income2021 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2021income.csv') %>%
  mutate(year = 2021)

# SMOKING RATES:
smoking2016 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2016_smoking.csv')
 %>%
  mutate(year = 2016)
smoking2017 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2017_smoking.csv')
 %>%
  mutate(year = 2017)
smoking2018 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2018_smoking.csv')
 %>%
  mutate(year = 2018)
smoking2019 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2019_smoking.csv')
 %>%
  mutate(year = 2019)
smoking2020 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2020_smoking.csv')
```

```
 %>%
  mutate(year = 2020)
smoking2021 <- read.csv(file = '/Users/derianlee/Desktop/project_csv/2021_smoking.csv')
 %>%
  mutate(year = 2021)
```

Part 2: Tidying

Luckily, the raw data from countyhealthrankings.org is already tidy so no tidying will be done on the datasets themselves. However, pivot_longer will be used to build summary statistics (see "Part 4: Wrangling").

Part 3: Joining / Merging

DOCUMENTATION:

First we will use bind_rows() to join the datasets for our explanatory variables vertically.

The 'deaths' dataset is created by stacking the deaths2016 - deaths2021 vertically. This dataset now has information about deaths from 2016 - 2021.

The same is done for the 'income' and 'smoking' datasets.

After that, we will use left_join() to join the deaths, income, and smoking datasets horizontally by the "County" variable. This will create our mega dataset: health_data

Let's do it!

```
# Average infant deaths by county from 2016 - 2021
deaths <- bind_rows(deaths2016, deaths2017, deaths2018, deaths2019, deaths2020, deaths20
21)

# Average income by county from 2016 - 2021
income <- bind_rows(income2016, income2017, income2018, income2019, income2020, income20
21)

# Average smoking rates by county from 2016 - 2021
smoking <- bind_rows(smoking2016, smoking2017,smoking2018,smoking2019, smoking2020, smok
ing2021)

# Join combined datasets by the County variable. Make sure years are matching so that th
e order remains intact
health_data <- deaths %>%
  left_join(income, by = "County") %>%
  filter(year.x == year.y) %>%
  left_join(smoking, by = "County")  %>%
  filter(year.y == year)
```

TOTAL OBSERVATIONS:

The 'deaths' dataset has 1524 observations, formed by the sum of the deaths2016 - deaths2021 datasets (254 observations in each).

The ‘income’ dataset has 1524 observations, formed by the sum of the income2016 - income2021 datasets (254 observations in each).

The ‘smoking’ dataset has 1524 observations, formed by the sum of the smoking2016 - smoking2021 (254 observations in each).

UNIQUE IDs:

There are no unique IDs in the three datasets mentioned above.

IDs that appear in some datasets, not others:

There are no dropped IDs in all datasets, since all of the datasets have a complete list of counties in Texas. No observations were dropped or added and therefore, the process of joining datasets was stress-free (no issues encountered).

Part 4: Wrangling

Wrangling in this project includes: (1) Data prep (2) Making summary statistics

1. Data prep: Step by step

First, let’s only keep only the data we need:

```
cleaned_health_data <- health_data %>%
  # Select only the relevant columns
  select(County, County.Value.x, County.Value.y, County.Value, year) %>%
  # Rename columns for interpretability
  rename(
    "Deaths" = County.Value.x,
    "Income" = County.Value.y,
    "Smokers"= County.Value,
    "Year" = year
  ) %>%
  # Remove rows that have missing observations for death count
  filter(Deaths != "")

head(cleaned_health_data)
```

```
##     County Deaths   Income Smokers Year
## 1 Anderson      6 $42,500      18% 2016
## 2 Angelina      5 $42,100      19% 2016
## 3 Atascosa      5 $52,800      15% 2016
## 4  Bastrop      6 $52,900      16% 2016
## 5     Bell      8 $51,000      18% 2016
## 6    Bexar      6 $50,700      13% 2016
```

Next, let’s remove or swap unwanted symbols (, $ % :):

```
cleaned_health_data <- cleaned_health_data %>%
  # Remove commas in the death count (ex: 1,000 -> 1000)
  mutate(Deaths= gsub(",", "",Deaths)) %>%
  # Remove dollar symbol for income
  mutate(Income = gsub("\\$", "", Income)) %>%
  # Remove commas for income
  mutate(Income = gsub("\\,", "", Income)) %>%
  # Remove percent symbol for smoking rate
  mutate(Smokers = gsub("\\%", "", Smokers))

head(cleaned_health_data)
```

```
##     County Deaths Income Smokers Year
## 1 Anderson      6 42500       18 2016
## 2 Angelina      5 42100       19 2016
## 3 Atascosa      5 52800       15 2016
## 4  Bastrop      6 52900       16 2016
## 5     Bell      8 51000       18 2016
## 6    Bexar      6 50700       13 2016
```

Next, let’s convert numeric variables from character to double data-type.

```
cleaned_health_data <- cleaned_health_data %>%
  # Turn the following columns into numeric variables
  mutate_at("Deaths", as.numeric) %>%
  mutate_at("Income", as.numeric) %>%
  mutate_at("Smokers", as.numeric)

head(cleaned_health_data)
```

```
##     County Deaths Income Smokers Year
## 1 Anderson      6  42500      18 2016
## 2 Angelina      5  42100      19 2016
## 3 Atascosa      5  52800      15 2016
## 4  Bastrop      6  52900      16 2016
## 5     Bell      8  51000      18 2016
## 6    Bexar      6  50700      13 2016
```

Next, let’s make our variables more interpretable.

```
cleaned_health_data <- cleaned_health_data %>%
  # Smoking rate is a percentage out of 100
  mutate(`SmokeRate` = Smokers / 100) %>%
  # Delete the original column
  select(-Smokers) %>%
  # Model income in thousands of USD
  mutate(Income = Income / 1000)

head(cleaned_health_data)
```

```
##     County Deaths Income Year SmokeRate
## 1 Anderson      6   42.5 2016      0.18
## 2 Angelina      5   42.1 2016      0.19
## 3 Atascosa      5   52.8 2016      0.15
## 4  Bastrop      6   52.9 2016      0.16
## 5     Bell      8   51.0 2016      0.18
## 6    Bexar      6   50.7 2016      0.13
```

Lastly, we'll remove outliers from our response variable (infant mortality):

```
# REMOVE OUTLIERS

# Store outliers
outliers <- boxplot(cleaned_health_data$Deaths, plot = FALSE)$out

# Remove outliers
cleaned_health_data <- subset(cleaned_health_data, !Deaths %in% outliers)
```

2. Making summary statistics

```
# Table 1: Average income and smoking rate from 2016 - 2021
table1 <- cleaned_health_data %>%
  # Group data by year
  group_by(Year) %>%
  # Return average income and smoking rate by year
  summarize(`Income (thousands, USD)` = mean(`Income`),
            `Smoking Rate` = mean(SmokeRate))

knitr::kable(table1, caption = "<center><strong>Average Income and Smoking Rate in Texa
s: 2016 - 2021</strong></center>")
```

Table:

Average Income and Smoking Rate in Texas: 2016 - 2021

Year	Income (thousands, USD)	Smoking Rate
2016	50.86410	0.1583333
2017	52.79630	0.1592593
2018	53.25833	0.1575000
2019	55.17564	0.1576923
2020	56.73718	0.1506410
2021	61.47973	0.1782432

From 2016 - 2021, it appears that the average income (in thousands, $) has steadily increased in Texas, potentially due to urbanization. In that same time frame, smoking rates have been relatively constant, with a relatively large increase from 2020 to 2021. This could be a psychological impact of COVID-19 where more people are seeking tobacco as an adaptation to stress. It is still unclear how the change in income and smoking rate will affected infant mortality.

```
# Table 2: How have infant mortality rates changed over time in Harris, Dallas, and Trav
is county?
table2 <- cleaned_health_data %>%
  # Look only at Harris, Dallas and Travis
  filter(County == "Harris" | County == "Dallas" | County == "Travis") %>%
  # Select only relevant columns
  select(County, `Deaths`, Year) %>%
  # Arrange counties alphabetically
  arrange(County) %>%
  # Change format
  pivot_wider(names_from = Year,
              values_from = `Deaths`)

knitr::kable(table2, caption = "<center><strong>Infant Deaths per Thousand, by County</s
trong></center>")
```

Table:

Infant Deaths per Thousand, by County

County	2016	2017	2018	2019	2020	2021
Dallas	7	7	7	7	6	6
Harris	6	6	6	6	6	6
Travis	5	5	4	4	4	4

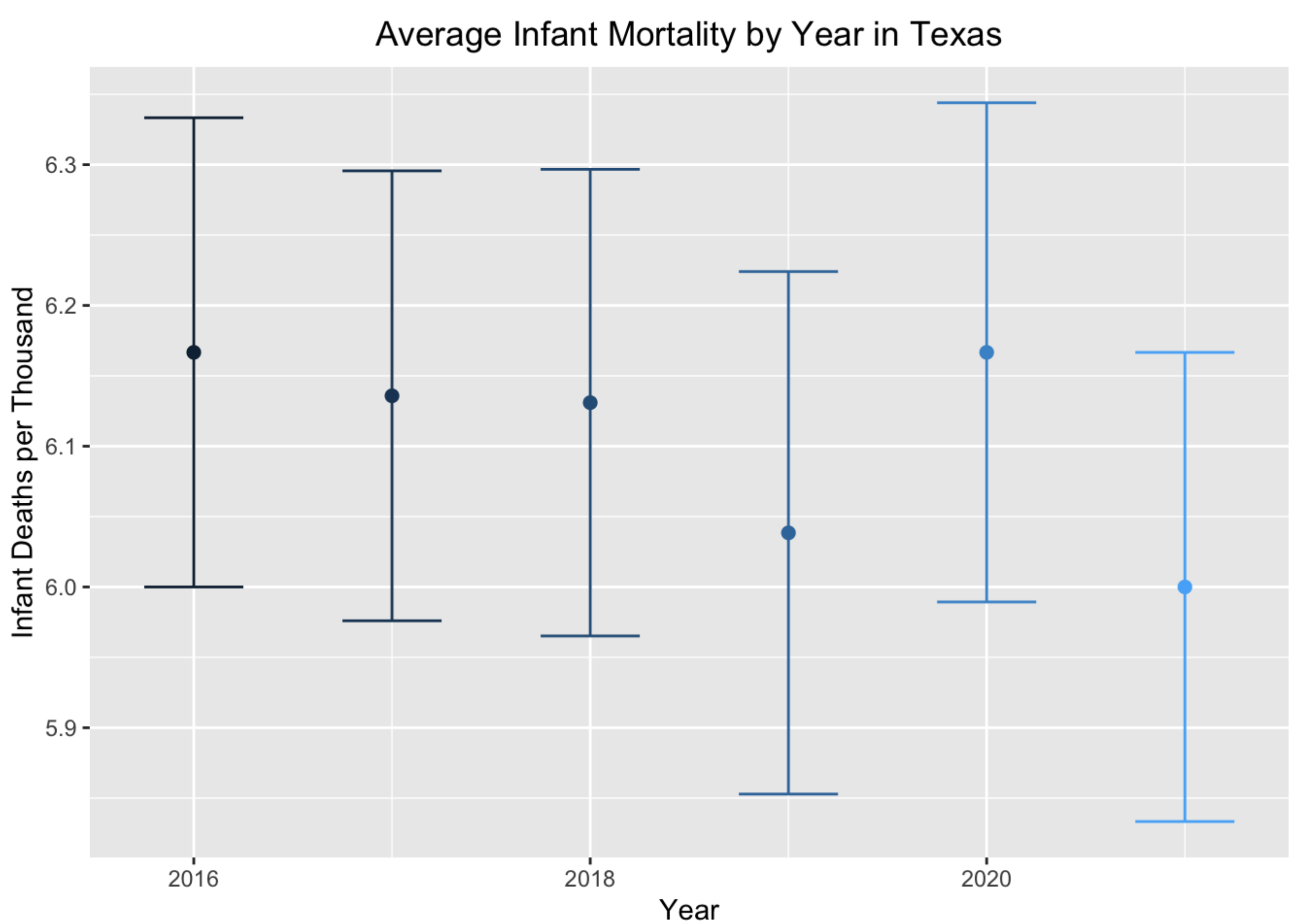
In this table, pivot_wider was used to transform the data from long to wide format. The ‘year’ variable contains numeric values from 2016 - 2021; these observations replaced the original ‘year’ variable, as seen above. New observations were then pulled from the ‘Death Rate’ variable, such that each year would now show the death rate for a given county.

In all three counties, infant mortality rates have been relatively unchanging. Dallas and Travis have a slight decrease in infant mortality but the difference is small. Harris county is stagnant. This data is surprising, given that COVID-19 started in December of 2019.

Part 5: Creating Visualizations

```
# Visual 1: How has infant mortality changed over the years?
cleaned_health_data %>%
  # Plot deaths against year
  ggplot(aes(x = Year, y = Deaths, color = Year)) +
  # Plot the average deaths
  geom_point(stat = "summary", fun = "mean", size = 2) +
  # Error bars
  geom_errorbar(stat = "summary", width = 0.5) +
  # Change theme to align title
  theme(plot.title = element_text(hjust = 0.5)) +
  # Change theme to remove legend
  theme(legend.position = "none") +
  # Rename axis
  scale_y_continuous(name = "Infant Deaths per Thousand") +
  # Title
  labs(title = "Average Infant Mortality by Year in Texas")
```

```
## No summary function supplied, defaulting to `mean_se()`
```



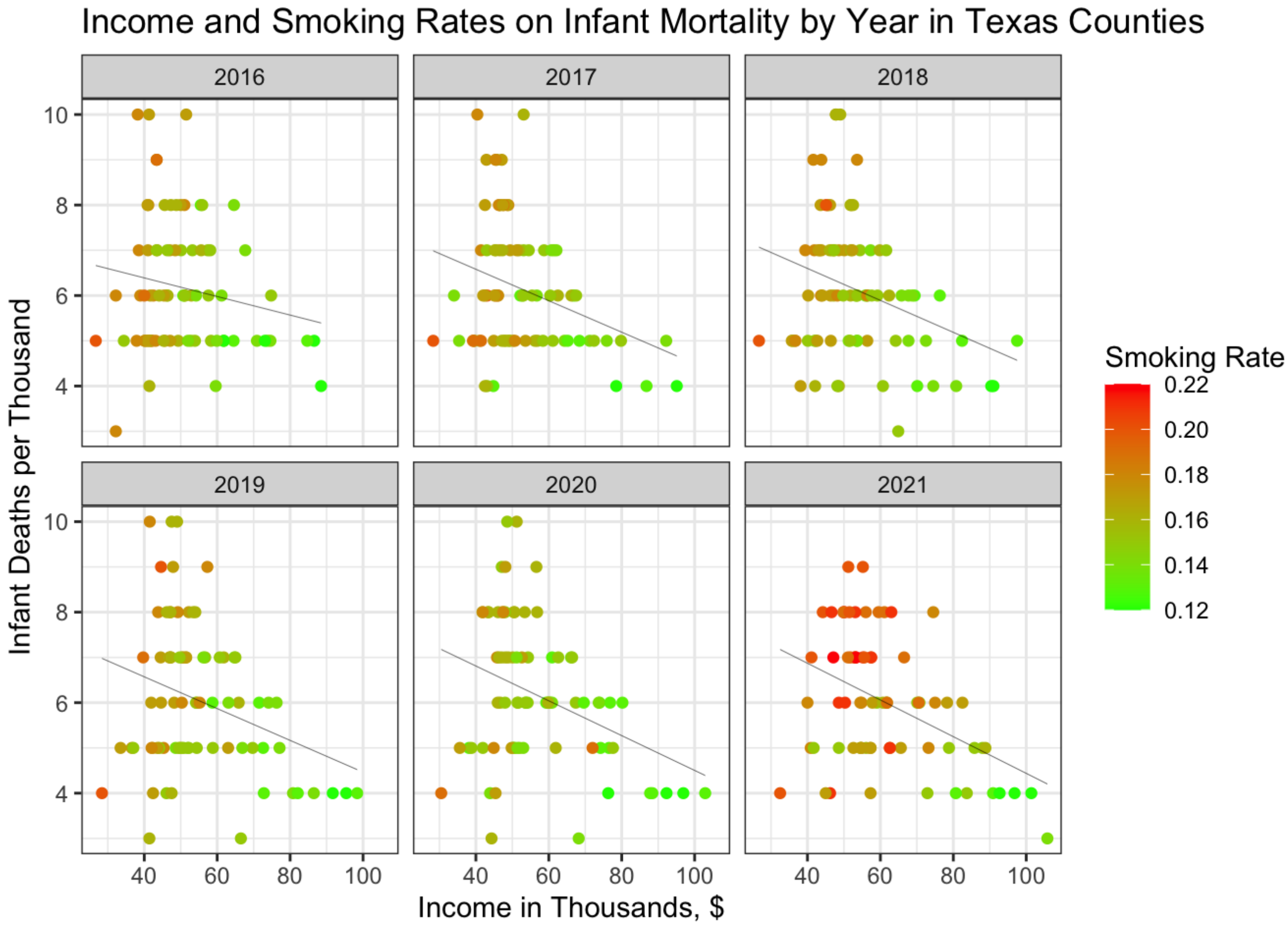
The first part of this project asked: how has infant mortality changed from 2016 - 2021 in Texas?

In terms of spread, mortality rates have been relatively constant from 2016 - 2021. In terms of mean, infant mortality was relatively constant from 2016 - 2018, but decreased in 2019. The year 2020 was COVID's most violent year, and infant mortality may have increased in 2020 for that reason. Still, the increase was shockingly less than expected. The subsequent decrease in infant mortality in the following year may represent increased vaccine efficacy and COVID protocol, but the relatively low death rate during that year still raises questions.

Overall, infant mortality has barely changed in Texas despite the pandemic, a pattern consistent with other publicly available data. Very surprising.

```
# Visual 2: How does income and smoking rates affect infant mortality?
cleaned_health_data %>%
  # Plot deaths against income, color by smoking rate
  ggplot(aes(x = Income, y = Deaths, color = SmokeRate)) +
  # Scatterplot
  geom_point() +
  # Custom gradient for visibility
  scale_color_gradient(low = "green", high = "red") +
  # Adding line of best fit
  geom_smooth(method = "lm", color = "black", se = F, size = 0.1) +
  # Facet by year
  facet_wrap(~Year) +
  # Adding labels
  labs(x = "Income in Thousands, $",
       y = "Infant Deaths per Thousand",
       title = "Income and Smoking Rates on Infant Mortality by Year in Texas Counties",
       color = "Smoking Rate") +
  # Changing theme
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The second part of the project asked: what are some risk factors for infant mortality, on the county level?

AVERAGE INCOME:

From 2016 - 2021, the distribution of average income appears to have right shifted, meaning the median income has increased over these 5 years. When looking at the effects of income on infant deaths, it appears that there is a negative correlation between the two: counties with higher median income report a lower number of infant deaths. This makes sense: households with more income can afford better access to healthcare and food. While this trend doesn't seem to apply to 20,000 - $50,000 income domain, this is likely due to lack of data at this income-level and thus conclusions shouldn't be drawn from this domain.

SMOKING RATE:

Studies on the individual reveal that women who smoke during pregnancy largely increase their chances of premature infant death. This project however, looks at counties, not individuals as the unit of analysis.

It appears that counties with lower household tend to have higher rates of smoking. Therefore, collinearity makes it difficult to draw conclusions, especially due to skewed income distributions.

Visually speaking, it does NOT appear that counties who have higher rates of smoking experience higher rates of infant mortality. Although 2021 could be a case against this, it is too difficult to tell visually. Further analysis will need to be done on the effects of smoking on infant mortality on the county level. If these county studies show a detrimental effect, this would strengthen pre-existing studies done on smoking and infant death on the individual level.

CONCLUSIONS:

VISUALLY SPEAKING:

1. COVID-19 had little to no effect on infant mortality
2. Counties with higher median income report a lower number of infant deaths than counties with a lower median income
3. The smoking rate of a given county does not affect that county's infant mortality rate