Coronavirus EDA

Data Science I (STAT 301-1)

Derex Wangmang

Contents

Introduction

My Exploratory Data Analysis focused on visualizing how the coronavirus exacerbated existing inequalities. I explored the following questions:

- What is the overall trend of coronavirus cases and deaths in the Chicagoland area?
- How does median income relate to coronavirus case and death rates?
- How do medical conditions, specifically diabetes and asthma rates, relate to the rates of coronavirus cases and deaths?

To answer these questions, I gathered a variety of data:

- COVID-19 Cases, Tests, and Deaths by ZIP Code: This dataset includes provides information on COVID-19 statistics separated by ZIP code in the Chicagoland area (City of Chicago, 2020). These statistics include zip code, week number, weekly cases, weekly tests, percent tested positive weekly, and weekly deaths.
- Public Health Statistics - Diabetes: This dataset provides information about the number of hospital discharges for those with diabetes between 2000 and 2011 (Illinois Department of Public Health, 2012). Variables are separated by ZIP code, the number of diabetes-related hospitalizations in each year, the estimated crude rate in each year, and the confidence intervals for the estimates.
- Public Health Statistics - Asthma: This dataset provides information about the number of hospital discharges for those with asthma between 2000 and 2011 (Illinois Department of Public Health, 2012). Variables are separated by ZIPs code, the number of asthma-related hospitalizations in each year, the estimated crude rate in each year, and the confidence intervals for the estimates.
- Median Income Data for Zip Codes: This is not a dataset, but rather a useful API (*Basic usage of tidycensus*). Through `tidycensus`, I retrieved the estimate of median income and the margin of error for that estimate per ZIP code.

Loading Packages

```
library(tidyverse)
```

I use `tidyverse`.

Reading in Processed Data

```
cases <- read_csv("data/processed/COVID-19_ZIP_Before_Metric_Change.csv")
diabetes <- read_csv("data/processed/Public_Health_Statistics-_Diabetes_hospitalizations_in_Chicago__20
asthma <- read_csv("data/processed/Public_Health_Statistics_-_Asthma_hospitalizations_in_Chicago__by_yea
med_income <- read_csv("data/processed/median_income_ZIP.csv")

med_income <- med_income %>%
  arrange(estimate)
```

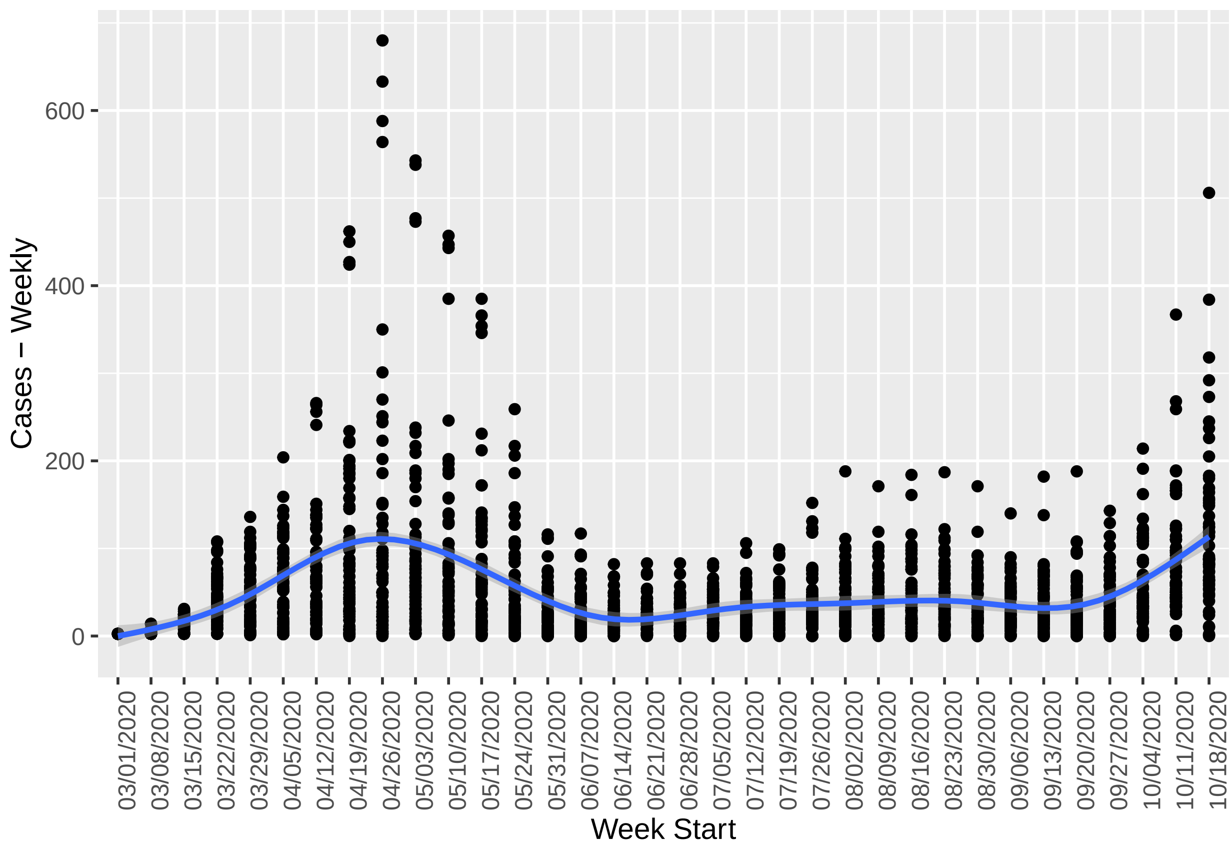
While I read and processed multiple datasets, the ones I use throughout this report include: * COVID-19 Cases: Includes coronavirus case data, including infection rate and death rate, separated by ZIP code and week. * Median Income: Includes the median income from different ZIP codes. * Diabetes Rates: Includes diabetes rate from different years, separated by ZIP code. * Asthma Rates: Includes asthma rate from different years, separated by ZIP code.

Visualizing Cases over Time

I wanted to identify general coronavirus trends in the Chicagoland area. The case and death rates both peaked around 04/26. Later in the pandemic, the case rate increased while the death rate remained stable, suggesting that with increased experience, health professionals decreased the death rate.

Cases over Weeks

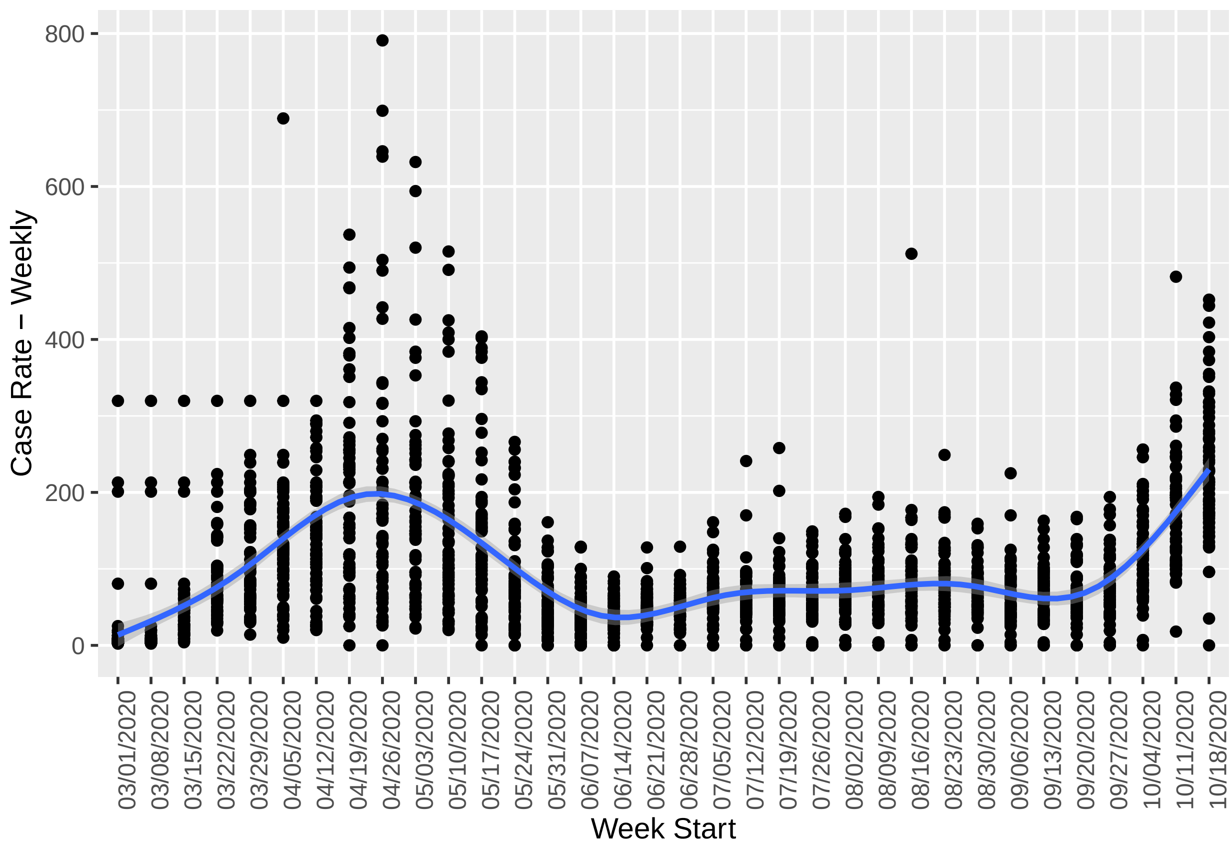
```
ggplot(cases) +
  geom_point(aes(`Week Start`, `Cases - Weekly`)) +
  # Difficult to fit a line with date, so developed a fitted line according to the week number
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Cases - Weekly`)) +
  theme(axis.text.x = element_text(angle = 90))
```



This is a visualization of the number of cases per week over time. These case numbers are absolute, rather than relative, not accounting for the population size within their respective ZIP codes. The fitted line takes into account cases from all ZIP codes.

Case Rate over Weeks

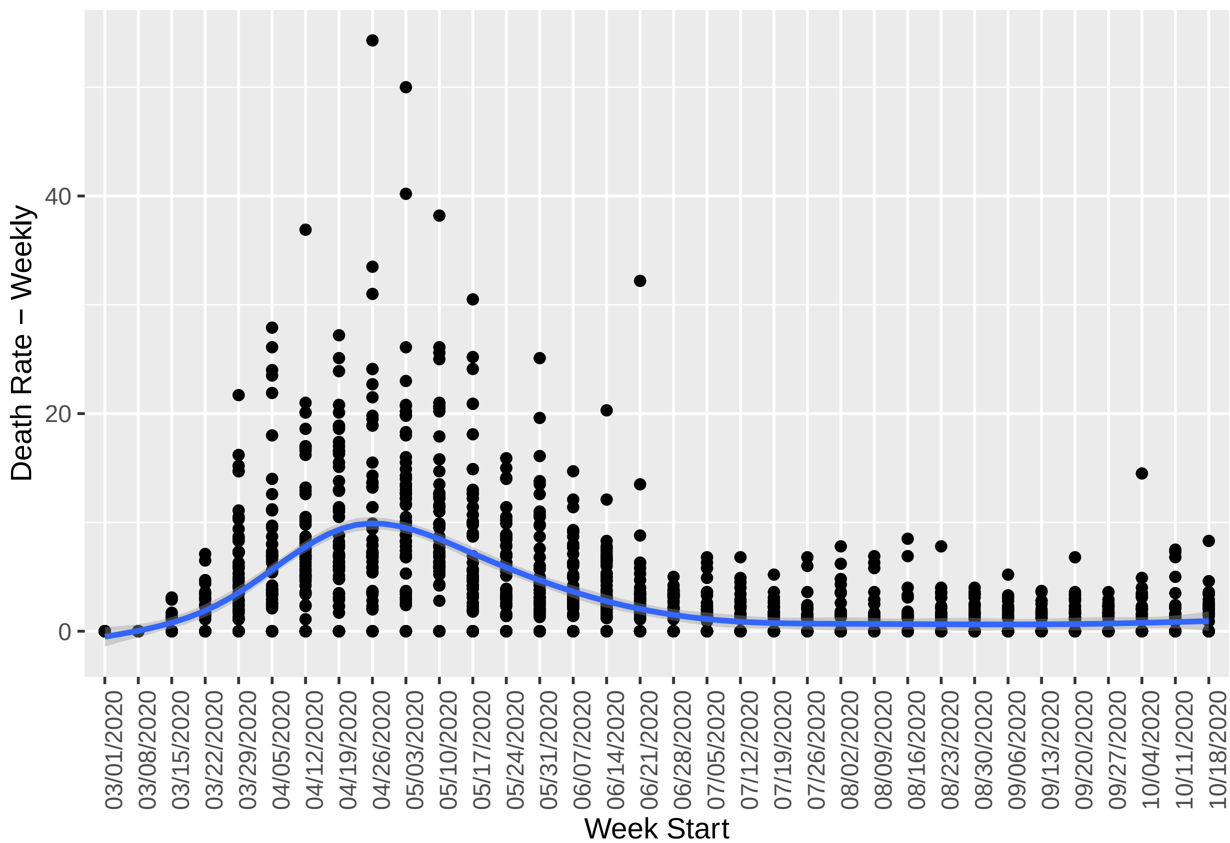
```r
ggplot(cases) +
  geom_point(aes(`Week Start`, `Case Rate - Weekly`)) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Case Rate - Weekly`)) +
  theme(axis.text.x = element_text(angle = 90))
```



This is a visualization of the case rate per week over time. To calculate the case rate, the population within the ZIP code is divided by the number of cases that week, multiplied by 100,000. This case rate ensures the number of cases is relative to its population size, making it easier to compare the proportion of cases for ZIP codes of different sizes. While the case rate range is different, the overall trend looks fairly similar to the one from the prior plot. The downward trend following the peak in the week of 04/26 may be a result of Mayor Lightfoot's Stay-at-Home Executive Order issued 05/01. Further information may be found here) (COVID-19 Orders, 2020).

Death Rate over Weeks

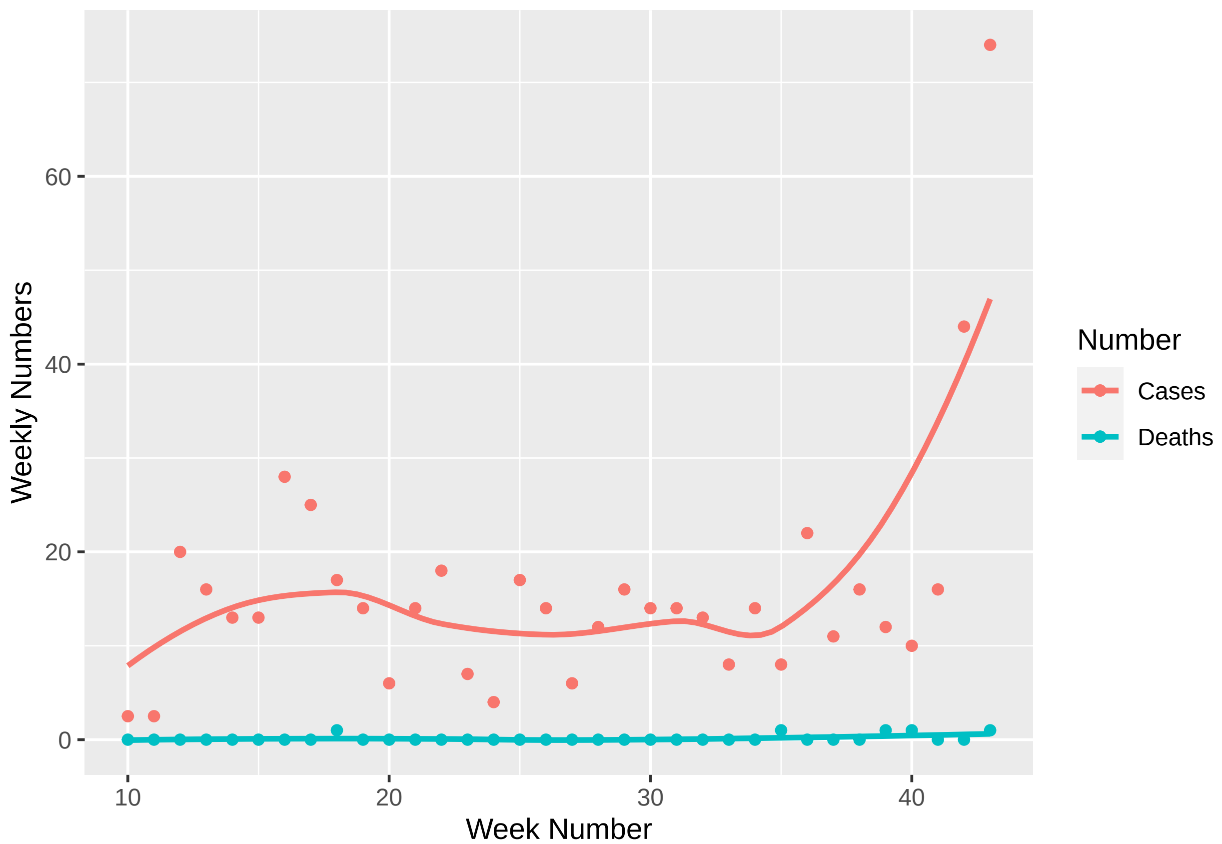
```
ggplot(cases) +
  geom_point(aes(`Week Start`, `Death Rate - Weekly`)) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Death Rate - Weekly`)) +
  theme(axis.text.x = element_text(angle = 90))
```



This is a visualization of the death rate over time. I expected the death rate to peak within two weeks after the peak of the case rate. However, the death rate peaked 04/26, the same week during which the case rate spiked. Additionally, while the case rate trends upwards after the week of 09/20, the death rate seemed to remain consistent, suggesting that with more experience, the healthcare system was better equipped to treat coronavirus patients.

Rates over Weeks in 60605

```
zip_60605 <- filter(cases, `ZIP Code` == 60605)
ggplot(zip_60605) +
  geom_point(aes(`Week Number`, `Cases - Weekly`, color = "Cases")) +
  geom_smooth(aes(`Week Number`, `Cases - Weekly`, color = "Cases"), se = FALSE) +
  geom_point(aes(`Week Number`, `Deaths - Weekly`, color = "Deaths")) +
  geom_smooth(aes(`Week Number`, `Deaths - Weekly`, color = "Deaths"), se = FALSE) +
  ylab("Weekly Numbers") +
  scale_color_discrete(name = "Number", labels = c("Cases", "Deaths"))
```



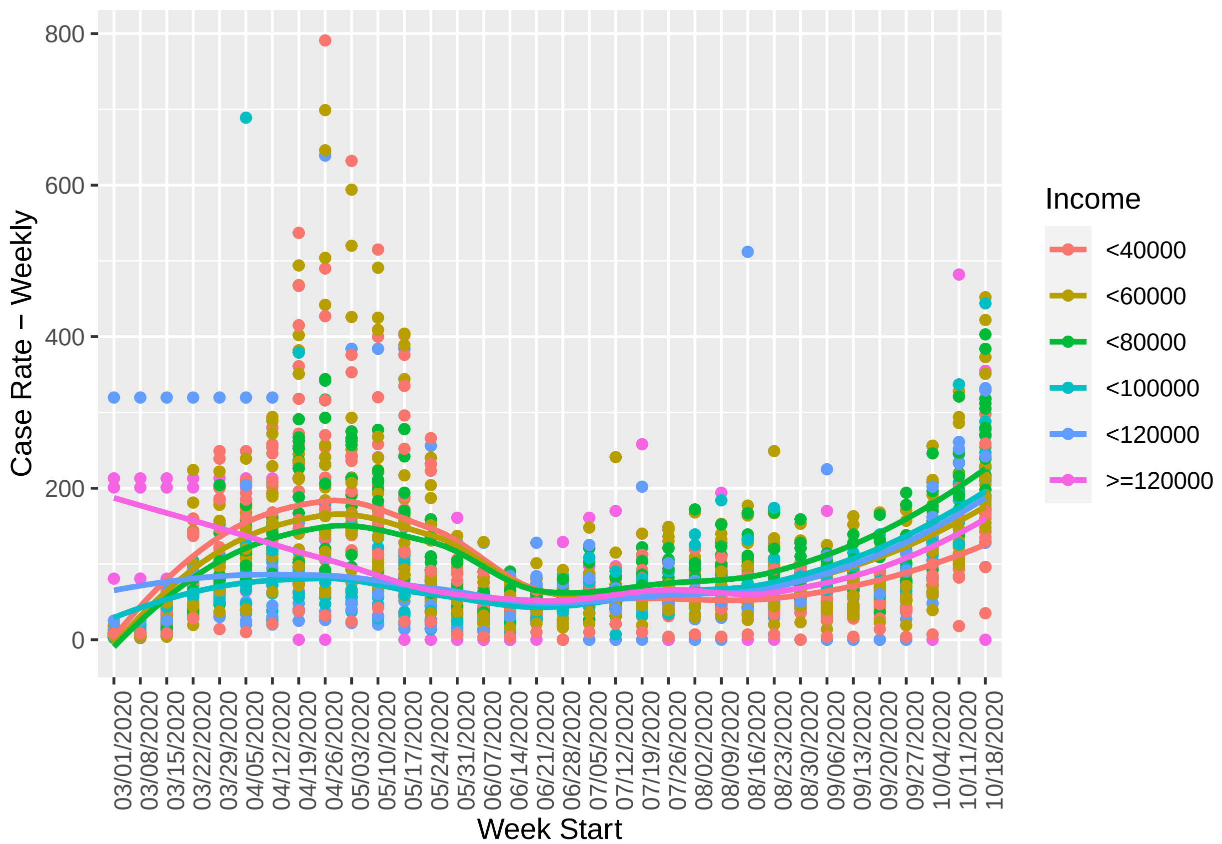
This is a visualization of the death rate over time in the ZIP code 60605. Since these values were within the same ZIP code and population is constant, I decided to plot the absolute value, rather than the rate. While the weekly cases were around 5 - 15 cases, the weekly deaths remained a near constant value of 0.

Visualizing Cases over Time with Income

Continuing my exploration, I sought to review correlations between income and coronavirus rates. I expected that the case rate between populations may be the same, but the death rate would not be. This would arise as a result of more access to resources, including insurance and healthcare.

Rates over Weeks Based on Income Levels

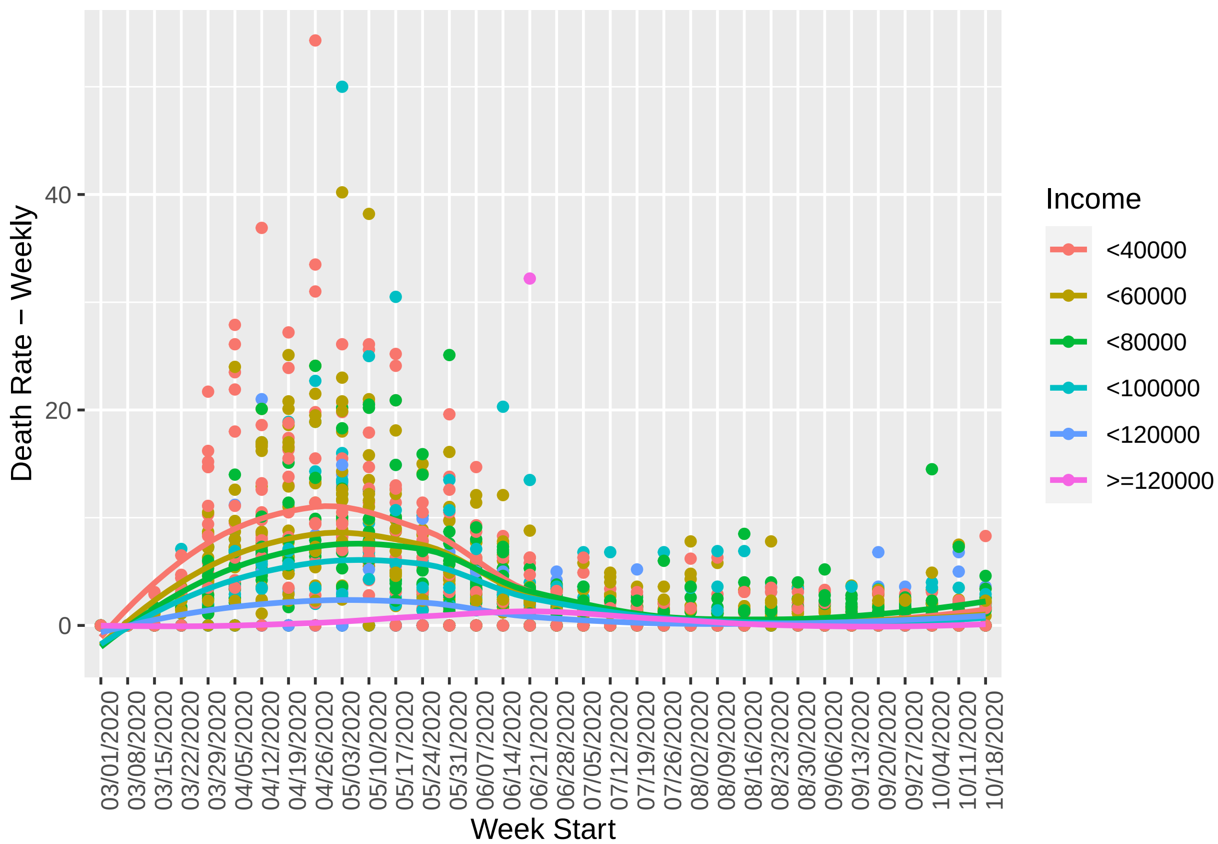
```
cases_with_income <- cases %>%
  left_join(med_income, by = "ZIP Code")

cases_with_income <- cases_with_income %>%
  mutate(`income_bracket` = cut(estimate, breaks = c(-Inf, 40000, 60000, 80000, 100000, 120000, Inf),
                              labels = c("<40000", "<60000","<80000","<100000","<120000", ">=120000"))

cases_with_income %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Case Rate - Weekly`, color = factor(income_bracket))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Case Rate - Weekly`, color = factor(in
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Income")
```



This is a visualization of the number of cases per week over time, with the legend displaying the grouped median income brackets. It appears that that the median incomes from below \$40000 until \$80000 have the same trend with a peak around 04/26. Populations within ZIP codes with higher median incomes have a smaller maximum, or no maximum there at all.

Rates over Weeks Based on Aggregate Income Levels

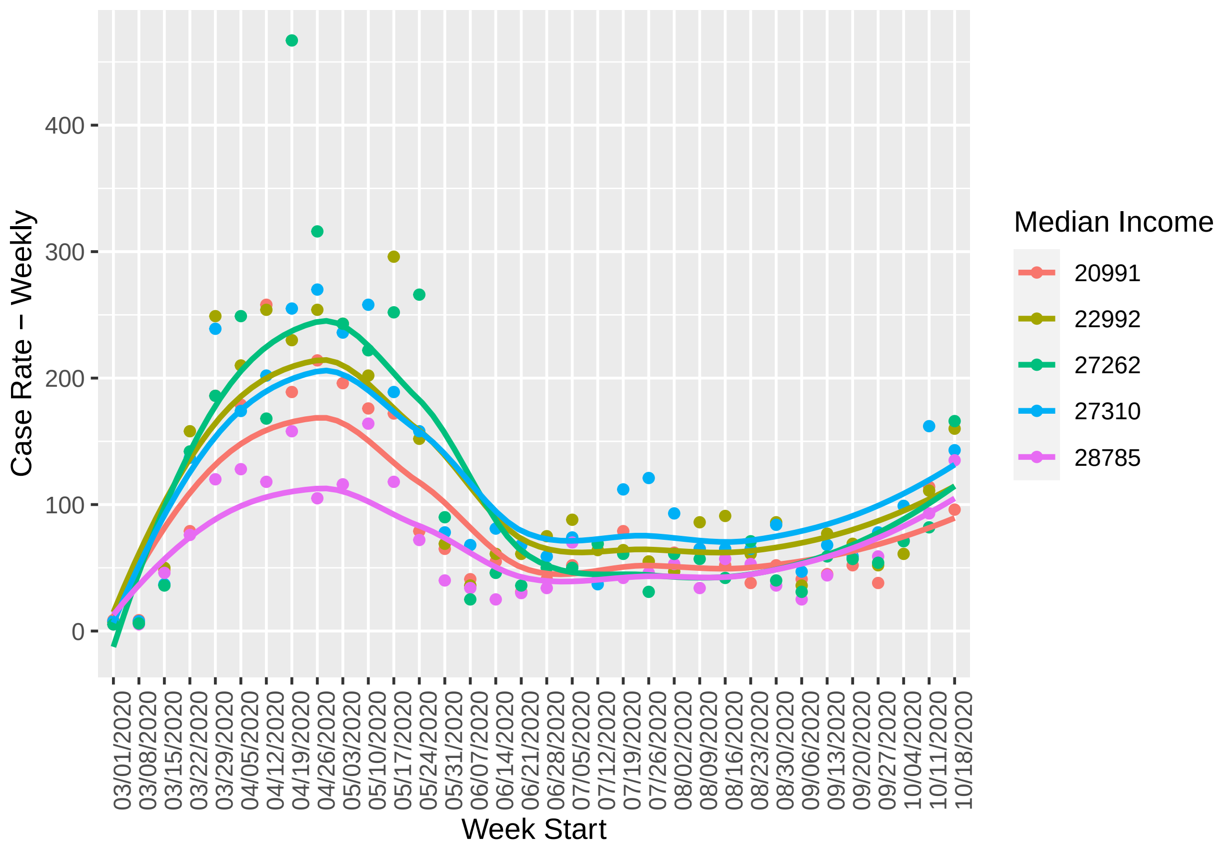
```
cases_with_income %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Death Rate - Weekly`, color = factor(income_bracket))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Death Rate - Weekly`, color = factor(i
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Income")
```



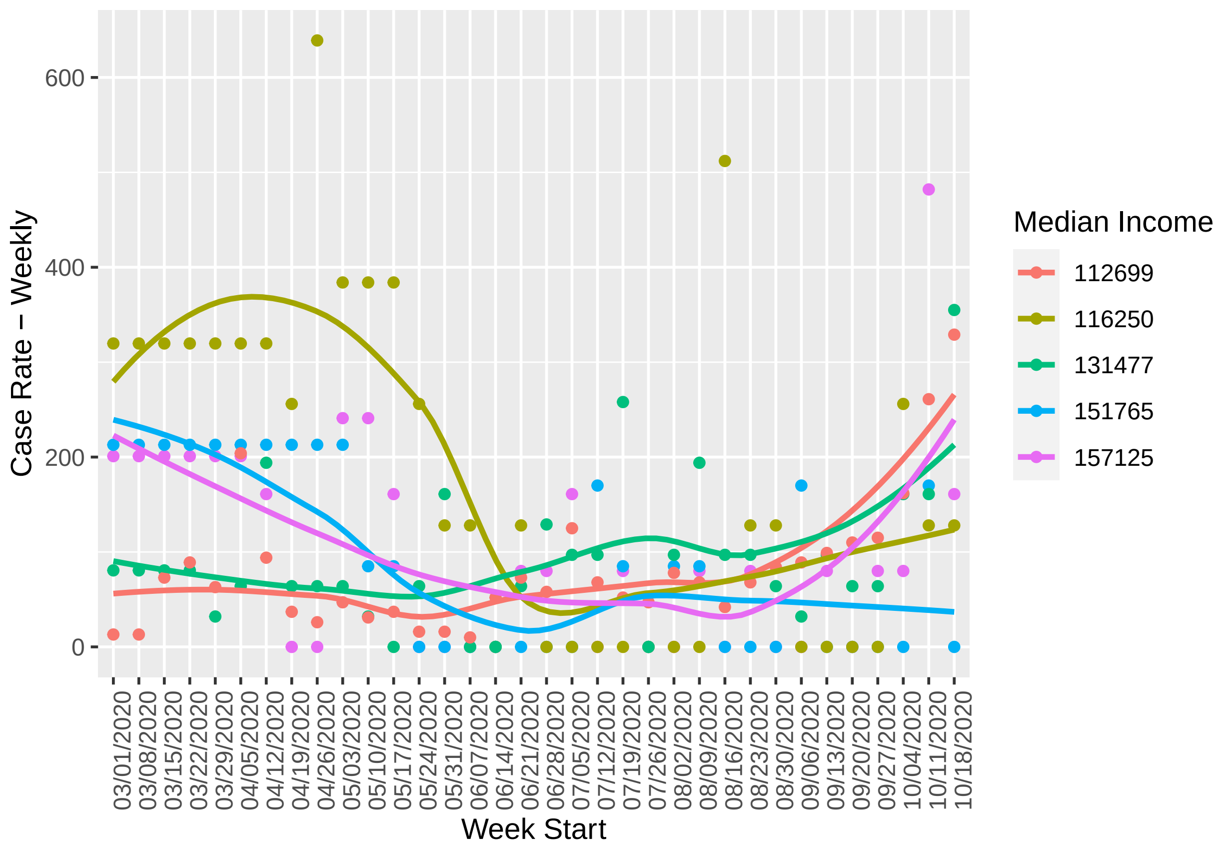
This is a visualization of the weekly death rate over time, with the legend displaying the aggregate median income. It appears that the weekly death rate had the same overall trends as the case rates. However, the higher the income level, the lower the death rate. This is especially clear when the death rate peaked around 04/26. At that point, the lowest income had the highest peak. For the same week, the rate decreases as the income increases. Those with the highest incomes of greater than or equal to 120000 have near 0 weekly deaths all throughout the pandemic.

Case Rates Split by Income Levels

```
cases_with_income %>%
  filter(`ZIP Code` %in% head(med_income, 5)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Case Rate - Weekly`, color = factor(`estimate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Case Rate - Weekly`, color = factor(`e
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Median Income")
```



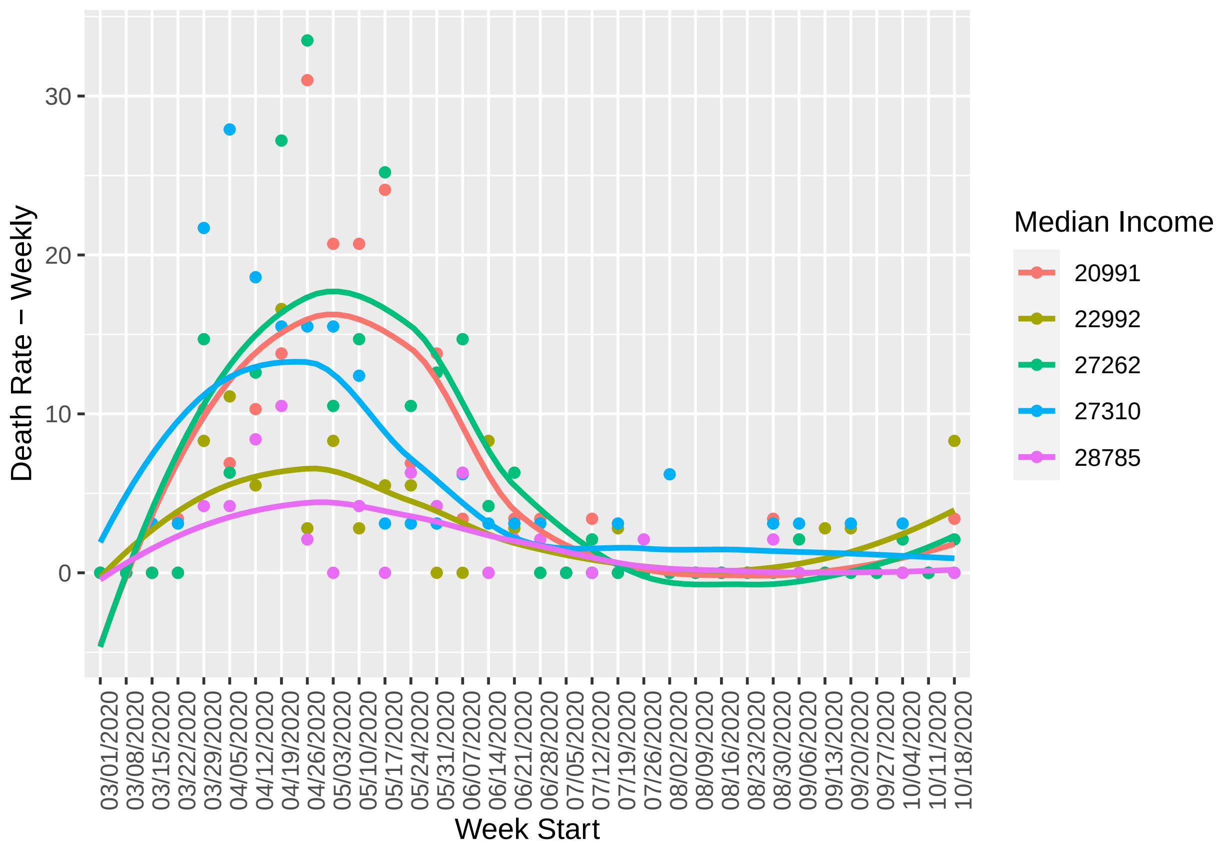

```
cases_with_income %>%
  filter(`ZIP Code` %in% tail(med_income, 5)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Case Rate - Weekly`, color = factor(`estimate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Case Rate - Weekly`, color = factor(`es
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Median Income")
```



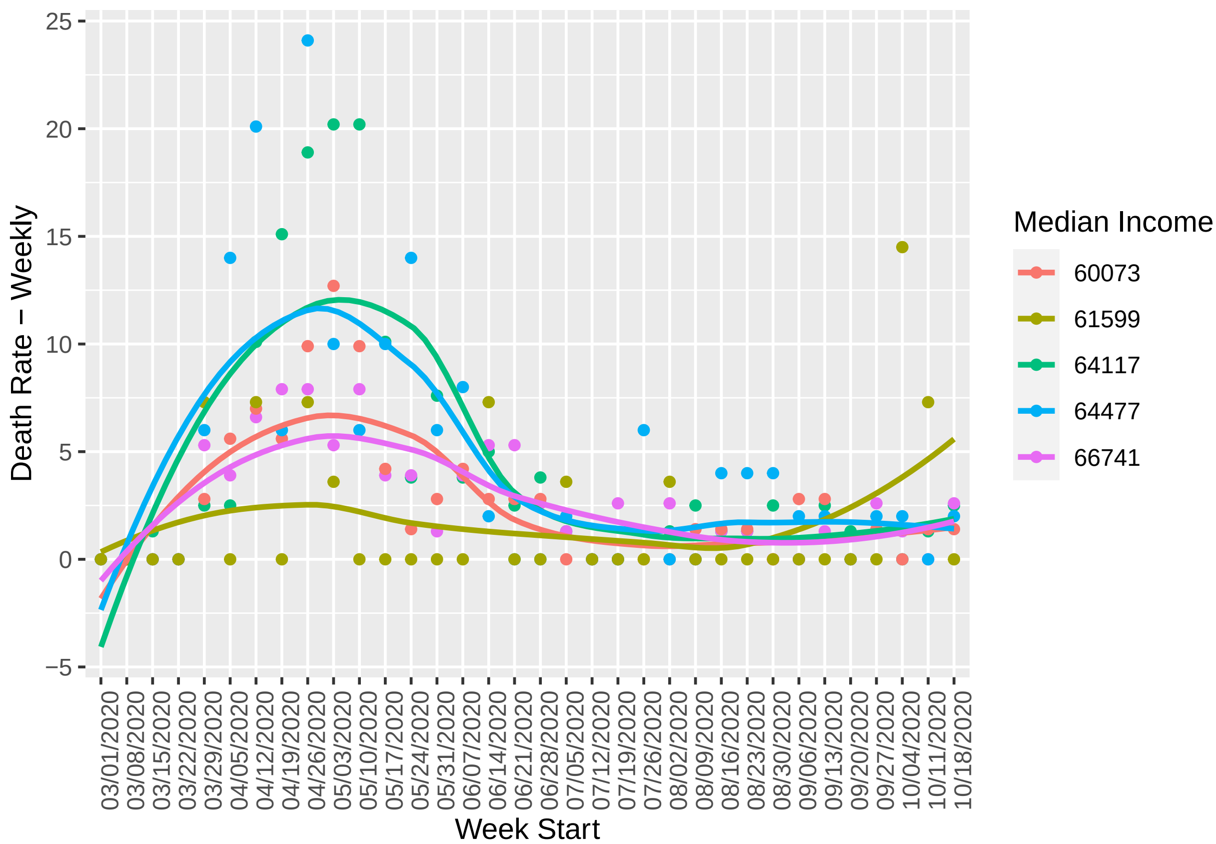
These are visualizations of the case rates among the 5 ZIP codes with the lowest median income levels and the highest median income levels, respectively. The case rates among the lowest income ZIP codes seem to follow the same trend, with a peak in the week of 04/26 and approximately stable levels until 09/02. However, the case rate for those with the highest income levels is much more different. There are many data that remain constant throughout the weeks, which is unrealistic and most likely a result of inaccurate or incomplete data collection.

Death Rates Split by Income Levels

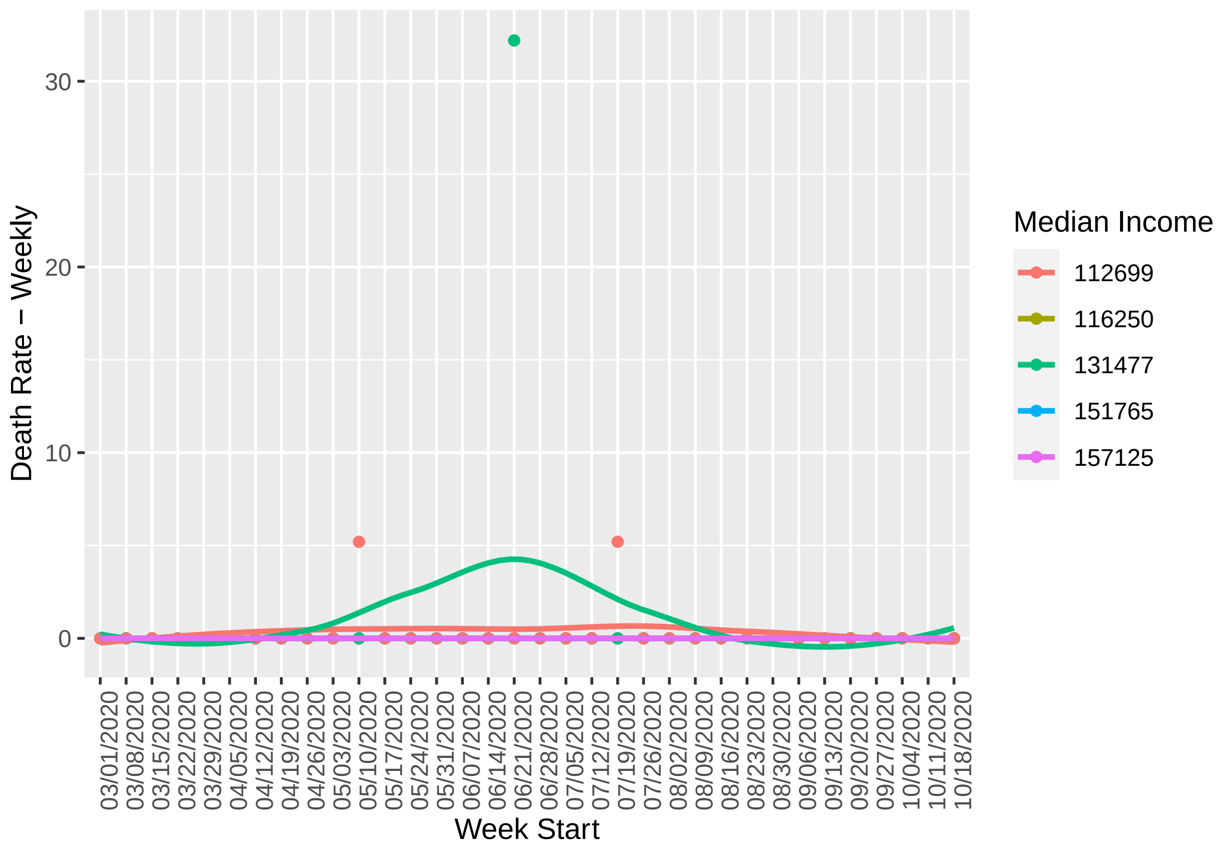
```
cases_with_income %>%
  filter(`ZIP Code` %in% head(med_income, 5)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Death Rate - Weekly`, color = factor(`estimate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Death Rate - Weekly`, color = factor(`e
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Median Income")
```




```
cases_with_income %>%
  filter(`ZIP Code` %in% med_income[30:34,]$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Death Rate - Weekly`, color = factor(`estimate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Death Rate - Weekly`, color = factor(`e
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Median Income")
```




```
cases_with_income %>%
  filter(`ZIP Code` %in% tail(med_income, 5)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Death Rate - Weekly`, color = factor(`estimate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Death Rate - Weekly`, color = factor(`e
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Median Income")
```



For the ZIP codes with the lowest income levels, the death rate trend follows the previous case rate trends. The ZIP codes with the incomes of 60 - 67k, have a similar trend. However, for the ZIP codes within the highest income levels, the death rates remained at a near constant 0 throughout the pandemic. This is most likely a result of a compound of different factors:

- Biased data: It's possible that the original data may not be correct. There are questionable trends among the case rates, which affect the death rate as well.
- White collar work: People within these high income ZIP codes are most likely performing the majority of their work from a remote setting, decreasing their contact with others. That led to a decreased case rate and subsequently death rate.
- Increased access to resources: The few who became sick most likely had insurance and access to numerous resources, including hospitals and healthcare.

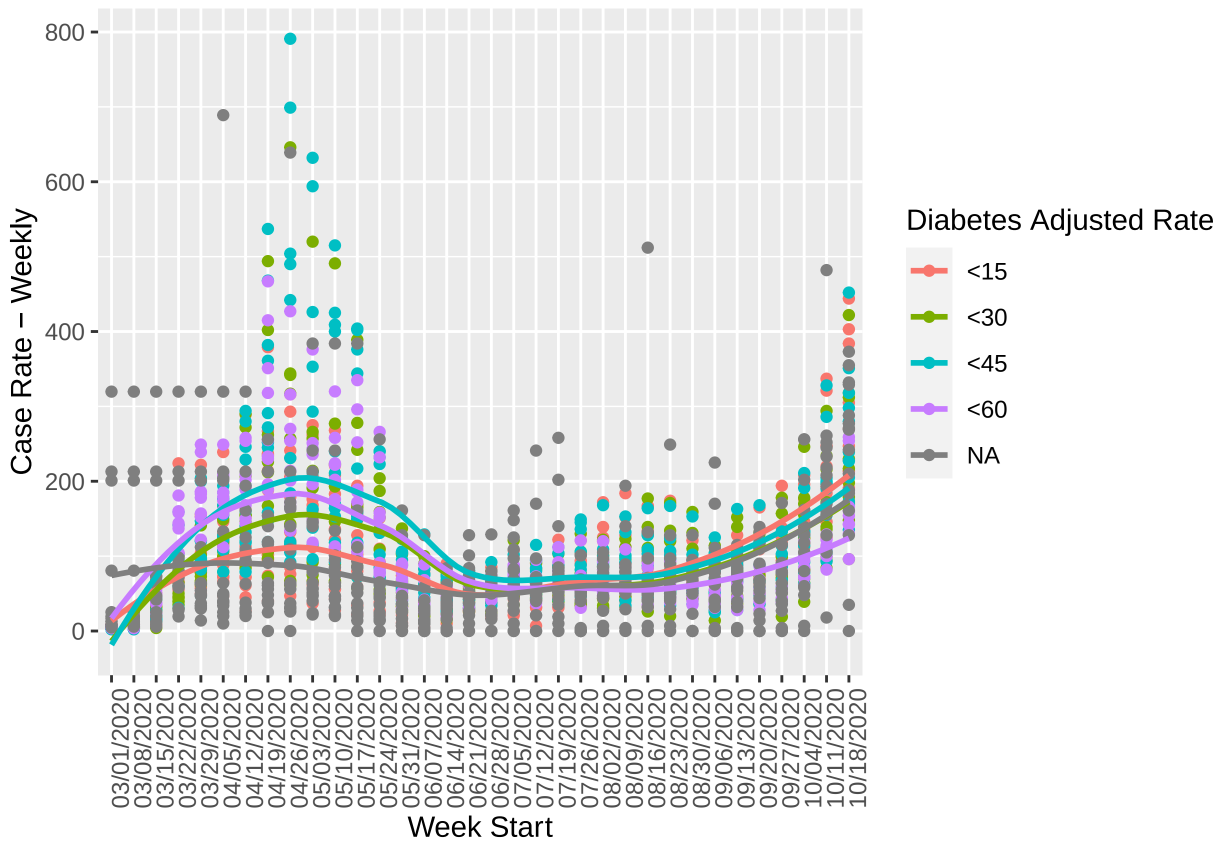
Visualizing Cases over Time with Diabetes

I wanted to identify how the coronavirus affected those with diabetes. I expected the case rate to be the same for both, but the death rate to be higher for those with higher rates of diabetes compared to the those with lower rates of diabetes. This is due to the fact that medical conditions such as diabetes may lead to complications in treatment. However, since diabetes is not a respiratory disease, that may explain why the death rates seemed similar across both populations of different diabetes rates.

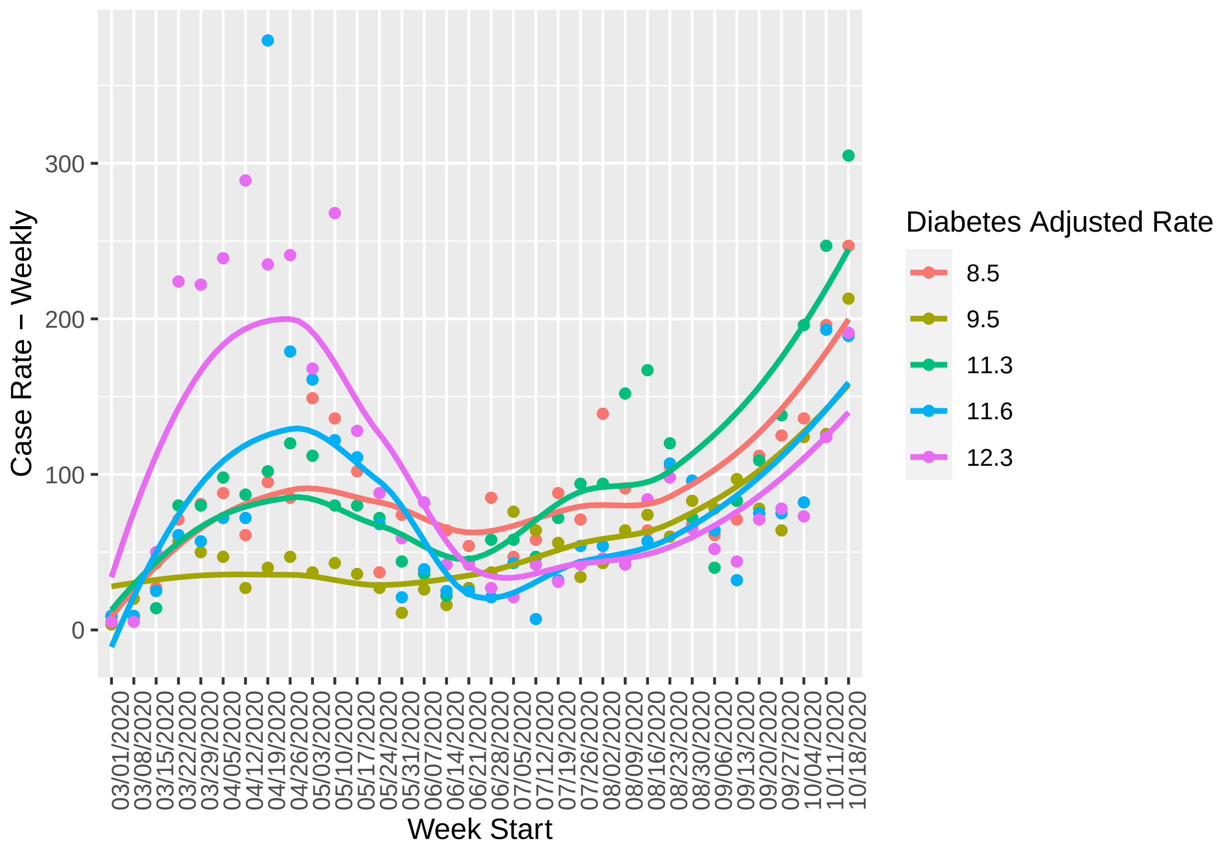
Case Rate over Weeks Based on Diabetes Rate

```
compressed_diabetes <- diabetes %>%
  select(`ZIP code (or aggregate)`, `Adjusted Rate 2011`) %>%
  rename(`ZIP Code` = `ZIP code (or aggregate)`,
         `Diabetes Adjusted Rate` = `Adjusted Rate 2011`) %>%
```

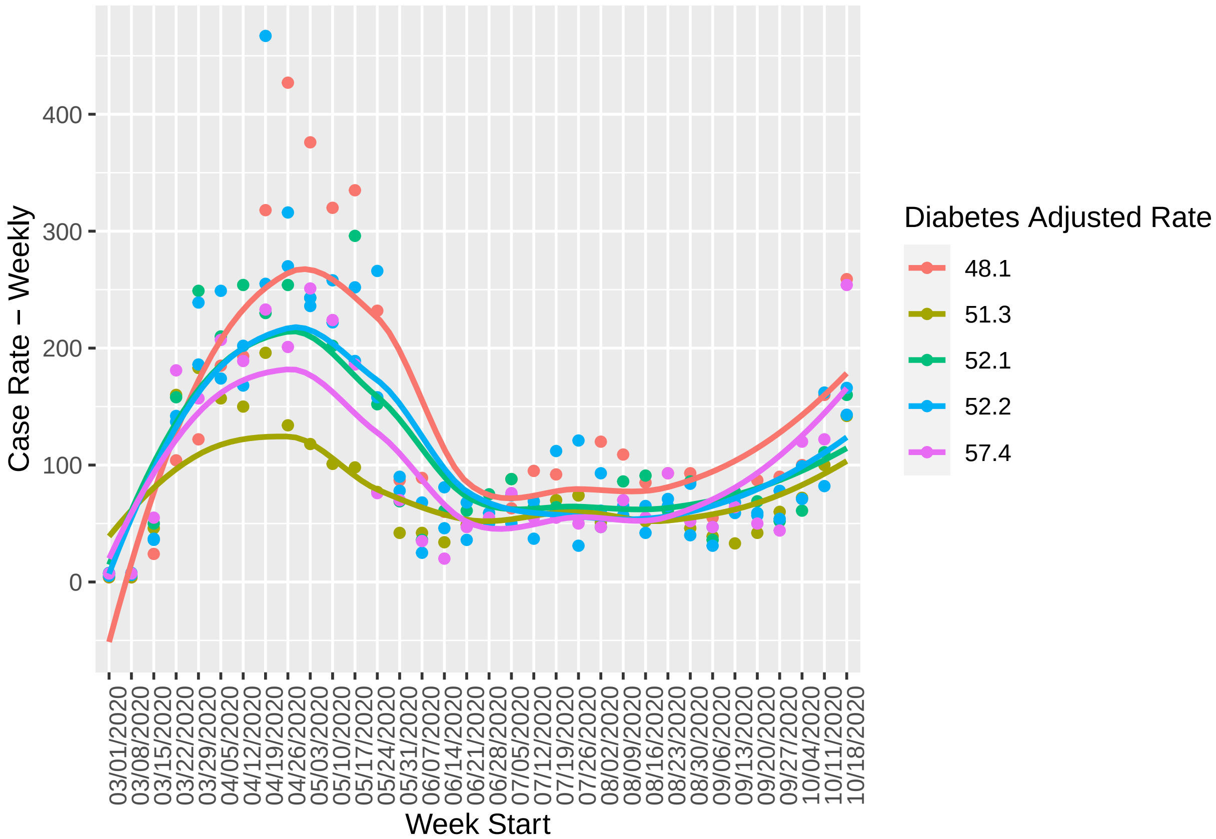
```
  arrange(`Diabetes Adjusted Rate`)

cases_with_diabetes <- cases %>%
  left_join(compressed_diabetes, by = "ZIP Code") %>%
  arrange(as.numeric(`Diabetes Adjusted Rate`)) %>%
  mutate(`diabetes_brackets` = cut(`Diabetes Adjusted Rate`,
                                   breaks = c(-Inf, 15, 30, 45, 60),
                                   labels = c("<15", "<30", "<45", "<60")))

cases_with_diabetes %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Case Rate - Weekly`, color = factor(`diabetes_brackets`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Case Rate - Weekly`, color = factor(`di
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Diabetes Adjusted Rate")
```




```
cases_with_diabetes %>%
  filter(`ZIP Code` %in% head(compressed_diabetes, 5)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Case Rate - Weekly`, color = factor(`Diabetes Adjusted Rate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Case Rate - Weekly`, color = factor(`Di
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Diabetes Adjusted Rate")
```



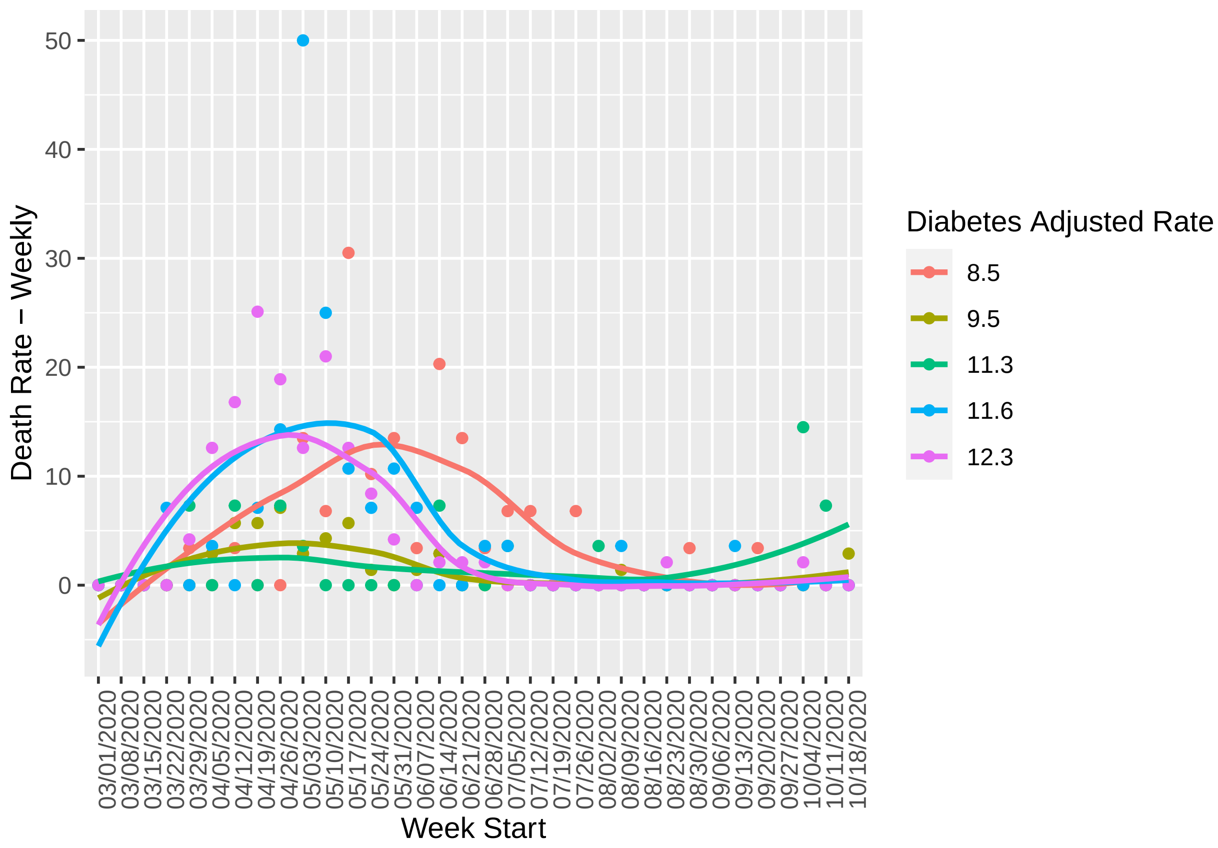

```
cases_with_diabetes %>%
  filter(`ZIP Code` %in% tail(compressed_diabetes, 6)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Case Rate - Weekly`, color = factor(`Diabetes Adjusted Rate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Case Rate - Weekly`, color = factor(`Di
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Diabetes Adjusted Rate")
```



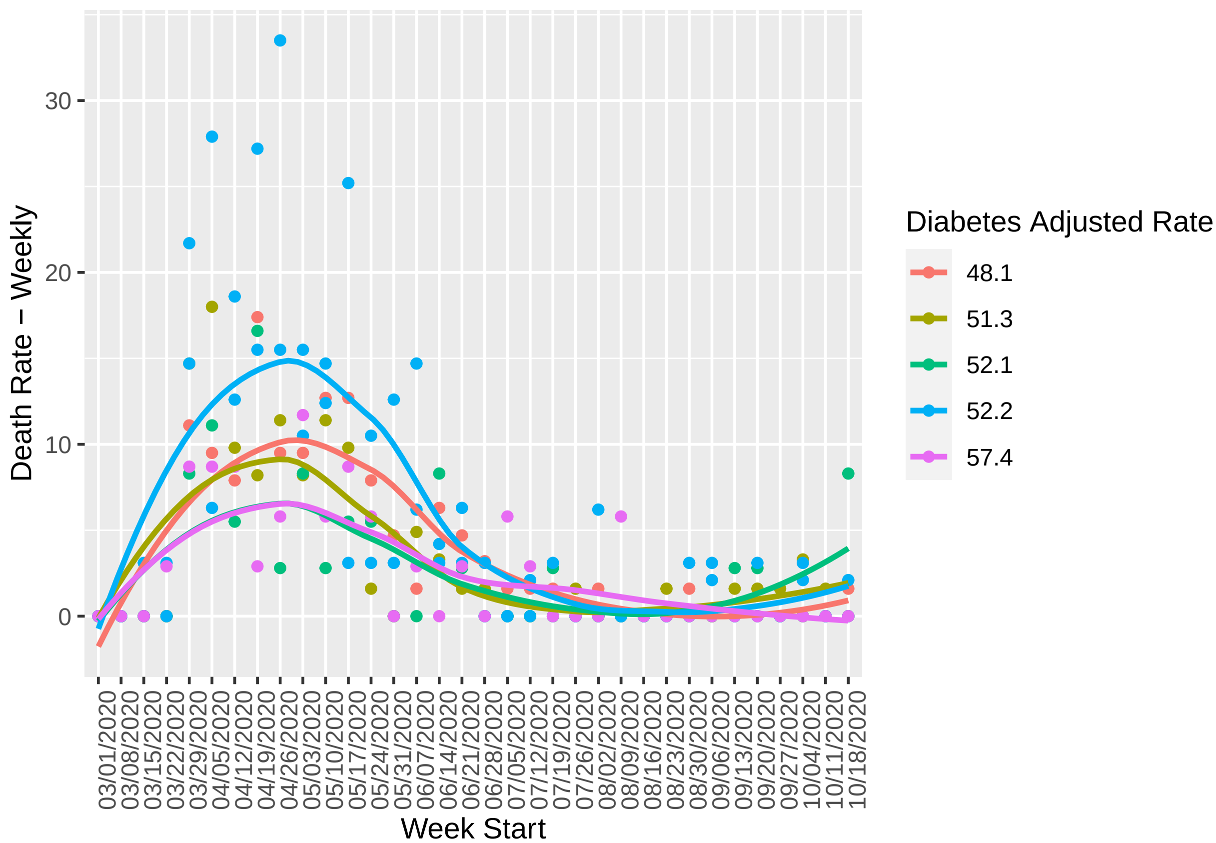
This is a visualization of the overall case rate in ZIP codes along with the rates within the lowest and highest diabetes adjusted rate. The adjusted rate is per 10,000 people and accounts for differences in population age distributions to allow for comparisons between multiple populations. The `NA` values within the first plot indicate that there was no data available for the diabetes adjusted rate within those ZIP codes. ZIP codes with the lowest diabetes rates all have a local maximum at around 04/26, a downwards slope, and an upwards trend following 07/05. For those with the highest rates of diabetes, there is the same trend, albeit with lower heights. The graphs only appear different due to the scale of y axis.

Death Rate over Weeks Based on Diabetes Rate

```
cases_with_diabetes %>%
  filter(`ZIP Code` %in% head(compressed_diabetes, 5)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Death Rate - Weekly`, color = factor(`Diabetes Adjusted Rate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Death Rate - Weekly`, color = factor(`I
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Diabetes Adjusted Rate")
```




```
cases_with_diabetes %>%
  filter(`ZIP Code` %in% tail(compressed_diabetes, 6)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Death Rate - Weekly`, color = factor(`Diabetes Adjusted Rate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Death Rate - Weekly`, color = factor(`D
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Diabetes Adjusted Rate")
```



This is a visualization of the death rate in ZIP codes with the lowest and highest diabetes adjusted rate. The ZIP codes with the lowest rates presented the same trends as the ZIP codes with the highest rate, peaking around 04/26 and flattening out for the most part. These graphs only seem different to the y scale. Diabetes does not appear to be a significant factor for the coronavirus death trend.

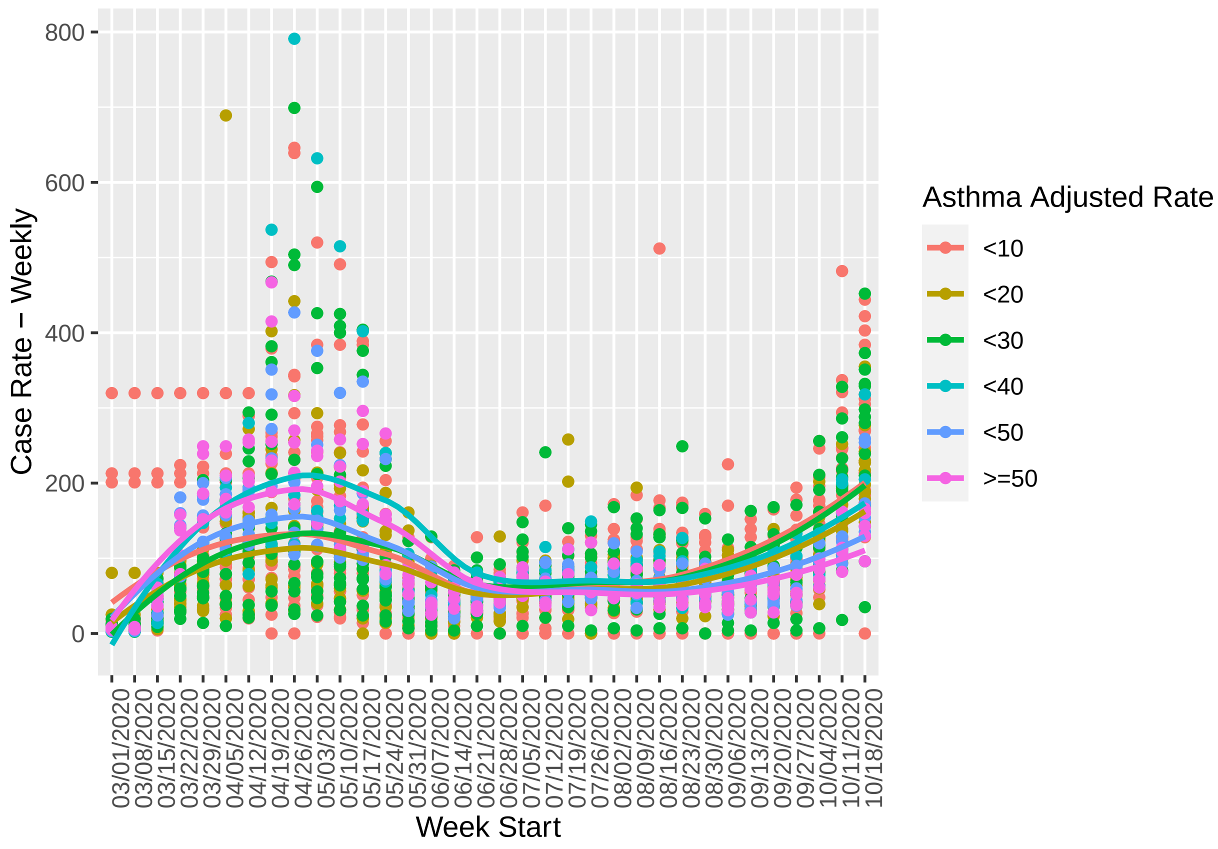
Visualizing Cases over Time with Asthma

I wanted to identify how the coronavirus affected those with asthma. I expected the case rate to be the same for both, but the death rate to be higher for those with higher rates of asthma compared to the those with lower rates of asthma. Since asthma is a respiratory disease, and the coronavirus affects the respiratory system, I expected a positive correlation. The trends aligned with my expectations.

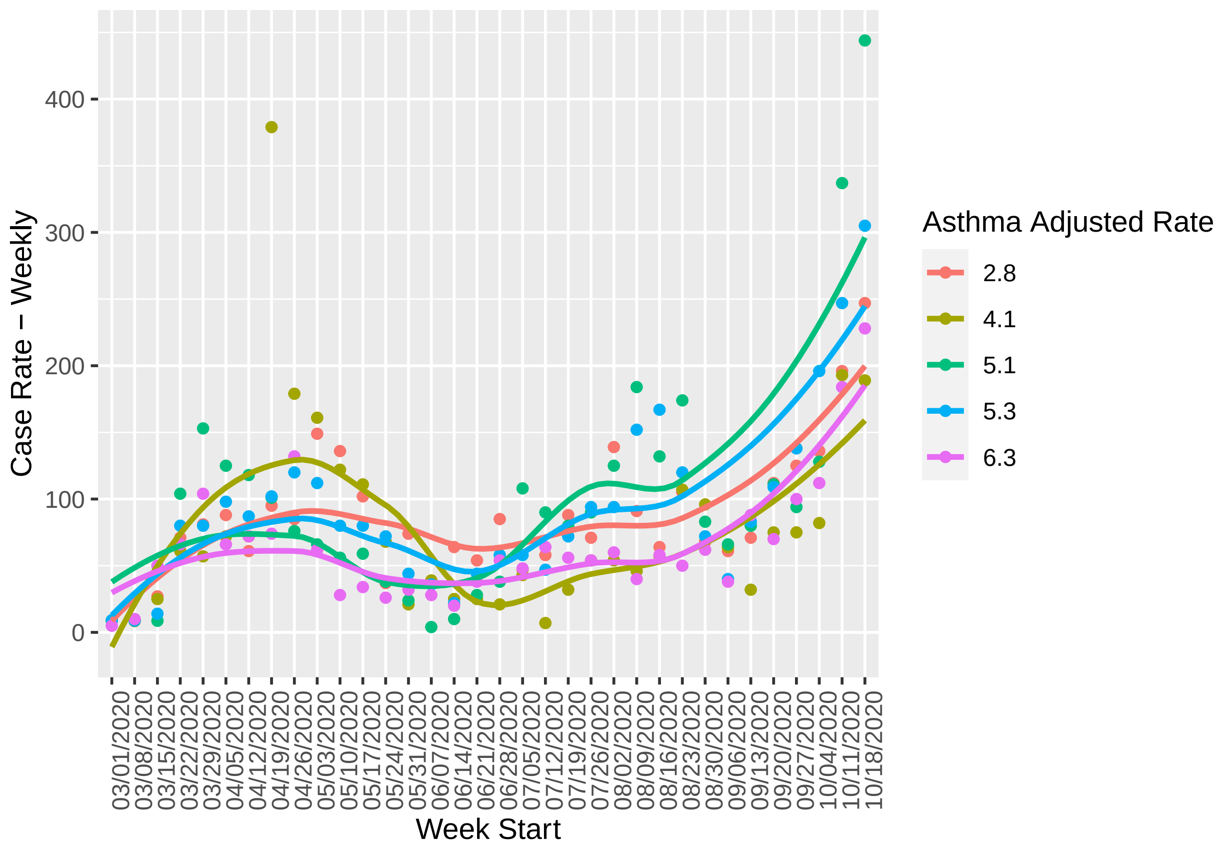
Case Rate over Weeks Based on Asthma Rate

```r
compressed_asthma <- asthma %>%
  select(`ZIP code (or aggregate)`, `Adjusted Rate 2011 (Ages 5-64 only)`) %>%
  rename(`ZIP Code` = `ZIP code (or aggregate)`,
         `Asthma Adjusted Rate` = `Adjusted Rate 2011 (Ages 5-64 only)`) %>%
  arrange(`Asthma Adjusted Rate`)

cases_with_asthma <- cases %>%
  left_join(compressed_asthma, by = "ZIP Code") %>%
  arrange(as.numeric(`Asthma Adjusted Rate`)) %>%
  mutate(`asthma_brackets` = cut(`Asthma Adjusted Rate`,
                              breaks = c(-Inf, 10, 20, 30, 40, 50, Inf),
                              labels = c("<10", "<20", "<30", "<40", "<50", ">=50")))
```

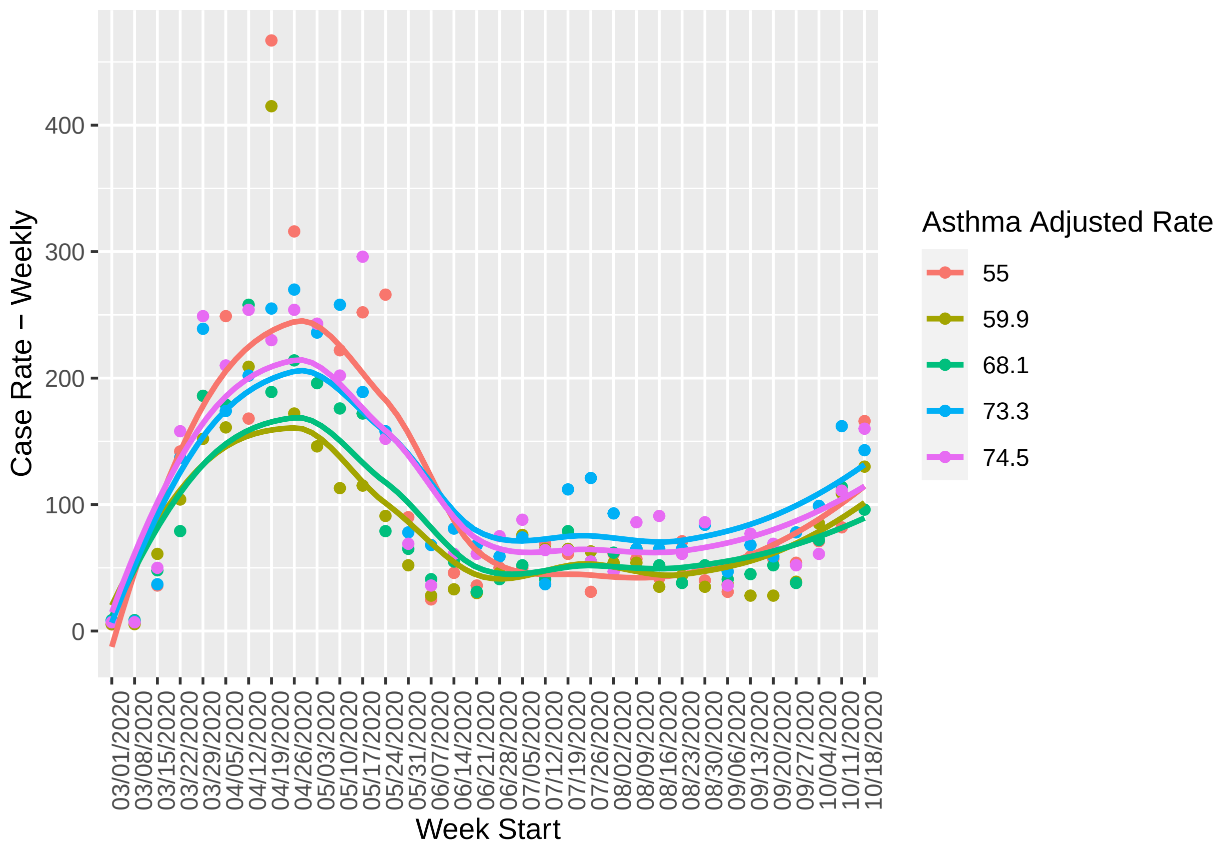
```
cases_with_asthma %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Case Rate - Weekly`, color = factor(`asthma_brackets`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Case Rate - Weekly`, color = factor(`as
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Asthma Adjusted Rate")
```




```
cases_with_asthma %>%
  filter(`ZIP Code` %in% head(compressed_asthma, 5)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Case Rate - Weekly`, color = factor(`Asthma Adjusted Rate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Case Rate - Weekly`, color = factor(`As
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Asthma Adjusted Rate")
```



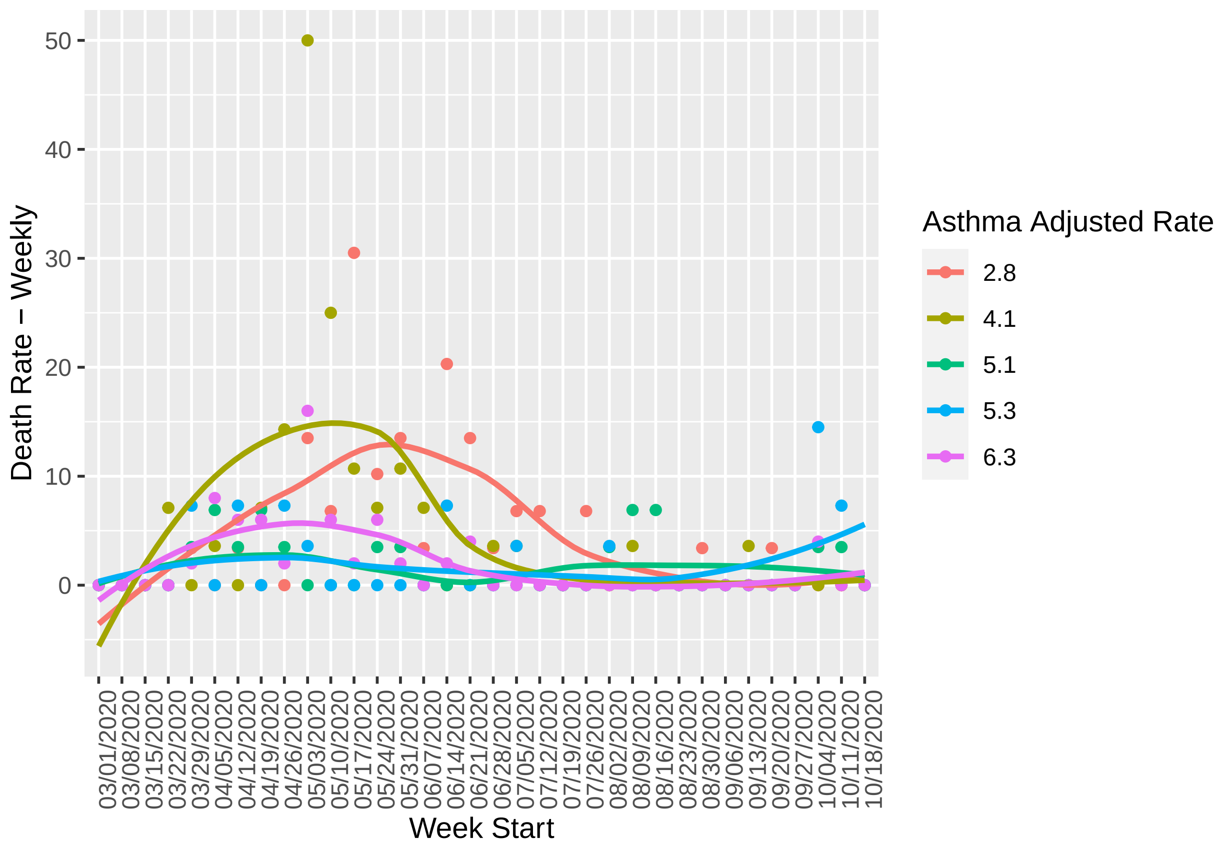

```
cases_with_asthma %>%
  filter(`ZIP Code` %in% tail(compressed_asthma, 5)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Case Rate - Weekly`, color = factor(`Asthma Adjusted Rate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Case Rate - Weekly`, color = factor(`As
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Asthma Adjusted Rate")
```



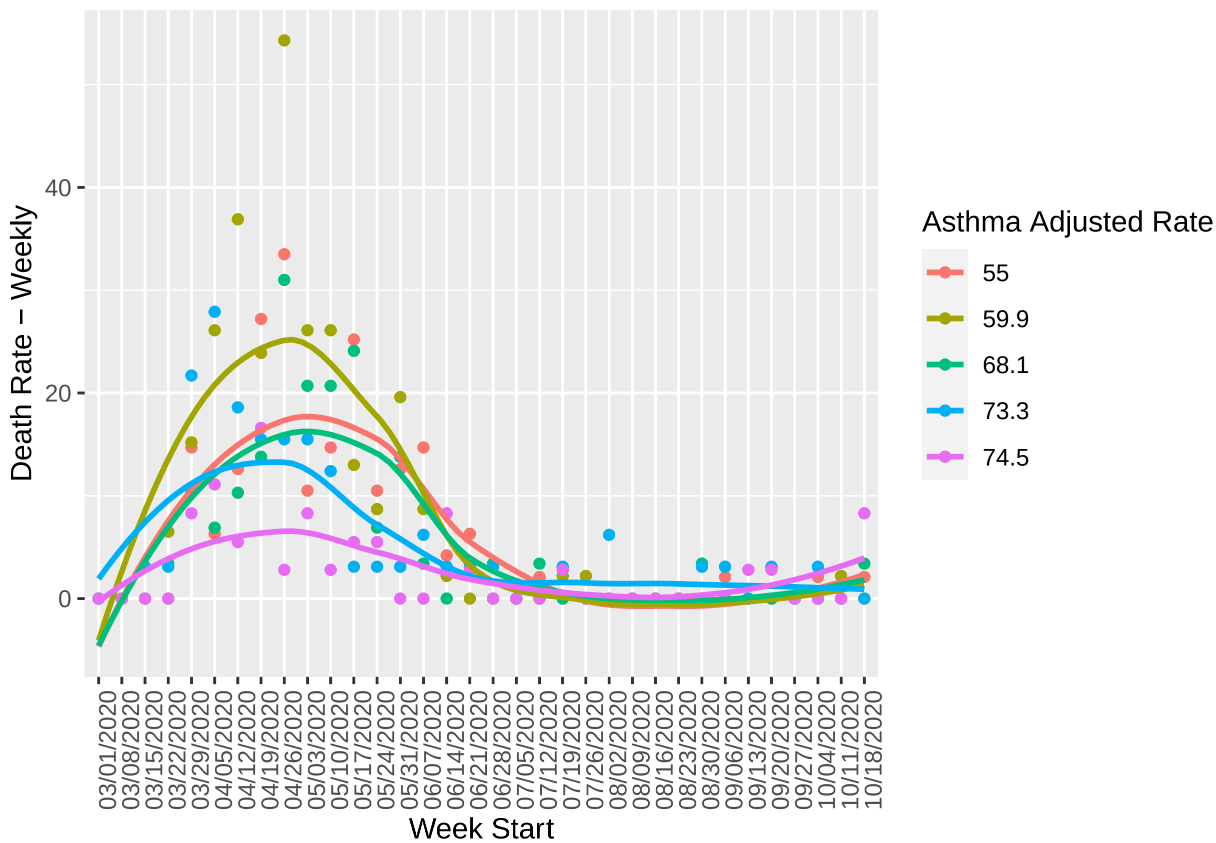
This is a visualization of the case rate in ZIP codes with aggregate asthma rates, along with the lowest and highest asthma adjusted rate. Since this is examining the case rate, I expected all populations to have around the same case rate given that asthma would worsen the effects of the coronavirus, but not increase the chances of receiving it. Those with the lowest asthma rates had a small bump around 04/19, but remained somewhat steady before increasing the week of 08/16. Those with the highest rates had the same peak from other visualizations, at around 04/26, before flattening and starting to pick up. It is potentially likely that asthma rates are correlated with income levels, as those with high incomes live in more environmentally friendly areas and those with low incomes may live in more polluted areas, leading to higher rates of asthma. As shown above, ZIP codes lower income levels had higher case rates.

Death Rate over Weeks Based on Asthma Rate

```
cases_with_asthma %>%
  filter(`ZIP Code` %in% head(compressed_asthma, 5)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Death Rate - Weekly`, color = factor(`Asthma Adjusted Rate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Death Rate - Weekly`, color = factor(`
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Asthma Adjusted Rate")
```




```
cases_with_asthma %>%
  filter(`ZIP Code` %in% tail(compressed_asthma, 5)$`ZIP Code`) %>%
  ggplot() +
  geom_point(aes(`Week Start`, `Death Rate - Weekly`, color = factor(`Asthma Adjusted Rate`))) +
  geom_smooth(aes(`Week Number` - min(cases$`Week Number`) + 1, `Death Rate - Weekly`, color = factor(`A
  theme(axis.text.x = element_text(angle = 90)) +
  labs(color = "Asthma Adjusted Rate")
```



This is a visualization of the death rate in ZIP codes with the lowest and highest asthma adjusted rate. As with the diabetes plot above, those with the lowest rates of asthma tended to have a flatter curve, with a smaller death rate.Those with the highest rates of asthma had a spike at around 04/26 before returning to near 0. This is one of the clearest visualizations demonstrating the correlation between a medical condition and COVID.

Conclusion

In my research, I found these general trends:

- There existed a local maximum for the case and death rate during the week of 04/26. As coronavirus cases trended upwards after 09/20, the death rate remained relatively stable, suggesting that healthcare had built on their previous experience to lower the death rate.
- ZIP codes with higher incomes tended to have less cases. This may be a result of the fact that generally, people within those high income areas perform more white-collar work and thus transitioned to working from home, leading to less contact with people and less cases overall.
- ZIP codes with higher incomes tended to have less deaths. This is most likely a result of insurance and increased access to healthcare, leading to less deaths.
- ZIP codes for those with high and low diabetes rates had the same case and death rates. This does not align with my initial thoughts. While the case rate may not be affected by medical conditions, I expected the death rate to be, as medical conditions may complicate survival rate and lead to exacerbated effects of the coronavirus.
- ZIP codes with high asthma rates climbed significantly faster than those with high asthma rates, starting from 07/26. It is unclear why, but it is possible that asthma levels are inversely related with income, suggesting that the high asthma areas are also low income areas, and thus have less access to healthcare.
- ZIP codes with high asthma rates had higher death rates than the death rates for ZIP codes with low

asthma rates. This is supported by the fact that the coronavirus is a respiratory disease, and asthma affects the respiratory system.

My analysis revealed correlations among a few socioeconomic factors. However, there are more socioeconomic factors and datasets online available to explore. Drawing from and analyzing those datasets could provide further insight into coronavirus trends and how to potentially mitigate its spread.

Citations

Basic usage of tidycensus. https://walker-data.com/tidycensus/articles/basic-usage.html.

City of Chicago. (2020, October 15). *COVID-19 Cases, Tests, and Deaths by ZIP Code.* Chicago Data Portal. https://data.cityofchicago.org/Health-Human-Services/COVID-19-Cases-Tests-and-Deaths-by-ZIP-Code/yhhz-zm2v.

City of Chicago. (2020, September 10). *COVID-19 Testing Sites. Chicago Data Portal.* https://data.cityofchicago.org/Health-Human-Services/COVID-19-Testing-Sites/thdn-3grx.

COVID-19 Orders. (2020). https://www.chicago.gov/city/en/sites/covid-19/home/health-orders.html.

How to obtain median income data for zip codes. Reddit. (2020). https://www.reddit.com/r/datasets/comments/hixfeo/how_to_obtain_median_income_data_for_zip_codes/.

Illinois Department of Public Health (IDPH). (2012, August 6). *Public Health Statistics - Diabetes hospitalizations in Chicago, 2000 - 2011.* Chicago Data Portal. https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Diabetes-hospitalizations/vekt-28b5.

Illinois Department of Public Health (IDPH). (2012, September 17). *Public Health Statistics - Asthma hospitalizations in Chicago, by year, 2000 - 2011.* Chicago Data Portal. https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Asthma-hospitalizations-i/vazh-t57q.