SDS 323 Project

Samuel Higgins

5/7/2020

Abstract

For this project, I wanted to look at Covid-19 cases over time for a few Texas counties and compare them through data visualization. I wanted to know which counties are experiencing a high number of cases and which counties are experiencing a low number of cases. Cases are ultimately dependent on the number of testing that's done. What I found was that some counties are not on the same page in terms of cases reported. Bexar county for example started recording cases early February, while other counties, like Zapata county, started recording cases early April. There are potential factors that may explain this, such as a lack of testing kits due to a delay or, in the case of San Antonio, the early exposure to the virus.

Introduction

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

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

```
cvd <- read.csv("https://github.com/zhamuelh/samprojects2/raw/master/Data%20sets/counties_covid19.csv")
head(cvd)
```

```
##         date    county      state  fips cases deaths
## 1 2020-01-21 Snohomish Washington 53061     1      0
## 2 2020-01-22 Snohomish Washington 53061     1      0
## 3 2020-01-23 Snohomish Washington 53061     1      0
## 4 2020-01-24      Cook   Illinois 17031     1      0
## 5 2020-01-24 Snohomish Washington 53061     1      0
## 6 2020-01-25    Orange California  6059     1      0
```

```
txcv <- cvd %>%
  filter(state == "Texas") %>% pivot_wider(names_from = state, values_from = county) %>%
  rename(county_tx = Texas)
head(txcv)
```

```
## # A tibble: 6 x 5
##   date        fips cases deaths county_tx
##   <fct>      <int> <int>  <int> <fct>
```

```
## 1 2020-02-12 48029     1      0 Bexar
## 2 2020-02-13 48029     2      0 Bexar
## 3 2020-02-14 48029     2      0 Bexar
## 4 2020-02-15 48029     2      0 Bexar
## 5 2020-02-16 48029     2      0 Bexar
## 6 2020-02-17 48029     2      0 Bexar
```

```
tail(txcv)
```

```
## # A tibble: 6 x 5
##   date        fips cases deaths county_tx
##   <fct>      <int> <int>  <int> <fct>
## 1 2020-05-05 48497    27      2 Wise
## 2 2020-05-05 48499    11      0 Wood
## 3 2020-05-05 48501     2      0 Yoakum
## 4 2020-05-05 48503     4      1 Young
## 5 2020-05-05 48505     7      0 Zapata
## 6 2020-05-05 48507     1      0 Zavala
```

I guess you could say my question for this analysis is essentially "how's everybody doing" in regards to the current pandemic in Texas. I will mainly look at Covid-19 cases over time and mortality over time for a few counties. This study could potentially be a good precursor to future studies. One such study could be evaluating changes in certain populations throughout the duration of the pandemic. Another future study could look at predicting the rate of Covid-19 spread on variables like population density and the number of open grocery stores in a city.

Methods

For this project, I imported two data sets. The first data set (annotated as cvd in the introduction chunk) contains the number of Covid-19 cases and deaths over time at the county level in the U.S. This data was released publicly on the New York Times' GitHub account. The second data set (annnotated as txp) contains population data for each Texas county as well as variables for the number of people in poverty and median income for that county. This data set was compiled from data obtained from the U.S Census Bureau. I tidyed the 'cvd' set to only display Texas data at the county level (txcv). Later I joined a summarised version of 'txcv' (txcv_cases) to 'txp' to obtain a table (txcv2) that displays summarised information on Covid-19 cases and deaths for each county (which will be outdated as the days progress). With 'txcv' I created time series plots for a few counties visualizing Covid-19 cases and deaths over time. 'Txcv2' was created as a summary table (Table 1), but also to calculate incidence for each county.

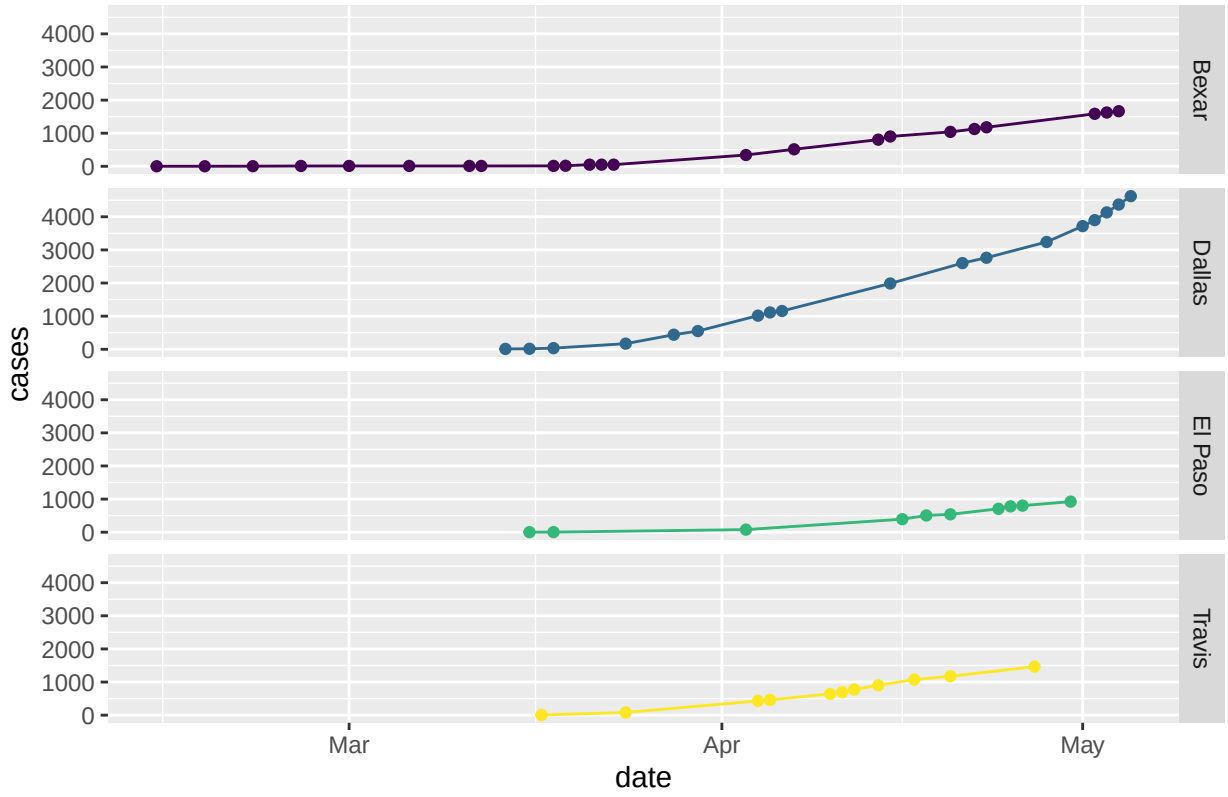
Time Series Data Visualization: Texas Counties

```
txcv$date <- as.Date(txcv$date)

# Time series plot for Travis, El Paso, Bexar, Dallas Counties
txcv %>% filter(county_tx == c("Travis", "El Paso", "Bexar", "Dallas")) %>%
  ggplot(aes(x = date, y = cases )) +
  geom_line(aes(color = county_tx)) +
  geom_point(aes(color = county_tx),na.rm = T) +
  facet_grid("county_tx") +
  theme(legend.position = "none") +
  ggtitle("Figure 1. Texas: Covid-19 Cases Feb 12 - May 5, 2020 ") +
```

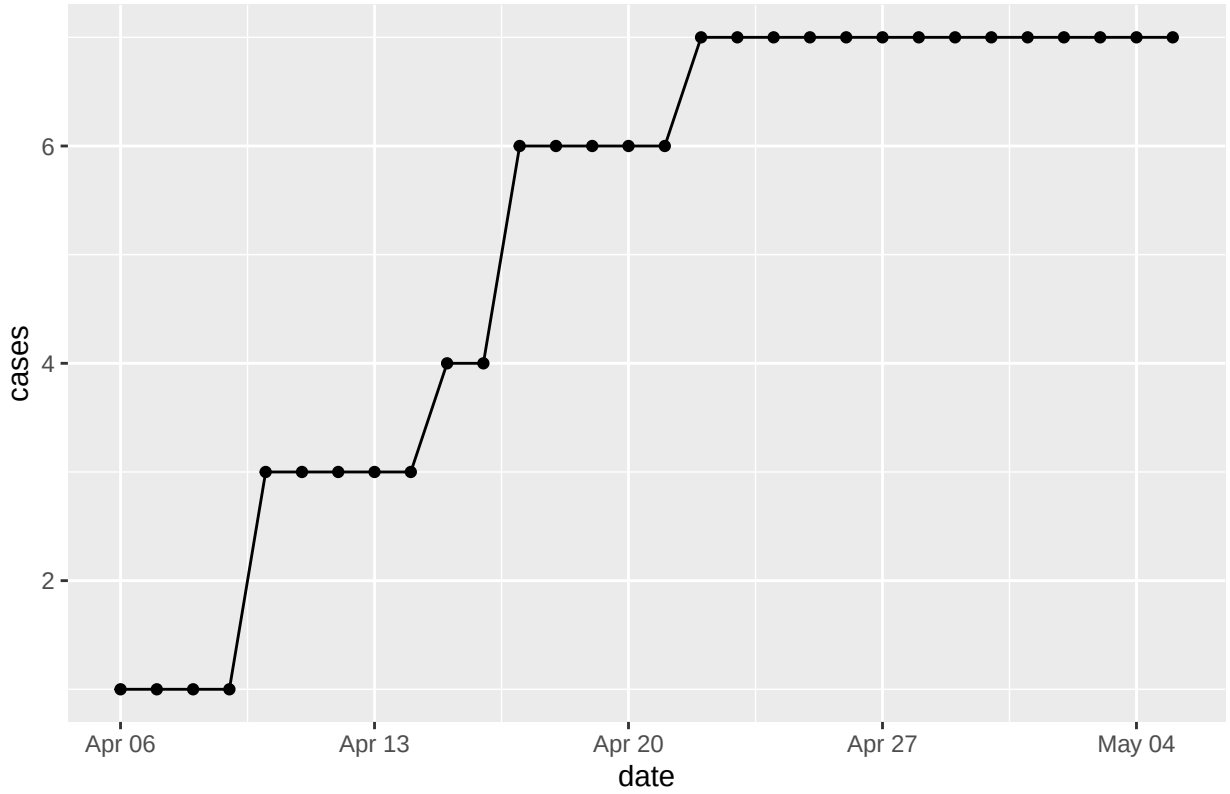
```
  scale_colour_viridis_d()
```

Figure 1. Texas: Covid−19 Cases Feb 12 − May 5, 2020



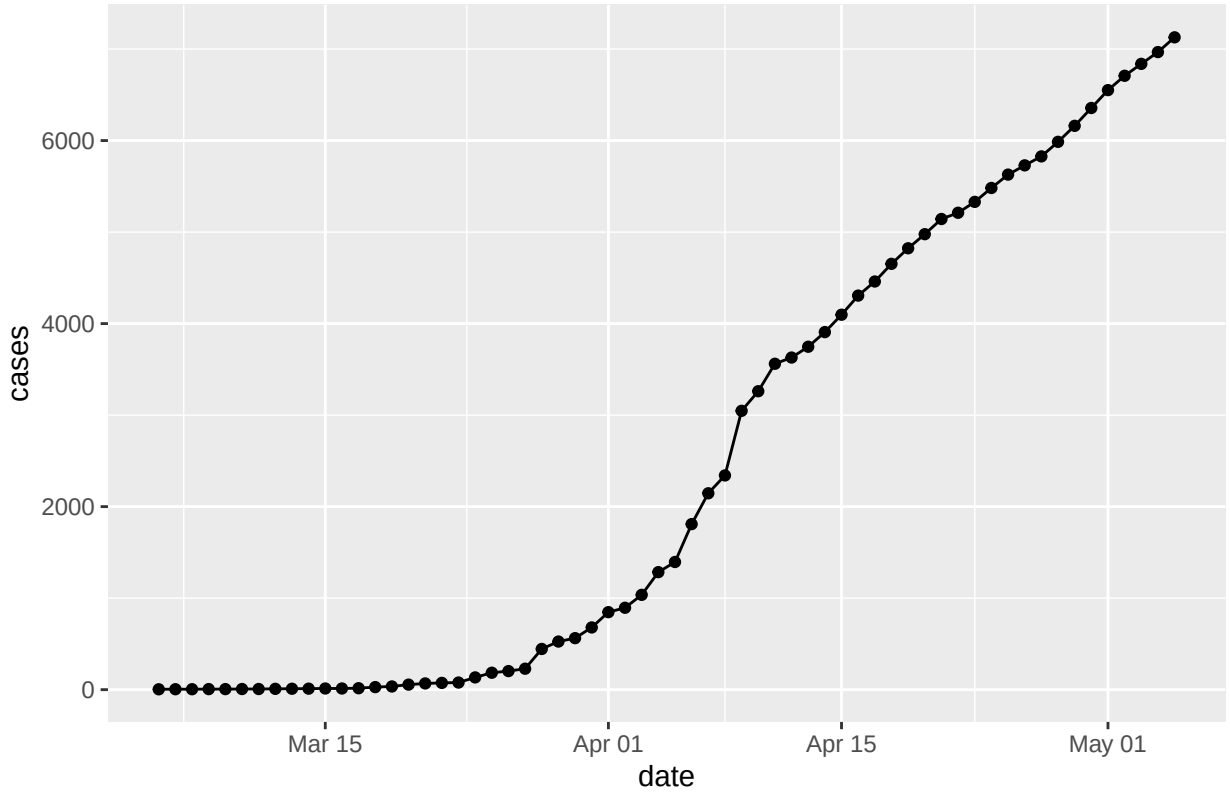

```
#...for Zapata County
txcv %>% filter(county_tx == "Zapata") %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line() + geom_point() +
  ggtitle("Figure 2. Zapata County Covid-19 Cases Apr 6 - May 5, 2020")
```

Figure 2. Zapata County Covid−19 Cases Apr 6 − May 5, 2020




```
#...for Harris County
txcv %>% filter(county_tx == "Harris") %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line() + geom_point() +
  ggtitle("Figure 3. Harris County Covid-19 Cases Mar 5 - May 5, 2020")
```

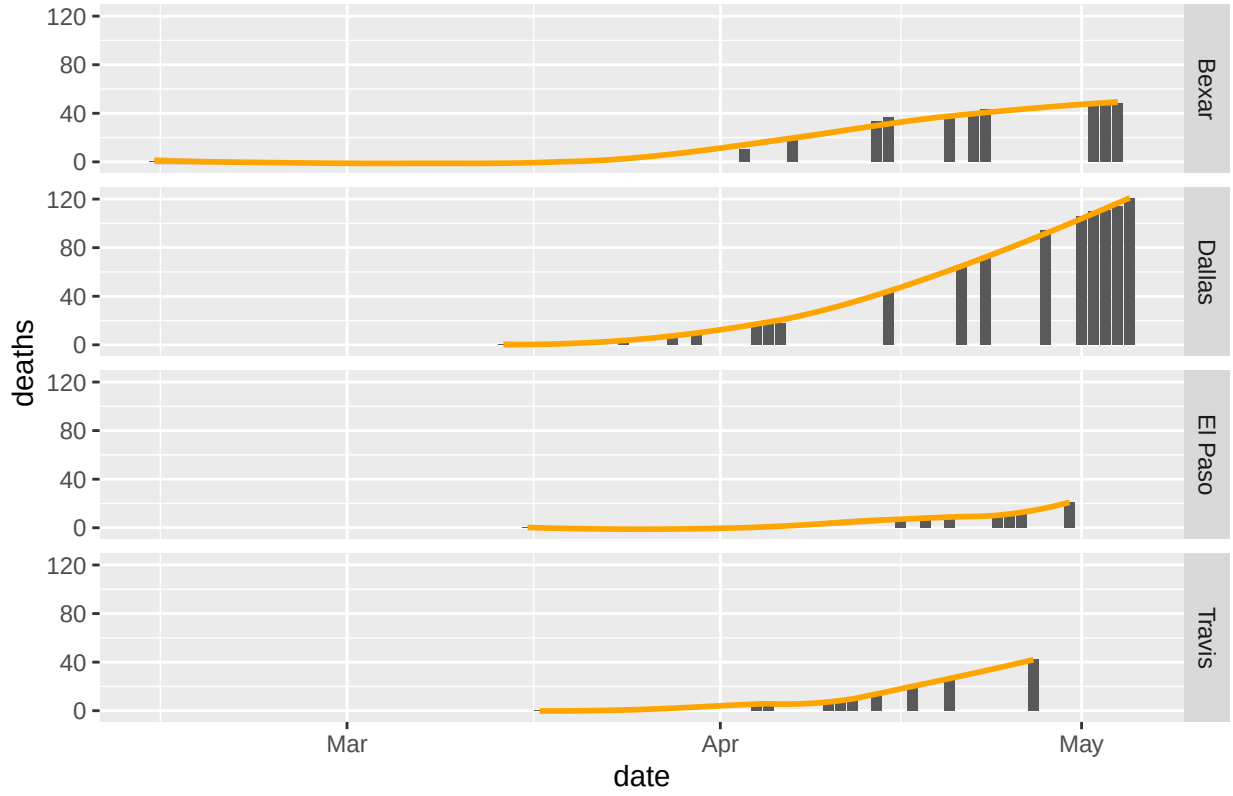
Figure 3. Harris County Covid−19 Cases Mar 5 − May 5, 2020




```
# Covid-19 Mortality over time
txcv %>% filter(county_tx == c("Travis", "El Paso", "Bexar", "Dallas")) %>%
  ggplot(aes(x = date, y = deaths)) +
  geom_bar(stat = "identity") +
  facet_grid("county_tx") + stat_smooth(color = "orange") +
  ggtitle("Figure 4. Texas: Covid-19 Mortality Feb 12 - May 5, 2020")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

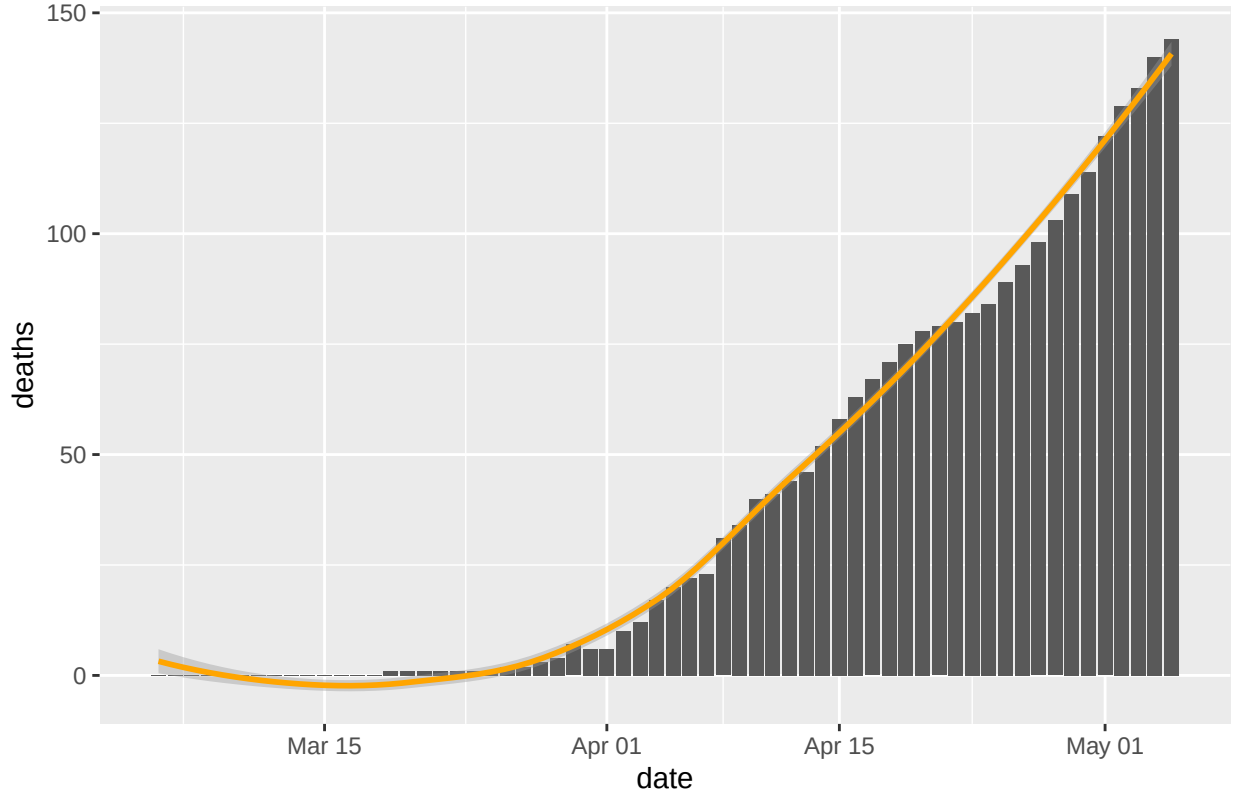
Figure 4. Texas: Covid−19 Mortality Feb 12 − May 5, 2020




```
# Harris County
txcv %>% filter(county_tx == "Harris") %>%
  ggplot(aes(x = date, y = deaths)) +
  geom_bar(stat = "identity") +
  stat_smooth(color = "orange") +
  ggtitle("Figure 5. Harris County Mortality Mar 19 - May 5, 2020")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Figure 5. Harris County Mortality Mar 19 − May 5, 2020



Results

Figure 1 is a time series plot of Covid-19 cases over time, faceted by county. Here we see that Bexar county starts recording cases Februaruy 12th, and as of May 5th had a case count of 1689. Dallas, El Paso, and Travis record cases on the 14th, 16th, and 17th of March, respectively. As of May 5th, Dallas had 4623 cases, El Paso had 1080 cases, and Travis had 1876 cases of the coronavirus. Highlighted in Figure 3, Harris county, which contains the Houston metropolis, had the highest number of recorded cases on May 5th at 7128 cases. Figure 2 displays Zapata county, which recorded its first case April 6th, and had 7 cases on May 5th (more about Zapata county will be discussed in the conclusion).

Figure 4 is a time series plot of deaths over time for the counties displayed in Figure 1. On May 5th, Travis had 58 Covid-19 related deaths, El Paso had 22, Bexar had 52, and Dallas had 121. Like with the number of cases, Harris county also had the highest number of deaths, at 144 as of May 5th.

Slopes for plots

```
tx_trav <- txcv %>% filter(county_tx == "Travis")
lm1 <- lm(cases ~ date, data = tx_trav)
summary(lm1)
```

```
##
## Call:
## lm(formula = cases ~ date, data = tx_trav)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -185.90 -112.21   10.94   75.47  304.72
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.053e+05  1.902e+04  -37.09   <2e-16 ***
## date         3.845e+01  1.036e+00   37.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 118.6 on 52 degrees of freedom
## Multiple R-squared:  0.9636, Adjusted R-squared:  0.9629
## F-statistic:  1378 on 1 and 52 DF,  p-value: < 2.2e-16
```

```
tx_elp <- txcv %>% filter(county_tx == "El Paso")
lm2 <- lm(cases ~ date, data = tx_elp)
summary(lm2)
```

```
##
## Call:
## lm(formula = cases ~ date, data = tx_elp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -175.07 -117.15  -17.81  104.78  232.58
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.949e+05  1.902e+04  -20.76   <2e-16 ***
## date         2.152e+01  1.036e+00   20.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 118.6 on 52 degrees of freedom
## Multiple R-squared:  0.8925, Adjusted R-squared:  0.8904
## F-statistic: 431.8 on 1 and 52 DF,  p-value: < 2.2e-16
```

```
tx_bex <- txcv %>% filter(county_tx == "Bexar")
lm3 <- lm(cases ~ date, data = tx_bex)
summary(lm3)
```

```
##
## Call:
## lm(formula = cases ~ date, data = tx_bex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -370.25 -221.46    4.35  198.15  454.52
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.600e+05  2.015e+04  -17.86   <2e-16 ***
## date         1.964e+01  1.099e+00   17.88   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 244.1 on 82 degrees of freedom
## Multiple R-squared:  0.7959, Adjusted R-squared:  0.7934
## F-statistic: 319.8 on 1 and 82 DF,  p-value: < 2.2e-16
```

```
tx_dal <- txcv %>% filter(county_tx == "Dallas")
lm4 <- lm(cases ~ date, data = tx_dal)
summary(lm4)
```

```
##
## Call:
## lm(formula = cases ~ date, data = tx_dal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -349.68 -243.21  -55.84  123.85  861.81
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.471e+06  4.628e+04  -31.80   <2e-16 ***
## date         8.023e+01  2.521e+00   31.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 313.1 on 55 degrees of freedom
## Multiple R-squared:  0.9485, Adjusted R-squared:  0.9476
## F-statistic:  1013 on 1 and 55 DF,  p-value: < 2.2e-16
```

```
tx_har <- txcv %>% filter(county_tx == "Harris")
lm5 <- lm(cases ~ date, data = tx_har)
summary(lm5)
```

```
##
## Call:
## lm(formula = cases ~ date, data = tx_har)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1263.5  -670.1   227.6   423.2  1613.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.468e+06  9.841e+04  -25.08   <2e-16 ***
## date         1.346e+02  5.361e+00   25.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 755.4 on 60 degrees of freedom
## Multiple R-squared:  0.9131, Adjusted R-squared:  0.9116
## F-statistic: 630.1 on 1 and 60 DF,  p-value: < 2.2e-16
```

Results (cont.)

In the chunk above, I computed the slopes for the plots in the figures that were discussed earlier. For Travis county, for every (one) day, Covid-19 cases inreases by 3.845 on average.

For El Paso county, Covid-19 cases increases by 2.152 on averge, every day. Bexar county, had an average slope of 1.964 for Covid-19 cases per day. Dallas county had an average slope of 8.023 for Covid-19 cases per day. Finally, for Harris county, Covid-19 cases increases by 134.6 on average for every day.

Conclusion

To conclude, most if not all of our "metropolitan counties" are affected by the Covid-19 pandemic to some varying degree. Harris county is definitely taking the brunt of the pandemic, along with Dallas county. For our smaller counties, there may not be a full picture as of yet. As mentioned previously, Zapata county has only recorded 7 cases since April 6th, which could be indicative of a lack of testing, or it could just mean that the disease is spreading a lot more slowly there. There are counties that have recorded less cases than Zapata, if you filter 'txcv2' to where you return the minimum number of cases for a county, you will get multiple counties that have only 1 case recorded. All of these counties have less than 20,000 people, some less than 10,000. Although most of this is just comparison, I believe this report can provide potential foresight for Covid-19 relief for our state.

Covid-19 Summary and Population Measures

```
txp <- read.csv("https://github.com/zhamuelh/samprojects2/raw/master/Data%20sets/txcounty_poverty.csv")

txp <- txp %>% rename(county_tx = County.Name) %>% rename(population_est  = Population.Estimates) %>%
  rename(n_poverty = Number.in.Poverty) %>% rename(med_income = Median.Household.income) %>%
  select(county_tx, population_est, n_poverty, med_income)

txcv_cases <- txcv %>% group_by(county_tx) %>% summarise(current_cases = max(cases),
                                                         total_deaths = max(deaths))
txcv2 <- left_join(txcv_cases, txp, by = "county_tx")
```

```
## Warning: Column `county_tx` joining factors with different levels, coercing to
## character vector
```

```
txcv2 <- txcv2 %>% mutate(incidence = current_cases/population_est) %>% na.omit #Clay County omitted du

txcv2 %>% filter(current_cases == min(current_cases))
```

```
## # A tibble: 25 x 7
##    county_tx current_cases total_deaths population_est n_poverty med_income
##    <chr>             <int>        <int>          <int>     <int>      <int>
##  1 Bailey                1            0           7256      1087      45051
##  2 Brewster              1            0           9128      1475      45670
##  3 Briscoe               1            0           1535       245      44409
##  4 Brooks                1            0           7242      2183      30116
##  5 Childress             1            0           7007      1180      42291
##  6 Cochran               1            0           2931       609      42873
##  7 Coke                  1            0           3340       469      44804
##  8 Coleman               1            0           8483      1604      39484
##  9 Collings~             1            0           2943       502      39536
## 10 Concho                1            0           2044       702      40987
## # ... with 15 more rows, and 1 more variable: incidence <dbl>
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

Table 1: Table 1. Covid-19 Summary and Population Measures per County

county_tx	current_cases	total_deaths	population_est	n_poverty	med_income	incidence
Anderson	34	0	58854	8778	45969	0.0005777
Andrews	21	0	19232	1922	84946	0.0010919
Angelina	79	0	92353	15052	46653	0.0008554
Aransas	2	0	23031	4642	46912	0.0000868
Armstrong	2	0	1949	197	57210	0.0010262
Atascosa	19	1	50276	7803	50594	0.0003779
Austin	13	0	31724	3031	59942	0.0004098
Bailey	1	0	7256	1087	45051	0.0001378
Bandera	6	0	22874	3016	53008	0.0002623
Bastrop	98	2	88296	10673	61883	0.0011099
Bee	6	0	33039	6607	41806	0.0001816
Bell	210	3	356587	44865	54292	0.0005889
Bexar	1689	52	1991779	334215	54210	0.0008480
Blanco	6	0	11835	1176	64832	0.0005070
Bosque	5	0	18882	2721	48460	0.0002648
Bowie	104	10	97488	16794	47339	0.0010668
Brazoria	570	7	383058	36342	74225	0.0014880
Brazos	219	17	228292	49310	50113	0.0009593
Brewster	1	0	9128	1475	45670	0.0001096
Briscoe	1	0	1535	245	44409	0.0006515
Brooks	1	0	7242	2183	30116	0.0001381
Brown	38	6	39470	4778	47097	0.0009628
Burleson	14	0	18381	2686	53626	0.0007617
Burnet	24	0	48190	5099	56696	0.0004980
Caldwell	23	0	42956	5850	52588	0.0005354

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

```
kable(head(txcv2, n = 25), caption = "Table 1. Covid-19 Summary and Population Measures per County") %>%
  kable_styling(latex_options = c("striped", "condensed"))
```

```
kable(tail(txcv2, n = 25)) %>% kable_styling(latex_options = c("striped", "condensed"))
```

county_tx	current_cases	total_deaths	population_est	n_poverty	med_income	incidence
Trinity	9	0	14643	2581	40486	0.0006146
Tyler	7	0	22474	3584	44041	0.0003115
Upshur	15	0	40998	6455	48554	0.0003659
Uvalde	7	0	27691	5999	39725	0.0002528
Val Verde	13	0	50624	8950	44276	0.0002568
Van Zandt	16	1	56596	8598	51152	0.0002827
Victoria	143	5	92059	13855	51646	0.0015534
Walker	298	2	75055	14525	43681	0.0039704
Waller	33	0	54349	7776	59807	0.0006072
Washington	147	18	35801	4566	54332	0.0041060
Webb	400	17	281964	69860	44919	0.0014186
Wharton	41	0	41141	6909	48040	0.0009966
Wheeler	11	0	5059	721	48018	0.0021743
Wichita	70	2	133814	21221	47477	0.0005231
Wilbarger	1	0	12548	1905	44735	0.0000797
Willacy	13	1	21475	6369	33171	0.0006054
Williamson	333	11	571610	35829	87817	0.0005826
Wilson	34	3	51077	5400	71207	0.0006657
Winkler	3	0	7696	1069	56781	0.0003898
Wise	27	2	69647	6643	66387	0.0003877
Wood	11	0	45229	6625	48384	0.0002432
Yoakum	2	0	9109	1028	61560	0.0002196
Young	4	1	18304	2468	49301	0.0002185
Zapata	7	0	13903	4529	33160	0.0005035
Zavala	1	0	12338	3701	30076	0.0000811