ACT01 - Data Mining

Kyle Scott, Lakshya Rathore, Nadine Rose, Bailey LaRea

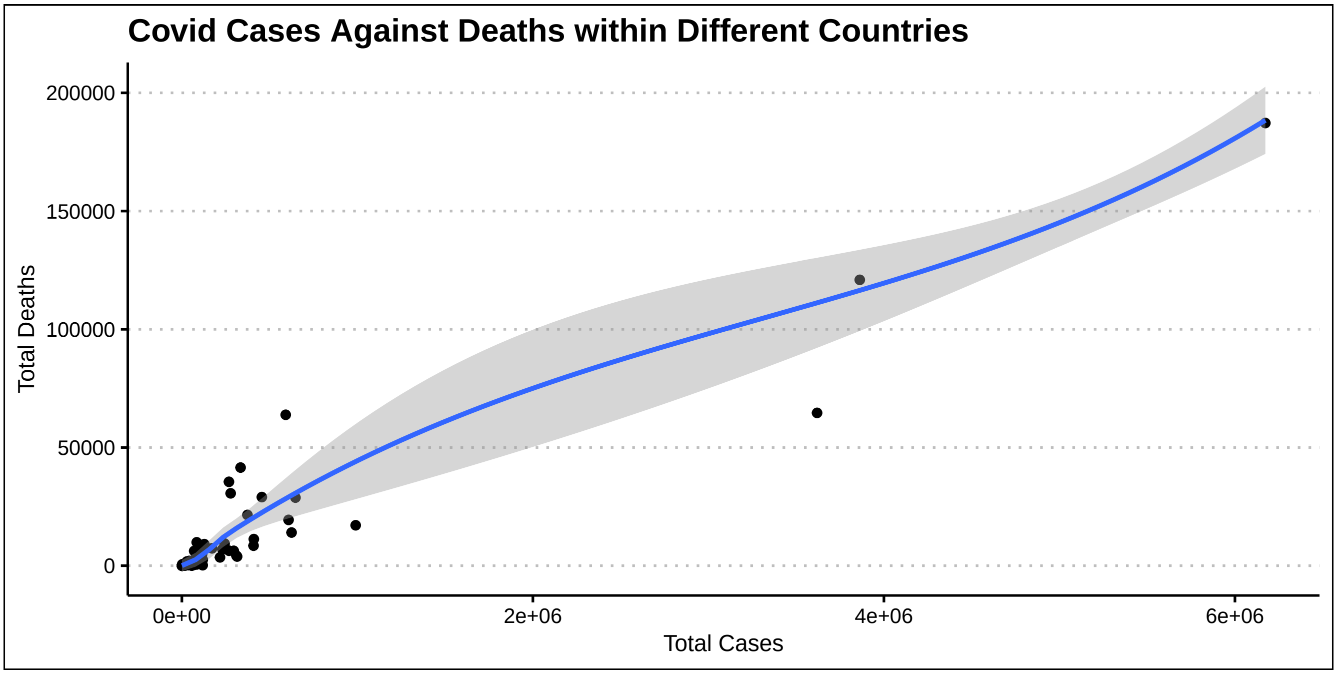
Question 1.1

```r
library(readr)


## read in the covid csv
covid <- read_csv("COVID_08312020.csv")
head(covid)
```

```
## # A tibble: 6 x 6
##   Country     `Total Cases` `Total Deaths` TOTCases_1M `TOTDeath_!M` TotalTested
##   <chr>               <dbl>          <dbl>       <dbl>         <dbl>       <dbl>
## 1 Afghanistan         38162           1402         977            36      102598
## 2 Albania              9380            280        3260            97       57618
## 3 Angola               2624            107          79             3       64747
## 4 Argentina          408426           8457        9023           187     1242269
## 5 Armenia             43750            877       14760           296      205450
## 6 Australia           25670            611        1005            24     6167592
```

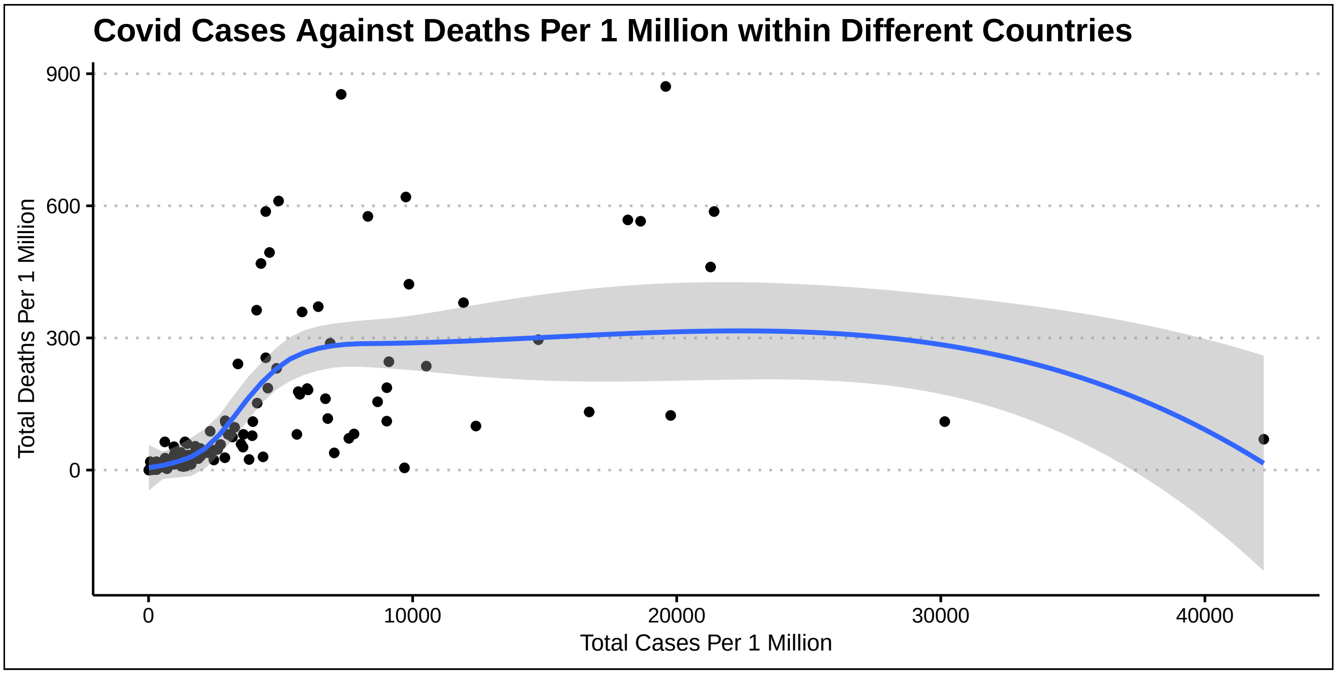
Question 1.2

```
library(ggplot2)

library(ggthemes)

## create a scatterplot with TotalCases and TotalDeaths

scplt1 <- ggplot(data = covid,

                 aes(x = `Total Cases`, y = `Total Deaths`)) +

          geom_point() +

          geom_smooth(method = "loess") +

          ggtitle(label = "Covid Cases Against Deaths within Different Countries") +

          theme_clean()

scplt1
```



Question 1.3

```
library(ggplot2)

library(ggthemes)

## create a scatterplot with TOTCases_1M and TOTDeath_MPOP

scplt2 <- ggplot(data = covid,

                 aes(x = `TOTCases_1M`, y = `TOTDeath_!M`)) +

          geom_point() +

          geom_smooth(method = "loess") +

          ggtitle(label = "Covid Cases Against Deaths Per 1 Million within Different Countries") +

          xlab("Total Cases Per 1 Million") +

          ylab("Total Deaths Per 1 Million") +

          theme_clean()

scplt2
```



Question 1.4

```r
library(psych)

library(kableExtra)

tbl <- describe(covid, skew = TRUE)

tbl <- tbl[-1,]

variance <- (tbl$sd)^2

variance <- round(variance, digits = 3)

tbl <- cbind(tbl, variance)

tbl <- tbl[, c(8, 3, 5, 14, 4, 9, 11)]

kable(tbl, caption = "Summary Statistics for Covid Data", linesep = "\\addlinespace", digits = 3,
      booktabs = T, format = 'pandoc')
```

Table 1: Summary Statistics for Covid Data

	min	mean	median	variance	sd	max	skew
Total Cases	355	181486.137	24367	4.767454e+11	690467.501	6173236	6.689
Total Deaths	1	6091.115	411	4.393447e+08	20960.550	187224	6.207
TOTCases_1M	11	4177.388	1789	3.814673e+07	6176.304	42230	3.000
TOTDeath_!M	0	115.187	34	3.215569e+04	179.320	871	2.181
TotalTested	120	3141261.633	404944	1.280726e+14	11316914.780	90410000	6.192

Question 1.5

```
library(kableExtra)
spear <- cor(x = covid[-1], method = "spearman")
kable(spear, caption = "Spearman Correlation", linesep = "\\addlinespace", digits = 3,
      booktabs = T, format = 'pandoc')
```

Table 2: Spearman Correlation

	Total Cases	Total Deaths	TOTCases_1M	TOTDeath_!M	TotalTested
Total Cases	1.000	0.919	0.736	0.720	0.736
Total Deaths	0.919	1.000	0.643	0.795	0.669
TOTCases_1M	0.736	0.643	1.000	0.889	0.457
TOTDeath_!M	0.720	0.795	0.889	1.000	0.449
TotalTested	0.736	0.669	0.457	0.449	1.000

```
pear <- cor(x = covid[-1], method = "pearson")
kable(pear, caption = "Pearson Correlation", linesep = "\\addlinespace", digits = 3,
      booktabs = T, format = 'pandoc')
```

Table 3: Pearson Correlation

	Total Cases	Total Deaths	TOTCases_1M	TOTDeath_!M	TotalTested
Total Cases	1.000	0.940	0.307	0.362	0.659
Total Deaths	0.940	1.000	0.310	0.526	0.620
TOTCases_1M	0.307	0.310	1.000	0.524	0.130
TOTDeath_!M	0.362	0.526	0.524	1.000	0.190
TotalTested	0.659	0.620	0.130	0.190	1.000

Question 1.6

```
library(kableExtra)
Nassif <- c("11/40", "0.275", "18/25", "0.72", "29/65", "0.446")
Yan <- c("16/55", "0.291", "17/23", "0.739", "33/78", "0.423")

sp <- data.frame()
sp <- rbind(Nassif, Yan)

kable(sp, caption = "Percentage of students attending class on Zoom",
      booktabs = T) %>%
  add_header_above(c(" ", "2020" = 2, "2021" = 2, "Combined" = 2)) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

Table 4: Percentage of students attending class on Zoom

	2020		2021		Combined	
Nassif	11/40	0.275	18/25	0.72	29/65	0.446
Yan	16/55	0.291	17/23	0.739	33/78	0.423

An example is constructed where the percentage of students attending class via Zoom are observed as an estimate across years 2020 and 2021.

One can see that in both years, Yan's class has a higher percentage of students attending class via Zoom. However, the highest combined percentage of students attending class on Zoom is revealed to be the classes of Nassiff.

This exemplifies the Simpson's paradox since there is a similar trend in years 2020 and 2021. However, when the groups of data are combined the trend is different.

Textbook Questions

Question 2:

a) Regression, inference, n=500, p=3

b) Classification, inference, n=20, p=13

c) Regression, predictive, n=52 weeks, p=2

Question 4:

a) Classification:

i) Use classification to predict whether an NBA team will make the playoffs. Take in point differential for each team as the predictor and whether the team made the playoffs as the response (0 or 1, no playoffs or playoffs).

ii) Use classification to predict whether someone is at risk of heart disease. Take in prediction parameters such as familial history, cholesterol, weight, etc and use whether the person will have heart disease as the response (0 or 1, no heart disease or heart disease).

iii) Use classification to predict the hair color of a child. Take in the hair color of parents, grandparents, etc as predictors and take the hair color of the child as the response (0, 1, 2, 3, 4 or black, brown, blonde, red, white).

b) Regression:

i) Use regression to predict the value of a house in the future. Take in current price of the home, volatility of the housing market, trends in interest rates and values of neighboring homes in the area.

ii) Using regression to take parents and family genetics to best predict how tall the child will be. Take in the family height history of paternal and maternal sides of an individual.

iii) Using regression to predict the value of a stock. Pay attention to the opening and closing costs throughout the week to best predict what they will be next week or even next month.

c) Clustering Analysis:

i) Using clustering analysis to go on Twitter and clump similar hashtags together to figure out who belongs to what political party.

ii) Using a clustering algorithm to identify spam within emails.

iii) Use clustering to classify network traffic to a website.

Question 7:

a) Euclidean Distances from (0, 0, 0):

Observation 1: 3

Observation 2: 2

Observation 3: sqrt(10)

Observation 4: sqrt(5)

Observation 5: sqrt(2)

Observation 6: sqrt(3)

b) The shortest ED is sqrt(2) which belongs to observation 5, classified as "GREEN". Therefore the observation (0, 0, 0) is classified as "GREEN" when K=1.

c) The 3 shortest ED's are sqrt(2) (observation 5, "GREEN"), sqrt(3) (observation 6, "RED"), and 2 (observation 2, "RED"). Therefore by a vote of 2 to 1, (0, 0, 0) will be classified as "RED" when K=3.

d) If the Baye's decision boundary is highly non-linear, then it can be expected that the ideal value for K would be small. This is because a low value for K will have significantly more flexibility than a larger value for K. As K grows, the boundary would become more linear and because we have a highly non-linear boundary we should not expect K to be a large number.