Module 4 Manuscript: Monica’s Data

Carter Coleman

2021-09-16

The structure below is a possible setup for a data analysis project (including the course project). For a manuscript, adjust as needed.

# 1 Summary/Abstract

*Write a summary of your project.*

# 2 Introduction

## 2.1 General Background Information

*Provide enough background on your topic that others can understand the why and how of your analysis*

## 2.2 Description of data and data source

*Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.*

## 2.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

# 3 Methods and Results

*In most research papers, results and methods are separate. You can combine them here if you find it easier. You are also welcome to structure things such that those are separate sections.*

## 3.1 Data aquisition

Data was obtained from the CDC’s open source data archive on COVID-19 deaths during the COVID-19 pandemic.

## 3.2 Data import and cleaning

#load needed packages. make sure they are installed.

library(ggplot2) #for plotting  
library(broom) #for cleaning up output from lm()  
library(here) #for data loading/saving

## here() starts at C:/Data/Github/MADA/MONICACHAN-MADA-analysis2

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble 3.1.3 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1  
## v purrr 0.3.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggpubr)  
library(AICcmodavg)

#path to data #note the use of the here() package and not absolute paths

data\_location <- here::here("data","processed\_data","processeddata.rds")

#load data.

NYC\_Virus\_Deaths <- readRDS(data\_location)

#re-label missing data (NA) as 0. It is fair to assume no reported deaths as 0, as it is assumed death data is coming from hospitals, who must report all deaths.

NYC\_Virus\_Deaths %>%  
 replace\_na(list(COVID.19.Deaths = 0, Influenza.Deaths = 0, Pneumonia.Deaths = 0))

## # A tibble: 63 x 8  
## Year Month State Sex Age.Group COVID.19.Deaths Pneumonia.Deaths  
## <dbl> <dbl> <chr> <chr> <chr> <dbl> <dbl>  
## 1 2020 1 New York City All Sexes All Ages 0 428  
## 2 2020 1 New York City Male All Ages 0 223  
## 3 2020 1 New York City Female All Ages 0 205  
## 4 2020 2 New York City All Sexes All Ages 0 369  
## 5 2020 2 New York City Male All Ages 0 209  
## 6 2020 2 New York City Female All Ages 0 160  
## 7 2020 3 New York City All Sexes All Ages 2182 1610  
## 8 2020 3 New York City Male All Ages 1362 936  
## 9 2020 3 New York City Female All Ages 820 674  
## 10 2020 4 New York City All Sexes All Ages 14928 6553  
## # ... with 53 more rows, and 1 more variable: Influenza.Deaths <dbl>

## 3.3 Exploratory analysis

#I’m using basic R commands here. #Lots of good packages exist to do more. #For instance check out the tableone or skimr packages

#summarize data

data\_summary = summary(NYC\_Virus\_Deaths)

#look at summary

print(data\_summary)

## Year Month State Sex   
## Min. :2020 Min. : 1.000 Length:63 Length:63   
## 1st Qu.:2020 1st Qu.: 3.000 Class :character Class :character   
## Median :2020 Median : 6.000 Mode :character Mode :character   
## Mean :2020 Mean : 5.857   
## 3rd Qu.:2021 3rd Qu.: 8.000   
## Max. :2021 Max. :12.000   
##   
## Age.Group COVID.19.Deaths Pneumonia.Deaths Influenza.Deaths  
## Length:63 Min. : 0.0 Min. : 16.0 Min. : 0.00   
## Class :character 1st Qu.: 66.5 1st Qu.: 182.0 1st Qu.: 0.00   
## Mode :character Median : 226.0 Median : 329.0 Median : 0.00   
## Mean : 946.0 Mean : 589.9 Mean : 69.45   
## 3rd Qu.: 864.5 3rd Qu.: 546.0 3rd Qu.: 53.00   
## Max. :14928.0 Max. :6553.0 Max. :747.00   
## NA's :2 NA's :34

#do the same, but with a bit of trickery to get things into the #shape of a data frame (for easier saving/showing in manuscript)

summary\_df = data.frame(do.call(cbind, lapply(NYC\_Virus\_Deaths, summary)))

## Warning in (function (..., deparse.level = 1) : number of rows of result is not  
## a multiple of vector length (arg 1)

#save data frame table to file for later use in manuscript

summarytable\_file = here("results", "summarytable.rds")  
saveRDS(summary\_df, file = summarytable\_file)

#Once the data is loaded, we will want to alter the month variable a bit to make it “month during pandemic.” This will create a sequential variable data set, irrespective of year. This will help up later on.

NYC\_Virus\_Deaths$Month[NYC\_Virus\_Deaths$Year == "2021"& NYC\_Virus\_Deaths$Month == 1] <- 13  
NYC\_Virus\_Deaths$Month[NYC\_Virus\_Deaths$Year == "2021"& NYC\_Virus\_Deaths$Month == 2] <- 14  
NYC\_Virus\_Deaths$Month[NYC\_Virus\_Deaths$Year == "2021"& NYC\_Virus\_Deaths$Month == 3] <- 15  
NYC\_Virus\_Deaths$Month[NYC\_Virus\_Deaths$Year == "2021"& NYC\_Virus\_Deaths$Month == 4] <- 16  
NYC\_Virus\_Deaths$Month[NYC\_Virus\_Deaths$Year == "2021"& NYC\_Virus\_Deaths$Month == 5] <- 17  
NYC\_Virus\_Deaths$Month[NYC\_Virus\_Deaths$Year == "2021"& NYC\_Virus\_Deaths$Month == 6] <- 18  
NYC\_Virus\_Deaths$Month[NYC\_Virus\_Deaths$Year == "2021"& NYC\_Virus\_Deaths$Month == 7] <- 19  
NYC\_Virus\_Deaths$Month[NYC\_Virus\_Deaths$Year == "2021"& NYC\_Virus\_Deaths$Month == 8] <- 20  
NYC\_Virus\_Deaths$Month[NYC\_Virus\_Deaths$Year == "2021"& NYC\_Virus\_Deaths$Month == 9] <- 21

#make a scatterplot of data (Covid-19 Deaths, Months) #Make a separate data set for each of the sex-specific variables.

NVD\_All <- NYC\_Virus\_Deaths %>% dplyr::filter(Sex == "All Sexes")  
NVD\_Female <- NYC\_Virus\_Deaths %>% dplyr::filter(Sex == "Female")  
NVD\_Male <- NYC\_Virus\_Deaths %>% dplyr::filter(Sex == "Male")

3.3.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.1

#Plotting Virus Deaths as a function of #Time, separated by gender and by total population (ALL) ######################################

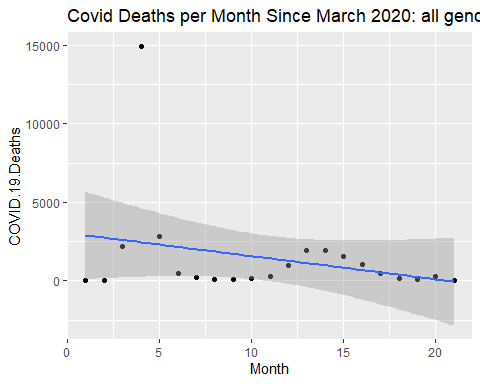
#Plot newly formed data sets for covid deaths with line of best fit for each gender and combination.

Covid\_Month\_All <- NVD\_All %>%   
 ggplot(aes(x=Month, y=COVID.19.Deaths)) +   
 geom\_point() +   
 geom\_smooth(method='lm') +   
 ggtitle("Covid Deaths per Month Since March 2020: all genders")  
Covid\_Month\_Female <- NVD\_Female %>% ggplot(aes(x=Month, y=COVID.19.Deaths)) + geom\_point() + geom\_smooth(method='lm')+ ggtitle("Covid Deaths per Month Since March 2020: Females")  
Covid\_Month\_Male <- NVD\_Male %>% ggplot(aes(x=Month, y=COVID.19.Deaths)) + geom\_point() + geom\_smooth(method='lm')+ ggtitle("Covid Deaths per Month Since March 2020: Males")

#look at each figure

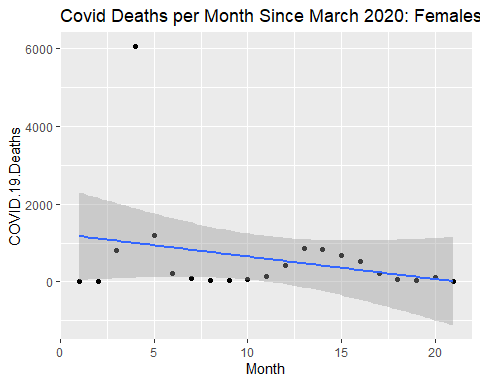
plot(Covid\_Month\_All)

## `geom\_smooth()` using formula 'y ~ x'



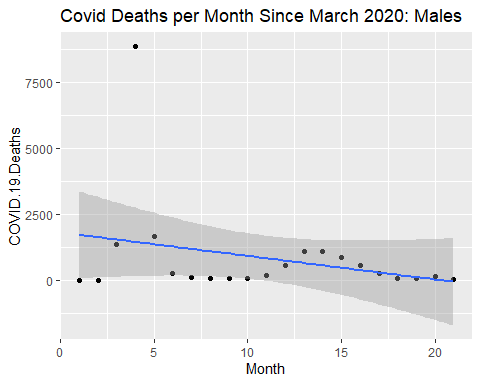
plot(Covid\_Month\_Female)

## `geom\_smooth()` using formula 'y ~ x'



plot(Covid\_Month\_Male)

## `geom\_smooth()` using formula 'y ~ x'

 ##save figures

#Total population  
figure\_file = here("results","COVID\_ALL\_figure.png")  
ggsave(filename = figure\_file, plot=Covid\_Month\_All)

## Saving 5 x 4 in image

## `geom\_smooth()` using formula 'y ~ x'

#Females  
figure\_file = here("results","COVID\_FEMALE\_figure.png")  
ggsave(filename = figure\_file, plot=Covid\_Month\_Female)

## Saving 5 x 4 in image  
## `geom\_smooth()` using formula 'y ~ x'

#Males  
figure\_file = here("results","COVID\_MALE\_figure.png")  
ggsave(filename = figure\_file, plot=Covid\_Month\_Male)

## Saving 5 x 4 in image  
## `geom\_smooth()` using formula 'y ~ x'

##Repeat Previous steps for Influenza

#Plot newly formed data sets for Influenza deaths with line of best fit for each gender and combination.

Influenza\_Month\_All <- NVD\_All %>% ggplot(aes(x=Month, y=Influenza.Deaths)) + geom\_point() + geom\_smooth(method='lm') + ggtitle("Flu Deaths per Month Since March 2020: all genders")  
Influenza\_Month\_Female <- NVD\_Female %>% ggplot(aes(x=Month, y=Influenza.Deaths)) + geom\_point() + geom\_smooth(method='lm')+ ggtitle("Flu Deaths per Month Since March 2020: Females")  
Influenza\_Month\_Male <- NVD\_Male %>% ggplot(aes(x=Month, y=Influenza.Deaths)) + geom\_point() + geom\_smooth(method='lm')+ ggtitle("Flu Deaths per Month Since March 2020: Males")

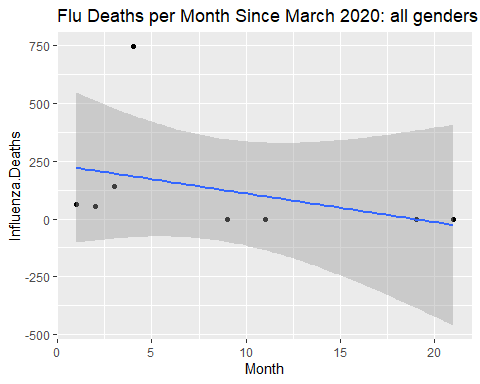
#look at each figure

plot(Influenza\_Month\_All)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 13 rows containing non-finite values (stat\_smooth).

## Warning: Removed 13 rows containing missing values (geom\_point).

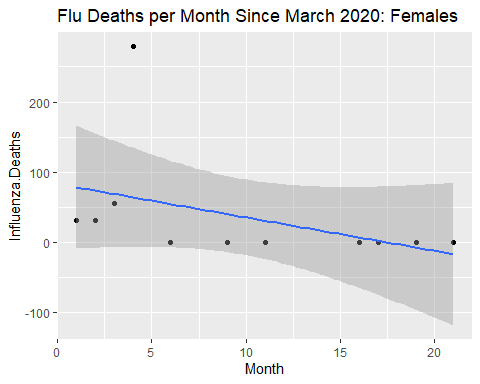


plot(Influenza\_Month\_Female)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 10 rows containing non-finite values (stat\_smooth).

## Warning: Removed 10 rows containing missing values (geom\_point).

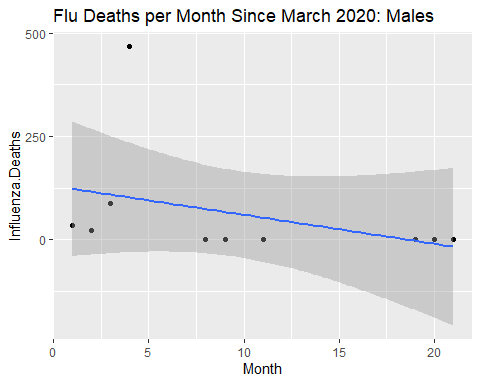


plot(Influenza\_Month\_Male)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 11 rows containing non-finite values (stat\_smooth).

## Warning: Removed 11 rows containing missing values (geom\_point).



##save figures

#Total population  
  
figure\_file = here("results","Flu\_ALL\_figure.png")  
ggsave(filename = figure\_file, plot=Influenza\_Month\_All)

## Saving 5 x 4 in image

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 13 rows containing non-finite values (stat\_smooth).

## Warning: Removed 13 rows containing missing values (geom\_point).

#Females  
figure\_file = here("results","Flu\_FEMALE\_figure.png")  
ggsave(filename = figure\_file, plot=Influenza\_Month\_Female)

## Saving 5 x 4 in image  
## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 10 rows containing non-finite values (stat\_smooth).

## Warning: Removed 10 rows containing missing values (geom\_point).

#Males  
figure\_file = here("results","Flu\_MALE\_figure.png")  
ggsave(filename = figure\_file, plot=Influenza\_Month\_Male)

## Saving 5 x 4 in image  
## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 11 rows containing non-finite values (stat\_smooth).

## Warning: Removed 11 rows containing missing values (geom\_point).

##Repeat Previous steps for Pneumonia

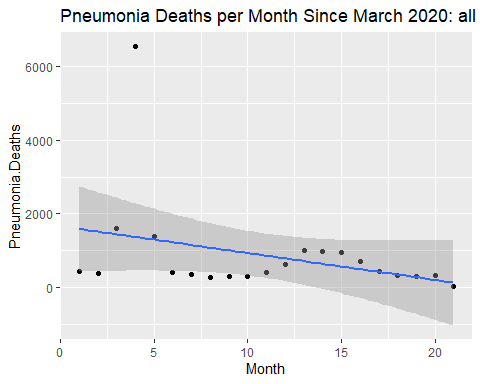
#Plot newly formed data sets for Influenza deaths with line of best fit for each gender and combination.

Pneumonia\_Month\_All <- NVD\_All %>% ggplot(aes(x=Month, y=Pneumonia.Deaths)) + geom\_point() + geom\_smooth(method='lm') + ggtitle("Pneumonia Deaths per Month Since March 2020: all genders")  
Pneumonia\_Month\_Female <- NVD\_Female %>% ggplot(aes(x=Month, y=Pneumonia.Deaths)) + geom\_point() + geom\_smooth(method='lm')+ ggtitle("Pneumonia Deaths per Month Since March 2020: Females")  
Pneumonia\_Month\_Male <- NVD\_Male %>% ggplot(aes(x=Month, y=Influenza.Deaths)) + geom\_point() + geom\_smooth(method='lm')+ ggtitle("Pneumonia Deaths per Month Since March 2020: Males")

#look at each figure

plot(Pneumonia\_Month\_All)

## `geom\_smooth()` using formula 'y ~ x'

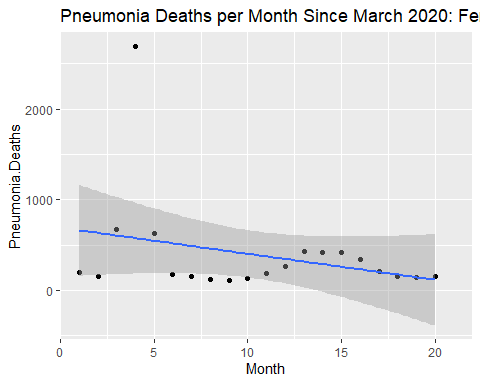


plot(Pneumonia\_Month\_Female)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat\_smooth).

## Warning: Removed 1 rows containing missing values (geom\_point).

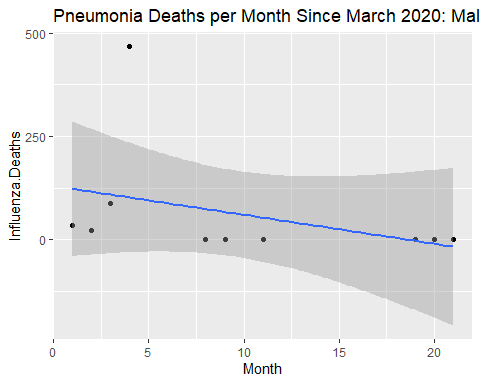


plot(Pneumonia\_Month\_Male)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 11 rows containing non-finite values (stat\_smooth).

## Warning: Removed 11 rows containing missing values (geom\_point).



##save figures

#Total population  
figure\_file = here("results","Pneumonia\_ALL\_figure.png")  
ggsave(filename = figure\_file, plot=Pneumonia\_Month\_All)

## Saving 5 x 4 in image

## `geom\_smooth()` using formula 'y ~ x'

#Females  
figure\_file = here("results","Pneumonia\_FEMALE\_figure.png")  
ggsave(filename = figure\_file, plot=Pneumonia\_Month\_Female)

## Saving 5 x 4 in image  
## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat\_smooth).

## Warning: Removed 1 rows containing missing values (geom\_point).

#Males  
figure\_file = here("results","Pneumonia\_MALE\_figure.png")  
ggsave(filename = figure\_file, plot=Pneumonia\_Month\_Male)

## Saving 5 x 4 in image  
## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 11 rows containing non-finite values (stat\_smooth).

## Warning: Removed 11 rows containing missing values (geom\_point).

##Finally, we will want to visualize how the three different virus deaths compare over time.

NVD\_ALL\_2 <- pivot\_longer(NVD\_All, cols=6:8, names\_to = "Virus\_Type", values\_to = "Deaths")  
  
Virus\_Death\_Plot <- NVD\_ALL\_2 %>%   
 ggplot(aes(x=Month, y=Deaths, color = Virus\_Type)) +   
 geom\_point() +   
 geom\_smooth(method = 'lm') +   
 ggtitle("Comparison of Virus Deaths Over Time")

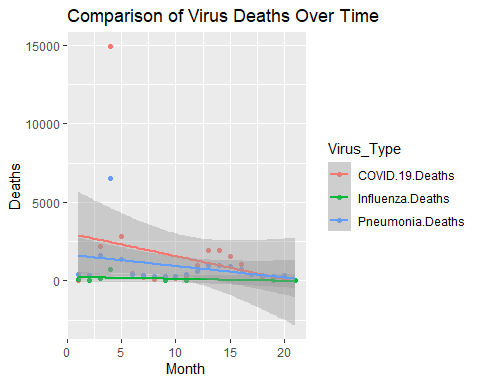
#View combined virus death plot

plot(Virus\_Death\_Plot)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 13 rows containing non-finite values (stat\_smooth).

## Warning: Removed 13 rows containing missing values (geom\_point).



#Save the figue

figure\_file = here("results","Virus\_Deaths\_figure.png")  
ggsave(filename = figure\_file, plot=Virus\_Death\_Plot)

## Saving 5 x 4 in image

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 13 rows containing non-finite values (stat\_smooth).

## Warning: Removed 13 rows containing missing values (geom\_point).

## 3.4 Full analysis

#Run an ANOVA test for the three different virus deaths over the entire provided time period.

anova\_one\_way <- aov(Deaths~Virus\_Type, data = NVD\_ALL\_2)  
summary(anova\_one\_way)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Virus\_Type 2 10200482 5100241 0.983 0.382  
## Residuals 47 243861744 5188548   
## 13 observations deleted due to missingness

# 4 place results from fit into a data frame with the tidy function

aov\_table <- broom::tidy(anova\_one\_way)

save fit results table

table\_file\_Covid = here("results", "Covid\_resulttable.rds")  
saveRDS(aov\_table, file = table\_file\_Covid)

#look at fit results

print(aov\_table)

## # A tibble: 2 x 6  
## term df sumsq meansq statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Virus\_Type 2 10200482. 5100241. 0.983 0.382  
## 2 Residuals 47 243861744. 5188548. NA NA

# 5 Discussion and Conclusions

Based on the reported p-value of 0.382, there is no significant difference between deaths resulting from COVID, Influenza, or Pneumonia.

Based on previous knowledge about the high virulence of COVID-19 compared to Influenza and Pneumonia and the results of the produced figures, one possible explanation for this finding is reporting bias. COVID was the main focus of reported deaths during the pandemic. If you compare the created figures, n for Influenza and Pneumonia deaths is quite low. Low n suggests that the ANOVA performed was biased due to low sample size.

This paper (Leek & Peng, 2015) discusses types of analyses.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*, 1314–1315. <https://doi.org/10.1126/science.aaa6146>