Corona_Case_Prediction

Steven Smith, PhD

3/18/2020

Contents

The 2019-2020 Coronavirus Pandemic Analysis

BACKGROUND & APPROACH

I wanted to track and trend the coronavirus outbreak on my own curiosity. There are some interesting questions that may fall out of this, as it is a very historic moment, including scientifically and analytically (we have a large amount of data being shared across the globe, analyzed in real-time). The world has come to a halt because of it.
This analysis attempts to answer the following questions (more to come): 1. What does the trend of the pandemic look like to date?
2. What are future case predictions based on historical model? 3. What interesting quirks or patterns emerge?

ASSUMPTIONS & LIMITATIONS: * This data is limited by the source. I realized early on that depending on source there were conflicting # of cases. Originally I was using JHU data… but this was always 'ahead' of the Our World In Data. I noticed that JHU's website was buggy- you clicked on the U.S. stats but it didn't reflect the U.S.. So I changed data sources to be more consistent with what is presented in the media (and Our World In Data has more extensive plots I can compare my own to). An interesting aside might be why the discrepancy? Was I missing something?
* Defintiions are important as is the idea that multiple varibales accumulate in things like total cases (more testing for example).

SOURCE RAW DATA: https://ourworldindata.org/coronavirus INPUT DATA LOCATION: github (https://github.com/sbs87coronavirus/data) OUTPUT DATA LOCATIOn: github (https://github.com/sbs87coronavirus/results)

PRE-ANALYSIS

The following sections are outside the scope of the 'analysis' but are still needed to prepare everything

UPSTREAM PROCESSING/ANALYSIS (N/A)

Not applicable - No analysis performed on remote server

```
# No analysis performed on remote server
```

SET UP ENVIORNMENT

Load libraries and set global variables

```
# clear previous enviornment
rm(list = ls())

##------------------------------------------
## LIBRARIES
##------------------------------------------
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --

## v tibble  2.0.1       v purrr   0.3.3
## v tidyr   0.8.3       v dplyr   0.8.0.1
## v readr   1.3.1       v stringr 1.4.0
## v tibble  2.0.1       v forcats 0.4.0

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

```
library(plyr)
```

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(plot.utils)
library(utils)

##------------------------------------------

##------------------------------------------
# GLOBAL VARIABLES
##------------------------------------------
working_dir<-"/Users/ssmith/coronavirus/" # don't forget trailing /
results_dir<-paste0(working_dir,"results/") # assumes diretory exists
Corona_Cases.fn<-paste0(working_dir,"data/","time_series_covid19_confirmed_global.csv")
Corona_Cases.US.fn<-paste0(working_dir,"data/time_series_covid19_confirmed_US.csv")
Corona_Cases.source_url<-"https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_
Corona_Cases.US.source_url<-"https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/cs
default_theme<-theme_bw()+theme(text = element_text(size = 14)) # fix this
##------------------------------------------
```

FUNCTIONS

List of functions 1. Function 1
2. Function 2

```
##------------------------------------------
## FUNCTION: prediction_model
##------------------------------------------
## --- //// ----
# Takes days vs log10 (case) linear model parameters and a set of days since 100 cases and outputs a da
## --- //// ----
prediction_model<-function(m=1,b=0,days=1){
  total_cases.log<-m*days+b
  total_cases<-10^total_cases.log
  prediction<-data.frame(Days_since_100=days,Total_confirmed_cases=total_cases,Total_confirmed_cases.log
  return(prediction)
}
##------------------------------------------

##------------------------------------------
## FUNCTION: make_long
##------------------------------------------
## --- //// ----
# Takes wide-format case data and converts into long format, using date and total cases as variable/val
## --- //// ----
make_long<-function(data_in,variable.name = "Date",
                   value.name = "Total_confirmed_cases",
                   id.vars=c("Province.State","Country.Region","Lat","Long","City")){

long_data<-melt(data_in,
                id.vars = id.vars,
                variable.name=variable.name,
                value.name=value.name)
return(long_data)

}
```

```
##------------------------------------------
```

READ IN DATA

- total number of cases. current source: https://github.com/CSSEGISandData (precvious source https://ourworldindata.org/coronavirus)

```
# Q: do we want to archive previous versions? Maybe an auto git mv?

##------------------------------------------
## Download and read in latest data from github
##------------------------------------------
download.file(Corona_Cases.source_url,destfile = Corona_Cases.fn)
Corona_Cases.raw<-read.csv(Corona_Cases.fn,header = T,stringsAsFactors = F)

download.file(Corona_Cases.US.source_url,destfile = Corona_Cases.US.fn)
Corona_Cases.US.raw<-read.csv(Corona_Cases.US.fn,header = T,stringsAsFactors = F)

head(Corona_Cases.US.raw)
```

```
##        UID iso2 iso3 code3 FIPS  Admin2           Province_State Country_Region
## 1       16   AS  ASM    16   60                   American Samoa             US
## 2      316   GU  GUM   316   66                             Guam             US
## 3      580   MP  MNP   580   69         Northern Mariana Islands             US
## 4      630   PR  PRI   630   72                      Puerto Rico             US
## 5      850   VI  VIR   850   78                   Virgin Islands             US
## 6 84001001   US  USA   840 1001 Autauga                  Alabama             US
##          Lat      Long_                 Combined_Key X1.22.20 X1.23.20 X1.24.20
## 1 -14.27100 -170.13200           American Samoa, US        0        0        0
## 2  13.44430  144.79370                     Guam, US        0        0        0
## 3  15.09790  145.67390 Northern Mariana Islands, US        0        0        0
## 4  18.22080  -66.59010              Puerto Rico, US        0        0        0
## 5  18.33580  -64.89630           Virgin Islands, US        0        0        0
## 6  32.53953  -86.64408         Autauga, Alabama, US        0        0        0
##   X1.25.20 X1.26.20 X1.27.20 X1.28.20 X1.29.20 X1.30.20 X1.31.20 X2.1.20
## 1        0        0        0        0        0        0        0       0
## 2        0        0        0        0        0        0        0       0
## 3        0        0        0        0        0        0        0       0
## 4        0        0        0        0        0        0        0       0
## 5        0        0        0        0        0        0        0       0
## 6        0        0        0        0        0        0        0       0
##   X2.2.20 X2.3.20 X2.4.20 X2.5.20 X2.6.20 X2.7.20 X2.8.20 X2.9.20 X2.10.20
## 1       0       0       0       0       0       0       0       0        0
## 2       0       0       0       0       0       0       0       0        0
## 3       0       0       0       0       0       0       0       0        0
## 4       0       0       0       0       0       0       0       0        0
## 5       0       0       0       0       0       0       0       0        0
## 6       0       0       0       0       0       0       0       0        0
##   X2.11.20 X2.12.20 X2.13.20 X2.14.20 X2.15.20 X2.16.20 X2.17.20 X2.18.20
## 1        0        0        0        0        0        0        0        0
## 2        0        0        0        0        0        0        0        0
## 3        0        0        0        0        0        0        0        0
## 4        0        0        0        0        0        0        0        0
## 5        0        0        0        0        0        0        0        0
## 6        0        0        0        0        0        0        0        0
##   X2.19.20 X2.20.20 X2.21.20 X2.22.20 X2.23.20 X2.24.20 X2.25.20 X2.26.20
## 1        0        0        0        0        0        0        0        0
## 2        0        0        0        0        0        0        0        0
## 3        0        0        0        0        0        0        0        0
## 4        0        0        0        0        0        0        0        0
## 5        0        0        0        0        0        0        0        0
## 6        0        0        0        0        0        0        0        0
##   X2.27.20 X2.28.20 X2.29.20 X3.1.20 X3.2.20 X3.3.20 X3.4.20 X3.5.20 X3.6.20
## 1        0        0        0       0       0       0       0       0       0
## 2        0        0        0       0       0       0       0       0       0
## 3        0        0        0       0       0       0       0       0       0
## 4        0        0        0       0       0       0       0       0       0
## 5        0        0        0       0       0       0       0       0       0
## 6        0        0        0       0       0       0       0       0       0
##   X3.7.20 X3.8.20 X3.9.20 X3.10.20 X3.11.20 X3.12.20 X3.13.20 X3.14.20 X3.15.20
## 1       0       0       0        0        0        0        0        0        0
## 2       0       0       0        0        0        0        0        0        0
## 3       0       0       0        0        0        0        0        0        0
## 4       0       0       0        0        0        0        0        0        0
## 5       0       0       0        0        0        0        0        0        0
## 6       0       0       0        0        0        0        0        0        0
##   X3.16.20 X3.17.20 X3.18.20 X3.19.20 X3.20.20 X3.21.20 X3.22.20 X3.23.20
## 1        0        0        0        0        0        0        0        0
## 2        3        3        5       12       14       15       27       29
## 3        0        0        0        0        0        0        0        0
## 4        5        5        5        5       14       21       23       31
## 5        1        2        2        3        3        6        6        7
## 6        0        0        0        0        0        0        0        0
##   X3.24.20 X3.25.20 X3.26.20 X3.27.20 X3.28.20 X3.29.20 X3.30.20 X3.31.20
## 1        0        0        0        0        0        0        0        0
## 2       32       37       45       51       55       56       58       69
## 3        0        0        0        0        0        0        0        2
## 4       39       51       64       79      100      127      174      239
## 5       17       17       17       19       22        0        0       30
## 6        1        4        6        6        6        6        6        7
##   X4.1.20
## 1       0
## 2      77
## 3       6
## 4     286
## 5      30
## 6       8
```

```
head(Corona_Cases.raw)
```

```
##   Province.State      Country.Region      Lat     Long X1.22.20 X1.23.20
## 1                        Afghanistan  33.0000  65.0000        0        0
## 2                            Albania  41.1533  20.1683        0        0
## 3                            Algeria  28.0339   1.6596        0        0
## 4                            Andorra  42.5063   1.5218        0        0
## 5                             Angola -11.2027  17.8739        0        0
## 6                Antigua and Barbuda  17.0608 -61.7964        0        0
##   X1.24.20 X1.25.20 X1.26.20 X1.27.20 X1.28.20 X1.29.20 X1.30.20 X1.31.20
## 1        0        0        0        0        0        0        0        0
## 2        0        0        0        0        0        0        0        0
## 3        0        0        0        0        0        0        0        0
## 4        0        0        0        0        0        0        0        0
## 5        0        0        0        0        0        0        0        0
## 6        0        0        0        0        0        0        0        0
##   X2.1.20 X2.2.20 X2.3.20 X2.4.20 X2.5.20 X2.6.20 X2.7.20 X2.8.20 X2.9.20
## 1       0       0       0       0       0       0       0       0       0
## 2       0       0       0       0       0       0       0       0       0
## 3       0       0       0       0       0       0       0       0       0
## 4       0       0       0       0       0       0       0       0       0
## 5       0       0       0       0       0       0       0       0       0
## 6       0       0       0       0       0       0       0       0       0
##   X2.10.20 X2.11.20 X2.12.20 X2.13.20 X2.14.20 X2.15.20 X2.16.20 X2.17.20
## 1        0        0        0        0        0        0        0        0
## 2        0        0        0        0        0        0        0        0
## 3        0        0        0        0        0        0        0        0
## 4        0        0        0        0        0        0        0        0
## 5        0        0        0        0        0        0        0        0
## 6        0        0        0        0        0        0        0        0
##   X2.18.20 X2.19.20 X2.20.20 X2.21.20 X2.22.20 X2.23.20 X2.24.20 X2.25.20
## 1        0        0        0        0        0        0        1        1
## 2        0        0        0        0        0        0        0        0
## 3        0        0        0        0        0        0        0        1
## 4        0        0        0        0        0        0        0        0
## 5        0        0        0        0        0        0        0        0
## 6        0        0        0        0        0        0        0        0
##   X2.26.20 X2.27.20 X2.28.20 X2.29.20 X3.1.20 X3.2.20 X3.3.20 X3.4.20 X3.5.20
## 1        1        1        1        1       1       1       1       1       1
## 2        0        0        0        0       0       0       0       0       0
## 3        1        1        1        1       1       3       5      12      12
## 4        0        0        0        0       0       1       1       1       1
## 5        0        0        0        0       0       0       0       0       0
## 6        0        0        0        0       0       0       0       0       0
##   X3.6.20 X3.7.20 X3.8.20 X3.9.20 X3.10.20 X3.11.20 X3.12.20 X3.13.20 X3.14.20
## 1       1       1       4       4        5        7        7        7       11
## 2       0       0       0       2       10       12       23       33       38
## 3      17      17      19      20       20       20       24       26       37
## 4       1       1       1       1        1        1        1        1        1
## 5       0       0       0       0        0        0        0        0        0
## 6       0       0       0       0        0        0        0        1        1
##   X3.15.20 X3.16.20 X3.17.20 X3.18.20 X3.19.20 X3.20.20 X3.21.20 X3.22.20
## 1       16       21       22       22       22       24       24       40
## 2       42       51       55       59       64       70       76       89
## 3       48       54       60       74       87       90      139      201
## 4        1        2       39       39       53       75       88      113
## 5        0        0        0        0        0        1        2        2
## 6        1        1        1        1        1        1        1        1
##   X3.23.20 X3.24.20 X3.25.20 X3.26.20 X3.27.20 X3.28.20 X3.29.20 X3.30.20
## 1       40       74       84       94      110      110      120      170
## 2      104      123      146      174      186      197      212      223
## 3      230      264      302      367      409      454      511      584
## 4      133      164      188      224      267      308      334      370
## 5        3        3        3        4        4        5        7        7
## 6        3        3        3        7        7        7        7        7
##   X3.31.20 X4.1.20
## 1      174     237
## 2      243     259
## 3      716     847
## 4      376     390
## 5        7       8
## 6        7       7
```

PROCESS DATA

- Convert to long format
- Fix date formatting/convert to numeric date
- Log10 transform total # cases

```
##--------------------------------------------
## Convert to long format
##--------------------------------------------
#JHU has a gross file format. It's in wide format with each column is the date in MM/DD/YY. So read thi
# Furthermore, the World and US level data is formatted differently, containing different columns, etc.

# prepare raw datasets for eventual combining
Corona_Cases.raw$City<-"NA" # US-level data has Cities
Corona_Cases.US.raw$Country_Region<-"US_state" # To differentiate from World-level stats
filter(Corona_Cases.raw,Country.Region=="US")
```

```
##   Province.State Country.Region     Lat     Long X1.22.20 X1.23.20 X1.24.20
## 1                            US 37.0902 -95.7129        1        1        2
##   X1.25.20 X1.26.20 X1.27.20 X1.28.20 X1.29.20 X1.30.20 X1.31.20 X2.1.20
## 1        2        5        5        5        5        5        7       8
##   X2.2.20 X2.3.20 X2.4.20 X2.5.20 X2.6.20 X2.7.20 X2.8.20 X2.9.20 X2.10.20
## 1       8      11      11      11      11      11      11      11       11
##   X2.11.20 X2.12.20 X2.13.20 X2.14.20 X2.15.20 X2.16.20 X2.17.20 X2.18.20
## 1       12       12       13       13       13       13       13       13
##   X2.19.20 X2.20.20 X2.21.20 X2.22.20 X2.23.20 X2.24.20 X2.25.20 X2.26.20
## 1       13       13       15       15       15       51       51       57
##   X2.27.20 X2.28.20 X2.29.20 X3.1.20 X3.2.20 X3.3.20 X3.4.20 X3.5.20 X3.6.20
## 1       58       60       68      74      98     118     149     217     262
##   X3.7.20 X3.8.20 X3.9.20 X3.10.20 X3.11.20 X3.12.20 X3.13.20 X3.14.20 X3.15.20
## 1     402     518     583      959     1281     1663     2179     2727     3499
##   X3.16.20 X3.17.20 X3.18.20 X3.19.20 X3.20.20 X3.21.20 X3.22.20 X3.23.20
## 1     4632     6421     7783    13677    19100    25489    33276    43847
##   X3.24.20 X3.25.20 X3.26.20 X3.27.20 X3.28.20 X3.29.20 X3.30.20 X3.31.20
## 1    53740    65778    83836   101657   121478   140886   161807   188172
##   X4.1.20 City
## 1  213372   NA
```

```
Corona_Cases.US.raw<-rename(Corona_Cases.US.raw,c("Province_State"="Province.State",
                                                  "Country_Region"="Country.Region",
                                                  "Long_"="Long",
                                                  "Admin2"="City"))

Corona_Cases<-rbind(make_long(select(Corona_Cases.US.raw,-c(UID,iso2,iso3,code3,FIPS,Combined_Key))),
make_long(Corona_Cases.raw))


##------------------------------------------
## Fix date formatting, convert to numeric date
##------------------------------------------
Corona_Cases$Date<-gsub(Corona_Cases$Date,pattern = "^X",replacement = "0") # leading 0 read in as X
Corona_Cases$Date<-gsub(Corona_Cases$Date,pattern = "20$",replacement = "2020") # ends in .20 and not 2
Corona_Cases$Date<-as.Date(Corona_Cases$Date,format = "%m.%d.%y")
Corona_Cases$Date.numeric<-as.numeric(Corona_Cases$Date)
filter(Corona_Cases,Country.Region=="US")
```

```
##    Province.State Country.Region     Lat     Long City       Date
## 1                             US 37.0902 -95.7129   NA 2020-01-22
## 2                             US 37.0902 -95.7129   NA 2020-01-23
## 3                             US 37.0902 -95.7129   NA 2020-01-24
## 4                             US 37.0902 -95.7129   NA 2020-01-25
## 5                             US 37.0902 -95.7129   NA 2020-01-26
## 6                             US 37.0902 -95.7129   NA 2020-01-27
## 7                             US 37.0902 -95.7129   NA 2020-01-28
## 8                             US 37.0902 -95.7129   NA 2020-01-29
## 9                             US 37.0902 -95.7129   NA 2020-01-30
## 10                            US 37.0902 -95.7129   NA 2020-01-31
## 11                            US 37.0902 -95.7129   NA 2020-02-01
## 12                            US 37.0902 -95.7129   NA 2020-02-02
## 13                            US 37.0902 -95.7129   NA 2020-02-03
## 14                            US 37.0902 -95.7129   NA 2020-02-04
## 15                            US 37.0902 -95.7129   NA 2020-02-05
## 16                            US 37.0902 -95.7129   NA 2020-02-06
## 17                            US 37.0902 -95.7129   NA 2020-02-07
## 18                            US 37.0902 -95.7129   NA 2020-02-08
## 19                            US 37.0902 -95.7129   NA 2020-02-09
## 20                            US 37.0902 -95.7129   NA 2020-02-10
## 21                            US 37.0902 -95.7129   NA 2020-02-11
## 22                            US 37.0902 -95.7129   NA 2020-02-12
## 23                            US 37.0902 -95.7129   NA 2020-02-13
## 24                            US 37.0902 -95.7129   NA 2020-02-14
## 25                            US 37.0902 -95.7129   NA 2020-02-15
## 26                            US 37.0902 -95.7129   NA 2020-02-16
## 27                            US 37.0902 -95.7129   NA 2020-02-17
## 28                            US 37.0902 -95.7129   NA 2020-02-18
## 29                            US 37.0902 -95.7129   NA 2020-02-19
## 30                            US 37.0902 -95.7129   NA 2020-02-20
## 31                            US 37.0902 -95.7129   NA 2020-02-21
## 32                            US 37.0902 -95.7129   NA 2020-02-22
## 33                            US 37.0902 -95.7129   NA 2020-02-23
## 34                            US 37.0902 -95.7129   NA 2020-02-24
## 35                            US 37.0902 -95.7129   NA 2020-02-25
## 36                            US 37.0902 -95.7129   NA 2020-02-26
## 37                            US 37.0902 -95.7129   NA 2020-02-27
## 38                            US 37.0902 -95.7129   NA 2020-02-28
## 39                            US 37.0902 -95.7129   NA 2020-02-29
## 40                            US 37.0902 -95.7129   NA 2020-03-01
## 41                            US 37.0902 -95.7129   NA 2020-03-02
## 42                            US 37.0902 -95.7129   NA 2020-03-03
## 43                            US 37.0902 -95.7129   NA 2020-03-04
## 44                            US 37.0902 -95.7129   NA 2020-03-05
## 45                            US 37.0902 -95.7129   NA 2020-03-06
## 46                            US 37.0902 -95.7129   NA 2020-03-07
## 47                            US 37.0902 -95.7129   NA 2020-03-08
## 48                            US 37.0902 -95.7129   NA 2020-03-09
## 49                            US 37.0902 -95.7129   NA 2020-03-10
## 50                            US 37.0902 -95.7129   NA 2020-03-11
## 51                            US 37.0902 -95.7129   NA 2020-03-12
## 52                            US 37.0902 -95.7129   NA 2020-03-13
## 53                            US 37.0902 -95.7129   NA 2020-03-14
## 54                            US 37.0902 -95.7129   NA 2020-03-15
## 55                            US 37.0902 -95.7129   NA 2020-03-16
## 56                            US 37.0902 -95.7129   NA 2020-03-17
## 57                            US 37.0902 -95.7129   NA 2020-03-18
## 58                            US 37.0902 -95.7129   NA 2020-03-19
## 59                            US 37.0902 -95.7129   NA 2020-03-20
## 60                            US 37.0902 -95.7129   NA 2020-03-21
## 61                            US 37.0902 -95.7129   NA 2020-03-22
## 62                            US 37.0902 -95.7129   NA 2020-03-23
## 63                            US 37.0902 -95.7129   NA 2020-03-24
## 64                            US 37.0902 -95.7129   NA 2020-03-25
## 65                            US 37.0902 -95.7129   NA 2020-03-26
## 66                            US 37.0902 -95.7129   NA 2020-03-27
## 67                            US 37.0902 -95.7129   NA 2020-03-28
## 68                            US 37.0902 -95.7129   NA 2020-03-29
## 69                            US 37.0902 -95.7129   NA 2020-03-30
## 70                            US 37.0902 -95.7129   NA 2020-03-31
## 71                            US 37.0902 -95.7129   NA 2020-04-01
##    Total_confirmed_cases Date.numeric
## 1                      1        18283
## 2                      1        18284
## 3                      2        18285
## 4                      2        18286
## 5                      5        18287
## 6                      5        18288
## 7                      5        18289
## 8                      5        18290
## 9                      5        18291
## 10                     7        18292
## 11                     8        18293
## 12                     8        18294
## 13                    11        18295
## 14                    11        18296
## 15                    11        18297
## 16                    11        18298
## 17                    11        18299
## 18                    11        18300
## 19                    11        18301
## 20                    11        18302
## 21                    12        18303
## 22                    12        18304
## 23                    13        18305
## 24                      13        18306
## 25                      13        18307
## 26                      13        18308
## 27                      13        18309
## 28                      13        18310
## 29                      13        18311
## 30                      13        18312
## 31                      15        18313
## 32                      15        18314
## 33                      15        18315
## 34                      51        18316
## 35                      51        18317
## 36                      57        18318
## 37                      58        18319
## 38                      60        18320
## 39                      68        18321
## 40                      74        18322
## 41                      98        18323
## 42                     118        18324
## 43                     149        18325
## 44                     217        18326
## 45                     262        18327
## 46                     402        18328
## 47                     518        18329
## 48                     583        18330
## 49                     959        18331
## 50                    1281        18332
## 51                    1663        18333
## 52                    2179        18334
## 53                    2727        18335
## 54                    3499        18336
## 55                    4632        18337
## 56                    6421        18338
## 57                    7783        18339
## 58                   13677        18340
## 59                   19100        18341
## 60                   25489        18342
## 61                   33276        18343
## 62                   43847        18344
## 63                   53740        18345
## 64                   65778        18346
## 65                   83836        18347
## 66                  101657        18348
## 67                  121478        18349
## 68                  140886        18350
## 69                  161807        18351
## 70                  188172        18352
## 71                  213372        18353
```

```
##-----------------------------------------
## log10 transform total # cases
##-----------------------------------------
Corona_Cases$Total_confirmed_cases.log<-log(Corona_Cases$Total_confirmed_cases,10)
##-----------------------------------------
```

```r
##------------------------------------------
## Compute # of days since 100th for US data
##------------------------------------------

# Find day that 100th case was found for Country/Province. NOTE: Non US countries may have weird provin
# TODO: consider city-level summary as well. This data may be sparse

Corona_Cases<-merge(Corona_Cases,ddply(filter(Corona_Cases,Total_confirmed_cases>100),c("Country.Region
Corona_Cases$Days_since_100<-Corona_Cases$Date.numeric-Corona_Cases$case100_date

# Filter df for US state-wide stats
Corona_Cases.US<-filter(Corona_Cases,Country.Region=="US_state" & Total_confirmed_cases>0)

# Preview
head(Corona_Cases)
```

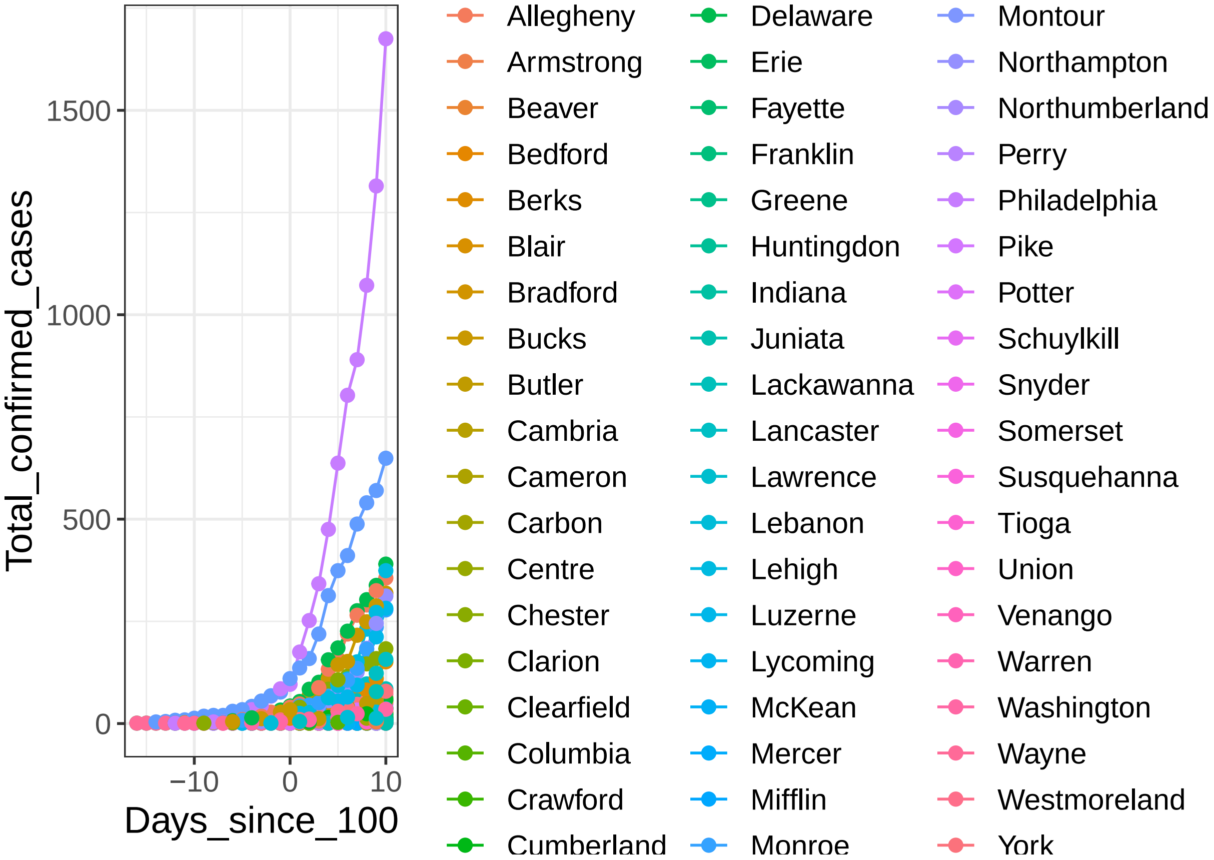
```
##   Province.State Country.Region Lat Long City       Date Total_confirmed_cases
## 1                   Afghanistan  33   65   NA 2020-03-23                    40
## 2                   Afghanistan  33   65   NA 2020-01-31                     0
## 3                   Afghanistan  33   65   NA 2020-01-29                     0
## 4                   Afghanistan  33   65   NA 2020-02-12                     0
## 5                   Afghanistan  33   65   NA 2020-03-25                    84
## 6                   Afghanistan  33   65   NA 2020-02-04                     0
##   Date.numeric Total_confirmed_cases.log case100_date Days_since_100
## 1        18344                  1.602060        18348             -4
## 2        18292                      -Inf        18348            -56
## 3        18290                      -Inf        18348            -58
## 4        18304                      -Inf        18348            -44
## 5        18346                  1.924279        18348             -2
## 6        18296                      -Inf        18348            -52
```

```r
head(Corona_Cases.US)
```

```
##   Province.State Country.Region      Lat      Long     City       Date
## 1        Alabama       US_state 32.99642 -87.12511     Bibb 2020-03-30
## 2        Alabama       US_state 33.98211 -86.56791   Blount 2020-03-30
## 3        Alabama       US_state 32.99642 -87.12511     Bibb 2020-03-31
## 4        Alabama       US_state 33.77484 -85.82630  Calhoun 2020-04-01
## 5        Alabama       US_state 33.67679 -85.52006 Cleburne 2020-04-01
## 6        Alabama       US_state 32.99642 -87.12511     Bibb 2020-04-01
##   Total_confirmed_cases Date.numeric Total_confirmed_cases.log case100_date
## 1                     2        18351                 0.3010300        18346
## 2                     5        18351                 0.6989700        18346
## 3                     3        18352                 0.4771213        18346
## 4                    11        18353                 1.0413927        18346
## 5                     6        18353                 0.7781513        18346
## 6                     3        18353                 0.4771213        18346
##   Days_since_100
## 1              5
## 2              5
## 3              6
## 4              7
## 5              7
## 6              7
```

```
#Corona_Cases.QC<-filter(Corona_Cases,Country.Region %in% countries_with_sufficient_data)

ggplot(filter(Corona_Cases.US,Province.State %in% c("Pennsylvania")),aes(x=Days_since_100,y=Total_confi
```



```
filter(Corona_Cases.US,Province.State %in% c("Pennsylvania") & Date=="2020-03-30") %>% arrange(Total_co
```

```
##    Province.State Country.Region      Lat      Long           City       Date
## 1    Pennsylvania       US_state 41.82148 -75.80072    Susquehanna 2020-03-30
## 2    Pennsylvania       US_state 40.42163 -77.97673     Huntingdon 2020-03-30
## 3    Pennsylvania       US_state 40.84785 -76.70798 Northumberland 2020-03-30
## 4    Pennsylvania       US_state 41.77255 -77.25433          Tioga 2020-03-30
## 5    Pennsylvania       US_state 40.40207 -77.26297          Perry 2020-03-30
## 6    Pennsylvania       US_state 41.81305 -79.26970         Warren 2020-03-30
## 7    Pennsylvania       US_state 41.43626 -78.20377        Cameron 2020-03-30
## 8    Pennsylvania       US_state 41.40323 -79.75845        Venango 2020-03-30
## 9    Pennsylvania       US_state 41.80939 -78.56478         McKean 2020-03-30
## 10   Pennsylvania       US_state 40.61117 -77.61071        Mifflin 2020-03-30
## 11   Pennsylvania       US_state 41.19266 -79.42414        Clarion 2020-03-30
## 12   Pennsylvania       US_state 40.77129 -77.06841         Snyder 2020-03-30
## 13   Pennsylvania       US_state 41.74472 -77.89560         Potter 2020-03-30
## 14   Pennsylvania       US_state 40.65241 -79.08963        Indiana 2020-03-30
## 15   Pennsylvania       US_state 39.97173 -79.02700       Somerset 2020-03-30
## 16   Pennsylvania       US_state 40.49527 -78.71377        Cambria 2020-03-30
## 17   Pennsylvania       US_state 41.78869 -76.51571       Bradford 2020-03-30
## 18   Pennsylvania       US_state 40.81666 -79.46291      Armstrong 2020-03-30
## 19   Pennsylvania       US_state 40.53359 -77.39975        Juniata 2020-03-30
## 20   Pennsylvania       US_state 41.00111 -78.47593     Clearfield 2020-03-30
## 21   Pennsylvania       US_state 41.34311 -77.06630       Lycoming 2020-03-30
## 22   Pennsylvania       US_state 41.68448 -80.10761       Crawford 2020-03-30
## 23   Pennsylvania       US_state 40.96189 -77.05996          Union 2020-03-30
## 24   Pennsylvania       US_state 40.47961 -78.34917          Blair 2020-03-30
## 25   Pennsylvania       US_state 41.04822 -76.40565       Columbia 2020-03-30
## 26   Pennsylvania       US_state 39.85747 -80.22357         Greene 2020-03-30
## 27   Pennsylvania       US_state 41.30249 -80.25817         Mercer 2020-03-30
## 28   Pennsylvania       US_state 39.87140 -77.21610          Adams 2020-03-30
## 29   Pennsylvania       US_state 40.99206 -80.33394       Lawrence 2020-03-30
## 30   Pennsylvania       US_state 41.03024 -76.66346        Montour 2020-03-30
## 31   Pennsylvania       US_state 41.64938 -75.29957          Wayne 2020-03-30
## 32   Pennsylvania       US_state 39.92041 -79.64291        Fayette 2020-03-30
## 33   Pennsylvania       US_state 39.92957 -77.72158       Franklin 2020-03-30
## 34   Pennsylvania       US_state 41.99254 -80.03302           Erie 2020-03-30
## 35   Pennsylvania       US_state 40.91545 -75.70685         Carbon 2020-03-30
## 36   Pennsylvania       US_state 40.92059 -77.82201         Centre 2020-03-30
## 37   Pennsylvania       US_state 40.16254 -77.26131     Cumberland 2020-03-30
## 38   Pennsylvania       US_state 40.19209 -80.24583     Washington 2020-03-30
## 39   Pennsylvania       US_state 40.36680 -76.45652        Lebanon 2020-03-30
## 40   Pennsylvania       US_state 40.70497 -76.21508     Schuylkill 2020-03-30
## 41   Pennsylvania       US_state 40.41377 -76.77993        Dauphin 2020-03-30
## 42   Pennsylvania       US_state 41.33155 -75.03208           Pike 2020-03-30
## 43   Pennsylvania       US_state 40.68255 -80.34922         Beaver 2020-03-30
## 44   Pennsylvania       US_state 40.91153 -79.91351         Butler 2020-03-30
## 45   Pennsylvania       US_state 39.92101 -76.73040           York 2020-03-30
## 46   Pennsylvania       US_state 40.31378 -79.46615   Westmoreland 2020-03-30
## 47   Pennsylvania       US_state 41.43565 -75.60379     Lackawanna 2020-03-30
## 48   Pennsylvania       US_state 40.41571 -75.92458          Berks 2020-03-30
## 49   Pennsylvania       US_state 40.03905 -76.24770      Lancaster 2020-03-30
## 50   Pennsylvania       US_state 39.97292 -75.74768        Chester 2020-03-30
## 51   Pennsylvania       US_state 41.17823 -75.98448        Luzerne 2020-03-30
## 52   Pennsylvania       US_state 41.05934 -75.34031         Monroe 2020-03-30
## 53   Pennsylvania       US_state 40.75183 -75.30472    Northampton 2020-03-30
## 54   Pennsylvania       US_state 40.61548 -75.59435         Lehigh 2020-03-30
## 55   Pennsylvania       US_state 40.33682 -75.10837          Bucks 2020-03-30
## 56   Pennsylvania       US_state 40.46810 -79.98168      Allegheny 2020-03-30
## 57   Pennsylvania       US_state 39.91680 -75.40244       Delaware 2020-03-30
## 58   Pennsylvania       US_state 40.21054 -75.36652     Montgomery 2020-03-30
## 59   Pennsylvania       US_state 40.00339 -75.13793   Philadelphia 2020-03-30
##    Total_confirmed_cases Date.numeric Total_confirmed_cases.log case100_date
## 1                      1        18351                 0.0000000        18343
## 2                      1        18351                 0.0000000        18343
## 3                      1        18351                 0.0000000        18343
## 4                      1        18351                 0.0000000        18343
## 5                      1        18351                 0.0000000        18343
## 6                      1        18351                 0.0000000        18343
## 7                      1        18351                 0.0000000        18343
## 8                      1        18351                 0.0000000        18343
## 9                      1        18351                 0.0000000        18343
## 10                     1        18351                 0.0000000        18343
## 11                     1        18351                 0.0000000        18343
## 12                     2        18351                 0.3010300        18343
## 13                     2        18351                 0.3010300        18343
## 14                     2        18351                 0.3010300        18343
## 15                     2        18351                 0.3010300        18343
## 16                     2        18351                 0.3010300        18343
## 17                     3        18351                 0.4771213        18343
## 18                     3        18351                 0.4771213        18343
## 19                     3        18351                 0.4771213        18343
## 20                     4        18351                 0.6020600        18343
## 21                     4        18351                 0.6020600        18343
## 22                     4        18351                 0.6020600        18343
## 23                     4        18351                 0.6020600        18343
## 24                     6        18351                 0.7781513        18343
## 25                     6        18351                 0.7781513        18343
## 26                     7        18351                 0.8450980        18343
## 27                     7        18351                 0.8450980        18343
## 28                     8        18351                 0.9030900        18343
## 29                    10        18351                 1.0000000        18343
## 30                    10        18351                 1.0000000        18343
## 31                    10        18351                 1.0000000        18343
## 32                    11        18351                 1.0413927        18343
## 33                    12        18351                 1.0791812        18343
## 34                    13        18351                 1.1139434        18343
## 35                    13        18351                 1.1139434        18343
## 36                    24        18351                 1.3802112        18343
## 37                    24        18351                 1.3802112        18343
## 38                    26        18351                 1.4149733        18343
## 39                    27        18351                 1.4313638        18343
## 40                    30        18351                 1.4771213        18343
## 41                    36        18351                 1.5563025        18343
## 42                    39        18351                 1.5910646        18343
## 43                    44        18351                 1.6434527        18343
## 44                    49        18351                 1.6901961        18343
## 45                    54        18351                 1.7323938        18343
## 46                    55        18351                 1.7403627        18343
## 47                    62        18351                 1.7923917        18343
## 48                    82        18351                 1.9138139        18343
## 49                    97        18351                 1.9867717        18343
## 50                   146        18351                 2.1643529        18343
## 51                   150        18351                 2.1760913        18343
## 52                   182        18351                 2.2600714        18343
## 53                   184        18351                 2.2648178        18343
## 54                   231        18351                 2.3636120        18343
## 55                   249        18351                 2.3961993        18343
## 56                   290        18351                 2.4623980        18343
## 57                   303        18351                 2.4814426        18343
## 58                   540        18351                 2.7323938        18343
## 59                  1072        18351                 3.0301948        18343
##    Days_since_100
## 1               8
## 2               8
## 3               8
## 4               8
## 5               8
## 6               8
## 7               8
## 8               8
## 9               8
## 10              8
## 11              8
## 12              8
## 13              8
## 14              8
## 15              8
## 16              8
## 17              8
## 18              8
## 19              8
## 20              8
## 21              8
## 22              8
## 23              8
## 24              8
## 25              8
## 26              8
## 27              8
## 28              8
## 29              8
## 30              8
## 31              8
## 32              8
## 33              8
## 34              8
## 35              8
## 36              8
## 37              8
## 38              8
## 39              8
## 40              8
## 41              8
## 42              8
## 43              8
## 44              8
## 45              8
## 46              8
## 47              8
## 48              8
## 49              8
## 50              8
## 51              8
## 52              8
## 53              8
## 54              8
## 55              8
## 56              8
## 57              8
## 58              8
## 59              8
```

ANALYSIS

Q1: What is the trend in total cases?

Plot # of cases vs time (US only) FUTURE: compare different countries

```
##-------------------------------------------
## Linear model for days since 100 cases vs log10(confirmed cases)
##-------------------------------------------

Corona_Cases.US<-filter(Corona_Cases,Country.Region=="US" & Total_confirmed_cases>0)
Corona_Cases.US.case100<-filter(Corona_Cases.US, Days_since_100>=0)
# linear model parameters
(model_fit<-lm(formula = Total_confirmed_cases.log~Days_since_100,data= Corona_Cases.US.case100 ))
```

```
##
## Call:
## lm(formula = Total_confirmed_cases.log ~ Days_since_100, data = Corona_Cases.US.case100)
##
## Coefficients:
##    (Intercept)  Days_since_100
##         2.1367          0.1184
```

```
N<-ddply(filter(Corona_Cases,Total_confirmed_cases>100),c("Country.Region"),summarise,n=length(Country.
#ddply(filter(Corona_Cases,Total_confirmed_cases>100 & Country.Region %in% N[N$Country.Region>2,"Countr

(slope<-model_fit$coefficients[2])
```

```
## Days_since_100
##      0.1183638
```

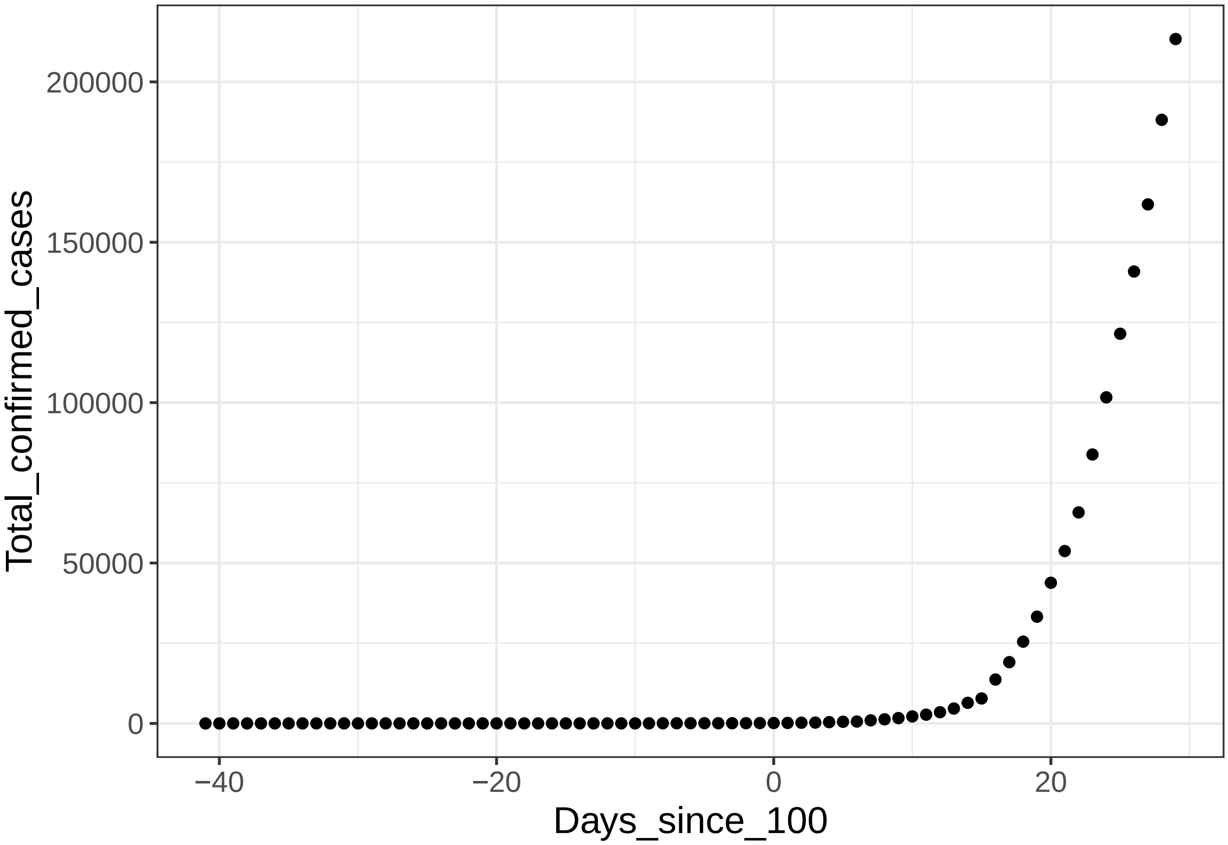
```
(intercept<-model_fit$coefficients[1])
```

```
## (Intercept)
##    2.136697
```

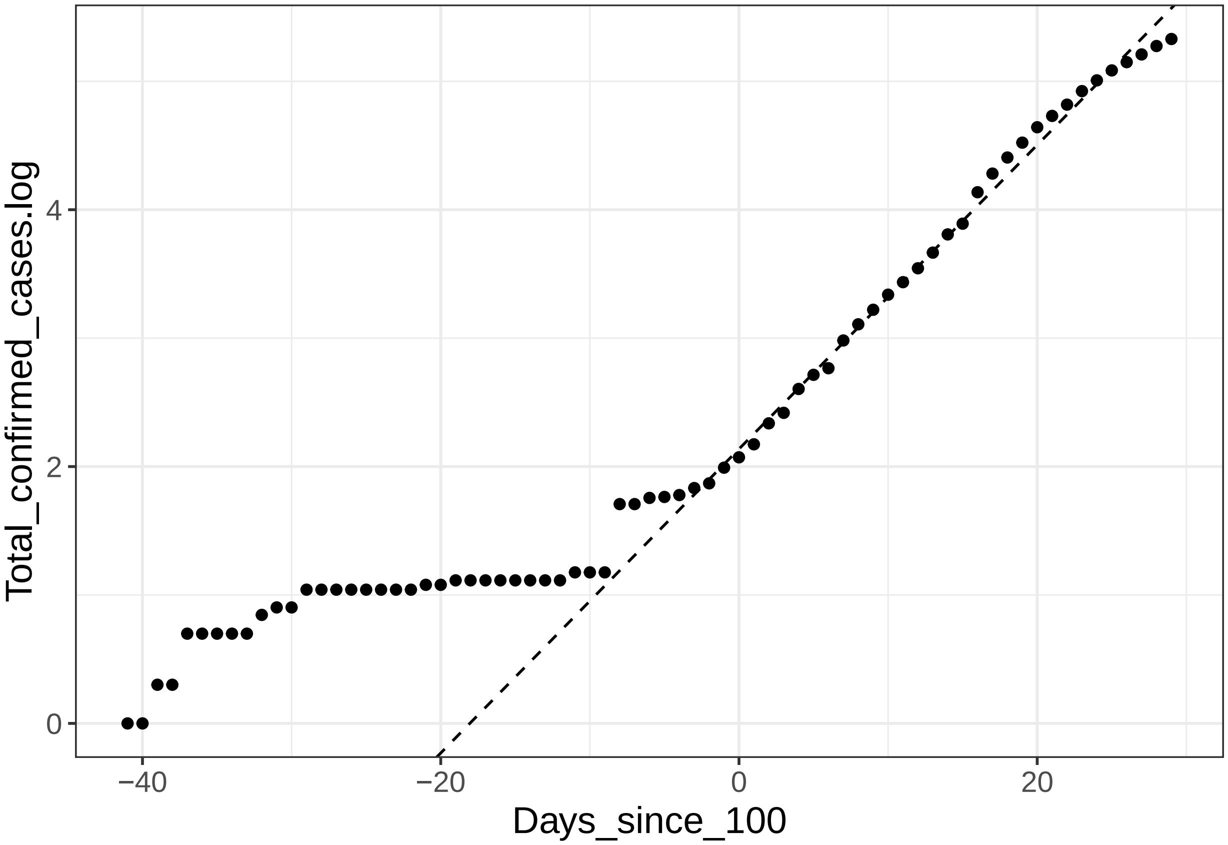
```
# Correlation coefficient
cor(x = Corona_Cases.US.case100$Days_since_100,y = Corona_Cases.US.case100$Total_confirmed_cases.log)
```

```
## [1] 0.9961539
```

```
(Corona_Cases.US.plot<-ggplot(Corona_Cases.US,aes(x=Days_since_100,y=Total_confirmed_cases))+
  default_theme+
  geom_point())
```



```
(Corona_Cases.US.log.plot<-ggplot(Corona_Cases.US,aes(x=Days_since_100,y=Total_confirmed_cases.log))+
  geom_abline(slope = slope,intercept = intercept,lty=2)+
   default_theme+
  geom_point())
```



```
write_plot(Corona_Cases.US.plot,wd = results_dir)
```

```
## [1] "/Users/ssmith/coronavirus/results/Corona_Cases.US.plot.png"
```

```
write_plot(Corona_Cases.US.log.plot,wd = results_dir)
```

```
## [1] "/Users/ssmith/coronavirus/results/Corona_Cases.US.log.plot.png"
```

Q2: What is the predicted number of cases?

What is the prediction of COVID-19 based on model thus far?

Additional questions:

WHy did it take to day 40 to start a log linear trend? How long will it be till x number of cases? When will the plateu happen? Are any effects noticed with social distancing? Delays

```
##-----------------------------------------
## Prediction and Prediction Accuracy
##-----------------------------------------

# What is the predict # of cases for the next few days?
# How is the model performing historically?

# Formula for # of cases by x days
paste0("log10_total_cases = ",slope,"*days + ",intercept)
```

```
## [1] "log10_total_cases = 0.118363785480949*days + 2.13669736383738"
```

```
paste0("total_cases = 10^(",slope,"*days + ",intercept,")")
```

```
## [1] "total_cases = 10^(0.118363785480949*days + 2.13669736383738)"
```

```
#Days untill... cases:
# 2.5k, 5k and 1M:
paste0("2.5k cases is ",(log(2.5E5,10) - intercept)/slope," days")
```

```
## [1] "2.5k cases is 27.5527065274501 days"
```

```
paste0("5k cases is ",(log(5E5,10)- intercept)/slope," days")
```

```
## [1] "5k cases is 30.0959674956661 days"
```

```
paste0("1M cases is ",(log(1E6,10)- intercept)/slope," days")
```

```
## [1] "1M cases is 32.639228463882 days"
```

```
today_num<-max(Corona_Cases.US$Days_since_100)
predicted_days<-today_num+c(1,2,3,7)

#mods = dlply(mydf, .(x3), lm, formula = y ~ x1 + x2)
#today:
Corona_Cases.US[Corona_Cases.US$Days_since_100==(today_num-1),]
```

```
##    Province.State Country.Region     Lat     Long   City     Date
## 24                            US 37.0902 -95.7129   NA 2020-03-31
##    Total_confirmed_cases Date.numeric Total_confirmed_cases.log case100_date
## 24                188172        18352                  5.274555        18324
##    Days_since_100
## 24             28
```

```
Corona_Cases.US[Corona_Cases.US$Days_since_100==today_num,]
```

```
##    Province.State Country.Region     Lat     Long   City     Date
## 43                            US 37.0902 -95.7129   NA 2020-04-01
##    Total_confirmed_cases Date.numeric Total_confirmed_cases.log case100_date
## 43                213372        18353                  5.329137        18324
##    Days_since_100
## 43             29
```

```
prediction_model(m = slope,b=intercept,days=predicted_days)
```

```
##   Days_since_100 Total_confirmed_cases Total_confirmed_cases.log
## 1             30              487091.9                  5.687611
## 2             31              639697.6                  5.805975
## 3             32              840114.5                  5.924338
## 4             36             2499157.6                  6.397794
```

```
Corona_Cases.US$type<-"Historical"
names(Corona_Cases)
```

```
##  [1] "Province.State"            "Country.Region"
##  [3] "Lat"                       "Long"
##  [5] "City"                      "Date"
##  [7] "Total_confirmed_cases"     "Date.numeric"
##  [9] "Total_confirmed_cases.log" "case100_date"
## [11] "Days_since_100"
```

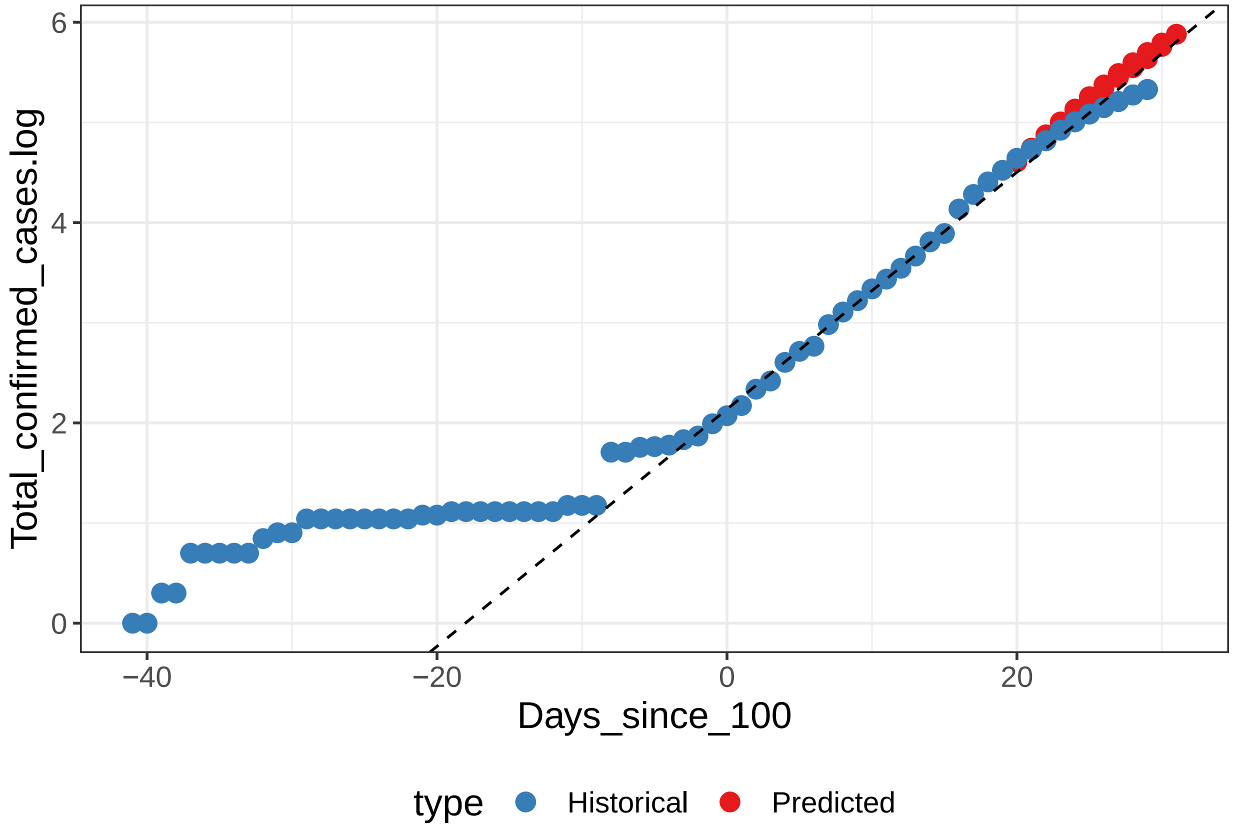
```
Corona_Cases_wprediction<-rbind.fill(Corona_Cases.US,data.frame(Code="USA",type="MAR26_prediction",predi

Corona_Cases.US.prediction<-Corona_Cases_wprediction
prediction_values<-prediction_model(m=slope,b=intercept,days = predicted_days)$Total_confirmed_cases

histoical_model<-data.frame(date=today_num,m=slope,b=intercept)

# model for previous y days
historical_model_predictions<-data.frame(day_x=NULL,Days_since_100=NULL,Total_confirmed_cases=NULL,Total
for(i in c(1,2,3,4,5,6,7,8,9,10)){
  #i<-1
day_x<-today_num-i # 1, 2, 3, 4
day_x_nextweek<-day_x+c(1,2,3)
model_fit_x<-lm(data = filter(Corona_Cases.US.case100,Days_since_100 < day_x),formula = Total_confirmed_
prediction_day_x_nextweek<-prediction_model(m = model_fit_x$coefficients[2],b = model_fit_x$coefficients
prediction_day_x_nextweek$type<-"Predicted"
acutal_day_x_nextweek<-filter(Corona_Cases.US,Days_since_100 %in% day_x_nextweek) %>% select(c(Days_sinc
acutal_day_x_nextweek$type<-"Historical"
historical_model_predictions.i<-data.frame(day_x=day_x,rbind(acutal_day_x_nextweek,prediction_day_x_next
historical_model_predictions<-rbind(historical_model_predictions.i,historical_model_predictions)
}

(historical_model_predictions.plot<-ggplot(rbind.fill(historical_model_predictions,data.frame(Corona_Cas
    geom_point(size=3)+
    default_theme+
    theme(legend.position = "bottom")+
      geom_abline(slope = slope,intercept =intercept,lty=2)+
    scale_color_manual(values = c("Historical"="#377eb8","Predicted"="#e41a1c")))
```



```
write_plot(historical_model_predictions.plot,wd=results_dir)
```

```
## [1] "/Users/ssmith/coronavirus/results/historical_model_predictions.plot.png"
```

```
##-------------------------------------------
## filter input_data1
##-------------------------------------------
input_data1.filter<-fitler(input_data1,col1=="foo")
##-------------------------------------------

##-------------------------------------------
## sub question 1
##-------------------------------------------
table(input_data1.filter$col<5)
##-------------------------------------------

##-------------------------------------------
## sub question 2
##-------------------------------------------
table(input_data1.filter$col<10)
##-------------------------------------------

##-------------------------------------------
## plot data
##-------------------------------------------
(input_data1.filter.plot<-ggplot(input_data1.filter,aes(x=col1,y=col2.log))+
   geom_point()+
   default_plot_theme)
```

```
write_plot(input_data1.filter.plot,wd=results_dir)
##-------------------------------------------
results_dir
```

CONCLUSION

A concluding remark(s) on the major findings, preferabbly to pointers where the data can be found.

Helps to have a bullet point for each analysis chunk or an answer to each of the above ‘questions’: * Answer 1. * Answer 2.

END

Cheatsheet: http://rmarkdown.rstudio.com> # TODO * mkdir the results dir if it doesn’t exist * make ggplot a dependency for plot.utils? * automated way of downloading daily data * fix plot_utils, add dataset and documentation * Auto git mv the new data?

Sandbox

```
##TODO:
# Geographical heatmap!
ggplot(data = Corona_Cases) +
    geom_sf()

reportTimeStamp = format(Sys.time(), "%Y-%m-%d (%a) %X")
titleStr        = paste("COVID-19 Deaths by Country/Region ", "[", reportTimeStamp, "]", sep="")
```