Masters Data Science Exam: README

Samantha Scott

17/06/2022

Question 1

Code

```
gc()
```

```
##          used (Mb) gc trigger (Mb) limit (Mb) max used (Mb)
## Ncells 463379 24.8     989484 52.9         NA   668911 35.8
## Vcells 865592  6.7    8388608 64.0      16384  1838985 14.1
```

```
library(pacman)
p_load(tidyverse, lubridate)

list.files('Question1/Code', full.names = T, recursive = T) %>% as.list() %>% walk(~source(.))
```

Loading Data

```
 library(readr)
 Deaths_by_cause <- read_csv("Question1/Data/Covid/Deaths_by_cause.csv")
```

```
## Rows: 7273 Columns: 36
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): Entity, Code, Number of executions (Amnesty International)
## dbl (33): Year, Deaths - Meningitis - Sex: Both - Age: All Ages (Number), De...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

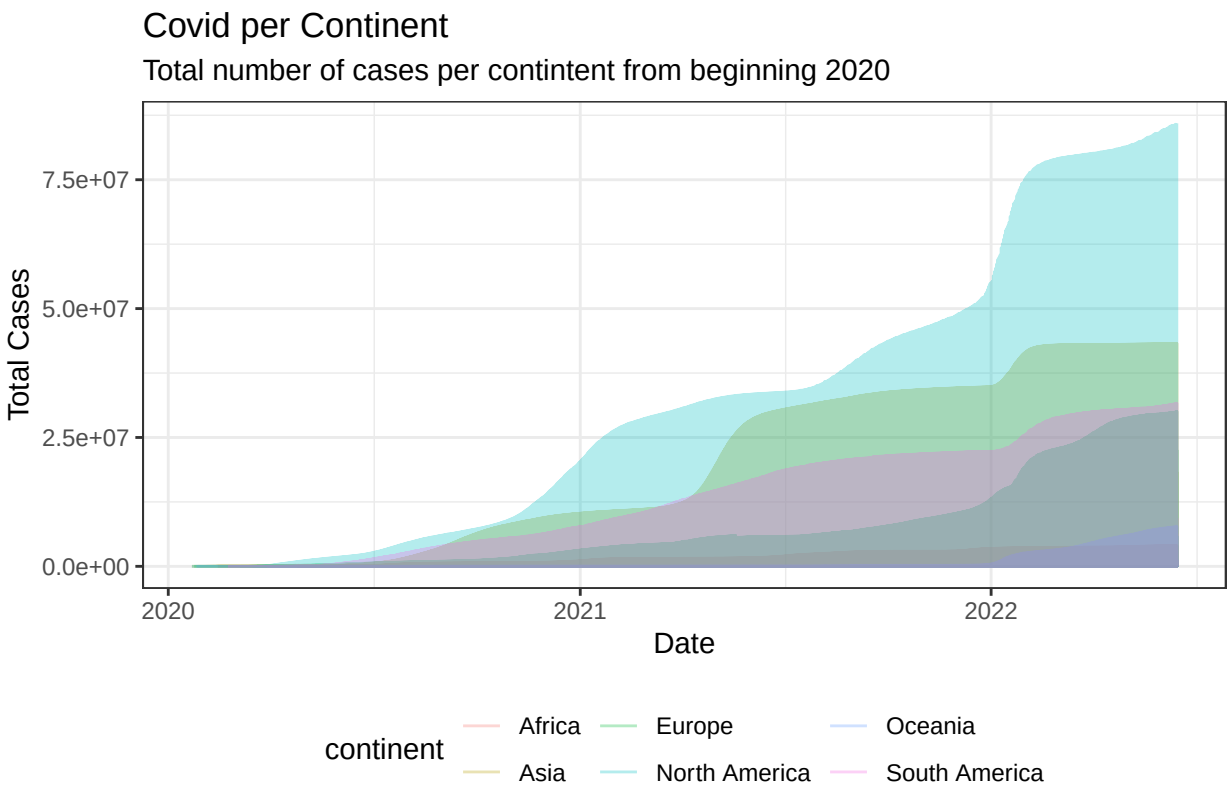
```
library(readr)
owid_covid_data <- read_csv("Question1/Data/Covid/owid-covid-data.csv")
```

```
## Rows: 194260 Columns: 67
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (4): iso_code, continent, location, tests_units
## dbl  (62): total_cases, new_cases, new_cases_smoothed, total_deaths, new_dea...
## date  (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Graph 1.1

```
g <- line_graph_continents(owid_covid_data)
```

```
## Adding missing grouping variables: `continent`
```



Note:OWID data used

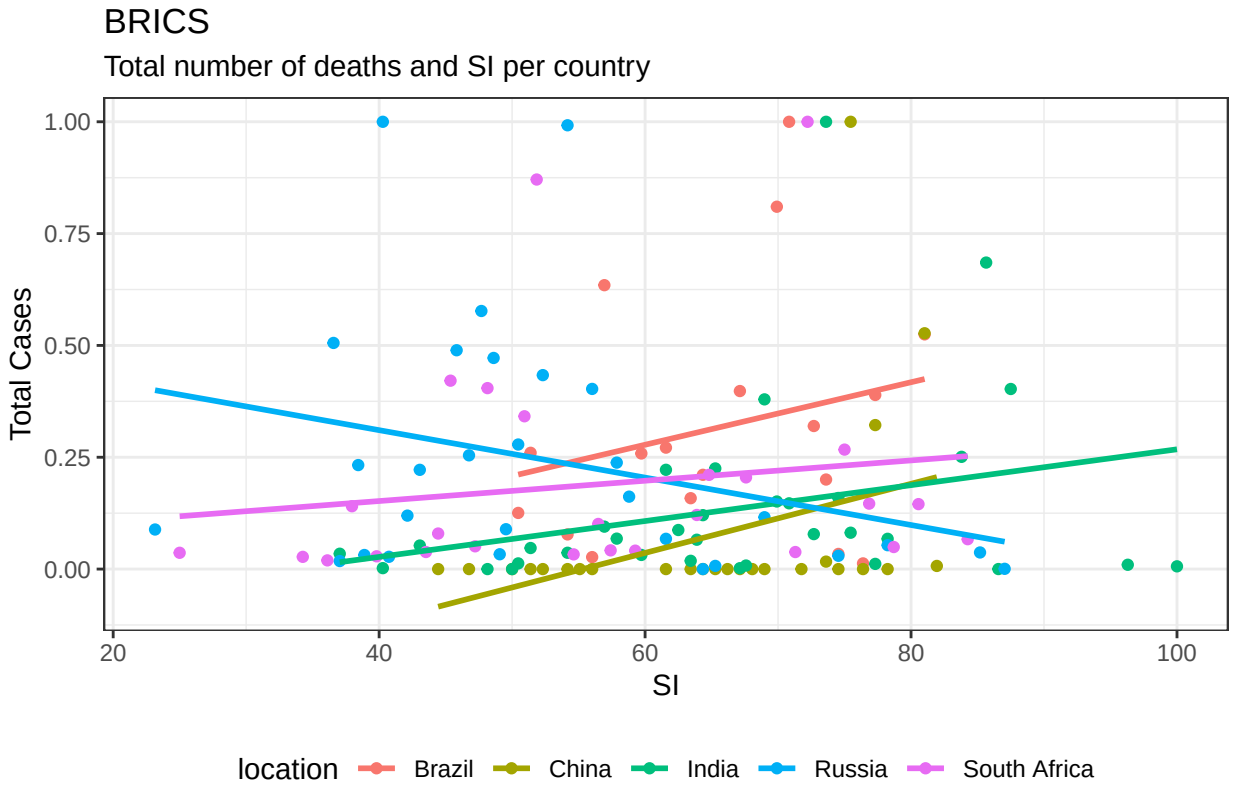
As seen in the graph, the continents with the highest number of total cases is Europe and North America. African countries are seen to have not been as greatly by COVID than North American, South American and European countries. In comparison, the number of total countries in Africa seem almost insignificant in comparison to countries in other continents.

However, to accurately interpret this graph, it is imperative that one take in to consideration the population of the continents.

Graph 1.2

```
g <- brics_SI_deaths(owid_covid_data)
```

```
## `summarise()` has grouped output by 'location'. You can override using the
## `.groups` argument.
## `geom_smooth()` using formula 'y ~ x'
```



Note:OWID data used A

seen in the figure above, BRICS countries are compared. When investigating the stringency index of these countries, it is evident that a stronger stringency index is positively correlated to total cases of each country.

Question 2

Code

```
gc()
```

```
##            used  (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
## Ncells  2397884 128.1    4033232 215.4         NA  4033232 215.4
## Vcells 17447972 133.2   31218135 238.2      16384 31110692 237.4
```

```
library(pacman)
p_load(tidyverse, lubridate)

list.files('Question2/Code', full.names = T, recursive = T) %>% as.list() %>% walk(~source(.))
```

Loading Data

```
library(readr)
london_weather <- read_csv("Question2/Data/London/london_weather.csv")
```

```
## Rows: 15341 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (10): date, cloud_cover, sunshine, global_radiation, max_temp, mean_temp...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

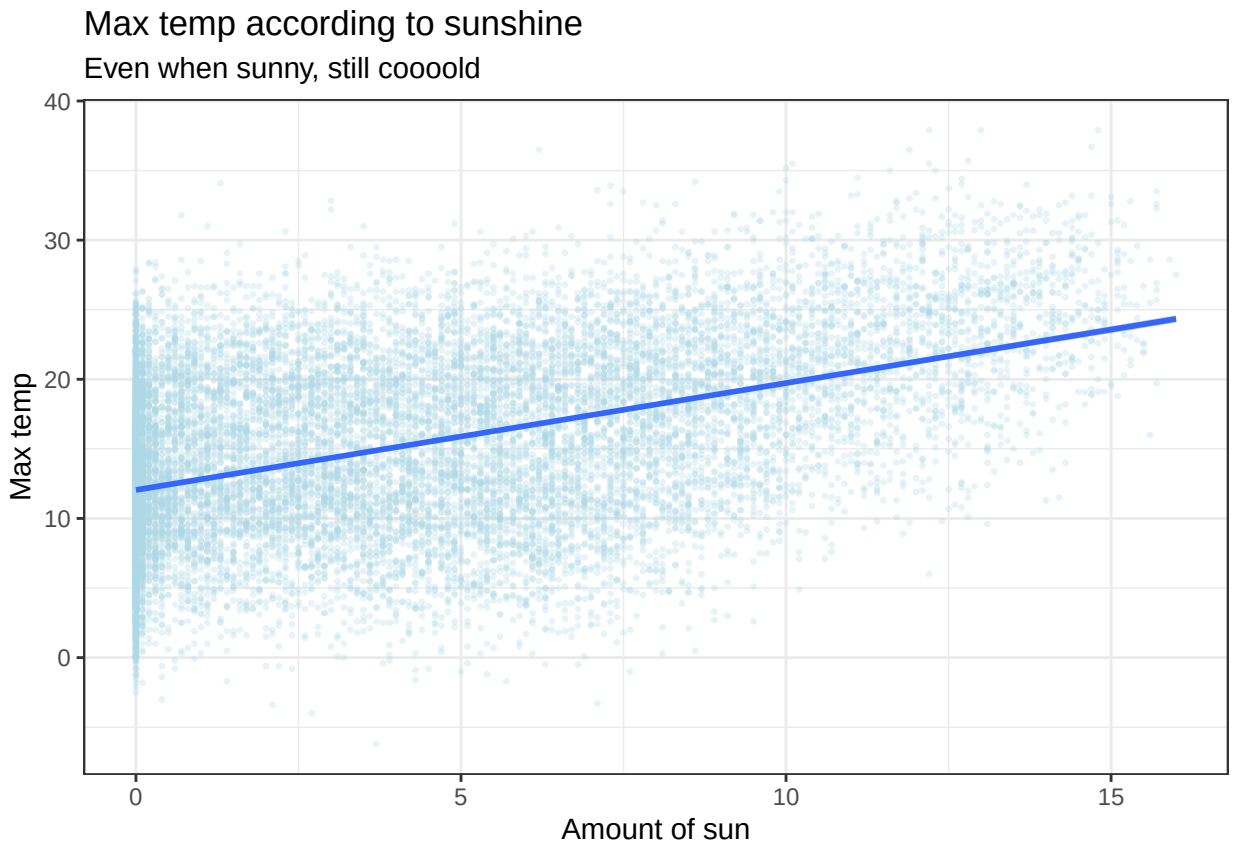
```
library(readr)
SUN_Hours <- read_csv("~/Desktop/20945043/Question2/Data/SUN Hours.csv")
```

```
## Rows: 20368 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): TmStamp
## dbl (16): RecNum, BattV_Min, TrackerWM_Avg, Tracker2WM_Avg, ShadowWM_Avg, Su...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Graph 2.1

```
g <- london_code(london_weather)
```

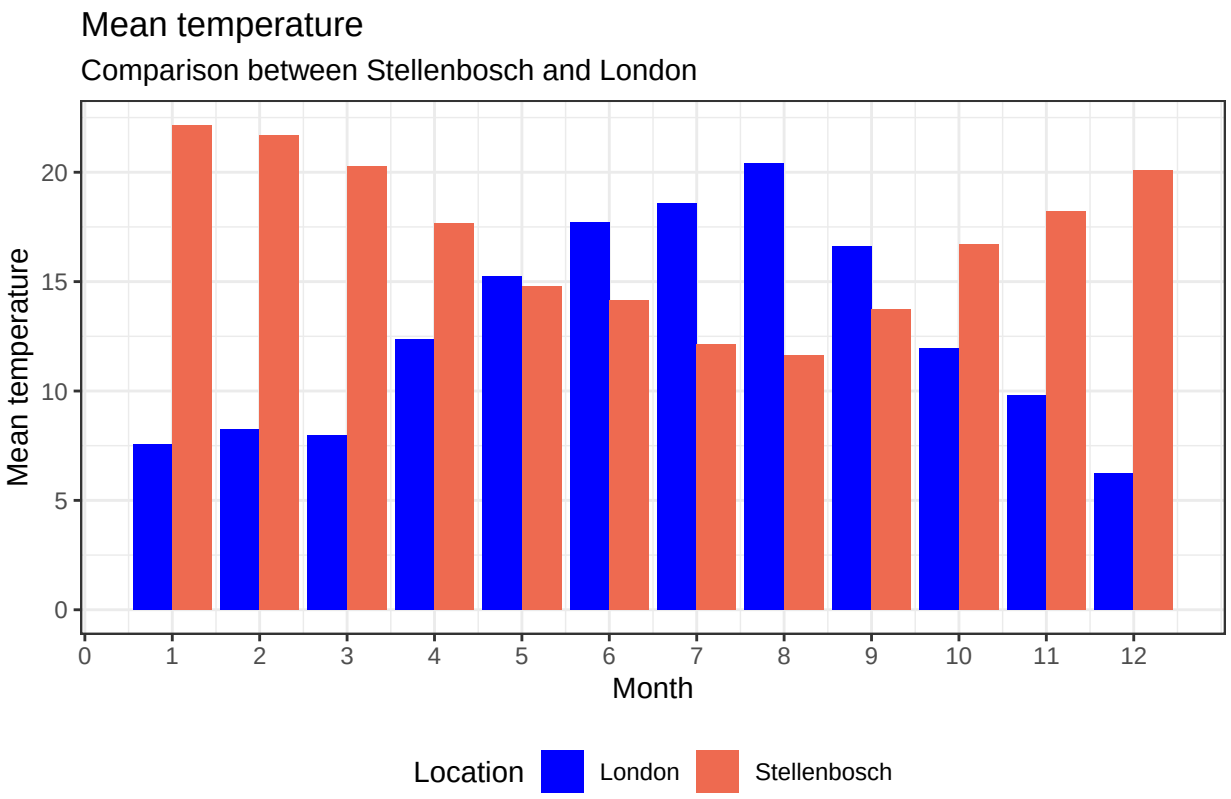
```
## `geom_smooth()` using formula 'y ~ x'
```



As seen in the graph above, there is a positive correlation between the maximum temperature and the amount of sun. However, this positive correlation is not very steep, indicating that a jump in the amount of sun only results in a small jump in the max temperature.

Graph 2.2

```
g <- lond_c(london_weather, SUN_Hours)
```



Note:Sauran External Data used

As depicted in the bar graph above, when compared, Stellenbosch winters are much warmer than London winters. And London summers are much cooler than Stellenbosch summers.

These two graphs suggest that moving to London is a bad idea.

Question 4

Code

```
gc()
```

```
##            used  (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
## Ncells  2442489 130.5    4033232 215.4         NA  4033232 215.4
## Vcells 18211022 139.0   31218135 238.2      16384 31110692 237.4
```

```
library(pacman)
p_load(tidyverse, lubridate)

list.files('Question4/Code', full.names = T, recursive = T) %>% as.list() %>% walk(~source(.))
```

Loading Data

```
library(readr)
credits <- read_csv("~/Desktop/20945043/Question4/Data/netflix/credits.csv")
```

```
## Rows: 77213 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): id, name, character, role
## dbl (1): person_id
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
library(readr)
titles <- read_csv("~/Desktop/20945043/Question4/Data/netflix/titles.csv")
```

```
## Rows: 5806 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): id, title, type, description, age_certification, genres, production...
## dbl (7): release_year, runtime, seasons, imdb_score, imdb_votes, tmdb_popula...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Graphs and Statistics

```
g <- netflix_code(titles)
```

```
## Adding missing grouping variables: `release_year`
## Selecting by ave_score
```

Figure 4.1

Table 1: IMDb Score: Movies vs Shows

	Movies	Shows
IMDb score	6.266980	7.017377

According to the data, shows did better than movies.

According to the accompanying data, the combination [‘scifi’, ‘family’, ‘fantasy’, ‘animation’, ‘action’] genre is most enjoyed out of all genre combinations, with an average IMDb score of 9.3.

```
g <- netflixcode(titles)
```

```
## ‘summarise()‘ has grouped output by ’type’. You can override using the
## ‘.groups‘ argument.
## ‘geom_smooth()‘ using formula ’y ~ x’
```

As seen in the figure, there is a positive correlation. Ratings are trustworthy.