COVID19_Johns_Hopkins

Carlos Barron

2022-11-08

```
knitr::opts_chunk$set(echo = TRUE)
webshot::install_phantomjs()
```

```
## It seems that the version of ‘phantomjs‘ installed is greater than or equal to the requested version
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
library(webshot)
library(lubridate)
```

```
## Loading required package: timechange
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(plotly)
```

```
##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
```

```
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(dplyr)
library(ggplot2)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

Introduction

After what was analyzed in the lectures, I'm interested to see if we can visualize a high level of comorbidity between smoking, exercising and COVID-19 within the dataset. Before I try to put a model in place, I need to find external datasets about smoking and physical exercise such that I may be able to make any further analysis. I will attempt to relate the datasets through the country column.

Data

COVID-19

The COVID-19 data for this report consists of 2 CSVs that you can find here.

Each one represents the confirmed cases and deaths worldwide.

Confirmed Cases

```
confirmed_global
```

```
## # A tibble: 603 x 3
## # Groups:   country [201]
##    country     year   cases
##    <chr>       <chr>  <dbl>
##  1 Afghanistan 2020   52330
##  2 Afghanistan 2021  158084
##  3 Afghanistan 2022  204610
##  4 Albania     2020   58316
##  5 Albania     2021  210224
##  6 Albania     2022  333161
##  7 Algeria     2020   99610
##  8 Algeria     2021  218432
##  9 Algeria     2022  270952
## 10 Andorra     2020    8049
## # ... with 593 more rows
```

Confirmed Deaths

```
confirmed_global
```

```
## # A tibble: 603 x 3
## # Groups:   country [201]
##    country     year   cases
##    <chr>       <chr>  <dbl>
##  1 Afghanistan 2020   52330
##  2 Afghanistan 2021  158084
##  3 Afghanistan 2022  204610
##  4 Albania     2020   58316
##  5 Albania     2021  210224
##  6 Albania     2022  333161
##  7 Algeria     2020   99610
##  8 Algeria     2021  218432
##  9 Algeria     2022  270952
## 10 Andorra     2020    8049
## # ... with 593 more rows
```

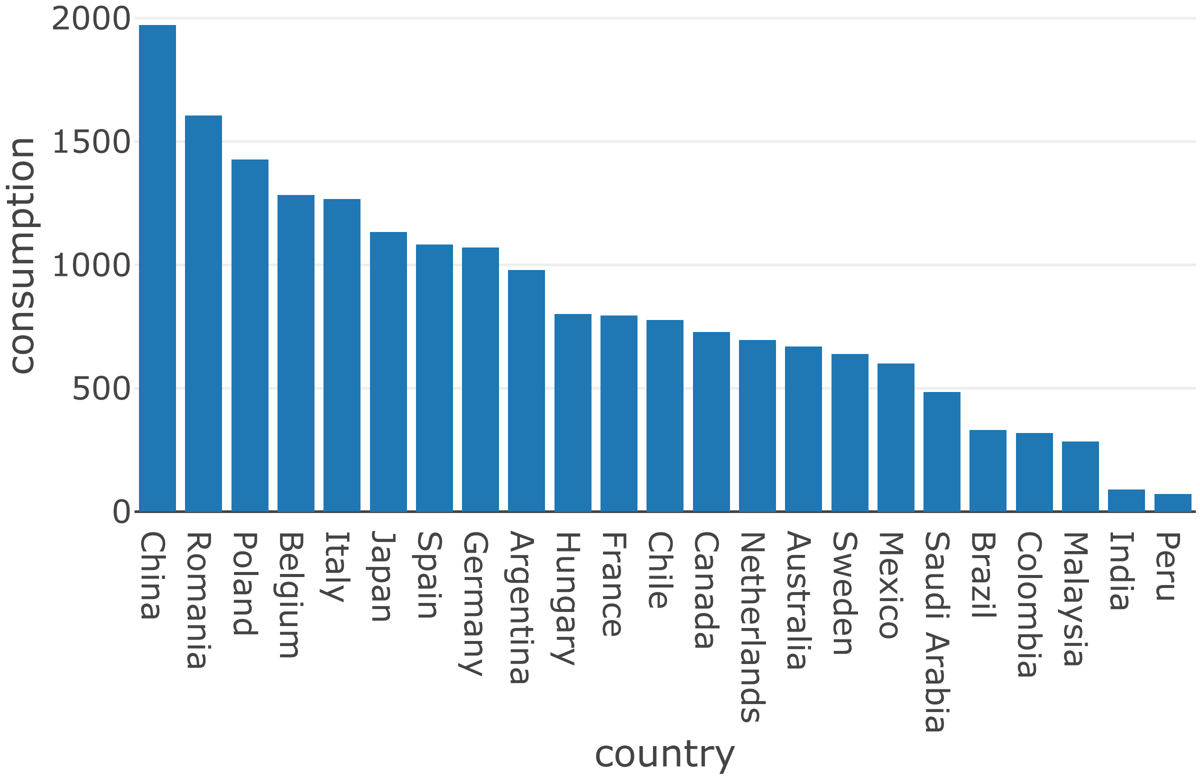
Tobacco Atlas

For cigarette consumption I will use the dataset avaialable throught theTobacco Atlas available here.

Fields

1. Country
2. Average daily number of cigarettes consumed per adult (15+ yr) smoker, 2019

```
avg_daily_cigar_consumption = avg_daily_cigar_consumption %>% filter(avg_daily_cigar_consumption$country

fig <- plot_ly(
  avg_daily_cigar_consumption,
  x = ~country,
  y = ~consumption,
  name = "Average Daily Cigar Consumption by Country",
  type = "bar",
  orientation="v"
) %>% layout(xaxis = list(categoryorder = "total descending"))
fig
```



Ipsos Global Advisor

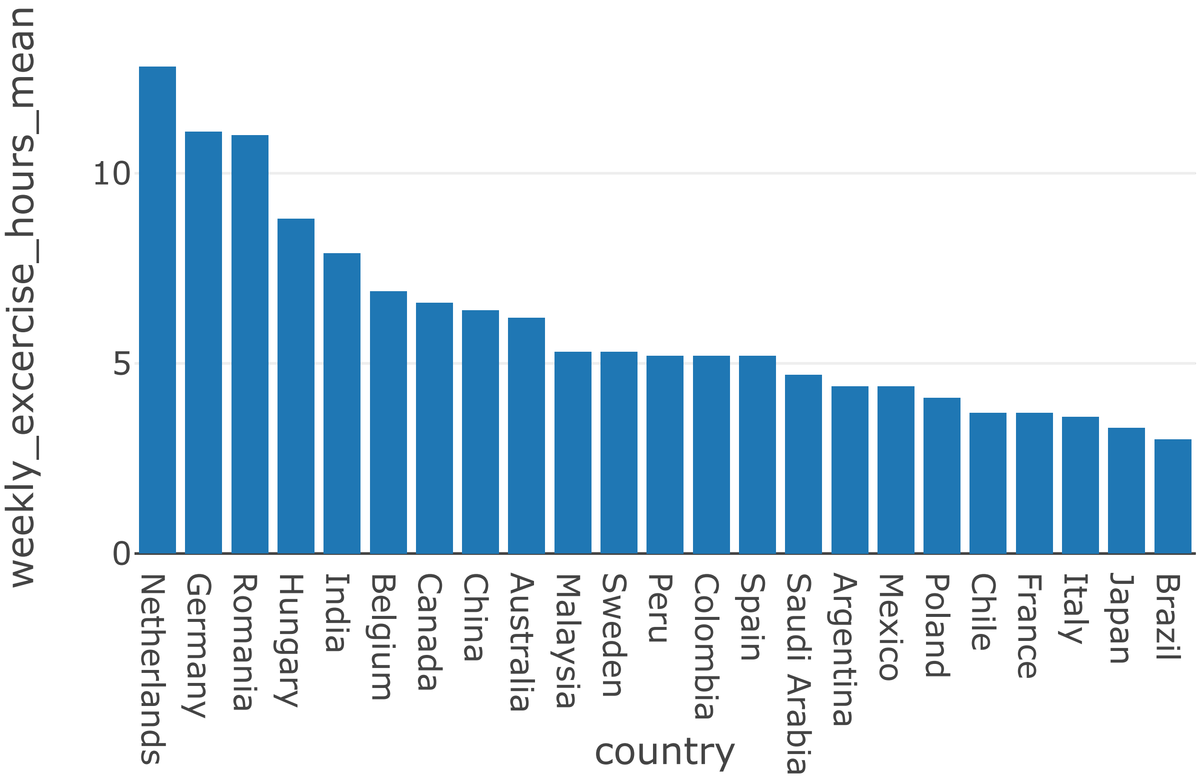
Global Views on Exercise and Team Sports

For the exercise information I will use the dataset available here.

Fields

1. Country
2. Mean Number of Hours Physical Excercise Per Week

```
weekly_excercise = weekly_excercise %>% filter(weekly_excercise$country %in% countries) %>% select(c(cou

fig <- plot_ly(
  weekly_excercise,
  x = ~country,
  y = ~weekly_excercise_hours_mean,
  name = "Weekly Excercise Hours Mean by Country",
  type = "bar",
  orientation="v"
) %>% layout(xaxis = list(categoryorder = "total descending"))
fig
```



United Nations

Department of Economic and Social Affairs, World Population Prospects 2022

For the age information I will use the dataset available here.

Fields

1. Country
2. Median Age

```
median_age_by_country = median_age_by_country %>% filter(median_age_by_country$year == 2020) %>% select
median_age_by_country = median_age_by_country %>% filter(median_age_by_country$country %in% countries) %

fig <- plot_ly(
  median_age_by_country,
  x = ~country,
  y = ~median_age,
  name = "Median Age by Country",
  type = "bar",
  orientation="v"
) %>% layout(xaxis = list(categoryorder = "total descending"))
fig
```

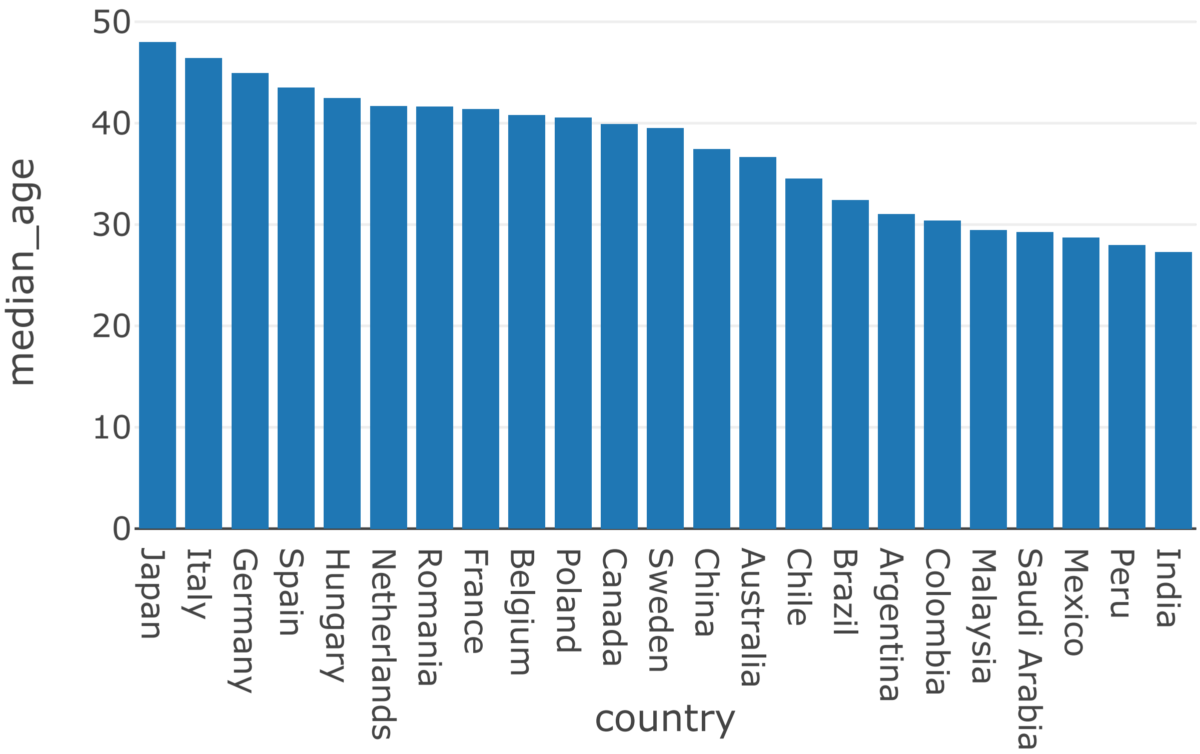


Table Joins

Now that we have the data loaded, let's join all of the different tables by country and year

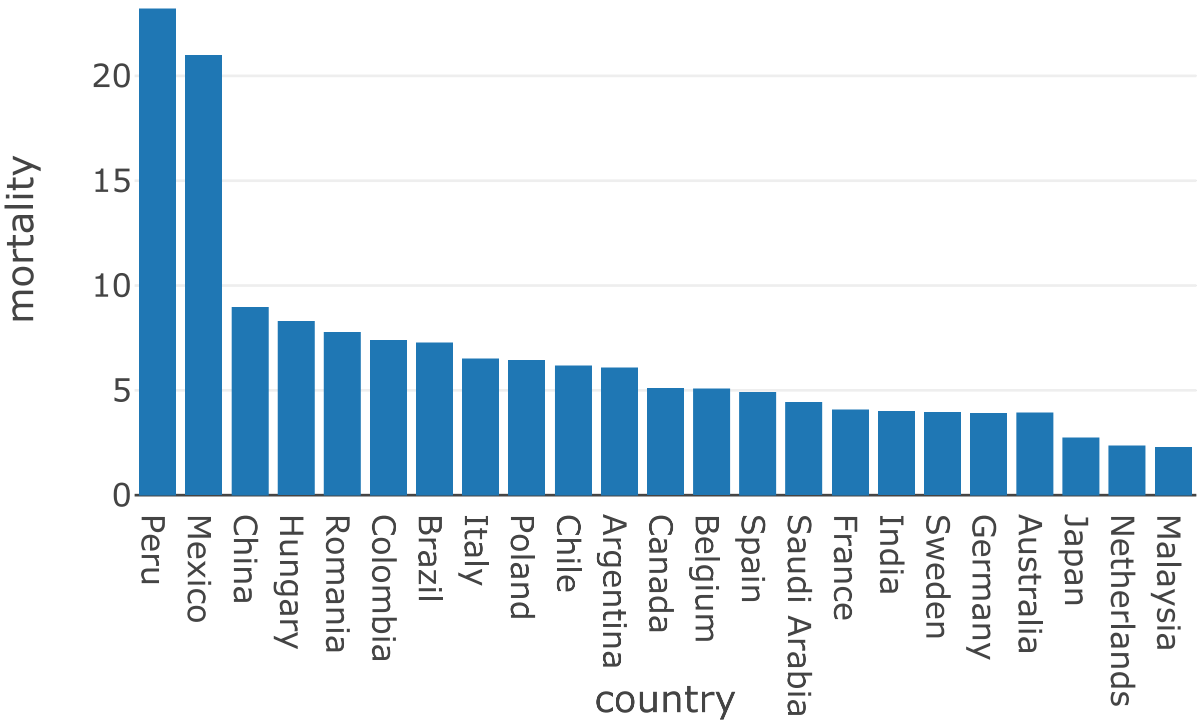
```
covid_stats = inner_join(
  confirmed_global, deaths_global, by = c('country','year')
) %>%  mutate(mortality = (
  deaths * 100) / cases
)

covid_stats = inner_join(
  covid_stats, avg_daily_cigar_consumption, by = c('country')
)
covid_stats = inner_join(
  covid_stats, weekly_excercise, by = c('country')
)
covid_stats = inner_join(
  covid_stats, median_age_by_country, by = c('country')
)
covid_stats
```

```
## # A tibble: 69 x 8
## # Groups:   country [23]
##    country   year     cases deaths mortality consumption weekly_excerc~1 media~2
##    <chr>     <chr>    <dbl>  <dbl>      <dbl>       <dbl>           <dbl>   <dbl>
##   1 Argentina 2020   1625514  43245      2.66         978.             4.4    31.0
##   2 Argentina 2021   5654408 117169      2.07         978.             4.4    31.0
##   3 Argentina 2022   9721718 130011      1.34         978.             4.4    31.0
##   4 Australia 2020     28425    909      3.20         668.             6.2    36.7
##   5 Australia 2021    425496   2253      0.529        668.             6.2    36.7
##   6 Australia 2022  10487217  15881      0.151        668.             6.2    36.7
##   7 Belgium   2020    646496  19528      3.02        1284.             6.9    40.8
##   8 Belgium   2021   2105343  28331      1.35        1284.             6.9    40.8
##   9 Belgium   2022   4624251  33000      0.714       1284.             6.9    40.8
##  10 Brazil    2020   7681032 195072      2.54         330.             3      32.4
## # ... with 59 more rows, and abbreviated variable names
## #   1: weekly_excercise_hours_mean, 2: median_age
```

Mortality

Let’s graph the mortality rate before the Vaccine came out (August 2021).

```
covid_stats_2020 <- filter(covid_stats, year == 2020)
fig <- plot_ly(
  covid_stats,
  x = ~country,
  y = ~mortality,
  name = "Mortality by Country",
  type = "bar",
  orientation="v"
) %>% layout(xaxis = list(categoryorder = "total descending"))
fig
```



Models

This is the model summary for cigarette consumption

```r
covid_model_consumption <- lm(deaths ~ consumption, data = covid_stats)
summary(covid_model_consumption)
```

```
##
## Call:
## lm(formula = deaths ~ consumption, data = covid_stats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -162110  -93661  -20662   38908  531415
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 194815.73   31870.81   6.113 5.64e-08 ***
## consumption   -113.89      33.39  -3.410   0.0011 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 130500 on 67 degrees of freedom
## Multiple R-squared:  0.1479, Adjusted R-squared:  0.1352
## F-statistic: 11.63 on 1 and 67 DF,  p-value: 0.001103
```

This is the model summary for weekly excercise mean

```
covid_model_excercise <- lm(deaths ~ weekly_excercise_hours_mean, data = covid_stats)
summary(covid_model_excercise)
```

```
##
## Call:
## lm(formula = deaths ~ weekly_excercise_hours_mean, data = covid_stats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -124917  -80723  -33363   12928  557159
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   162385      42162   3.851 0.000265 ***
## weekly_excercise_hours_mean   -10296       6415  -1.605 0.113182
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 138800 on 67 degrees of freedom
## Multiple R-squared:  0.03703,    Adjusted R-squared:  0.02266
## F-statistic: 2.576 on 1 and 67 DF,  p-value: 0.1132
```

And this is the model summary for the median age

```
covid_model_age <- lm(deaths ~ median_age, data = covid_stats)
summary(covid_model_age)
```

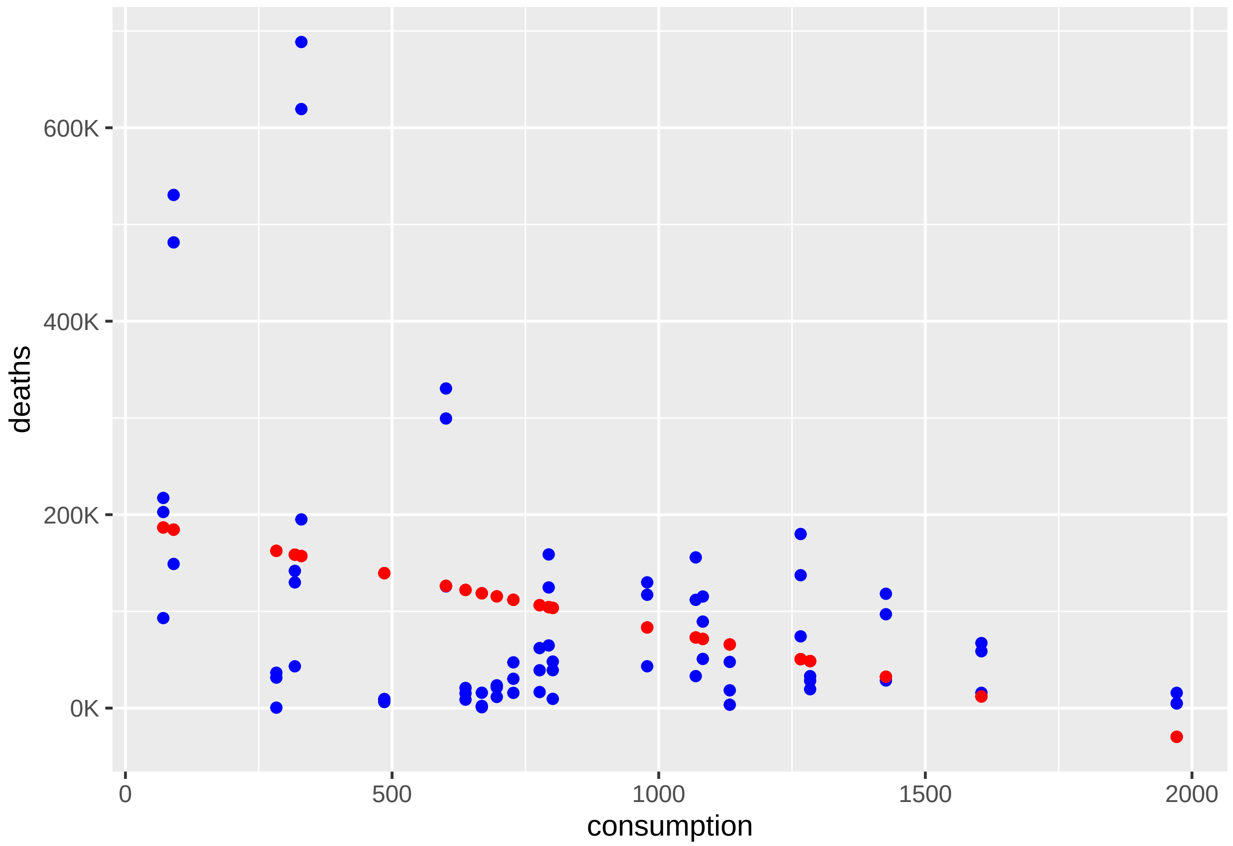
```
##
## Call:
## lm(formula = deaths ~ median_age, data = covid_stats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -162708  -72866  -31953   40166  549404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   402605      94665   4.253 6.69e-05 ***
## median_age     -8124       2508  -3.240  0.00186 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 131500 on 67 degrees of freedom
## Multiple R-squared:  0.1354, Adjusted R-squared:  0.1225
## F-statistic: 10.49 on 1 and 67 DF,  p-value: 0.001865
```

```
global_total_deaths_w_pred <- covid_stats %>% mutate(
  consumption_prediction = 194815.29 -(113.89 * consumption),
  excercise_prediction = 162385 -(10296 * weekly_excercise_hours_mean),
  age_prediction = 402604 -(8124 * median_age)
)
```

Graphing the models

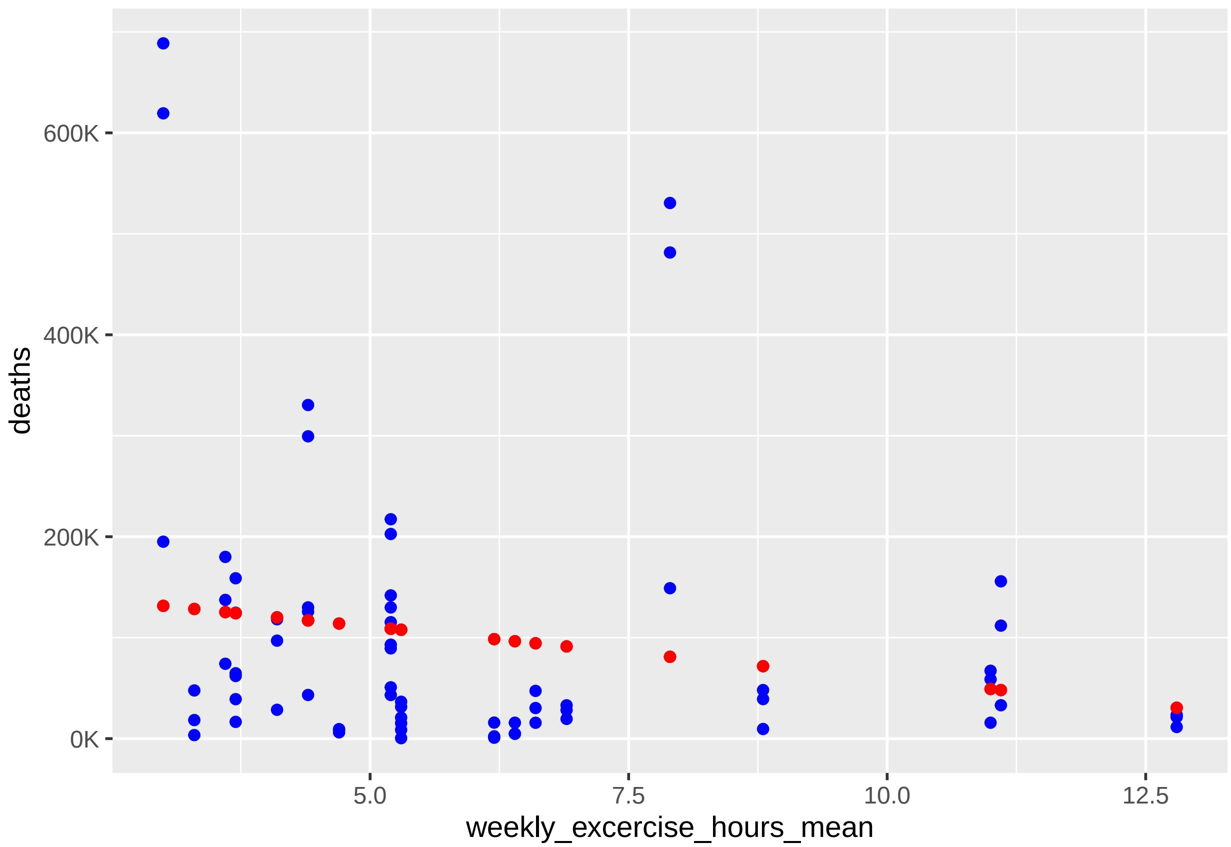
Cigarette Consumption

```
global_total_deaths_w_pred %>% ggplot() + geom_point(
  aes(x=consumption, y=deaths),
  color = "blue"
) + geom_point(
  aes(x=consumption, y=consumption_prediction),
  color = "red"
) + scale_y_continuous(labels = label_number(suffix = "K", scale = 1e-3))
```



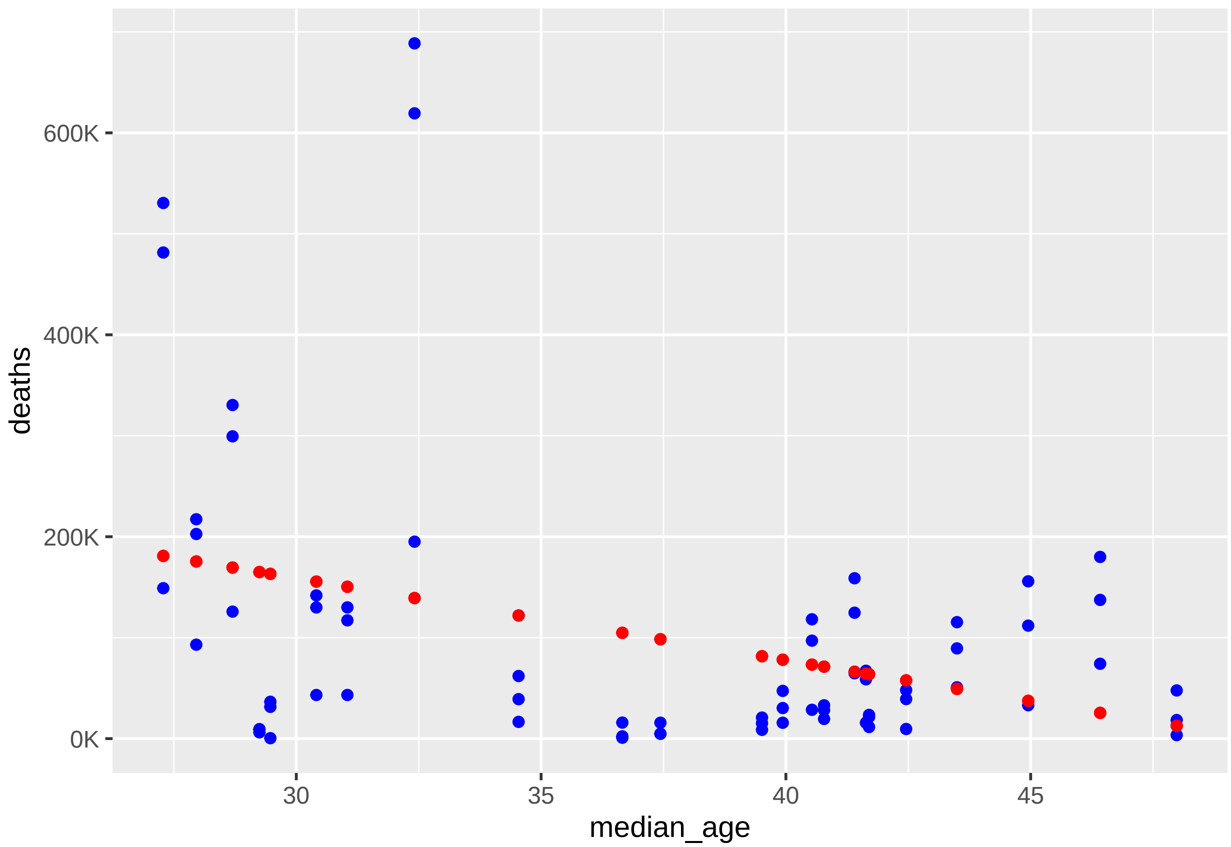
Weekly Excercise

```
global_total_deaths_w_pred %>% ggplot() + geom_point(
  aes(x=weekly_excercise_hours_mean, y=deaths),
  color = "blue"
) + geom_point(
  aes(x=weekly_excercise_hours_mean, y=excercise_prediction),
  color = "red"
) + scale_y_continuous(labels = label_number(suffix = "K", scale = 1e-3))
```



Age

```
global_total_deaths_w_pred %>% ggplot() + geom_point(
  aes(x=median_age, y=deaths),
  color = "blue"
) + geom_point(
  aes(x=median_age, y=age_prediction),
  color = "red"
) + scale_y_continuous(labels = label_number(suffix = "K", scale = 1e-3))
```



Bias & Conclusion

- The cigarette consumption and age are significant but my model had low R-squared values.
- The excercise doesn't look to have a significance and had a low R-squared value.
- There are obviously other very important factors to be considered such as public health governance, and the timeline of infections for each country during the pandemic

Finally, please find the session info below.

```
sessionInfo()
```

```
## R version 4.2.1 (2022-06-23)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
##
## other attached packages:
##  [1] scales_1.2.1     plotly_4.10.1    lubridate_1.9.0  timechange_0.1.1
##  [5] webshot_0.5.4    knitr_1.40       forcats_0.5.2    stringr_1.4.1
##  [9] dplyr_1.0.10     purrr_0.3.5      readr_2.1.3      tidyr_1.2.1
## [13] tibble_3.1.8     ggplot2_3.4.0    tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.4          bit64_4.0.5         vroom_1.6.0
##  [4] jsonlite_1.8.3      viridisLite_0.4.1   modelr_0.1.9
##  [7] assertthat_0.2.1    highr_0.9           googlesheets4_1.0.1
## [10] cellranger_1.1.0    yaml_2.3.6          pillar_1.8.1
## [13] backports_1.4.1     glue_1.6.2          digest_0.6.30
## [16] rvest_1.0.3         colorspace_2.0-3    htmltools_0.5.3
## [19] pkgconfig_2.0.3     broom_1.0.1         haven_2.5.1
## [22] processx_3.8.0      tzdb_0.3.0          googledrive_2.0.0
## [25] farver_2.1.1        generics_0.1.3      ellipsis_0.3.2
## [28] withr_2.5.0         lazyeval_0.2.2      cli_3.4.1
## [31] magrittr_2.0.3      crayon_1.5.2        readxl_1.4.1
## [34] evaluate_0.18       ps_1.7.2            fs_1.5.2
## [37] fansi_1.0.3         xml2_1.3.3          tools_4.2.1
## [40] data.table_1.14.4   hms_1.1.2           gargle_1.2.1
## [43] lifecycle_1.0.3     munsell_0.5.0       reprex_2.0.2
## [46] callr_3.7.3         compiler_4.2.1      rlang_1.0.6
## [49] grid_4.2.1          rstudioapi_0.14     htmlwidgets_1.5.4
## [52] crosstalk_1.2.0     labeling_0.4.2      rmarkdown_2.17
## [55] gtable_0.3.1        DBI_1.1.3           curl_4.3.3
## [58] R6_2.5.1            fastmap_1.1.0       bit_4.0.4
## [61] utf8_1.2.2          stringi_1.7.8       parallel_4.2.1
## [64] vctrs_0.5.0         dbplyr_2.2.1        tidyselect_1.2.0
## [67] xfun_0.34
```