COVID-19 Johns Hopkins Data Analysis

Carlos Barron

2022-11-16

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

```r
paged_table(confirmed_global, options = list(rows.print = 15, cols.print = 10))
```

```r
confirmed_global
```

```
## # A tibble: 603 x 5
## # Groups:   country [201]
##    country     year   cases cases_K cases_M
##    <chr>       <chr>  <dbl>   <dbl>   <dbl>
##  1 Afghanistan 2020   52330   52.3  0.0523
##  2 Afghanistan 2021  158084  158.   0.158
##  3 Afghanistan 2022  204724  205.   0.205
##  4 Albania     2020   58316   58.3  0.0583
##  5 Albania     2021  210224  210.   0.210
##  6 Albania     2022  333197  333.   0.333
##  7 Algeria     2020   99610   99.6  0.0996
##  8 Algeria     2021  218432  218.   0.218
##  9 Algeria     2022  270969  271.   0.271
## 10 Andorra     2020    8049    8.05 0.00805
## # ... with 593 more rows
```

Confirmed Deaths

```
paged_table(deaths_global, options = list(rows.print = 15, cols.print = 10))
```

```
deaths_global
```

```
## # A tibble: 603 x 5
## # Groups:   country [201]
##    country     year  deaths deaths_K deaths_M
##    <chr>       <chr>  <dbl>    <dbl>    <dbl>
##  1 Afghanistan 2020    2189    2.19  0.00219
##  2 Afghanistan 2021    7356    7.36  0.00736
##  3 Afghanistan 2022    7829    7.83  0.00783
##  4 Albania     2020    1181    1.18  0.00118
##  5 Albania     2021    3217    3.22  0.00322
##  6 Albania     2022    3594    3.59  0.00359
##  7 Algeria     2020    2756    2.76  0.00276
##  8 Algeria     2021    6276    6.28  0.00628
##  9 Algeria     2022    6881    6.88  0.00688
## 10 Andorra     2020      84    0.084 0.000084
## # ... with 593 more rows
```

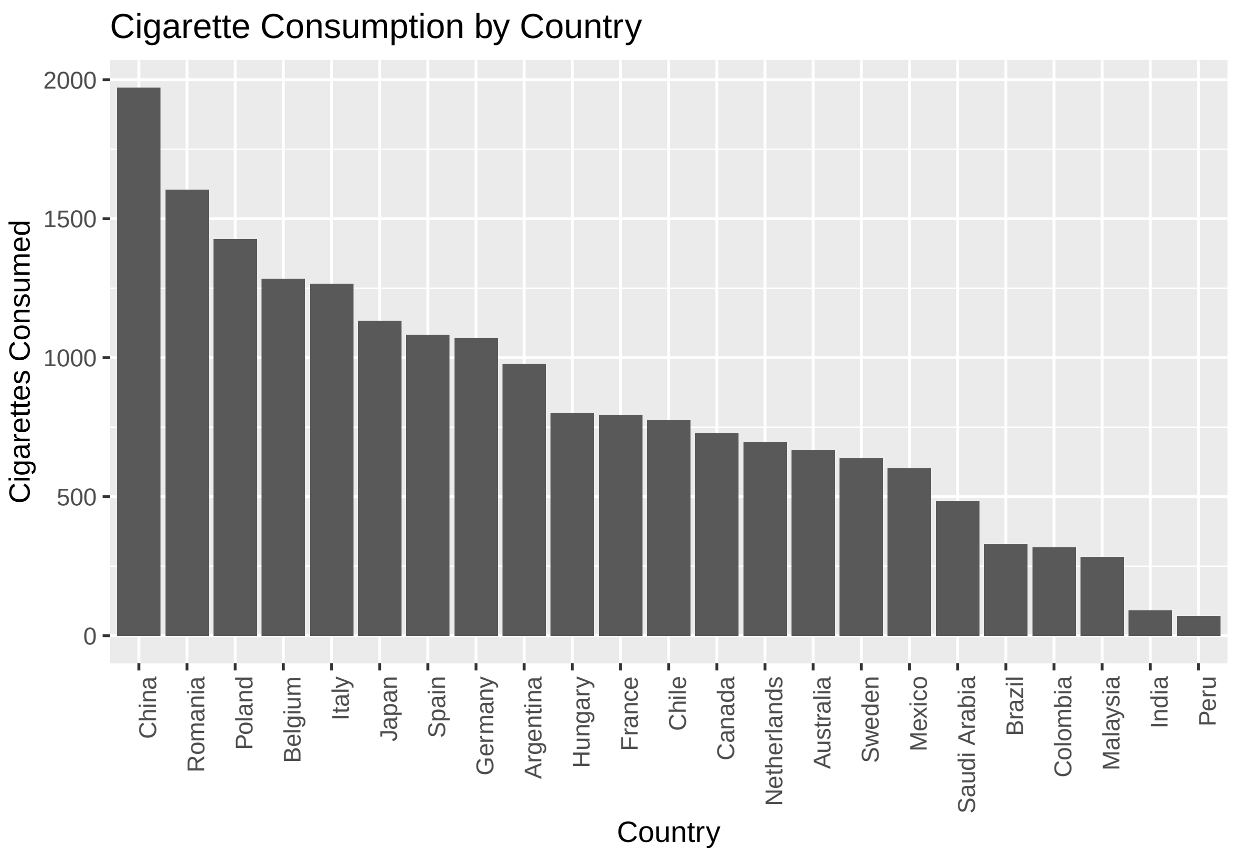
Tobacco Atlas

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_chart
```



Ipsos Global Advisor

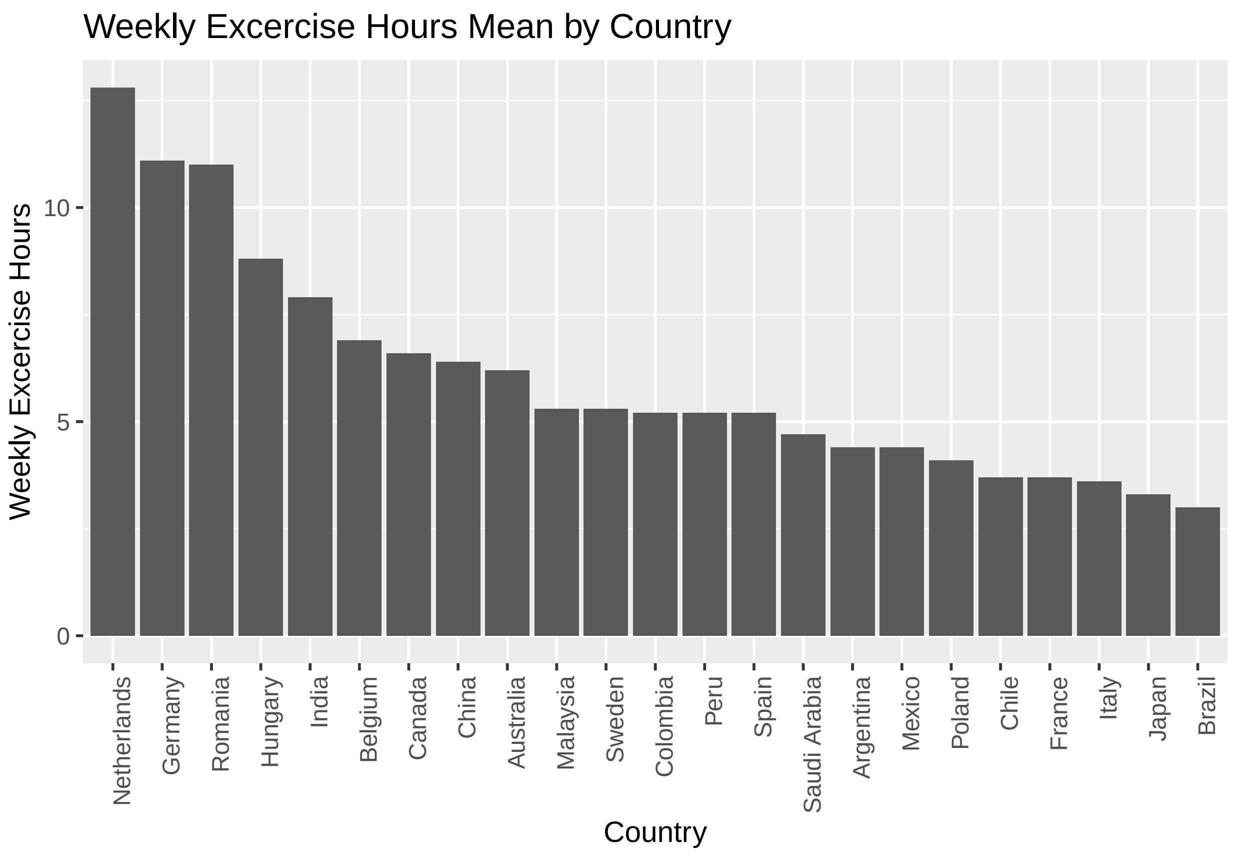
Global Views on Exercise and Team Sports

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_chart
```



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_chart
```

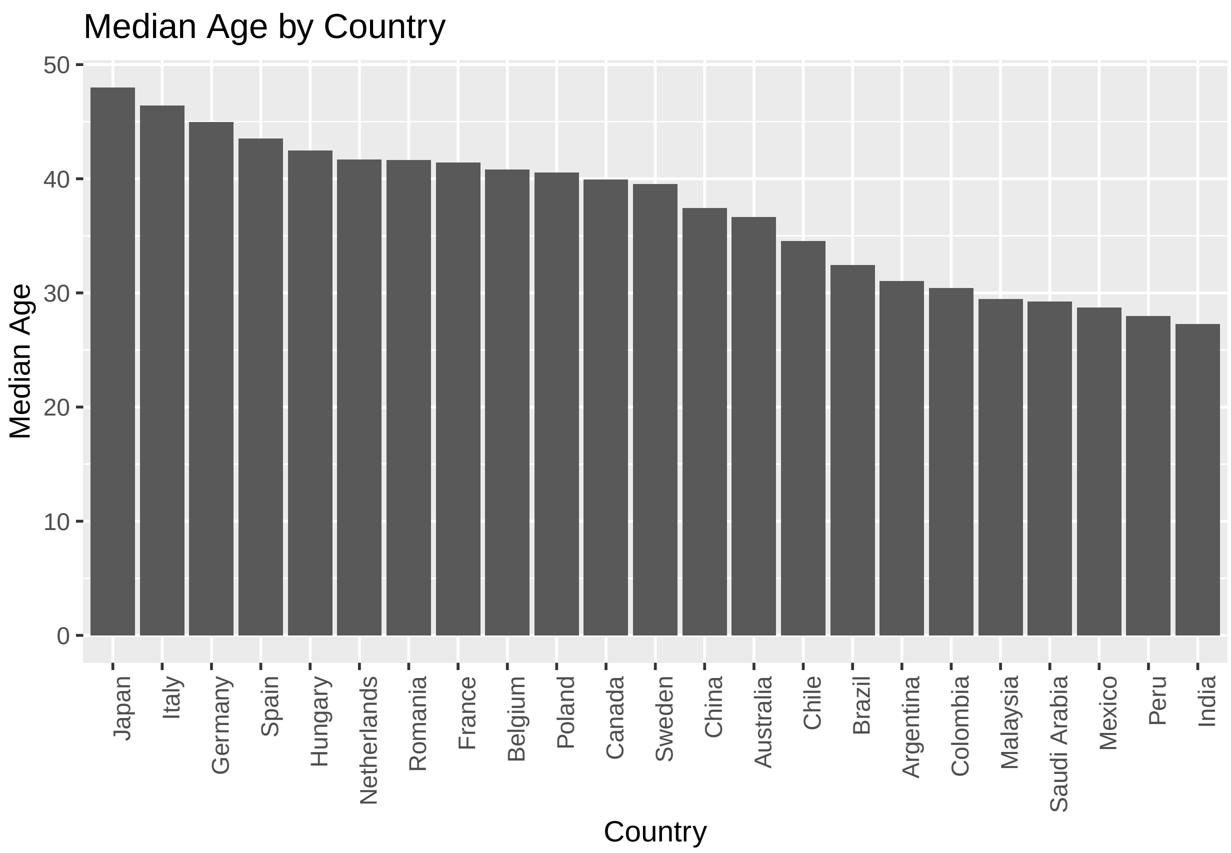


Table Joins

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

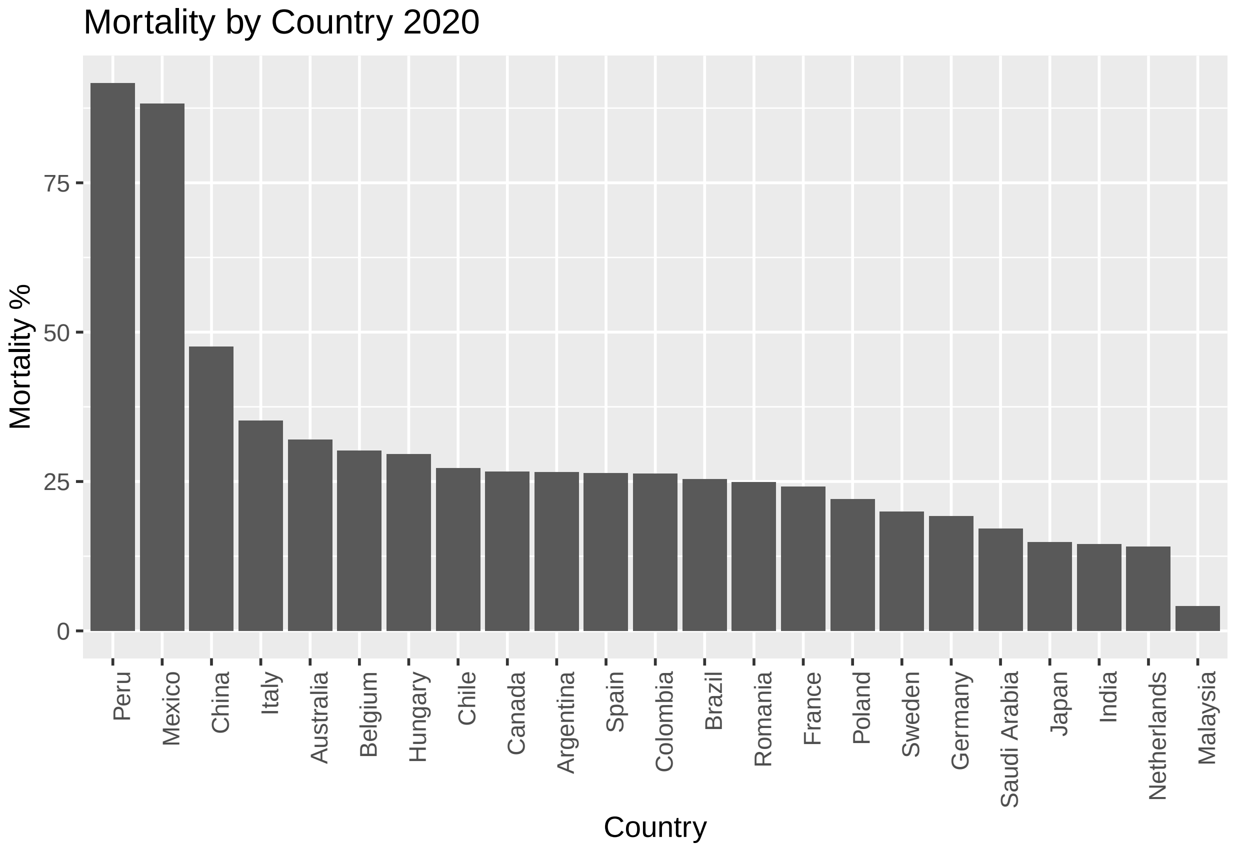
```
paged_table(covid_stats, options = list(rows.print = 15, cols.print = 10))
```

```
paged_table(covid_stats, options = list(rows.print = 15, cols.print = 10))
```

Mortality

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

```
mortality_chart
```

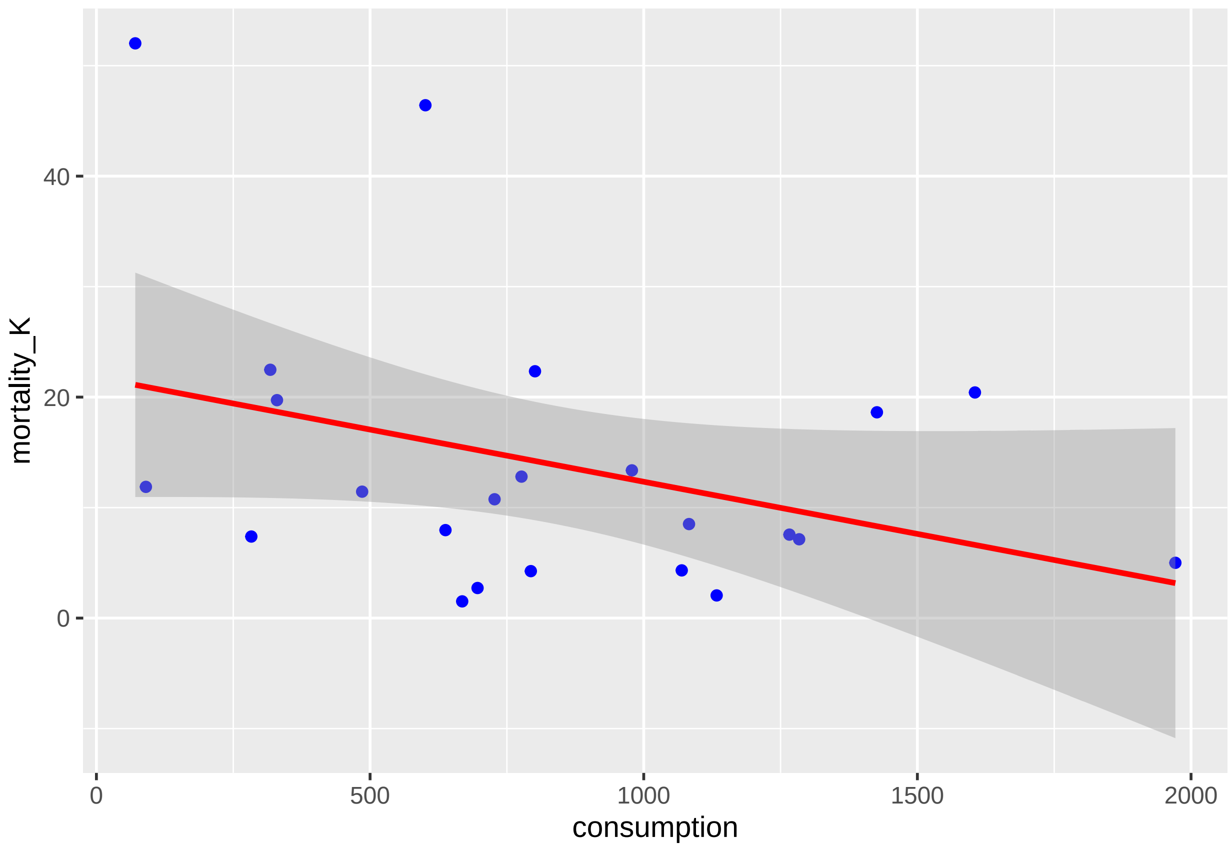


Models & Analysis

Let's start by making an analysis on how much mortality and cigarette consumption is related.

```
consumption_analysis_load
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



This is the model summary for cigarette consumption

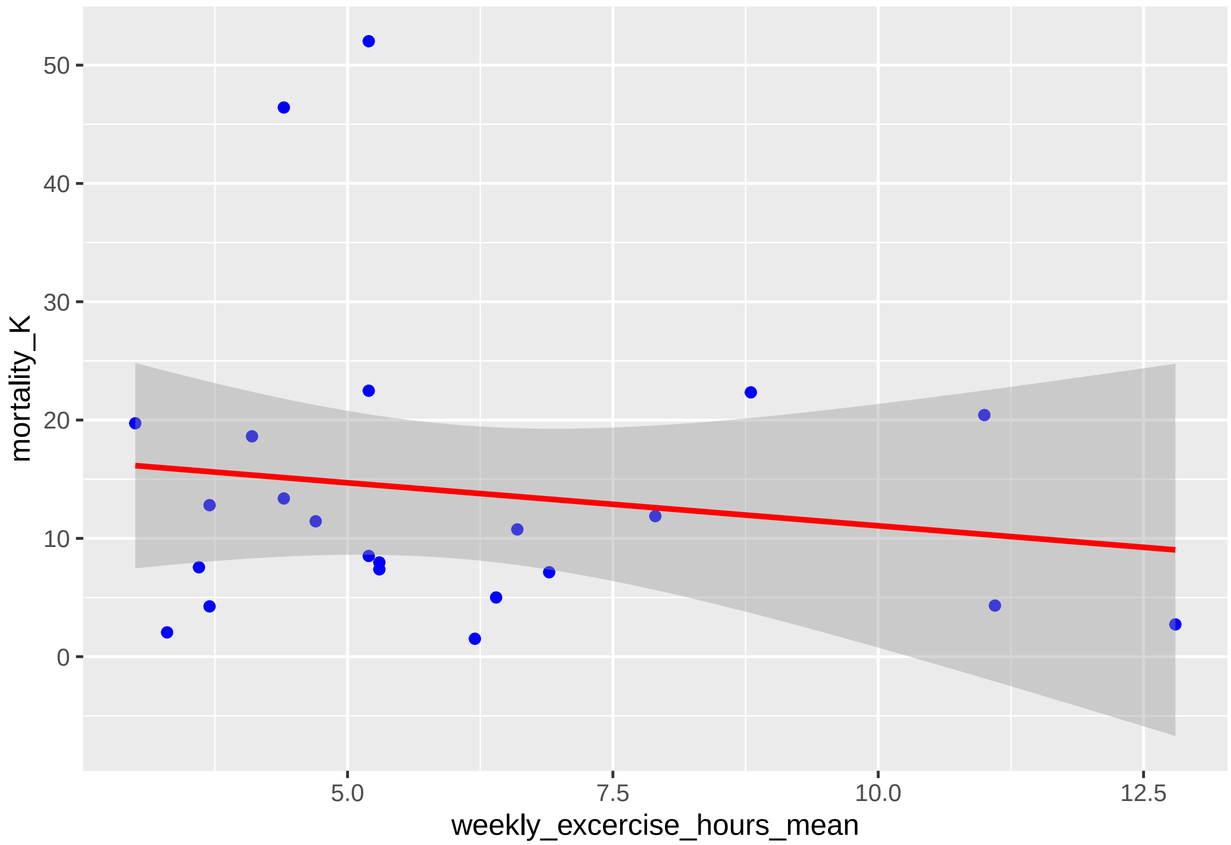
```
summary(covid_model_consumption)
```

```
##
## Call:
## lm(formula = mortality_K ~ consumption, data = covid_stats_2022)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.959  -8.413  -2.523   2.765  30.905
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.784773   5.217820   4.175 0.000428 ***
## consumption -0.009444   0.005467  -1.727 0.098791 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.34 on 21 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.08271
## F-statistic: 2.984 on 1 and 21 DF,  p-value: 0.09879
```

Now let's see how mortality and physical exercise is related.

```
excercise_chart_load
```

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



This is the model summary for weekly excercise mean

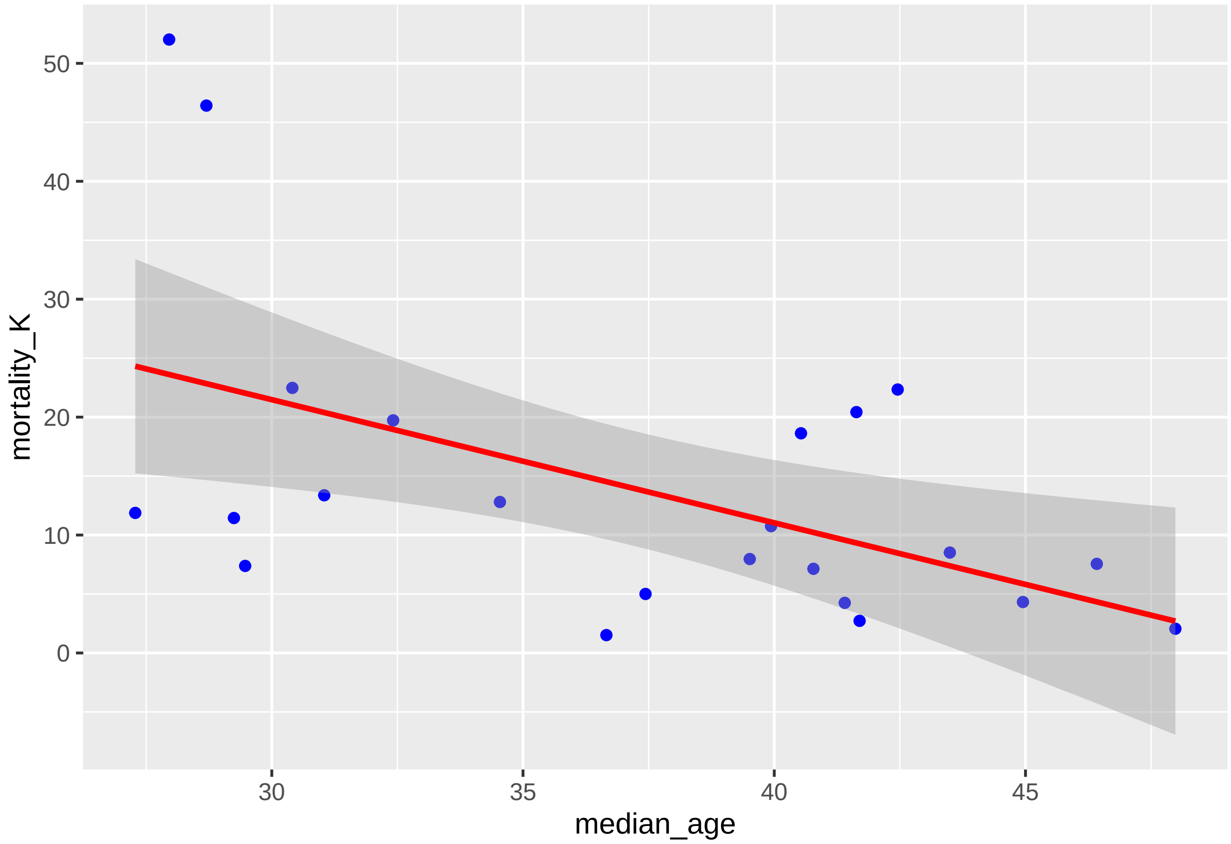
```
summary(covid_model_excercise)
```

```
##
## Call:
## lm(formula = mortality_K ~ weekly_excercise_hours_mean, data = covid_stats_2022)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.877  -6.804  -3.470   3.428  37.470
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  18.3289     6.8602   2.672   0.0143 *
## weekly_excercise_hours_mean  -0.7267     1.0437  -0.696   0.4939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.04 on 21 degrees of freedom
## Multiple R-squared:  0.02256,    Adjusted R-squared:  -0.02398
## F-statistic: 0.4847 on 1 and 21 DF,  p-value: 0.4939
```

Finally let's see how mortality and age is related.

```
age_chart_load
```

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



And this is the model summary for the median age

```
summary(covid_model_age)
```

```
##
## Call:
## lm(formula = mortality_K ~ median_age, data = covid_stats_2022)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.651  -6.777  -1.547   2.324  28.407
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  52.8103    14.0127   3.769  0.00113 **
## median_age   -1.0444     0.3712  -2.813  0.01041 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.24 on 21 degrees of freedom
## Multiple R-squared:  0.2737, Adjusted R-squared:  0.2391
## F-statistic: 7.915 on 1 and 21 DF,  p-value: 0.01041
```

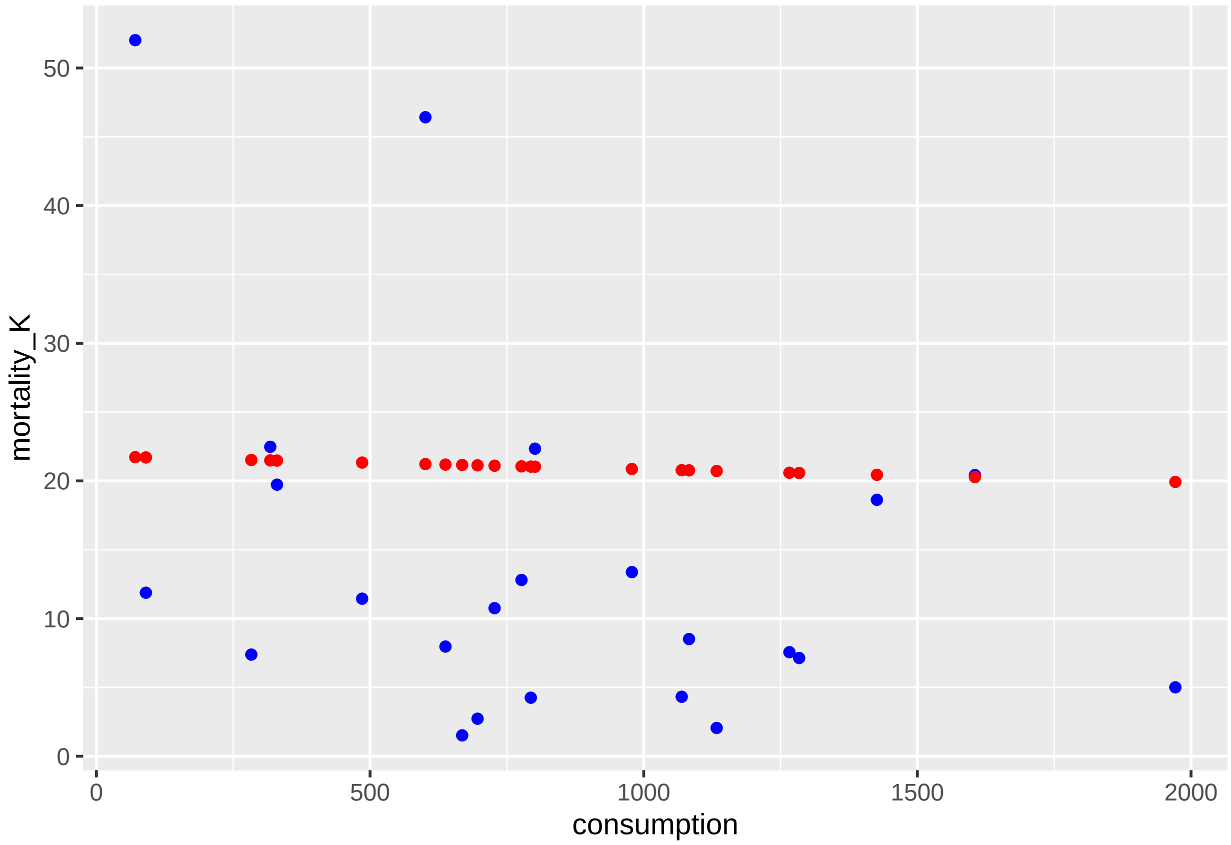
Let's build the prediction into our dataset.

```
global_total_deaths_w_pred <- covid_stats_2022 %>% mutate(
  consumption_prediction = 21.789989 -(0.0009448 * consumption),
  excercise_prediction = 18.3313 -(0.7268 * weekly_excercise_hours_mean),
  age_prediction = 52.8234 -(1.0447 * median_age)
)
```

Graphing the models

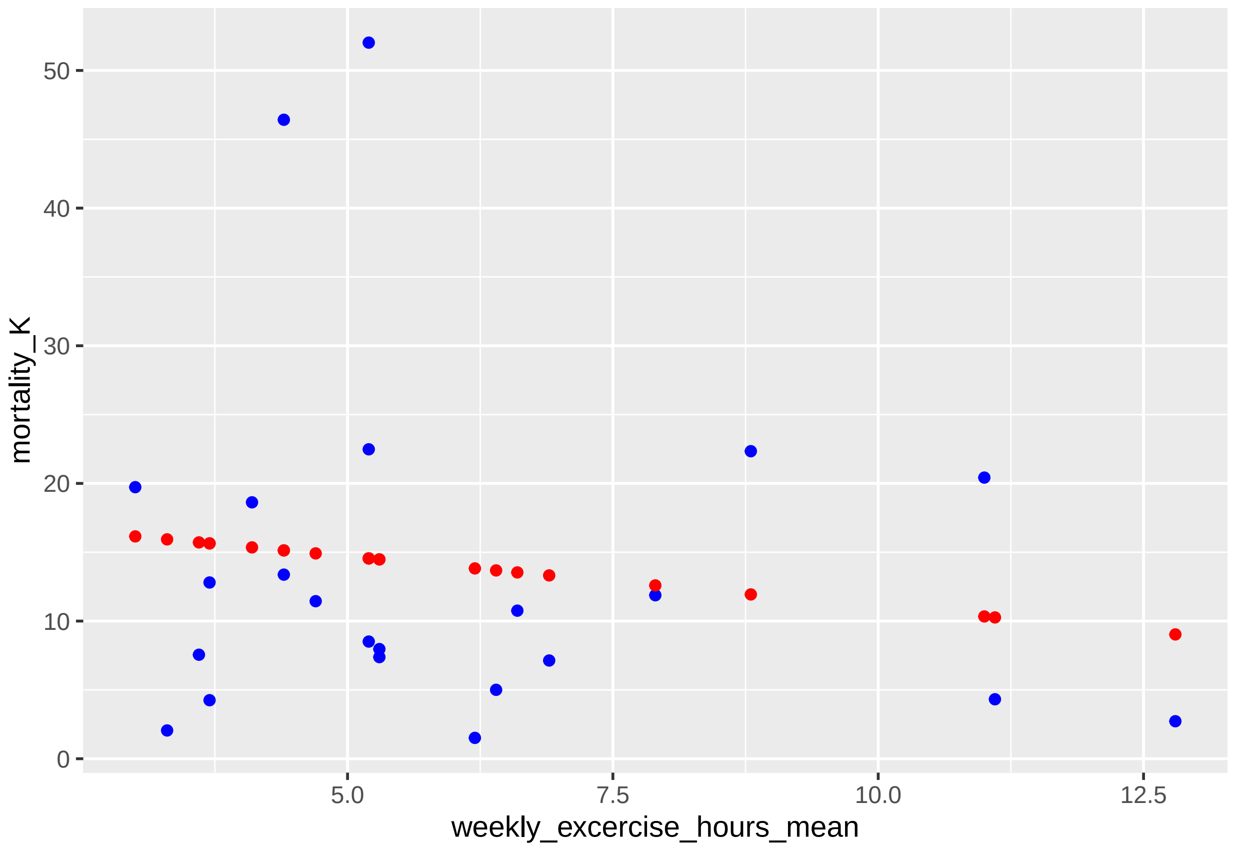
Cigarette Consumption

```
cig_model_chart
```



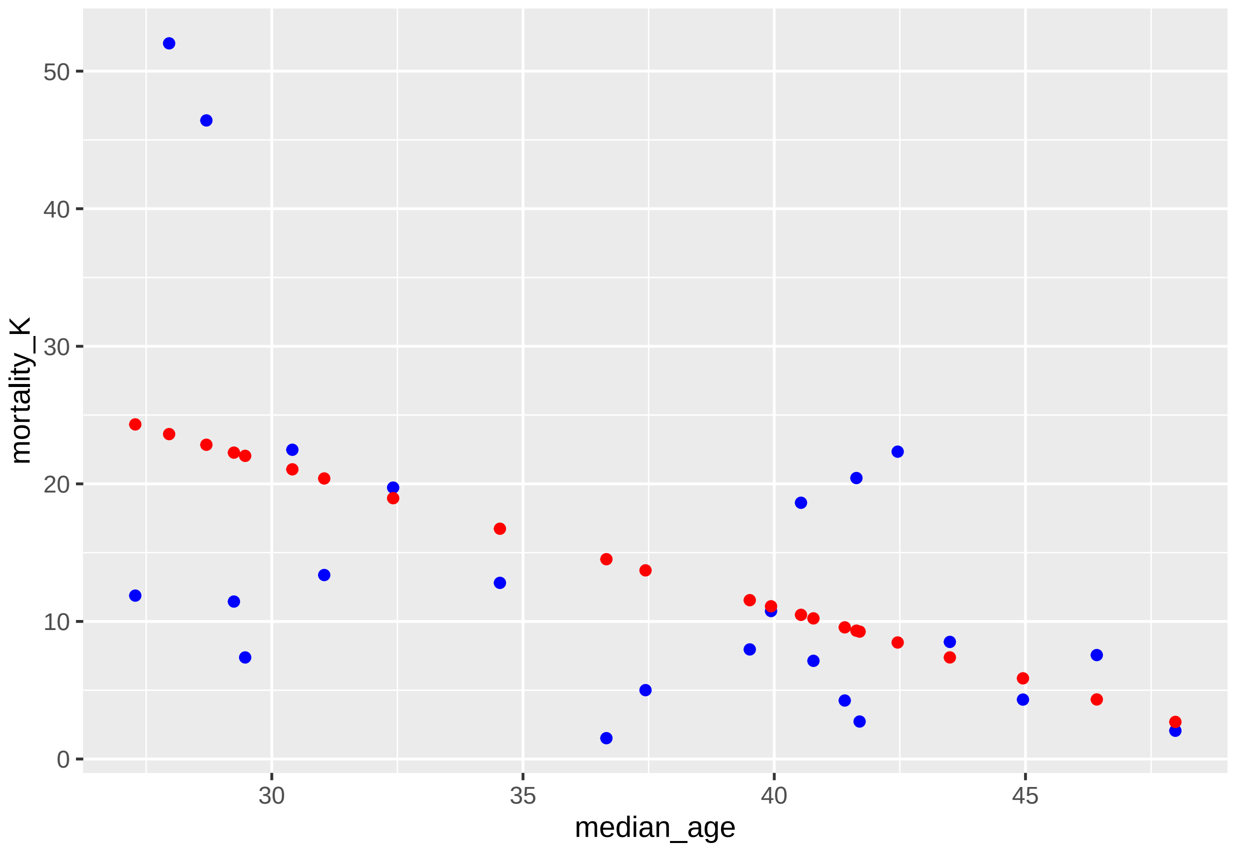
Weekly Excercise

```
excercise_model_chart
```



Age

```
age_model_chart
```



Bias & Conclusion

- After these iterations, the models are not statistically significant and had low R-squared values.
- For next iterations I would replace some of the data obtained (exercise and cigarette consumption)
- Cigarette consumption by adult might not be a great fit since it does not give a clear indication of the percentage of smokers by country.
- Getting information for the percentage of population who excercises might be tricky.
- An important factor to take into account is the timeline on which the vaccines started to roll out by country, this could have introduced noise in my implementation.
- From the list of countries in the implementation, all had different isolation strategies that should also be considered.
- The quality and saturation of the public health care system is another important factor to take into account.

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] rmarkdown_2.17   scales_1.2.1     lubridate_1.9.0  timechange_0.1.1
##  [5] knitr_1.40       webshot_0.5.4    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] lattice_0.20-45     assertthat_0.2.1    digest_0.6.30
##  [4] utf8_1.2.2          R6_2.5.1            cellranger_1.1.0
##  [7] backports_1.4.1     reprex_2.0.2        evaluate_0.18
## [10] highr_0.9           httr_1.4.4          pillar_1.8.1
## [13] rlang_1.0.6         googlesheets4_1.0.1 curl_4.3.3
## [16] readxl_1.4.1        rstudioapi_0.14     Matrix_1.4-1
## [19] splines_4.2.1       labeling_0.4.2      googledrive_2.0.0
## [22] bit_4.0.4           munsell_0.5.0       broom_1.0.1
## [25] compiler_4.2.1      modelr_0.1.9        xfun_0.34
## [28] pkgconfig_2.0.3     mgcv_1.8-40         htmltools_0.5.3
## [31] tidyselect_1.2.0    fansi_1.0.3         crayon_1.5.2
## [34] tzdb_0.3.0          dbplyr_2.2.1        withr_2.5.0
## [37] grid_4.2.1          nlme_3.1-157        jsonlite_1.8.3
## [40] gtable_0.3.1        lifecycle_1.0.3     DBI_1.1.3
## [43] magrittr_2.0.3      cli_3.4.1           stringi_1.7.8
## [46] vroom_1.6.0         farver_2.1.1        fs_1.5.2
## [49] xml2_1.3.3          ellipsis_0.3.2      generics_0.1.3
## [52] vctrs_0.5.0         tools_4.2.1         bit64_4.0.5
## [55] glue_1.6.2          hms_1.1.2           parallel_4.2.1
## [58] fastmap_1.1.0       yaml_2.3.6          colorspace_2.0-3
## [61] gargle_1.2.1        rvest_1.0.3         haven_2.5.1
```