The results below are generated from an R script.

```
# Analysis
# Scott Cohn + Ruja Kambli

# Libraries ------------------------------------------------------------------

library(dplyr)

##
## Attaching package:  'dplyr'
## The following objects are masked from 'package:stats':
##
##    filter, lag
## The following objects are masked from 'package:base':
##
##    intersect, setdiff, setequal, union

library(tidyverse) # duh.

## - Attaching packages ---------------------------- tidyverse 1.2.1 -
## v ggplot2 3.2.1    v readr  1.3.1
## v tibble  2.1.3    v purrr  0.3.3
## v tidyr   1.0.0    v stringr 1.4.0
## v ggplot2 3.2.1    v forcats 0.4.0
## - Conflicts ------------------------------ tidyverse_conflicts() -
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(ggplot2) # plotting
library(gridExtra) # plotting options

##
## Attaching package:  'gridExtra'
## The following object is masked from 'package:dplyr':
##
##    combine

library(ggsci)  # plot color palette
library(ggthemes) # Themes
library(bbplot) # plot style
library(readr) # import csv
library(lmtest) # BP test

## Loading required package:  zoo
##
## Attaching package:  'zoo'
## The following objects are masked from 'package:base':
##
##    as.Date, as.Date.numeric

library(scales) # Scale x-axis

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

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(MASS)
```

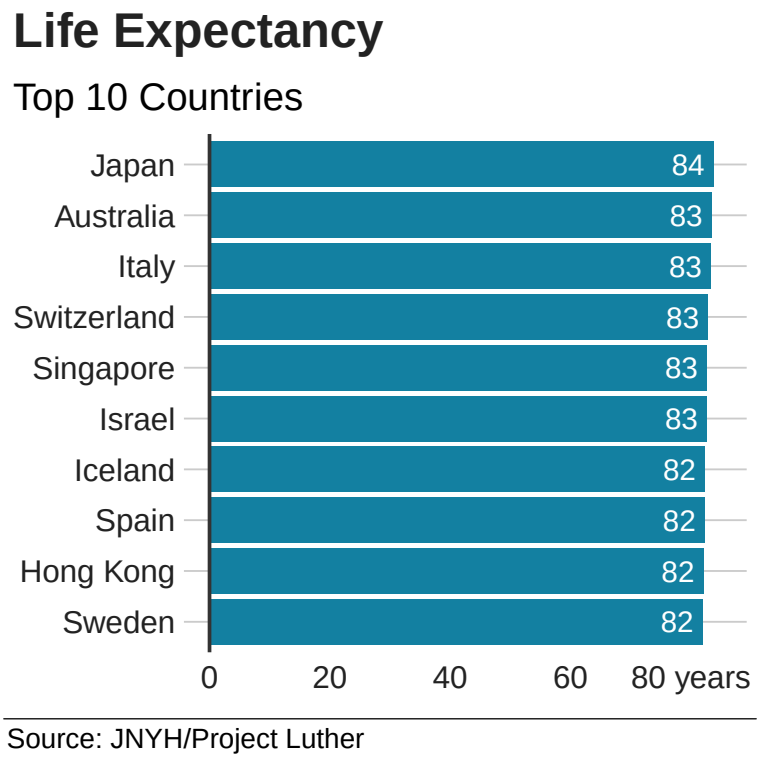
```
##
## Attaching package:  'MASS'
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(faraway) # Box-Cox transform / vif

# Colors
COLA <- c("#99d8c9","#66c2a4","#41ae76", "#238b45", "#005824")
COLB <- c("#4eb3d3", "#2b8cbe", "#0868ac","#084081")

# Import Data -------------------------------------------------------------
life_exp_full <- read_csv("data/life_exp_full.csv")
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   `Birth Rate` = col_double(),
##   `Cancer Rate` = col_double(),
##   `Dengue Cases` = col_double(),
##   EPI = col_double(),
##   GDP = col_double(),
##   `Health Expenditure` = col_double(),
##   `Heart Disease Rate` = col_double(),
##   Population = col_double(),
##   Area = col_double(),
##   `Pop Density` = col_double(),
##   `Stroke Rate` = col_double(),
##   `Life Expectancy` = col_double()
## )
```

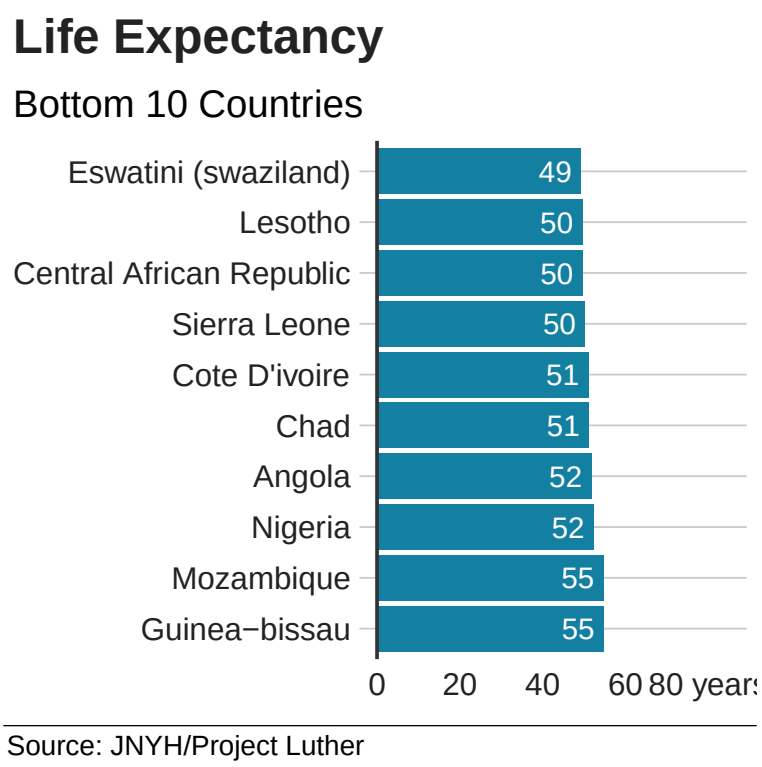
```
# Data Transformations ----------------------------------------------------

# Capitalize letters in Country var
# Not perfect, but good enough
simpleCap <- function(x) {
  s <- strsplit(x, " ")[[1]]
  paste(toupper(substring(s, 1,1)), substring(s, 2),
        sep = "", collapse = " ")
}

life_exp_full <- life_exp_full %>%
  mutate(Country = apply(life_exp_full, 1, simpleCap))

# Visualizations ----------------------------------------------------------

# Top 10 life exp by country
topten_lifeexp_country <- life_exp_full %>%
  arrange(desc(`Life Expectancy`)) %>%
```

```
  slice(1:10) %>%
  ggplot(aes(x = reorder(Country, `Life Expectancy`),
             y = `Life Expectancy`)) +
  geom_bar(stat = 'identity',
           fill = "#1380A1") +
  #scale_fill_d3() +
  coord_flip() +
  scale_y_continuous(
    limits = c(0, 85),
    breaks = seq(0, 80, by = 20),
    labels = c("0", "20", "40", "60", "80 years")
  ) +
  geom_hline(yintercept = 0,
             size = 1,
             color = "#333333") +
  geom_label(
    aes(label = round(`Life Expectancy`, 0)),
    hjust = 1,
    vjust = 0.5,
    colour = "white",
    fill = NA,
    label.size = NA,
    family = "Helvetica",
    size = 6
  ) +
  bbc_style() +
  labs(title = "Life Expectancy",
       subtitle = "Top 10 Countries")

# Save graph
finalise_plot(plot_name = topten_lifeexp_country,
              source = "Source: JNYH/Project Luther",
              save_filepath = "figures/topten_lifeexp_country.pdf",
              width_pixels = 640,
              height_pixels = 450)
```



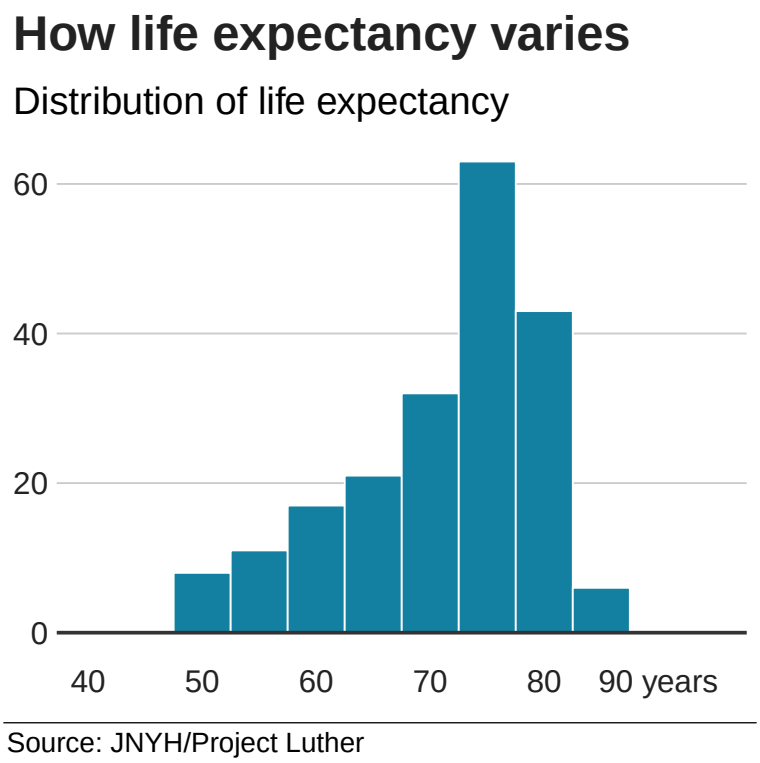

```
              #logo_image_path = "placeholder.png")

# Bottom 10 life exp by country
# life_exp_full %>% drop_na(`Life Expectancy`) %>% nrow() = 201 rows w/out NA
bottomten_lifeexp_country <- life_exp_full %>%
  drop_na(`Life Expectancy`) %>%
  arrange(desc(`Life Expectancy`)) %>%
  slice(192:201) %>%
  ggplot(aes(x = reorder(Country, -`Life Expectancy`),
             y = `Life Expectancy`)) +
  geom_bar(stat = 'identity',
           fill = "#1380A1") +
  #scale_fill_d3() +
  coord_flip() +
  scale_y_continuous(
    limits = c(0, 85),
    breaks = seq(0, 80, by = 20),
    labels = c("0", "20", "40", "60", "80 years")
  ) +
  geom_hline(yintercept = 0,
             size = 1,
             color = "#333333") +
  geom_label(
    aes(label = round(`Life Expectancy`, 0)),
    hjust = 1,
    vjust = 0.5,
    colour = "white",
    fill = NA,
    label.size = NA,
```

```
    family = "Helvetica",
    size = 6
  ) +
  bbc_style() +
  labs(title = "Life Expectancy",
       subtitle = "Bottom 10 Countries")

# Save graph
finalise_plot(plot_name = bottomten_lifeexp_country,
              source = "Source: JNYH/Project Luther",
              save_filepath = "figures/bottomten_lifeexp_country.pdf",
              width_pixels = 640,
              height_pixels = 450)
```



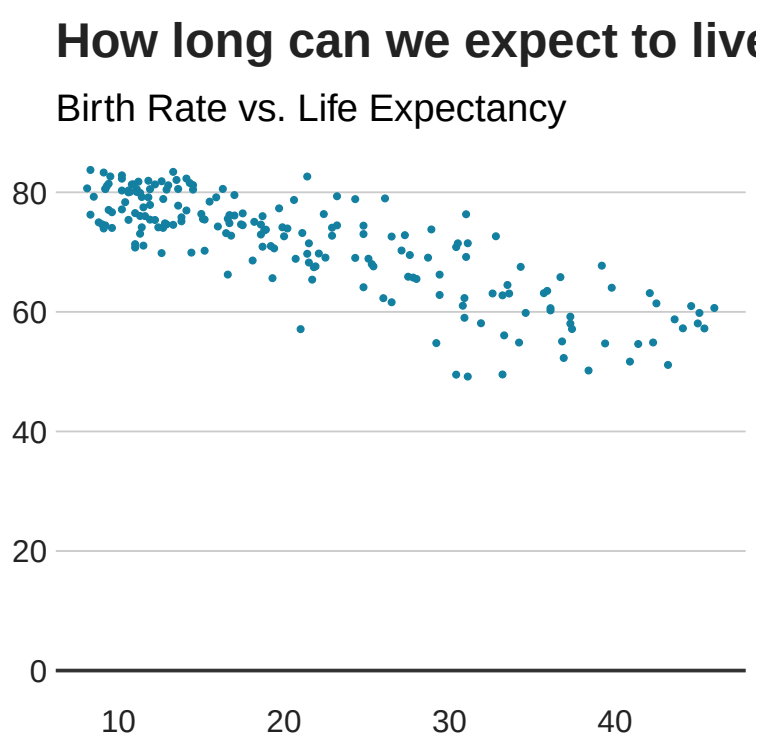

```
#logo_image_path = "placeholder.png")

# Distribution of Life Expectancy, Histogram
lifeexp_distro <- life_exp_full %>%
  ggplot(aes(x = `Life Expectancy`)) +
  geom_histogram(binwidth = 5,
                 color = "white",
                 fill = "#1380A1") +
  geom_hline(yintercept = 0,
             size = 1,
             color = "#333333") +
  bbc_style() +
  scale_x_continuous(
    limits = c(40, 95),
    breaks = seq(40, 90, by = 10),
```

```
    labels = c("40", "50", "60", "70", "80", "90 years")
  ) +
  labs(title = "How life expectancy varies",
       subtitle = "Distribution of life expectancy")

# Save graph
finalise_plot(plot_name = lifeexp_distro,
              source = "Source: JNYH/Project Luther",
              save_filepath = "figures/lifeexp_distro.pdf",
              width_pixels = 640,
              height_pixels = 450)

## Warning:  Removed 47 rows containing non-finite values (stat_bin).
## Warning:  Removed 2 rows containing missing values (geom_bar).
```

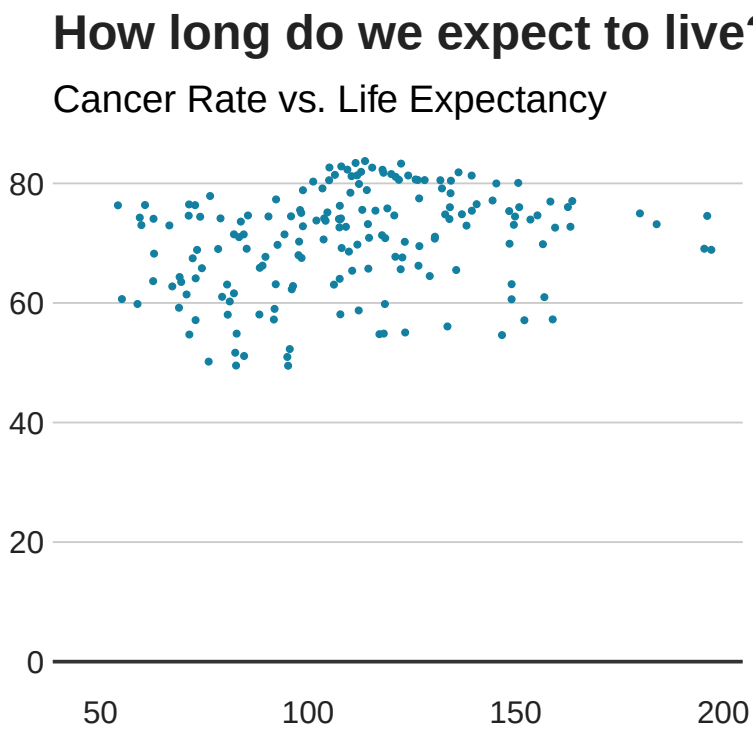



```
#logo_image_path = "placeholder.png")

# Life exp vs Birth Rate
life_exp_full %>%
  ggplot(aes(x = `Birth Rate`,
             y = `Life Expectancy`)) +
             xlab("Birth Rate per 1000 People") +       #RK I tried labelling these axes multiple times
             ylab("Life Expectancy") +
  geom_point(color = "#1380A1") +
  geom_hline(yintercept = 0,
             size = 1,
             color = "#333333") +
  #scale_color_d3() +
  bbc_style() +
```

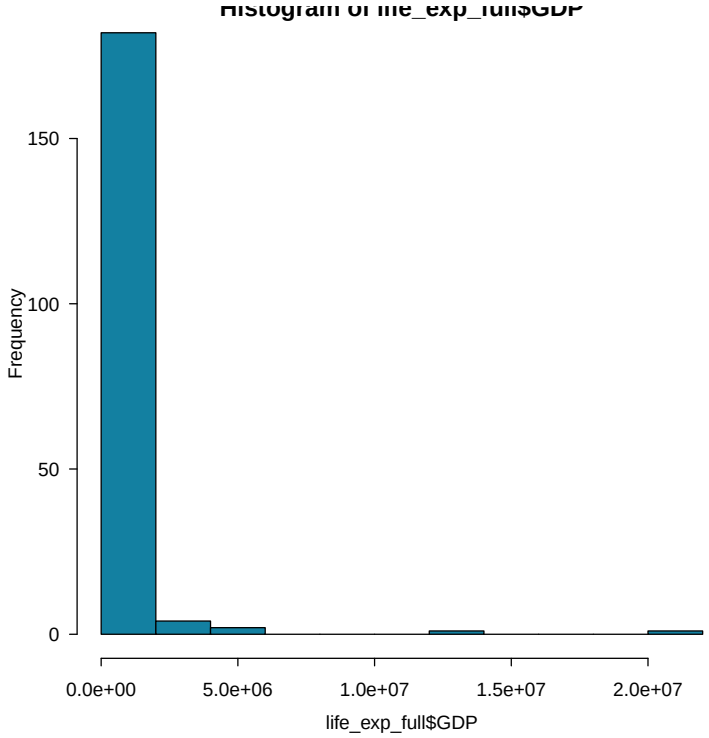
```
  labs(title = "How long can we expect to live?",
       subtitle = "Birth Rate vs. Life Expectancy",
       ylab = "Life Expectancy",
       xlab = "Births Per 1000 People")

## Warning:  Removed 54 rows containing missing values (geom_point).
```



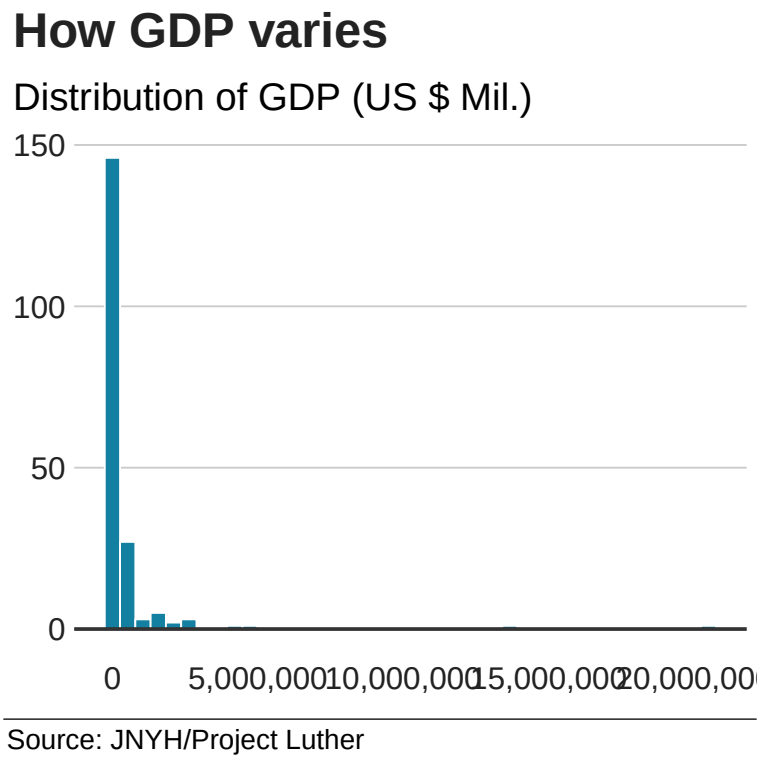

```
# Life vs Cancer
life_exp_full %>%
  ggplot(aes(x = `Cancer Rate`,
             y = `Life Expectancy`)) +
  geom_point(color = "#1380A1") +
  geom_hline(yintercept = 0,
             size = 1,
             colour = "#333333") +
  #scale_color_d3() +
  bbc_style() +
  labs(title = "How long do we expect to live?",
       subtitle = "Cancer Rate vs. Life Expectancy")

## Warning:  Removed 70 rows containing missing values (geom_point).
```




```
# GDP Distribution, Histogram
hist(life_exp_full$GDP, col = "#1380A1")
```



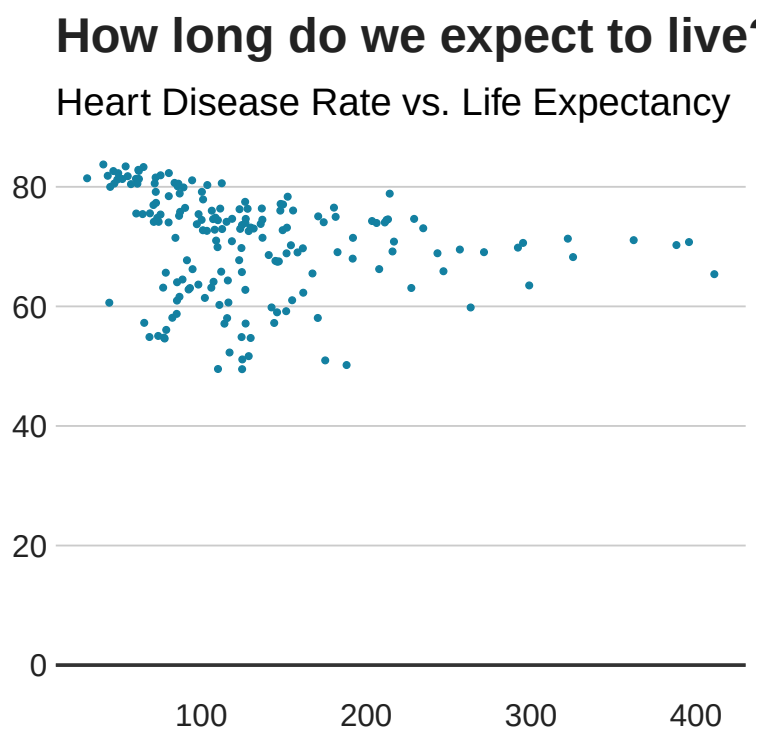
```r
GDP_distro <- life_exp_full %>%
  ggplot(aes(x = GDP)) +
  geom_histogram(
    color = "white",
    fill = "#1380A1",
    na.rm = TRUE,
    bins = 40
  ) +
  geom_hline(yintercept = 0,
             size = 1,
             color = "#333333") +
  bbc_style() +
  scale_x_continuous(labels = scales::comma) +
  labs(title = "How GDP varies",
       subtitle = "Distribution of GDP (US $ Mil.)")

# Save graph
finalise_plot(plot_name = GDP_distro,
              source = "Source: JNYH/Project Luther",
              save_filepath = "figures/GDP_distro.pdf",
              width_pixels = 640,
              height_pixels = 450)
```




```r
#logo_image_path = "placeholder.png")

# Life vs Heart Disease
life_exp_full %>%
  ggplot(aes(x = `Heart Disease Rate`,
             y = `Life Expectancy`)) +
```

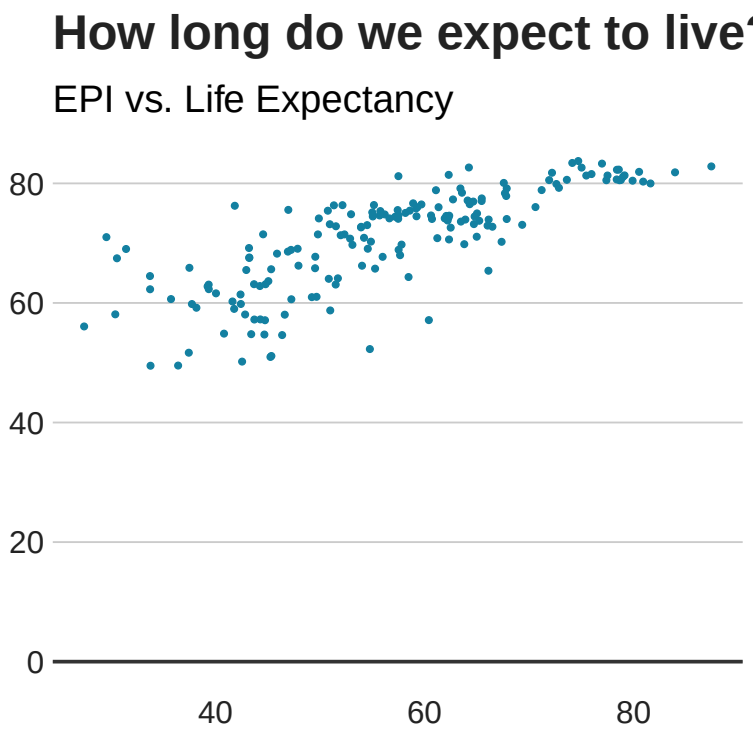
```
  geom_point(color = "#1380A1") +
  geom_hline(yintercept = 0,
             size = 1,
             colour = "#333333") +
  #scale_color_d3() +
  bbc_style() +
  labs(title = "How long do we expect to live?",
       subtitle = "Heart Disease Rate vs. Life Expectancy")
```

```
## Warning:  Removed 70 rows containing missing values (geom_point).
```

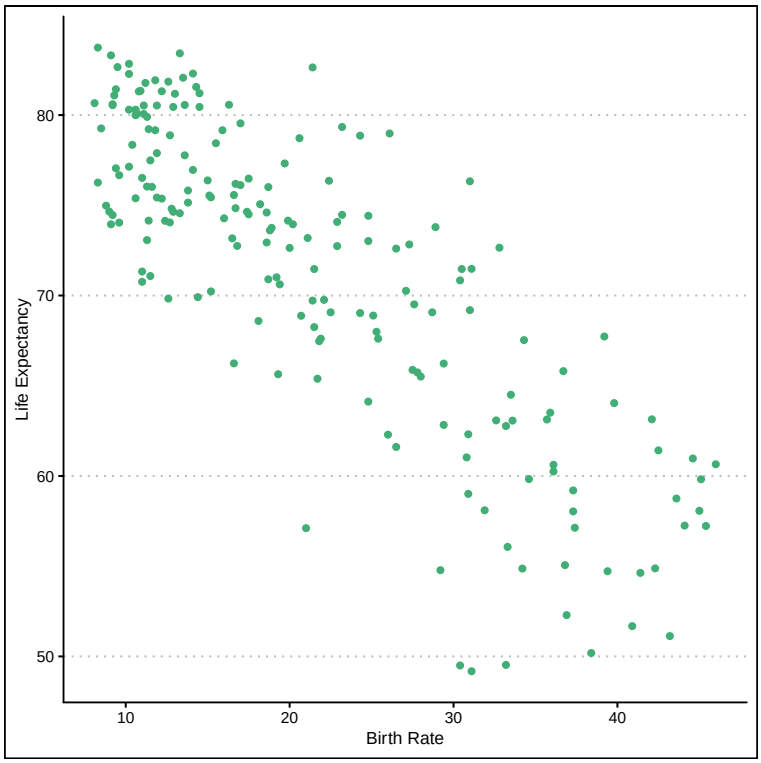



```
# Life vs EPI
life_exp_full %>%
  ggplot(aes(x = EPI,
             y = `Life Expectancy`)) +
  geom_point(color = "#1380A1") +
  geom_hline(yintercept = 0,
             size = 1,
             colour = "#333333") +
  #scale_color_d3() +
  bbc_style() +
  labs(title = "How long do we expect to live?",
       subtitle = "EPI vs. Life Expectancy")
```

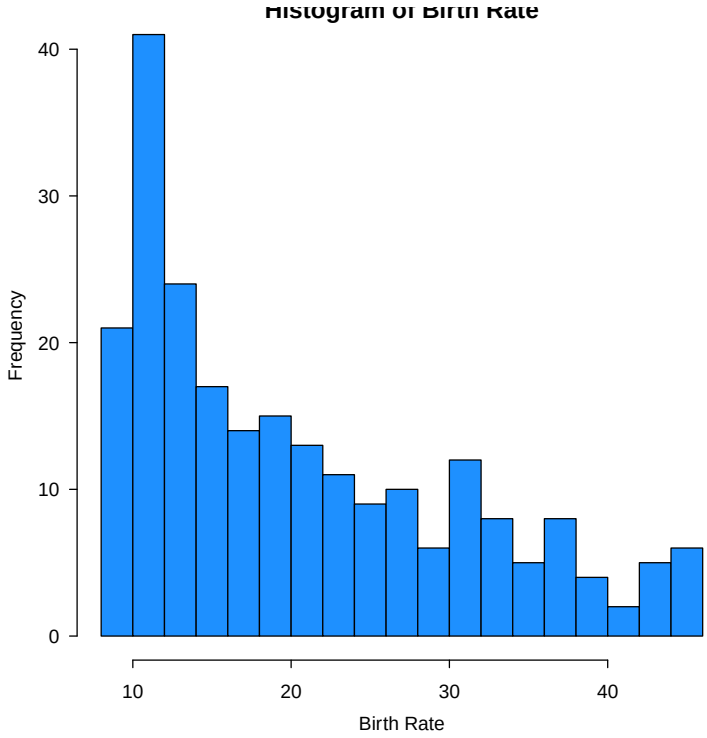
```
## Warning:  Removed 73 rows containing missing values (geom_point).
```



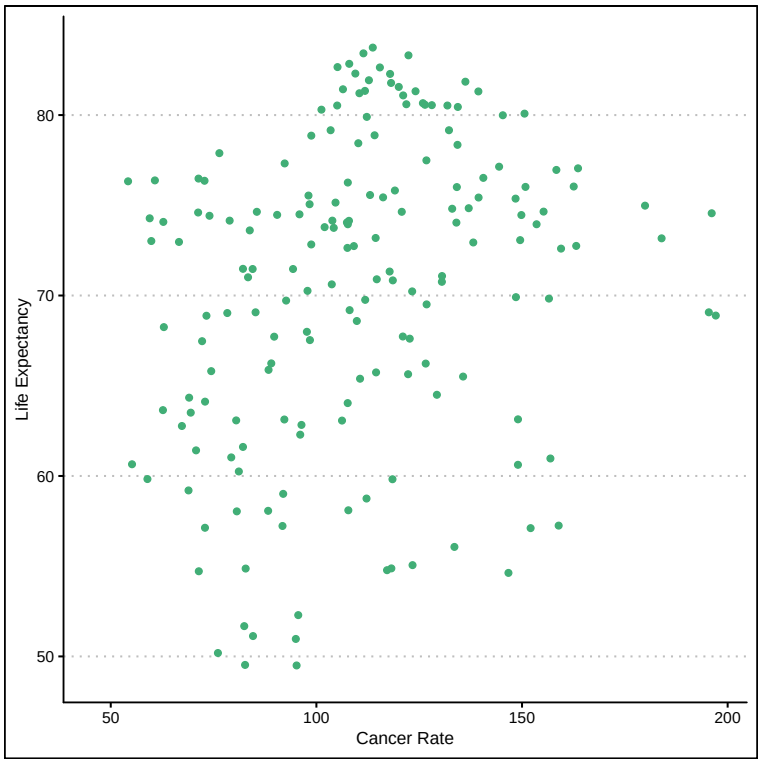

```
# Plot Variables v Life Expectancy ---------------------------------------

# Birth
life_exp_full %>%
  ggplot() +
  geom_point(
    aes(y = `Life Expectancy`, x = `Birth Rate`),
    color = COLA[3]) +
  theme_clean()

## Warning:  Removed 54 rows containing missing values (geom_point).
```



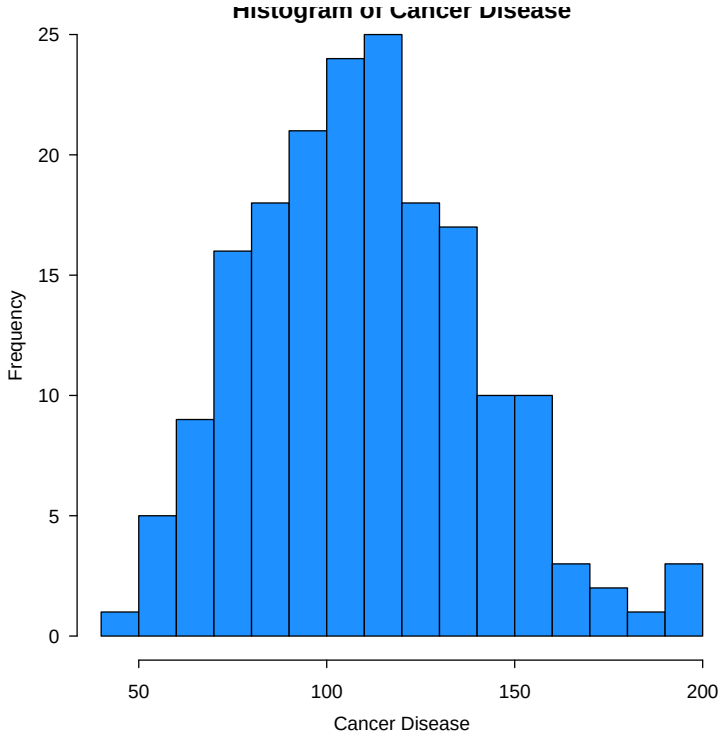

```
hist(life_exp_full$`Birth Rate`,
     xlab   = "Birth Rate",
     main   = "Histogram of Birth Rate",
     col    = "dodgerblue",
     border = "black",
     breaks = 20)
```



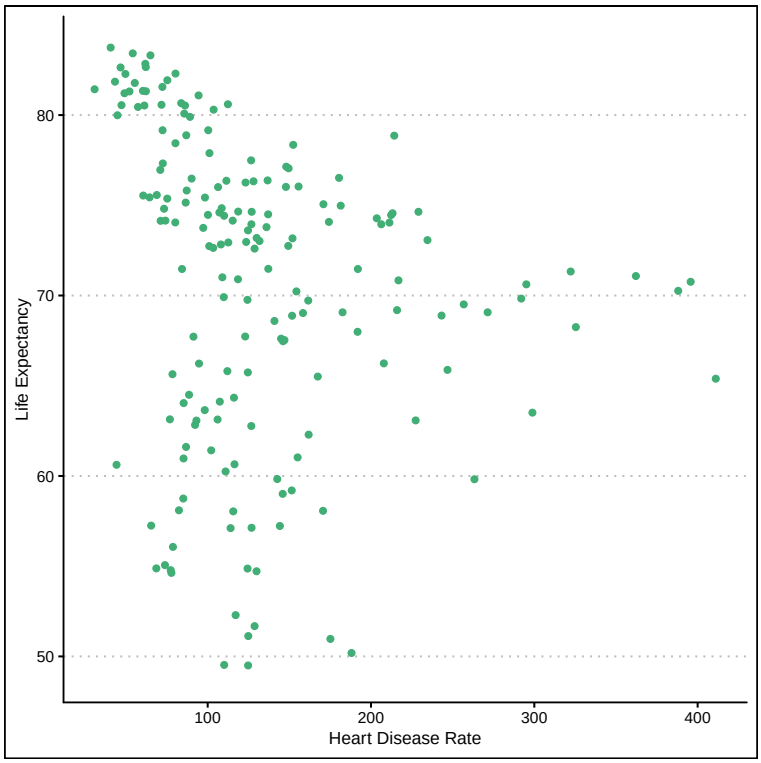

```
# Cancer
life_exp_full %>%
  ggplot() +
  geom_point(
    aes(y = `Life Expectancy`, x = `Cancer Rate`),
    color = COLA[3]) +
  theme_clean()

## Warning:  Removed 70 rows containing missing values (geom_point).
```



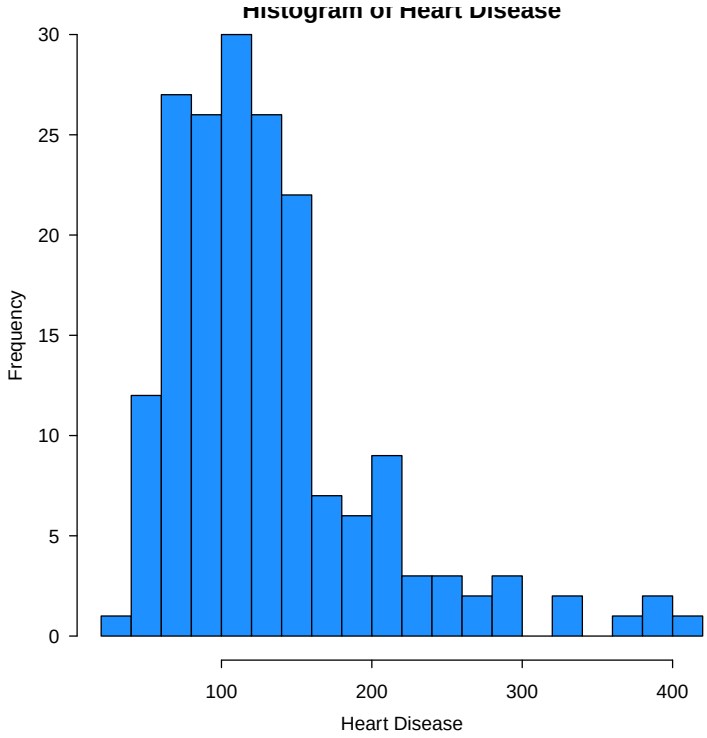

```
hist(life_exp_full$`Cancer Rate`,
     xlab   = "Cancer Disease",
     main   = "Histogram of Cancer Disease",
     col    = "dodgerblue",
     border = "black",
     breaks = 20)
```



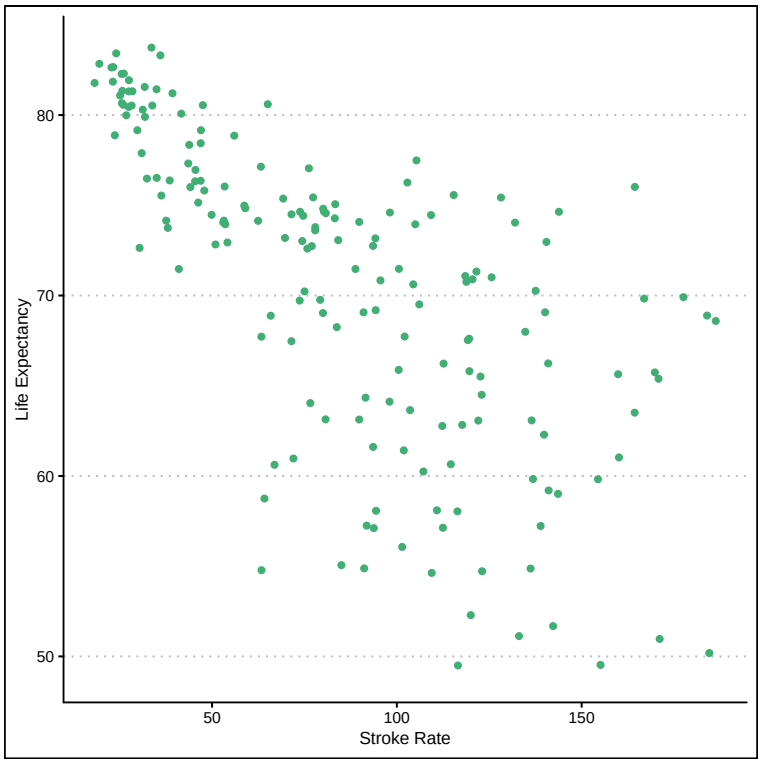

```
# Heart Disease
life_exp_full %>%
  ggplot() +
  geom_point(
    aes(y = `Life Expectancy`, x = `Heart Disease Rate`),
    color = COLA[3]) +
  theme_clean()

## Warning:  Removed 70 rows containing missing values (geom_point).
```



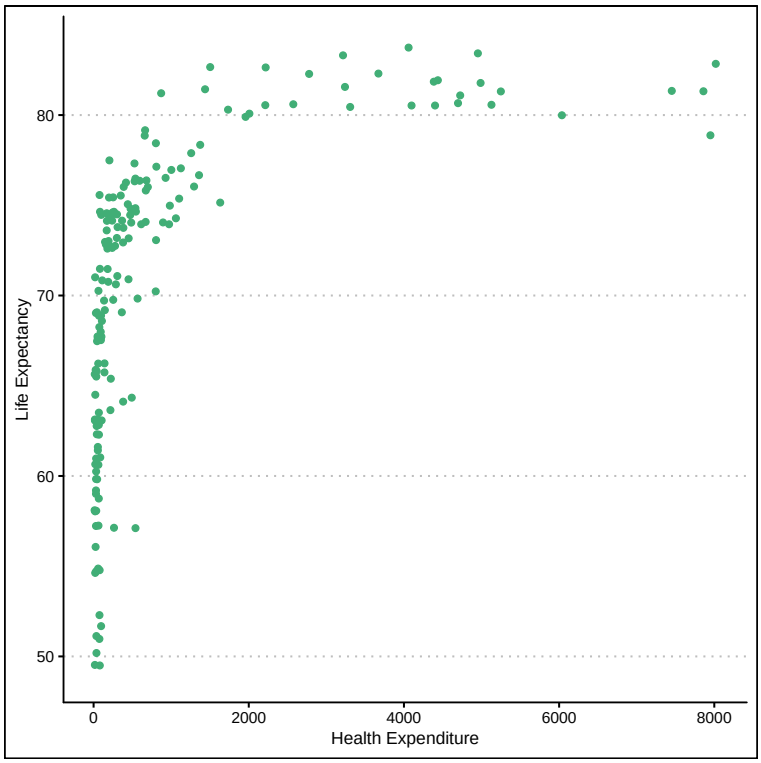

```
hist(life_exp_full$`Heart Disease Rate`,
     xlab   = "Heart Disease",
     main   = "Histogram of Heart Disease",
     col    = "dodgerblue",
     border = "black",
     breaks = 20)
```



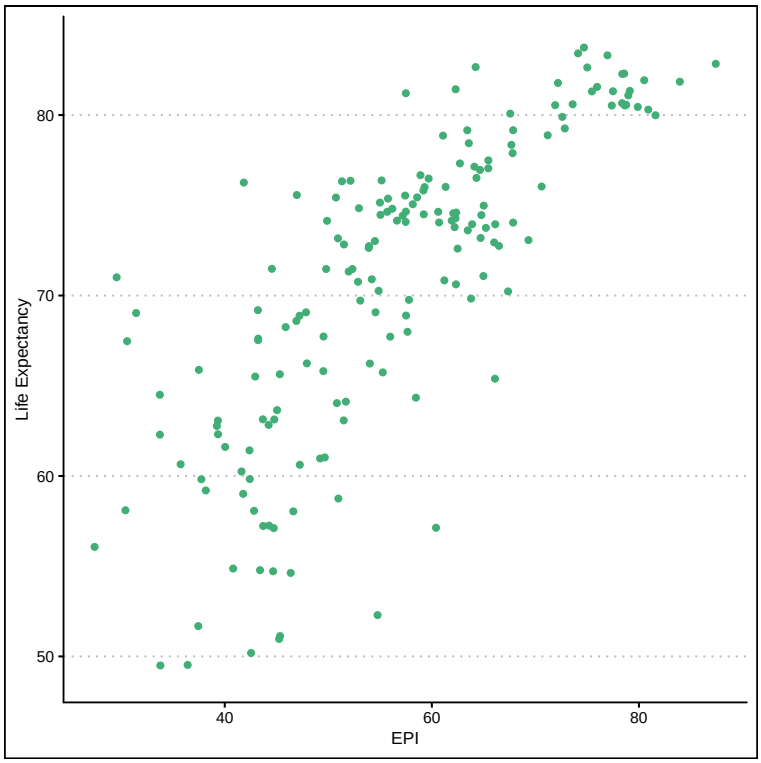

```
# Stroke
life_exp_full %>%
  ggplot() +
  geom_point(
    aes(y = `Life Expectancy`, x = `Stroke Rate`),
    color = COLA[3]) +
  theme_clean()

## Warning:  Removed 70 rows containing missing values (geom_point).
```



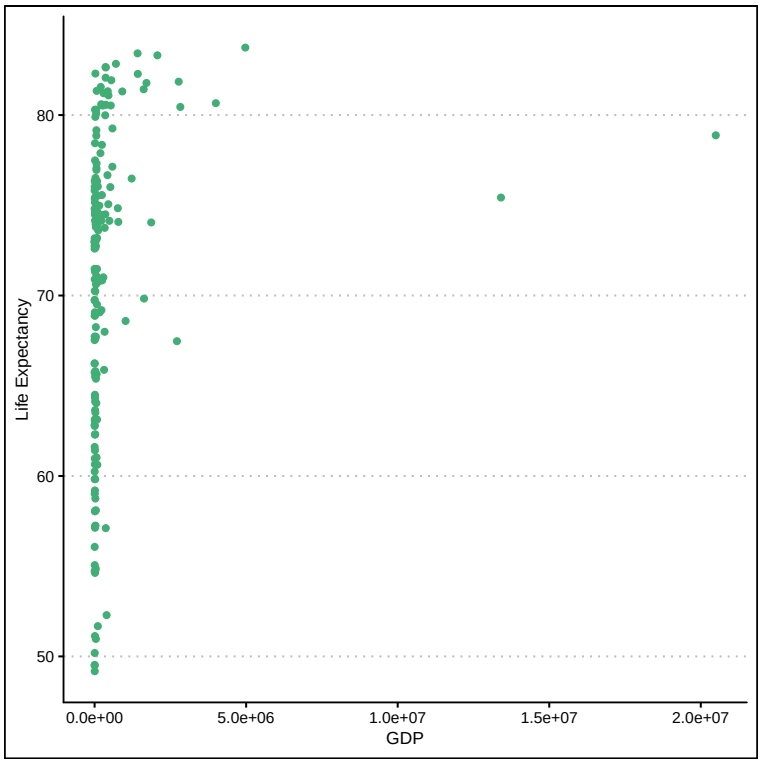

```
# Health Expenditure
life_exp_full %>%
  ggplot() +
  geom_point(
    aes(y = `Life Expectancy`, x = `Health Expenditure`),
    color = COLA[3]) +
  theme_clean()

## Warning:  Removed 76 rows containing missing values (geom_point).
```



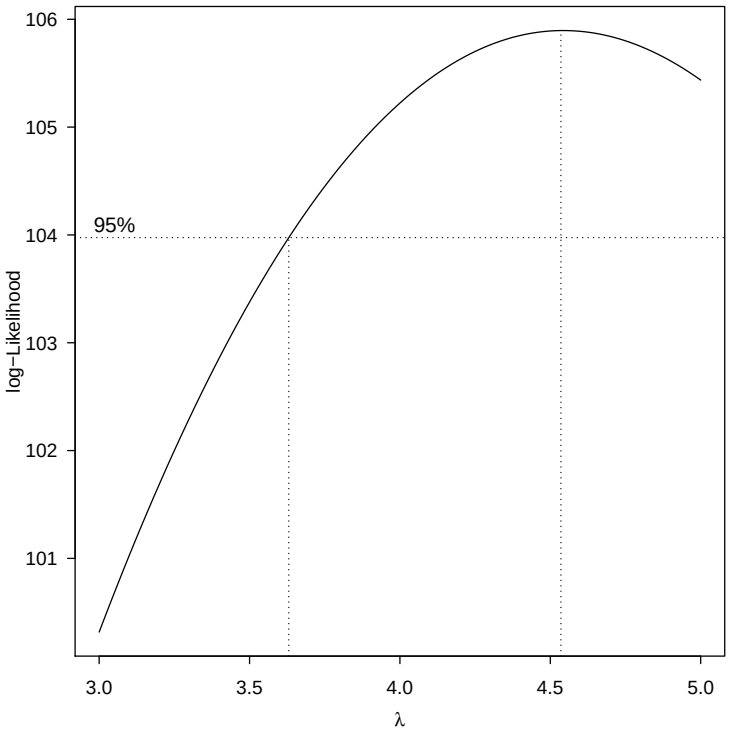

```
# EPI
life_exp_full %>%
  ggplot() +
  geom_point(
    aes(y = `Life Expectancy`, x = EPI),
    color = COLA[3]) +
  theme_clean()

## Warning:  Removed 73 rows containing missing values (geom_point).
```




```
# GDP
life_exp_full %>%
  ggplot() +
  geom_point(
    aes(y = `Life Expectancy`, x = GDP),
    color = COLA[3]) +
  theme_clean()

## Warning:  Removed 67 rows containing missing values (geom_point).
```




```r
# Regressions ------------------------------------------------------------

model_full <- lm(
  `Life Expectancy` ~ `Birth Rate` + `Cancer Rate` + `Heart Disease Rate` + `Stroke Rate` +  `Health Exp
  data = life_exp_full
)

model_red <- lm(
    `Life Expectancy` ~ `Birth Rate` + `Stroke Rate` +  `Health Expenditure` + EPI + GDP,
    data = model_full$model)
# Has violations of assumptions (see below)
# Note: data = model_full$model in reduced
#    model to avoid "models were not all fitted to the same size of dataset" error in ANOVA


# Model BC Full --- Box-Cox Full Transform
boxcox(model_full, plotit = TRUE, lambda = seq(3, 5, by = 0.1))
```




```
transform_bc_y <- ((life_exp_full$`Life Expectancy`)^4.4 - 1)/4.5

model_bc_full_transform <-
  lm(
    transform_bc_y ~ `Birth Rate` + `Cancer Rate` + `Heart Disease Rate` + `Stroke Rate` + `Health Exper
    data = life_exp_full
  )
# Here we see that lambda = 4.5 is both in the confidence interval, and is extremely close to the maxim
# This suggests a transformation of $\frac{y^\lamba - 1}{\lambda} = \frac{y^4.5 - 1}{4.5}$

# Testing Model Fit ------------------------------------------------------

# compare model full and model reduced.
anova(model_red, model_full)

## Analysis of Variance Table
##
## Model 1: `Life Expectancy` ~ `Birth Rate` + `Stroke Rate` + `Health Expenditure` +
##     EPI + GDP
## Model 2: `Life Expectancy` ~ `Birth Rate` + `Cancer Rate` + `Heart Disease Rate` +
##     `Stroke Rate` + `Health Expenditure` + EPI + GDP
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    158 1995.2
## 2    156 1963.5  2    31.704 1.2594 0.2867

# F = 0.8399
# Pr(>F) = 0.5018
# Failed to reject H0: that removed var are zero.

# Diagnostic Checks - Model Full -----------------------------------------
```

```r
# Model Summary and ANOVA
summary(model_full)
```

```
##
## Call:
## lm(formula = `Life Expectancy` ~ `Birth Rate` + `Cancer Rate` +
##     `Heart Disease Rate` + `Stroke Rate` + `Health Expenditure` +
##     EPI + GDP, data = life_exp_full)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0296  -1.7375   0.0473   2.1224   7.1708
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.637e+01  2.958e+00  25.817  < 2e-16 ***
## `Birth Rate`         -4.243e-01  3.881e-02 -10.931  < 2e-16 ***
## `Cancer Rate`        -1.507e-02  1.017e-02  -1.482    0.140
## `Heart Disease Rate`  2.311e-03  5.156e-03   0.448    0.655
## `Stroke Rate`        -5.894e-02  1.051e-02  -5.607 9.13e-08 ***
## `Health Expenditure` -2.225e-04  2.687e-04  -0.828    0.409
## EPI                   1.801e-01  4.050e-02   4.448 1.64e-05 ***
## GDP                   9.325e-08  1.539e-07   0.606    0.546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.548 on 156 degrees of freedom
##   (84 observations deleted due to missingness)
## Multiple R-squared:  0.8326,Adjusted R-squared:  0.8251
## F-statistic: 110.8 on 7 and 156 DF,  p-value: < 2.2e-16
```

```r
anova(model_full)
```

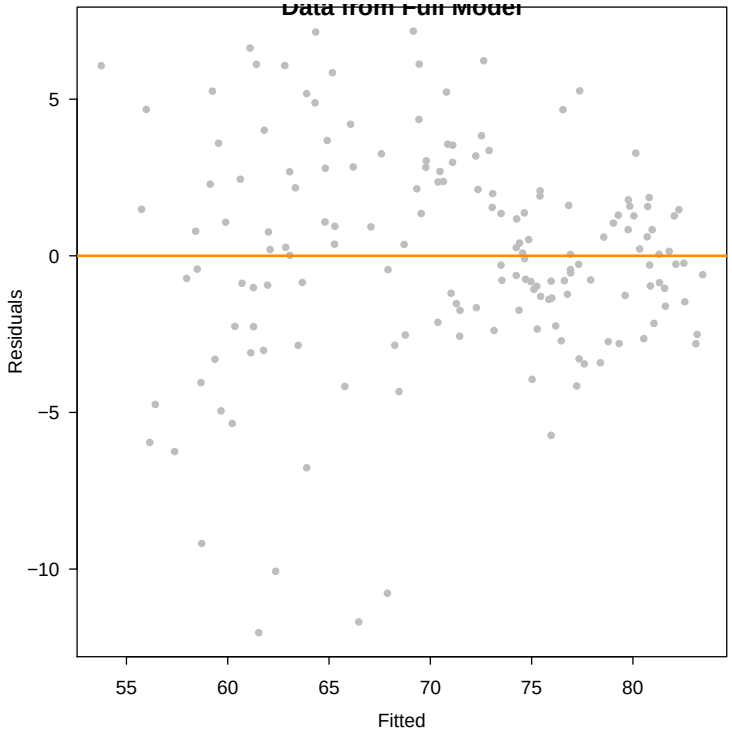
```
## Analysis of Variance Table
##
## Response: Life Expectancy
##                       Df Sum Sq Mean Sq  F value    Pr(>F)
## `Birth Rate`           1 8213.0  8213.0 652.5242 < 2.2e-16 ***
## `Cancer Rate`          1   29.2    29.2   2.3207   0.12969
## `Heart Disease Rate`   1  276.2   276.2  21.9426 6.072e-06 ***
## `Stroke Rate`          1  950.8   950.8  75.5445 4.620e-15 ***
## `Health Expenditure`   1   45.0    45.0   3.5773   0.06043 .
## EPI                    1  244.8   244.8  19.4518 1.914e-05 ***
## GDP                    1    4.6     4.6   0.3669   0.54557
## Residuals            156 1963.5    12.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 1 The regression function is linear (the relationship is linear).
# Yes
vif(model_full)
```

```
##         `Birth Rate`        `Cancer Rate` `Heart Disease Rate`        `Stroke Rate`
```

```
##             2.177726             1.214850             1.651303             2.617937
## `Health Expenditure`                  EPI                  GDP
##             2.712092             3.685143             1.226175
```

```r
# 2 The error terms have a constant variance

# Fitted vs Residuals --- model_full
plot(fitted(model_full), resid(model_full), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Data from Full Model")
abline(h = 0, col = "darkorange", lwd = 2)
```

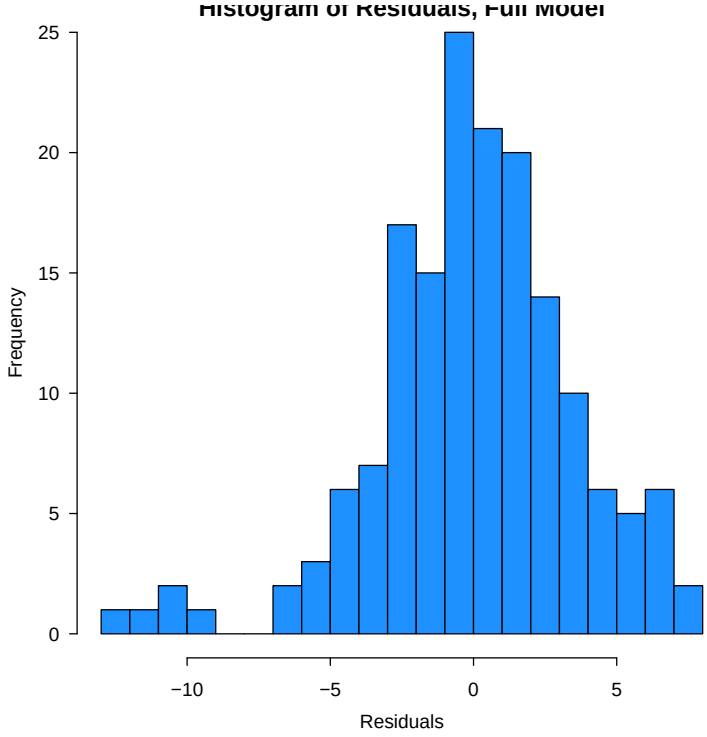



```r
# Looks like it has a inverse parabolic shape


# 3 The error terms are independent (there is no relationship among the error terms).

# Breusch-Pagan Test for Homoskedasticity
bptest(model_full)
```

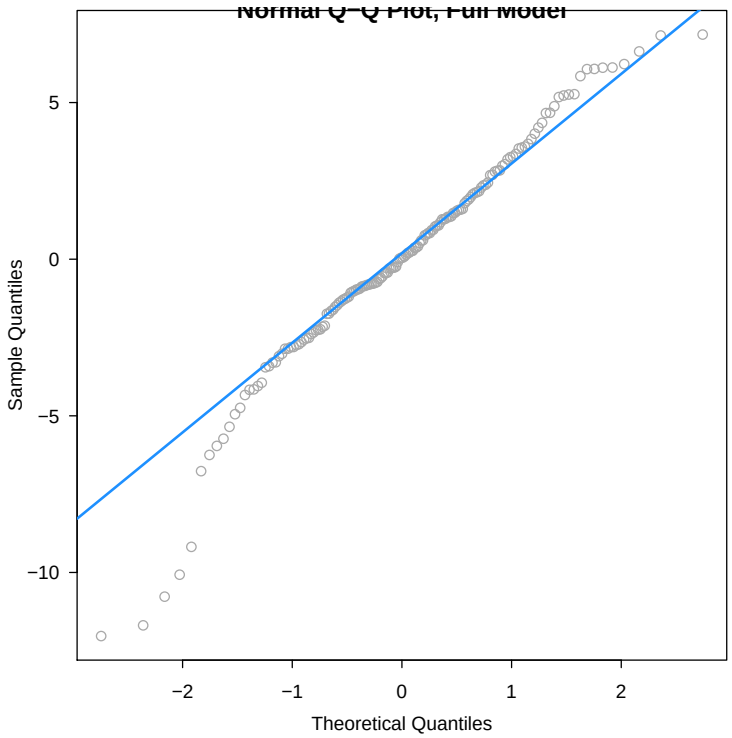
```
##
##  studentized Breusch-Pagan test
##
## data:  model_full
## BP = 20.003, df = 7, p-value = 0.005563
```

```r
# For model_full we see a small p-value, so we reject the null hypothesis of
#     homoskedasticity is rejected and heteroskedasticity assumed.
# The constant variance assumption is violated.
# This matches our findings with a fitted versus residuals plot.
```

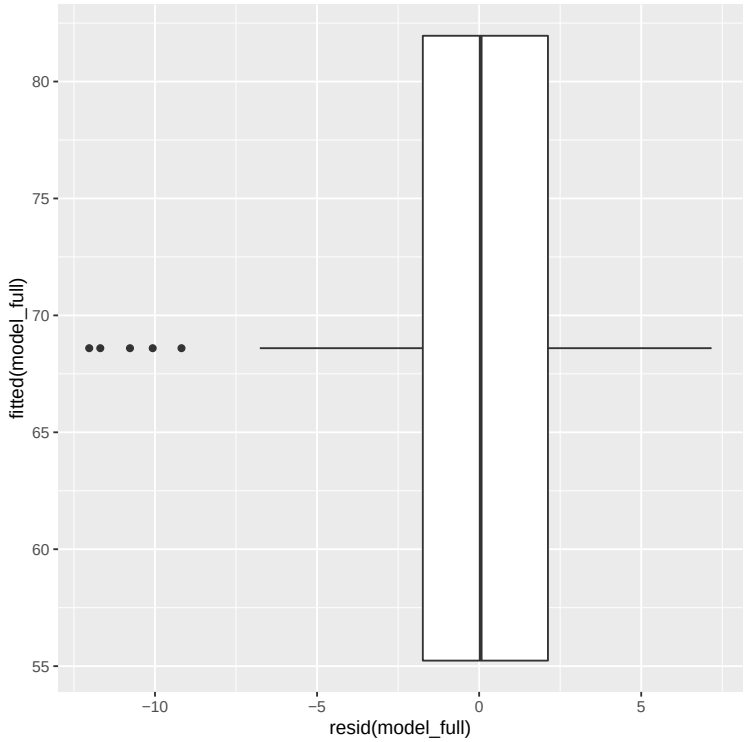
```r
# 4 The error terms are normally distributed

hist(resid(model_full),
     xlab   = "Residuals",
     main   = "Histogram of Residuals, Full Model",
     col    = "dodgerblue",
     border = "black",
     breaks = 20)
```




```r
# It does have a rough bell shape, however, it also has a semi-sharp peak.

# Q-Q Plot
qqnorm(resid(model_full), main = "Normal Q-Q Plot, Full Model", col = "darkgrey")
qqline(resid(model_full), col = "dodgerblue", lwd = 2)
```




```
# Deviates in smaller quantiles
# For Model Full, we have a suspect Q-Q plot.
# We would probably not believe the errors follow a normal distribution.

# Shapiro-Wilk Test
shapiro.test(resid(model_full))

##
##  Shapiro-Wilk normality test
##
## data:  resid(model_full)
## W = 0.96343, p-value = 0.0002576

# p = 7.152e-05
# A small p-value indicates we believe there is only a small probability
# the data could have been sampled from a normal distribution.

#RK 5 Outlier Check Via Boxplots- how should we deal with those outliers? Which countries are they?
outlierAssumption <- ggplot(model_full, aes(x=fitted(model_full), y=resid(model_full))) +
  geom_boxplot() +
  coord_flip()
outlierAssumption

## Warning:  Continuous x aesthetic - did you forget aes(group=...)?
```



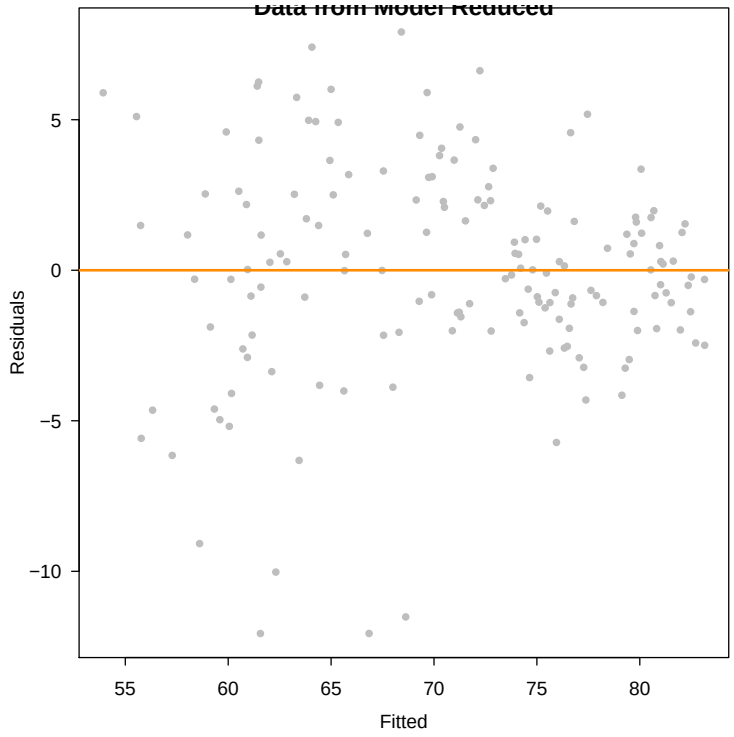

```
# 6 There is no important predictor that have been omitted from the model
# RK I think for this one since she's just looking for a logical explanation, we can say
# that there very well maybe be other factors that are contributing to the life expectancy
# of a country, but it is impossible to state them all. I'm going to put a better explanation
# and possible alternative predictors in the actual paper.


# Diagnostic Checks - Model Reduced ------------------------------------------------------

# Model Summary and ANOVA
summary(model_red)

##
## Call:
## lm(formula = `Life Expectancy` ~ `Birth Rate` + `Stroke Rate` +
##     `Health Expenditure` + EPI + GDP, data = model_full$model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0656  -1.8950   0.0142   2.1378   7.9142
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.523e+01  2.868e+00  26.228  < 2e-16 ***
## `Birth Rate`         -4.158e-01  3.763e-02 -11.052  < 2e-16 ***
## `Stroke Rate`        -5.863e-02  8.794e-03  -6.667 4.14e-10 ***
## `Health Expenditure` -2.300e-04  2.617e-04  -0.879    0.381
## EPI                   1.725e-01  3.872e-02   4.456 1.57e-05 ***
## GDP                   8.809e-08  1.542e-07   0.571    0.569
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.554 on 158 degrees of freedom
## Multiple R-squared:  0.8299,Adjusted R-squared:  0.8245
## F-statistic: 154.1 on 5 and 158 DF,  p-value: < 2.2e-16
```

```
anova(model_red)
```

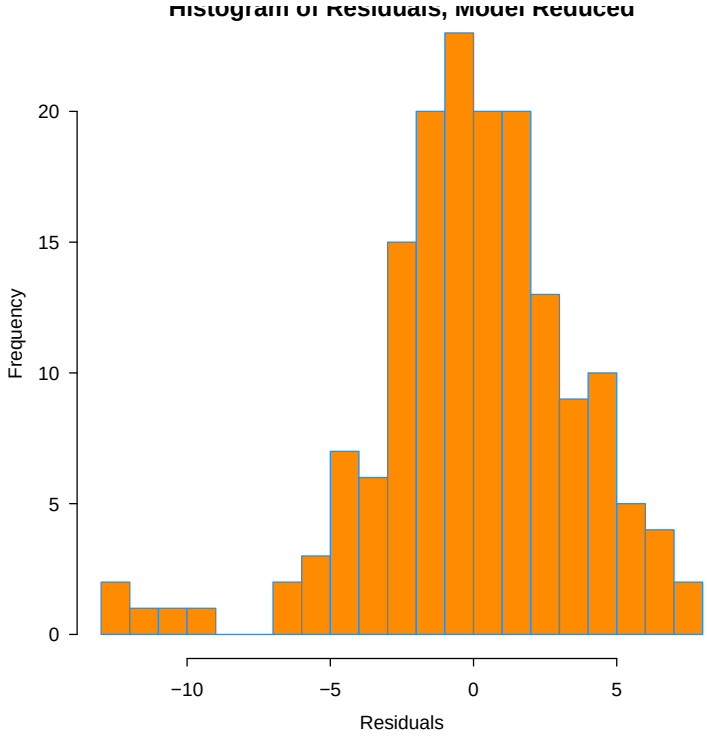
```
## Analysis of Variance Table
##
## Response: Life Expectancy
##                        Df Sum Sq Mean Sq  F value    Pr(>F)
## `Birth Rate`            1 8213.0  8213.0 650.3884 < 2.2e-16 ***
## `Stroke Rate`           1 1233.4  1233.4  97.6747 < 2.2e-16 ***
## `Health Expenditure`    1   34.0    34.0   2.6914    0.1029
## EPI                     1  247.5   247.5  19.5972 1.776e-05 ***
## GDP                     1    4.1     4.1   0.3265    0.5685
## Residuals             158 1995.2    12.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Fitted vs Residuals --- model_red
plot(fitted(model_red), resid(model_red), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Data from Model Reduced")
abline(h = 0, col = "darkorange", lwd = 2)
```

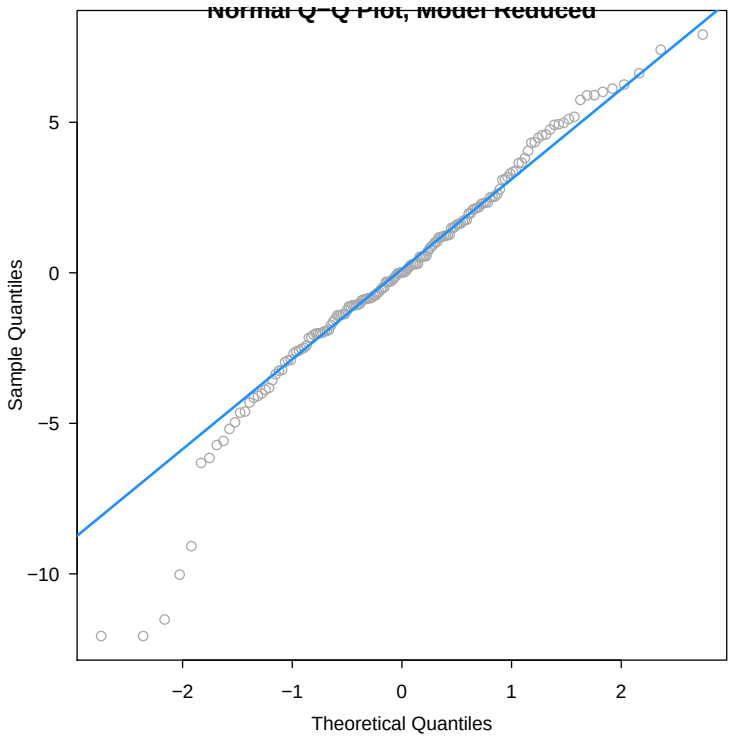



```
# Looks like it has a inverse parabolic shape

# Breusch-Pagan Test for Homoskedasticity
bptest(model_red)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_red
## BP = 18.747, df = 5, p-value = 0.002142

# For model_red we see a small p-value, so we reject the null of homoskedasticity.
# The constant variance assumption is violated.
# This matches our findings with a fitted versus residuals plot.

# Normality of errors
hist(resid(model_red),
     xlab   = "Residuals",
     main   = "Histogram of Residuals, Model Reduced",
     col    = "darkorange",
     border = "dodgerblue",
     breaks = 20)
```

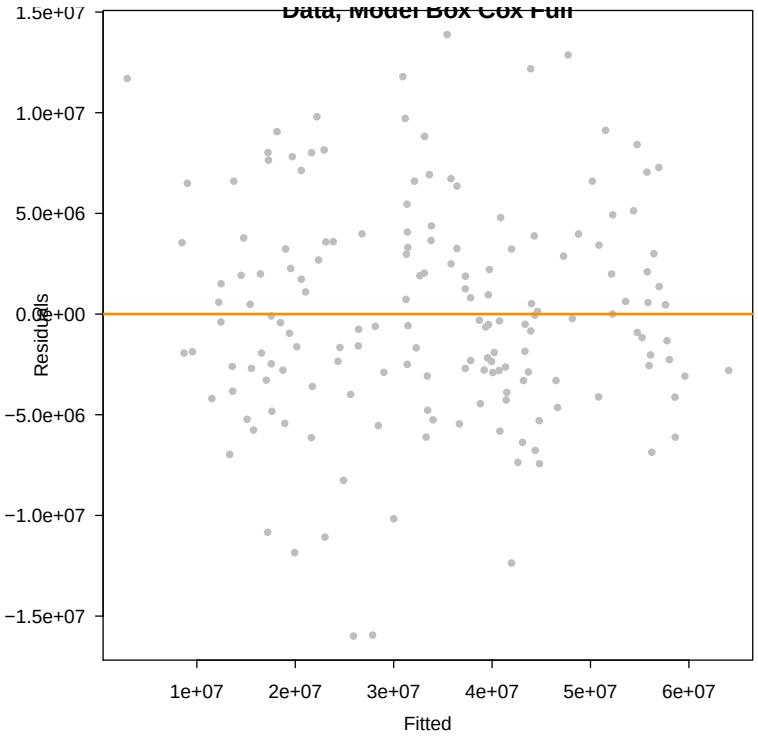



```
# It does have a rough bell shape, however, it also has a very sharp peak.

# Q-Q Plot
qqnorm(resid(model_red), main = "Normal Q-Q Plot, Model Reduced", col = "darkgrey")
qqline(resid(model_red), col = "dodgerblue", lwd = 2)
```

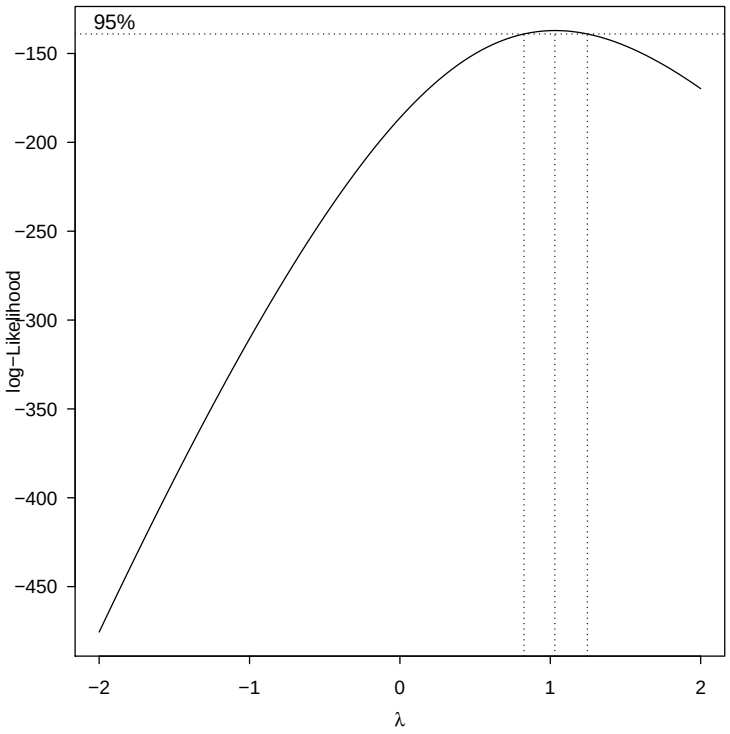



```
# Deviates in smaller quantiles
# For Model Reduced, we have a suspect Q-Q plot.
# We would probably not believe the errors follow a normal distribution.


# Shapiro-Wilk Test
shapiro.test(resid(model_red))

##
##  Shapiro-Wilk normality test
##
## data:  resid(model_red)
## W = 0.96242, p-value = 0.0002042

# p = 7.152e-05
# A small p-value indicates we believe there is only a small probability
# the data could have been sampled from a normal distribution.


# Diagnostic Checks - Model Box Cox Full Transform ------------------------------

# Model Summary and ANOVA
summary(model_bc_full_transform)

##
## Call:
## lm(formula = transform_bc_y ~ `Birth Rate` + `Cancer Rate` +
##     `Heart Disease Rate` + `Stroke Rate` + `Health Expenditure` +
##     EPI + GDP, data = life_exp_full)
##
```

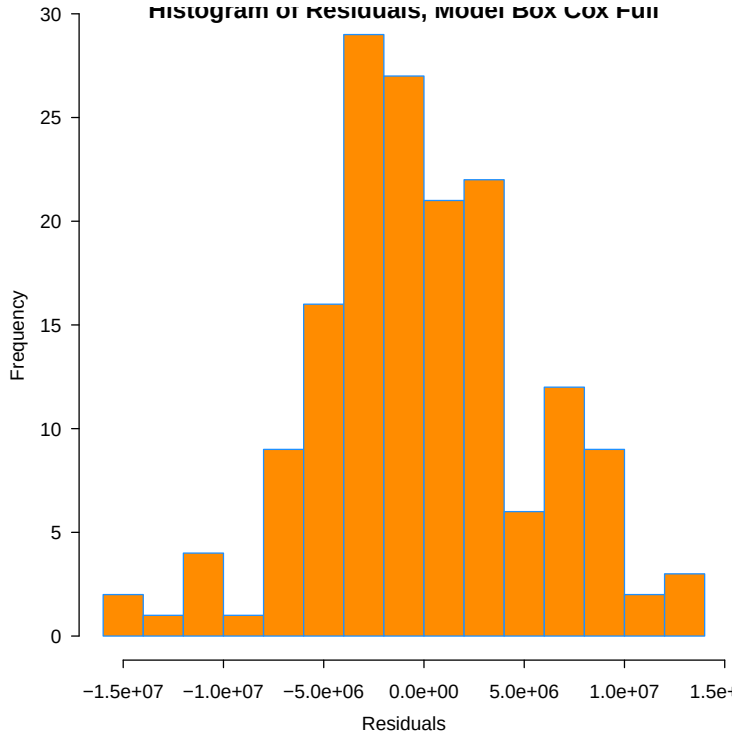
```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -15991413  -3077959   -466975   3338495  13883067
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.077e+07  4.604e+06   8.856 1.74e-15 ***
## `Birth Rate`         -6.679e+05  6.040e+04 -11.058  < 2e-16 ***
## `Cancer Rate`        -3.381e+04  1.582e+04  -2.137   0.0342 *
## `Heart Disease Rate` -1.181e+04  8.025e+03  -1.471   0.1432
## `Stroke Rate`        -8.940e+04  1.636e+04  -5.465 1.80e-07 ***
## `Health Expenditure`  7.125e+02  4.181e+02   1.704   0.0903 .
## EPI                   3.486e+05  6.303e+04   5.531 1.31e-07 ***
## GDP                   2.251e-02  2.396e-01   0.094   0.9253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5521000 on 156 degrees of freedom
##   (84 observations deleted due to missingness)
## Multiple R-squared:  0.8743,Adjusted R-squared:  0.8686
## F-statistic:   155 on 7 and 156 DF,  p-value: < 2.2e-16

anova(model_bc_full_transform)

## Analysis of Variance Table
##
## Response: transform_bc_y
##                        Df     Sum Sq    Mean Sq  F value    Pr(>F)
## `Birth Rate`            1 2.5564e+16 2.5564e+16 838.6101 < 2.2e-16 ***
## `Cancer Rate`           1 1.2635e+14 1.2635e+14   4.1449   0.04345 *
## `Heart Disease Rate`    1 2.2608e+15 2.2608e+15  74.1647 7.408e-15 ***
## `Stroke Rate`           1 3.2181e+15 3.2181e+15 105.5688 < 2.2e-16 ***
## `Health Expenditure`    1 9.3871e+14 9.3871e+14  30.7943 1.204e-07 ***
## EPI                     1 9.5687e+14 9.5687e+14  31.3900 9.320e-08 ***
## GDP                     1 2.6913e+11 2.6913e+11   0.0088   0.92526
## Residuals             156 4.7554e+15 3.0483e+13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Fitted vs Residuals
plot(
  fitted(model_bc_full_transform),
  resid(model_bc_full_transform),
  col = "grey",
  pch = 20,
  xlab = "Fitted",
  ylab = "Residuals",
  main = "Data, Model Box Cox Full"
)
abline(h = 0, col = "darkorange", lwd = 2)
```



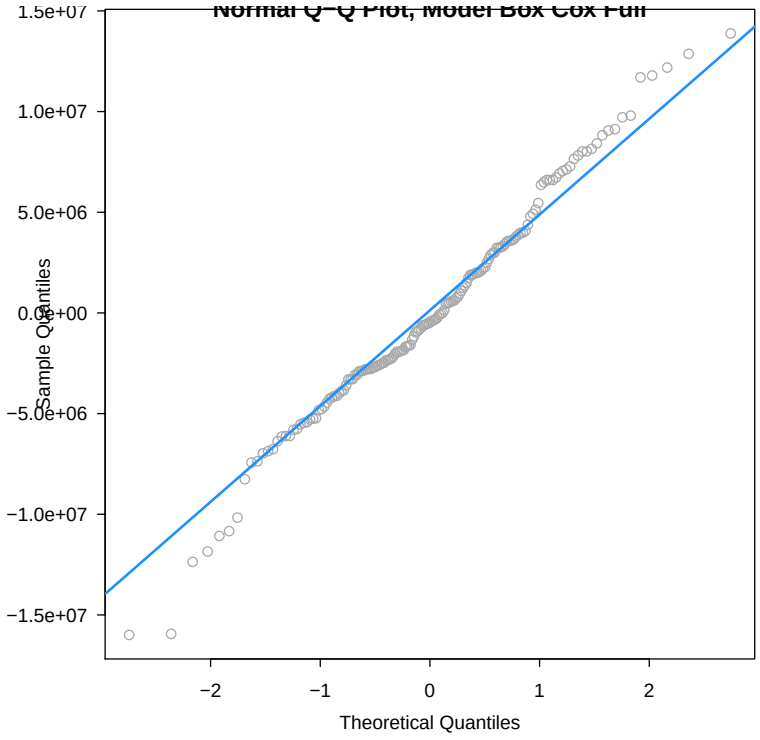

```
# Looks random so all set

# Breusch-Pagan
bptest(model_bc_full_transform)

##
##  studentized Breusch-Pagan test
##
## data:  model_bc_full_transform
## BP = 3.6047, df = 7, p-value = 0.824

# pass -- FTR null of homosked.

# Shapiro-Wilks
shapiro.test(resid(model_bc_full_transform))

##
##  Shapiro-Wilk normality test
##
## data:  resid(model_bc_full_transform)
## W = 0.98731, p-value = 0.1442

# pass
# A large p-value indicates we believe it is likely
# the data could have been sampled from a normal distribution.

# Box-cox
boxcox(model_bc_full_transform)
```



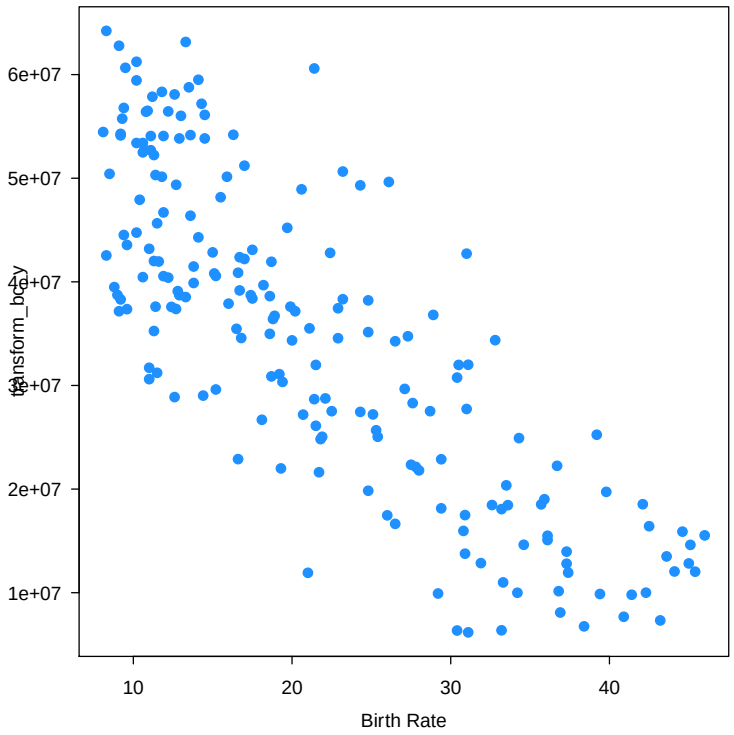

```
# pass

# Variation Inflation Factor
vif(model_bc_full_transform)

##         `Birth Rate`        `Cancer Rate` `Heart Disease Rate`        `Stroke Rate`
##             2.177726             1.214850             1.651303             2.617937
## `Health Expenditure`                  EPI                  GDP
##             2.712092             3.685143             1.226175

# All <5 so no multicollinearity problems

# Normality of errors
hist(
  resid(model_bc_full_transform),
  xlab   = "Residuals",
  main   = "Histogram of Residuals, Model Box Cox Full",
  col    = "darkorange",
  border = "dodgerblue",
  breaks = 20
)
```



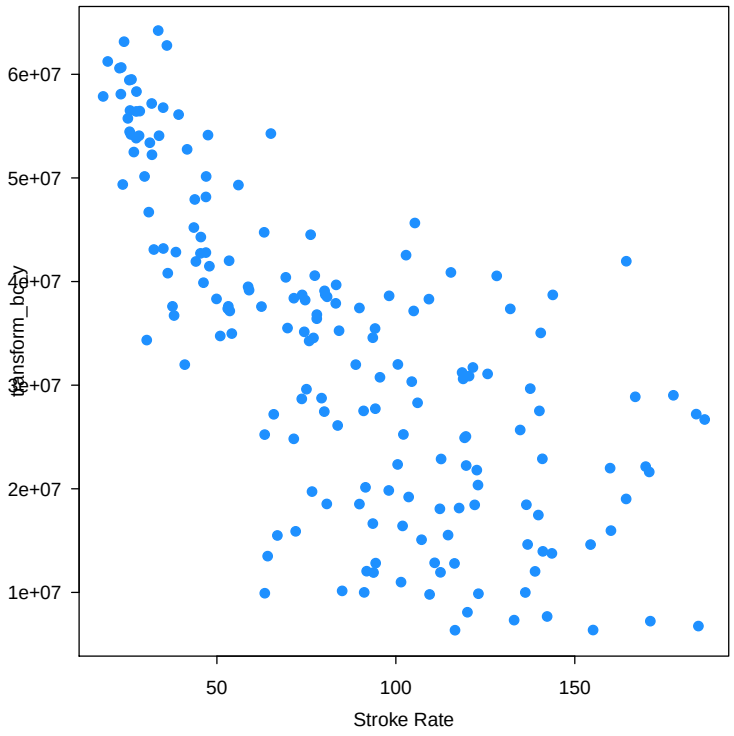

```r
# It does have a rough bell shape. Looks Good.

# Q-Q Plot
qqnorm(resid(model_bc_full_transform), main = "Normal Q-Q Plot, Model Box Cox Full", col = "darkgrey")
qqline(resid(model_bc_full_transform), col = "dodgerblue", lwd = 2)
```



```r
# Deviates in slightly smaller quantiles
# For Model BC, we have an okay Q-Q plot.
# We would probably believe the errors follow a mostly normal distribution.

# Linearity
plot(
  transform_bc_y ~ `Birth Rate`,
  data = life_exp_full,
  col = "dodgerblue",
  pch = 20,
  cex = 1.5
)
```




```r
# Linear (-)

plot(
  transform_bc_y ~ `Stroke Rate`,
  data = life_exp_full,
  col = "dodgerblue",
  pch = 20,
  cex = 1.5
)
```




```
# Linear (-) ish --- looks like it flairs out

plot(
  transform_bc_y ~ EPI,
  data = life_exp_full,
  col = "dodgerblue",
  pch = 20,
  cex = 1.5
)
# Linear (+)

# Diagnostic Checks - Model Box Cox Reduced Transform ----------------------------

# Model Summary and ANOVA
summary(model_bc_red_transform)
```

```
## Error in summary(model_bc_red_transform):  object 'model_bc_red_transform' not found
```

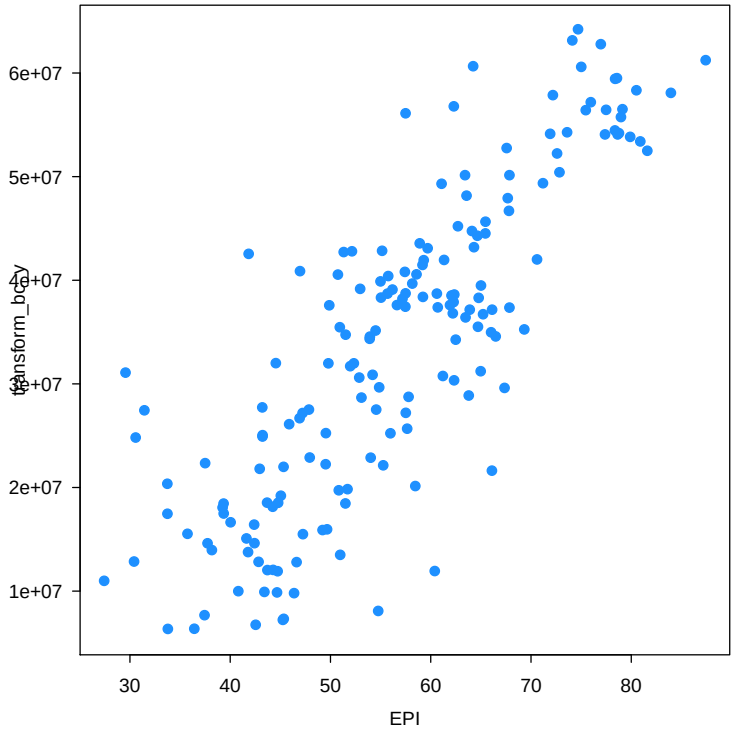
```
anova(model_bc_red_transform)
```

```
## Error in anova(model_bc_red_transform):  object 'model_bc_red_transform' not found
```

```
# Fitted vs Residuals
plot(
  fitted(model_bc_red_transform),
  resid(model_bc_red_transform),
  col = "grey",
  pch = 20,
  xlab = "Fitted",
  ylab = "Residuals",
  main = "Data from Model 10"
)
```

```
## Error in fitted(model_bc_red_transform):  object 'model_bc_red_transform' not found
```

```
abline(h = 0, col = "darkorange", lwd = 2)
```




```
# Looks random so all set

# Breusch-Pagan
bptest(model_bc_red_transform)
```

```
## Error in bptest(model_bc_red_transform):  object 'model_bc_red_transform' not found
```

```
# pass -- FTR null of homosked.

# Shapiro-Wilks
shapiro.test(resid(model_bc_red_transform))
```

```
## Error in resid(model_bc_red_transform):  object 'model_bc_red_transform' not found
```

```
# pass
# A large p-value indicates we believe it is likely
# the data could have been sampled from a normal distribution.

# Box-cox
boxcox(model_bc_red_transform)
```

```
## Error in boxcox(model_bc_red_transform):  object 'model_bc_red_transform' not found
```

```
# pass

# Variation Inflation Factor
vif(model_bc_red_transform)
```

```
## Error in vif(model_bc_red_transform):  object 'model_bc_red_transform' not found
```

```
# All <5 so no multicollinearity problems

# Normality of errors
hist(
  resid(model_bc_red_transform),
  xlab   = "Residuals",
  main   = "Histogram of Residuals, Model 10",
  col    = "darkorange",
  border = "dodgerblue",
  breaks = 20
)
```

```
## Error in resid(model_bc_red_transform):  object 'model_bc_red_transform' not found
```

```
# It does have a rough bell shape. Looks Good.

# Q-Q Plot
qqnorm(resid(model_bc_red_transform), main = "Normal Q-Q Plot, Model 13", col = "darkgrey")
```

```
## Error in resid(model_bc_red_transform):  object 'model_bc_red_transform' not found
```

```
qqline(resid(model_bc_red_transform), col = "dodgerblue", lwd = 2)
```

```
## Error in resid(model_bc_red_transform):  object 'model_bc_red_transform' not found
```

```
# Deviates in slightly smaller quantiles
# For Model BC, we have an okay Q-Q plot.
# We would probably believe the errors follow a mostly normal distribution.

# Linearity
plot(
  transform_bc_red_y ~ `Birth Rate`,
  data = life_exp_full,
  col = "dodgerblue",
  pch = 20,
  cex = 1.5
)
```

```
## Error in eval(predvars, data, env):  object 'transform_bc_red_y' not found
```

```
# Linear (-)

plot(
  transform_bc_red_y ~ `Stroke Rate`,
  data = life_exp_full,
  col = "dodgerblue",
  pch = 20,
  cex = 1.5
)
```

```
## Error in eval(predvars, data, env):  object 'transform_bc_red_y' not found
```

```
# Linear (-) ish --- looks like it flairs out

plot(
  transform_bc_red_y ~ EPI,
```

```
  data = life_exp_full,
  col = "dodgerblue",
  pch = 20,
  cex = 1.5
)
```

```
## Error in eval(predvars, data, env):  object 'transform_bc_red_y' not found
```

```
# Linear (+)
```

```
# End of File ------------------------------------------------------------
```

The R session information (including the OS info, R version and all packages used):

```
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.1
##
## Matrix products: default
## BLAS:   /System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versio
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/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] faraway_1.0.7   MASS_7.3-51.4   scales_1.0.0    lmtest_0.9-37   zoo_1.8-6
##  [6] bbplot_0.2      ggthemes_4.2.0  ggsci_2.9       gridExtra_2.3   forcats_0.4.0
## [11] stringr_1.4.0   purrr_0.3.3     readr_1.3.1     tidyr_1.0.0     tibble_2.1.3
## [16] ggplot2_3.2.1   tidyverse_1.2.1 dplyr_0.8.3     knitr_1.26
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3         lubridate_1.7.4    lattice_0.20-38    png_0.1-7
##  [5] statquotes_0.2.2   assertthat_0.2.1   zeallot_0.1.0      packrat_0.5.0
##  [9] R6_2.4.1           cellranger_1.1.0   backports_1.1.5    evaluate_0.14
## [13] httr_1.4.0         highr_0.8          pillar_1.4.2       rlang_0.4.1
## [17] lazyeval_0.2.2     readxl_1.3.1       minqa_1.2.4        rstudioapi_0.10
## [21] nloptr_1.2.1       Matrix_1.2-17      labeling_0.3       splines_3.6.1
## [25] lme4_1.1-21        tidytext_0.2.1     munsell_0.5.0      broom_0.5.2
## [29] compiler_3.6.1     janeaustenr_0.1.5  modelr_0.1.4       xfun_0.11
## [33] pkgconfig_2.0.3    tidyselect_0.2.5   ggpubr_0.2.4       crayon_1.3.4
## [37] withr_2.1.2        SnowballC_0.6.0    grid_3.6.1         nlme_3.1-140
## [41] jsonlite_1.6       gtable_0.3.0       lifecycle_0.1.0    magrittr_1.5
## [45] tokenizers_0.2.1   cli_1.1.0          stringi_1.4.3      ggsignif_0.6.0
## [49] xml2_1.2.0         ellipsis_0.3.0     generics_0.0.2     vctrs_0.2.0
## [53] cowplot_1.0.0      boot_1.3-22        wordcloud_2.6      RColorBrewer_1.1-2
## [57] tools_3.6.1        glue_1.3.1         hms_0.5.0          colorspace_1.4-1
```

```
## [61] rvest_0.3.4        haven_2.1.1
```

```
Sys.time()
```

```
## [1] "2019-12-02 11:27:12 EST"
```