STAT 135 Technical Report

05/24/2021

AUTHORS

Group H: Dasha Asienga, Gerardo Orellana, Parmis Rouzbahani, Mathew Mueller

TITLE

Prevalence of Alzheimer's Disease per Fertilizer and Herbicide Use, Economic Well-Being, and Life Expectancy

ABSTRACT

There have been significant correlates between exposure to nitrates and other fertilizers in the environment and the development of Alzheimer's and early onset dementia in adults. In an attempt to test these findings at a more global scale, we looked at the fertilizer used in different countries around the world and controlled for life expectancy, GDP, and level of development while testing for a relationship with Alzheimer's death rates. We conducted univariate analysis, bivariate analysis, and contemplated transforming the data using logarithms and square roots. After checking the necessary conditions for linear regression, we used multiple regression and concluded that fertilizer usage has no significant positive effects on the Alzheimer's death rate, while development and GDP are strongly correlated.

PROJECT AIMS

The purpose of our statistical analysis is to measure how much prevalence of Alzheimer's Disease is predicted by type of fertilizer and pesticide use, in addition to a country's economic status and health quality as well as whether it is developed or not. Prevalence of Alzheimer's Disease is measured by each country's Alzheimer's death rate, Fertilizer and Herbicide Use is measured by the kilograms per hectare of nitrogen, potash, and phosphate fertilizer in each country, Economic Well-Being is measured by the country's GDP per capita and whether it is developed or not, and Life Expectancy is measured by each country's life expectancy at birth. Our aim, thus, is to determine whether there is a relationship between usage of phosphate, potash, and phosphate fertilizers and Alzheimer's death rates in a country, after controlling for the effects of economic well-being (GDP per capita and developed vs not developed) and life expectancy.

INTRODUCTION

The agricultural industry uses large amounts of fertilizers to increase crop yield. These fertilizers enter the environment and affect animals, plants, and drinking water, posing a risk to environmental and human health. Some chemical fertilizers, specifically those with nitrates, have been found to trigger health problems in people with Alzheimer's Disease. In some cases, these fertilizers can trigger death. We decided to investigate these findings and use the data to evaluate the actual impact of fertilizer use on Alzheimer's deaths. We chose countries as our experimental units, and defined fertilizer use as the mass of fertilizer used per year in a country. Three fertilizers were chosen for investigation: potash, nitrogen, and phosphate. Our dependent variable, Alzheimer's death rates, was defined by the amount of people who died from Alzheimer's relative to the amount of people who had Alzheimer's in a given country.

There are many variables that are also related to Alzheimer's death rates, like socioeconomic status, access to healthcare, and other lifestyle factors which influence health. Therefore, all these variables would also impact Alzheimer's death rates, and we attempted to control for them in our investigation. We chose to analyze the impact of GDP, life expectancy, and level of development as variables which might confound the impact of fertilizer use on Alzheimer's deaths. By controlling for these variables, we were aiming to see the effect of fertilizer use on Alzheimer's deaths without their impact.

VARIABLES

nitrogen: Kilograms of nitrogen per hectare of cropland in 2015 (quantitative). potash: Kilograms of potash per hectare of cropland in 2015 (quantitative). phosphate: Kilograms of phosphate per hectare of cropland in 2015 (quantitative). life: Life expectancy in 2015 (quantitative). gdp: GDP Per Capita in 2015 (quantitative). alzh: Alzheimer's death rate (quantitative). developed: 0=not developed, 1=developed (qualitative).

nitrogen, potash, phosphate, life, developed, and gdp are our explanatory variables. alzh is our response variable.

DATA

The data set containing the amount of nitrogen, potash, and phosphate fertilizer used by the 195 countries of the world in 2015 was obtained from https://ourworldindata.org/fertilizers

The data set containing the life expectancy at birth for different countries in the world, which served as our measure for the health status of a country, was obtained from https://www.who.int/data/gho/data/indicators/indicator-details/GHO/life-expectancy-at-birth-(years)

The data set containing the GDP per capita for the 195 countries of the world, which served as one measure of the economic status of a country, was obtained from https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2019&most_recent_year_desc=false&start=1961

The data set containing whether the 195 countries in our data set are developed or not, which served as a second measure of the economic status of a country, was obtained from https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf

Lastly, the data set containing rate of death from Alzheimer's disease, our response variable, was obtained from https://www.worldlifeexpectancy.com/cause-of-death/alzheimers-dementia/by-country/

METHODS AND ANALYSIS:

UNIVARIATE EXPLORATION

Favstats of our Variables:

```
favstats (~potash, main ="Figure 1a: Favstats of Potash", data=ds)
```

```
##  min   Q1 median   Q3 max mean sd   n missing
##    0 1.32   9.45 25.8 109 18.6 24 159       0
```

```
favstats (~nitrogen, main ="Figure 1b: Favstats of Nitrogen", data=ds)
```

```
##  min   Q1 median   Q3 max mean   sd   n missing
##    0 8.75   43.3 83.5 400 59.4 64.3 159       0
```

```
favstats (~phosphate, main ="Figure 1c: Favstats of Phosphate", data=ds)
```

```
##  min   Q1 median   Q3  max mean   sd   n missing
##    0 3.13   12.9 25.5 94.1 18.7 20.5 159       0
```

```
favstats (~life, main ="Figure 1d: Favstats of Life Expectancy", data=ds)
```

```
##   min   Q1 median   Q3  max mean  sd   n missing
##  50.5 67.6   73.6 77.5 83.6 72.4 7.1 155       4
```

```
favstats (~gdp, main ="Figure 1e: Favstats of GDP", data=ds)
```

```
##  min   Q1 median    Q3    max  mean    sd   n missing
##  306 2213   5735 15180 102006 13021 18211 159       0
```

```
favstats (~alz, main ="Figure 1f: Favstats of Alzheimer's Death Rate", data=ds)
```

```
##  min Q1 median   Q3  max mean   sd   n missing
##  0.4 14   24.2 31.8 57.6   23 12.9 150       9
```

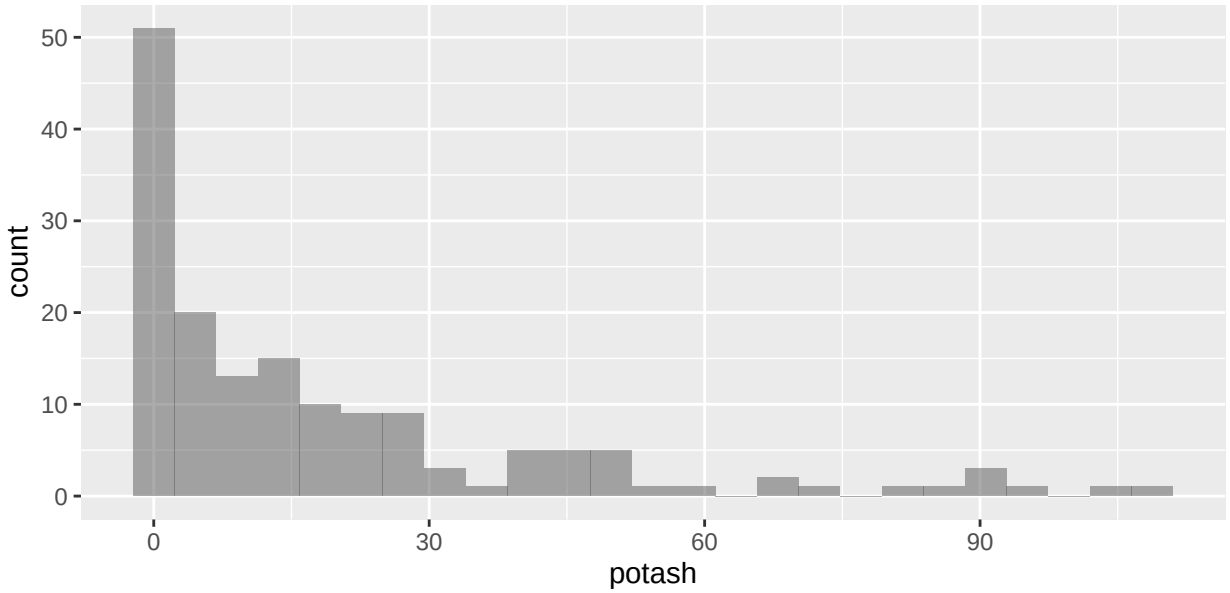
```
favstats (~developed, main ="Figure 1g: Favstats of Developed", data=ds)
```

```
##  min Q1 median Q3 max  mean    sd   n missing
##    0  0      0  0   1 0.195 0.397 159       0
```

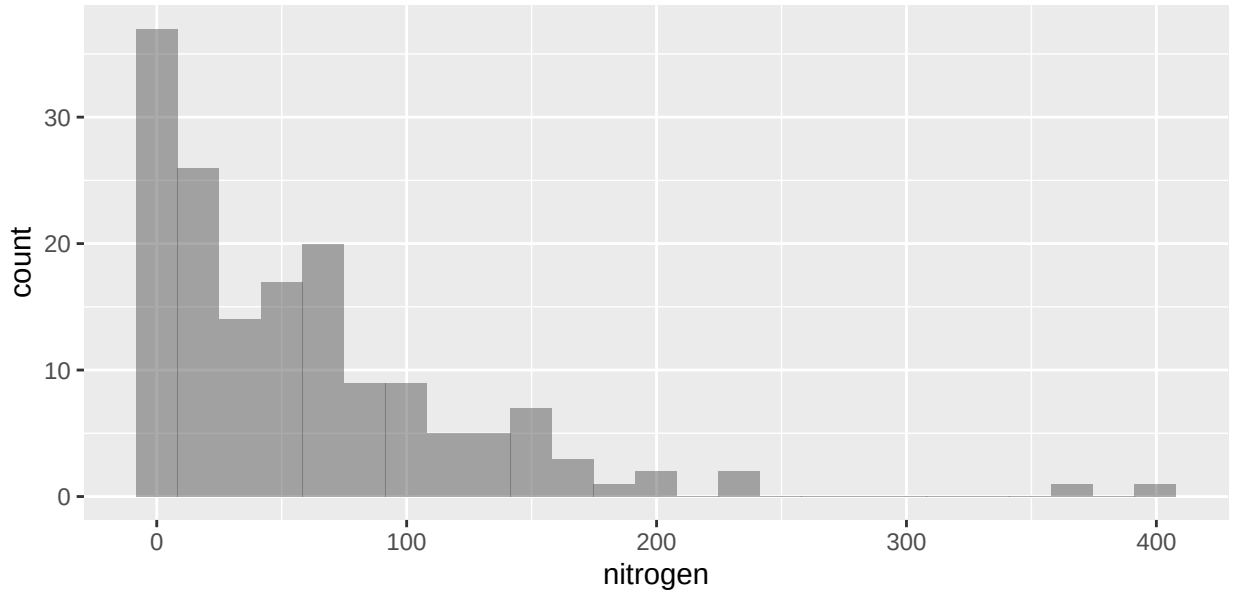
Because all of the data is skewed, as seen in the histograms below, we will use the median and IQR as measurements of center and spread for our data. The median was 9.45 kilograms per hectare for potash, 43.3 kilograms per hectare for nitrogen, 12.9 kilograms per hectare for phosphate, 73.6 years for Life Expectancy, 5735 for GDP per capita, and 24.2 for Alzheimer's death rate.

Histograms of our Quantitative Variables:

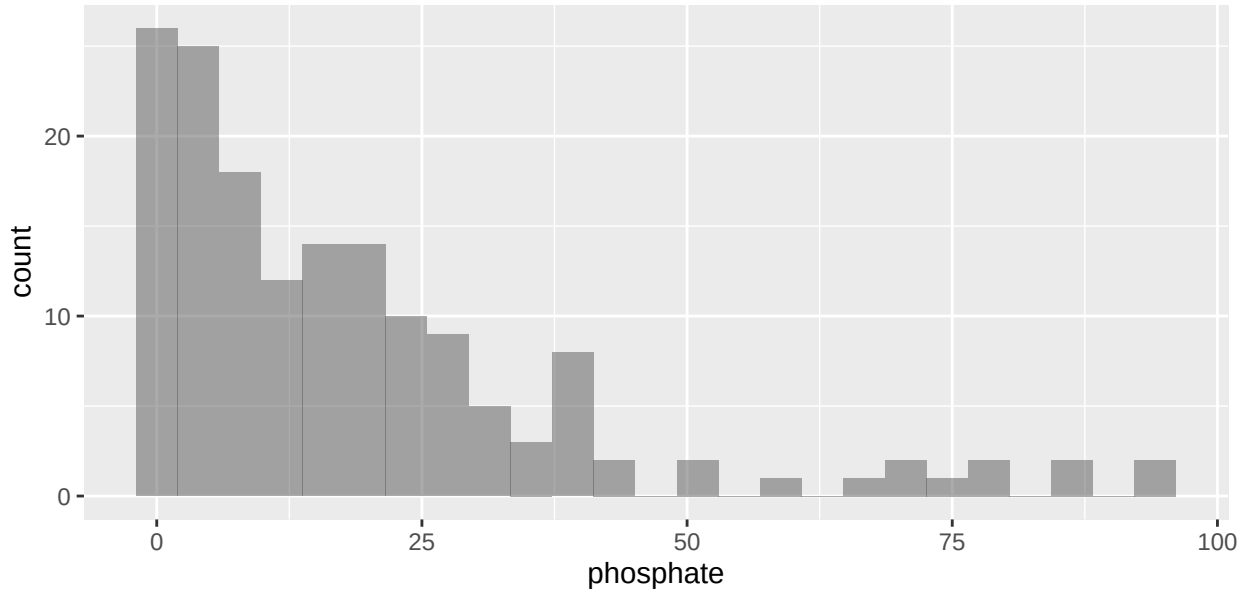
```
gf_histogram (~potash, main="Figure 2a: Histogram of Potash", data=ds)
```



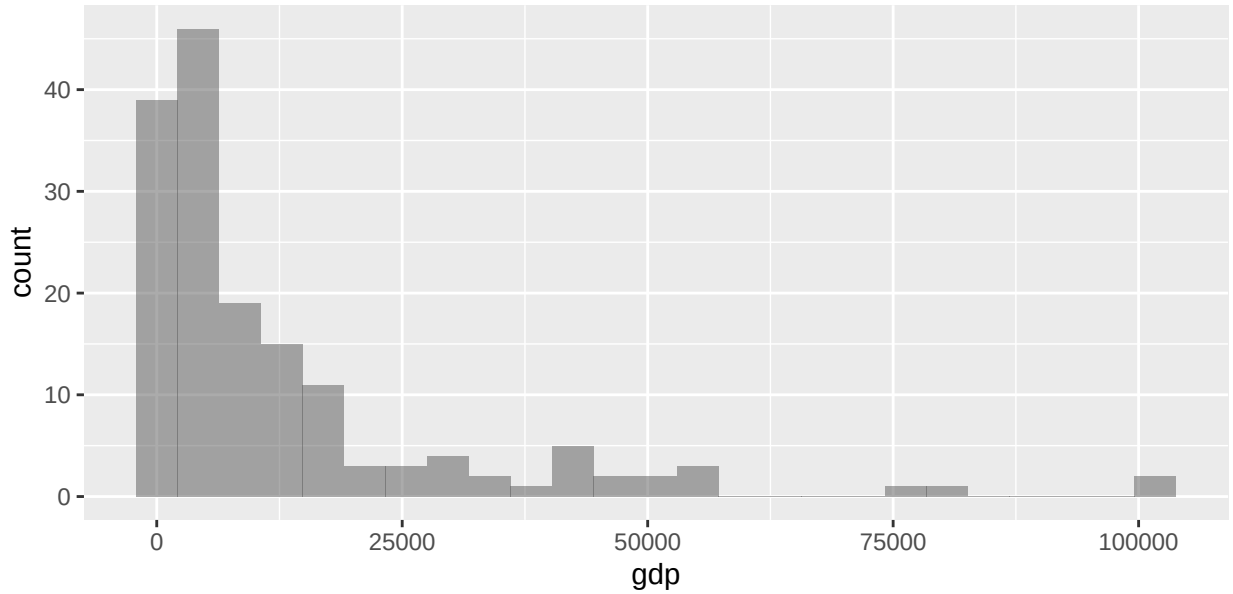

```
gf_histogram (~nitrogen, main="Figure 2b: Histogram of Nitrogen", data=ds)
```




```
gf_histogram (~phosphate, main="Figure 2c: Histogram of Phosphate", data=ds)
```

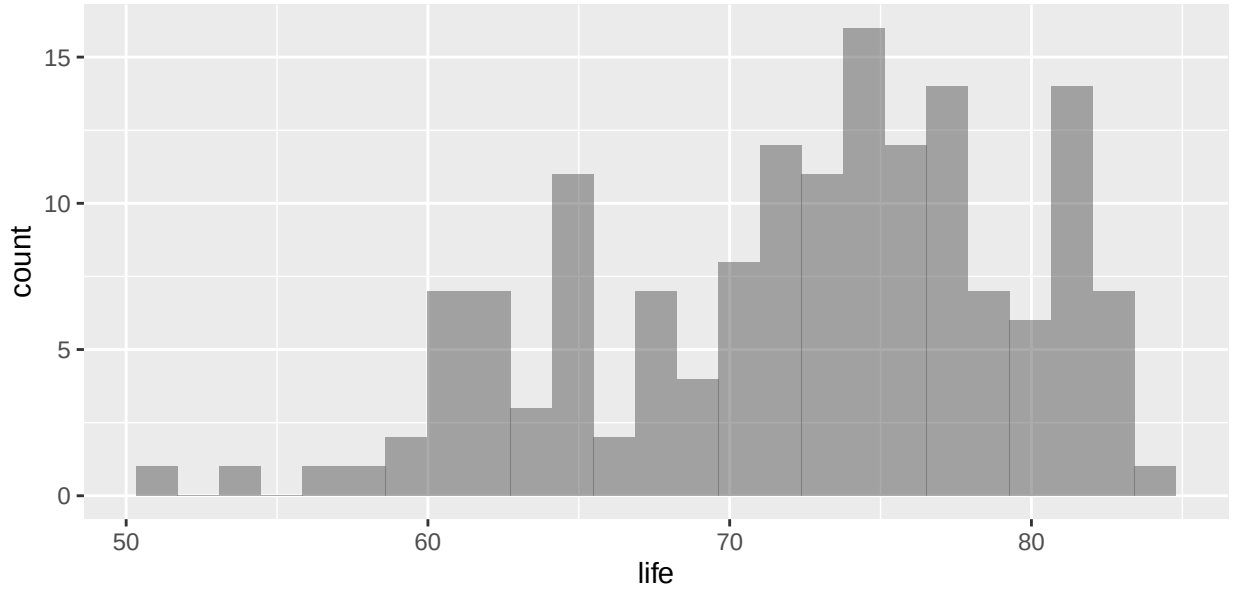



```
gf_histogram (~gdp, main="Figure 2d: Histogram of GDP", data=ds)
```



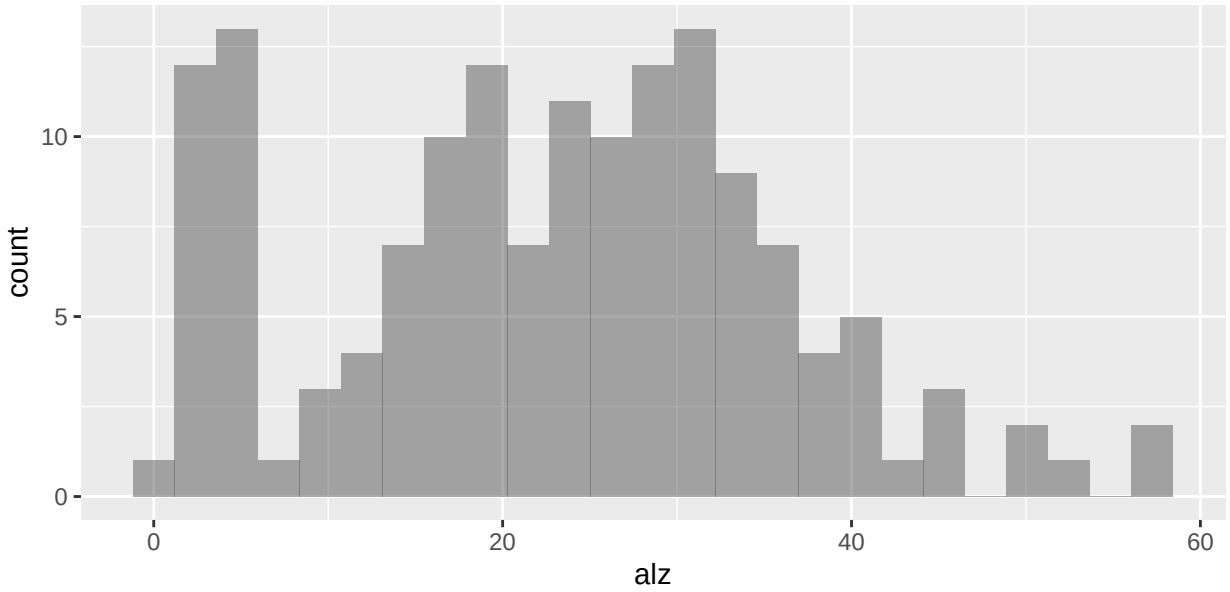

```
gf_histogram (~life, main="Figure 2e: Histogram of Life Expectancy", data=ds)
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```




```
gf_histogram (~alz, main="Figure 2f: Histogram of Alzheimer's Death Rate", data=ds)
```

```
## Warning: Removed 9 rows containing non-finite values (stat_bin).
```



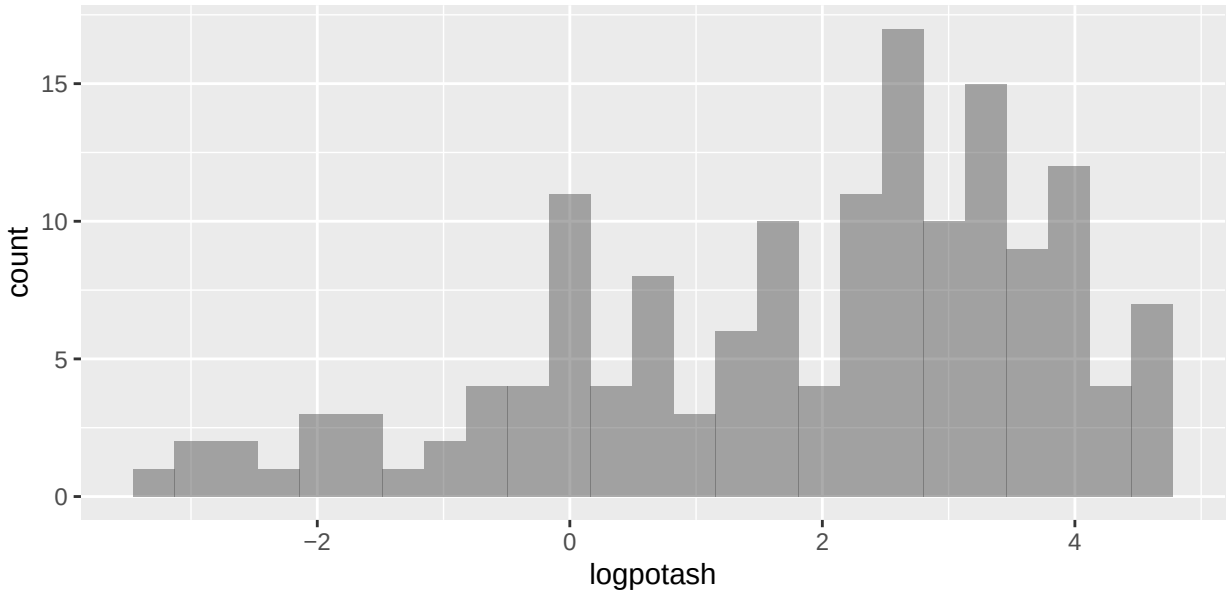
Potash, Nitrogen, Phosphate and Gdp were all very skewed to the right. Life was only slightly skewed to the left but relatively normal, and alzheimers death rates was relatively normal with the exception of it being bimodal (one high mode on the left).

Since the histograms for potash, phosphate, nitrogen, and gdp were skewed, the data was transformed using log and square root.

Histograms of Transformed Data:

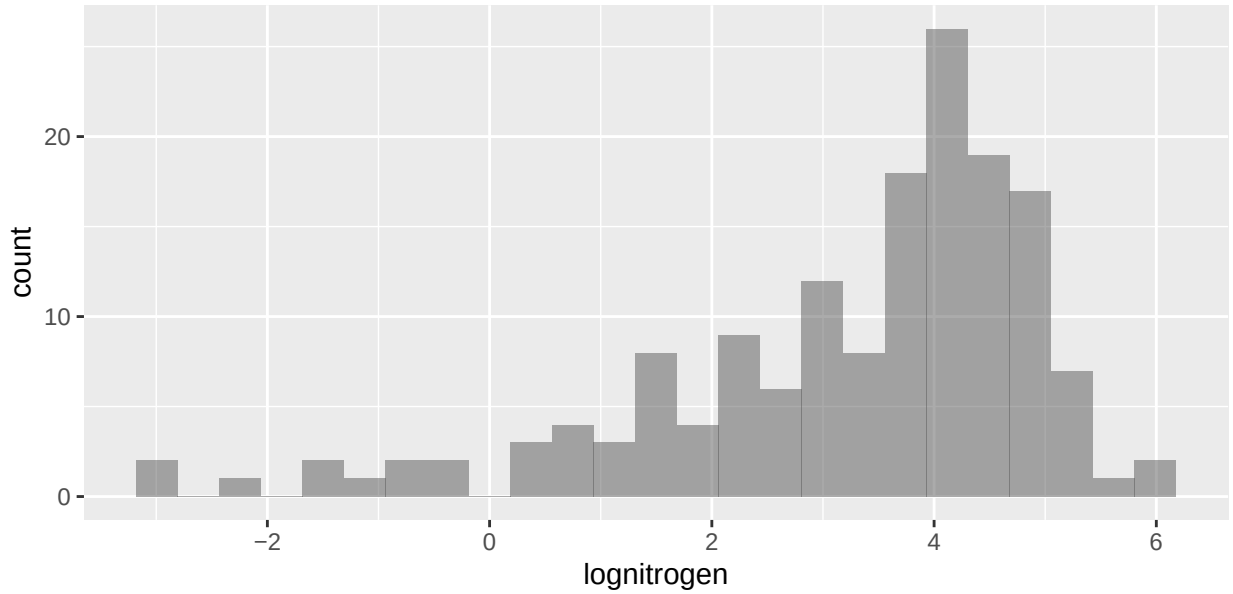
```
gf_histogram(~logpotash, main ="Figure 3a: Histogram of Logpotash", data=ds)
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```



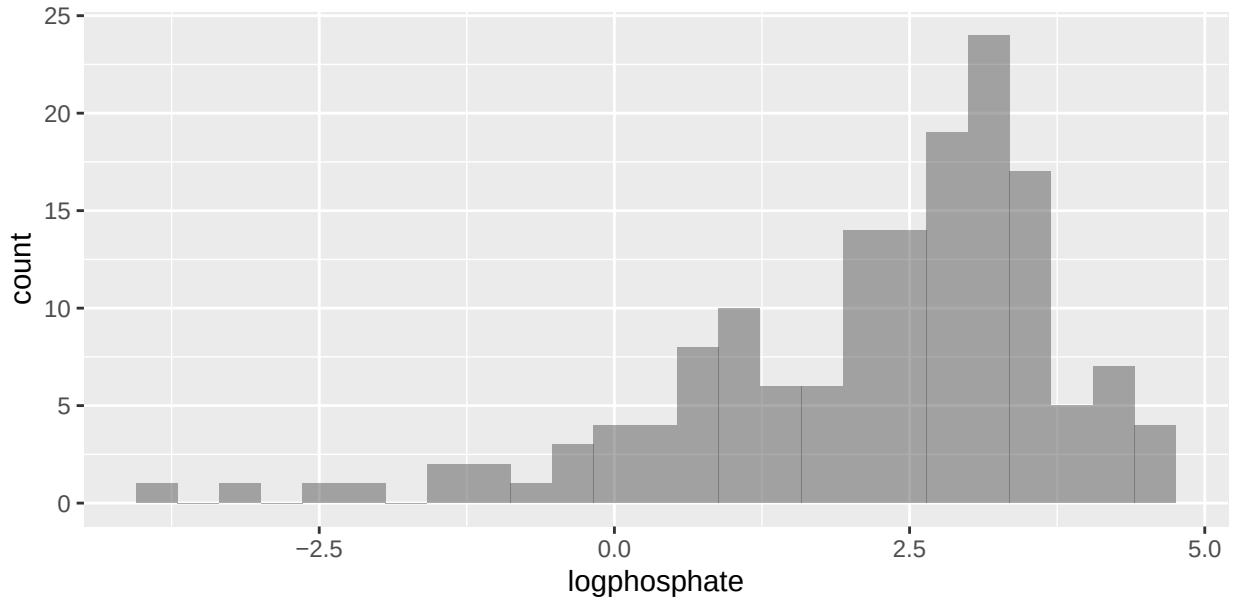

```
gf_histogram (~lognitrogen, main ="Figure 3b: Histogram of Lognitrogen", data=ds)
```

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

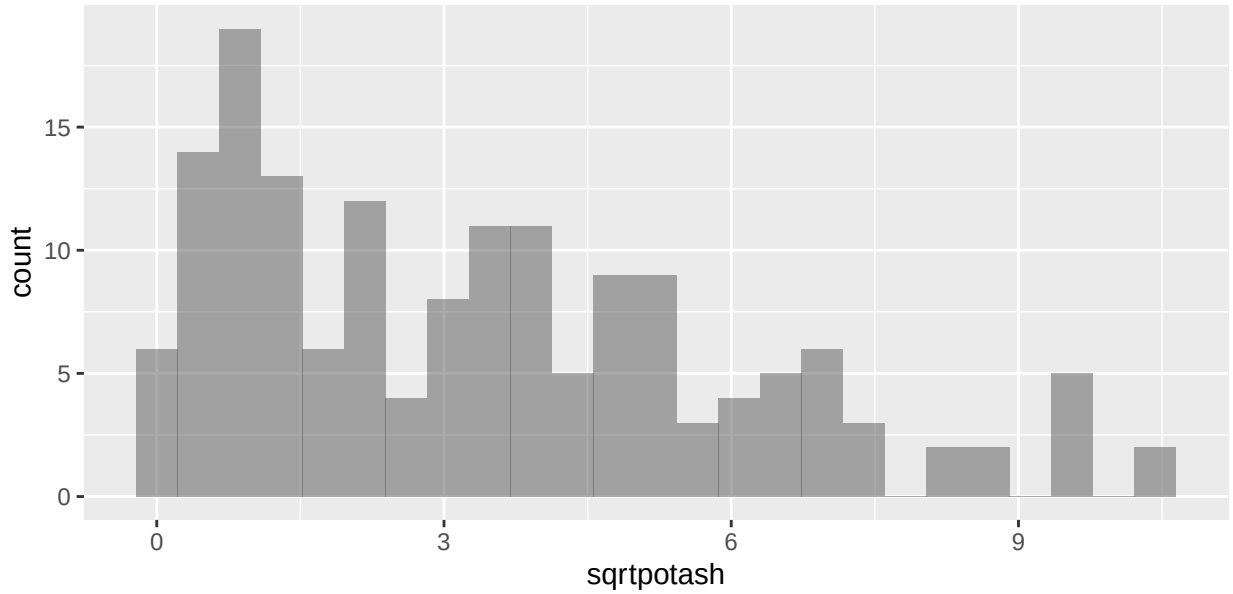



```
gf_histogram (~logphosphate, main ="Figure 3c: Histogram of Logphosphate", data=ds)
```

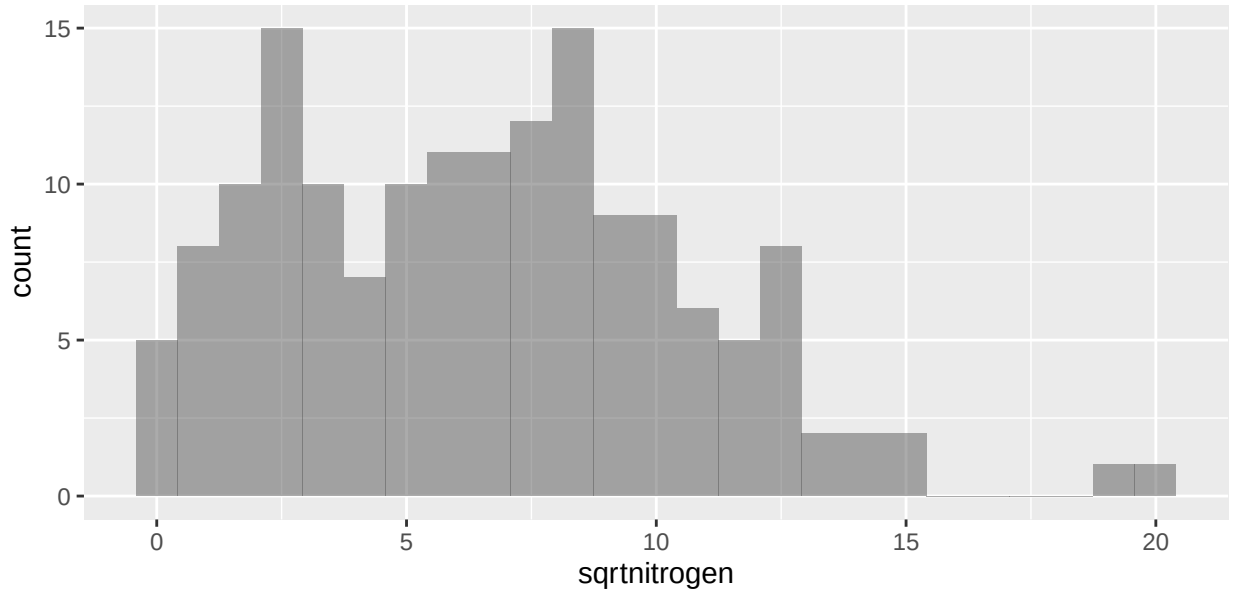
```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```



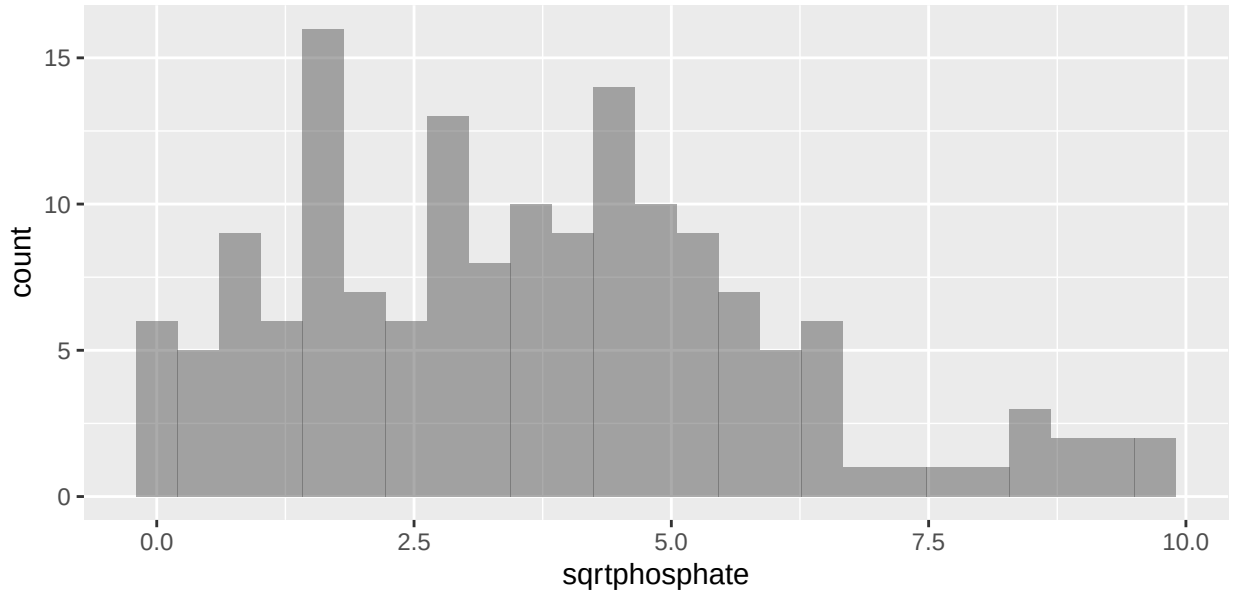

```
gf_histogram(~sqrtpotash, main ="Figure 3d: Histogram of Sqrtpotash", data=ds)
```



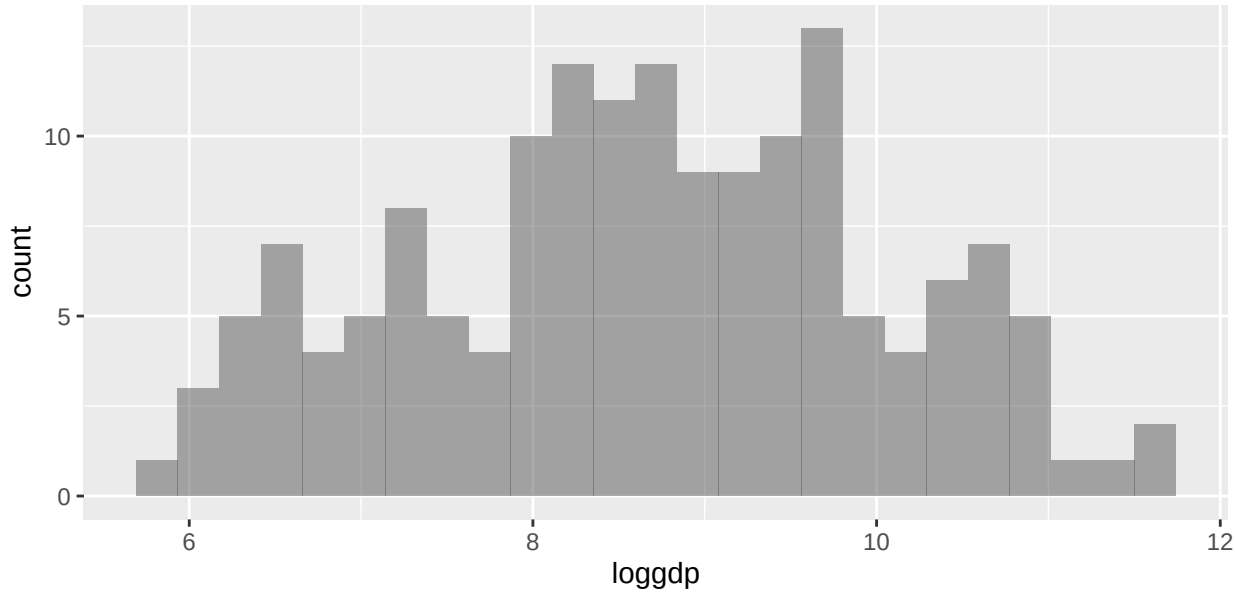

```
gf_histogram (~sqrtnitrogen, main ="Figure 3e: Histogram of Sqrtnitrogen", data=ds)
```




```
gf_histogram (~sqrtphosphate, main ="Figure 3f: Histogram of Sqrtphosphate", data=ds)
```




```
gf_histogram (~loggdp, main ="Figure 3g: Histogram of Loggdp", data=ds)
```



The histograms after the transformations all look more normal, and the favstats of the transformed data are recorded below. We can now use the mean and standard deviation, as recorded, as measures of center and spread.

```
favstats(~logpotash, main ="Figure 4a: Favstats of Logpotash", data=ds)
```

```
##   min    Q1 median   Q3  max mean  sd   n missing
##  -Inf 0.277   2.25 3.25 4.69 -Inf NaN 159       0
```

```
favstats (~lognitrogen, main ="Figure 4b: Favstats of Lognitrogen", data=ds)
```

```
##   min   Q1 median   Q3  max mean  sd   n missing
##  -Inf 2.17   3.77 4.43 5.99 -Inf NaN 159       0
```

```
favstats (~logphosphate, main ="Figure 4c: Favstats of Logphosphate", data=ds)
```

```
##   min   Q1 median   Q3  max mean  sd   n missing
##  -Inf 1.14   2.56 3.24 4.54 -Inf NaN 159       0
```

```
favstats(~sqrtpotash, main ="Figure 4d: Favstats of Sqrtpotash", data=ds)
```

```
##  min   Q1 median   Q3  max mean   sd   n missing
##    0 1.15   3.07 5.08 10.4 3.44 2.61 159       0
```

```
favstats (~sqrtnitrogen, main ="Figure 4e: Favstats of Sqrtnitrogen", data=ds)
```

```
##  min   Q1 median   Q3 max mean   sd   n missing
##    0 2.96   6.58 9.14  20 6.57 4.04 159       0
```

```
favstats (~sqrtphosphate, main ="Figure 4f: Favstats of Sqrtphosphate", data=ds)
```

```
##  min   Q1 median   Q3 max mean   sd   n missing
##    0 1.77    3.6 5.05 9.7 3.68 2.28 159       0
```

```
favstats (~loggdp, main ="Figure 4g: Favstats of Loggdp", data=ds)
```

```
##   min  Q1 median   Q3  max mean   sd   n missing
##  5.72 7.7   8.65 9.63 11.5 8.66 1.36 159       0
```

Our Qualitative Variable (developed):

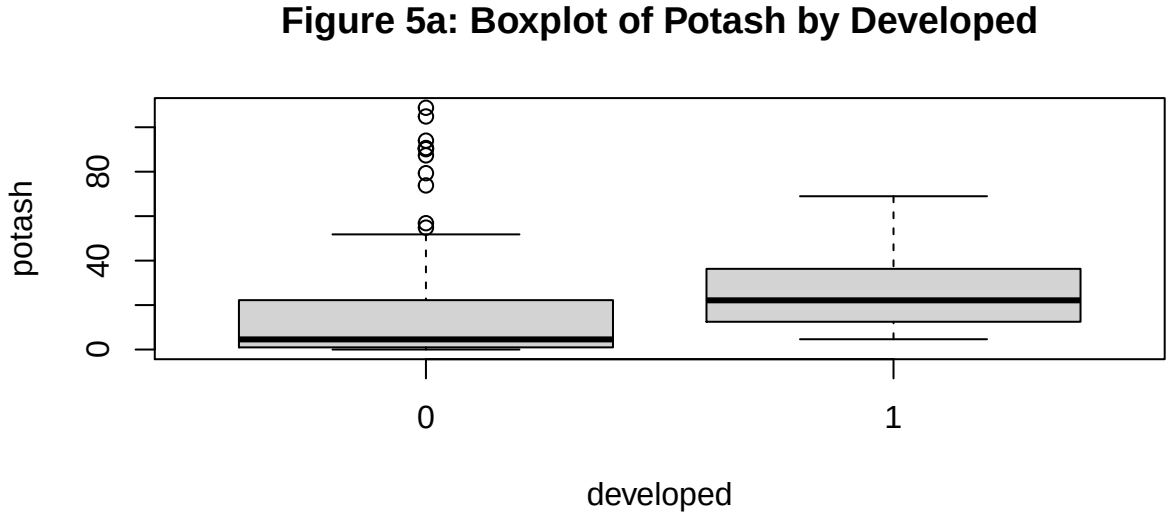
```
tally (~developed, data=ds)
```

```
## developed
##   0   1
## 128  31
```

There were 31 developed countries in our model and 128 countries that are not developed. Even so, there was a significant difference between developed countries and non-developed countries for all the variables as seen in the boxplots below.

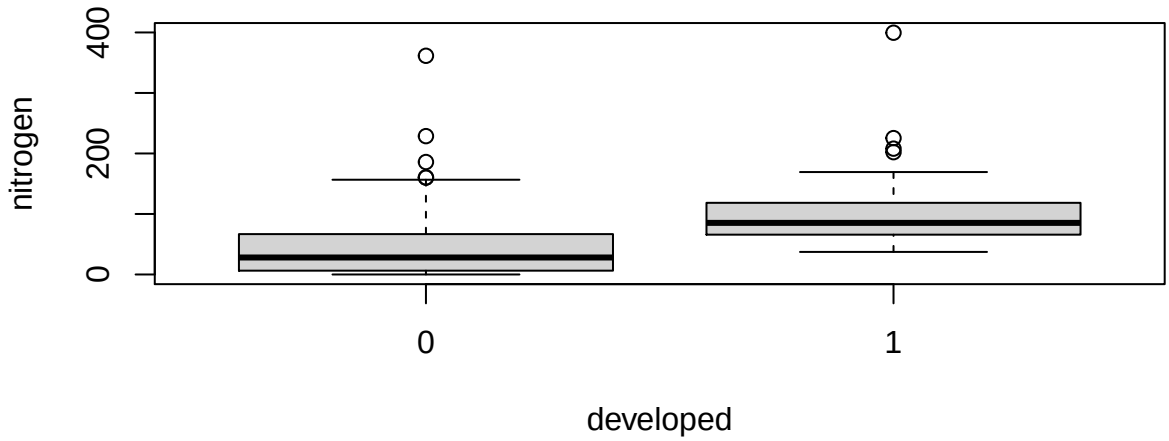
Box and Whisker plots of our Qualitative Variable against our Quantitative Variables.

```
boxplot(potash ~ developed, main ="Figure 5a: Boxplot of Potash by Developed", data=ds)
```



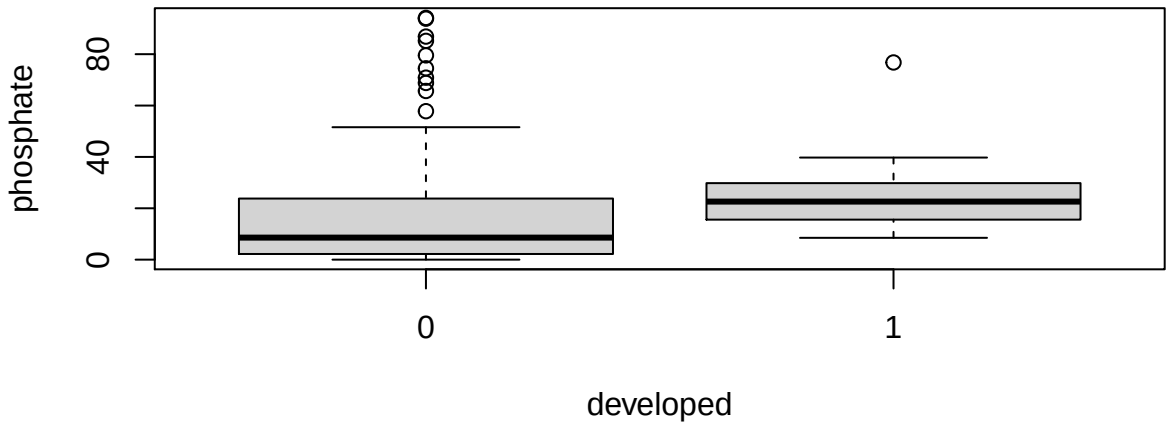

```
boxplot(nitrogen ~ developed, main ="Figure 5b: Boxplot of Nitrogen by Developed", data=ds)
```

Figure 5b: Boxplot of Nitrogen by Developed



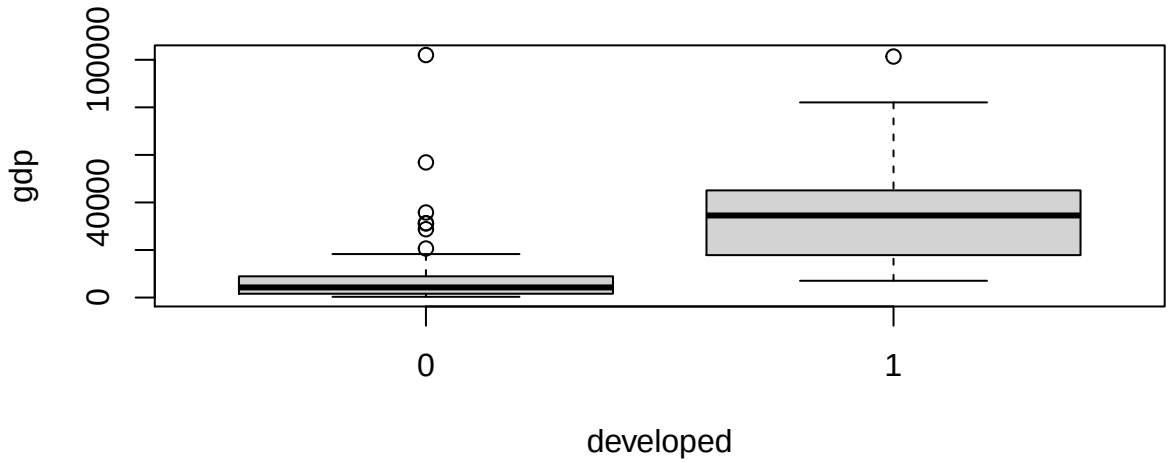

```
boxplot(phosphate ~ developed, main ="Figure 5c: Boxplot of Phosphate by Developed", data=ds)
```

Figure 5c: Boxplot of Phosphate by Developed



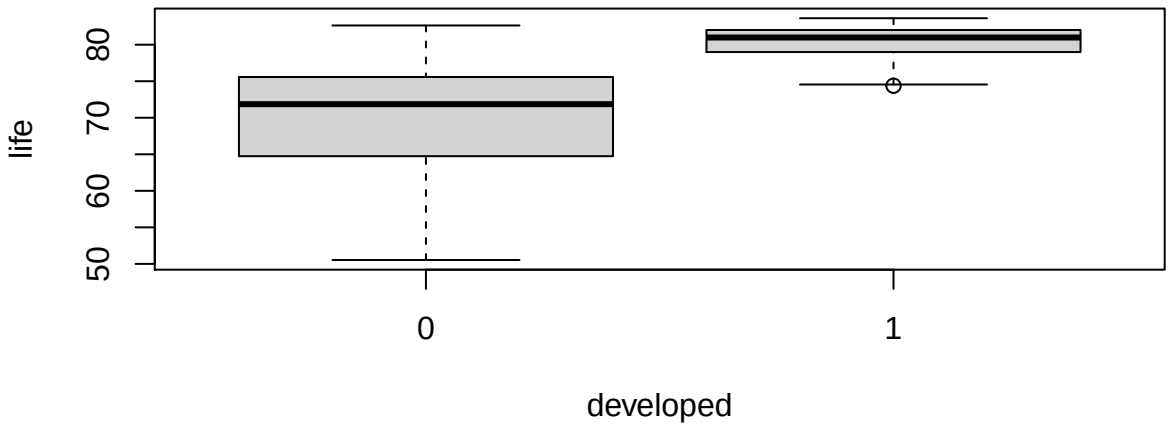

```
boxplot(gdp ~ developed, main ="Figure 5d: Boxplot of GDP by Developed", data=ds)
```

Figure 5d: Boxplot of GDP by Developed



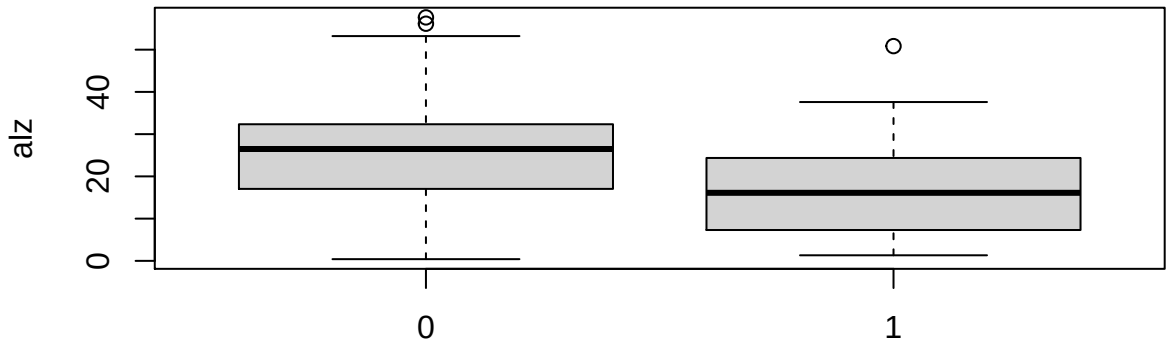
```
boxplot(life ~ developed, main ="Figure 5e: Boxplot of Life Expectancy by Developed", data=ds)
```

Figure 5e: Boxplot of Life Expectancy by Developed




```
boxplot(alz ~ developed, main ="Figure 5f: Boxplot of Alzheimer's Death Rates by Developed", data=ds)
```

Figure 5f: Boxplot of Alzheimer's Death Rates by Developed



The median of potash use, nitrogen use, phosphate use, GDP, and Life expectancy in developed countries was higher than in countries that are not developed. The median for Alzheimer's death rate was lower in developed countries. Specific comparisons can be seen in the favstats below.

Favstats of our Quantitative Variable by whether the country is developed or not:

```
favstats(potash ~ developed, main = "Figure 6a: Favstats of Potash by Developed", data=ds)
```

```
##   developed  min    Q1 median   Q3 max mean   sd   n missing
## 1         0 0.00  0.97   4.56 21.7 109 16.7 25.1 128       0
## 2         1 4.64 12.50  22.13 36.3  69 26.2 17.3  31       0
```

```
favstats(nitrogen ~ developed, main = "Figure 6b: Favstats of Nitrogen by Developed", data=ds)
```

```
##   developed  min    Q1 median    Q3 max  mean   sd   n missing
## 1         0  0.0  6.57   28.1  65.6 361  47.7 56.5 128       0
## 2         1 37.3 65.80   85.3 118.4 400 107.6 72.3  31       0
```

```
favstats(phosphate ~ developed, main = "Figure 6c: Favstats of Phosphate by Developed", data=ds)
```

```
##   developed min    Q1 median   Q3  max mean   sd   n missing
## 1         0 0.0  2.21   8.59 23.5 94.1 17.4 21.8 128       0
## 2         1 8.5 15.58  22.61 29.8 76.8 24.2 13.1  31       0
```

```
favstats(gdp ~ developed, main = "Figure 6d: Favstats of GDP by Developed", data=ds)
```

```
##   developed  min    Q1 median    Q3    max  mean    sd   n missing
## 1         0  306  1605   4256  8869 102006  7506 11649 128       0
## 2         1 7056 17845  34524 45075 101376 35792 22508  31       0
```

```
favstats(life ~ developed, main = "Figure 6e: Favstats of Life Expectancy by Developed", data=ds)
```

```
##   developed  min   Q1 median   Q3  max mean   sd   n missing
## 1         0 50.5 64.8   71.9 75.6 82.6 70.4 6.52 124       4
## 2         1 74.4 79.0   81.0 82.0 83.6 80.2 2.63  31       0
```

```
favstats(alz ~ developed, main = "Figure 6f: Favstats of Alzheimer's Death Rate by Developed", data=ds)
```

```
##   developed  min   Q1 median   Q3  max mean   sd   n missing
## 1         0 0.40 17.0   26.5 32.3 57.6 24.5 13.0 119       9
## 2         1 1.32  7.3   16.1 24.4 50.8 17.5 11.5  31       0
```

In general, there appears to be a significant difference across all the statistics between countries labeled as developed vs countries labeled as not developed. We will run further analysis on this information as needed.

BIVARIATE ASSOCIATIONS

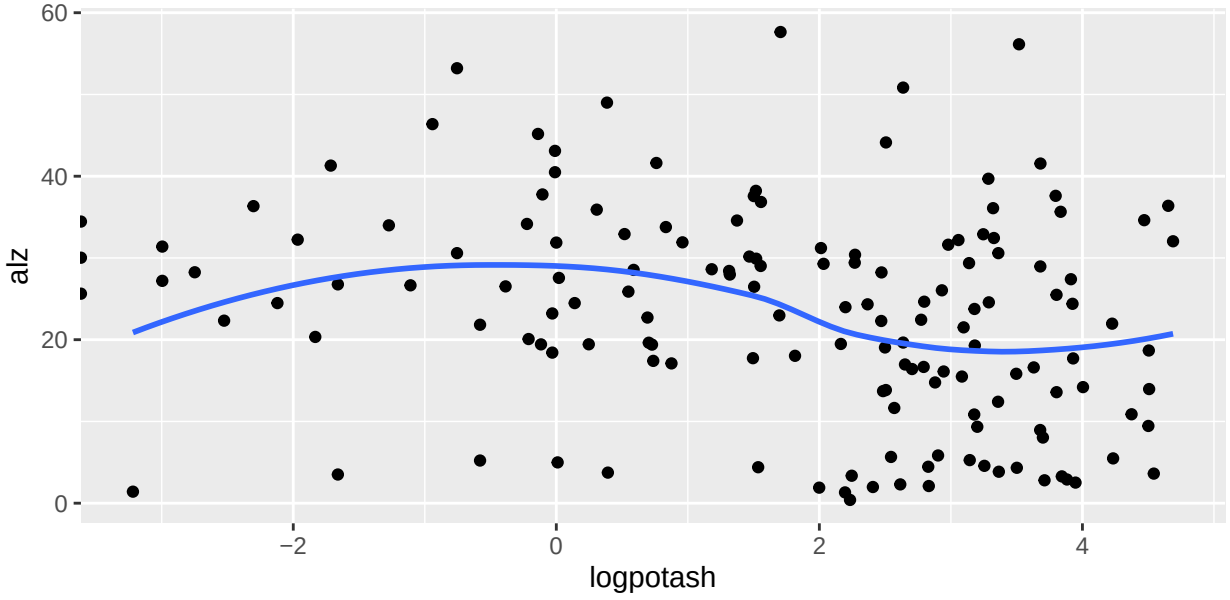
Scatterplots of our Quantitative Predictor Variables against our Response Variable:

```
gf_point(alz ~ logpotash, main ="Figure 7a: Scatterplot of alz by logpotash",data=ds) %>% gf_smooth(se=
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 12 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```




```
cor (alz ~ logpotash, data=ds)
```

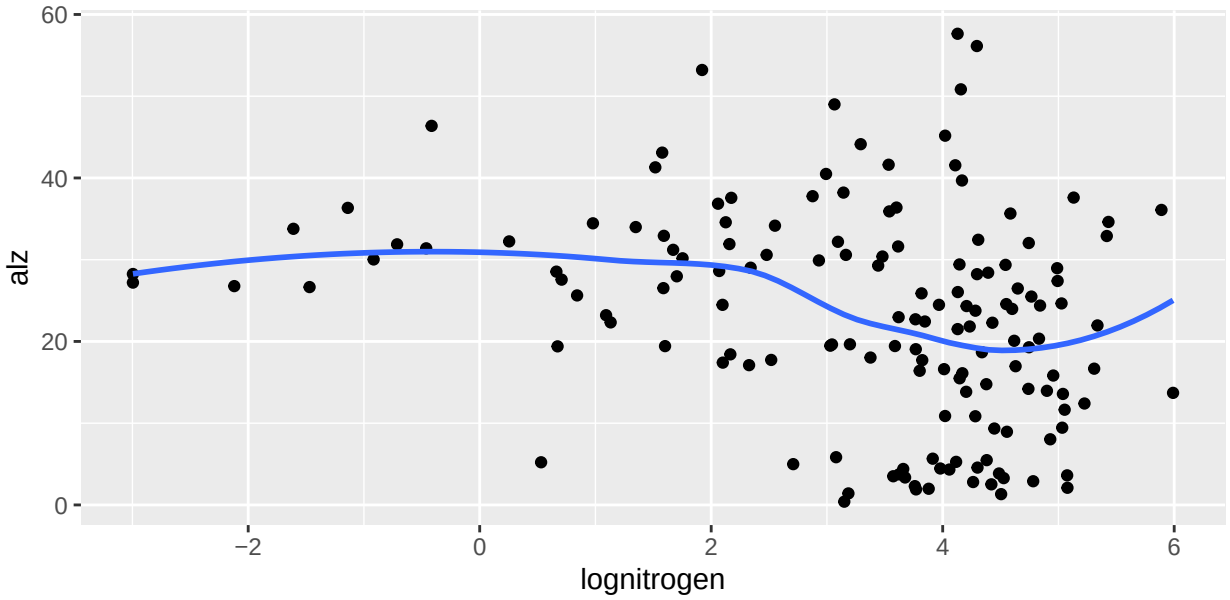
```
## [1] NA
```

```
gf_point(alz ~ lognitrogen, main="Figure 7b: Scatterplot of alz by lognitrogen", data=ds)%>% gf_smooth(
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 9 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```




```
cor (alz ~lognitrogen, data=ds)
```

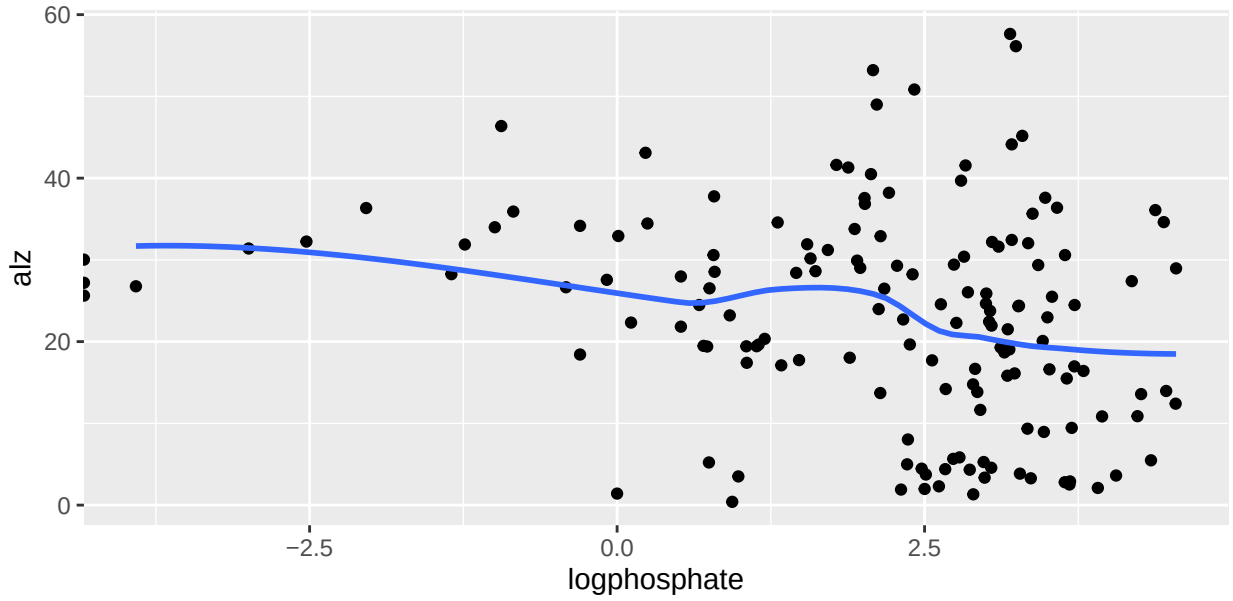
```
## [1] NA
```

```
gf_point(alz ~logphosphate, main= "Figure 7c: Scatterplot of alz by logphosphate", data=ds) %>% gf_smoo
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 12 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```




```
cor (alz ~logphosphate, data=ds)
```

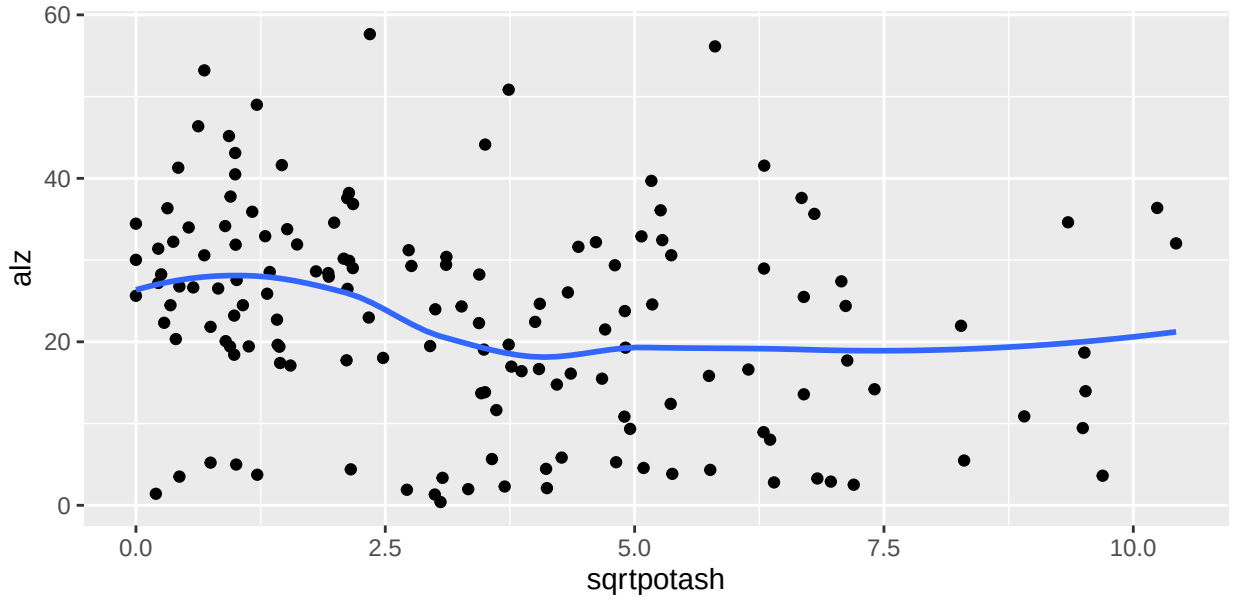
```
## [1] NA
```

```
gf_point(alz ~sqrtpotash, main ="Figure 7d: Scatterplot of alz by sqrtpotash",data=ds) %>% gf_smooth(se=
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 9 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```




```
cor (alz ~sqrtpotash, data=ds)
```

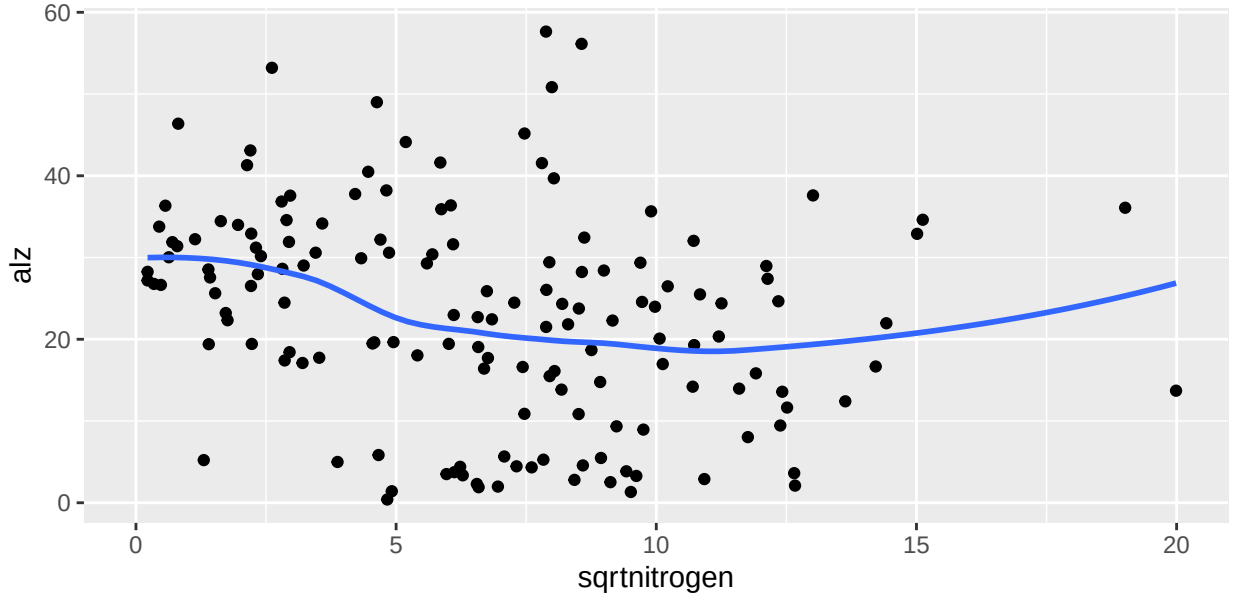
```
## [1] NA
```

```
gf_point(alz ~sqrtnitrogen, main="Figure 7e: Scatterplot of alz by sqrtnitrogen", data=ds)%>% gf_smooth
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 9 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```




```
cor (alz ~sqrtnitrogen, data=ds)
```

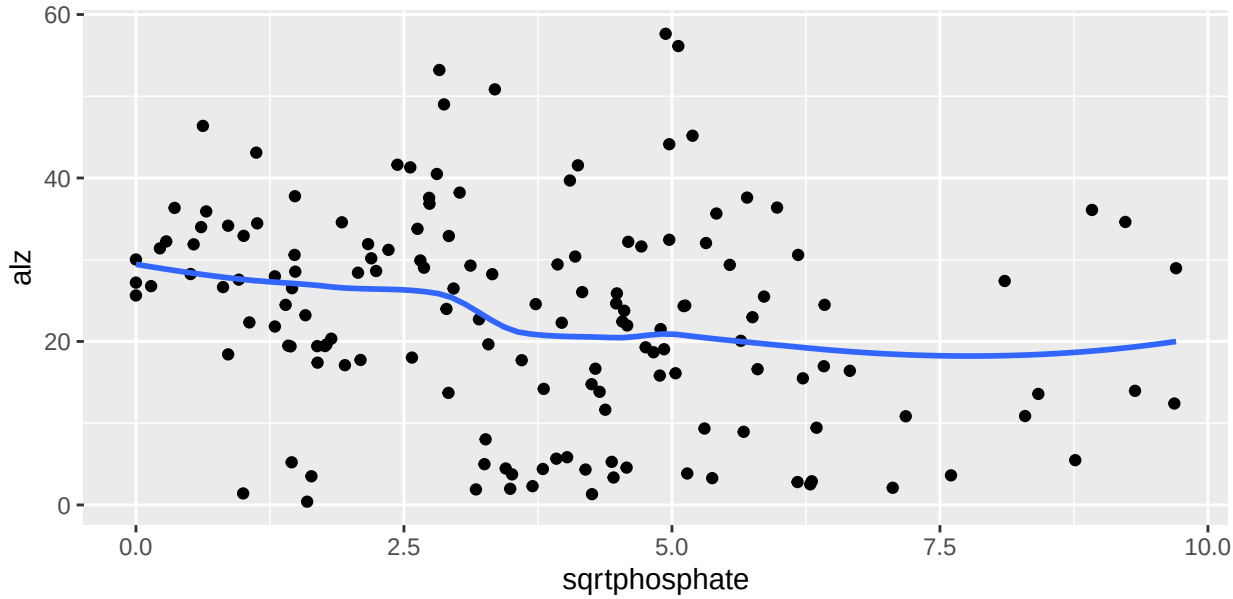
```
## [1] NA
```

```
gf_point(alz ~sqrtphosphate, main= "Figure 7f: Scatterplot of alz by sqrtphosphate", data=ds) %>% gf_sm
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 9 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```



```
cor (alz ~sqrtphosphate, data=ds)
```

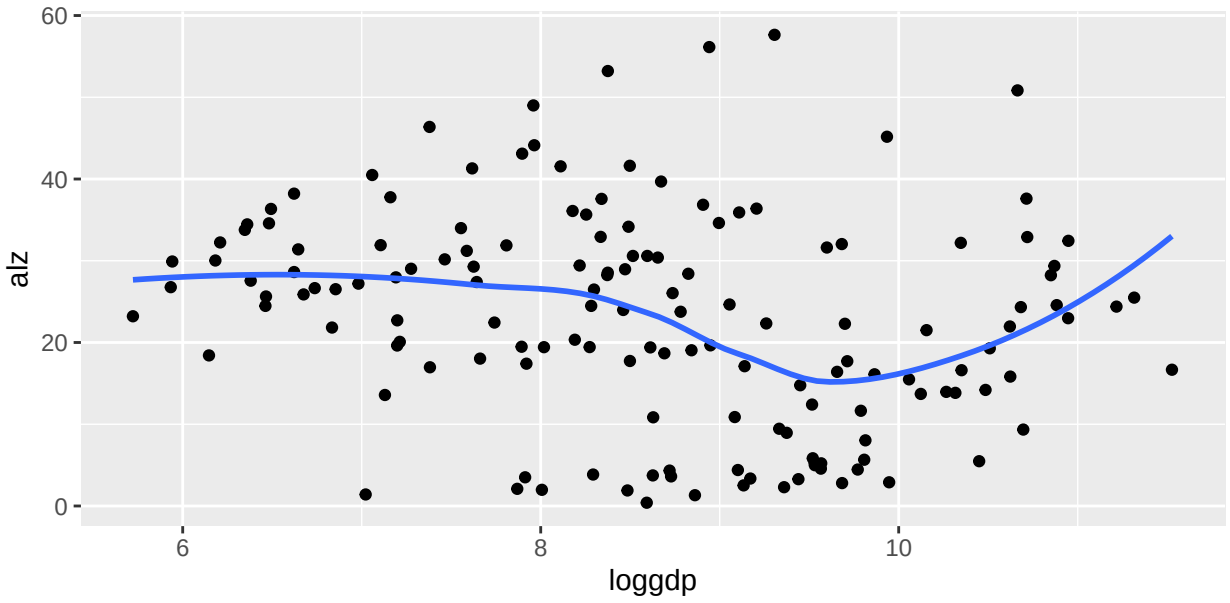
```
## [1] NA
```

```
gf_point(alz ~loggdp, main="Figure 7g: Scatterplot of alz by loggdp", data=ds) %>% gf_smooth(se=FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 9 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```




```
cor (alz ~loggdp, data=ds)
```

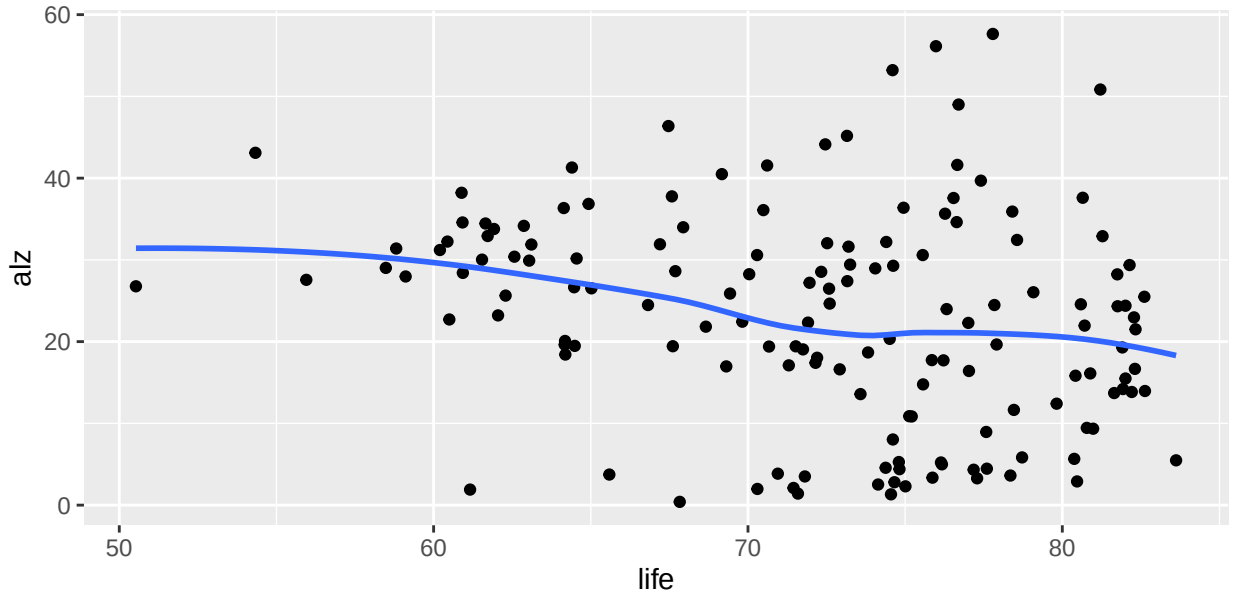
```
## [1] NA
```

```
gf_point(alz ~life, main="Figure 7h: Scatterplot of alz by life", data=ds) %>% gf_smooth(se=FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 10 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 10 rows containing missing values (geom_point).
```




```
cor (alz ~loggdp, data=ds)
```

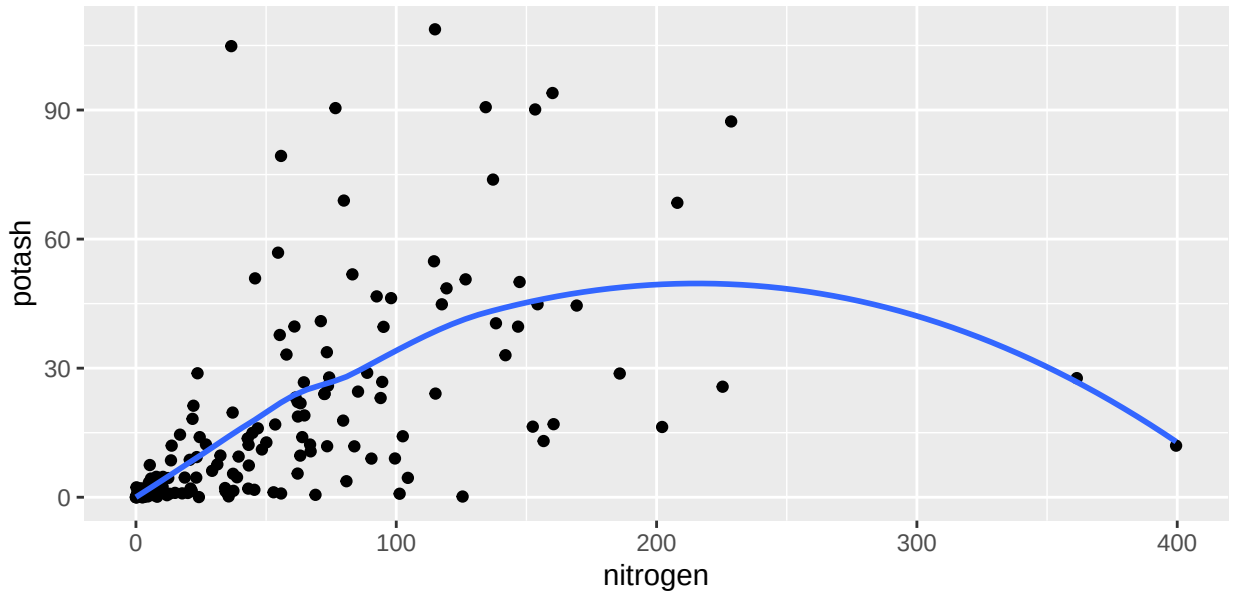
```
## [1] NA
```

There seems to be a slight relationship between our predictor variables and the response variable. We need to analyze the strength in our linear model in order to find significant predictors.

Before proceeding with the linear model, we decided to run the potash, phosphate, and nitrogen, as well as life and gdp, against each other to test for colinearity.

```
gf_point(potash ~ nitrogen, main ="Figure 8a: Scatterplot of potash by nitrogen",data=ds) %>% gf_smooth
```

```
## `geom_smooth()` using method = 'loess'
```

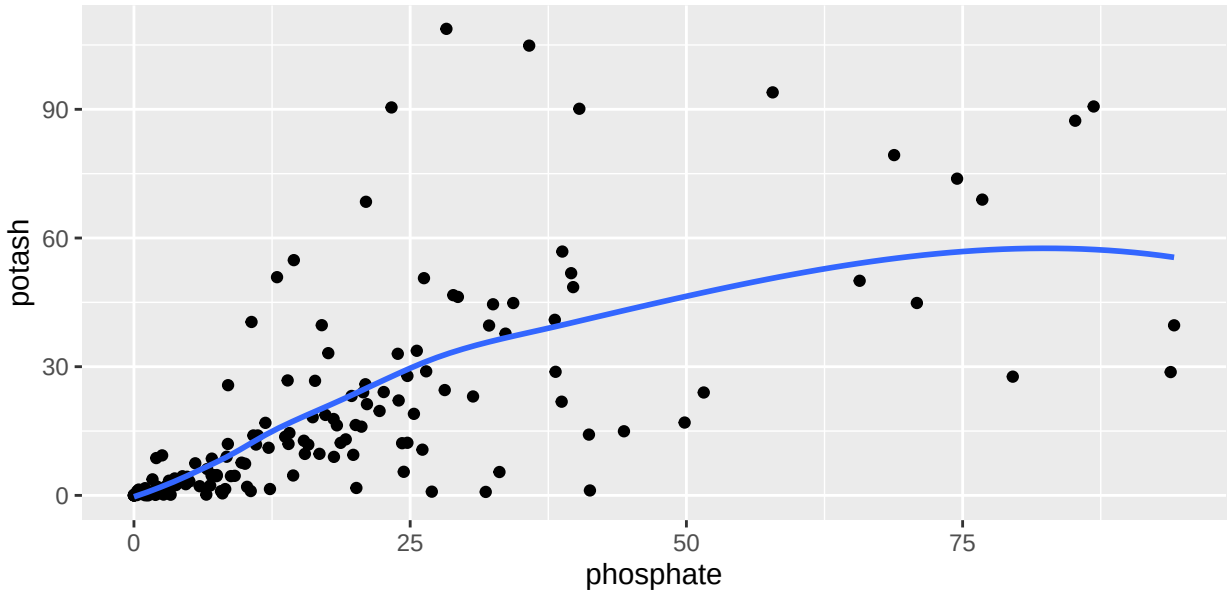



```
cor (potash ~ nitrogen, data=ds)
```

```
## [1] 0.516
```

```
gf_point(potash ~ phosphate, main="Figure 8b: Scatterplot of potash by phosphate", data=ds)%>% gf_smooth
```

```
## `geom_smooth()` using method = 'loess'
```

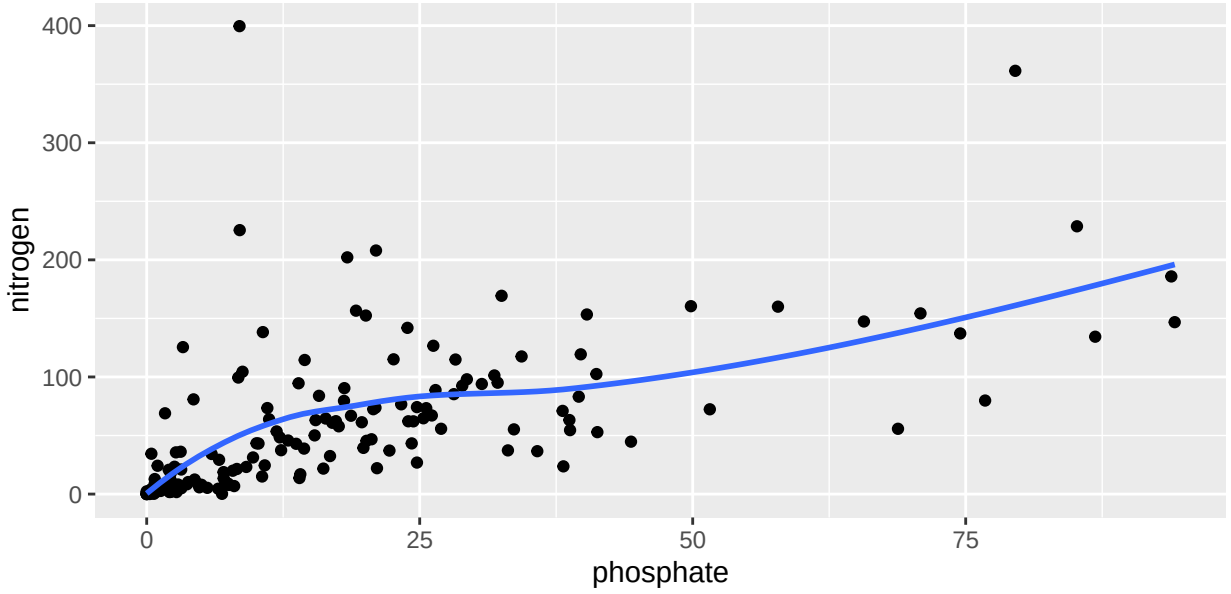



```
cor (potash ~ phosphate, data=ds)
```

```
## [1] 0.685
```

```
gf_point(nitrogen ~ phosphate, main= "Figure 8c: Scatterplot of nitrogen by phosphate", data=ds) %>% gf_
```

```
## `geom_smooth()` using method = 'loess'
```

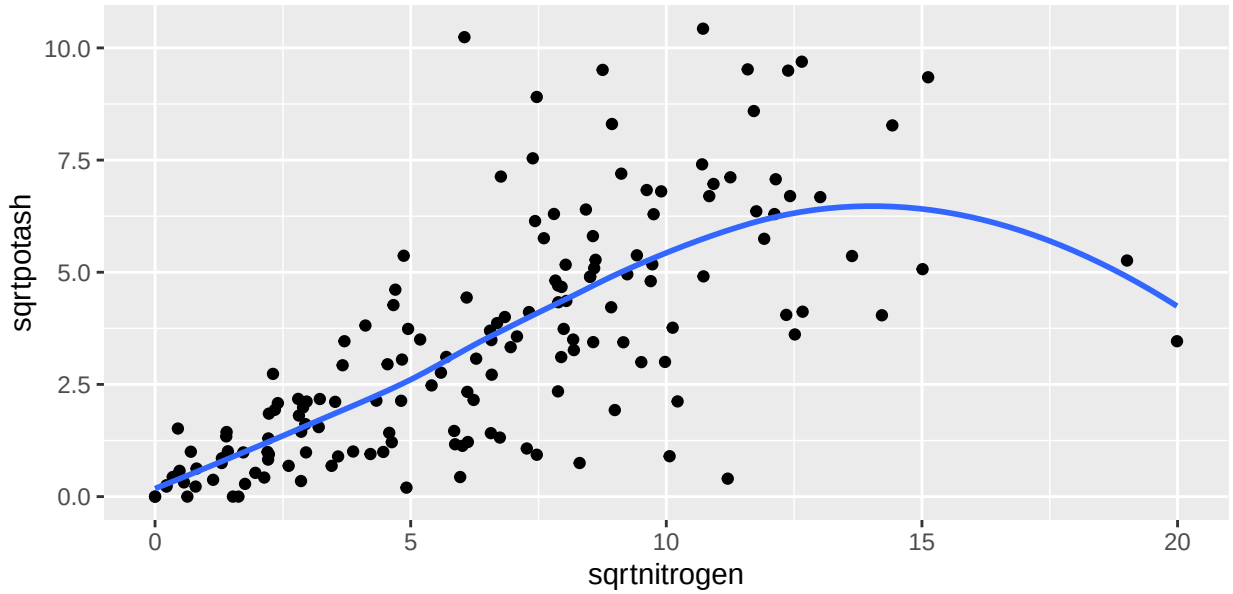



```
cor (nitrogen ~ phosphate, data=ds)
```

```
## [1] 0.6
```

```
gf_point(sqrtpotash ~ sqrtnitrogen, main ="Figure 8d: Scatterplot of sqrtpotash by sqrtnitrogen",data=ds
```

```
## `geom_smooth()` using method = 'loess'
```

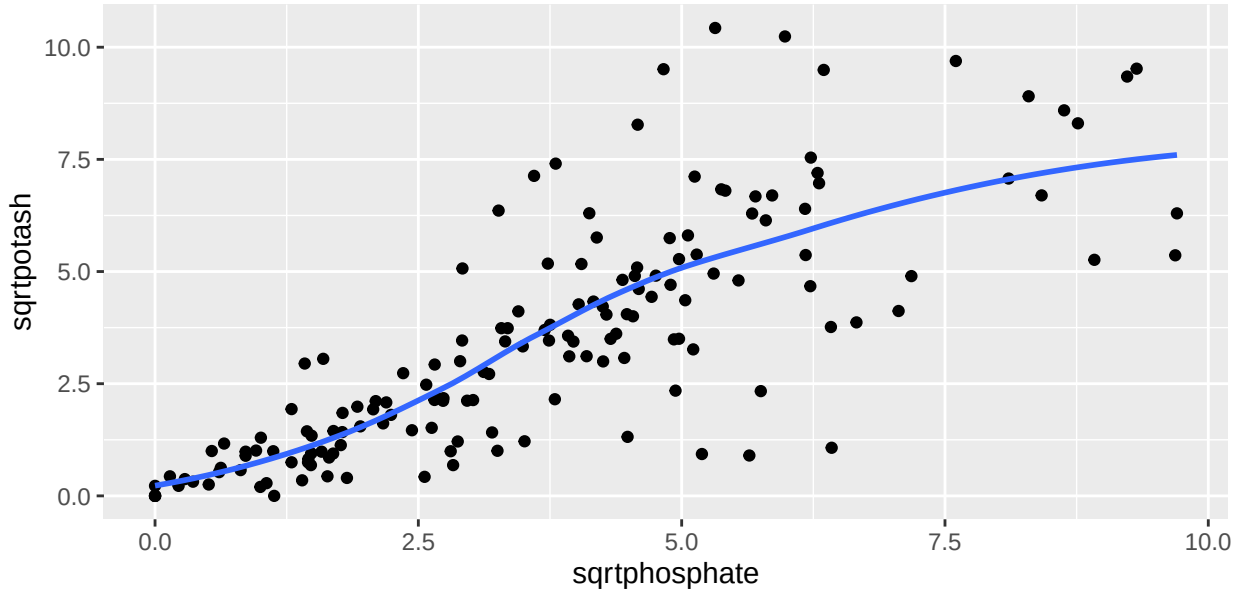



```
cor (sqrtpotash ~ sqrtnitrogen, data=ds)
```

```
## [1] 0.698
```

```
gf_point(sqrtpotash ~ sqrtphosphate, main="Figure 8e: Scatterplot of sqrtpotash by sqrtphosphate", data=
```

```
## `geom_smooth()` using method = 'loess'
```

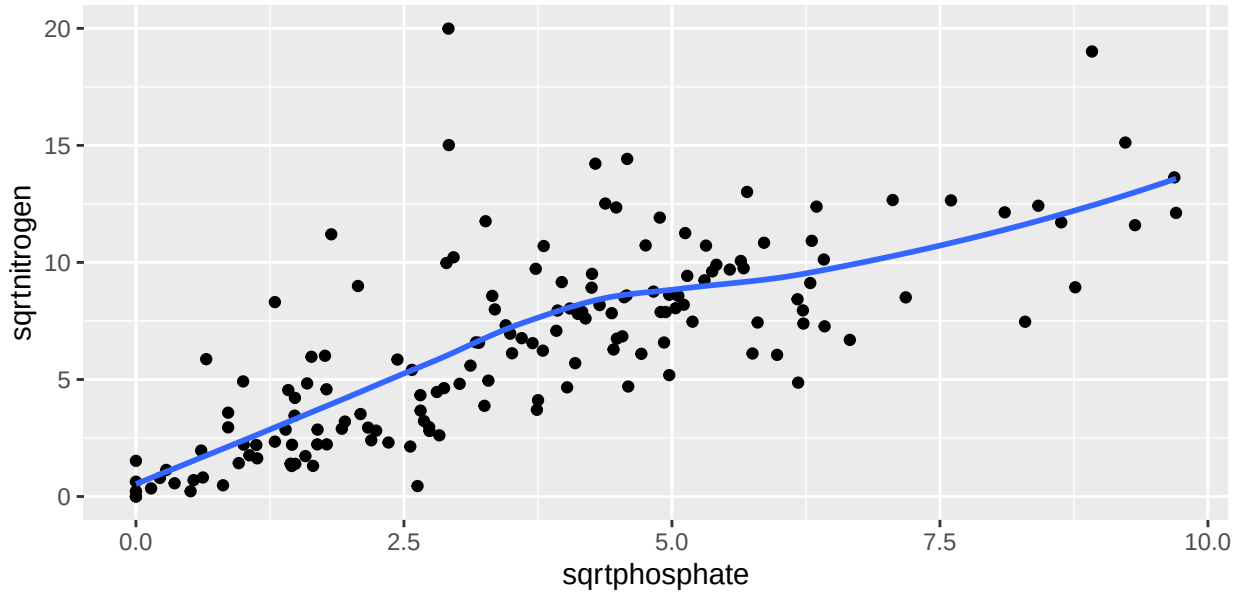



```
cor (sqrtpotash ~ sqrtphosphate, data=ds)
```

```
## [1] 0.796
```

```
gf_point(sqrtnitrogen ~ sqrtphosphate, main= "Figure 8f: Scatterplot of sqrtnitrogen by sqrtphosphate",
```

```
## `geom_smooth()` using method = 'loess'
```

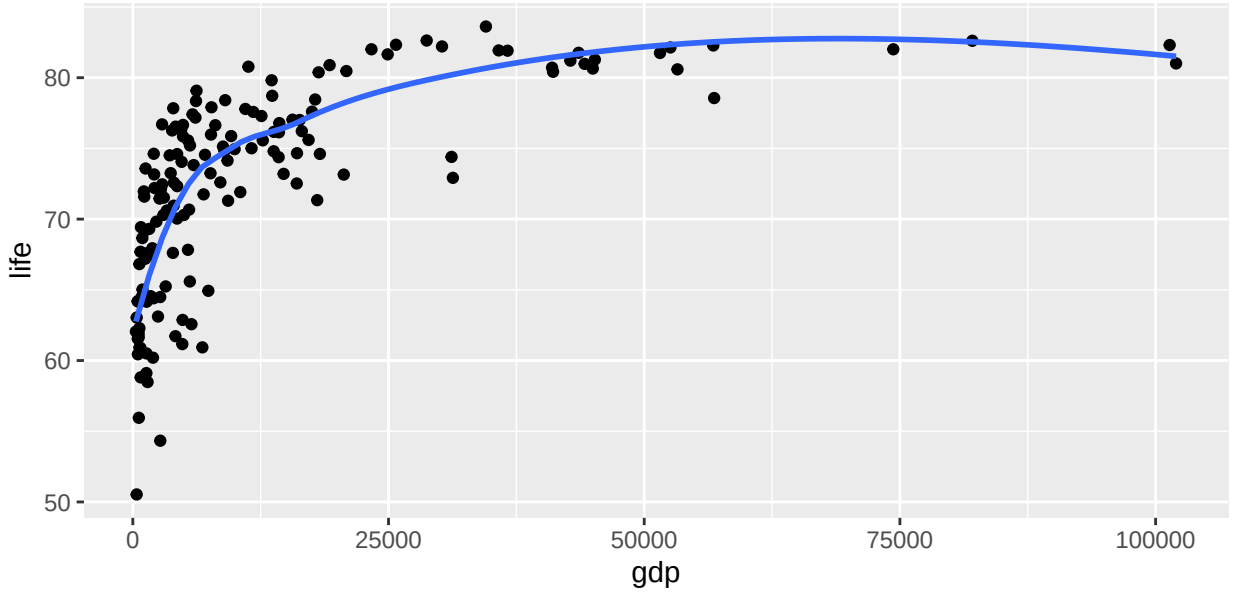



```
cor (sqrtnitrogen ~ sqrtphosphate, data=ds)
```

```
## [1] 0.743
```

```
gf_point(life ~ gdp, main="Figure 8g: Scatterplot of life by gdp", data=ds) %>% gf_smooth(se=FALSE)
```

```
## `geom_smooth()` using method = 'loess'
## Warning: Removed 4 rows containing non-finite values (stat_smooth).
## Warning: Removed 4 rows containing missing values (geom_point).
```




```
cor (life ~ gdp, data=ds)
```

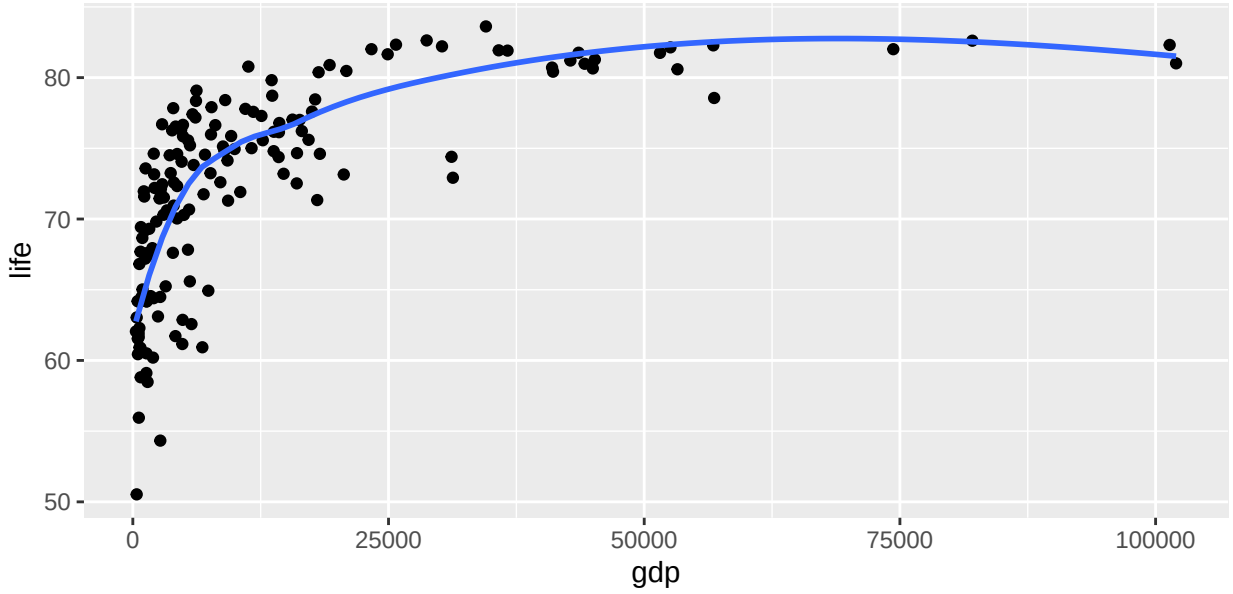
```
## [1] NA
```

```
gf_point(life ~ gdp, main="Figure 8h: Scatterplot of life by potash", data=ds) %>% gf_smooth(se=FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 4 rows containing non-finite values (stat_smooth).

## Warning: Removed 4 rows containing missing values (geom_point).
```




```
cor (life ~ potash, data=ds)
```

```
## [1] NA
```

Phosphate was fairly highly correlated with potash and nitrogen. The square root of potash was highly correlated with the square root of nitrogen and potash, and the square roots of potash and nitrogen were highly correlated as depicted by the r values. This suggests that we may end up only needing one of the fertilizers in the final multiple regression model.

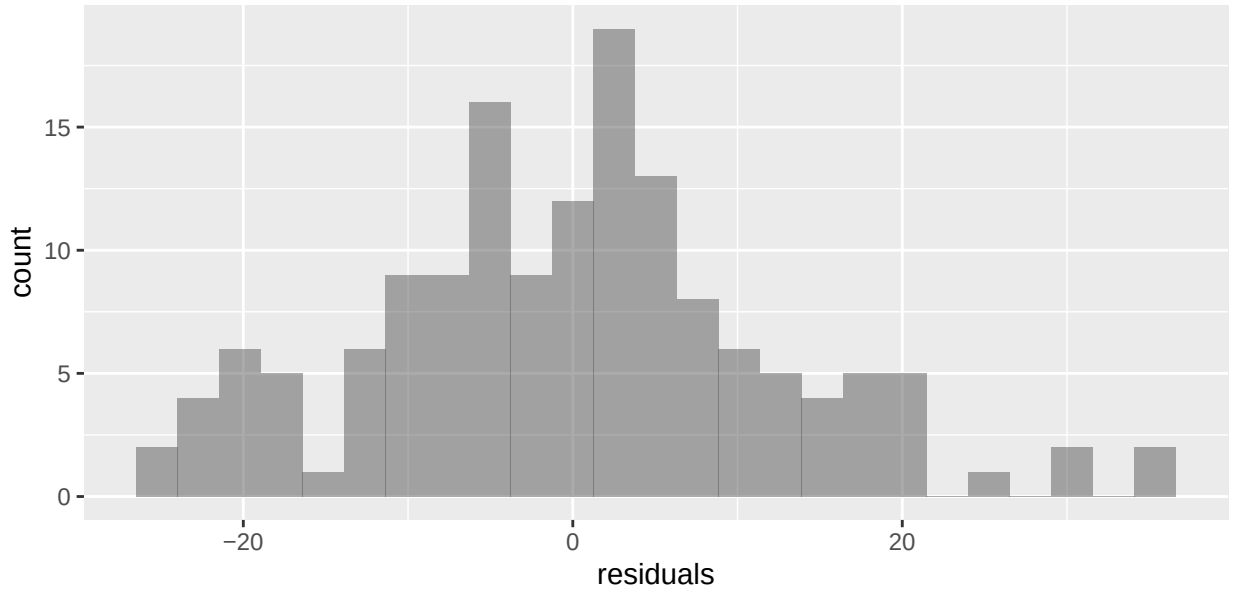
MULTIPLE LINEAR REGRESSION

Linear Model 1:

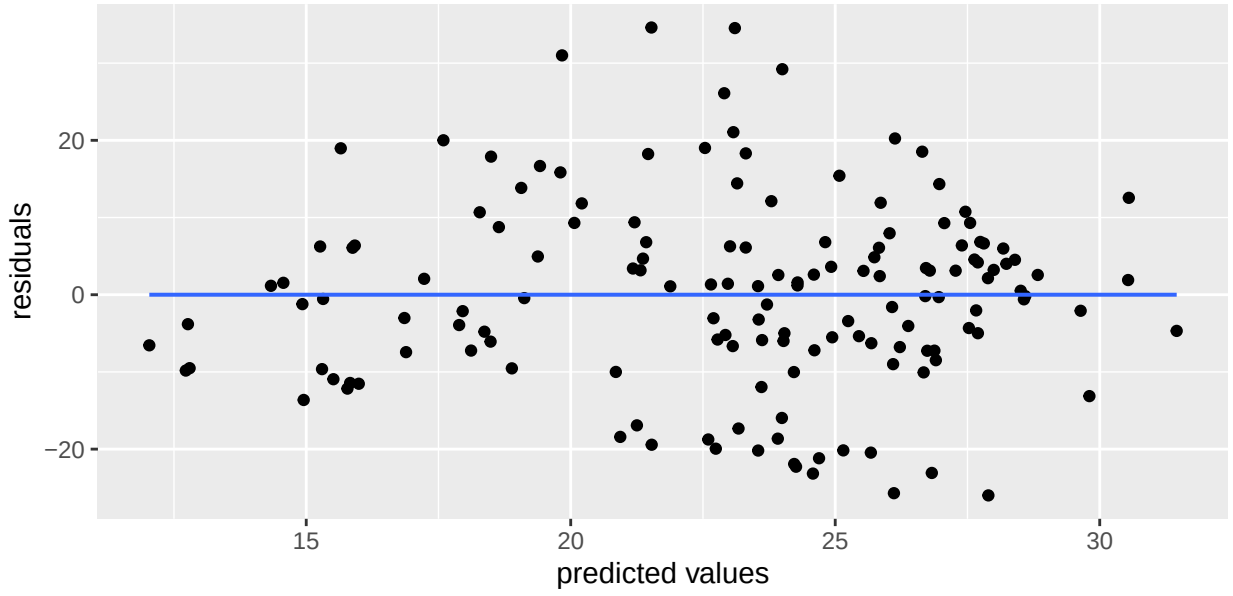
```
lm1 <- lm(alz ~ potash + nitrogen + phosphate + life + gdp + developed, data=ds)
msummary(lm1)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47.9024256 14.4321162    3.32   0.0011 **
## potash      -0.0503514  0.0599331   -0.84   0.4022
## nitrogen    -0.0044983  0.0214752   -0.21   0.8344
## phosphate   -0.0392607  0.0753247   -0.52   0.6030
## life        -0.3267336  0.2144381   -1.52   0.1298
## gdp          0.0001937  0.0000909    2.13   0.0349 *
## developed   -8.3894931  3.6629761   -2.29   0.0235 *
##
## Residual standard error: 12.5 on 142 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.115,  Adjusted R-squared:  0.0771
## F-statistic: 3.06 on 6 and 142 DF,  p-value: 0.00756
```

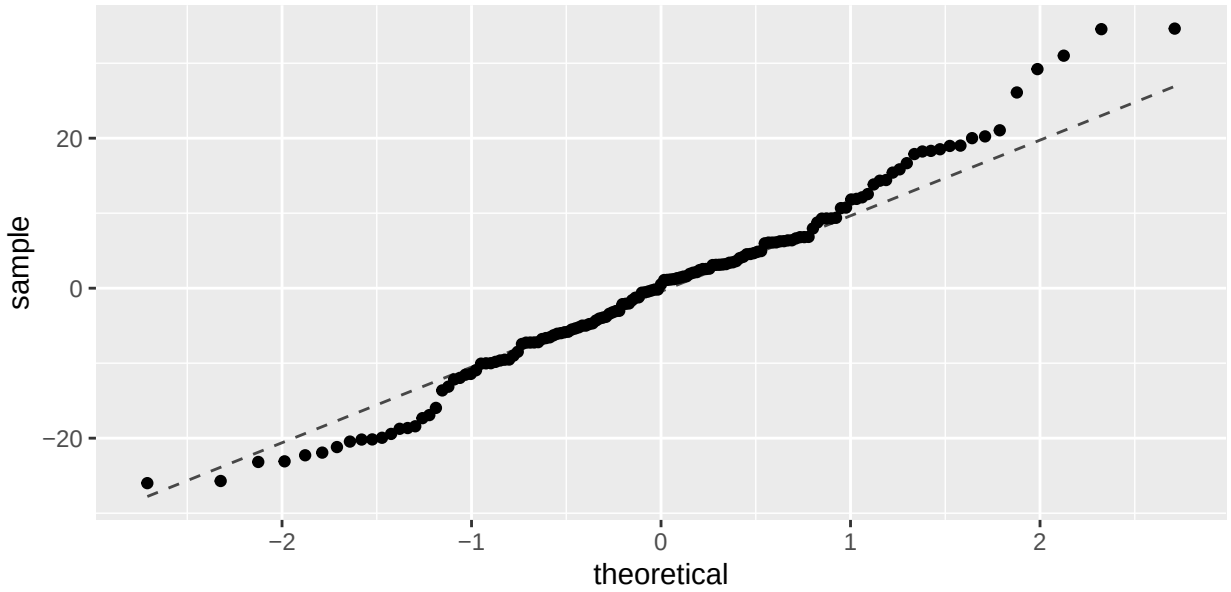
```
gf_histogram(~ residuals(lm1), nint=10,  xlab="residuals")
```




```
gf_point(residuals(lm1) ~ fitted(lm1), xlab = "predicted values", ylab = "residuals") %>%
  gf_lm()
```




```
gf_qq(~resid(lm1)) %>%
  gf_qqline
```



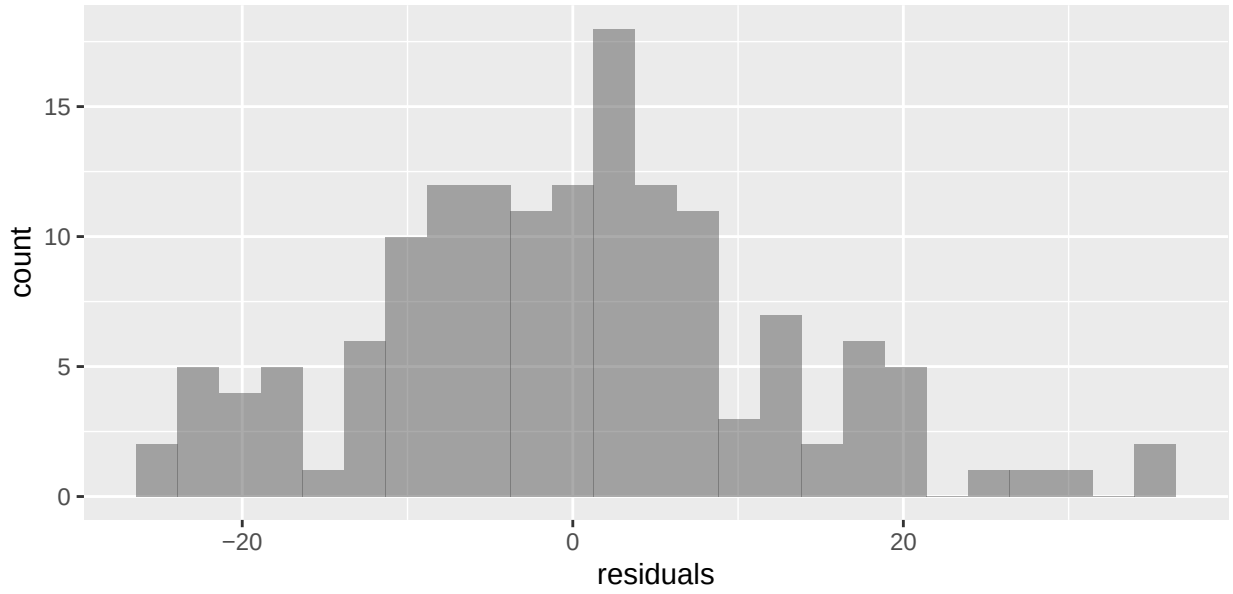
This is the first linear model. The main/ significant predictors for alzheimers that we can see in this model is whether or not a nation is classified as developed or not as well as gdp. Our adjusted r-squared is 0.0771. This is not that great. In looking at the residuals, however, the errors are normally distributed, but there is slight heteroscedasticity.

Next we will create a linear model using our square root transformed data which helped with the heteroscedasticity more than a log transformation.

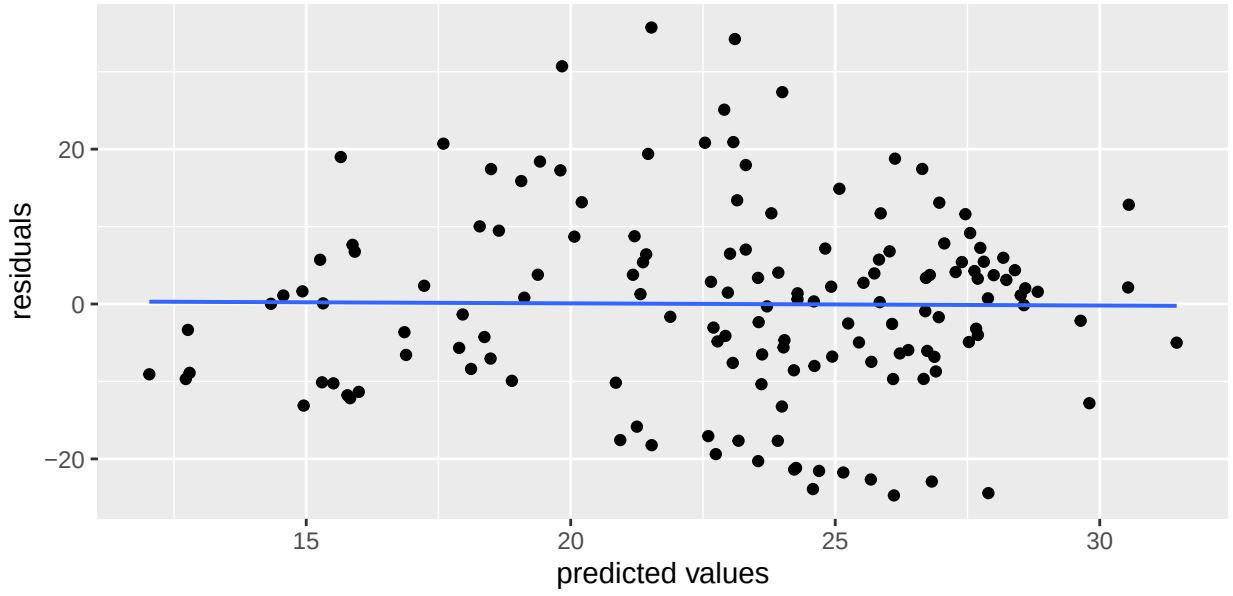
```
lms <- lm(alz ~ sqrtpotash + sqrtnitrogen + sqrtphosphate + life + gdp + developed, data=ds)
msummary(lms)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   44.4510572 14.5426446    3.06   0.0027 **
## sqrtpotash    -0.7209874  0.6679362   -1.08   0.2822
## sqrtnitrogen  -0.3711560  0.4174798   -0.89   0.3755
## sqrtphosphate  0.0632696  0.8232230    0.08   0.9388
## life          -0.2440719  0.2236202   -1.09   0.2769
## gdp            0.0002060  0.0000903    2.28   0.0239 *
## developed     -7.8526423  3.6067639   -2.18   0.0311 *
##
## Residual standard error: 12.4 on 142 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.128,  Adjusted R-squared:  0.0909
## F-statistic: 3.47 on 6 and 142 DF,  p-value: 0.00316
```

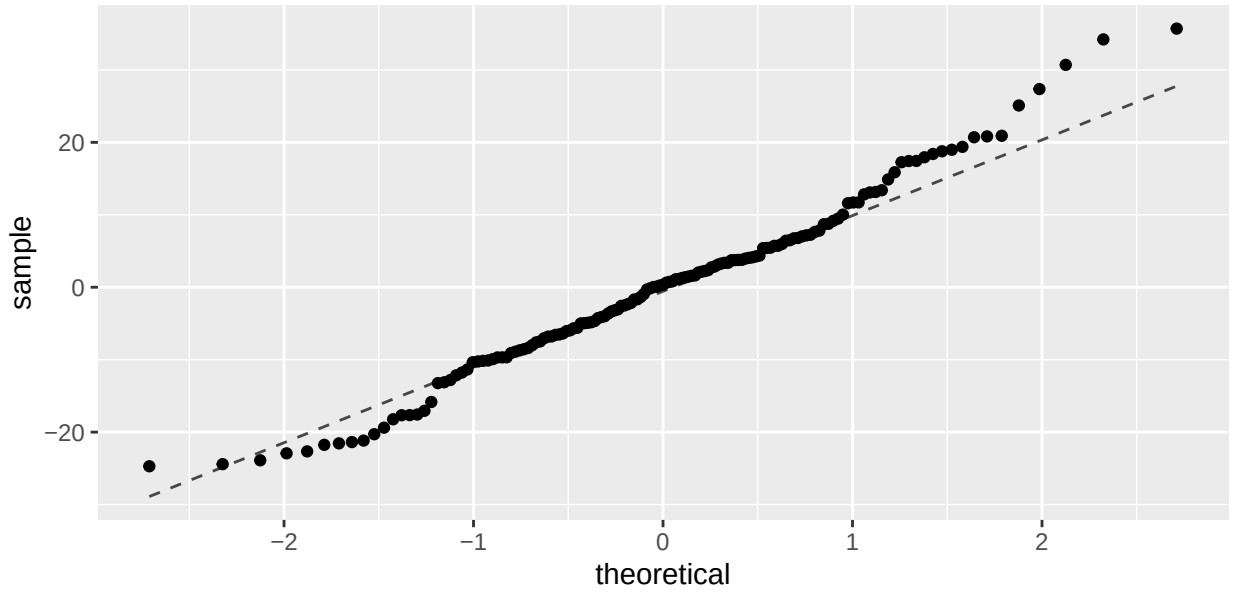
```
gf_histogram(~ residuals(lms), nint=10,  xlab="residuals")
```




```
gf_point(residuals(lms) ~ fitted(lm1), xlab = "predicted values", ylab = "residuals") %>%
  gf_lm()
```




```
gf_qq(~resid(lms)) %>%
  gf_qqline
```



The adjusted R-squared of this model is slightly improved, and it is now 0.0909. GDP and developed are still the only significant predictors, but the errors are normally and randomly distributed with equal variance, and the relationship is linear with the exception of some outliers.

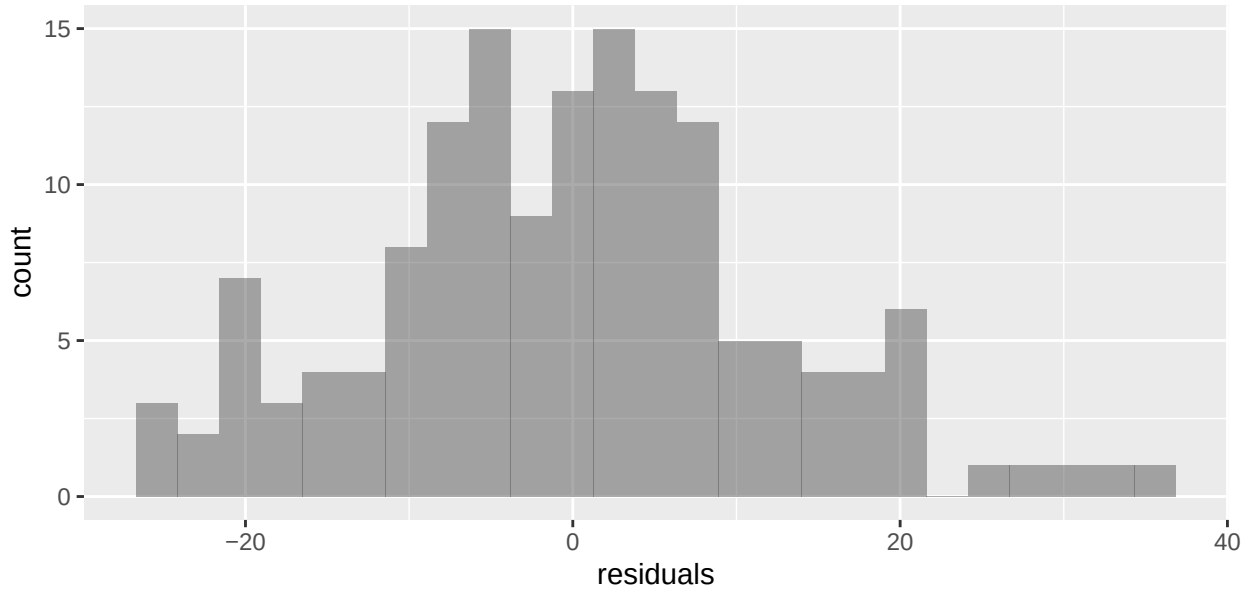
Linear Model 2:

This model has both potash and life expectancy (only 1 is significant)

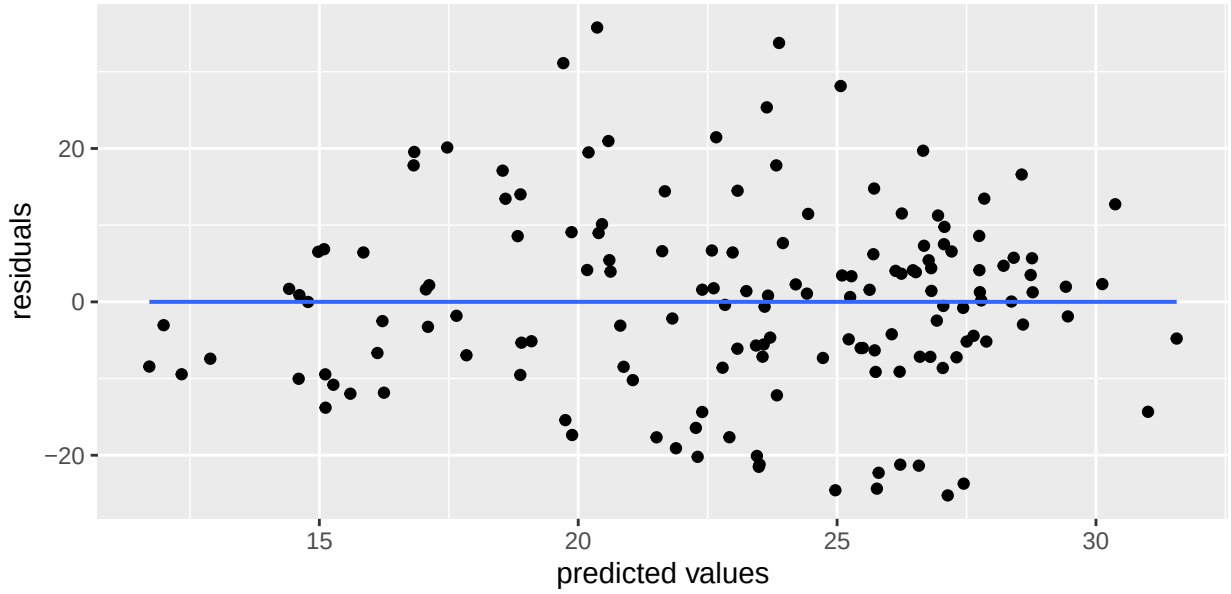
```
lm2 <- lm(alz ~ sqrtpotash+ life + gdp + developed, data=ds)
msummary(lm2)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.7106296 14.0635194    3.32   0.0011 **
## sqrtpotash  -0.9708491  0.4812168   -2.02   0.0455 *
## life        -0.2929965  0.2109877   -1.39   0.1671
## gdp          0.0002033  0.0000892    2.28   0.0242 *
## developed   -8.2724816  3.5527210   -2.33   0.0213 *
##
## Residual standard error: 12.3 on 144 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.122,  Adjusted R-squared:  0.098
## F-statistic: 5.02 on 4 and 144 DF,  p-value: 0.000815
```

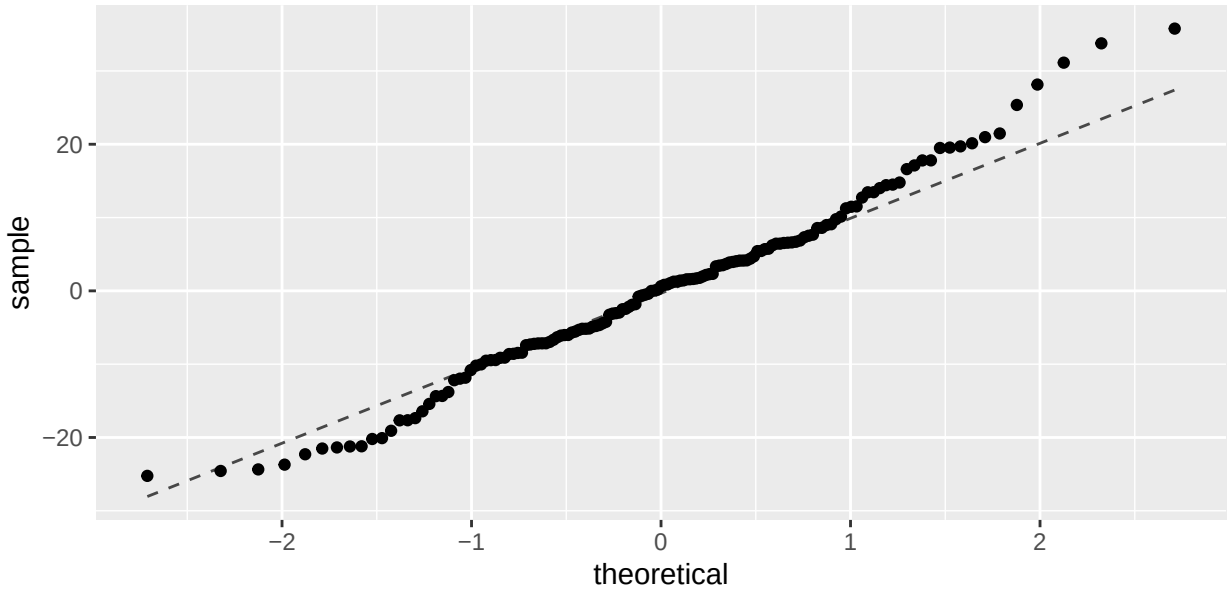
```
gf_histogram(~ residuals(lm2), nint=10,  xlab="residuals")
```




```
gf_point(residuals(lm2) ~ fitted(lm2), xlab = "predicted values", ylab = "residuals") %>%
  gf_lm()
```




```
gf_qq(~resid(lm2)) %>%
  gf_qqline
```



In this model, potash is now significant along with GDP and developed. The adjusted R-squared us now 0.098.

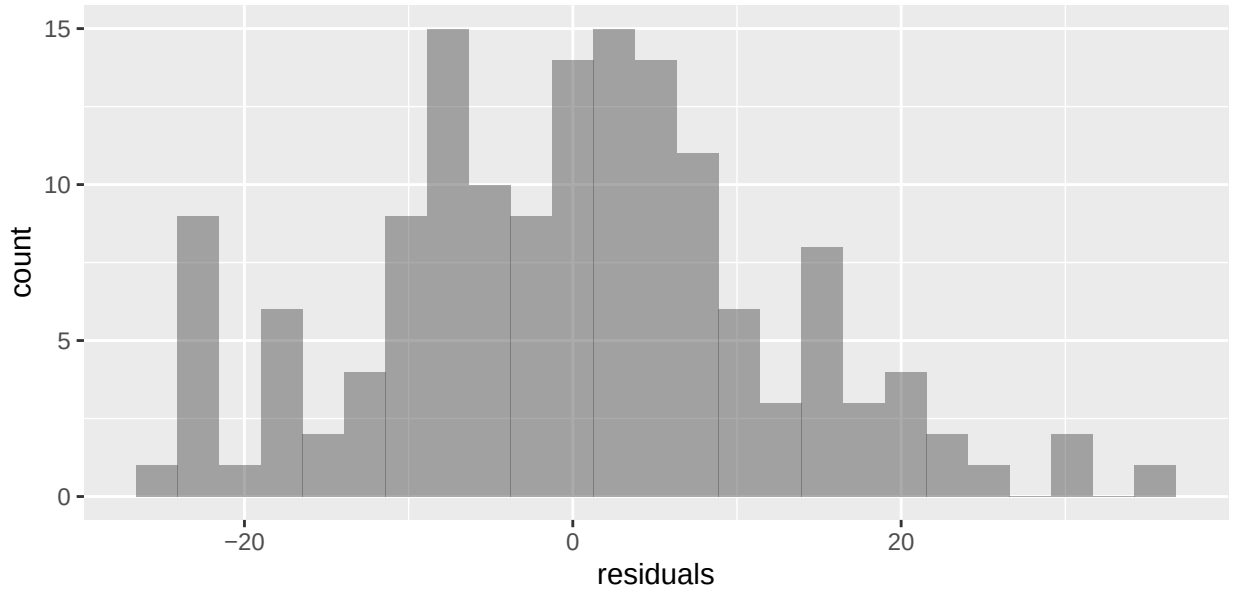
Linear Model 3:

This model has just potash, with no life expectancy.

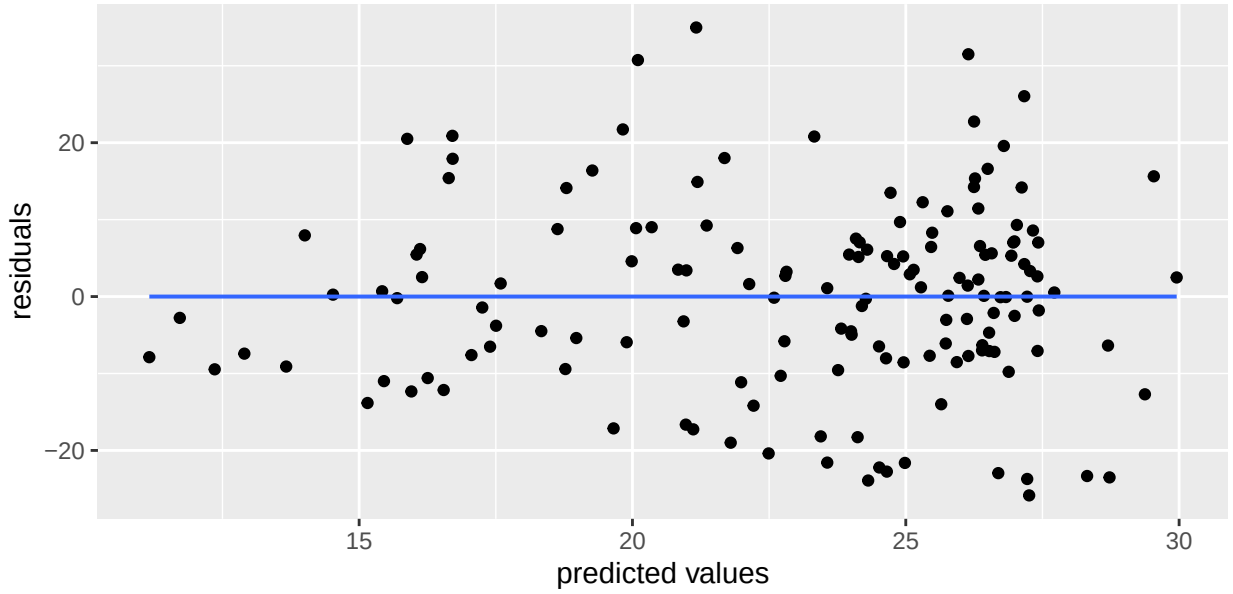
```
lm3 <- lm(alz ~ sqrtpotash + gdp + developed, data=ds)
msummary(lm3)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.3274829  1.6952466   16.12   <2e-16 ***
## sqrtpotash  -1.2783138  0.4249909   -3.01   0.0031 **
## gdp          0.0001649  0.0000849    1.94   0.0539 .
## developed   -9.5063055  3.4401542   -2.76   0.0065 **
##
## Residual standard error: 12.3 on 146 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.11,   Adjusted R-squared:  0.0922
## F-statistic: 6.05 on 3 and 146 DF,  p-value: 0.000657
```

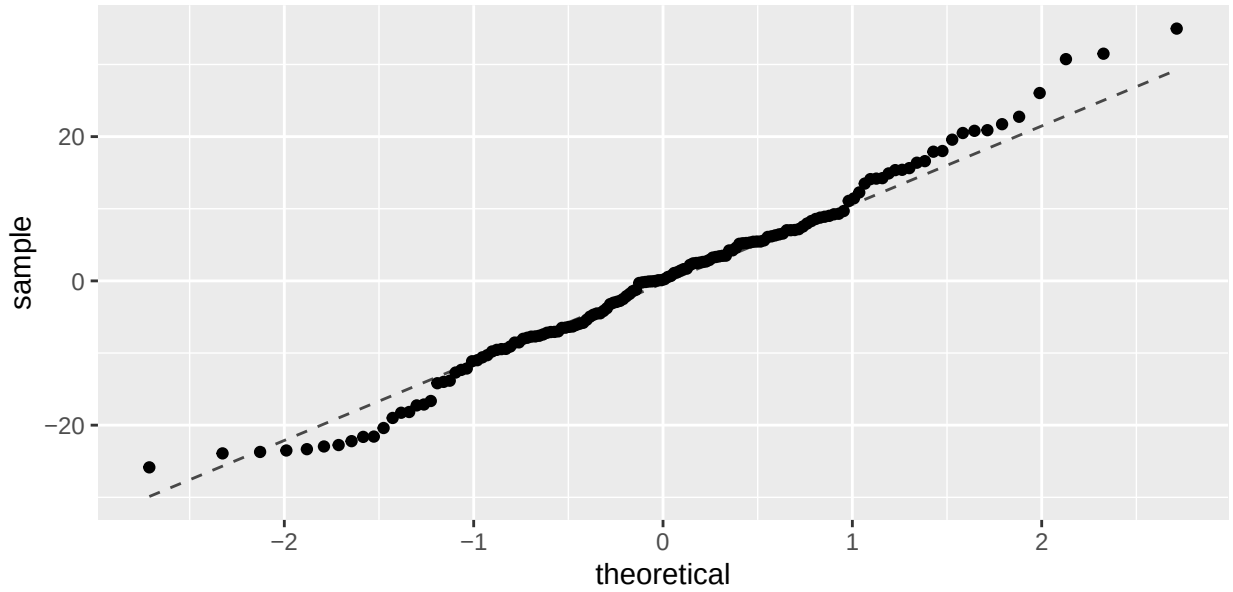
```
gf_histogram(~ residuals(lm3), nint=10,  xlab="residuals")
```




```
gf_point(residuals(lm3) ~ fitted(lm3), xlab = "predicted values", ylab = "residuals") %>%
  gf_lm()
```




```
gf_qq(~resid(lm3)) %>%
  gf_qqline
```



The p-value of GDP is not exactly below the alpha value of 0.05, however, it is almost significant. The adjusted R-squared value is 0.0922. This is our final model.

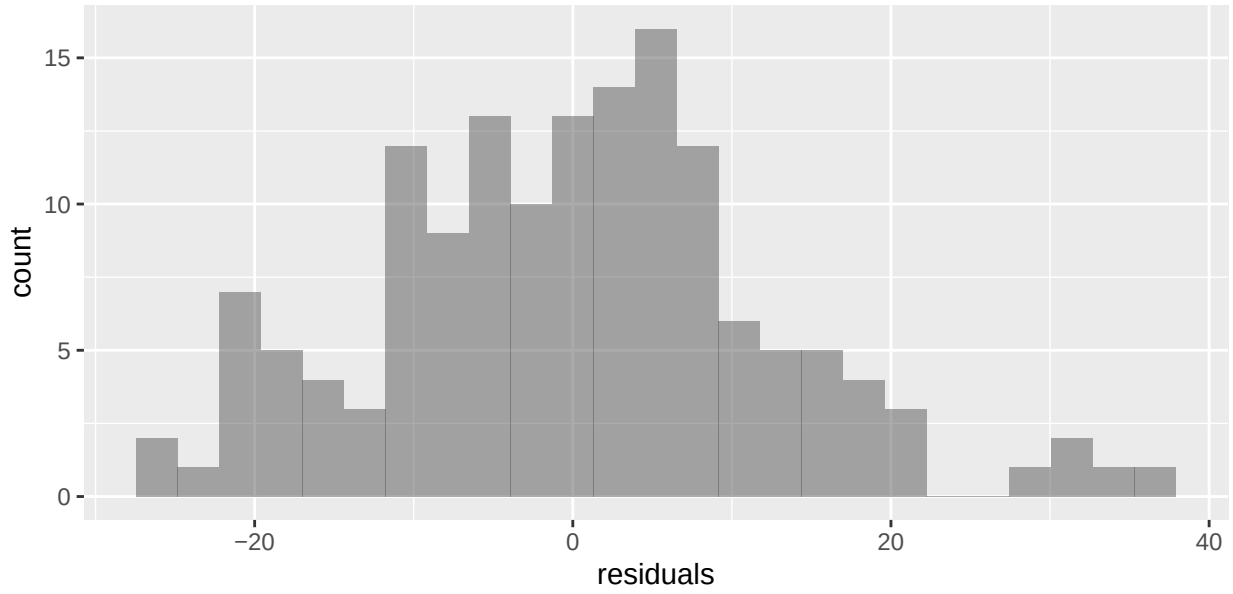
Linear Model 4:

This model has just life expectancy, with no potash.

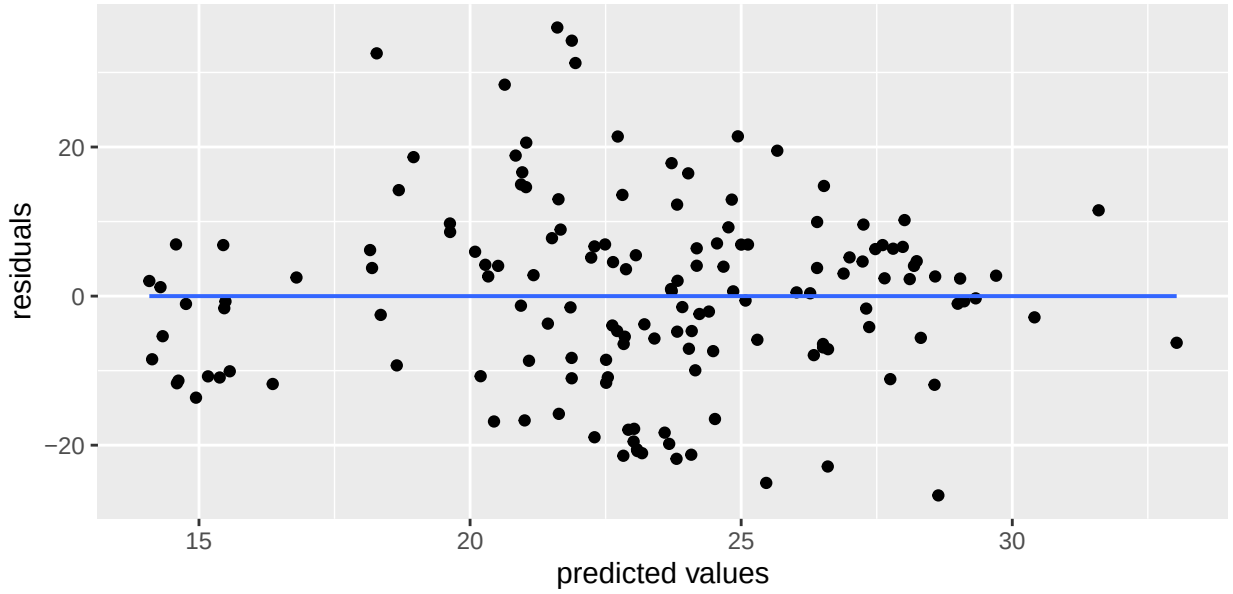
```
lm4 <- lm(alz ~ life + gdp + developed, data=ds)
msummary(lm4)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 57.7784056 13.0857904    4.42  0.00002 ***
## life        -0.4911126  0.1887091   -2.60    0.010 *
## gdp          0.0001848  0.0000897    2.06    0.041 *
## developed   -7.5248040  3.5705507   -2.11    0.037 *
##
## Residual standard error: 12.5 on 145 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.0975, Adjusted R-squared:  0.0789
## F-statistic: 5.22 on 3 and 145 DF,  p-value: 0.00188
```

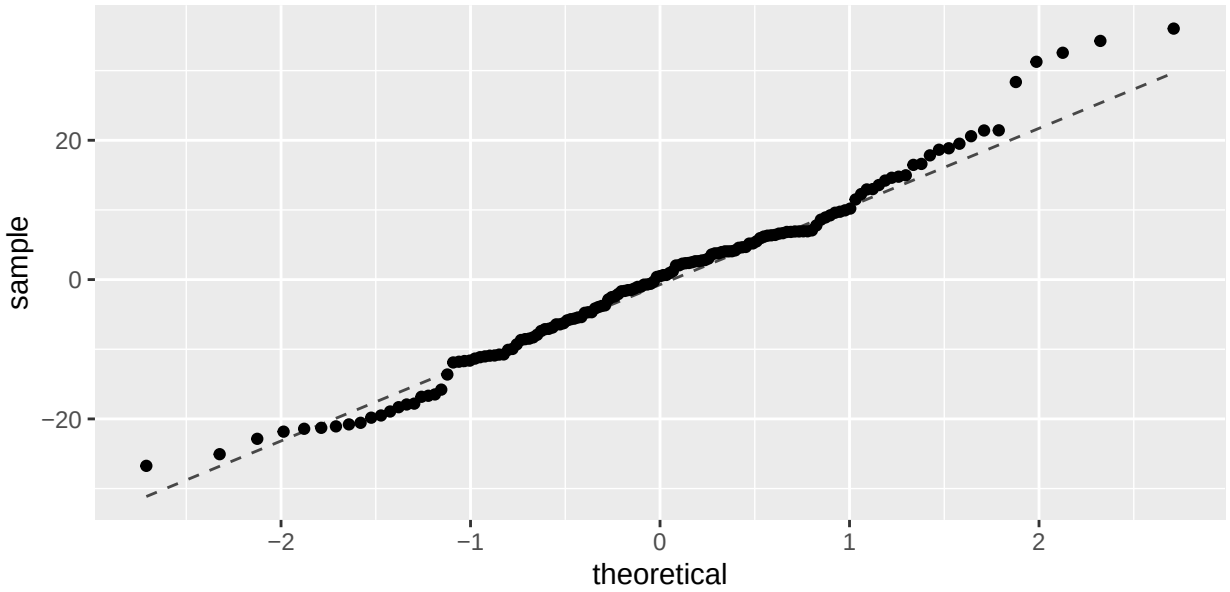
```
gf_histogram(~ residuals(lm4), nint=10,  xlab="residuals")
```




```
gf_point(residuals(lm4) ~ fitted(lm4), xlab = "predicted values", ylab = "residuals") %>%
  gf_lm()
```




```
gf_qq(~resid(lm4)) %>%
  gf_qqline
```



Potash was a better predictor than life expectancy and yielded a higher R-squared. We have decided to use linear model 3 as our final linear model.

MULTIPLE LINEAR EQUATION

> Alzheimer's Death Rate = 27.3274829 - 1.2783138 sqrtpotash + 0.0001649 GDP - 9.5063055 developed

According to our model and project aims, this means that a one unit increase in sqrtpotash, after controlling for the effects of GDP and whether a country is developed or not, results in a decrease of 1.2783138 in Alzheimer's death rates.

RESULTS

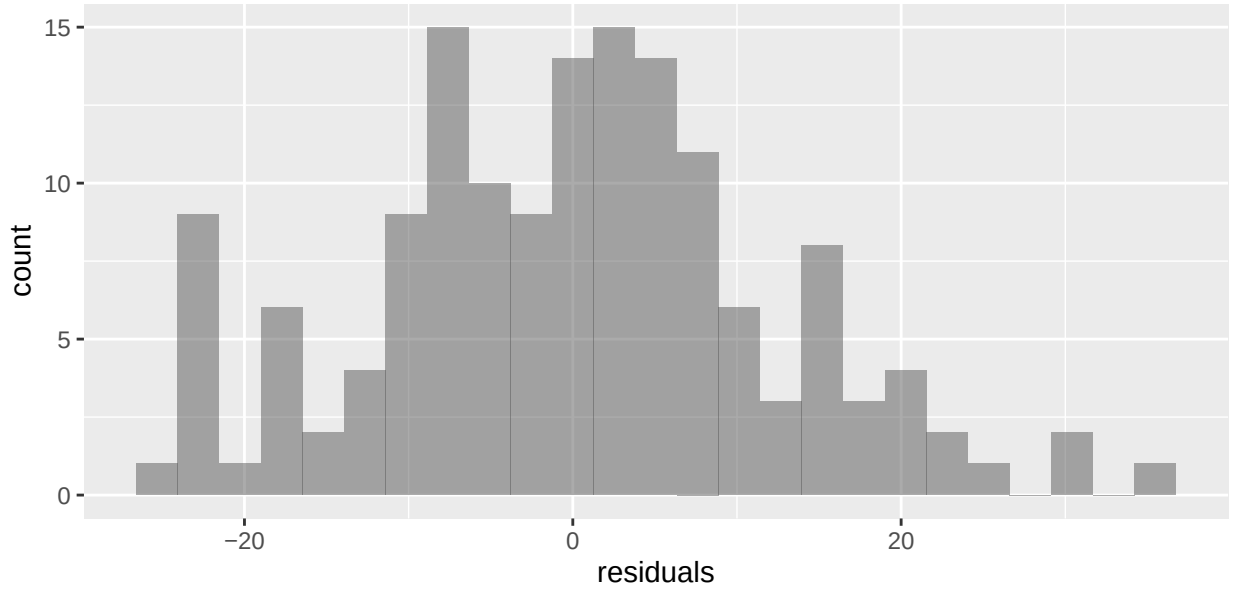
Our final linear model, linear model 3, only had potash, gdp, and whether a country was developed or not as significant predictors. The final model ended up only having potash as a predictor, rather than the other 2 fertilizers, because predictor v predictor scatterplots found potash, nitrogen, and phosphate to all be highly correlated and collinear. Having all 3 in the model reduced the significance of the other fertilizers as they are predictors of each other. The strongest fertilizer predictor ended up, therefore, being potash.

```
lm3 <- lm(alz ~ sqrtpotash + gdp + developed, data=ds)
msummary(lm3)
```

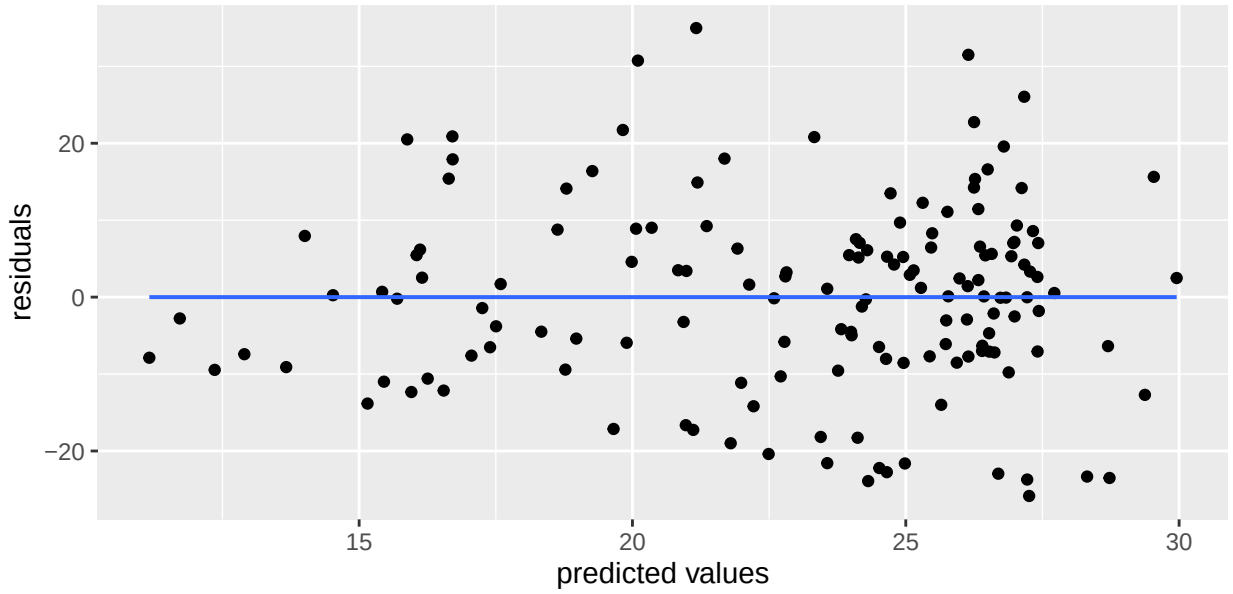
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.3274829  1.6952466   16.12   <2e-16 ***
## sqrtpotash  -1.2783138  0.4249909   -3.01   0.0031 **
## gdp          0.0001649  0.0000849    1.94   0.0539 .
## developed   -9.5063055  3.4401542   -2.76   0.0065 **
##
## Residual standard error: 12.3 on 146 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.11,   Adjusted R-squared:  0.0922
## F-statistic: 6.05 on 3 and 146 DF,  p-value: 0.000657
```

DIAGNOSTICS OF THE FINAL MODEL

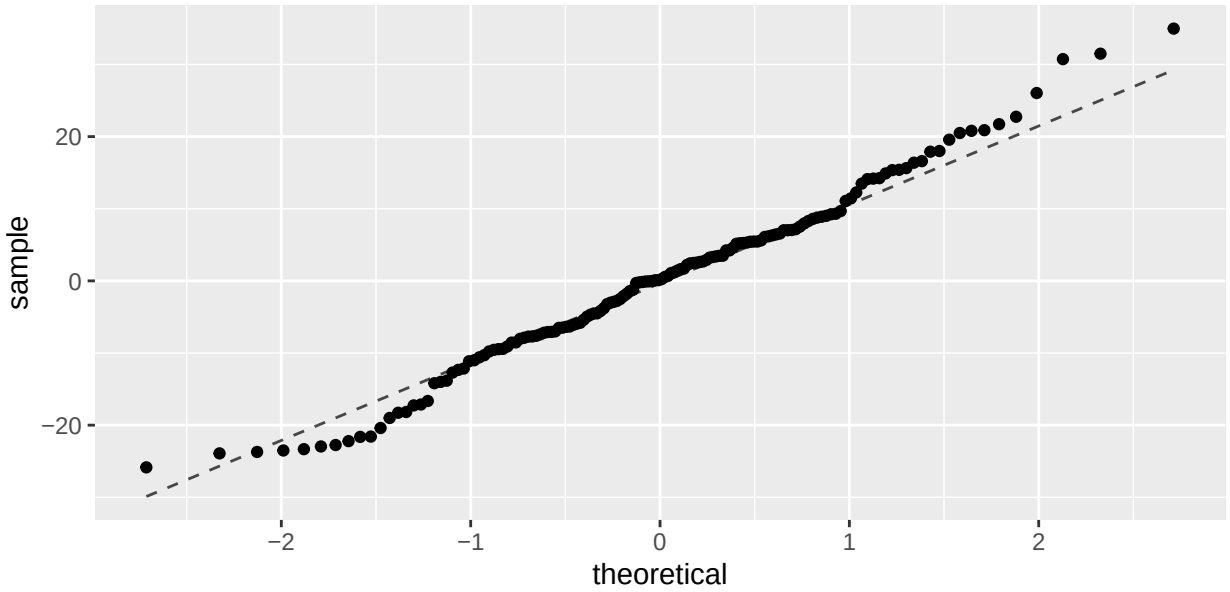
```
gf_histogram(~ residuals(lm3), nint=10,  xlab="residuals")
```




```
gf_point(residuals(lm3) ~ fitted(lm3), xlab = "predicted values", ylab = "residuals") %>%
  gf_lm()
```




```
gf_qq(~resid(lm3)) %>%
  gf_qqline
```



All the conditions for a linear regression model were met by our final multiple regression model, linear model 3. Most of the points lay on the qq plot with the exception of the extreme ends. This suggested that there may have been some few outliers in the model, but generally, the linear condition had been satisfied. We assumed that the independence condition was satisfied because all of the data sets that we used had collected their data randomly. The histogram of the residuals was relatively normal, satisfying the Normally Distributed Errors condition. Lastly, the residuals in the residual plot were randomly distributed and there were no apparent shapes or patterns, satisfying the Equal Variance or the 'Does the Plot Thicken?' condition. Therefore, this model is fit to be used and we can proceed with caution as we only had an R^squared value of 9.22%.

CONCLUSION

The final model would predict that for every increase in fertilizer use, there would be a decrease in the death rate of alzheimers. This result however is a bit misleading as the amount of fertilizer used in countries with high GDP's and with very strong medical infrastructure is incredibly high, and these countries disproportionately affect the linear model as their strong health systems may offset the effects of the high fertilizer use. As such, boxplots of the developed countries versus the non-developed countries against all the predictors showed this. Fertilizer use for potash, phosphate, and nitrogen was highest in developed countries, but so was life and GDP. On the flip side, Alzheimer's death rates were lower in developed countries.

Ultimately our results are more or less inconclusive and there are various further tests that could be run. One such model would be to run two separate analysis tests, one with developed countries and one that consists only of countries that are not considered developed, which would allow us to see just how impactful something like development is in this case and if Alzheimer's death rate is positively correlated with fertilizer use when countries don't have as robust of healthcare systems.