Data Analysis Project - Alcohol Consumption/Health Expenditure/Status vs. life expectancy in the World

Ankit Gubiligari

2023-06-08

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
LED <- read.csv("C:/Users/AnkitGG/Desktop/LED.csv")
head(LED)
```

```
##       Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 1 Afghanistan 2015 Developing            65.0             263            62
## 2 Afghanistan 2014 Developing            59.9             271            64
## 3 Afghanistan 2013 Developing            59.9             268            66
## 4 Afghanistan 2012 Developing            59.5             272            69
## 5 Afghanistan 2011 Developing            59.2             275            71
## 6 Afghanistan 2010 Developing            58.8             279            74
##   Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 1    0.01              71.279624          65    1154 19.1                83
## 2    0.01              73.523582          62     492 18.6                86
## 3    0.01              73.219243          64     430 18.1                89
## 4    0.01              78.184215          67    2787 17.6                93
## 5    0.01               7.097109          68    3013 17.2                97
## 6    0.01              79.679367          66    1989 16.7               102
##   Polio Total.expenditure Diphtheria HIV.AIDS       GDP Population
## 1     6              8.16         65      0.1 584.25921   33736494
## 2    58              8.18         62      0.1 612.69651     327582
## 3    62              8.13         64      0.1 631.74498   31731688
## 4    67              8.52         67      0.1 669.95900    3696958
## 5    68              7.87         68      0.1  63.53723    2978599
## 6    66              9.20         66      0.1 553.32894    2883167
##   thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1                 17.2               17.3                           0.479
## 2                 17.5               17.5                           0.476
## 3                 17.7               17.7                           0.470
## 4                 17.9               18.0                           0.463
## 5                 18.2               18.2                           0.454
## 6                 18.4               18.4                           0.448
##   Schooling
## 1      10.1
## 2      10.0
## 3       9.9
## 4       9.8
## 5       9.5
## 6       9.2
```

```
LED_2010 <- LED[LED$Year == 2010, ]
head(LED_2010)
```

```
##                 Country Year     Status Life.expectancy Adult.Mortality
## 6           Afghanistan 2010 Developing            58.8             279
## 22              Albania 2010 Developing            76.2              91
## 38              Algeria 2010 Developing            74.7             119
## 54               Angola 2010 Developing            49.6             365
## 70  Antigua and Barbuda 2010 Developing            75.6             138
## 86            Argentina 2010 Developing            75.5             121
##    infant.deaths Alcohol percentage.expenditure Hepatitis.B Measles  BMI
## 6             74    0.01               79.67937          66    1989 16.7
## 22             1    5.28               41.82276          99      10 54.3
## 38            21    0.45              430.71759          95     103 53.9
## 54            78    7.80              191.65374          77    1190  2.4
## 70             0    7.84             1983.95694          98       0 44.4
## 86            10    8.15              187.61095          94      17 59.8
##    under.five.deaths Polio Total.expenditure Diphtheria HIV.AIDS        GDP
## 6                102    66              9.20         66      0.1   553.3289
## 22                 1    99              5.34         99      0.1   494.3588
## 38                24    95              5.12         95      0.1  4463.3947
## 54               121    81              3.39         77      2.5  3529.5348
## 70                 0    99              5.63         98      0.1 12126.8761
## 86                11    95              6.55         94      0.1  1276.2650
##    Population thinness..1.19.years thinness.5.9.years
## 6     2883167                 18.4               18.4
## 22     291321                  1.4                1.5
## 38   36117637                  5.9                5.8
## 54   23369131                  9.1                9.0
## 70         NA                  3.3                3.3
## 86   41223889                  1.0                0.9
##    Income.composition.of.resources Schooling
## 6                            0.448       9.2
## 22                           0.725      12.5
## 38                           0.714      13.6
## 54                             0.488      9.0
## 70                             0.783     14.1
## 86                             0.802     16.8
```

```
names(LED_2010)
```

```
##  [1] "Country"                         "Year"
##  [3] "Status"                          "Life.expectancy"
##  [5] "Adult.Mortality"                 "infant.deaths"
##  [7] "Alcohol"                         "percentage.expenditure"
##  [9] "Hepatitis.B"                     "Measles"
## [11] "BMI"                             "under.five.deaths"
## [13] "Polio"                           "Total.expenditure"
## [15] "Diphtheria"                      "HIV.AIDS"
## [17] "GDP"                             "Population"
## [19] "thinness..1.19.years"            "thinness.5.9.years"
## [21] "Income.composition.of.resources" "Schooling"
```

```
LED_2010 <- select(LED_2010, Country, Year, Status, Life.expectancy, Adult.Mortality, Alcohol, percenta
str(LED_2010)
```

```
## 'data.frame':    183 obs. of  7 variables:
##  $ Country               : chr  "Afghanistan" "Albania" "Algeria" "Angola" ...
##  $ Year                  : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ Status                : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ Life.expectancy       : num  58.8 76.2 74.7 49.6 75.6 75.5 73.5 81.9 84 71.1 ...
##  $ Adult.Mortality       : int  279 91 119 365 138 121 132 64 75 13 ...
##  $ Alcohol               : num  0.01 5.28 0.45 7.8 7.84 ...
##  $ percentage.expenditure: num  79.7 41.8 430.7 191.7 1984 ...
```

```
colSums(is.na(LED_2010))
```

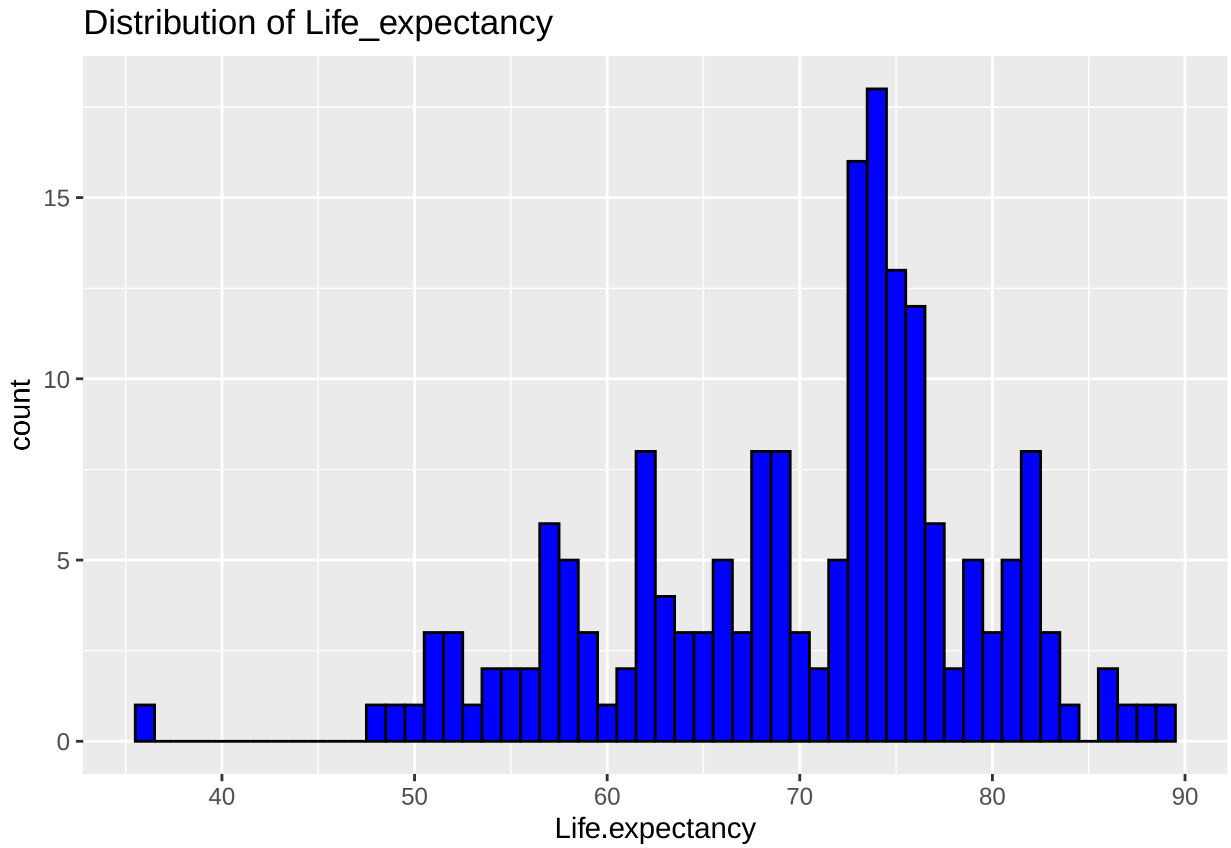
```
##                Country                   Year                 Status
##                      0                      0                      0
##        Life.expectancy        Adult.Mortality                Alcohol
##                      0                      0                      1
## percentage.expenditure
##                      0
```

```
LED_2010_imputed <- LED_2010 %>%
                    mutate(
                           Alcohol = ifelse(is.na(Alcohol), mean(Alcohol, na.rm = TRUE), Alcohol),
  )
#Convert status to categorical
LED_2010_imputed$Status <- as.factor(LED_2010_imputed$Status)
```

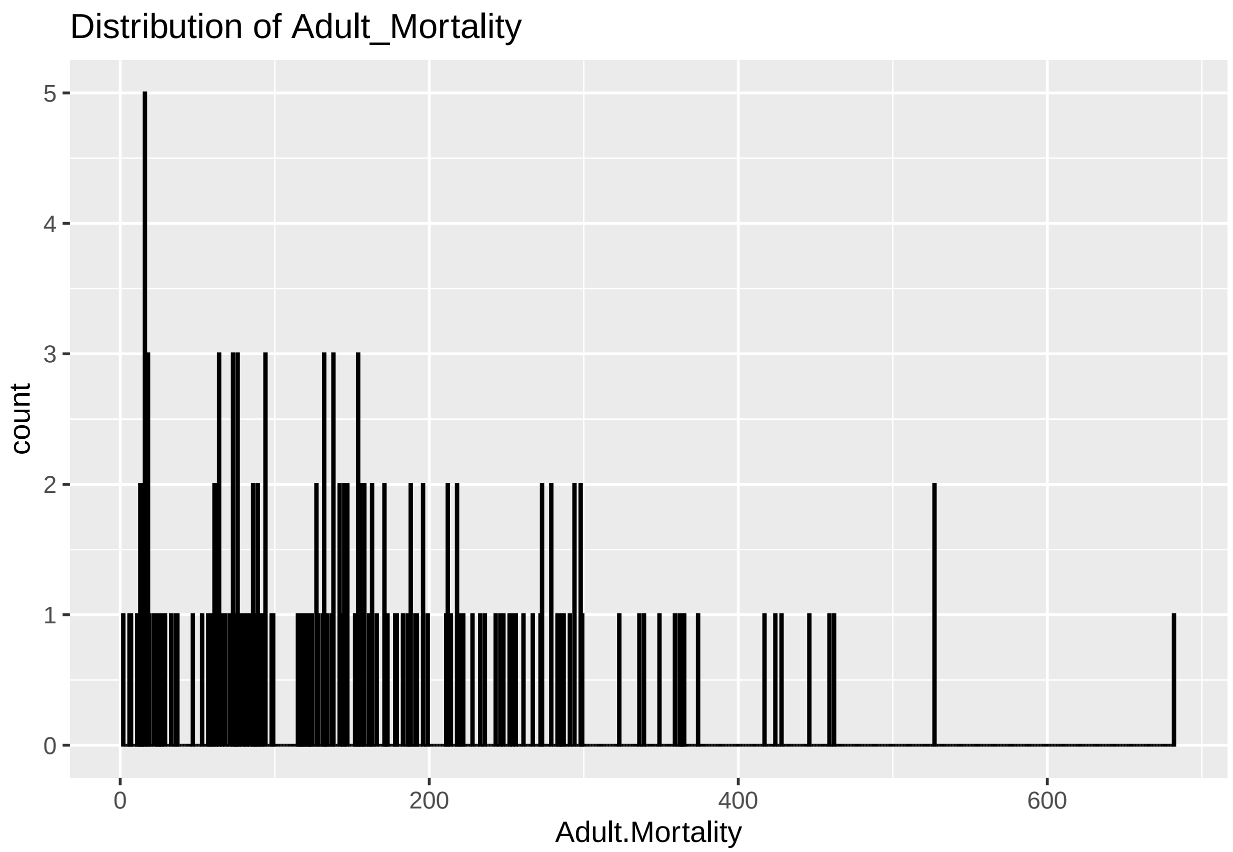
```
summary(LED_2010_imputed)
```

```
##    Country               Year              Status    Life.expectancy
##  Length:183         Min.   :2010   Developed : 32   Min.   :36.30
##  Class :character   1st Qu.:2010   Developing:151   1st Qu.:63.45
##  Mode  :character   Median :2010                      Median :72.80
##                     Mean   :2010                      Mean   :70.05
##                     3rd Qu.:2010                      3rd Qu.:75.80
##                     Max.   :2010                      Max.   :89.00
##  Adult.Mortality    Alcohol        percentage.expenditure
##  Min.   :  2.0   Min.   : 0.010   Min.   :    0.00
##  1st Qu.: 73.5   1st Qu.: 1.405   1st Qu.:   20.52
##  Median :142.0   Median : 4.230   Median :  129.23
##  Mean   :161.9   Mean   : 4.944   Mean   :  768.22
##  3rd Qu.:221.5   3rd Qu.: 7.925   3rd Qu.:  585.21
##  Max.   :682.0   Max.   :14.970   Max.   :15268.06
```

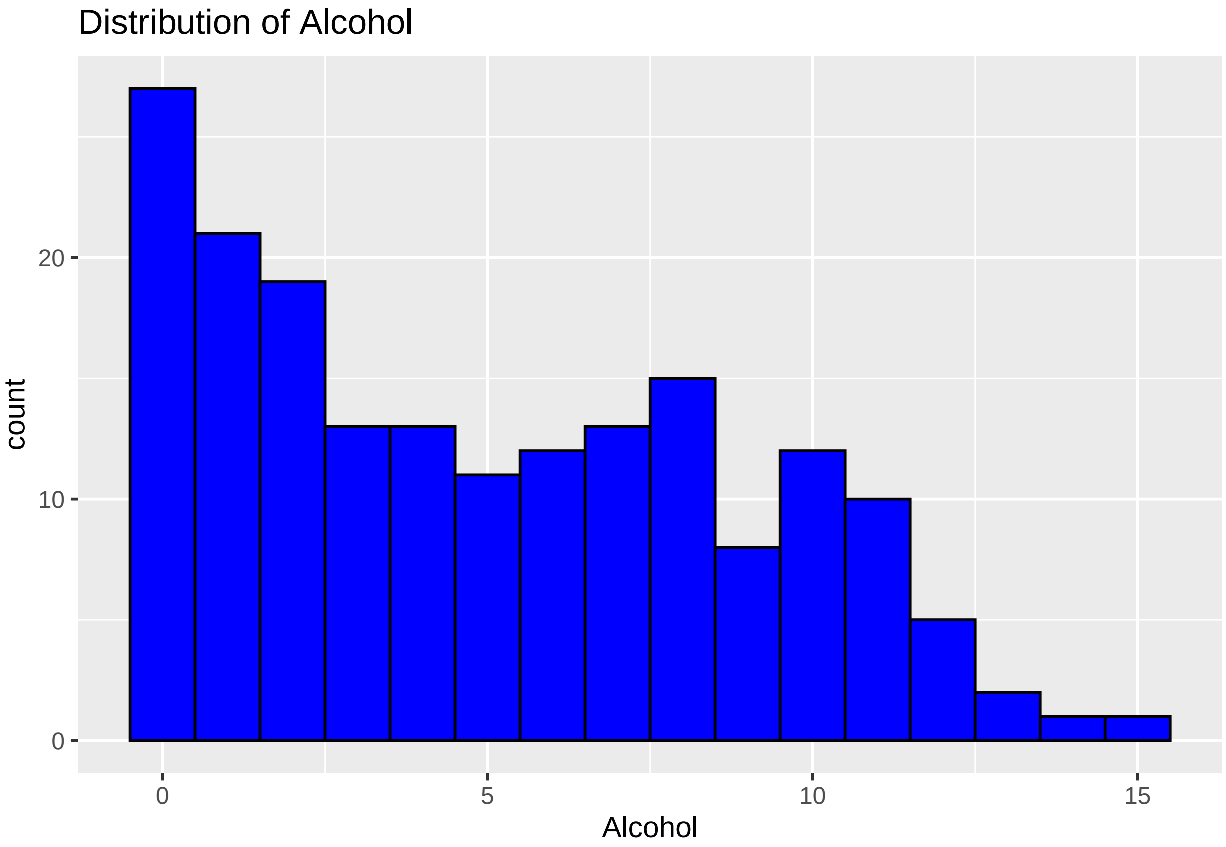
```
par(mfrow = c(3,2), mar = c(4, 4, 2, 1))
# For 'Life_expectancy'
ggplot(LED_2010_imputed, aes(x=Life.expectancy)) +
  geom_histogram(binwidth=1, fill="blue", color="black") +
  ggtitle("Distribution of Life_expectancy")
```



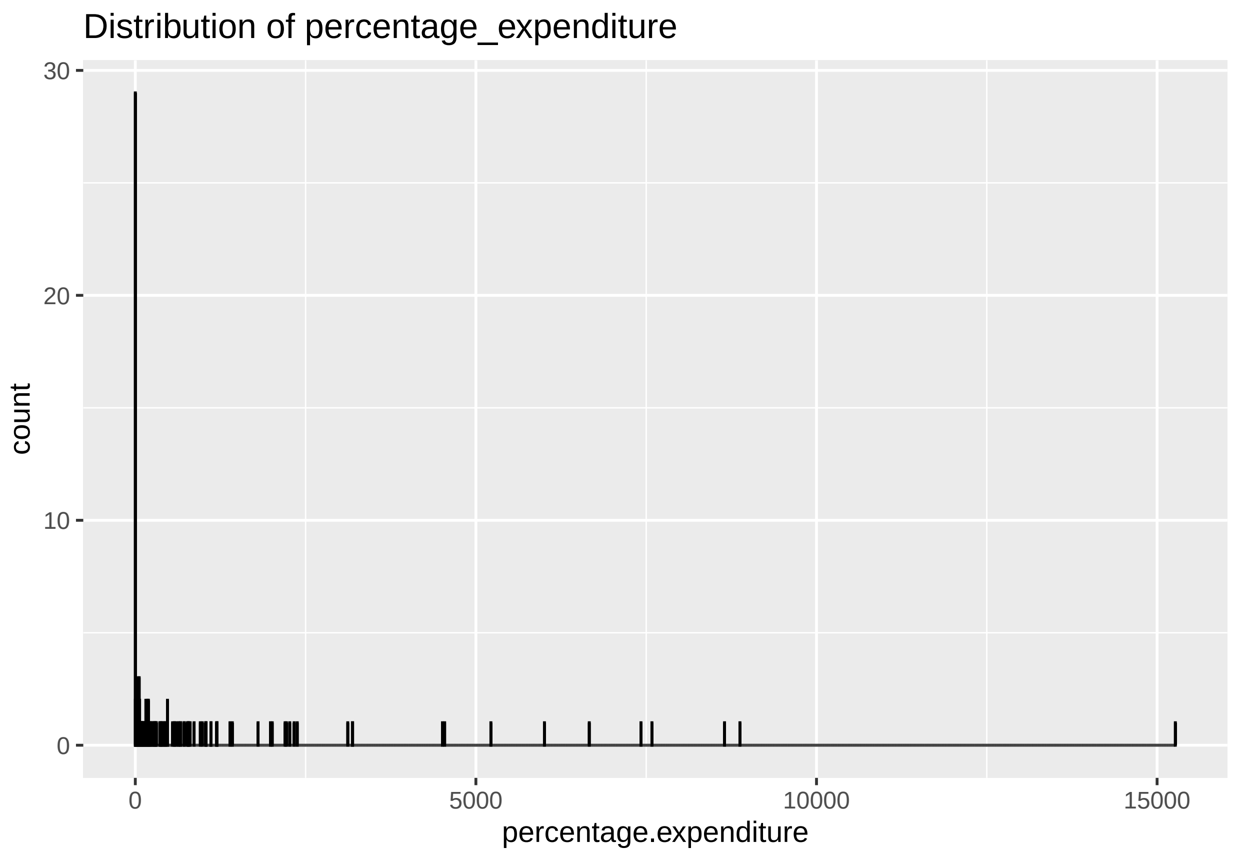
```
# For 'Adult_Mortality'
ggplot(LED_2010_imputed, aes(x=Adult.Mortality)) +
  geom_histogram(binwidth=1, fill="blue", color="black") +
  ggtitle("Distribution of Adult_Mortality")
```



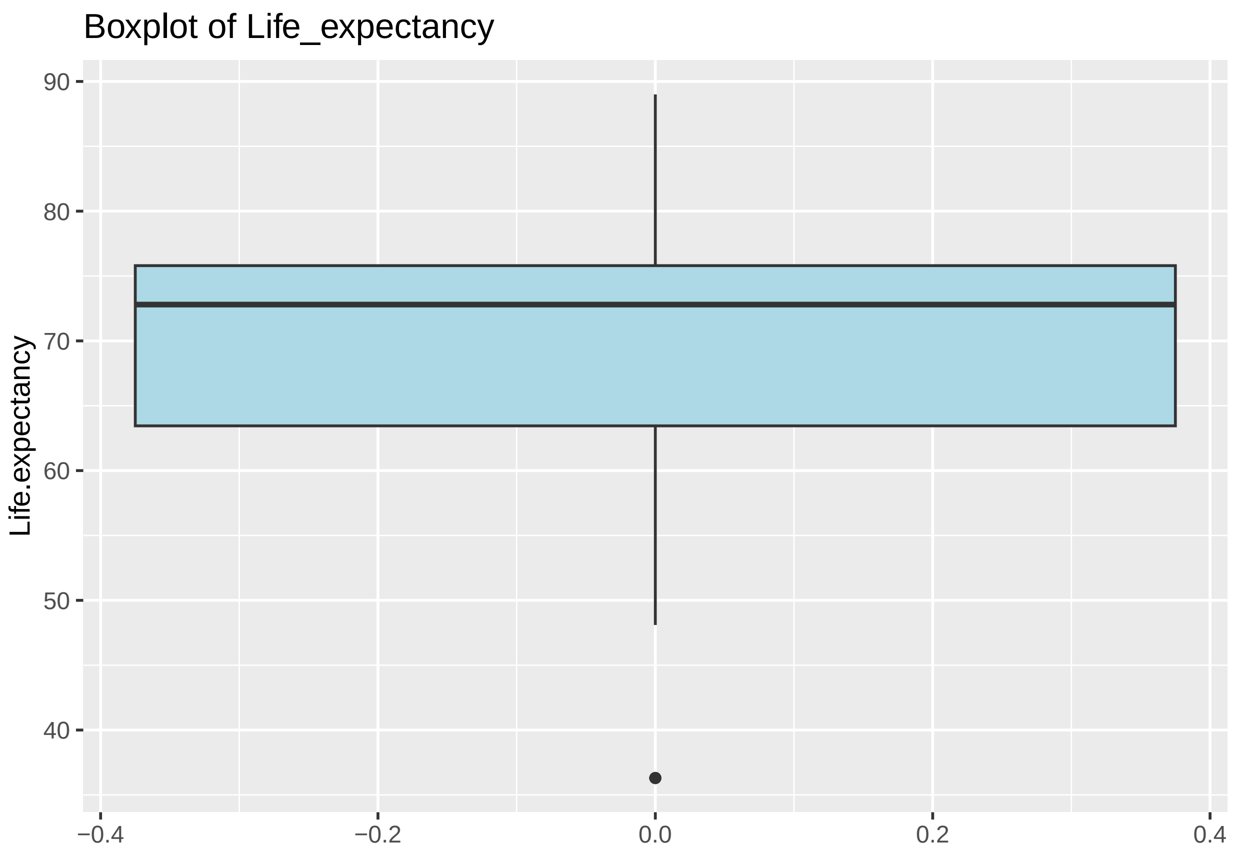
```
# For 'Alcohol'
ggplot(LED_2010_imputed, aes(x=Alcohol)) +
  geom_histogram(binwidth=1, fill="blue", color="black") +
  ggtitle("Distribution of Alcohol")
```



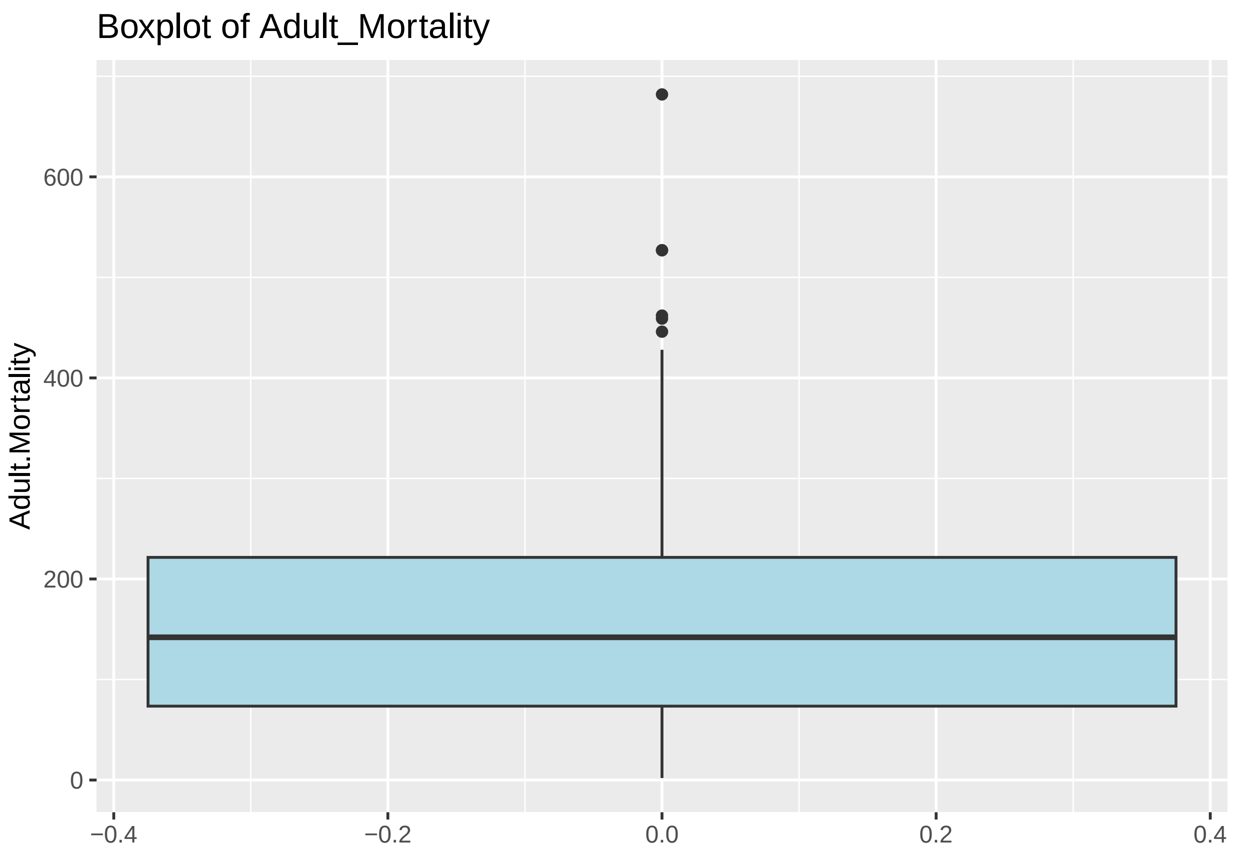
```
# For 'percentage_expenditure'
ggplot(LED_2010_imputed, aes(x=percentage.expenditure)) +
  geom_histogram(binwidth=1, fill="blue", color="black") +
  ggtitle("Distribution of percentage_expenditure")
```



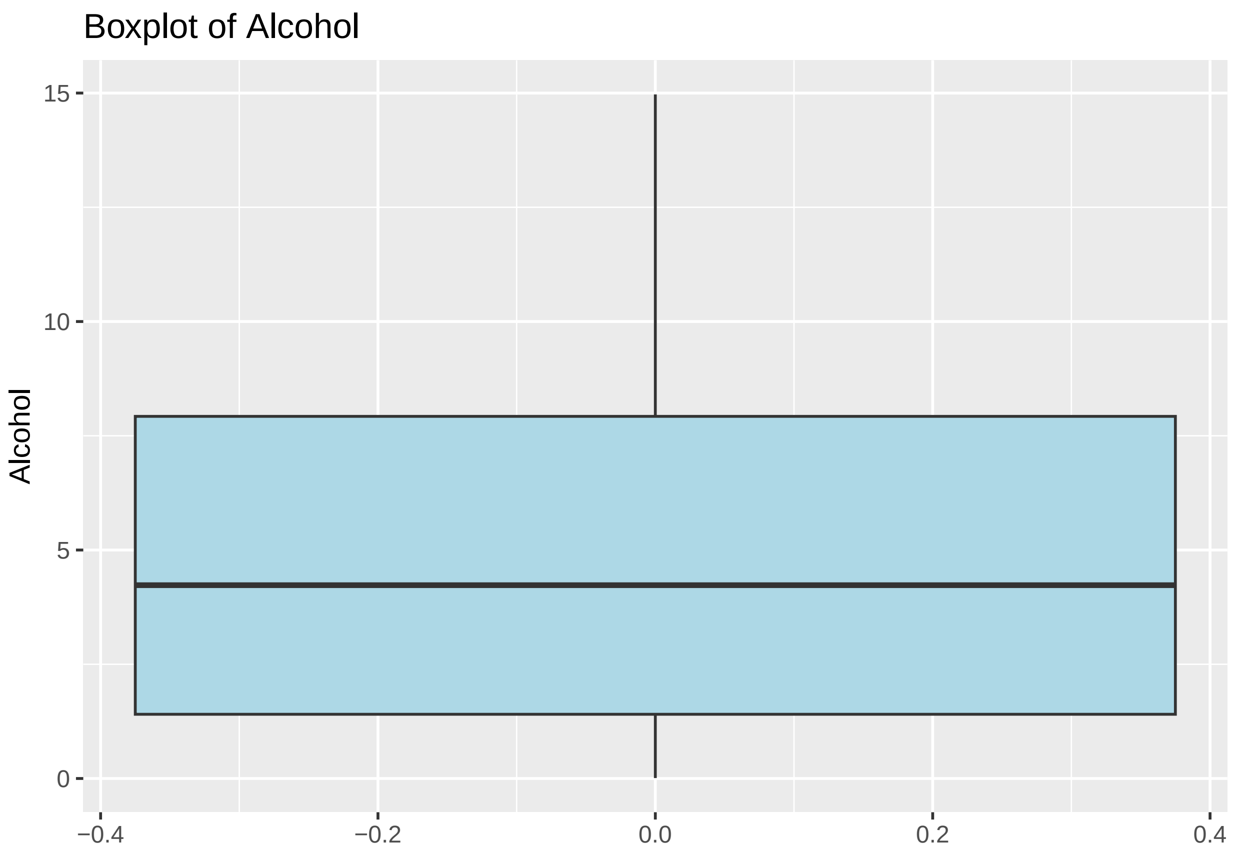
```
# For 'Life_expectancy'
ggplot(LED_2010_imputed, aes(y=Life.expectancy)) +
  geom_boxplot(fill="lightblue") +
  ggtitle("Boxplot of Life_expectancy")
```



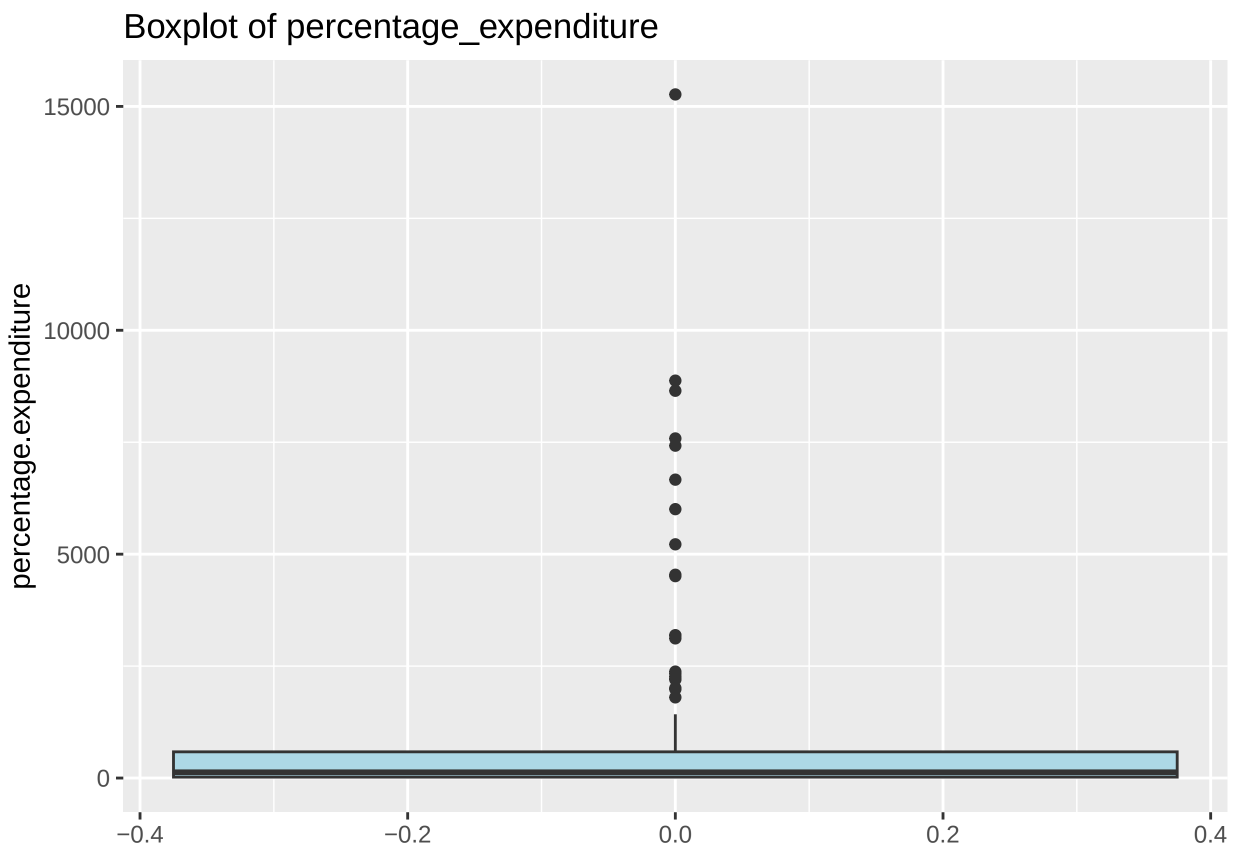
```
# For 'Adult_Mortality'
ggplot(LED_2010_imputed, aes(y=Adult.Mortality)) +
  geom_boxplot(fill="lightblue") +
  ggtitle("Boxplot of Adult_Mortality")
```



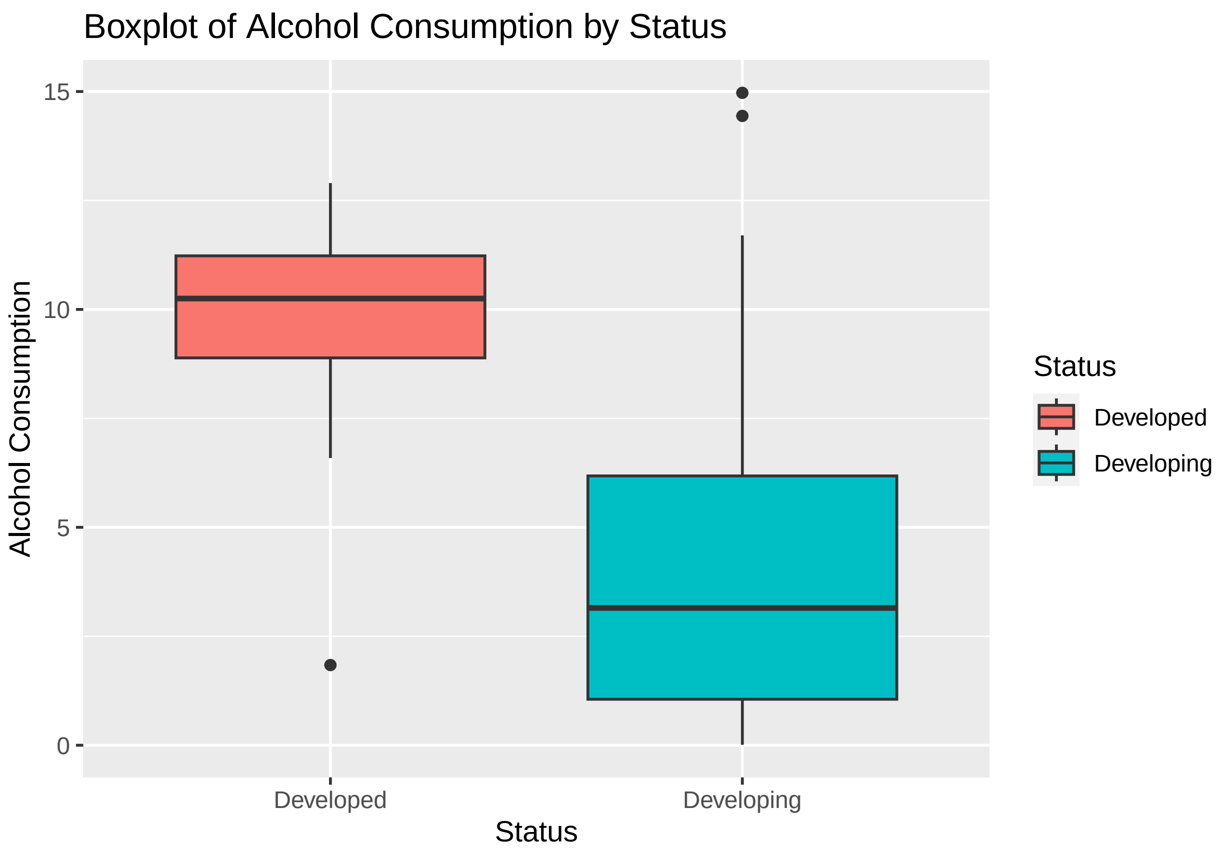
```
# For 'Alcohol'
ggplot(LED_2010_imputed, aes(y=Alcohol)) +
  geom_boxplot(fill="lightblue") +
  ggtitle("Boxplot of Alcohol")
```



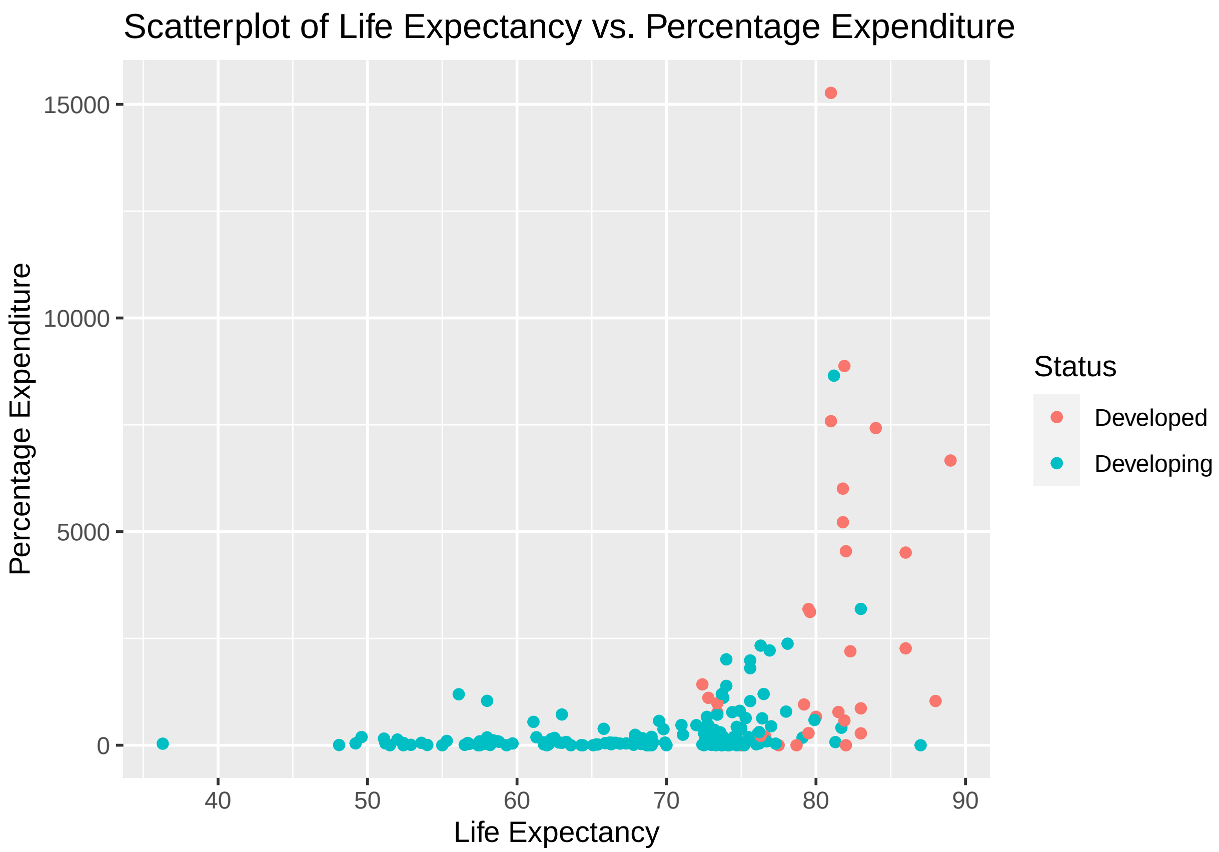
```
# For 'percentage_expenditure'
ggplot(LED_2010_imputed, aes(y=percentage.expenditure)) +
  geom_boxplot(fill="lightblue") +
  ggtitle("Boxplot of percentage_expenditure")
```



```r
# Create boxplot of 'Alcohol' by 'Status'
ggplot(LED_2010_imputed, aes(x=Status, y=Alcohol, fill=Status)) +
  geom_boxplot() +
  labs(title="Boxplot of Alcohol Consumption by Status",
       x="Status",
       y="Alcohol Consumption",
       fill="Status")
```



```
# Generate scatterplot
ggplot(LED_2010_imputed, aes(x=Life.expectancy, y=percentage.expenditure, color=Status)) +
  geom_point() +
  labs(title="Scatterplot of Life Expectancy vs. Percentage Expenditure",
       x="Life Expectancy",
       y="Percentage Expenditure",
       color="Status")
```



1. Investigating the effect of changes in alcohol consumption on life expectancy in developed vs underdeveloped countries:

```
# Fit the linear regression model
model_alcohol <- lm(Life.expectancy ~ Alcohol + Status, data=LED_2010_imputed)

# Show the summary of the model
summary(model_alcohol)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Alcohol + Status, data = LED_2010_imputed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.311  -5.003   1.487   5.943  17.069
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        76.4154     2.3232  32.893  < 2e-16 ***
## Alcohol             0.3802     0.1877   2.025   0.0443 *
## StatusDeveloping   -9.9940     1.9117  -5.228 4.72e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8.005 on 180 degrees of freedom
## Multiple R-squared:  0.2678, Adjusted R-squared:  0.2596
## F-statistic: 32.91 on 2 and 180 DF,  p-value: 6.589e-13
```

2. Investigating the effect of changes in health expenditure (percentage expenditure) on life expectancy in developed vs underdeveloped countries:

```
# Fit the linear regression model
model_health_expenditure <- lm(Life.expectancy ~ percentage.expenditure + Status, data=LED_2010_imputed)

# Show the summary of the model
summary(model_health_expenditure)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ percentage.expenditure + Status,
##     data = LED_2010_imputed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.270  -5.057   1.269   6.053  19.469
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            77.2527723  1.7032827  45.355  < 2e-16 ***
## percentage.expenditure  0.0010641  0.0003582   2.971  0.00337 **
## StatusDeveloping       -9.7215699  1.7560035  -5.536 1.08e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.904 on 180 degrees of freedom
## Multiple R-squared:  0.2861, Adjusted R-squared:  0.2782
## F-statistic: 36.07 on 2 and 180 DF,  p-value: 6.734e-14
```