Lab Formative Assessment 2

DSC1105 Exploratory Data Analysis

Sindayen, Lorenzo Danilo V.

2026-01-29

GitHub Link:
https://github.com/ChewyGnome/DSC1105

Dataset

The dataset that will be used in this report is the built-in mpg dataset included in the ggplot2 package.

```
data(mpg)
```

Data Inspection

Structure of the dataset:

```
class(mpg)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

First 6 rows of the dataset:

```
head(mpg)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

Data Wrangling and Transformation

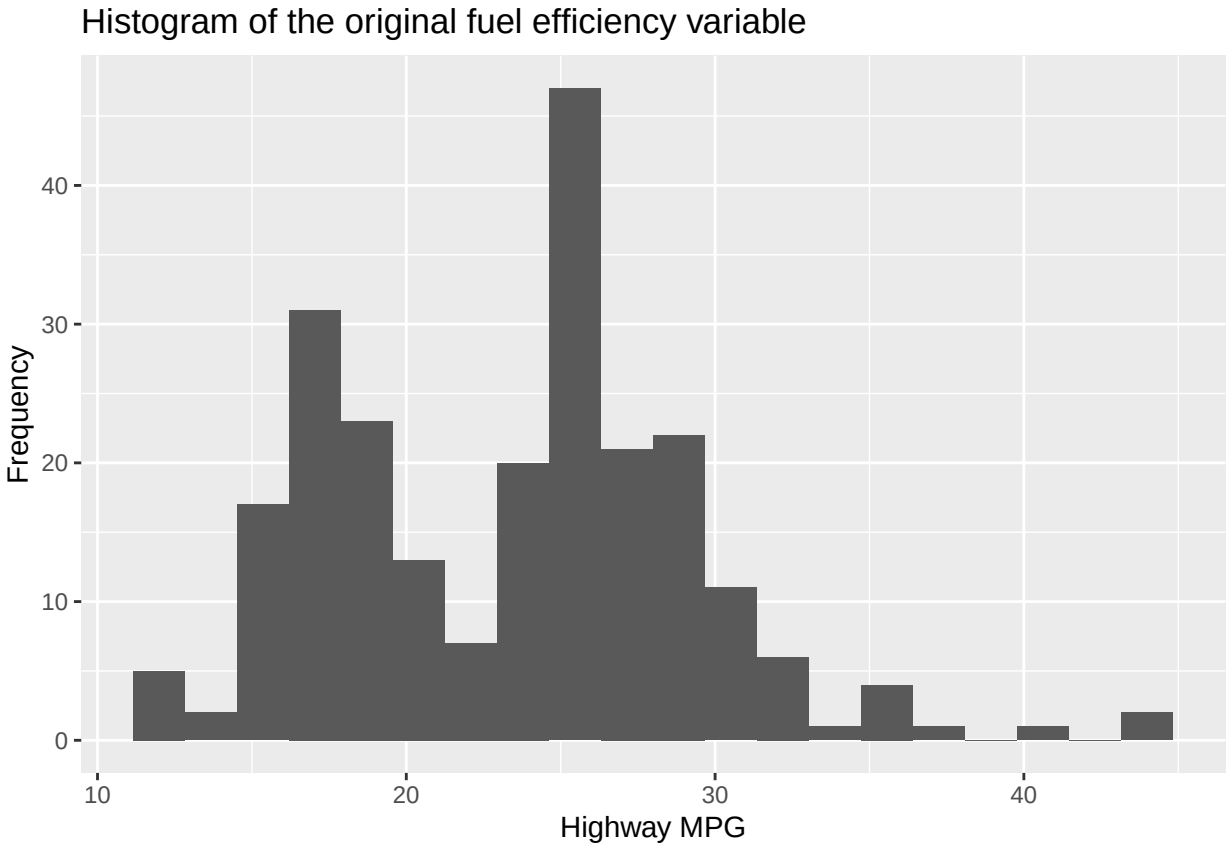
```
mpg_wrangle <- mpg %>%
  separate(
    trans,
    into = c("transmission_type", "gears"),
    sep = "\\(",
    remove = FALSE
  ) %>%
  mutate(gears = str_remove(gears, "\\)"))

mpg_wrangle <- mpg_wrangle %>%
  mutate(
    log_hwy = log(hwy),
    sqrt_displ = sqrt(displ)
  )

mpg_wrangle %>%
  select(trans, transmission_type, gears, log_hwy, sqrt_displ) %>%
  head()
```

```
## # A tibble: 6 x 5
##   trans      transmission_type gears log_hwy sqrt_displ
##   <chr>      <chr>             <chr>   <dbl>      <dbl>
## 1 auto(l5)   auto              l5       3.37       1.34
## 2 manual(m5) manual            m5       3.37       1.34
## 3 manual(m6) manual            m6       3.43       1.41
## 4 auto(av)   auto              av       3.40       1.41
## 5 auto(l5)   auto              l5       3.26       1.67
## 6 manual(m5) manual            m5       3.26       1.67
```

Visualization

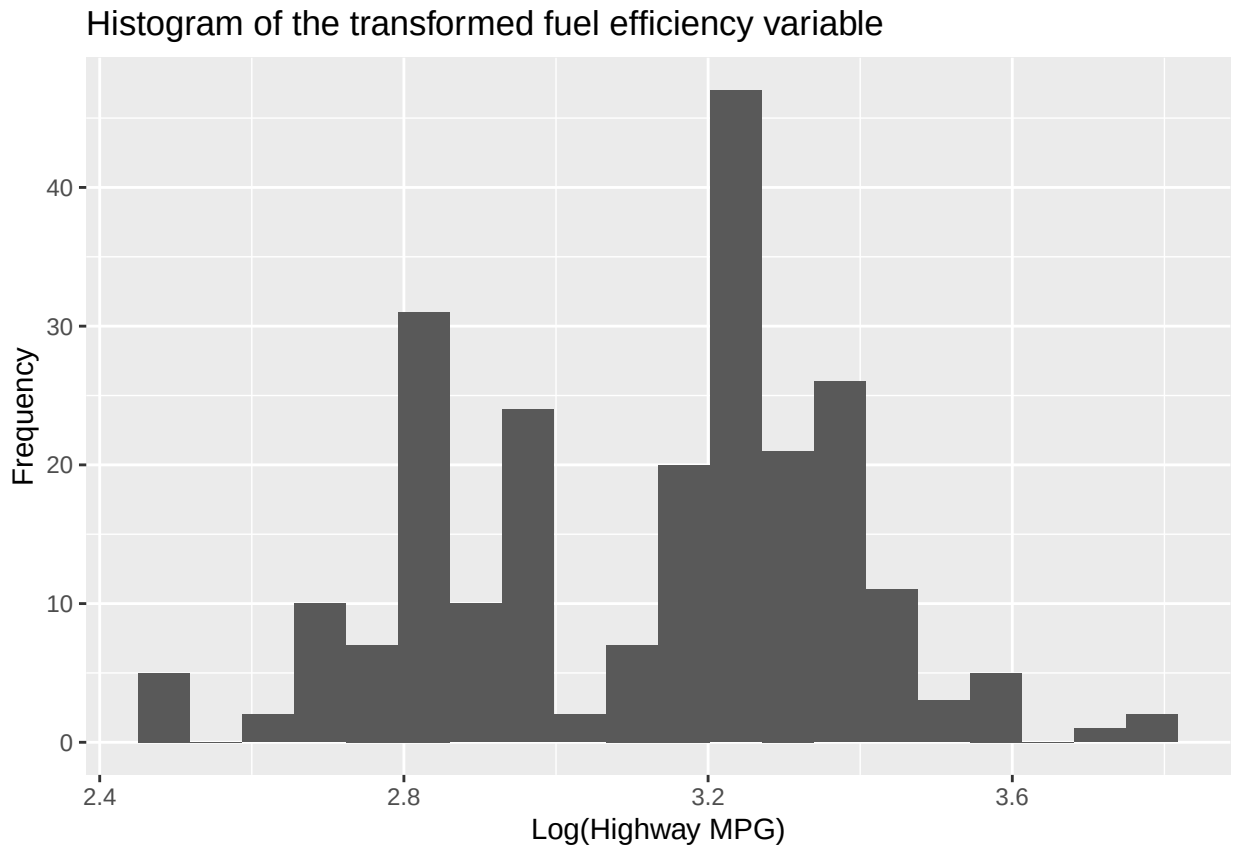
Histogram of Original Highway Fuel Efficiency

```
ggplot(mpg_wrangle, aes(x = hwy)) +
  geom_histogram(bins = 20) +
  labs(
    title = "Histogram of the original fuel efficiency variable",
    x = "Highway MPG",
    y = "Frequency"
  )
```



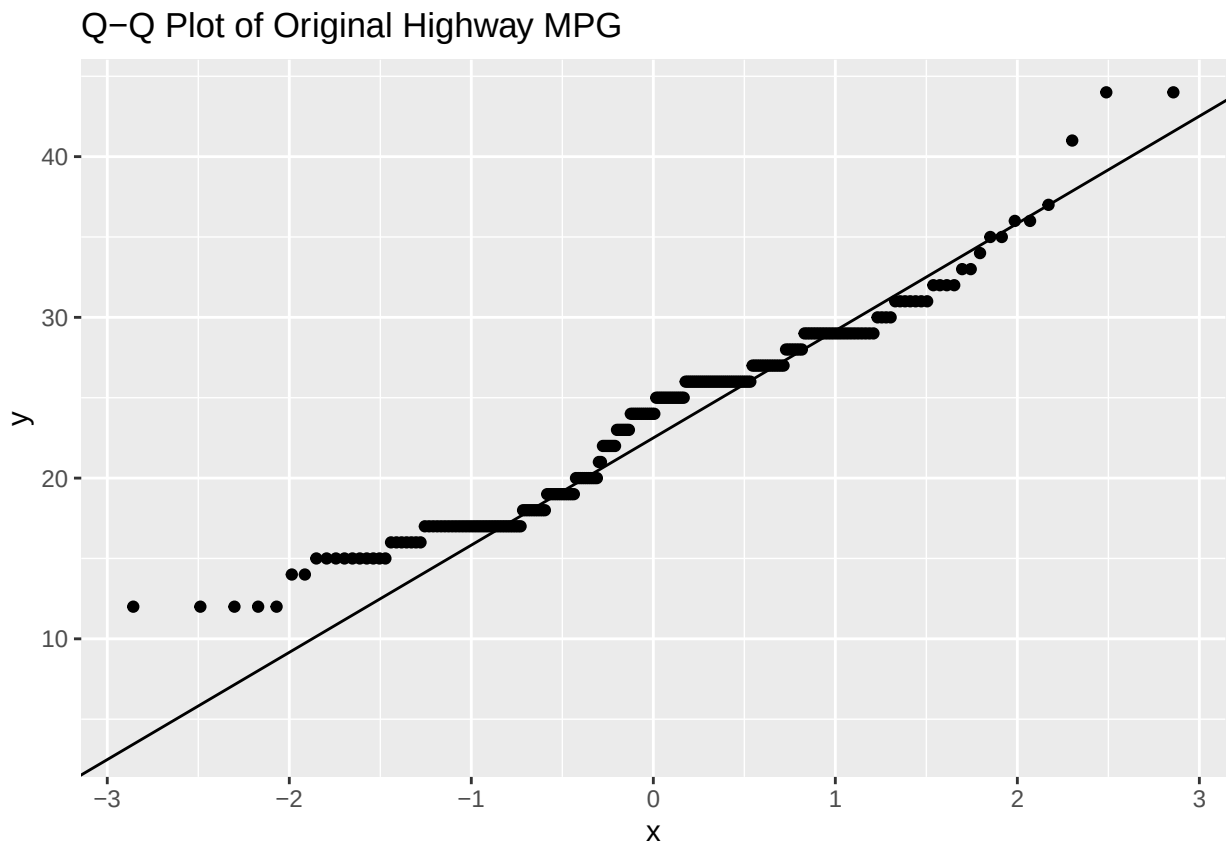
Histogram of the Transformed Highway Fuel Efficiency

```
ggplot(mpg_wrangle, aes(x = log_hwy)) +
  geom_histogram(bins = 20) +
  labs(
    title = "Histogram of the transformed fuel efficiency variable",
    x = "Log(Highway MPG)",
    y = "Frequency"
  )
```



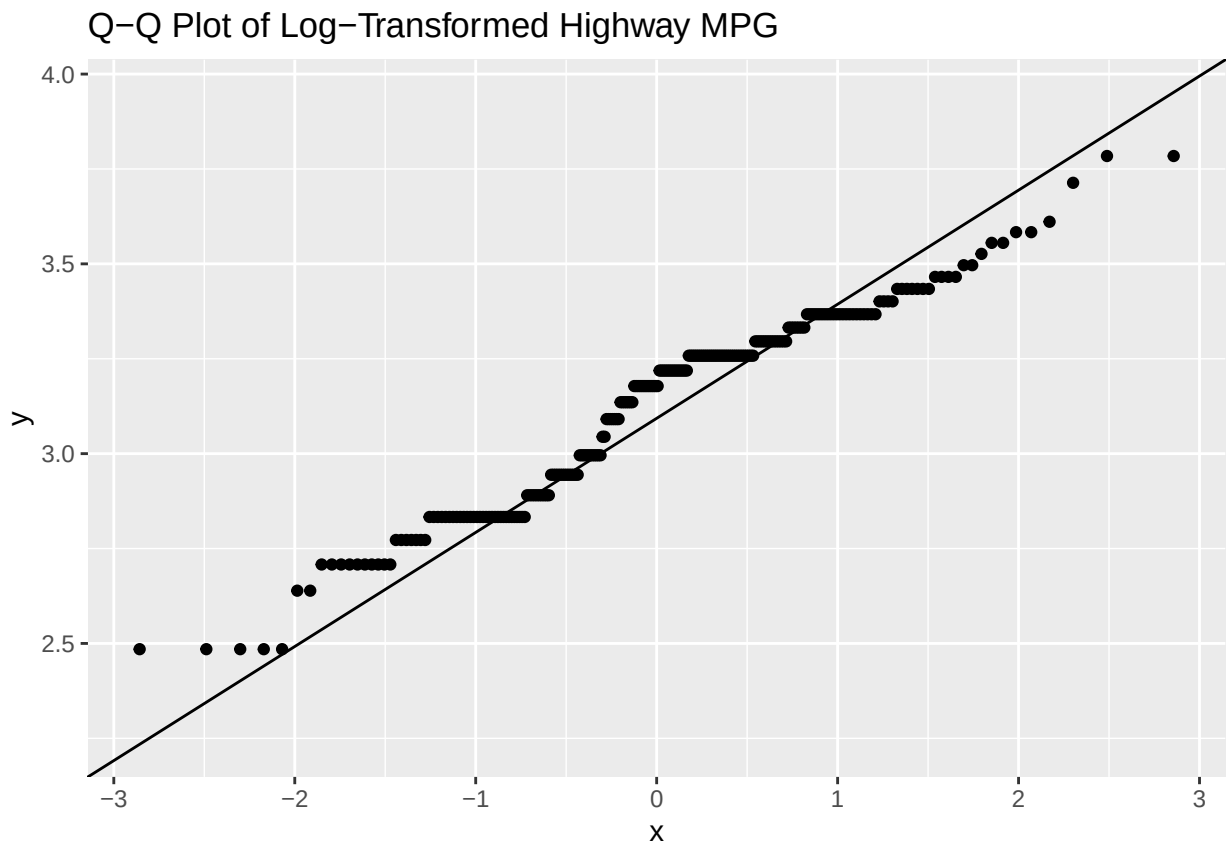
Q-Q plot of Original Highway MPG

```
ggplot(mpg_wrangle, aes(sample = hwy)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Q-Q Plot of Original Highway MPG")
```



Q-Q plot of Log-Transformed Highway MPG

```
ggplot(mpg_wrangle, aes(sample = log_hwy)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Q-Q Plot of Log-Transformed Highway MPG")
```



Interpretation

In the data provided, the original highway fuel efficiency exhibits a right skewness in the distribution of data, evident in the both the histogram and the Q-Q plot. However, after the application of the logarithmic transformation, it can be seen that the distribution of the data became a lot more balance. This could also be seen in the Q-Q plot of the transformed variable, where it can be seen that the alignment of the points are a lot more standardized, indicating a decreased deviation of the results.

Reflection

A log transformation was chosen for highway fuel efficiency due to the fuel economy data showing a skewed distribution of the dataset. The square root transformation of engine displacement helps reduce the influence of large values while preserving interpretability. Concluding this, the transformation of the variable had a massive improvement in normalizing the results. Improvements such as this could be crucial in simplifying the constraints of some certain datasets that rely on consistency and a constant variance.