Data Exploration Wifi - Load Performance vs Energy Efficiency

10/22/2020

```
options(digits = 3, warn = -1)

# Load Libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.3
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(e1071)
library(ggplot2)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
##
## Attaching package: 'gridExtra'
```

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

This sample contains **30 observations** of several load performance features, collected over **Wifi connection**.

Factors (Perfume.js metrics):

1. **fcpAvg** - First Contentful Paint (ms)
2. **fpAvg** - First Paint (ms)
3. **totalLoadTimeAvg** (ms) - The time taken for the entire page to be loaded (the time taken until the onload callback is called)
4. **fcpNormalized** - Total time (ms) - Average request plus response time (network only)
5. **fpNormalized** - Normalized size of the header
6. **totalLoadTimeNormalized** (ms) - Time to First Byte (TTFB) - Normalized amount of time it takes for the server to send the first payload to the client
7. **loadMean** - Performance score from the perspective of load speed
8. **performanceLvl** - Ordinal performance score

Numerical outcome (dependent variable): **batteryStatsAvg** - Mean energy efficiency measured with battery stats (Joules)

All averages represent the mean of the values measured for 11 runs, for each subject.

The **loadMean** (the aggregated performance score from the perspective of load timing) is an average score of the normalized load speed metrics (i.e., fp, fcp, totalLoadTime).

```
load_df =  read_csv("./LoadSpeed_Wifi.csv", col_names = TRUE)
```

```
##
## -- Column specification ------------------------------------------------------------------------------------------------
-------
## cols(
##   X1 = col_character(),
##   subject = col_character(),
##   totalLoadTimeAvg = col_double(),
##   fcpAvg = col_double(),
##   fpAvg = col_double(),
##   totalLoadTimeNormalized = col_double(),
##   fcpNormalized = col_double(),
##   fpNormalized = col_double(),
##   loadMean = col_double(),
##   batteryStatsAvg = col_double(),
##   performanceLvl = col_character(),
##   website = col_character()
## )
```

```
load_df = select(load_df, -c(X1))

View(load_df)
```

Summary of energy efficiency:

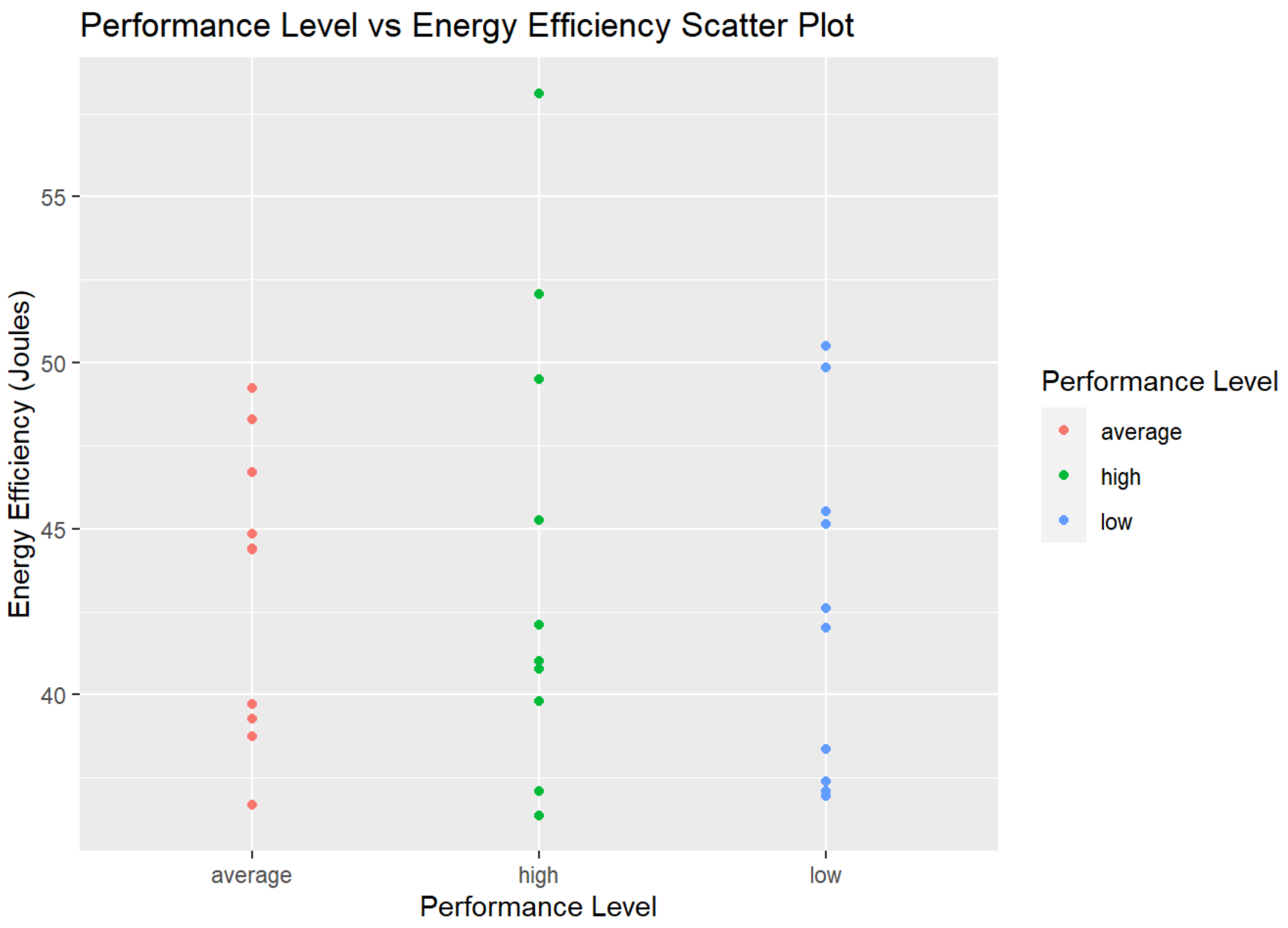
```
summary(load_df$batteryStatsAvg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    36.4    38.9    42.4    43.3    46.4    58.1
```

Exploring the data

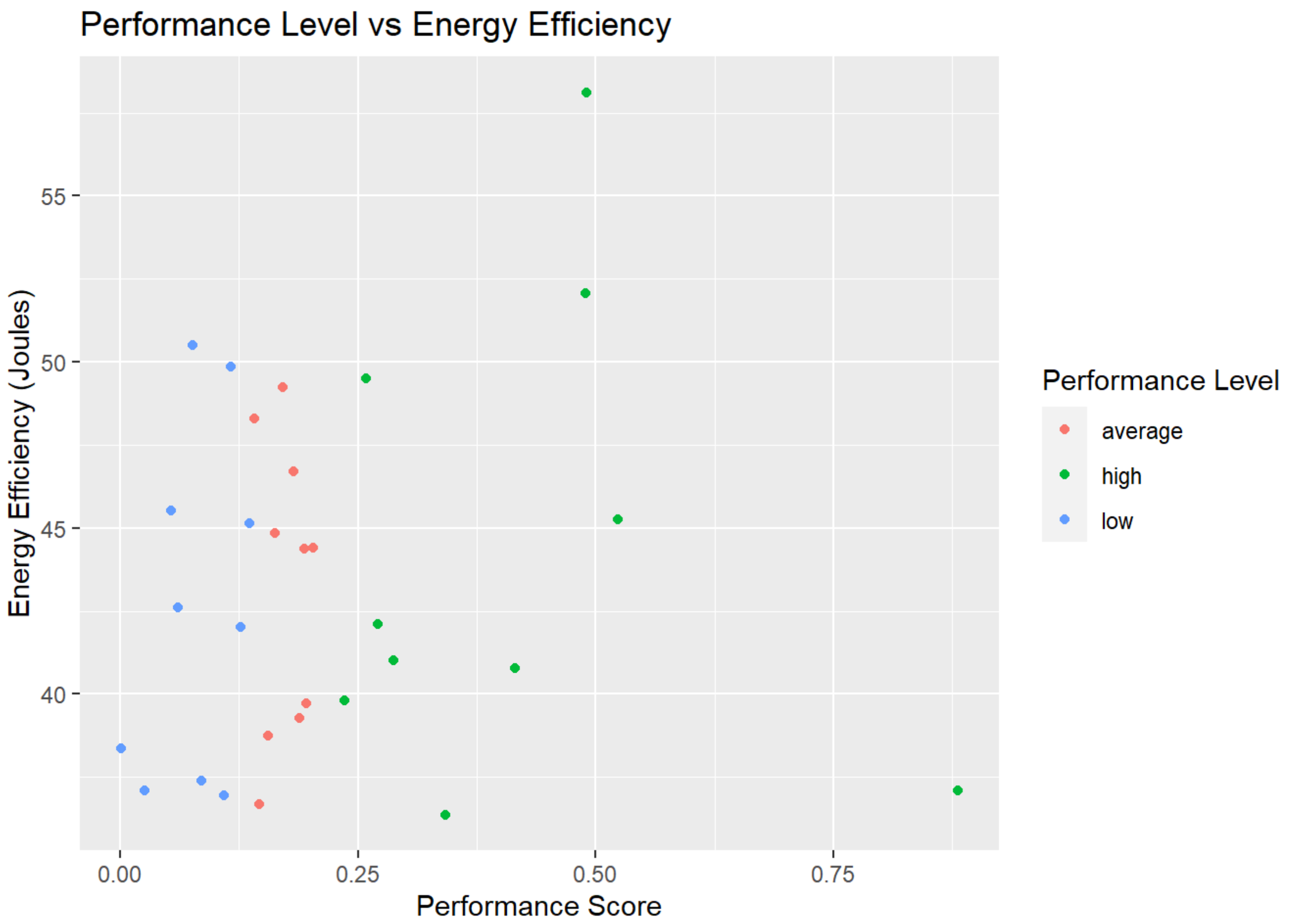
Performance Levels vs Energy Efficiency Scatter Plot

```
ggplot(data.frame(y = load_df$batteryStatsAvg), aes(x = load_df$performanceLvl, y = y, sample = y, color = load_df$performan
ceLvl)) +
  labs(title="Performance Level vs Energy Efficiency Scatter Plot", x="Performance Level",  y = "Energy Efficiency (Joules)"
, color = "Performance Level")+
  geom_point()
```



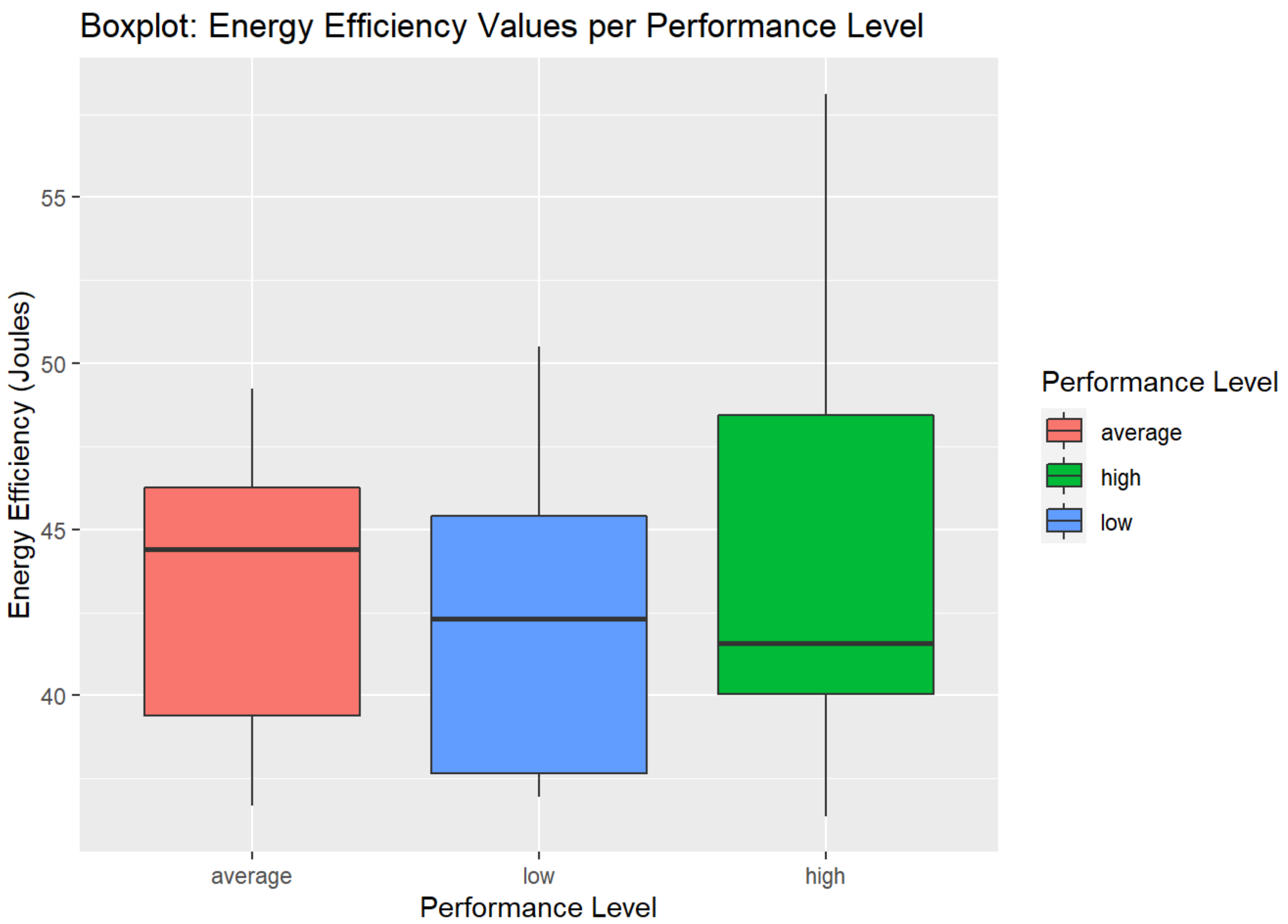
Performance score vs Energy Efficiency Scatter Plot

```
ggplot(data.frame(y = load_df$batteryStatsAvg), aes(x = load_df$loadMean, y=y, sample = y, color = load_df$performanceLvl))
+
  labs(title="Performance Level vs Energy Efficiency", x="Performance Score",  y = "Energy Efficiency (Joules)", color = "Pe
rformance Level")+
  geom_point()
```



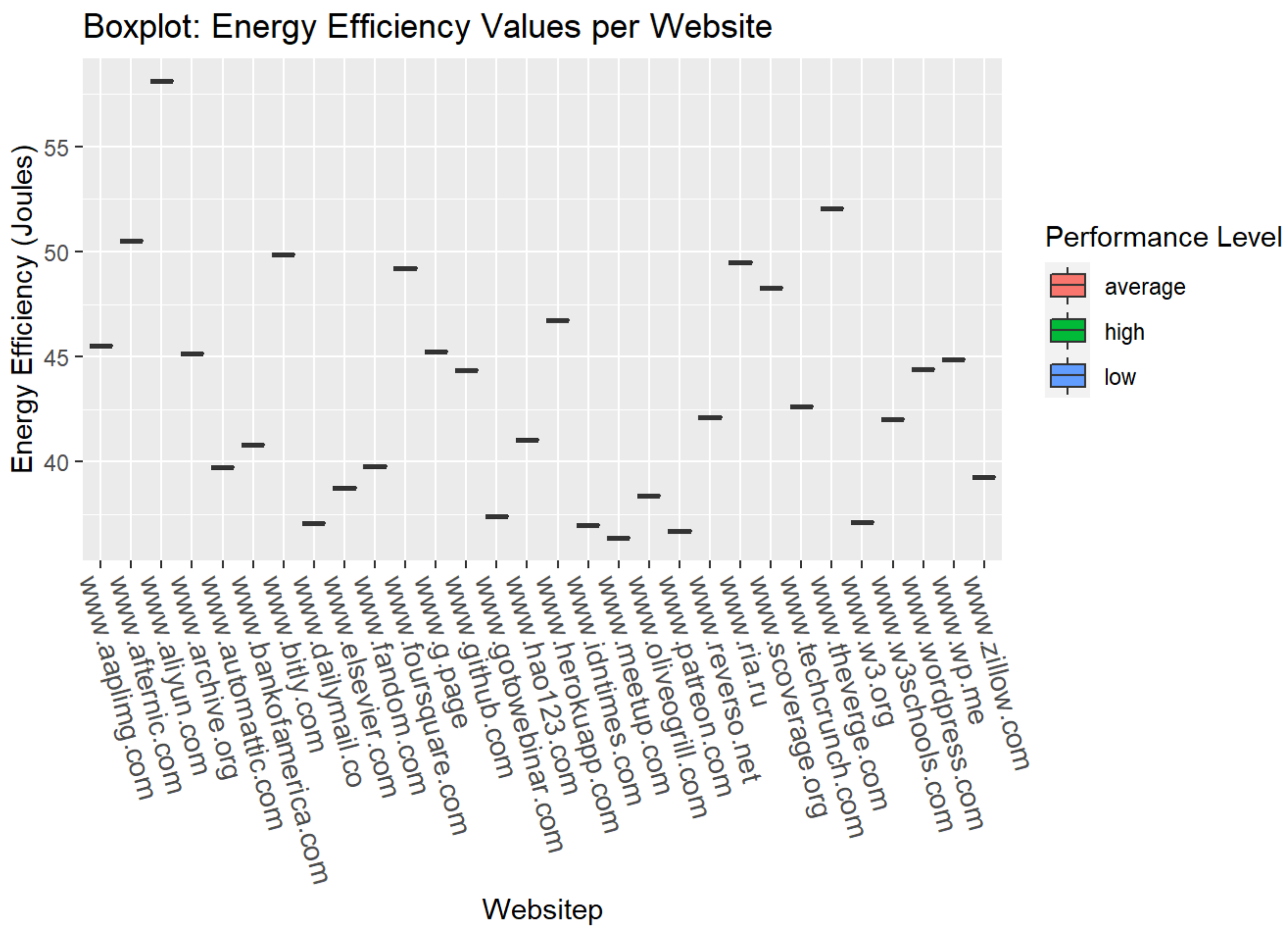
Box plot: Energy Efficiency vs Performance Level

```
ggplot(data.frame(y = load_df$batteryStatsAvg), aes(x = rev(load_df$performanceLvl), y=y, sample = y, fill = (load_df$perfor
manceLvl))) +
  labs(title="Boxplot: Energy Efficiency Values per Performance Level", x="Performance Level",  y = "Energy Efficiency (Joul
es)", fill = "Performance Level")+
  geom_boxplot() + scale_x_discrete(breaks=c("high", "average", "low"),
                                    labels=c("low", "average", "high"))
```



Box Plot: Website vs Energy Efficiency

```
ggplot(data.frame(y = load_df$batteryStatsAvg), aes(x = load_df$website, y=y, sample = y, fill = load_df$performanceLvl)) +
  theme(axis.text.x = element_text(angle= -75, hjust = 0, size = 11))+
  labs(title="Boxplot: Energy Efficiency Values per Website", x="Websitep",  y = "Energy Efficiency (Joules)", fill = "Perfo
rmance Level")+
  geom_boxplot() + scale_x_discrete(limits= rev(levels(load_df$website)))
```



Investigating the normality of the dependent variable

We already know the dependent variable, i.e., energy efficiency is normal. This has already been analyzed in Data_transformation_Wifi.Rmd, where we considered the reciprocal of the energy efficiency, as this was the least skewed distribution.

```
batteryStatsReciprocal <- 1 / load_df$batteryStatsAvg
```

Hypothesis testing

$H_0 : \mu1_{high} = \mu1_{average} = \mu1_{low}$

i.e.,: The mean energy efficiency does not significantly differ among web apps having different load speed levels

$H_a : \mu1_{high} \neq \mu1_{average} \vee \mu1_{average} \neq \mu_{low} \vee \mu1_{low} \neq \mu_{high}$

i.e.,: The mean energy efficiency significantly differs among web apps having different levels of load speed for at least one pair of load speed levels.

Test used and motivation: We have one factor, >2 treatments, hence we perform Analysis of Variance (ANOVA) to test for effect of the treatments (load speed levels) on the mean energy efficiency.

One-way ANOVA can be used in the analysis to test the effect of each performance category (i.e., load speed, load speed) on energy efficiency. In this case, we assume: - The dependent variable (energy efficiency) is continuous - TRUE - The samples are independent - TRUE (guaranteed by the experimental design) - Normal distribution of the dependent variable between the independent groups - TRUE - Residuals (i.e., errors in the sample) should be normally distributed - TO BE CHECKED AFTER FITTING THE MODEL - Homoscedasticity (variance between groups should be the same) - TO BE CHECKED AFTER FITTING THE MODEL

```
# quantitative variable, representing the average power consumed in each round of the experiment
efficiency = as.numeric(batteryStatsReciprocal)
# the treatment, categorical variable with 3 levels - high, average, low
performance = as.factor(load_df$performanceLvl)
summary(efficiency) # obtain a numeric summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0172  0.0215  0.0236  0.0234  0.0257  0.0275
```

```
summary(performance) # listed as categorical variable, with the number of observations at each level
```

```
## average    high     low
##      10      10      10
```

```
load_df.aov <- aov(efficiency~performance, data = load_df)
summary(load_df.aov)
```

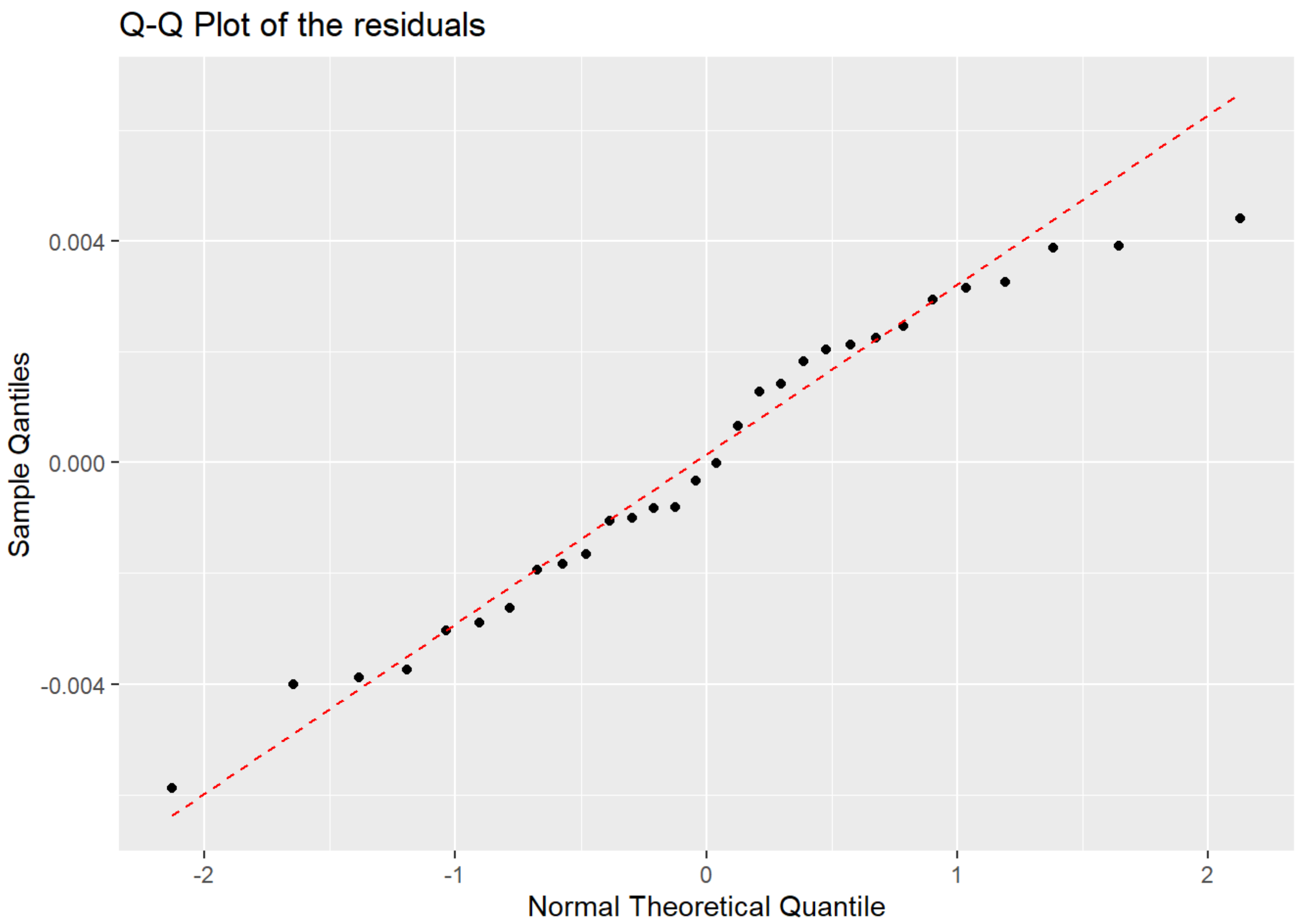
```
##             Df   Sum Sq  Mean Sq F value Pr(>F)
## performance  2 2.60e-06 1.30e-06    0.16   0.86
## Residuals   27 2.24e-04 8.29e-06
```

Outcome: The case factor is giving a p-value higher than 0.05, meaning that we cannot reject the null hypothesis, stating that the mean energy efficiency does not significantly differ among web apps having different load speed levels. Hence, there is no statistical significance between the variables.

Check normality of residuals

QQ-plot

```
ggplot(data.frame(y = residuals(load_df.aov)), aes(sample = y)) + stat_qq() + stat_qq_line(col="red", lty=2) + ylab("Sample Qantiles") + xlab("Normal Theoretical Quantile") + ggtitle("Q-Q Plot of the residuals")
```



Shapiro-Wilk normality test

Hypothesis:

H_0: energy efficiency data is normally distributed

H_1: energy efficiency data is NOT normally distributed

```
shapiro.test(residuals(load_df.aov))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(load_df.aov)
## W = 1, p-value = 0.4
```

Outcome: The p-value for testing H_0 is greater than 0.05, hence we cannot reject the null hypothesis that energy efficiency data is normally distributed.

Check homoscedacity

The spread of the data points between the groups must be the same.

Levene's test with one independent variable

Hypothesis:

H_0: the population variances are equal

H_1: the population variances are NOT equal

```
leveneTest(efficiency~performance, data = load_df)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2    0.41   0.67
##       27
```

Outcome:

The p-value of Levene's test is higher than the significance level (0.05), hence the differences in sample variances are likely to have occurred based on random sampling from a population with equal variances. Thus, we cannot reject the null hypothesis of equal variances.

Hence, the assumptions of ANOVA are all valid.

We conclude that there is no evidence of statistical significance between performance and energy efficiency.