Cognition & Communication - Alex Presa Hughes

Alex Presa Hughes

2024-12-24

This document contains all the code used to perform statistical analyses on the data collected for Alex's experiment, which was conducted at Aarhus University, Denmark. The experiment consisted of data scraping initially to collect the relevant headlines and then analyse them using Python packages to get word count, readability, polarity and subjectivity. Once that was done, I proceeded with the collection of participants. In total, 24 participants volunteered and their data is within the data frame I import further on. The experimenter gathered 24 headlines and put them into three groups and participants went through three rounds and in each one they were initially shown 4 headlines and shortly after they saw all 8 of the specific group in sequence and they had to remember which ones they had seen, which were of course the initial 4 headlines. The goal of the experiment was to determine if there is an influence of word count, readability, polarity anf subjectivity on reaction time and recall accuracy when participants were asked whether they had previously seen the headlines they were being shown. The experimenter also checked if participants took longer to give their correct or incorrect responses and if participants took longer to respond when presented with a previously seen or unseen headline.

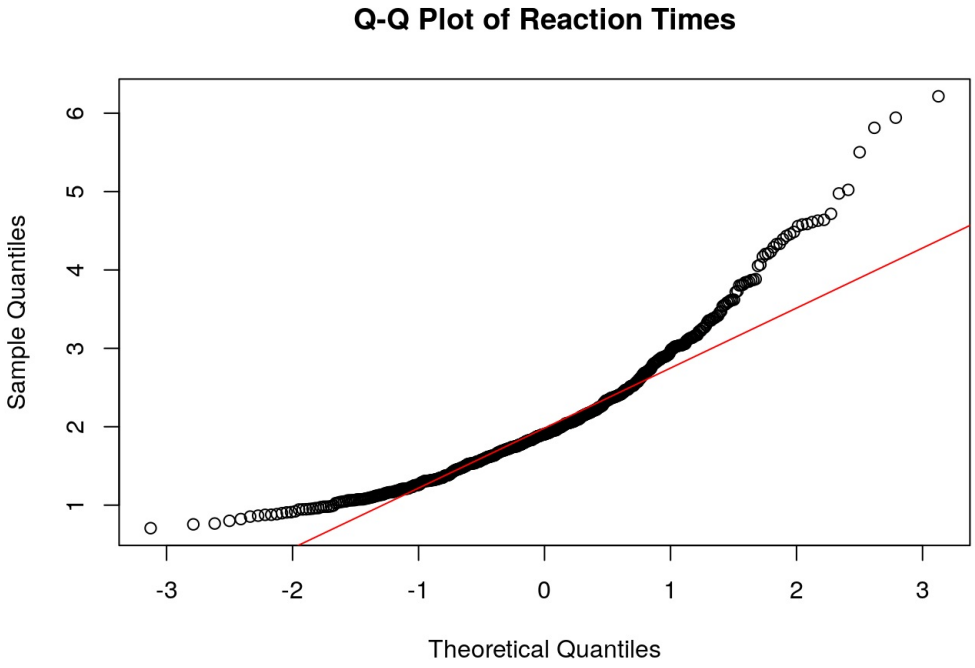
```
# Here I am installing some basic packages to proceed with the data analysis. I also import the data frame and el
iminate a surplus column which is not needed.

library(readr)
library(dplyr)
library(tidyverse)
library(sandwich)
library(lme4)
library(car)
```

```
# My data frame is contained within a csv file.

df <- read_delim("/work/AlexPresa-Hughes#2584/Experiment_Data_Headlines.csv", delim = ",")
df <- select(df,-round_number)
summary(df)

# Now I will remove any reaction time data points more than 3 standard deviations away from the mean. This will e
nsure that outliers will be removed. To do this, I will use the mean and standard deviation.

mean_rt <- mean(df$reaction_time)
sd_rt <- sd(df$reaction_time)

df <- df[abs(df$reaction_time - mean_rt) <= 3 * sd_rt, ]

df$n_reaction_time <- df$reaction_time
df$n_number_of_words <- df$number_of_words
df$n_readability <- df$readability
df$n_polarity <- df$polarity
df$n_subjectitivty <- df$subjectivity
```

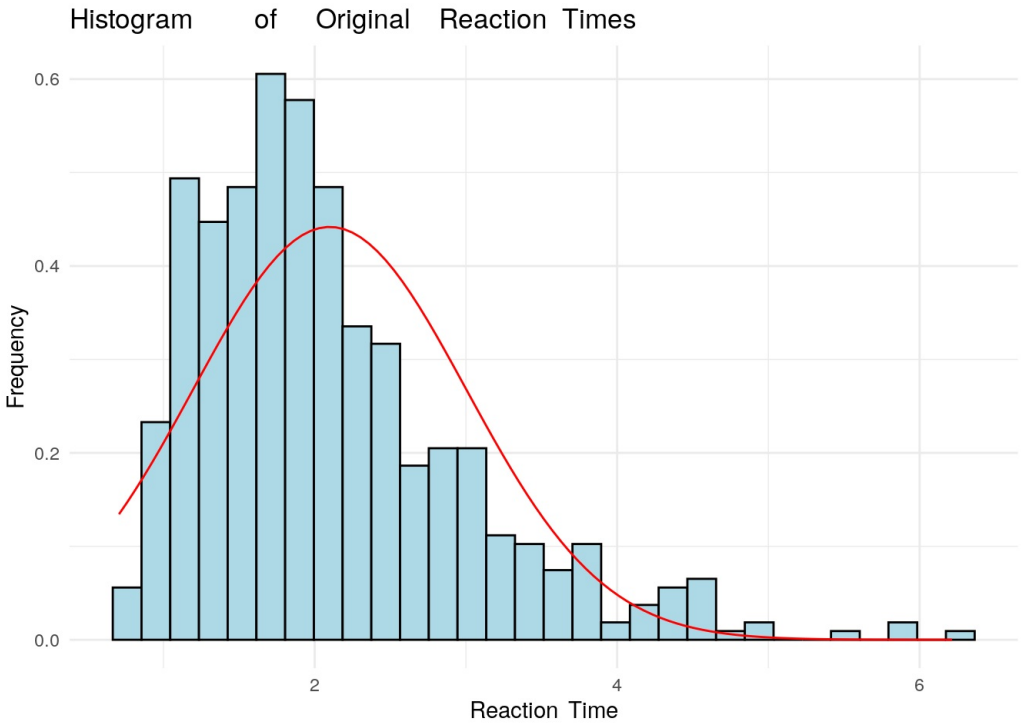
```
# I am also standardizing the independent variables so they are on the same scale because number of words varies
between 7 and 18, readability between 0 and 100, polarity between -1 and 1 and subjectivity between 0 and 1. If I
did not standardize them, it would affect the validity of the results as the different scales of the variables wo
uld distort the results.

df$number_of_words <- scale(df$number_of_words)
df$readability <- scale(df$readability)
df$polarity <- scale(df$polarity)
df$subjectivity <- scale(df$subjectivity)

# Now I will check for normality with a Shapiro-Wilk test that gives many measures relevant when determining if d
ata is normally distributed, such a skewness, kurtosis and a p-value. The skewness (skew.2SE) and kurtosis (kurt.
2SE) values typically have to be below 1 to indicate normality and the p-value has to be above the threshold of 0
.05 to indicate normality.

round(pastecs::stat.desc(cbind(df$reaction_time), basic = FALSE, norm = TRUE), digits = 4)
```

```
##                 V1
## median       1.9030
## mean         2.0993
## SE.mean      0.0380
## CI.mean.0.95 0.0746
## var          0.8153
## std.dev      0.9029
## coef.var     0.4301
## skewness     1.2891
## skew.2SE     6.2710
## kurtosis     2.0623
## kurt.2SE     5.0251
## normtest.W   0.9087
## normtest.p   0.0000
```

```
# Now I will create a QQ-plot of reaction times to visually inspect the distribution of the data points compared to those of a normal distribution and therefore determine if there are any tails.

qqnorm(df$reaction_time, main = "Q-Q Plot of Reaction Times")
qqline(df$reaction_time, col = "red")
```




```
# Below I am going to calculate the mean and standard deviation of reaction time to be able to construct a theoretical normal distribution. Then I will plot the reaction time data in a histogram, with the hypothetical normal distribution against it to see if they align.

mean_reaction_time <- mean(df$reaction_time, na.rm = TRUE)
sd_reaction_time <- sd(df$reaction_time, na.rm = TRUE)
ggplot(df, aes(x = reaction_time)) +
geom_histogram(aes(y = ..density..),bins = 30, color = "black", fill = "lightblue") +
stat_function(fun=dnorm,args = list(mean = mean_reaction_time, sd = sd_reaction_time), color="red")
+
theme_minimal() +
labs(title = "Histogram of Original Reaction Times",
x = "Reaction Time",
y = "Frequency")
```




```
# When looking at the QQ-plot, it is evident that points on both sides deviate from the red line and, in the hist
ogram, reaction time does not follow a normal distribution. Furthermore, a p-value below 0.05 means that it does
not resemble a normal distribution. The high skewness and kurtosis values indicate that the data does not follow
the bell shape characteristic of a normal distribution .It can be seen that the data without transformations is v
ery far from being normally distributed, therefore I will apply various transformations to see if it makes my dat
a normal. I will also standardize the data to evaluate the data points based on their distance from the mean meas
ured in standard deviations. The dependent variable should be transformed to try to make it indicate normality.

# I will apply a logarithmic transformation as my data has a noticeable right skew. This will hopefully deal with
that.

df$log_rt <- log(df$reaction_time) # Logarithmic transformation

round(pastecs::stat.desc(df$log_rt, basic = FALSE, norm = TRUE), digits = 4)
```

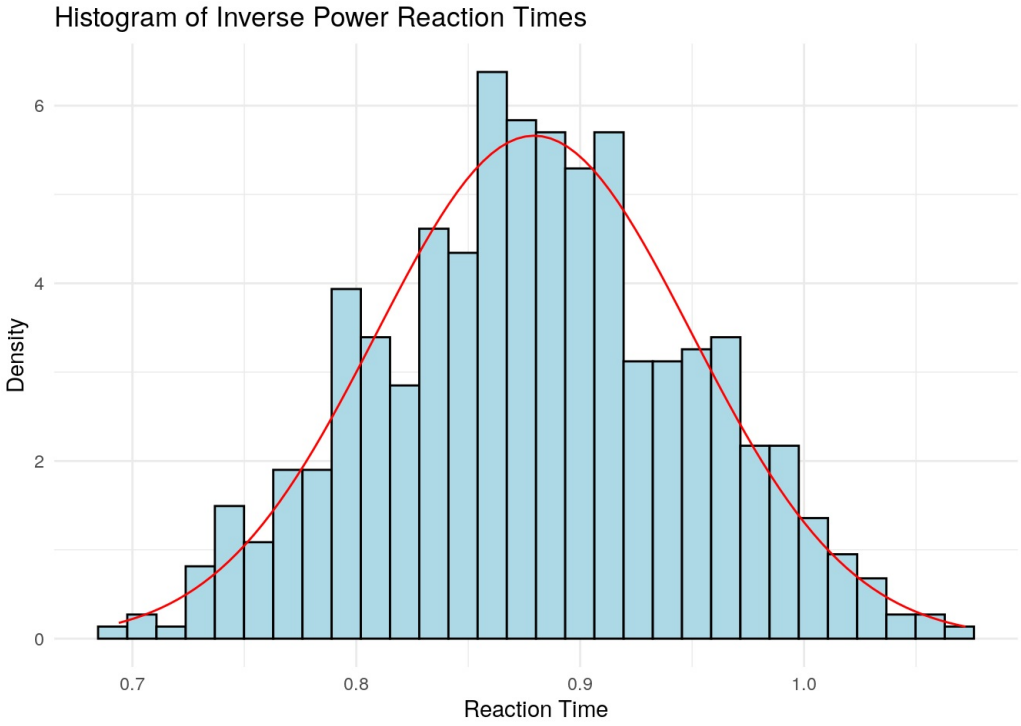
```
##       median         mean      SE.mean CI.mean.0.95          var      std.dev
##       0.6434       0.6586       0.0170       0.0333       0.1627       0.4033
##     coef.var     skewness     skew.2SE     kurtosis     kurt.2SE   normtest.W
##       0.6123       0.1912       0.9304      -0.2894      -0.7051       0.9945
##   normtest.p
##       0.0397
```

```
# The logarithmic transformation yields a relatively small p-value (p < 0.05), so we can definitely conclude that
this transformation is not appropriate.

# I will apply an inverse transformation to see if this yields different results.

df$inverse_rt <- 1/(df$reaction_time) # Inverse transformation

round(pastecs::stat.desc(df$inverse_rt, basic = FALSE, norm = TRUE), digits = 4)
```

```
##       median         mean      SE.mean CI.mean.0.95          var      std.dev
##       0.5255       0.5600       0.0094       0.0184       0.0496       0.2227
##     coef.var     skewness     skew.2SE     kurtosis     kurt.2SE   normtest.W
##       0.3976       0.7878       3.8327       0.4437       1.0812       0.9581
##   normtest.p
##       0.0000
```

```
# The inverse transformation yields a tiny p-value (p < 0.05), so we can definitely conclude that this transforma
tion is not appropriate.

# I will have to seek other transformations as these have not worked.
```

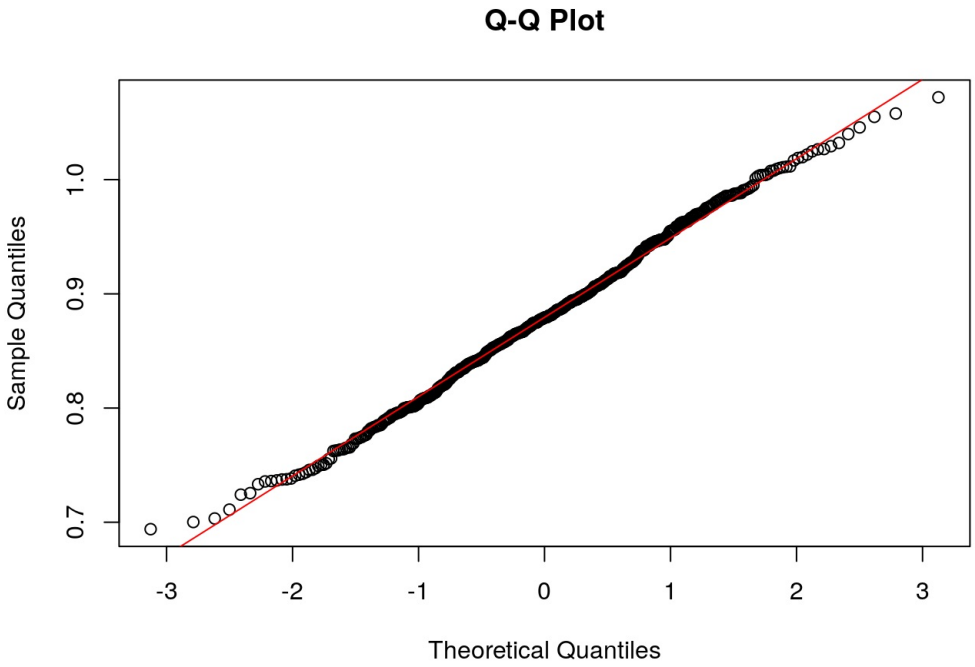
```
# Below I am transforming the reaction time data by employing a inverse power transformation, specifically, I cal
culate the inverse of reaction time to the power of 0.2. It also should be borne in mind that I removed outliers
more than 3 standard deviations away.

df$inverse_power_rt <- 1/(df$reaction_time)^0.2 # Inverse power transformation

mean_reaction_time <- mean(df$inverse_power_rt, na.rm = TRUE)
sd_reaction_time <- sd(df$inverse_power_rt, na.rm = TRUE)

ggplot(df, aes(x = df$inverse_power_rt)) +
    geom_histogram(aes(y = ..density..), bins = 30, color = "black", fill = "lightblue") +
    stat_function(fun = dnorm, args = list(mean = mean_reaction_time, sd = sd_reaction_time), color = "red") +
    theme_minimal() +
    labs(title = paste("Histogram of Inverse Power Reaction Times"),
         x = "Reaction Time",
         y = "Density")
```




```
round(pastecs::stat.desc(df$inverse_power_rt, basic = FALSE, norm = TRUE), digits = 4)
```

```
##       median         mean      SE.mean CI.mean.0.95          var      std.dev
##       0.8792       0.8794       0.0030       0.0058       0.0050       0.0705
##     coef.var     skewness     skew.2SE     kurtosis     kurt.2SE   normtest.W
##       0.0801       0.0096       0.0465      -0.3569      -0.8698       0.9967
##   normtest.p
##       0.2960
```

```
qqnorm(df$inverse_power_rt, main = paste("Q-Q Plot"))
qqline(df$inverse_power_rt, col = "red")
```




```
# The histogram shows that the distribution of reaction time with this transformation is similar to that of a hyp
othetical normal distribution. Moreover, the QQ-plot further confirms this as the data points do not significantl
y stray away from the red line which represents the data points of a normal distribution. It can also be seen tha
t the p-value is not below the significance threshold of 0.05 (p-value = 0.2960), so we fail to reject the null h
ypothesis. This suggests that the transformed reaction time data follows a normal distribution, enabling me to em
ploy parametric tests. I will therefore stick to this transformation in the following analyses.
```

```
# I will calculate the correlation between reaction time and the independent variables using Pearson as my depend
ent variable is normally distributed. However, my data has many ties (repetitions of values) in the independent v
ariables and this could affect the results of Pearson's test. Alternatively, Kendall correlation could be used to
deal with these ties, despite this test being used more frequently with non-normal data.

# Below I am going to use Pearson correlation test to determine the correlation between the dependent variable an
d the independent variable.

cor.test(df$number_of_words, df$inverse_power_rt, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$number_of_words and df$inverse_power_rt
## t = -1.6192, df = 563, p-value = 0.106
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.14973009  0.01448684
## sample estimates:
##         cor
## -0.06808275
```

```
cor.test(df$readability, df$inverse_power_rt, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$readability and df$inverse_power_rt
## t = 1.9072, df = 563, p-value = 0.057
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.002383212  0.161541371
## sample estimates:
##        cor
## 0.08012077
```

```
cor.test(df$polarity, df$inverse_power_rt, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$polarity and df$inverse_power_rt
## t = 0.95425, df = 563, p-value = 0.3404
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.0424447  0.1222672
## sample estimates:
##        cor
## 0.04018423
```

```
cor.test(df$subjectivity, df$inverse_power_rt, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$subjectivity and df$inverse_power_rt
## t = -0.61358, df = 563, p-value = 0.5397
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.10810823  0.05675874
## sample estimates:
##         cor
## -0.02585053
```

```
# Here I have tried to ascertain whether there is a relationship between the dependent variable and the independe
nt variables, values, orientation and significance with Pearson's correlation.

# The results indicate that there is not a significant correlation between reaction time and word count, readabil
ity, polarity and subjectivity, with p-values of 0.106, 0.057, 0.3404 and 0.5397, respectively.

shapiro.test(df$number_of_words)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$number_of_words
## W = 0.87518, p-value < 2.2e-16
```

```
shapiro.test(df$readability)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$readability
## W = 0.88932, p-value < 2.2e-16
```

```
shapiro.test(df$polarity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$polarity
## W = 0.83414, p-value < 2.2e-16
```

```
shapiro.test(df$subjectivity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$subjectivity
## W = 0.77397, p-value < 2.2e-16
```

```
# As can be seen, the none of the independent variables used in the correlation are normal and this violates a vi
tal assumption of Pearson's correlation.

# Due to the violation of normality of the independent variables it would be a good idea to run Kendall correlati
on tests to deal with ties that may distort the values. Furthermore, Kendall's correlation does not require the i
ndependent variables to be normal. I will do this now.
```

```
# Here I will run Kendall correlation tests, which should deal with ties very well and yield more accurate result
s. This test is usually used with non-normal data (p-value < 0.05), but it is still suitable in this situation.

cor.test(df$number_of_words, df$inverse_power_rt, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  df$number_of_words and df$inverse_power_rt
## z = -1.4912, p-value = 0.1359
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##         tau
## -0.04529217
```

```
cor.test(df$readability, df$inverse_power_rt, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  df$readability and df$inverse_power_rt
## z = 2.2918, p-value = 0.02192
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## 0.06627694
```

```
cor.test(df$polarity, df$inverse_power_rt, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  df$polarity and df$inverse_power_rt
## z = -0.32033, p-value = 0.7487
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##          tau
## -0.009897624
```

```
cor.test(df$subjectivity, df$inverse_power_rt, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  df$subjectivity and df$inverse_power_rt
## z = -0.90697, p-value = 0.3644
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.0276022
```

```
# These tests are aimed at determining the values, orientation and significance between the dependent variable an
d the independent variables. The results of this test show that there is not a significant correlation between nu
mber count, polarity and subjectivity, each one individually, and reaction time, with p-values of 0.1359, 0.7487
and 0.3644, respectively.

# The variables are continuous and the observations are independent of each other. Therefore, I will be able to r
un this correlation. The independent variables do not have to be normally distributed.

# In the specific case of readability, namely, average number of syllables per word and sentence length, it can b
e seen that there is a significant correlation with respect to reaction time because of the p-value below the sig
nificance threshold of 0.05 (p-value = 0.02192). This correlation is not very large, as can be reflected in the t
au value of 0.066, so readability has a significant but weak correlation in relation to reaction time. This relat
ionship is positive, but I have employed an inverse transformation, essentially changing the direction of the cor
relation. Therefore, undoing the inverse transformation, the real direction of the correlation is negative. As a
result, I will conclude that as readability increases, reaction time decreases.
```

```
# Under this comment, I will continue by implementing a linear regression model with inverse power reaction time as the outcome variable and word count, readability, polarity and subjectivity as predictor variables. I will do this because the dependent variable is normally distributed, meeting this vital assumption. However, the nature of my data, i.e. the presence of many ties, will distort the results, making the model unrealiable. I will explore a different type of model which deals with this further on.

linear_model_data <- lm(inverse_power_rt ~ number_of_words + readability + polarity + subjectivity, data = df)

summary(linear_model_data)
```

```
##
## Call:
## lm(formula = inverse_power_rt ~ number_of_words + readability +
##     polarity + subjectivity, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.195795 -0.045987 -0.002792  0.046593  0.188054
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.879414   0.002953 297.826   <2e-16 ***
## number_of_words -0.005427   0.003020  -1.797   0.0728 .
## readability      0.008257   0.003747   2.203   0.0280 *
## polarity         0.002382   0.002994   0.796   0.4265
## subjectivity     0.003041   0.003727   0.816   0.4148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07019 on 560 degrees of freedom
## Multiple R-squared:  0.01487,    Adjusted R-squared:  0.007838
## F-statistic: 2.114 on 4 and 560 DF,  p-value: 0.07774
```

```
# As can be seen in the results, the only predictor variable out of the four that has a significant effect on the outcome variable is readability. The overall model explains an infinitesimal portion of the total variance, as can be seen in the adjusted R-squared value of 0.007838. Furthermore, the overall p-value is marginally significant (below 0.1), but not significant at the 0.05 level and the F-statistic has a value of 2.114, meaning that the model with the four independent variables fails to explain reaction time significantly better than the intercept alone would do. However, of course, the results are not of great value due to the fact that each participant registered multiple responses. Therefore, a mixed-effects model seems suitable in this case to account for individual differences between participants (participants as a source of variance).

# This model has certain assumptions that have to be met. However, this model does not account for participants as a source of variance as they registered multiple responses and it violates the assumption of independence of observations. Therefore, this model is not valid.
```

```
# I will now employ a mixed-effects model to deal with individual participants as a source of variance. Individual participants will be treated as a random factor. Mixed-effect models do not strictly require the dependent variable to be normally distributed, although in my case is it (p-value = 0.2960).

# Here I am going to create a linear mixed-effects model with the transformed reaction time (inverse_power_rt) as the outcome variable and the other four variables (number_of_words, readability, polarity and subjectivity) as predictor variables, and individual participants as a source of variance, an error from the previous model that will be corrected now.

model <- lmer(inverse_power_rt ~ number_of_words + readability + polarity + subjectivity + (1|ID), data = df)

# I will view the summary of the model.

summary(model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: inverse_power_rt ~ number_of_words + readability + polarity +
##     subjectivity + (1 | ID)
##    Data: df
##
## REML criterion at convergence: -1462.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.05387 -0.63841  0.07128  0.64238  2.94304
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 0.001249 0.03535
##  Residual             0.003717 0.06096
## Number of obs: 565, groups:  ID, 24
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)      0.878966   0.007658 114.783
## number_of_words -0.005588   0.002624  -2.130
## readability      0.008436   0.003256   2.591
## polarity         0.002371   0.002601   0.912
## subjectivity     0.003369   0.003239   1.040
##
## Correlation of Fixed Effects:
##             (Intr) nmbr__ rdblty polrty
## nmbr_f_wrds  0.000
## readability  0.000 -0.147
## polarity     0.000  0.142  0.009
## subjectivty  0.000 -0.103  0.607 -0.049
```

```
# Below I will transform the t-values into p-values.

t_values <- c(114.783, -2.130, 2.591, 0.912, 1.040)
dof <- 560
p_values <- 2 * pt(abs(t_values), df = dof, lower.tail = FALSE)
p_values
```

```
## [1] 0.000000000 0.033606928 0.009819175 0.362161145 0.298788705
```

```
# The residuals have to be normally distributed because this is an assumption for mixed-effects models, therefore
the p-value of the residuals has to be above the 0.05 threshold.

residuals_model <- resid(model)

# I will run a Shapiro-Wilk test to obtain the p-value.

shapiro.test(residuals_model)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals_model
## W = 0.99748, p-value = 0.5534
```

```
# Below is a QQ-plot of the residuals to see if they align with those of a theoretical normal distribution visual
ly.

qqnorm(residuals_model)
qqline(residuals_model, col = "red")
```