Personality Data Analysis: Portfolio Project 1

Esben Kran Christensen

9/21/2019

Contents

Packages: pacman, bookdown, tidyverse, rticles, RColorBrewer, extrafont, pastec, gridExtra

1 Abstract

This paper investigates the properties of 2019's Cognitive Science Bachelor students at Aarhus University. The students answered and performed 40 questions and actions to accommodate a personality analysis sheet developed by Mikkel Wallentin. In this paper, the data was analyzed using several different data visualization techniques as well as statistical theory.

2 Introduction

The process of analysis performed consists of five different steps: 1. Ocular dominance and lung size. 2. Noise level preferences between females and males. 3. Frequency analysis of held breath times. 4. Frequency analysis of balloon related variables. 5. Correlation between ability to hold breath and shoe size.

3 Methods

The RStudio IDE and the R programming language is used to perform the analyses of the personality data. Using the aggregated personality data from the 1st semester students, the analyses can be visualized to answer or shed light on the analysis targets outlined above.

4 Analysis

We will start by importing the data into a dataframe in R:

```
df <- read.csv("NEW_CogSciPersonalityTest2019.csv")
```

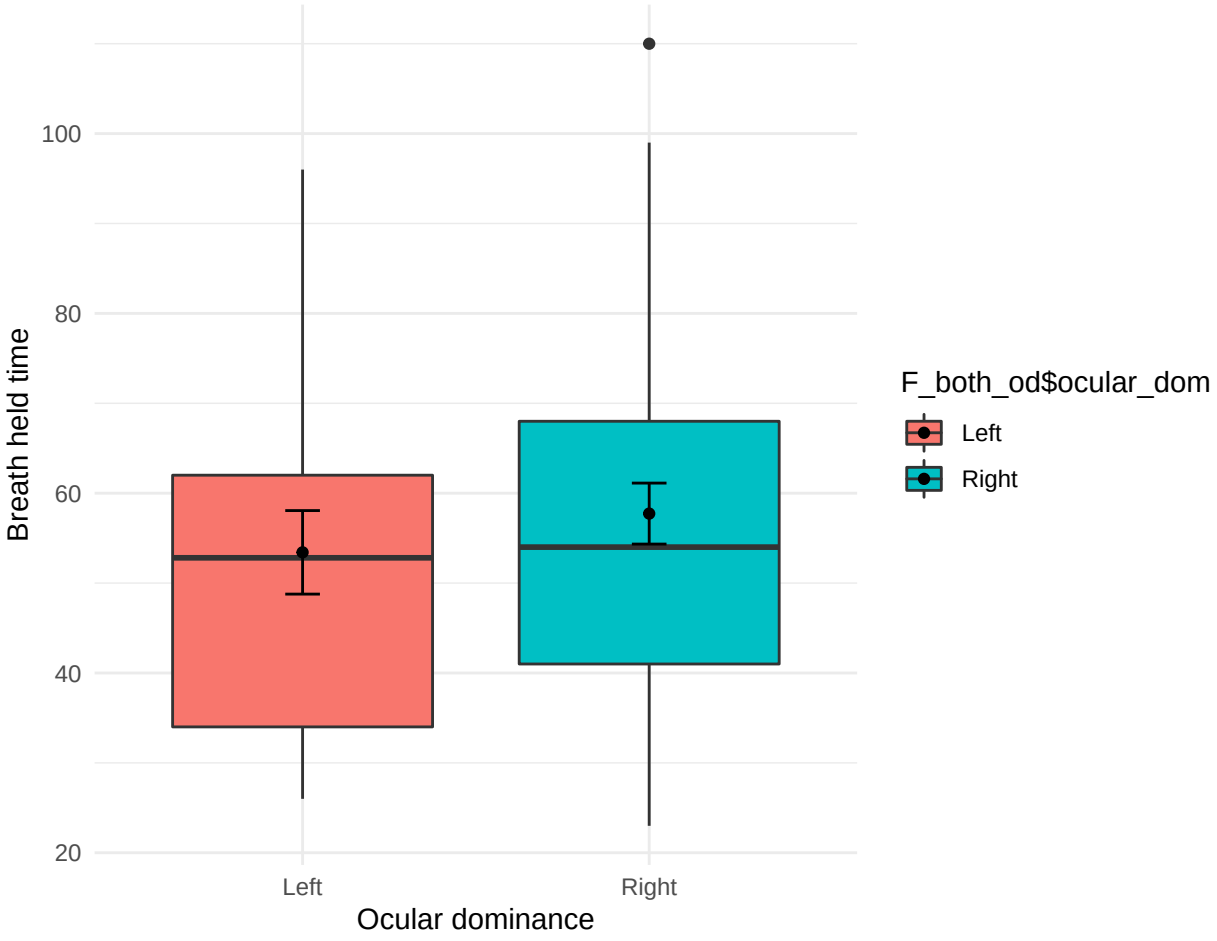


Figure 1: Plotting breath_hold by ocular_dom

4.1 Ocular dominance and lung size

Who can hold their breath longer on average, those with right or left ocular dominance? By using the ggplot2 package in RStudio and performing a summary analysis with the dplyr::summarise function, the differences in ability to hold their breath between 1st semester students with right ocular dominance and left ocular dominance can be analyzed.

We'll start by removing the rows where the ocular dominance is equals to "Both" as it is irrelevant to analyzing differences between left and right ocular dominance and the probability of not having any ocular dominance ("Both") is below 1% in healthy adults (see Eser et al. 2008).

```
F_both_od <- df[df$ocular_dom != "Both", ]
```

Secondly, we'll plot the mean breath_hold according to the ocular dominance to see how they match up visually:

```
ggplot(F_both_od, aes(x = F_both_od$ocular_dom, y = F_both_od$breath_hold, fill = F_both_od$ocular_dom)
    geom_boxplot() + geom_point(stat = "summary", fun.y = mean, ) + geom_errorbar(show.legend = F,
    stat = "summary", fun.data = mean_se, width = 0.1) + labs(x = "Ocular dominance",
    y = "Breath held time") + theme_minimal()
```

The dot with the black error bars is the mean with mean standard error error bars appended.

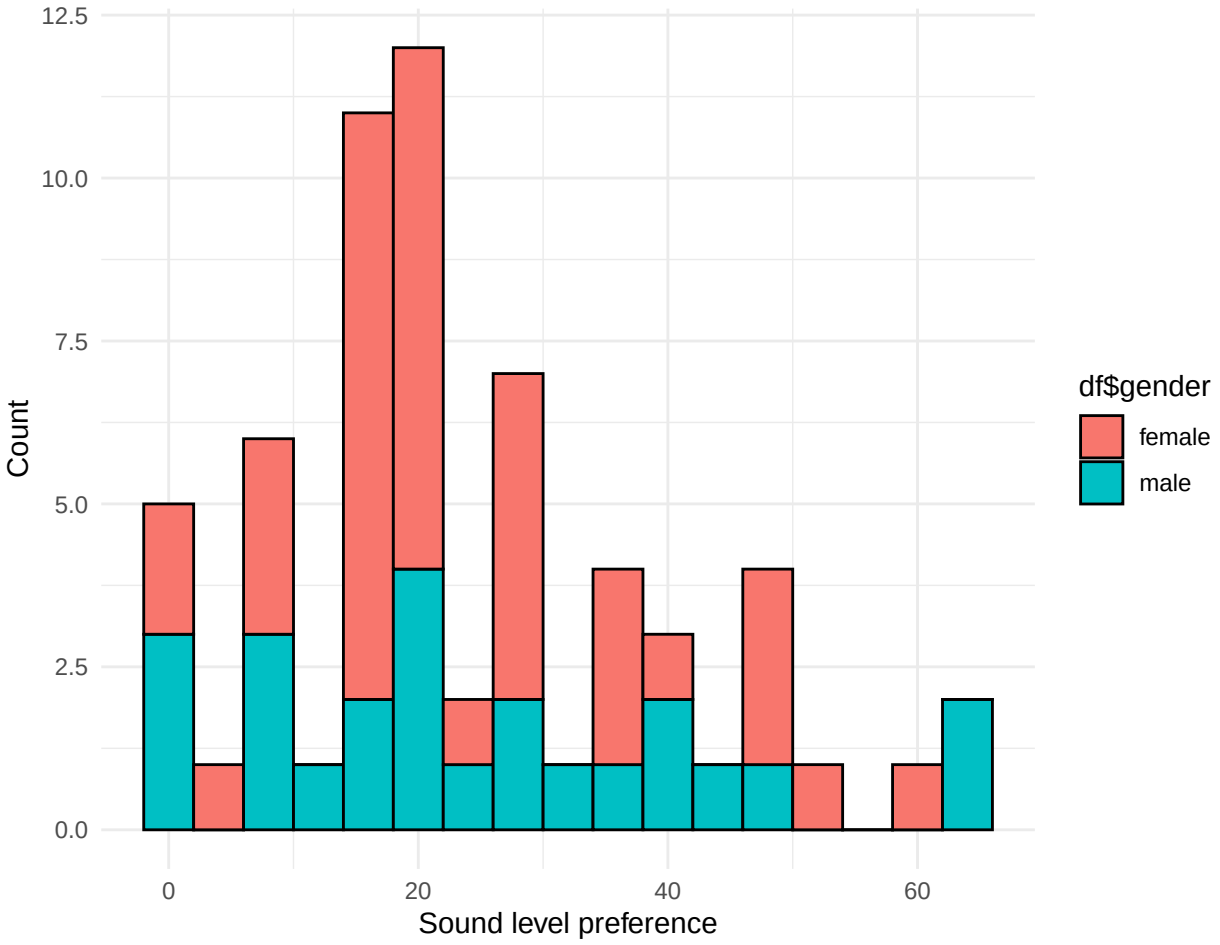


Figure 2: Histogram of sound level preference

By looking at the boxplot and seeing that the SEM error bars overlap, we can conclude from a simplification of Cumming's (2007) rules on statistical significance that the P-value is *much* greater than 0.05 which means the difference is not statistically significant.

We can therefore conclude that there is no large difference between groups of differing ocular dominance and the ability to hold their breath.

4.2 Noise level preferences between females and males

Who likes silence versus noise the best? Males or females? By using the same process as in the previous step, we can perform an analysis of the gender dissonance in noise preferences. But first, we will plot the data to find out if there are any outliers

```
ggplot(df) + geom_histogram(binwidth = 4, aes(x = df$sound_level_pref, y = ..count..,
    fill = df$gender), colour = "black") + theme_minimal() + labs(x = "Sound level preference",
    y = "Count")
```

As there are no significant outliers, we will commence the analysis with the dplyr::summarise_each function:

```
df %>% group_by(gender) %>% summarise_each(funs(mean, sd), noise = sound_level_pref,
    noise = sound_level_pref)
```

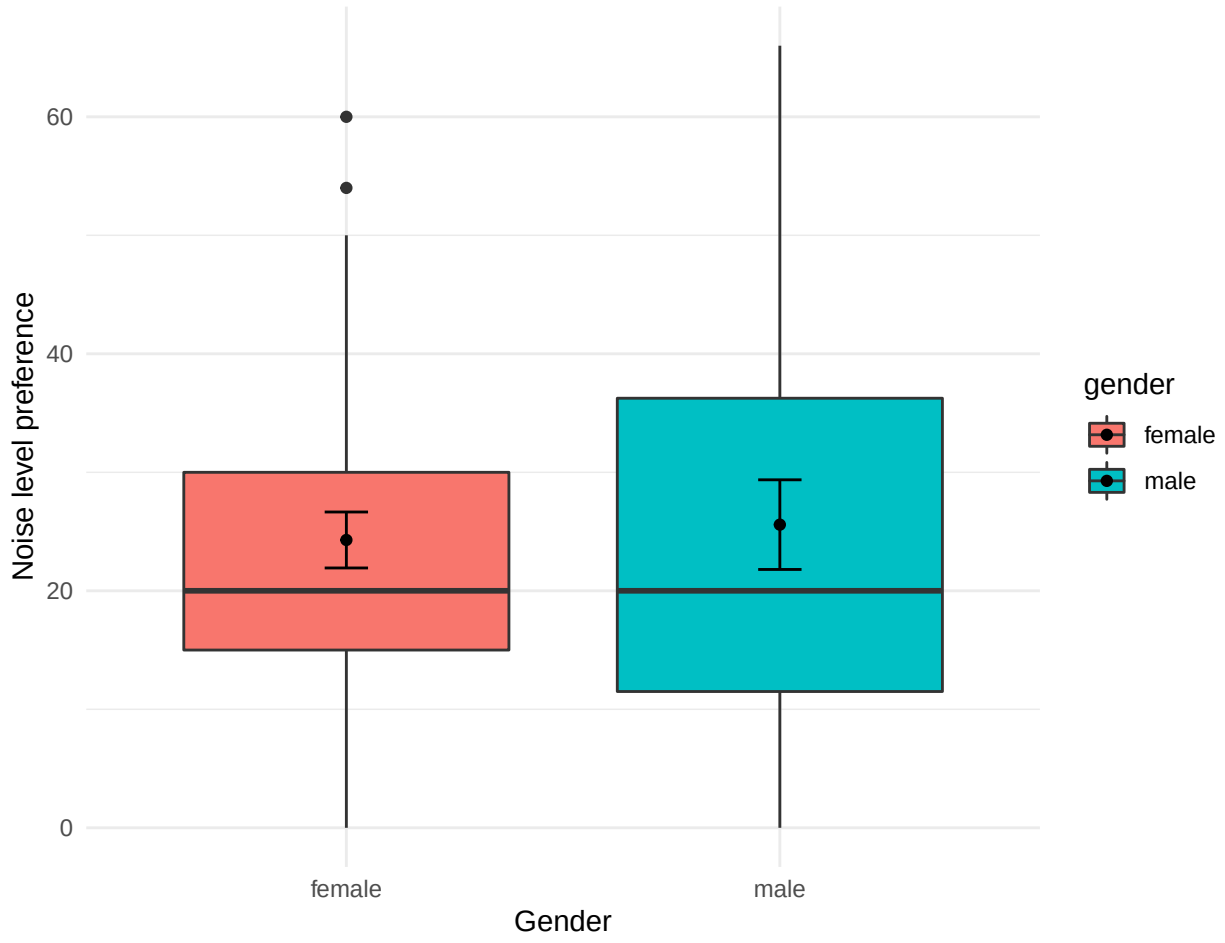


Figure 3: Noise level preference by gender

```
> # A tibble: 2 x 3
>   gender noise_mean noise_sd
>   <fct>       <dbl>    <dbl>
> 1 female       24.3     14.6
> 2 male         25.6     18.5
```

Through a purely numeric argumentation, we can see that the gender difference is completely negligible through mean and standard deviance differences between males and females. By trying to plot the gender differences in a bar plot with an error bars layer in ggplot2, we can see that no real differences are visible except larger center aggregation of values in female data (which could be seen as quite interesting from cultural bias perspective).

```
ggplot(df, aes(x = gender, y = sound_level_pref, fill = gender)) + geom_boxplot() +
    geom_errorbar(stat = "summary", fun.data = mean_se, width = 0.1) + geom_point(stat = "summary",
    fun.y = mean, ) + labs(x = "Gender", y = "Noise level preference") + theme_minimal()
```

The visual explanation is the same as in the previous analysis process step: As the SEM error bars overlap, we can conclude from a simplification of Cumming's (2007) rules on statistical significance that the P-value is *much* greather than 0.05 which means the difference is not statistically significant.

4.3 Frequency analysis of held breath times

Is the breath_hold data normally distributed? With visual and numeric support, an analysis to determine this is executed using ggplot2 plotting powers. We will define this plotting as a function to ease comprehension and for later usage.

```
e_freq_distribution <- function(dataframe, x, title) {
    ggplot(dataframe, aes(x = x)) + geom_histogram(binwidth = 2, colour = "black",
        aes(y = ..density.., fill = ..count..)) + scale_fill_gradient("Count") +
        stat_function(fun = dnorm, color = "orangered2", size = 1, args = list(mean = mean(x),
            sd = sd(x))) + labs(title = title, subtitle = "Using dnorm() to create normal distribution
        x = "Time", y = "Density") + theme_minimal()
}
```

And using the custom function with the variable inputs as well as a qq-plot:

```
ggplot1 <- e_freq_distribution(df, df$breath_hold, "Held breath time frequency density")
ggplot2 <- ggplot(df, aes(sample = df$breath_hold)) + stat_qq() + stat_qq_line() +
    labs(y = "Breath hold", title = "QQ-plot of breath hold times") + theme_minimal()
ggplot3 <- ggplot(df, aes(x = df$breath_hold, y = ..density.., fill = gender)) +
    geom_density(alpha = 0.7) + theme_minimal() + labs(x = "Breath hold")

grid.arrange(ggplot1, ggplot2, ggplot3, nrow = 2, ncol = 2)
```

With the quantile-quantile plot (QQ-plot), we can see how much the breath_hold data (y axis) resembles a theoretically perfect normal distribution (x axis) quantile for quantile. The linear function indicates the perfect normal distribution. We can see that the breath_hold data correlates roughly to this linear function which indicates that the data *might* follow a normal distribution though it is doubtful as it still deviates quite a lot.

Therefore, to be completely sure, we can test it using the Shapiro-Wilk test that returns the p-value for the probability that the data rejects the null-hypothesis that it follows a normal distribution:

With the Shapiro-Wilk test output of 0.035 (normtest.p = 3.5%), we can see that we will have to reject the null-hypothesis that the breath_hold data follows a normal distribution. But let's keep this thought and hypothesize that the reason for the failed Shapiro-Wilk normality test is because the data is binomial because of the difference between female and male breath_hold times (see figure 3.4):

```
stat.desc(df[df$gender == "male", c("breath_hold")], df$breath_hold, basic = F,
    norm = T)
```

```
>   skewness   skew.2SE   kurtosis   kurt.2SE normtest.W normtest.p
> -0.2381634 -0.2521524 -1.2139052 -0.6613290  0.9450407  0.2110520
```

```
stat.desc(df[df$gender == "female", c("breath_hold")], df$breath_hold, basic = F,
    norm = T)[]
```

```
>    skewness    skew.2SE    kurtosis    kurt.2SE  normtest.W  normtest.p
> 1.211480885 1.582317996 2.563814041 1.709892516 0.912328039 0.005788547
```

Using this data, we might be able to conclude that the male data for breath_hold follows a normal distribution (`normtest.p` > 0.05) while the female data definitely does not (`normtest.p` < 0.05). A probable reason for these differences and results might be the sample size. But as we are analyzing this specific sample's distribution, the conclusions are correct in relation to this exact population.

4.4 Frequency analysis of balloon related variables

Is the balloon and balloon_balance data normally distributed? With visual and numeric support, an analysis to determine this is performed using ggplot2 plotting capabilities and statistics analysis.

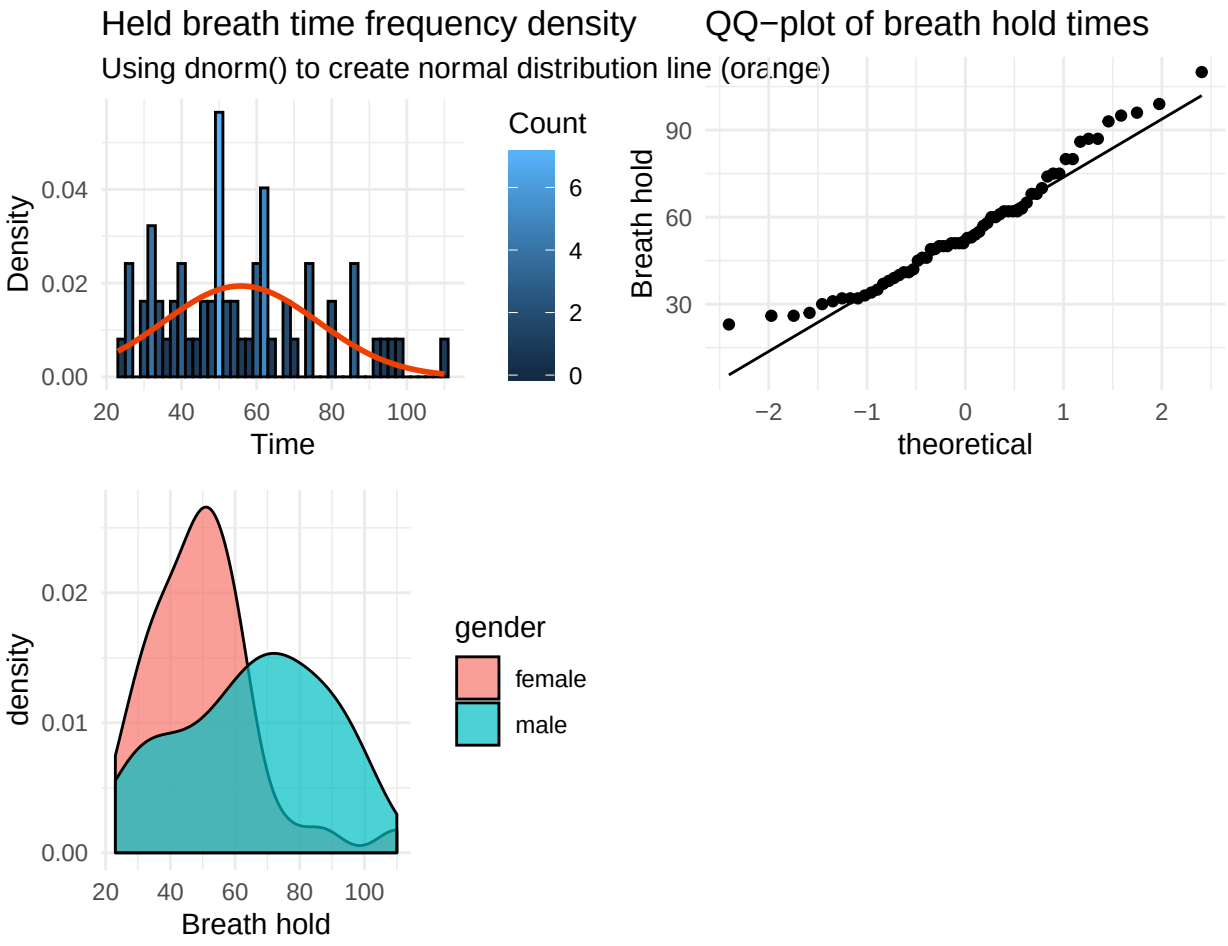
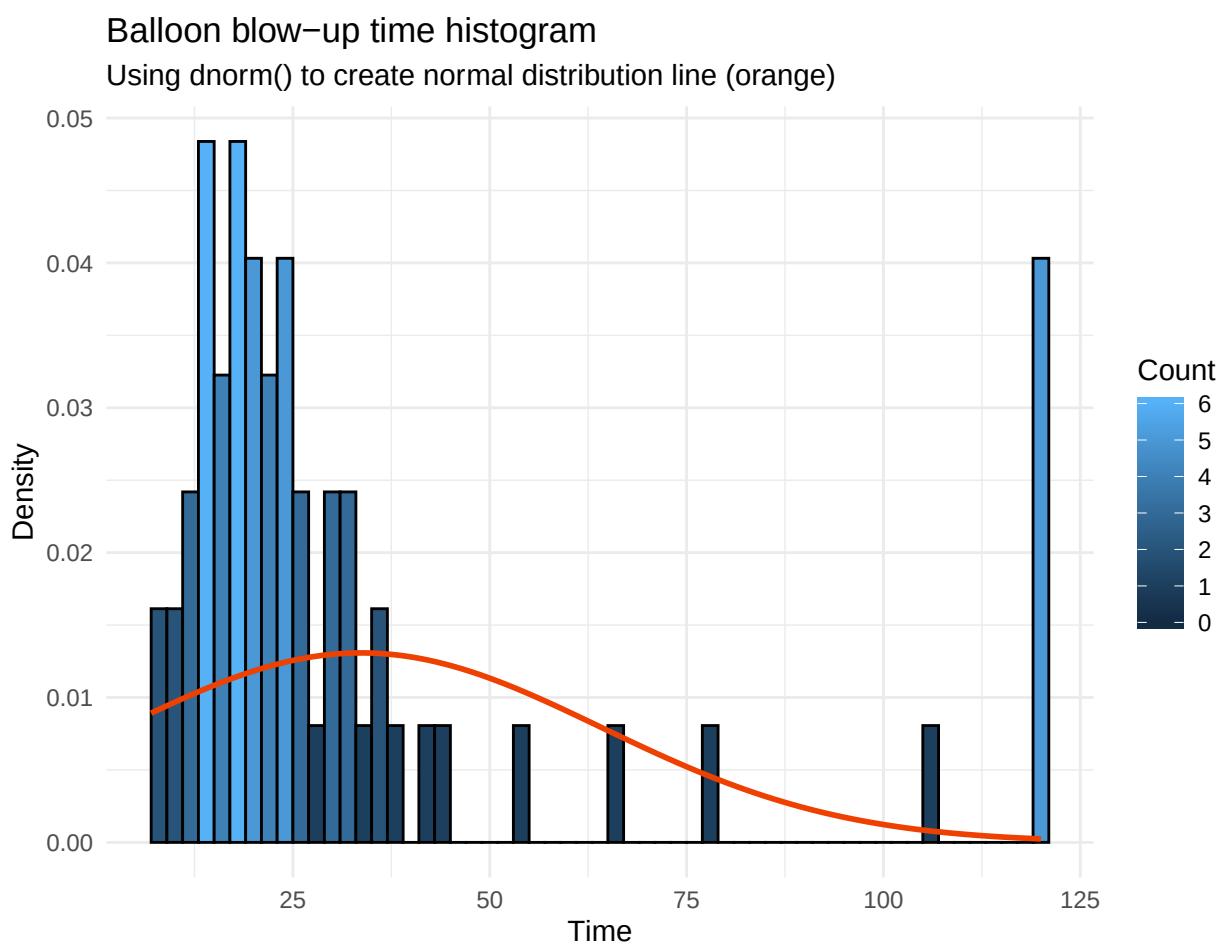


Figure 4: Plotting a frequency distribution, density plot and QQ-plot beside each other

```
e_freq_distribution(df, df$balloon, "Balloon blow-up time histogram")
```



We can see that the normal distribution's kurtosis and skew is drastically affected by the outlier in the higher end of the scale. We will try to identify numerically that this outlier has a large effect on the data:

```
getmode <- function(v) {
    uniqv <- unique(v)
    return(uniqv[which.max(tabulate(match(v, uniqv)))])
}

mode <- getmode(df$balloon)
names(mode) <- c("mode")
print(mode)
```

```
> mode
>   15
```

```
round(stat.desc(df$balloon, basic = F, norm = T), digits = 3)
```

```
>       median         mean      SE.mean CI.mean.0.95          var
>       22.500       33.677        3.874        7.746      930.403
>      std.dev     coef.var     skewness     skew.2SE     kurtosis
>       30.503        0.906        2.010        3.306        2.852
>     kurt.2SE   normtest.W   normtest.p
>        2.379        0.660        0.000
```

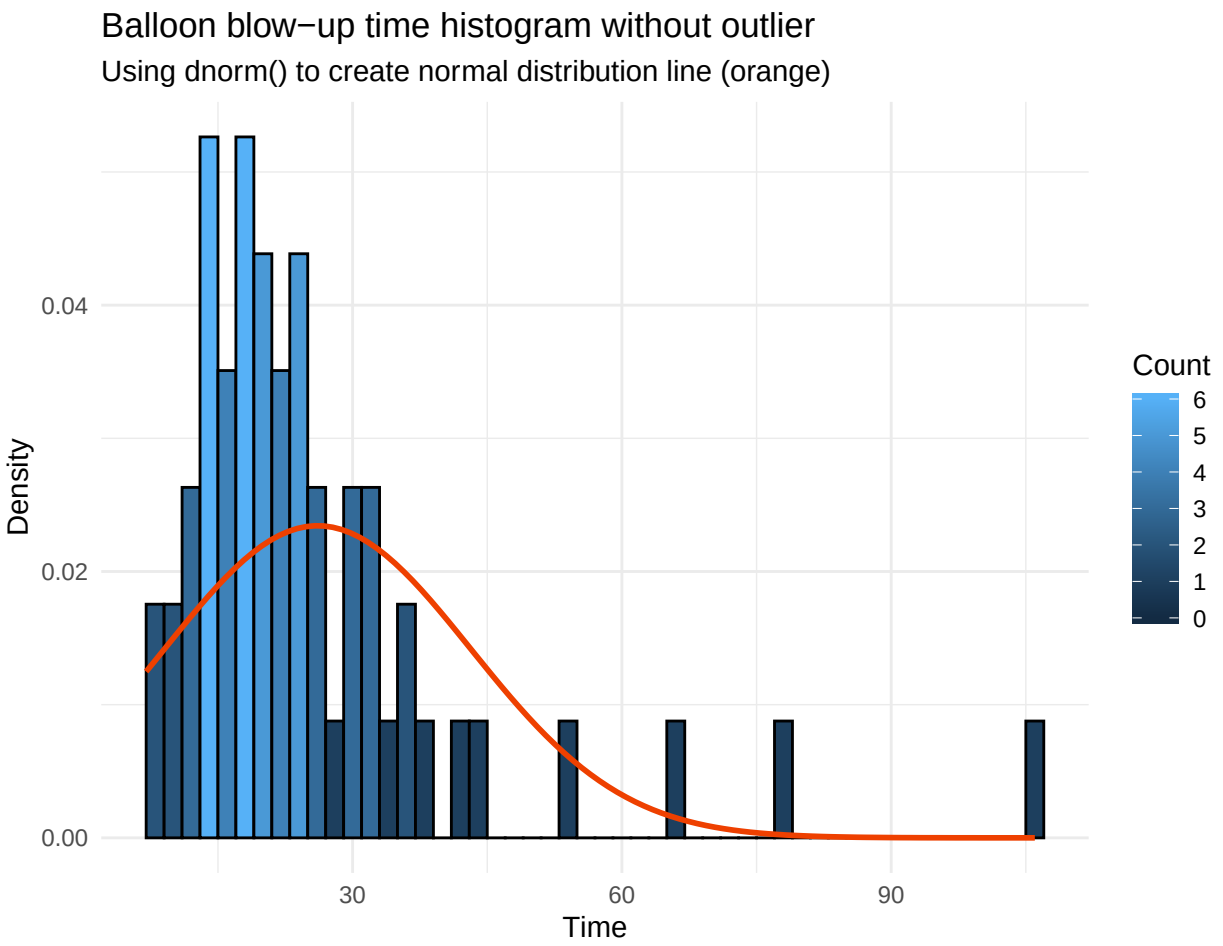


Figure 5: Removing outliers (-2.58 < z < 2.58) and re-plotting the histogram

Using the summarise function, we can see that the mean, mode and median are significantly different, signifying that there's an outlier that has a strong influence on especially the mean (33.7 s) compared to the median and the mode (22.5 s and 15 s respectively). Additionally, the stat.desc shows us that the normal distribution has a major skew.2SE (2 standard error) of 3.306 (is 1 for standard normal distribution) which shows us that the data includes values that skews the data (outlier).

We will try to exclude the outlier by excluding any values deviating by 2.58 standard deviations from the mean and re-plotting the histogram. By using the value of 2.58 we practically exclude results that are below a 1% chance of existing in normally distributed populations - a z-score under 2.58 = 99.012% probability of being under the distribution curve.

```
sd_b <- sd(df$balloon)
b_258sd_df <- df[df$balloon < mean(df$balloon) + 2.58 * sd_b & df$balloon >
    mean(df$balloon) - 2.58 * sd_b, ]
e_freq_distribution(b_258sd_df, b_258sd_df$balloon, "Balloon blow-up time histogram without outlier")
```

Here we can see how the distribution fits a normal distribution better visually but would still have a skew if plotted without dnorm() function. We can do a numeric summary using the stat.desc once more to see if we have more normally distributed data after removing the outlier:

```
round(stat.desc(b_258sd_df$balloon, basic = F, norm = T), digits = 3)
```

```
>       median         mean      SE.mean CI.mean.0.95          var
>       22.000       26.105        2.255        4.517      289.793
>      std.dev     coef.var     skewness     skew.2SE     kurtosis
>       17.023        0.652        2.536        4.009        7.919
>     kurt.2SE   normtest.W   normtest.p
>        6.354        0.740        0.000
```

The skew has increased unexpectedly. The reason for this is probably the fact that the previous skew was caused by the data having a lower boundary because the data is time-based and starts from 0. As there are no possible values below zero, the skew is therefore expected and we cannot say that the data is normally distributed (normtest.p < 0.05) even though it might look that way on the graph. Additionally, it is a positive skew so this could have been analyzed from the first skewness calculation.

4.5 Correlation between ability to hold breath and shoe size

Shoe size could tell us something about general body size, which could also be connected to one's ability to hold one's breath. In other words, we predict that there is a positive relation between shoe size and how long time CogSci students can hold their breath. By plotting the two sets of data against each other on a scatter plot and analyzing differences between males and females, we can see how shoe size affects the ability to hold one's breath.

```
p_load(gridExtra)

plot1 <- ggplot(df, aes(x = df$shoesize, y = df$breath_hold)) + geom_point(show.legend = F,
    aes(color = df$gender)) + geom_smooth(show.legend = F, formula = y ~ x,
    method = "lm", color = "#2c3e50") + labs(title = "1: Correlation between shoe size\nand breath hold
    subtitle = "Blue = male, red = female", x = "Shoe Size", y = "Breath hold time") +
    theme_grey()

plot2 <- ggplot(df, aes(x = df$shoesize, y = df$breath_hold, color = df$gender)) +
    geom_point(show.legend = F) + geom_smooth(show.legend = F, formula = y ~
    x, method = "lm") + labs(title = "2: Gender difference", subtitle = "Blue = male, red = female",
    x = "Shoe Size", y = "Breath hold time") + theme_grey()

grid.arrange(plot1, plot2, ncol = 2)
```

The graphs show a correlation between shoe size and ability to hold breath but we can't talk about a difinite correlation between the two variables. On the differences between genders, females generally have a lower ability to hold their breath compared to males when looking at the graph's scatterplot and if we calculate the means: $mean_f=$49.6528947 and $mean_m=$65.4583333, we can see that the likely reason for the linear regression's slope in visualization 1 is this difference in mean and in shoe size between genders.

The reason we can see this is that visualization 2 shows us that there is no apparent correlation between shoe size and breath_hold in the male regression. Additionally the error visualization in the graphing of the linear regression shows us that the variance is too large to conclude direct correlation.

But beyond just looking at the graphs, we can analyze the correlation between the two variables and see if it's statistically significant. By setting up a correlation matrix, we can see if there's any good correlation between the values. balloon, sound_level_pref, hours_music_per_week and tongue_twist have also been added because it would be interesting to see if there's a correlation between any of these in addition to the other two.

This is done by melting a data frame with the variable correlations into a three-variable data frame and inputting it in the geom_tile() function that creates color tiles based on correlation (0:1).

```
p_load(reshape2)
cormat <- df[, c("hours_music_per_week", "sound_level_pref", "shoesize", "breath_hold",
```

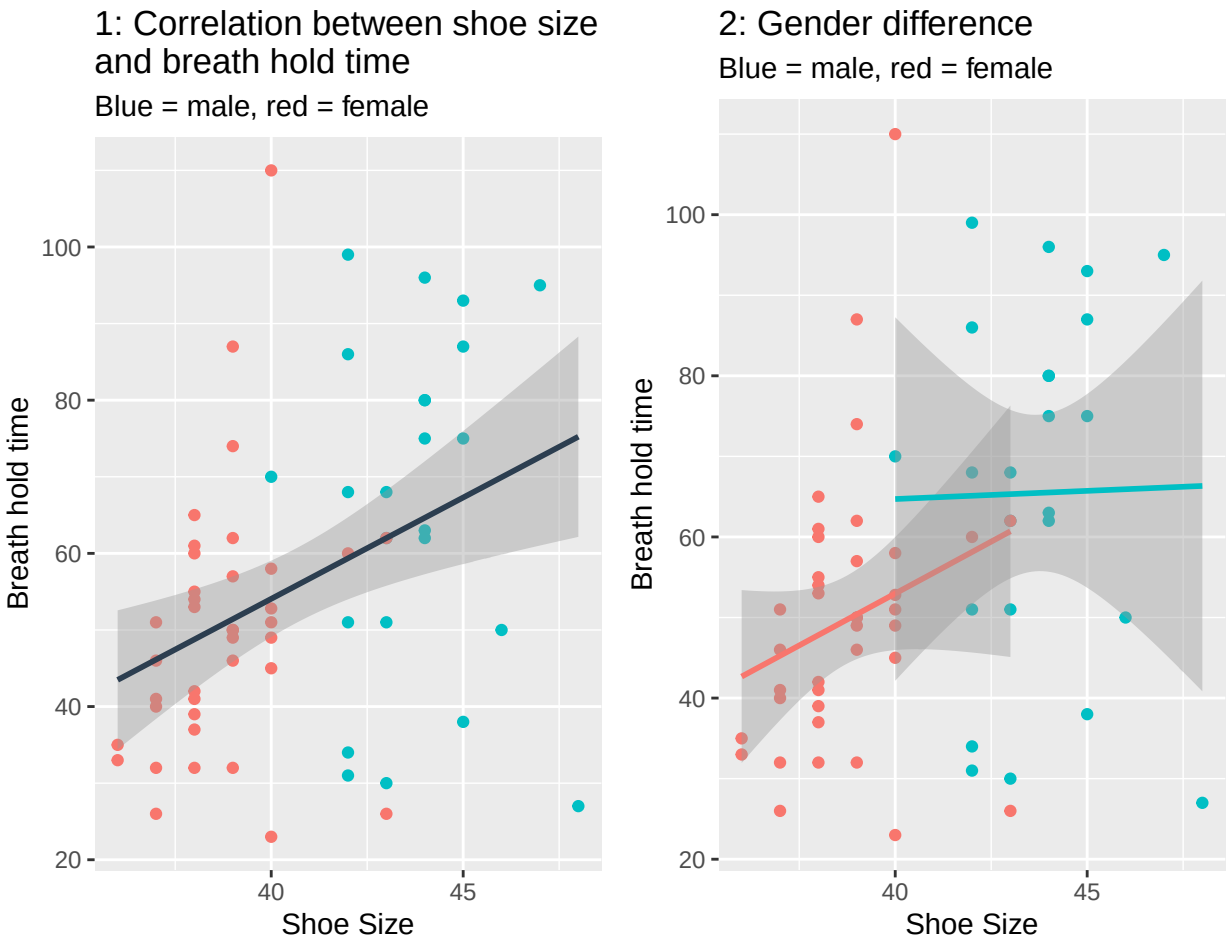
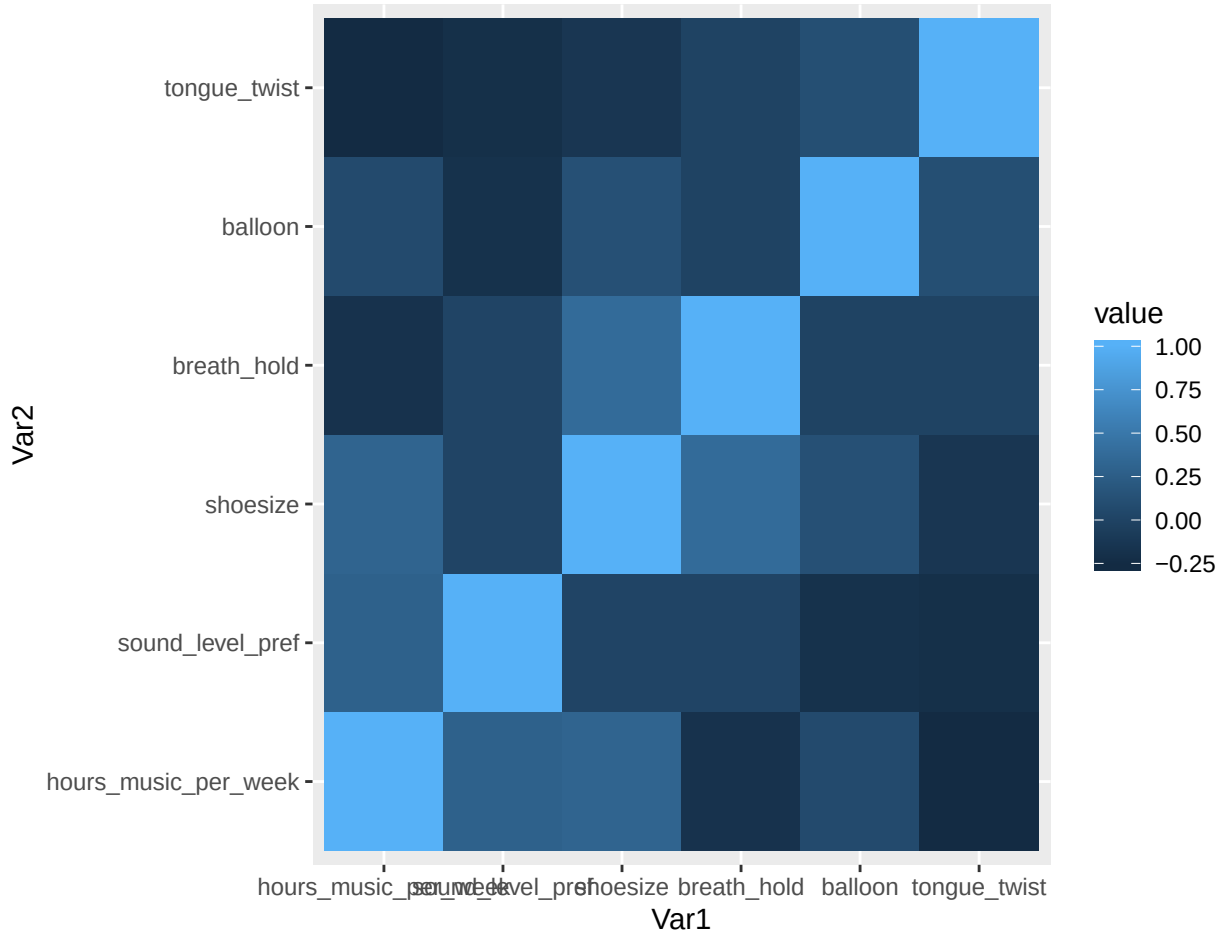


Figure 6: Correlation between shoe size and breath hold time with regression lines

```
    "balloon", "tongue_twist")]
cormat <- round(cor(cormat), 2)
melted_cormat <- melt(cormat)

ggplot(melted_cormat, aes(Var1, Var2, fill = value)) + geom_tile()
```



Interestingly, we see a small correlation between hours_music_per_week and sound_level_pref as well as shoesize and hours_music_per_week. These are definitely not statistically significant ($\rho = cor(hours, sound\_level) = 0.2848$) but a hypothesis could have been that the hours of music correlates to preference for louder environments though this has just been disproven in our sample.

Besides being able to see the correlations of these variables on the geom_tile, to analyze the correlation between shoesize and breath_hold, we will calculate this correlation using the cor() function:

$\rho = cor(shoesize, breath\_hold) = 0.3826$

As it equals 0.38, there is no significant correlation ($\rho > 0.95$). We can thereby conclude that there's no significant relation between shoesize and breath_hold but that the gender influences the statistical probability of having a good skill in holding one's breath.

5 Appendix

1: Improved readability prototype

```
> [1] "nice"
```

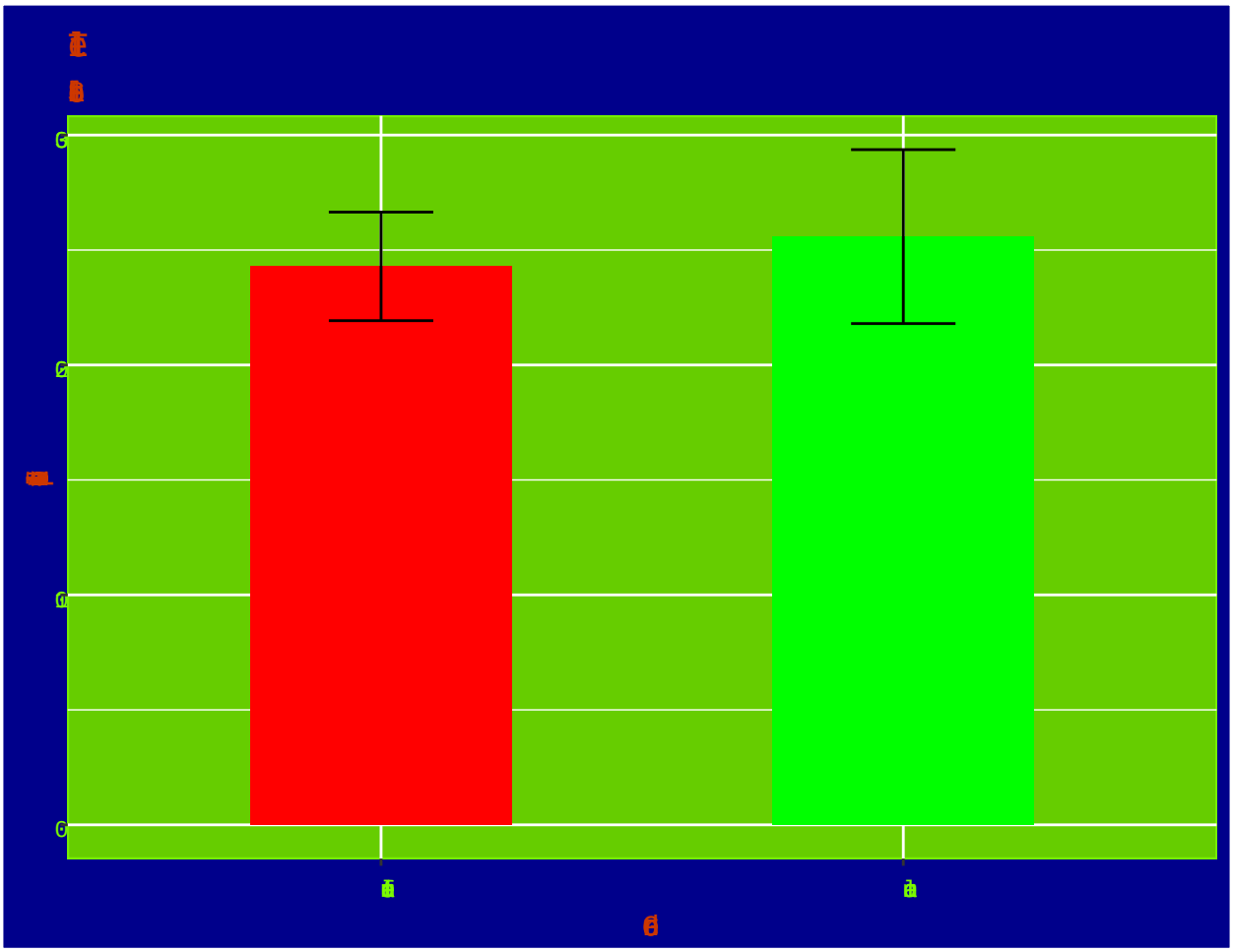


Figure 7: Improved readability prototype

References

Eser, Ilker, Frank Schwendeman, Daniel S. Durrie, and Jason E. Stahl. 2008. "Association Between Ocular Dominance and Refraction." *Journal of Refractive Surgery* 24 (7): 685–89. https://doi.org/10.3928/1081597X-20080901-07.