Portfolio exam - Part 1 | Methods 1 E2021, CogSci @AU

Laurits Lyngbæk

21/9/2021

Deadline: Wednesday 29/9/2021 h23:59

This is an individual portfolio assignment

Upload your Portfolio 1 assignment to the dedicated link ion Brightspace, under "Assignments". **Remember to upload the HTML knit of the markdown, and not the markdown (.Rmd) itself. No PDF knits, please.**

Please write your name in the `author` field above.

Introduction

The goal of this exam is to write a short data mining report on the CogSci Intro Week Personality Test Data in which you answer the following questions **in prose, code and graphs**.

First of all, let's start by looking at the setup chunk. If you need to load packages or set your working directory, do so here:

```
pacman::p_load(tidyverse)
pacman::p_load(pastecs)
```

Now you have to import the personality data. Once you have done so, use the `head()` function to print the first 10 lines of the data set.

```
df <- read_csv("personality_data_cleaned_2021.csv")
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 48 Columns: 51
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (37): timestamp, student_number, name, gender, native_Danish, handednes...
## dbl  (13): ...1, shoesize, choose_rand_num, 2D4D, balloon, balloon_balance, ...
## date  (1): birth_day
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df %>%
head()
```

```
## # A tibble: 6 x 51
##    ...1 timestamp  student_number name  birth_day  shoesize gender native_Danish
##   <dbl> <chr>      <chr>          <chr> <date>        <dbl> <chr>  <chr>
## 1     1 2021/08/2~ 202105598      Corn~ 1999-11-12       37 female Yes
## 2     2 2021/08/2~ 202104529      Tilde 2000-09-02       37 female Yes
## 3     3 2021/08/2~ 202108998      Rebe~ 2001-06-26       38 female Yes
## 4     4 2021/08/2~ 202109723      Sara~ 2000-04-26       37 female Yes
## 5     5 2021/08/2~ 202104528      Maja~ 2000-09-02       37 female Yes
## 6     6 2021/08/2~ 202106904      Vlada 2002-01-25       36 female No
## # ... with 43 more variables: handedness <chr>, choose_rand_num <dbl>,
## #   touch_floor <chr>, touch_hands <chr>, 2D4D <dbl>, balloon <dbl>,
## #   balloon_balance <dbl>, breathhold <dbl>, bad_choices <chr>,
## #   tongue_twist <dbl>, romberg_open <dbl>, romberg_closed <dbl>,
## #   ling_animal <chr>, ling_direct <chr>, ling_demonstr <chr>,
## #   ling_place <chr>, ling_abstract <chr>, ling_pronoun <chr>, ling_math <chr>,
## #   ling_activity <chr>, ling_adjective <chr>, ling_kiki <chr>, ...
```

Once you are done loading the data, you can start working on the questions below.

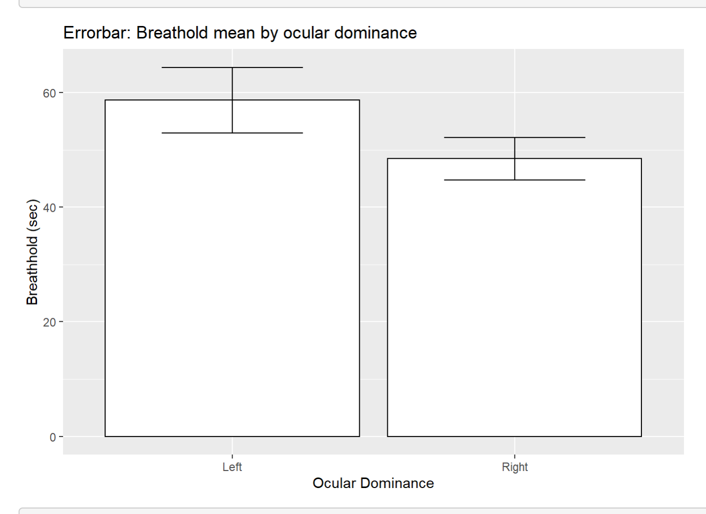
Question 1

Who can hold their breath the longest on average — those with right or left ocular dominance? Plot the data using `ggplot2` to find out. The plots should include error bars depicting the standard error of the mean: you can add these using the `geom_errorbar()` function and specifying `stat = "summary", fun.data = "mean_se"`. Then use the `mean()` and `sd()` functions *within a tidyverse pipe* to make a summary data set, in which you show mean and standard deviation of the two eye dominance groups.

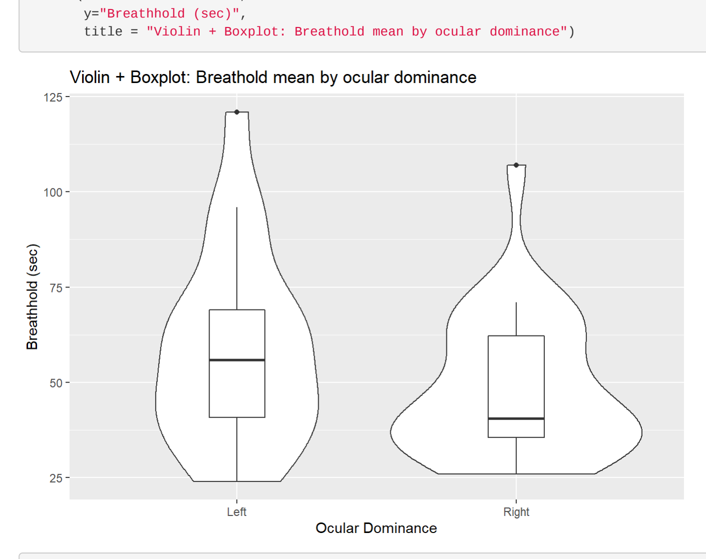
If there are people that answered other things than "Right" or "Left", then filter them out.

Bonus question: If you feel brave, you can instead try making a boxplot (`geom_boxplot()`) or a violin plot (`geom_violin()`) which are better at representing the actual distribution of the data (compared to a bar plot, which only depicts mean and standard deviation).

```
filter_df1<- df %>%
  filter(ocular_dom == "Right"| ocular_dom == "Left")

summary_df1<- filter_df1 %>%
  group_by(ocular_dom) %>%
  summarise(mean = mean(breathhold),
            sd = sd(breathhold),
            se = sd/sqrt(n()),
            max=mean+se,
            min=mean-se
            )


summary_df1 %>%
  ggplot(aes(x=ocular_dom,y=mean,ymin=min,ymax=max))+
  geom_col(fill="white",col = 'black')+
  geom_errorbar(width=0.5)+
  labs(x="Ocular Dominance",
       y="Breathhold (sec)",
       title = "Errorbar: Breathold mean by ocular dominance")
```




```
filter_df1 %>%
  ggplot(aes(x=ocular_dom,y=breathhold))+
  geom_violin()+
  geom_boxplot(width=0.2)+
  labs(x="Ocular Dominance",
       y="Breathhold (sec)",
       title = "Violin + Boxplot: Breathold mean by ocular dominance")
```




```
summary_df1
```

```
## # A tibble: 2 x 6
##   ocular_dom  mean    sd    se   max   min
##   <chr>      <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Left        58.7  25.6  5.72  64.4  53.0
## 2 Right       48.5  18.9  3.71  52.2  44.8
```

Explain your results in plain terms here (max 3 sentences):

The errorbars are narrow, which indicates that the datas mean closely approximates the true populations mean.

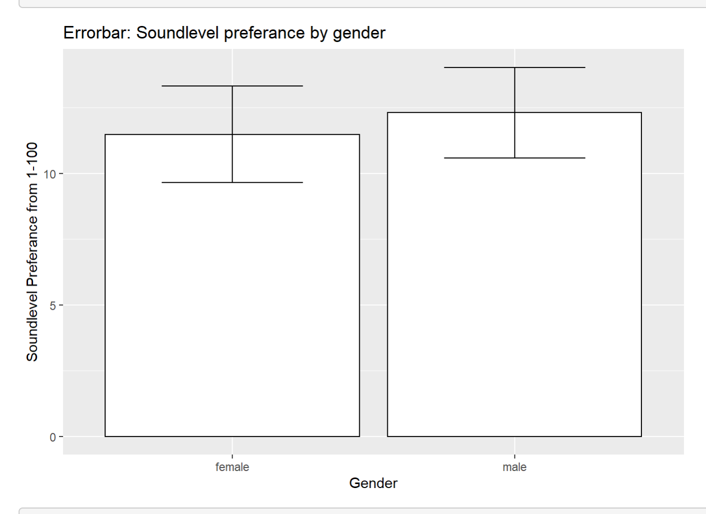
The box+violinplot indicates that the data has a few outliers.

Question 2

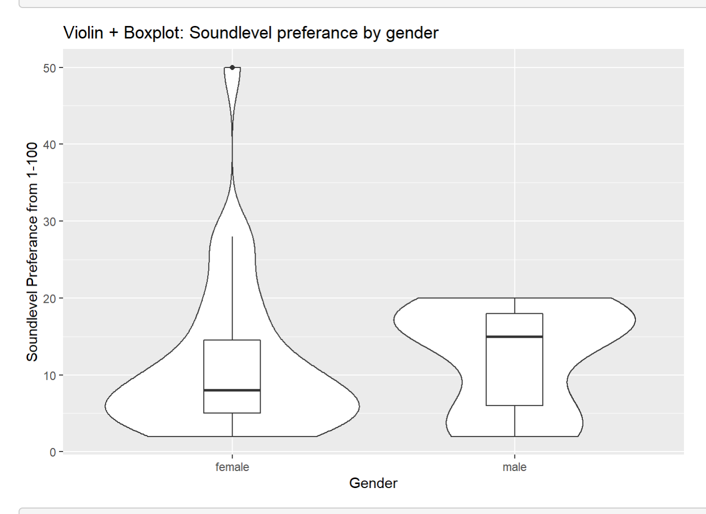
Who likes silence vs. noise best – by gender? Also in this case you should plot the data using `ggplot2` (including error bars depicting the standard error of the mean), then use the `mean()` and `sd()` functions to find mean and standard deviation of the two genders (still making a summary data set with tidyverse and pipes).

Bonus question: If you feel brave, you can instead try making a boxplot (`geom_boxplot()`) or a violin plot (`geom_violin()`) which are better at representing the actual distribution of the data (compared to a bar plot, which only depicts mean and standard deviation).

```
filter_df2 <- df %>%
  filter(!is.na(sound_level_pref))

summary_df2<- filter_df2 %>%
  group_by(gender) %>%
  summarise(mean = mean(sound_level_pref),
            sd = sd(sound_level_pref),
            se = sd/sqrt(n()),
            max=mean+se,
            min=mean-se
            )

summary_df2 %>%
  ggplot(aes(x=gender,y=mean,ymin=min,ymax=max))+
  geom_col(fill="white",col = 'black')+
  geom_errorbar(width=0.5)+
  labs(x="Gender",
       y="Soundlevel Preferance from 1-100",
       title = "Errorbar: Soundlevel preferance by gender")
```




```
filter_df2 %>%
  ggplot(aes(x=gender,y=sound_level_pref))+
  geom_violin()+
  geom_boxplot(width=0.2)+
  labs(x="Gender",
       y="Soundlevel Preferance from 1-100",
       title = "Violin + Boxplot: Soundlevel preferance by gender")
```




```
summary_df2
```

```
## # A tibble: 2 x 6
##   gender  mean    sd    se   max   min
##   <chr>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 female  11.5 10.2   1.84  13.3  9.65
## 2 male    12.3  6.89  1.72  14.0 10.6
```

Explain your results in plain terms here (max 3 sentences):

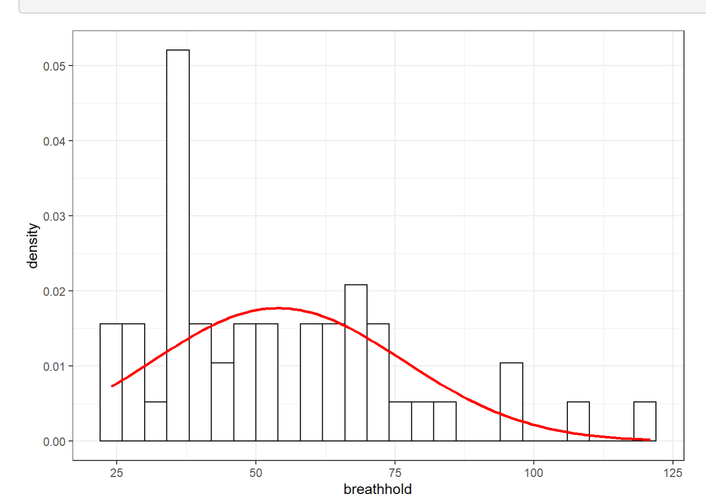
The errorbars overlap rather much, indicating that the population might have the same mean for both genders.

The box+violinplot indicates that only the female data has outliers.

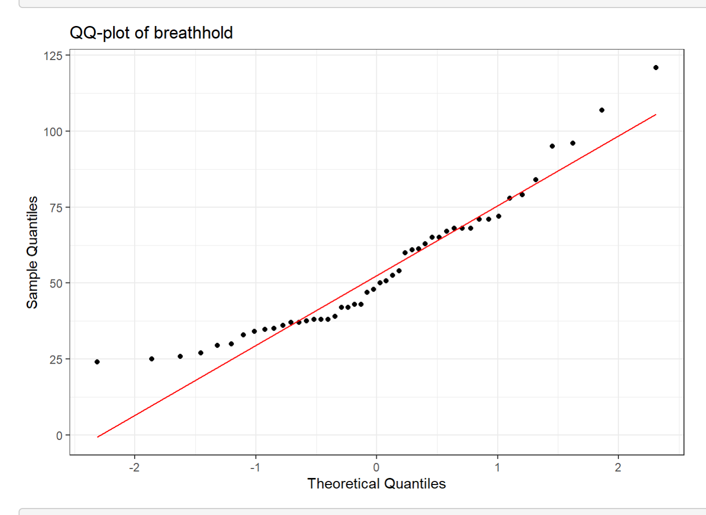
Question 3

Is the `breathhold` variable normally distributed? Provide both visual (histogram and QQ-plot) and numeric (Shapiro-Wilk test and skewness/kurtosis values) support for your answer.

```
ggplot(df, aes(breathhold)) +
 geom_histogram(aes(y = ..density..), binwidth=4, colour = "black", fill = "white") +
 stat_function(fun = dnorm, args = list(mean = mean(df$breathhold),
 sd = sd(df$breathhold)), colour = "red", size = 1) +
 theme_bw()
```




```
df %>%
  ggplot(aes(sample = breathhold))+
  stat_qq()+
  stat_qq_line(color="red")+
   labs(x = "Theoretical Quantiles",
        y = "Sample Quantiles",
        title = "QQ-plot of breathhold")+
  theme_bw()
```




```
round(pastecs::stat.desc(cbind(Breathhold=df$breathhold), basic = FALSE, norm = TRUE,desc= FALSE), digits = 2)
```

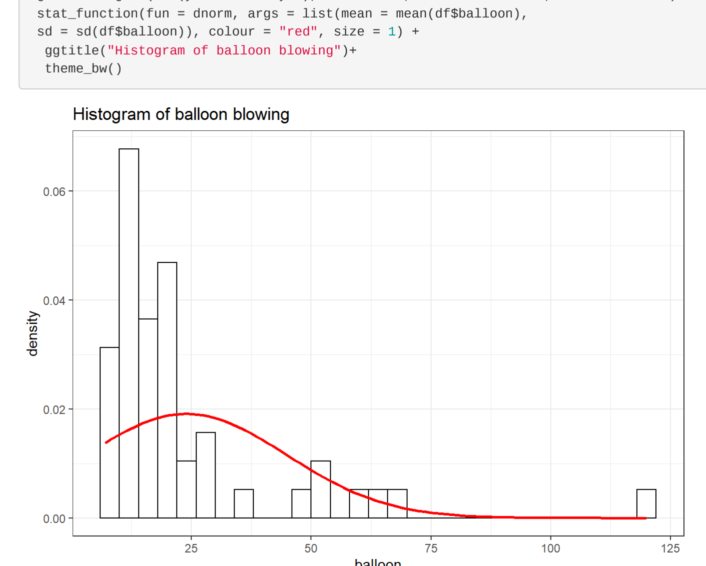
```
##            Breathhold
## skewness         0.89
## skew.2SE         1.29
## kurtosis         0.31
## kurt.2SE         0.23
## normtest.W       0.92
## normtest.p       0.00
```

Question 4

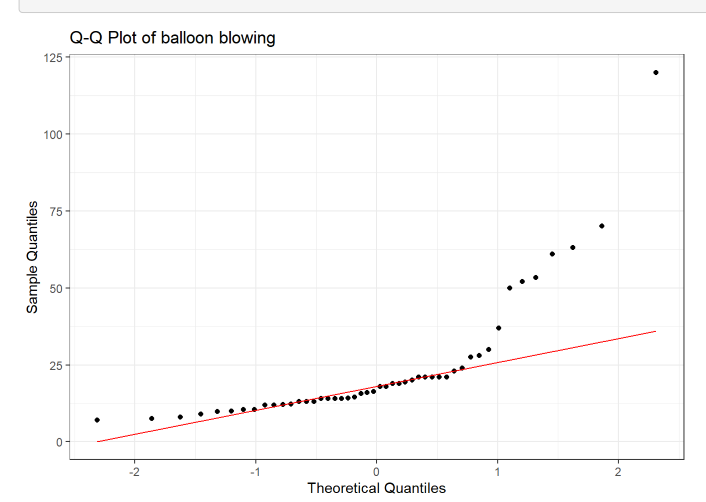
Are the two balloon reaction time variables (`balloon` and `balloon_balance`) normally distributed? Provide visual (histogram and QQ-plot) and numeric (Shapiro-Wilk test and skewness/kurtosis values) support for your answer.

If they are not, then discuss your results below.

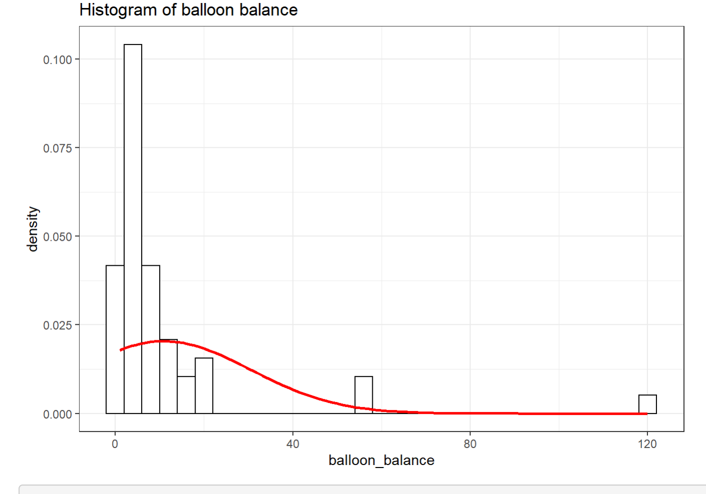
```
ggplot(df, aes(balloon)) +
 geom_histogram(aes(y = ..density..), binwidth=4, colour = "black", fill = "white") +
 stat_function(fun = dnorm, args = list(mean = mean(df$balloon),
 sd = sd(df$balloon)), colour = "red", size = 1) +
  ggtitle("Histogram of balloon blowing")+
  theme_bw()
```



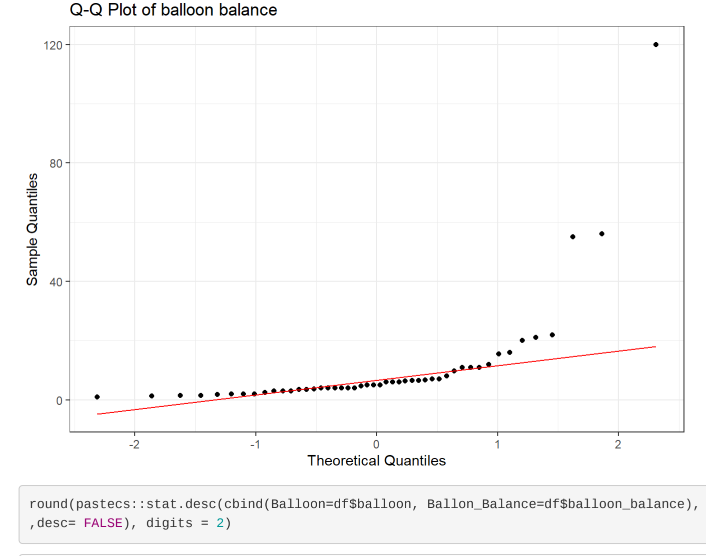

```
ggplot(df, aes(sample = balloon)) +
 stat_qq() +
 stat_qq_line(colour = "red") +
 labs(x = "Theoretical Quantiles", y = "Sample Quantiles") +
 ggtitle("Q-Q Plot of balloon blowing") +
 theme_bw()
```




```
ggplot(df, aes(balloon_balance)) +
 geom_histogram(aes(y = ..density..), binwidth=4, colour = "black", fill = "white") +
 stat_function(fun = dnorm, args = list(mean = mean(df$balloon_balance),
 sd = sd(df$balloon_balance)), colour = "red", size = 1) +
  ggtitle("Histogram of balloon balance")+
  theme_bw()
```




```
ggplot(df, aes(sample = balloon_balance)) +
 stat_qq() +
 stat_qq_line(colour = "red") +
 labs(x = "Theoretical Quantiles", y = "Sample Quantiles") +
 ggtitle("Q-Q Plot of balloon balance") +
 theme_bw()
```




```
round(pastecs::stat.desc(cbind(Balloon=df$balloon, Ballon_Balance=df$balloon_balance), basic = FALSE, norm = TRUE
,desc= FALSE), digits = 2)
```

```
##            Balloon Ballon_Balance
## skewness      2.55           4.08
## skew.2SE      3.72           5.95
## kurtosis      7.55          18.33
## kurt.2SE      5.60          13.59
## normtest.W    0.67           0.46
## normtest.p    0.00           0.00
```

Explain your results in plain terms here (max 3 sentences):

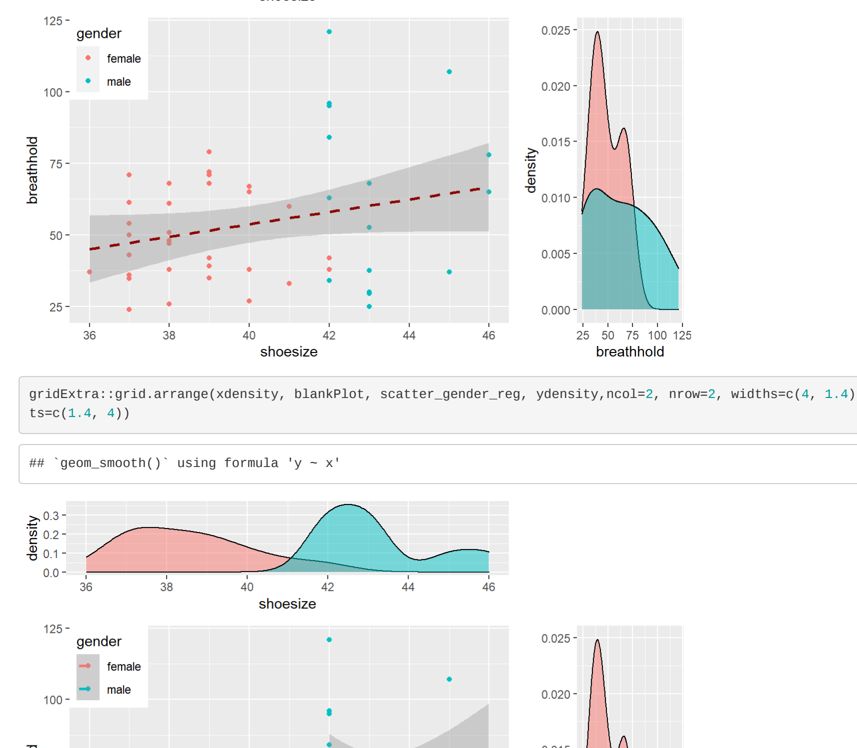
The data is not normally distributed, which we can deduct in two ways: Visually we can see that the data is skewed from the histogram and QQ-plot. Numerically we can see that skew.2SE and kurt.2SE is way above 1 for both data sets, and therefore not normally distributed data.

Question 5

Shoe size could tell us something about general body size, which could also be connected to one's ability to hold your 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. Try plotting the two sets of data against each other using a *scatter plot* (hint: both variables are continuous variables). You can make a scatter plot in `ggplot2` using the `geom_point()` function and plotting one variable on each axis. Use grouping in your plot to distinguish the relationship between shoe size and holding breath for males and females, since we expect males and females to have different show sizes. You can for instance use the `color` parameter within the `aes()` function to color by gender.

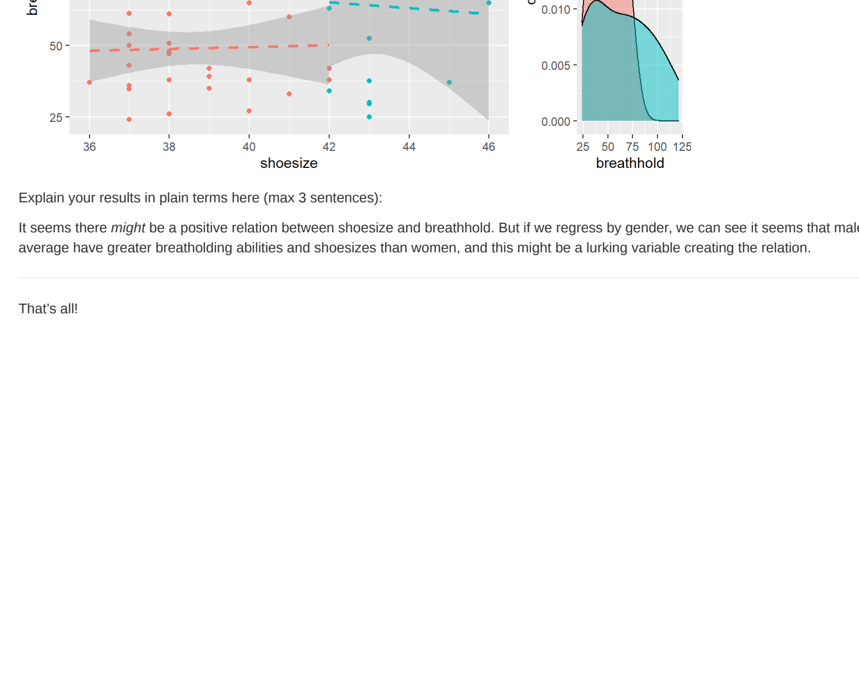
```
scatterplot<- ggplot(df,aes(shoesize,breathhold))+
  geom_point(aes(color=gender))+
  geom_smooth(method=lm,linetype="dashed",color="darkred")+
  theme(legend.position=c(0,1), legend.justification=c(0,1))

scatter_gender_reg<- ggplot(df,aes(shoesize,breathhold,color=gender))+
  geom_point()+
  geom_smooth(method=lm,linetype="dashed")+
  theme(legend.position=c(0,1), legend.justification=c(0,1))


xdensity <- ggplot(df, aes(shoesize, fill=gender)) +
  geom_density(alpha=.5) +
  theme(legend.position = "none")

ydensity <- ggplot(df, aes(breathhold, fill=gender)) +
  geom_density(alpha=.5) +
  theme(legend.position = "none")

blankPlot <- ggplot()+geom_blank(aes(1,1))+
  theme(plot.background = element_blank(),
   panel.grid.major = element_blank(),
   panel.grid.minor = element_blank(),
   panel.border = element_blank(),
   panel.background = element_blank(),
   axis.title.x = element_blank(),
   axis.title.y = element_blank(),
   axis.text.x = element_blank(),
   axis.text.y = element_blank(),
   axis.ticks = element_blank()
     )

gridExtra::grid.arrange(xdensity, blankPlot, scatterplot, ydensity,ncol=2, nrow=2, widths=c(4, 1.4), heights=c(1.
4, 4))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
gridExtra::grid.arrange(xdensity, blankPlot, scatter_gender_reg, ydensity,ncol=2, nrow=2, widths=c(4, 1.4), heigh
ts=c(1.4, 4))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Explain your results in plain terms here (max 3 sentences):

It seems there *might* be a positive relation between shoesize and breathhold. But if we regress by gender, we can see it seems that males on average have greater breatholding abilities and shoesizes than women, and this might be a lurking variable creating the relation.

That's all!