Homework 2

Huanzhang Xia

Table of contents

Link to the Github repository

❗ Due: Tue, Feb 14, 2023 @ 11:59pm

Please read the instructions carefully before submitting your assignment.

1. This assignment requires you to only upload a `PDF` file on Canvas
2. Don't collapse any code cells before submitting.
3. Remember to make sure all your code output is rendered properly before uploading your submission.

Please add your name to the author information in the frontmatter before submitting your assignment

For this assignment, we will be using the Abalone dataset from the UCI Machine Learning Repository. The dataset consists of physical measurements of abalone (a type of marine snail) and includes information on the age, sex, and size of the abalone.

We will be using the following libraries:

```r
rm(list=ls())
library(readr)
library(tidyr)
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(purrr)
library(cowplot)
```

Question 1

30 points

EDA using `readr`, `tidyr` and `ggplot2`

1.1 (5 points)

Load the "Abalone" dataset as a tibble called `abalone` using the URL provided below. The `abalone_col_names` variable contains a vector of the column names for this dataset (to be consistent with the R naming pattern). Make sure you read the dataset with the provided column names.

```r
url <- "http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data"

abalone_col_names <- c(
  "sex",
  "length",
```

```r
    "diameter",
    "height",
    "whole_weight",
    "shucked_weight",
    "viscera_weight",
    "shell_weight",
    "rings"
  )

  abalone <- read.csv(url)
  names(abalone)[1:9] <- abalone_col_names
```

1.2 (5 points)

Remove missing values and `NA`s from the dataset and store the cleaned data in a tibble called `df`. How many rows were dropped?

```r
  df <- tibble(abalone)
  drop_na(df)
```

```
# A tibble: 4,176 x 9
   sex   length diameter height whole_weight shucked_wei~1 visce~2 shell~3 rings
   <chr>  <dbl>    <dbl>  <dbl>        <dbl>         <dbl>   <dbl>   <dbl> <int>
 1 M      0.35     0.265  0.09         0.226        0.0995  0.0485   0.07      7
 2 F      0.53     0.42   0.135        0.677        0.256   0.142    0.21      9
 3 M      0.44     0.365  0.125        0.516        0.216   0.114    0.155    10
 4 I      0.33     0.255  0.08         0.205        0.0895  0.0395   0.055     7
 5 I      0.425    0.3    0.095        0.352        0.141   0.0775   0.12      8
 6 F      0.53     0.415  0.15         0.778        0.237   0.142    0.33     20
 7 F      0.545    0.425  0.125        0.768        0.294   0.150    0.26     16
 8 M      0.475    0.37   0.125        0.509        0.216   0.112    0.165     9
 9 F      0.55     0.44   0.15         0.894        0.314   0.151    0.32     19
10 F      0.525    0.38   0.14         0.606        0.194   0.148    0.21     14
# ... with 4,166 more rows, and abbreviated variable names 1: shucked_weight,
#   2: viscera_weight, 3: shell_weight
```

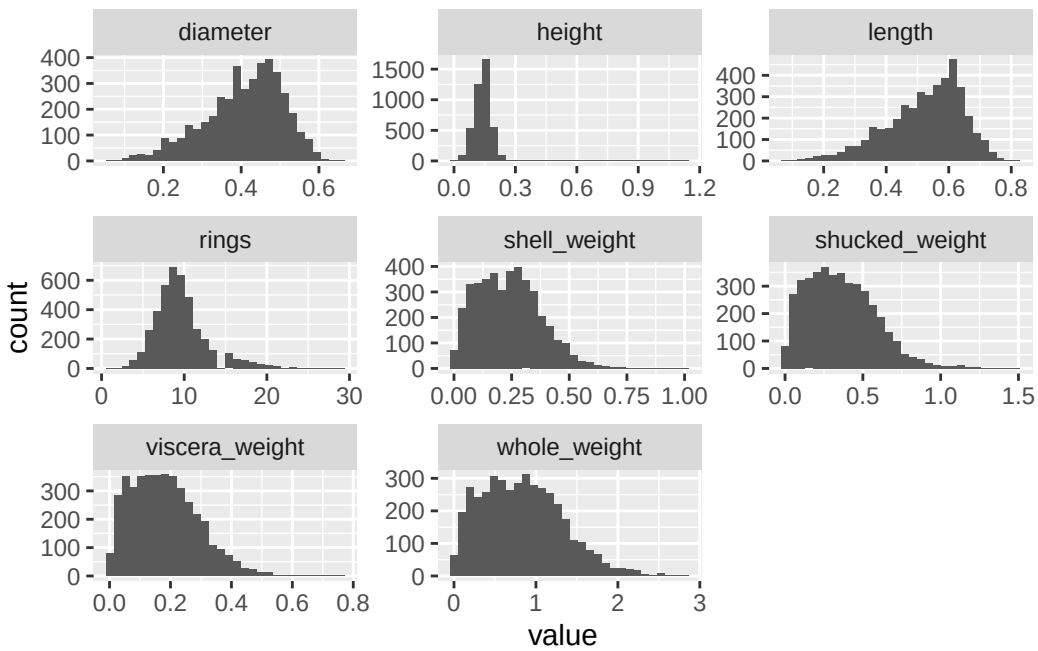
No row is dropped.

1.3 (5 points)

Plot histograms of all the quantitative variables in a **single plot** [1]

```
df%>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

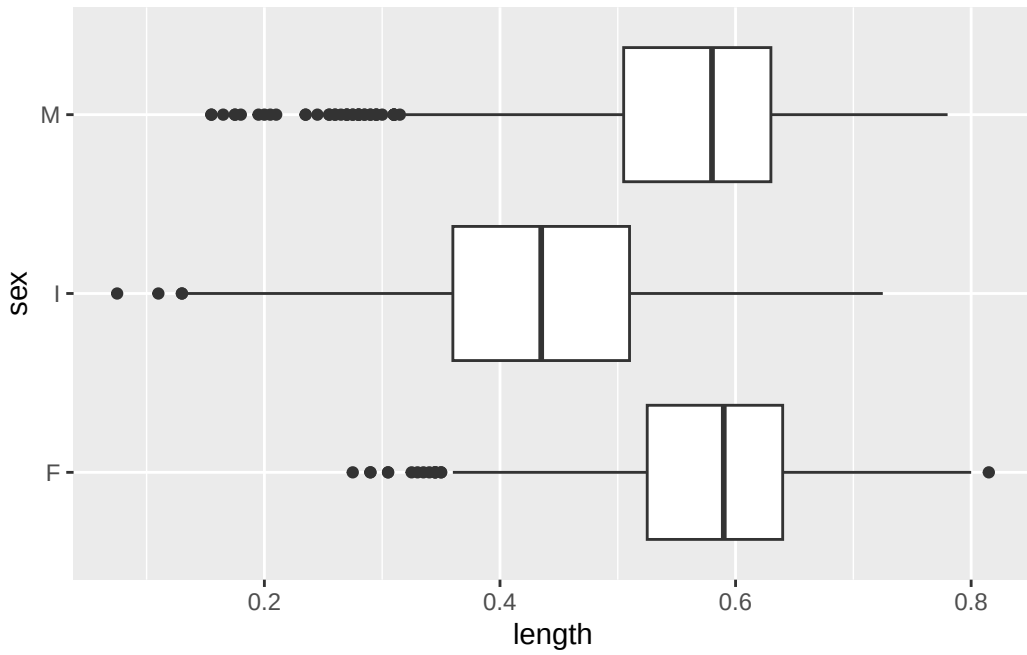


1.4 (5 points)

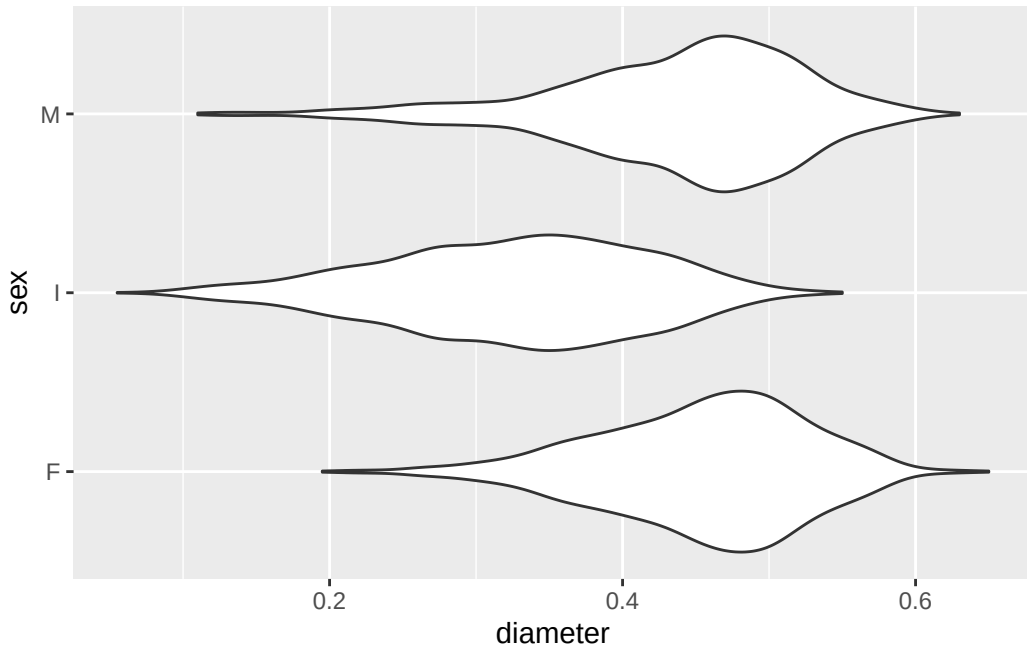
Create a boxplot of `length` for each `sex` and create a violin-plot of of `diameter` for each `sex`. Are there any notable differences in the physical appearences of abalones based on your analysis here?

[1] You can use the `facet_wrap()` function for this. Have a look at its documentation using the help console in R

```r
ggplot(df,aes(y=sex,x=length))+geom_boxplot()
```




```r
ggplot(df, aes(x=diameter, y=sex)) +
  geom_violin()
```

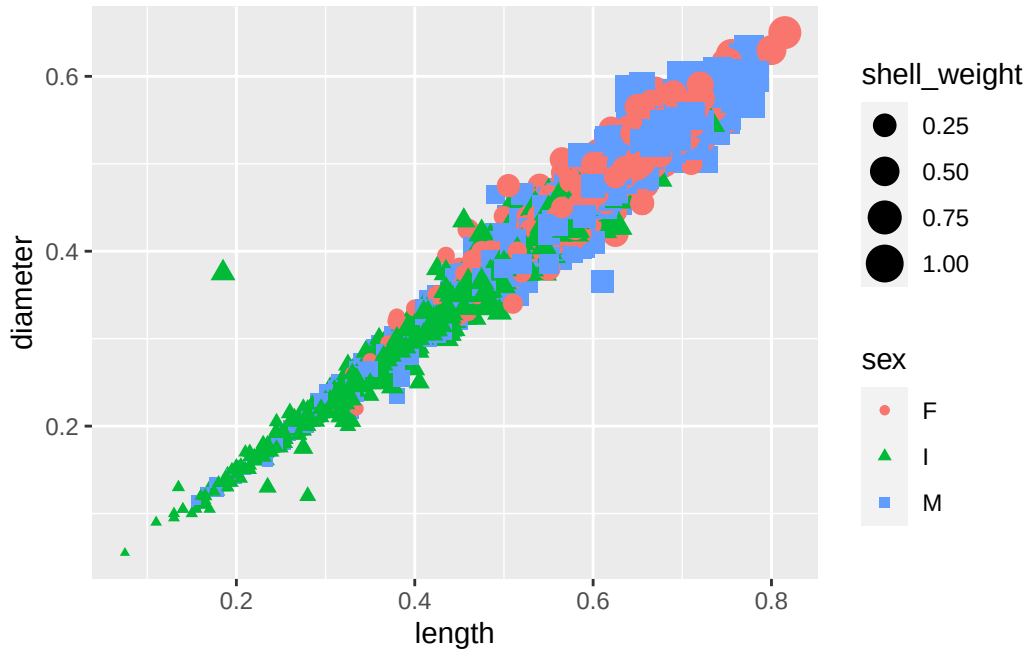


1.5 (5 points)

Create a scatter plot of `length` and `diameter`, and modify the shape and color of the points based on the `sex` variable. Change the size of each point based on the `shell_weight` value for each observation. Are there any notable anomalies in the dataset?

Inter sex are smaller in length and diameter. The higher the shell weight, the higher the diameter and length.

```
ggplot(df, aes(x=length, y=diameter,color=sex,shape=sex,size=shell_weight)) +
  geom_point()
```



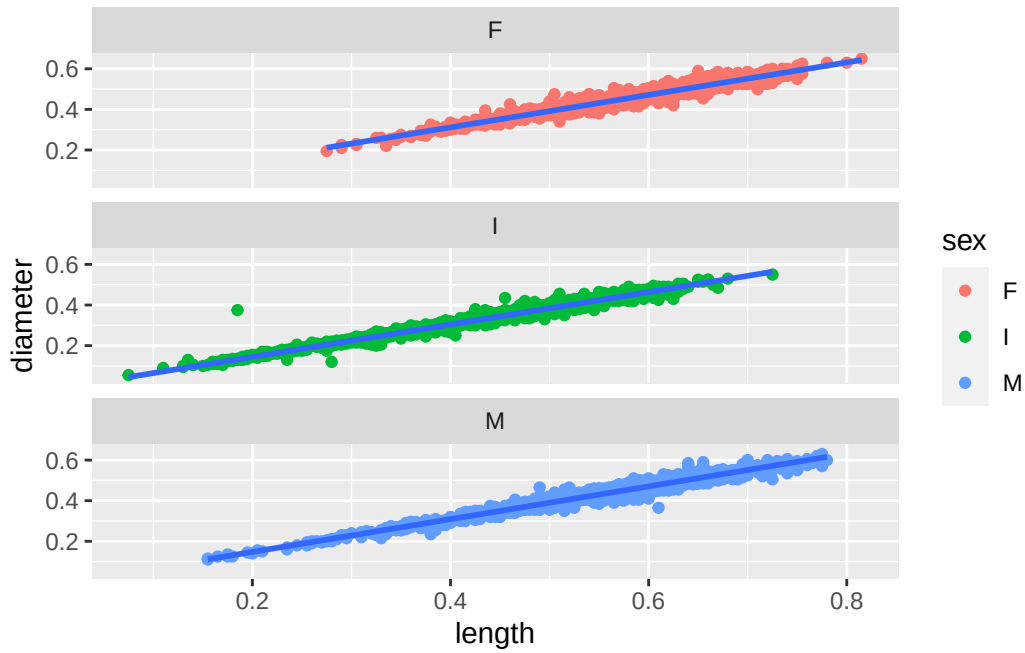
1.6 (5 points)

For each `sex`, create separate scatter plots of `length` and `diameter`. For each plot, also add a **linear** trendline to illustrate the relationship between the variables. Use the `facet_wrap()` function in R for this, and ensure that the plots are vertically stacked **not** horizontally. You should end up with a plot that looks like this: [2]

```
ggplot(df,aes(x=length,y=diameter))+geom_point(aes(color=sex))+geom_smooth(method=lm,se=FA
```

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

[2]Plot example for 1.6



—

Question 2

40 points

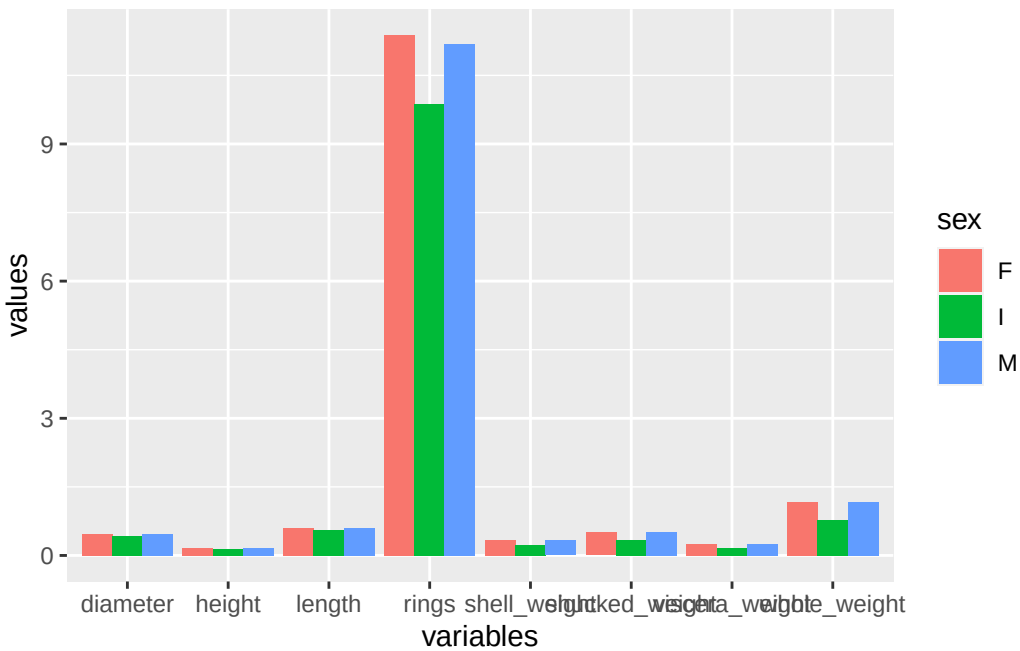
More advanced analyses using `dplyr`, `purrrr` and `ggplot2`

2.1 (10 points)

Filter the data to only include abalone with a length of at least 0.5 meters. Group the data by `sex` and calculate the mean of each variable for each group. Create a bar plot to visualize the mean values for each variable by `sex`.

```
dffilter <- df%>%
  filter(length>=0.5)%>%
  group_by(sex)%>%
  summarise(across(everything(),mean))
dffilterp <- dffilter%>%pivot_longer(cols=c("length",
```

```
  "diameter",
  "height",
  "whole_weight",
  "shucked_weight",
  "viscera_weight",
  "shell_weight",
  "rings"),names_to = "variables",values_to = "values")
ggplot(dffilterp,aes(x=values,y=variables,fill=sex))+geom_col(position=position_dodge())+c
```



2.2 (15 points)

Implement the following in a **single command**:

1. Temporarily create a new variable called `num_rings` which takes a value of:

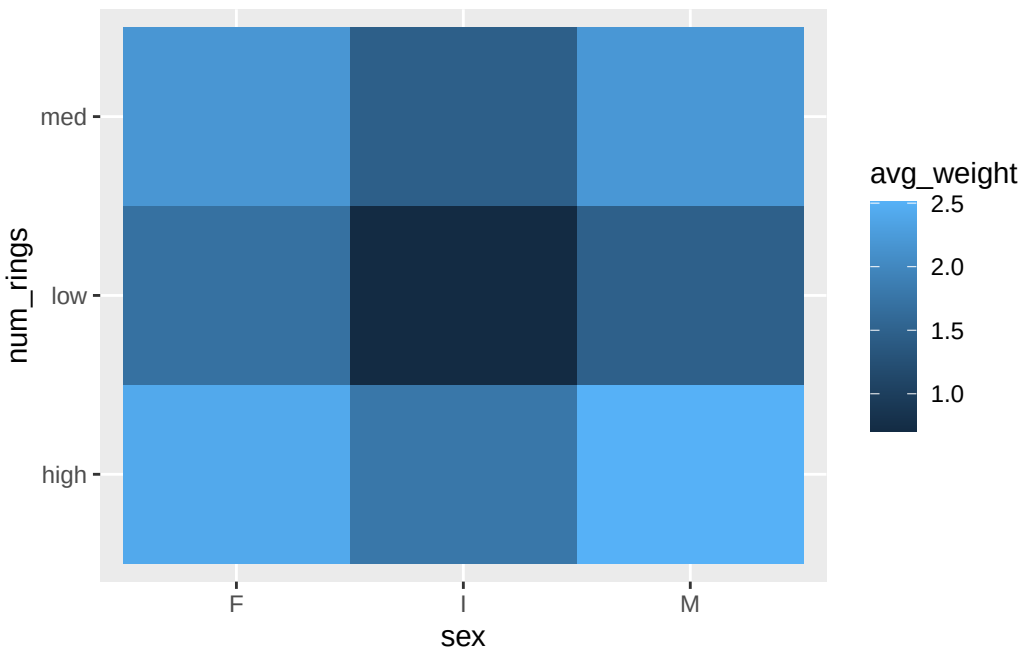
- `"low"` if `rings < 10`
- `"high"` if `rings > 20`, and
- `"med"` otherwise

2. Group `df` by this new variable and `sex` and compute `avg_weight` as the average of the `whole_weight + shucked_weight + viscera_weight + shell_weight` for each combination of `num_rings` and `sex`.

3. Use the `geom_tile()` function to create a tile plot of `num_rings` vs `sex` with the color indicating of each tile indicating the `avg_weight` value.

```
df%>%
  mutate(num_rings=ifelse(rings<10 | rings>20,case_when(rings < 10 ~ 'low',rings > 20 ~ 'h
  group_by(num_rings,sex)%>%
  summarize(avg_weight=mean(whole_weight + shucked_weight + viscera_weight + shell_weight)
  ggplot()+geom_tile(aes(y=num_rings,x=sex,fill=avg_weight))
```

```
`summarise()` has grouped output by 'num_rings'. You can override using the
`.groups` argument.
```



2.3 (5 points)

Make a table of the pairwise correlations between all the numeric variables rounded to 2 decimal points. Your final answer should look like this [3]

```
table <- df%>%
  keep(is.numeric)%>%
  cor()
round(table, 2)
```

```
               length diameter height whole_weight shucked_weight
length           1.00     0.99   0.83         0.93           0.90
diameter         0.99     1.00   0.83         0.93           0.89
height           0.83     0.83   1.00         0.82           0.77
whole_weight     0.93     0.93   0.82         1.00           0.97
shucked_weight   0.90     0.89   0.77         0.97           1.00
viscera_weight   0.90     0.90   0.80         0.97           0.93
shell_weight     0.90     0.91   0.82         0.96           0.88
rings            0.56     0.58   0.56         0.54           0.42
               viscera_weight shell_weight rings
length                   0.90         0.90  0.56
diameter                 0.90         0.91  0.58
height                   0.80         0.82  0.56
whole_weight             0.97         0.96  0.54
shucked_weight           0.93         0.88  0.42
viscera_weight           1.00         0.91  0.50
shell_weight             0.91         1.00  0.63
rings                    0.50         0.63  1.00
```

2.4 (10 points)

Use the `map2()` function from the `purrr` package to create a scatter plot for each *quantitative* variable against the number of `rings` variable. Color the points based on the `sex` of each abalone. You can use the `cowplot::plot_grid()` function to finally make the following grid of plots.

```
#df%>%
 # keep(is.numeric) %>%
  #map2()
```

—

[3] Table for 2.3

Question 3

30 points

Linear regression using `lm`

3.1 (10 points)

Perform a simple linear regression with `diameter` as the covariate and `height` as the response. Interpret the model coefficients and their significance values.

The intercept is -0.003784, meaning that when diameter is 0, the height is -0.003784, which is meaningless. The slope is 0.351346, meaning that per 1 increase in the diameter, we expect an increase of 0.351346 in height. Both terms are significant.

```r
model1 <- lm(height~diameter,data=df)
summary(model1)
```

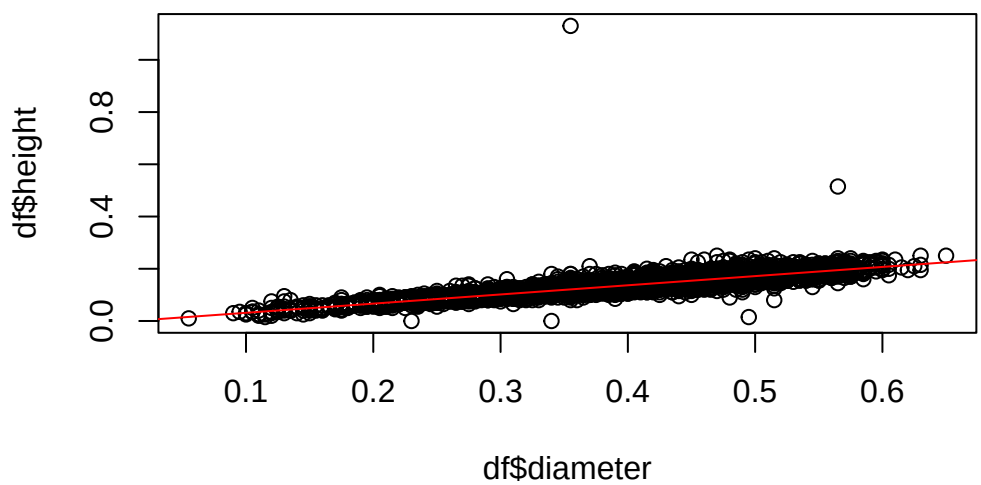
```
Call:
lm(formula = height ~ diameter, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.15513 -0.01044 -0.00148  0.00852  1.00906 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.003784   0.001512  -2.502   0.0124 *  
diameter     0.351346   0.003602  97.540   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0231 on 4174 degrees of freedom
Multiple R-squared:  0.6951,	Adjusted R-squared:  0.695 
F-statistic:  9514 on 1 and 4174 DF,  p-value: < 2.2e-16
```

3.2 (10 points)

Make a scatterplot of `height` vs `diameter` and plot the regression line in `color="red"`. You can use the base `plot()` function in R for this. Is the linear model an appropriate fit for this relationship? Explain.

Yes. The data is mostly linear.

```
plot(df$diameter,df$height)
abline(model1,col="red")
```

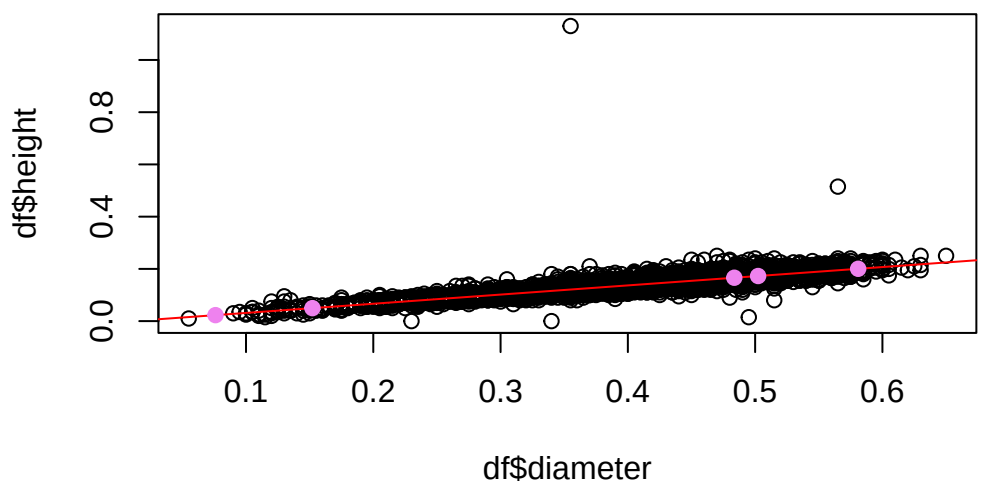


3.3 (10 points)

Suppose we have collected observations for "new" abalones with `new_diameter` values given below. What is the expected value of their `height` based on your model above? Plot these new observations along with your predictions in your plot from earlier using `color="violet"`

```
new_diameters <- c(
  0.15218946,
  0.48361548,
  0.58095513,
```

```
  0.07603687,
  0.50234599,
  0.83462092,
  0.95681938,
  0.92906875,
  0.94245437,
  0.01209518
)

new_data <- data.frame(new_diameters)
newdata <- new_data%>%
  rename(diameter=new_diameters)
height <- predict(model1,newdata=newdata)
plot(df$diameter,df$height)
abline(model1,col="red")
points(x=new_diameters,y=height,pch=19,col="violet")
```



—

Appendix

Session Information

Print your R session information using the following command

```r
sessionInfo()
```

```
R version 4.2.1 (2022-06-23)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Big Sur ... 10.16

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices datasets  utils     methods   base

other attached packages:
[1] cowplot_1.1.1 purrr_1.0.1   dplyr_1.1.0   ggplot2_3.4.1 tidyr_1.3.0
[6] readr_2.1.4

loaded via a namespace (and not attached):
 [1] pillar_1.8.1     compiler_4.2.1   tools_4.2.1      digest_0.6.31
 [5] lattice_0.20-45  nlme_3.1-157     jsonlite_1.8.4   evaluate_0.20
 [9] lifecycle_1.0.3  tibble_3.1.8     gtable_0.3.1     mgcv_1.8-40
[13] pkgconfig_2.0.3  rlang_1.0.6      Matrix_1.4-1     cli_3.6.0
[17] yaml_2.3.7       xfun_0.37        fastmap_1.1.0    withr_2.5.0
[21] knitr_1.42       generics_0.1.3   vctrs_0.5.2      hms_1.1.2
[25] grid_4.2.1       tidyselect_1.2.0 glue_1.6.2       R6_2.5.1
[29] fansi_1.0.4      rmarkdown_2.20   farver_2.1.1     tzdb_0.3.0
[33] magrittr_2.0.3   splines_4.2.1    scales_1.2.1     ellipsis_0.3.2
[37] htmltools_0.5.4  colorspace_2.1-0 renv_0.16.0-53   labeling_0.4.2
[41] utf8_1.2.3       munsell_0.5.0
```