# Lab 2 Assignment

```
library(ISLR)
library(MASS)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

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

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

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

## Lab Assignment

### Q1. Discuss this example with your group. (no submission needed)

*Galton's data on the heights of parents and their children, by child.*

This data set lists the individual observations for 934 children in 205 families on which Galton (1886) based his cross-tabulation.

Suppose you are tasked with building a machine learning algorithm that predicts the *son's height* Y, using the *father's height* X.

```
library(HistData)

#?GaltonFamilies
head(GaltonFamilies)
```

```
##   family father mother midparentHeight children childNum gender childHeight
## 1    001   78.5   67.0           75.43        4        1   male        73.2
## 2    001   78.5   67.0           75.43        4        2 female        69.2
## 3    001   78.5   67.0           75.43        4        3 female        69.0
## 4    001   78.5   67.0           75.43        4        4 female        69.0
## 5    002   75.5   66.5           73.66        4        1   male        73.5
## 6    002   75.5   66.5           73.66        4        2   male        72.5
```

```r
table(GaltonFamilies$family)
```

```
##
##  001  002  003  004  005  006  007  008  009  010  011  012  013  014  015  016
##    4    4    2    5    6    1    6    3    1    1    8    1    2    2    3    9
##  017  018  019  020  021  022  023  024  025  026  027  028  029  030  031  032
##    6    3    1    8    3    3    7    1    2    5    3    6    3    1    6    5
##  033  034  035  036  037  038  039  040  041  042  043  044  045  046  047  048
##    5    1    5    4    4    6    2    5    1    6    2    2    3    8    4    3
##  049  050  051  052  053  054  055  056  057  058  059  060  061  062  063  064
##    7    2    2    5    9    4    5    5    5    7    1    2    4    6    1    5
##  065  066  067  068  069  070  071  072  073  074  075  076  077  078  079  080
##    1   11    4    5    8    5    6    7    3    2    7    7    4    5    8    1
##  081  082  083  084  085  086  087  088  089  090  091  092  093  094  095  096
##    4    9    8    4    5    4    4    4    8    7    3    2    4    2    3    5
##  097  098  099  100  101  102  103  104  105  106  107  108  109  110  111  112
##   10    1    8    3    4    6    7    4    6    7    9    7    7    4    1    3
##  113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128
##    1    6    7    3    1    3    5   11    8    4    5    9    3    4    1    2
##  129  130  131  132  133  134  135  136 136A  137  138  139  140  141  142  143
##    3   11    2    2    7    4    8   10    8    4    5    1   10    8    4    1
##  144  145  146  147  148  149  150  151  152  153  154  155  156  157  158  159
##    4    8    6    1    1    5    1    2    1    5    1    7    4    1   10    5
##  160  161  162  163  164  165  166  167  168  169  170  171  172  173  174  175
##    1    8    6    5    4    3   11    4    8    3    5    1    8    9    5    6
##  176  177  178  179  180  181  182  183  184  185  186  187  188  189  190  191
##    8    5    1    2    6    7    1    4    1   15    4    1    4    5    9    2
##  192  193  194  195  196  197  198  199  200  201  202  203  204
##    6    6    2    3    4    5    7    7    1    2    2    3    2
```

```r
set.seed(1983)
galton_heights <- GaltonFamilies %>%
  filter(gender == "male") %>%
  group_by(family) %>% #grouped data frame
  sample_n(1) %>%
  ungroup() %>% #removes grouping
  select(father, childHeight) %>%
  rename(son = childHeight)

head(galton_heights)
```

```
## # A tibble: 6 x 2
##   father   son
##    <dbl> <dbl>
## 1   78.5  73.2
## 2   75.5  72.5
## 3   75    71
## 4   75    70.5
## 5   75    69
## 6   74    76.5
```

(i) Let's split the data into training and testing.

```r
#?slice
#slice() lets you index rows by their (integer) locations.
```

```
#It allows you to select, remove, and duplicate rows.

y <- galton_heights$son
test_index <- createDataPartition(y, times = 1, p = 0.3, list = FALSE) #30% for the test set

train_set <- galton_heights %>% slice(-test_index)
test_set <- galton_heights %>% slice(test_index)
```

(ii) The *best guess(prediction)* would be the mean;

```
(yhat <- mean(train_set$son))
```

```
## [1] 69.14839
```

```
# then the error would be (yhat-ybar)

mean((yhat - test_set$son)^2)
```

```
## [1] 7.738353
```

(iii) Can we do better? Yes.

$$f(x) = E(Y|X = x) = \beta_0 + \beta_1 x$$

```
fit <- lm(son ~ father, data = train_set)
fit$coef
```

```
## (Intercept)      father
##  37.6101619   0.4570757
```

$$\hat{f}(x) = 43.7 + 0.36x$$

We can see that this does indeed provide an improvement over our guessing approach.

```
y_hat <- fit$coef[1] + fit$coef[2]*test_set$father
mean((y_hat - test_set$son)^2)
```

```
## [1] 6.634095
```

(iv) Predict

```
y_hat <- predict(fit, test_set)
mean((y_hat - test_set$son)^2)
```

```
## [1] 6.634095
```

Learn more;

```
?predict.lm
?predict.glm
```

many machine learning algorithms have a predict function.

## Q2: Coefficient Interpretation

Suppose we have the following model to estimate GPA for Georgetown undergraduate students.

$$GPA = \beta_0 + \beta_1 * Female + \beta_2 Age$$

Where *Age* is a continuous variable measured in years and gender is encoded as:

- Female = 1 if student is female
- Female = 0 if student is male

a) What is the interpretation of the $\beta_0$ coefficient?

b) What is the interpretation of the $\beta_1$ coefficient?

c) Suppose $\beta_0 = 1.5$ and $\beta_1 = .5$ and $\beta_2 = 1.5$. What is the expected GPA of a student who is male and 20 years old?

d) Suppose you believe that males have lower GPAs than females. Further, suppose you believe that GPA tends to decrease as students get older. What would this imply about the signs of the coefficients of the model?

e) Suppose we change the units of measurement for *Age* from years to days. What effect would that have on $\beta_2$? This is an important reminder about interpreting the absolute magnitude of the coefficients.

## Question 3

Data Science Question: Are the variables contributing for predicting “Valence” of the songs is same for both musicians; Taylor Swift and John Legend?

Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

```r
Artists <- read.csv("Artists.csv",header = TRUE)
head(Artists)
```

```
##   X  artist_name Valence danceability energy loudness speechiness acousticness
## 1 1 Taylor Swift   0.328        0.594  0.713   -5.314      0.0503     0.000328
## 2 2 Taylor Swift   0.408        0.516  0.777   -4.908      0.0375     0.001080
## 3 3 Taylor Swift   0.299        0.645  0.593   -6.506      0.0288     0.034400
## 4 4 Taylor Swift   0.767        0.584  0.557   -6.371      0.0342     0.012900
## 5 5 Taylor Swift   0.132        0.440  0.528   -7.809      0.0317     0.017100
## 6 6 Taylor Swift   0.642        0.642  0.695   -5.620      0.0281     0.000443
##   liveness   tempo                                track_name
## 1   0.1140 129.958         State Of Grace (Taylor's Version)
## 2   0.0761 125.047                    Red (Taylor's Version)
## 3   0.1300 109.984            Treacherous (Taylor's Version)
## 4   0.0576 154.008 I Knew You Were Trouble (Taylor's Version)
## 5   0.2340 185.972           All Too Well (Taylor's Version)
## 6   0.0753 103.984                     22 (Taylor's Version)
##              album_name album_release_year
## 1 Red (Taylor's Version)               2021
## 2 Red (Taylor's Version)               2021
## 3 Red (Taylor's Version)               2021
## 4 Red (Taylor's Version)               2021
## 5 Red (Taylor's Version)               2021
## 6 Red (Taylor's Version)               2021
```

Create a new variable with the name " danceable" and add it to the “Artists” data set. Fill this variable with values “High”,“Medium”,“Low” according to the danceability value such that 0.8-1 is “High”, 0.5-0.79 is " Medium" and 0.0-0.49 is “Low”.

```r
Artists$danceable<-rep(0, nrow(Artists))
```

```
Artists1 <- within(Artists, {

  danceable[danceability>= 0.800 & danceability <= 1] <- "High"
  danceable[danceability>= 0.500 & danceability < 0.8] <- "Medium"
  danceable[danceability <= 0.499] <- "Low"
} )

head(Artists1)
```

```
##   X  artist_name Valence danceability energy loudness speechiness acousticness
## 1 1 Taylor Swift   0.328        0.594  0.713   -5.314      0.0503     0.000328
## 2 2 Taylor Swift   0.408        0.516  0.777   -4.908      0.0375     0.001080
## 3 3 Taylor Swift   0.299        0.645  0.593   -6.506      0.0288     0.034400
## 4 4 Taylor Swift   0.767        0.584  0.557   -6.371      0.0342     0.012900
## 5 5 Taylor Swift   0.132        0.440  0.528   -7.809      0.0317     0.017100
## 6 6 Taylor Swift   0.642        0.642  0.695   -5.620      0.0281     0.000443
##   liveness   tempo                                 track_name
## 1   0.1140 129.958          State Of Grace (Taylor's Version)
## 2   0.0761 125.047                     Red (Taylor's Version)
## 3   0.1300 109.984             Treacherous (Taylor's Version)
## 4   0.0576 154.008 I Knew You Were Trouble (Taylor's Version)
## 5   0.2340 185.972            All Too Well (Taylor's Version)
## 6   0.0753 103.984                      22 (Taylor's Version)
##               album_name album_release_year danceable
## 1 Red (Taylor's Version)               2021    Medium
## 2 Red (Taylor's Version)               2021    Medium
## 3 Red (Taylor's Version)               2021    Medium
## 4 Red (Taylor's Version)               2021    Medium
## 5 Red (Taylor's Version)               2021       Low
## 6 Red (Taylor's Version)               2021    Medium
```

## Just do the analysis for John Legend.

a. Fit a linear regression models separately for Artists: *John Legend* , to predict *Valence* of the songs on all the quantitative variables in the data set and "danceable" variable you created earlier .

Remember to train your model on a training set and validate the model using a validation set.

b.Check the multicollinearity c. In cooperate interaction terms c. what is the best model? with or without the ineractions?