Modern Data Mining, HW 2

Group Member Mahika Calyanakoti Group Member Andrew Raines
Group Member Graham Branscom

Due: 11:59 PM, Sunday, 02/25

Contents

Overview

Principle Component Analysis is widely used in data exploration, dimension reduction, data visualization. The aim is to transform original data into uncorrelated linear combinations of the original data while keeping the information contained in the data. High dimensional data tends to show clusters in lower dimensional view.

Clustering Analysis is another form of EDA. Here we are hoping to group data points which are close to each other within the groups and far away between different groups. Clustering using PC's can be effective. Clustering analysis can be very subjective in the way we need to summarize the properties within each group.

Both PCA and Clustering Analysis are so called unsupervised learning. There is no response variables involved in the process.

For supervised learning, we try to find out how does a set of predictors relate to some response variable of the interest. Multiple regression is still by far, one of the most popular methods. We use a linear model as a working model for its simplicity and interpretability. It is important that we use domain knowledge as much as we can to determine the form of the response as well as the function format of the factors on the other hand.

Important Notice: This homework encompasses material from three modules. You will have a period of three weeks to complete it. Please manage your time accordingly.

0.1 Objectives

- PCA
- SVD
- Clustering Analysis
- Linear Regression

0.2 Review materials

- Study Module 2: PCA
- Study Module 3: Clustering Analysis
- Study Module 4: Multiple regression (Including Simple regression as well)

0.3 Data needed

- `NLSY79.csv`
- `brca_subtype.csv`
- `brca_x_patient.csv`

1 Case study 1: Self-esteem

Self-esteem generally describes a person's overall sense of self-worthiness and personal value. It can play significant role in one's motivation and success throughout the life. Factors that influence self-esteem can be inner thinking, health condition, age, life experiences etc. We will try to identify possible factors in our data that are related to the level of self-esteem.

In the well-cited National Longitudinal Study of Youth (NLSY79), it follows about 13,000 individuals and numerous individual-year information has been gathered through surveys. The survey data is open to public here. Among many variables we assembled a subset of variables including personal demographic variables in different years, household environment in 79, ASVAB test Scores in 81 and Self-Esteem scores in 81 and 87 respectively.

The data is store in `NLSY79.csv`.

Here are the description of variables:

Personal Demographic Variables

- Gender: a factor with levels "female" and "male"
- Education05: years of education completed by 2005
- HeightFeet05, HeightInch05: height measurement. For example, a person of 5'10 will be recorded as HeightFeet05=5, HeightInch05=10.
- Weight05: weight in lbs.
- Income87, Income05: total annual income from wages and salary in 2005.
- Job87 (missing), Job05: job type in 1987 and 2005, including Protective Service Occupations, Food Preparation and Serving Related Occupations, Cleaning and Building Service Occupations, Entertainment Attendants and Related Workers, Funeral Related Occupations, Personal Care and Service Workers, Sales and Related Workers, Office and Administrative Support Workers, Farming, Fishing and Forestry Occupations, Construction Trade and Extraction Workers, Installation, Maintenance and Repairs Workers, Production and Operating Workers, Food Preparation Occupations, Setters, Operators and Tenders, Transportation and Material Moving Workers

Household Environment

- Imagazine: a variable taking on the value 1 if anyone in the respondent's household regularly read magazines in 1979, otherwise 0
- Inewspaper: a variable taking on the value 1 if anyone in the respondent's household regularly read newspapers in 1979, otherwise 0
- Ilibrary: a variable taking on the value 1 if anyone in the respondent's household had a library card in 1979, otherwise 0
- MotherEd: mother's years of education
- FatherEd: father's years of education
- FamilyIncome78

Variables Related to ASVAB test Scores in 1981

Test	Description
AFQT	percentile score on the AFQT intelligence test in 1981
Coding	score on the Coding Speed test in 1981
Auto	score on the Automotive and Shop test in 1981
Mechanic	score on the Mechanic test in 1981
Elec	score on the Electronics Information test in 1981
Science	score on the General Science test in 1981
Math	score on the Math test in 1981
Arith	score on the Arithmetic Reasoning test in 1981
Word	score on the Word Knowledge Test in 1981
Parag	score on the Paragraph Comprehension test in 1981
Numer	score on the Numerical Operations test in 1981

Self-Esteem test 81 and 87

We have two sets of self-esteem test, one in 1981 and the other in 1987. Each set has same 10 questions. They are labeled as `Esteem81` and `Esteem87` respectively followed by the question number. For example, `Esteem81_1` is Esteem question 1 in 81.

The following 10 questions are answered as 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree

- Esteem 1: "I am a person of worth"
- Esteem 2: "I have a number of good qualities"
- Esteem 3: "I am inclined to feel like a failure"
- Esteem 4: "I do things as well as others"

- Esteem 5: “I do not have much to be proud of”
- Esteem 6: “I take a positive attitude towards myself and others”
- Esteem 7: “I am satisfied with myself”
- Esteem 8: “I wish I could have more respect for myself”
- Esteem 9: “I feel useless at times”
- Esteem 10: “I think I am no good at all”

1.1 Data preparation

Load the data. Do a quick EDA to get familiar with the data set. Pay attention to the unit of each variable. Are there any missing values?

We saw that there were no rows with missing values. We also changed the datatypes appropriately: Gender, Job05,Imagazine, Inewspaper, and Ilibrary to factors. We also made sure to remove rows with negative Income87 or negative HeightFeet05.

```
## 'data.frame':    2431 obs. of  46 variables:
##  $ Subject       : int  2 6 7 8 9 13 16 17 18 20 ...
##  $ Gender        : chr  "female" "male" "male" "female" ...
##  $ Education05   : int  12 16 12 14 14 16 13 13 13 17 ...
##  $ Income87      : int  16000 18000 0 9000 15000 2200 27000 20000 28000 27000 ...
##  $ Job05         : chr  "4700 TO 4960: Sales and Related Workers" "10 TO 430: Executive, Administrati
##  $ Income05      : int  5500 65000 19000 36000 65000 8000 71000 43000 120000 64000 ...
##  $ Weight05      : int  160 187 175 246 180 235 160 188 173 130 ...
##  $ HeightFeet05  : int  5 5 5 5 5 6 5 5 5 5 ...
##  $ HeightInch05  : int  2 5 9 3 6 0 4 10 9 4 ...
##  $ Imagazine     : int  1 0 1 1 1 1 1 1 1 1 ...
##  $ Inewspaper    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Ilibrary      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ MotherEd      : int  5 12 12 9 12 12 12 12 12 12 ...
##  $ FatherEd      : int  8 12 12 6 10 16 12 15 16 18 ...
##  $ FamilyIncome78: int  20000 35000 8502 7227 17000 20000 48000 15000 4510 50000 ...
##  $ Science       : int  6 23 14 18 17 16 13 19 22 21 ...
##  $ Arith         : int  8 30 14 13 21 30 17 29 30 17 ...
##  $ Word          : int  15 35 27 35 28 29 30 33 35 28 ...
##  $ Parag         : int  6 15 8 12 10 13 12 13 14 14 ...
##  $ Number        : int  29 45 32 24 40 36 49 35 48 39 ...
##  $ Coding        : int  52 68 35 48 46 30 58 58 61 54 ...
##  $ Auto          : int  9 21 13 11 13 21 11 18 21 18 ...
##  $ Math          : int  6 23 11 4 13 24 17 21 23 20 ...
##  $ Mechanic      : int  10 21 9 12 13 19 11 19 16 20 ...
##  $ Elec          : int  5 19 11 12 15 16 10 16 17 13 ...
##  $ AFQT          : num  6.84 99.39 47.41 44.02 59.68 ...
##  $ Esteem81_1    : int  1 2 2 1 1 1 2 2 2 1 ...
##  $ Esteem81_2    : int  1 1 1 1 1 1 2 2 2 1 ...
##  $ Esteem81_3    : int  4 4 3 3 4 4 3 3 3 3 ...
##  $ Esteem81_4    : int  1 2 2 2 1 1 2 2 2 1 ...
##  $ Esteem81_5    : int  3 4 3 3 1 4 3 3 3 3 ...
##  $ Esteem81_6    : int  3 2 2 2 1 1 2 2 2 2 ...
##  $ Esteem81_7    : int  1 2 2 3 1 1 3 2 2 1 ...
##  $ Esteem81_8    : int  3 4 2 3 4 4 3 3 3 3 ...
##  $ Esteem81_9    : int  3 3 3 3 4 4 3 3 3 3 ...
##  $ Esteem81_10   : int  3 4 3 3 4 4 3 3 3 3 ...
##  $ Esteem87_1    : int  2 1 2 1 1 1 1 2 1 1 ...
##  $ Esteem87_2    : int  1 1 2 1 1 1 1 2 1 1 ...
```

```
##  $ Esteem87_3    : int  4 4 4 3 4 4 4 3 4 4 ...
##  $ Esteem87_4    : int  1 1 2 1 1 1 2 2 1 4 ...
##  $ Esteem87_5    : int  2 4 4 4 4 4 4 3 4 4 ...
##  $ Esteem87_6    : int  2 1 2 2 1 1 2 2 1 1 ...
##  $ Esteem87_7    : int  2 2 2 1 1 2 2 2 2 1 ...
##  $ Esteem87_8    : int  3 3 4 2 4 4 4 3 4 3 ...
##  $ Esteem87_9    : int  3 2 3 2 4 4 3 3 3 4 ...
##  $ Esteem87_10   : int  4 4 4 2 4 4 4 3 4 4 ...
```

```
##     Subject         Gender           Education05      Income87
##  Min.   :    2   Length:2431        Min.   : 6.0   Min.   :   -2
##  1st Qu.: 1592   Class :character   1st Qu.:12.0   1st Qu.: 4500
##  Median : 3137   Mode  :character   Median :13.0   Median :12000
##  Mean   : 3504                      Mean   :13.9   Mean   :13399
##  3rd Qu.: 4668                      3rd Qu.:16.0   3rd Qu.:19000
##  Max.   :12140                      Max.   :20.0   Max.   :59387
##     Job05              Income05         Weight05    HeightFeet05
##  Length:2431        Min.   :    63   Min.   : 81   Min.   :-4.00
##  Class :character   1st Qu.: 22650   1st Qu.:150   1st Qu.: 5.00
##  Mode  :character   Median : 38500   Median :180   Median : 5.00
##                     Mean   : 49415   Mean   :183   Mean   : 5.18
##                     3rd Qu.: 61350   3rd Qu.:209   3rd Qu.: 5.00
##                     Max.   :703637   Max.   :380   Max.   : 8.00
##   HeightInch05     Imagazine       Inewspaper       Ilibrary       MotherEd
##  Min.   : 0.00   Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   : 0.0
##  1st Qu.: 2.00   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:11.0
##  Median : 5.00   Median :1.000   Median :1.000   Median :1.00   Median :12.0
##  Mean   : 5.32   Mean   :0.718   Mean   :0.861   Mean   :0.77   Mean   :11.7
##  3rd Qu.: 8.00   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.00   3rd Qu.:12.0
##  Max.   :11.00   Max.   :1.000   Max.   :1.000   Max.   :1.00   Max.   :20.0
##     FatherEd    FamilyIncome78     Science         Arith           Word
##  Min.   : 0.0   Min.   :    0   Min.   : 0.0   Min.   : 0.0   Min.   : 0.0
##  1st Qu.:10.0   1st Qu.:11167   1st Qu.:13.0   1st Qu.:13.0   1st Qu.:23.0
##  Median :12.0   Median :20000   Median :17.0   Median :19.0   Median :28.0
##  Mean   :11.8   Mean   :21252   Mean   :16.3   Mean   :18.6   Mean   :26.6
##  3rd Qu.:14.0   3rd Qu.:27500   3rd Qu.:20.0   3rd Qu.:25.0   3rd Qu.:32.0
##  Max.   :20.0   Max.   :75001   Max.   :25.0   Max.   :30.0   Max.   :35.0
##      Parag          Number         Coding          Auto           Math
##  Min.   : 0.0   Min.   : 0.0   Min.   : 0.0   Min.   : 0.0   Min.   : 0.0
##  1st Qu.:10.0   1st Qu.:29.0   1st Qu.:38.0   1st Qu.:10.0   1st Qu.: 9.0
##  Median :12.0   Median :36.0   Median :48.0   Median :14.0   Median :14.0
##  Mean   :11.2   Mean   :35.5   Mean   :47.1   Mean   :14.3   Mean   :14.3
##  3rd Qu.:14.0   3rd Qu.:44.0   3rd Qu.:57.0   3rd Qu.:18.0   3rd Qu.:20.0
##  Max.   :15.0   Max.   :50.0   Max.   :84.0   Max.   :25.0   Max.   :25.0
##     Mechanic         Elec           AFQT         Esteem81_1     Esteem81_2
##  Min.   : 0.0   Min.   : 0.0   Min.   :  0.0   Min.   :1.00   Min.   :1.00
##  1st Qu.:11.0   1st Qu.: 9.0   1st Qu.: 31.9   1st Qu.:1.00   1st Qu.:1.00
##  Median :14.0   Median :12.0   Median : 57.0   Median :1.00   Median :1.00
##  Mean   :14.4   Mean   :11.6   Mean   : 54.7   Mean   :1.42   Mean   :1.42
##  3rd Qu.:18.0   3rd Qu.:15.0   3rd Qu.: 78.2   3rd Qu.:2.00   3rd Qu.:2.00
##  Max.   :25.0   Max.   :20.0   Max.   :100.0   Max.   :4.00   Max.   :4.00
##    Esteem81_3     Esteem81_4     Esteem81_5     Esteem81_6     Esteem81_7
##  Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00
##  1st Qu.:3.00   1st Qu.:1.00   1st Qu.:3.00   1st Qu.:1.00   1st Qu.:1.00
```

```
##  Median :4.00   Median :2.00   Median :4.00   Median :2.00   Median :2.00
##  Mean   :3.51   Mean   :1.57   Mean   :3.46   Mean   :1.62   Mean   :1.75
##  3rd Qu.:4.00   3rd Qu.:2.00   3rd Qu.:4.00   3rd Qu.:2.00   3rd Qu.:2.00
##  Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00
##    Esteem81_8     Esteem81_9    Esteem81_10    Esteem87_1     Esteem87_2
##  Min.   :1.00   Min.   :1.00   Min.   :1.0   Min.   :1.00   Min.   :1.0
##  1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.0   1st Qu.:1.00   1st Qu.:1.0
##  Median :3.00   Median :3.00   Median :3.0   Median :1.00   Median :1.0
##  Mean   :3.13   Mean   :3.16   Mean   :3.4   Mean   :1.38   Mean   :1.4
##  3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.0   3rd Qu.:2.00   3rd Qu.:2.0
##  Max.   :4.00   Max.   :4.00   Max.   :4.0   Max.   :4.00   Max.   :4.0
##    Esteem87_3     Esteem87_4    Esteem87_5     Esteem87_6     Esteem87_7
##  Min.   :1.00   Min.   :1.0   Min.   :1.00   Min.   :1.00   Min.   :1.00
##  1st Qu.:3.00   1st Qu.:1.0   1st Qu.:3.00   1st Qu.:1.00   1st Qu.:1.00
##  Median :4.00   Median :1.0   Median :4.00   Median :2.00   Median :2.00
##  Mean   :3.58   Mean   :1.5   Mean   :3.53   Mean   :1.59   Mean   :1.72
##  3rd Qu.:4.00   3rd Qu.:2.0   3rd Qu.:4.00   3rd Qu.:2.00   3rd Qu.:2.00
##  Max.   :4.00   Max.   :4.0   Max.   :4.00   Max.   :4.00   Max.   :4.00
##    Esteem87_8    Esteem87_9    Esteem87_10
##  Min.   :1.0   Min.   :1.00   Min.   :1.00
##  1st Qu.:3.0   1st Qu.:3.00   1st Qu.:3.00
##  Median :3.0   Median :3.00   Median :3.00
##  Mean   :3.1   Mean   :3.06   Mean   :3.37
##  3rd Qu.:4.0   3rd Qu.:4.00   3rd Qu.:4.00
##  Max.   :4.0   Max.   :4.00   Max.   :4.00
##
##
##                                                                                  56
##                      10 TO 430: Executive, Administrative and Managerial Occupations
##                                                                                 377
##                                    1000 TO 1240: Mathematical and Computer Scientists
##                                                                                  64
## 1300 TO 1560: Engineers, Architects, Surveyers, Engineering and Related Technicians
##                                                                                  53
##                                                      1600 TO 1760: Physical Scientists
##                                                                                   4
##                                    1800 TO 1860: Social Scientists and Related Workers
##                                                                                   6
##                           1900 TO 1960: Life, Physical and Social Science Technicians
##                                                                                   7
##                               2000 TO 2060: Counselors, Sociala and Religious Workers
##                                                                                  41
##                               2100 TO 2150: Lawyers, Judges and Legal Support Workers
##                                                                                  15
##                                                                 2200 TO 2340: Teachers
##                                                                                 120
##                                 2400 TO 2550: Education, Training and Library Workers
##                                                                                  29
##                  2600 TO 2760: Entertainers and Performers, Sports and Related Workers
##                                                                                  24
##                                       2800 TO 2960: Media and Communications Workers
##                                                                                  13
##                            3000 TO 3260: Health Diagnosing and Treating Practitioners
```

```
##                                                                    74
##                       3300 TO 3650: Health Care Technical and Support Occupations
##                                                                    99
##                                    3700 TO 3950: Protective Service Occupations
##                                                                    54
##                     4000 TO 4160: Food Preparation and Serving Related Occupations
##                                                                    68
##                          4200 TO 4250: Cleaning and Building Service Occupations
##                                                                    67
##                         4300 TO 4430: Entertainment Attendants and Related Workers
##                                                                    10
##                                   4500 TO 4650: Personal Care and Service Workers
##                                                                    42
##                                       4700 TO 4960: Sales and Related Workers
##                                                                   205
##                                     500 TO 950: Management Related Occupations
##                                                                   108
##                          5000 TO 5930: Office and Administrative Support Workers
##                                                                   360
##                          6000 TO 6130: Farming, Fishing and Forestry Occupations
##                                                                     9
##                          6200 TO 6940: Construction Trade and Extraction Workers
##                                                                   135
##                       7000 TO 7620: Installation, Maintenance and Repairs Workers
##                                                                   108
##                                   7700 TO 7750: Production and Operating Workers
##                                                                    49
##                                     7800 TO 7850: Food Preparation Occupations
##                                                                     4
##                                    7900 TO 8960: Setters, Operators and Tenders
##                                                                   112
##                        9000 TO 9750: Transportation and Material Moving Workers
##                                                                   117
##                                                          9990: Uncodeable
##                                                                     1
```

1.2 Self esteem evaluation

Let concentrate on Esteem scores evaluated in 87.

0. First do a quick summary over all the `Esteem` variables. Pay attention to missing values, any peculiar numbers etc. How do you fix problems discovered if there is any? Briefly describe what you have done for the data preparation.

We extracted the columns that began with "Esteem87" and then ran summary stats for those columsn. We then checked for NA values; we saw that there are no missing values. For all Esteem87 columns, the min is 1 and the max is 4, which is expected. So there are no issues with these variables.

```
##    Esteem87_1     Esteem87_2     Esteem87_3     Esteem87_4     Esteem87_5
##  Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00
##  1st Qu.:1.00   1st Qu.:1.00   1st Qu.:3.00   1st Qu.:1.00   1st Qu.:3.00
##  Median :1.00   Median :1.00   Median :4.00   Median :1.00   Median :4.00
##  Mean   :1.37   Mean   :1.39   Mean   :3.59   Mean   :1.49   Mean   :3.55
##  3rd Qu.:2.00   3rd Qu.:2.00   3rd Qu.:4.00   3rd Qu.:2.00   3rd Qu.:4.00
##  Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00
##    Esteem87_6     Esteem87_7     Esteem87_8     Esteem87_9    Esteem87_10
```

```
##  Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00
##  1st Qu.:1.00   1st Qu.:1.00   1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00
##  Median :2.00   Median :2.00   Median :3.00   Median :3.00   Median :3.00
##  Mean   :1.59   Mean   :1.71   Mean   :3.11   Mean   :3.08   Mean   :3.37
##  3rd Qu.:2.00   3rd Qu.:2.00   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00
##  Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00
```

```
##  [1] Esteem87_1  Esteem87_2  Esteem87_3  Esteem87_4  Esteem87_5  Esteem87_6
##  [7] Esteem87_7  Esteem87_8  Esteem87_9  Esteem87_10
## <0 rows> (or 0-length row.names)
```

1. Please note that higher scores on Esteem questions 1, 2, 4, 6, and 7 indicate higher self-esteem, whereas higher scores on the remaining questions suggest lower self-esteem. To maintain consistency, consider reversing the scores of certain Esteem questions. For example, if the esteem data is stored in `data.esteem`, you can use the code `data.esteem[, c(1, 2, 4, 6, 7)] <- 5 - data.esteem[, c(1, 2, 4, 6, 7)]` to invert the scores.

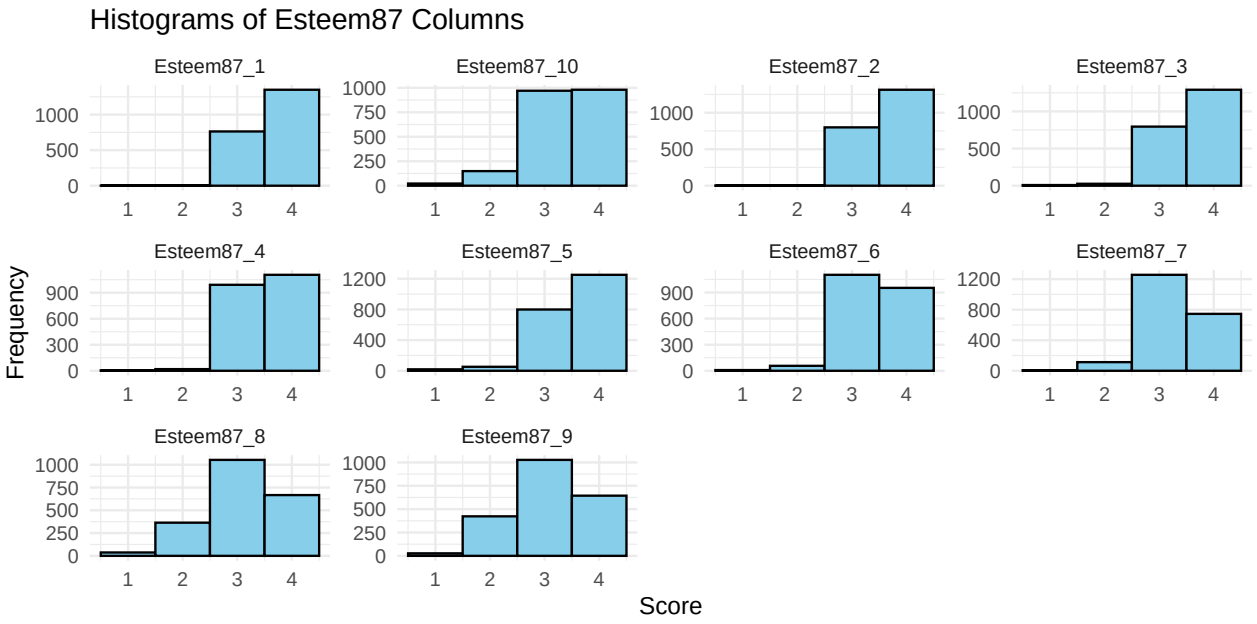
Now, for all Esteem87 columns, a lower score implies higher self esteem.

```
##    Esteem87_1     Esteem87_2     Esteem87_3     Esteem87_4     Esteem87_5
##  Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00
##  1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00
##  Median :4.00   Median :4.00   Median :4.00   Median :4.00   Median :4.00
##  Mean   :3.63   Mean   :3.61   Mean   :3.59   Mean   :3.51   Mean   :3.55
##  3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00
##  Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00
##    Esteem87_6     Esteem87_7     Esteem87_8     Esteem87_9    Esteem87_10
##  Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00
##  1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00
##  Median :3.00   Median :3.00   Median :3.00   Median :3.00   Median :3.00
##  Mean   :3.41   Mean   :3.29   Mean   :3.11   Mean   :3.08   Mean   :3.37
##  3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00
##  Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00
```

2. Write a brief summary with necessary plots about the 10 esteem measurements.

Esteem87_1 through Esteem87_5 all have a median of 4. And the rest have a median of 3. They all have a min of 1 and a max of 4. The mean valyes are in the summary below. We created a histogram for all 10 measurements.

```
##    Esteem87_1     Esteem87_2     Esteem87_3     Esteem87_4     Esteem87_5
##  Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00
##  1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00
##  Median :4.00   Median :4.00   Median :4.00   Median :4.00   Median :4.00
##  Mean   :3.63   Mean   :3.61   Mean   :3.59   Mean   :3.51   Mean   :3.55
##  3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00
##  Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00
##    Esteem87_6     Esteem87_7     Esteem87_8     Esteem87_9    Esteem87_10
##  Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00
##  1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.00
##  Median :3.00   Median :3.00   Median :3.00   Median :3.00   Median :3.00
##  Mean   :3.41   Mean   :3.29   Mean   :3.11   Mean   :3.08   Mean   :3.37
##  3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00
##  Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00   Max.   :4.00
```



3. Do esteem scores all positively correlated? Report the pairwise correlation table and write a brief summary.

From the pairwise correlation table, all scores are positively correlated. This implies that for all scores, one score increasing is correlated to another score increasing. The same goes if they are both decreasing.

```
##             Esteem87_1 Esteem87_2 Esteem87_3 Esteem87_4 Esteem87_5 Esteem87_6
## Esteem87_1       1.000      0.708      0.443      0.517      0.389      0.460
## Esteem87_2       0.708      1.000      0.447      0.545      0.402      0.484
## Esteem87_3       0.443      0.447      1.000      0.400      0.554      0.418
## Esteem87_4       0.517      0.545      0.400      1.000      0.368      0.494
## Esteem87_5       0.389      0.402      0.554      0.368      1.000      0.411
## Esteem87_6       0.460      0.484      0.418      0.494      0.411      1.000
## Esteem87_7       0.375      0.410      0.350      0.416      0.378      0.610
## Esteem87_8       0.270      0.281      0.353      0.298      0.375      0.423
## Esteem87_9       0.234      0.252      0.354      0.279      0.351      0.372
## Esteem87_10      0.299      0.327      0.463      0.354      0.444      0.443
##             Esteem87_7 Esteem87_8 Esteem87_9 Esteem87_10
## Esteem87_1       0.375      0.270      0.234       0.299
## Esteem87_2       0.410      0.281      0.252       0.327
## Esteem87_3       0.350      0.353      0.354       0.463
## Esteem87_4       0.416      0.298      0.279       0.354
## Esteem87_5       0.378      0.375      0.351       0.444
## Esteem87_6       0.610      0.423      0.372       0.443
## Esteem87_7       1.000      0.401      0.360       0.392
## Esteem87_8       0.401      1.000      0.432       0.440
## Esteem87_9       0.360      0.432      1.000       0.581
## Esteem87_10      0.392      0.440      0.581       1.000
```

4. PCA on 10 esteem measurements. (centered but no scaling)

    a) Report the PC1 and PC2 loadings. Are they unit vectors? Are they orthogonal?

We extracted the first two columns from the rotation part of prcomp to get PC1 and PC2 loadings. We then checked if they are unit vectors by checking their lengths. They are both length of 1, so they are unit vectors. The dot product of the two columns is very close to 0, so they are considered to be orthogonal.

```
##                PC1    PC2
## Esteem87_1   0.229 -0.379
## Esteem87_2   0.240 -0.378
## Esteem87_3   0.280 -0.152
## Esteem87_4   0.254 -0.324
## Esteem87_5   0.307 -0.135
## Esteem87_6   0.317 -0.206
## Esteem87_7   0.300 -0.157
## Esteem87_8   0.396  0.307
## Esteem87_9   0.400  0.577
## Esteem87_10 0.381  0.271
```

```
## PC1 PC2
##   1   1
```

```
## [1] -4.16e-17
```

```
b) Are there good interpretations for PC1 and PC2? (If loadings are all negative, take the positive loa
```

PC1 is the weighted sum of all Esteem87 values since all loadings are positive. PC2 represents ((the weighted sum of Esteem87_8 through Esteem87_10) - (the weighted sum of Esteem87_1 through Esteem87_7)) since Esteem87_8 through Esteem87_10 loadings are positive and the others' loadings are negative.


```
##                PC1    PC2
## Esteem87_1   0.229 -0.379
## Esteem87_2   0.240 -0.378
## Esteem87_3   0.280 -0.152
## Esteem87_4   0.254 -0.324
## Esteem87_5   0.307 -0.135
## Esteem87_6   0.317 -0.206
## Esteem87_7   0.300 -0.157
## Esteem87_8   0.396  0.307
## Esteem87_9   0.400  0.577
## Esteem87_10 0.381  0.271
```


```
c) How is the PC1 score obtained for each subject? Write down the formula.
```

PC1 = (Esteem87_1 * 0.229) + (Esteem87_2 * 0.240) + (Esteem87_3 * 0.280) + (Esteem87_4 * 0.254) + (Esteem87_5 * 0.307) + (Esteem87_6 * 0.317) + (Esteem87_7 * 0.300) + (Esteem87_8 * 0.396) + (Esteem87_9 * 0.400) + (Esteem87_10 * 0.381)

```
d) Are PC1 scores and PC2 scores in the data uncorrelated?
```

Yes, they are uncorrelated. In the x section of prcomp, we ran a correlation table. We found that the cor between PC1 and PC2 scores is near 0 (7.58e-16), meaning they are uncorrelated.

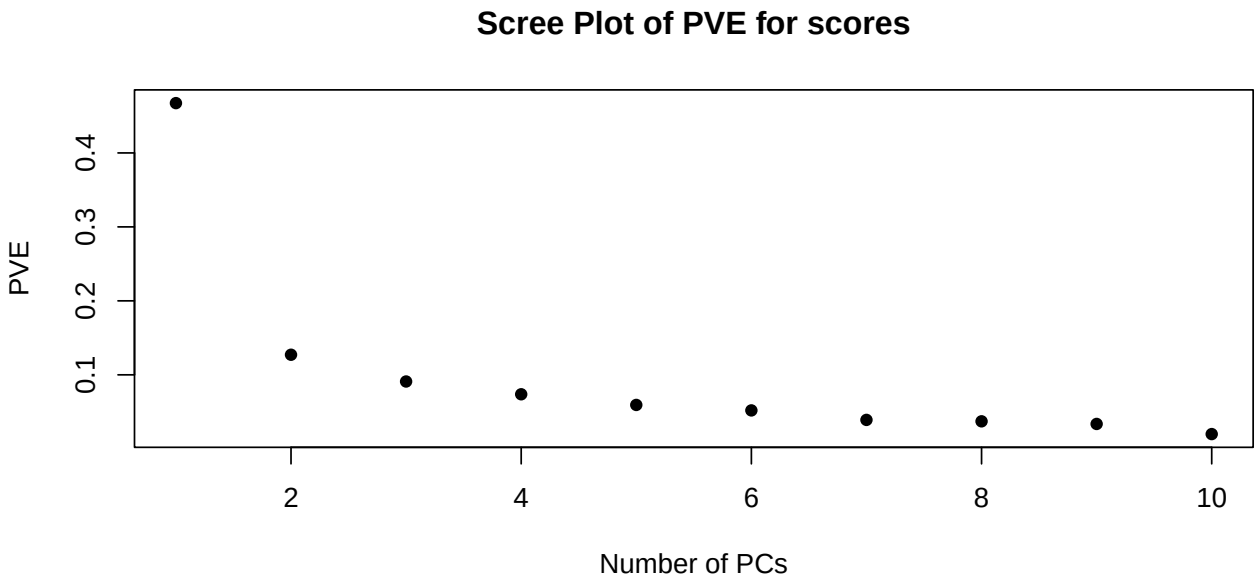
```
##             PC1       PC2       PC3       PC4       PC5       PC6       PC7
## PC1    1.00e+00  1.16e-15 -7.67e-16 -1.31e-15  7.94e-15 -6.16e-16  8.49e-15
## PC2    1.16e-15  1.00e+00 -3.08e-15  4.87e-15 -8.04e-15  1.84e-16 -7.09e-15
## PC3   -7.67e-16 -3.08e-15  1.00e+00  6.87e-16  2.25e-15  2.45e-15  1.75e-15
## PC4   -1.31e-15  4.87e-15  6.87e-16  1.00e+00  2.92e-16 -1.48e-15  6.94e-15
## PC5    7.94e-15 -8.04e-15  2.25e-15  2.92e-16  1.00e+00 -4.92e-15  6.40e-15
## PC6   -6.16e-16  1.84e-16  2.45e-15 -1.48e-15 -4.92e-15  1.00e+00  8.63e-15
## PC7    8.49e-15 -7.09e-15  1.75e-15  6.94e-15  6.40e-15  8.63e-15  1.00e+00
## PC8   -3.40e-15 -9.85e-16  3.50e-16 -4.25e-15 -7.77e-15 -5.99e-15  8.96e-15
## PC9    1.38e-14 -1.45e-14  2.74e-15  3.06e-16  6.76e-15 -2.31e-15  2.41e-15
## PC10   1.23e-14 -1.72e-14  3.52e-15  1.08e-15  2.01e-14 -4.18e-16 -1.57e-14
##             PC8       PC9      PC10
## PC1   -3.40e-15  1.38e-14  1.23e-14
```

```
## PC2  -9.85e-16 -1.45e-14 -1.72e-14
## PC3   3.50e-16  2.74e-15  3.52e-15
## PC4  -4.25e-15  3.06e-16  1.08e-15
## PC5  -7.77e-15  6.76e-15  2.01e-14
## PC6  -5.99e-15 -2.31e-15 -4.18e-16
## PC7   8.96e-15  2.41e-15 -1.57e-14
## PC8   1.00e+00 -1.11e-14 -2.02e-14
## PC9  -1.11e-14  1.00e+00  4.05e-16
## PC10 -2.02e-14  4.05e-16  1.00e+00
```

`e) Plot PVE (Proportion of Variance Explained) and summarize the plot.`

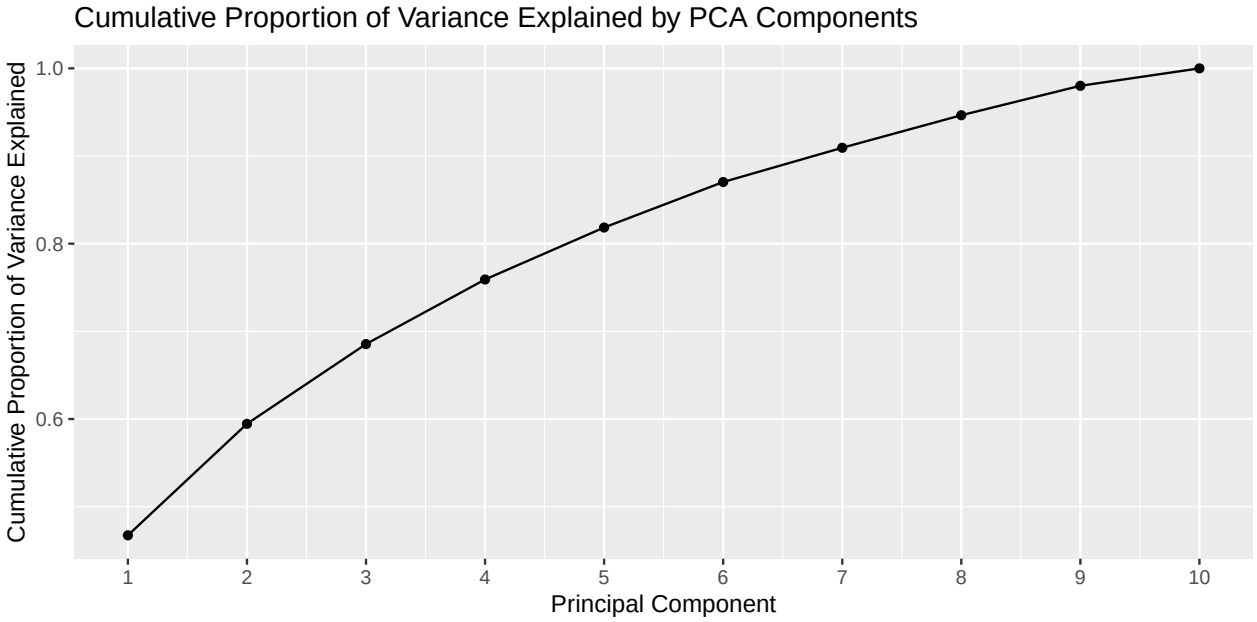
By plotting the PVE for each PC, we see that the PVEs exponentially decrease as the PCs increase, with the elbow at PC2. We can only consider PC1 and PC2 to capture the majority of the data in 2 dimensions.

```
##         pve
## PC1  0.4673
## PC2  0.1272
## PC3  0.0910
## PC4  0.0737
## PC5  0.0593
## PC6  0.0519
## PC7  0.0391
## PC8  0.0371
## PC9  0.0336
## PC10 0.0199
```



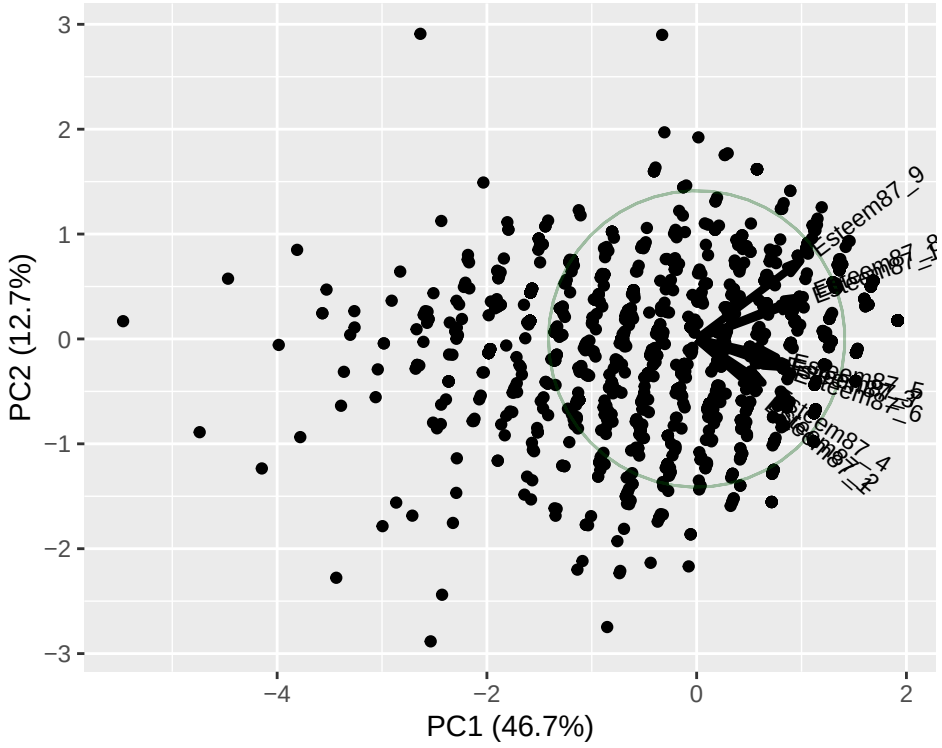
`f) Also plot CPVE (Cumulative Proportion of Variance Explained). What proportion of the variance in the`

About 60% of the variance is explained by the first two principal components. (Note that the y-axis of this visualization doesn't start at 0).



g) PC's provide us with a low dimensional view of the self-esteem scores. Use a biplot with the first tw

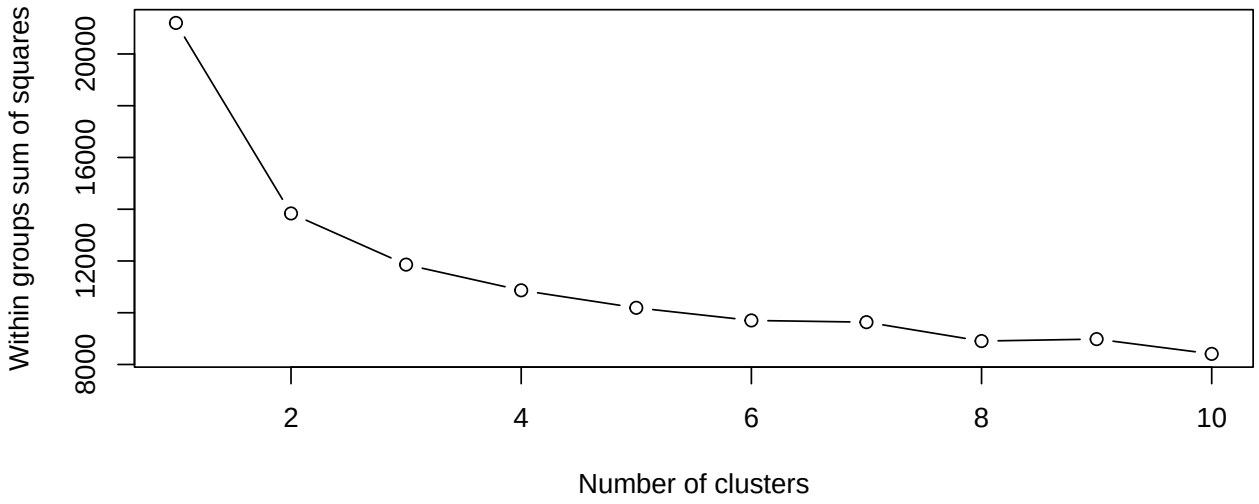
PC2 splits the variables into two groups, one with Esteem87_8, 9, 10 and one with the rest; this indicates that there is a group of questions that is negatively correlated with PC2 and one that is positively correlated. PC1 is highly correlated with all esteem questions.



5. Apply k-means to cluster subjects on the original esteem scores

    a) Find a reasonable number of clusters using within sum of squared with elbow rules.

We would say that the elbow in the data with regards to WSS can be seen at 3 clusters.



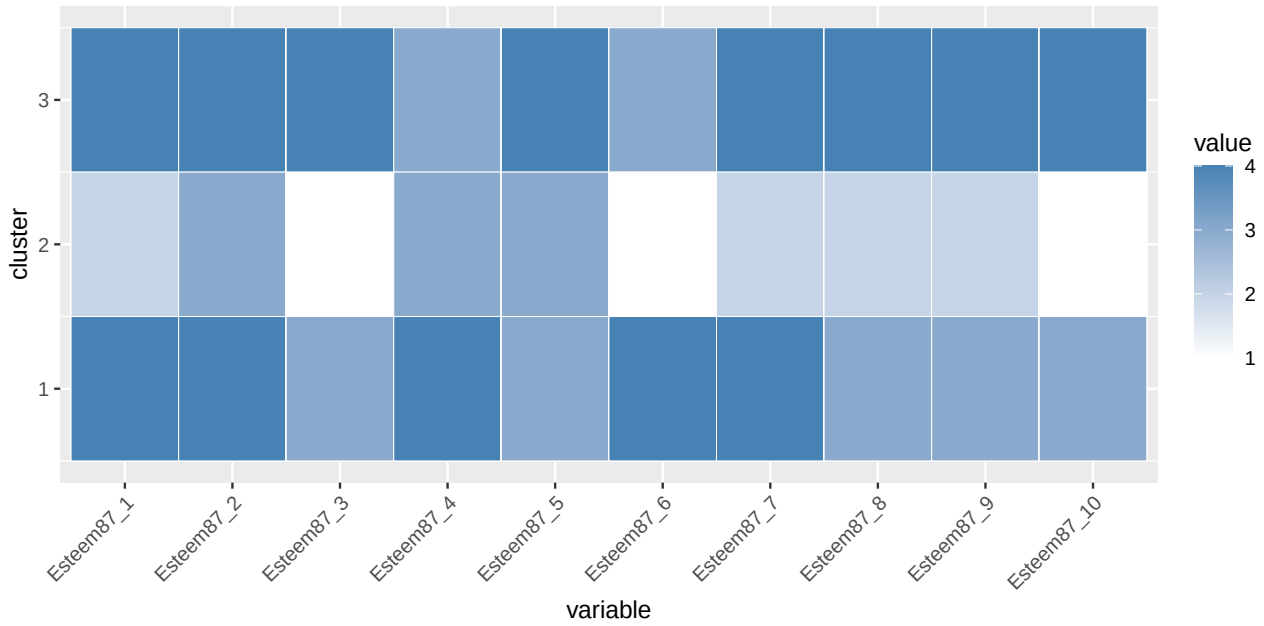

```
##   Esteem87_1 Esteem87_2 Esteem87_3 Esteem87_4 Esteem87_5 Esteem87_6 Esteem87_7
## 1      0.573      0.538     0.0939      0.120    -0.0594     -0.125     -0.251
## 2     -1.142     -1.168    -0.8062     -0.868    -0.7204     -0.780     -0.638
## 3      0.562      0.617     0.6649      0.699     0.7177      0.829      0.807
##   Esteem87_8 Esteem87_9 Esteem87_10
## 1     -0.462     -0.395      -0.336
## 2     -0.450     -0.459      -0.557
## 3      0.815      0.765       0.805
```

b) Can you summarize common features within each cluster?

For each cluster we found the average esteem question response (among each of the 10 questions) and then plotted a heatmap with the resulting data. We can observe that cluster 3 contains people that responded highly to most of the esteem questions, followed by those in cluster 1, and finally those in cluster 2 with the lowest average ratings. Thus, it seems that clusters 3 and 1 contains people with relatively low self-esteem, and cluster 2 contains high self-esteem respondents. In particular, cluster 1 has high answer averages for questions 1, 2, 4, 6, and 7, whereas cluster 3 has higher average responses for 1, 2, 3, 5, and 7 thorugh 10.

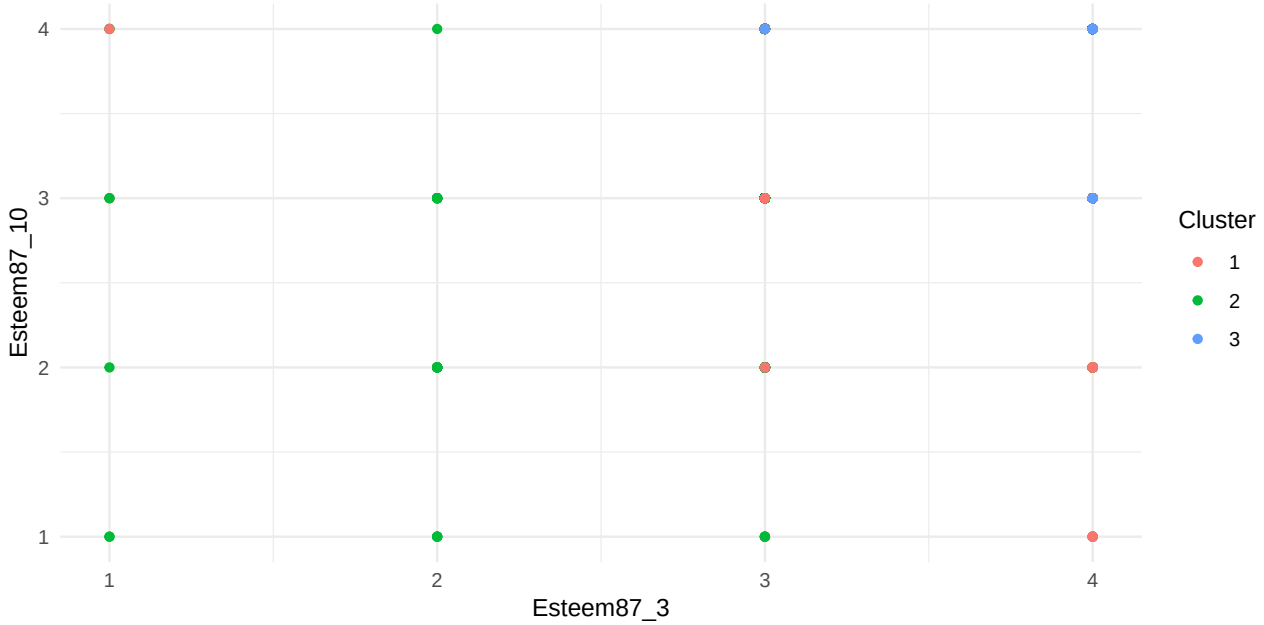
```
##   cluster Esteem87_1 Esteem87_2 Esteem87_3 Esteem87_4 Esteem87_5 Esteem87_6
## 1       1       3.91       3.88       3.64       3.57       3.51       3.34
## 2       2       3.07       3.03       3.15       3.05       3.12       2.97
## 3       3       3.91       3.92       3.95       3.88       3.97       3.88
##   Esteem87_7 Esteem87_8 Esteem87_9 Esteem87_10
## 1       3.15       2.77       2.79        3.15
## 2       2.92       2.78       2.74        3.00
## 3       3.76       3.71       3.65        3.90
```

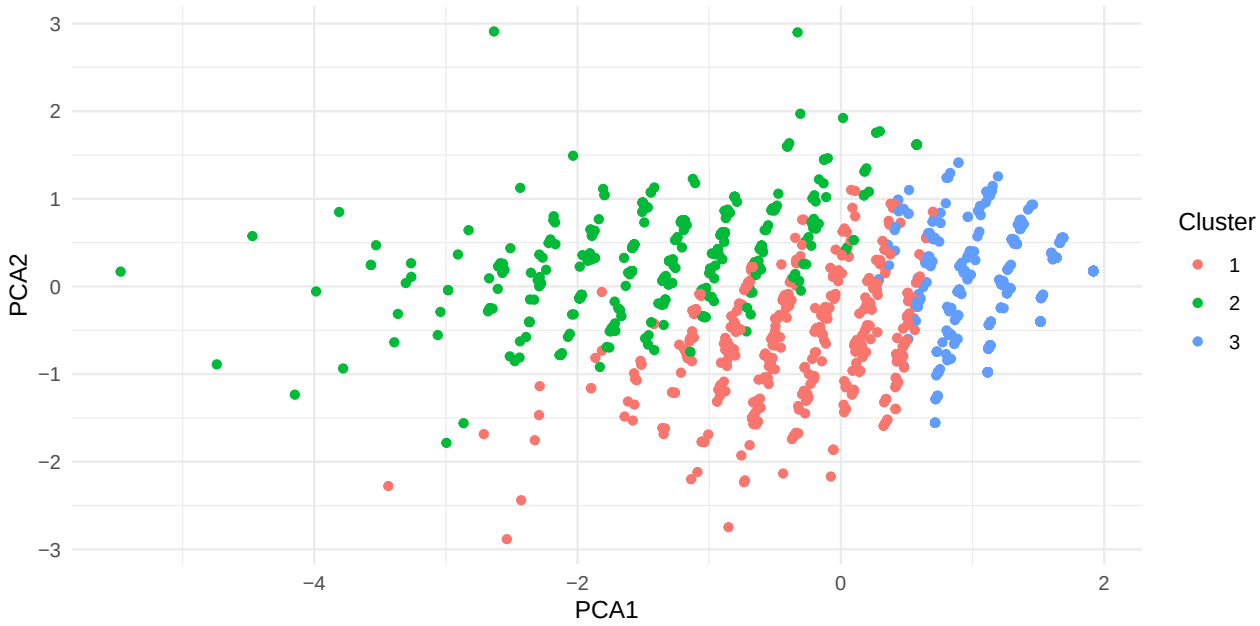
```
## Warning in melt.data.table(data.table(clustered_df), id.vars = "cluster"):
## 'measure.vars' [Esteem87_1, Esteem87_2, Esteem87_3, Esteem87_4, ...] are not
## all of the same type. By order of hierarchy, the molten data value column will
## be of type 'double'. All measure variables not of type 'double' will be coerced
## too. Check DETAILS in ?melt.data.table for more on coercion.
```



c) Can you visualize the clusters with somewhat clear boundaries? You may try different pairs of variabl

Using the two variables with the most drastic between-cluster color difference from the heatmap above, we can see a low-resolution separation between the three clusters in our first plot. Using the two principal components with the most explained variance, we see a higher resolution visualization of the three clusters (though without clear boundaries). We can see that cluster 1 generally has the highest PC1 values, followed by cluster 3, then cluster 2. Among the PC2 values, we see that cluster 3 tends to have lower PC2 values. We also see that the red dots, cluster 1, seems to be more clumped have reduced within SS (less intra-cluster distance from its center), whereas cluster 2 seems to be more spread apart and have more intra-cluster distance from its center.





6. We now try to find out what factors are related to self-esteem? PC1 of all the Esteem scores is a good variable to summarize one's esteem scores. We take PC1 as our response variable.

   a) Prepare possible factors/variables:

   - EDA the data set first.

```
##    pca$x[, 1]
##  Min.   :-5.47
##  1st Qu.:-1.14
##  Median :-0.06
##  Mean   : 0.00
##  3rd Qu.: 1.12
##  Max.   : 1.92
```

```
##  [1] "Subject"        "Gender"         "Education05"    "Income87"
##  [5] "Job05"          "Income05"       "Weight05"       "HeightFeet05"
##  [9] "HeightInch05"   "Imagazine"      "Inewspaper"     "Ilibrary"
## [13] "MotherEd"       "FatherEd"       "FamilyIncome78" "Science"
## [17] "Arith"          "Word"           "Parag"          "Number"
## [21] "Coding"         "Auto"           "Math"           "Mechanic"
## [25] "Elec"           "AFQT"
```

```
##     Subject         Gender      Education05    Income87
##  Min.   :    2   female:1008   Min.   : 6   Min.   : 3.91
##  1st Qu.: 1560   male  :1113   1st Qu.:12   1st Qu.: 8.99
##  Median : 3100                 Median :13   Median : 9.55
##  Mean   : 3457                 Mean   :14   Mean   : 9.30
##  3rd Qu.: 4596                 3rd Qu.:16   3rd Qu.: 9.90
##  Max.   :12140                 Max.   :20   Max.   :10.99
##
##                                                          Job05
##  10 TO 430: Executive, Administrative and Managerial Occupations:337
##  5000 TO 5930: Office and Administrative Support Workers        :323
##  4700 TO 4960: Sales and Related Workers                        :185
##  6200 TO 6940: Construction Trade and Extraction Workers        :111
```

```
##  2200 TO 2340: Teachers                                          :105
##  9000 TO 9750: Transportation and Material Moving Workers        :104
##  (Other)                                                         :956
##     Income05        Weight05    HeightFeet05   HeightInch05   Imagazine
##  Min.   : 4.14   Min.   : 82   Min.   :2.00   Min.   : 0.00   0: 569
##  1st Qu.:10.13   1st Qu.:152   1st Qu.:5.00   1st Qu.: 2.00   1:1552
##  Median :10.60   Median :180   Median :5.00   Median : 5.00
##  Mean   :10.50   Mean   :184   Mean   :5.21   Mean   : 5.32
##  3rd Qu.:11.07   3rd Qu.:210   3rd Qu.:5.00   3rd Qu.: 8.00
##  Max.   :13.46   Max.   :380   Max.   :8.00   Max.   :11.00
##
##  Inewspaper Ilibrary    MotherEd       FatherEd  FamilyIncome78     Science
##  0: 285     0: 482   Min.   : 0.0   Min.   : 0   Min.   :    0   Min.   : 2.0
##  1:1836     1:1639   1st Qu.:11.0   1st Qu.:10   1st Qu.:12000   1st Qu.:13.0
##                      Median :12.0   Median :12   Median :20000   Median :17.0
##                      Mean   :11.8   Mean   :12   Mean   :21789   Mean   :16.5
##                      3rd Qu.:12.0   3rd Qu.:14   3rd Qu.:28000   3rd Qu.:20.0
##                      Max.   :20.0   Max.   :20   Max.   :75001   Max.   :25.0
##
##      Arith           Word          Parag          Number       Coding
##  Min.   : 3.0   Min.   : 4.0   Min.   : 0.0   Min.   : 3   Min.   : 2.0
##  1st Qu.:13.0   1st Qu.:23.0   1st Qu.:10.0   1st Qu.:29   1st Qu.:38.0
##  Median :19.0   Median :29.0   Median :12.0   Median :37   Median :49.0
##  Mean   :18.9   Mean   :26.9   Mean   :11.4   Mean   :36   Mean   :47.6
##  3rd Qu.:25.0   3rd Qu.:32.0   3rd Qu.:14.0   3rd Qu.:44   3rd Qu.:58.0
##  Max.   :30.0   Max.   :35.0   Max.   :15.0   Max.   :50   Max.   :84.0
##
##       Auto           Math         Mechanic         Elec           AFQT
##  Min.   : 0.0   Min.   : 1.0   Min.   : 2.0   Min.   : 1.0   Min.   :  0.0
##  1st Qu.:10.0   1st Qu.: 9.0   1st Qu.:11.0   1st Qu.: 9.0   1st Qu.: 34.3
##  Median :14.0   Median :14.0   Median :15.0   Median :12.0   Median : 58.9
##  Mean   :14.5   Mean   :14.6   Mean   :14.7   Mean   :11.8   Mean   : 55.9
##  3rd Qu.:19.0   3rd Qu.:20.0   3rd Qu.:19.0   3rd Qu.:15.0   3rd Qu.: 78.7
##  Max.   :25.0   Max.   :25.0   Max.   :25.0   Max.   :20.0   Max.   :100.0
##
##    Esteem_PC1
##  Min.   :-5.47
##  1st Qu.:-1.14
##  Median :-0.06
##  Mean   : 0.00
##  3rd Qu.: 1.12
##  Max.   : 1.92
##
```

- Personal information: gender, education (05), log(income) in 87, job type in 87. One way to summari

- Household environment: Imagazine, Inewspaper, Ilibrary, MotherEd, FatherEd, FamilyIncome78. Do set

```
## 'data.frame':    2121 obs. of  28 variables:
##  $ Subject       : int  2 6 8 9 13 16 17 18 20 21 ...
##  $ Gender        : Factor w/ 2 levels "female","male": 1 2 1 2 2 1 2 2 1 1 ...
##  $ Education05   : int  12 16 14 14 16 13 13 13 17 16 ...
##  $ Income87      : num  9.68 9.8 9.1 9.62 7.7 ...
```

```
##  $ Job05         : Factor w/ 31 levels "","10 TO 430: Executive, Administrative and Managerial Occupa
##  $ Income05      : num  8.61 11.08 10.49 11.08 8.99 ...
##  $ Weight05      : int  160 187 246 180 235 160 188 173 130 150 ...
##  $ HeightFeet05  : int  5 5 5 5 6 5 5 5 5 5 ...
##  $ HeightInch05  : int  2 5 3 6 0 4 10 9 4 2 ...
##  $ Imagazine     : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 2 ...
##  $ Inewspaper    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Ilibrary      : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ MotherEd      : int  5 12 9 12 12 12 12 12 12 12 ...
##  $ FatherEd      : int  8 12 6 10 16 12 15 16 18 18 ...
##  $ FamilyIncome78: int  20000 35000 7227 17000 20000 48000 15000 4510 50000 40000 ...
##  $ Science       : int  6 23 18 17 16 13 19 22 21 15 ...
##  $ Arith         : int  8 30 13 21 30 17 29 30 17 29 ...
##  $ Word          : int  15 35 35 28 29 30 33 35 28 33 ...
##  $ Parag         : int  6 15 12 10 13 12 13 14 14 13 ...
##  $ Number        : int  29 45 24 40 36 49 35 48 39 49 ...
##  $ Coding        : int  52 68 48 46 30 58 58 61 54 64 ...
##  $ Auto          : int  9 21 11 13 21 11 18 21 18 19 ...
##  $ Math          : int  6 23 4 13 24 17 21 23 20 25 ...
##  $ Mechanic      : int  10 21 12 13 19 11 19 16 20 21 ...
##  $ Elec          : int  5 19 12 15 16 10 16 17 13 8 ...
##  $ AFQT          : num  6.84 99.39 44.02 59.68 72.31 ...
##  $ Esteem_PC1    : num  -0.34 0.42 -1.04 1.92 1.62 ...
##  $ BMI           : num  29.3 31.1 43.6 29.1 31.9 ...
```

- You may use PC1 of ASVAB as level of intelligence

```
##    pc1_asvab
##  Min.   :-42.2
##  1st Qu.:-13.8
##  Median : -1.5
##  Mean   :  0.0
##  3rd Qu.: 11.9
##  Max.   : 59.6
```

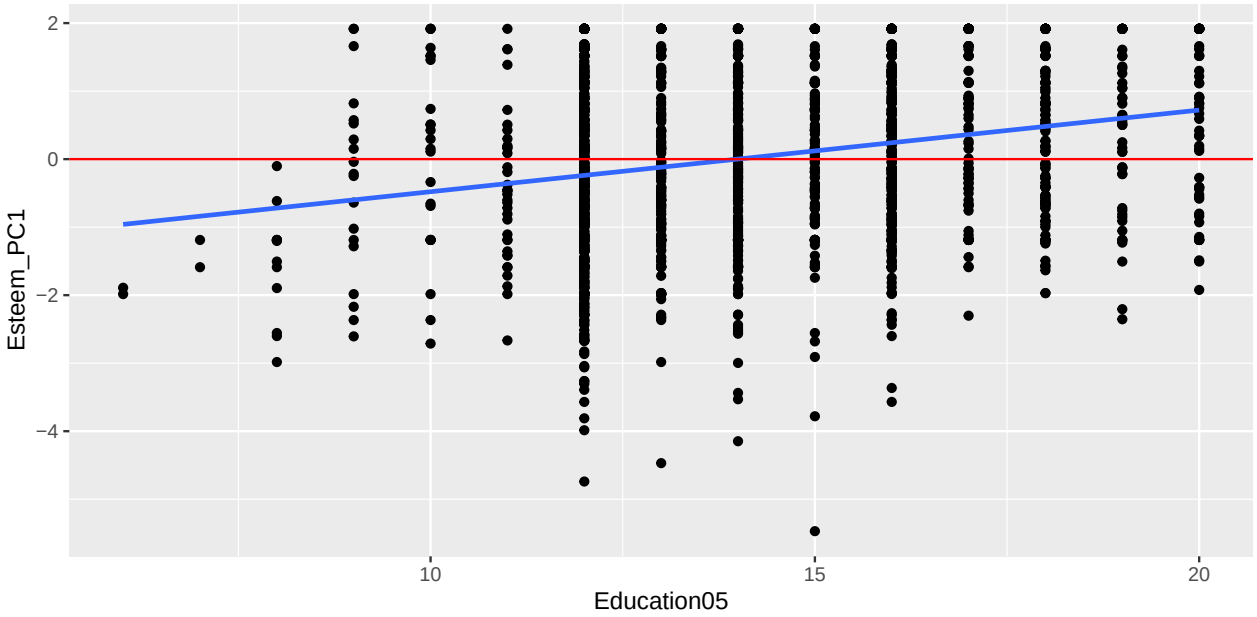
b) Run a few regression models between PC1 of all the esteem scores and suitable variables listed in a

Our intial thought is to find invidiual variables that have the largest effect on Esteem_PC1. We do indi

```
##
## Call:
## lm(formula = Esteem_PC1 ~ Education05, data = merged_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.591 -0.949 -0.025  1.051  2.515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.6796     0.1576   -10.7   <2e-16 ***
## Education05   0.1200     0.0111    10.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

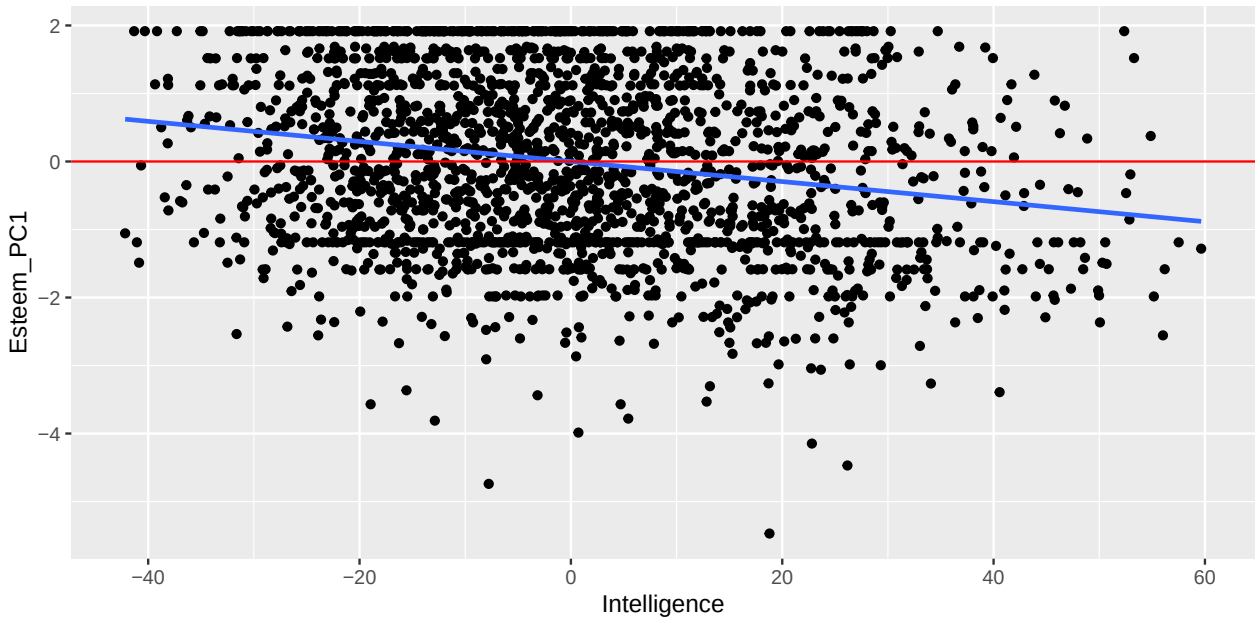
```
## Residual standard error: 1.26 on 2119 degrees of freedom
## Multiple R-squared:  0.0524, Adjusted R-squared:  0.0519
## F-statistic:  117 on 1 and 2119 DF,  p-value: <2e-16
```

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



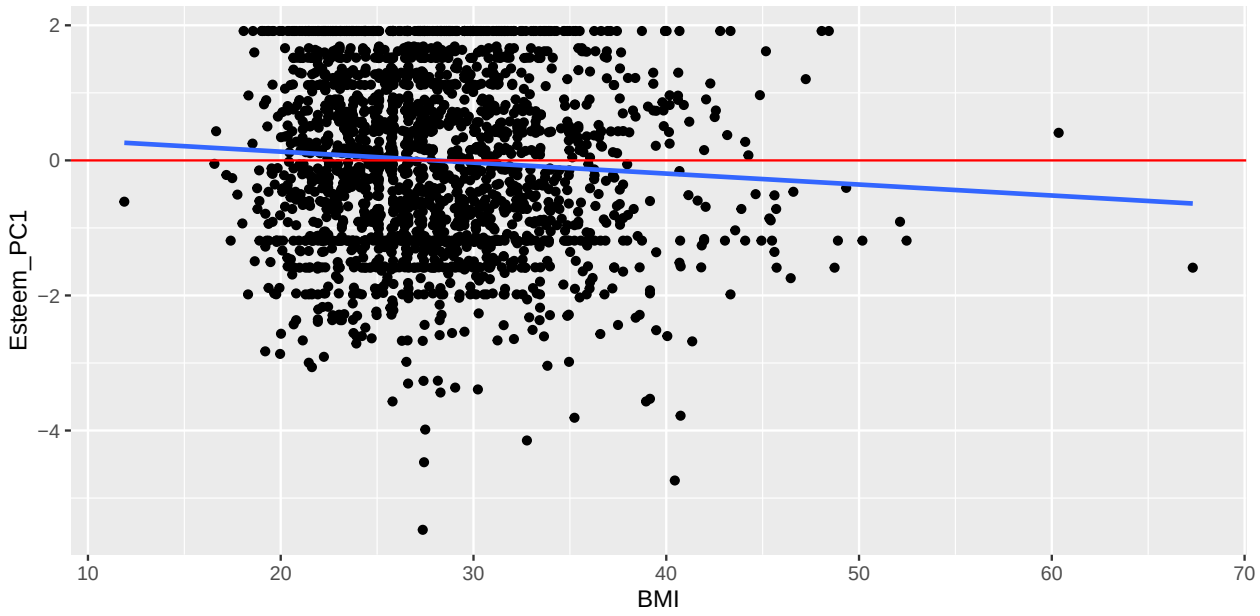

```
##
## Call:
## lm(formula = Esteem_PC1 ~ Intelligence, data = merged_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.193 -0.987 -0.033  1.069  2.689
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.16e-17   2.75e-02    0.00        1
## Intelligence -1.48e-02   1.49e-03   -9.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.27 on 2119 degrees of freedom
## Multiple R-squared:  0.0443, Adjusted R-squared:  0.0438
## F-statistic: 98.2 on 1 and 2119 DF,  p-value: <2e-16
```

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




```
##
## Call:
## lm(formula = Esteem_PC1 ~ BMI, data = merged_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.480 -1.076 -0.039  1.125  2.248
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.45368    0.14742    3.08   0.0021 **
## BMI         -0.01623    0.00517   -3.14   0.0017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.29 on 2118 degrees of freedom
## Multiple R-squared:  0.00463,    Adjusted R-squared:  0.00416
## F-statistic: 9.84 on 1 and 2118 DF,  p-value: 0.00173
```

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



We can see from above that that Education has the largest effect on esteem. The B-value for education has a p-value <2e-16 ***.

We now want to run multiple regressions on two different sets of variables that could effect esteem: 1) household variables, and 2) personal variables.

```
##
## Call:
## lm(formula = Esteem_PC1 ~ Imagazine + Inewspaper + Ilibrary +
##     MotherEd + FatherEd + FamilyIncome78, data = merged_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.520 -0.982 -0.039  1.079  2.627
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.25e+00   1.35e-01   -9.27   <2e-16 ***
## Imagazine1      1.37e-01   6.78e-02    2.03   0.0430 *
## Inewspaper1     2.46e-01   8.80e-02    2.79   0.0052 **
## Ilibrary1       1.03e-01   6.90e-02    1.49   0.1365
## MotherEd        4.40e-02   1.37e-02    3.20   0.0014 **
## FatherEd        1.99e-02   1.02e-02    1.96   0.0497 *
## FamilyIncome78  4.70e-06   2.12e-06    2.21   0.0270 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.27 on 2113 degrees of freedom
## Multiple R-squared:  0.0488, Adjusted R-squared:  0.0461
## F-statistic: 18.1 on 6 and 2113 DF,  p-value: <2e-16
```

We can see that Inewspaper1 and MotherEd have the lowest p-values from these variables. For this, we used an alpha cutoff os 0.01.

We can see from above that Intelligence, Education05, Income87, and Income 05 have the lowest p-values among these variables. This indicates that these factors have the largest impact on Esteem_PC1. We used

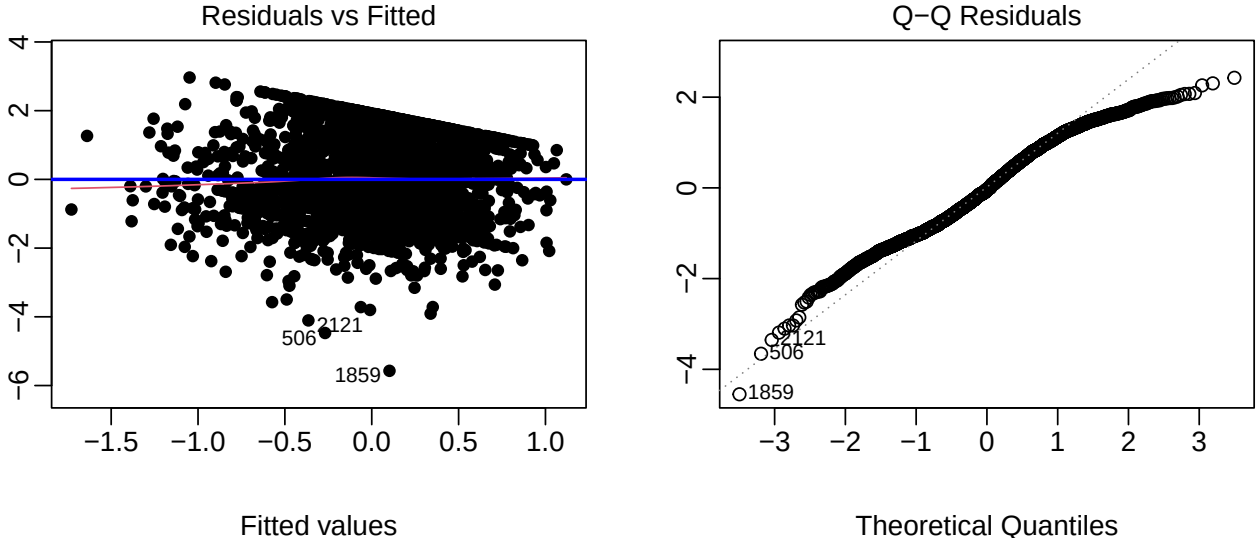
an alpha cutoff of 0.01. We did not include any job categories since their p-values across the categories vary so much.

```
- How did you land this model? Run a model diagnosis to see if the linear model assumptions are reasor
```

We can combine the most signficant variables from both multiple regressions above: Inewspaper1 and MotherEd, along with Intelligence, Education05, Income87, and Income 05. We create a final multiple regression fit with these 6 variables:

```
##
## Call:
## lm(formula = Esteem_PC1 ~ Inewspaper + MotherEd + Intelligence +
##     Education05 + Income87 + Income05, data = merged_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.572 -0.957 -0.054  1.002  2.965
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.63291    0.38718  -11.97  < 2e-16 ***
## Inewspaper1   0.19554    0.08332    2.35   0.0190 *
## MotherEd      0.02999    0.01175    2.55   0.0108 *
## Intelligence -0.00498    0.00171   -2.91   0.0036 **
## Education05   0.06478    0.01310    4.95  8.1e-07 ***
## Income87      0.18215    0.02868    6.35  2.6e-10 ***
## Income05      0.14371    0.03069    4.68  3.0e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.23 on 2113 degrees of freedom
## Multiple R-squared:  0.11,   Adjusted R-squared:  0.107
## F-statistic: 43.4 on 6 and 2113 DF,  p-value: <2e-16
```

We now test our linear model assumptions (linearity, equal variances/homoscedasticity, normality). Based on our residuals plot, there is a general spread across our values. There is a negative sloping line at the top, but the rest of the residuals are well-spread. In the Q-Q plot, we see the points generally fit a linear shape.




```
##
```

```
## Call:
## lm(formula = Esteem_PC1 ~ Inewspaper + MotherEd + Intelligence +
##     Education05 + Income87 + Income05, data = merged_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.572 -0.957 -0.054  1.002  2.965
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.63291    0.38718  -11.97  < 2e-16 ***
## Inewspaper1   0.19554    0.08332    2.35   0.0190 *
## MotherEd      0.02999    0.01175    2.55   0.0108 *
## Intelligence -0.00498    0.00171   -2.91   0.0036 **
## Education05   0.06478    0.01310    4.95  8.1e-07 ***
## Income87      0.18215    0.02868    6.35  2.6e-10 ***
## Income05      0.14371    0.03069    4.68  3.0e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.23 on 2113 degrees of freedom
## Multiple R-squared:  0.11,   Adjusted R-squared:  0.107
## F-statistic: 43.4 on 6 and 2113 DF,  p-value: <2e-16
```

```
  - Write a summary of your findings. In particular, explain what and how the variables in the model af
```

Summary: We found from our multiple regression models that the variables with the lowest p-values, which had the highest impact on the Esteem_PC1 variable, were Intelligence, Education05, Income87, Income05, Inewspaper1, and MotherEd. We evaluated the p-values for each of these variables, which were derived using the f-statistic. We used an alpha value of 0.01 as our cutoff. We then ran a multiple regression with these 6 variables, and tested our linear model assumptions of linearity, equal variances (homoscedasticity), and normality. In order to do this, we looked at our residual plot and the Q-Q residuals. Roughly speaking, the linearity, homoscedasticity and normality assumptions are satisfied. We can see in our residual plot that we have a decent split of points above and below 0, and we also have a decent spread. In the Q-Q plot, we see the points generally fit a linear shape. Thus, we conclude that our multiple regression model fits the linear model assumptions. In terms of how these values impact the Esteem_PC1, considering that our esteem scores have an inverse relationship (lower scores imply higher esteem), we see that all of variables have positive "Estimate" values in summary(fit6), except for Intelligence, which has a negative "Estimate" value. This tells us that for higher values of the Education05, Income87, Income05, Inewspaper1, and MotherEd variables, this corresponds to lower esteem, whereas higher scores for Intelligence correspond to higher Esteem_PC1 scores.

2 Case study 2: Breast cancer sub-type

The Cancer Genome Atlas (TCGA), a landmark cancer genomics program by National Cancer Institute (NCI), molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. The genome data is open to public from the Genomic Data Commons Data Portal (GDC).

In this study, we focus on 4 sub-types of breast cancer (BRCA): basal-like (basal), Luminal A-like (lumA), Luminal B-like (lumB), HER2-enriched. The sub-type is based on PAM50, a clinical-grade luminal-basal classifier. (We had hoped to download the data for control groups for each type of the cancer. But failed to do so. Please let us know if you find the appropriate data.)

- Luminal A cancers are low-grade, tend to grow slowly and have the best prognosis.
- Luminal B cancers generally grow slightly faster than luminal A cancers and their prognosis is slightly

worse.

- HER2-enriched cancers tend to grow faster than luminal cancers and can have a worse prognosis, but they are often successfully treated with targeted therapies aimed at the HER2 protein.
- Basal-like breast cancers or triple negative breast cancers do not have the three receptors that the other sub-types have so have fewer treatment options.

We will try to use mRNA expression data alone without the labels to classify 4 sub-types. Classification without labels or prediction without outcomes is called unsupervised learning. We will use K-means and spectrum clustering to cluster the mRNA data and see whether the sub-type can be separated through mRNA data.

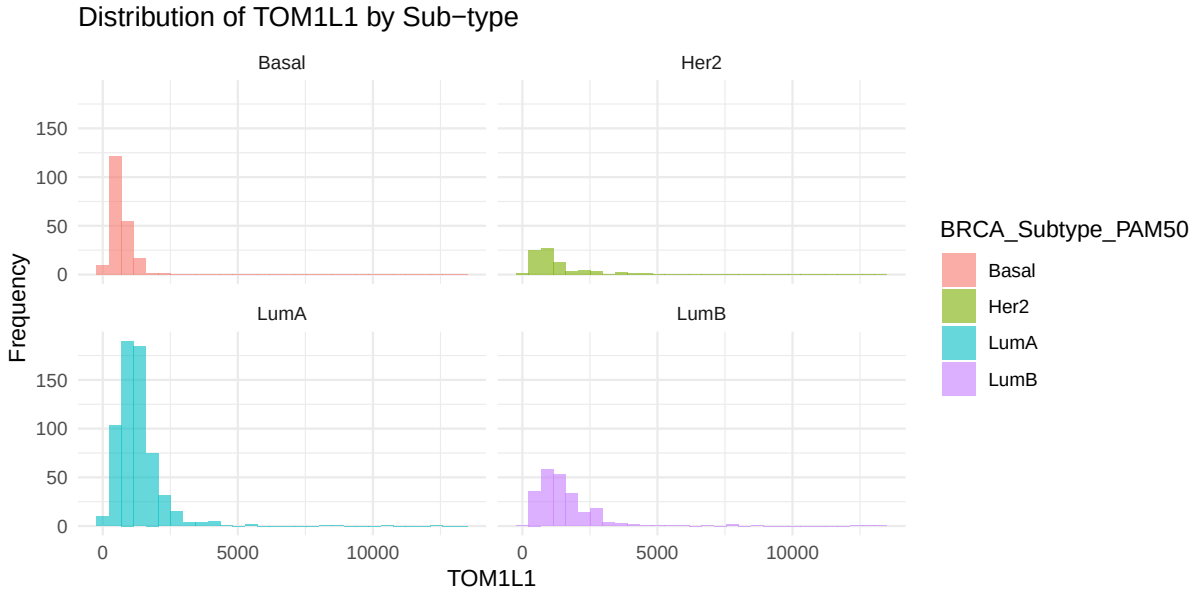
We first read the data using `data.table::fread()` which is a faster way to read in big data than `read.csv()`.

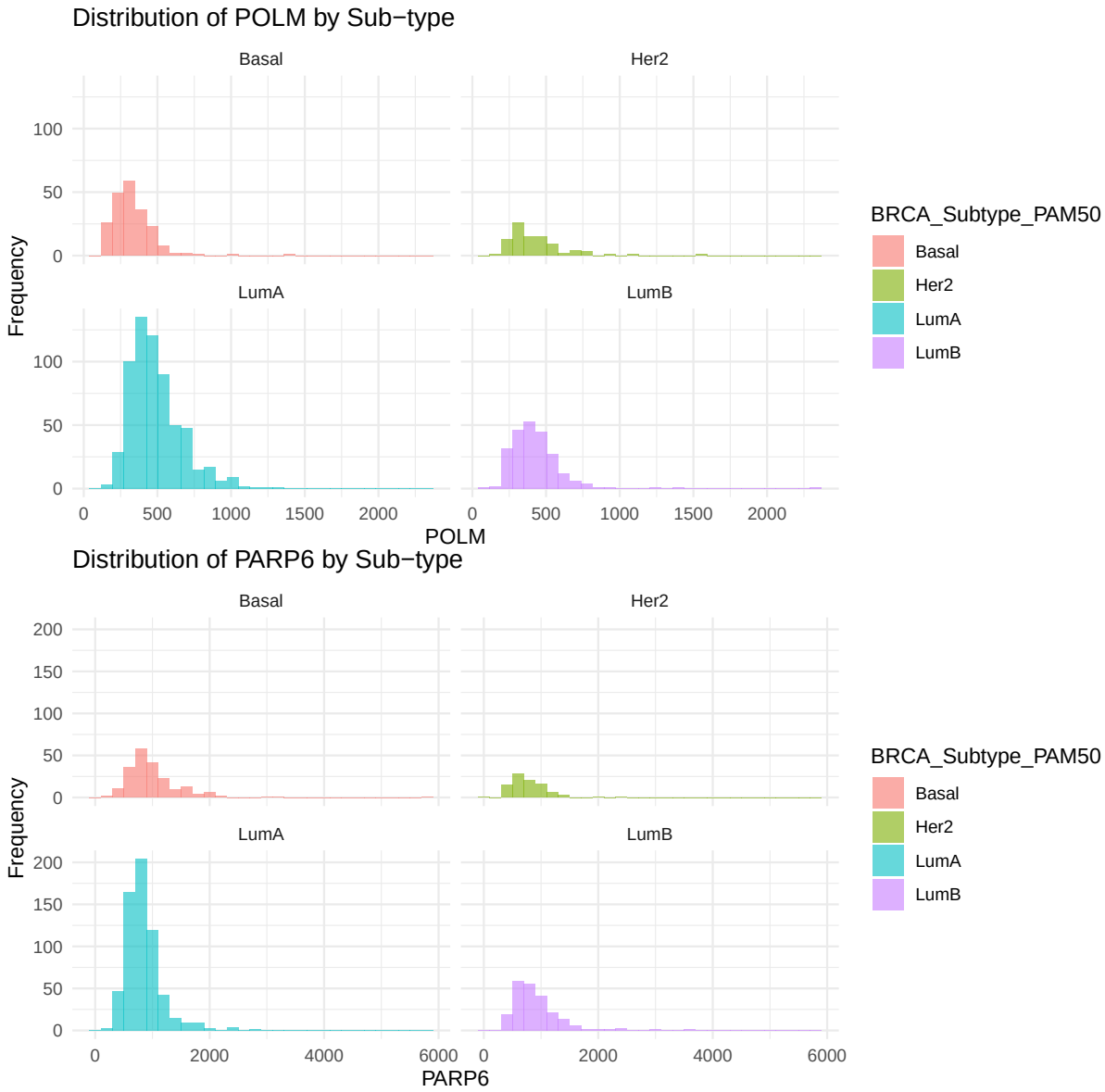
1. Summary and transformation

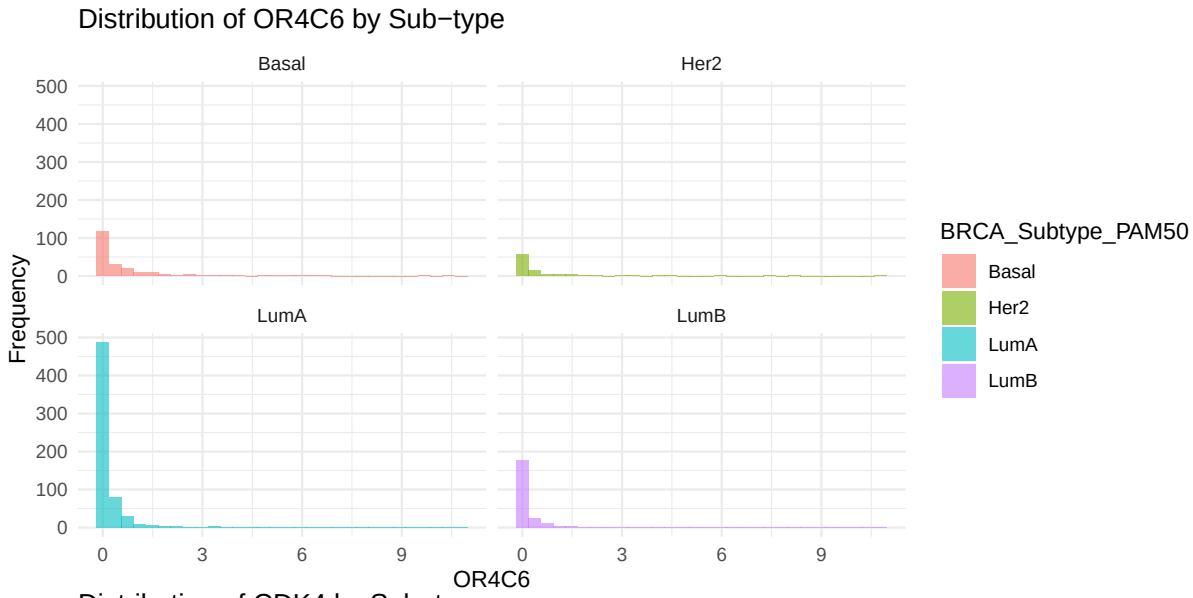
   a) How many patients are there in each sub-type? There are 208 Basal, 91 Her2, 628 LumA, and 233 Lumb patients.

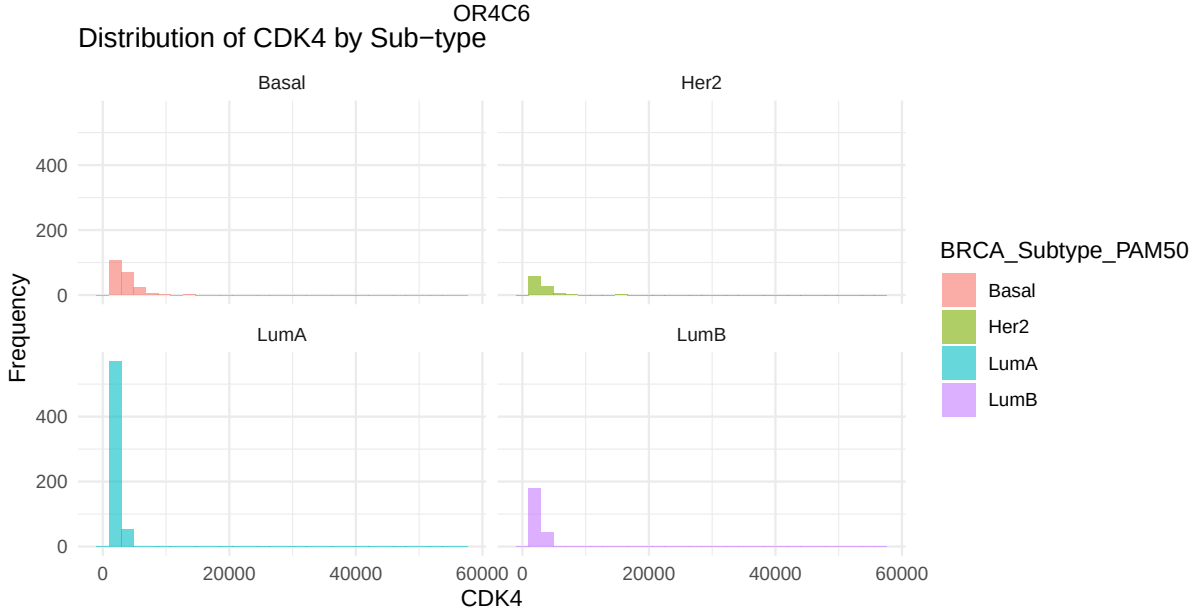
```
## brca_subtype
## Basal  Her2  LumA  LumB
##   208    91   628   233
```

   b) Randomly pick 5 genes and plot the histogram by each sub-type.




Distribution of POLM by Sub−type
Basal
Her2
LumA
LumB
Frequency
POLM
BRCA_Subtype_PAM50
Basal
Her2
LumA
LumB
Distribution of PARP6 by Sub−type
PARP6





c) Clean and transform the mRNA sequences by first remove gene with zero count and no variability and then apply logarithmic transform.

```
## [1]  1160 19947
```

We see that there were originally 19,947 rows, and our removal process resulted in 19,669, so we elimitated 288 rows (patients).

```
## [1]  1160 19669
```

2. Apply kmeans on the transformed dataset with 4 centers (4 clusters) and output the discrepancy table between the real sub-type `brca_subtype` and the cluster labels.

```
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
##
## brca_subtype   1   2   3   4
##        Basal  17   1   3 187
##        Her2    9  43  23  16
```

```
##         LumA   86 395 147   0
##         LumB   22 104 105   2
```

3. Spectrum clustering: to scale or not to scale?

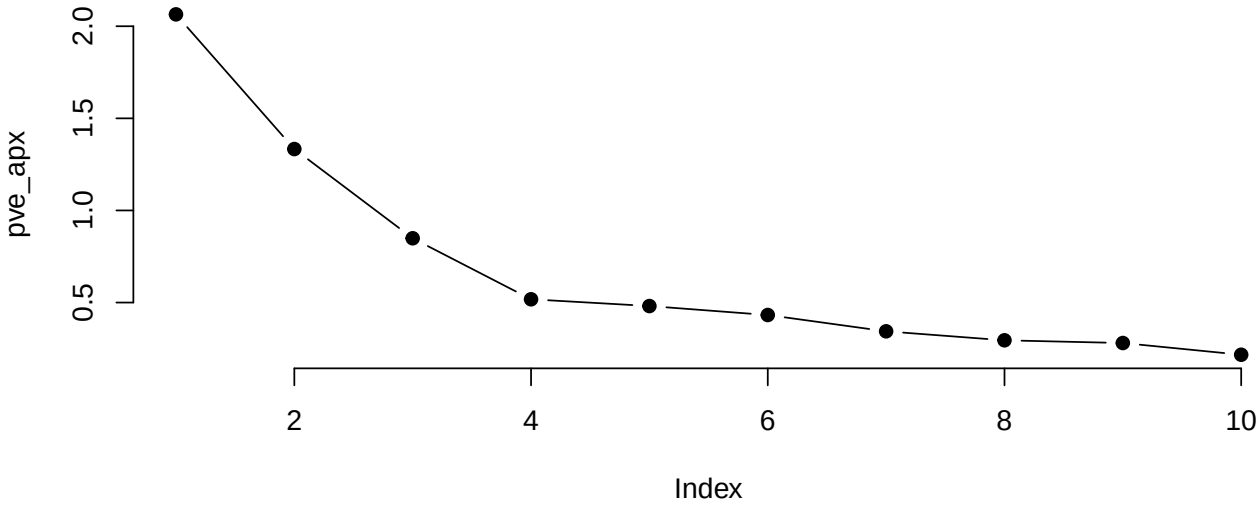
    a) Apply PCA on the centered and scaled dataset. How many PCs should we use and why? You are encouraged to use `irlba::irlba()`. **In order to do so please review the section about SVD in PCA module.**

We first calculated the PCA using irlba based on the SVD of the scaled and centered data. We then plotted the PVE and saw that the elbow was located at 4 PCs.

`b) Plot PC1 vs PC2 of the centered and scaled data and PC1 vs PC2 of the centered but unscaled data sid`

We first calculate the centered and unscaled data and their PCs using irlba. We then found the PVEs for this data, and we are also going to use 4 PCs because the elbow is located there.

```
## [1] "d"     "u"     "v"     "iter"  "mprod"
```




```
##  [1] 2.064 1.333 0.849 0.518 0.481 0.432 0.344 0.295 0.280 0.217
```

We now plot the PC1 and PC2 of both datasets. Start w/ scaled data:

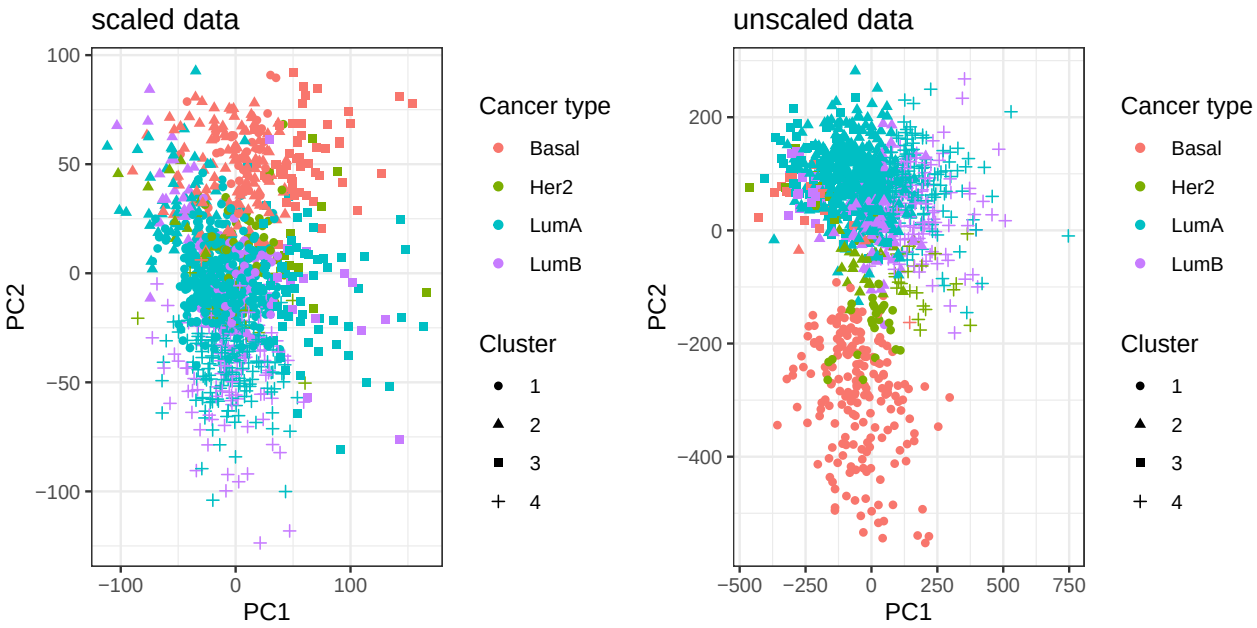
```
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
##
## brca_subtype   1   2   3   4
##         Basal  15 174   1  18
##         Her2    9  19  43  20
##         LumA   89   0 401 138
##         LumB   21   3 178  31
```

Now for unscaled data:

```
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
##
## brca_subtype   1   2   3   4
##         Basal 187   1   3  17
##         Her2   16  43  23   9
##         LumA    0 395 147  86
##         LumB    2 104 105  22
```

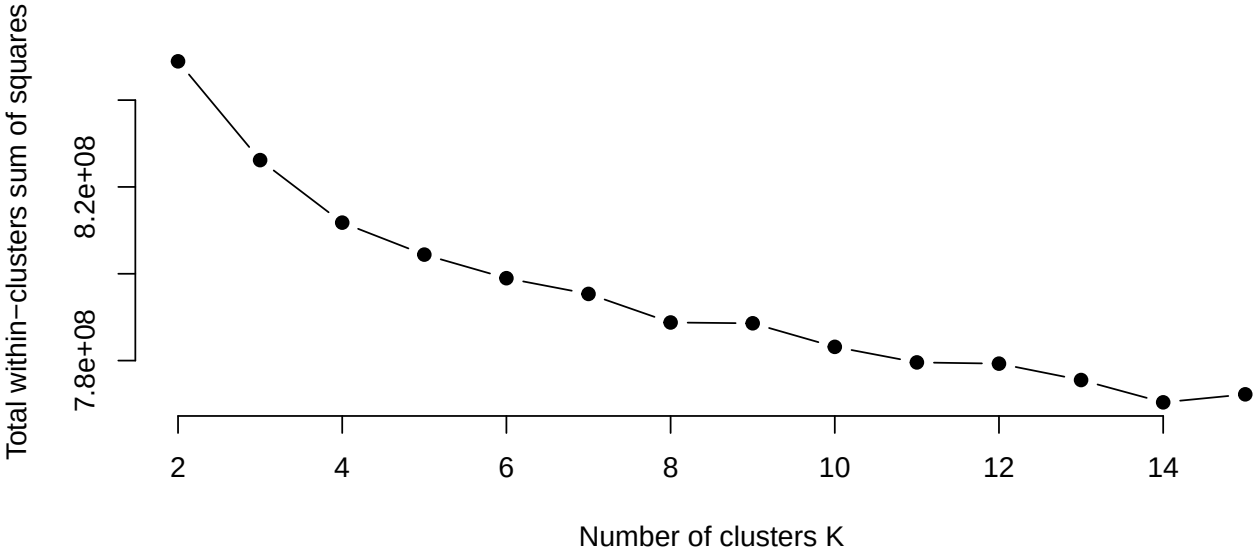


After plotting the PC2 vs. PC1 for the scaled and unscaled/scaled data, we see that the clusters for the unscaled data are less overlapping and more distinct than the scaled data.

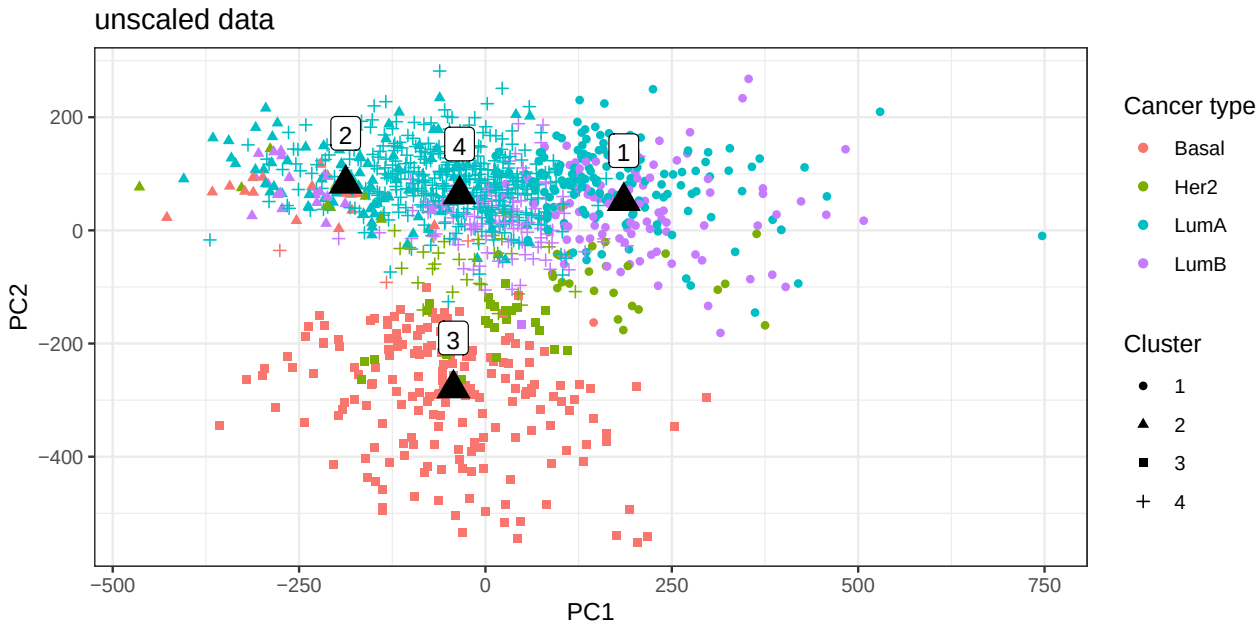
4. Spectrum clustering: center but do not scale the data

   a) Use the first 4 PCs of the centered and unscaled data and apply kmeans. Find a reasonable number of clusters using within sum of squared with the elbow rule.

We plotted the WCSS for each value of k and we found, using the elbow rule, that the optimal number of clusters should be $k = 4$.



b) Choose an optimal cluster number and apply kmeans. Compare the real sub-type and the clustering labe




```
##
## brca_subtype   1   2   3   4
##        Basal   1  16 188   3
##        Her2   27   9  24  31
##        LumA  148  85   0 395
##        LumB  101  22   1 109
```

Compared the the actual sub-types, the third cluster is fairly differentiated, while clusters 1, 2, and 4 are of low quality because their sub-type distributions are not concentrated.

c) Compare the clustering result from applying kmeans to the original data and the clustering result f

The sub-type distributions of the clusters after PCA are slightly more concentrated relative to the original data. PCA may be constructing some features that capture a greater amount of variance than the original features, hence the slightly more concentrated distributions.

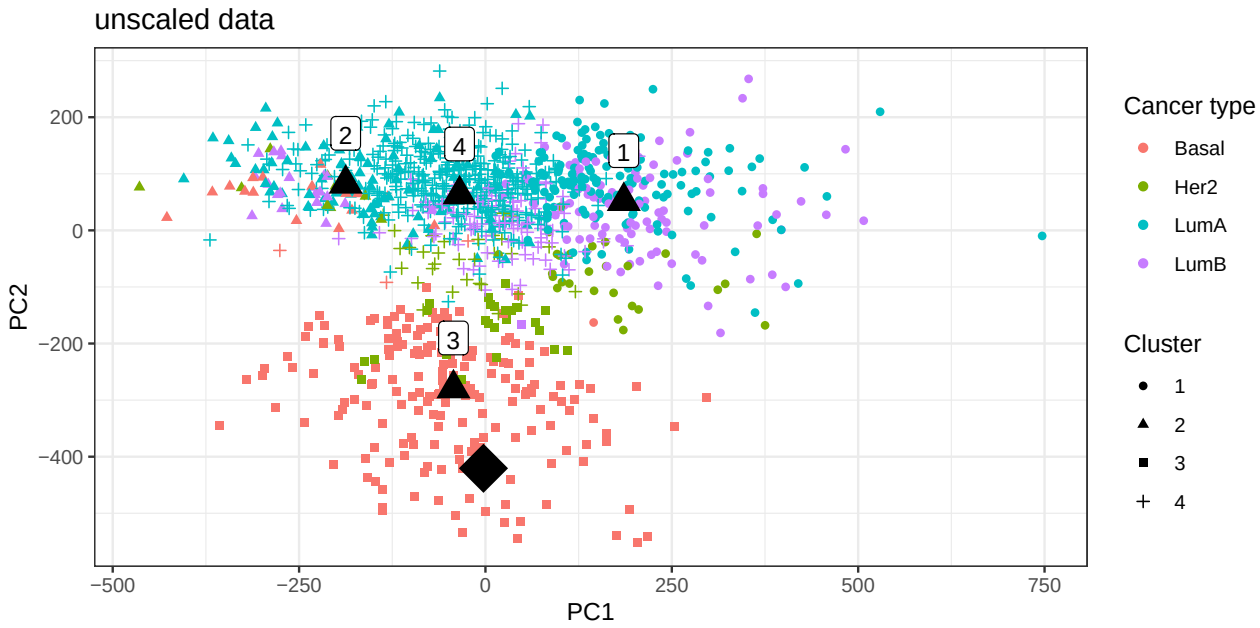
```
##
## brca_subtype   1   2   3   4
##        Basal  17   1   3 187
##        Her2    9  43  23  16
##        LumA   86 395 147   0
##        LumB   22 104 105   2
```

```
##
## brca_subtype   1   2   3   4
##        Basal   1  16 188   3
##        Her2   27   9  24  31
##        LumA  148  85   0 395
##        LumB  101  22   1 109
```

d) Now we have an x patient with breast cancer but with unknown sub-type. We have this patient's mRNA s

```
##         A1BG          A2M         NAT1         NAT2 RP11-986E7.7        AADAC
##       -0.106       -0.823       -3.483        0.417        2.809      -22.158
##         AAMP        AANAT         AARS         ABAT
##       -0.492       16.141        1.269       -2.913
```

```
## [1]   -2.65 -420.51   16.23   23.56
```




```
##   1   2   3   4
## 513 609 151 490
```

We can see that the euclidean distance to the first cluster centroid is the smallest, so we predict that the x_patient has the Basal sub-type, since we see that it is in cluster 1.

3 Case Study: Fuel Efficiency in Automobiles

What determines how fuel efficient a car is? Are Japanese cars more fuel efficient? To answer these questions we will build various linear models using the `Auto` dataset from the book `ISLR`. The original dataset contains information for about 400 different cars built in various years. To get the data, first install the package ISLR which has been done in the first R-chunk. The `Auto` dataset should be loaded automatically. Original data source is here: https://archive.ics.uci.edu/ml/datasets/auto+mpg

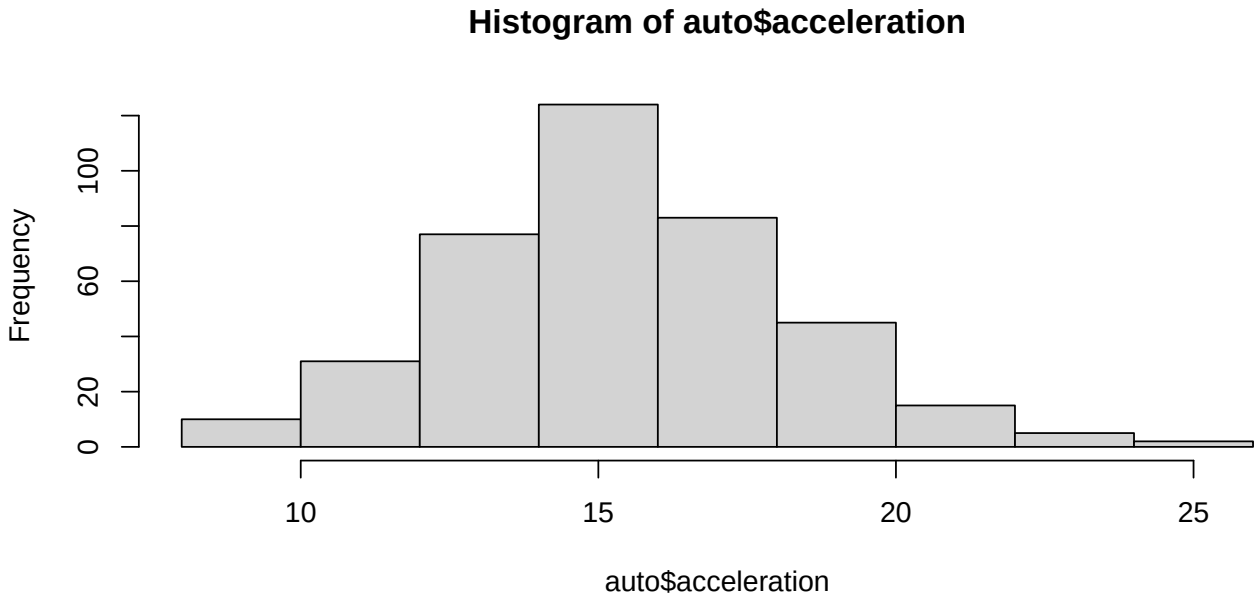
Get familiar with this dataset first. Tip: you can use the command `?ISLR::Auto` to view a description of the dataset. Our response variable will me `MPG`: miles per gallon.

3.1 EDA

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 1
```

```
##       mpg          cylinders      displacement   horsepower        weight
##  Min.   : 9.0   Min.   :3.00   Min.   : 68   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.0   1st Qu.:4.00   1st Qu.:105   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.8   Median :4.00   Median :151   Median : 93.5   Median :2804
##  Mean   :23.4   Mean   :5.47   Mean   :194   Mean   :104.5   Mean   :2978
```

```
##  3rd Qu.:29.0   3rd Qu.:8.00   3rd Qu.:276   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.6   Max.   :8.00   Max.   :455   Max.   :230.0   Max.   :5140
##
##   acceleration       year        origin                     name
##  Min.   : 8.0   Min.   :70   Min.   :1.00   amc matador       :  5
##  1st Qu.:13.8   1st Qu.:73   1st Qu.:1.00   ford pinto        :  5
##  Median :15.5   Median :76   Median :1.00   toyota corolla    :  5
##  Mean   :15.5   Mean   :76   Mean   :1.58   amc gremlin       :  4
##  3rd Qu.:17.0   3rd Qu.:79   3rd Qu.:2.00   amc hornet        :  4
##  Max.   :24.8   Max.   :82   Max.   :3.00   chevrolet chevette:  4
##                                             (Other)           :365
```



a) Explore the data, list the variables with clear definitions. Set each variable with its appropriate class. For example `origin` should be set as a factor.

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : int  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : Factor w/ 5 levels "3","4","5","6",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ displacement: int  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : int  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 1
```

```
##       mpg       cylinders  displacement   horsepower        weight
##  Min.   : 9.0   3:  4     Min.   : 68   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.0   4:199     1st Qu.:105   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.5   5:  3     Median :151   Median : 93.5   Median :2804
##  Mean   :23.3   6: 83     Mean   :194   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.0   8:103     3rd Qu.:276   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.0             Max.   :455   Max.   :230.0   Max.   :5140
##
##   acceleration       year    origin                  name
##  Min.   : 8.0   Min.   :70   1:245   amc matador       :  5
```

```
##  1st Qu.:13.8   1st Qu.:73   2: 68   ford pinto        :  5
##  Median :15.5   Median :76   3: 79   toyota corolla    :  5
##  Mean   :15.5   Mean   :76           amc gremlin       :  4
##  3rd Qu.:17.0   3rd Qu.:79           amc hornet        :  4
##  Max.   :24.8   Max.   :82           chevrolet chevette:  4
##                                      (Other)           :365
```

The variables are:

mpg (miles per gallon) - int: fuel efficiency metric cylinders - factor: number of cylinders in the car, 4-8 displacement - int: engine displacement in cubic inches horsepower - int: the rate at which work is done by the car (power metric) weight - int: the weight of the car in lbs acceleration - num: time to accelerate from 0 to 60 mph in seconds year - num: the model year of the car, modulo 100 origin - factor: either 1 American, 2 European, or 3 Japanese

b) How many cars are included in this data set?

There are 392 cars included in this data set, each with 9 variables.

```
## [1] 392   9
```

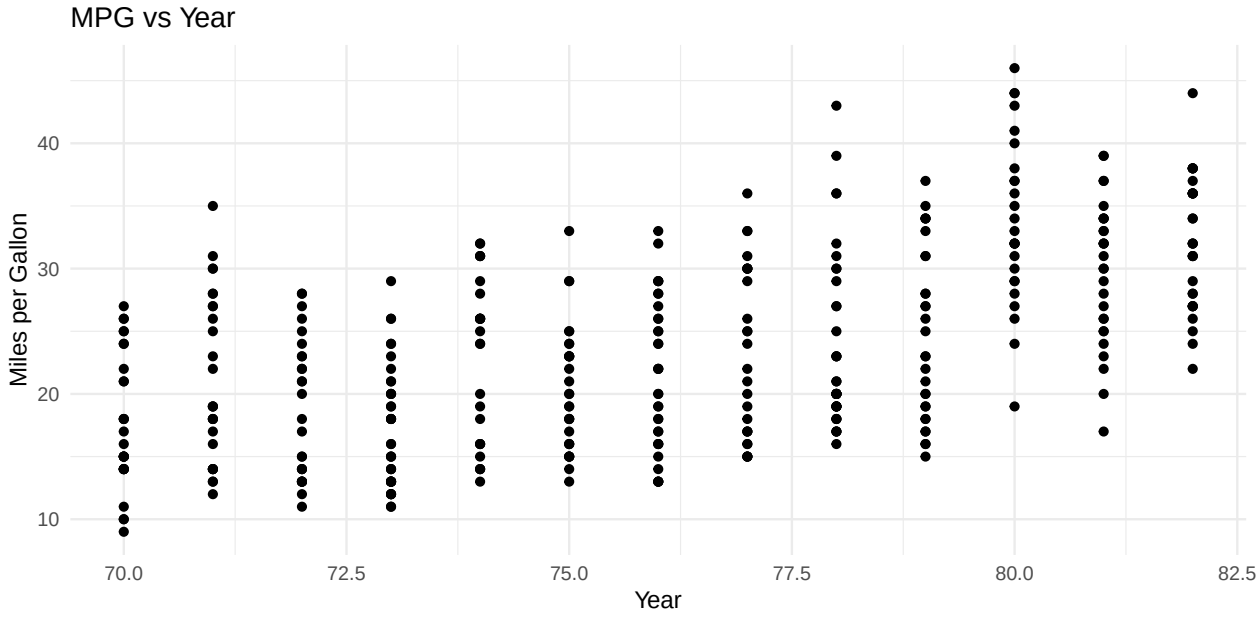
c) EDA, focus on pairwise plots and summary statistics. Briefly summarize your findings and any peculiarities in the data.

```
##       mpg       cylinders  displacement   horsepower        weight
##  Min.   : 9.0   3:  4     Min.   : 68   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.0   4:199     1st Qu.:105   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.5   5:  3     Median :151   Median : 93.5   Median :2804
##  Mean   :23.3   6: 83     Mean   :194   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.0   8:103     3rd Qu.:276   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.0             Max.   :455   Max.   :230.0   Max.   :5140
##
##   acceleration       year    origin                  name
##  Min.   : 8.0   Min.   :70   1:245   amc matador       :  5
##  1st Qu.:13.8   1st Qu.:73   2: 68   ford pinto        :  5
##  Median :15.5   Median :76   3: 79   toyota corolla    :  5
##  Mean   :15.5   Mean   :76           amc gremlin       :  4
##  3rd Qu.:17.0   3rd Qu.:79           amc hornet        :  4
##  Max.   :24.8   Max.   :82           chevrolet chevette:  4
##                                      (Other)           :365
```

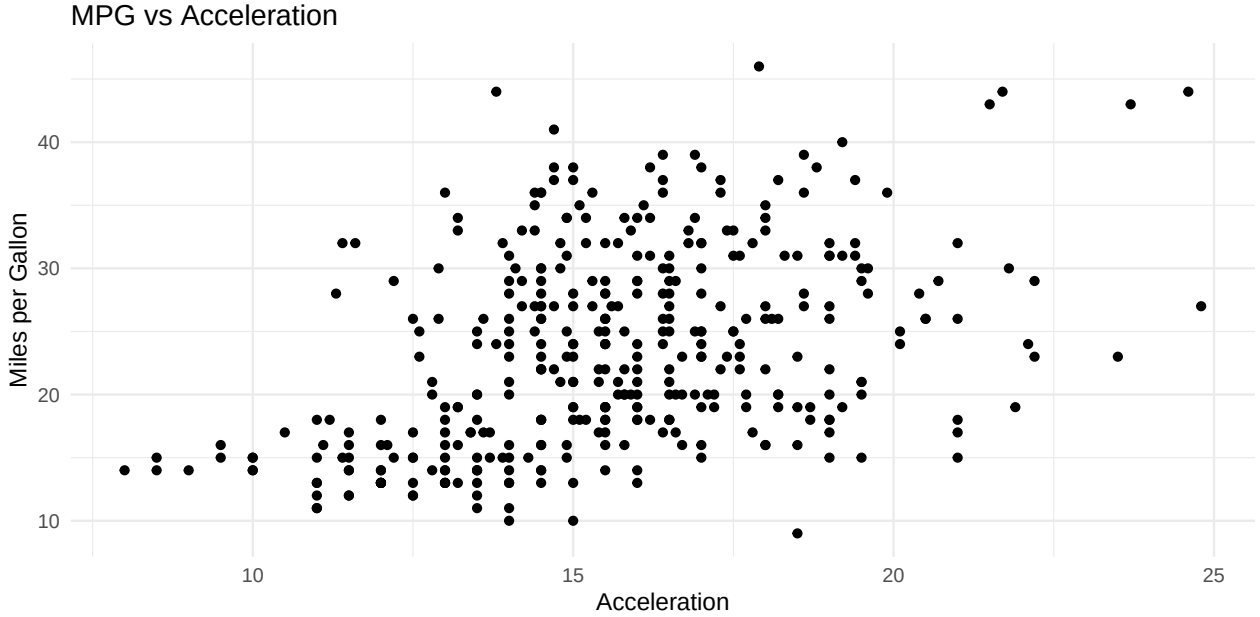
```
##       mpg       cylinders  displacement   horsepower        weight
##  Min.   : 9.0   3:  0     Min.   : 68   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.0   4:199     1st Qu.:105   1st Qu.: 75.0   1st Qu.:2225
##  Median :23.0   5:  3     Median :151   Median : 92.5   Median :2822
##  Mean   :23.3   6: 83     Mean   :196   Mean   :104.5   Mean   :2984
##  3rd Qu.:29.0   8:103     3rd Qu.:302   3rd Qu.:129.0   3rd Qu.:3622
##  Max.   :46.0             Max.   :455   Max.   :230.0   Max.   :5140
##
##   acceleration       year    origin                  name
##  Min.   : 8.0   Min.   :70   1:245   amc matador       :  5
##  1st Qu.:13.9   1st Qu.:73   2: 68   ford pinto        :  5
##  Median :15.5   Median :76   3: 75   toyota corolla    :  5
##  Mean   :15.6   Mean   :76           amc gremlin       :  4
##  3rd Qu.:17.1   3rd Qu.:79           amc hornet        :  4
##  Max.   :24.8   Max.   :82           chevrolet chevette:  4
##                                      (Other)           :361
```

The data set said that the number of cylinders would be between 4-8, but there are 4 cars with 3 cylinders (so we remove these out of range cars). We expect cars to have an even number of cylinders but there are 3 cars with 5. We decided to keep these values because there are cars manufactured with 5 cylinders, but we note that they are abnormal. We also see an outlier in mpg, where the minimum is just 9. We see most cars are American, followed by some Japanese then some European. We have 304 different factor categories for name. The years are between '70 and '82. The car weights range from 1600 to 5140 lbs, with an average of 2804. There don't seem to be any crazy outliers in horsepower, displacement, or weight. We think that though some cars have higher accelerations, they are not worth removing from the set since they aren't too far off from the other data points.

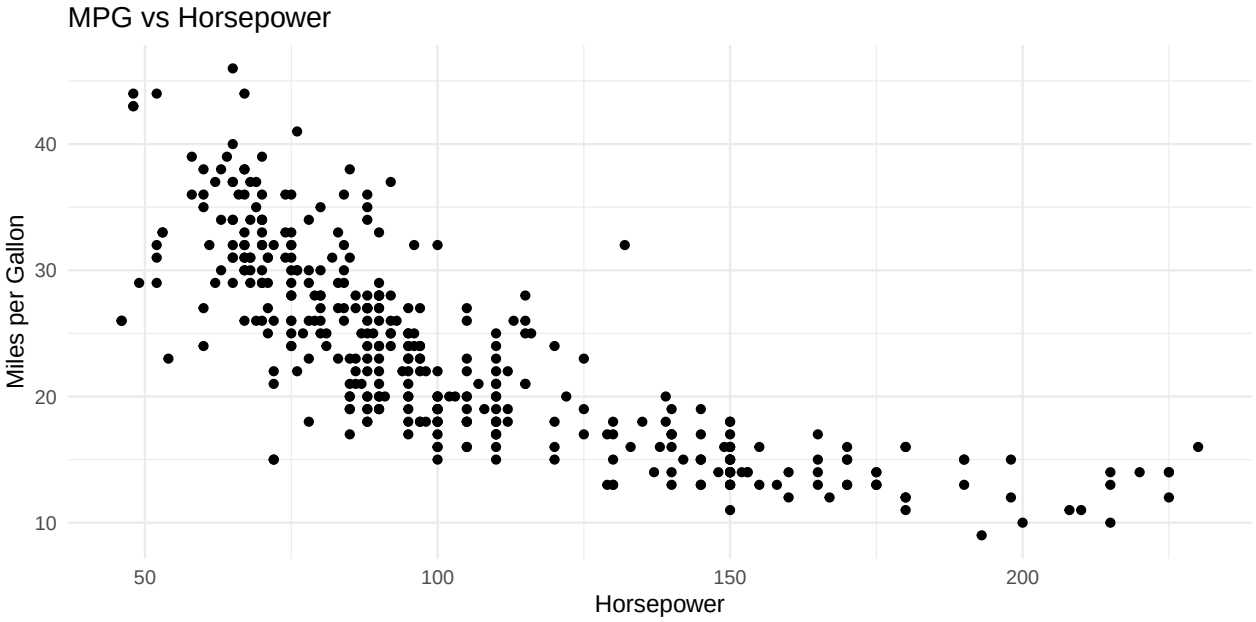
We see from the plot below that more recent cars have better fuel efficiency.



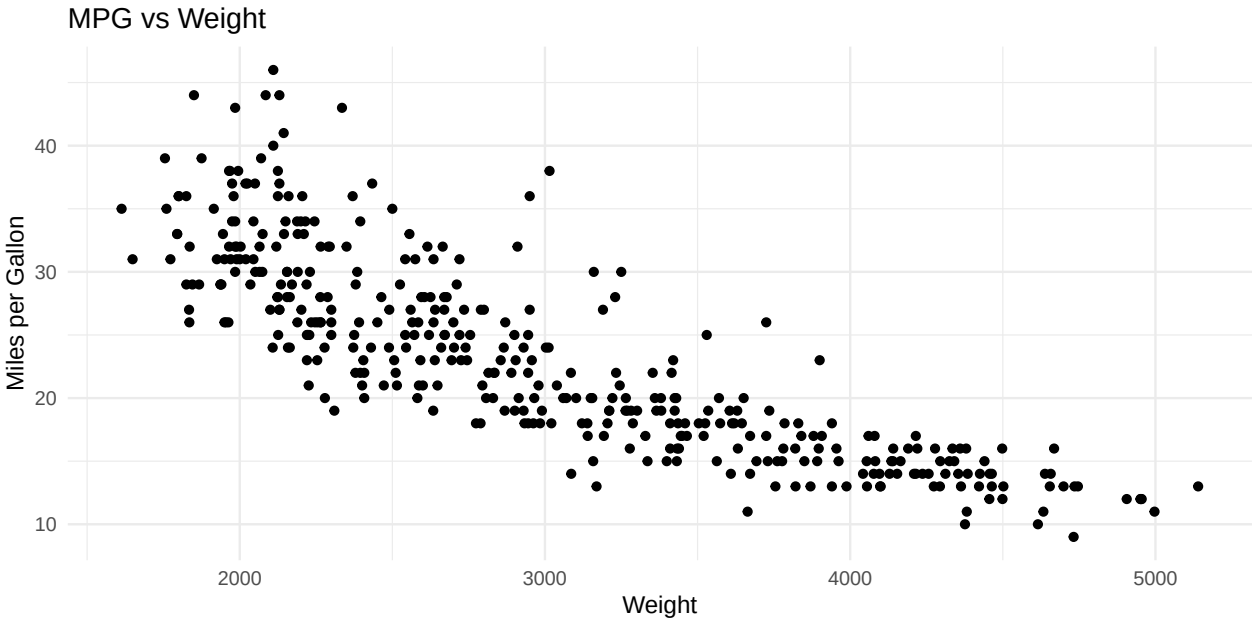
From the plot below, we see a general positive trend between acceleration and MPG, but not too significant.



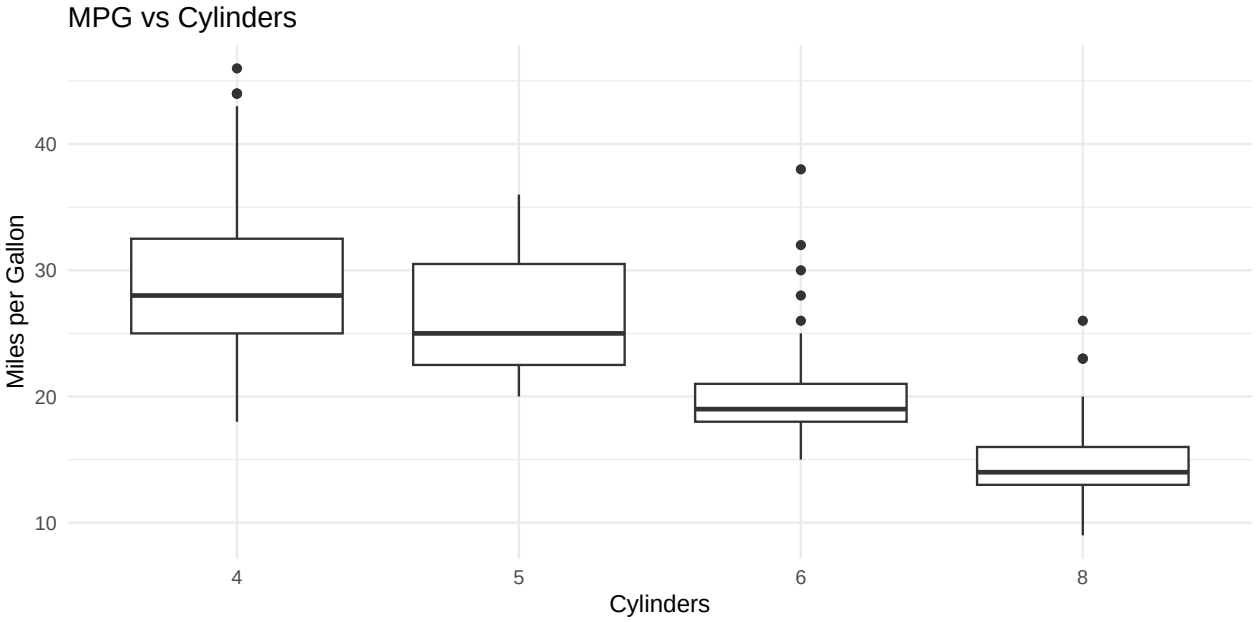
From the plot below, we see a clear decreasing trend that implies that higher horsepower reduces MPG fuel efficiency.



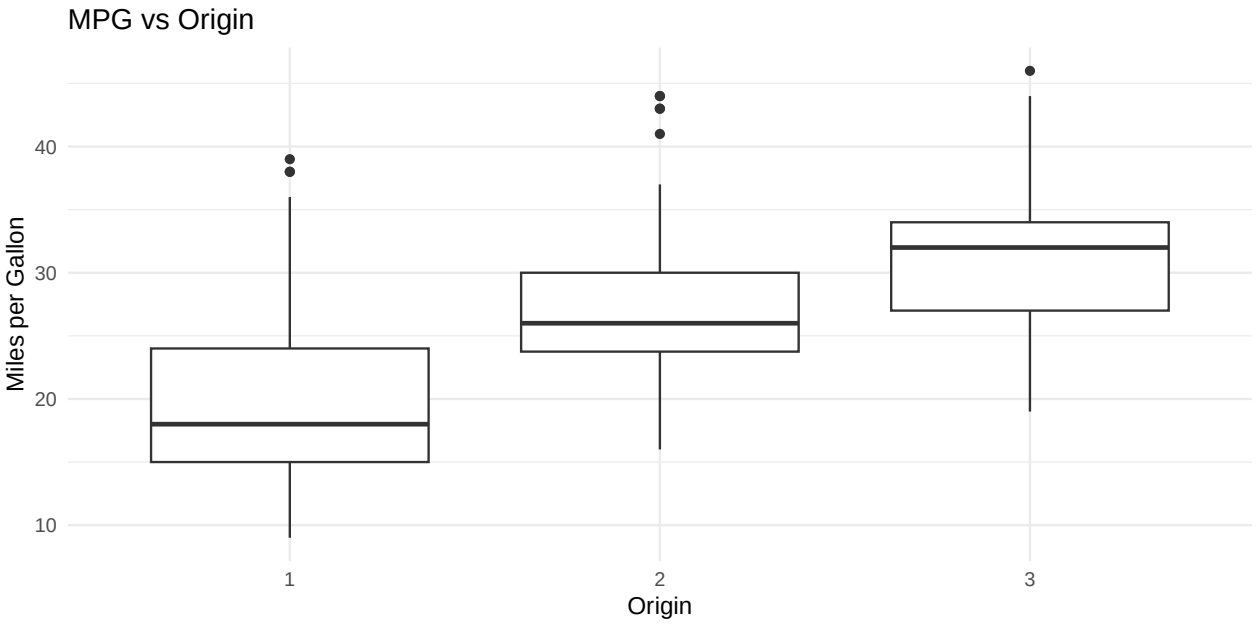
We see from the plot below that increased weight corresponds to decreased fuel efficiency for the car.



We see from the plot below that more cylinders correspond to lower MPG's (lower fuel efficiency) on average.



We see from the plot below that the cars with the higheest fuel efficiency are Japanese cars, followed by European, then American.



3.2 What effect does `time` have on `MPG`?

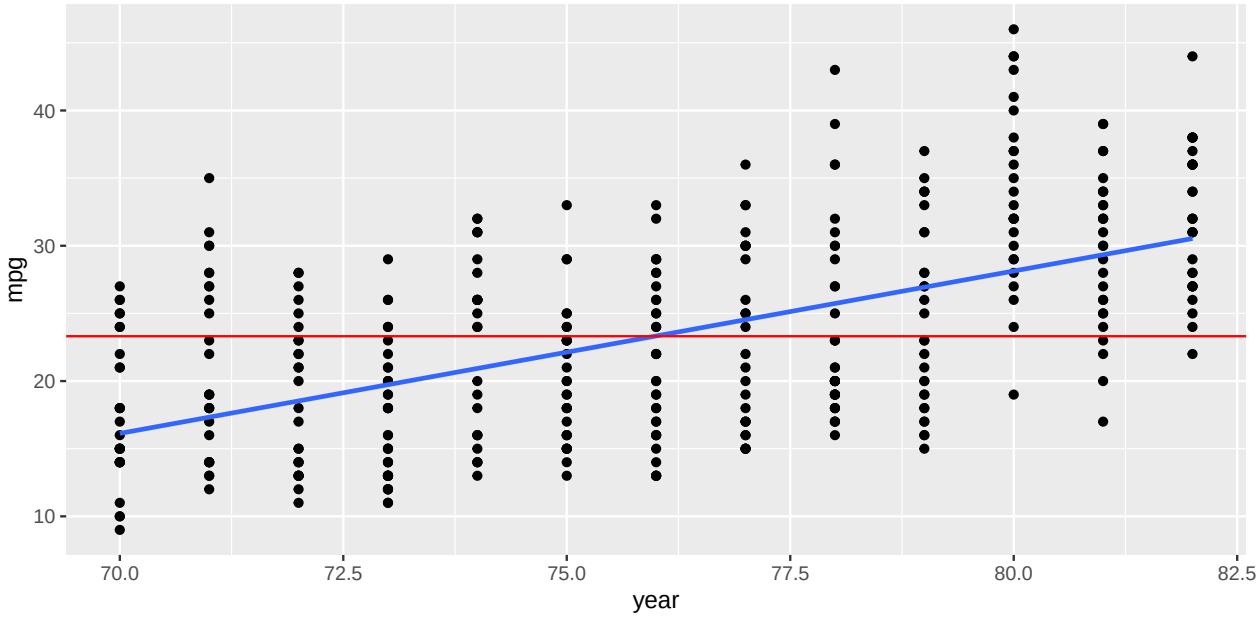
a) Start with a simple regression of `mpg` vs. `year` and report R's `summary` output. Is `year` a significant variable at the .05 level? State what effect `year` has on `mpg`, if any, according to this model.

After running a linear regression of mpg vs year, we see that there is a positive relationship between year and fuel efficiency. Based on the summary of our fit, we see a p-value of <2e-16, which suggests that at the 0.05 level, year IS a significant factor on mpg. Higher year implies higher mpg. We see a beta value of 1.1977.

```
##
## Call:
## lm(formula = mpg ~ year, data = auto)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.335  -5.531  -0.532   4.870  17.865
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -67.8987     6.7131   -10.1   <2e-16 ***
## year          1.2004     0.0882    13.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.4 on 386 degrees of freedom
## Multiple R-squared:  0.324,  Adjusted R-squared:  0.322
## F-statistic:  185 on 1 and 386 DF,  p-value: <2e-16
```

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



b) Add `horsepower` on top of the variable `year` to your linear model. Is `year` still a significant variable at the .05 level? Give a precise interpretation of the `year`'s effect found here.

Based on the summary of our lm for year and horsepower, we see that the p-value is still less than alpha = 0.05. It is <2e-16, suggesting that year IS still a significant variable on mpg. Our precise interpretation of year's effect is that its beta value is 0.61917, which implies that a single unit increase in year increases the mpg by 0.61917. Since we have such a low p-value, this disproves the null hypothesis that year doesn't impact mpg.

```
##
## Call:
## lm(formula = mpg ~ year + horsepower, data = auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.01  -3.08  -0.36   2.54  14.97
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.83423    5.39528   -1.82    0.069 .
## year         0.61836    0.06682    9.25   <2e-16 ***
## horsepower  -0.13238    0.00637  -20.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.4 on 385 degrees of freedom
## Multiple R-squared:  0.681,  Adjusted R-squared:  0.68
## F-statistic:  412 on 2 and 385 DF,  p-value: <2e-16
```

c) The two 95% CI's for the coefficient of year differ among (a) and (b). How would you explain the difference to a non-statistician?

To find the 95% CI for the coefficient of the year between (a) and (b) we do +/- 2 x the standard error for year. For (a), this would be 1.1977 +/- 2 * 0.0886, and for (b) it would be 0.61917 +/- 2 * 0.06690. In the second case, we have tighter range/bound for our 95% confidence interval, meaning we are more certain on our estimate for the beta value of year with a 95% confidence. In the first case, we have a higher estimate, and thus the range of values that we believe beta to be in have higher values. In both cases, since we have a very low p-value, we know that year does in fact have an impact on mpg in our linear regression.

d) Create a model with interaction by fitting `lm(mpg ~ year * horsepower)`. Is the interaction effect significant at .05 level? Explain the year effect (if any).

As we see from the summary of the fit with interaction below the interactio effect IS significant at the 0.05 level, since we see very small p-values of <2e-16. Year alone still does have an effect, since the beta value is 2.16e+00, so there is a positive relationship where if year increases by a unit, then mpg increases this much. We also see that in terms of interaction, we see a negative value of -1.60e-02.

```
##
## Call:
## lm(formula = mpg ~ year * horsepower, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.636  -2.476  -0.408   2.453  14.095
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.25e+02   1.21e+01  -10.27   <2e-16 ***
## year             2.17e+00   1.62e-01   13.40   <2e-16 ***
## horsepower       1.06e+00   1.16e-01    9.13   <2e-16 ***
## year:horsepower -1.61e-02   1.56e-03  -10.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.9 on 384 degrees of freedom
## Multiple R-squared:  0.75,   Adjusted R-squared:  0.748
## F-statistic:  385 on 3 and 384 DF,  p-value: <2e-16
```

3.3 Categorical predictors

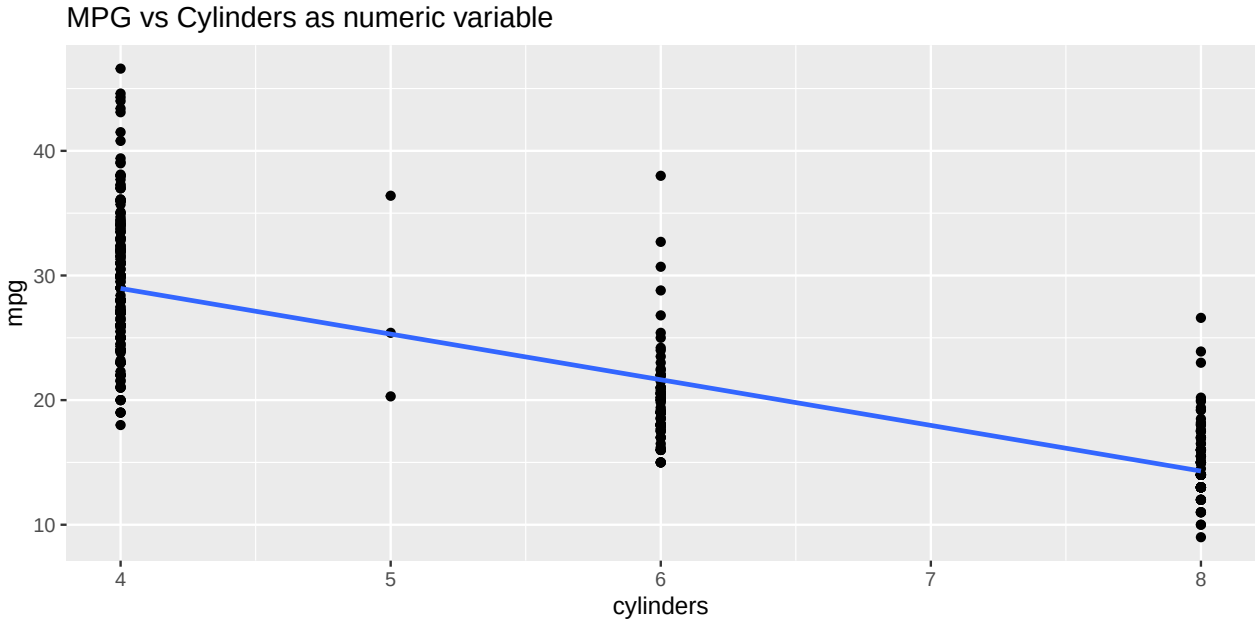
Remember that the same variable can play different roles! Take a quick look at the variable `cylinders`, and try to use this variable in the following analyses wisely. We all agree that a larger number of cylinders will lower mpg. However, we can interpret `cylinders` as either a continuous (numeric) variable or a categorical variable.

a) Fit a model that treats `cylinders` as a continuous/numeric variable. Is `cylinders` significant at the 0.01 level? What effect does `cylinders` play in this model?

The following is a linear model of mpg vs cylinders as a numeric variable. The p-value is very small, <2e-16, which suggests that cylinders is significant at the 0.01 level. Specifically, the beta value is -3.558, which suggests that increased number of cylinders by 1 unit on average implies decreased mpg by that amount on average.

```
##
## Call:
## lm(formula = mpg ~ cylinders, data = auto2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.961  -3.135  -0.635   2.572  17.639
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   43.613      0.825    52.9   <2e-16 ***
## cylinders     -3.663      0.143   -25.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.78 on 386 degrees of freedom
## Multiple R-squared:  0.628,  Adjusted R-squared:  0.627
## F-statistic:  653 on 1 and 386 DF,  p-value: <2e-16
```

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

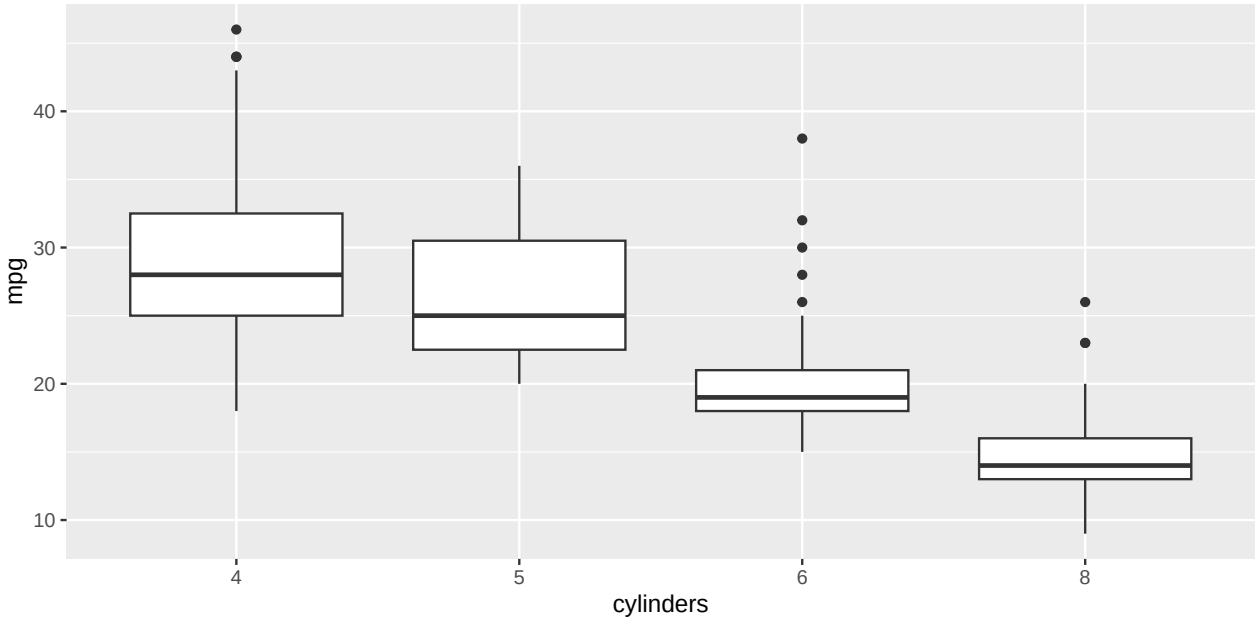



```
## mapping: yintercept = ~mean(mpg)
## geom_hline: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

b) Fit a model that treats `cylinders` as a categorical/factor. Is `cylinders` significant at the .01 level? What is the effect of `cylinders` in this model? Describe the `cylinders` effect over `mpg`.

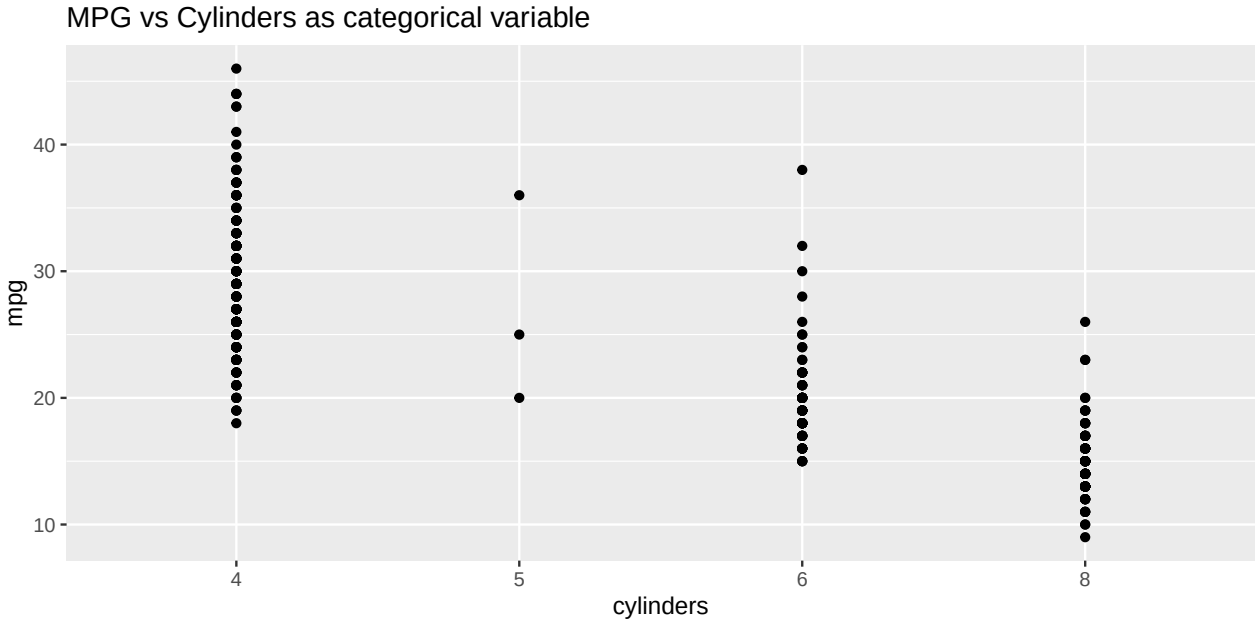
Here, we create a boxplot to show the spread of mgp given the number of cylinders.

```
## # A tibble: 4 x 4
##   cylinders  mean    sd     n
##   <fct>     <dbl> <dbl> <int>
## 1 4          29.1  5.61   199
## 2 5          27    8.19     3
## 3 6          19.8  3.75    83
## 4 8          14.9  2.72   103
```




```
##
## Call:
## lm(formula = mpg ~ cylinders, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.111  -2.854  -0.783   2.146  18.217
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.111      0.329   88.47   <2e-16 ***
## cylinders5    -2.111      2.700   -0.78     0.43
## cylinders6    -9.327      0.607  -15.38   <2e-16 ***
## cylinders8   -14.256      0.563  -25.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.64 on 384 degrees of freedom
## Multiple R-squared:  0.647,  Adjusted R-squared:  0.644
## F-statistic:  234 on 3 and 384 DF,  p-value: <2e-16
```

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




```
## mapping: yintercept = ~mean(mpg)
## geom_hline: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

```
## Anova Table (Type II tests)
##
## Response: mpg
##           Sum Sq  Df F value Pr(>F)
## cylinders  15133   3     234 <2e-16 ***
## Residuals   8274 384
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that there are p-values that are very small, <2e-16, such as in cylinders6 and cylinders8, as well as in the Anova test, which suggests that these cylinders categories do have a significant effect on mpg. We see the biggest magnitude beta values for cylinders8, followed by cylinders6. These are negative value, -14 and -9 respectively, suggesting that more cylinders implies less fuel efficiency.

c) What are the fundamental differences between treating `cylinders` as a continuous and categorical variable in your models?

When we treat cylinders as a continuous variable, we see one beta value returned by our lm model, whereas if we treat it as a categorical variable, we get separate beta values for each category, where some categories have greater magnitudes (such as cylinders8), suggesting a stronger relationship for these specific categories. Treating "cylinders" as continuous assumes a linear relationship and may not capture potential non-linear relationships or differences in the effect of each additional cylinder. When we treat cylinders as a categorical variable, we see lower RSE and higher R^2 values. We overall thus think that treating it as a categorical variable is better.

d) Can you test the null hypothesis: fit0: `mpg` is linear in `cylinders` vs. fit1: `mpg` relates to `cylinders` as a categorical variable at .01 level?

We ran an anova analysis for the two different models, where fit0 is represented by mpg_lm3, and fit1 is represented by mpg_lm4. We see that our p-value is very small, 3.7e-06. Thus, at the 0.01 level, we can reject the null hypothesis. We thus think that we should treat cylinders as a categorical variable.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cylinders
## Model 2: mpg ~ cylinders
##   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
## 1    386 8832
## 2    384 8274  2       557 12.9 3.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.4 Results

Final modeling question: we want to explore the effects of each feature as best as possible. You may explore interactions, feature transformations, higher order terms, or other strategies within reason. The model(s) should be as parsimonious (simple) as possible unless the gain in accuracy is significant from your point of view.

a) Describe the final model. Include diagnostic plots with particular focus on the model residuals and diagnoses.

For our final model, we are choosing to disregard acceleration, since as we see in the first part of this case, there isn't a very strong/clear relationship between acceleration and MPG. We include all of the other variables, since as our plots above show, there are clear relationships between all of the variables and MPG. We also chose to not include name since there are 304 categories, and we can't gain meaningful information about trends from this data (just about 5 cars in each name category).

```
##
## Call:
## lm(formula = mpg ~ acceleration, data = auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17.81  -5.83  -1.08   4.95  22.79
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.79       2.05    2.33     0.02 *
## acceleration     1.19       0.13    9.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.06 on 386 degrees of freedom
## Multiple R-squared:  0.179,  Adjusted R-squared:  0.177
## F-statistic:   84 on 1 and 386 DF,  p-value: <2e-16
```

Thus, our final model includes the variables year, cylinders (as a categorical variable), displacement, weight, horsepower, and origin. We use interactions for year and origin, since we thought the model year could have different effects in different countries. We can see that all of the p-values, with the except of a some of the cylinders categories, are very small, and certainly less than alpha = 0.01. This tells us that these variables do have significant impact on the mpg. If we look at the last line in the summary, F-statistic: 200 on 11 and 376 DF, p-value: <2e-16. These metrics confirm that our regression model includes meaningful predictor variables.

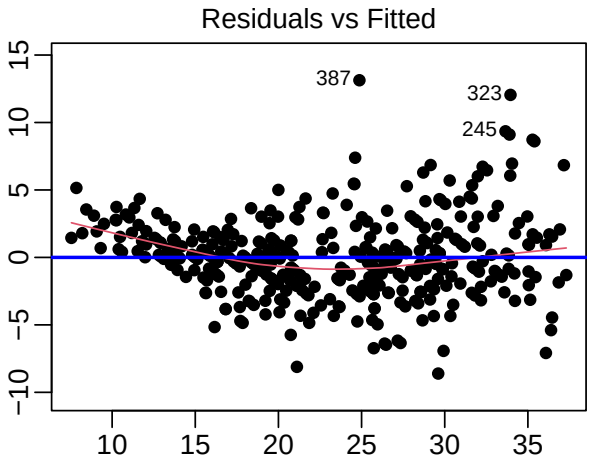
```
##
## Call:
## lm(formula = mpg ~ year * origin + cylinders + displacement +
##     weight + horsepower, data = auto)
```

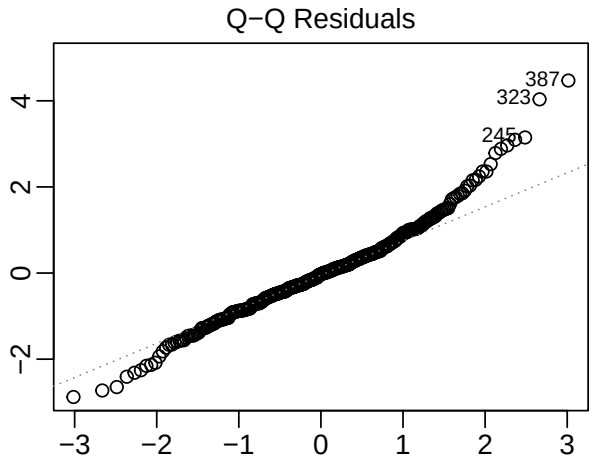
```
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.61  -1.74  -0.09   1.45  13.13
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.94e+00   4.98e+00    0.39  0.69618
## year          5.19e-01   6.20e-02    8.38  1.0e-15 ***
## origin2      -4.18e+01   9.67e+00   -4.32  2.0e-05 ***
## origin3      -2.32e+01   8.93e+00   -2.60  0.00981 **
## cylinders5   -1.09e+00   1.86e+00   -0.59  0.55766
## cylinders6   -3.65e+00   6.77e-01   -5.39  1.2e-07 ***
## cylinders8   -1.89e+00   1.19e+00   -1.58  0.11398
## displacement  1.44e-02   7.09e-03    2.03  0.04291 *
## weight       -5.54e-03   5.42e-04  -10.22  < 2e-16 ***
## horsepower   -3.64e-02   1.05e-02   -3.48  0.00055 ***
## year:origin2  5.70e-01   1.27e-01    4.50  9.0e-06 ***
## year:origin3  3.31e-01   1.15e-01    2.88  0.00414 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.02 on 376 degrees of freedom
## Multiple R-squared:  0.854,  Adjusted R-squared:  0.85
## F-statistic:  200 on 11 and 376 DF,  p-value: <2e-16
```

When we use Anova to test our model, we also see very small p-values, but displacement has the highest, which indicates it might not be the best predictor among the others. Overall, our model seems to use good quality predictors.

```
## Anova Table (Type II tests)
##
## Response: mpg
##              Sum Sq  Df F value  Pr(>F)
## year           1955   1  214.72 < 2e-16 ***
## origin          246   2   13.53 2.1e-06 ***
## cylinders       443   3   16.21 6.3e-10 ***
## displacement     38   1    4.13 0.04291 *
## weight          950   1  104.40 < 2e-16 ***
## horsepower      110   1   12.13 0.00055 ***
## year:origin     212   2   11.66 1.2e-05 ***
## Residuals      3423 376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residuals Plot: We plotted the Residuals vs Fitted plot and the Q-Q Residuals plot to test our linear model assumptions of equal variance (homoscedasticity), linearty, and normality. We see that the residuals are evenly distributed with even variance, with an even number of points above and below 0, and in the Q-Q plot we see the points fit a linear trend. Thus our model passes the model assumptions tests.






```
##
## Call:
## lm(formula = mpg ~ year * origin + cylinders + displacement +
##     weight + horsepower, data = auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.61  -1.74  -0.09   1.45  13.13
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.94e+00   4.98e+00    0.39  0.69618
## year          5.19e-01   6.20e-02    8.38  1.0e-15 ***
## origin2      -4.18e+01   9.67e+00   -4.32  2.0e-05 ***
## origin3      -2.32e+01   8.93e+00   -2.60  0.00981 **
## cylinders5   -1.09e+00   1.86e+00   -0.59  0.55766
## cylinders6   -3.65e+00   6.77e-01   -5.39  1.2e-07 ***
## cylinders8   -1.89e+00   1.19e+00   -1.58  0.11398
## displacement  1.44e-02   7.09e-03    2.03  0.04291 *
## weight       -5.54e-03   5.42e-04  -10.22  < 2e-16 ***
## horsepower   -3.64e-02   1.05e-02   -3.48  0.00055 ***
## year:origin2  5.70e-01   1.27e-01    4.50  9.0e-06 ***
## year:origin3  3.31e-01   1.15e-01    2.88  0.00414 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.02 on 376 degrees of freedom
## Multiple R-squared:  0.854,  Adjusted R-squared:  0.85
## F-statistic:  200 on 11 and 376 DF,  p-value: <2e-16
```

b) Summarize the effects found.

Our lm model shows many significant effects on mpg based on the provided coefficients in the table. The intercept term indicates an estimated mpg of approximately 1.94. Year has a positive effect on mpg, with each additional year leading to an increase of approximately 0.52 mpg. Among the different cylinder counts, cars with 6 cylinders show the most substantial negative effect on mpg, with an estimated decrease of about 3.65 mpg compared to the reference category. Weight has a strong negative effect on mpg, with an estimated

decrease of approximately 0.0055 mpg for each additional pound. Horsepower also negatively impacts mpg, with each additional unit of horsepower associated with a decrease of approximately 0.0364 mpg. There are also significant interaction effects between year and origin, indicating that the relationship between year and mpg varies across different origins. Overall, the model has a high adjusted R-squared value of 0.85, suggesting that it explains a good portion of the variance in the data.

```
##
## Call:
## lm(formula = mpg ~ year * origin + cylinders + displacement +
##     weight + horsepower, data = auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.61  -1.74  -0.09   1.45  13.13
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.94e+00   4.98e+00    0.39  0.69618
## year          5.19e-01   6.20e-02    8.38  1.0e-15 ***
## origin2      -4.18e+01   9.67e+00   -4.32  2.0e-05 ***
## origin3      -2.32e+01   8.93e+00   -2.60  0.00981 **
## cylinders5   -1.09e+00   1.86e+00   -0.59  0.55766
## cylinders6   -3.65e+00   6.77e-01   -5.39  1.2e-07 ***
## cylinders8   -1.89e+00   1.19e+00   -1.58  0.11398
## displacement  1.44e-02   7.09e-03    2.03  0.04291 *
## weight       -5.54e-03   5.42e-04  -10.22  < 2e-16 ***
## horsepower   -3.64e-02   1.05e-02   -3.48  0.00055 ***
## year:origin2  5.70e-01   1.27e-01    4.50  9.0e-06 ***
## year:origin3  3.31e-01   1.15e-01    2.88  0.00414 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.02 on 376 degrees of freedom
## Multiple R-squared:  0.854,  Adjusted R-squared:  0.85
## F-statistic:  200 on 11 and 376 DF,  p-value: <2e-16
```

c) Predict the `mpg` of the following car: A red car built in the US in 1983 that is 180 inches long, has eight cylinders, displaces 350 cu. inches, weighs 4000 pounds, and has a horsepower of 260. Also give a 95% CI for your prediction.

```
##   cylinders displacement weight origin horsepower year
## 1         8          350   4000      1        260   83
```

```
## $fit
##    fit lwr  upr
## 1 16.6  14 19.2
##
## $se.fit
## [1] 1.31
##
## $df
## [1] 376
##
## $residual.scale
## [1] 3.02
```

Based on the prediction above with our final model, we can see that the prediction is 16.6 mpg +/- 2 * 1.31 for a 95% CI, which gives us lower and upper bounds of 14 and 19.2 mpg.

4 Simple Regression through simulations (Optional)

4.1 Linear model through simulations

This exercise is designed to help you understand the linear model using simulations. In this exercise, we will generate (x_i, y_i) pairs so that all linear model assumptions are met.

Presume that $\mathbf{x}$ and $\mathbf{y}$ are linearly related with a normal error $\boldsymbol{\varepsilon}$, such that $\mathbf{y} = 1 + 1.2\mathbf{x} + \boldsymbol{\varepsilon}$. The standard deviation of the error ε_i is $\sigma = 2$.

We can create a sample input vector ($n = 40$) for $\mathbf{x}$ with the following code:

```
# Generates a vector of size 40 with equally spaced values between 0 and 1, inclusive
x <- seq(0, 1, length = 40)
```

4.1.1 Generate data

Create a corresponding output vector for $\mathbf{y}$ according to the equation given above. Use `set.seed(1)`. Then, create a scatterplot with (x_i, y_i) pairs. Base R plotting is acceptable, but if you can, please attempt to use `ggplot2` to create the plot. Make sure to have clear labels and sensible titles on your plots.

4.1.2 Understand the model

i. Find the LS estimates of β_0 and β_1, using the `lm()` function. What are the true values of β_0 and β_1? Do the estimates look to be good?

ii. What is your RSE for this linear model fit? Is it close to $\sigma = 2$?

iii. What is the 95% confidence interval for β_1? Does this confidence interval capture the true β_1?

iv. Overlay the LS estimates and the true lines of the mean function onto a copy of the scatterplot you made above.

4.1.3 diagnoses

i. Provide residual plot where fitted $\mathbf{y}$-values are on the x-axis and residuals are on the y-axis.

ii. Provide a normal QQ plot of the residuals.

iii. Comment on how well the model assumptions are met for the sample you used.

4.2 Understand sampling distribution and confidence intervals

This part aims to help you understand the notion of sampling statistics and confidence intervals. Let's concentrate on estimating the slope only.

Generate 100 samples of size $n = 40$, and estimate the slope coefficient from each sample. We include some sample code below, which should guide you in setting up the simulation. Note: this code is easier to follow but suboptimal; see the appendix for a more optimal R-like way to run this simulation.

```
# Inializing variables. Note b_1, upper_ci, lower_ci are vectors
x <- seq(0, 1, length = 40)
n_sim <- 100              # number of simulations
b1 <- 0                   # n_sim many LS estimates of beta_1 (=1.2). Initialize to 0 for now
upper_ci <- 0             # upper bound for beta_1. Initialize to 0 for now.
lower_ci <- 0             # lower bound for beta_1. Initialize to 0 for now.
t_star <- qt(0.975, 38)   # Food for thought: why 38 instead of 40? What is t_star?
```

```r
# Perform the simulation
for (i in 1:n_sim){I l
  y <- 1 + 1.2 * x + rnorm(40, sd = 2)
  lse <- lm(y ~ x)
  lse_output <- summary(lse)$coefficients
  se <- lse_output[2, 2]
  b1[i] <- lse_output[2, 1]
  upper_ci[i] <- b1[i] + t_star * se
  lower_ci[i] <- b1[i] - t_star * se
}
results <- as.data.frame(cbind(se, b1, upper_ci, lower_ci))

# remove unecessary variables from our workspace
rm(se, b1, upper_ci, lower_ci, x, n_sim, b1, t_star, lse, lse_out)
```

i. Summarize the LS estimates of β_1 (stored in `results$b1`). Does the sampling distribution agree with theory?

ii. How many of your 95% confidence intervals capture the true β_1? Display your confidence intervals graphically.