Modern Data Mining, HW 2

Jose Cervantez Bethany Hsiao Rob Kuan

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?

Answer: There are no missing values as there are no null values.

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.

Answer: For all of the Esteem variables, the minimum value is 1 and the maximum value is 4, which is as expected. There are no missing values. There is nothing to fix.

```
##    Esteem87_1     Esteem87_2    Esteem87_3     Esteem87_4    Esteem87_5
##  Min.   :1.00   Min.   :1.0   Min.   :1.00   Min.   :1.0   Min.   :1.00
##  1st Qu.:1.00   1st Qu.:1.0   1st Qu.:3.00   1st Qu.:1.0   1st Qu.:3.00
##  Median :1.00   Median :1.0   Median :4.00   Median :1.0   Median :4.00
##  Mean   :1.38   Mean   :1.4   Mean   :3.58   Mean   :1.5   Mean   :3.53
##  3rd Qu.:2.00   3rd Qu.:2.0   3rd Qu.:4.00   3rd Qu.:2.0   3rd Qu.:4.00
##  Max.   :4.00   Max.   :4.0   Max.   :4.00   Max.   :4.0   Max.   :4.00
##    Esteem87_6     Esteem87_7     Esteem87_8    Esteem87_9     Esteem87_10
##  Min.   :1.00   Min.   :1.00   Min.   :1.0   Min.   :1.00   Min.   :1.00
##  1st Qu.:1.00   1st Qu.:1.00   1st Qu.:3.0   1st Qu.:3.00   1st Qu.:3.00
##  Median :2.00   Median :2.00   Median :3.0   Median :3.00   Median :3.00
##  Mean   :1.59   Mean   :1.72   Mean   :3.1   Mean   :3.06   Mean   :3.37
##  3rd Qu.:2.00   3rd Qu.:2.00   3rd Qu.:4.0   3rd Qu.:4.00   3rd Qu.:4.00
##  Max.   :4.00   Max.   :4.00   Max.   :4.0   Max.   :4.00   Max.   :4.00
```

```
## [1] 0
```

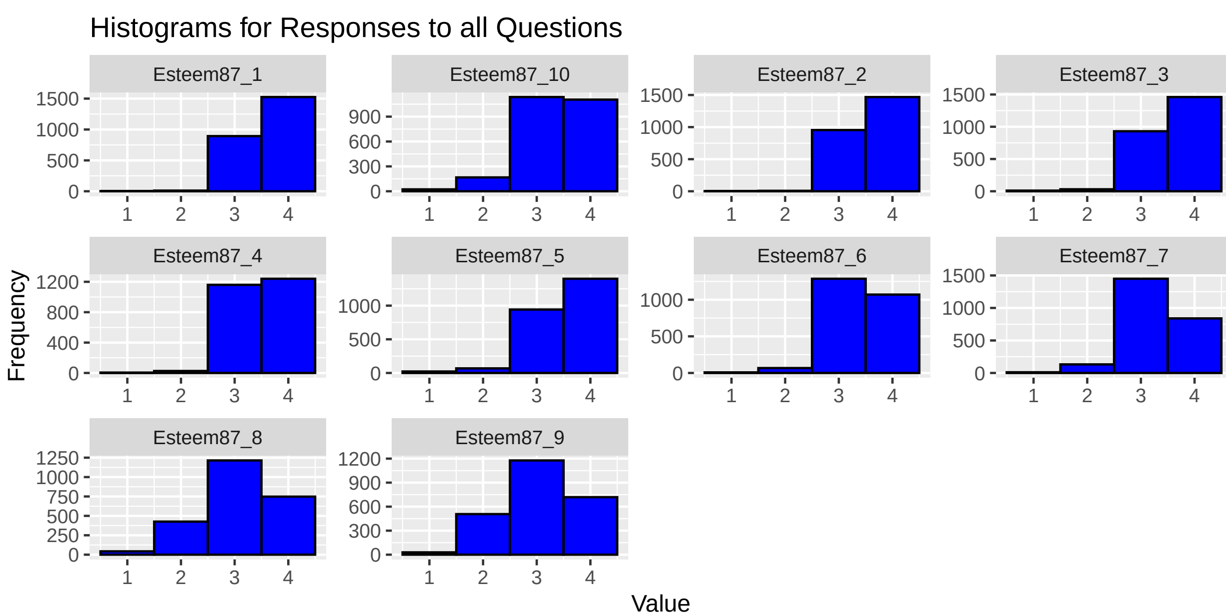
1. Please note that higher scores on Esteem questions 1, 2, 4, 6, and 7 indicate higher self-esteem, whereas higher scores on the remaining questions suggest lower self-esteem. To maintain consistency, consider reversing the scores of certain Esteem questions. For example, if the esteem data is stored 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.

Answer: See .rmd document.

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

Answer: After correcting the responses such that higher scores correspond to higher self-esteem, we see that all of the questions have more weight towards the right, meaning that most participants gave responses that indicate higher self-esteem. However, questions 6 through 9 have a mass that is more centered around the middle, indicating that respondents have slightly lower self-esteem for these questions.

```
##    Esteem87_1      Esteem87_2     Esteem87_3      Esteem87_4     Esteem87_5
##  Min.   :1.00   Min.   :1.0   Min.   :1.00   Min.   :1.0   Min.   :1.00
##  1st Qu.:3.00   1st Qu.:3.0   1st Qu.:3.00   1st Qu.:3.0   1st Qu.:3.00
##  Median :4.00   Median :4.0   Median :4.00   Median :4.0   Median :4.00
##  Mean   :3.62   Mean   :3.6   Mean   :3.58   Mean   :3.5   Mean   :3.53
##  3rd Qu.:4.00   3rd Qu.:4.0   3rd Qu.:4.00   3rd Qu.:4.0   3rd Qu.:4.00
##  Max.   :4.00   Max.   :4.0   Max.   :4.00   Max.   :4.0   Max.   :4.00
##    Esteem87_6      Esteem87_7      Esteem87_8     Esteem87_9      Esteem87_10
##  Min.   :1.00   Min.   :1.00   Min.   :1.0   Min.   :1.00   Min.   :1.00
##  1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.0   1st Qu.:3.00   1st Qu.:3.00
##  Median :3.00   Median :3.00   Median :3.0   Median :3.00   Median :3.00
##  Mean   :3.41   Mean   :3.28   Mean   :3.1   Mean   :3.06   Mean   :3.37
##  3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.0   3rd Qu.:4.00   3rd Qu.:4.00
##  Max.   :4.00   Max.   :4.00   Max.   :4.0   Max.   :4.00   Max.   :4.00
```

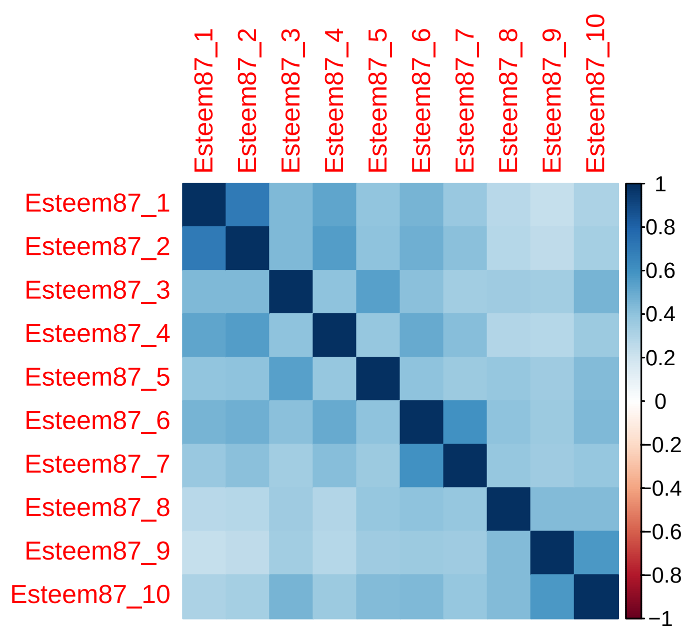


3. Do esteem scores all positively correlated? Report the pairwise correlation table and write a brief summary.

Answer: All esteem scores are positively correlated; the minimum value in each row is positive. Questions that are adjacent to each other (e.g., questions 1 and 2 or questions 6 and 7) tend to have stronger correlations than questions that are further away from each other.

```
## Esteem87_1  Esteem87_2  Esteem87_3  Esteem87_4  Esteem87_5  Esteem87_6
##      0.236       0.259       0.343       0.287       0.354       0.364
## Esteem87_7  Esteem87_8  Esteem87_9 Esteem87_10
##      0.343       0.273       0.236       0.312
```

```
##   Esteem87_1       Esteem87_2       Esteem87_3       Esteem87_4
## Min.   :0.236   Min.   :0.259   Min.   :0.343   Min.   :0.287
## 1st Qu.:0.328   1st Qu.:0.348   1st Qu.:0.365   1st Qu.:0.369
## Median :0.424   Median :0.427   Median :0.427   Median :0.415
## Mean   :0.474   Mean   :0.486   Mean   :0.476   Mean   :0.475
## 3rd Qu.:0.512   3rd Qu.:0.534   3rd Qu.:0.457   3rd Qu.:0.523
## Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000
##   Esteem87_5       Esteem87_6       Esteem87_7       Esteem87_8
## Min.   :0.354   Min.   :0.364   Min.   :0.343   Min.   :0.273
## 1st Qu.:0.381   1st Qu.:0.409   1st Qu.:0.372   1st Qu.:0.309
## Median :0.401   Median :0.453   Median :0.389   Median :0.385
## Mean   :0.468   Mean   :0.508   Mean   :0.465   Mean   :0.425
## 3rd Qu.:0.428   3rd Qu.:0.502   3rd Qu.:0.419   3rd Qu.:0.425
## Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000
##   Esteem87_9      Esteem87_10
## Min.   :0.236   Min.   :0.312
## 1st Qu.:0.303   1st Qu.:0.372
## Median :0.353   Median :0.437
## Mean   :0.421   Mean   :0.475
## 3rd Qu.:0.414   3rd Qu.:0.456
## Max.   :1.000   Max.   :1.000
```



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

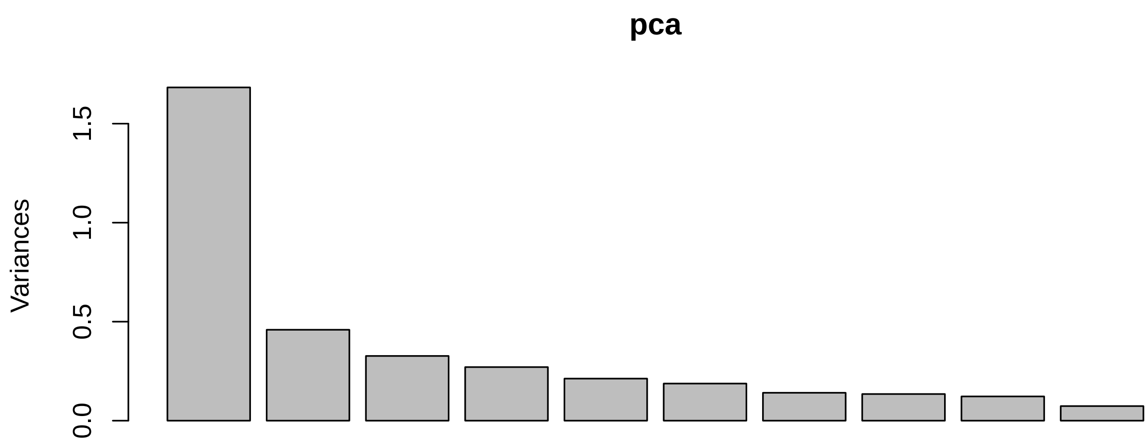
   a) Report the PC1 and PC2 loadings. Are they unit vectors? Are they orthogonal?

Answer: The PC1 loading is (-0.235, -0.244, 0.279, -0.261, 0.312, -0.313, -0.299, 0.393, 0.398, 0.376). The PC2 loading is (0.374, 0.367, -0.149, 0.321, -0.131, 0.209, 0.163, 0.332, 0.578, 0.260). As shown in the table below, PC1 and PC2 both are unit vectors and are orthogonal to each other.

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

	PC1	PC2
Esteem87_1	0.235	-0.374
Esteem87_2	0.244	-0.367
Esteem87_3	0.279	-0.149
Esteem87_4	0.261	-0.321
Esteem87_5	0.312	-0.131
Esteem87_6	0.313	-0.209
Esteem87_7	0.299	-0.163
Esteem87_8	0.393	0.332
Esteem87_9	0.398	0.578
Esteem87_10	0.376	0.260

	PC1	PC2
PC1	1	0
PC2	0	1



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

Answer: We can interpret PC1 as the difference between the total score for Questions 3, 5, 8, 9, 10 and the total score for Questions, 1, 2, 4, 6, 7. We do not see a good interpretation for PC2.

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

Answer:

$$PC1 = -0.235 \times (Q1_score) + -0.244 \times (Q2_score) + 0.279 \times (Q3_score) + -0.261 \times (Q4_score) + 0.312 \times (Q5_score) + -0.313 \times (Q6_score$$

.

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

Answer: Yes, the PC1 and PC2 scores are uncorrelated. As shown in the table, the covariance of PC1 and PC2 is 0, implying that their correlation is 0.

	PC1	PC2
PC1	1.68	0.000
PC2	0.00	0.459

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

Answer: The plot shows that PC1 explains almost 50% of the variation and PC2 explains a little over 10% of the variance. We can also use this scree plot to determine the number of PCs that would be appropriate to use with our data. Using the elbow method, we see that using 2 PCs would be appropriate.

```
![](hw2_sp2024_files/figure-latex/unnamed-chunk-8-1.pdf)<!-- -->
```

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

Answer: Approximately 60% of variance in the data is explained by the first 2 PCs.

```
![](hw2_sp2024_files/figure-latex/unnamed-chunk-9-1.pdf)<!-- -->
```

g) PC's provide us with a low dimensional view of the self-esteem scores. Use a biplot with the first t

Answer: The biplot shows that PC1 roughly corresponds to the difference between the total scores for Questions 3, 5, 8, 9, 10 and the total scores for Questions, 1, 2, 4, 6, 7; PC2 does not have a clear interpretation. We also see that the scores for Questions 1, 2, and 4 are highly correlated; the scores for Questions 3 and 5 are highly correlated; the scores for Questions 6 and 7 are highly correlated; and the scores for Questions 8 and 10 are also highly correlated. This supports are prior analysis from our above EDA, where we saw that questions that are closer to being "adjacent" with each other are more correlated with each other than are questions that are far away from each other. There still does not seem to be a clear interpretation of PC2, but PC2 seems to represent the rough difference between the total scores for questions 8, 9, 10 and questions 1-7.

```
![](hw2_sp2024_files/figure-latex/unnamed-chunk-10-1.pdf)<!-- -->
```

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.

Answer: Using the elbow method, it appears that 2 is a reasonable number of clusters.

```
![](hw2_sp2024_files/figure-latex/unnamed-chunk-11-1.pdf)<!-- -->
```

b) Can you summarize common features within each cluster?

Answer: Cluster 1 is characterized by higher scores for questions 3, 5, 8, 9, and 10. Cluster 2 is characterized by higher scores for questions 1, 2, 4, 6, and 7. This seems to roughly correspond to PC1, which can be interpreted as the approximate difference betwen the total score for questions 3, 5, 8, 9, and 10 and the total score for questions 1, 2, 4, 6, and 7.

```
| Group.1| Esteem87_1| Esteem87_2| Esteem87_3| Esteem87_4| Esteem87_5| Esteem87_6| Esteem87_7| Esteem87_
|-------:|----------:|----------:|----------:|----------:|----------:|----------:|----------:|----------
|       1|       3.89|       3.89|       3.90|       3.82|       3.90|       3.78|       3.63|       3.5
|       2|       3.37|       3.33|       3.28|       3.20|       3.18|       3.06|       2.96|       2.0
```

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

Answer: When we cluster by PC1 and PC2, we see a very clear boundary given by PC1. Specifically, group 1 is almost entirely to the left of PC1 = 0, while group 2 is almost entirely to the right of PC2 = 0.

```
![](hw2_sp2024_files/figure-latex/unnamed-chunk-13-1.pdf)<!-- -->
```

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.
- Personal information: gender, education (05), log(income) in 87, job type in 87. One way to summarize one's weight and height is via Body Mass Index which is defined as the body mass divided by the square of the body height, and is universally expressed in units of kg/m^2. Note, you need to create BMI first. Then may include it as one possible predictor.
- Household environment: Imagazine, Inewspaper, Ilibrary, MotherEd, FatherEd, FamilyIncome78. Do set indicators `Imagazine`, `Inewspaper` and `Ilibrary` as factors.
- You may use PC1 of ASVAB as level of intelligence

Answer: See .rmd file for code. I created a dataframe that contains all of the above listed variables. Income87 has some negative values, so I use 0 for any negative values. Since we cannot take the log of 0, I use log1p, which takes the log(1 + [value]). I also merge in columns for PC1 of the Esteem scores and PC1 of the ASVAB scores, for which I ran PCA for the ASVAB data.

```
##     Gender      Education05    LogIncome87    Imagazine Inewspaper Ilibrary
##  female:1199   Min.   : 6.0   Min.   : 0.00   0: 686    0: 339     0: 559  
##  male  :1232   1st Qu.:12.0   1st Qu.: 8.41   1:1745    1:2092     1:1872  
##                Median :13.0   Median : 9.39                                
##                Mean   :13.9   Mean   : 8.13                                
##                3rd Qu.:16.0   3rd Qu.: 9.85                                
##                Max.   :20.0   Max.   :10.99                                
##     MotherEd       FatherEd     FamilyIncome78    PC1_Esteem     PC1_asvab  
##  Min.   : 0.0   Min.   : 0.0   Min.   :    0   Min.   :3.95   Min.   :  0.0  
##  1st Qu.:11.0   1st Qu.:10.0   1st Qu.:11167   1st Qu.:7.48   1st Qu.: 60.9  
##  Median :12.0   Median :12.0   Median :20000   Median :7.98   Median : 89.1  
##  Mean   :11.7   Mean   :11.8   Mean   :21252   Mean   :7.88   Mean   : 85.4  
##  3rd Qu.:12.0   3rd Qu.:14.0   3rd Qu.:27500   3rd Qu.:8.39   3rd Qu.:112.2  
##  Max.   :20.0   Max.   :20.0   Max.   :75001   Max.   :9.74   Max.   :144.2  
```

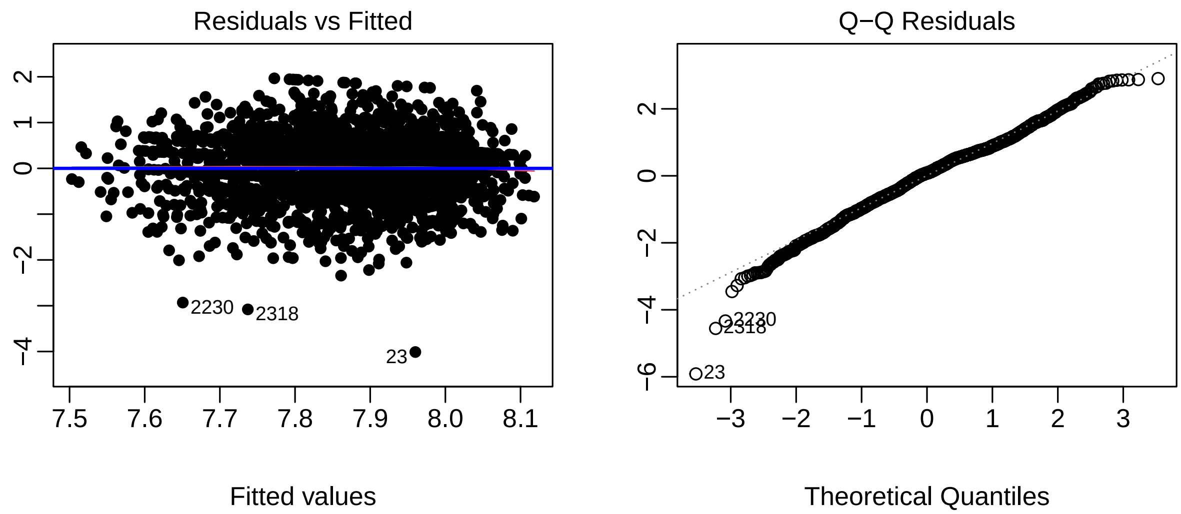
```
b)  Run a few regression models between PC1 of all the esteem scores and suitable variables listed in a

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

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

Answer:

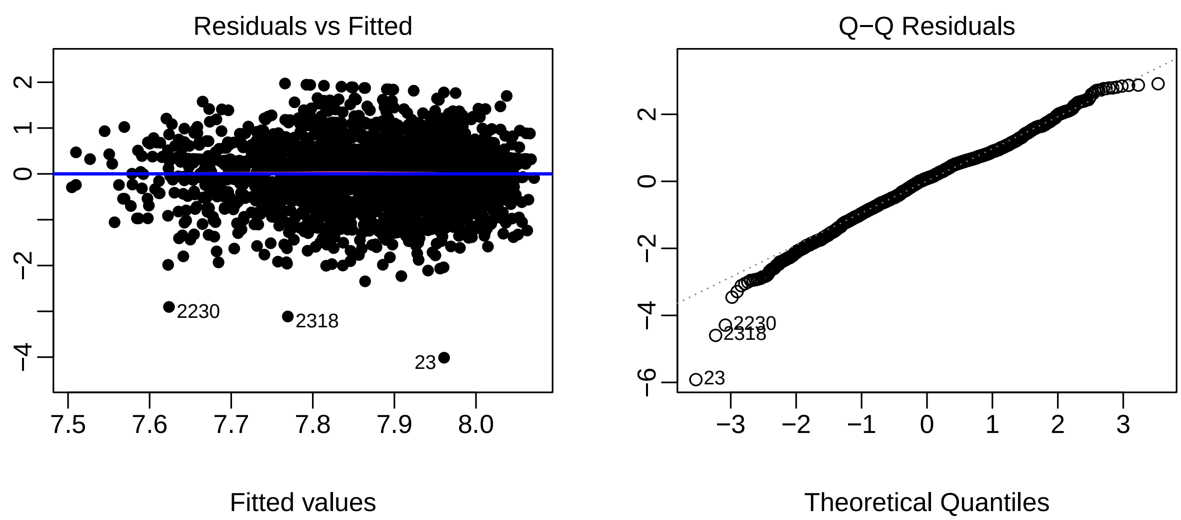
(i) model4 is my final best model. This model uses PC1 of the esteem scores as the dependent variables and the following as explanatory variables: gender, education level in 2005, log income in 1987, family income in 1978, and PC1 of ASVAB scores. I landed this model by running multiple regression models and looking into the residual and QQ plots. In all the models I ran, the plots provided evidence that the linearity and homoscedasity assumptions are met because the residuals follow a symmetric pattern around h=0 and are evenly distributed within a band. The QQ plot also provided evidence of the normality assumption being met due to the presence of a well fitted straight line. I decided not to use statistical significance as a criterion because the significance did not seem very informative – in the model that incorporated all variables, the intercept, IMagazine, and PC1 of the ASVAB scores were statistically significant. Intuitively, whether a family reads a magazine or not should not affect one's self-esteem. Reading a magazine could be correlated to income, and if this is the case, then we would not need to include this redudant variable. Additionally, Ward, Greenhill, & Bakke (2010) show that statistically significant models can actually have very low predictive power; thus, choosing variables based on statistical significance does not seem like a well-informed criterion.

(ii) When we use model4 as the final best model, we see that gender, education, income, family income, and intelligence all affect one's self-esteem. Specifically, being male and having more education, a higher income in 1987, a higher family income in 1978, and a higher score on the ASVAB as captured by PC1 is associated with having a higher self-esteem, as captured by PC1 of the 1987 esteem scores.

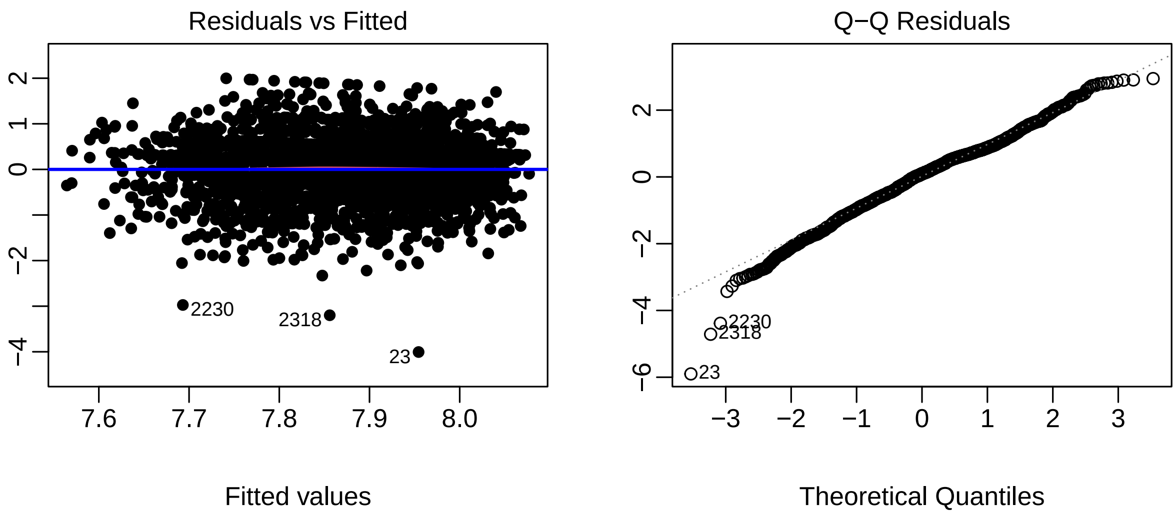
```
##
## Call:
## lm(formula = PC1_Esteem ~ Gender + Education05 + LogIncome87 +
##     Imagazine + Inewspaper + Ilibrary + MotherEd + FatherEd +
##     FamilyIncome78 + PC1_asvab, data = data87)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.011 -0.434  0.056  0.446  1.966
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.45e+00   9.34e-02   79.73  < 2e-16 ***
## Gendermale      1.25e-02   2.80e-02    0.45  0.65468
## Education05     2.59e-03   7.09e-03    0.37  0.71431
## LogIncome87     5.46e-03   4.41e-03    1.24  0.21535
## Imagazine1      3.53e-02   3.41e-02    1.03  0.30078
## Inewspaper1     8.83e-02   4.37e-02    2.02  0.04330 *
## Ilibrary1      -2.56e-02   3.48e-02   -0.74  0.46123
## MotherEd        3.55e-03   7.08e-03    0.50  0.61623
## FatherEd        4.13e-04   5.27e-03    0.08  0.93756
## FamilyIncome78  1.22e-06   1.10e-06    1.11  0.26769
## PC1_asvab       2.16e-03   5.79e-04    3.74  0.00019 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.679 on 2420 degrees of freedom
## Multiple R-squared:  0.026,  Adjusted R-squared:  0.022
## F-statistic: 6.47 on 10 and 2420 DF,  p-value: 6.56e-10
```



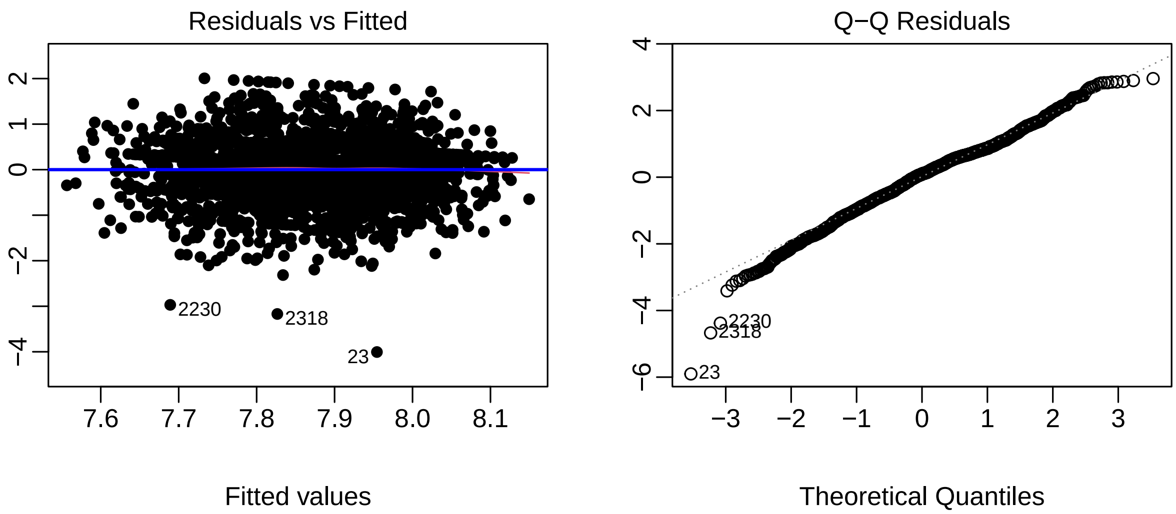

```
##
## Call:
## lm(formula = PC1_Esteem ~ Gender + Education05 + LogIncome87 +
##     Inewspaper + PC1_asvab, data = data87)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.012 -0.428  0.059  0.447  1.972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.457714   0.087216   85.51  < 2e-16 ***
## Gendermale  0.016207   0.027854    0.58    0.561
## Education05 0.004496   0.006857    0.66    0.512
## LogIncome87 0.006116   0.004389    1.39    0.164
## Inewspaper1 0.103231   0.041417    2.49    0.013 *
## PC1_asvab   0.002438   0.000548    4.45  9.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.679 on 2425 degrees of freedom
## Multiple R-squared:  0.0245, Adjusted R-squared:  0.0225
## F-statistic: 12.2 on 5 and 2425 DF,  p-value: 1.1e-11
```




```
##
## Call:
## lm(formula = PC1_Esteem ~ Gender + Education05 + LogIncome87 +
##     PC1_asvab, data = data87)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.006 -0.424  0.063  0.450  1.997
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.510986   0.084647   88.73   <2e-16 ***
## Gendermale  0.016824   0.027882    0.60     0.55
## Education05 0.005190   0.006859    0.76     0.45
## LogIncome87 0.006401   0.004392    1.46     0.15
## PC1_asvab   0.002711   0.000538    5.04    5e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.679 on 2426 degrees of freedom
## Multiple R-squared:  0.022,  Adjusted R-squared:  0.0204
## F-statistic: 13.7 on 4 and 2426 DF,  p-value: 5.22e-11
```




```
##
## Call:
## lm(formula = PC1_Esteem ~ Gender + Education05 + LogIncome87 +
##     FamilyIncome78 + PC1_asvab, data = data87)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.005 -0.424  0.063  0.449  2.005
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.51e+00   8.46e-02   88.77  < 2e-16 ***
## Gendermale     1.48e-02   2.79e-02    0.53     0.59
## Education05    3.87e-03   6.90e-03    0.56     0.58
## LogIncome87    5.81e-03   4.40e-03    1.32     0.19
## FamilyIncome78 1.79e-06   1.06e-06    1.70     0.09 .
## PC1_asvab      2.54e-03   5.47e-04    4.65  3.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.679 on 2425 degrees of freedom
## Multiple R-squared:  0.0232, Adjusted R-squared:  0.0212
## F-statistic: 11.5 on 5 and 2425 DF,  p-value: 5.34e-11
```



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()`.

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

1. Summary and transformation

a) How many patients are there in each sub-type?

- Basal: 208
- Her2: 91
- LumA: 628
- LumB: 233

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

- See plot below

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

- See cleaning procedures below.

d) Apply PCA to the transformed data. How many PCs should we use and why?

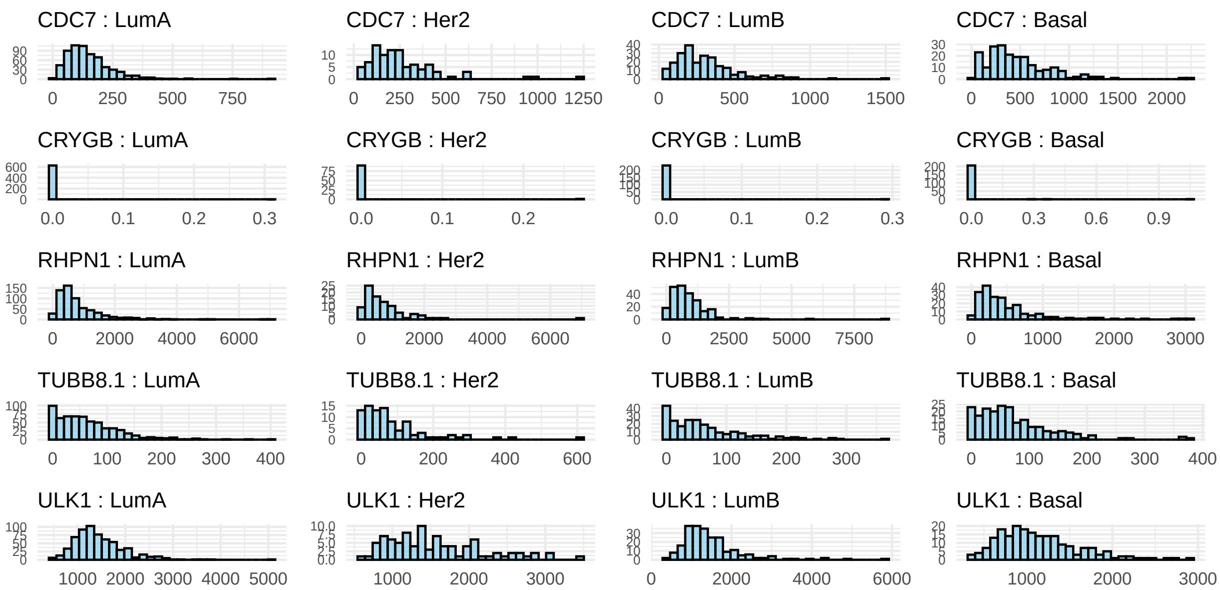
- According to the scree plot, it would be best to use 4 PCs. The plot shows a clear elbow after the 4th PC, and the proportion of variance explained looks significantly lower after the 4th PC.

```
################
## Histograms ##
################

set.seed(124)
random_genes <- sample(colnames(brca)[-1], 5, replace = FALSE)

## Histogram function
plot_gene_histograms <- function(data, genes) {
  plots <- list()
  for (gene in genes) {
    for (subtype in unique(data$BRCA_Subtype_PAM50)) {
      subset_data <- data[data$BRCA_Subtype_PAM50 == subtype, ]
      p <- ggplot(subset_data, aes_string(x=gene)) +
        geom_histogram(bins=30, fill="skyblue", color="black", alpha=0.7) +
        labs(title=paste(gene, ":", subtype), x=NULL, y=NULL) + # Remove redundant axis titles
        theme_minimal() +
        theme(legend.position="none", # Hide legend
              plot.title=element_text(size=10), # Reduce title size
              axis.text.x=element_text(size=8), # Reduce axis text size
              axis.text.y=element_text(size=6)) # Reduce axis text size
      plots[[paste(gene, subtype)]] <- p
    }
  }
  return(plots)
}

# Generate and arrange plots
plots <- plot_gene_histograms(brca, random_genes)
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
```

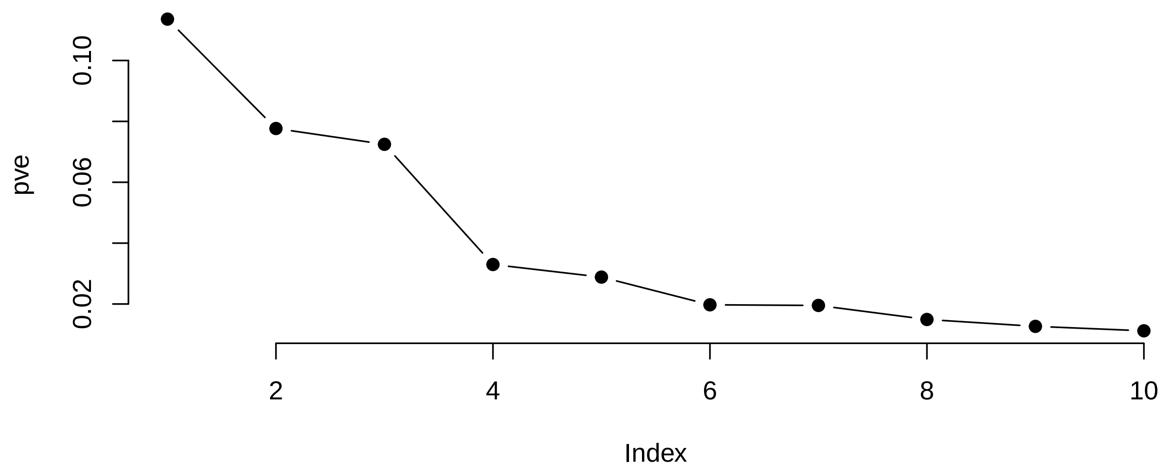
```
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
plot_grid <- do.call(patchwork::wrap_plots, c(plots, ncol = 4))

# Display the arranged plot grid
plot_grid
```

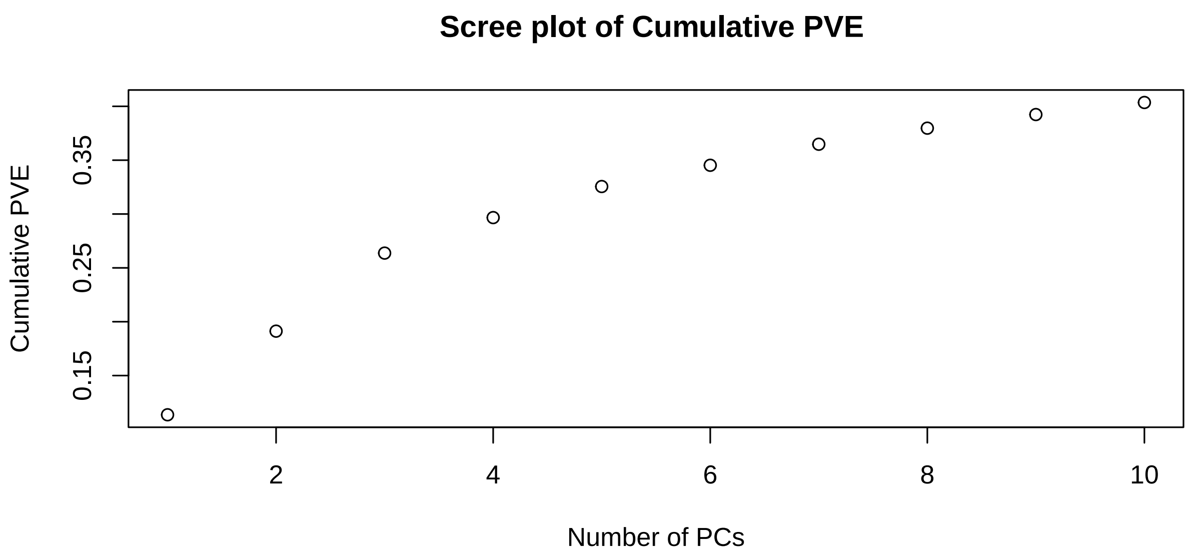



```r
################
### Cleaning ###
################

library(dplyr)

## Filter out genes with only zero counts and no variability
# genes_to_keep <- brca %>%
#   select(-BRCA_Subtype_PAM50) %>%
#   summarise(across(everything(), ~any(. != 0) & sd(.) != 0)) %>%
#   select_if(~ . == TRUE) %>%
#   names()

#length(genes_to_keep)
#
# brca_clean <- brca %>%
#   select(BRCA_Subtype_PAM50, all_of(genes_to_keep))
#
# # Apply logarithmic transformation
# brca_transformed <- brca_clean %>%
#   mutate(across(-BRCA_Subtype_PAM50, log1p))
#
# # Save the transformed data
#
```

```
# saveRDS(brca_transformed, "data/brca_transformed.rds")


brca_transformed <- readRDS("data/brca_transformed.rds")

################
##### PCA ######
################

brca_pca <- prcomp(brca_transformed %>% select(-BRCA_Subtype_PAM50), scale. = T, center=TRUE)
brca_original <- prcomp(brca %>% select(-BRCA_Subtype_PAM50), scale. = F, center=TRUE)
#
# brca_pca$rotation <- brca_pca$rotation[, 1:20]
# brca_pca$x <- brca_pca$x[, 1:20]
# saveRDS(brca_pca, "data/brca_pca.rds")

brca_pca <- readRDS("data/brca_pca.rds")


# Scree Plot of PVE
pve <- summary(brca_pca)$importance[2, 1:10]
plot(pve, type="b", pch = 19, frame = FALSE)
```




```
# Scree Plot of CPVE
plot(summary(brca_pca)$importance[3, 1:10], ylab="Cumulative PVE", xlab="Number of PCs", main="Scree pl
```



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.

```
##
## brca_subtype   1   2   3   4
##        Basal   1 190  17   0
##        Her2   41  18   9  23
##        LumA  350   0  71 207
##        LumB   36   2  22 173
```

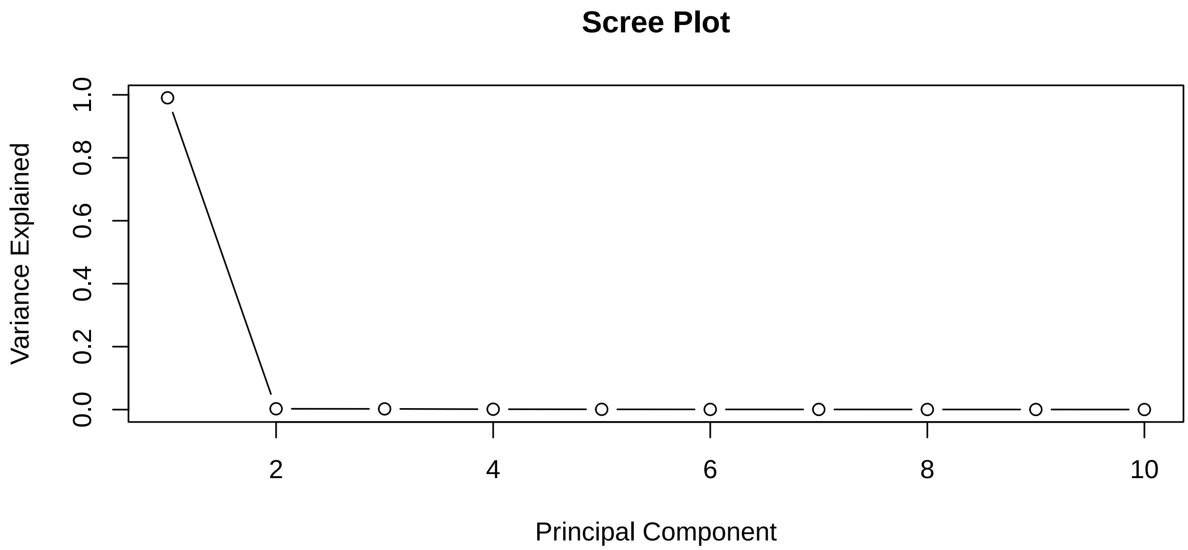
3. Spectrum clustering: to scale or not to scale?

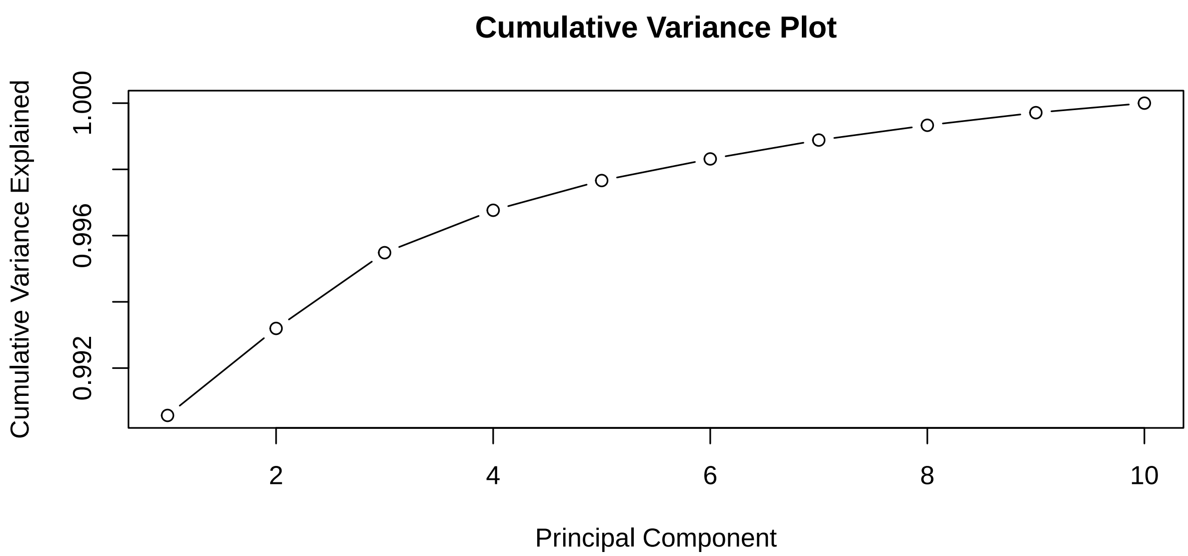
   a) Apply PCA on the centered and scaled dataset. How many PCs should we use and why? You are encouraged to use `irlba::irlba()`. **In order to do so please review the section about SVD in PCA module.**

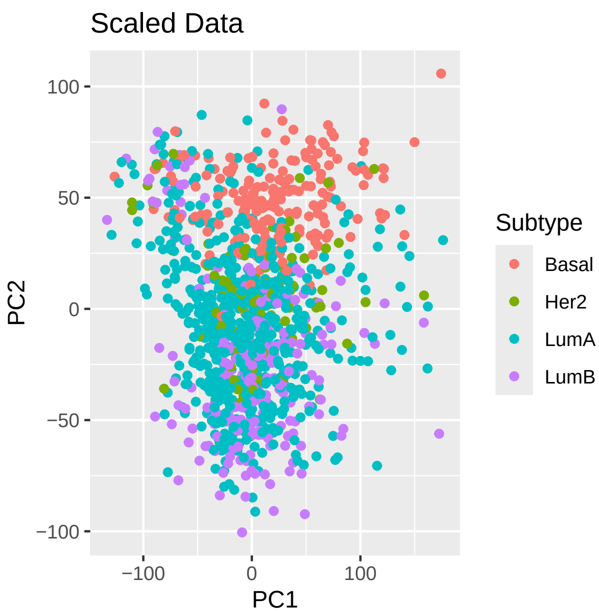
   - According to the scree plots, it would be best to use 2 PCs.

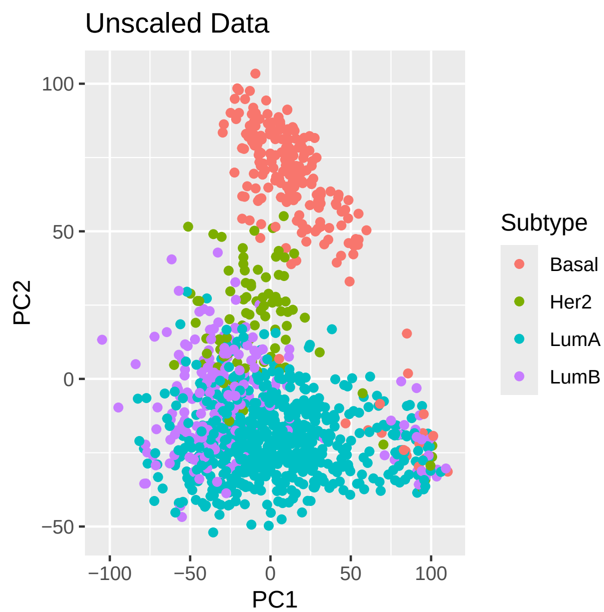
   b) Plot PC1 vs PC2 of the centered and scaled data and PC1 vs PC2 of the centered but unscaled data side by side. Should we scale or not scale for clustering process? Why? (Hint: to put plots side by side, use `gridExtra::grid.arrange()` or `ggpubr::ggrrange()` or `egg::ggrrange()` for ggplots; use `fig.show="hold"` as chunk option for base plots)

   - We should definitely unscale the data. The plot of the centered and scaled data does not show a clear separation between the clusters, while the plot of the centered and unscaled data shows a clear separation between the clusters.


Scree Plot
Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
Principal Component


Cumulative Variance Plot
Cumulative Variance Explained
0.992
0.996
1.000
2
4
6
8
10
Principal Component





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.

- I chose to use 4 clusters because the elbow method suggests that 4 clusters would be appropriate.

b) Choose an optimal cluster number and apply kmeans. Compare the real sub-type and the clustering label as follows: Plot scatter plot of PC1 vs PC2. Use point color to indicate the true cancer type and point shape to indicate the clustering label. Plot the kmeans centroids with black dots. Summarize how good is clustering results compared to the real sub-type.

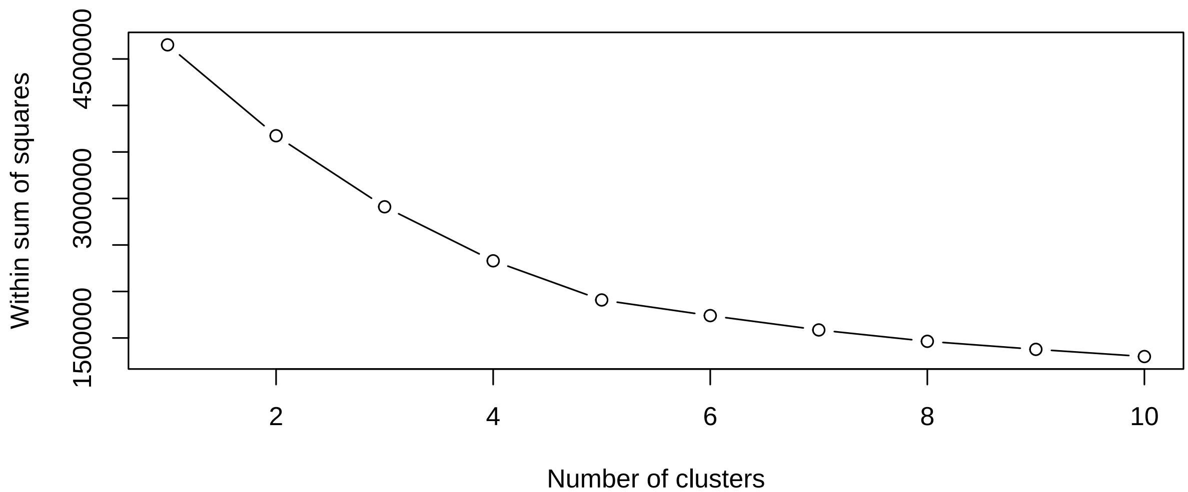
- According to the plot, the clustering results are respectable (in my limited opinion). I do see some overlap between the clusters, but the true gene types have a lot of overlap which will make it difficult to cluster the data.

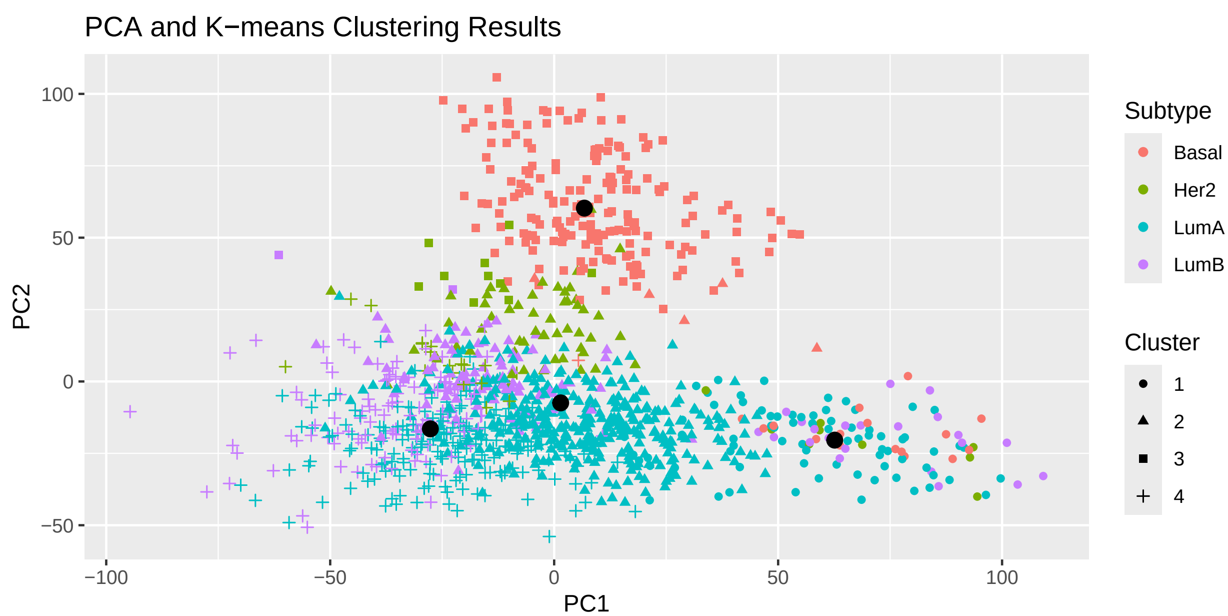
c) Compare the clustering result from applying kmeans to the original data and the clustering result from applying kmeans to 4 PCs. Does PCA help in kmeans clustering? What might be the reasons if PCA helps?

- The average silhouette score for the PCA-reduced data is higher than that for the original data, suggesting that PCA does help in kmeans clustering. PCA helps in kmeans clustering because it reduces the dimensionality of the data, which can help to reduce noise and make the clusters more distinct.

d) Now we have an x patient with breast cancer but with unknown sub-type. We have this patient's mRNA sequencing data. Project this x patient to the space of PC1 and PC2. (Hint: Rmemeber we remove some gene with no counts or no variablity, take log and centered, then find its PC1 to PC4 score) Plot this patient in the plot in b) with a black dot as well. Calculate the Euclidean distance between this patient and each of the centroid of the cluster. (Don't forget the clusters are obtained by using 4 PC's) Can you tell which sub-type this patient might have?

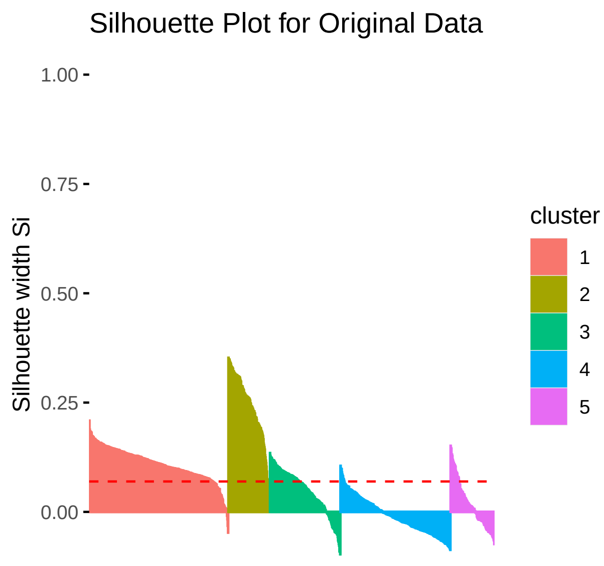
- See below for the graph. I would estimate the sub-type would be LumB, as the patient is closest to cluster 3, which is the cluster that is closest to the LumB sub-type.

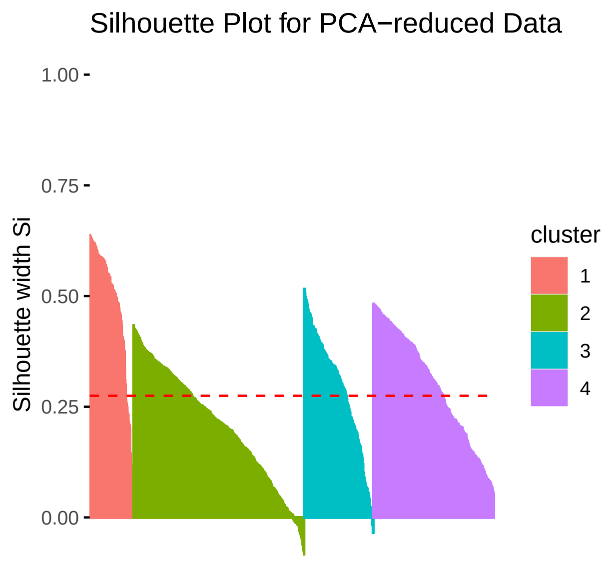





```
##   cluster size ave.sil.width
## 1       1  398          0.11
## 2       2  119          0.24
## 3       3  202          0.05
## 4       4  315         -0.01
## 5       5  126          0.02
```

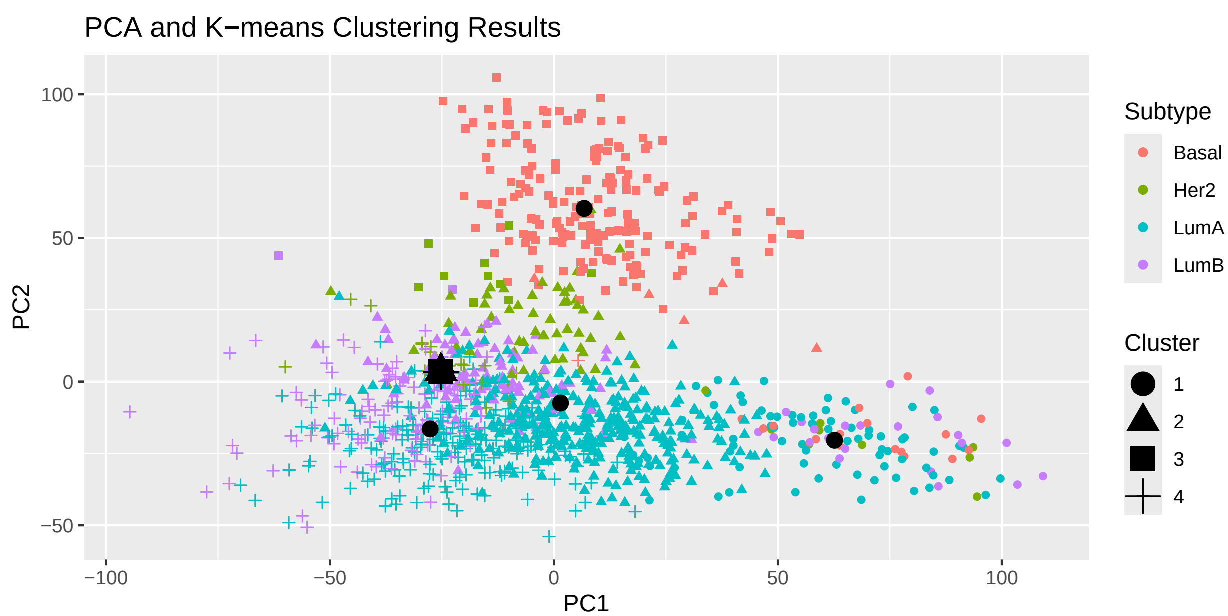
```
##   cluster size ave.sil.width
## 1       1  125          0.47
## 2       2  489          0.21
## 3       3  198          0.29
## 4       4  348          0.29
```






```
## [1] "Average silhouette score for original data: 0.0695609574429012"
```

```
## [1] "Average silhouette score for PCA-reduced data: 0.274817535026146"
```




```
##    1    2    3    4
## 41.1 24.1 66.4 54.4
```

```
## [1] "The patient is closest to cluster: 2"
```

3 Case Study: Fuel Efficiency in Automobiles

Linda will refine this case study by the following Monday, Feb 12th)

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

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.

Answer:

Variables and definitions:

- `mpg`: miles per gallon
- `cylinders`: number of cylinders between 4 and 8
- `displacement`: engine displacement (cubic inches)
- `horsepower`: engine horsepower
- `weight`: vehicle weight (lbs)
- `acceleration`: time to accelerate from 0 to 60 mph (sec)
- `year`: model year (e.g. 70 for 1970)
- `origin`: origin of car (1. American, 2. European, 3. Japanese)
- `name`: vehicle name

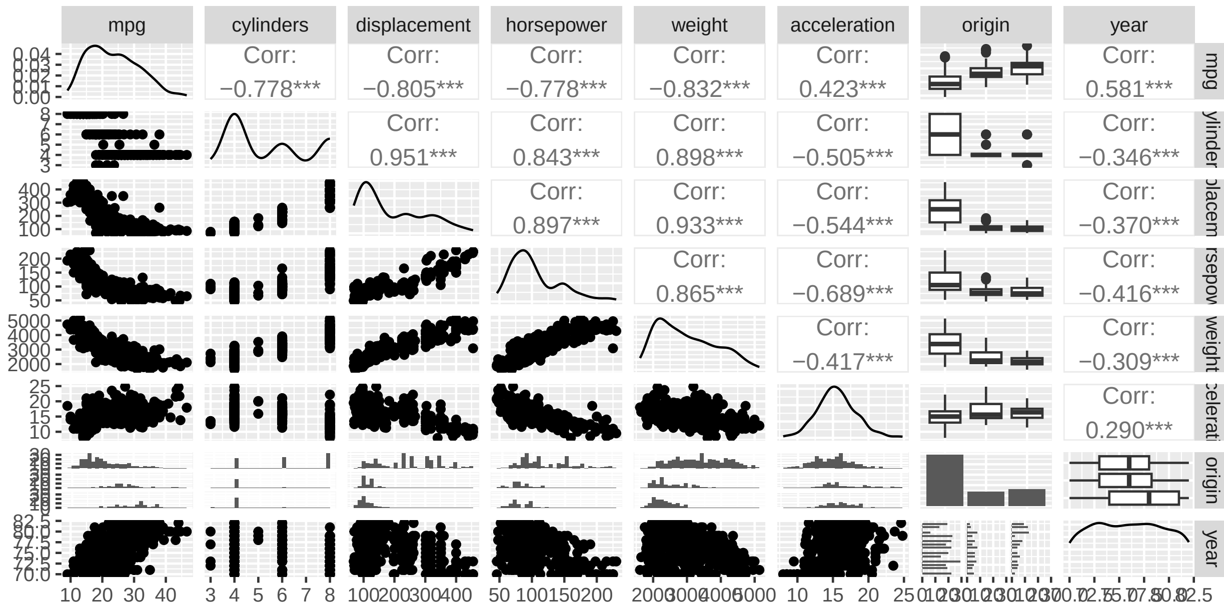
Look to .rmd document for the code on EDA and setting variable to appropriate class.

---

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

Answer: There are 392 cars in the dataset (with 392 unique rows). There are also 301 unique car names in the dataset (that span different model years).

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



Answer:

Summary of variables and findings based on pairwise plots:

- MPG: The distribution of `mpg` is right-skewed, with a few cars having very high `mpg` values. MPG is negatively correlated with `cylinders`, `displacement`, `horsepower`, and `weight`, but positively correlated with `acceleration`. American cars have the lowest `mpg` on average, while Japanese cars have the highest `mpg` on average.
- Cylinders: The distribution of `cylinders` is right-skewed. `cylinders` is positively correlated with `displacement`, `horsepower`, and `weight`, but negatively correlated with `mpg` and `acceleration`. American cars have the highest average number of `cylinders`.
- Displacement: The distribution of `displacement` is right-skewed. `displacement` is positively correlated with `cylinders`, `horsepower`, and `weight`, but negatively correlated with `mpg` and `acceleration`. American cars have the highest average `displacement`.
- Horsepower: The distribution of `horsepower` is right-skewed. `horsepower` is positively correlated with `cylinders`, `displacement`, and `weight`, but negatively correlated with `mpg` and `acceleration`. American cars have the highest average `horsepower`.
- Weight: The distribution of `weight` is right-skewed. `weight` is positively correlated with `cylinders`, `displacement`, and `horsepower`, but negatively correlated with `mpg` and `acceleration`. American cars have the highest average `weight`.
- Acceleration: The distribution of `acceleration` is approximately normal. `acceleration` is positively correlated with `mpg`, but negatively correlated with `cylinders`, `displacement`, `horsepower`, and `weight`.
- Origin: There are more American cars in this dataset. American cars tend to have the highest `cylinders`, `displacement`, `horsepower`, `weight`, and lowest `mpg` on average. Japanese cars tend to have the lowest `cylinders`, `displacement`, `horsepower`, `weight`, and highest `mpg` on average. European cars tend to have similar characteristics to Japanese cars, but not to American cars.
- Year: The distribution of `year` is approximately uniform. `year` is positively correlated with `mpg` and `acceleration`, but negatively correlated with `cylinders`, `displacement`, `horsepower`, and `weight`.

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.

```
##
## Call:
## lm(formula = mpg ~ year, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.021  -5.441  -0.441   4.974  18.209
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -70.0117     6.6452   -10.5   <2e-16 ***
## year          1.2300     0.0874    14.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.36 on 390 degrees of freedom
## Multiple R-squared:  0.337,  Adjusted R-squared:  0.335
## F-statistic:  198 on 1 and 390 DF,  p-value: <2e-16
```

Answer: Yes, year is a significant variable at the .05 level. The coefficient of year is 1.23, which means that for each additional year, the `mpg` of a car increases by 1.23 on average. Note that this does not imply a causal link between `year` and `mpg`, but an association between the two variables.

---

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.

```
##
## Call:
## lm(formula = mpg ~ year + horsepower, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.077  -3.078  -0.431   2.588  15.315
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.73917    5.34903   -2.38    0.018 *
## year          0.65727    0.06626    9.92   <2e-16 ***
## horsepower   -0.13165    0.00634  -20.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.39 on 389 degrees of freedom
## Multiple R-squared:  0.685,  Adjusted R-squared:  0.684
## F-statistic:  424 on 2 and 389 DF,  p-value: <2e-16
```

Answer: Yes, year is still a significant variable at the .05 level. The coefficient of year is 0.65, which means that for each additional year, the `mpg` of a car increases by 0.75 on average, **holding `horsepower` constant**. This is a smaller effect than the one found in the previous model, which suggests that `horsepower` is a confounding variable in the relationship between `year` and `mpg`, and is correlated with power variables.

---

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?

Answer: The difference in the 95% CI's for the coefficient of year between the two models is due to the presence of `horsepower` in the second model. In the first model, there is less precision and more noise in the model since the effect of `horsepower` isn't taken into account. This results in a wider confidence interval for the coefficient of `year` in the first model. In the second model, including `horsepower` in the model makes the model more precise, and so the we can estimate a narrower confidence interval for the coefficient of `year` is narrower.

---

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).

```
##
## Call:
## lm(formula = mpg ~ year * horsepower, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.349  -2.451  -0.456   2.406  14.444
##
## Coefficients:
##                  Estimate Std. Error t value           Pr(>|t|)
## (Intercept)     -126.60885   12.11726  -10.45 <0.0000000000000002 ***
## year               2.19198    0.16135   13.59 <0.0000000000000002 ***
## horsepower         1.04567    0.11537    9.06 <0.0000000000000002 ***
## year:horsepower   -0.01596    0.00156  -10.22 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.9 on 388 degrees of freedom
## Multiple R-squared:  0.752,  Adjusted R-squared:  0.75
## F-statistic:  393 on 3 and 388 DF,  p-value: <0.0000000000000002
```

Answer: Yes, the interaction effect is significant at the .05 level. The direct effect (coefficient of year) has now increased to 2.19, and the interaction effect is -0.016. This means that for each additional year, the `mpg` of a car increases by 2.19 on average, but this effect is reduced by 0.016 for each additional horsepower. This suggests that the effect of `year` on `mpg` is moderated by `horsepower`.

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?

```
##
## Call:
## lm(formula = mpg ~ cylinders, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.241  -3.183  -0.633   2.549  17.917
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   42.916      0.835    51.4 <0.0000000000000002 ***
## cylinders     -3.558      0.146   -24.4 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.91 on 390 degrees of freedom
## Multiple R-squared:  0.605,  Adjusted R-squared:  0.604
## F-statistic:  597 on 1 and 390 DF,  p-value: <0.0000000000000002
```

Answer: Yes, `cylinders` is significant at the .01 level. The coefficient of `cylinders` is -3.56, which means that for each additional cylinder, the `mpg` of a car decreases by 3.56 on average.

However, the interpretation of the model is difficult, becuase the intercept represents zero cylinders, which is not a meaningful value. Also, cylinders only take on discrete values.

---

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`.

```
##
## Call:
## lm(formula = mpg ~ cylinders, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.284  -2.904  -0.963   2.344  18.027
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   20.550      2.349    8.75 < 0.0000000000000002 ***
## cylinders4     8.734      2.373    3.68              0.00027 ***
## cylinders5     6.817      3.589    1.90              0.05825 .
## cylinders6    -0.577      2.405   -0.24              0.81071
```

```
## cylinders8    -5.587      2.395   -2.33             0.02015 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.7 on 387 degrees of freedom
## Multiple R-squared:  0.641,  Adjusted R-squared:  0.638
## F-statistic:  173 on 4 and 387 DF,  p-value: <0.0000000000000002
```

Answer: If `cylinders` is treated as a categorical variable, the OLS regression will treat each factor as a separate indicator variable. Only `cylinders4` is significant at the .01 level. The coefficient of `cylinders4` is 8.73, which means that the `mpg` of a car with 4 cylinders is 8.73 higher on average than a car with 3 cylinders.

It makes sense that only `cyliners4` is significant, since the other levels of `cylinders` have very few observations.

---

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

Answer: The fundamental difference between treating `cylinders` as a continuous and categorical variable is that the continuous model assumes a linear relationship between `cylinders` and `mpg`, while the categorical model assumes that each level of `cylinders` has a different effect on `mpg`.

The other main difference is that the continuous model is difficult to interpret, since the intercept represents zero cylinders, which is not a meaningful value. The categorical model is easier to interpret, since each level of `cylinders` has a separate coefficient.

Personally, we think it is best to use `cylinders` as a categorical model (where linearity is not assumed), with `cylinders4` as the reference level, since it is the most commonly observed level.

---

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?

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cylinders
## Model 2: mpg ~ cylinders
##   Res.Df  RSS Df Sum of Sq    F      Pr(>F)
## 1    390 9416
## 2    387 8544  3       871 13.2 0.000000034 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

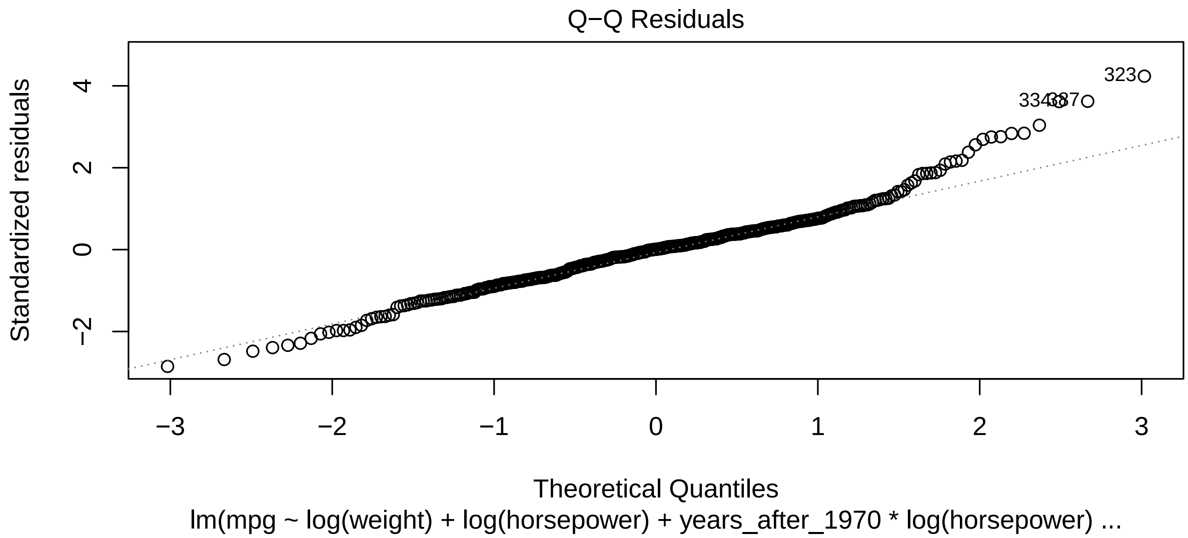
Answer: Yes, we can test the null hypothesis that `mpg` is linear in `cylinders` vs. `mpg` relates to `cylinders` as a categorical variable at the .01 level using an ANOVA test. The p-value of the test is less than .01 with an F-statistic of 13.2, so we reject the null hypothesis and conclude that the model using `cylinders` as a categorical variable explains more of the variation and is significantly different from the model modeling `cylinders` as a linear relationship with `mpg`.

3.4 Results

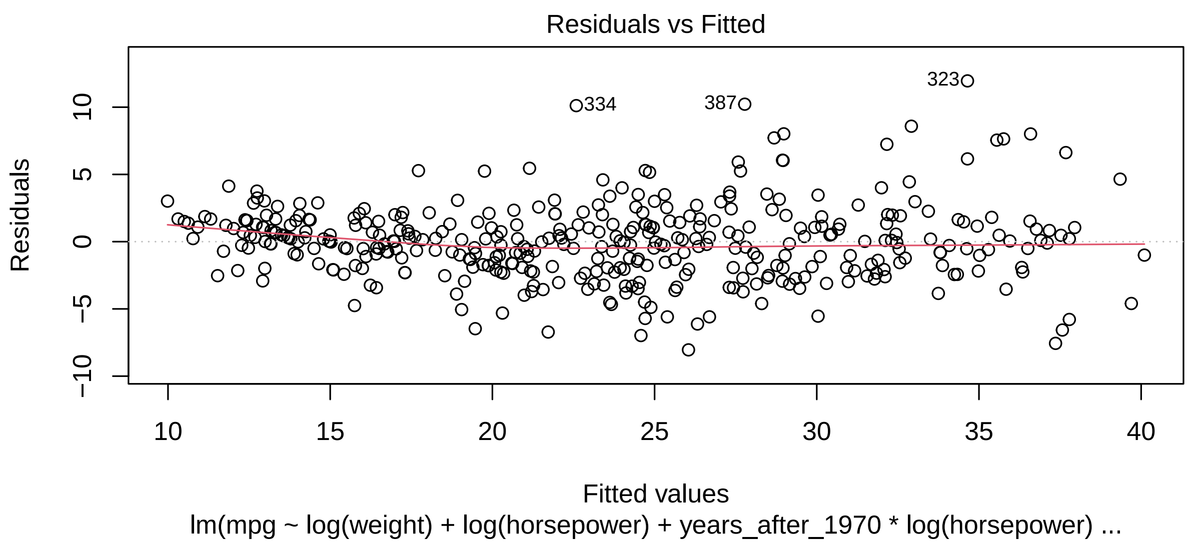
Final modeling question: we want to explore the effects of each feature as best as possible. You may explore interactions, feature transformations, higher order terms, or other strategies within reason. The model(s) should be as parsimonious (simple) as possible unless the gain in accuracy is significant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

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

```
##
## Call:
## lm(formula = mpg ~ log(weight) + log(horsepower) + years_after_1970 *
##     log(horsepower), data = auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.045 -1.866  0.019  1.457 11.954
##
## Coefficients:
##                                  Estimate Std. Error t value
## (Intercept)                       145.298      6.064   23.96
## log(weight)                       -16.943      1.063  -15.94
## log(horsepower)                     1.825      1.034    1.77
## years_after_1970                    5.801      0.551   10.52
## log(horsepower):years_after_1970   -1.116      0.121   -9.21
##                                             Pr(>|t|)
## (Intercept)                      <0.0000000000000002 ***
## log(weight)                      <0.0000000000000002 ***
## log(horsepower)                                0.078 .
## years_after_1970                 <0.0000000000000002 ***
## log(horsepower):years_after_1970 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.84 on 387 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.868
## F-statistic:  642 on 4 and 387 DF,  p-value: <0.0000000000000002
```




```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(ols_7)
## W = 1, p-value = 0.0000003
```




```
##
##  studentized Breusch-Pagan test
##
## data:  ols_7
## BP = 25, df = 4, p-value = 0.00004
```

Answer: The final model is a parsimonious linear regression model with `mpg` as the response variable and `log(weight)`, `log(horsepower)`, `years_after_1970`, and an interaction term between `log(horsepower)` and `years_after_1970` as the predictor variables. With only three transformed variables and an interaction term, the model has an adjusted R-squared of 0.868 (this is compared to a 0.90 adjusted R-squared in a full model with all variables, log terms, squared terms and two-way interactions).

However, the model's residuals are not normally distributed, as shown by the QQ plot and the Shapiro-Wilk test. The residuals are also heteroscedastic, as shown by the residual plot and the Breusch-Pagan test. However, even with non-normal residuals and heteroscedasticity, this only causes problems for inference and calculation of standard errors/confidence intervals; OLS is still a unbiased estimator so it is still a valid model for prediction.

---

b) Summarize the effects found.

Answer:

- `log(weight)` has a negative effect on `mpg`, with a coefficient of -17, which means that (roughly) for every 1% increase in weight, `mpg` decreases by 17
- `log(horsepower)` has a positive effect on `mpg`, with a coefficient of 80, which means that for every 1% increase in horsepower, `mpg` increase by 80. However, one should not interpret this in isolation, because there is an interaction term with `year` that changes the effect of `log(horsepower)` on `mpg` depending on the year of the car.
- `years_after_1970` has a positive effect on `mpg`, with a coefficient of 5.8, which means that for every year increase in the car's model year, `mpg` increases by 5.8. However, this effect is not constant, as it interacts with `log(horsepower)`.
- The interaction term between `years_after_1970` and `log(horsepower)` has a negative effect on `mpg`, with a coefficient of -1.1, which means that for every 1% increase in horsepower `mpg` decreases by 1.1 AND this effect multiplicatively increases for each year after 1970. In essence, the effect of `log(horsepower)` on `mpg` is less negative for older cars and more negative for newer cars in the dataset.

---

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.

```
##    fit  lwr  upr
## 1 9.65 6.95 12.3
```

Answer: The predicted `mpg` of the new car is 9.65 with a 95% confidence interval of (6.95, 12.3).

The extremely low predicted `mpg` is likely due to the fact that the car is high in all of the variables that are negatively associated with `mpg` (weight, horsepower, displacement and cylinders). However, we should also be extremely careful, the model is extrapolating outside of the range of the data, and should not be trusted. For example, the data has no values for cars with a horsepower over 230 and no values for cars made in 1983, so the model is making predictions based on the assumption that the relationships between the variables are constant outside of the range of the data, which is not a safe assumption.

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 $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$, using the `lm()` function. What are the true values of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$? 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 $\boldsymbol{\beta}_1$? Does this confidence interval capture the true $\boldsymbol{\beta}_1$?

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?

# 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 $\boldsymbol{\beta}_1$ (stored in `results$b1`). Does the sampling distribution agree with theory?

ii. How many of your 95% confidence intervals capture the true $\boldsymbol{\beta}_1$? Display your confidence intervals graphically.