Modern Data Mining, HW 2

Group Member 1 - Jenea Adams Group Member 2 - Annan Timon

Due: 11:59 PM, Sunday, 02/12

Contents

1 Packages

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 models 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.

1.1 Objectives

- PCA
- SVD
- Clustering Analysis
- Linear Regression

1.2 Review materials

- Study Module 2: PCA
- Study Module 3: Clustering Analysis
- Study Module 4: Multiple regression

1.3 Data needed

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

2 Case study 1: Self-seteem

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”

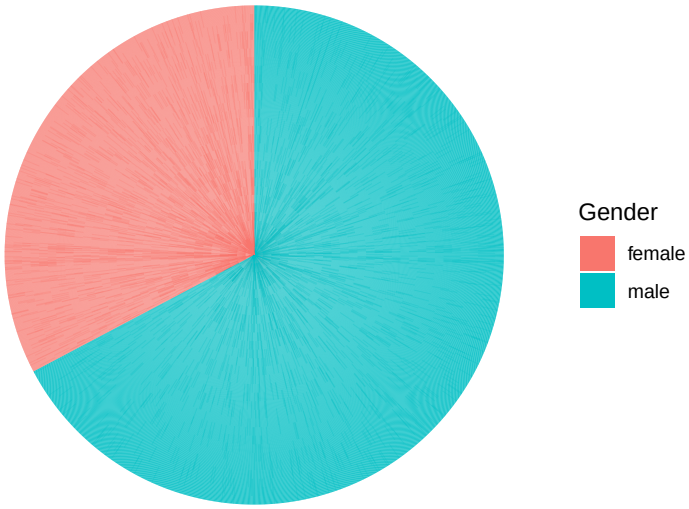
2.1 Data preparation

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?

```
# temp <- read.csv('data/NLSY79.csv', header = T, stringsAsFactors = F)
# # missing values? real variables vs. factors? are varable values reasonable?
str(nlsy)
summary(nlsy)

levels(as.factor(nlsy$Job05))
table(as.factor(nlsy$Job05))
```

2.1.1 Subject demographics

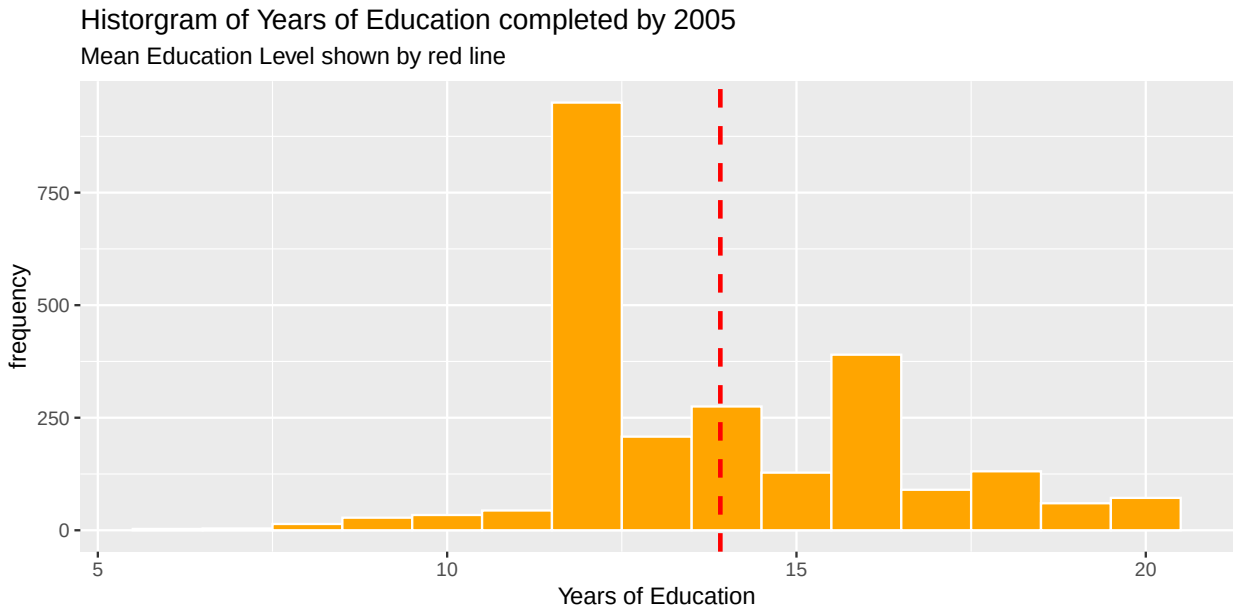
```
gender_pie = nlsy %>%
  ggplot(aes(x = "", y = Gender, fill = Gender)) +
  geom_bar(stat ="identity", width = 1) +
  coord_polar("y", start=0) +
  theme_void()

gender_pie
```



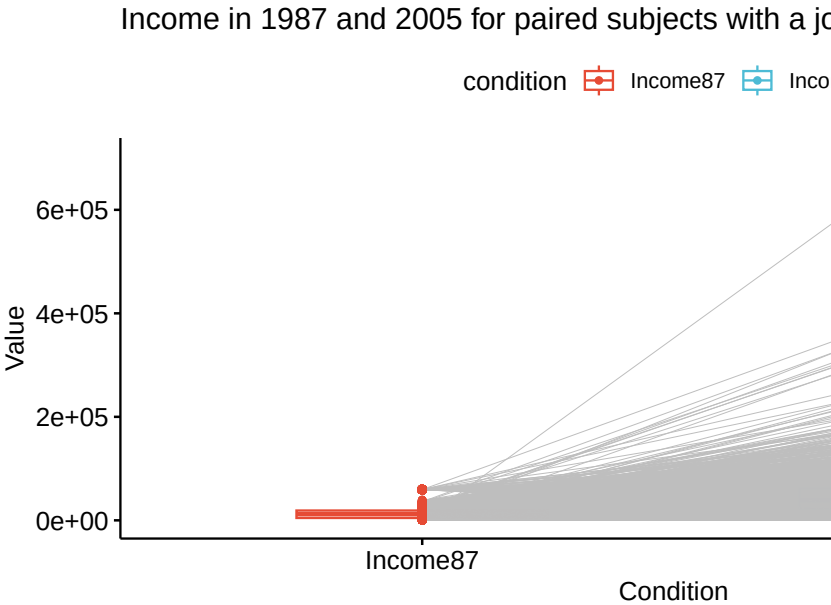

```
edu_hist = nlsy %>%
  ggplot() +
  geom_histogram(aes(x = Education05), bins = 15, color ="white", fill = "orange") +
  labs(title = "Historgram of Years of Education completed by 2005", subtitle = "Mean Education Level sh
  geom_vline(aes(xintercept=mean(Education05)),
             color = "red", linetype = "dashed", size = 1)
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
```

```
edu_hist
```

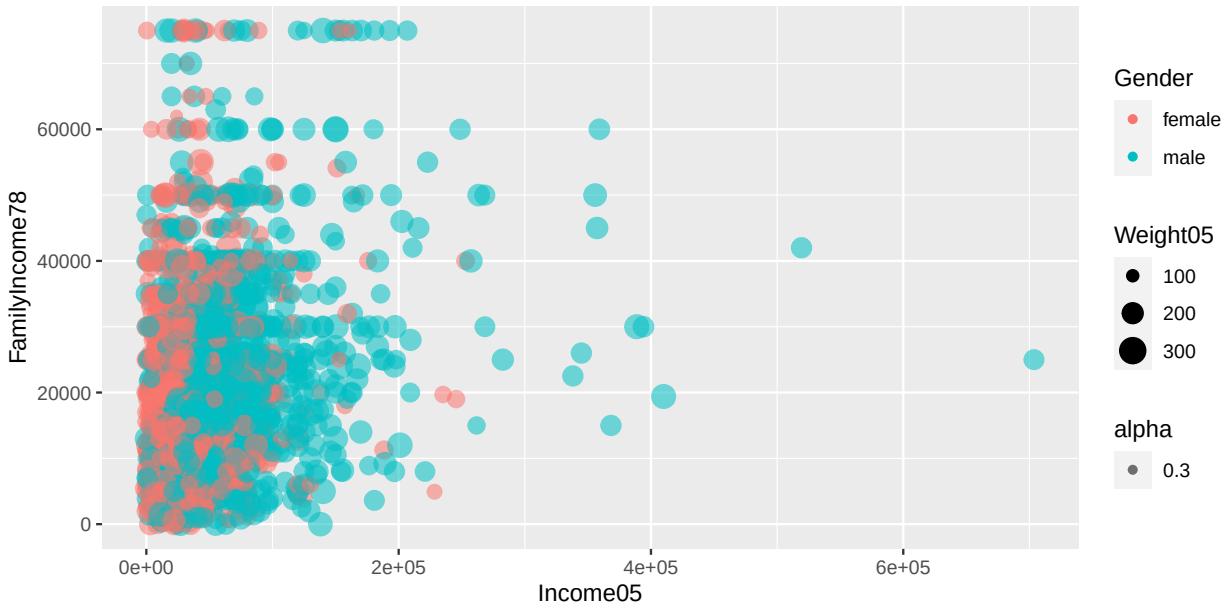


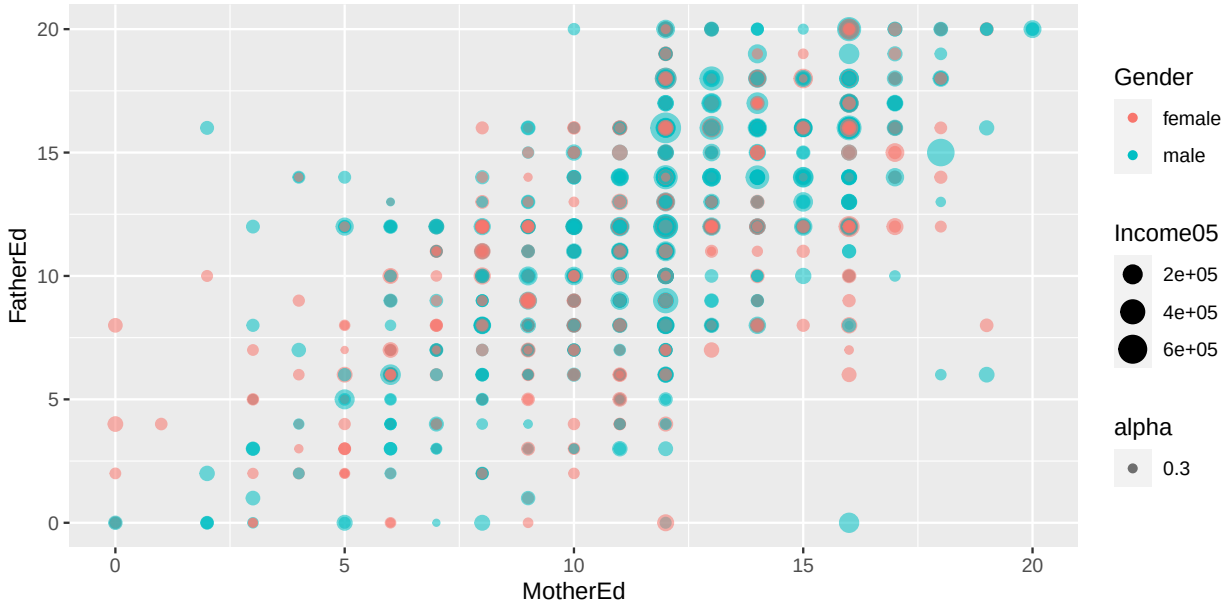
```
job_bar = nlsy %>%
  # filter(!Job05=="") %>%
  ggplot(aes(x = Job05, fill = Job05)) +
  geom_bar() +
  theme(legend.position = "none",
        # adjust for margins around the plot; t: top; r: right; b: bottom; l: left,
        axis.text.x = element_text(angle = -90, vjust = 0, hjust = 0))

job_bar
```

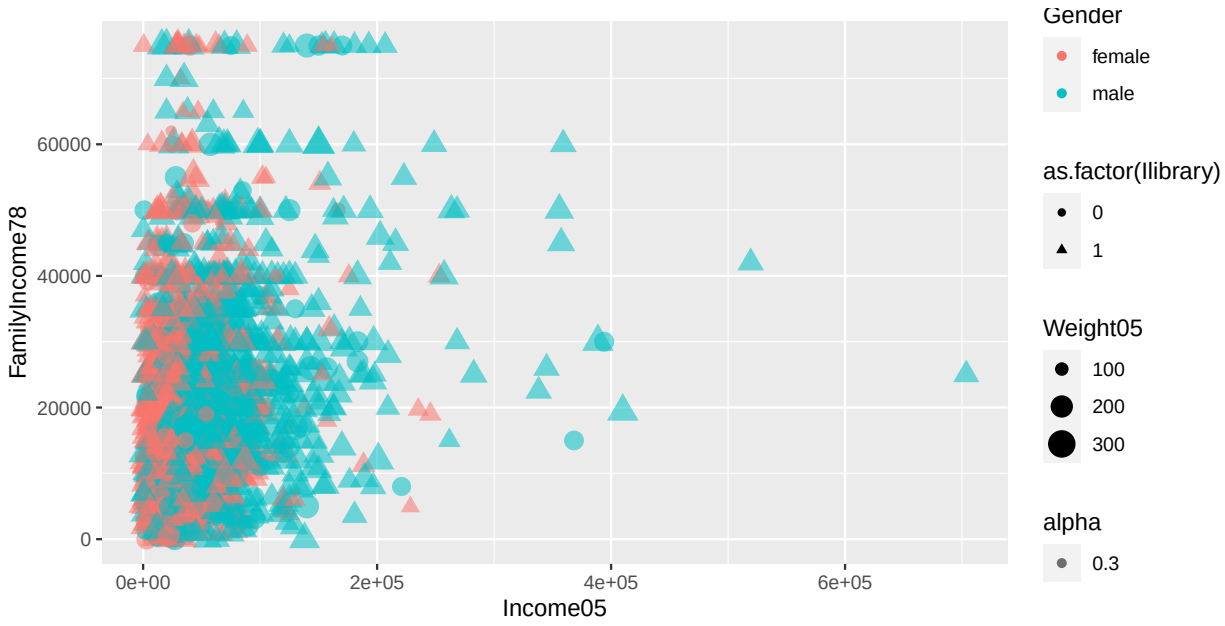


paired visualization of income in 87 and income in 05

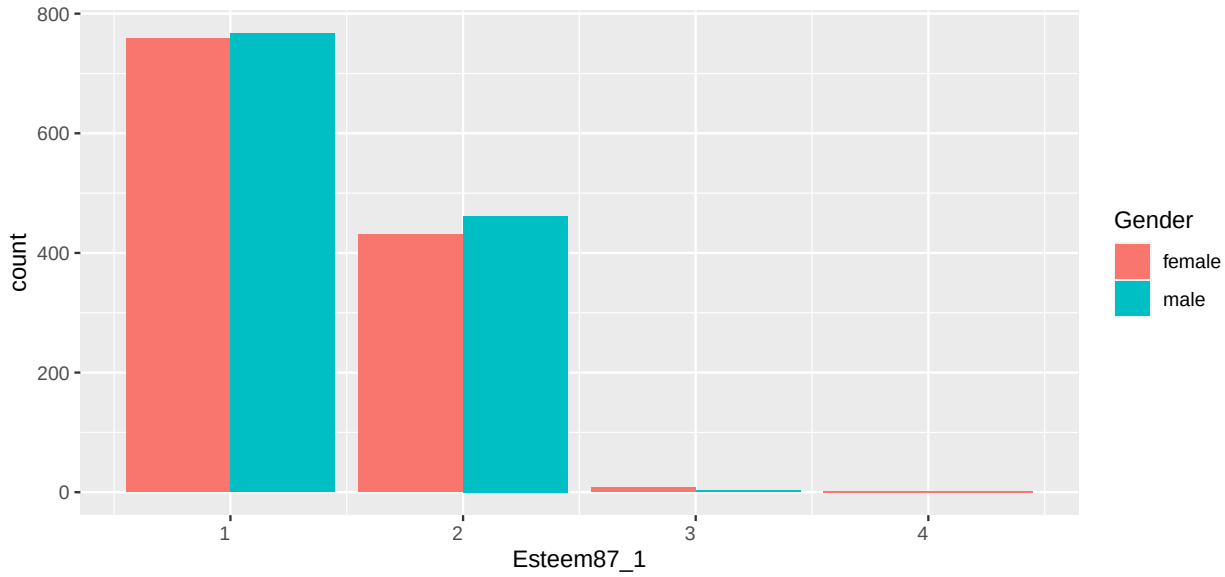
Family income to 87 and 2005 income + gender + weight

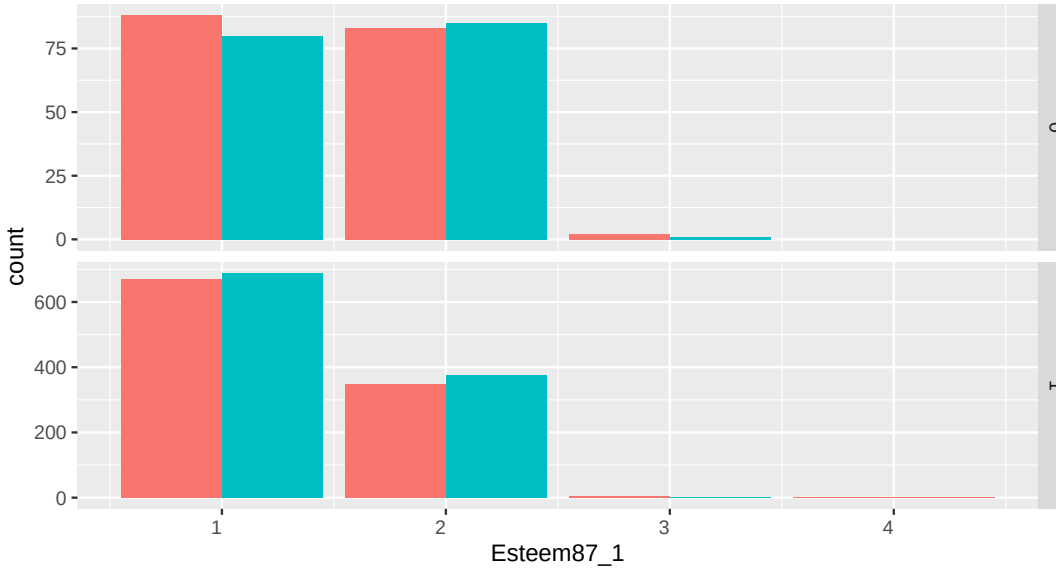

FamilyIncome78
60000
40000
20000
0
0e+00
2e+05
4e+05
6e+05
Income05
Gender
female
male
Weight05
100
200
300
alpha
0.3


FatherEd
20
15
10
5
0
0
5
10
15
20
MotherEd
Gender
female
male
Income05
2e+05
4e+05
6e+05
alpha
0.3

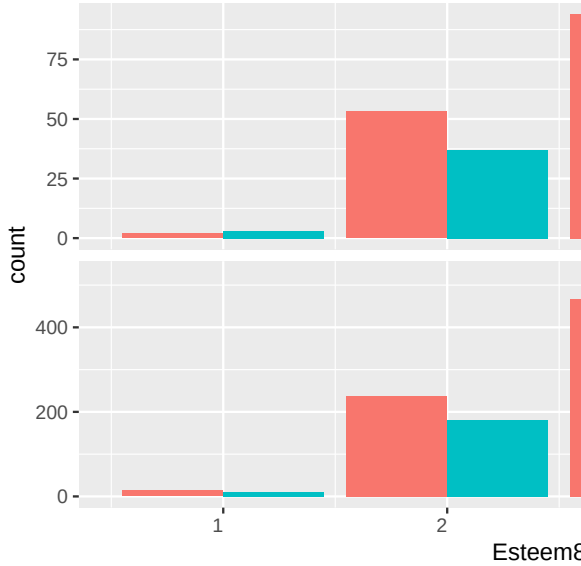


How many people believe they are a person of worth in 87, and how is this affected by gender and income in 05?

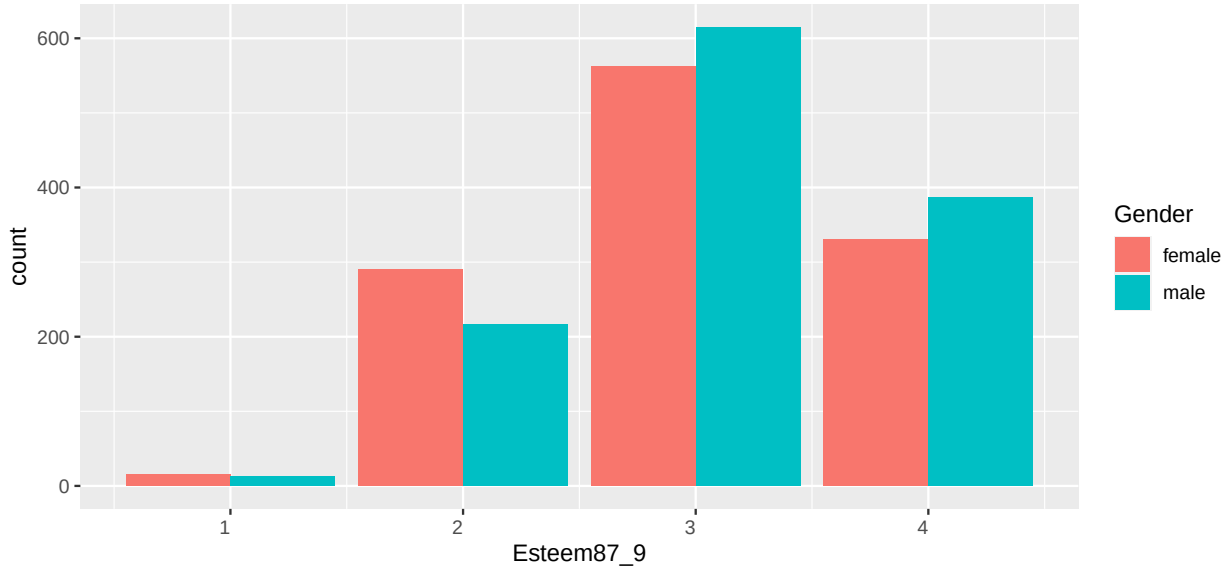


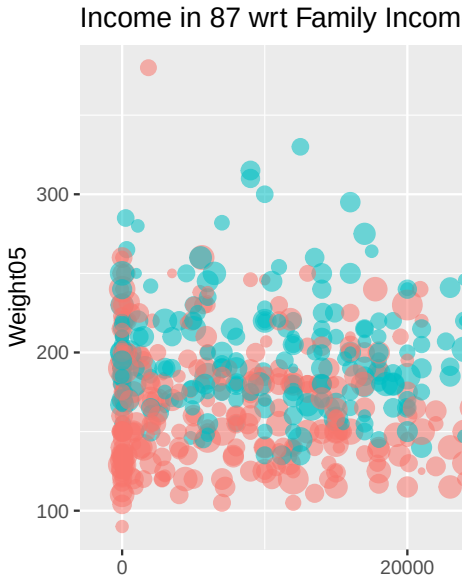


by parents reading the newspaper



feeling useless: 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree





Of those who agreed with feeling useless in 87, what was their income in 87 and weight

2.1.2 Missing values

From looking at the jobs barplots, it's posisble those with missing values could come from missing job types. These people may be unemployed, which is different than having a job that was "uncodeable". Therefore, I'm deciding not to remove thise entries as it could still be informative given there are several individuals who report income as 0

2.2 Self esteem evaluation

Let concentrate on Esteem scores evaluated in 87.

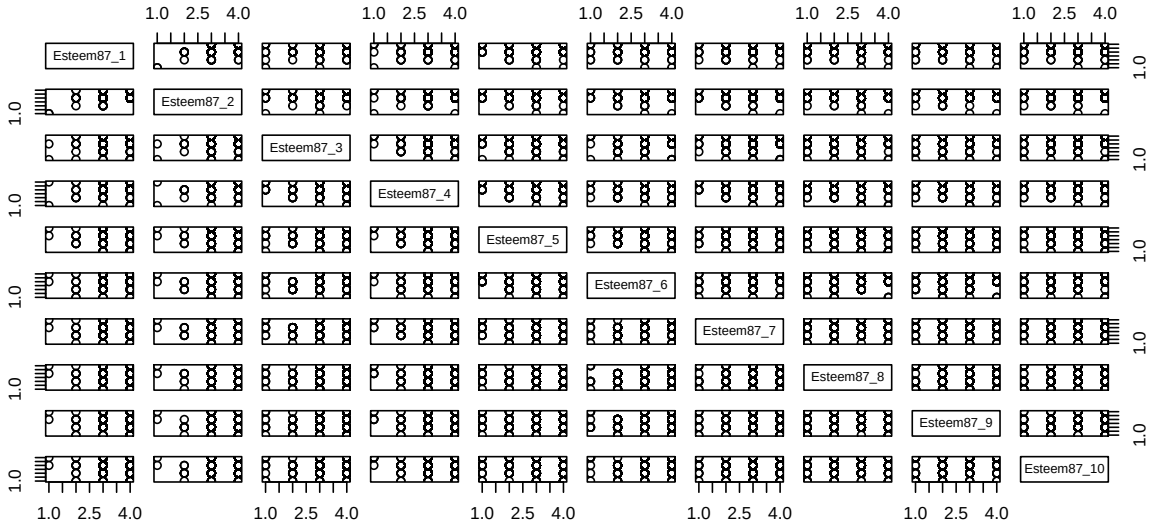
The following 10 questions are answered as 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree

- 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” _

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.

There appear to be no blanks or NAs

1. Reverse Esteem 1, 2, 4, 6, and 7 so that a higher score corresponds to higher self-esteem. (Hint: if we store the esteem data in `data.esteem`, then `data.esteem[, c(1, 2, 4, 6, 7)] <- 5 - data.esteem[, c(1, 2, 4, 6, 7)]` to reverse the score.)
2. Write a brief summary with necessary plots about the 10 esteem measurements.
3. Are esteem scores all positively correlated? Report the pairwise correlation table and write a brief summary.



The esteem scores are all positively correlated.

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

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

```
data.esteem.centered = scale(data.esteem, center = T, scale = F)
data.esteem.centered <- as.data.frame(data.esteem.centered)
```

```
round(sapply(data.esteem.centered,mean), 3) # col mean with 3 decimals
sapply(data.esteem, mean) # col mean
sapply(data.esteem.centered, sd) #col sd
sapply(data.esteem, sd) # col sd
```

Here we see the new mean centered at 0 but the standard deviation is the same as the uncentered data

All loadings are perpendicular and with unit 1

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

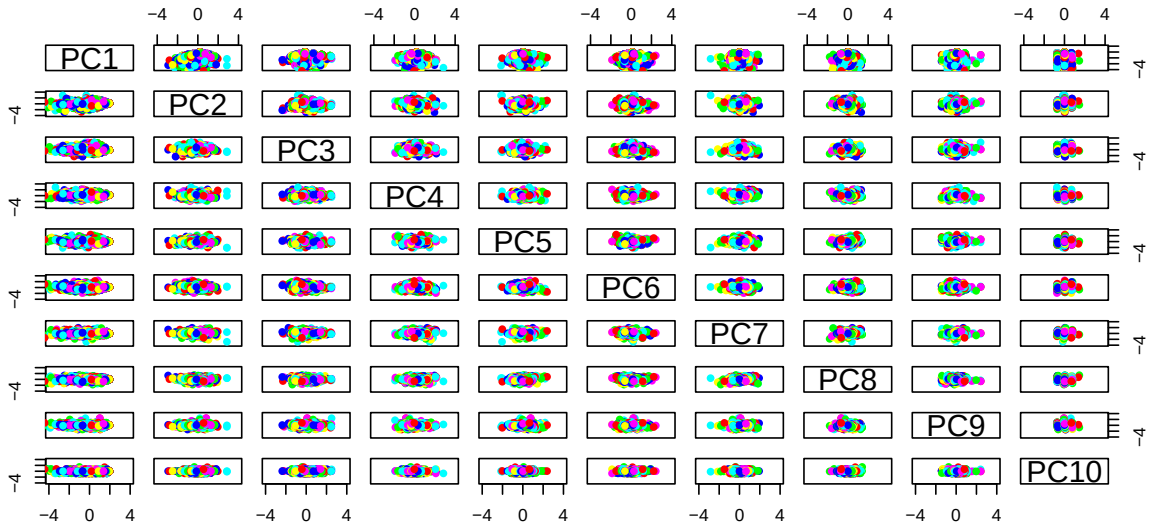
Loadings determine the contribution of each variable to the PCs.
PC1 loadings are all positively correlated with the esteem features, given their positive loadings. PC2 had loadings which are mostly negatively correlated with the esteem features, besides the final three.

These loadings could indicate that the 6 loadings in PC1 with values over 0.300 contribute more to PC1. Also Esteem87_9's loading in PC2 is one of three positively correlated with PC2 and also has the strongest contribution to the PC.

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

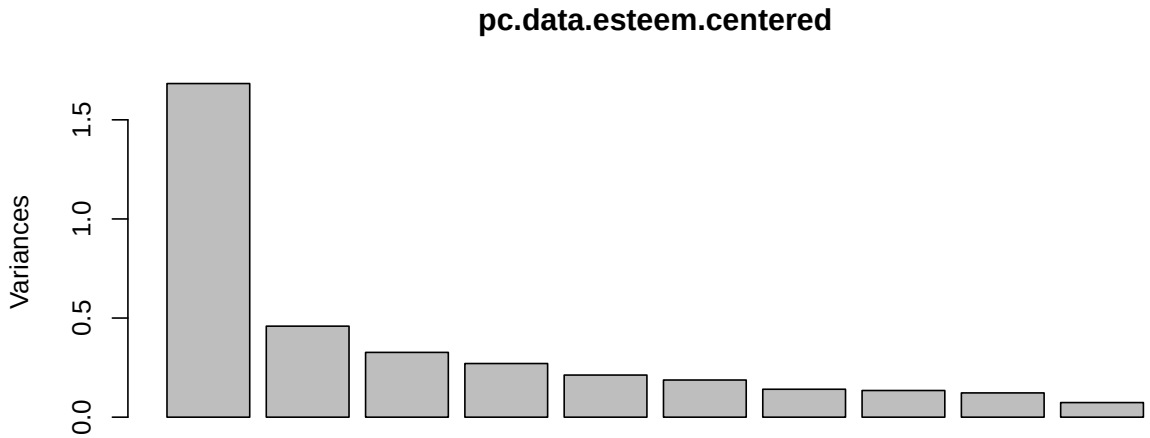
PC1 = 0.235 x Esteem87_1 + 0.244 x Esteem87_2 + 0.279 x Esteem87_3 + 0.261 x Esteem87_4 + 0.312 x Esteem87_5 + 0.313 x Esteem87_6 + 0.299 x Esteem87_7 + 0.393 x Esteem87_8 + 0.398 x Esteem87_9 + 0.376 x Esteem87_10

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

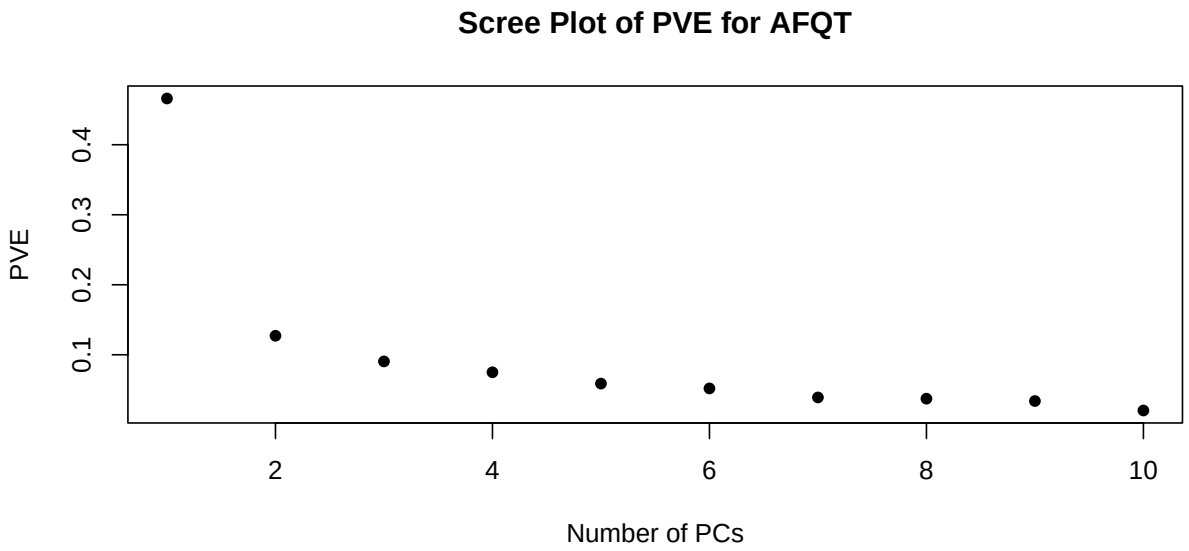


There appears to be no correlation between the PCs

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

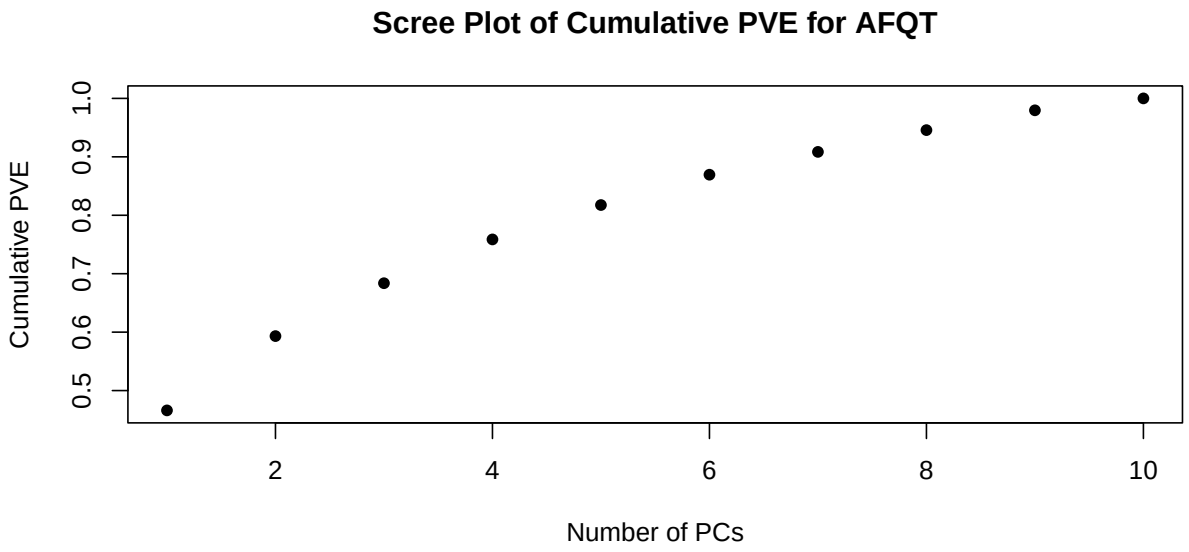


PC1 captures a large percetnage of the variance and each subsequent princial component explaines lessa variance than the preceeding one.



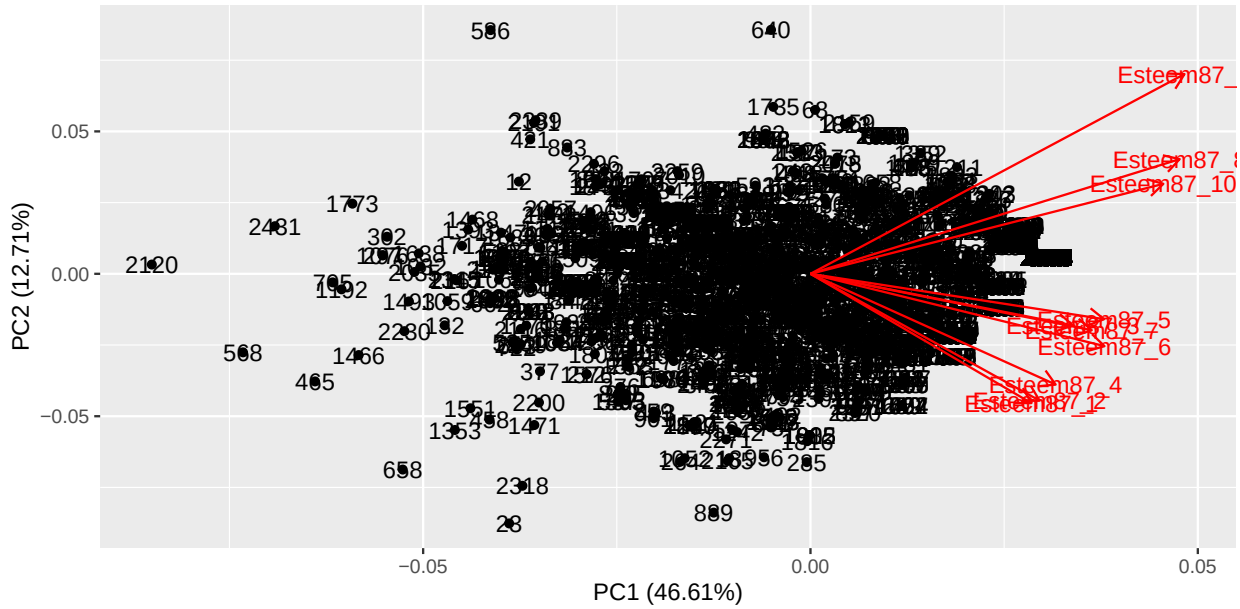
PC1 captures about 40% of the variance and PC2 about 10%

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



Cumulatively, about 60% of the variance in this data is explained by the first 2 PCs

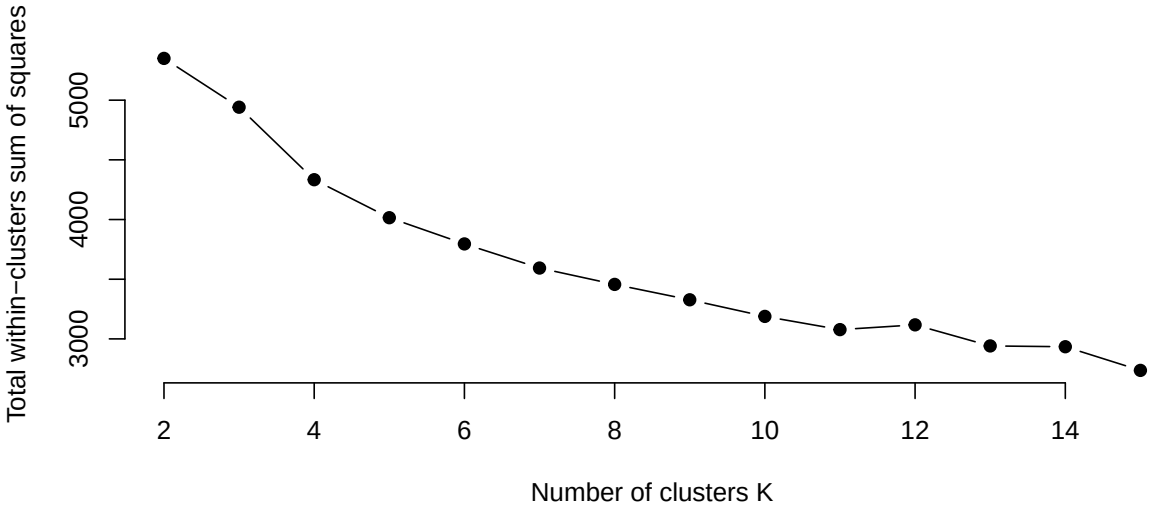
g) PC's provide us with a low dimensional view of the self-esteem scores. Use a biplot with the first tw

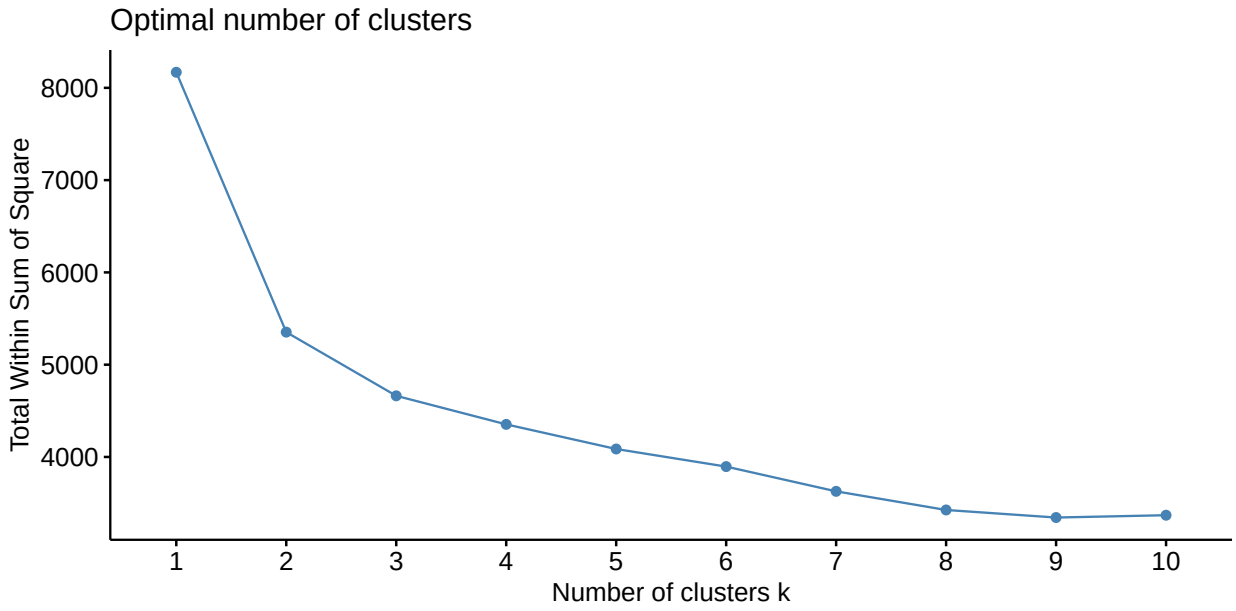


This biplot shows that the PC1 loadings are similarin magnitudes and signs. It also shows that PC2 shoudl capture the differences in responses to questions 8, 9 and 10 from the other questions. This also means questions 8, 9, and 10 are more correlated than the other questions (features) .

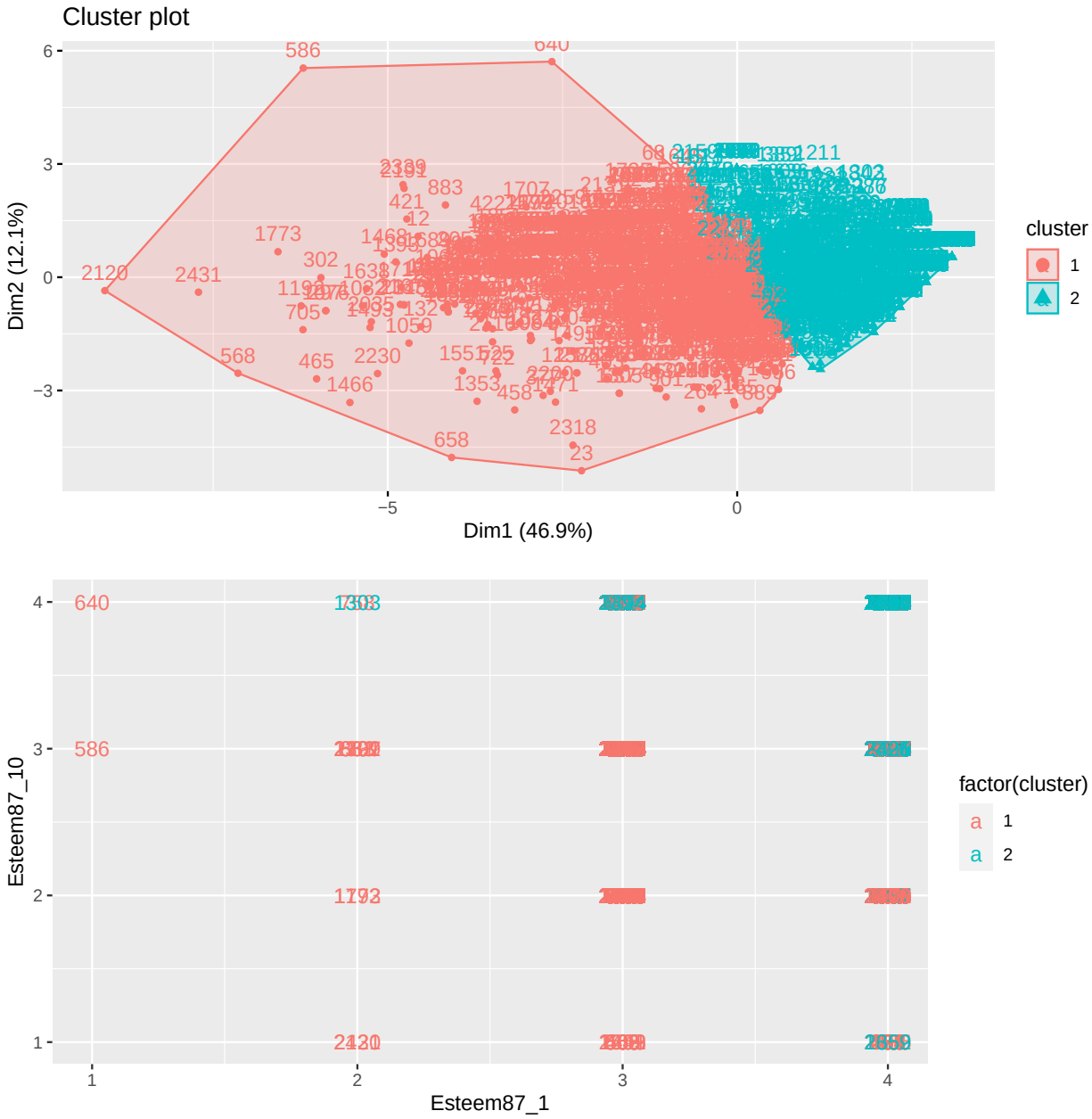
5. Apply k-means to cluster subjects on the original esteem scores

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



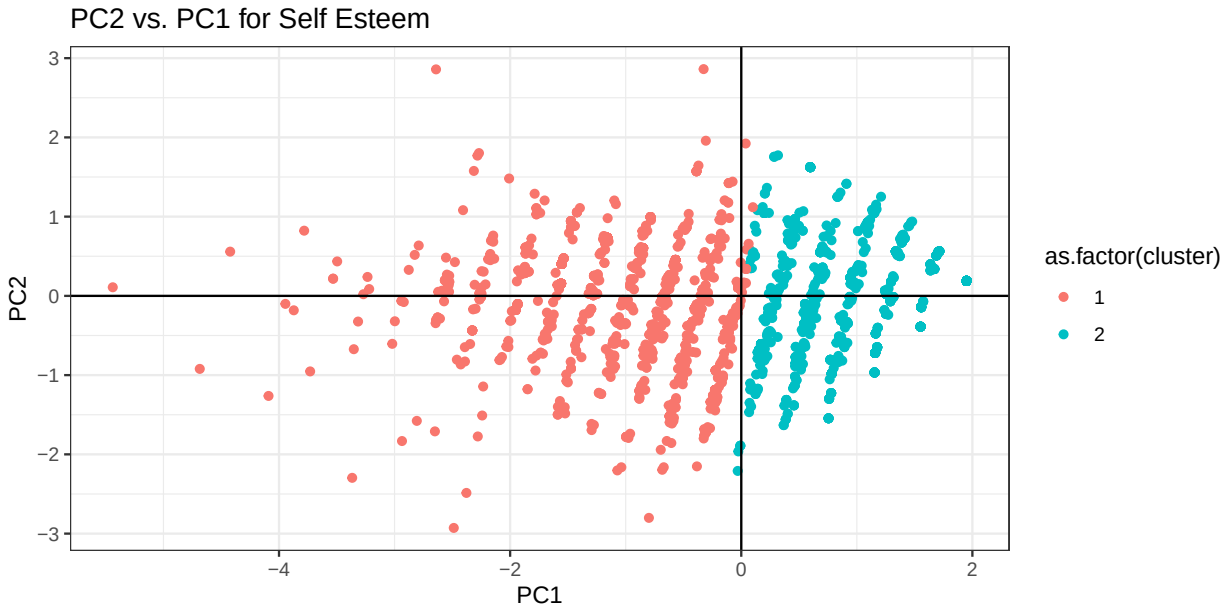


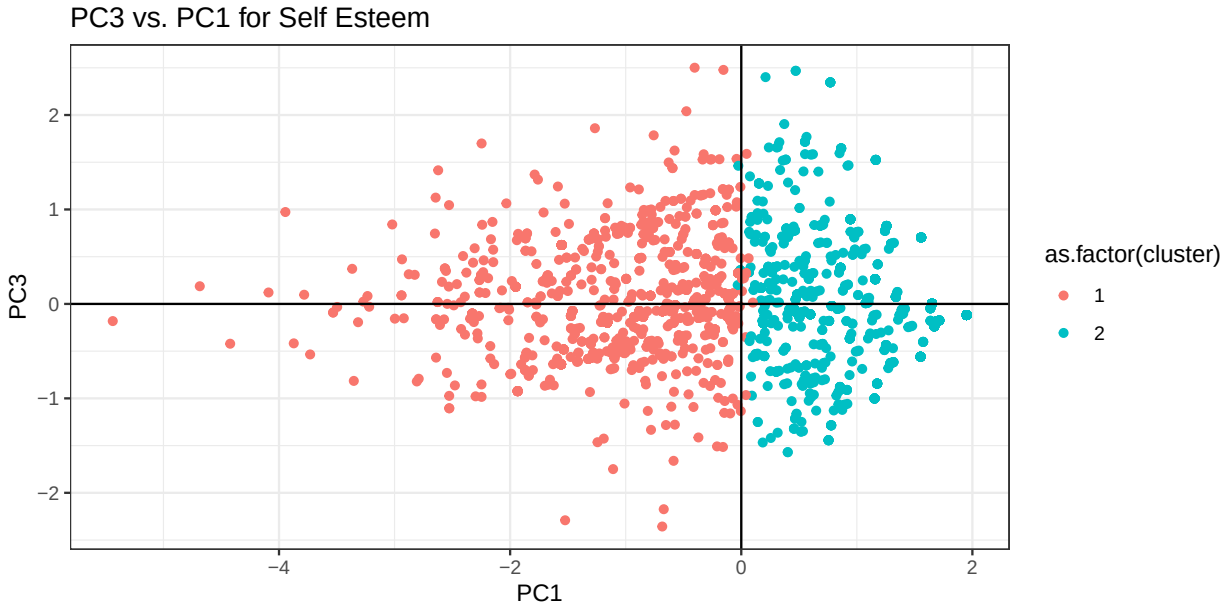
b) Can you summarize common features within each cluster?

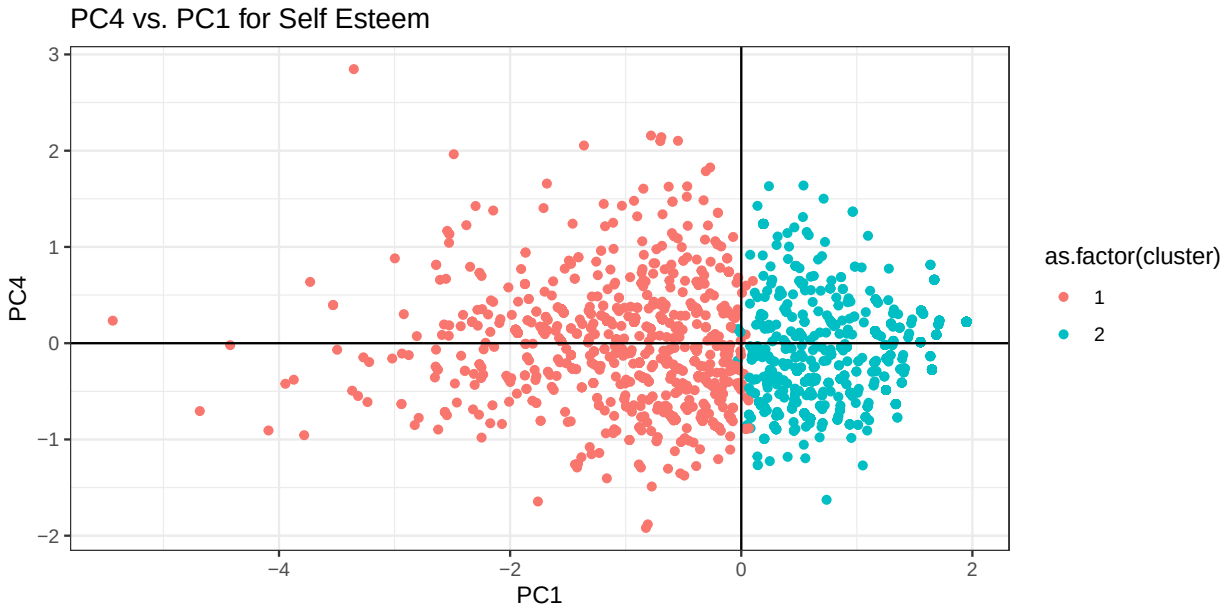


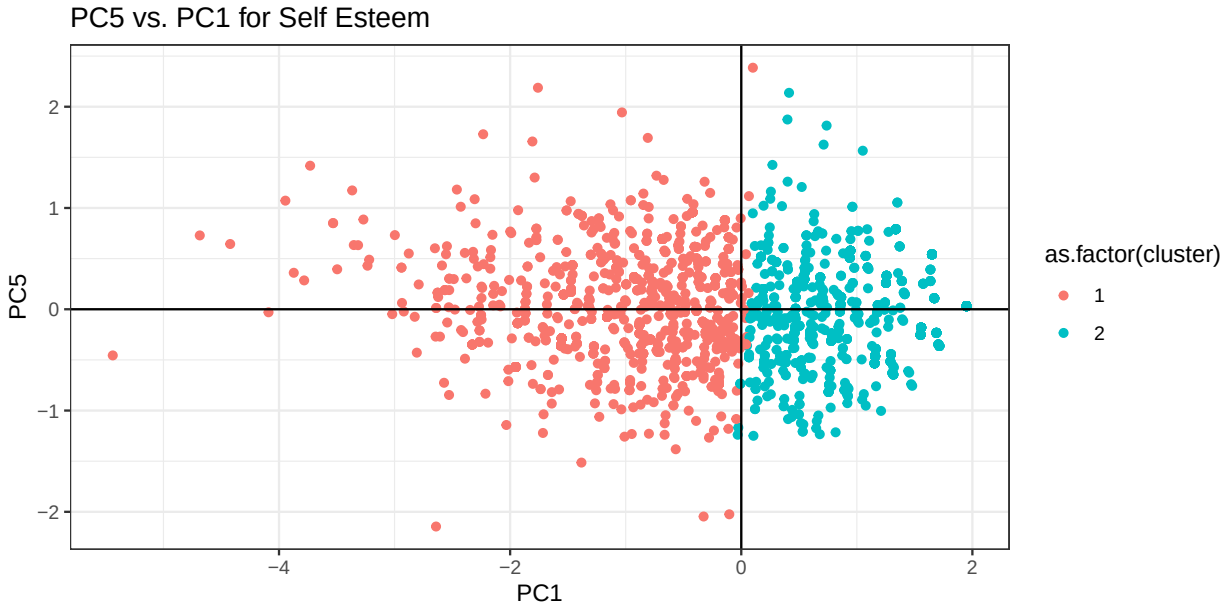
this shows that those with the highest self esteem are mostly in cluster 1 while others are in cluster 2. These are people who agree with a positive statement and disagree with a negative stateent.

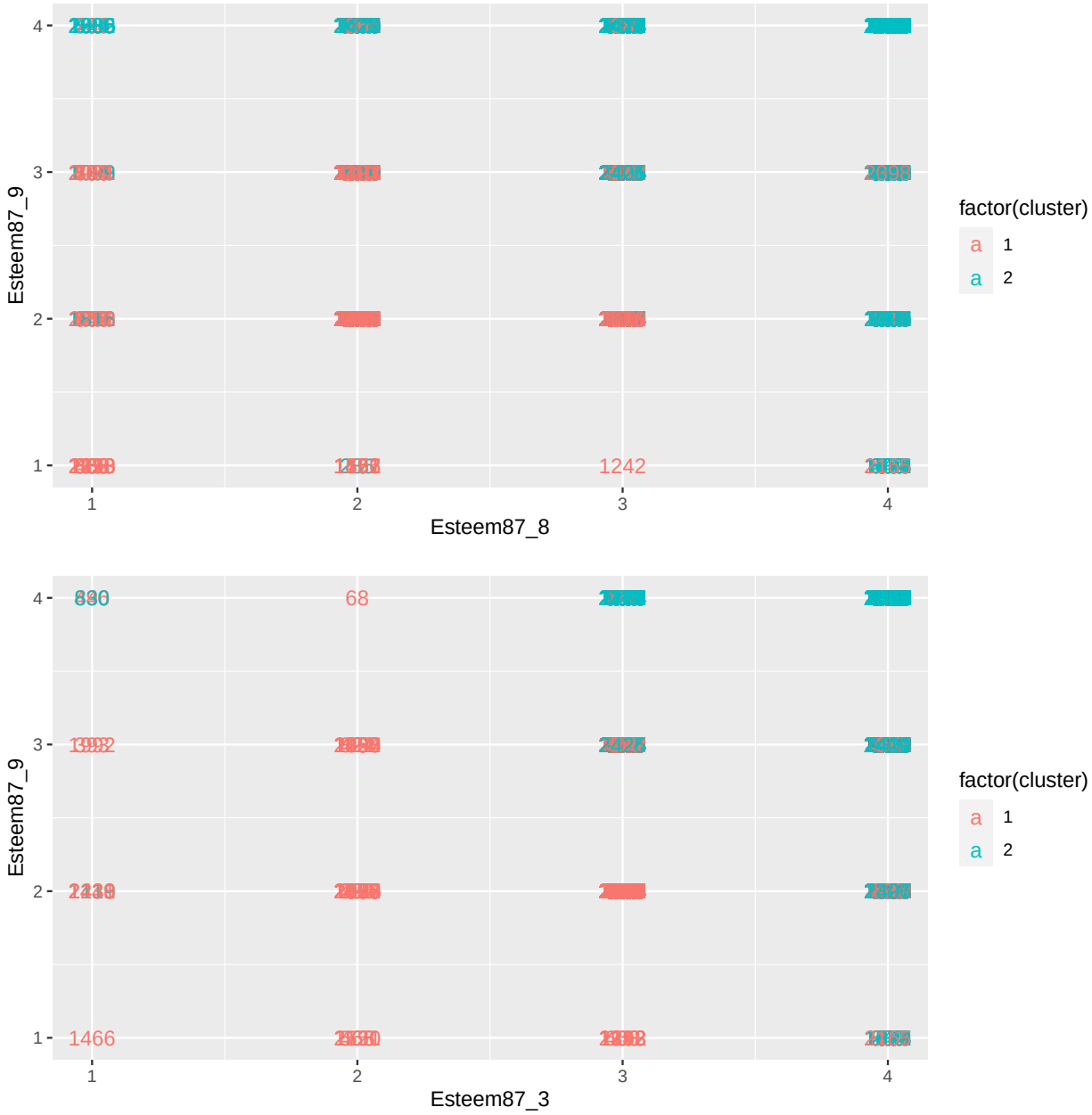
c) Can you visualize the clusters with somewhat clear boundaries? You may try different pairs of variabl


PC2 vs. PC1 for Self Esteem
PC2
PC1
3
2
1
0
−1
−2
−3
−4
−2
0
2
as.factor(cluster)
1
2


PC3 vs. PC1 for Self Esteem
PC3
PC1
2
1
0
−1
−2
−4
−2
0
2
as.factor(cluster)
1
2


PC4 vs. PC1 for Self Esteem
PC4
PC1
3
2
1
0
−1
−2
−4
−2
0
2
as.factor(cluster)
1
2


PC5 vs. PC1 for Self Esteem
PC5
PC1
2
1
0
−1
−2
−4
−2
0
2
as.factor(cluster)
1
2



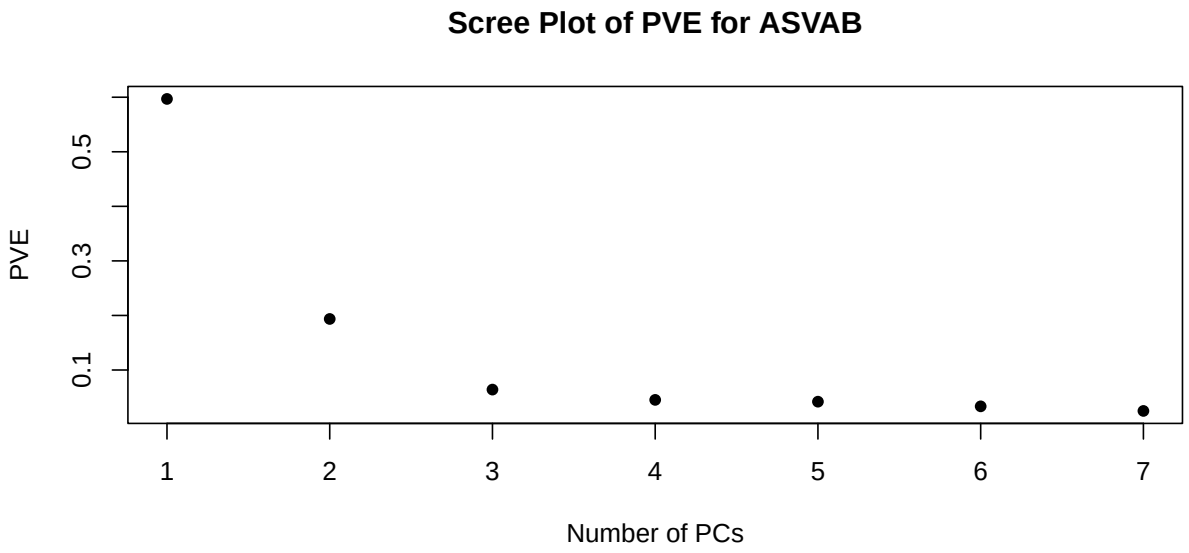
6. We now try to find out what factors are related to self-esteem? PC1 of all the Esteem scores is a good variable to summarize one's esteem scores. We take PC1 as our response variable.

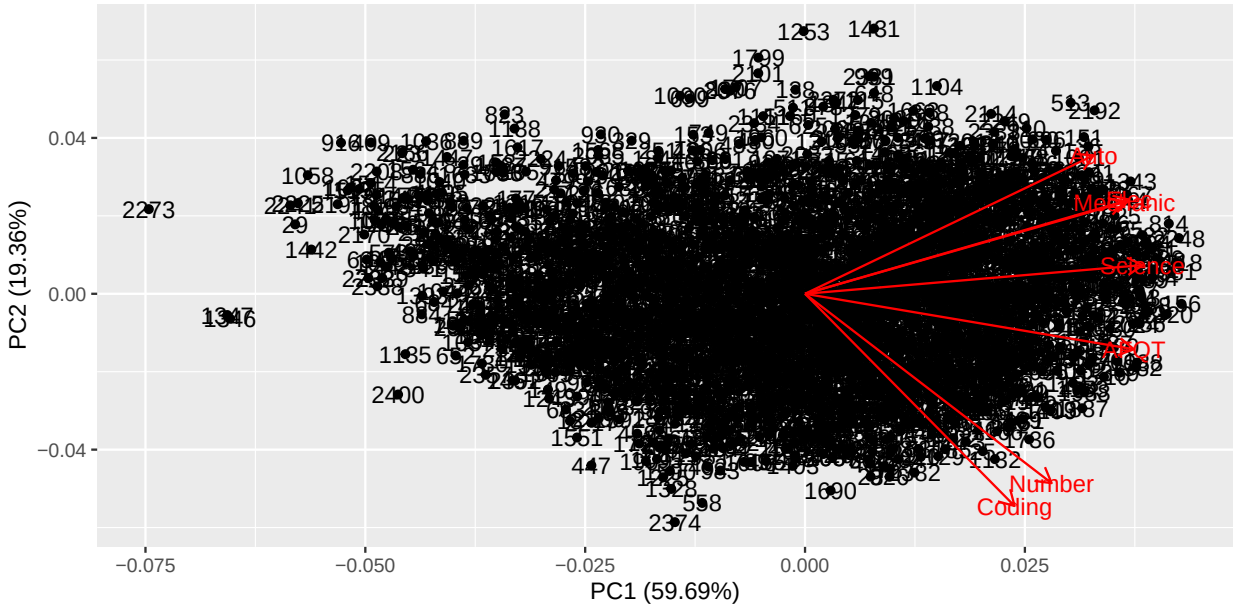
   a) Prepare possible factors/variables:

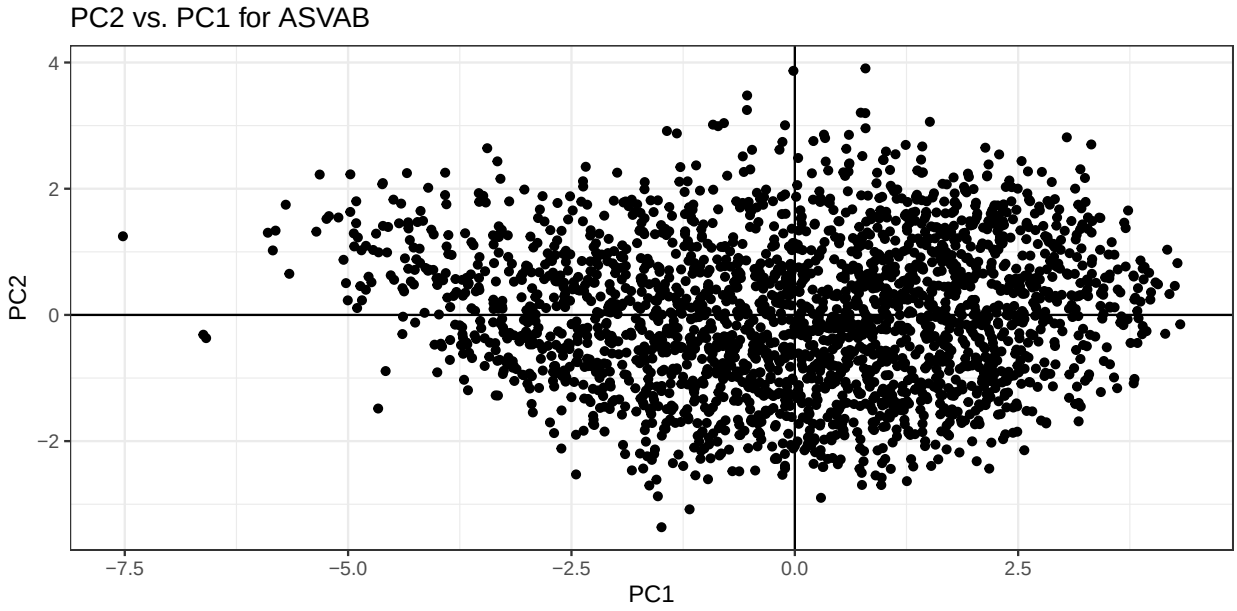
   - EDA the data set first.
   - Personal information: gender, education (05), log(income) in 87, job type in 87. Weight05 (lb) and HeightFeet05 together with Heightinch05. One way to summarize one's weight and height is via Body Mass Index which is defined as the body mass divided by the square of the 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 **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







2.2.1 regression models

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

bmi

intelligence

Gender

Since we saw slight diffferences in esteem scores across gender and newspaper indicators, we can run a model against these features

These results may warrant further investigation into a potential linear relationship between Males, newspaper usage at home, and esteem scores.

Personal information: gender, education (05), log(income) in 87, job type in 87.

combing Personal info

Here we see the effect of Male gender become insignificant when we control for log(income) in 87.

These models are only explaining less than 1% of the variation in the esteem data, however. Let's look at household environment features

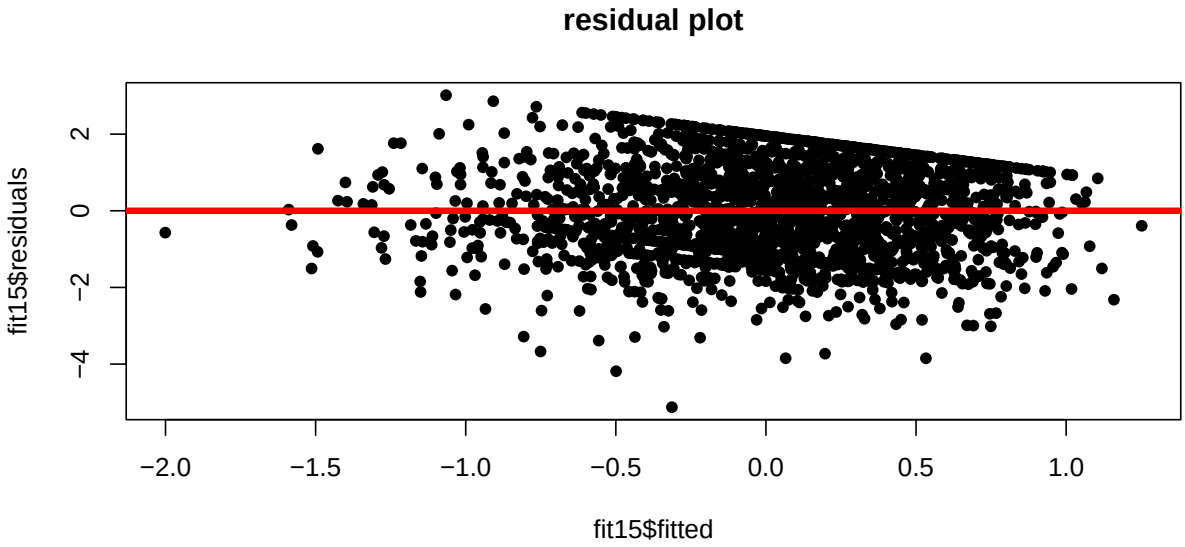
```
- Household environment: Imagazine, Inewspaper, Ilibrary, MotherEd, FatherEd, FamilyIncome78. Do set in
```

Combining more personal info with home environment

fit15 seems to be the best model so far

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

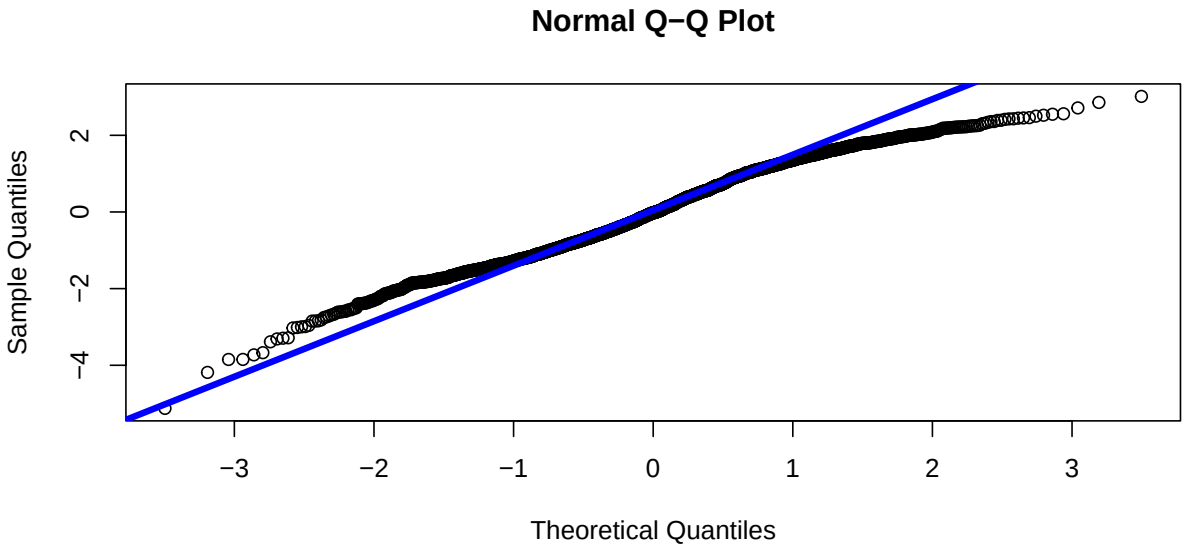
We are choosing `fit15` as our best model given the relatively higher R^2 value and features spanning home envrionment and personal demographics which have a significant linear relationship with PC1.



Residual Plot

There appear to be heteroscedasticity within the model as the variance is not equally distributed across all values of x. In other words, the variance does not appear to be constant, which does not support a linear model assumption.

Check for Normality



The points in the qqplot deviate significantly from the reference line, indicating the data may not be entirely normally distributed.

Taken together, we do not believe there is enough evidence that the assumptios of the linear model are met.

2.2.2 Summary

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

Given the opposing outcomes of the linear model diagnosis which indicate the linear model assumptions may not be completely met, interpretation of this analysis should proceed with caution, as a linear model may not be the best model for predicting PC1 scores of self esteem. With that in mind, featyures such as male gender, income, education, intelligence, family income, and certain leadership-related job positions seem to have a positive correlation with higher self esteem scores. Features such as BMI and some STEM-related fields appear to have a negative correlation with high self esteem scores. Most of the persional background features such as education, gender, and income appear to have the stronge linear realtionships with self esteem scores. Together, the p-values and β coefficients give a complementary picture of the magnitude and directions of linear relationships between these features and show that personal perceptions may be more shaped by personal demographics than family environment.

3 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.

- 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()`.

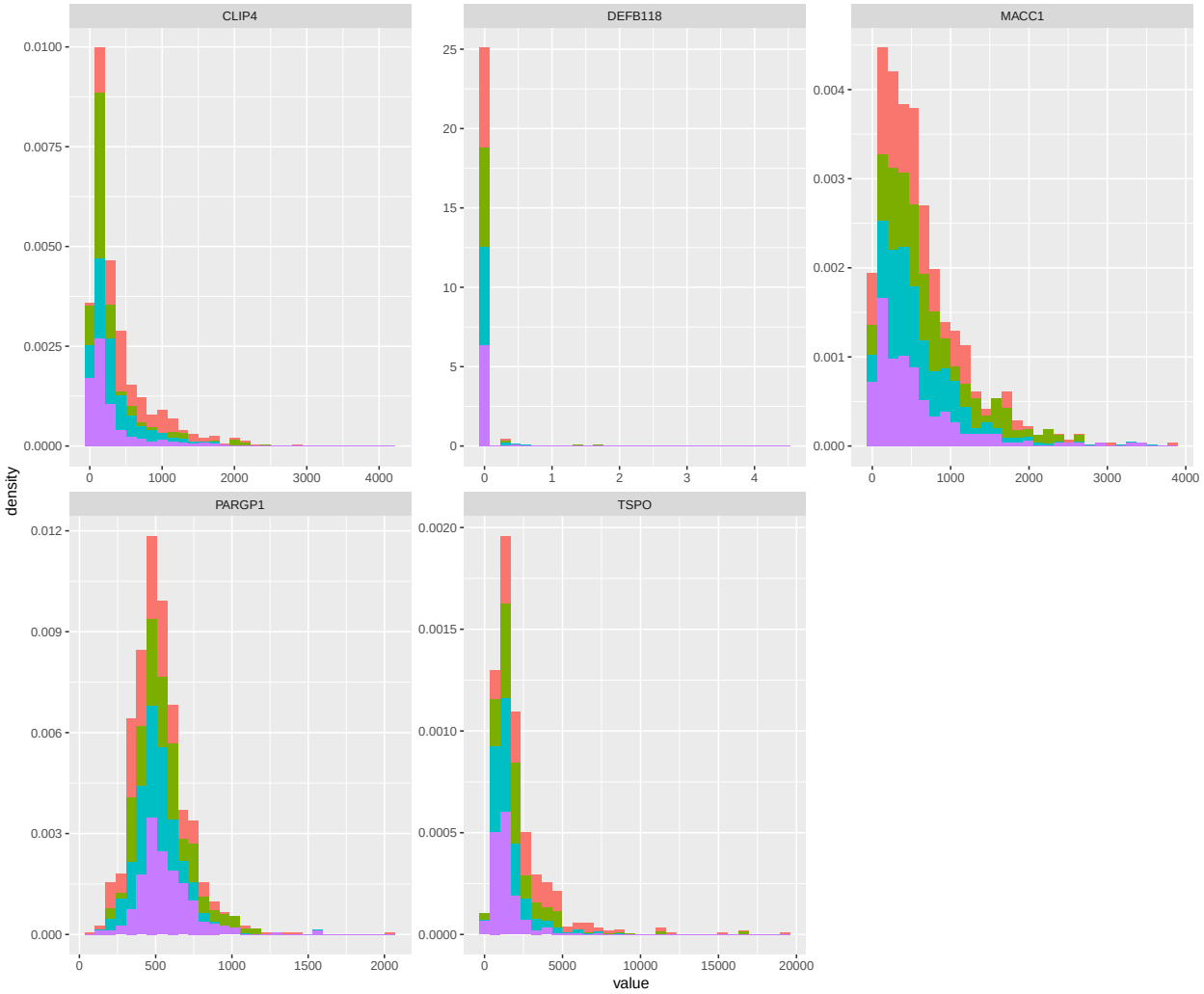
1. Summary and transformation

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

Table 3: Patients in each sub-type

Basal	Her2	LumA	LumB
208	91	628	233

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



c) Remove gene with zero count and no variability. Then apply logarithmic transform.

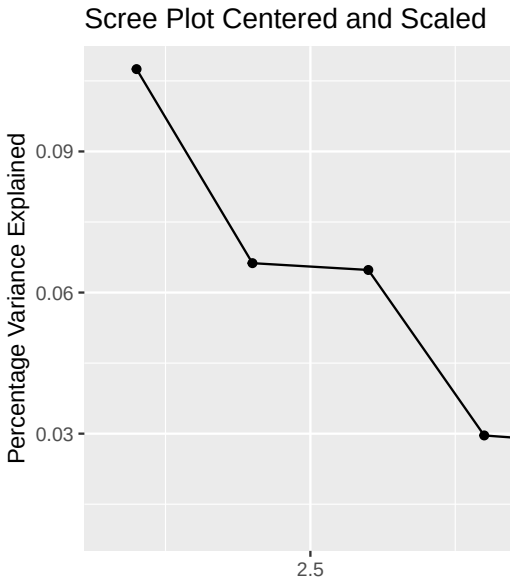
278 genes were removed

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

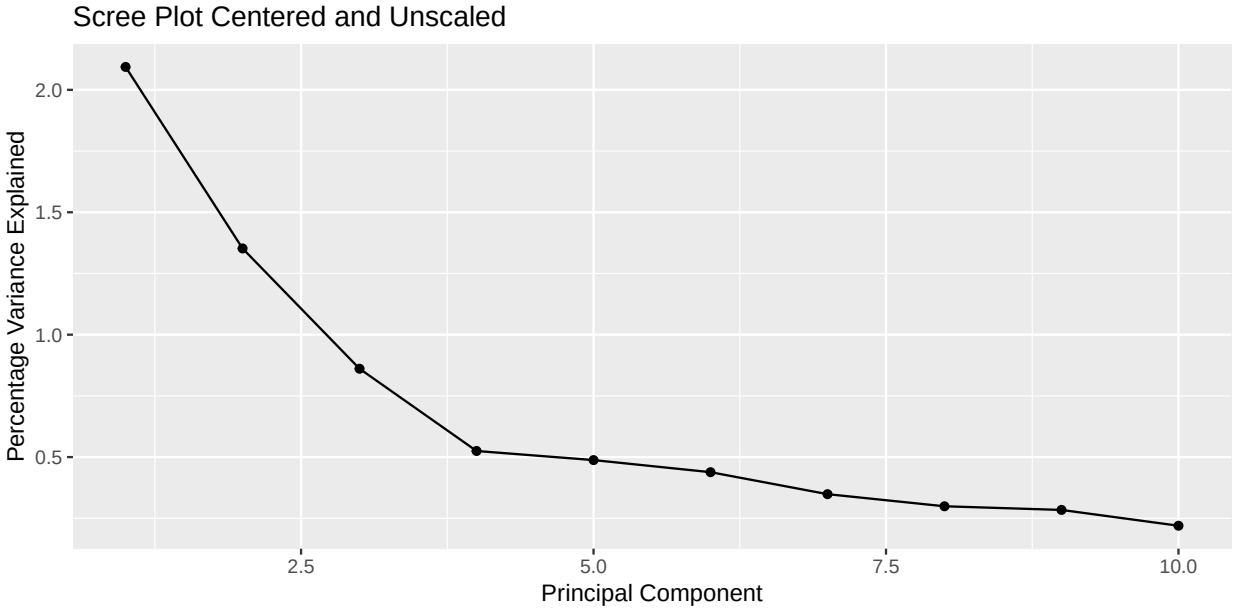
	1	2	3	4
Basal	1	17	3	187
Her2	39	9	26	17
LumA	392	82	154	0
LumB	98	22	111	2

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

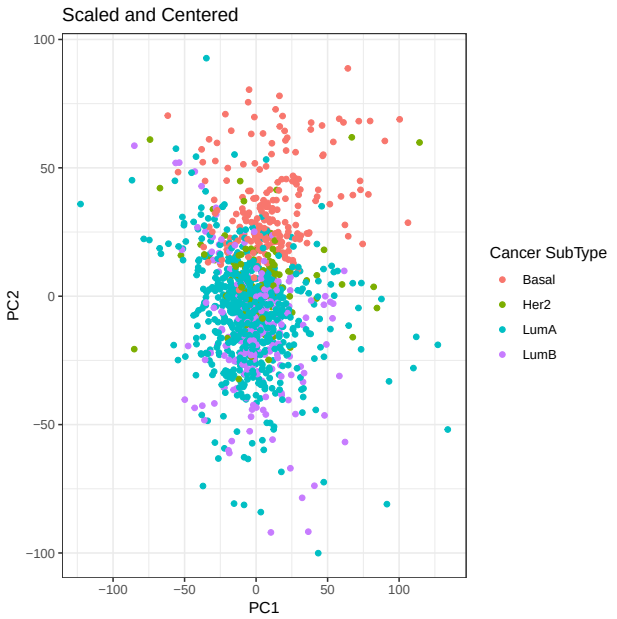
a) Apply PCA on the centered and scaled dataset. How many PCs should we use and why? You are encouraged to use `irlba::irlba()`.



We will use 2 PCs for the centered and scaled and 4 for the centered but not scaled



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)

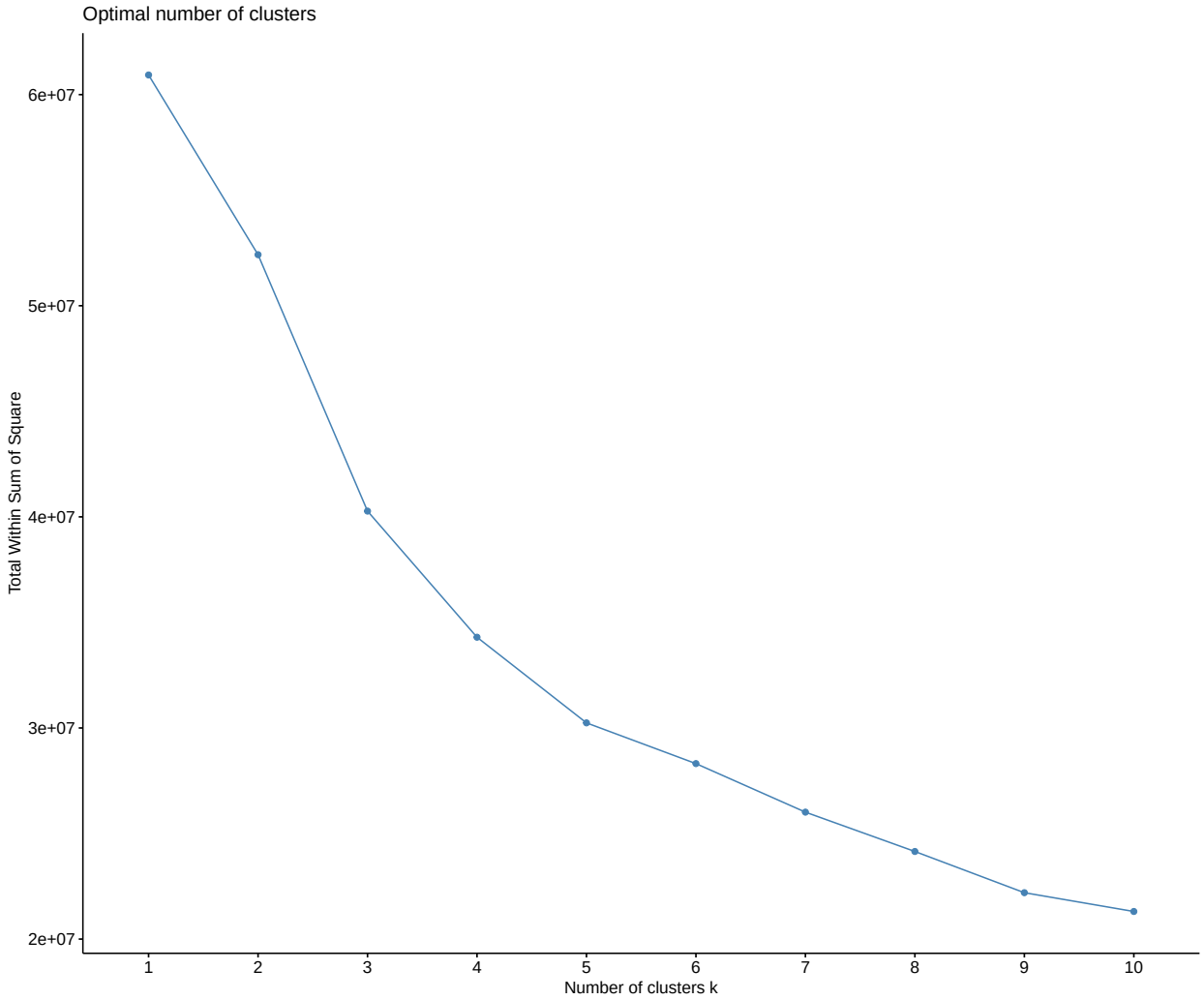


We should not scale the because scaling reduces the separation of clusters

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

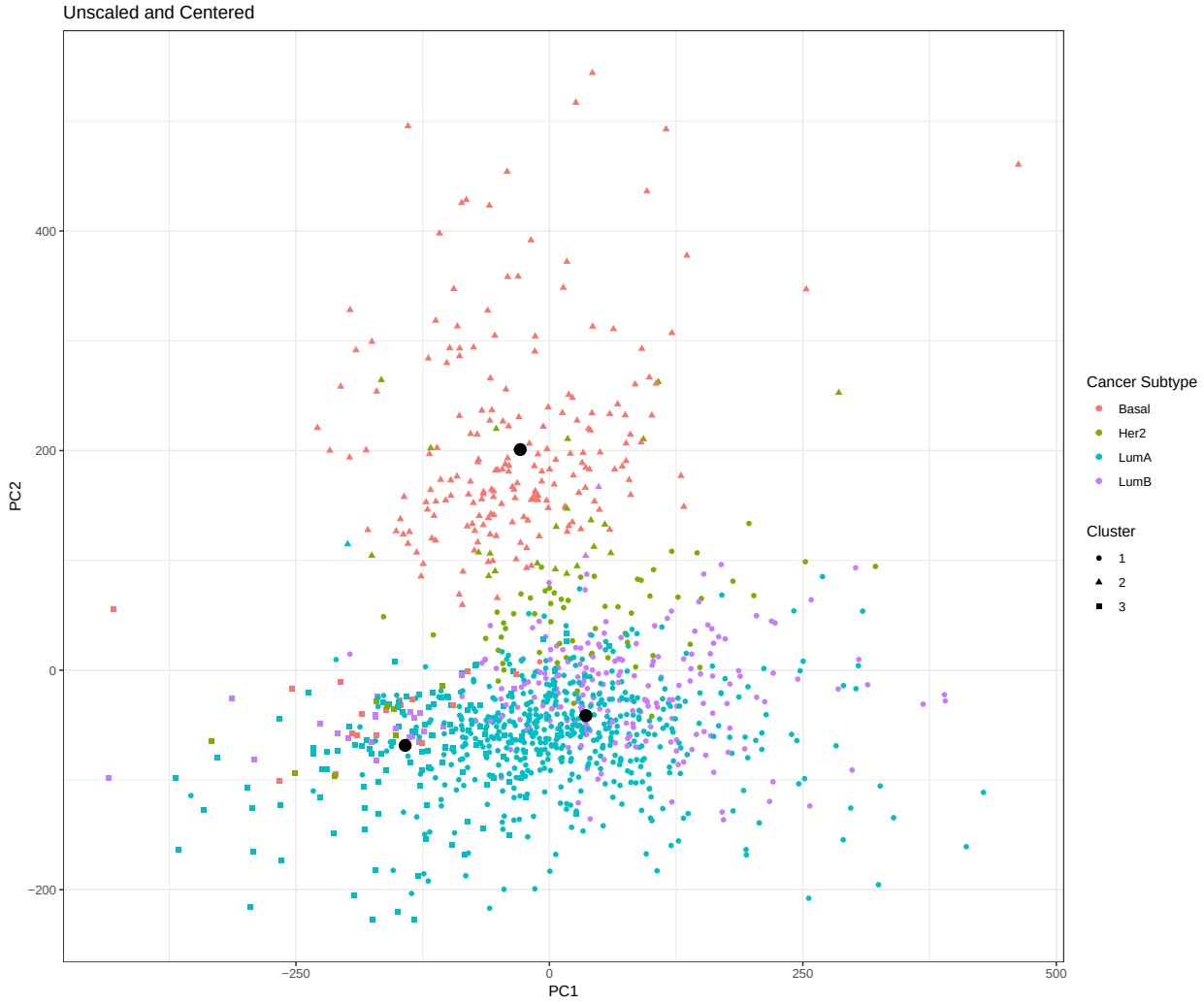
a) Use the first 4 PCs of the centered and unscaled data and apply kmeans. Find a reasonable number of clusters using within sum of squared with the elbow rule.

We choose 3 clusters after looking at the elbow plot



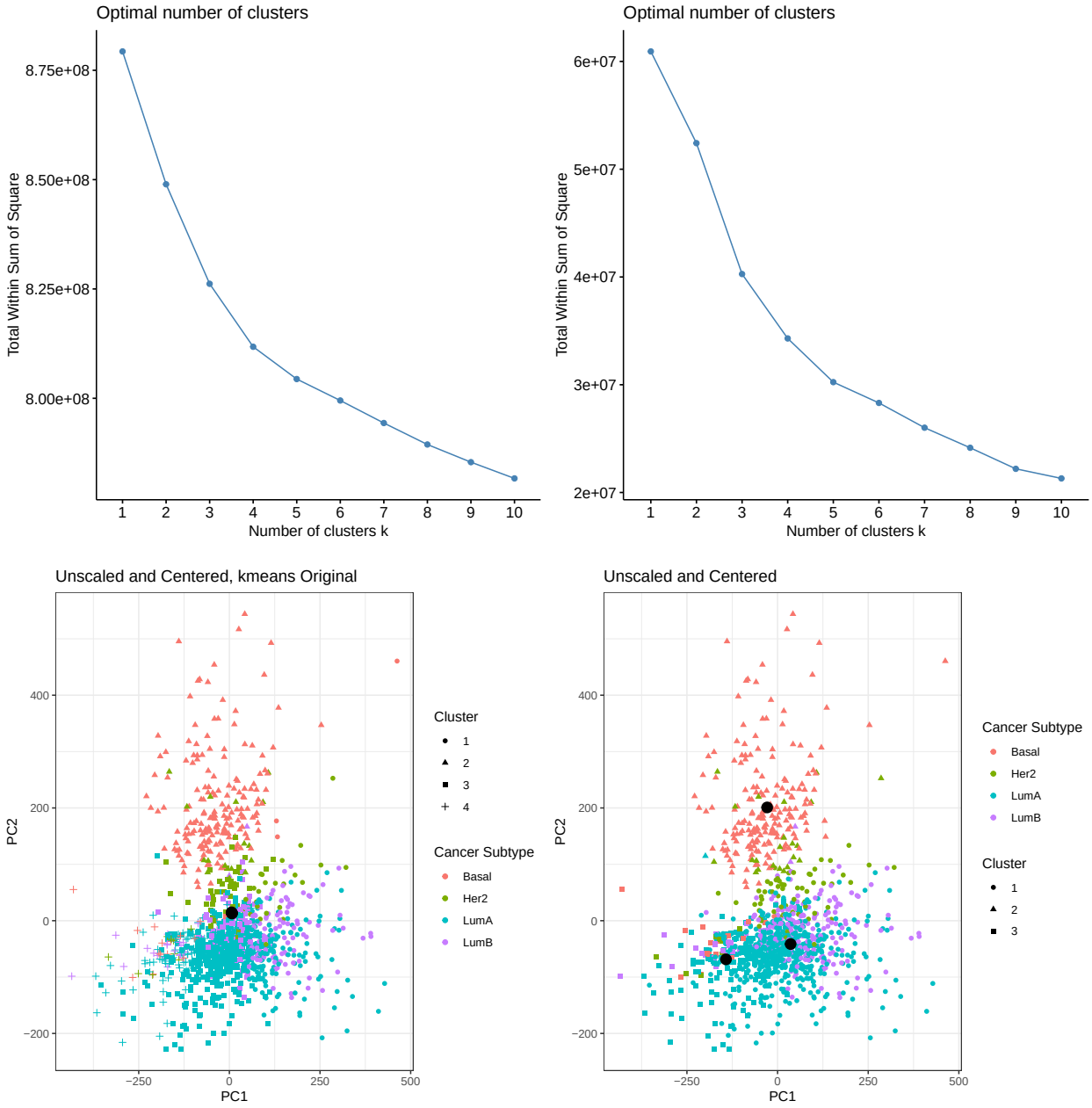
b) Choose an optimal cluster number and apply kmeans. Compare the real sub-type and the clustering label as follows: Plot scatter plot of PC1 vs PC2. Use point color to indicate the true cancer type and point shape to indicate the clustering label. Plot the kmeans centroids with black dots. Summarize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

Clustering is able to separate basal fairly well; however, it doesn't do as well for Luminal A, Lumnial B and HER2



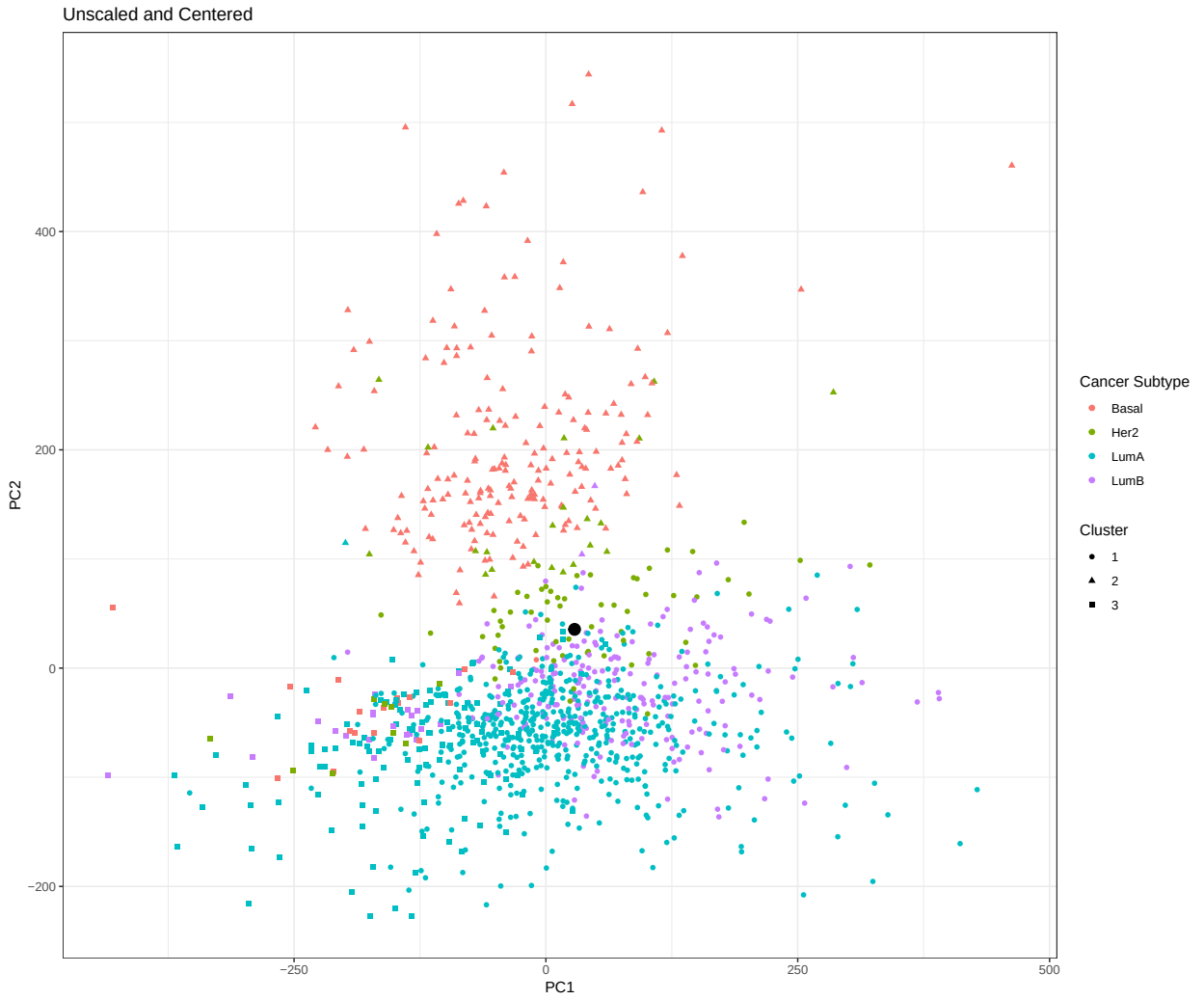
c) Compare the clustering result from applying kmeans to the original data and the clustering result from applying kmeans to 4 PCs. Does PCA help in kmeans clustering? What might be the reasons if PCA helps?

PCA yields better clustering than applying kmeans to the original data. PCA helps because it is able to capture the variance of the data.



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: remember we remove some gene with no counts or no variablity, take log and centered) Plot this patient in the plot in iv) with a black dot. Calculate the Euclidean distance between this patient and each of centroid of the cluster. Can you tell which sub-type this patient might have?

It is difficult to tell which subtype it is since Luminal A, Lumnial B and HER2 are tightly clustered together, but we can definitely rule ot basal. It is Probably LiminalB or HER2



4 Case study 3: Auto data set

This question utilizes the `Auto` dataset from ISLR. The original dataset contains 408 observations about cars. It is similar to the CARS dataset that we use in our lectures. To get the data, first install the package ISLR. The `Auto` dataset should be loaded automatically. We'll use this dataset to practice the methods learn so far. Original data source is here: https://archive.ics.uci.edu/ml/datasets/auto+mpg

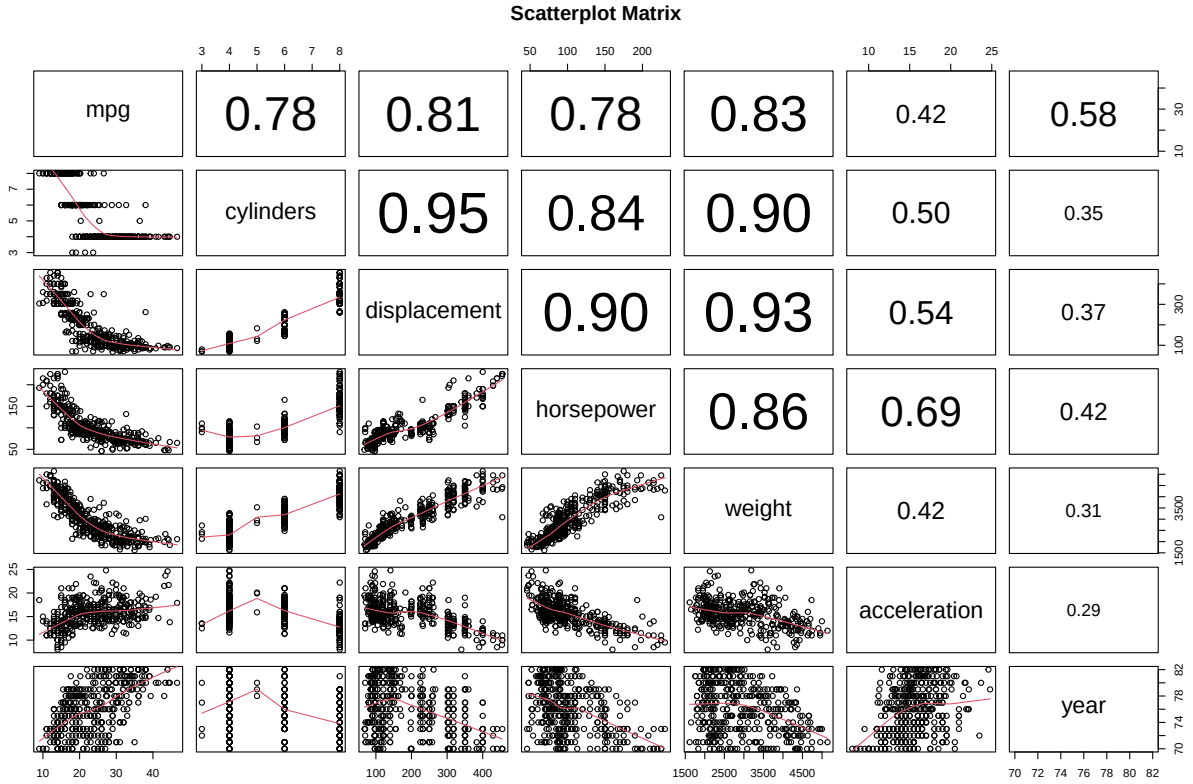
Get familiar with this dataset first. Tip: you can use the command `?ISLR::Auto` to view a description of the dataset.

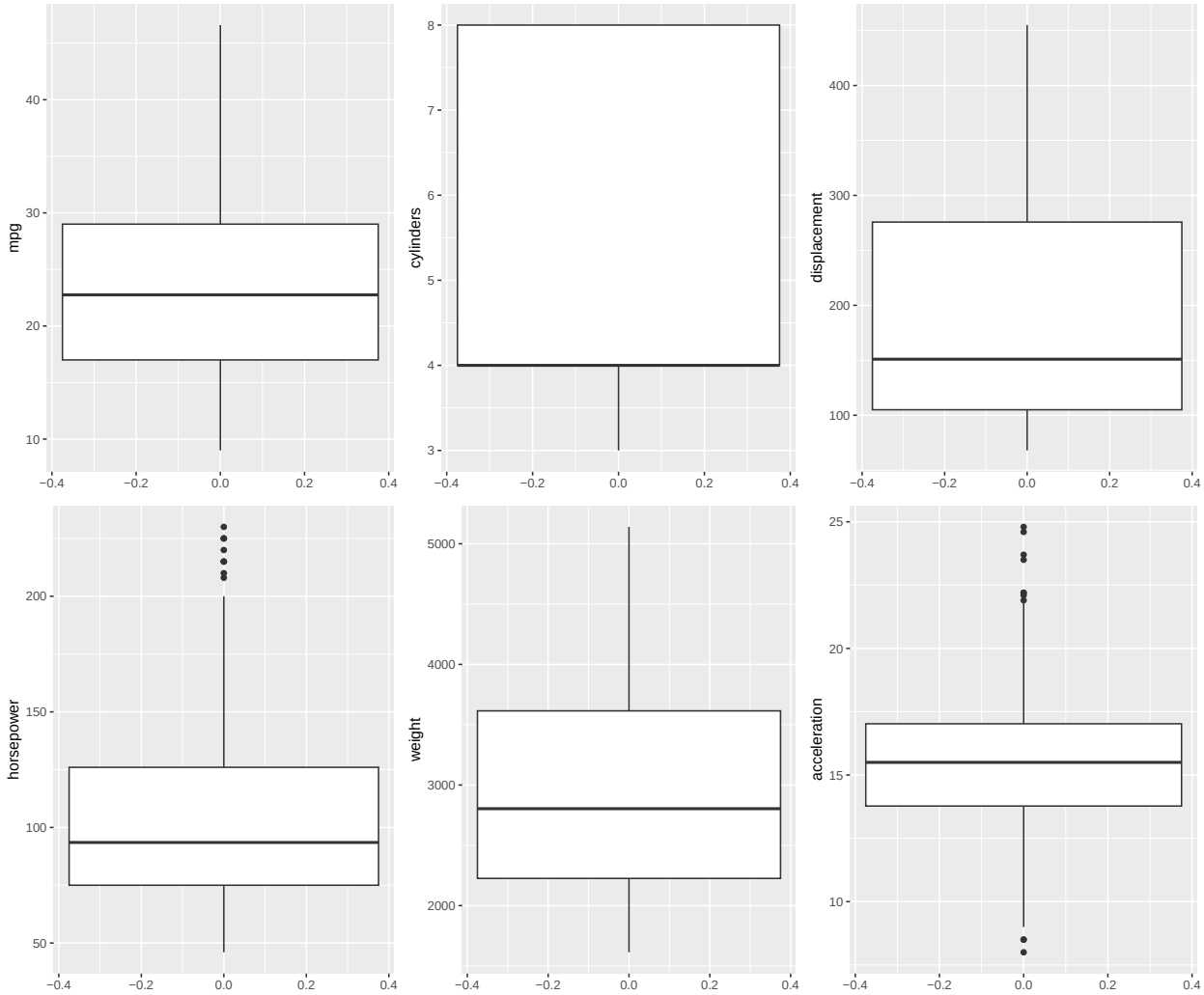
4.1 EDA

Explore the data, with particular focus on pairwise plots and summary statistics. Briefly summarize your findings and any peculiarities in the data.

- mpg summary Min. 1st Qu. Median Mean 3rd Qu. Max. 9.0 17.0 22.8 23.4 29.0 46.6
- cylinders summary Min. 1st Qu. Median Mean 3rd Qu. Max. 3.00 4.00 4.00 5.47 8.00 8.00

- displacement summary Min. 1st Qu. Median Mean 3rd Qu. Max. 68 105 151 194 276 455
- horsepower summary Min. 1st Qu. Median Mean 3rd Qu. Max. 46.0 75.0 93.5 104.5 126.0 230.0
- weight summary Min. 1st Qu. Median Mean 3rd Qu. Max. 1613 2225 2804 2978 3615 5140
- acceleration summary Min. 1st Qu. Median Mean 3rd Qu. Max. 8.0 13.8 15.5 15.5 17.0 24.8
- year summary
  total 13 years
  from 1970-1982
- Origin of car
  American: 245
  European: 68
  Japanese: 79
- Auto names
  unique auto names: 301





4.2 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.

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Feb 12, 2023 - 14:59:59 % Requires LaTeX packages: dcolumn

Year is significant at the 0.01 level. Our model is saying that for every year that goes by, there is about a 1.230 increase in the mpg of a car.

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. (Table 4)_

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Feb 12, 2023 - 14:59:59 % Requires LaTeX packages: dcolumn

Year is significant at the 0.01 level. Our model is saying that for every year that passes by, there is about a .657 increase in the mpg of a car. This effect size decreases from the previous one since we added horsepower to the dataset. (Table 5)

Table 5:

	Dependent variable:
	mpg
year	1.230^{***}
	(1.060,1.400)
Constant	-70.000^{***}
	(-83.000,-57.000)
Observations	392
R^2	0.337
Adjusted R^2	0.335
Residual Std. Error	6.360 (df = 390)
F Statistic	198.000^{***} (df = 1; 390)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6:

	Dependent variable:
	mpg
year	0.657^{***}
	(0.527,0.787)
horsepower	-0.132^{***}
	(-0.144,-0.119)
Constant	-12.700^{**}
	(-23.200,-2.250)
Observations	392
R^2	0.685
Adjusted R^2	0.684
Residual Std. Error	4.390 (df = 389)
F Statistic	424.000^{***} (df = 2; 389)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

The confidence intervals got a lot smaller going from (i) to (ii). Since we added more information to the model (`horspower`) this reduces some of the variability that we see when we examine year alone. This reduction in conifidence interval means that we are likely getting more precise.

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).

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Feb 12, 2023 - 14:59:59 % Requires LaTeX packages: dcolumn

Table 7:

	Dependent variable:
	mpg
year	2.190***
	(1.880,2.510)
horsepower	1.050***
	(0.820,1.270)
year:horsepower	−0.016***
	(-0.019,-0.013)
Constant	−127.000***
	(-150.000,-103.000)
Observations	392
R^2	0.752
Adjusted R^2	0.750
Residual Std. Error	3.900 (df = 388)
F Statistic	393.000*** (df = 3; 388)
Note:	*p<0.1; **p<0.05; ***p<0.01

All of the variables are significant at the 0.01 level. Year is an extremely significant variable. Our model is saying that for every year that passes by, there is about a 2.190 increase in the mpg of a car. This effect size increases dramatically from the previous models. (Table 6)

4.3 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?

Table 8:

	Dependent variable:
	mpg
cylinders	-3.560^{***}
	(-3.840,-3.270)
Constant	42.900^{***}
	(41.300,44.600)
Observations	392
R^2	0.605
Adjusted R^2	0.604
Residual Std. Error	4.910 (df = 390)
F Statistic	597.000^{***} (df = 1; 390)
Note:	*p<0.1; **p<0.05; ***p<0.01

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Feb 12, 2023 - 14:59:59 % Requires LaTeX packages: dcolumn

Cylinders is significant at the 0.01 level. Our model is saying that for every 1 cylinder added, there is about a 3.560 increase in the mpg of a car. (Table 7)

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`.

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Feb 12, 2023 - 14:59:59 % Requires LaTeX packages: dcolumn

Only 4 Cylinders is significant at the 0.01 level. Our model is saying that for every 1 cylinder added, there is about a 3.560 increase in the mpg of a car. (Table 7)

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

From a practical sense it's not feasible to consider cylinders as a continuous variable because it'll out put results that don't make sense. It will assume that the more cylinders you have, the lower your mpg will be. However, considering cylinders as a categorical variable allows you to see that having different numbers of cylinders is not a linear relationship.

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?

Yes you can using anova(H_0, H_1). There is strong evidence of rejecting the null hypothesis that fit0: `mpg` is linear in `cylinders` vs. fit1: `mpg` relates to `cylinders` as a categorical variable

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cylinders
## Model 2: mpg ~ factor(cylinders)
```

Table 9:

	Dependent variable:
	mpg
factor(cylinders)4	8.730^{***}
	(4.080,13.400)
factor(cylinders)5	6.820^{*}
	(-0.217,13.900)
factor(cylinders)6	-0.577
	(-5.290,4.140)
factor(cylinders)8	-5.590^{**}
	(-10.300,-0.894)
Constant	20.600^{***}
	(15.900,25.200)
Observations	392
R^2	0.641
Adjusted R^2	0.638
Residual Std. Error	4.700 (df = 387)
F Statistic	173.000^{***} (df = 4; 387)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
##   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
## 1    390 9416
## 2    387 8544  3       871 13.2 3.4e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.4 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.

In this final model, we consider both cylinders and origin as categorical variables. We remove acceleration and we assume interaction between year-horsepower and origin-displacemnt

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Feb 12, 2023 - 15:00:00

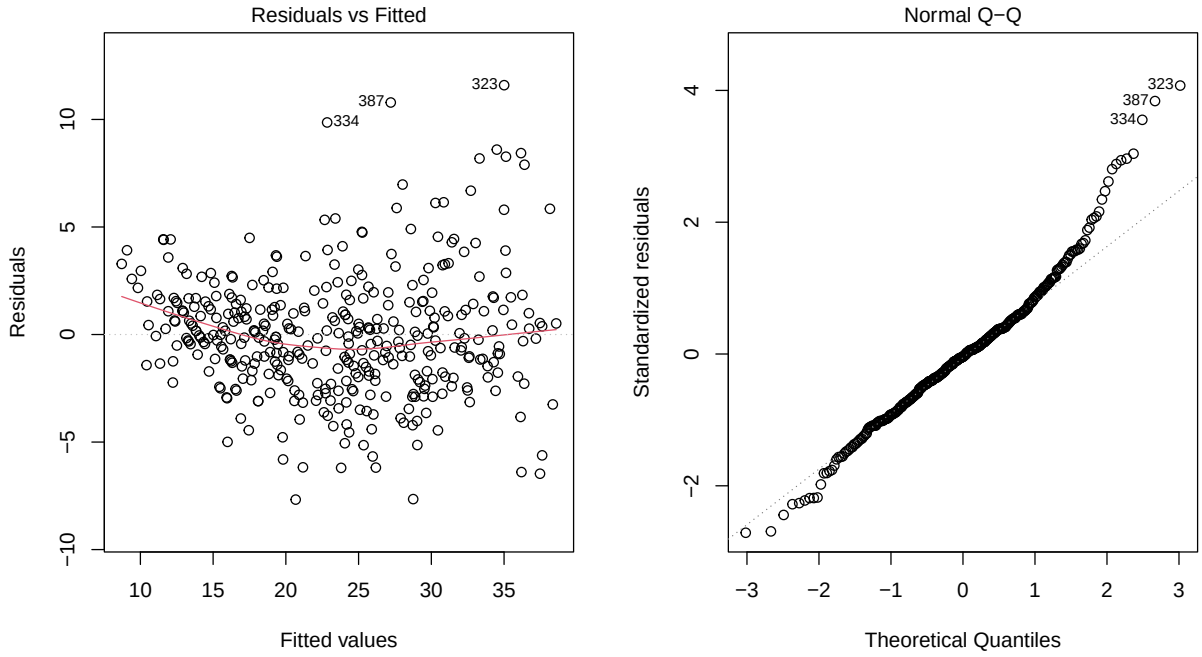


Table 10:

	Dependent variable:				
	mpg				
	(1)	(2)	(3)	(4)	
cylinders	−0.490		0.086		
	(−1.120, 0.140)		(−0.502, 0.674)		
factor(cylinders)4		6.720***		6.250***	
		(3.480, 9.960)		(3.240, 9.260)	(3.
factor(cylinders)5		7.080***		6.250***	
		(2.150, 12.000)		(1.670, 10.800)	(1.
factor(cylinders)6		3.350*		4.370**	
		(−0.225, 6.930)		(1.040, 7.690)	(1.
factor(cylinders)8		5.100**		6.520***	
		(0.966, 9.230)		(2.670, 10.400)	(2.
displacement	0.024***	0.019***	0.009	0.007	
	(0.009, 0.039)	(0.005, 0.033)	(−0.005, 0.023)	(−0.007, 0.020)	(−0
horsepower	−0.018	−0.035***	0.807***	0.686***	
	(−0.045, 0.009)	(−0.061, −0.009)	(0.622, 0.992)	(0.505, 0.866)	(0.
weight	−0.007***	−0.006***	−0.006***	−0.005***	
	(−0.008, −0.005)	(−0.007, −0.005)	(−0.007, −0.004)	(−0.006, −0.004)	(−0.
acceleration	0.079	0.026	−0.031	−0.062	
	(−0.113, 0.272)	(−0.156, 0.208)	(−0.208, 0.147)	(−0.233, 0.109)	(−0
year	0.777***	0.737***	1.860***	1.690***	
	(0.676, 0.879)	(0.641, 0.833)	(1.600, 2.120)	(1.430, 1.940)	(1.
originEuropean	2.630***	1.760***	2.400***	1.770***	
	(1.520, 3.740)	(0.683, 2.850)	(1.390, 3.410)	(0.766, 2.770)	(0.
originJapanese	2.850***	2.620***	2.360***	2.260***	
	(1.770, 3.940)	(1.580, 3.650)	(1.360, 3.350)	(1.300, 3.230)	(1.
year:horsepower			−0.011***	−0.010***	
			(−0.014, −0.009)	(−0.012, −0.007)	(−0.
Constant	−18.000***	−22.100***	−98.400***	−90.500***	
	(−27.100, −8.790)	(−31.000, −13.200)	(−118.000, −78.700)	(−109.000, −71.600)	(−109
Observations	392	392	392	392	
R^2	0.824	0.847	0.854	0.869	
Residual Std. Error	3.310 (df = 383)	3.100 (df = 380)	3.020 (df = 382)	2.870 (df = 379)	2.87

Note:

b) Summarize the effects found.

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Feb 12, 2023 - 15:00:00

Table 11:

	Dependent variable:
	mpg
factor(cylinders)4	6.110^{***}
	(3.130, 9.090)
factor(cylinders)5	6.100^{***}
	(1.540, 10.700)
factor(cylinders)6	4.230^{**}
	(0.927, 7.540)
factor(cylinders)8	6.410^{***}
	(2.570, 10.200)
year	1.680^{***}
	(1.430, 1.930)
horsepower	0.682^{***}
	(0.502, 0.862)
weight	-0.005^{***}
	$(-0.006, -0.004)$
originEuropean	1.770^{***}
	(0.768, 2.770)
originJapanese	2.260^{***}
	(1.300, 3.220)
displacement	0.007
	$(-0.006, 0.021)$
year:horsepower	-0.010^{***}
	$(-0.012, -0.007)$
Constant	-90.800^{***}
	$(-110.000, -71.900)$
Observations	392
R^2	0.868
Residual Std. Error	2.870 (df = 380)
Note:	*p<0.1; **p<0.05; ***p<0.01

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.

From these specifications we predict that the mpg will be 2.91 with a CI of [-4.37, 10.2]

5 [Case Study 4] Simple Regression through simulations

5.1 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.

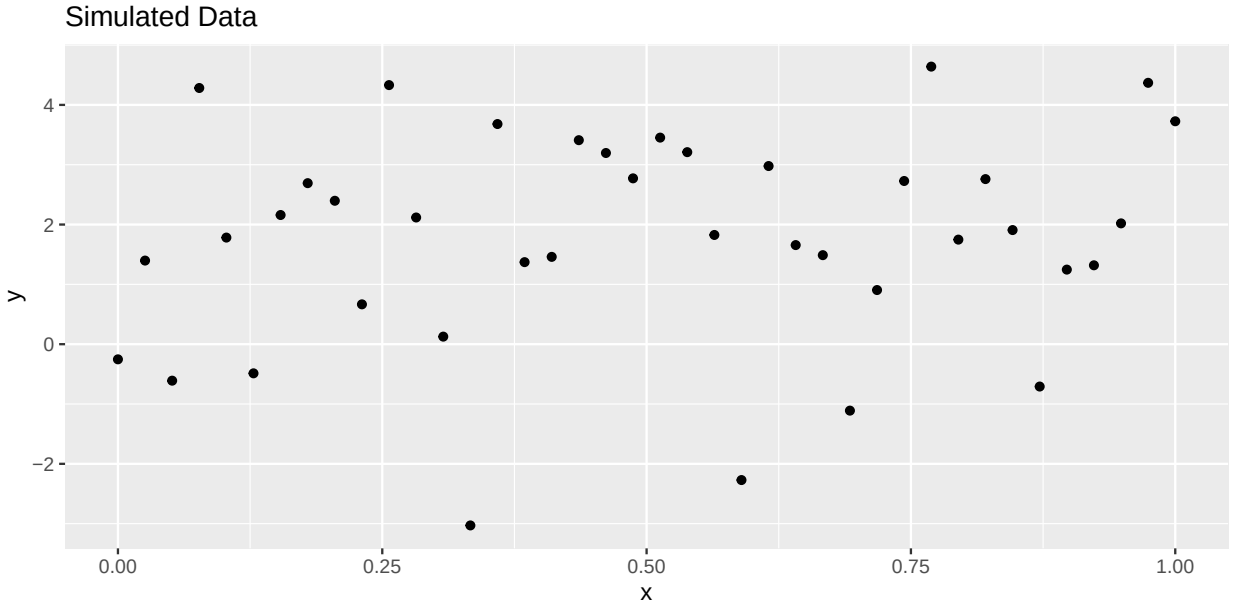
Presume that $\mathbf{x}$ and $\mathbf{y}$ are linearly related with a normal error $\boldsymbol{\varepsilon}$, such that $\mathbf{y} = 1 + 1.2\mathbf{x} + \boldsymbol{\varepsilon}$. The standard deviation of the error ε_i is $\sigma = 2$.

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)
x
```

5.1.1 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.



5.1.2 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?

LS estimate of β_0 is 1.3 and β_1 is 0.906.

The true value of β_0 is 1

The true value of β_1 is 1.2

The estimates appears to be slightly off from the true values, but they are relatively close.

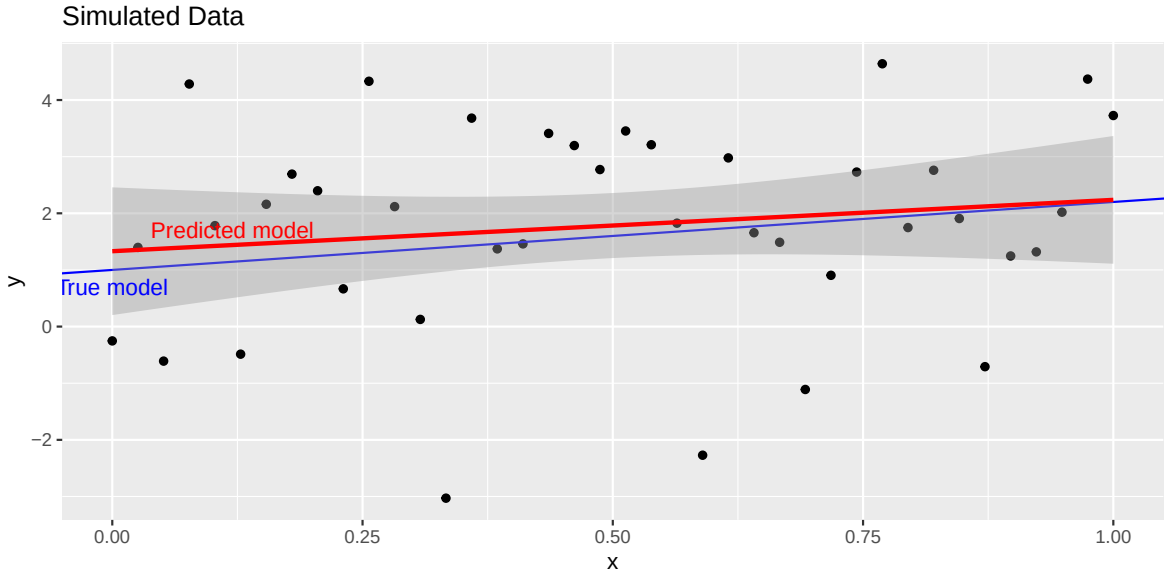
ii. What is your RSE for this linear model fit? Is it close to $\sigma = 2$?

The RSE for this model is 1.79 which is pretty close to 2, slightly smaller.

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

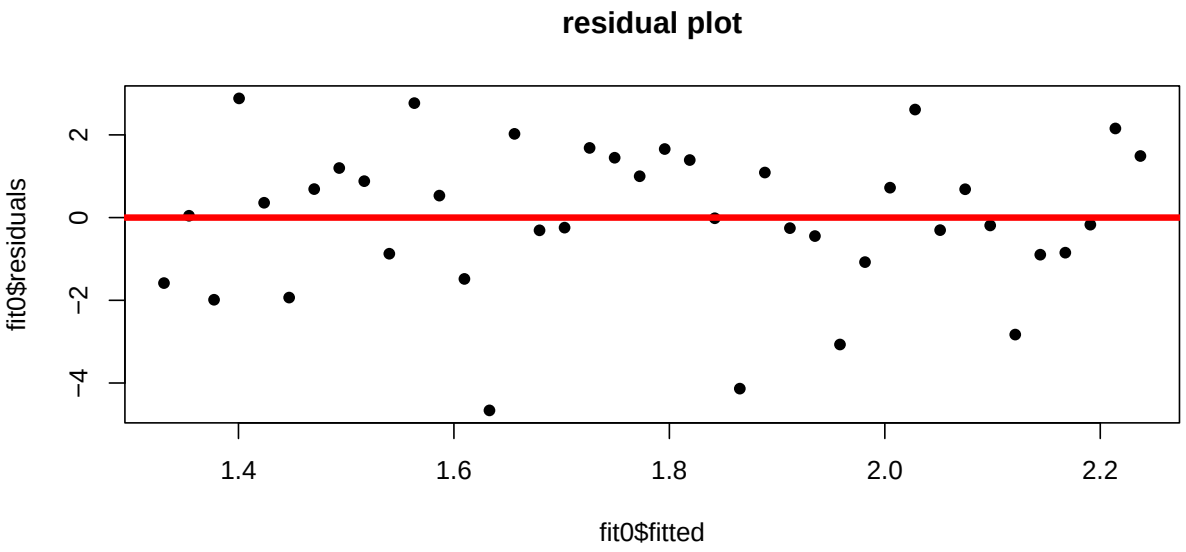
The 95% confidence interval for β_1 is 0.906 +/- 0.959 = [-0.053, 1.865]. This confidence interval does capture the true β_1 of 1.2. If we had more samples (higher N), this confidence interval would become narrower and more precise around the true value.

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

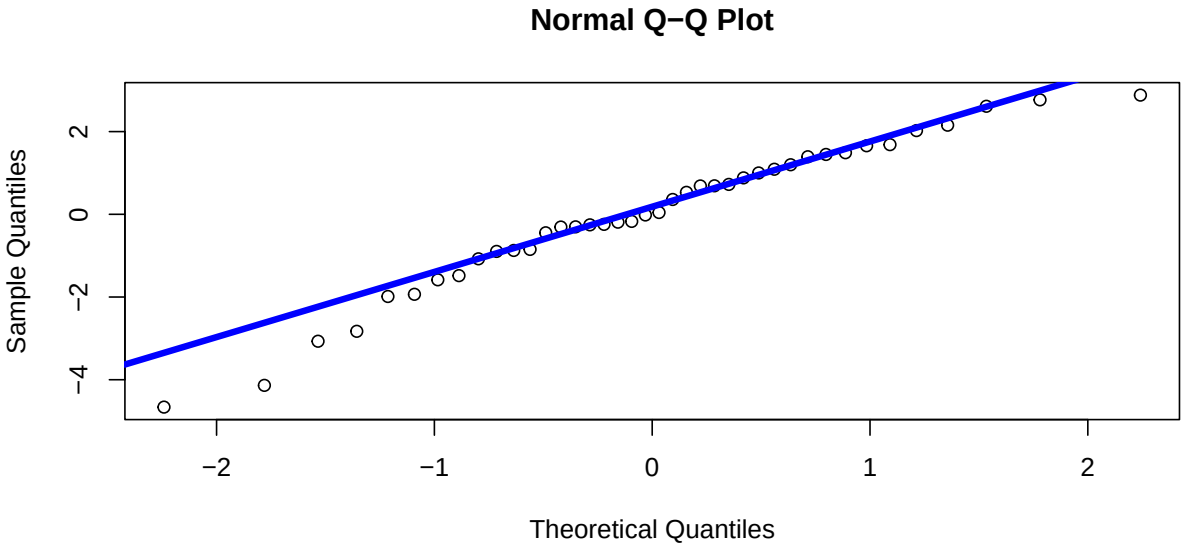


5.1.3 diagnoses

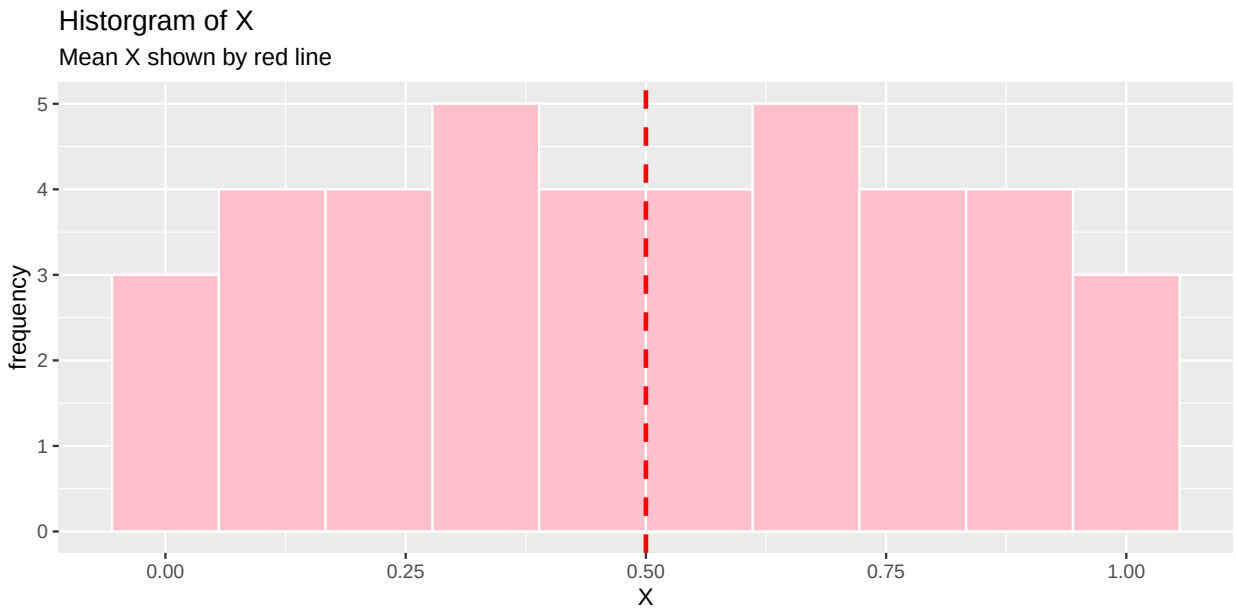
i. Provide residual plot where fitted **y**-values are on the x-axis and residuals are on the y-axis.



ii. Provide a normal QQ plot of the residuals.



iii. Comment on how well the model assumptions are met for the sample you used. *We observe relatively favorable homoscedastic data as well as quantiles which mostly follow a normal reference line. Another way to qualitatively check for normality is with a histogram.*



From this we can see the data is relatively normal and most likely follows linear model assumptions, but, conservatively, interpretation should proceed with caution as the distribution of the data may not be unimodal.

5.2 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.

```
## Warning in rm(se, b1, upper_ci, lower_ci, x, n_sim, t_star, lse, lse_out):
## object 'lse_out' not found
```

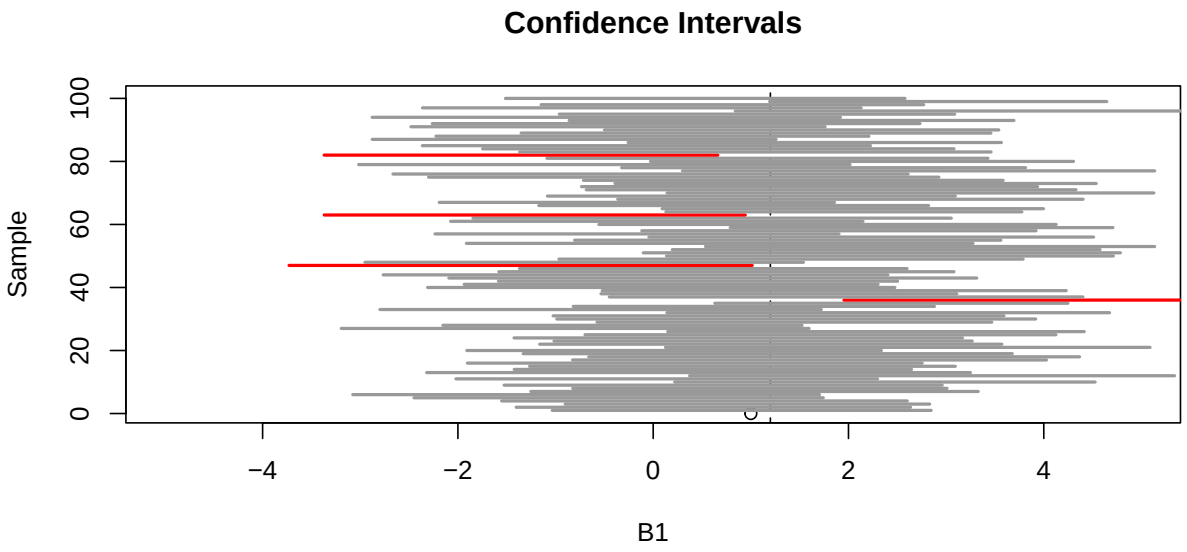
i. Summarize the LS estimates of β_1 (stored in `results$b1`). Does the sampling distribution agree with theory?

The mean of the sampling distribution of LS estimates for β_1 is 1.15, which is very close too the true β_1 of 1.2. Therefore, this sampling distribution shows strong support for the theory.

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

Given that this is the 95% confidnce interval, we expect that 95% of our 100 confidence intervals will capture the true value of β_1

Currently we see this is closer to 94%, which is relatively close. Whis intervals don't cover the treu value?



Given the previously established 94% proportion of tru-value confidence intervals, the 6 red intervals out of hte 100 total intervals are an accurate deficiton of the 6% of intervals which do not contain the true value of β_1