HW5

Graded

Student

Sangwon Ji

Total Points

95 / 95 pts

Question 1

Q1 **50** / 50 pts

1.1 a **15** / 15 pts

✔ **– 0 pts** Correct

– 4 pts Incorrect PC2

1.2 b **10** / 10 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect Outlier

1.3 c **5** / 5 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect Number of PC Use

– 5 pts Missing Answer

– 3 pts Missing Plot

1.4 d **20** / 20 pts

✔ **– 0 pts** Correct

– 20 pts Missing Answer

Question 2

Q2 **0** / 0 pts

2.1 **a** **0** / 0 pts

✔ **– 0 pts** Correct

– 0 pts Incorrect

2.2 **b** **Resolved** **0** / 0 pts

✔ **– 0 pts** Correct

– 0 pts Incorrect

Regrade Request **Submitted on: May 03**

Hi ! I think in the solution it says the answer is number i, but I think it's indicating number ii in the context. Would help a lot if you check this please ! Thank you in advance.

corrected

Reviewed on: May 03

2.3 **c** **0** / 0 pts

✔ **– 0 pts** Correct

– 0 pts Incorrect

Question 3

Q3 **15** / 15 pts

✔ **– 0 pts** Correct

– 3 pts Incorrect XXFXX

– 3 pts Incorrect XXBXX

– 3 pts Incorrect XXDXX

– 3 pts Incorrect XXAXX

– 3 pts Incorrect XXCXX

– 15 pts Missing Answer

Question 4

Q4 **30** / 30 pts

4.1 **a** **10** / 10 pts

✔ **– 0 pts** Correct

– 5 pts Missing Variables

4.2 **b** **20** / 20 pts

✔ **– 0 pts** Correct

– 10 pts Missing Diagnostics

Question 5

Late Points **0** / 0 pts

✔ **– 0 pts** On Time

– 9.5 pts 1 Day late (-10%)

– 19 pts 2 Days late (-20%)

– 28.5 pts 3 Days late (-30%)

Question assigned to the following page: 1.1

HW5

STAT 131A Spring 2023

The first question is going to use multiple variable visualization tools (PCA) to learn about an unknown dataset. This dataset comes from the daily measures of sensors in a urban waste water treatment plant [https://archive.ics.uci.edu/ml/datasets/Water+Treatment+Plant]. The measurements are all continuous, and are described in `VariableDescriptions.txt` file that comes with the homework. However, these are not "intuitive" variables, since they are measurements of various chemical properties of the water, so we don't expect you to understand the measurements. But we will use some of our visualization techniques to get a sense of the data, even though we don't have an intuitive sense of the meaning of the variables.

There are also variables that are related to the date in which the observations were collected (e.g. `Date`, `Month`, `Season`). For simplicity, we have removed days with NA values in any of the sensors, though this is not ideal for data analysis.

First we will provide you with some code to read in the data, and we will set up some factors and colors for the date-related variables.

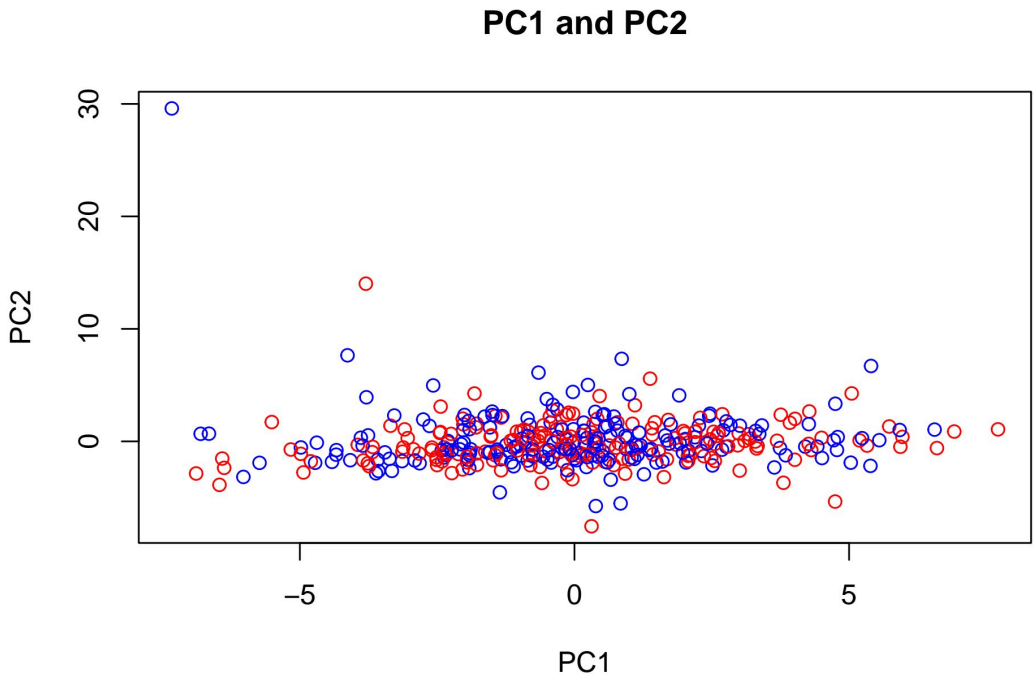
```
water<-read.csv(file = "water-treatment-cleaned.csv", header = TRUE, stringsAsFactors = FALSE)
water$Month<-factor(water$Month,levels=month.name)
water$Day<-factor(water$Day,levels=weekdays(x=as.Date(seq(7), origin="1950-01-01")))
water$Year<-factor(water$Year,levels=c(90,91),labels=c("1990","1991"))
colDays<-palette()
names(colDays)<-levels(water$Day)
library(RColorBrewer)
colMonths<-c("coral4",brewer.pal(11, "Spectral"))
names(colMonths)<-levels(water$Month)
colYear<-c("blue","green")
names(colYear)<-levels(water$Year)
colSeason<-c("Blue","Green","Red","Brown")
names(colSeason)<-c("Winter","Spring","Summer","Fall")
```

Question 1: PCA

a. (15 points) Perform a PCA of this data and plot a scatterplot of the samples based on the first 2 principal coordinates.

```
# add code here for pca and scatterplot
PCAwater <- prcomp((water[-c(1:5)]), center = T, scale = T)
plot(PCAwater$x[,1], PCAwater$x[,2], col = c("red", "blue"), xlab = "PC1", ylab = "PC2", main = "PC1 an
```

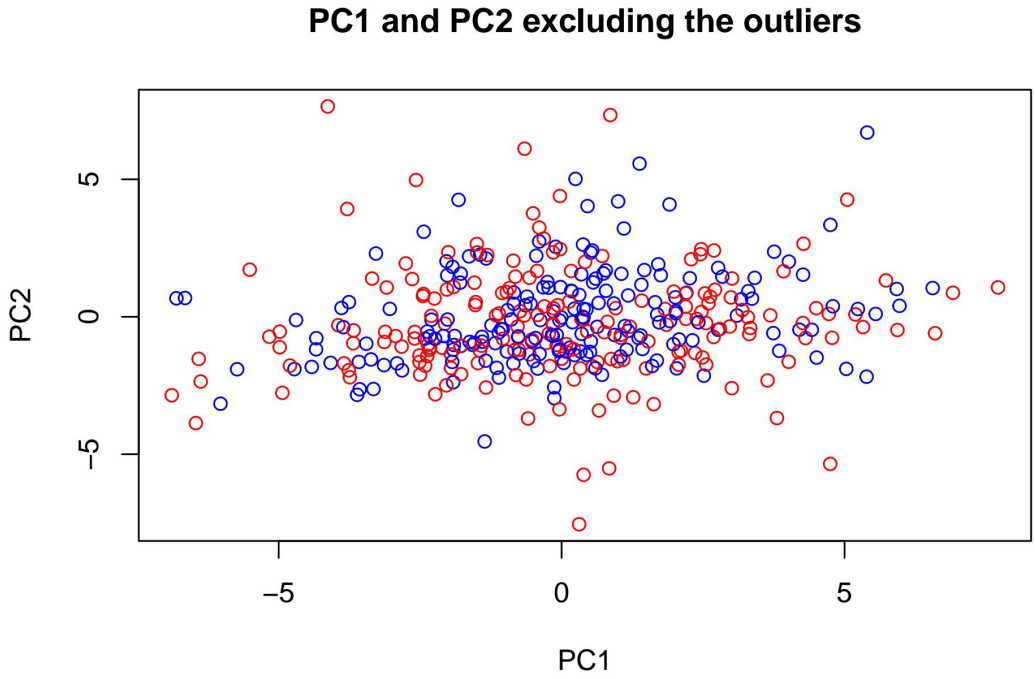
Questions assigned to the following page: 1.1 and 1.2



b. (10 points) There are 1-2 observations that seem perhaps far away from the other points and might be influencing our visualization or PCA. Identify them, and remove them and redo the PCA and the scatterplot. In your R code, print out the date of the observation(s) you remove.

```
# add code here for pca and scatterplot
PCAwaters <- PCAwater$x [,1:2]
# The outliers seems to be those higher than 10.
outliers <- PCAwaters[,2] > 10
pc1 <- PCAwaters[, 1][-c(4, 109)]
pc2 <- PCAwaters[, 2][-c(4, 109)]
# print the date outlying points
plot(pc1, pc2, col = c("red", "blue"), xlab = "PC1", ylab= "PC2",
     main ="PC1 and PC2 excluding the outliers")
```

Questions assigned to the following page: 1.2 and 1.3




```
# Removed observation
water [4,1]
```

```
## [1] "1990-03-13"
```

```
water [109, 1]
```

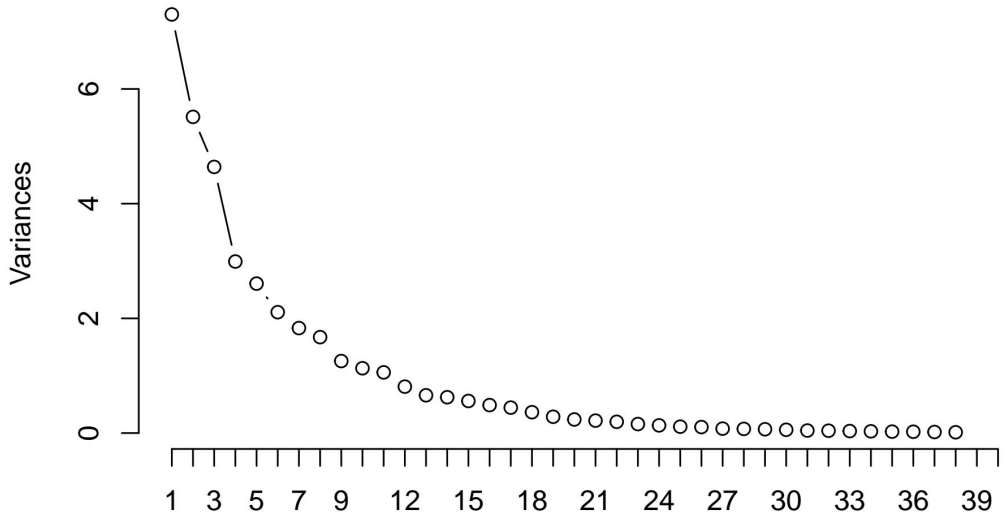
```
## [1] "1990-04-29"
```

If you are interested, you can use the function `identify` to find these points (see help of `identify`). This is an interactive feature in R, but you can use it to find the points, and then once you find them, you can hard-code in your code which ones they are. This is just for interest – you do not have to find them in this way.

c. (5 points) Plot a scree plot of the principal components and comment on how many PC components you would use. Use the PCA after removing the points (i.e. from question b)

```
# add code here for scree plot
waters <- water[-c(4,109),]
PCAwaters2 <-  prcomp((waters[-c(1:5)]), center = T, scale = T)
screeplot(PCAwaters2, type = "line", npcs = 40, main = "PC Components")
```

Question assigned to the following page: 1.3




```
summary(PCAwaters2)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.7017 2.3480 2.1542 1.72947 1.61440 1.45176 1.35308
## Proportion of Variance 0.1921 0.1451 0.1221 0.07871 0.06859 0.05546 0.04818
## Cumulative Proportion  0.1921 0.3372 0.4593 0.53800 0.60659 0.66205 0.71023
##                            PC8     PC9    PC10    PC11   PC12    PC13    PC14
## Standard deviation     1.29318 1.12059 1.06314 1.02878 0.8997 0.81187 0.79051
## Proportion of Variance 0.04401 0.03305 0.02974 0.02785 0.0213 0.01735 0.01644
## Cumulative Proportion  0.75424 0.78729 0.81703 0.84488 0.8662 0.88353 0.89997
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.74816 0.69819 0.66609 0.60317 0.53339 0.48633 0.46588
## Proportion of Variance 0.01473 0.01283 0.01168 0.00957 0.00749 0.00622 0.00571
## Cumulative Proportion  0.91470 0.92753 0.93921 0.94878 0.95627 0.96249 0.96820
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.44337 0.39635 0.36604 0.33382 0.32286 0.27805 0.26996
## Proportion of Variance 0.00517 0.00413 0.00353 0.00293 0.00274 0.00203 0.00192
## Cumulative Proportion  0.97338 0.97751 0.98104 0.98397 0.98671 0.98875 0.99066
##                           PC29   PC30    PC31    PC32    PC33   PC34    PC35
## Standard deviation     0.25770 0.2387 0.21068 0.20254 0.18693 0.1746 0.15777
## Proportion of Variance 0.00175 0.0015 0.00117 0.00108 0.00092 0.0008 0.00066
## Cumulative Proportion  0.99241 0.9939 0.99508 0.99616 0.99708 0.9979 0.99854
##                           PC36    PC37    PC38
## Standard deviation     0.15201 0.13571 0.11890
## Proportion of Variance 0.00061 0.00048 0.00037
```

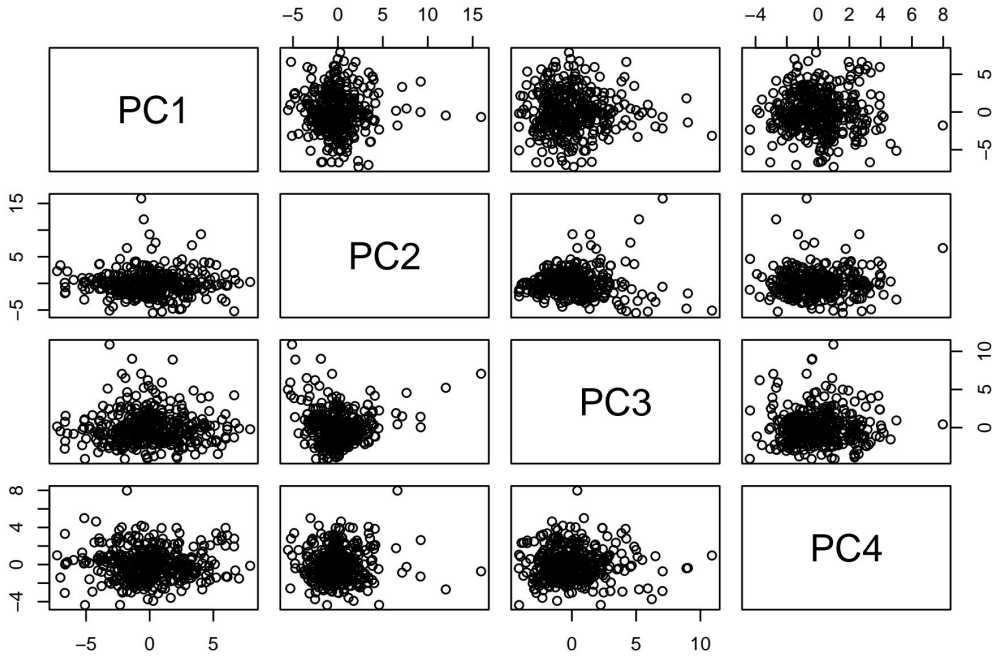
Question assigned to the following page: 1.4

```
## Cumulative Proportion  0.99914 0.99963 1.00000
```

My answer is that I would consider using 7 pc components.

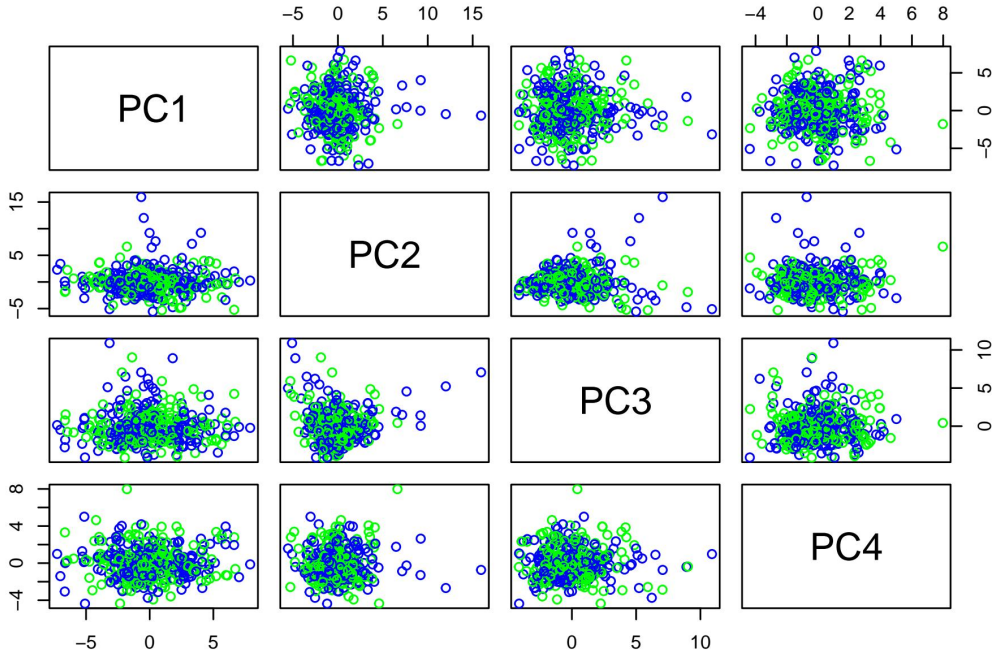
d. (20 points) Plot a pairs plot of the first 4 PCA components and color code them with the categorical variables `Year`, `Season`,`Month`, `Day` using the colors defined above (i.e. 4 separate plots). Do you see evidence that there are strong patterns in the data due to these categorical variables? You should use the PCA after removing the points (i.e. from question b)

```
# add code here for pairs plot of the first 4 components
pairs(PCAwaters2$x[, c(1:4)])
```



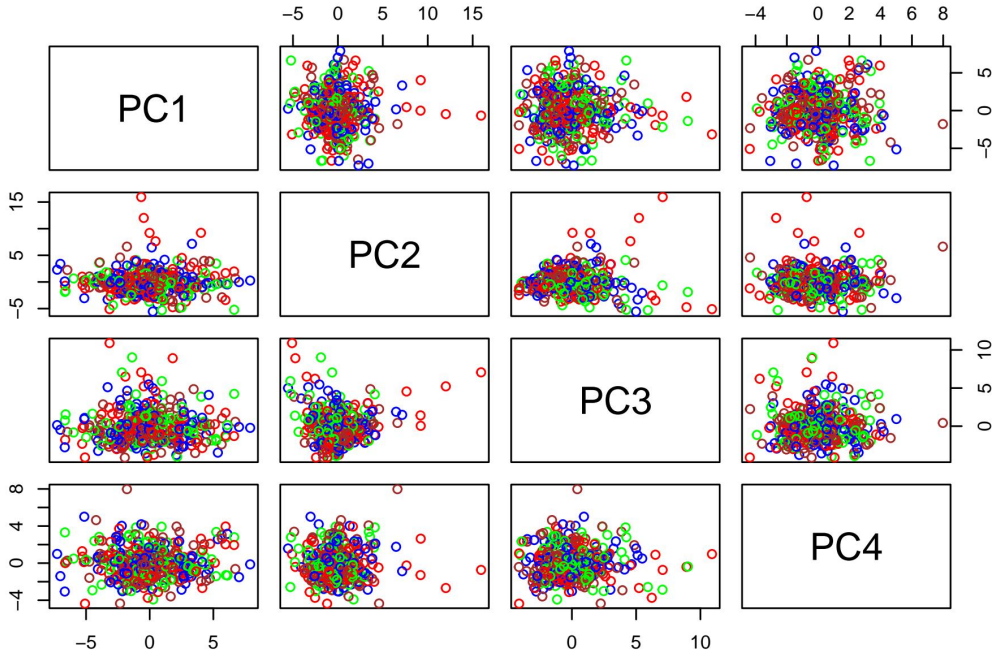

```
pairs(PCAwaters2$x[, c(1:4)], col = colYear)
```

Question assigned to the following page: 1.4



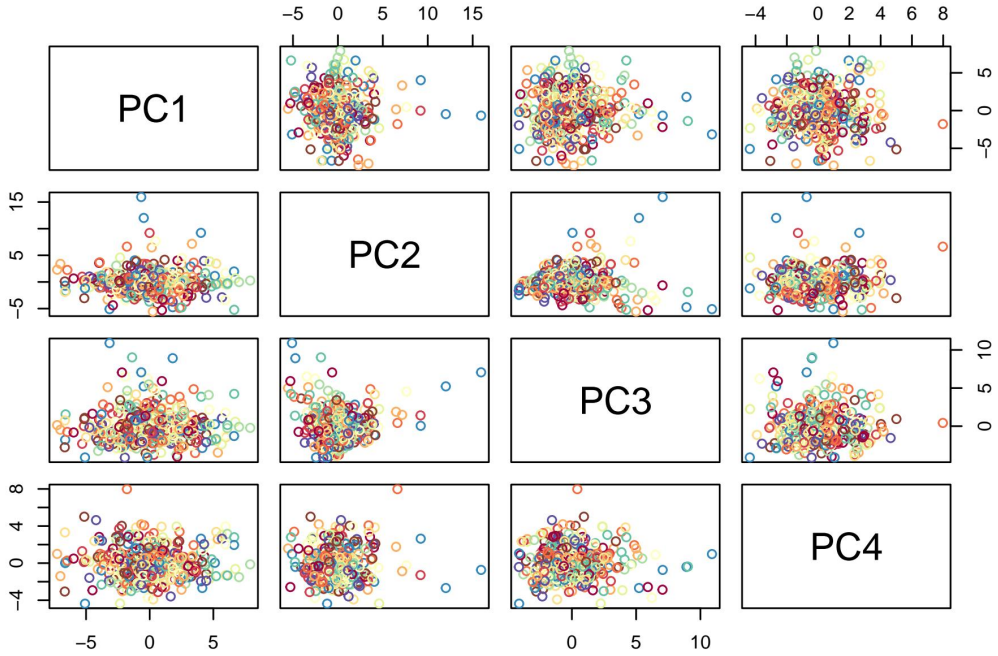

```
pairs(PCAwaters2$x[, c(1:4)], col = colSeason)
```

Question assigned to the following page: 1.4



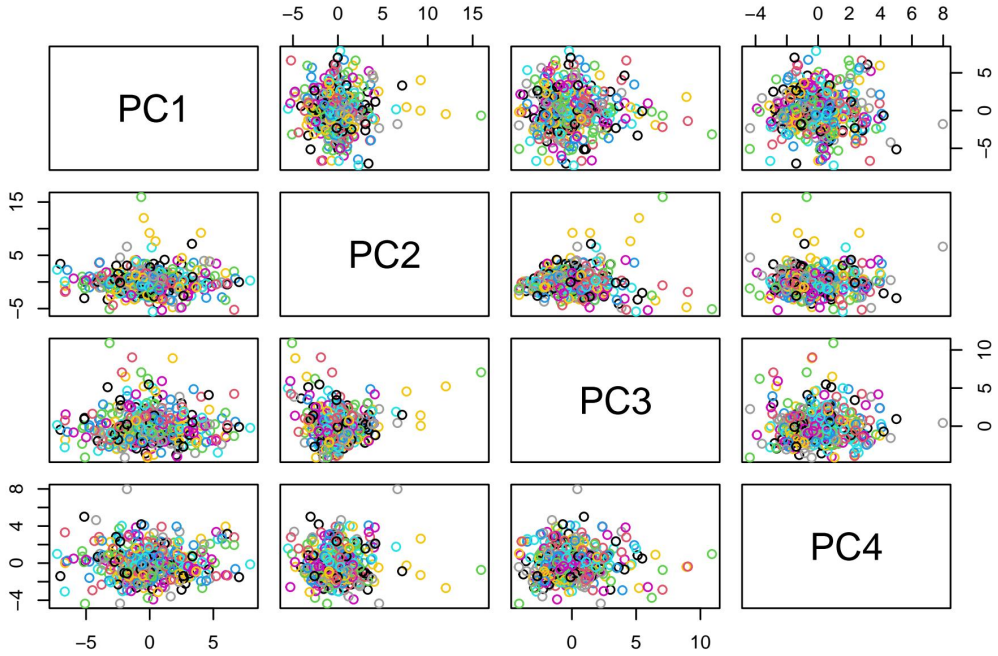

```
pairs(PCAwaters2$x[, c(1:4)], col = colMonths)
```

Question assigned to the following page: 1.4




```
pairs(PCAwaters2$x[, c(1:4)], col = colDays)
```

Questions assigned to the following page: 2.1 and 1.4



My answer is that plotting the plots, we can't conclude that there is evidence that it changes due to categorical variables. It just looks similar to each other, which we can conclude that there is much evidence that it causes differences.

Question 2: Regression

a. (2 points) I have a dataset containing average hourly earnings in dollars (wage) and years of education (educ) for 526 individuals. I fit a simple linear regression equation with wage as response and educ as the explanatory variable. This gave me the following equation:

wage = -0.90485 + 0.54136 * (educ).

Which among the following is the correct interpretation for this equation? Give reasons for your answer.

(i) For every additional four years of education, the average hourly wage increases by 4*0.54 = 2.16 dollars.

(ii) For every additional year of education, the average hourly wage increases by 54%.

(iii) For every 1% increase in the number of years of education, the average hourly wage increases by 0.54%.

My answer is (i). For every additional four years of education, the average hourly wage increases by $4 0.54 = 2.16$ *dollars. According to the sentence, the coefficient of education is 0.54 which indicates, with increase of 0.54 in education, wage will follow up by 0.54, which means with 4 years of education, it will follow with the same equation given, 4* 0.54. It seems to follow the lin-lin model.

Questions assigned to the following page: 2.2, 2.3, and 3

b. (2 points) For the same dataset as in the previous part, I fit a simple linear regression equation with log(wage) as response and educ as the explanatory variable. This gave me the following equation:

log(wage) = 0.583773 + 0.082744 * (educ).

Which among the following is the correct interpretation for this equation? Give reasons for your answer.

(i) For every additional year of education, the average hourly wage increases by 0.0827 dollars.

(ii) For every additional year of education, the average hourly wage increases by 8.27 percent.

(iii) For every additional year of education, the average hourly wage increases by 0.0827 percent.

My answer is (ii) For every additional year of education, the average hourly wage increases by 8.27 percent. The coefficient is 0.082744, which converted into percentage will be 0.082744 * 100 = 8.27 percent. It seems to follow the log-lin model as we have wage converted to log.

c. (2 points) I have a dataset on the salaries of the CEOs of 209 firms (variable name is salary) along with the sales of the firm (variable is sales). The dataset is from the year 1990. Salary is in thousands of dollars and Sales is in millions of dollars. I fit a simple linear regression with log(salary) as the response variable and log(sales) as the explanatory variable and this gave me the equation:

log(salary) = 4.822 + 0.25667 * log(sales).

Which among the following is the correct interpretation for this equation? Give reasons for your answer.

(i) For a 1 percent increase in sales, the CEO salary increases by 0.257 percent on average.

(ii) For a 1 million dollar increase in firm sales, the CEO salary increases by 25.667 thousand dollars on average.

(iii) For a 1 million dollar increase in firm sales, the CEO salary increases by 2.57 percent.

My answer is (i) For a 1 percent increase in sales, the CEO salary increases by 0.257 percent on average. The coefficient given is 0.25667, which means that with 1 percent increase in sales, it follows up by 0.257 percentage. Following the log-log model.

Question 3: Regression output

(15 points) The following is a the output of running `lm` on a subset of the imdb dataset you will work with in a future homework. The below output above has five missing values which are indicated by XXAXX-XXFXX Using only the available information in the above summary, fill in the missing values. I give you space below for R code, but this is just for using R as a *calculator* – i.e. show your computations in the R code space by correctly typing (manually) the values printed in the output below.

```
# director facebook likes t-value and p-value for XXAXX, XXBXX
XXAXX <- sum((1.079e-05)/(1.095e-05))
XXBXX <- 2* pt (0.985, 558, lower = FALSE)
XXFXX <- pf(34.16, 13, 558, lower.tail = FALSE)
XXFXX
```

```
## [1] 1.397271e-62
```

My answer is that starting with XXAXX = 0.9853881, XXBXX = 0.325051, XXCXX = 13, XXDXX = 558, XXFXX = 1.397271e-62

```
summary(lmMoviesSmall2)
```

Question assigned to the following page: 3

```
Call:
lm(formula = imdb_score ~ ., data = moviesSmall2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6138 -0.4630  0.0876  0.5490  1.9408

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.425e+01  9.482e+00   4.667 3.84e-06 ***
num_critic_for_reviews     2.540e-03  5.055e-04   5.025 6.78e-07 ***
duration                   1.111e-02  1.710e-03   6.497 1.81e-10 ***
director_facebook_likes    1.079e-05  1.095e-05   XXAXX    XXBXX
actor_3_facebook_likes     9.128e-05  5.226e-05   1.747  0.08126 .
actor_1_facebook_likes     8.848e-05  3.153e-05   2.807  0.00518 **
gross                      1.662e-10  6.693e-10   0.248  0.80399
num_voted_users            3.746e-06  4.309e-07   8.694  < 2e-16 ***
cast_total_facebook_likes -7.583e-05  3.127e-05  -2.425  0.01564 *
num_user_for_reviews      -7.565e-04  1.575e-04  -4.804 2.01e-06 ***
budget                    -4.223e-09  1.025e-09  -4.122 4.33e-05 ***
title_year                -1.973e-02  4.727e-03  -4.175 3.46e-05 ***
actor_2_facebook_likes     6.026e-05  3.242e-05   1.859  0.06362 .
movie_facebook_likes      -1.150e-06  2.366e-06  -0.486  0.62723
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8023 on 558 degrees of freedom
Multiple R-squared:  0.4432,    Adjusted R-squared:  0.4302
F-statistic: 34.16 on XXCXX and XXDXX DF,  p-value: XXFXX
```

Question 4: Regression with categorical predictors

Consider the data in `ceodata_num.csv` which consists of 209 firms and has data on the salary of the CEO, sales of the firm, and the firm type. The data is from the year 1990.

```
ceodata<-read.csv("ceodata_num.csv")
ceodata
```

```
##     salary   sales FirmType
## 1     1095 27595.0        3
## 2     1001  9958.0        3
## 3     1122  6125.9        3
## 4      578 16246.0        3
## 5     1368 21783.2        3
## 6     1145  6021.4        3
## 7     1078  2266.7        3
## 8     1094  2966.8        3
## 9     1237  4570.2        3
## 10     833  2830.0        3
## 11     567   596.8        3
## 12     933 19773.0        3
## 13    1339 40047.0        3
## 14     937  2513.8        3
## 15    2011  1580.6        3
```

No questions assigned to the following page.

```
## 16    1585  6754.0        3
## 17     905  1066.3        3
## 18    1058  3199.9        3
## 19     922  1452.7        3
## 20    1220  8995.0        3
## 21    1022  1212.3        3
## 22     759  2824.2        3
## 23    1414  7621.0        3
## 24    1041  4418.3        3
## 25    1688 12343.0        3
## 26    2983 57662.0        3
## 27    1160  4319.7        3
## 28    3844 20604.0        3
## 29     476   611.3        3
## 30    1492 12431.7        3
## 31    1024  8169.0        3
## 32    1593 20659.0        3
## 33     427  3072.1        3
## 34     829  1669.1        3
## 35     797  2401.2        3
## 36     577  3043.6        3
## 37    1342  6222.2        3
## 38    1774  7727.9        3
## 39     709   719.5        3
## 40     860  3921.3        3
## 41    1336  5155.1        3
## 42     516   649.2        3
## 43     931 10885.0        3
## 44     815  1651.9        3
## 45    1681 12915.0        3
## 46     568 11436.0        3
## 47     775  2210.3        3
## 48    1188  3737.8        3
## 49     782  1976.7        3
## 50    1170  2576.0        3
## 51    1469  6309.1        3
## 52     916  2940.5        3
## 53    1070  1072.6        3
## 54     894  1534.0        3
## 55     829  2158.8        3
## 56     780  1736.0        3
## 57    2327  3598.8        3
## 58     717  1413.6        3
## 59    1368 13538.0        3
## 60    2028 69018.0        3
## 61    1195  6285.0        3
## 62     256   526.3        3
## 63     775  3195.6        3
## 64    1407  2762.8        3
## 65     543  1873.1        3
## 66     874  4648.1        3
## 67    1287 16951.0        3
## 68    1248  3506.9        2
## 69     875  5333.1        2
```

No questions assigned to the following page.

```
## 70     925  3296.2        2
## 71     798  2584.0        2
## 72     760   834.4        2
## 73     600  4068.7        2
## 74     991  2518.6        2
## 75    1570 10465.0        2
## 76     911  2682.4        2
## 77    1360  2688.2        2
## 78     700  6682.3        2
## 79     741  4879.9        2
## 80    1097  2772.1        2
## 81     953  1320.4        2
## 82     441  3408.5        2
## 83     595  1919.7        2
## 84    1067 19020.5        2
## 85    1298  4778.3        2
## 86    1798 24332.0        2
## 87    4143  2678.4        2
## 88    1336  4481.0        2
## 89    1750 14932.1        2
## 90     912  2626.4        2
## 91    1892  2659.5        2
## 92     833  1073.2        2
## 93    1142  2577.3        2
## 94    1159  4293.8        2
## 95    1283 18164.0        2
## 96    2109  9944.4        2
## 97    1039 12719.0        2
## 98     992  1931.6        2
## 99    1253  2993.2        2
## 100    721  2127.2        2
## 101   1351  4211.9        2
## 102   1391  8489.6        2
## 103   1245 12222.8        2
## 104   1550 11213.4        2
## 105   2150  5869.6        2
## 106   1846  6193.8        2
## 107    573  3053.2        2
## 108   6640  8946.0        2
## 109    959  4045.2        2
## 110    612  3618.9        2
## 111   1820  1796.1        2
## 112   1411  6703.1        2
## 113   1026  2170.3        2
## 114   1287  4674.0        4
## 115    800  3372.0        4
## 116   1115  8205.0        4
## 117   1631  4617.0        4
## 118   1910  6964.0        4
## 119    996  6623.0        1
## 120    918 29797.0        1
## 121   1261  5225.5        1
## 122   1053  3639.0        1
## 123   1221 97649.9        1
```

No questions assigned to the following page.

```
## 124   1738 10743.6        1
## 125   3142 10236.3        1
## 126   1900 17802.7        1
## 127    427  6158.7        1
## 128   1700  6775.2        1
## 129    360   298.7        1
## 130    459   785.3        1
## 131   1340  1368.6        1
## 132    729  2066.1        1
## 133    223   181.5        1
## 134   2101 10300.0        1
## 135   1082 11232.0        1
## 136   1781  5191.6        1
## 137    791   976.7        1
## 138   2092  7671.5        1
## 139   1573  6406.0        1
## 140   1045  2917.4        1
## 141   1694  3322.9        1
## 142    453   175.2        1
## 143   1130  1686.3        1
## 144   1334   514.1        1
## 145   1344  3032.7        1
## 146   1585  4686.9        1
## 147   1946  8388.1        1
## 148   1619  7632.8        1
## 149   1620 17453.8        1
## 150    967  5781.1        1
## 151   1431  6896.2        1
## 152   1231  6503.1        1
## 153    770  2715.6        1
## 154   1594  2761.9        1
## 155   1568  5181.4        1
## 156    995  1323.0        1
## 157   1077  5296.0        1
## 158   1161  7177.0        1
## 159   1401 12183.5        1
## 160   1127  7863.5        1
## 161   3068  3825.6        1
## 162    730  1110.6        1
## 163    729  1391.3        1
## 164  11233  6047.9        1
## 165    949 18908.0        1
## 166   3646  3921.5        1
## 167   1502  3453.8        1
## 168    807  1558.5        1
## 169    713   962.8        1
## 170   1489 25848.0        1
## 171    736   631.5        1
## 172   1226  1728.9        1
## 173    543  2669.3        1
## 174  14822  2159.2        1
## 175    890  2612.6        1
## 176   1627  8270.3        1
## 177   2408 44323.0        1
```

Question assigned to the following page: 4.1

```
## 178   2248   764.7        1
## 179    787  5167.5        4
## 180    474  2159.3        4
## 181    439  2617.1        4
## 182    465  2367.7        4
## 183    594  2744.0        4
## 184    688  2357.9        4
## 185    607  5262.0        4
## 186    634  5738.9        4
## 187    532  2714.9        4
## 188    441  3306.7        4
## 189    694  3532.5        4
## 190    520  3681.5        4
## 191    757  3982.1        4
## 192    668  2996.0        4
## 193    803  4178.6        4
## 194    500  2616.3        4
## 195    552  2064.5        4
## 196    412  2225.5        4
## 197   1100  9470.1        4
## 198    959  3783.2        4
## 199    333  2388.7        4
## 200    503  3705.2        4
## 201    448  1411.7        4
## 202    732  4800.1        4
## 203    720  1771.9        4
## 204    808  7198.5        4
## 205    930  1509.1        4
## 206    525  1097.1        4
## 207    658  4542.6        4
## 208    555  2023.0        4
## 209    626  1442.5        4
```

`salary`: Salary of the CEO in thousands of dollars `sales`: Sales of company in millions of dollars `FirmType`: the type of company coded as numeric values 1-4 which correspond to:

- 1=consumer product
- 2=finance
- 3=industry
- 4=utility

a (10 points) Fit a regression in R with `sales` and `FirmType` as a predictor of `salary`. Interpret each of the coefficient estimates given by R, except for the intercept.

```
# Code for fitting regression here
Firmtype.f = factor(ceodata$FirmType)
sum <- summary(lm(salary ~sales +Firmtype.f, data=ceodata))
sum
```

```
##
## Call:
## lm(formula = salary ~ sales + Firmtype.f, data = ceodata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1620.3  -425.8  -162.1    71.2 13173.5
```

Questions assigned to the following page: 4.1 and 4.2

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.621e+03  1.866e+02   8.689  1.2e-15 ***
## sales        1.249e-02  8.833e-03   1.414  0.15882
## Firmtype.f2 -3.496e+02  2.625e+02  -1.332  0.18444
## Firmtype.f3 -5.867e+02  2.374e+02  -2.471  0.01429 *
## Firmtype.f4 -9.396e+02  2.842e+02  -3.305  0.00112 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1336 on 204 degrees of freedom
## Multiple R-squared:  0.07097,    Adjusted R-squared:  0.05275
## F-statistic: 3.896 on 4 and 204 DF,  p-value: 0.004517
```

```
sum[["coefficients"]]
```

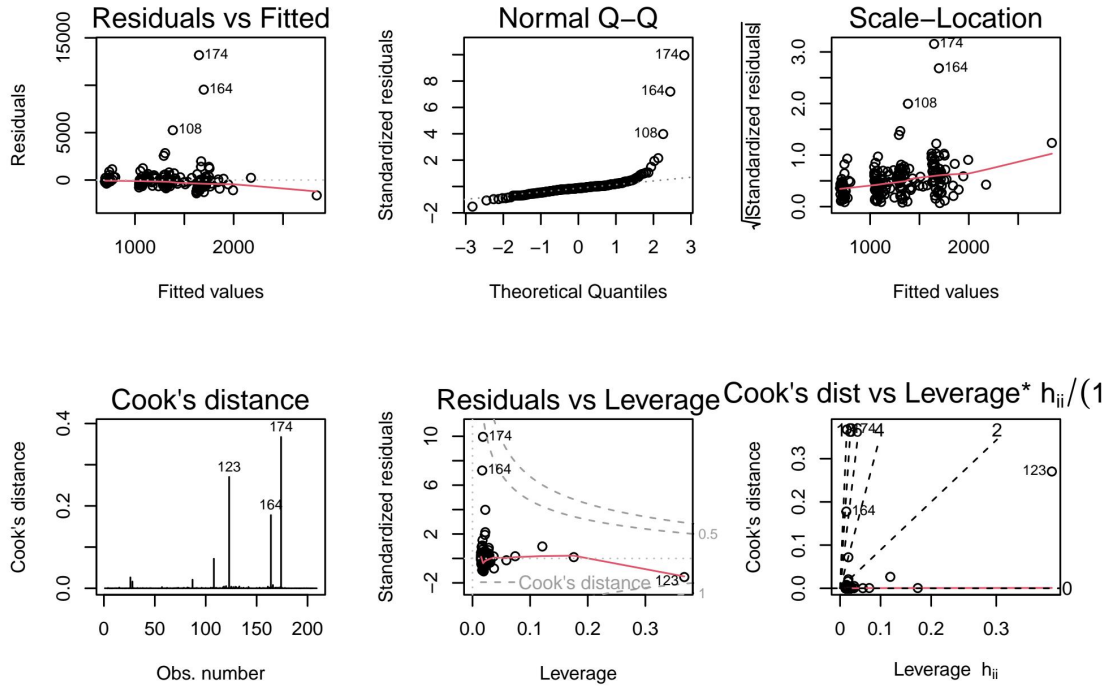
```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) 1621.48367778 1.866189e+02  8.688745 1.195467e-15
## sales          0.01249148 8.832727e-03  1.414227 1.588198e-01
## Firmtype.f2 -349.57142455 2.624980e+02 -1.331711 1.844412e-01
## Firmtype.f3 -586.68641503 2.374153e+02 -2.471140 1.428783e-02
## Firmtype.f4 -939.55250768 2.842436e+02 -3.305448 1.120224e-03
```

My answer is that every increase of one million dollars in sales would also increase in average salary by 12.49 dollars. For type two firms, which is the finance firms, will decrease in 349571.42 dollars in average sales as 1 unit increase releative to the consumer product. For type three firms, the industry firms, will decrease in 586686.42 dollars in average sales as 1 unit increases, and for lasty the last type four firm, the utility firms, there will be an 939552.51 dollars decrease in average sales for every increase of one unit. The industry and utility firms stand out in p-value as they have less than 0.05 p-values which is statistically significant.

b. (20 points) Run diagnostics on the model you found in part a, and determine whether there are any problems with this model that should be addressed. If so, explain next steps you might take to try to improve this model.

```
# Code for diagnostics
par(mfrow =c(2,3))
plot(lm(salary ~sales + Firmtype.f, data=ceodata), which = 1:6)
```

Question assigned to the following page: 4.2



My answer is that there are problems with this model that should be addressed. First looking into the Residuals vs Fitted plot, the points are clustering. Looking at the Normal Q-Q plot, the right side of the residuals are not aligned which seems that it has problems, therefore not normal. For the Scale-Location plot, as it shows that residual points are in cone shape, which indicates an heteroscedasticity. Looking at the Cook's distance and Residuals vs Leverage, there seems to be an outlier, which is affecting the whole fit. There are some plots that are not normal, therefore there are problems with the data. To solve these kinds of problems, we should try to remove the outliers, and try to get a better model in whole.