Data Report on the US Natality dataset

Name: Keerti Chalasani

Abstract:

In this report, I will conduct an analysis on the US Natality data from the Center for Disease Control and Prevention(CDC)

```
#load the data
rm(list = ls())
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

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

```
#birth<- read_csv("~/Desktop/FALL2019/STAT448/DCIR04/Nat2015parts/nat15p1.csv")

birth <- read_csv("https://uofi.box.com/shared/static/k0oo6r8h4sdgv5nj3fnceuici4eisne2.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   MAGE_IMPFLG = col_logical(),
##   MAGE_REPFLG = col_logical(),
##   MAR_P = col_character(),
##   MAR_IMP = col_logical(),
##   FAGERPT_FLG = col_logical(),
##   WIC = col_character(),
##   CIG_REC = col_character(),
##   RF_PDIAB = col_character(),
##   RF_GDIAB = col_character(),
##   RF_PHYPE = col_character(),
##   RF_GHYPE = col_character(),
##   RF_EHYPE = col_character(),
##   RF_PPTERM = col_character(),
##   RF_INFTR = col_character(),
##   RF_REDRG = col_character(),
##   RF_ARTEC = col_character(),
##   RF_CESAR = col_character(),
##   IP_GON = col_character(),
##   IP_SYPH = col_character(),
##   IP_CHLAM = col_character()
##   # ... with 42 more columns
## )
```

```
## See spec(...) for full column specifications.
```

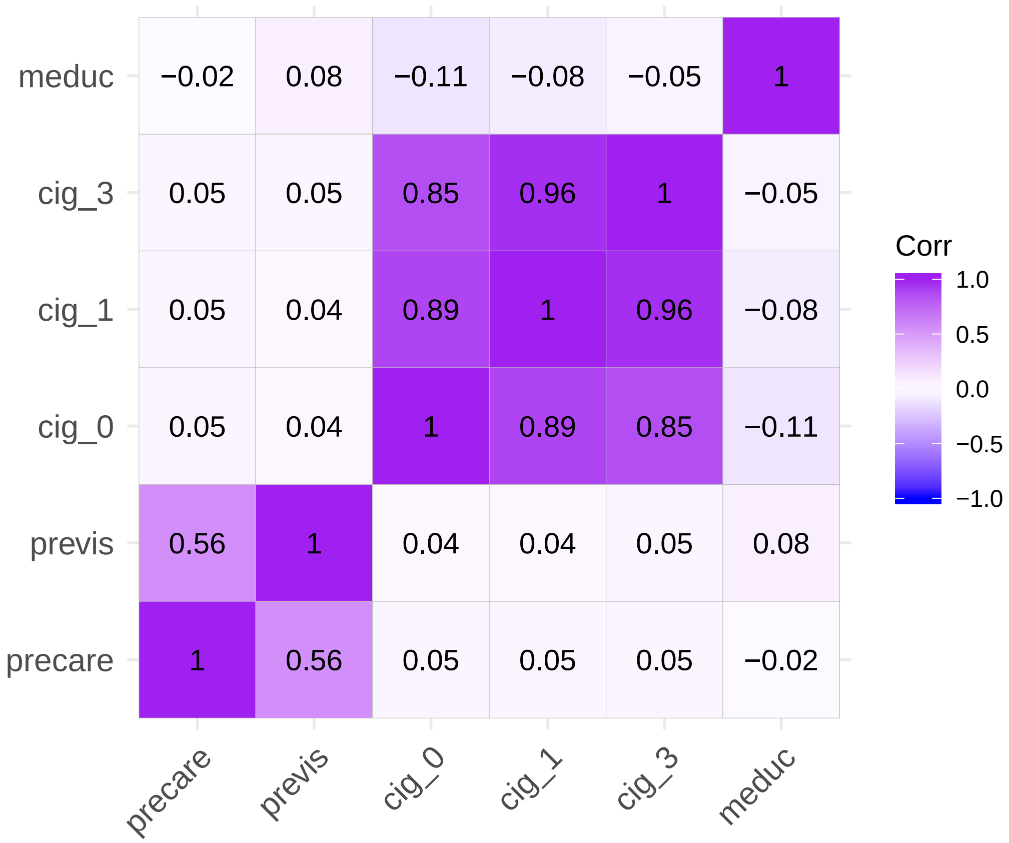
```
colnames(birth) <- tolower(colnames(birth))
dim(birth)
```

```
## [1] 200000    241
```

Introduction:

In this analysis, I will be looking at the US Natality Data for the year for 2015 which consists of 200,000 and 226 rows. The Natality data consists of many variables including but not limited to, the number of prenatal visits, cigarette use, level of education of the mother, age, height, race, BMI, and weight gain during pregnancy. For the first cycle of analysis, I was curious to see if the education level has an impact on if a mother seeks prenatal care or if the mother smokes cigarettes. I am assuming that if the mother obtained a higher education or even graduated high school she would be more aware of seeking out prenatal care and would know that one should not smoke cigarettes while pregnant. I will be using the variables, precare(month prenatal care began), previs(number of prenatal visits), cig_0(cigarettes before pregnancy),cig_1(cigarettes during the first trimester), cig_3(cigarettes during the third trimester) and meduc(education level of the mother). With these variables, I am going to create a correlation matrix and will be using principal component analysis(PCA).

```
library(ggcorrplot)
corr_birth = cor(select(birth, precare, previs, cig_0, cig_1, cig_3, meduc))
ggcorrplot(corr_birth,lab=TRUE, colors = c("blue", "white", "purple"))
```



Analysis:

As you can see from the analysis if a woman was smoking cigarettes before pregnancy it was highly likely that she was smoking cigarettes during the first trimester and the third trimester with a correlation of .89 and .85 respectively. From the correlation matrix, we can also see that the education level and cigarette use have a negative correlation. We can also see that education level and visits for prenatal care have a very small correlation that is almost zero.

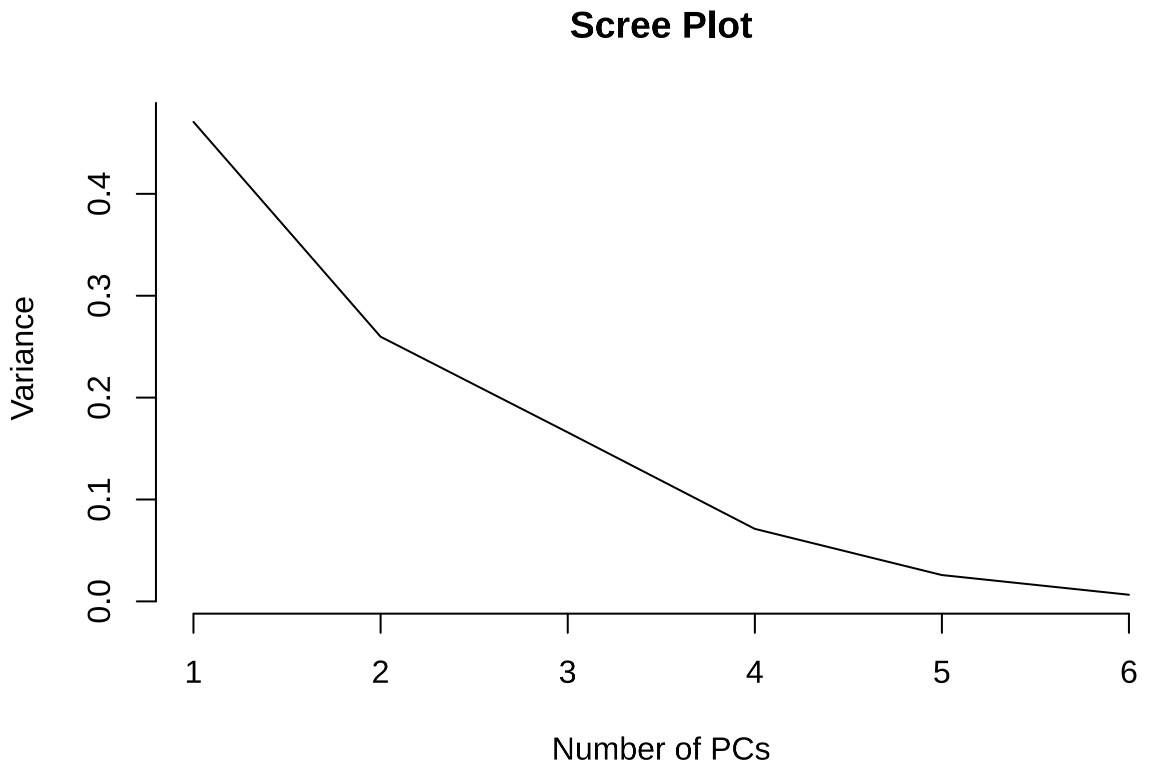
```
pcad <- prcomp(~ precare+ previs+ cig_0+ cig_1+ cig_3+meduc, data = birth, scale = TRUE)
eval<-pcad$sdev^2
eval/sum(eval)
```

```
## [1] 0.470654936 0.259814610 0.165957260 0.071207395 0.025851418 0.006514382
```

```
eval[eval>mean(eval)]
```

```
## [1] 2.823930 1.558888
```

```
# R code
plot(1:length(pcad$sdev),pcad$sdev^2/sum(pcad$sdev^2), type="l", axes=FALSE,xlab="Number of PCs", ylab=
axis(1, seq(1,length(pcad$sdev),1))
axis(2, seq(0,round(max(eval/sum(eval)),1),0.1) )
```



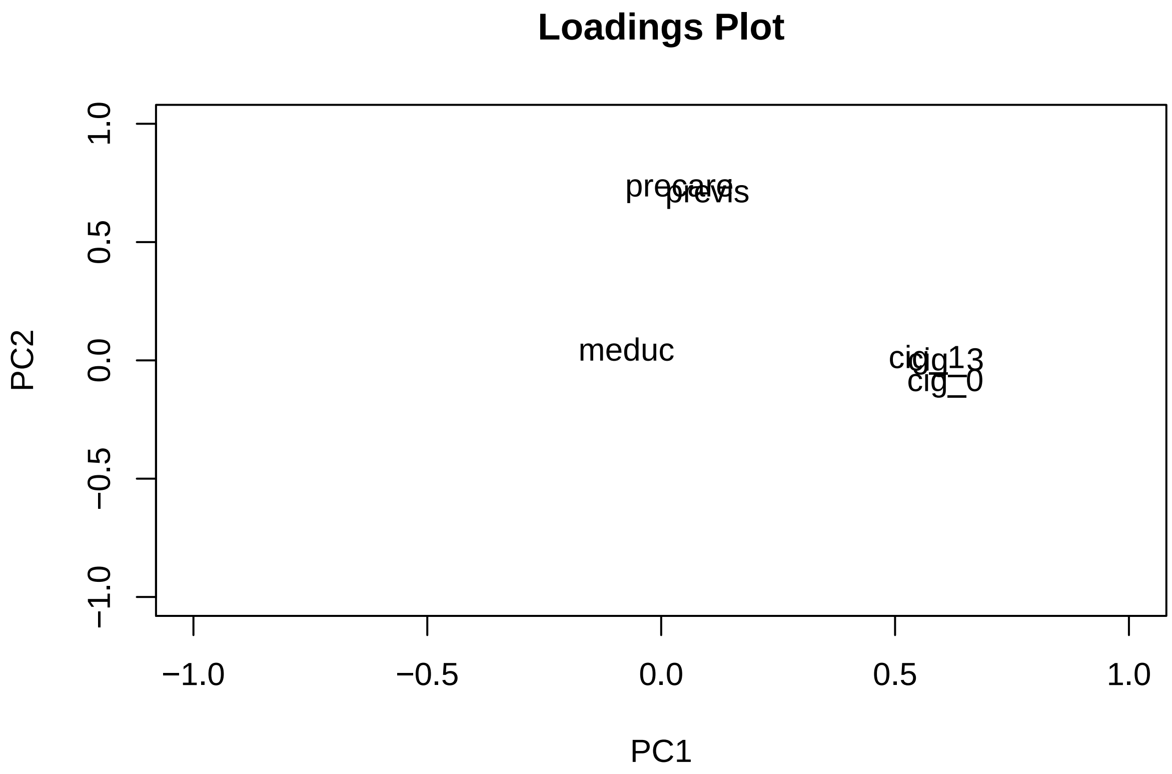
This scree plot shows that there should be four factors generated by the analysis. We will keep the eigenvalues above the scree 'elbow' which is at 4. We would like to keep four components.

```
## putting the eigenvectors as a table
evec <- pcad$rotation
evec
```

```
##                 PC1         PC2          PC3          PC4          PC5
## precare  0.06674474  0.69747611  0.142794003  0.698991226 -0.008993646
## previs   0.05731658  0.70630683 -0.024726840 -0.705130725  0.004070711
## cig_0    0.56248796 -0.05204259 -0.007509176 -0.011495485 -0.807153327
## cig_1    0.58340718 -0.04574955 -0.053945106  0.008030230  0.246739967
## cig_3    0.57452933 -0.04259176 -0.078182710  0.007376419  0.533769662
## meduc   -0.07361622  0.08967846 -0.984844890  0.118113860 -0.051140856
##                  PC6
## precare  0.003531509
## previs  -0.002853147
## cig_0    0.170920543
## cig_1   -0.770511431
## cig_3    0.614023609
## meduc   -0.007259370
```

The eigenvectors are both positive and negative in direction and not all of them are near zero. Some of the eigenvectors are around .7 while some are closer to zero at -.007. With this, we can interpret these components as having contrasting behavior.

```
#Loadings Plot
plot(evec, type = "n", xlim = c(-1,1), ylim = c(-1,1), main = "Loadings Plot")
text(jitter(evec, amount = 0.05), labels = rownames(evec))
```



The loading plot shows that the mother's education is in the center which means it isn't correlated with anything else. There is one cluster on this loading plot and that cluster contains cigarette use which is expected since cigarette use before and during pregnancy has a positive correlation. while prenatal care and month of prenatal care are also in a cluster which shows that there is a correlation between them and ther far away from the origin.

Conclusion:

In conclusion, the main part of this analysis was to see if there is a relationship of some sort with the level of education and cigarette use/seeking prenatal care. I looked at multiple variables such as pre-care(month prenatal care began), previs(number of prenatal visits), cig_0(cigarettes before pregnancy),cig_1(cigarettes during the first trimester), cig_3(cigarettes during the third trimester) and meduc(education level of the mother) and preformed PCA analysis including a correlation matrix to be able to visualize the correlation between the variables. In this report, I found that there isn't a relationship between a mother's education level and cigarette use and that there isn't a relationship with the mother's education level and seeking prenatal care. This was interesting to look at, I would be curious to see what other variables were correlated with cigarette use during pregnancy and if there were ways to stop cigarette use during pregnancy by doing further analysis on this topic.

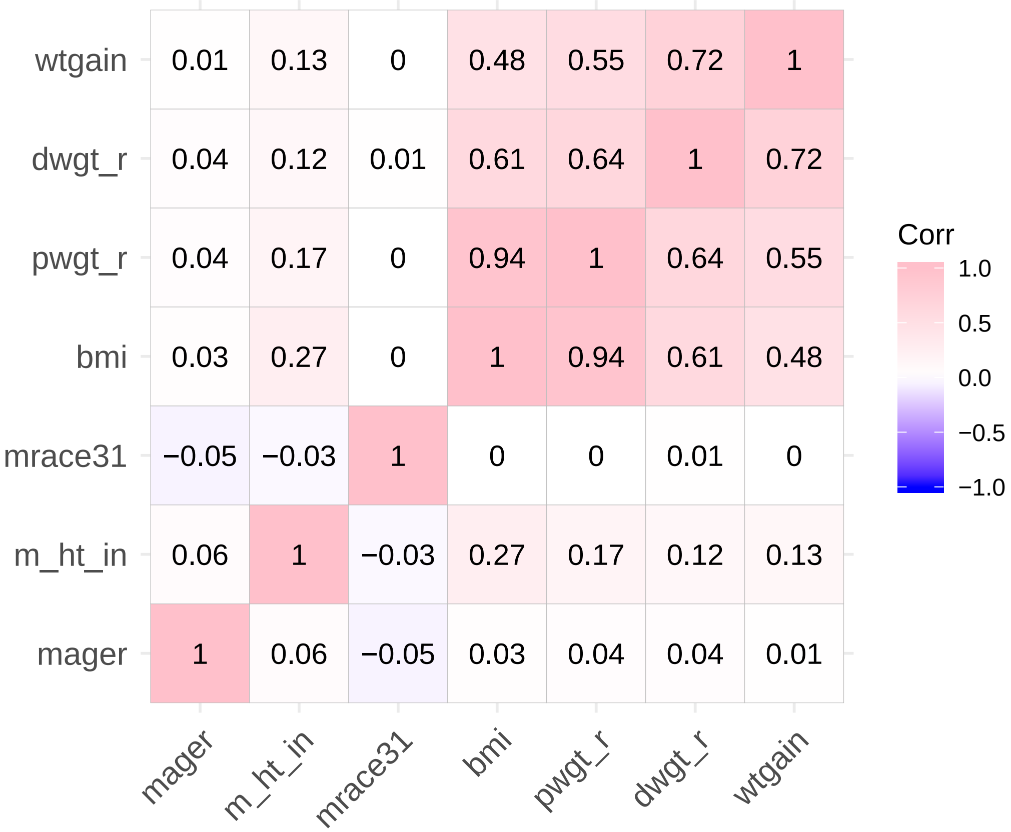
Abstract:

In this report, I will conduct an analysis on the US Natality data from the Center for Disease Control and Prevention(CDC)

Introduction:

In this analysis, I will be looking at the US Natality Data for the year for 2015 which consists of 200,000 and 226 rows. The Natality data consists of many variables including but not limited to, the number of prenatal visits, cigarette use, level of education of the mother, age, height, race, BMI, and weight gain during pregnancy. For the second cycle of analysis, I wanted to see if age, height, race affected the amount of weight that was gained during pregnancy. I also want to see if pre-pregnancy weight and BMI are correlated with weight gain. From prior knowledge, I'm assuming that there might be a positive relationship between these variables. I will be using the variables, mager(monther's age), m_ht_in(mother's height in inches), mrace31(mothers race),bmi(mother's bmi), pwgt_r(pre-pregnancy weight) and dwgt_r(delivery weight), wtgain(weight gain). With these variables, I am going to create a correlation matrix and will be using principal component analysis(PCA).

```r
library(ggcorrplot)
corr_weight = cor(select(birth, mager,m_ht_in,mrace31, bmi, pwgt_r, dwgt_r, wtgain))
ggcorrplot(corr_weight,lab=TRUE,colors = c("blue", "white", "pink"))
```



	mager	m_ht_in	mrace31	bmi	pwgt_r	dwgt_r	wtgain
wtgain	0.01	0.13	0	0.48	0.55	0.72	1
dwgt_r	0.04	0.12	0.01	0.61	0.64	1	0.72
pwgt_r	0.04	0.17	0	0.94	1	0.64	0.55
bmi	0.03	0.27	0	1	0.94	0.61	0.48
mrace31	−0.05	−0.03	1	0	0	0.01	0
m_ht_in	0.06	1	−0.03	0.27	0.17	0.12	0.13
mager	1	0.06	−0.05	0.03	0.04	0.04	0.01

As you can see from the correlation matrix BMI and weight gain during pregnancy have a slight positive correlation of .48, however, I thought it would be more than that. It is expected that pre-pregnancy weight and BMI would have a strong positive correlation which is true since the correlation is .94. Weight gain and pre-pregrnacy weight also have a positive correlation of 0.55 while age and weight gain don't really have a correlation. Race doesn't really have an effect on any of the variables with a very small correlation or 0 for all of them.

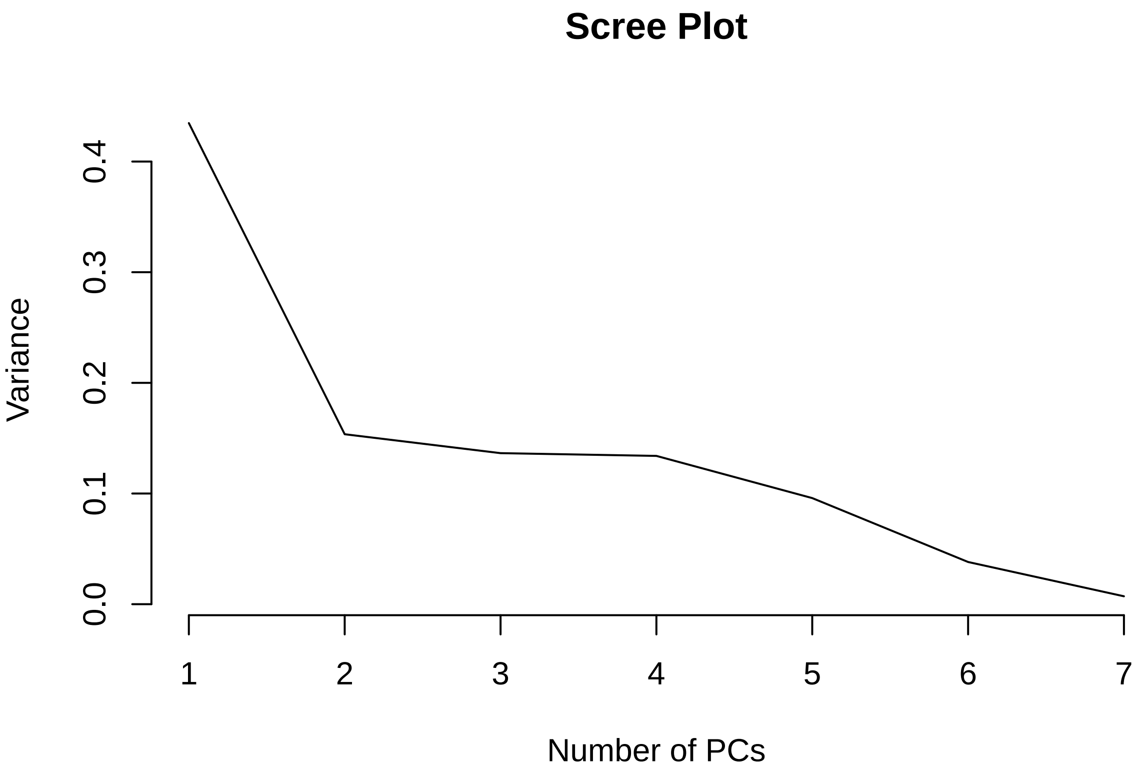
```
pcad2 <- prcomp(~   mager+ +m_ht_in+mrace31+ bmi+ pwgt_r+ dwgt_r +wtgain, data = birth, scale = TRUE)
evalue<-pcad2$sdev^2
evalue/sum(evalue)
```

```
## [1] 0.434760965 0.153574356 0.136491928 0.133996201 0.095885062 0.038138210
## [7] 0.007153279
```

```
evalue[evalue>mean(evalue)]
```

```
## [1] 3.043327 1.075020
```

```
# R code
plot(1:length(pcad2$sdev),pcad2$sdev^2/sum(pcad2$sdev^2), type="l", axes=FALSE,xlab="Number of PCs", yl
axis(1, seq(1,length(pcad2$sdev),1))
axis(2, seq(0,round(max(evalue/sum(evalue)),1),0.1) )
```



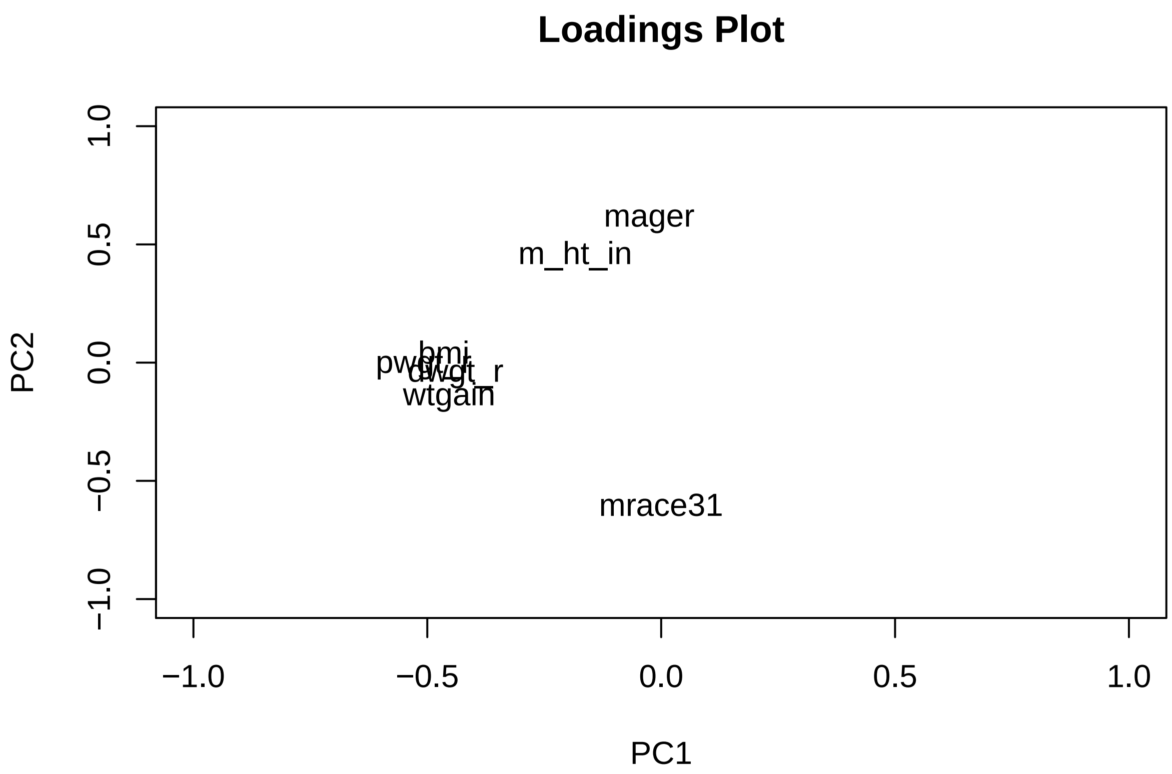
This scree plot shows that there should be four factors generated by the analysis. We will keep the eigenvalues above the scree 'elbow' which is around 2. We would like to keep two components.

```
# R code
## eigenvectors as a table
evector <- pcad2$rotation
evector
```

```
##                  PC1         PC2         PC3          PC4         PC5
## mager   -0.035754538  0.64556563 -0.40812490  0.643542675  0.02014255
## m_ht_in -0.171275536  0.42884214 -0.44171051 -0.708577430 -0.28649776
## mrace31 -0.001123519 -0.61416077 -0.77933566  0.121712251 -0.01884027
## bmi     -0.513262312  0.02398301 -0.04917513 -0.109143153  0.48754617
## pwgt_r  -0.523723110 -0.02629615  0.01664318 -0.005038567  0.44253061
## dwgt_r  -0.484888604 -0.09548116  0.09546888  0.174235490 -0.33924716
## wtgain  -0.443332817 -0.10845576  0.13839191  0.163283062 -0.60707093
##                 PC6          PC7
## mager   -0.02809094  0.007660996
## m_ht_in  0.03029842 -0.081055540
## mrace31 -0.01614173 -0.003439230
## bmi     -0.04252791  0.694367553
## pwgt_r  -0.17044613 -0.706990107
## dwgt_r   0.77456891 -0.035333397
## wtgain  -0.60598255  0.100646515
```

The eigenvectors are both positive and negative in direction and not all of them are near zero. Some of the eigenvectors are around .64 while some are closer to zero at -.001. With this, we can interpret these components as having contrasting behavior.

```
#Loadings Plot
plot(evector, type = "n", xlim = c(-1,1), ylim = c(-1,1), main = "Loadings Plot")
text(jitter(evector, amount = 0.05), labels = rownames(evector))
```



The loading plot shows that variables, pre-pregnancy weight, delivery weight, BMI, and weight gain are clustered together and they are relatively far away from the center which tells us that they are all correlated. The mother's height is slightly above this cluster which shows that it is slightly correlated with these variables more so than age and race which are farther away from the cluster.

Conclusion:

In conclusion, the main part of this analysis was to see if there was any correlation between age, height, race, and weight gained during pregnancy. I was curious to see if these variables affected weight gain in any way since very women's body is different and weight gained during pregnancy is supposedly arbitrary. I also want to see if pre-pregnancy weight and BMI are correlated with weight gain. I used the variables mager(monther's age), m_ht_in(mother's height in inches), mrace31(mothers race),bmi(mother's bmi), pwgt_r(pre-pregnancy weight) and dwgt_r(delivery weight), wtgain(weight gain) and preformed PCA analysis including a correlation matrix to be able to visualize the correlation between the variables. In this report, I found that there isn't a relationship between race, age, and weight gain. The correlation for these variables is extremely low or just zero. However, weight gain is correlated with bmi, pre-pregnancy weight, and delivery weight which is expected. It would be interesting to do further analysis on this subject by looking at the weight gain from a women's previous pregnancy or genetics to see if weight gain during pregnancy is a hereditary thing or if it truly is arbitrary.