# Project 1: PCA on a Birth Weight Dataset

Trevor Carpenter, Christina De Cesaris, Michelle Tran
STA 135: Multivariate Data Analysis
Professor Maxime Pouokam
University of California, Davis
May 4, 2021

# I. Introduction

Although low birth weighted infants can be healthy, under improper care or other circumstances, it can cause serious long-term health problems. A low birth weight is defined as being below 2.5kg. It can be caused by a multitude of factors —poor socioeconomic situation, premature birth, a mother's pre-existing health conditions, and more. In this project, we will highlight how smoking and other factors can cause low birth weights in infants in comparison to non-smoking parents.

We want to find out what variables related to smoking and birth data are related to eachother and how strong those relationship is. In order to answer this, we will be using principal component analysis to analyze the data. PCA is a statistical technique that reduces a dimensionality of a dataset to make it interpretable while also preserving the variability of the data. We can use it to see what variables group together, identify possible outliers, recognize correlations between variables, and more.

# II. Summary of Data

Before conducting PCA onto a dataset, we must first examine the data and decide on what variables we want to specifically look at by looking at different plots and summaries.

| Length | Birthweight | Headcirc | Gestation | smoker | mage | mnocig | mheight | mppwt | fage | fedyrs | fnocig | fheight |
|--------|-------------|----------|-----------|--------|------|--------|---------|-------|------|--------|--------|---------|
| 56 | 4.55 | 34 | 44 | 0 | 20 | 0 | 162 | 57 | 23 | 10 | 35 | 179 |
| 53 | 4.32 | 36 | 40 | 0 | 19 | 0 | 171 | 62 | 19 | 12 | 0 | 183 |
| 58 | 4.10 | 39 | 41 | 0 | 35 | 0 | 172 | 58 | 31 | 16 | 25 | 185 |
| 53 | 4.07 | 38 | 44 | 0 | 20 | 0 | 174 | 68 | 26 | 14 | 25 | 189 |
| 54 | 3.94 | 37 | 42 | 0 | 24 | 0 | 175 | 66 | 30 | 12 | 0 | 184 |
| 51 | 3.93 | 38 | 38 | 0 | 29 | 0 | 165 | 61 | 31 | 16 | 0 | 180 |

The full dataset contains a total of 42 observations of 16 variables. The first six observations are provided for visualization purposes.

| smoker | mean(Birthweight) | median(Birthweight) | sd(Birthweight) | max(Birthweight) | min(Birthweight) |
|--------|-------------------|---------------------|-----------------|------------------|------------------|
| 0 | 3.51 | 3.38 | 0.52 | 4.55 | 2.65 |
| 1 | 3.13 | 3.18 | 0.63 | 4.57 | 1.92 |

This study focus on the dependent variable of birth weight. To obtain a better understanding of the differences in birth weight between smoking and non-smoking mothers, the chart above groups babies by their mother's smoking status and calculates the respective birth weight summary statistics. From these numbers, babies with non-smoking mothers appear to have high mean and median bith weights.

| smoker | Birthweight | Length | Headcirc | Gestation | mage | fage | mnocig | fnocig | fedyrs |
|--------|-------------|--------|----------|-----------|------|------|--------|--------|--------|
| 0 | 3.51 | 51.80 | 35.05 | 39.45 | 24.30 | 27.50 | 0 | 9.7 | 13.70 |
| 1 | 3.13 | 50.91 | 34.18 | 38.95 | 26.68 | 30.18 | 18 | 24.0 | 13.64 |

While birth weights are our main variable of interest, it is important to see if a mother's smoking status affects other characteristics of the baby. Other demographics of the father are also provided. From this chart, babies from non-smoking mothers have slightly higher average birth weight, gestation time, head circumference, and length. It is also important to look at the other demographics in play. For one, notice that the average age of smoking mothers is higher than that of non-smoking mothers for this data set, so while we can say smoking is correlated to low birth weight, it is likely that other factors and intersections of these factors also play a role.

Parallel Coordinate Plot of Select Variables



Parallel Coordinate Plot of Baby's Features
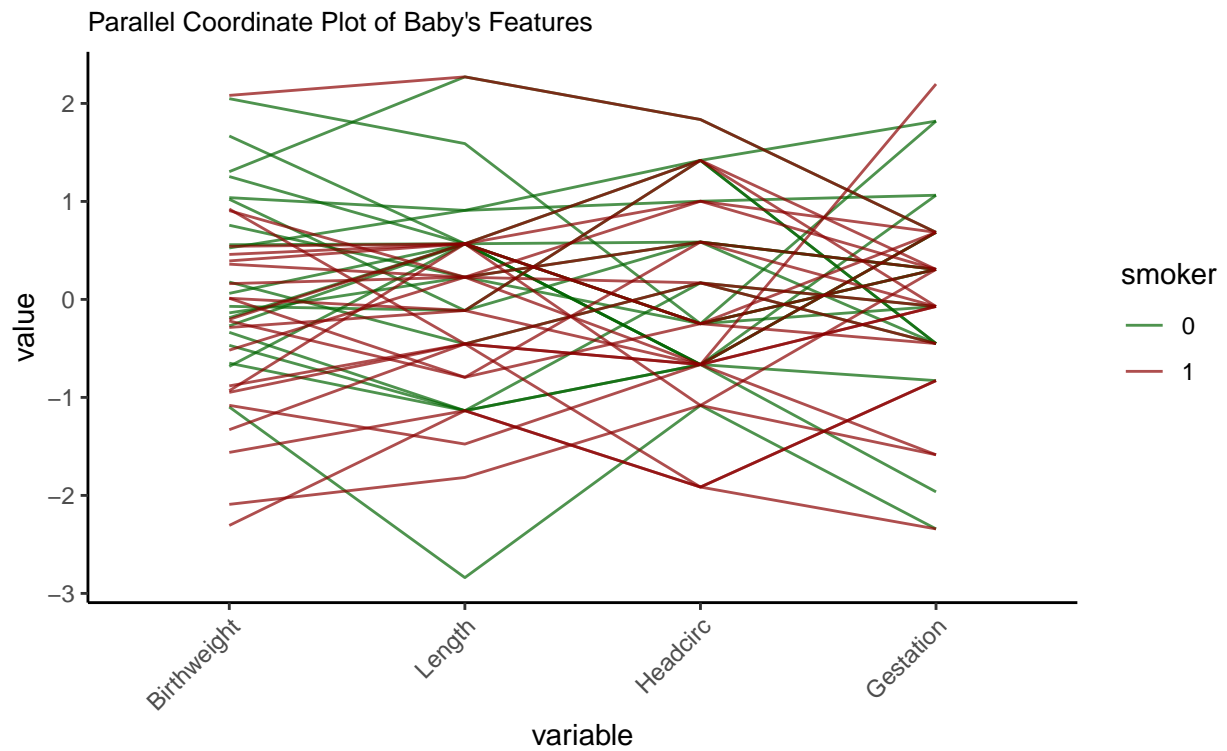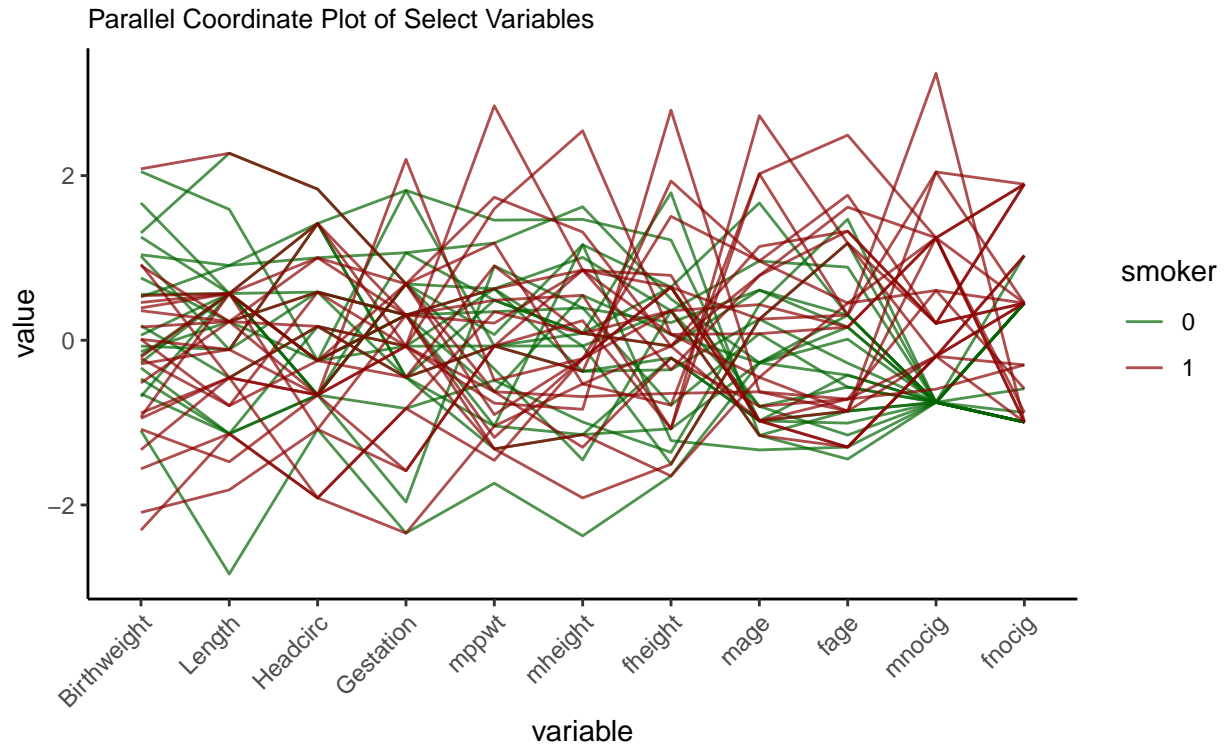
The parallel plots above seeks to demonstrate the relationship between variables. In the case of these plots, smoker = 0 represents a non smoking mother.

The second plot is the first plot but only containing the features specific to the babies, as the second plot is complicated. The values on the y-axis have been normalized, and from the plot alone, it appears that smoking mothers have babies with lower birth weight, length, head circumference, and gestation periods on average. This graph however also reveals there are outliers. It can be seen that there is a baby born to a non-smoking mother which is substantially small in length. The second plot also indicates a relationship between mothers and fathers who smoke. A baby who's mother smokes is more likely to have a father who smokes.



Birth Weights by Mother's Smoking Status
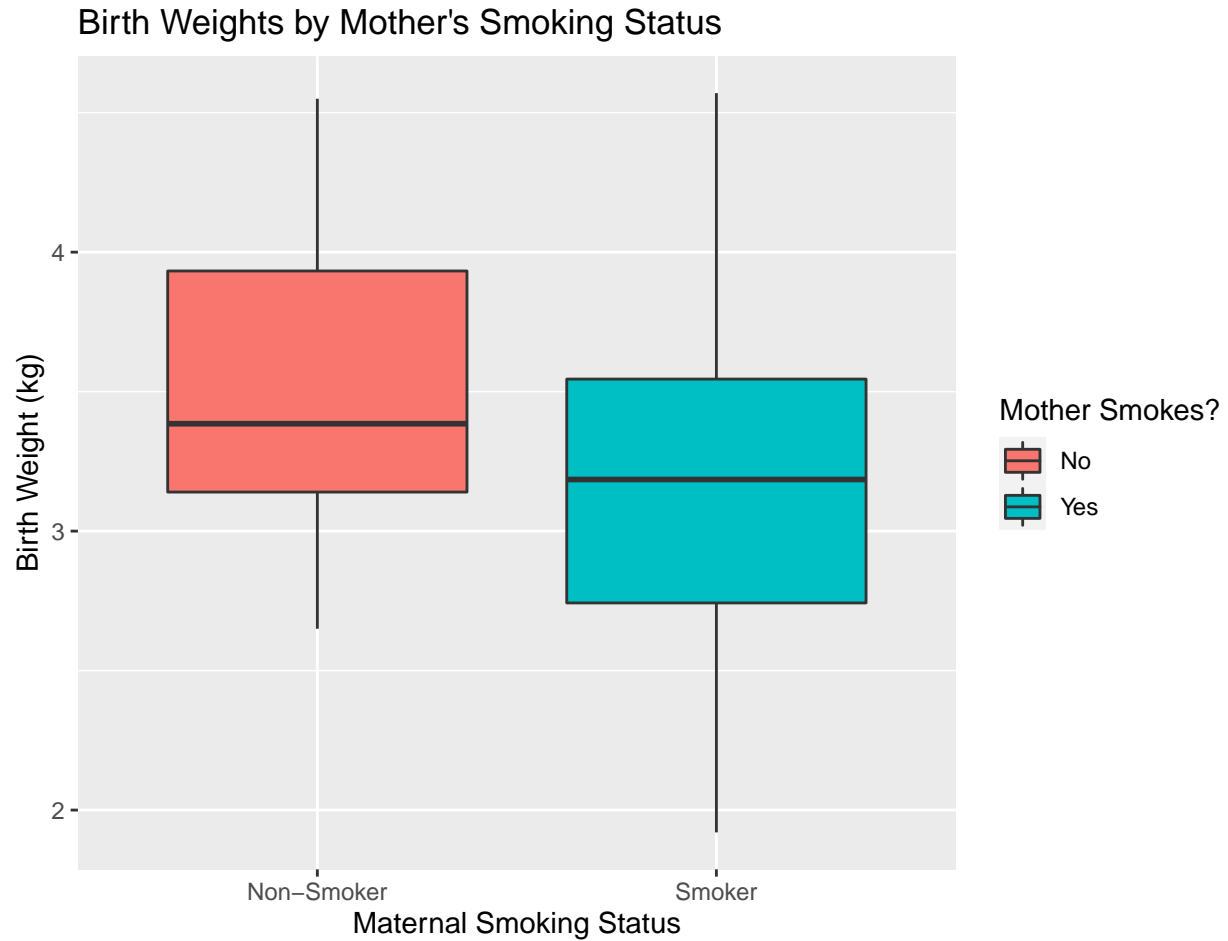
## Birth Weights: Smoking vs Non Smoking Mothers
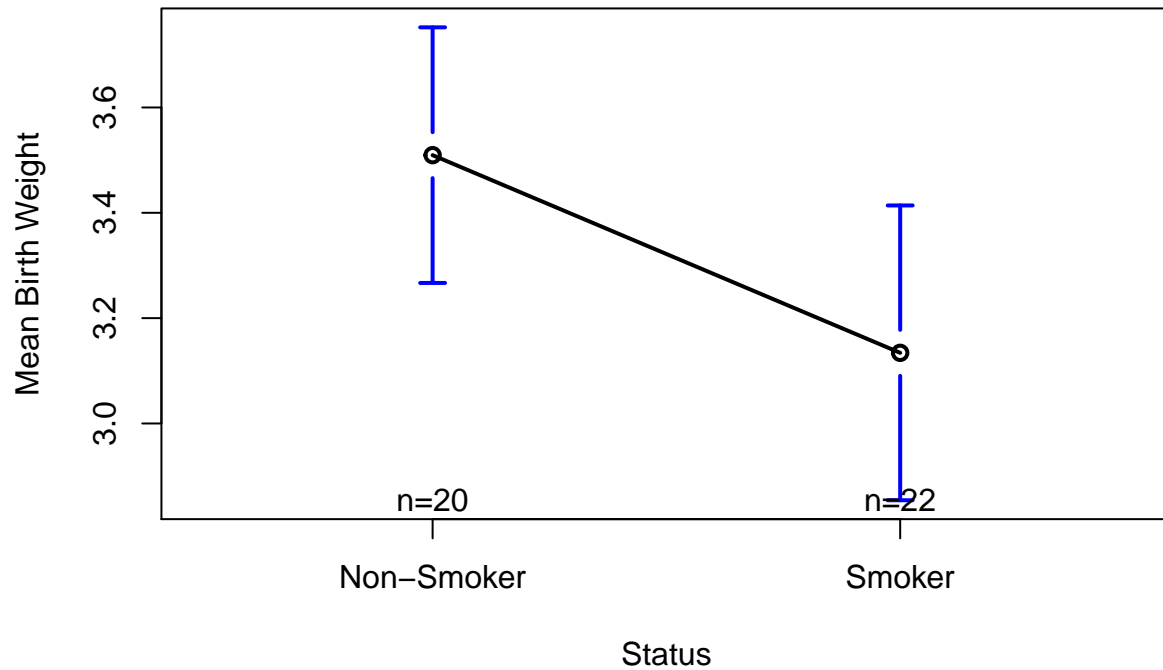


```
## NULL
```

The boxplot and meansplot essentially tell us what has already been established: mother's who smoke have lighter babies on average. The confidence intervals present in both of these graphs raise questions to whether the difference between baby weights in smoking and non-smoking mothers is significant, however, the exploration of this question is beyond the current scope of the class.
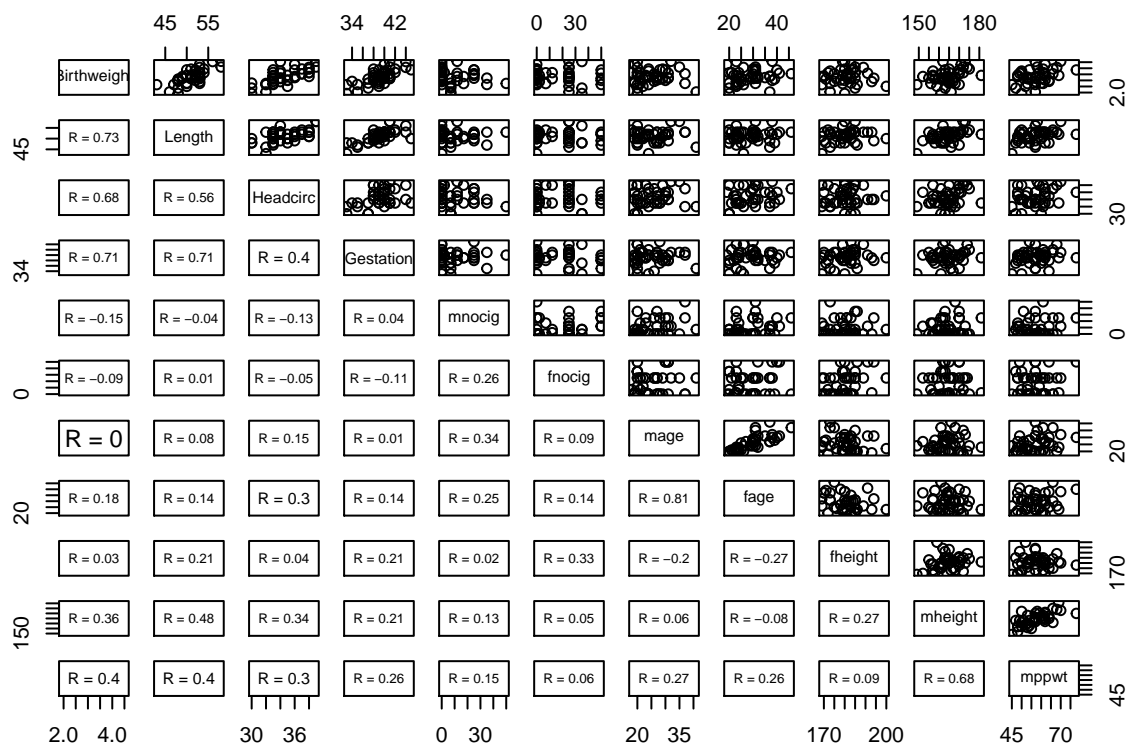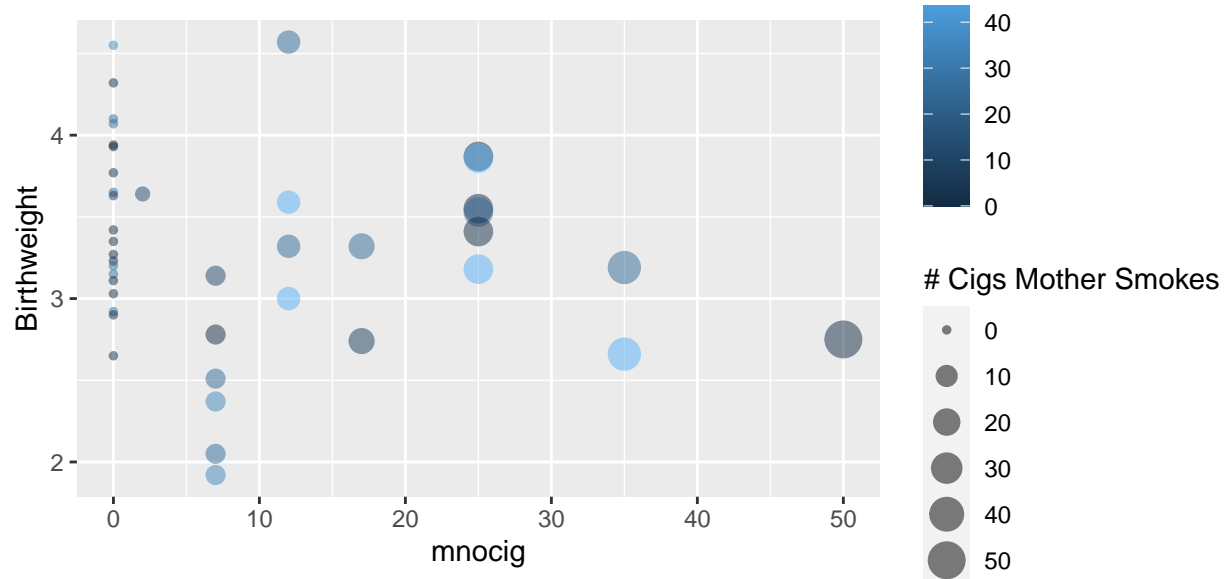
## NULL
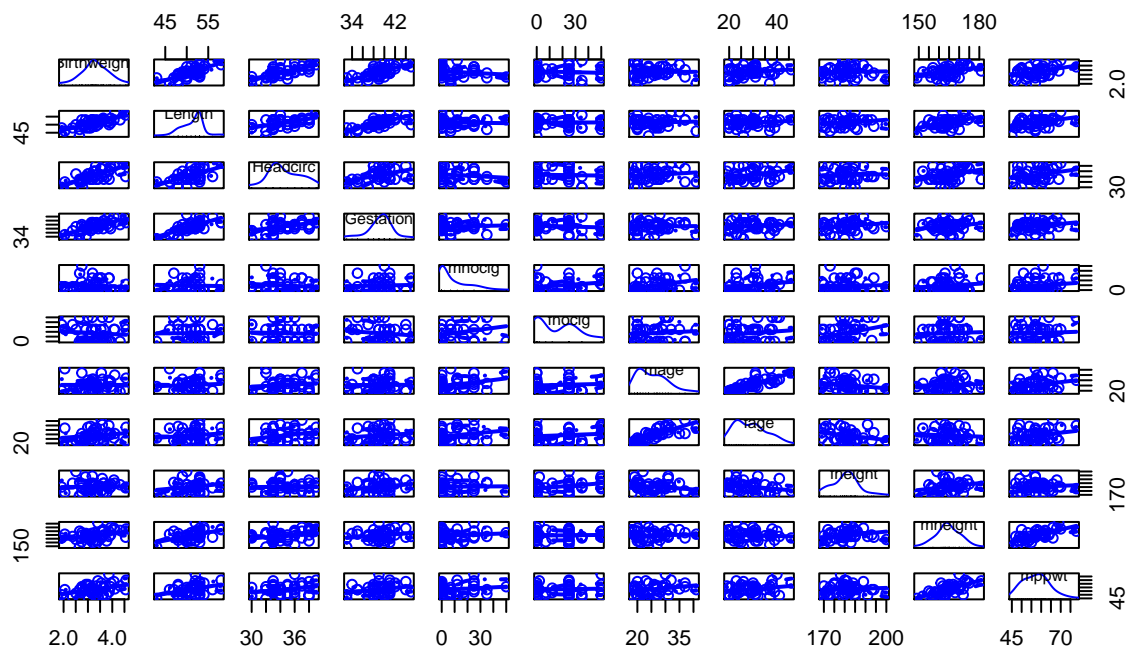
## NULL

The correlation scatter plots show a weak negative correlation between the number of cigarettes a mother smokes and the baby's birth weight. There is a weak correlation between the number of cigarettes the mother smokes and the baby's other features as well. It appears the gestation time and length are most highly correlated with birth weight.



The plot above explores the relationship between parental smoking habits and birth weight. The number of cigarettes smoked by the father do not seem to be correlated with the baby's birth weight directly, however, mothers who's husbands smoke appear to smoke more than those who's husbands do not smoke. Ultimately, the mother's smoking habits appear negatively correlated to baby weight.
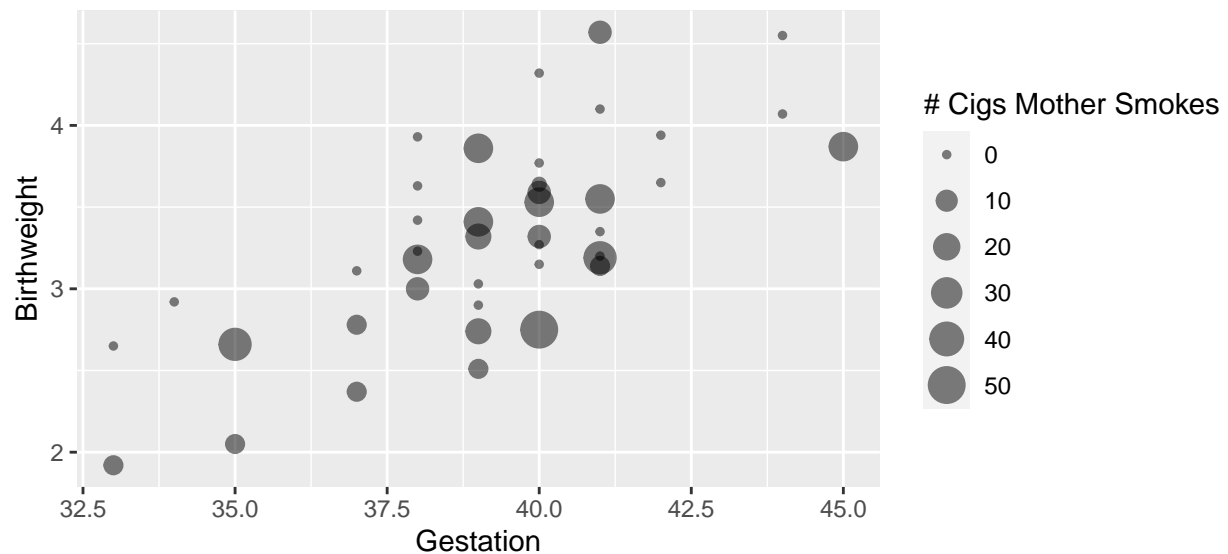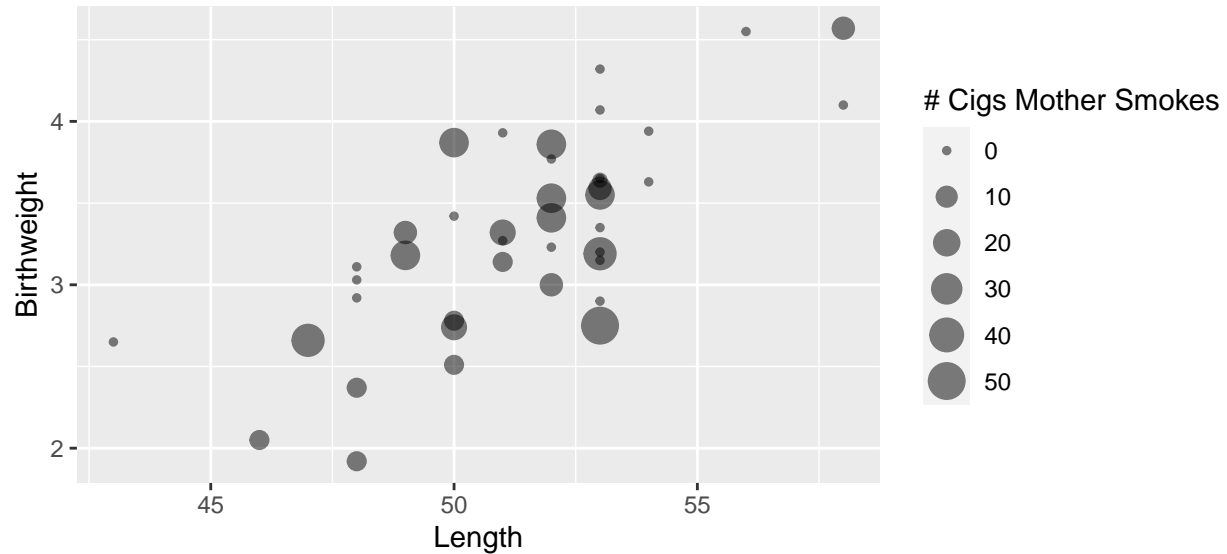
Following up on the high correlation between birth weight and gestation period, it does seem that gestation period has a higher correlation between birth weight than the mother's smoking status. Babies with longer gestation periods tend to weight more. It also appears that non-smoking mothers have higher gestation periods and birth weights than smoking mothers, but this difference is subtle.

When it comes to birth weight and length, a similar relationship is observed. It can be noted that non-smoking mothers have heavier and longer babies than smoking mothers. However, the relationship between length and weight is much more correlated than that of the number of cigarettes the mother smokes and either of the variables.

When conducting the analysis, we also want to know if we want to use the covariance matrix or the correlation matrix. The covariance matrix can be used for PCA when the scales of all varaibles are the same. In the case where variables differ in scale, it is best to use the correlation matrix for consistency. For this data set, the correlation matrix will be used when conducting PCA as units such as number of cigarrets smoked and baby weight in kg are far from on the same scale.

## III. Analysis

If we compared babies according the their mother's smoking status, we can see what factors influence each other in both groups and compare them to see if they are different or similar. If there is a difference, it shows that a mother's smoking status does play a role in influencing their infant's growth. However, if they are similar, then something other than the mother's smoking status is causing a difference in an infant's growth.

When analyzing the data via PCA, we found that both datasets only needed the first 6 PCs to achieve over 90% explanation of the variance. The majority of the variance is explained by the first PC at about 40% as seen from both scree plots below. From there it tapers off and after the 6th principal component, it reaches a plateau and each of the remaining PCs explains less than 3% of the total variance.

## Scree Plot for Non−Smoking Mothers Dataset



## Scree Plot for Smoking Mothers Dataset
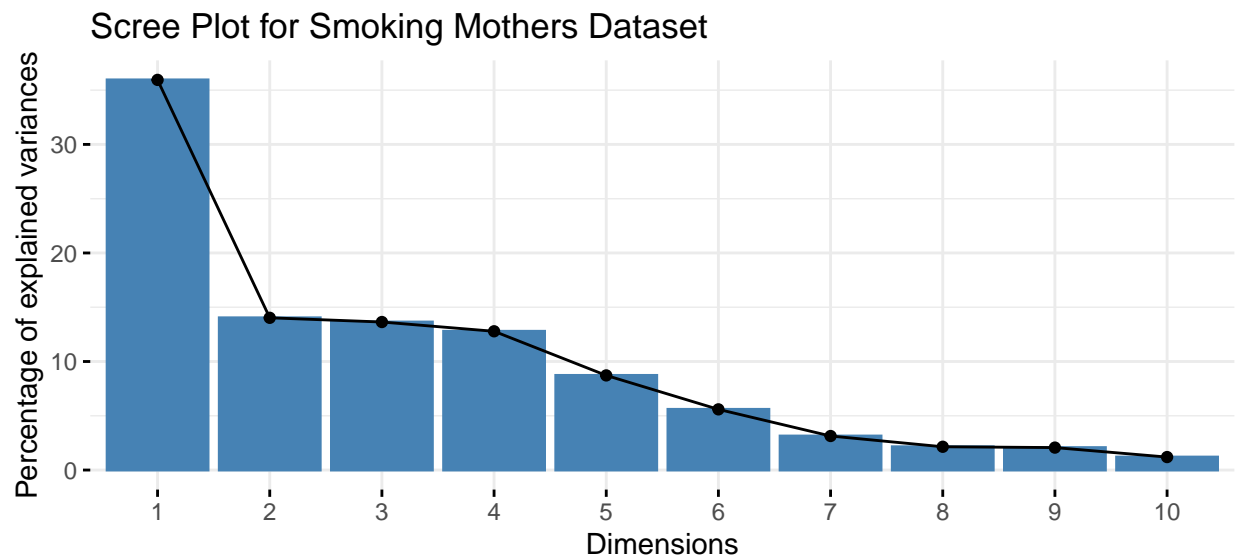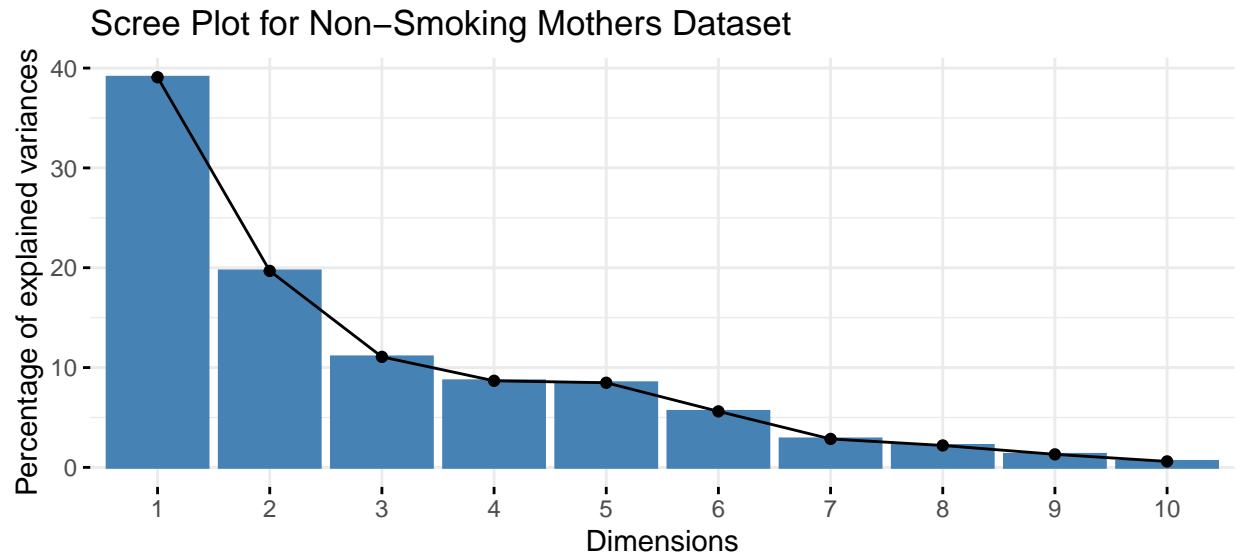


The non-smoking mothers dataset shows that the first PC is based mainly on the positive correlation of body measurements of the baby, mother, and father; specifically, the gestation period, the parents' heights, and the infant's weight and length are the most significant here. This means that genetics and longer gestation period is correlated with a healthier baby. In comparison to the smoking mothers dataset, although they are similar, there is more emphasis on the parent's age than height. Overall, both confirm that gestation plays an important role in a healthier baby.

For the second PC, in the non-smoking dataset, it was positively influenced by the parent's age. This highlights that both datasets agree that the parent's age is important in determining a healthy baby. For the smoking data set, the second PC negatively correlates the mother's height, pregnancy weight, and number of cigarettes together. This correlation is saying that an increase in these values means an unhealthy baby. But in PC 1 for the non-smoking dataset, a mother's height meant it would positively impact the baby's health. This confirms that smoking status does in fact effect a baby's health.

Non-Smoking Mothers Loadings

|             | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 |
|-------------|--------|--------|--------|--------|--------|--------|
| Length      | 0.413  |        |        | 0.219  | 0.295  |        |
| Birthweight | 0.380  |        |        | 0.510  |        |        |
| Headcirc    | 0.297  | 0.310  | -0.479 | 0.110  | -0.134 | -0.296 |
| Gestation   | 0.396  |        | 0.152  | 0.194  | 0.257  | 0.483  |
| mage        |        | 0.589  | 0.296  |        |        | -0.284 |
| mheight     | 0.384  | 0.173  | -0.115 | -0.125 | -0.419 | -0.198 |
| mppwt       | 0.306  | 0.147  | 0.363  | -0.356 | -0.440 | 0.118  |
| fage        | -0.133 | 0.526  | 0.367  | 0.240  |        | 0.317  |
| fedyrs      | -0.122 | 0.432  | -0.370 | -0.242 | 0.481  |        |
| fnocig      | 0.234  | -0.192 | 0.466  | -0.179 | 0.420  | -0.576 |
| fheight     | 0.325  |        | -0.141 | -0.593 | 0.203  | 0.323  |

Smoking Mothers Loadings

|             | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 |
|-------------|--------|--------|--------|--------|--------|--------|
| Length      | 0.384  | 0.135  | 0.271  |        |        |        |
| Birthweight | 0.408  | 0.235  | 0.241  |        |        |        |
| Headcirc    | 0.326  | 0.414  |        | 0.194  | 0.134  | 0.146  |
| Gestation   | 0.336  | 0.238  | 0.171  | -0.379 | -0.280 |        |
| mage        | 0.282  | -0.138 | -0.540 |        |        | -0.214 |
| mnocig      | 0.191  | -0.398 | -0.136 | -0.131 | -0.261 | 0.808  |
| mheight     | 0.179  | -0.478 | 0.341  | 0.302  | 0.225  |        |
| mppwt       | 0.322  | -0.363 | 0.111  | 0.265  | 0.270  | -0.200 |
| fage        | 0.339  | 0.177  | -0.469 | 0.182  |        | -0.100 |
| fedyrs      | 0.284  | -0.340 |        | -0.244 | -0.359 | -0.379 |
| fnocig      |        | 0.118  | -0.190 | 0.678  | -0.265 | 0.190  |
| fheight     | -0.106 |        | 0.365  | 0.285  | -0.707 | -0.203 |

For PCs 3-6, the most notable similarity between the two groups is that they agree that the father's number of cigarettes has a negative influence on the health of the baby. In the smoking dataset, the parent's number of cigarettes predominantly impacts the baby's health negatively. Whereas in the non-smoking dataset, the significant variables seem relatively spread out across the PCs and it becomes harder to interpret them.

Overall, because there is a difference in which variables impact the PCs, it highlights that a mother's smoking status does impact an infant's growth.

It is further notable that smoking is not beneficial for babies based on the PC analysis of the dataset as a whole. When the Principle Components are found on the dataset using the correlation matrix, we find the following:

|                        | Comp.1    | Comp.2    | Comp.3    | Comp.4    | Comp.5    | Comp.6    | Comp.7    |
|------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Standard deviation     | 1.912632  | 1.604867  | 1.3756776 | 1.0653791 | 1.0094043 | 0.9227677 | 0.7300199 |
| Proportion of Variance | 0.281397  | 0.198123  | 0.1455761 | 0.0873102 | 0.0783767 | 0.0655000 | 0.0409945 |
| Cumulative Proportion  | 0.281397  | 0.479520  | 0.6250961 | 0.7124063 | 0.7907830 | 0.8562830 | 0.8972775 |

|  | Comp.8 | Comp.9 | Comp.10 | Comp.11 | Comp.12 | Comp.13 |
|---|---|---|---|---|---|---|
| Standard deviation | 0.6070519 | 0.5729325 | 0.4785064 | 0.4067690 | 0.3659491 | 0.3320844 |
| Proportion of Variance | 0.0283471 | 0.0252501 | 0.0176129 | 0.0127278 | 0.0103014 | 0.0084831 |
| Cumulative Proportion | 0.9256246 | 0.9508748 | 0.9684877 | 0.9812155 | 0.9915169 | 1.0000000 |

The number of PCs needed using the correlation matrix is only the first 8. Using only these 8 we are able to explain 92.56% of the variance of the data.

This is further visualized using the following scree plot:
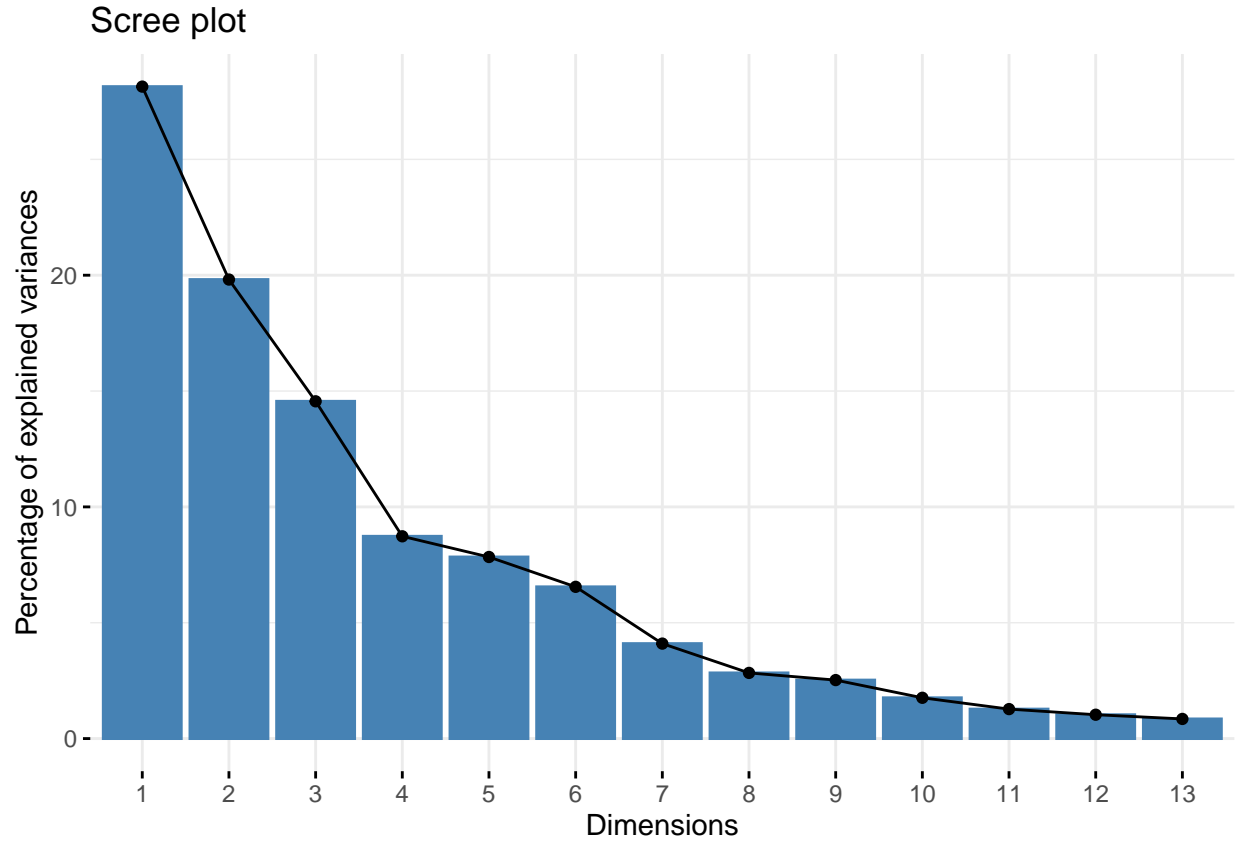


In this plot we are able to see that the elbow curve at the 8th dimension reflects the calculations results showing a need of only 8 dimensions to preserve at least 90% of the variance of the data.

Based on the loadings with a cutoff value of 0.15:

|  | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 |
|---|---|---|---|---|---|---|---|
| Length | 0.444 | | | | | | |
| Birthweight | 0.447 | 0.166 | | 0.212 | | | |
| Headcirc | 0.389 | | | 0.203 | | | -0.767 |
| Gestation | 0.382 | | | 0.272 | 0.432 | 0.163 | 0.387 |
| smoker | | -0.445 | 0.329 | | 0.157 | 0.280 | -0.232 |
| mage | | -0.476 | -0.289 | | | -0.230 | 0.166 |
| mnocig | | -0.474 | 0.224 | | 0.279 | 0.405 | |
| mheight | 0.317 | | 0.304 | -0.493 | -0.254 | | |
| mppwt | 0.340 | | | -0.442 | -0.327 | | 0.280 |
| fage | 0.183 | -0.423 | -0.321 | 0.261 | -0.188 | -0.153 | |
| fedyrs | | -0.222 | -0.293 | -0.381 | 0.512 | -0.340 | -0.175 |
| fnocig | | -0.231 | 0.439 | 0.375 | -0.319 | -0.347 | |
| fheight | | | 0.497 | | 0.306 | -0.610 | |

|  | Comp.8 | Comp.9 | Comp.10 | Comp.11 | Comp.12 | Comp.13 |
|---|---|---|---|---|---|---|
| Length | 0.224 | 0.543 | 0.396 | 0.458 | | |
| Birthweight | -0.335 | | -0.236 | -0.276 | 0.660 | -0.155 |
| Headcirc | | -0.168 | | | -0.334 | |
| Gestation | | -0.196 | | -0.289 | -0.530 | |
| smoker | | -0.202 | 0.564 | -0.287 | 0.177 | |
| mage | 0.362 | 0.178 | -0.201 | | | -0.592 |
| mnocig | | | -0.544 | 0.382 | | |
| mheight | 0.254 | 0.264 | | -0.465 | | 0.319 |
| mppwt | -0.301 | -0.439 | 0.203 | 0.315 | | -0.190 |
| fage | | -0.246 | | | | 0.660 |
| fedyrs | -0.429 | 0.202 | 0.207 | | | |
| fnocig | -0.485 | 0.303 | | | -0.206 | |
| fheight | 0.288 | -0.311 | -0.165 | | 0.217 | |

we can see that when the physical features of the baby are all positive, as seen in the first four features of Comp. 1,

| Length | Birthweight | Headcirc | Gestation |
|---|---|---|---|
| 0.4442198 | 0.4465604 | 0.3891197 | 0.3822305 |

that means that these features all correlate with eachother, meaning that a long gestation period is well correlated with a larger and healthier baby.

That said, when birthweight is positive and another score in the component is negative, as seen with smoker in Comp. 2:

| Birthweight | smoker |
|---|---|
| 0.1655971 | -0.4446873 |

that is a sign that the attribute of being a smoker is not beneficial for the baby, since it will impact the

physical features.

As seen based on the PC analysis, these correlations become weaker as we iterate through the components This means that the later components contradicting that birth weight and smoking are actually the same sign score, such as in Comp.11:

| Birthweight | smoker |
|---|---|
| -0.2764383 | -0.2871868 |

are negligible, since the later components explain very little proportion of the overall data variance. Comp. 11 itself only explains about 1.2% of the variance, so findings such as this are far less relevant as in Comp. 2 which explains almost 20% of the variance.

We can also see that the number of cigarettes smoked per day by mothers and fathers are both, as expected, well correlated with the smoking status of the mother. This is also visible with the mother and father's ages, as seen by the data in Comp.2:

| smoker | mage | mnocig | fage | fnocig |
|---|---|---|---|---|
| -0.4446873 | -0.4756988 | -0.4735367 | -0.4231726 | -0.2310304 |

Based on these Principle Components of the correlation matrix, it is clear that a larger gestational period is a beneficial thing for a baby's health, however a greater count of smoked cigarettes is not. It is also clear that the father's smoking status, as shown by the number of cigarettes he smokes per day, is well correlated with the mothers.

What is further interesting about this analysis is the lack of correlation of the physical attributes of the parents with the physical attributes of their newborn in comparison to that of the smoking features. While in Comp.1 we see that the physical features of the child and the physical features of the mother are similar scores:

| Birthweight | Headcirc | Gestation | mheight | mppwt |
|---|---|---|---|---|
| 0.4465604 | 0.3891197 | 0.3822305 | 0.3169582 | 0.3404202 |

we see only three principle components later in Comp.4 a contradiction:

| Birthweight | Headcirc | Gestation | mheight | mppwt |
|---|---|---|---|---|
| 0.2124146 | 0.2030967 | 0.2719839 | -0.4928765 | -0.4417855 |

This contradiction may be in a later component, however it is still a component that explains 8.7% of the overall variance of the data, and is relatively significant. It shows a reduced importance of the mother's physical features in explaining the physical features of her child, and a comparative increase in the impact of smoking.

# IV. Interpretation

Based on the analysis we can understand an estimation of what the physical features of a new baby not considered in this analysis will be using only a few features of the parents. For example, if the father is older he is more likely to smoke multiple cigarettes per day. In this case, based on the Principle Component Analysis we could expect that the mother also smokes multiple per day, and as such is marked as a smoker in the data. If this is in fact true we can predict that the gestation period will be shorter and the head circumference, birth weight, and length of this newborn baby will all be lower than expected.

We would also see that the physical features of the father and mother such as their height and weight do not factor into the features of their child nearly as much as their actions do when it comes to smoking or not. Based on only these Principle Components we are able to understand how the smoking features of the father and mother have a direct impact on the features of their baby, and subsequently the healthiness of their baby.

Let's also take for an example with the baby features as follows:

| length | birthweight | headcirc | gestation | smoker | motherage | mnocig | mheight | mppwt | fage |
|--------|-------------|----------|-----------|--------|-----------|--------|---------|-------|------|
| 61 | 5.1 | 36 | 43 | 0 | 43 | 7 | 165 | 64 | 38 |

| fedyrs | fnocig | fheight |
|--------|--------|---------|
| 19 | 45 | 189 |

From the PC analysis, we can understand how well this new data point matches with the data we have previously seen. The number of cigarettes smoked by the mother is 7, yet she is marked as a nonsmoker, which is unlike any data we have previously seen and doesn't really make sense. 7 is around the average amount for the mothers in our dataset to smoke, however the father is on the upper end of the dataset. Other physical characteristics of the parents are average based on what we have previously seen. Based on all of this, we would expect the baby to physically be smaller than the average child. However, the baby is in the 75th percentile for all of its physical characteristics (gestational period, length, birthweight, and head circumference), which is highly unusual especially giving the parent data. Because of this, we would conclude that this child is an outlier and does not match with the rest of the data in this Principle Component Analysis study.

# V. Conclusion

Generally, this analysis shows that while physical features of the parents do explain the physical features of their children, features regarding their smoking habits seems to be a stronger influence. We also see that the smoking actions of the father may correlate with the smoking actions of the mother, meaning that it is likely if one smokes so does the other. This is interesting because it disproves any assumption that the actions of the father do not influence the characteristics of his child.

In future analysis of this topic, it would be interesting to include alcohol consumption as another feature, as it has also been shown to have impacts on child gestation and birth features. We could also further analyze any correlation between the mother and father's non-physical features with each other, such as seeing if the more positive actions of the father in activities such as exercise or sleep influence the actions of the mother in similar ways that smoking does. Father's education level is also a feature in this dataset that does not correlate with other features and wasn't heavily analyzed, however maybe the addition of Mother's education level would show some deeper results.

## Appendix: R Script

```r
knitr::opts_chunk$set(echo = F, warning = F, message = F)
library(factoextra)
library(knitr)

library(dplyr)
library(MASS)


#graphics
library(ggplot2)
library(gplots)
library(DT)
library(GGally)
library(viridis)
library(hrbrthemes)
library(car)

#data
smok = read.csv("data/Birthweight_reduced_kg_R.csv")
smok = smok[,2:14]

new_print_loadings = function (x, digits = 3L, cutoff = 0.1, sort = FALSE, ...)
{
  ## code from the original print.loadings function, without printing the variances
    Lambda <- unclass(x)
    p <- nrow(Lambda)
    factors <- ncol(Lambda)
    if (sort) {
        mx <- max.col(abs(Lambda))
        ind <- cbind(1L:p, mx)
        mx[abs(Lambda[ind]) < 0.5] <- factors + 1
        Lambda <- Lambda[order(mx, 1L:p), ]
    }
    fx <- setNames(format(round(Lambda, digits)), NULL)
    nc <- nchar(fx[1L], type = "c")
    fx[abs(Lambda) < cutoff] <- strrep(" ", nc)
    return(fx)
}

new_print_sum = function (x, digits = 3L, loadings = x$print.loadings, cutoff = x$cutoff,
    ...)
{
    vars <- x$sdev^2
    vars <- vars/sum(vars)
    return(rbind(`Standard deviation` = x$sdev, `Proportion of Variance` = vars,
        `Cumulative Proportion` = cumsum(vars)))
}

#full data, provided for exploritory purposes

full = smok %>% dplyr::select(smoker,
         Birthweight,
```

```r
          Length,
          Headcirc,
          Gestation,
          mppwt,
          mheight,
          fheight,
          mage,
          fage,
          mnocig,
          fnocig,
          fedyrs)


# Mean Other baby and family demos

demo_stat = smok %>%
  group_by(smoker) %>%
  dplyr::select(Birthweight,
        Length,
        Headcirc,
        Gestation,
        mage,
        fage,
        mnocig,
        fnocig,
        fedyrs) %>%
  summarize_all(mean)%>%
  mutate_if(is.numeric,
          funs(round(.,2)))

#Birthweight only summary stats
bw_stat = smok %>%
  group_by(smoker) %>%
  dplyr::select(Birthweight) %>%
  summarise(mean(Birthweight),
          median(Birthweight),
          sd(Birthweight),
          max(Birthweight),
          min(Birthweight))%>%
  mutate_if(is.numeric,funs(round(.,2)))



# Full Data
full_dt=datatable(smok,
          class = 'cell-border stripe',
          caption = htmltools::tags$caption(style = 'caption-side: bottom; text-align: center;','Table
    htmltools::em('Full Dataset')))
```

```r
#demographics
demo_names=c("Mother Smokes?","Birth Wt kg","Length cm",'Head Circ (cm)',"Gestation wks","Mother's Age"

demo_dt=datatable(demo_stat,
          class = 'cell-border stripe',
          caption = htmltools::tags$caption(style = 'caption-side: bottom; text-align: center;',
   htmltools::em('Smoking Vs Non-Smoking Averages')),
   colnames = c(demo_names),
   rownames = c("No","Yes"),
   options = list(sDom  = 't'))



# Birthweight Stats
bw_names=c("Mother Smokes?","Mean","Median","Standard Deviation","Max","Min")

#object displays table when run
bw_dt=datatable(bw_stat,
          class = 'cell-border stripe',
          caption = htmltools::tags$caption(style = 'caption-side: bottom; text-align: center;',
   htmltools::em('Birth Weight Summary Statistics')),
   colnames = c(bw_names),
   rownames = c("No","Yes"),
   options = list(sDom  = 't'))

kable(head(smok))


kable(bw_stat)
kable(demo_stat)
full$smoker=as.factor(full$smoker)
parallelplot1=ggparcoord(full, columns=2:12,
          groupColumn = 1,
          order = 2:13,
          showPoints = F,
          title = "Parallel Coordinate Plot of Select Variables",
          alphaLines = .7) +
  theme_classic()+
  theme(plot.title = element_text(size=10),
     axis.text.x = element_text( angle = 45, hjust = 1))+ scale_color_manual(values=c( "darkgreen", "da
   scale_fill_discrete(name = "Mother Smokes?",labels = c("No","Yes"))

parallelplot1

full$smoker=as.factor(full$smoker)
parallelplot2=ggparcoord(full, columns=2:5,
          groupColumn = 1,
          order = 2:13,
          showPoints = F,
          title = "Parallel Coordinate Plot of Baby's Features",
          alphaLines = .7) +
  theme_classic()+
  theme(plot.title = element_text(size=10),
```

```r
        axis.text.x = element_text( angle = 45, hjust = 1))+  scale_color_manual(values=c( "darkgreen", "d
    scale_fill_discrete(name = "Mother Smokes?",labels = c("No","Yes"))

parallelplot2

boxplot=ggplot(smok,
        aes(factor(smoker),
            Birthweight,
             fill= factor(smoker)))+
  geom_boxplot()+
  labs(title="Birth Weights by Mother's Smoking Status",
       x="Maternal Smoking Status",
       y = "Birth Weight (kg)")+
  scale_x_discrete(labels=c("Non-Smoker",
                             "Smoker"))+
  scale_fill_discrete(name = "Mother Smokes?",
                       labels = c("No",
                                   "Yes"))

boxplot
#means plot with confidence intervals
mean_plot=plotmeans(Birthweight~smoker,
         data=smok,
         xlab="Status",
         ylab="Mean Birth Weight",
         main="Birth Weights: Smoking vs Non Smoking Mothers",
         legends=c("Non-Smoker","Smoker"),
         lwd=2,
         barwidth=2)
mean_plot
panel_cor <- function(x, y){ #used to generate lower panel of pearson correlation coefficents for pairs
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y, use="complete.obs"), digits=2)
  txt <- paste0("R = ", r)
  cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor )
}

pairs_graph=pairs(Birthweight~Length+Headcirc+Gestation+mnocig+fnocig+mage+fage+fheight+mheight+mppwt,da
      lower.panel = panel_cor)

pairs_graph
corroMat = scatterplotMatrix(~Birthweight+Length+Headcirc+Gestation+mnocig+fnocig+mage+fage+fheight+mhe
diagonal = "histogram")

corroMat

mcig_bw=ggplot(smok, aes(x=mnocig, y=Birthweight, size = mnocig, color =fnocig)) +
    geom_point(alpha=0.5)+
    scale_size(name="# Cigs Mother Smokes")+
    labs(color='# Cigs Father Smokes')
```

```r
gest_birthwt=ggplot(smok, aes(x=Gestation, y=Birthweight, size = mnocig)) +
    geom_point(alpha=0.5)+
    scale_size(name="# Cigs Mother Smokes")


len_bw=ggplot(smok,
       aes(x=Length,
           y=Birthweight,
           size = mnocig)) +
    geom_point(alpha=0.5)+
    scale_size(name="# Cigs Mother Smokes")
mcig_bw
gest_birthwt

len_bw
######[If the overall goal is to compare babies according with their
###### mother's smoking status, how would you proceed ?]

#split the data
nonsmoke = subset(smok, smoker == 0, select = -c(5, 7))
smoke = subset(smok, smoker == 1, select = -c(5))

#pca and summary
##nonsmoking mothers
pca.nonsmoke =  princomp(nonsmoke, cor = TRUE)
t = summary(pca.nonsmoke, loadings = TRUE, cutoff = 0.1)$loadings


##smoking mothers
pca.smoke = princomp(smoke, cor = TRUE)
u = summary(pca.smoke, loadings = TRUE, cutoff = 0.1)$loadings

#scree plots for nonsmoking vs. smoking mothers datasets
library(factoextra)

fviz_eig(pca.nonsmoke, main = "Scree Plot for Non-Smoking Mothers Dataset")
fviz_eig(pca.smoke, main = "Scree Plot for Smoking Mothers Dataset")

kable(new_print_loadings(t, cutoff = 0.1)[,1:6])
kable(new_print_loadings(u, cutoff = 0.1)[,1:6])
## PCA Analysis on whole data
c = princomp(smok, cor = TRUE)
l = summary(c, loadings = TRUE, cutoff = 0.15)$loadings
kable(new_print_sum(c, cutoff = 0.15)[,1:7])
kable(new_print_sum(c, cutoff = 0.15)[,8:13])
fviz_eig(c, ncp = 13)
kable(new_print_loadings(l, cutoff = 0.15)[,1:7])
kable(new_print_loadings(l, cutoff = 0.15)[,8:13])
vars = l[,1][1:4]
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))

vars = c(l[,2]['Birthweight'], l[, 2]['smoker'])
```

```r
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))

vars = c(l[,11]['Birthweight'], l[, 11]['smoker'])
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))

vars = c(l[,2][5:7], l[, 2][10], l[, 2][12])
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))
vars = c(l[,1][2:4], l[,1][8:9])
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))
vars = c(l[,4][2:4], l[,4][8:9])
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))

kable(matrix(c(61, 5.1, 36, 43, 0, 43, 7, 165, 64, 38), ncol = 10),  col.names = c("length", "birthweigl
kable(matrix(c( 19, 45, 189), ncol = 3), col.names = c("fedyrs", "fnocig", "fheight"))
```