Homework 5

Akshay Kumar Varanasi (av32826)

This homework is due on Mar. 5, 2019 at 4:00pm. Please submit as a PDF file on Canvas.

For this homework, you will work with a dataset collected by John Holcomb from the North Carolina State Center for Health and Environmental Statistics. This data set contains 1409 birth records from North Carolina in 2001.

```
NCbirths <- read_csv("http://wilkelab.org/classes/SDS348/data_sets/NCbirths.csv
")
```

```
## Parsed with column specification:
## cols(
##   Plural = col_integer(),
##   Sex = col_integer(),
##   MomAge = col_integer(),
##   Weeks = col_integer(),
##   Gained = col_integer(),
##   Smoke = col_integer(),
##   BirthWeightGm = col_double(),
##   Low = col_integer(),
##   Premie = col_integer(),
##   Marital = col_integer()
## )
```

```
head(NCbirths)
```

```
## # A tibble: 6 x 10
##   Plural   Sex MomAge Weeks Gained Smoke BirthWeightGm   Low Premie Marital
##    <int> <int>  <int> <int>  <int> <int>         <dbl> <int>  <int>   <int>
## 1      1     1     32    40     38     0         3147.     0      0       0
## 2      1     2     32    37     34     0         3289.     0      0       0
## 3      1     1     27    39     12     0         3912.     0      0       0
## 4      1     1     27    39     15     0         3856.     0      0       0
## 5      1     1     25    39     32     0         3430.     0      0       0
## 6      1     1     28    43     32     0         3317.     0      0       0
```

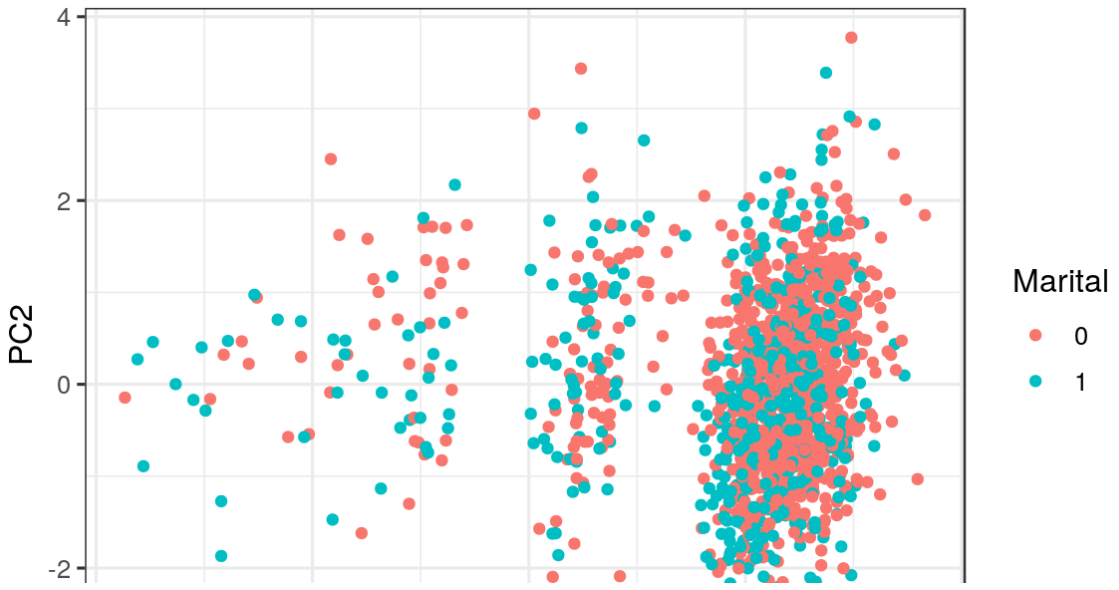
The column contents are as follows:

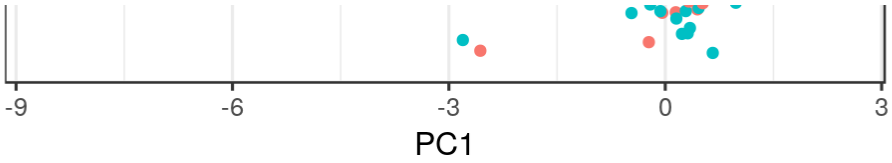
- **Plural**: 1=single birth, 2=twins, 3=triplets.
- **Sex**: Sex of the baby 1=male 2=female.
- **MomAge**: Mother's age (in years).

- **Weeks**: Completed weeks of gestation.
- **Gained**: Weight gained during pregnancy (in pounds).
- **Smoke**: Mother is a smoker: 1=yes, 0=no.
- **BirthWeightGm**: Birth weight in grams.
- **Low**: Indicator for low birth weight, 1=2500 grams or less, 0=otherwise.
- **Premie**: Indicator for premature birth, 1=36 weeks or sooner, 0=otherwise.
- **Marital**: Marital status: 0=married or 1=not married.

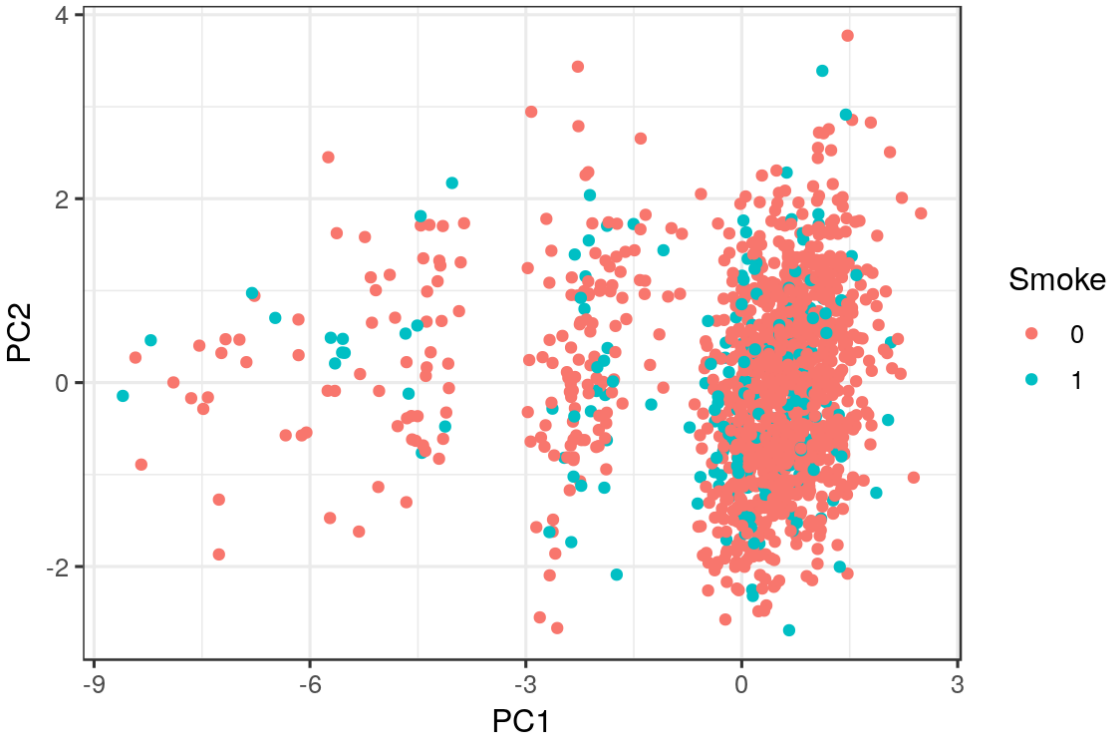
Problem 1 (3 pts): We are interested in assessing the relationships between the variables in the dataset `NCbirths` and the mothers' marital status, the mothers' smoking habits, and plural births. Perform a principal components analysis (PCA) on the dataset `NCbirths`. Remove the columns `Marital`, `Smoke`, and `Plural` prior to performing PCA. Create a scatterplot of PC1 vs. PC2. First, color each point bythe mother's marital status, then color each point by the mother's smoking habit, and then color each point by the indicator of plural births. What do you observe? Visually, and without doing any calculations, do the different types of births cluster together in principal-component space? Do the smokers or married mothers cluster together?

```
NCbirths %>%
  select(-Marital, -Smoke, -Plural) %>%
  scale() %>%
  prcomp() ->
pca

NCbirths_pca <- data.frame(NCbirths, pca$x)
NCbirths_pca$Plural <-factor(NCbirths_pca$Plural)
NCbirths_pca$Marital <-factor(NCbirths_pca$Marital)
NCbirths_pca$Smoke <-factor(NCbirths_pca$Smoke)

ggplot(NCbirths_pca, aes(x = PC1, y = PC2, color = Marital)) + geom_point()
```

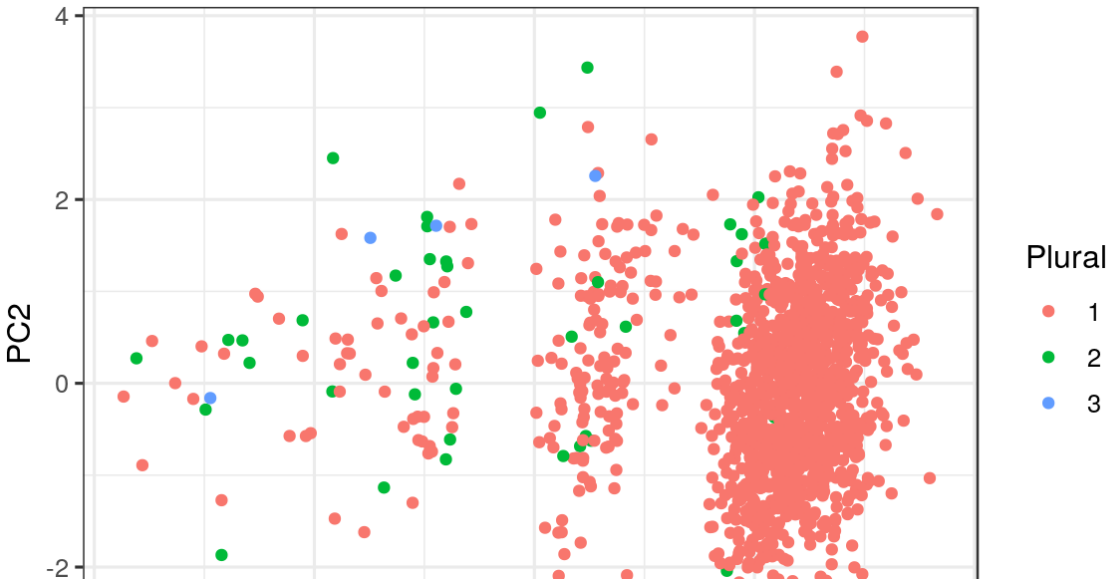


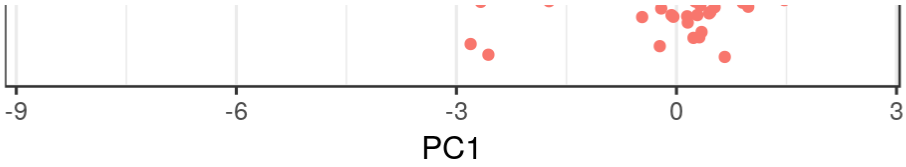



```
ggplot(NCbirths_pca, aes(x = PC1, y = PC2, color = Smoke)) + geom_point()
```




```
ggplot(NCbirths_pca, aes(x = PC1, y = PC2, color = Plural)) + geom_point()
```

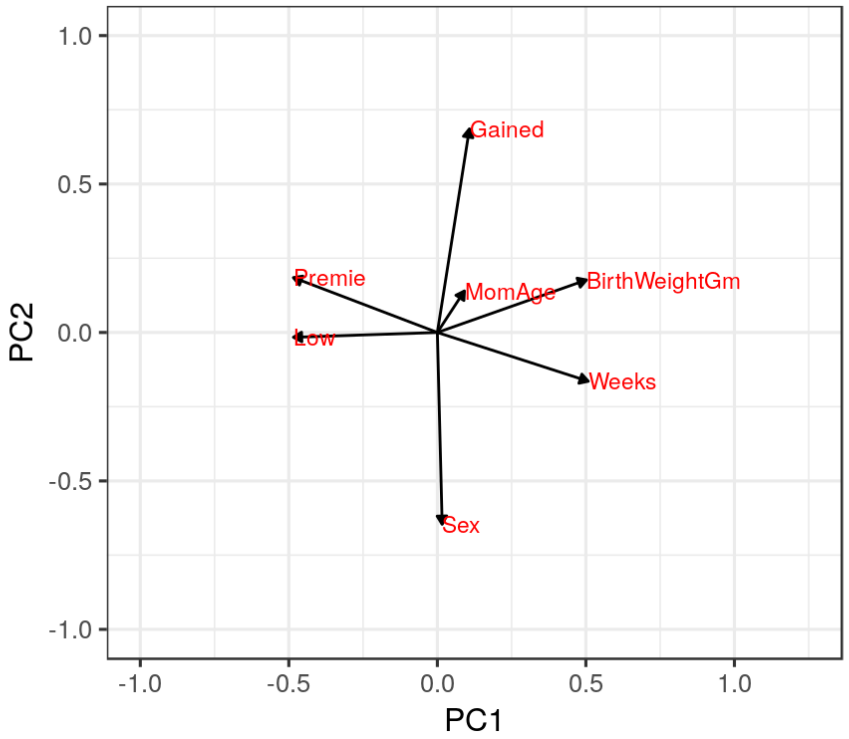




Even in PC space, there are hardly any clusters or trends seen. Looking at plot with color based on Plural, we see that plural equal to one is clustered. It is difficult to conclude anything clearly whether we see that smokers or married mothers cluster together or not. It looks like they don't.

Problem 2 (4 pts): Now visualize the rotation matrix of the PCA obtained under Problem 1.

```
# capture the rotation matrix in a data frame
rotation_data <- data.frame(pca$rotation, variable = row.names(pca$rotation))
# define a pleasing arrow style
arrow_style <- arrow(
length = unit(0.05, "inches"),
type = "closed"
)
# now plot, using geom_segment() for arrows and geom_text for labels
ggplot(rotation_data) +
geom_segment(aes(xend = PC1, yend = PC2), x = 0, y = 0, arrow = arrow_style)+
geom_text(aes(x = PC1, y = PC2, label = variable), hjust = 0, size = 3, color
= "red") +
xlim(-1., 1.25) +
ylim(-1., 1.) +
coord_fixed() # fix aspect ratio to 1:1
```



Given the plots from Problem 1 and the arrow plot you made, how do you interpret PC1 and PC2? What does PC1 tell you about a data point? What does PC2 tell you about a data point?

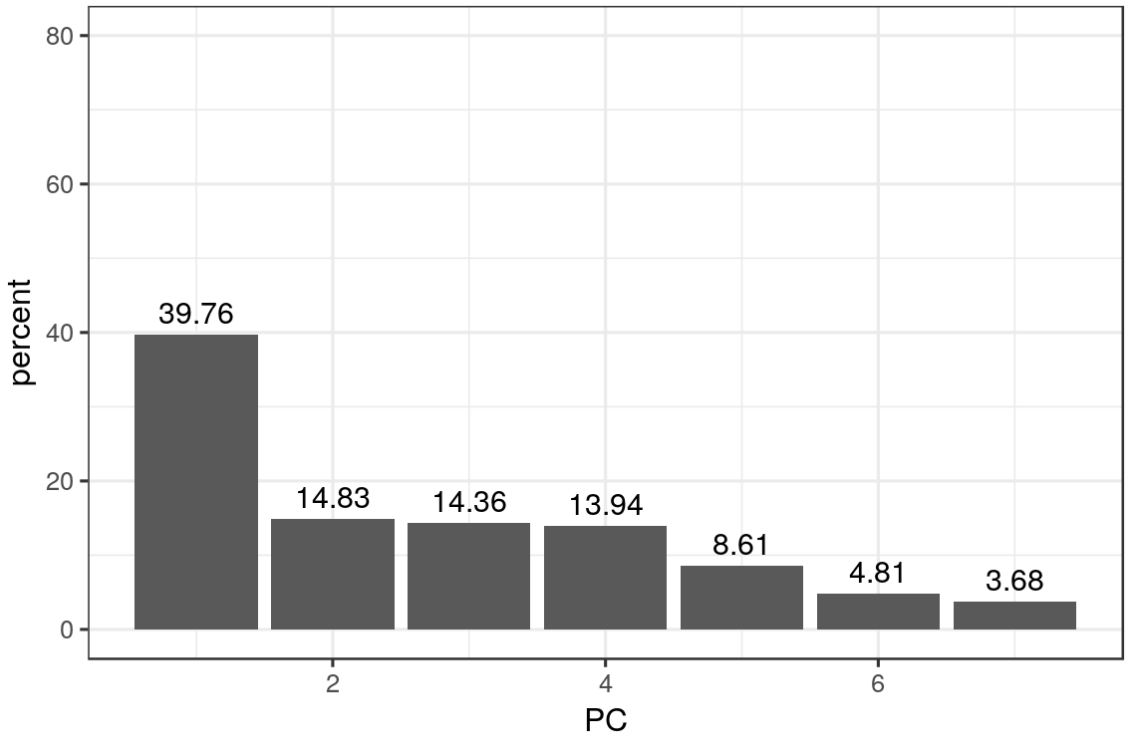
PC1 depends on Premie, Birth weight, Low and weeks while PC2 depends on Gained, sex and Mom age primarily. So PC1 tells mainly about the baby conditions while PC2 about Mother's condition.

Problem 3 (3 pts): Create a bar plot that shows the percent variance explained by each principal component. State how much variance is explained by each of the principal components 1 through 4.

```
NC_percent <- 100*pca$sdev^2 / sum(pca$sdev^2)
NC_percent
```

```
## [1] 39.756987 14.832899 14.362641 13.943168  8.609349  4.813317  3.681639
```

```
perc_data <- data.frame(percent = NC_percent, PC = 1:length(NC_percent))
ggplot(perc_data, aes(x = PC, y = percent)) +
  geom_col() +
  geom_text(aes(label = round(percent, 2)), size = 4, vjust = -0.5) +
  ylim(0, 80)
```



Variance explained by PC 1 is 39.76 % while PC2, PC3 and PC4 explain almost equal variance of around 14 %.