Homework 2. PCA. (60 Points)

Nikhil Ambati

2023-10-01

Part 1. PCA vs Linear Regression (6 points).

Let's say we have two 'features': let one be x and another y. Recall that in linear regression, we are looking to get a model like:

$$y_i = \beta_0 + \beta_1 * x_i + \varepsilon_i$$

after the fitting, for each data point we would have:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i + r_i$$

where r_i is residual. It can be rewritten as:

$$\hat{\beta}_0 + r_i = y_i - \hat{\beta}_1 * x_i \quad (1)$$

The first principal component z_1 calculated on (x, y) is

$$z_{i1} = \phi_{i1} y_i + \phi_{i2} x_i$$

Dividing it by ϕ_{i1}:

$$\frac{z_{i1}}{\phi_{i1}} = y_i + \frac{\phi_{i2}}{\phi_{i1}} x_i \quad (2)$$

There is a functional resemblance between equations (1) and (2) (described linear relationship between y and x). Is the following true:

$$\hat{\beta}_0 + r_i = \frac{z_{i1}}{\phi_{i1}}$$

$$\frac{\phi_{i2}}{\phi_{i1}} = -\hat{\beta}_1$$

Answer: *(just yes or no)*

Soln: YES.

What is the difference between linear regression coefficients optimization and first PCA calculations?

Answer:

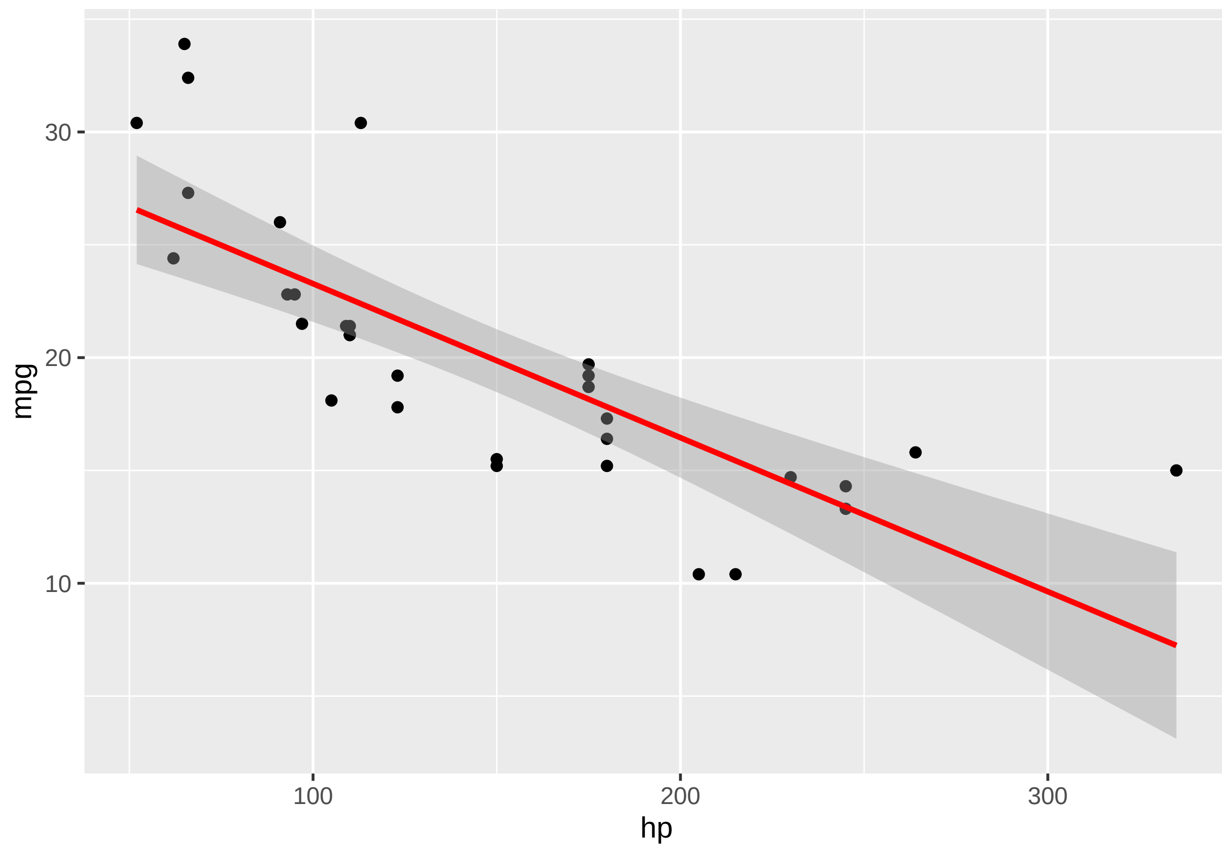
PCA calculations aim to identify a sequence of linear combinations of variables that exhibit maximal variance and are mutually uncorrelated. In contrast, linear regression involves both independent and dependent variables. The process of finding the relationship between these variables, while minimizing the difference between predicted and actual values, is known as linear regression coefficients optimization. PCA is predominantly utilized for feature extraction, whereas linear regression and coefficient optimization are used for estimate the impact of predictors on the dependent variable.

```r
#Regression Line below for mtcars data set
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.1
```

```r
scatter_plot<-ggplot(data=mtcars,aes(x = hp, y = mpg))+
  geom_point()+
  geom_smooth(method="lm",color ="red")
print(scatter_plot)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
#Scree plot I have given below  in the below document.
```

Part 2. PCA Exercise (27 points).

In this exercise we will study UK Smoking Data (`smoking.R`, `smoking.rda` or `smoking.csv`):

Description

Survey data on smoking habits from the UK. The data set can be used for analyzing the demographic characteristics of smokers and types of tobacco consumed.

Format

A data frame with 1691 observations on the following 12 variables.

`gender` - Gender with levels Female and Male.

`age` - Age.

`marital_status` - Marital status with levels Divorced, Married, Separated, Single and Widowed.

`highest_qualification` - Highest education level with levels A Levels, Degree, GCSE/CSE, GCSE/O Level, Higher/Sub Degree, No Qualification, ONC/BTEC and Other/Sub Degree

`nationality` - Nationality with levels British, English, Irish, Scottish, Welsh, Other, Refused and Unknown.

`ethnicity` - Ethnicity with levels Asian, Black, Chinese, Mixed, White and Refused Unknown.

`gross_income` - Gross income with levels Under 2,600, 2,600 to 5,200, 5,200 to 10,400, 10,400 to 15,600, 15,600 to 20,800, 20,800 to 28,600, 28,600 to 36,400, Above 36,400, Refused and Unknown.

`region` - Region with levels London, Midlands & East Anglia, Scotland, South East, South West, The North and Wales

`smoke` - Smoking status with levels No and Yes

`amt_weekends` - Number of cigarettes smoked per day on weekends.

`amt_weekdays` - Number of cigarettes smoked per day on weekdays.

`type` - Type of cigarettes smoked with levels Packets, Hand-Rolled, Both/Mainly Packets and Both/Mainly Hand-Rolled

Source National STEM Centre, Large Datasets from stats4schools, https://www.stem.org.uk/resources/elibrary/resource/28452/large-datasets-stats4schools.

Obtained from https://www.openintro.org/data/index.php?data=smoking

Read and Clean the Data

2.1 Read the data from smoking.R or smoking.rda (3 points) > hint: take a look at source or load functions > there is also smoking.csv file for a refference

```
# load libraries
library(ggplot2)
library(plyr)
library(dplyr)
library(caret)
library(tibble)
library(plotly)
library(ggplot2)
library(ggplot2)
library(devtools)
install_github("vqv/ggbiplot")
library(ggbiplot)
```

```
# Load data
data_smokes <- source("C:\\Users\\Nikhil\\Sem 2\\SDM 2\\HomeWork2\\smoking.R")
data_smoke <- data.frame(do.call(cbind, data_smokes), check.names = FALSE)
colnames(data_smoke) <- sub("value\\.", "", colnames(data_smoke))
```

Take a look into data

```r
# place holder
head(data_smoke,  n =8)
```

```
##   gender age marital_status highest_qualification nationality ethnicity
## 1   Male  38       Divorced      No Qualification     British     White
## 2 Female  42         Single      No Qualification     British     White
## 3   Male  40        Married                Degree     English     White
## 4 Female  40        Married                Degree     English     White
## 5 Female  39        Married          GCSE/O Level     British     White
## 6 Female  37        Married          GCSE/O Level     British     White
## 7   Male  53        Married                Degree     British     White
## 8   Male  44         Single                Degree     English     White
##        gross_income    region smoke amt_weekends amt_weekdays    type visible
## 1   2,600 to 5,200 The North    No           NA           NA           FALSE
## 2      Under 2,600 The North   Yes           12           12 Packets   FALSE
## 3 28,600 to 36,400 The North    No           NA           NA           FALSE
## 4 10,400 to 15,600 The North    No           NA           NA           FALSE
## 5   2,600 to 5,200 The North    No           NA           NA           FALSE
## 6 15,600 to 20,800 The North    No           NA           NA           FALSE
## 7     Above 36,400 The North   Yes            6            6 Packets   FALSE
## 8 10,400 to 15,600 The North    No           NA           NA           FALSE
```

There are many fields there so for this exercise lets only concentrate on smoke, gender, age, marital_status, highest_qualification and gross_income.

Create new data.frame with only these columns.

```r
# place holder
data_smoke_filter<-select(data_smoke,c('smoke','gender','age','marital_status','highest_qualification',
```

2.2 Omit all incomplete records.(3 points)

```r
# place holder
data_smoke_filter<-na.omit(data_smoke_filter)
data_filter_final<- data_smoke_filter
```

2.3 For PCA feature should be numeric. Some of fields are binary (`gender` and `smoke`) and can easily be converted to numeric type (with one and zero). Other fields like `marital_status` has more than two categories, convert them to binary (e.g. is_married, is_devorced). Several features in the data set are ordinal (`gross_income` and `highest_qualification`), convert them to some king of sensible level (note that levels in factors are not in order). (3 points)

```r
# place holder
# Assuming your column is numeric

data_smoke_filter$gender<-ifelse(data_smoke_filter$gender=='Female',1,0)
data_smoke_filter$smoke<-ifelse(data_smoke_filter$smoke=='Yes',1,0)
data_smoke_filter <- data_smoke_filter %>%
  mutate(is_separated = as.integer(marital_status == "Separated"),
         is_married = as.integer(marital_status == "Married"),
         is_divorced = as.integer(marital_status=="Divorced"),
         is_widowed = as.integer(marital_status == "Widowed"),
```

```r
        is_single = as.integer(marital_status == "Single"))



data_smoke_filter$highest_qualification_bin <- factor(data_smoke_filter$highest_qualification, levels =
data_smoke_filter$highest_qualification_bin<-as.numeric(data_smoke_filter$highest_qualification_bin)
data_smoke_filter$gross_income_bin <- factor(data_smoke_filter$gross_income,levels=c("Under 2,600", "2,
),labels = c(0, 0, 1, 1, 2, 2, 3, 4, 0))
data_smoke_filter$gross_income_bin<-as.numeric(data_smoke_filter$gross_income_bin)
data_smoke_filter <- subset(data_smoke_filter, select = -c(marital_status, highest_qualification, gross
data_smoke_filter <- na.omit(data_smoke_filter)
```

2.4. Do PCA on all columns except smoking status. (3 points)

```r
# place holder
pca1 <- prcomp(select(data_smoke_filter, -c("smoke")),scale.=TRUE)
summary(pca1)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     1.4567 1.3136 1.0844 1.0683 1.0283 0.90964 0.76477
## Proportion of Variance 0.2358 0.1917 0.1306 0.1268 0.1175 0.09194 0.06499
## Cumulative Proportion  0.2358 0.4275 0.5582 0.6850 0.8024 0.89438 0.95937
##                            PC8       PC9
## Standard deviation     0.60471 1.934e-15
## Proportion of Variance 0.04063 0.000e+00
## Cumulative Proportion  1.00000 1.000e+00
```

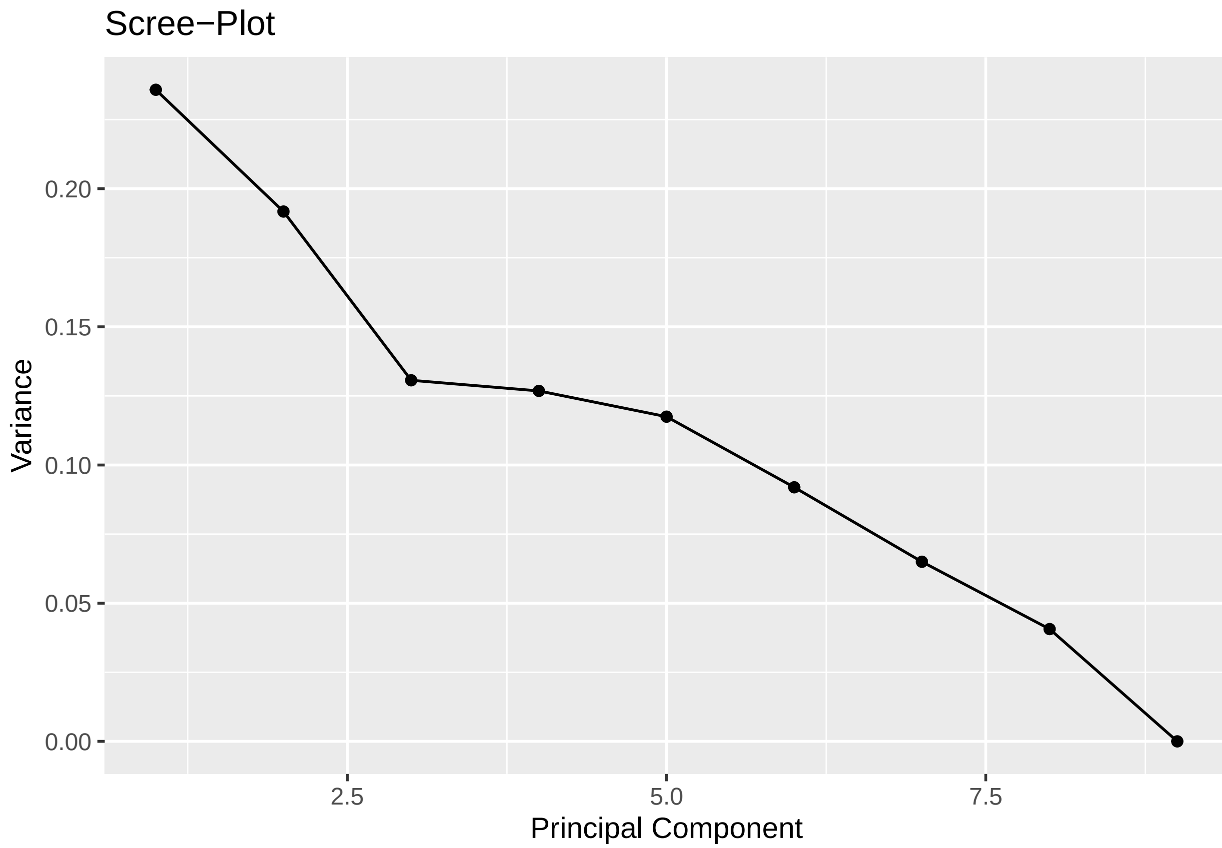
2.5 Make a scree plot (3 points)

```r
# place holder
var <- (pca1$sdev)^2 / sum(pca1$sdev^2)
round(var,3)
```

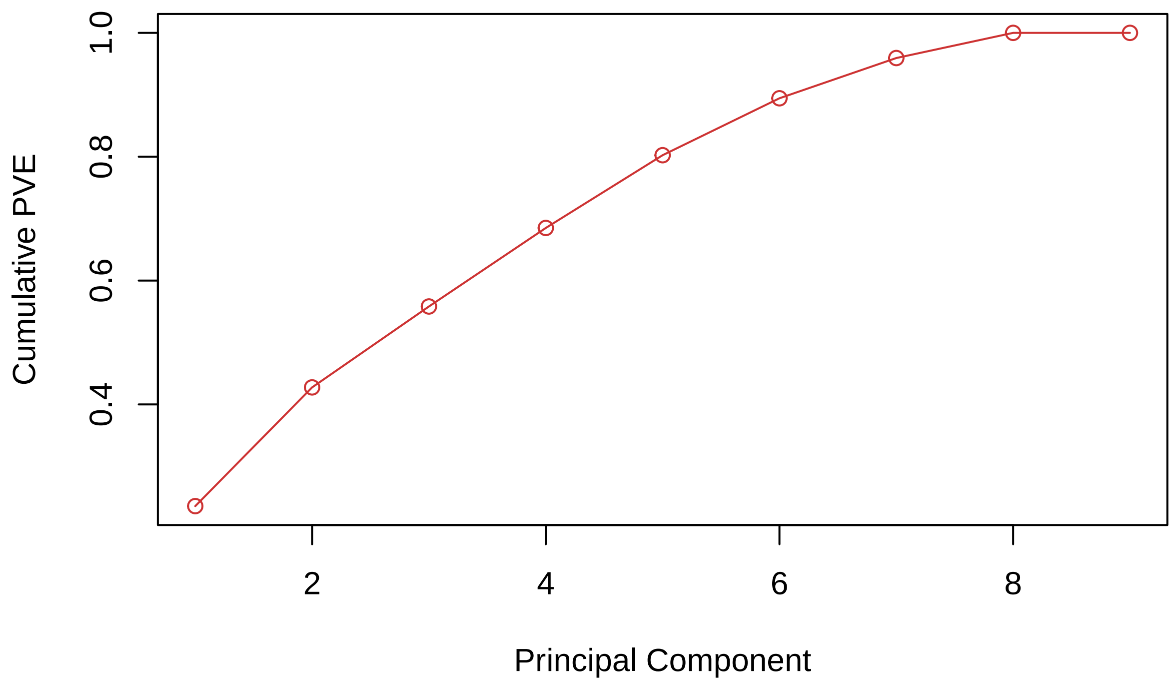
```
## [1] 0.236 0.192 0.131 0.127 0.117 0.092 0.065 0.041 0.000
```

```r
qplot(c(1:9), var) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance") +
  ggtitle("Scree-Plot") +
  scale_y_continuous(breaks = seq(0, 0.30, 0.05))
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```r
plot(cumsum(var), type = "o", ylab = "Cumulative PVE",
    xlab = "Principal Component", col = "brown3")
```

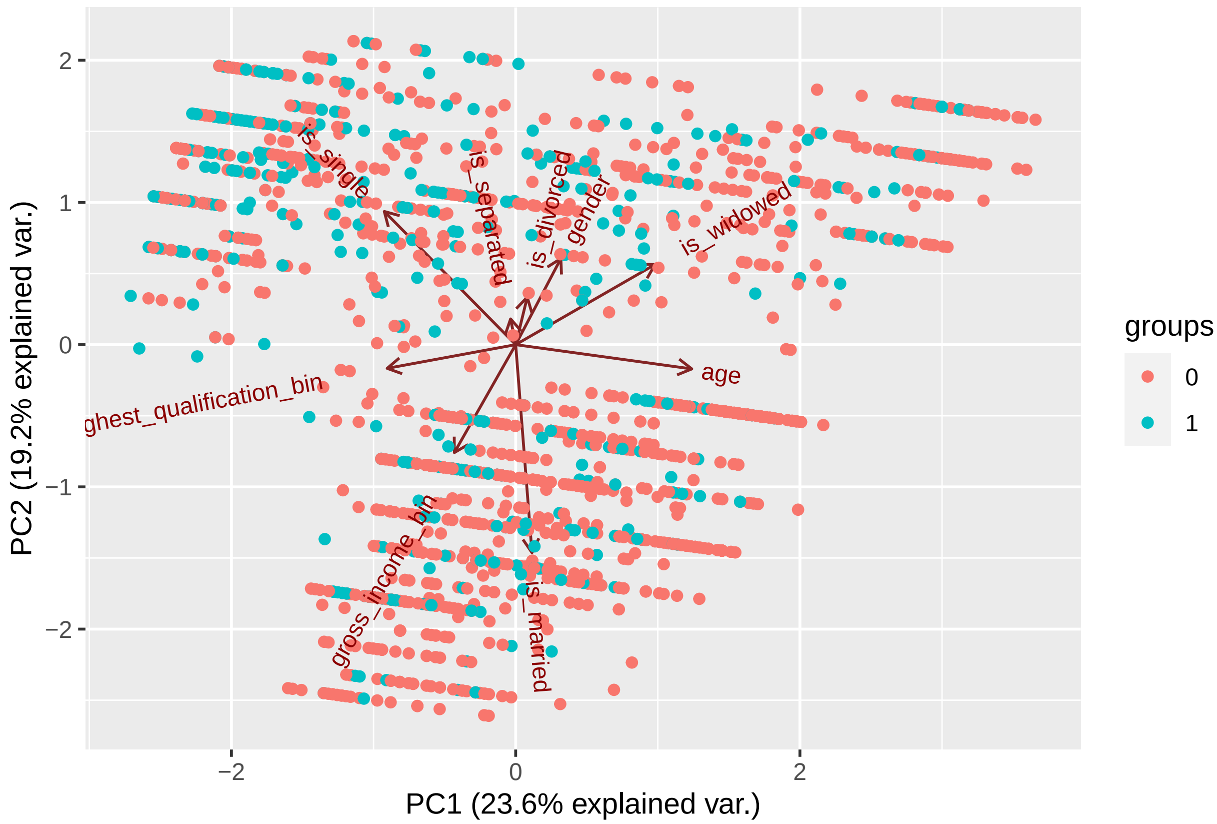


Comment on the shape, if you need to reduce dimensions home many would you choose

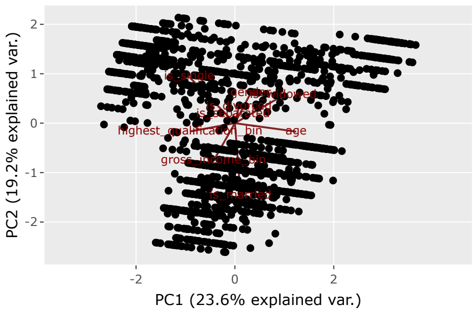
```
I've observed a uniform distribution of variance across all PCs. Upon analyzing the plot using the elbo
a sharp bend is noticeable at PC8. Consequently, utilizing all eight PC's captures approximately 90% of
variance in predictors.
```

2.6 Make a biplot color points by smoking field. (3 points)

```r
# place holder
ggbiplot(pca1, scale=0, groups = as.factor(data_smoke_filter$smoke))
```




```r
ggplotly(ggbiplot(pca1, scale = 0),groups = as.factor(data_smoke_filter$smoke))
```



Comment on observed biplot.

The variables ’is_separated,’ ’is_divorced,’ and ’gender’ exhibit a highly positve correlation. This sug

discernible statistical relationship, indicating that individuals who are identified as separated or di
tend to be more likely associated with the female gender. Furthermore, there exists a noteworthy negati
between 'age' and 'highest_qualification_bin.' This mean that on average, as individuals' ages increase
their categorized highest qualifications tend to decrease, or vice versa.

Can we use first two PC to discriminate smoking?

The initial two principal components encapsulate approximately 44% of the total variance.
This observation prompts the consideration that relying solely on the first two principal
components may not be sufficiently discriminative for distinguishing smoking behavior.

2.7 Based on the loading vector can we name PC with some descriptive name? (3 points)

The variables 'is_separated' and 'is_divorced' shows a positive correlation,suggesting a close relation
We can group them under 'was_married.

2.8 May be some of splits between categories or mapping to numerics should be revisited, if so what will you do differently? (3 points)

As I mentioned above 'is_separated' and 'is_divorced' shows a positive correlation so would considered
Also I would remove is_widowed as it give very less value to dataset which we have.

2.9 Follow your suggestion in 2.10 and redo PCA and biplot (3 points)

```
# I re filtered everything.
data_filter_final$gender<-ifelse(data_filter_final$gender=='Female',1,0)
data_filter_final$smoke<-ifelse(data_filter_final$smoke=='Yes',1,0)
data_filter_final <- data_filter_final %>%
  mutate(was_married = as.integer(marital_status == "Divorced"),
         is_single = as.integer(marital_status == "Single"),
         is_married = as.integer(marital_status == "Married"),
         is_widowed = as.integer(marital_status == "Widowed"),
         was_married = as.integer(marital_status == "Separated"))
data_filter_final$highest_qualification_bin <- factor(data_filter_final$highest_qualification, levels =
data_filter_final$highest_qualification_bin<-as.numeric(data_filter_final$highest_qualification_bin)
data_filter_final$gross_income_bin <- factor(data_filter_final$gross_income,levels=c("Under 2,600", "2,
),labels = c(0, 0, 1, 1, 2, 2, 3, 4, 0))
data_filter_final$gross_income_bin<-as.numeric(data_filter_final$gross_income_bin)
data_filter_final <- subset(data_filter_final, select = -c(marital_status, highest_qualification, gross
data_filter_final <- na.omit(data_filter_final)
#Redo the PCA
pca_new <- prcomp(select(data_filter_final, -c("smoke", "is_widowed")),scale.=TRUE)
pca_new
```

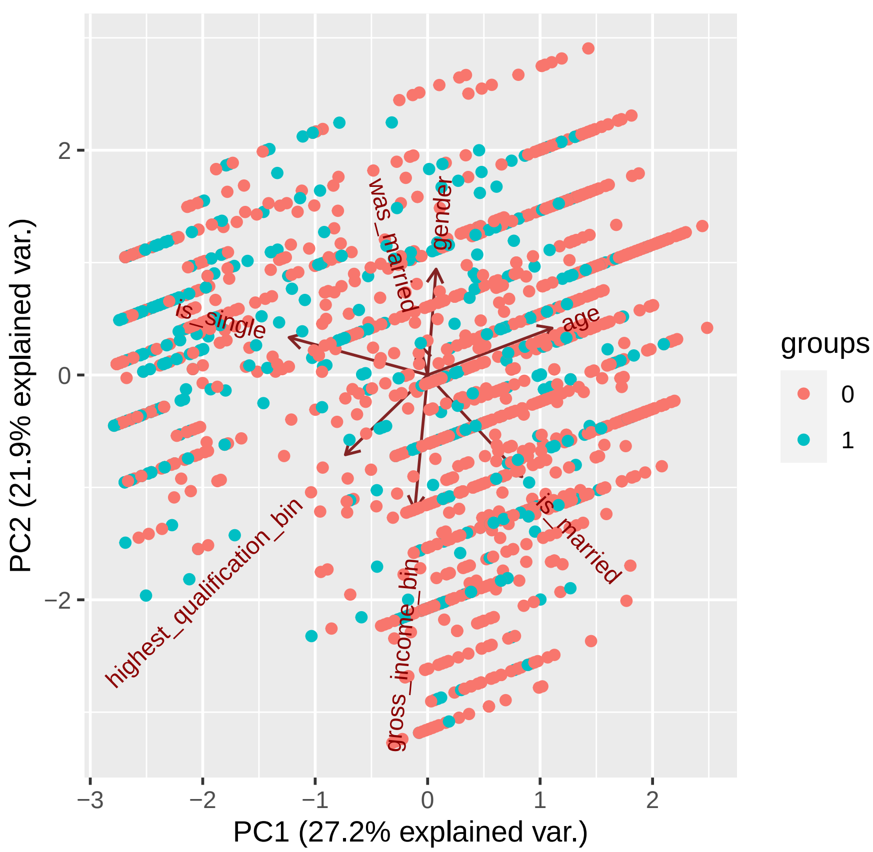
```
## Standard deviations (1, .., p=7):
## [1] 1.3807409 1.2392759 1.0426732 1.0313068 0.7873814 0.7456400 0.4806665
##
## Rotation (n x k) = (7 x 7):
##                                   PC1        PC2        PC3         PC4
## gender                     0.03758545  0.4710199  0.5020167  0.38663588
## age                        0.55300345  0.2084446 -0.2135074 -0.17315951
## was_married               -0.03473495  0.1333606 -0.5434424  0.75080454
```

```
## is_single                  -0.61586498  0.1670137 -0.0465703 -0.33430557
## is_married                  0.41866755 -0.4523506  0.3990228  0.09072697
## highest_qualification_bin -0.36569379 -0.3546354  0.3675523  0.36860504
## gross_income_bin          -0.05713034 -0.5988532 -0.3325381  0.03097978
##                                    PC5         PC6         PC7
## gender                    -0.546056771 -0.25379626 -0.11237883
## age                       -0.280260555  0.56945100 -0.41514820
## was_married                0.227823468 -0.06540290 -0.25649620
## is_single                  0.001778694 -0.08222639 -0.68711236
## is_married                 0.311737291 -0.29010466 -0.52086947
## highest_qualification_bin -0.040279711  0.68259628 -0.04446697
## gross_income_bin          -0.687434895 -0.22399980 -0.06186252
```

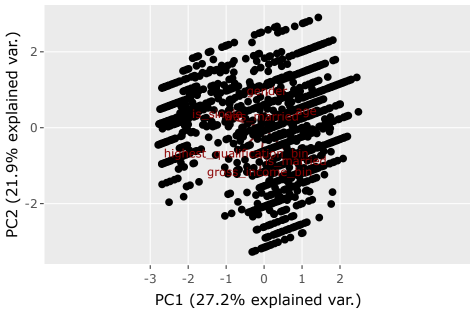
```r
summary(pca_new)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.3807 1.2393 1.0427 1.0313 0.78738 0.74564 0.48067
## Proportion of Variance 0.2723 0.2194 0.1553 0.1519 0.08857 0.07943 0.03301
## Cumulative Proportion  0.2723 0.4918 0.6471 0.7990 0.88757 0.96699 1.00000
```

```r
ggbiplot(pca_new, scale=0, groups = as.factor(data_filter_final$smoke))
```



```r
ggplotly(ggbiplot(pca_new, scale = 0),groups = as.factor(data_filter_final$smoke))
```



Part 3. Freestyle. (27 points).

Get the data set from your final project (or find something suitable). The data set should have at least four variables and it shouldn't be used in class PCA examples: iris, mpg, diamonds and so on).

- Convert a columns to proper format (9 points)
- Perform PCA (3 points)
- Make a skree plot (3 points)
- Make a biplot (3 points)
- Discuss your observations (9 points)

```
#The dataset contains student information, including gender, race/ethnicity, parental education, lunch
data_freestyle <- read.csv("C:\\Users\\Nikhil\\Sem 2\\SDM 2\\HomeWork2\\Student_performance.csv")
data_freestyle<- na.omit(data_freestyle)

#1.Convert a columns to proper format
# 1st convertion change gender into binary
data_freestyle$gender<-ifelse(data_freestyle$gender=='female',1,0)
#In the race.ethnicity column we have Indian and Chinese which are Asians so combining them into 1. Als
data_freestyle$race.ethnicity <- ifelse(data_freestyle$race.ethnicity %in% c("Indian", "Chinese"), "Asia
data_freestyle$race.ethnicity <- ifelse(data_freestyle$race.ethnicity %in% c("African American", "Black
#Catagories asians as 0, White as 1, and black as 2
data_freestyle$race.ethnicity <- ifelse(data_freestyle$race.ethnicity == "Asian", 0,
                              ifelse(data_freestyle$race.ethnicity == "White", 1,
                                     ifelse(data_freestyle$race.ethnicity == "Black", 2, NA)))
#Remove lunch, parental.level.of.education columns as it doesn't show any signifance
data_freestyle <- subset(data_freestyle, select = -c(lunch))
#Catagories test.preparation.course into binary
data_freestyle$test.preparation.course <- ifelse(data_freestyle$test.preparation.course == "none", 0, 1)
#Change the parental.level education into categories
data_freestyle$parental_level <- factor(data_freestyle$parental.level.of.education, levels = c(
  "some college", "master's degree", "associate's degree", "high school", "some high school", "bachelor
), labels = c(1, 1, 1, 0, 0, 2))
data_freestyle <- subset(data_freestyle, select = -c(parental.level.of.education))
data_freestyle <- data_freestyle %>%
  mutate(
    gender = as.numeric(gender),
    race.ethnicity = as.numeric(race.ethnicity),
    test.preparation.course = as.numeric(test.preparation.course),
    math.score = as.integer(math.score),
    parental_level = as.numeric(as.factor(parental_level))
  )
data_freestyle <- na.omit(data_freestyle)

#2.Performed PCA for all except overall score
pca2 <- prcomp(select(data_freestyle, -c("Overall.Result")),scale.=TRUE)
pca2
```

```
## Standard deviations (1, .., p=5):
## [1] 1.1502193 1.0228647 0.9998746 0.9564999 0.8462282
##
## Rotation (n x k) = (5 x 5):
##                                  PC1         PC2         PC3          PC4
```

```
## gender                    -0.4437943 -0.05158954  0.57421409 -0.593249394
## race.ethnicity             0.4473532  0.45907576 -0.09615780 -0.673754193
## test.preparation.course    0.3611190 -0.65474870  0.47152708 -0.027849446
## math.score                 0.6788621 -0.08221222  0.08340099  0.001032737
## parental_level            -0.1079862 -0.59256498 -0.65706799 -0.439697380
##                                  PC5
## gender                    -0.3445555
## race.ethnicity             0.3548712
## test.preparation.course    0.4666731
## math.score                -0.7248659
## parental_level            -0.1101526
```

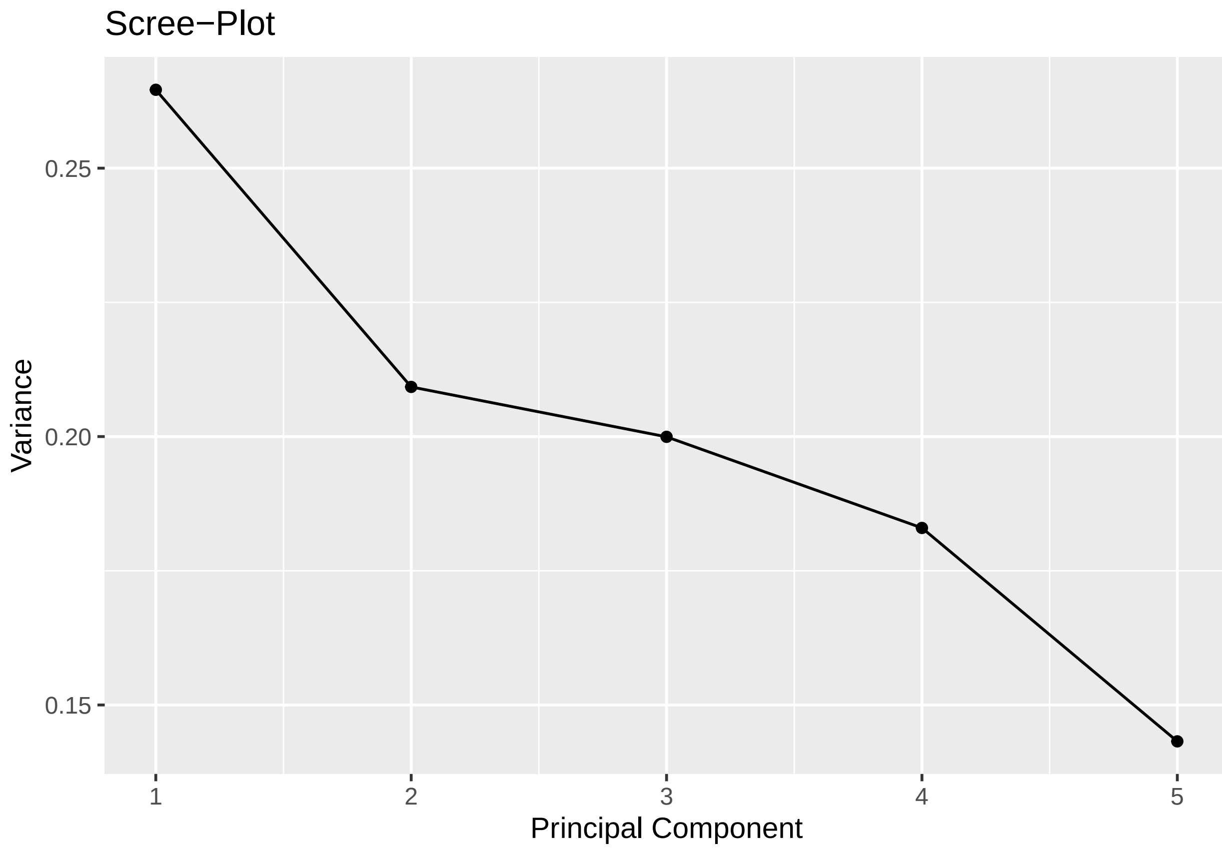
```r
summary(pca2)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5
## Standard deviation     1.1502 1.0229 0.9999 0.9565 0.8462
## Proportion of Variance 0.2646 0.2092 0.1999 0.1830 0.1432
## Cumulative Proportion  0.2646 0.4738 0.6738 0.8568 1.0000
```

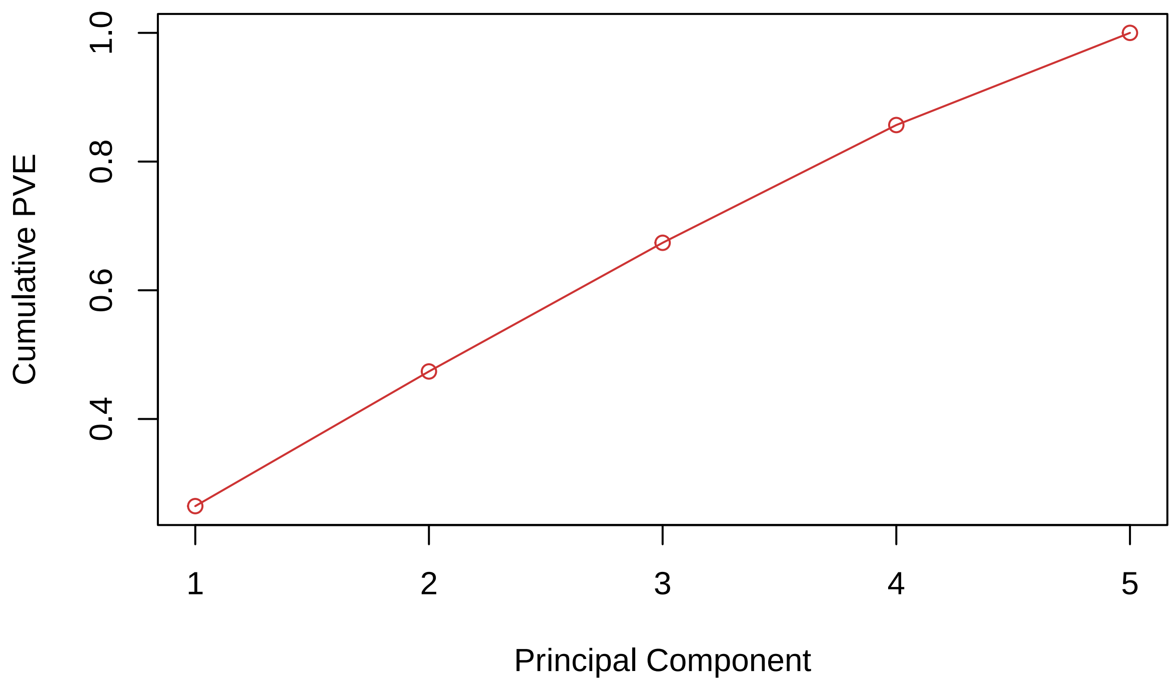
```r
#3. Made a scree plot
var <- (pca2$sdev)^2 / sum(pca2$sdev^2)
round(var,3)
```

```
## [1] 0.265 0.209 0.200 0.183 0.143
```

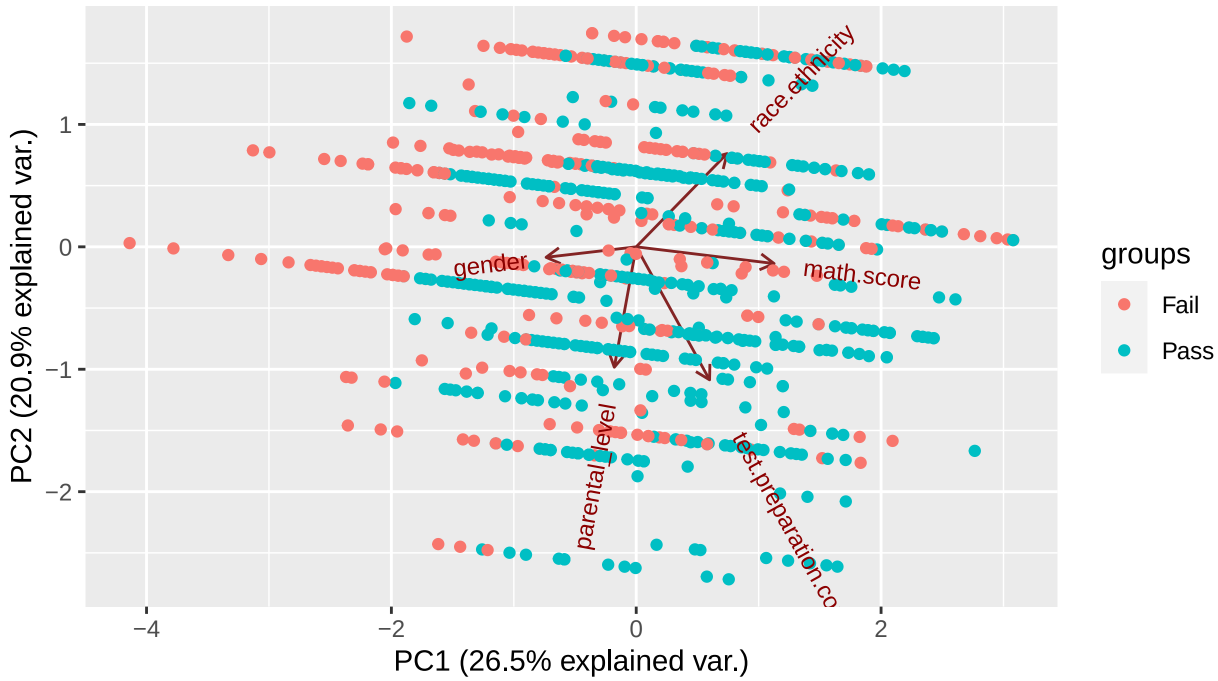
```r
qplot(c(1:5), var) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance") +
  ggtitle("Scree-Plot") +
  scale_y_continuous(breaks = seq(0, 0.30, 0.05))
```




```r
plot(cumsum(var), type = "o", ylab = "Cumulative PVE",
    xlab = "Principal Component", col = "brown3")
```




```r
#4.Make a big plot
ggbiplot(pca2, scale=0, groups = as.factor(data_freestyle$Overall.Result))
```



#5.Discuss your observations (9 points) I consider four features to be particularly crucial for our dataset as we can see a sharp bend after that. Furthermore, there seems to be a negative correlation between gender and math scores. In my assessment, the existing category divisions adequately capture the essence of the data, and I don't see the necessity for introducing new categories. However, I believe that incorporating additional features could enhance the model's ability to accurately predict pass or fail outcomes.