Homework 2. PCA.

Nisarg Negi

2022-10-03

Part 1. PCA vs Linear Regression (10 points).

Lets say we have two 'features', let one be x and another y. Recall that In linear regression we are looking to get 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:

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

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

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

Dividing it by ϕ_{i1}:

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

There is a functional resemblance between two last equation (described linear relationship between y and x). Is following true:

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

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

Answer: Yes

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

Answer: In first PCA calculation, error squares are minimized by taking PCA orthogal to the straight such that the variance is zero while in linear regression, error squares are minimized in y direction.

(here should be the answere. help yourself with a plot)

Part 2. PCA Exercise (45 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 (https://www.stem.org.uk/resources/elibrary/resource/28452/large-datasets-stats4schools).

Obtained from https://www.openintro.org/data/index.php?data=smoking (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 > hint: take a look at source or load functions > there is also smoking.csv file for a refference

```
# Load libraries
library(readr)
library(dplyr)
library(tidyr)
library(data.table)
library(data.table)
library(plotly)
library(lubridate)
library(ggbiplot)
library(caret)
```

```
# Load data
load("smoking.rda")
```

Take a look into data

```
# place holder
head(smoking)
```

gen… <fct>	…	marital_status <int><fct>	highest_qualification <fct>	nationality <fct>	ethnicity <fct>	gross_income <fct>
Male	38	Divorced	No Qualification	British	White	2,600 to 5,200
Female	42	Single	No Qualification	British	White	Under 2,600
Male	40	Married	Degree	English	White	28,600 to 36,400
Female	40	Married	Degree	English	White	10,400 to 15,600
Female	39	Married	GCSE/O Level	British	White	2,600 to 5,200
Female	37	Married	GCSE/O Level	British	White	15,600 to 20,800

6 rows | 1-8 of 12 columns

```
summary(smoking)
```

```
##     gender         age          marital_status       highest_qualification
##  Female:965   Min.   :16.00   Divorced :161    No Qualification :586
##  Male  :726   1st Qu.:34.00   Married  :812    GCSE/O Level     :308
##               Median :48.00   Separated: 68    Degree           :262
##               Mean   :49.84   Single   :427    Other/Sub Degree :127
##               3rd Qu.:65.50   Widowed  :223    Higher/Sub Degree:125
##               Max.   :97.00                    A Levels         :105
##                                                (Other)          :178
##    nationality    ethnicity             gross_income
##  English :833   Asian  :  41   5,200 to 10,400 :396
##  British :538   Black  :  34   10,400 to 15,600:268
##  Scottish:142   Chinese:  27   2,600 to 5,200  :257
##  Other   : 71   Mixed  :  14   15,600 to 20,800:188
##  Welsh   : 66   Refused:  13   20,800 to 28,600:155
##  Irish   : 23   Unknown:   2   Under 2,600     :133
##  (Other) : 18   White  :1560   (Other)         :294
##                      region    smoke       amt_weekends    amt_weekdays
##  London                 :182   No :1270   Min.   : 0.00   Min.   : 0.00
##  Midlands & East Anglia:443   Yes: 421   1st Qu.:10.00   1st Qu.: 7.00
##  Scotland               :148              Median :15.00   Median :12.00
##  South East             :252              Mean   :16.41   Mean   :13.75
##  South West             :157              3rd Qu.:20.00   3rd Qu.:20.00
##  The North              :426              Max.   :60.00   Max.   :55.00
##  Wales                  : 83              NA's   :1270    NA's   :1270
##                        type
##                         :1270
##  Both/Mainly Hand-Rolled:  10
##  Both/Mainly Packets    :  42
##  Hand-Rolled            :  72
##  Packets                : 297
##
##
```

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

Create new data.frame with only these columns.

```
# place holder
df1<-smoking
df1<-select(df1,smoke, gender, age, marital_status, highest_qualification,gross_income)
df1
```

smo… <fct>	gender <fct>	a… <int>	marital_status <fct>	highest_qualification <fct>	gross_income <fct>
No	Male	38	Divorced	No Qualification	2,600 to 5,200
Yes	Female	42	Single	No Qualification	Under 2,600
No	Male	40	Married	Degree	28,600 to 36,400

smo… <fct>	gender <fct>	a… <int>	marital_status <fct>	highest_qualification <fct>	gross_income <fct>
No	Female	40	Married	Degree	10,400 to 15,600
No	Female	39	Married	GCSE/O Level	2,600 to 5,200
No	Female	37	Married	GCSE/O Level	15,600 to 20,800
Yes	Male	53	Married	Degree	Above 36,400
No	Male	44	Single	Degree	10,400 to 15,600
Yes	Male	40	Single	GCSE/CSE	2,600 to 5,200
Yes	Female	41	Married	No Qualification	5,200 to 10,400

1-10 of 1,691 rows Previous **1** 2 3 4 5 6 … 170 Next

2.2 Omit all incomplete records.

```
# place holder
df1 %>% drop_na()
```

smo… <fct>	gender <fct>	a… <int>	marital_status <fct>	highest_qualification <fct>	gross_income <fct>
No	Male	38	Divorced	No Qualification	2,600 to 5,200
Yes	Female	42	Single	No Qualification	Under 2,600
No	Male	40	Married	Degree	28,600 to 36,400
No	Female	40	Married	Degree	10,400 to 15,600
No	Female	39	Married	GCSE/O Level	2,600 to 5,200
No	Female	37	Married	GCSE/O Level	15,600 to 20,800
Yes	Male	53	Married	Degree	Above 36,400
No	Male	44	Single	Degree	10,400 to 15,600
Yes	Male	40	Single	GCSE/CSE	2,600 to 5,200
Yes	Female	41	Married	No Qualification	5,200 to 10,400

1-10 of 1,691 rows Previous **1** 2 3 4 5 6 … 170 Next

```
str(df1)
```

```
## tibble [1,691 × 6] (S3: tbl_df/tbl/data.frame)
##  $ smoke               : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 2 1 2 2 ...
##  $ gender              : Factor w/ 2 levels "Female","Male": 2 1 2 1 1 1 2 2 2 1 ...
##  $ age                 : int [1:1691] 38 42 40 40 39 37 53 44 40 41 ...
##  $ marital_status      : Factor w/ 5 levels "Divorced","Married",..: 1 4 2 2 2 2 2 4 4 2 ...
##  $ highest_qualification: Factor w/ 8 levels "A Levels","Degree",..: 6 6 2 2 4 4 2 2 3 6 ...
##  $ gross_income        : Factor w/ 10 levels "10,400 to 15,600",..: 3 9 5 1 3 2 7 1 3 6 ...
```

```
df1=df1[!grepl("Unknown|Refused", df1$gross_income),]
df9<-df1
df1
```

smo… <fct>	gender <fct>	a… <int>	marital_status <fct>	highest_qualification <fct>	gross_income <fct>
No	Male	38	Divorced	No Qualification	2,600 to 5,200
Yes	Female	42	Single	No Qualification	Under 2,600
No	Male	40	Married	Degree	28,600 to 36,400
No	Female	40	Married	Degree	10,400 to 15,600
No	Female	39	Married	GCSE/O Level	2,600 to 5,200
No	Female	37	Married	GCSE/O Level	15,600 to 20,800
Yes	Male	53	Married	Degree	Above 36,400
No	Male	44	Single	Degree	10,400 to 15,600
Yes	Male	40	Single	GCSE/CSE	2,600 to 5,200
Yes	Female	41	Married	No Qualification	5,200 to 10,400

1-10 of 1,565 rows Previous **1** 2 3 4 5 6 … 157 Next

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 (i.e. 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). (5 points)

```
# place holder
#ohe <- dummyVars(" ~ smoke + gender + marital_status + highest_qualification + gross_income", d
ata #= df1, fullRank = T)
#ohe_df <- data.frame(predict(ohe, newdata = df1))
#ohe_df
#ohe_df <- ohe_df[ - as.numeric(which(apply(ohe_df, 2, var) == 0))]
df2 <- data.frame(
  # convert binary from boolean format to numeric
  smoker =as.numeric(df1$smoke=="Yes"),
  male = as.numeric(df1$gender=="Male"),
  age=as.numeric(df1$age),
  # marital_status to multiple binary categories:
  divorced=as.numeric(df1$marital_status=="Divorced"),
  married=as.numeric(df1$marital_status=="Married"),
  separated=as.numeric(df1$marital_status=="Separated"),
  single=as.numeric(df1$marital_status=="Single"),
  widowed=as.numeric(df1$marital_status=="Widowed"))

df2$education <- revalue(df1$highest_qualification, c(
  "No Qualification"=0,

  "GCSE/CSE"=1,
  "GCSE/O Level"=1,
  "ONC/BTEC"=1,
  "A Levels"=1,

  "Other/Sub Degree"=1,
  "Higher/Sub Degree"=1,

  "Degree"=2
  ))
df2$gross_income <- revalue(df1$gross_income,c(
  "Under 2,600"=0,
  "2,600 to 5,200"=1,
  "5,200 to 10,400"=2,
  "10,400 to 15,600"=3,
  "15,600 to 20,800"=4,
  "20,800 to 28,600"=5,
  "28,600 to 36,400"=6,
  "Above 36,400"=7))


#convert to numeric
for(col in colnames(df2)){
  df2[col] <- as.numeric(df2[[col]])
}

head(df2)
```

smoker <dbl>	male <dbl>	a… <dbl>	divorced <dbl>	married <dbl>	separated <dbl>	single <dbl>	widowed <dbl>	education <dbl>

	smoker <dbl>	male <dbl>	a… <dbl>	divorced <dbl>	married <dbl>	separated <dbl>	single <dbl>	widowed <dbl>	education <dbl>
1	0	1	38	1	0	0	0	0	3
2	1	0	42	0	0	0	1	0	3
3	0	1	40	0	1	0	0	0	2
4	0	0	40	0	1	0	0	0	2
5	0	0	39	0	1	0	0	0	1
6	0	0	37	0	1	0	0	0	1

6 rows | 1-10 of 11 columns

```
apply(df2, 2, var)
```

```
##       smoker         male          age     divorced      married    separated
##   0.18944461   0.24569834 340.21001038   0.08774748   0.24995302   0.04041983
##       single      widowed    education gross_income
##   0.18593024   0.11201066   0.81168300   5.83401861
```

2.4. Do PCA on all columns except smoking status. (5 points)

```
# place holder
#pca_df1 <- subset(df1,select=-c(smoke,marital_status,highest_qualification,marital_status))
pr.out <- prcomp(df2[-1],scale = TRUE)
pr.out
```

```
## Standard deviations (1, .., p=9):
## [1] 1.448106e+00 1.277947e+00 1.085353e+00 1.032554e+00 1.007889e+00
## [6] 9.522279e-01 8.542440e-01 6.110390e-01 2.298863e-15
##
## Rotation (n x k) = (9 x 9):
##                       PC1        PC2         PC3          PC4         PC5
## male         -0.069045898  0.2219838 -0.25632568 -0.306080000 -0.60565274
## age           0.592691108 -0.0571466 -0.01327458 -0.072309268 -0.13601513
## divorced      0.009009441 -0.2277011  0.73052957 -0.419257856  0.13197378
## married       0.136937598  0.7514015  0.00608842  0.008375545  0.04035204
## separated    -0.034922482 -0.1198604  0.25780441  0.829196554 -0.29578157
## single       -0.484572029 -0.3450197 -0.42461118 -0.104711862  0.06865573
## widowed       0.432758406 -0.4044061 -0.26348416 -0.004631138 -0.08786250
## education     0.421621057 -0.1522312 -0.14456439 -0.046717500 -0.03787620
## gross_income  0.155058229  0.1043154 -0.25165427  0.156536362  0.70305370
##                       PC6         PC7         PC8           PC9
## male         -0.56031284 -0.31537829 -0.07961684 -2.075081e-16
## age           0.01800593 -0.05823202  0.78602250 -2.377316e-16
## divorced     -0.29039180 -0.08004172 -0.02602050  3.602669e-01
## married       0.15502032  0.14146485 -0.03384155  6.080456e-01
## separated    -0.28151060 -0.03568364  0.05087594  2.445143e-01
## single       -0.07024729  0.21385593  0.35283087  5.244233e-01
## widowed       0.28506267 -0.39457239 -0.41155972  4.070396e-01
## education    -0.36327024  0.75176413 -0.27826386 -1.390551e-16
## gross_income -0.52911030 -0.32075259  0.01083054  3.939220e-17
```

2.5 Make a scree plot (5 points)

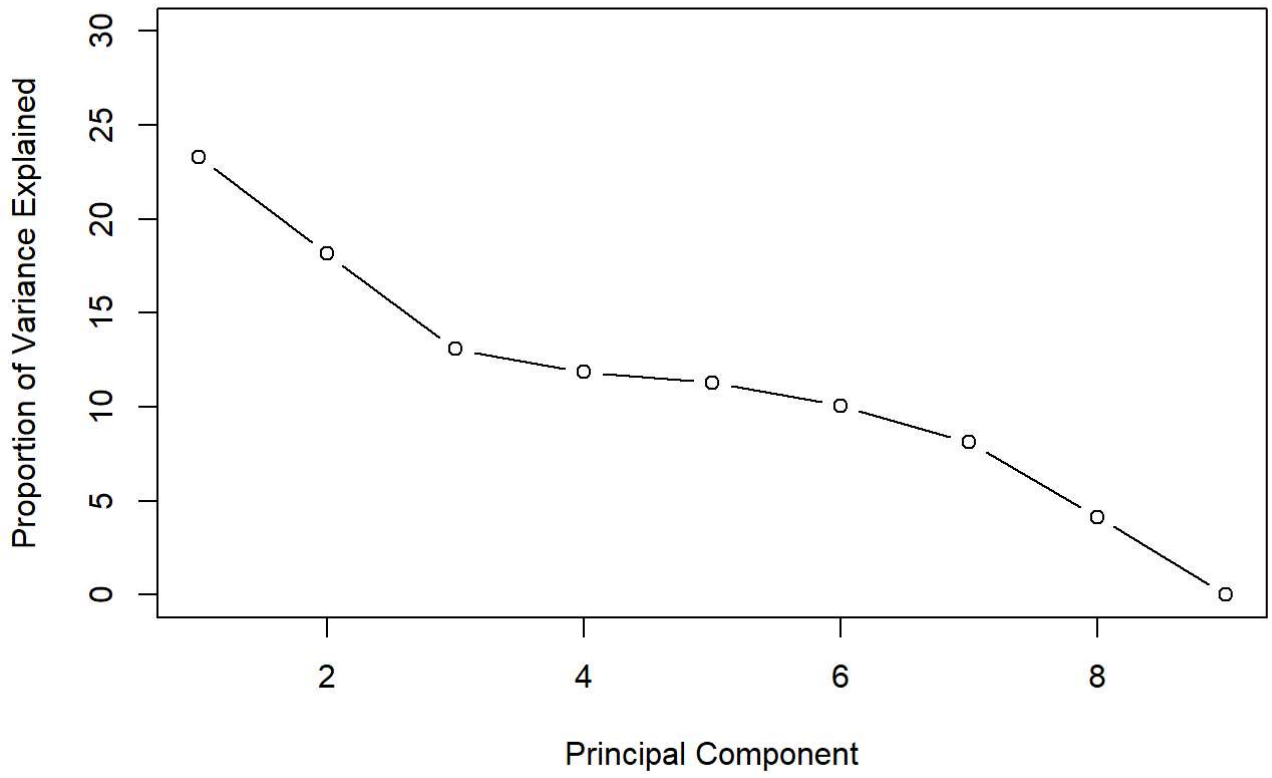
```
# place holder
#calculate total variance explained by each principal component
#var_explained = 100* pr.out$sdev^2 / sum(pr.out$sdev^2)

#create scree plot
#library(ggplot2)

#qplot(c(1:19), var_explained) +
#  geom_line() +
#  xlab("Principal Component") +
#  ylab("Variance Explained") +
#  ggtitle("Scree Plot") +
#  ylim(0, 1)
pr.var <- pr.out$sdev^2
pr.var
```

```
## [1] 2.097011e+00 1.633150e+00 1.177992e+00 1.066167e+00 1.015841e+00
## [6] 9.067380e-01 7.297328e-01 3.733687e-01 5.284773e-30
```

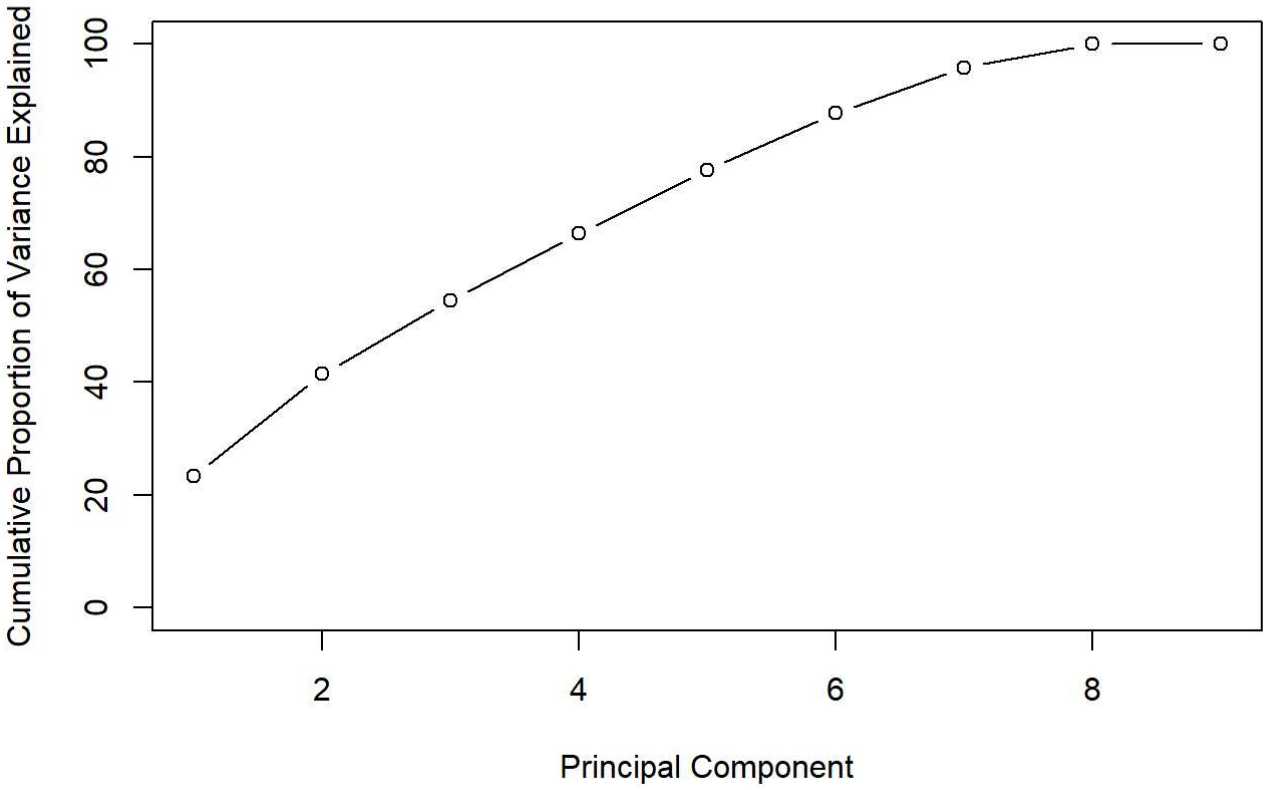
```
pve <- 100 * pr.var / sum(pr.var)
pve
```

```
## [1] 2.330012e+01 1.814611e+01 1.308880e+01 1.184630e+01 1.128712e+01
## [6] 1.007487e+01 8.108142e+00 4.148541e+00 5.871970e-29
```

```
plot(pve, xlab="Principal Component", ylab="Proportion of Variance Explained", ylim=c(0,30),type
='b')
```




```
plot(cumsum(pve), xlab="Principal Component", ylab="Cumulative Proportion of Variance Explained"
, ylim=c(0,100),type='b')
```

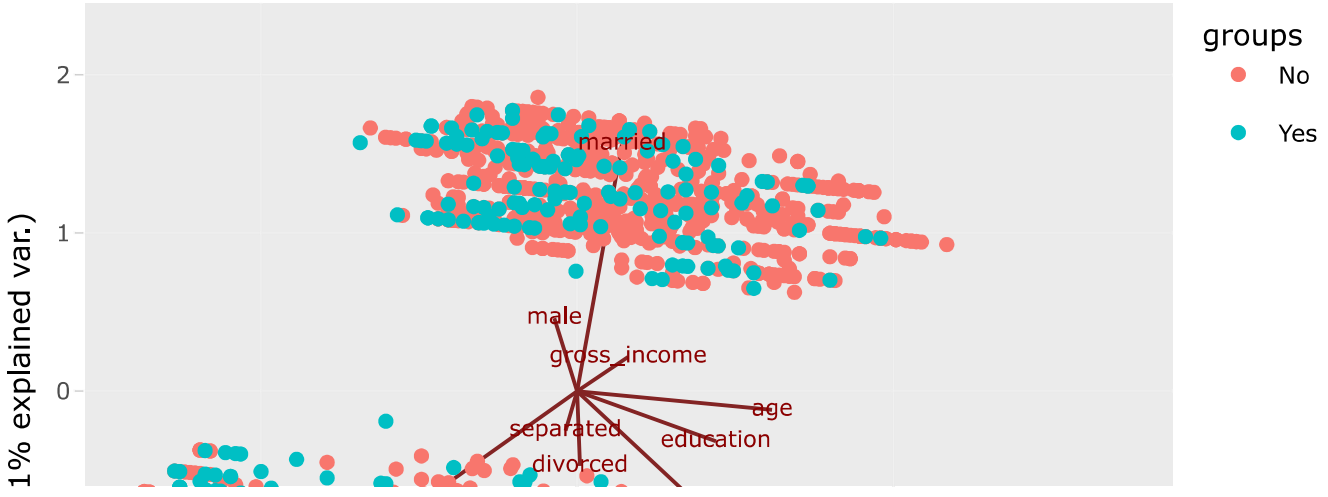


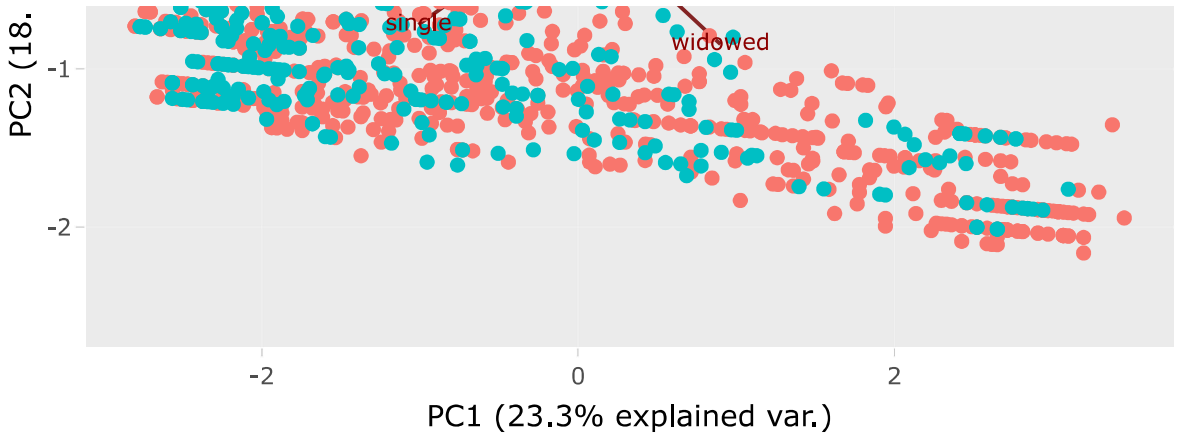
Comment on the shape, if you need to reduce dimensions home many would you choose

```
The Scree plots have a single elbow and the eigen values level off here.
We need just to Principal componenst as at they in total describe more than 40% of the variance
in the dataset.
```

2.6 Make a biplot color points by smoking field. (5 points)

```
# place holder
#biplot(pr.out, scale=0, groups=factor(ohe_df$smoke))
ggplotly(ggbiplot(pr.out, scale = 0, groups=factor(df1$smoke)))
```





Comment on observed biplot.

```
PCA 1 explains 23.3% of the variability of the dataset and PC2 explains 18.1% of the dataset. He
nce the first to principal components explain 41% percent of the variability of the dataset.
In biplot, datapoints are represented as dots and the arrows/vectors contain information on the
loadings.
Red dots are the people who smoke while blue dots are the people who dont.
Length of these vectors interpret to how well the variables are represented in the graph and are
directly proportional to their variability.
The first to PC interpret well about the features marital_status, but dont perform well about th
e rest of the features in the dataset.
The angle between 2 vectors shows represents collinearity. Matital_status single and married are
strongly negative collinear. Age and education shows the most positive colliearity.
```

Can we use first two PC to discriminate smoking?

```
Yes, we can use the first two to disciminate smoking.
```

2.7 Based on the loading vector can we name PC with some descriptive name? (5 points)

```
PC1 = PC_marital_status_single
PC2 = PC_Marital_status_married_Gender_male
```

2.8 May be some of splits between categories or mapping to numerics should be revisited, if so what will you do differently? (5 points)

```
Looking at the biplot I can tell that the major difference is made by marital_status, that too,
there should only be 2 categories among it, first is weather a person is married and second shou
ld be everyone else. So if I do the splits agains, I will combine everything other than married
together.
```

2.9 Follow your suggestion in 2.10 and redo PCA and biplot (5 points)

```
df1<-smoking
df1<-select(df1,smoke, gender, age, marital_status, highest_qualification,gross_income)


df1 %>% drop_na()
```

smo… <fct>	gender <fct>	a… <int>	marital_status <fct>	highest_qualification <fct>	gross_income <fct>
No	Male	38	Divorced	No Qualification	2,600 to 5,200
Yes	Female	42	Single	No Qualification	Under 2,600
No	Male	40	Married	Degree	28,600 to 36,400
No	Female	40	Married	Degree	10,400 to 15,600
No	Female	39	Married	GCSE/O Level	2,600 to 5,200
No	Female	37	Married	GCSE/O Level	15,600 to 20,800
Yes	Male	53	Married	Degree	Above 36,400
No	Male	44	Single	Degree	10,400 to 15,600
Yes	Male	40	Single	GCSE/CSE	2,600 to 5,200
Yes	Female	41	Married	No Qualification	5,200 to 10,400

```
df1=df1[!grepl("Unknown|Refused", df1$gross_income),]
df9<-df1


df3 <- data.frame(
  # convert binary from boolean format to numeric
  smoker =as.numeric(df9$smoke=="Yes"),
  male = as.numeric(df9$gender=="Male"),
  age=as.numeric(df9$age),

  married=as.numeric(df9$marital_status=="Married" ))

df3$education <- revalue(df9$highest_qualification, c(
  "No Qualification"=0,

  "GCSE/CSE"=1,
  "GCSE/O Level"=1,
  "ONC/BTEC"=1,
  "A Levels"=1,

  "Other/Sub Degree"=1,
  "Higher/Sub Degree"=1,

  "Degree"=2
  ))
df3$gross_income <- revalue(df9$gross_income,c(
  "Under 2,600"=0,
  "2,600 to 5,200"=1,
  "5,200 to 10,400"=2,
  "10,400 to 15,600"=3,
  "15,600 to 20,800"=4,
  "20,800 to 28,600"=5,
  "28,600 to 36,400"=6,
  "Above 36,400"=7))


#convert to numeric
for(col in colnames(df9)){
  df9[col] <- as.numeric(df9[[col]])
}

head(df9)
```

smoke <dbl>	gender <dbl>	a… <dbl>	marital_status <dbl>	highest_qualification <dbl>	gross_income <dbl>
1	2	38	1	6	3
2	1	42	4	6	9
1	2	40	2	2	5

smoke <dbl>	gender <dbl>	a... <dbl>	marital_status <dbl>	highest_qualification <dbl>	gross_income <dbl>
1	1	40	2	2	1
1	1	39	2	4	3
1	1	37	2	4	2

6 rows

```
apply(df9, 2, var)
```

```
##                 smoke                gender                   age
##             0.1894446             0.2456983           340.2100104
##        marital_status highest_qualification          gross_income
##             1.6044279             3.8807367             5.8340186
```

```
#PCA except smoking
pr.out <- prcomp(df9[-1],scale = TRUE)
pr.out
```

```
## Standard deviations (1, .., p=5):
## [1] 1.1741713 1.0237360 1.0055189 0.9550250 0.8063159
##
## Rotation (n x k) = (5 x 5):
##                                PC1        PC2         PC3         PC4
## gender                -0.09489883 -0.7838009  0.04549135 -0.61200669
## age                    0.68255652 -0.1365162  0.07092028  0.06803896
## marital_status         0.21074943  0.3842261 -0.69003281 -0.57601092
## highest_qualification  0.66405755 -0.2180501 -0.06234255  0.17769926
## gross_income           0.19930643  0.4145407  0.71614871 -0.50739816
##                                PC5
## gender                -0.005362163
## age                    0.711210904
## marital_status        -0.004593319
## highest_qualification -0.689940635
## gross_income          -0.134577559
```

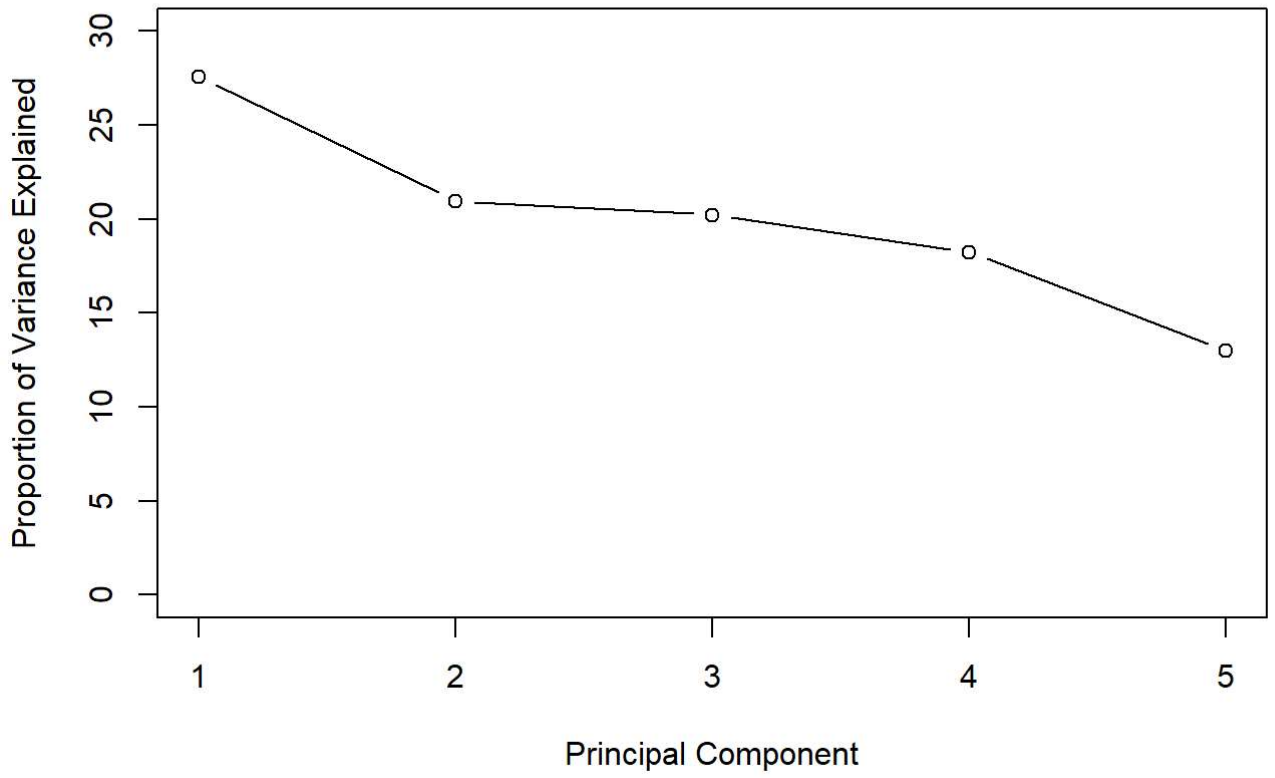
```
#SCREE
pr.var2 <- pr.out$sdev^2
pr.var2
```

```
## [1] 1.3786783 1.0480353 1.0110682 0.9120727 0.6501454
```

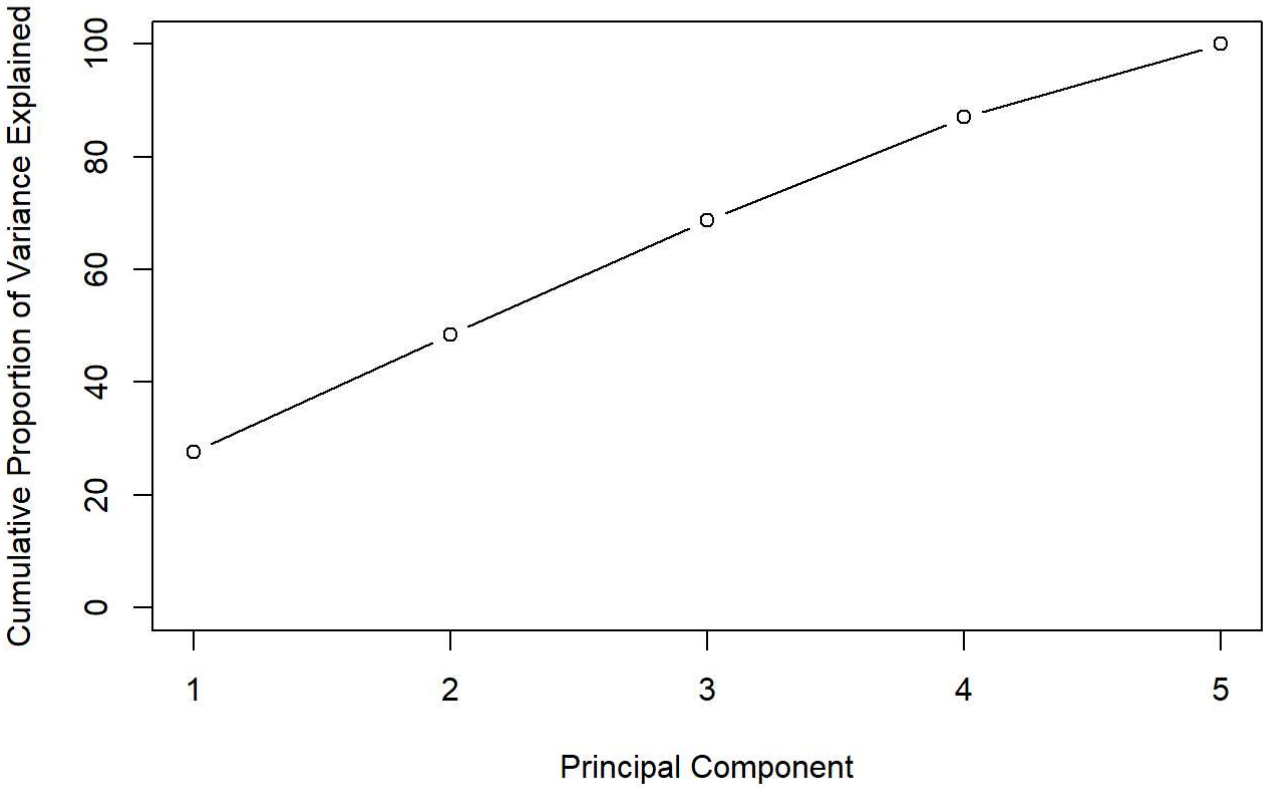
```
pve2 <- 100 * pr.var2 / sum(pr.var2)
pve2
```

```
## [1] 27.57357 20.96071 20.22136 18.24145 13.00291
```

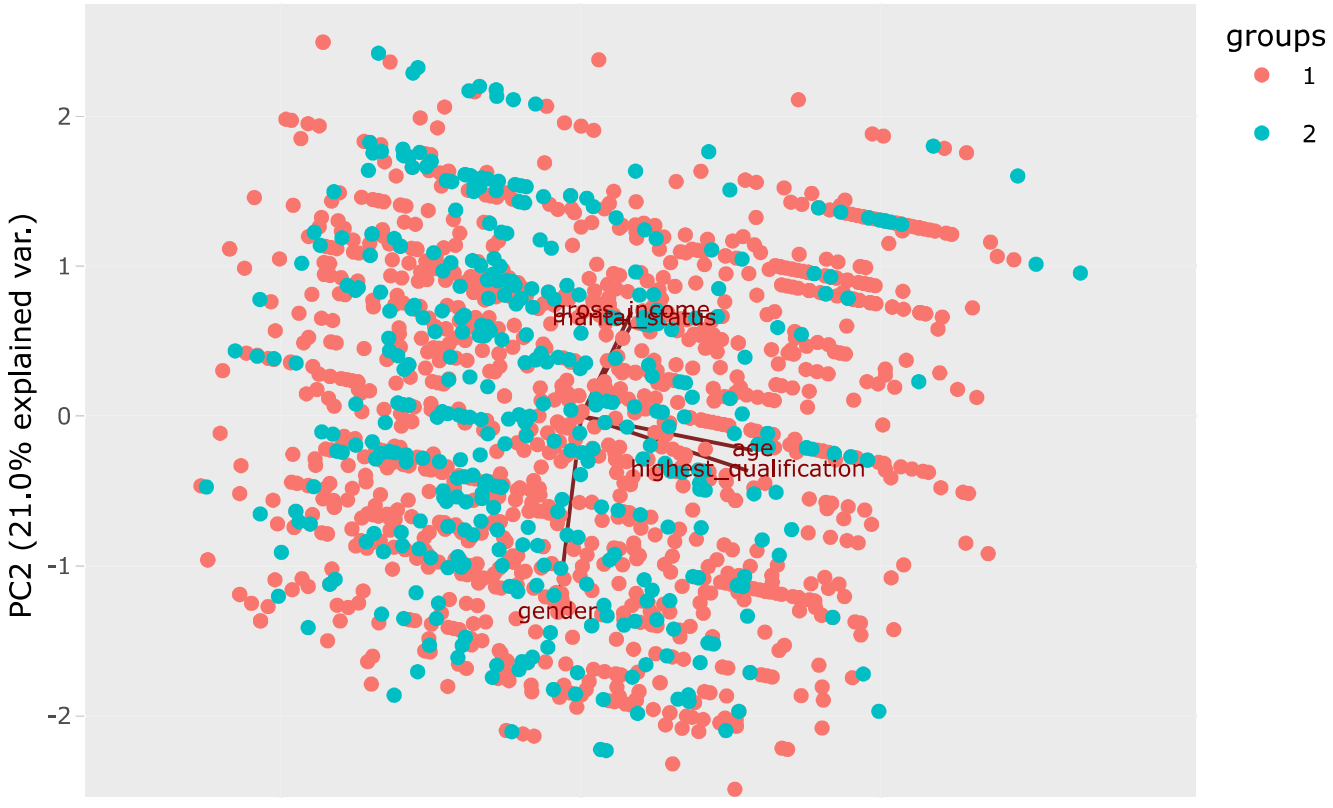
```
plot(pve2, xlab="Principal Component", ylab="Proportion of Variance Explained", ylim=c(0,30),type='b')
```

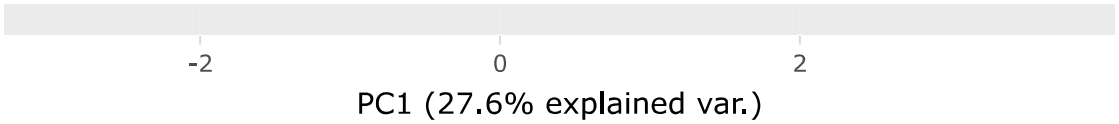



```
plot(cumsum(pve2), xlab="Principal Component", ylab="Cumulative Proportion of Variance Explained", ylim=c(0,100),type='b')
```




```
#Biplot
ggplotly(ggbiplot(pr.out, scale = 0, groups=factor(df9$smoke)))
```





Part 3. Freestyle. (45 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 (15 points)
- Perform PCA (5 points)
- Make a skree plot (5 points)
- Make a biplot (5 points)
- Discuss your observations and how PCA can be used in your final project. (15 points)

```
sales <- read.csv(file = "car_data.csv")
head(sales)
```

	User.ID <int>	Gender <chr>	Age <int>	AnnualSalary <int>	Purchased <int>
1	385	Male	35	20000	0
2	681	Male	40	43500	0
3	353	Male	49	74000	0
4	895	Male	40	107500	1
5	661	Male	25	79000	0
6	846	Female	47	33500	1

6 rows

```
summary(sales)
```

```
##     User.ID          Gender               Age         AnnualSalary
##  Min.   :   1.0   Length:1000        Min.   :18.00   Min.   : 15000
##  1st Qu.: 250.8   Class :character   1st Qu.:32.00   1st Qu.: 46375
##  Median : 500.5   Mode  :character   Median :40.00   Median : 72000
##  Mean   : 500.5                      Mean   :40.11   Mean   : 72689
##  3rd Qu.: 750.2                      3rd Qu.:48.00   3rd Qu.: 90000
##  Max.   :1000.0                      Max.   :63.00   Max.   :152500
##    Purchased
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.402
##  3rd Qu.:1.000
##  Max.   :1.000
```

```
colnames(sales)
```

```
## [1] "User.ID"      "Gender"       "Age"          "AnnualSalary" "Purchased"
```

```
df1<-sales
df1<-select(df1,Purchased, User.ID, Gender, Age, AnnualSalary, Purchased)
df1
```

Purchased <int>	User.ID <int>	Gender <chr>	Age <int>	AnnualSalary <int>
0	385	Male	35	20000
0	681	Male	40	43500
0	353	Male	49	74000
1	895	Male	40	107500
0	661	Male	25	79000
1	846	Female	47	33500
1	219	Female	46	132500
0	588	Male	42	64000
0	85	Female	30	84500
0	465	Male	41	52000

1-10 of 1,000 rows Previous **1** 2 3 4 5 6 … 100 Next

```
df1 %>% drop_na()
```

Purchased <int>	User.ID <int>	Gender <chr>	Age <int>	AnnualSalary <int>
0	385	Male	35	20000
0	681	Male	40	43500
0	353	Male	49	74000
1	895	Male	40	107500
0	661	Male	25	79000
1	846	Female	47	33500
1	219	Female	46	132500
0	588	Male	42	64000
0	85	Female	30	84500

Purchased <int>	User.ID <int>	Gender <chr>	Age <int>	AnnualSalary <int>
0	465	Male	41	52000

1-10 of 1,000 rows Previous **1** 2 3 4 5 6 … 100 Next

```
str(df1)
```

```
## 'data.frame':    1000 obs. of  5 variables:
##  $ Purchased   : int  0 0 0 1 0 1 1 0 0 0 ...
##  $ User.ID     : int  385 681 353 895 661 846 219 588 85 465 ...
##  $ Gender      : chr  "Male" "Male" "Male" "Male" ...
##  $ Age         : int  35 40 49 40 25 47 46 42 30 41 ...
##  $ AnnualSalary: int  20000 43500 74000 107500 79000 33500 132500 64000 84500 52000 ...
```

```
df2 <- data.frame(
  Purshase = df1$Purchased=="Yes",
  male = df1$Gender=="Male",
  age=df1$Age,
  salary = df1$AnnualSalary )


#convert to numeric
for(col in colnames(df2)){
  df2[col] <- as.numeric(df2[[col]])
}

head(df2)
```

	Purshase <dbl>	male <dbl>	age <dbl>	salary <dbl>
1	0	1	35	20000
2	0	1	40	43500
3	0	1	49	74000
4	0	1	40	107500
5	0	1	25	79000
6	0	0	47	33500

6 rows

```
apply(df2, 2, var)
```

```
##     Purshase         male          age       salary
## 0.000000e+00 2.499940e-01 1.146414e+02 1.189446e+09
```

```
pr.out <- prcomp(df2[-1],scale = TRUE)
pr.out
```

```
## Standard deviations (1, .., p=3):
## [1] 1.1030650 0.9751905 0.9122781
##
## Rotation (n x k) = (3 x 3):
##               PC1       PC2        PC3
## male   -0.4358527 0.8930603  0.1116949
## age     0.6485078 0.2255723  0.7270178
## salary  0.6240754 0.3893077 -0.6774728
```

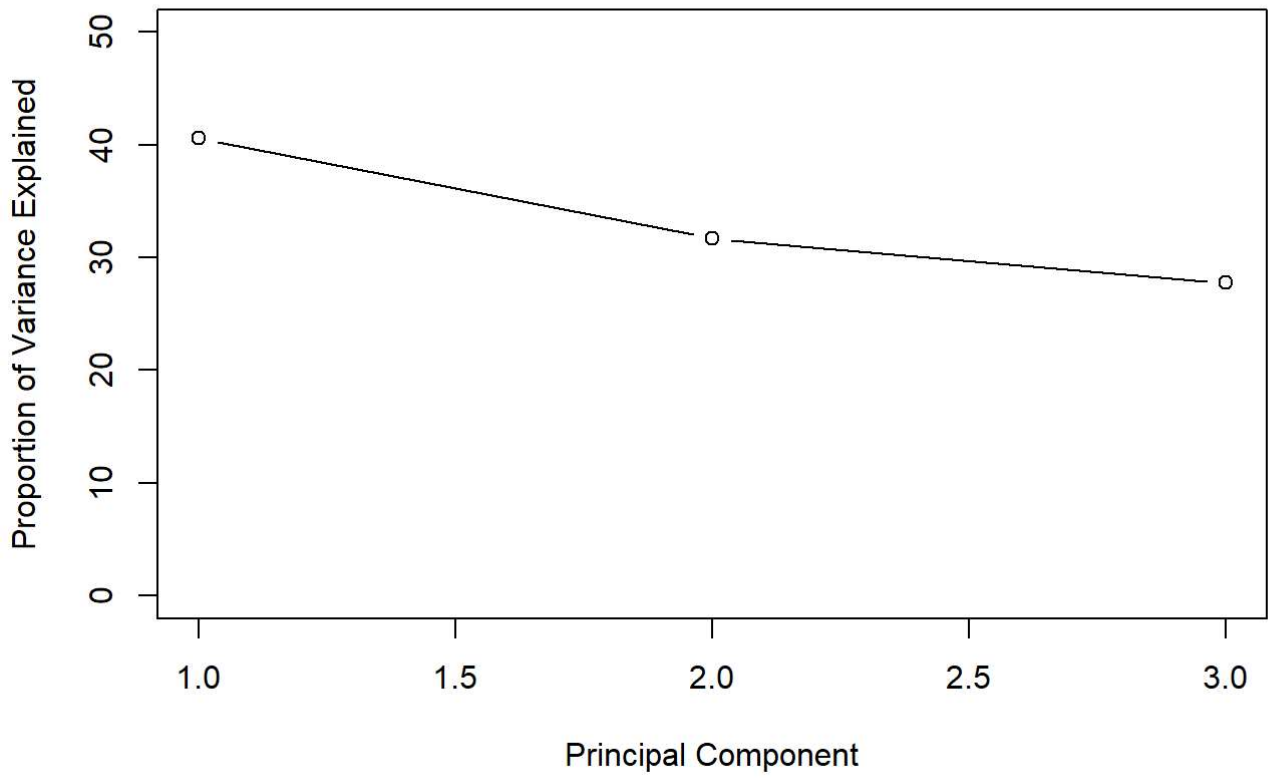
```
pr.var <- pr.out$sdev^2
pr.var
```

```
## [1] 1.2167523 0.9509964 0.8322513
```

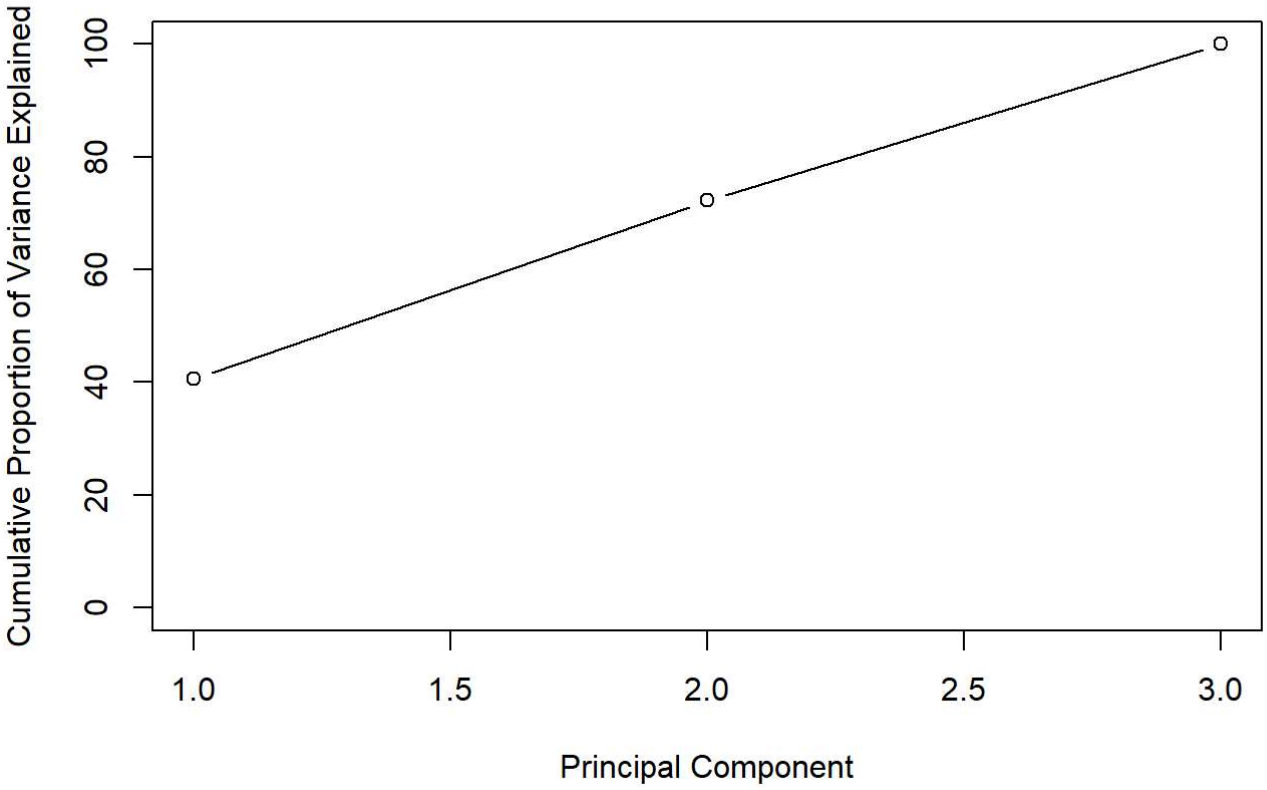
```
pve <- 100 * pr.var / sum(pr.var)
pve
```

```
## [1] 40.55841 31.69988 27.74171
```

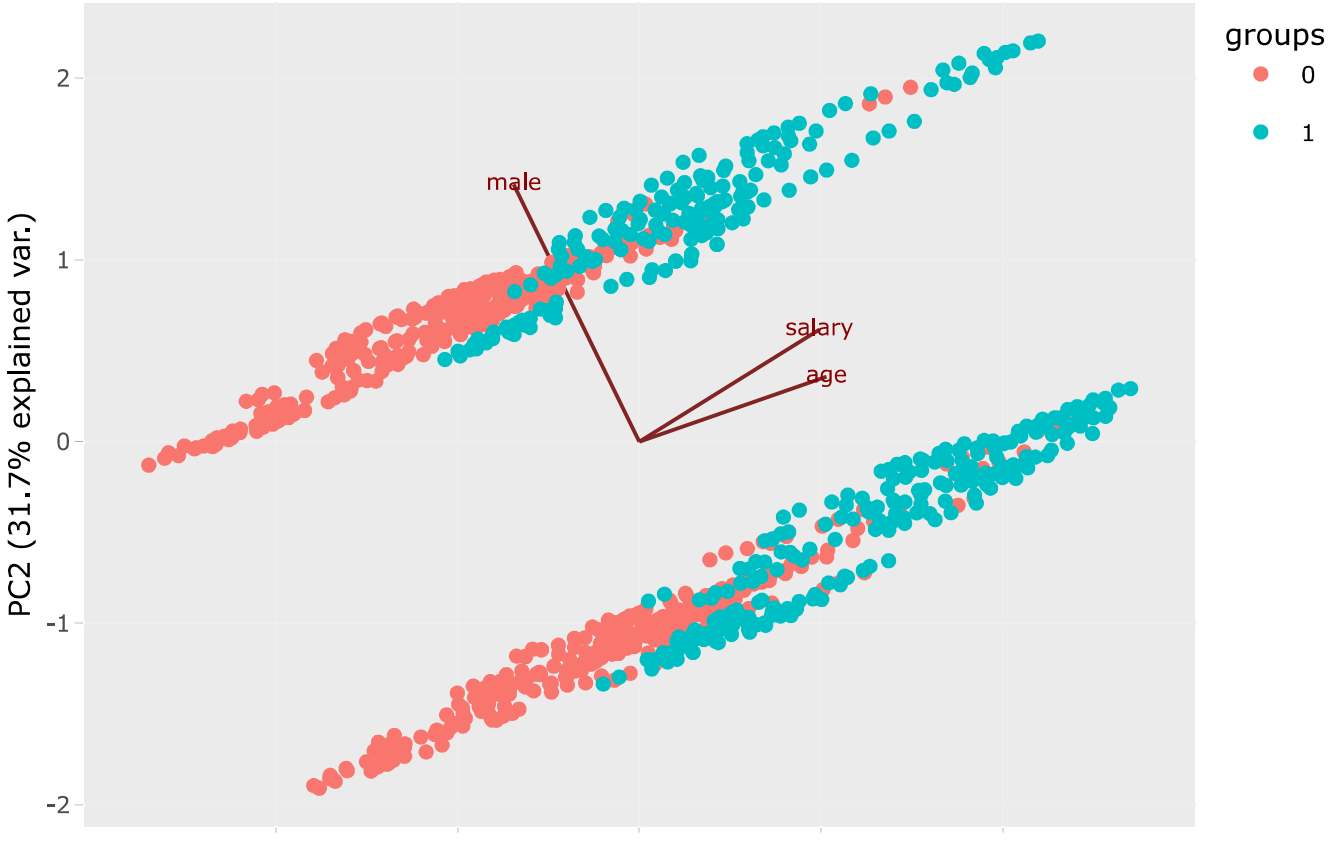
```
plot(pve, xlab="Principal Component", ylab="Proportion of Variance Explained", ylim=c(0,50),type
='b')
```

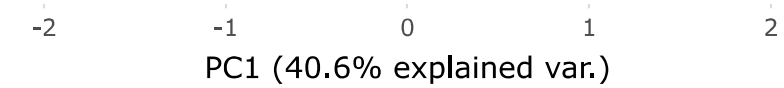



```
plot(cumsum(pve), xlab="Principal Component", ylab="Cumulative Proportion of Variance Explained"
, ylim=c(0,100),type='b')
```




```
ggplotly(ggbiplot(pr.out, scale = 0, groups=factor(df1$Purchase)))
```





The dataset is from kaggle and use to find if a person will purchase a car or not.
From the scree plot we can tell that PC1 explains 40% of the variance and PC2 explains 31% of th e variance. So, in all we two PC's we were able to find tha variability in more than 70% of the dataset.
In the biplot we can see that the green dots are for people who purchased a car and red for peop le who did not. We can tell that
the most variability is shown by the feature gender, but rest of the loading vectors are almost orthogonal to it. at PC1 it shows
-.43 variability and at PC2 it shows .89 variability. Gender here has been the best feature at t he 2 PC's for variability.