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

> hint: take a look at source or load functions there is also smoking.csv file for a refference

```r
source("smoking.R")
```

```r
# load libraries
library(tibble)
library(readr)
library(dplyr)
library(broom)
library(ggplot2)
library(ggbiplot)
library(fastDummies)
library(plotly)
```

```r
# Load data
data1 = data.frame(source("smoking.R"))
```

Take a look into data

```r
head(data1)
```

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

```
data2 = data1[, c("value.smoke", "value.gender", "value.age", "value.marital_status",
"value.highest_qualification", "value.gross_income")]
```

```
data3 = na.omit(data2)
```

```
unique(data3$value.marital_status)
```

```
## [1] Divorced  Single    Married   Widowed   Separated
## Levels: Divorced Married Separated Single Widowed
```

```
unique(data3$value.gross_income)
```

```
##  [1] 2,600 to 5,200   Under 2,600      28,600 to 36,400 10,400 to 15,600
##  [5] 15,600 to 20,800 Above 36,400     5,200 to 10,400  Refused
##  [9] 20,800 to 28,600 Unknown
## 10 Levels: 10,400 to 15,600 15,600 to 20,800 ... Unknown
```

```
unique(data3$value.highest_qualification)
```

```
## [1] No Qualification  Degree            GCSE/O Level      GCSE/CSE
## [5] Other/Sub Degree  Higher/Sub Degree ONC/BTEC          A Levels
## 8 Levels: A Levels Degree GCSE/CSE GCSE/O Level ... Other/Sub Degree
```

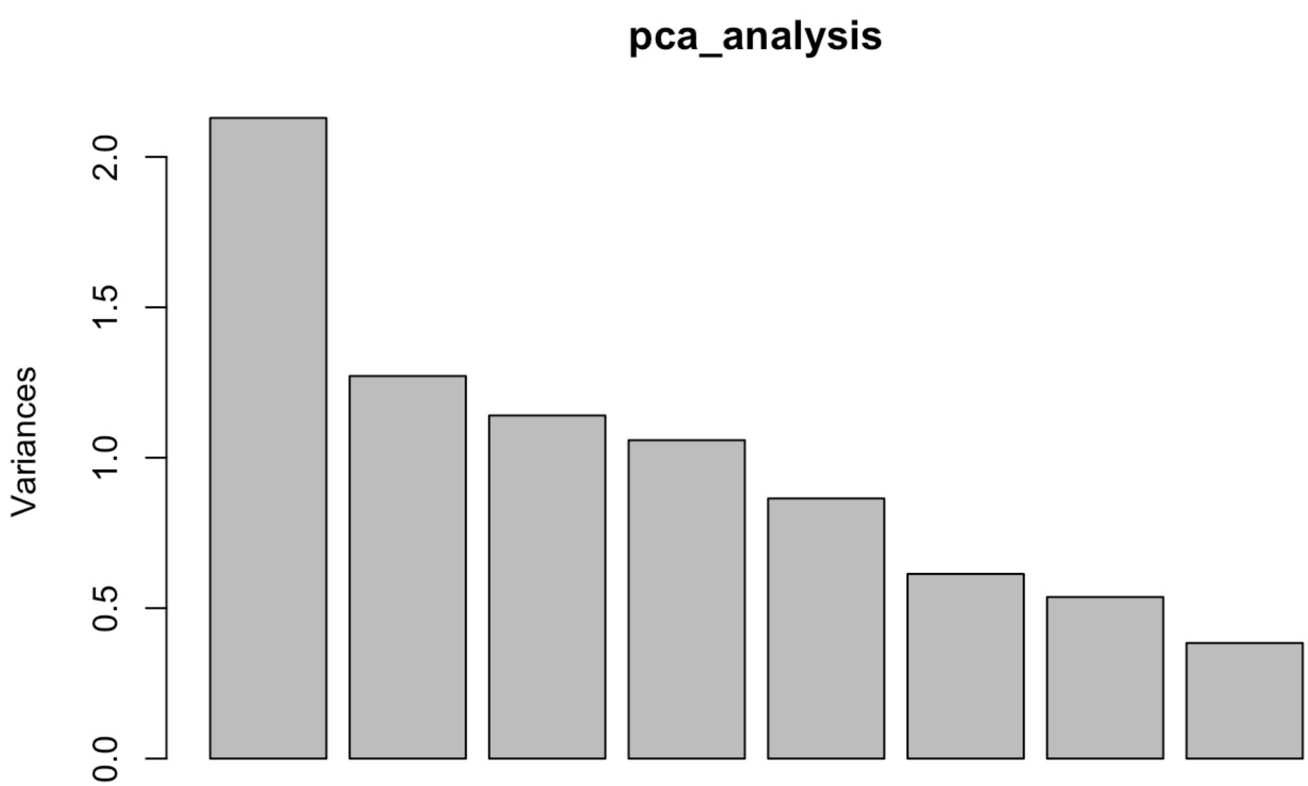
```
data3$value.gender = as.numeric(data3$value.gender == "Female")
data3$value.smoke = as.numeric(data3$value.smoke == "No")
data3 = data3 %>%
  mutate(
    value.highest_qualification = case_when(
      value.highest_qualification == "No Qualification" ~ 1,
      value.highest_qualification == "GCSE/O Level" ~ 2,
      value.highest_qualification == "GCSE/CSE" ~ 3,
      value.highest_qualification == "Other/Sub Degree" ~ 4,
      value.highest_qualification == "Higher/Sub Degree" ~ 5,
      value.highest_qualification == "ONC/BTEC" ~ 6,
      value.highest_qualification ==  "A Levels" ~ 7,
      value.highest_qualification == "Degree" ~ 8,
      TRUE ~ NA
    )
  )

data3 = data3 %>%
  mutate(
    value.gross_income = case_when(
      grepl("^Unknown", value.gross_income) ~ 1,
      grepl("^Under", value.gross_income) ~ 2,
      grepl("^2,600 to 5,200", value.gross_income) ~ 3,
      grepl("^5,200 to 10,400", value.gross_income) ~ 4,
      grepl("^10,400 to 15,600", value.gross_income) ~ 5,
      grepl("^15,600 to 20,800", value.gross_income) ~ 6,
      grepl("^28,600 to 36,400", value.gross_income) ~ 7,
      grepl("^Above", value.gross_income) ~ 8,
      grepl("^Refused", value.gross_income) ~ 9,
      grepl("^20,800 to 28,600", value.gross_income) ~ 10,
      TRUE ~ NA_integer_
    )
  )
data4 = dummy_cols(data3, select_columns = 'value.marital_status')
```

PCA on all columns except smoking status

```
data5 = data4 %>%select(-value.smoke, -value.marital_status, -value.marital_status_Ma
rried)
pca_analysis = prcomp(data5, scale = T)
summary(pca_analysis)
```

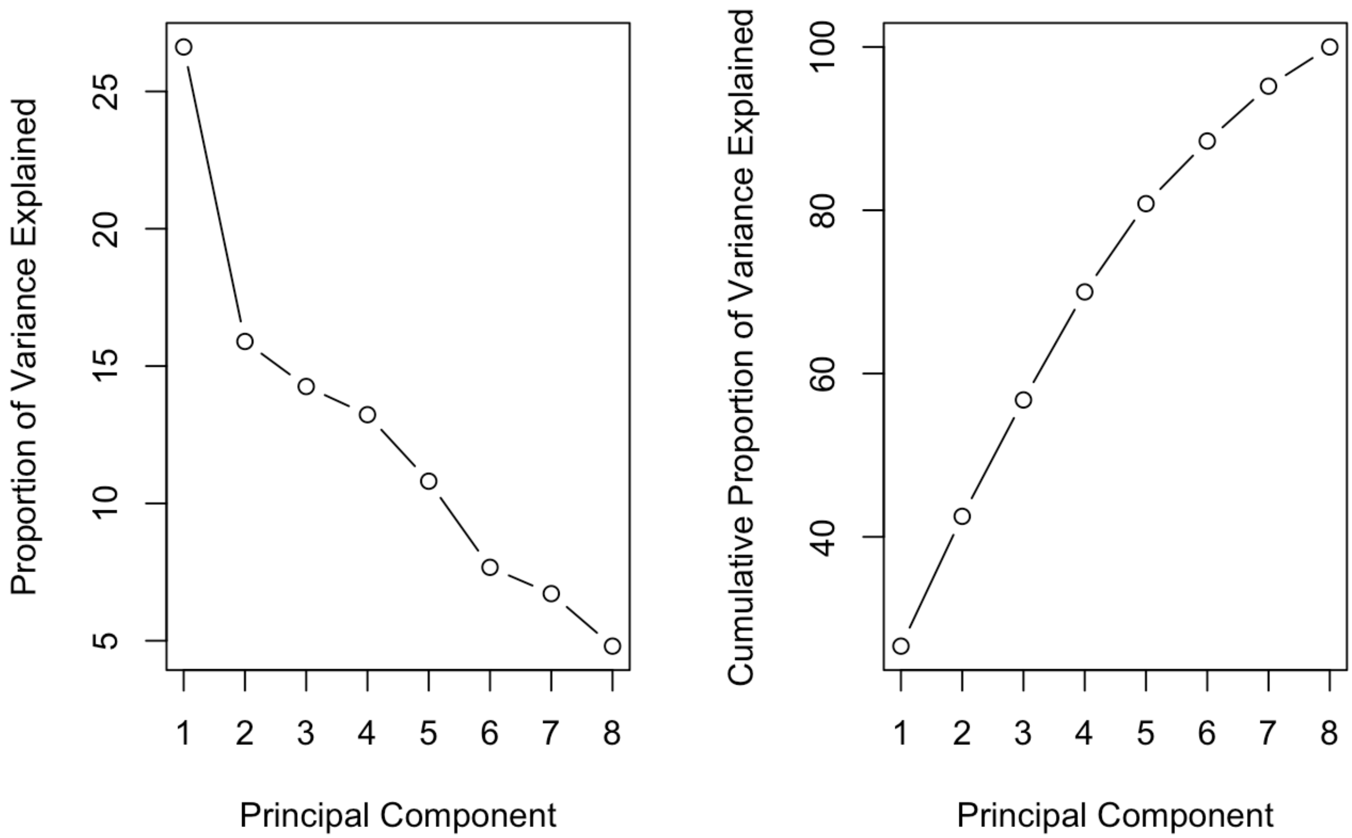
```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     1.4593 1.1277 1.0679 1.0288 0.9299 0.78347 0.73290
## Proportion of Variance 0.2662 0.1590 0.1426 0.1323 0.1081 0.07673 0.06714
## Cumulative Proportion  0.2662 0.4251 0.5677 0.7000 0.8081 0.88482 0.95197
##                            PC8
## Standard deviation     0.61989
## Proportion of Variance 0.04803
## Cumulative Proportion  1.00000
```

```
plot(pca_analysis)
```



scree plot

```
pr.var = pca_analysis$sdev^2
pve = 100 * pr.var / sum(pr.var)
par(mfrow = c(1, 2))
plot(pve, xlab = "Principal Component",
    ylab = "Proportion of Variance Explained",
    type = "b")
plot(cumsum(pve), xlab = "Principal Component",
    ylab = "Cumulative Proportion of Variance Explained",
     type = "b")
```



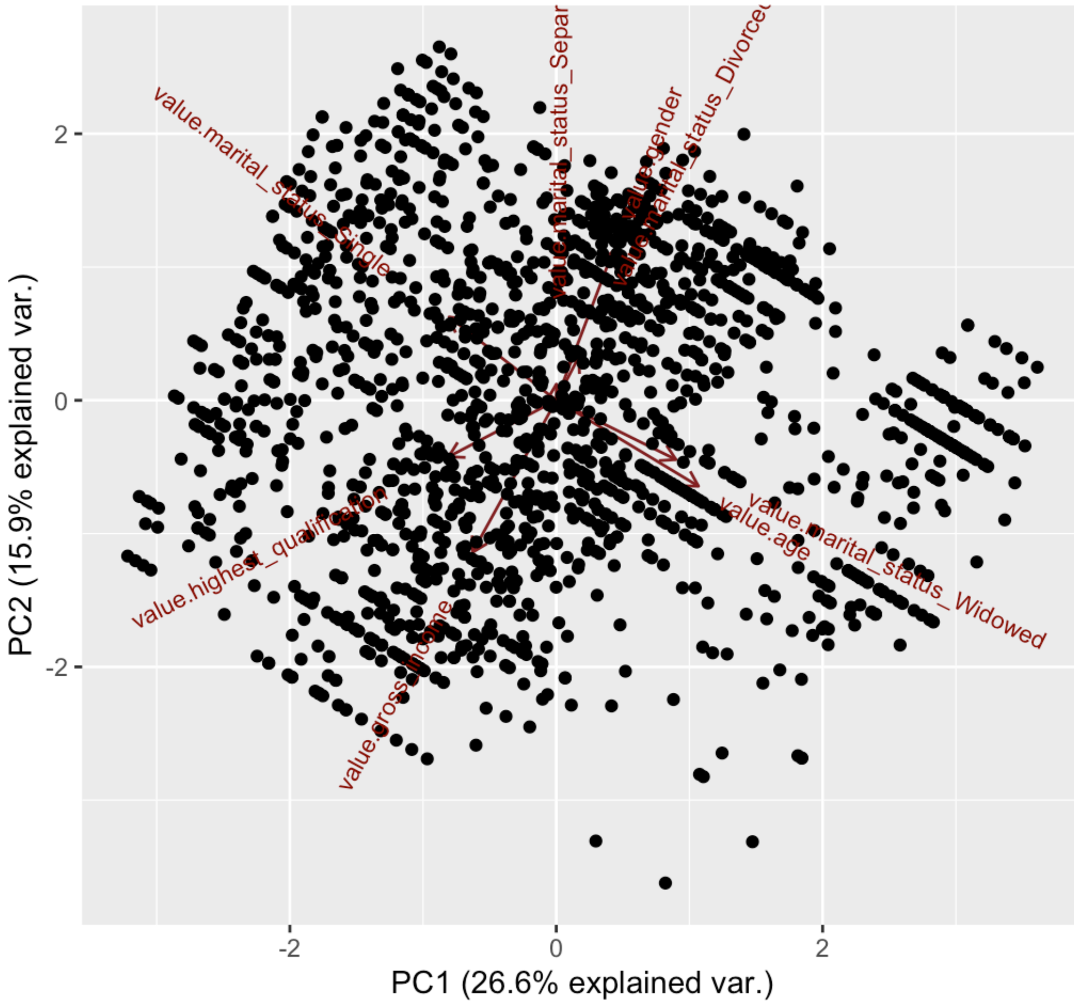
6-

elbow method choice.

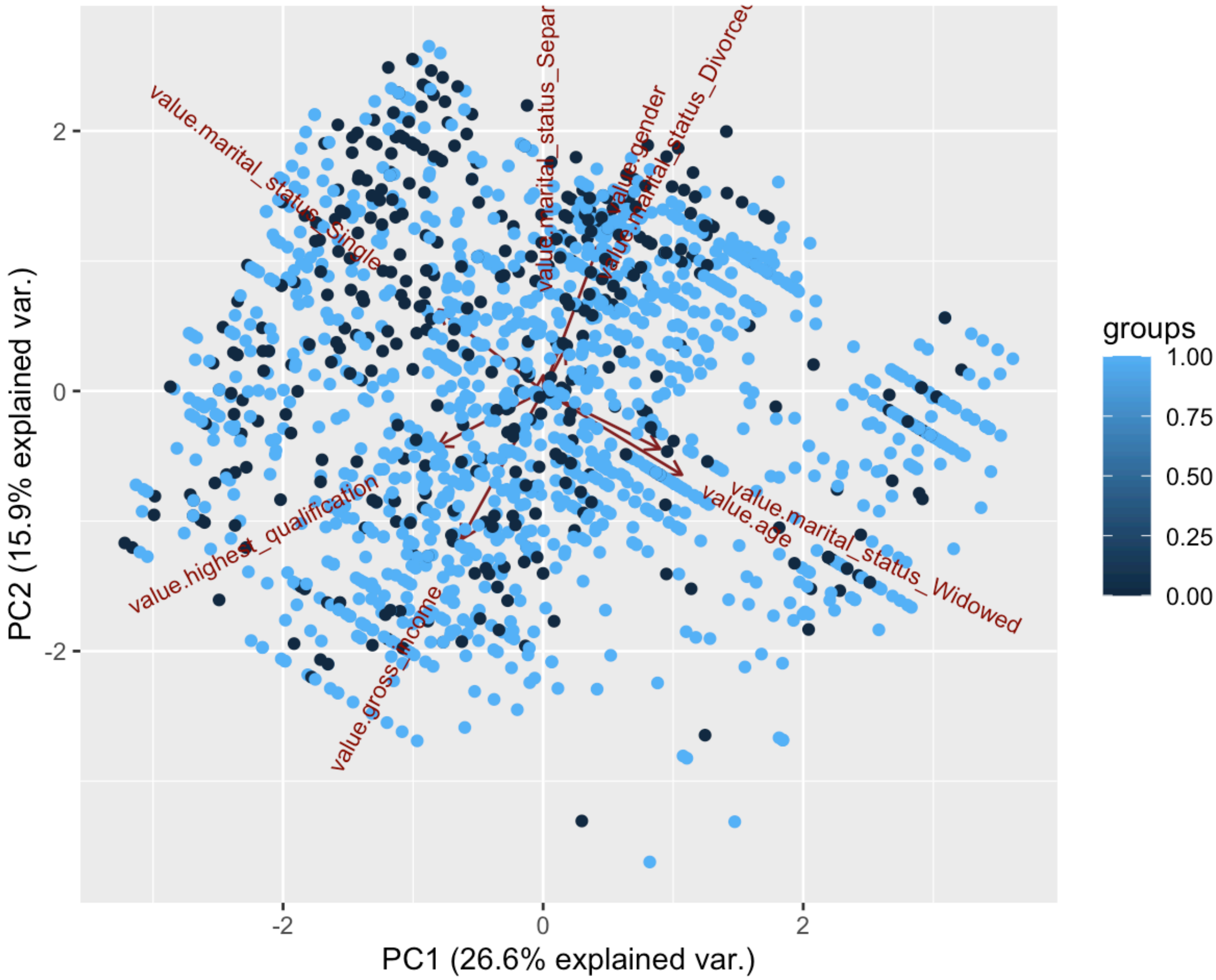
We have 8 PCA from which we can decide how many captures the most variance in data. By elbow method, we can decide the 6PCA will be good choice. Retaining the information and reducing the dimensionality. But if we have to capture 90+% of variance then we have to keep till PC7 which again depends on the situation and needs. Here, pc1 - pc6 will be a good choice as it covers 88% of data variance.

biplot color points by smoking field

```
# biplot without smoking field
ggbiplot(pca_analysis, scale = 0, labels=rownames(pca_analysis$x))
```




```
ggbiplot(pca_analysis, scale = 0, labels=rownames(pca_analysis$x), groups = data3$val
ue.smoke)
```



The principle component analysis (PCA) observed biplot explains on the connections between the original variables and the principal components.Here we have 8 PCA and 8 features builds this biplot explaining us the contributions to each PC by examining the arrows that represent those features.

Notably, PC1 and PC2 seem to be the most significant factors, as shown by the long arrows connected to the values of “value.age,” “value.gender,” and “value.marital_status_Single” for PC1 and “value.gross_income” and “value.marital_status_Separated” for PC2, respectively.

Closely grouped data points around these arrows imply that specific subsets of the dataset have shared traits relating to these key factors. who are older and have never been married may group together.

Additionally, the division of observations along PC1 and PC2 raises the possibility of a possible differentiation in the data depending on elements like age and marital status. ‘value.highest_qualification’ stands out as a significant contribution to PC3, whereas PC4 is less significant but still helps to comprehend the data structure.

Finally, I can conclude the disussion by saying the gender, marital is not correlated with the gross income, while they both are correlated eliminating one doesn’t matter in this context. Whereas considering age and as well as the widow is interestingly highly correlated from the biplot which has been plotted and it is in the same direction as in PCA1 where 26% of variance is being explained.

we cannot use first two PC to discriminate between smoking as I can see from the first two PCs that the data based on smoke is not accurately categorised. The biplot generated using the first two PCs shows the contribution of the initial factors to these PCs. The scores on the first two PCs determine the locations of the dots in the scatter plot, each of which represents an observation. It was also possible to qualitatively assess if the first two PCs can discriminate between smokers and non-smokers by colouring the scatter plot points according to whether or not they are smokers.And also, the contrubution of first 2 pc is just 41%

```
unique(data3$value.marital_status)
```

```
## [1] Divorced  Single    Married   Widowed   Separated
## Levels: Divorced Married Separated Single Widowed
```

```
data3$value.marital_status = factor(data3$value.marital_status, levels = unique(data3
$value.marital_status), labels = c(4L, 2L, 1L, 5L, 3L), ordered = TRUE)
```

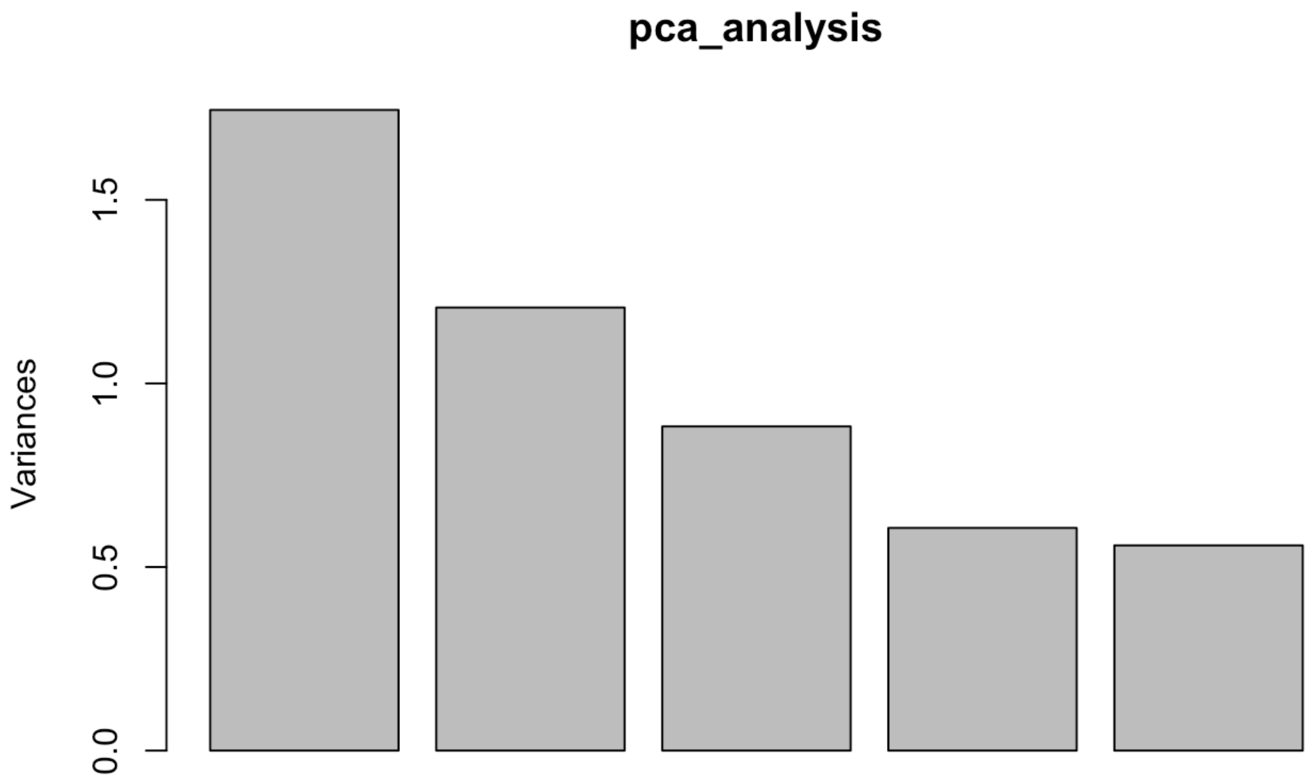
```
data3$value.highest_qualification = as.numeric(data3$value.highest_qualification)
data3$value.marital_status= as.numeric(data3$value.marital_status)
sapply(data3,class)
```

```
##                value.smoke               value.gender
##                  "numeric"                  "numeric"
##                  value.age       value.marital_status
##                  "integer"                  "numeric"
## value.highest_qualification         value.gross_income
##                  "numeric"                  "numeric"
```

```
data_revisit <- data3 %>%select(-value.smoke)
pca_analysis <- prcomp(data_revisit, scale = T)
summary(pca_analysis)
```

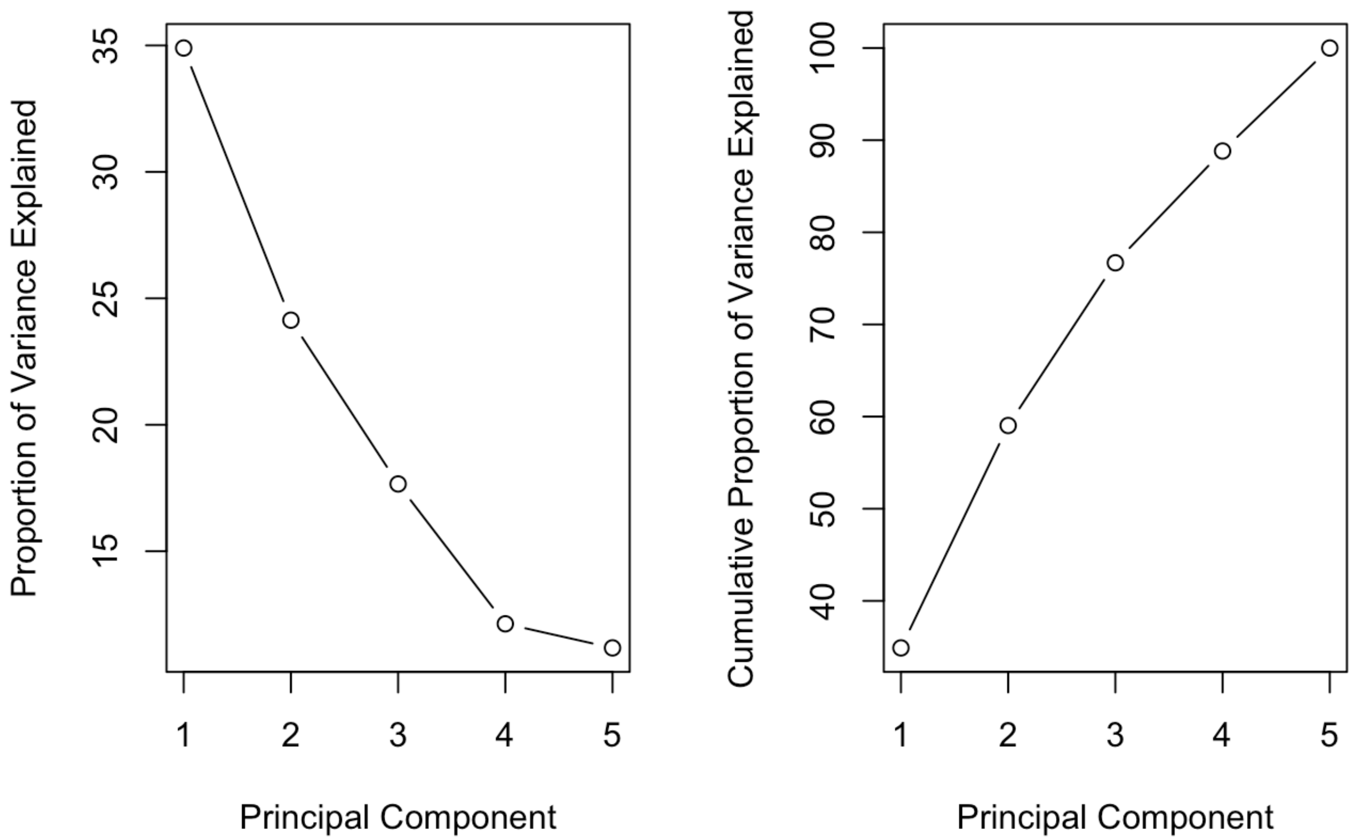
```
## Importance of components:
##                          PC1    PC2    PC3    PC4    PC5
## Standard deviation     1.321 1.0985 0.9397 0.7788 0.7476
## Proportion of Variance 0.349 0.2414 0.1766 0.1213 0.1118
## Cumulative Proportion  0.349 0.5903 0.7669 0.8882 1.0000
```

```
plot(pca_analysis)
```



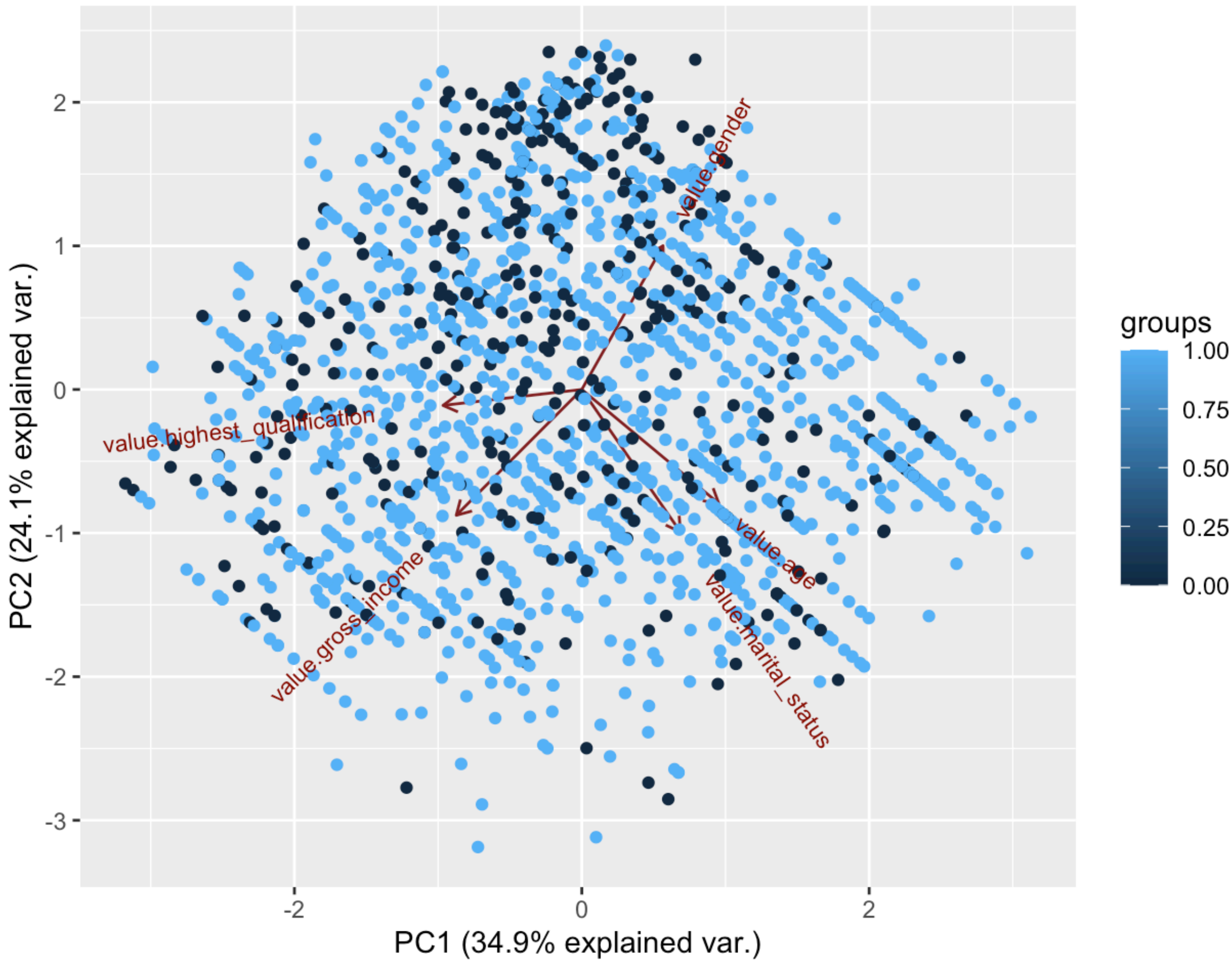
skee plot

```
pr.var = pca_analysis$sdev^2
pve <- 100 * pr.var / sum(pr.var)
par(mfrow = c(1, 2))
plot(pve, xlab = "Principal Component",
    ylab = "Proportion of Variance Explained",
    type = "b")
plot(cumsum(pve), xlab = "Principal Component",
    ylab = "Cumulative Proportion of Variance Explained",
     type = "b")
```



biplot

```
ggbiplot(pca_analysis, scale = 0, groups = data3$value.smoke)
```



#With the change made using marital column as a ordinal feature with respect to smoking grouping I guess we can see good changes and also, the pc1 and pc2 contrubute 59% of total variance in data.With this changes lesser PC analysis and less features explains the same contrubution.

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)

```
data.loan = read.csv("loan_train.csv")
```

```
head(data.loan)
```

```
##   Gender Married Dependents    Education Self_Employed Applicant_Income
## 1   Male      No          0     Graduate            No           584900
## 2   Male     Yes          1     Graduate            No           458300
## 3   Male     Yes          0     Graduate           Yes           300000
## 4   Male     Yes          0 Not Graduate            No           258300
## 5   Male      No          0     Graduate            No           600000
## 6   Male     Yes          2     Graduate           Yes           541700
##   Coapplicant_Income Loan_Amount Term Credit_History  Area Status
## 1                  0    15000000  360              1 Urban      Y
## 2             150800    12800000  360              1 Rural      N
## 3                  0     6600000  360              1 Urban      Y
## 4             235800    12000000  360              1 Urban      Y
## 5                  0    14100000  360              1 Urban      Y
## 6             419600    26700000  360              1 Urban      Y
```

```
data.loan1 <- data.loan[,c("Gender", "Married", "Education","Self_Employed","Applican
t_Income", "Loan_Amount","Area","Status")]
```

```
str(data.loan1)
```

```
## 'data.frame':    614 obs. of  8 variables:
##  $ Gender          : chr  "Male" "Male" "Male" "Male" ...
##  $ Married         : chr  "No" "Yes" "Yes" "Yes" ...
##  $ Education       : chr  "Graduate" "Graduate" "Graduate" "Not Graduate" ...
##  $ Self_Employed   : chr  "No" "No" "Yes" "No" ...
##  $ Applicant_Income: int  584900 458300 300000 258300 600000 541700 233300 303600
400600 1284100 ...
##  $ Loan_Amount     : int  15000000 12800000 6600000 12000000 14100000 26700000 950
0000 15800000 16800000 34900000 ...
##  $ Area            : chr  "Urban" "Rural" "Urban" "Urban" ...
##  $ Status          : chr  "Y" "N" "Y" "Y" ...
```

#* Convert a columns to proper format (9 points) #PRE-PROCESSING #Converting the gender and married status to numeric

```
data.loan1$Gender <- as.numeric(data.loan1$Gender == "Male")
data.loan1$Married <- as.numeric(data.loan1$Married == "Yes")
data.loan1$Education <- as.numeric(data.loan1$Education == "Graduate")
data.loan1$Self_Employed <- as.numeric(data.loan1$Self_Employed == "Yes")
data.loan1$Status <- as.numeric(data.loan1$Status == "Y")
```

```
data.loan1$Area = factor(data.loan1$Area, levels = unique(data.loan1$Area), labels =
c(3L, 1L, 2L), ordered = TRUE)
```

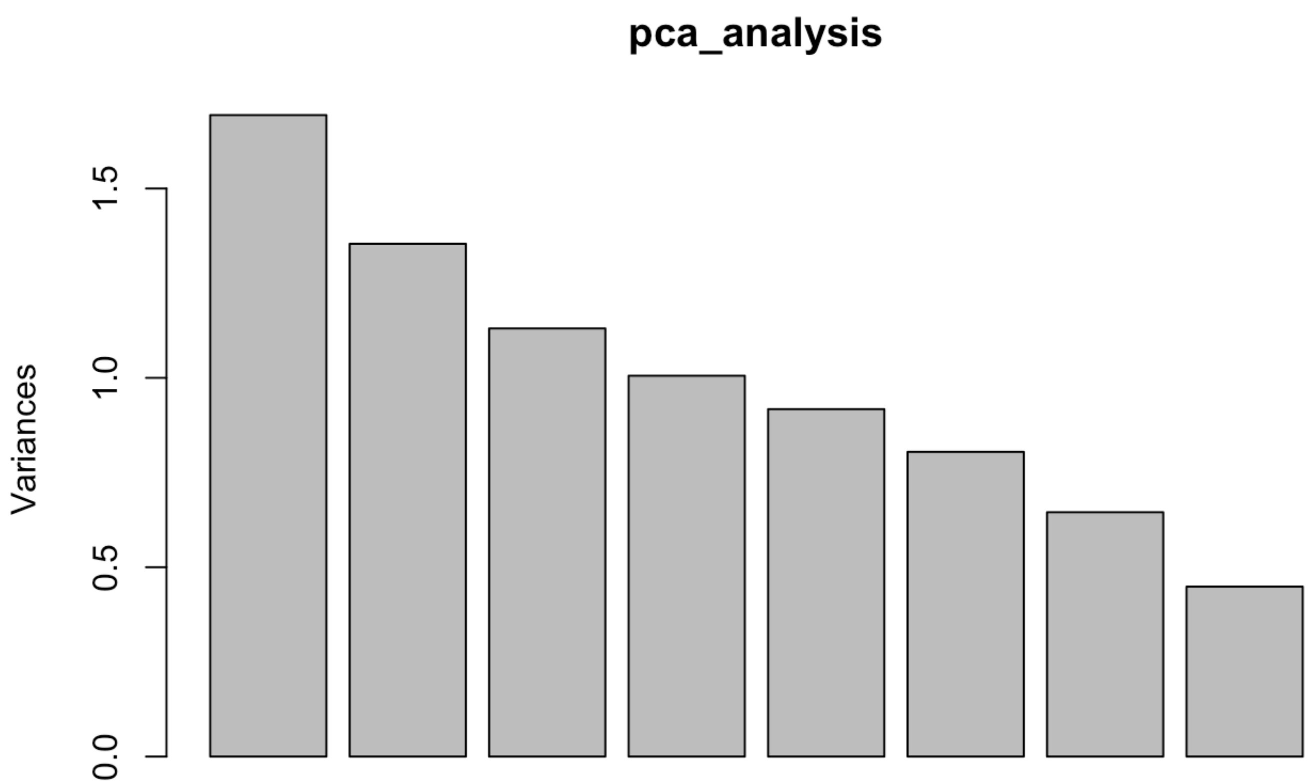
```
data.loan1$Area = as.numeric(data.loan1$Area)
sapply(data.loan1, class)
```

```
##           Gender          Married        Education    Self_Employed
##        "numeric"        "numeric"        "numeric"        "numeric"
## Applicant_Income      Loan_Amount             Area           Status
##        "integer"        "integer"        "numeric"        "numeric"
```

```
pca_analysis <- prcomp(data.loan1, scale = T)
summary(pca_analysis)
```

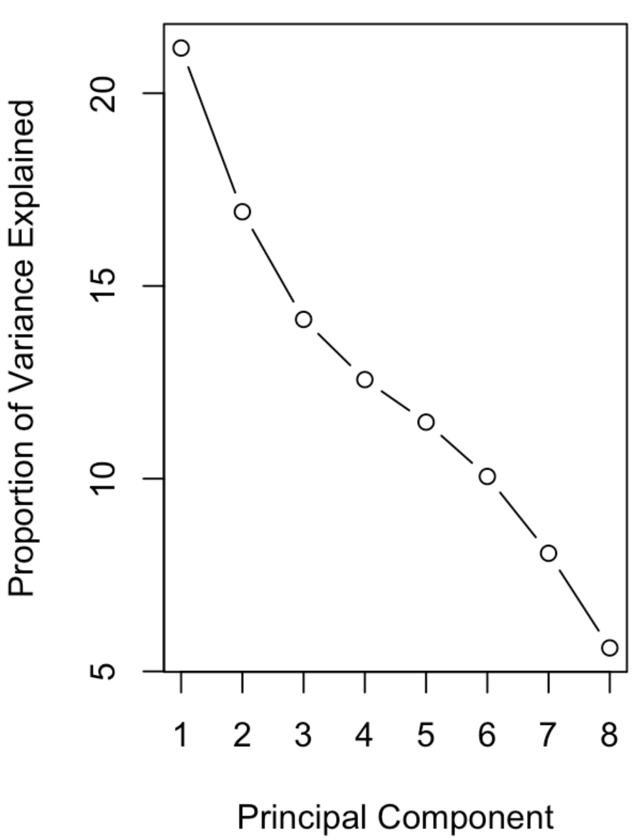
```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5    PC6     PC7    PC8
## Standard deviation     1.3015 1.1636 1.0633 1.0028 0.9578 0.8970 0.80323 0.6699
## Proportion of Variance 0.2117 0.1693 0.1413 0.1257 0.1147 0.1006 0.08065 0.0561
## Cumulative Proportion  0.2117 0.3810 0.5223 0.6480 0.7627 0.8633 0.94390 1.0000
```

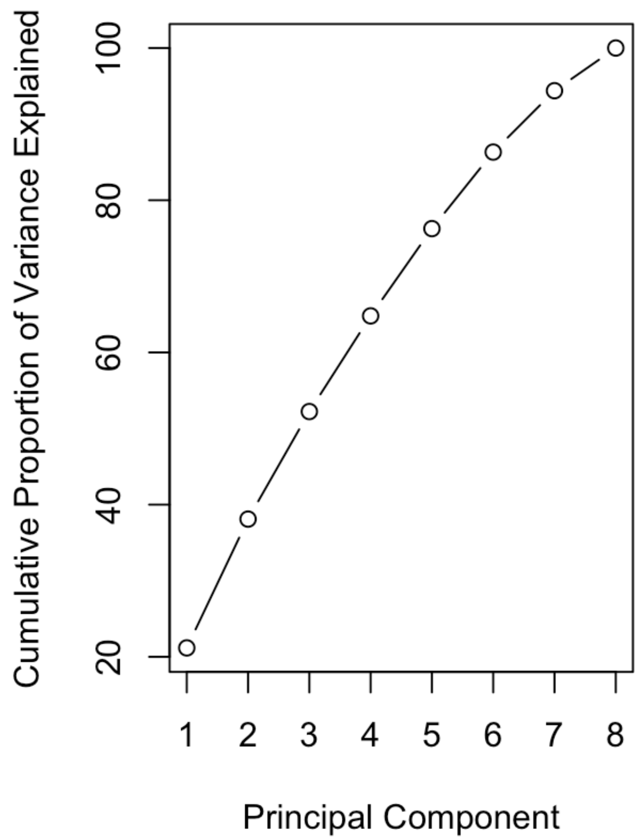
```
plot(pca_analysis)
```



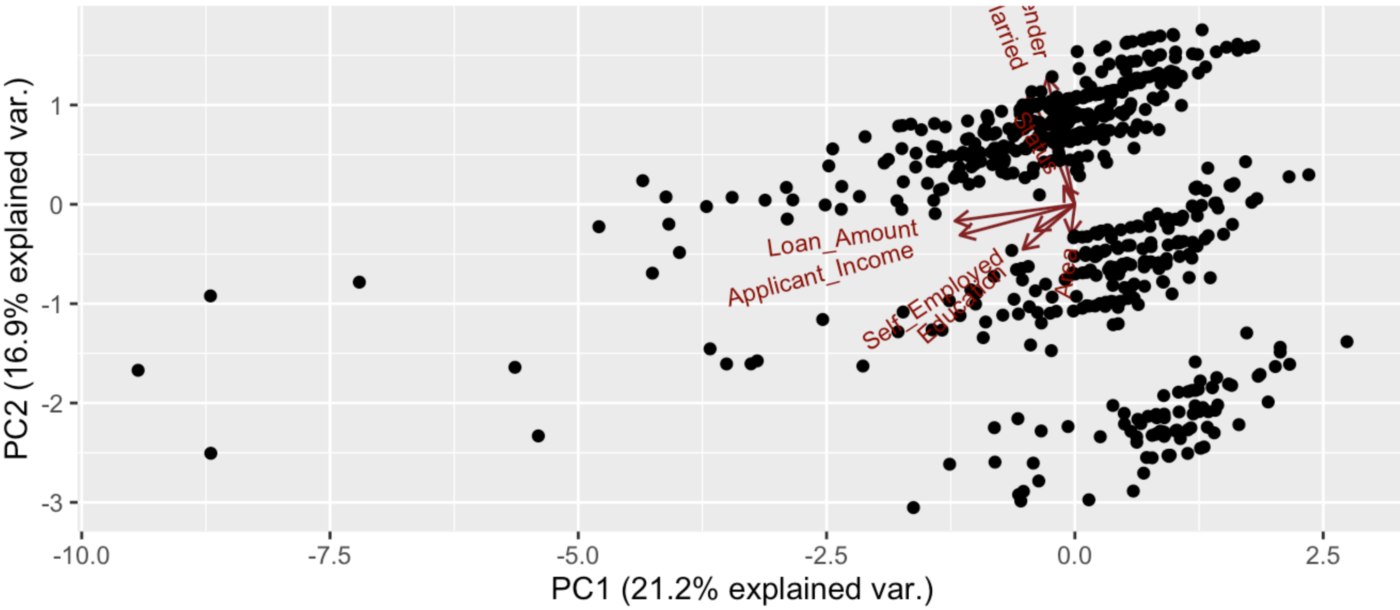
#* Make a skree plot (3 points)

```
pr.var = pca_analysis$sdev^2
pve <- 100 * pr.var / sum(pr.var)
par(mfrow = c(1, 2))
plot(pve, xlab = "Principal Component",
    ylab = "Proportion of Variance Explained",
    type = "b")
plot(cumsum(pve), xlab = "Principal Component",
    ylab = "Cumulative Proportion of Variance Explained",
     type = "b")
```






```
ggbiplot(pca_analysis, scale = 0, labels=rownames(pca_analysis$x))
```


PC2 (16.9% explained var.)
PC1 (21.2% explained var.)
1
0
-1
-2
-3
-10.0
-7.5
-5.0
-2.5
0.0
2.5
Loan_Amount
Applicant_Income
Self_Employed
Education
Status
Area