Project Report

Data Analysis & Visualization

INFX 512

University of Louisiana at Lafayette

Name: Jamal Ahmadov

School ID: C00423516

Submission date: May 2, 2020

Contents

Dataset

This data set contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. The data set was obtained from the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Census+Income). The initial data set had 48,842 instances and 14 attributes that are both categorical and numeric. Of those 48,842 instances, 3,620 of them contain missing values. To have a consistent analysis, they are removed from the data set, leaving 45,222 instances. Table 1 lists the variable (column) names appearing in dataset and their corresponding descriptions.

Table 1. Features of data set and their descriptions.

age	Age of the worker
workclass	Class of worker (e.g. state government, never worked)
education_level	Level of education (e.g. 11th grade, Masters degree)
education.num	Number of years of education
marital.status	Marital status (e.g. never married, divorced)
occupation	Major occupation (e.g. tech-support, sales, craft-repair)
relationship	Relationship (e.g. wife, own-child, husband)
race	Race (e.g. white, black)
sex	Sex (i.e. male or female)
capital.gain	Capital gains
capital.loss	Capital losses
hours.per.week	Hours worked per week
native.country	Native country
income	Annual income (i.e. above $50k or below $50k)

The analysis will be focused on determining variables that affect the income of individuals most. Based on these features, the prediction model will be developed to determine if an individual makes over $50k a year.

First few lines of dataset were obtained through the "head" command (Figure 1).

```
> head(censusdata)
  age         workclass education_level education.num      marital.status         occupation   relationship   race
1  39         State-gov       Bachelors            13       Never-married       Adm-clerical  Not-in-family  White
2  50  Self-emp-not-inc       Bachelors            13  Married-civ-spouse    Exec-managerial        Husband  White
3  38           Private         HS-grad             9            Divorced  Handlers-cleaners  Not-in-family  White
4  53           Private            11th             7  Married-civ-spouse  Handlers-cleaners        Husband  Black
5  28           Private       Bachelors            13  Married-civ-spouse     Prof-specialty           Wife  Black
6  37           Private         Masters            14  Married-civ-spouse    Exec-managerial           Wife  White
      sex capital.gain capital.loss hours.per.week native.country income
1    Male         2174            0             40  United-States  <=50K
2    Male            0            0             13  United-States  <=50K
3    Male            0            0             40  United-States  <=50K
4    Male            0            0             40  United-States  <=50K
5  Female            0            0             40           Cuba  <=50K
6  Female            0            0             40  United-States  <=50K
```

Figure 1. The snapshot of the dataset.

Analysis

In Table 2, the summary statistics of continuous variables were provided using summary() function in R. The insights about average working population can be obtained by analyzing the range between 1st and 3rd quartile (i.e. 50 % of the data). Within this range, the age of individuals is 28-47 while the weekly hours they work are 40-45. The education number ranges from 9 to 13 which corresponds to high school and bachelor's degrees, respectively. The capital gain and capital loss are 0 for most of the individuals, only having positive values in the last quartile.

Table 2. Summary statistics of continuous variables.

	Age	Education.num	Capital gain	Capital.loss	Hours.per.week
Minimum	17	1	0	0	1
1st quartile	28	9	0	0	40
Median	37	10	0	0	40
Mean	38.6	10.1	1101	89	40.9
3rd quartile	47	13	0	0	45
Maximum	90	16	99999	4356	99

The percentage of different groups in each category are given in Table 3. 74 % of individuals are employed in the private sector. The most famous occupations are administrative/clerical, executive/managerial, sales and crafts/repair. The third of individuals are high school graduates (33%) followed by people with some college experience (22%), bachelor's degree (17%) and master's degree (6%). 47 % are married while 14 % are divorced and 32 % have never been

married. In terms of racial classification, white and black races constitute 86 % and 9 % of the data set followed by a small percentage of other races. 68 % of individuals are male and 32 % are female. The native country of majority is United States (91%) followed by Mexico (3%) and Philippines (1%). Only 25 % of individuals earn more than $50,000 a year.

Table 3. Percentage of different groups in each category.

Variable	Percentage of groups
workclass	Private (74%), Self-emp-not-inc (8%), Local-gov (7%), Self-emp-inc (4%), State-gov (4%), Federal-gov (3%), Without-pay (0%).
education.level	HS-grad (33%), Some-college (22%), Bachelors (17%), Masters (6%), 11th (4%), Assoc-voc (4%), 10th (3%), Assoc-acdm (3%), Prof-school (2%), 7th-8th (2%), Doctorate (1%), 5th-6th (1%), 9th (1%), 12th (1%), 1st-4th (0%), Preschool (0%).
marital.status	Married-civ-spouse (47%), Never-married (32%), Divorced (14%), Separated (3%), Widowed (3%), Married-spouse-absent (1%), Married-AF-spouse (0%).
occupation	Exec-managerial (13%), Craft-repair (13%), Prof-specialty (13%), Sales (12%), Adm-clerical (12%), Other-service (11%), Machine-op-inspct (7%), Transport-moving (5%), Handlers-cleaners (5%), Farming-fishing (3%), Tech-support (3%), Protective-serv (2%), Priv-house-serv (1%), Armed-Forces (0%).
relationship	Husband (41%), Not-in-family (26%), Own-child (15%), Unmarried (11%), Wife (5%), Other-relative (3%).
race	White (86%), Black (9%), Asian-Pac-Islander (3%), Amer-Indian-Eskimo (1%), Other (1%).
sex	Male (68%), Female (32%).
native.country	United-States (91%), Mexico (2%), Philippines (1%), Others (6%).
income	<=50K (75%), >50K (25%).

Next, the relationships between the variables will be analyzed. Since there are both continuous and categorical variables, the analysis is divided into 3 parts: continuous vs. continuous, categorical

vs. categorical and continuous vs. categorical. Different techniques will be applied for each part of analysis.

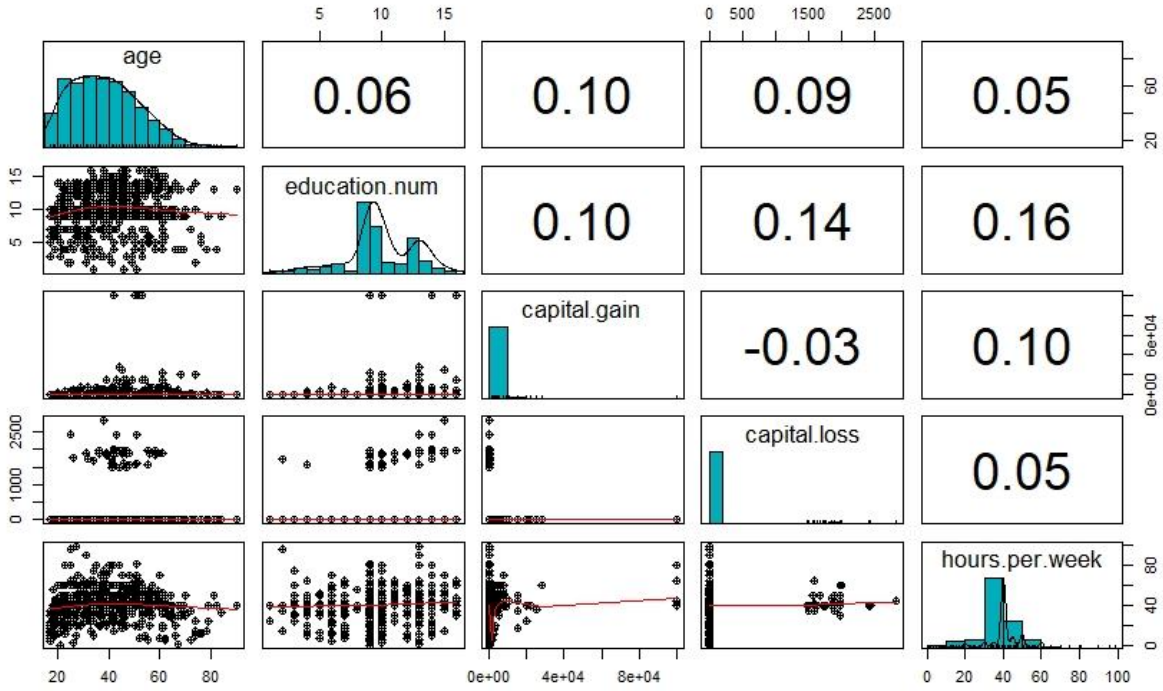
Exploratory analysis

Correlation between continuous variables

Pearson's correlation coefficient was estimated to determine the degree of correlation between the continuous variables. It can be seen that there is a very weak correlation between variables which is shown by low r values (-0.03 - 0.15). As there are many instances in the dataset, 1,000 instances were randomly chosen from data set to generate a pair plot. The Pearson's correlation coefficient was also shown on the plot. As it is seen the r values are similar to the ones obtained from the whole dataset.

```
> r=round(cor(censusdata[,c(1,4,10,11,12)],method = "pearson"),2)
> r
                age education.num capital.gain capital.loss hours.per.week
age            1.00           0.04         0.08         0.06           0.10
education.num  0.04           1.00         0.13         0.08           0.15
capital.gain   0.08           0.13         1.00        -0.03           0.08
capital.loss   0.06           0.08        -0.03         1.00           0.05
hours.per.week 0.10           0.15         0.08         0.05           1.00
```

```
> set.seed(1)
> sampled=sample(nrow(censusdata),size=1000)
> censusdata_sampled=censusdata[sampled,]
> library(psych)
>  pairs.panels(censusdata_sampled[,c(1,4,10,11,12)],method="pearson",hist.col
="#00AFBB",ellipses = FALSE, pch = 10)
```



Dependency between categorical variables

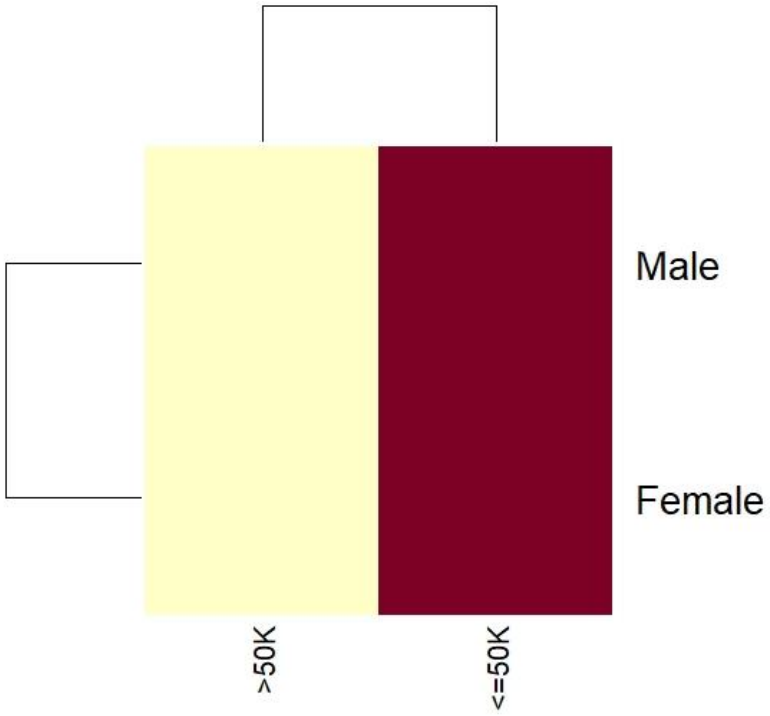
First, we create a contingency table with the table() function that cross-classifies the number of rows that are in the categories specified by the two categorical variables. We also visualize it using a heatmap() function. The Chi-Squared test are then applied to determine if two categorical variables are independent. The null hypothesis with this test is that the two categories are independent. The alternative hypothesis is that there is some dependency between the two categories. As the main focus of analysis is to predict income, we analyze the dependency between the income and other categorical variables.

1. **Sex vs. income**

As it can be seen from the contingency table the proportion of the male individuals who make over $50K a year is greater than the proportion of female individuals in the same category. Also, the p-value obtained from Chi-Squared is very small and we should reject the hypothesis that these variables are independent.

```
> table(censusdata$sex,censusdata$income)

          <=50K  >50K
```

```
Female 13026  1669
Male   20988  9539
```




```
> chisq.test(table(censusdata$sex,censusdata$income))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(censusdata$sex, censusdata$income)
X-squared = 2104.1, df = 1, p-value < 2.2e-16
```

2. Education level vs. income

The higher the education level an individual obtained the larger number of individuals in that education category makes over $50K a year. It is a reasonable observation since the higher level of education is correlated with more qualified and thus, high-paid jobs. Again, the p-value obtained from Chi-Squared is very small and we should reject the hypothesis that these variables are independent.

```
> table(censusdata$education_level,censusdata$income)

              <=50K  >50K
  10th          1141    82
  11th          1530    89
  12th           534    43
```

```
1st-4th         214     8
5th-6th         427    22
7th-8th         768    55
9th             638    38
Assoc-acdm     1109   398
Assoc-voc      1455   504
Bachelors      4392  3178
Doctorate       145   399
HS-grad       12367  2416
Masters        1121  1393
Preschool        71     1
Prof-school     193   592
Some-college   7909  1990
```


```
> chisq.test(table(censusdata$education_level,censusdata$income))

        Pearson's Chi-squared test

data:  table(censusdata$education_level, censusdata$income)
X-squared = 5996, df = 15, p-value < 2.2e-16
```

3. Race vs. income

According to the contingency table, the greater proportion of individuals from a white race makes over $50K a year than other races. The Asian-Pac-Islander group have a similar proportion although the number of individuals in this group is relatively small. Also, the p-value obtained from Chi-Squared test is very small which means that we should reject the hypothesis that these variables are independent.

```
> table(censusdata$race,censusdata$income)

                     <=50K  >50K
   Amer-Indian-Eskimo   382    53
   Asian-Pac-Islander   934   369
   Black               3694   534
   Other                308    45
   White              28696 10207
```

```
> chisq.test(table(censusdata$race,censusdata$income))

        Pearson's Chi-squared test

data:  table(censusdata$race, censusdata$income)
X-squared = 452.3, df = 4, p-value < 2.2e-16
```

4. Marital status vs. income

It can be seen that the larger proportion of married individuals makes over $50K a year than other groups. These individuals have more family needs and responsibilities when compared to the other groups in this category. That is why, they might need to earn more to make ends meet. Again, the p-value obtained from Chi-Squared test is very small which means that we should reject the hypothesis that these variables are independent.

```
> table(censusdata$marital.status,censusdata$income)

                        <=50K  >50K
   Divorced              5642   655
   Married-AF-spouse       18    14
   Married-civ-spouse   11491  9564
   Married-spouse-absent  498    54
   Never-married        13897   701
   Separated             1312    99
```

```
   Widowed                1156    121
```

```
> chisq.test(table(censusdata$marital.status,censusdata$income))

        Pearson's Chi-squared test

data:  table(censusdata$marital.status, censusdata$income)
X-squared = 9109.2, df = 6, p-value < 2.2e-16
```

Continuous vs. categorical variables

In this section, we generate a box plot to visualize a continuous variable together with an income to show how univariate statistics of the continuous variables change with respect to different levels of the income. The mean value of each level was estimated and shown in the plot. The summary statistics of each level was also provided. We use a two-sample t-test to evaluate whether the means of two groups are different. It will also show whether an income is affected by a particular continuous variable.

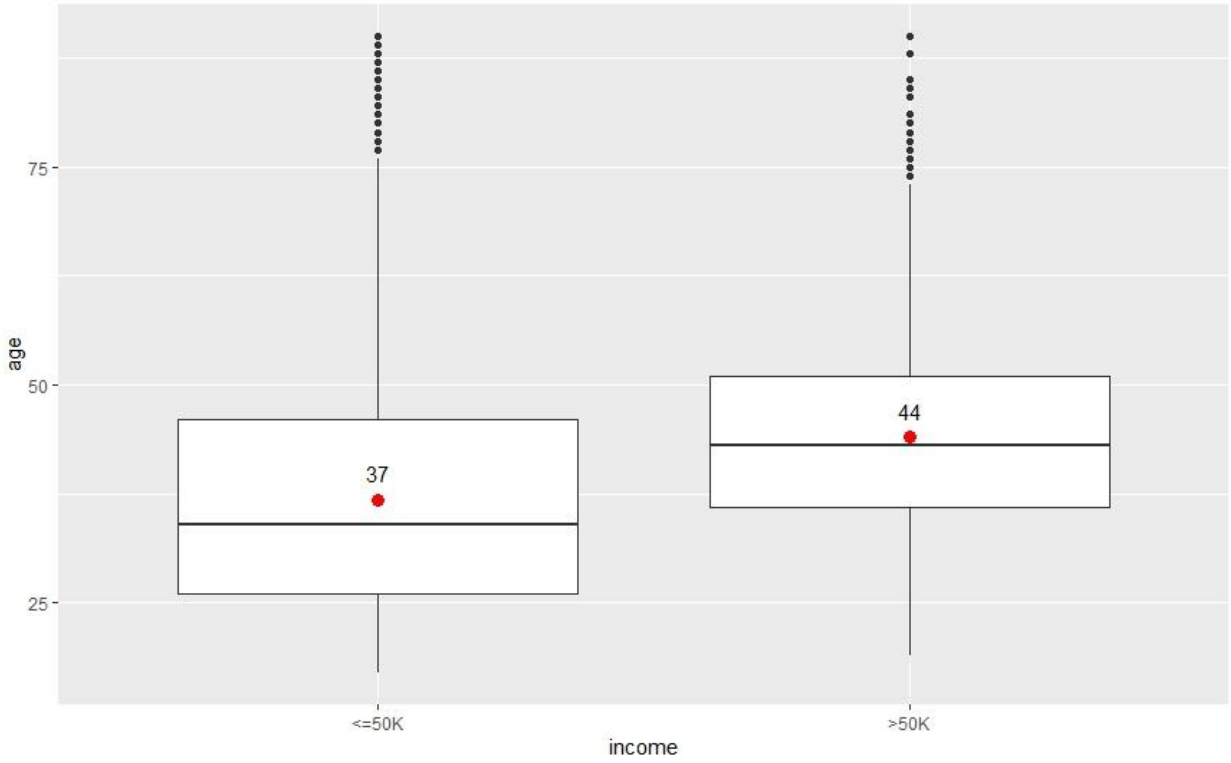
1. **Age vs. income**

As it is seen the age of individuals who make over $50K a year is statistically higher than the other group. It is a reasonable observation since the employees tend to earn more as they get experience and become more qualified. Since a p-value obtained from t-test is very small we can conclude that the difference between these two populations is statistically significant.

	Age	
	<=50K	>50K
Minimum	17	19
1st quartile	26	36
Median	34	43
Mean	37	44
3rd quartile	46	51
Maximum	90	90

```
> means=aggregate(age~income,censusdata,mean)
> means

  income      age
1  <=50K 36.74943
2   >50K 44.00607
```

```
> means$age=round(means$age,0)
> bp1=ggplot(data=censusdata,mapping = aes(x=income,y=age))+geom_boxplot()+st
at_summary(fun.y=mean, geom="point", color="red", size=3)+geom_text(data = me
ans, aes(label = age, y = age + 3))
> bp1
```



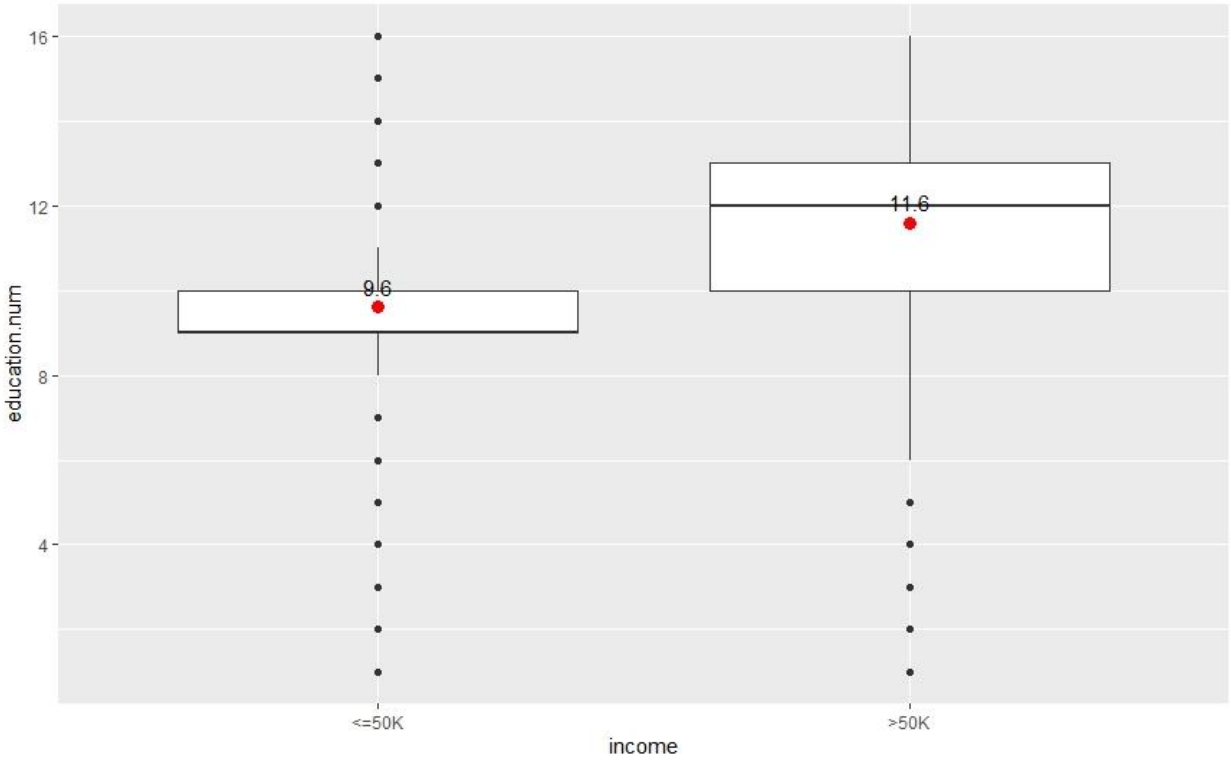

```
> t1=t.test((censusdata$age~censusdata$income))
> t1

        Welch Two Sample t-test

t = -59.35, df = 24884, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
```

2. **Education number vs. income**

In the previous section, we showed that the higher education level is correlated to the number of i ndividuals who makes over $50K a year. Accordingly, it can be seen that the education number f or those individuals is relatively higher than the individuals who makes less than $50K a year. Ag ain, a p-value obtained from t-test is very small and hence, we can conclude that the difference be tween these two populations is statistically significant.

	Education.num	
	<=50K	>50K
Minimum	1	1
1st quartile	9	10
Median	9	12
Mean	9.7	11.6
3rd quartile	10	13
Maximum	16	16

```
> means=aggregate(education.num~income,censusdata,mean)
> means$education.num=round(means$education.num,1)
> means
  income education.num
1  <=50K           9.6
2   >50K          11.6
> bp2=ggplot(data=censusdata,mapping = aes(x=income,y=education.num))+geom_bo
xplot()+stat_summary(fun.y=mean, geom="point", color="red", size=3)+geom_text
(data = means, aes(label = education.num, y = education.num + 0.5))
> bp2
```




```
> t2=t.test((censusdata$education.num~censusdata$income))
> t2

        Welch Two Sample t-test
```

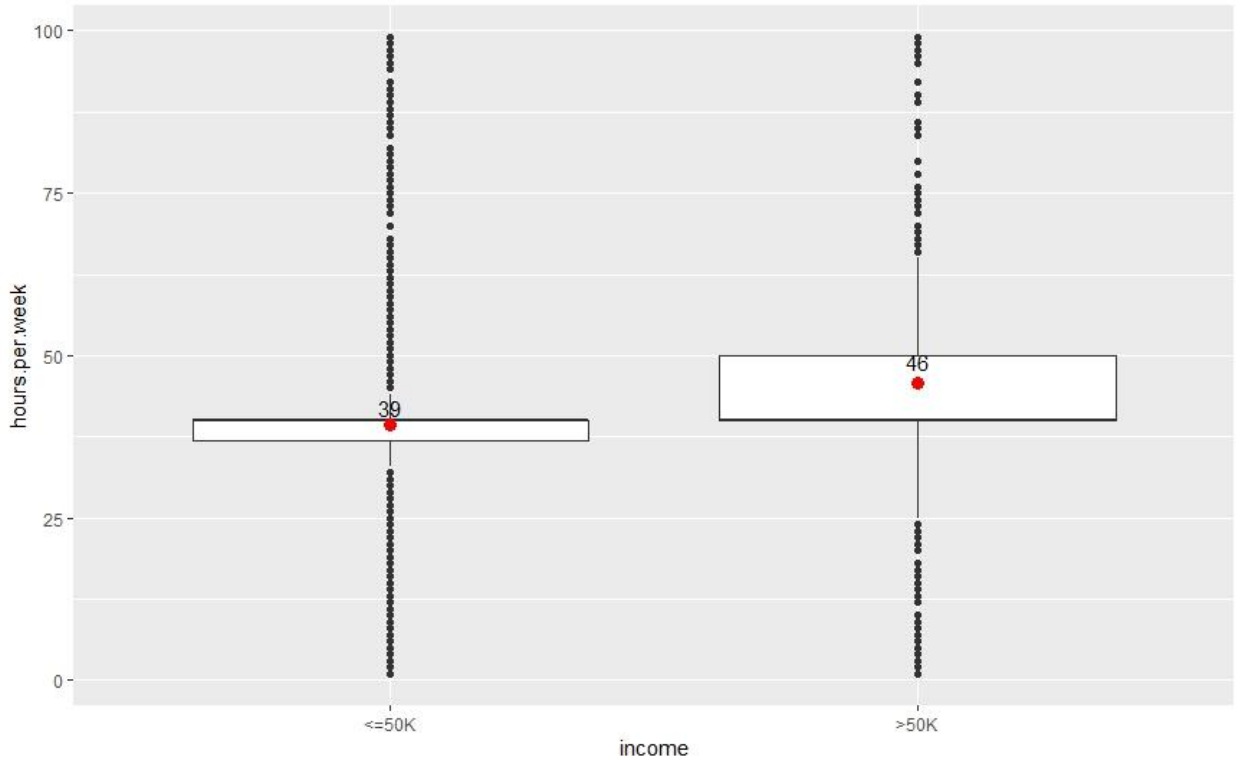
```
t = -75.89, df = 19494, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
```

3. **Hours per week vs. income**

It can be seen that the individuals who make over $50K a year have a slightly higher weekly work hours than the other group. Since a p-value obtained from t-test is very small we can conclude that the difference between these two populations is statistically significant.

	Hours.per.week	
	<=50K	>50K
Minimum	1	1
1st quartile	37	40
Median	40	40
Mean	39.4	45.7
3rd quartile	40	50
Maximum	99	99

```
> means=aggregate(hours.per.week~income,censusdata,mean)
> means$hours.per.week=round(means$hours.per.week,0)
> means
  income hours.per.week
1  <=50K              39
2   >50K              46
> bp3=ggplot(data=censusdata,mapping = aes(x=income,y=hours.per.week))+geom_b
oxplot()+stat_summary(fun.y=mean, geom="point", color="red", size=3)+geom_tex
t(data = means, aes(label = hours.per.week, y = hours.per.week + 3))
> bp3
```




```
> t3=t.test((censusdata$hours.per.week~censusdata$income))
> t3

        Welch Two Sample t-test

t = -52.26, df = 20994, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
```

Feature selection

Some similar categories that had fewer instances were merged to obtain a clear representation of the dataset. In the education level feature, 1^{st}-4^{th},5^{th}-6^{th},7^{th}-8^{th} and 9^{th} categories were merged into the one category named "1^{st}-9^{th}". In the marital status feature, all married categories were merged into one category named "Married". All countries in native.country feature were merged into a category named "Other country", except US, Mexico, Philippines and India.

```
> levels(censusdata_mod$education_level)
 [1] " 10th" " 11th" " 12th" " 1st-9th" "Assoc-acdm" "Assoc-voc" " Bachelors"
 [8] "Doctorate" "HS-grad" "Masters" "Preschool" "Prof-school" "Some-college"
> levels(censusdata_mod$marital.status)
[1] " Divorced"    " Married"     " Never-married" " Separated"    " Widowed"
> levels(censusdata_mod$native.country)
[1] " India"   " Mexico"   " Other country" " Philippines"   " United-States"
```

One hot encoding process is used to convert categorical variables into a form that can be inputted to the prediction models. In this process, each category in the categorical feature becomes a new feature of the dataset. These are the dummy variables which takes only the value 0 or 1 to indicate the absence or presence of category for a given instance. For instance, the race feature is converted into 5 dummy variables as shown.

```
> library(fastDummies)
> race_dummy=dummy_cols(censusdata_mod$race)
> colnames(race_dummy)=c("Race","race: Amer-Indian-Eskimo","race: Asian-Pac-I
slander","race: Black","race: Other","race: White")
```

	Race	race: Amer-Indian-Eskimo	race: Asian-Pac-Islander	race: Black	race: Other	race: White
1	White	0	0	0	0	1
2	White	0	0	0	0	1
3	White	0	0	0	0	1
4	Black	0	0	1	0	0
5	Black	0	0	1	0	0
6	White	0	0	0	0	1
7	Black	0	0	1	0	0

The new dataset contains 63 features including both continuous variables and dummy variables formed from categorical variables. The last column is a dummy variable for the income feature which will be used as a response variable. For this feature, 1 corresponds to the instances where income is over $50k a year and 0 when income is equal to or less than $50k a year. As "capital-gain" and "capital-loss" have highly-skewed distributions, a logarithmic transformation was applied on these variables so that the extreme values do not negatively affect the performance of any learning algorithm. Using a logarithmic transformation significantly reduces the range of values caused by outliers. Additionally, scaling was applied to the continuous variables to ensure that each feature is treated equally when applying algorithms.

```
> censusdata_scaled=censusdata_dummies
> censusdata_scaled[,3]=log(censusdata_scaled[,3]+1)
> censusdata_scaled[,4]=log(censusdata_scaled[,4]+1)
> censusdata_scaled[,1:5]=scale(censusdata_scaled[,1:5],center = TRUE, scale =
TRUE)
```

To run the algorithms faster, the number of variables was reduced further. Since education.num and education_level variables represent the same information, the education_level variable was omitted to reduce the number of dummy variables in the data set.

```
> censusdata_final=censusdata_scaled[,-c(13:25)]
```

Subset selection

First, the "best subset" selection method was performed. The method was found to be unstable for this dataset since there are a lot of variables and it considers all of the possible models. Therefore, the forward and backward selection methods were used as they are path dependent models and are faster to implement.

```
> regfit.forward=regsubsets(income~.,censusdata_final,nvmax=49,method = "forw
ard")
> regfit.summary=summary(regfit.forward)
> which.max(regfit.summary$adjr2)
[1] 32
```

```
                      age                 education.num                  capital.gain                  capital.loss
                        1                             2                             3                             4
             hours.per.week       `workclass: Federal-gov`         `workclass: Local-gov`          `workclass: Private`
                        5                             6                             7                             8
   `workclass: Self-emp-inc` `workclass: Self-emp-not-inc`             `ms: Married`             `ms: Separated`
                        9                            10                            14                            16
             `ms: Widowed`            `ocp: Adm-clerical`          `ocp: Craft-repair`         `ocp: Exec-managerial`
                       17                            18                            20                            21
      `ocp: Farming-fishing`           `ocp: Other-service`        `ocp: Priv-house-serv`          `ocp: Prof-specialty`
                       22                            25                            26                            27
      `ocp: Protective-serv`                  `ocp: Sales`          `ocp: Tech-support`                `rel: Husband`
                       28                            29                            30                            32
        `rel: Not-in-family`               `rel: Unmarried`                  `rel: Wife`      `race: Amer-Indian-Eskimo`
                       33                            36                            37                            38
                `race: White`                  `sex: Female`      `natcount: Other country`       `natcount: Philippines`
                       42                            43                            47                            48
```

```
> which.min(regfit.summary$cp)
[1] 31
> which.min(regfit.summary$bic)
[1] 25
> which(regfit.summary$outmat[25,]=="*")
```

```
                      age                 education.num                  capital.gain                  capital.loss
                        1                             2                             3                             4
             hours.per.week       `workclass: Federal-gov`          `workclass: Private`      `workclass: Self-emp-inc`
                        5                             6                             8                             9
`workclass: Self-emp-not-inc`                `ms: Married`          `ocp: Adm-clerical`          `ocp: Craft-repair`
                       10                            14                            18                            20
       `ocp: Exec-managerial`         `ocp: Farming-fishing`          `ocp: Other-service`         `ocp: Priv-house-serv`
                       21                            22                            25                            26
        `ocp: Prof-specialty`         `ocp: Protective-serv`                  `ocp: Sales`           `ocp: Tech-support`
                       27                            28                            29                            30
              `rel: Husband`           `rel: Not-in-family`                  `rel: Wife`                 `race: White`
                       32                            33                            37                            42
                `sex: Female`
                       43
```

```
> regfit.backward=regsubsets(income~.,censusdata_final,nvmax=49,method = "bac
kward")
> regfit.summary2=summary(regfit.backward)
> which.max(regfit.summary2$adjr2)
[1] 34
```

```
                      age                 education.num                  capital.gain                  capital.loss
                        1                             2                             3                             4
             hours.per.week       `workclass: Federal-gov`         `workclass: Local-gov`          `workclass: Private`
                        5                             6                             7                             8
   `workclass: Self-emp-inc`          `workclass: State-gov`             `ms: Divorced`                `ms: Married`
                        9                            11                            13                            14
         `ms: Never-married`            `ocp: Adm-clerical`          `ocp: Craft-repair`         `ocp: Exec-managerial`
                       15                            18                            20                            21
      `ocp: Farming-fishing`           `ocp: Other-service`        `ocp: Priv-house-serv`          `ocp: Prof-specialty`
                       22                            25                            26                            27
      `ocp: Protective-serv`                  `ocp: Sales`          `ocp: Tech-support`                `rel: Husband`
                       28                            29                            30                            32
        `rel: Not-in-family`           `rel: Other-relative`               `rel: Own-child`             `rel: Unmarried`
                       33                            34                            35                            36
`race: Amer-Indian-Eskimo`      `race: Asian-Pac-Islander`                 `race: Black`                  `sex: Female`
                       38                            39                            40                            43
 `natcount: Other country`         `natcount: Philippines`
                       47                            48
```

```
> which.min(regfit.summary2$cp)
[1] 33
> which.min(regfit.summary2$bic)
[1] 25
> which(regfit.summary2$outmat[25,]=="*")
```

```
                     age             education.num             capital.gain             capital.loss
                       1                         2                        3                        4
          hours.per.week  `workclass: Federal-gov`    `workclass: Local-gov`      `workclass: Private`
                       5                         6                        7                        8
`workclass: Self-emp-inc`           `ms: Married`      `ocp: Adm-clerical`       `ocp: Craft-repair`
                       9                        14                       18                       20
   `ocp: Exec-managerial`    `ocp: Farming-fishing`     `ocp: Other-service`    `ocp: Prof-specialty`
                      21                        22                       25                       27
   `ocp: Protective-serv`           `ocp: Sales`      `ocp: Tech-support`           `rel: Husband`
                      28                        29                       30                       32
     `rel: Not-in-family`    `rel: Other-relative`        `rel: Own-child`       `rel: Unmarried`
                      33                        34                       35                       36
           `sex: Female`
                      43
```

The training set Mean Squared Error (MSE) is generally an underestimate of the test MSE. This is because when we fit a model to the training data using least squares, we specifically estimate the regression coefficients such that the training RSS is minimized. In particular, the training RSS decreases as we add more features to the model, but the test error may not. Therefore, the training RSS and R2R2 may not be used for selecting the best model unless we adjust for this underestimation. The evaluation criteria, including adjusted R^2, Mallow's Cp criteria and Bayesian Information Criteria (BIC) are used to select the best subset. As it can be seen they gave different results for the best subset. It was decided to choose the subset with 25 variables obtained from the forward selection method. It includes most important variables related to the income variable that were discussed in the previous section.

```
> censusdata_final=censusdata_final[,c(1:6,8:10,14,18,20:22,25:30,32,33,37,42
,43,50)]
```

Classification

In this section, classification techniques, including logistic regression, linear discriminant analysis, quadratic discriminant analysis and k-nearest neighbors are employed to predict whether an individual makes over $50K a year. Two different sets of predictions were performed for each technique. The first case uses the previously selected subset as the predictors. The second case utilizes only five continuous variables, namely age, education number, capital gain, capital loss and hours per week to predict a response variable. To evaluate the classification technique, the accuracy is estimated by dividing correctly predicted instances by the number of all instances in the test dataset. k-fold cross validation was performed to compare the models. In practice, one typically performs k-fold cross-validation using k = 5 or k = 10, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high

variance. For this analysis, 10-fold cross validation was chosen, and the accuracy was averaged over 10-folds.

Logistic regression

The logistic regression was performed using two different sets of predictors. “For” loop was written to perform cross-validation. In the j^{th} fold, the elements of folds that equal j are in the test set, and the remainder are in the training set. The predictions were made for each model size using predict(), the accuracy was computed on the appropriate subset, and stored in the appropriate slot in the matrix accuracy. The logistic regression showed a good performance with an average accuracy of 0.84. Interestingly, when the five continuous variables were used as predictors, the average accuracy is 0.8 which is still very high. It shows that these variables play a major role in prediction of the income parameter.

```
> k=10
> set.seed (1)
> folds=sample (1:k,nrow(censusdata_final),replace =TRUE)
> accuracy =matrix (NA ,k,1, dimnames =list(NULL , paste (1:1) ))
> for(j in 1:k){
+ log_reg=glm(income~.,data=censusdata_final[folds!=j,],family=binomial)
+ test_fit=predict(log_reg,censusdata_final[folds==j,],type="response")
+ log_reg.pred=rep(0,nrow(censusdata_final[folds==j,]))
+ log_reg.pred[test_fit>0.5]=1
+ accuracy[j,1]=sum(log_reg.pred==censusdata_final[folds==j,26])/nrow(censusd
ata_final[folds==j,])
+ }
> round(accuracy,2)
         1
 [1,] 0.84
 [2,] 0.84
 [3,] 0.84
 [4,] 0.85
 [5,] 0.85
 [6,] 0.84
 [7,] 0.83
 [8,] 0.83
 [9,] 0.84
[10,] 0.84

> round(mean(accuracy),2)
```

```
[1] 0.84
```

```
> for(j in 1:k){
+ log_reg=glm(income~age+education.num+capital.gain+capital.loss+hours.per.we
ek,data=censusdata_final[folds!=j,],family=binomial)
+ test_fit=predict(log_reg,censusdata_final[folds==j,],type="response")
+ log_reg.pred=rep(0,nrow(censusdata_final[folds==j,]))
+ log_reg.pred[test_fit>0.5]=1
+ accuracy[j,1]=sum(log_reg.pred==censusdata_final[folds==j,26])/nrow(censusd
ata_final[folds==j,])
+ }
> round(accuracy,2)
         1
 [1,] 0.80
 [2,] 0.80
 [3,] 0.81
 [4,] 0.80
 [5,] 0.81
 [6,] 0.81
 [7,] 0.80
 [8,] 0.80
 [9,] 0.80
[10,] 0.80
> round(mean(accuracy),2)
[1] 0.8
```

Linear Discriminant Analysis (LDA)

Discriminant analysis models the distribution of the predictors *X* separately in each of the response classes, and then uses Bayes' theorem to convert these into estimates for the probability of the response category given the value of *X*. The covariances among the predictor variables *X* across all levels of *Y* are assumed to be equal for Linear Discriminant Analysis (LDA).

The LDA showed a similar performance to the logistic regression. It displayed an average accuracy of 0.84 for the selected subset of predictors and 0.8 when 5 continuous variables were used as the predictors. In addition, LDA performed slightly better for some of the folds than logistic regression.

```
> k=10
> set.seed (1)
> folds=sample (1:k,nrow(censusdata_final),replace =TRUE)
> accuracy =matrix (NA ,k,1, dimnames =list(NULL , paste (1:1) ))
```

```
> for(j in 1:k){
+  lda.fit=lda(income~.,data=censusdata_final[folds!=j,])
+  lda.pred=predict(lda.fit,censusdata_final[folds==j,])
+  accuracy[j,1]=sum(lda.pred$class==censusdata_final[folds==j,45])/nrow(cens
usdata_final[folds==j,])
+ }
> round(accuracy,2)
         1
 [1,] 0.85
 [2,] 0.84
 [3,] 0.84
 [4,] 0.85
 [5,] 0.84
 [6,] 0.84
 [7,] 0.83
 [8,] 0.83
 [9,] 0.84
[10,] 0.84
> round(mean(accuracy),2)
[1] 0.84
```

```
> for(j in 1:k){
+  lda.fit=lda(income~age+education.num+capital.gain+capital.loss+hours.per.w
eek,data=censusdata_final[folds!=j,])
+  lda.pred=predict(lda.fit,censusdata_final[folds==j,])
+  accuracy[j,1]=sum(lda.pred$class==censusdata_final[folds==j,26])/nrow(cens
usdata_final[folds==j,])
+ }
> round(accuracy,2)
         1
 [1,] 0.80
 [2,] 0.80
 [3,] 0.80
 [4,] 0.81
 [5,] 0.80
 [6,] 0.80
 [7,] 0.80
 [8,] 0.79
 [9,] 0.80
[10,] 0.80
> round(mean(accuracy),2)
[1] 0.8
```

Quadratic Discriminant Analysis (QDA)

Quadratic discriminant analysis (QDA) is a variant of LDA that allows for non-linear separation of data. For QDA, an individual covariance matrix is estimated for every class of observations.

The accuracy of classification using QDA is lower than LDA with an average accuracy of 0.77 for the first set of predictors and 0.78 for the second set of predictors. On the contrary to the logistic regression and LDA, the accuracy of classification when only 5 continuous variables were used as the predictors is higher than the accuracy with the selected subset for QDA. It can be due to the fact that when the number of predictors is large, the number of parameters estimated with QDA becomes very large since a separate covariance matrix for each class should be estimated. This can lead to a high variance when using QDA. Generally, LDA is a much less flexible classifier than QDA, and has substantially lower variance.

```
> for(j in 1:k){
+     qda.fit=qda(income~.,data=censusdata_final[folds!=j,])
+     qda.pred=predict(qda.fit,censusdata_final[folds==j,])
+     accuracy[j,1]=sum(qda.pred$class==censusdata_final[folds==j,26])/nrow(c
ensusdata_final[folds==j,])
+     }
> round(accuracy,2)
         1
 [1,] 0.78
 [2,] 0.77
 [3,] 0.78
 [4,] 0.78
 [5,] 0.78
 [6,] 0.76
 [7,] 0.77
 [8,] 0.77
 [9,] 0.78
[10,] 0.76

> round(mean(accuracy),2)
[1] 0.77

> for(j in 1:k){
+  qda.fit=qda(income~age+education.num+capital.gain+capital.loss+hours.per.w
eek,data=censusdata_final[folds!=j,])
+  qda.pred=predict(qda.fit,censusdata_final[folds==j,])
```

```
+  accuracy[j,1]=sum(qda.pred$class==censusdata_final[folds==j,26])/nrow(cens
usdata_final[folds==j,])
+ }
> round(accuracy,2)
         1
 [1,] 0.78
 [2,] 0.78
 [3,] 0.77
 [4,] 0.78
 [5,] 0.78
 [6,] 0.78
 [7,] 0.78
 [8,] 0.77
 [9,] 0.78
[10,] 0.77
> round(mean(accuracy),2)
[1] 0.78
```

K-Nearest Neighbors

K-Nearest Neighbors (KNN) was performed on both subsets of predictors. Since the number of observations is high, the optimum k was determined by estimating the accuracy of classification between k=10 and k=100. Due to the computational constraints, the procedure was performed for only k=10, 20, 30, 40, 50, 60, 70, 80, 90, 100. At each k value, the accuracy was averaged over 10 folds. Although the accuracy for different k values are very similar, the highest accuracy was obtained at k=50 which is 0.846. The same procedure was repeated when 5 continuous variables were used as the predictors. Again, the highest accuracy was obtained at k=50 which is equal to 0.8148.

```
> a=10
> set.seed (1)
> folds=sample (1:a,nrow(censusdata_final),replace =TRUE)
> accuracy =matrix (NA ,a,10, dimnames =list(NULL , paste (1:10) ))
> for(j in 1:a){
+ for (i in seq(10,100,10)){
+ knn.pred=knn(censusdata_final[folds!=j,1:25],censusdata_final[folds==j,1:25
],censusdata_final[folds!=j,26],k=i)
+ accuracy[j,i/10]=sum(knn.pred==censusdata_final[folds==j,26])/nrow(censusda
ta_final[folds==j,])
+ }
+ }
> round(accuracy,2)
```

```
        10   20   30   40   50   60   70   80   90  100
 [1,] 0.84 0.85 0.85 0.85 0.85 0.85 0.85 0.85 0.85 0.85
 [2,] 0.84 0.84 0.84 0.84 0.84 0.84 0.84 0.84 0.84 0.84
 [3,] 0.84 0.85 0.85 0.85 0.85 0.85 0.85 0.85 0.85 0.85
 [4,] 0.85 0.85 0.85 0.85 0.85 0.85 0.85 0.85 0.85 0.85
 [5,] 0.85 0.85 0.86 0.86 0.86 0.85 0.85 0.85 0.85 0.85
 [6,] 0.84 0.84 0.84 0.84 0.84 0.84 0.84 0.84 0.84 0.84
 [7,] 0.84 0.84 0.84 0.84 0.85 0.84 0.84 0.84 0.84 0.84
 [8,] 0.83 0.83 0.84 0.83 0.84 0.84 0.83 0.83 0.83 0.83
 [9,] 0.84 0.85 0.85 0.85 0.85 0.85 0.85 0.84 0.84 0.84
[10,] 0.83 0.84 0.84 0.84 0.84 0.84 0.84 0.84 0.84 0.84

> round(apply(accuracy ,2, mean),3)
   10    20    30    40    50    60    70    80    90   100
0.840 0.844 0.845 0.845 0.846 0.845 0.845 0.845 0.845 0.844

> for(j in 1:a){
+     for (i in seq(10,100,10)){
+     knn.pred=knn(censusdata_final[folds!=j,1:5],censusdata_final[folds==j,1
:5],censusdata_final[folds!=j,26],k=i)
+     accuracy[j,i/10]=sum(knn.pred==censusdata_final[folds==j,26])/nrow(cens
usdata_final[folds==j,])
+     }
+     }
        10   20   30   40   50   60   70   80   90  100
 [1,] 0.81 0.80 0.81 0.81 0.81 0.80 0.80 0.80 0.80 0.81
 [2,] 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81
 [3,] 0.81 0.82 0.82 0.82 0.82 0.82 0.82 0.81 0.82 0.81
 [4,] 0.81 0.81 0.81 0.81 0.82 0.82 0.82 0.82 0.82 0.82
 [5,] 0.82 0.82 0.82 0.82 0.82 0.82 0.82 0.82 0.82 0.82
 [6,] 0.81 0.82 0.82 0.82 0.82 0.82 0.82 0.82 0.82 0.82
 [7,] 0.80 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.80 0.80
 [8,] 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81
 [9,] 0.81 0.81 0.81 0.82 0.82 0.82 0.82 0.82 0.82 0.82
[10,] 0.81 0.81 0.82 0.82 0.82 0.82 0.82 0.82 0.82 0.82
> round(apply(accuracy ,2, mean),4)
    10     20     30     40     50     60     70     80     90    100
0.8100 0.8119 0.8140 0.8141 0.8148 0.8147 0.8145 0.8139 0.8136 0.8137
```

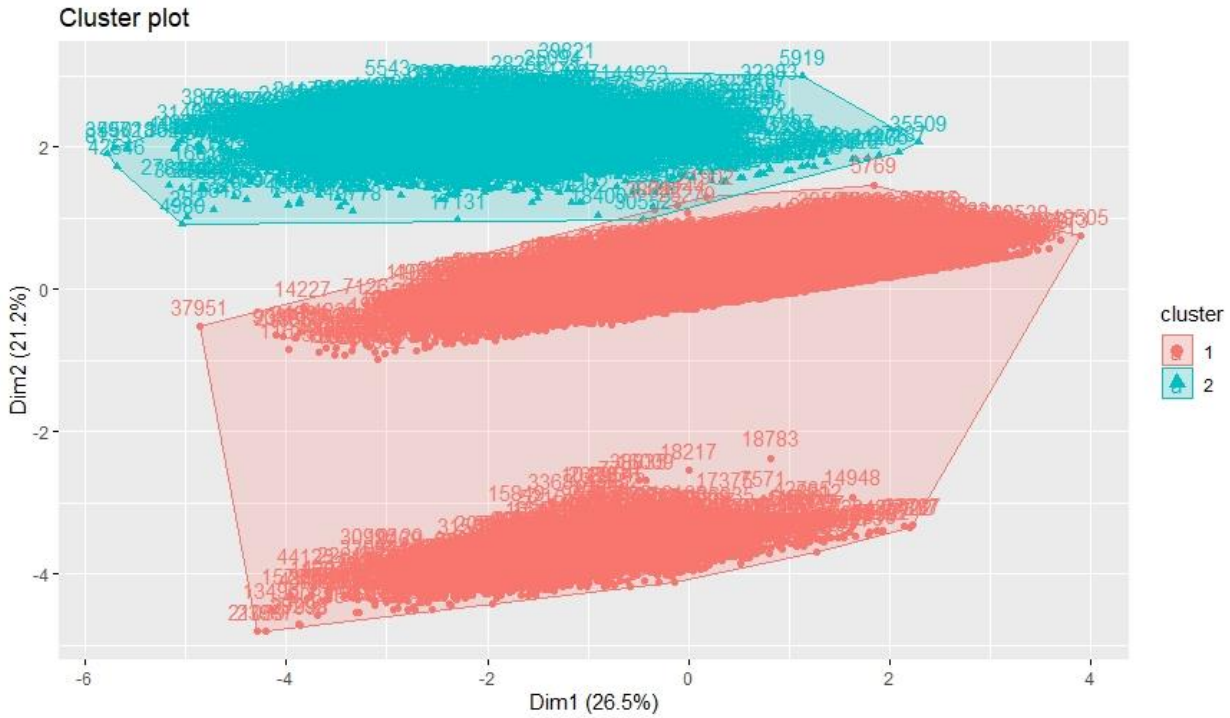
K-means clustering

In the previous section, the five continuous variables showed a good performance in predicting the income variable. Therefore, we decided to check whether they can be used to cluster the data into 2 groups, namely individuals who makes over $50K and equal to or less than $50K a year.

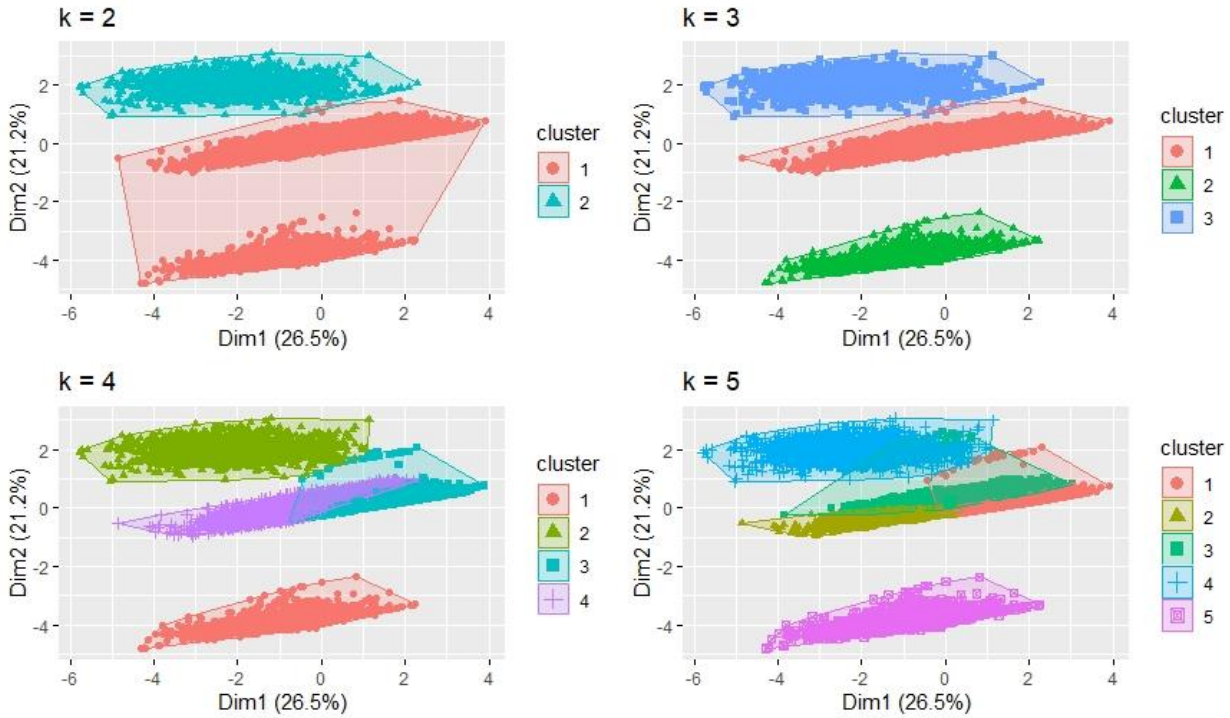
K-means clustering was used since the number of clusters was known a priori. The instances were classified in multiple groups, such that objects within the same cluster are as similar as possible, and objects from different clusters are as dissimilar as possible. In k-means clustering, each cluster is represented by its centroid which corresponds to the mean of points assigned to the cluster. The k-means clustering works by defining clusters so that the total variation within cluster is minimized.

The kmeans function in R was used. It also has nstart option that attempts multiple initial configurations and reports on the best one. nstart = 25 or 50 is usually recommended. Our groupings resulted in 2 cluster sizes of 41437 and 3785. However, our original dataset contains 11,208 instances with income > $50K and 34,014 instances with income = < $50K. Since there are more than two dimensions (variables) fviz_cluster() function performs principal component analysis (PCA) and plot the data points according to the first two principal components that explain the majority of the variance. The same process was executed for 3, 4, and 5 clusters and the results are shown in the figure below.

```
> library(tidyverse)
> library(cluster)
> library(factoextra)
> k2=kmeans(censusdata_final[,1:5], centers = 2, nstart = 25)
> k2
K-means clustering with 2 clusters of sizes 41437, 3785
> fviz_cluster(k2, data = censusdata_final[,1:5])
```

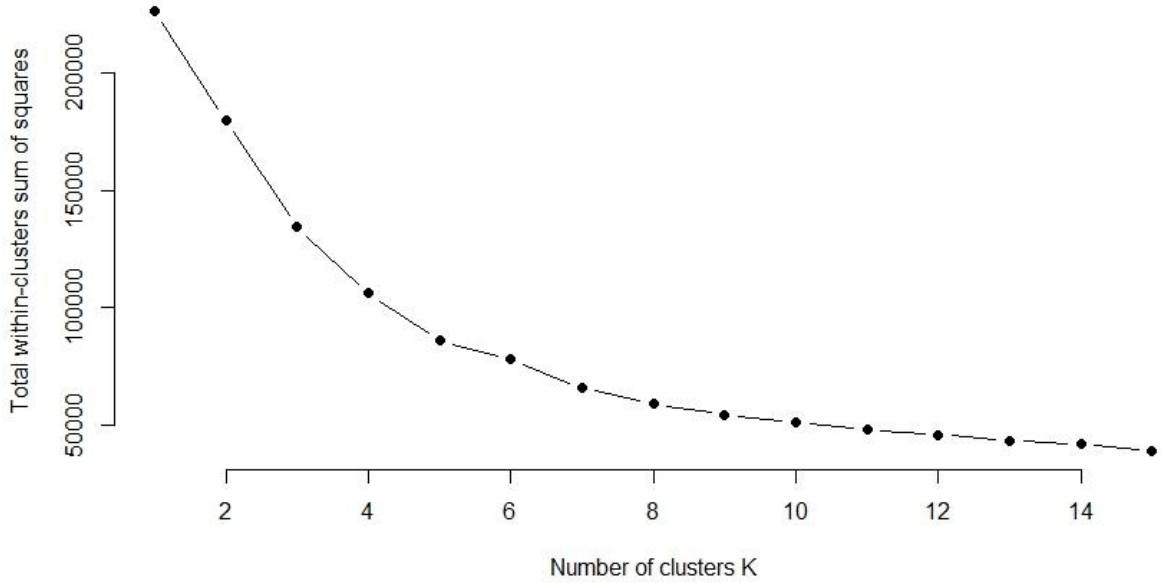



```
> k3 <- kmeans(censusdata_final[,1:5], centers = 3, nstart = 25)
> k4 <- kmeans(censusdata_final[,1:5], centers = 4, nstart = 25)
> k5 <- kmeans(censusdata_final[,1:5], centers = 5, nstart = 25)
> p1 <- fviz_cluster(k2, geom = "point", data = censusdata_final[,1:5]) + ggt
itle("k = 2")
> p2 <- fviz_cluster(k3, geom = "point",  data = censusdata_final[,1:5]) + gg
title("k = 3")
> p3 <- fviz_cluster(k4, geom = "point",  data = censusdata_final[,1:5]) + gg
title("k = 4")
> p4 <- fviz_cluster(k5, geom = "point",  data = censusdata_final[,1:5]) + gg
title("k = 5")
> library(gridExtra)
> grid.arrange(p1, p2, p3, p4, nrow = 2)
```



To determine the optimal number of clusters, the “Elbow” method was applied. For each k, total within-cluster sum of square (wss) was calculated. Next, the curve of wss according to the numbe r of clusters k was plotted. The location of a bend (knee) in the plot is generally considered as an i ndicator of the appropriate number of clusters. The figure shows that 5 is the optimal number of c lusters as it appears to be the bend in the knee (or elbow). It means that the original income varia ble can also be grouped in more than two levels.

```
set.seed(1)
wss <- function(k) {
    kmeans(censusdata_final[,1:5], k, nstart = 10 )$tot.withinss
}
k.values <- 1:15
plot(k.values, wss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```




```
> censusdata_cluster=as.data.frame(censusdata_final[,1:5])
```

Principal Component Analysis (PCA)

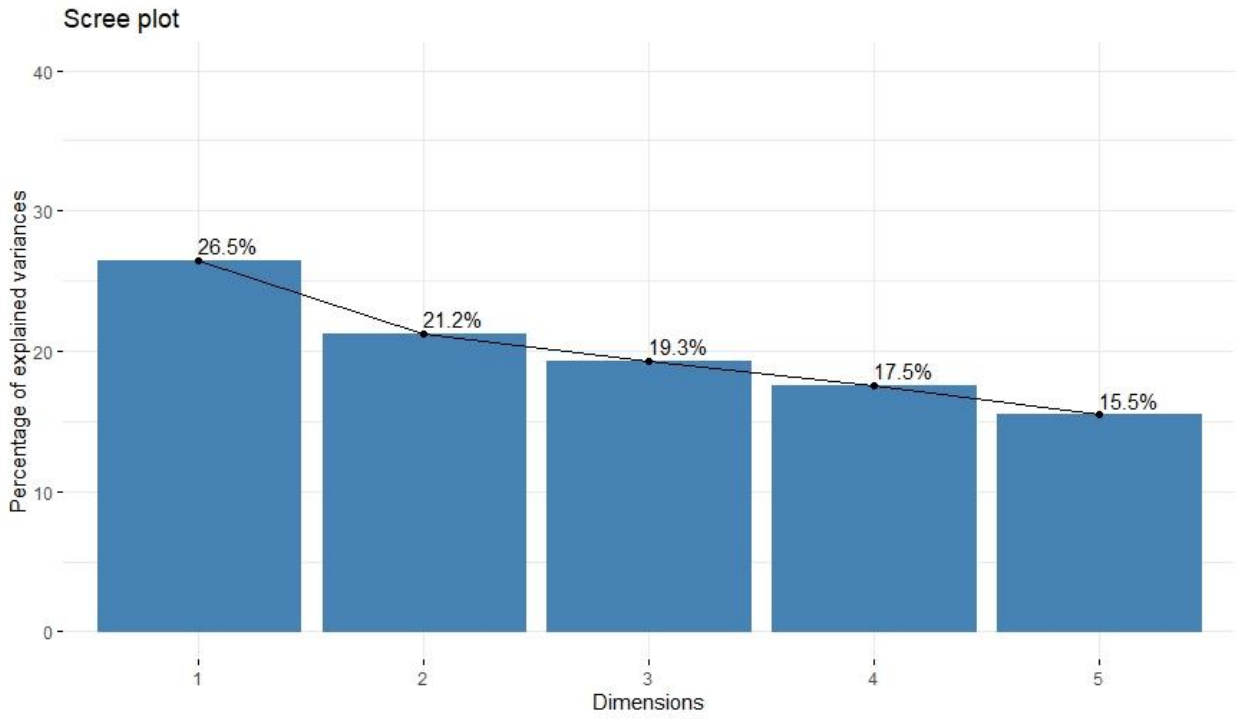
Principal component analysis is used to derive necessary information from a multivariate dataset and to convey this information as a set of fewer transformed variables called principal componen ts. The goal of PCA is to determine directions (i.e. principal components) along which the variati on in the data is the largest. Strictly speaking, PCA reduces the dimensionality of a multivariate d ata to a few principal components. This helps to visualize data graphically with minimal loss of i nformation.

PCA was performed in the previous section when K-means clustering was applied to the dataset. Here, it will be analyzed in more detail and the variables contributing to the different components will be shown.

The amount of variation possessed by each principal component is measured by eigenvalues. Eig envalues are large for the first principal components and small for the subsequent principal comp onents. It means that the first principal component corresponds to the direction with the largest a mount of variation in the dataset. An eigenvalue > 1 demonstrates that principal component acco unts for more variance than one of the initial variables in standardized data. This can be chosen a

s cutoff criteria to retain the principal components. Another method to determine the number of p rincipal components is to look at a Scree Plot. This is the plot of eigenvalues ordered from largest to the smallest. Instead of eigenvalues, the percentage of explained variances is given in the y-axi s of the figure below. It can be seen that the first three principal components explain 67% of the v ariation which is an acceptably large percentage.

```
> library("FactoMineR")
> library("factoextra")
> census_pca=PCA(censusdata_cluster, scale.unit = TRUE, graph = FALSE)
> eig.val=get_eigenvalue(census_pca)
> eig.val
       eigenvalue     variance.percent cumulative.variance.percent
Dim.1  1.3246750         26.49350                     26.49350
Dim.2  1.0620043         21.24009                     47.73359
Dim.3  0.9639465         19.27893                     67.01252
Dim.4  0.8762469         17.52494                     84.53746
Dim.5  0.7731272         15.46254                    100.00000

> fviz_eig(census_pca, addlabels = TRUE, ylim = c(0, 50))
```



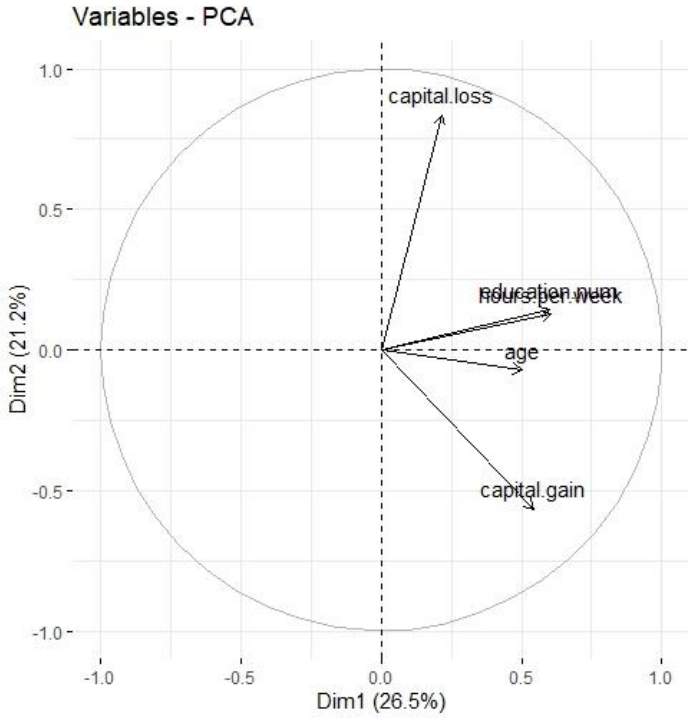
The correlation between a variable and a principal component (PC) is used as the coordinates of t he variable on the PC. In the correlation plot, positively correlated variables are grouped together , and negatively correlated variables are positioned on opposite sides of the plot origin. The dista nce between variables and the origin measures the quality of the variables on the factor map. The

factor map is a view of the projection of the observed variables projected into the plane covered by the first two principal components. The variables that are away from the origin are well repres ented on the factor map. It is shown that education number, hours per week and age are grouped t ogether and can be considered as positively correlated. As it is expected capital gain and capital l oss are on the opposite sides of the plot origin which shows their negative correlation. In addition , the distance between these two variables and origin is larger than the distance for other variable s.

```
var=get_pca_var(census_pca)
head(var$coord, 5)
```

```
                Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
age            0.5018645 -0.07125837  0.78041406  0.02222755 -0.3653958
education.num  0.6009703  0.14271129 -0.55628742  0.28706278 -0.4760330
capital.gain   0.5445787 -0.56647471 -0.02376183  0.36866464  0.4960465
capital.loss   0.2159306  0.83617157  0.16563234  0.31722937  0.3551374
hours.per.week 0.6070006  0.12838845 -0.13208352 -0.74619014  0.2020418
```

```
fviz_pca_var(census_pca, col.var = "black")
```



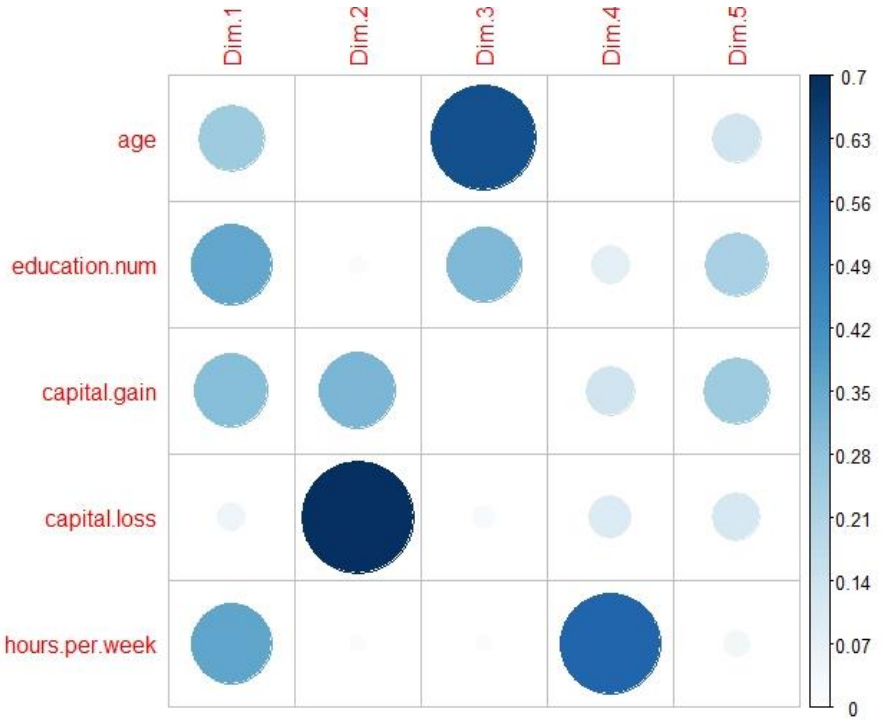
The cos2 parameter describes the quality of representation of the variables on factor map. A high cos2 corresponds to a good representation of the variable on the principal component. On a facto r map, the variable is positioned close to the circumference of the correlation circle. A low cos2 demonstrates that the variable is not thoroughly represented by the principal component. In this c

ase, the variable is close to the center of the circle. The cos2 of variables on all the dimensions w as visualized using the correlation plot. The age, education number, hours per week and capital g ain has similar high cos2 values which display a good representation of these variables on PC1. F or PC2, capital loss has a very high cos2 value followed by capital gain while other variables do not have a good representation on this dimension.

```
library("corrplot")
head(var$cos2, 5)
```

```
                Dim.1       Dim.2        Dim.3        Dim.4      Dim.5
age            0.25186796 0.005077755 0.6090461001 0.0004940641 0.13351412
education.num  0.36116532 0.020366513 0.3094556913 0.0824050372 0.22660744
capital.gain   0.29656599 0.320893592 0.0005646245 0.1359136153 0.24606218
capital.loss   0.04662601 0.699182897 0.0274340715 0.1006344754 0.12612254
hours.per.week 0.36844973 0.016483593 0.0174460556 0.5567997244 0.04082089
```

```
corrplot(var$cos2, is.corr=FALSE)
```



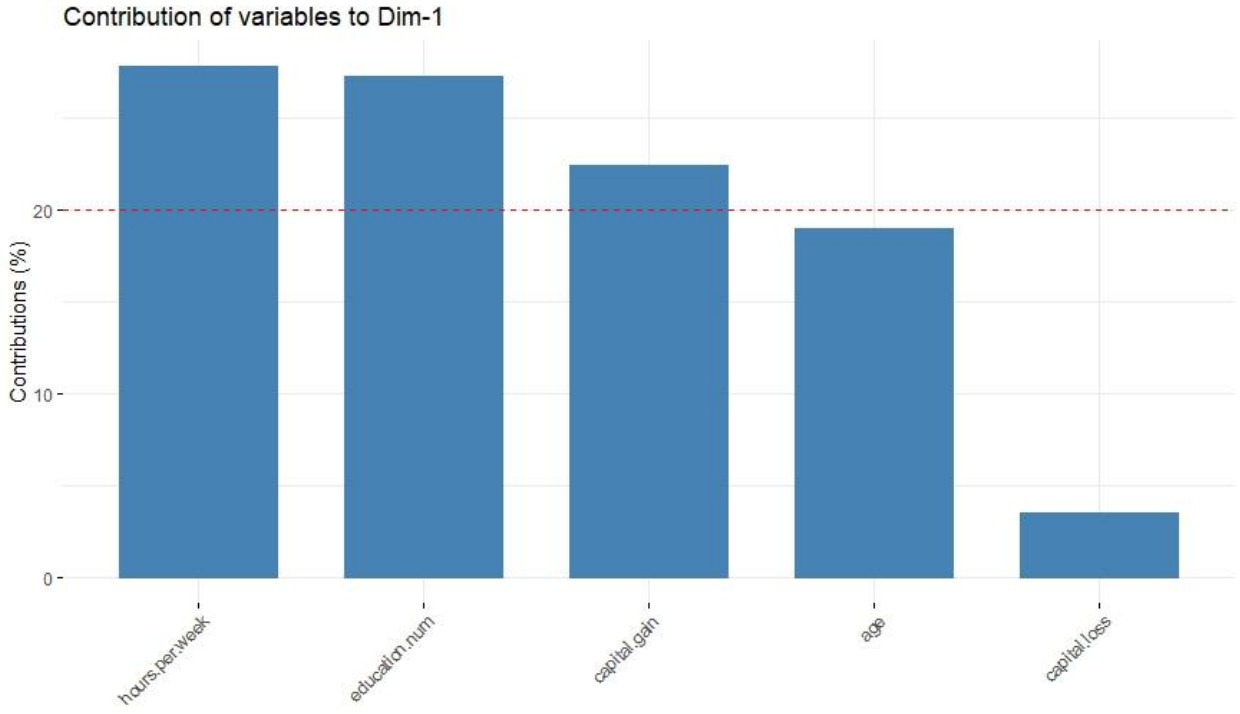
The contribution of variable in accounting for the variability in a given principal component is sh own in percentage. The most important variables in explaining the variability in the data set are t he ones that are correlated with PC1 and PC2. The variables that correlate with the last dimensio ns or do not correlate with any PC are variables with the lower contribution. The contribution of variables was extracted, and a bar plot was drawn. The red dashed line on the graph below indic ates the expected average contribution. The expected value would be 1/length(variables) = 1/5 =

20% if the contribution of the variables were uniform. For a given component, a variable with a c ontribution larger than this cutoff could be considered as important in contributing to the principa l component. These are hours per week, education number and capital gain for PC1 and capital g ain and capital loss for PC2. In addition, the most important (or, contributing) variables were hig hlighted on the correlation plot.

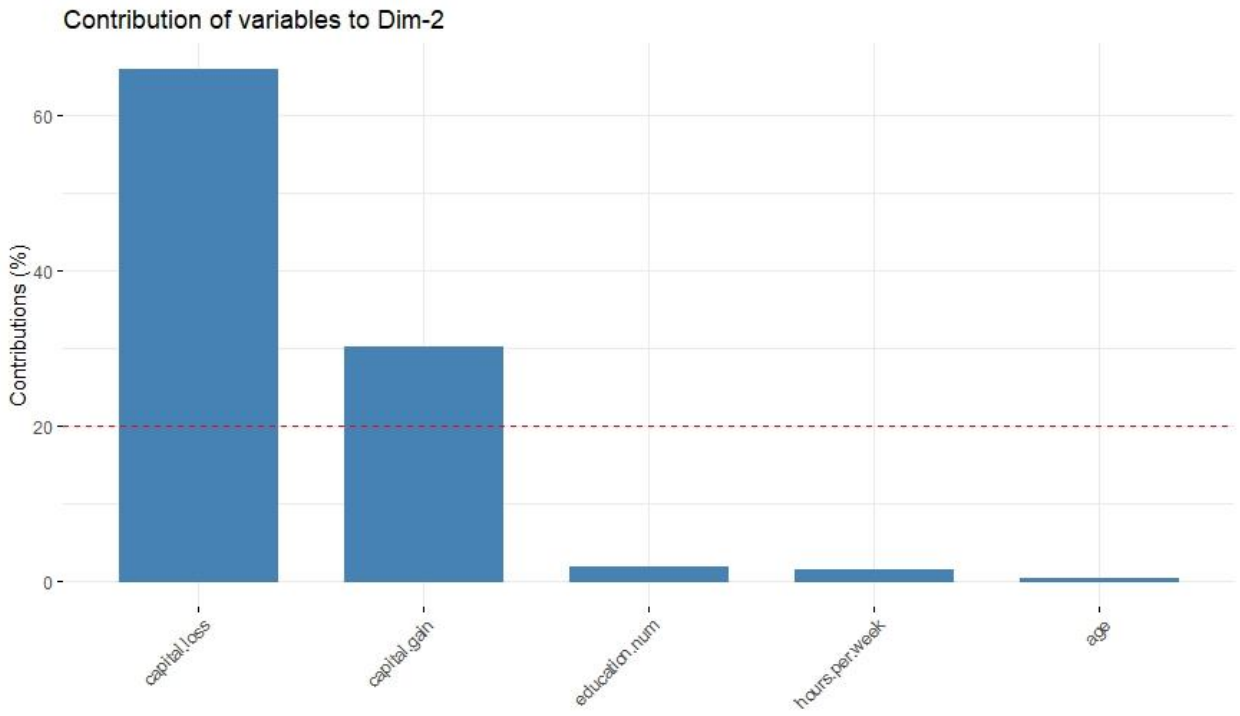
```
head(var$contrib, 5)
```

```
                 Dim.1      Dim.2       Dim.3       Dim.4     Dim.5
age           19.013566  0.4781294 63.18255971  0.05638412 17.269361
education.num 27.264447  1.9177429 32.10299301  9.40431694 29.310500
capital.gain  22.387830 30.2158453  0.05857425 15.51088086 31.826869
capital.loss   3.519808 65.8361613  2.84601586 11.48471664 16.313299
hours.per.week 27.814349  1.5521210  1.80985717 63.54370144  5.279971
```

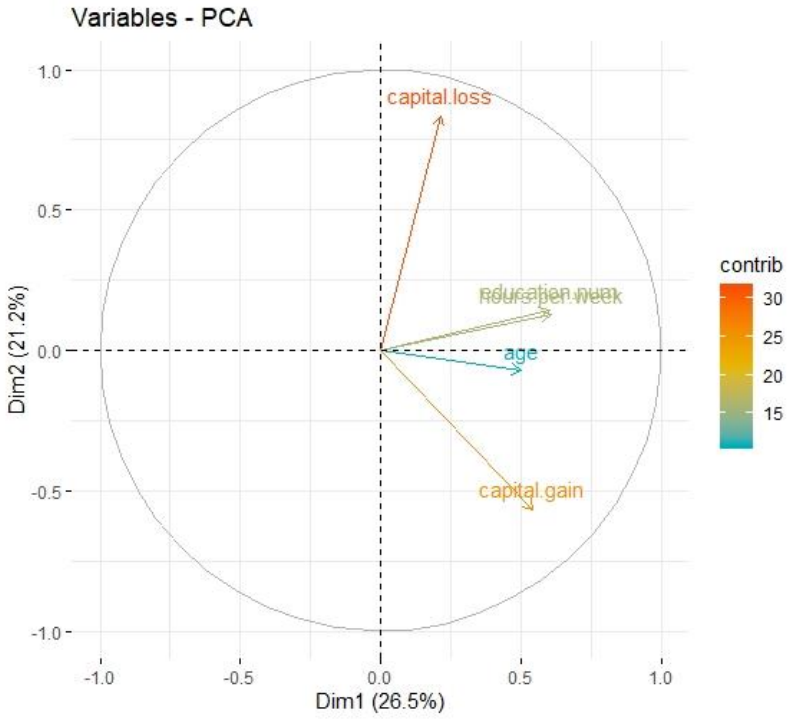
```
fviz_contrib(census_pca, choice = "var", axes = 1, top = 10)
```




```
fviz_contrib(census_pca, choice = "var", axes = 2, top = 10)
```




```
> fviz_pca_var(census_pca, col.var = "contrib",gradient.cols = c("#00AFBB", "
#E7B800", "#FC4E07"))
```



Summary

The census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau were analyzed. The classification techniques were applied to predict whether an individual makes over $50K a year. Two sets of predictors were used. First set of predictors w as obtained from the forward subset selection method while the second set included only 5 contin uous variables from the dataset. The summary of prediction results is given in the table below for both first and second set of predictors (average accuracy (1) and average accuracy (2), respectivel y). K-means clustering and Principal Component Analysis (PCA) was performed on the dataset o f 5 continuous variables. The optimum number of clusters obtained from the "elbow" method wa s 5. PC1 and PC2 explain the 47 % of the variation that is slightly low. It is related to the fact that the continuous variables in our dataset are not highly correlated and the PCA method is useful wh en the variables within the data set are highly correlated.

Classification technique	**Average accuracy (1)**	**Average accuracy (2)**
Logistic Regression	84 %	80 %
LDA	84 %	80 %
QDA	77 %	78 %
K-Nearest Neighbors	85 %	82 %

References

1. https://archive.ics.uci.edu/ml/datasets/Census+Income
2. http://michaelminn.net/tutorials/r-categorical/
3. http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software
4. http://uc-r.github.io/discriminant_analysis
5. https://thatdatatho.com/2018/02/12/linear-vs-quadratic-discriminant-analysis/
6. http://www.sthda.com/english/wiki/one-way-anova-test-in-r
7. https://uc-r.github.io/kmeans_clustering
8. http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/