Zadanie1

Łukasz Berwid

1/6/2020

R Markdown

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 contains demographic and employment related variables. here (http://mlr.cs.umass.edu/ml/machine-learning-databases/census-income-mld/census-income.data.html).

Data URL here (http://mlr.cs.umass.edu/ml/machine-learning-databases/census-income/census-income.data).

1. Download and load data

Download data

```
URL <- 'http://mlr.cs.umass.edu/ml/machine-learning-databases/census-income/census-income.data'
CSVFilePath <- 'Zadanie1.csv'
download.file(url=URL, destfile=CSVFilePath, method="libcurl")
```

Load data

```
CSVFilePath <- 'Zadanie1.csv'
censusIncomeDataFrame <- read.csv(file=CSVFilePath, strip.white=TRUE)
colnames(censusIncomeDataFrame) <- c("age", "workclass", "fnlwgt", "education", "education-num", "marital-status"
, "occupation", "relationship", "race", "sex", "capital-gain", "capital-loss", "hours-per-week", "native-country"
, "class")

kable(summary(censusIncomeDataFrame))
```

age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	cla
Min. :17.00	Private :22696	Min. : 12285	HS-grad :10501	Min. : 1.00	Divorced : 4443	Prof-specialty :4140	Husband :13193	Amer-Indian-Eskimo: 311	Female:10771	Min. : 0	Min. : 0.00	Min. : 1.00	United-States:29169	<=5
1st Qu.:28.00	Self-emp-not-inc: 2541	1st Qu.: 117832	Some-college: 7291	1st Qu.: 9.00	Married-AF-spouse : 23	Craft-repair :4099	Not-in-family : 8304	Asian-Pac-Islander: 1039	Male :21789	1st Qu.: 0	1st Qu.: 0.00	1st Qu.:40.00	Mexico : 643	>5(
Median :37.00	Local-gov : 2093	Median : 178363	Bachelors : 5354	Median :10.00	Married-civ-spouse :14976	Exec-managerial:4066	Other-relative: 981	Black : 3124	NA	Median : 0	Median : 0.00	Median :40.00	? : 583	NA
Mean :38.58	? : 1836	Mean : 189782	Masters : 1723	Mean :10.08	Married-spouse-absent: 418	Adm-clerical :3769	Own-child : 5068	Other : 271	NA	Mean : 1078	Mean : 87.31	Mean :40.44	Philippines : 198	NA
3rd Qu.:48.00	State-gov : 1297	3rd Qu.: 237054	Assoc-voc : 1382	3rd Qu.:12.00	Never-married :10682	Sales :3650	Unmarried : 3446	White :27815	NA	3rd Qu.: 0	3rd Qu.: 0.00	3rd Qu.:45.00	Germany : 137	NA
Max. :90.00	Self-emp-inc : 1116	Max. :1484705	11th : 1175	Max. :16.00	Separated : 1025	Other-service :3295	Wife : 1568	NA	NA	Max. :99999	Max. :4356.00	Max. :99.00	Canada : 121	NA
NA	(Other) : 981	NA	(Other) : 5134	NA	Widowed : 993	(Other) :9541	NA	NA	NA	NA	NA	NA	(Other) : 1709	NA

2.

Columns that contains missing values

```
cols_with_missing_names <- colnames(censusIncomeDataFrame)[apply(censusIncomeDataFrame, MARGIN = 2, function(a) a
ny(a=='?'))]
NameList <- cols_with_missing_names
idx <- match(NameList, names(censusIncomeDataFrame))
kable(colSums(censusIncomeDataFrame[,c(idx)]  == '?'), row.names = NA, col.names = 'missing count')
```

	missing count
workclass	1836
occupation	1843
native-country	583

Total Number of missing values

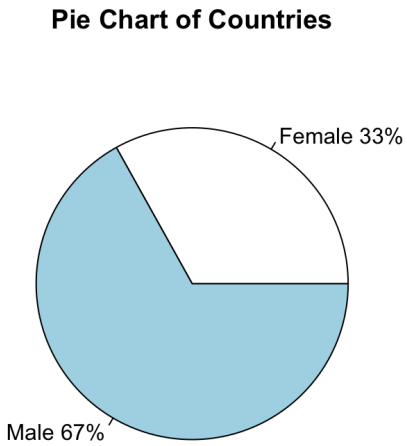
```
kable(length(censusIncomeDataFrame[censusIncomeDataFrame=='?']), col.names = 'missing count')
```

missing count
4262

3. Gender and age distribution

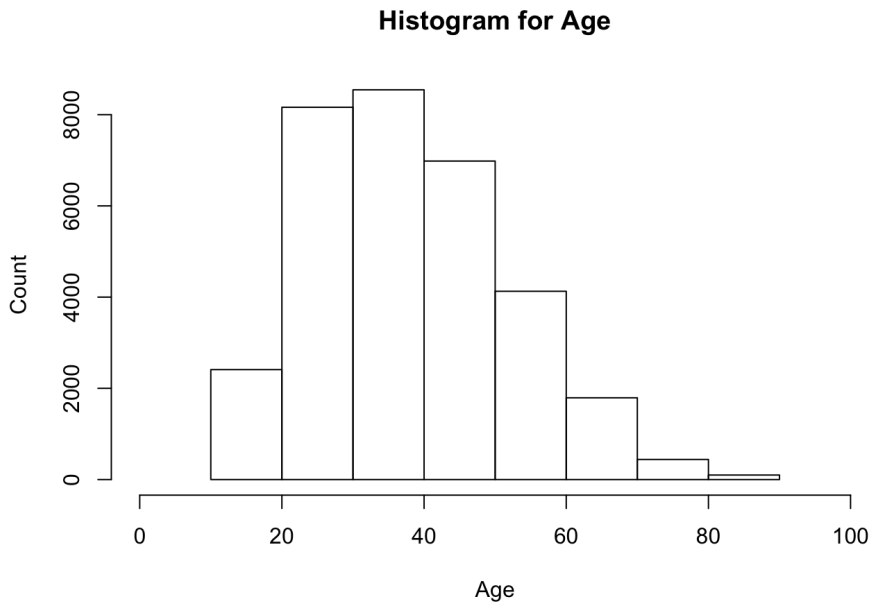
Gender distribution chart

```
slices <- aggregate(censusIncomeDataFrame$sex,by=list(censusIncomeDataFrame$sex),FUN=length)
lbls <- c('Female', 'Male')
pct <- round(slices$x/sum(slices$x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices$x, labels = lbls, main="Pie Chart of Countries")
```



Age distribution historgram

```
hist(censusIncomeDataFrame$age,
     main="Histogram for Age",
     xlab="Age",
     ylab="Count",
     xlim=c(0,100),
     breaks=10)
```



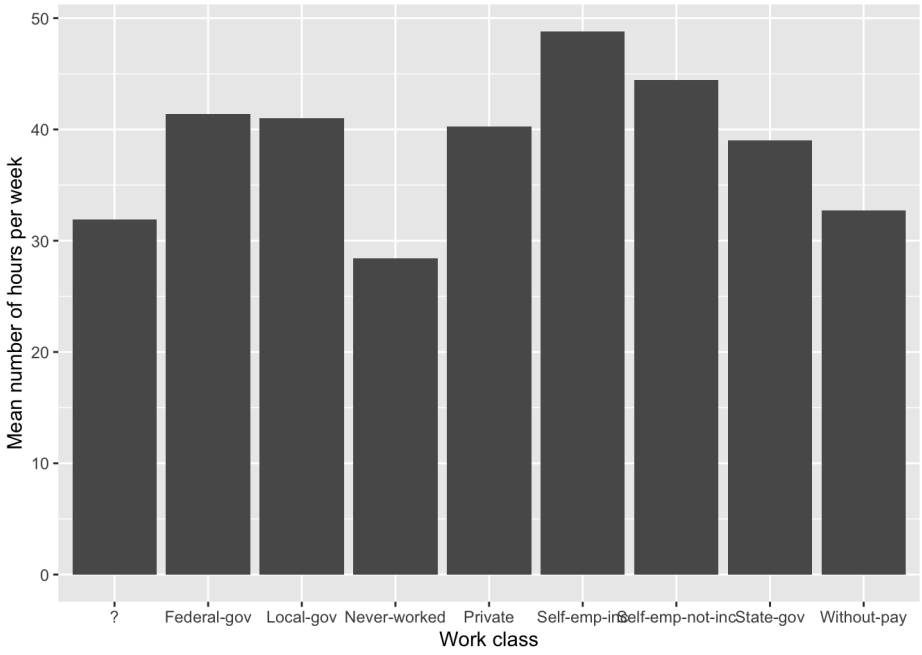
4. Table showing percetage count of native americans

```
natives_count <- length(which(censusIncomeDataFrame$`native-country` == "United-States")) + length(which(censusIn
comeDataFrame$`native-country` == "OutlyingUS(Guam-USVI-etc)"))
natives_percentage <- (natives_count / count(censusIncomeDataFrame)) * 100

natives_percentage <- data.frame(total_count=count(censusIncomeDataFrame), natives_count=natives_count, natives_p
ercentage=round(natives_percentage,2))

colnames(natives_percentage) <- rbind("total_count", "natives_count", "natives_percentage")
kable(natives_percentage)
```

total_count	natives_count	natives_percentage
32560	29169	89.59

5. Number of workhours per week by workclass

```
slices <- aggregate(censusIncomeDataFrame$`hours-per-week`,by=list(censusIncomeDataFrame$workclass),FUN=mean)
colnames(slices)<-rbind("workclass", "hours-per-week")

ggplot(slices, aes(x=slices$workclass, y=slices$`hours-per-week`)) +
  geom_bar(stat = "identity") +
  xlab("Work class") +
  ylab("Mean number of hours per week")
```

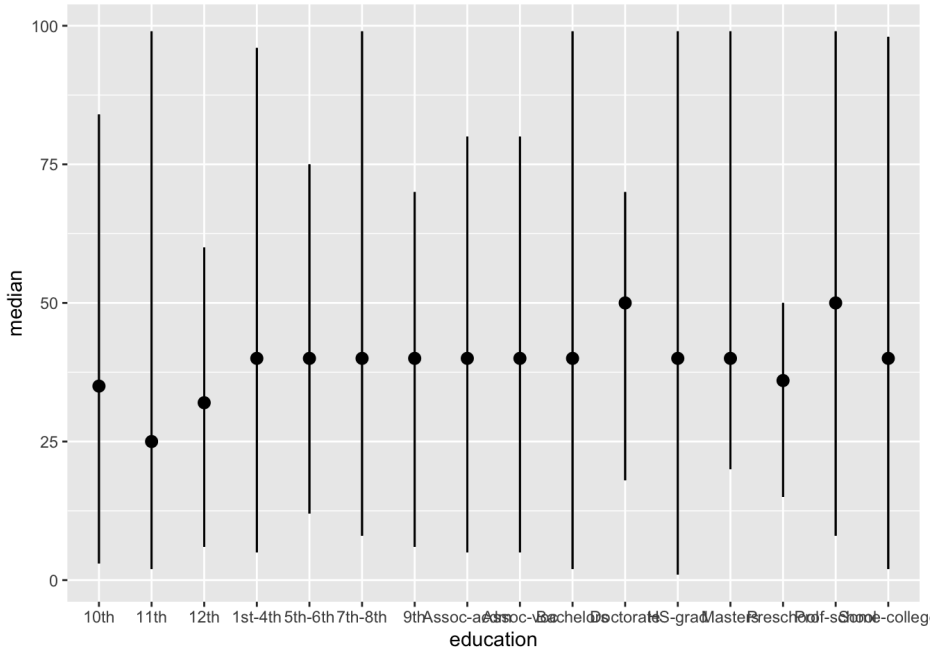


6. Distribution of average number of hours worked in private sector, for the group of people under age 30

```
under30_private <-subset(censusIncomeDataFrame, age < 30 & workclass == 'Private')

under30_private_summary <- under30_private %>%
    group_by(`education`) %>%
    summarise(count=n(), min = min(`hours-per-week`), max = max(`hours-per-week`), median = median(`hours-per-wee
k`))

kable(under30_private_summary)
```

education	count	min	max	median
10th	287	3	84	35
11th	494	2	99	25
12th	190	6	60	32
1st-4th	28	5	96	40
5th-6th	81	12	75	40
7th-8th	82	8	99	40
9th	123	6	70	40
Assoc-acdm	218	5	80	40
Assoc-voc	282	5	80	40
Bachelors	1062	2	99	40
Doctorate	11	18	70	50
HS-grad	2513	1	99	40
Masters	107	20	99	40
Preschool	11	15	50	36
Prof-school	32	8	99	50
Some-college	2154	2	98	40

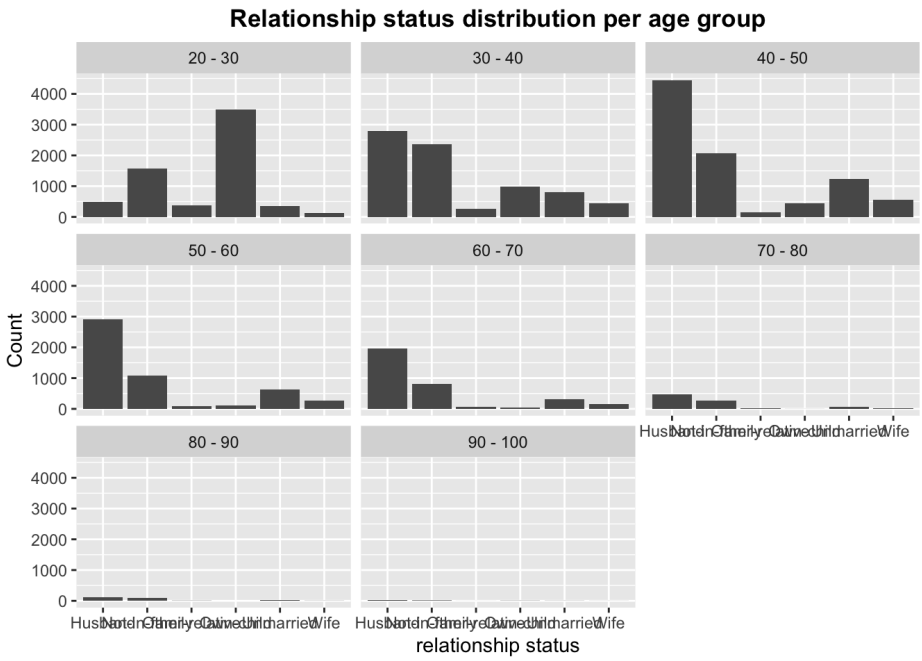
```
ggplot(under30_private_summary, aes(x = education, y = median, ymin = min, ymax = max)) +
    geom_linerange() +
    geom_pointrange()
```



7. Interesting plot

Relationship distribution by age group

```
censusIncomeDataFrame$age_group <- paste((round(censusIncomeDataFrame$age, -1)),  "-",  (round(censusIncomeDataFr
ame$age, -1) + 10))

suppressWarnings(ggplot(censusIncomeDataFrame, aes(relationship)) +
  ggtitle("Relationship status distribution per age group") +
  theme(plot.title = element_text(hjust = 0.5, face="bold")) +
  geom_histogram(stat="count") +
  xlab("relationship status") +
  ylab("Count") +
  facet_wrap("age_group"))
```

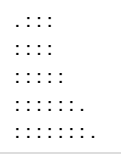
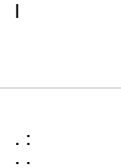
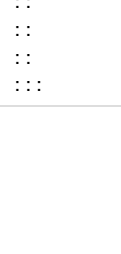
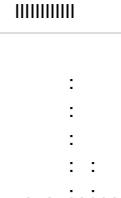


Data Summary

```
suppressWarnings(dfSummary(censusIncomeDataFrame, plain.ascii = FALSE, style = "grid", graph.magnif = 0.75, vali
d.col = FALSE))
```

Data Frame Summary

censusIncomeDataFrame
Dimensions: 32560 x 16
Duplicates: 24

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	age [integer]	Mean (sd) : 38.6 (13.6) min < med < max: 17 < 37 < 90 IQR (CV) : 20 (0.4)	73 distinct values		0 (0%)
2	workclass [factor]	1. ? 2. Federal-gov 3. Local-gov 4. Never-worked 5. Private 6. Self-emp-inc 7. Self-emp-not-inc 8. State-gov 9. Without-pay	1836 (5.6%) 960 (2.9%) 2093 (6.4%) 7 (0.0%) 22696 (69.7%) 1116 (3.4%) 2541 (7.8%) 1297 (4.0%) 14 (0.0%)		0 (0%)
3	fnlwgt [integer]	Mean (sd) : 189781.8 (105549.8) min < med < max: 12285 < 178363 < 1484705 IQR (CV) : 119223 (0.6)	21647 distinct values		0 (0%)
4	education [factor]	1. 10th 2. 11th 3. 12th 4. 1st-4th 5. 5th-6th 6. 7th-8th 7. 9th 8. Assoc-acdm 9. Assoc-voc 10. Bachelors [6 others]	933 (2.9%) 1175 (3.6%) 433 (1.3%) 168 (0.5%) 333 (1.0%) 646 (2.0%) 514 (1.6%) 1067 (3.3%) 1382 (4.2%) 5354 (16.4%) 20555 (63.1%)		0 (0%)
5	education-num [integer]	Mean (sd) : 10.1 (2.6) min < med < max: 1 < 10 < 16 IQR (CV) : 3 (0.3)	16 distinct values		0 (0%)
6	marital-status [factor]	1. Divorced 2. Married-AF-spouse 3. Married-civ-spouse 4. Married-spouse-absent 5. Never-married 6. Separated 7. Widowed	4443 (13.7%) 23 (0.1%) 14976 (46.0%) 418 (1.3%) 10682 (32.8%) 1025 (3.1%) 993 (3.0%)		0 (0%)
7	occupation [factor]	1. ? 2. Adm-clerical 3. Armed-Forces 4. Craft-repair 5. Exec-managerial 6. Farming-fishing 7. Handlers-cleaners 8. Machine-op-inspct 9. Other-service 10. Priv-house-serv [5 others]	1843 (5.7%) 3769 (11.6%) 9 (0.0%) 4099 (12.6%) 4066 (12.5%) 994 (3.1%) 1370 (4.2%) 2002 (6.1%) 3295 (10.1%) 149 (0.5%) 10964 (33.7%)		0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
8	relationship [factor]	1. Husband 2. Not-in-family 3. Other-relative 4. Own-child 5. Unmarried 6. Wife	13193 (40.5%) 8304 (25.5%) 981 (3.0%) 5068 (15.6%) 3446 (10.6%) 1568 (4.8%)	IIIIIIII IIIII III II	0 (0%)
9	race [factor]	1. Amer-Indian-Eskimo 2. Asian-Pac-Islander 3. Black 4. Other 5. White	311 (1.0%) 1039 (3.2%) 3124 (9.6%) 271 (0.8%) 27815 (85.4%)	 I IIIIIIIIIIIIIII	0 (0%)
10	sex [factor]	1. Female 2. Male	10771 (33.1%) 21789 (66.9%)	IIIIII IIIIIIIIIII	0 (0%)
11	capital-gain [integer]	Mean (sd) : 1077.6 (7385.4) min < med < max: 0 < 0 < 99999 IQR (CV) : 0 (6.9)	119 distinct values	: : : : :	0 (0%)
12	capital-loss [integer]	Mean (sd) : 87.3 (403) min < med < max: 0 < 0 < 4356 IQR (CV) : 0 (4.6)	92 distinct values	: : : : :	0 (0%)
13	hours-per-week [integer]	Mean (sd) : 40.4 (12.3) min < med < max: 1 < 40 < 99 IQR (CV) : 5 (0.3)	94 distinct values	: : : : . . . : : . .	0 (0%)
14	native-country [factor]	1. ? 2. Cambodia 3. Canada 4. China 5. Columbia 6. Cuba 7. Dominican-Republic 8. Ecuador 9. El-Salvador 10. England [32 others]	583 (1.8%) 19 (0.1%) 121 (0.4%) 75 (0.2%) 59 (0.2%) 95 (0.3%) 70 (0.2%) 28 (0.1%) 106 (0.3%) 90 (0.3%) 31314 (96.2%)	 IIIIIIIIIIIIIIIII	0 (0%)
15	class [factor]	1. <=50K 2. >50K	24719 (75.9%) 7841 (24.1%)	IIIIIIIIIIIII IIII	0 (0%)
16	age_group [character]	1. 20 - 30 2. 30 - 40 3. 40 - 50 4. 50 - 60 5. 60 - 70 6. 70 - 80 7. 80 - 90 8. 90 - 100	6411 (19.7%) 7638 (23.5%) 8884 (27.3%) 5119 (15.7%) 3350 (10.3%) 872 (2.7%) 238 (0.7%) 48 (0.1%)	III IIII IIIII III II	0 (0%)