Report on Data Anaylisis Project

June 2018, Helsinki University.

conducted by *Neli Noykova*

The data

The data for this project are taken from [**Statistics Finland**] (https://tilastokeskus.fi/tup/mikroaineistot/aineistot_en.html). These are individual level combined employer-employee data, or so-called FLEED (Finnish Longitudinal Employer-Employee Data). The original data contain information on population of working age, which can be combined with enterprise and establishment level data. Here we use two adapted for studying purposes data sets in Excel format (files **fleed_tyo.csv** and **fleed_yritys.csv**).

Employees

The file **fleed_tyo.csv** consists of data about 15-70 years old employees in Finland for the period 1990-2010. Because of data protection reasons only 15 years period is taken into account, and the years are numbered from 1 to 15. The file is fully described (in Finnish) here. The sample data involve information about 89312 persons and 18 variables, describing person's basic characteristics, family, living, employment relationships, periods of unemployment, income and education. The variables, listed in this file, are:

vuosi - Year, given in integer values.
shtun - Encrypted personal identity code, given in integer values.
syrtun - Encrypted enterprise code, related to the employment relationship during the last week of the year), integer values, with missing values.
sukup - Gender, 2 different integer values.
syntyv - Year of Birth, integer values.
kieli - Native language, defined as factor variable with 3 levels.
peas - Family status, 7 different integer values: 1 -head, 2 - spouse, 3 - child, 4 - head of cohabiting family, 5 - spouse of cohabiting family, 9 - unknown, 0 - not belonging to a family.
a7lkm - Number of children aged under 7 in family, integer values.
a18lkm - Number of children aged under 18 in family, integer values.
ktutk - Education, integer values.
sose - Socio-economic group, 9 different integer values.
ptoim1 - Main activity (TVM=employment relationship during the last week of the year), 7 different integer values.
tyokk - Months in employment, 13 different integer values.
tyke - Number of unemployment months, 13 different integer values.
toimiala - Industry (TVM=employment relationship during the last week of the year). It is defined as factor variable with 23 levels.
svatva - Earned income total in state taxation, integer values.
tyotu - Earned income, integer values.
suuralue12 - Major region based on the 2012 regional division, 5 different integer values.

```
data_workers <- read.csv(file="fleed_tyo.csv", header=TRUE, sep=",")
str(data_workers)
```

```
## 'data.frame':    89312 obs. of  18 variables:
##  $ vuosi     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ shtun     : int  2 3 5 7 8 9 11 12 13 14 ...
##  $ syrtun    : int  887 NA NA 4963 6639 8749 2506 7777 NA 6932 ...
```

```
##  $ sukup     : int  2 2 1 2 2 1 2 2 2 1 ...
##  $ syntyv    : int  1945 1927 1930 1952 1947 1950 1949 1957 1928 1946 ...
##  $ kieli     : Factor w/ 3 levels "9","fi","sv": 3 2 2 2 2 2 2 2 2 2 ...
##  $ peas      : int  2 2 1 0 1 1 5 2 0 1 ...
##  $ a7lkm     : int  0 0 0 NA 1 0 0 0 NA 0 ...
##  $ a18lkm    : int  0 0 0 NA 1 0 0 2 NA 1 ...
##  $ ktutk     : int  38 52 43 63 40 NA 34 32 NA 62 ...
##  $ sose      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ ptoim1    : int  11 24 24 11 11 11 11 11 24 11 ...
##  $ tyokk     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ tyke      : int  NA NA NA NA NA NA NA 7 NA NA ...
##  $ toimiala  : Factor w/ 23 levels "","A","B","C",..: 17 1 1 14 14 5 5 17 1 16 ...
##  $ svatva    : int  19000 16000 18000 54000 20000 22000 13000 8000 14000 37000 ...
##  $ tyotu     : int  19000 NA 9000 53000 20000 22000 12000 6000 NA 37000 ...
##  $ suuralue12: int  1 4 1 1 1 2 2 3 4 2 ...
```

Employers

The other file, **fleed_yritys.csv**, involves data about the corresponding employers during the same period. It is described in Finnish here.
The sample data involve information about 66878 companies and 6 variables, describing the different companies where employees have been working during the observed period. The variables, listed in this file, are:

vuosi - Year, given in integer values.
syrtun - Encrypted enterprise code, related to the employment relationship during the last week of the year), integer values, with missing values.
oty - Type of owner, integer values.
toimiala - Industry (TVM=employment relationship during the last week of the year). It is defined as factor variable with 22 levels.
SLHKY - Group of the company according the number of employees working there, 9 different integer values corresponding to 9 different groups - 1 with the smallest number of employees (< 4,5), and 9 with the biggest number of employees (>=9 999,5).
sllvy - Group of the company according its turnover, 9 different integer values, corresponding to 9 different groups - 1 with the smallest turnover (< 1000), and 9 with the biggest turnover (>=9 200 000 000).

```r
data_firms <- read.csv(file="fleed_yritys.csv", header=TRUE, sep=",")
str(data_firms)
```

```
## 'data.frame':    66878 obs. of  6 variables:
##  $ vuosi   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ syrtun  : int  1 3 6 14 16 17 20 24 27 28 ...
##  $ oty     : int  1 1 1 2 1 5 1 1 3 1 ...
##  $ toimiala: Factor w/ 22 levels "","A","B","C",..: 10 7 8 16 7 8 8 5 16 10 ...
##  $ SLHKY   : int  3 6 4 7 1 2 1 5 5 1 ...
##  $ sllvy   : int  5 7 6 1 2 6 5 6 1 3 ...
```

How the data about employees and employers are connected?

Both tables are connected via the variable **syrtun**, which is an encrypted enterprise code describing the employment relationship during the last week of the year.

The goal

The goal of this project is to find out how the employee's family status and the size of the employing company (measured by its turnover) are related to the size of the earned income.

Simplifying assumptions

1. The data should be restricted to the year **vuosi**=2 (the last number of my student number).
2. We assume that the whole earned income of each employee has come from a linked company during the investigated year.
3. We assume that the different employees are "independent of each other".
4. The missing values can be omited.

Data wrangling

1. Take a subset for year **vuosi** = 2 from both original tables.

```
data_workers2 <- subset(data_workers, vuosi==2)
dim(data_workers2)
```

```
## [1] 5855   18
```

```
data_firms2 <- subset(data_firms, vuosi== 2)
dim(data_firms2)
```

```
## [1] 3699    6
```

2. Merge both data frames by their intersection via the column **syrtun**.

```
data1<-merge(data_workers2, data_firms2, by="syrtun")
dim(data1)
```

```
## [1] 1925   23
```

3. We have to convert the integer values of some variables to factors. Since these variables are categorized to several groups, I think it is logical to investigate them as categorical ones. For example the groups in **sllvy** are divided according the companies turnovers in increasing order, but even in this case we are not allowed to compare directly for example the values 1 and 9 (for first and ninth group) since we are not guaranteed that there is such a numerical proportion between both groups. The numbers from 1 to 9 are just labels of different categories.

```
data1[,'peas']<-factor(data1[,'peas'])
levels(data1$peas)
```

```
## [1] "0" "1" "2" "3" "4" "5" "9"
```

```
data1[,'a7lkm']<-factor(data1[,'a7lkm'])
levels(data1$a7lkm)
```

```
## [1] "0" "1" "2" "3" "4" "5"
```

```
data1[,'a18lkm']<-factor(data1[,'a18lkm'])
levels(data1$a18lkm)
```

```
## [1] "0" "1" "2" "3" "4" "5" "6" "7" "8"
```

```r
data1[,'SLHKY']<-factor(data1[,'SLHKY'])
levels(data1$SLHKY)
```

```
## [1] "1" "2" "3" "4" "5" "6" "7" "8" "9"
```

```r
data1[,'sllvy']<-factor(data1[,'sllvy'])
levels(data1$sllvy)
```

```
## [1] "1" "2" "3" "4" "5" "6" "7" "8" "9"
```

4. Remove the missing values.

```r
data2<-na.omit(data1)
dim(data2)
```

```
## [1]  0 23
```

We see that since there are some missing data for all employees corresponding to year 2, if we remove all missing values, we end up with empty subset.

Therefore at this stage I remove only the missing values, involved directly in task description.

```r
data<-data1[!is.na(data1$svatva),]
dim(data)
```

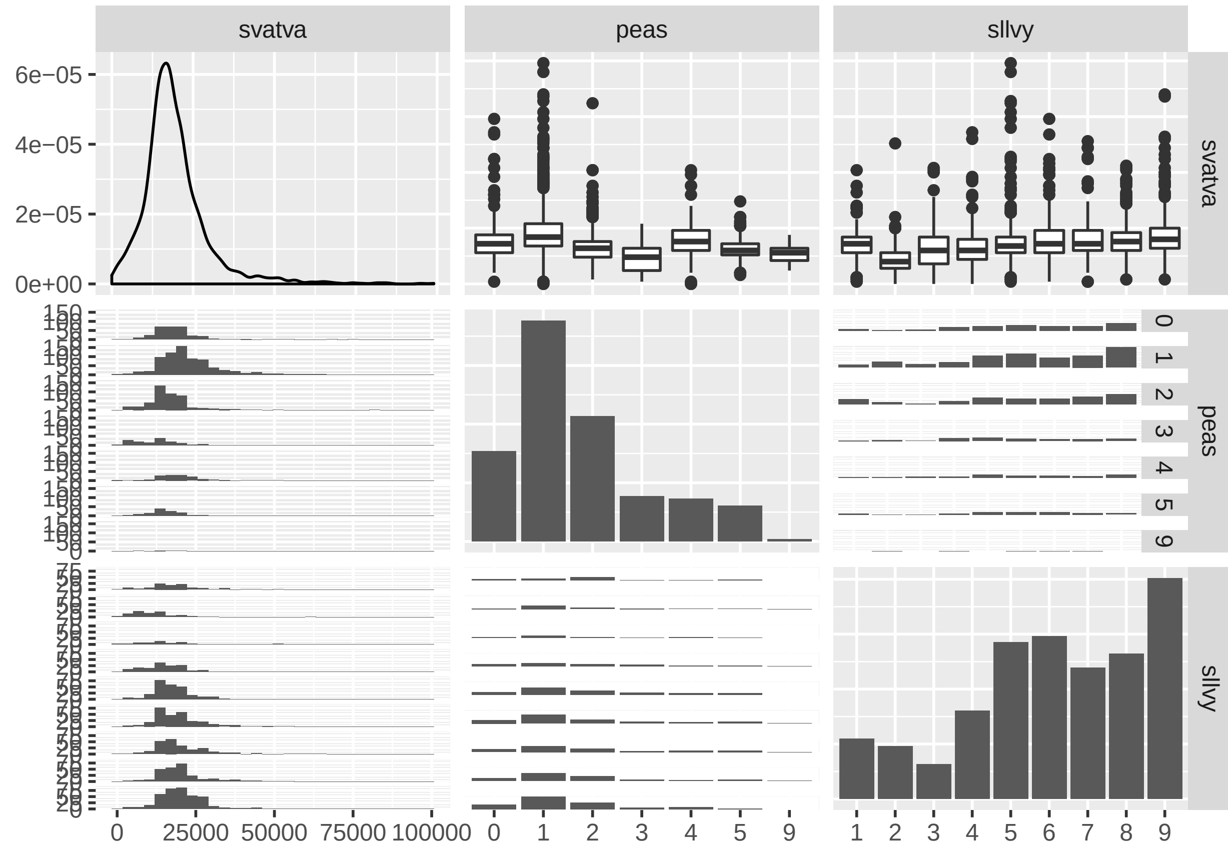
```
## [1] 1916   23
```

First look at the within correlation using ggpairs

To get a glimpse of the three targeted data variables we plot all investigated variables against each other.

```r
library(ggplot2)
library(GGally)
var<-data[,c("svatva", "peas", "sllvy")]
assignInNamespace("ggally_cor", ggally_cor, "GGally")
ggpairs(var, upper = list(continuous = wrap("cor", size = 10)), lower = list(continuous = "smooth"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We observe that the total earned income in almost all groups in both categoracal variables - family status and company turnover - deviates from normal distribution.

Univariate analyses

Univariate exploratory data analysis of total earned income svatva

The aim at this first analysis step is to achieve some preliminary assessments about the population distribution of the variables **svatva**, **SLLVY** and **peas** using the data of the observed sample for year=2. The variable **svatva** is numerical, while **svatva** and **SLLVY** are categorical.

First we analyse **svatva**.

```
attach(data)
summary(svatva)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   14000   18000   19434   23000   99000
```

We see that the mean and median are quite near, which suggests normal distribution.

The distribution of **svatva** can be visualized using so called **steam-and-leaf plot**. This plot is a special textual graph (table) where each data value is split into a “stem” (the first digit or digits) and a “leaf” (usually the last digit).

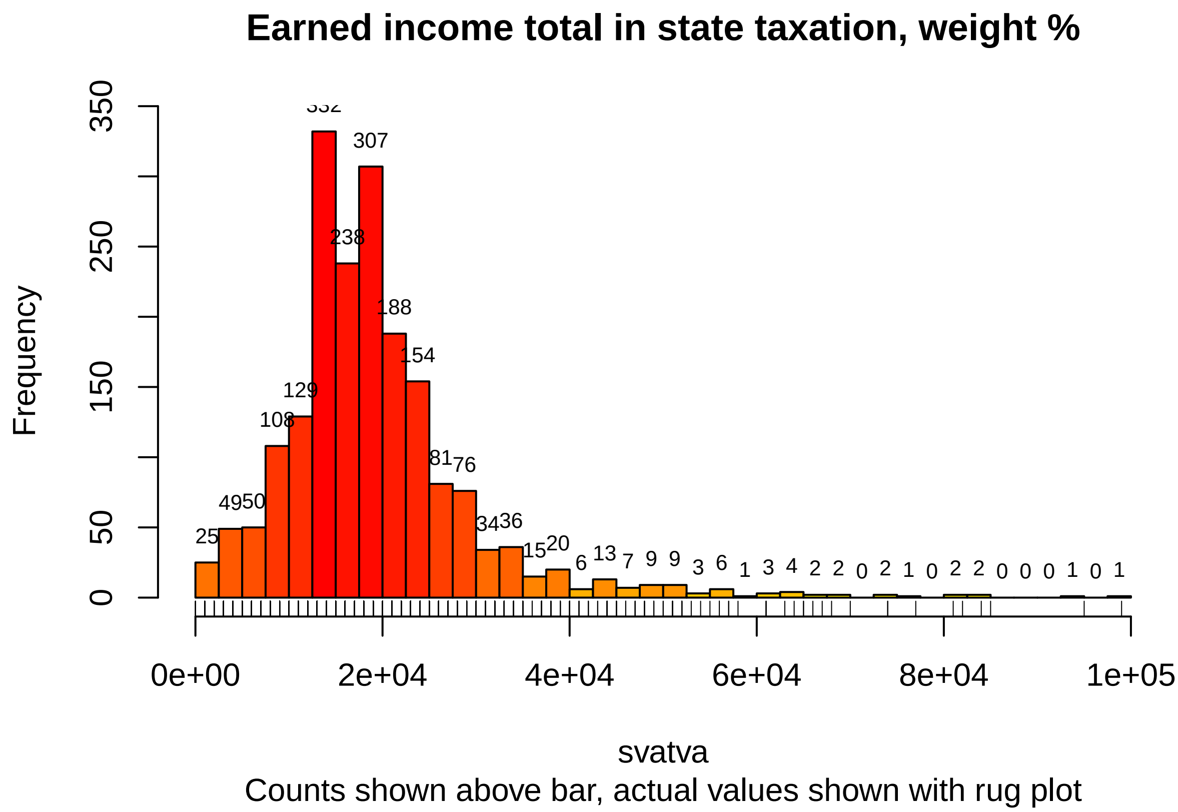
```
stem(svatva)
```

```
##
```

```
##   The decimal point is 4 digit(s) to the right of the |
##
##   0 | 0000111111111222222222222333333333333334444444444444
##   0 | 55555555555555555555555666666666666666666666667777777777777777777777+61
##   1 | 00000000000000000000000000000000000000011111111111111111111111111111+283
##   1 | 55555555555555555555555555555555555555555555555555555555555555555555+525
##   2 | 00000000000000000000000000000000000000000000000000000000000000000000+295
##   2 | 55555555555555555555555555555555555555555556666666666666666666666666+97
##   3 | 00000000000000000000000011111111111111112222222222222222233333333333+5
##   3 | 555555555666677777777777888889999999
##   4 | 0000000011122233444444
##   4 | 5555566667778889999
##   5 | 0011112222234
##   5 | 56666778
##   6 | 11134
##   6 | 55678
##   7 | 044
##   7 | 7
##   8 | 124
##   8 | 5
##   9 |
##   9 | 59
```

The stem-and-leaf plot suggests that comparing to normal distribution this one is skewed towards positive values. Next the distribution is shown in more details using a **histogram plot**.

```r
h<-hist(svatva, breaks=seq(0, 101000, by=2500), plot=F)
#str(h)

plot(h, col = heat.colors(length(h$mids))[length(h$count)-rank(h$count)+1],
     ylim = c(0, max(h$count)+5),
     main="Earned income total in state taxation, weight %",
     sub="Counts shown above bar, actual values shown with rug plot")
rug(svatva)
#cex - size of the text on the figure
text(h$mids, h$count, h$count, cex=0.7, cex.main =0.7, cex.sub=0.7, pos=3)
```




```r
rm(h)
```

We see that there are some unusually high values on the right tail of the histogram. Next we investigate this part of the histogram, the employees who have earned total income **svatva**>48500.

```r
data[svatva>48500,]
```

```
##     syrtun vuosi.x shtun sukup syntyv kieli peas a71km a181km ktutk sose
## 20     117       2  5371     1   1940    fi    1     0      0    74   NA
## 22     135       2  4018     1   1944    fi    1     0      0    54   NA
## 79     386       2   836     1   1926    fi    1     0      0    NA   NA
## 98     505       2  3020     1   1954    fi    1     1      5    63   NA
## 109    553       2  1269     1   1941    fi    1     0      1    43   NA
## 134    655       2  8248     1   1946    fi    1     1      2    74   NA
## 170    819       2  5176     1   1946    fi    1     0      1    44   NA
## 194    966       2  2112     2   1953    fi    0  <NA>   <NA>    76   NA
## 283   1405       2  4721     1   1946    fi    1     0      2    40   NA
## 327   1538       2  5202     1   1938    fi    4     0      0    84   NA
## 417   1960       2  3345     2   1945    fi    1     0      0    NA   NA
## 438   2084       2  4533     1   1932    fi    1     0      0    NA   NA
## 484   2373       2   217     1   1945    fi    1     0      1    73   NA
## 519   2558       2  8343     1   1954    fi    0  <NA>   <NA>    43   NA
## 612   2975       2  1338     1   1930    fi    1     0      0    74   NA
## 753   3653       2  5132     1   1941    sv    0  <NA>   <NA>    74   NA
## 807   3872       2  5699     1   1925    fi    1     0      0    NA   NA
## 811   3893       2  2083     1   1947    fi    1     0      2    43   NA
## 841   4062       2  1500     2   1942    fi    2     0      0    63   NA
```

```
## 880    4247        2  7485      1   1945    fi    1     0      2    44   NA
## 944    4626        2   130      1   1935    fi    1     0      0    54   NA
## 977    4850        2  5574      1   1951    fi    1     0      2    74   NA
## 988    4931        2  1604      2   1945    fi    0  <NA>   <NA>    43   NA
## 990    4940        2  7784      1   1944    fi    1     1      2    63   NA
## 996    4963        2     7      2   1952    fi    0  <NA>   <NA>    63   NA
## 1236   5850        2   282      1   1951    fi    1     0      1    34   NA
## 1241   5879        2  7405      1   1939    fi    1     0      0    63   NA
## 1360   6226        2   865      1   1944    fi    1     0      3    63   NA
## 1365   6226        2  1231      1   1944    fi    1     0      1    76   NA
## 1379   6328        2  7281      1   1941    fi    1     0      0    63   NA
## 1381   6332        2  8010      1   1940    fi    1     0      1    74   NA
## 1412   6505        2  7929      1   1946    fi    1     0      2    NA   NA
## 1426   6537        2  3260      1   1949    fi    4     0      2    43   NA
## 1443   6618        2  1307      1   1947    sv    1     0      1    43   NA
## 1450   6637        2  1296      2   1936    fi    2     0      0    76   NA
## 1497   6894        2   768      1   1953    fi    1     0      0    43   NA
## 1599   7400        2  2852      1   1931    fi    0  <NA>   <NA>    73   NA
## 1616   7485        2  2350      1   1936    sv    1     0      0    NA   NA
## 1625   7533        2  6662      1   1947    fi    1     0      1    34   NA
## 1633   7581        2  6999      1   1941    fi    1     0      3    44   NA
## 1738   8135        2  7438      1   1950    fi    1     0      3    74   NA
## 1746   8135        2  3246      1   1946    fi    1     0      0    74   NA
## 1801   8486        2  5679      1   1943    fi    1     0      0    76   NA
## 1822   8591        2  1876      2   1940    fi    2     0      1    76   NA
## 1853   8733        2  3080      1   1948    fi    1     0      2    86   NA
##      ptoim1 tyokk tyke toimiala.x svatva tyotu suuralue12 vuosi.y oty
## 20       11    NA   NA          M  74000 69000          1       2   1
## 22       11    NA   NA          M  56000 56000          1       2   1
## 79       11    NA   NA          G  99000    NA          4       2   1
## 98       11    NA   NA          F  70000 69000          4       2   1
## 109      11    NA   NA          D  57000 56000          2       2   1
## 134      11    NA   NA          D  52000 50000          1       2   2
## 170      11    NA   NA          G  52000    NA          4       2   1
## 194      11    NA   NA          P  52000 52000          2       2   1
## 283      11    NA   NA          G  64000 64000          1       2   5
## 327      11    NA   NA          M  51000 49000          1       2   1
## 417      11    NA   NA          F  51000 41000          2       2   1
## 438      11    NA   NA          G  49000 48000          2       2   1
## 484      11    NA   NA          K  49000 49000          2       2   1
## 519      11    NA   NA          M  74000 56000          1       2   1
## 612      11    NA   NA          D  50000 50000          1       2   1
## 753      11    NA   NA          D  56000 56000          2       2   1
## 807      11    NA   NA          A  63000    NA          2       2   1
## 811      11    NA   NA          L  61000 31000          2       2   1
## 841      11    NA   NA          D  51000 50000          1       2   1
## 880      11    NA   NA          F  50000    NA          1       2   1
## 944      11    NA   NA          D  53000 52000          3       2   1
## 977      11    NA   NA          M  56000 55000          1       2   1
## 988      11    NA   NA          G  56000 56000          1       2   5
## 990      11    NA   NA          M  55000 54000          1       2   1
## 996      11    NA   NA          M  67000 67000          1       2   5
## 1236     11    NA   NA          F  52000    NA          1       2   1
## 1241     11    NA   NA          G  57000 57000          1       2   5
```

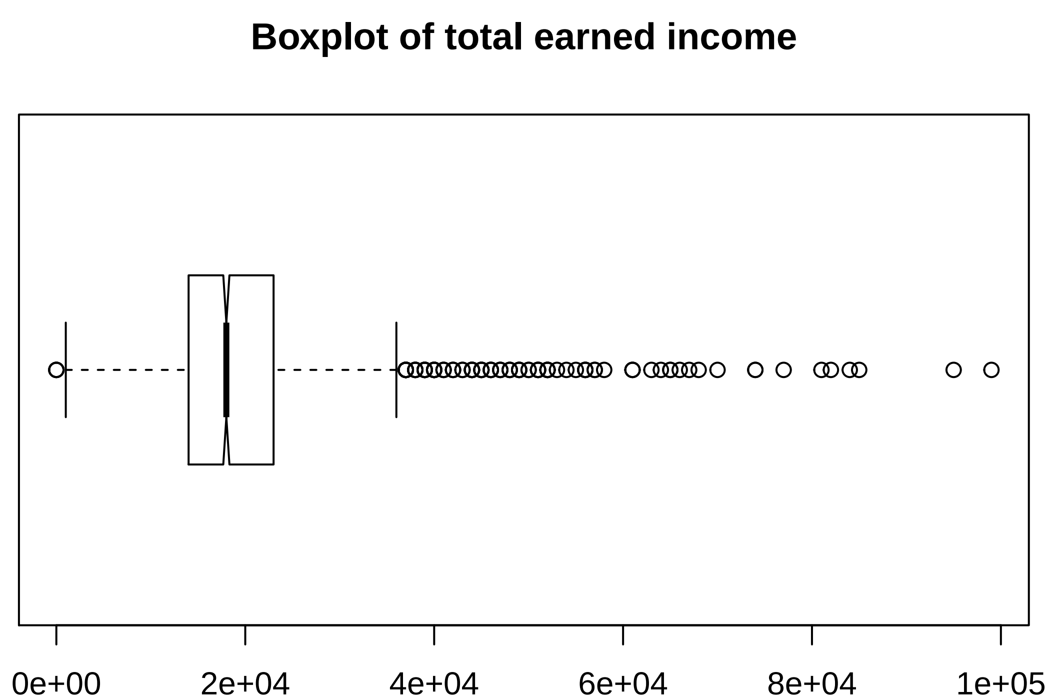
```
## 1360     11    NA   NA          D  66000 65000          1       2   2
## 1365     11    NA   NA          E  85000 61000          4       2   2
## 1379     11    NA   NA          M  61000 61000          1       2   1
## 1381     11    NA   NA          M  77000 40000          1       2   1
## 1412     11    NA   NA          K  52000 46000          1       2   1
## 1426     11    NA   NA          G  49000 40000          2       2   1
## 1443     11    NA   NA          G  95000 69000          1       2   1
## 1450     11    NA   NA          P  51000    NA          2       2   1
## 1497     11    NA   NA          D  49000 48000          4       2   1
## 1599     11    NA   NA          M  68000 58000          1       2   1
## 1616     11    NA   NA          I  65000    NA          1       2   1
## 1625     11    NA   NA          G  54000 52000          1       2   5
## 1633     11    NA   NA          F  58000 58000          1       2   1
## 1738     11    NA   NA          D  84000 84000          1       2   2
## 1746     11    NA   NA          D  61000 61000          1       2   2
## 1801     11    NA   NA          G  82000    NA          4       2   1
## 1822     11    NA   NA          G  81000    NA          1       2   1
## 1853     11    NA   NA          D  65000 62000          2       2   1
##      toimiala.y SLHKY sllvy
## 20            M     4     6
## 22            M     5     6
## 79            G     2     5
## 98            F     4     5
## 109           D     4     5
## 134           D     9     9
## 170           G     3     6
## 194           P     5     5
## 283           G     5     7
## 327           M     1     1
## 417           F     5     6
## 438           G     7     9
## 484           K     5     6
## 519           M     4     5
## 612           D     9     9
## 753           D     9     9
## 807           A     1     2
## 811           L     9     7
## 841           D     9     8
## 880           F     1     3
## 944           D     9     8
## 977           M     3     5
## 988           G     4     7
## 990           M     2     5
## 996           M     4     6
## 1236          F     1     3
## 1241          G     5     7
## 1360          D     9     9
## 1365          D     9     9
## 1379          M     7     7
## 1381          M     3     5
## 1412          K     9     8
## 1426          G     4     6
## 1443          G     2     5
## 1450          P     1     3
```

```
## 1497          D     6     6
## 1599          M     2     4
## 1616          I     1     4
## 1625          G     3     6
## 1633          F     9     9
## 1738          D     9     9
## 1746          D     9     9
## 1801          G     2     5
## 1822          G     3     5
## 1853          D     9     9
```

We observe that these are 45 observations, which are consistent with the total earned income of other employees. There is no evidence for a data entry error. The interesting observation is that 34 of these employees work in private domestic companies, and 42 are native Finnish speakers (other 3 are Swedish speaking). Only 7 of them are women.

We continue to investigate this distribution using **other exploratory graphics**.

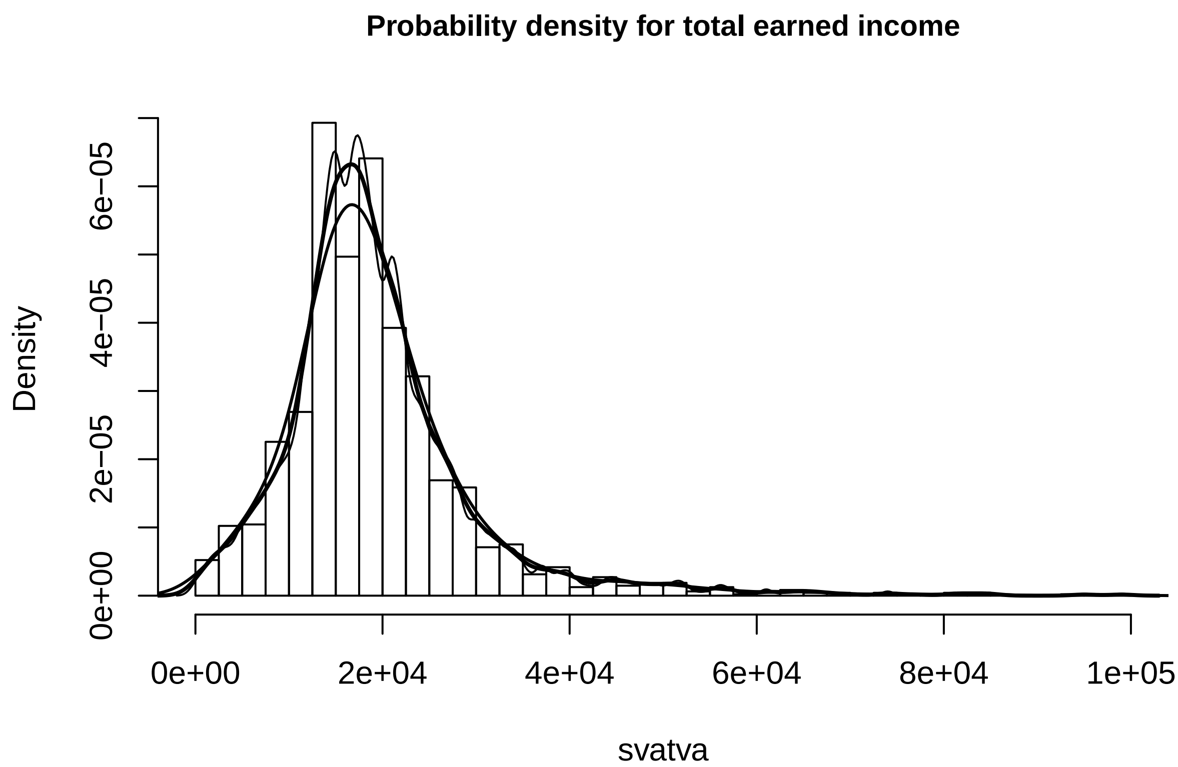
```r
boxplot(svatva, notch=T, horizontal=T,main="Boxplot of total earned income")
```



This **boxplot** does not show clear evidence that the distribution is right skewed. The employees with total earned income higher than 10 000 are shown as outliers (more than the estimated maximum of the distribution (1.5 times IQR larger than third quartile)). Also there is one outlier from the left (less than estimated minimum of the distribution (1.5 times IQR smaller than first quartile)). This does not means that these aoutliers are not part from the investigated population.

In order to be able to compare the histograms with other histograms, using different number of observations, we use **probability density** instead frequencies when plot the histogram. These densities are computed using 3 different kernel density estimators, which are smoothed continuous approximation to the histogram.

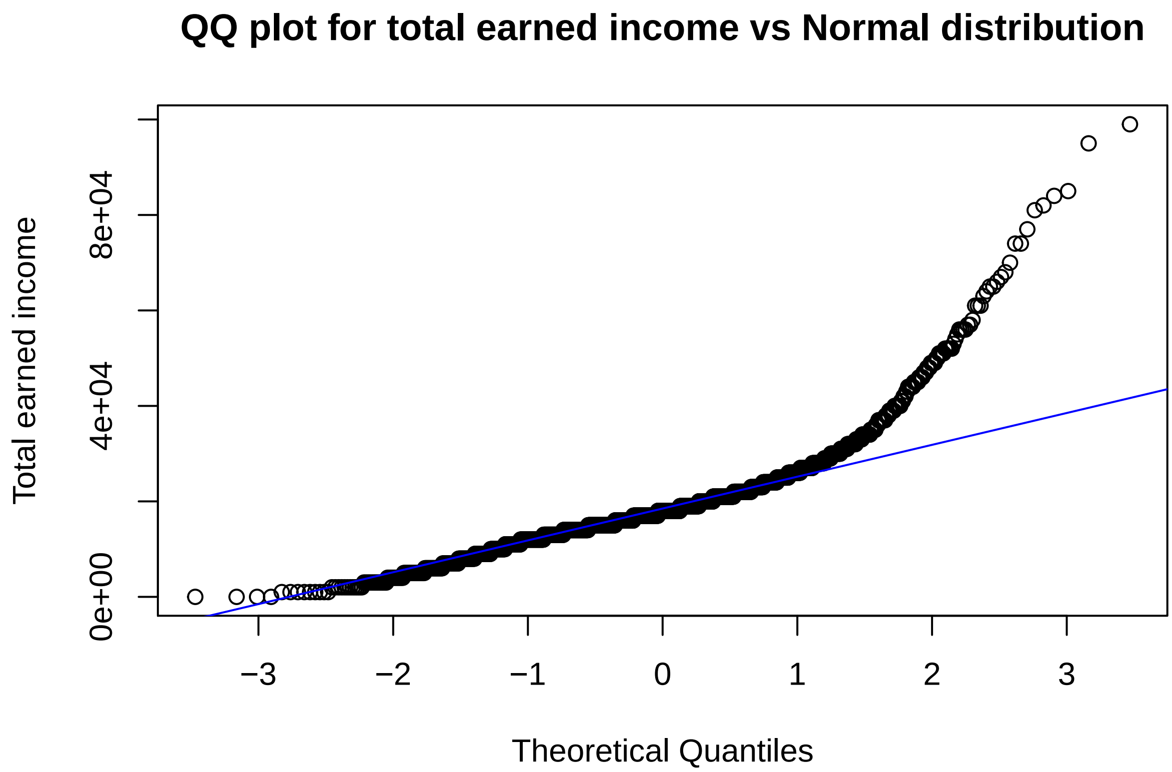
```r
hist(svatva, freq=F, breaks=seq(0, 101000, by=2500), cex.main =0.9, main="Probability density for total
#density - computes kernel density estimates.
#These are smoothed continuous approximation to the histogram.
lines(density(svatva),lwd=2)
lines(density(svatva, adj=.5),lwd=1)
lines(density(svatva, adj=2),lwd=1.5)
```



We again see that this distribution has quite long right tail.

Next we compare the actual distribution to the theoretical one using **quantile-quantile plot**.

```r
qqnorm(svatva, main="QQ plot for total earned income vs Normal distribution",ylab="Total earned income")
qqline(svatva, col=4)
```



This plot shows clearly that the distribution is not normal, especially at the right tail where the values are too high.

Point estimation. Inference of the mean. We have used *summary()* function to compute the descriptive statistics, including also the median and mean of this sample. Here we can try to make some inferences about the population mean (point estimate). For such small sample we assume t-distribution. For the null hyphotesis H0 here we assume the mean mu= 18000. We set confidence interval 99%.

```
t.test(svatva, mu=18000, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  svatva
## t = 6.059, df = 1915, p-value = 1.645e-09
## alternative hypothesis: true mean is not equal to 18000
## 99 percent confidence interval:
##  18823.60 20043.84
## sample estimates:
## mean of x
##  19433.72
```

The best estimate of the mean is 19433.72 for total earned income of the company. With only 1% to be wrong we state that the true mean is between 18823.60 and 20043.84 total earned income. Since the p-value is very small (almost 0), we cannot reject H0.

Conclusion: We observe clear deviation from normal distribution of total earned income. Therefore we expect problems in the further multivariate analysis.

Univariate exploratory data analysis of the group of the company according its turnover sllvy and family status peas.

We are interested to find the relative frequencies of different categories in both variables. For this purpose we first count the number of occurrences of each type using the *table* function, and next compute the proportions of the different classes.

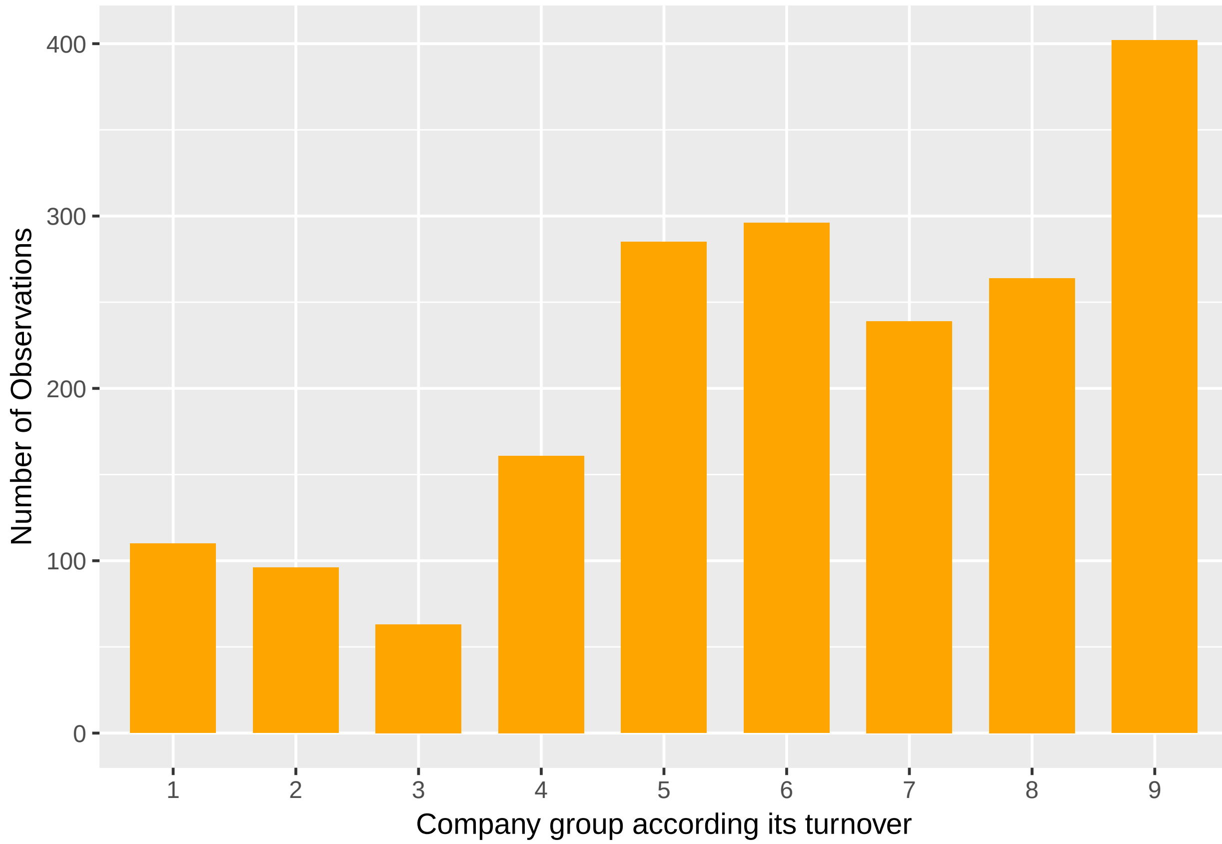
Analysis of sllvy - company groups according their turnover.

```
counts_sllvy <- table(sllvy)
proportions_sllvy<-counts_sllvy / sum(counts_sllvy)
proportions_sllvy
```

```
## sllvy
##          1          2          3          4          5          6
## 0.05741127 0.05010438 0.03288100 0.08402923 0.14874739 0.15448852
##          7          8          9
## 0.12473904 0.13778706 0.20981211
```

The most common category corresponds to the sample mode, which in this case is the category 9, companies with turnover >= 200 000 000. Since the values are nominal, is is not very correct to compute the median directly from them. Next we display the counts on bar plot:

```
library(ggplot2)
library(GGally)
bar_sllvy <- ggplot(data, aes(x = sllvy))
bar_sllvy <- bar_sllvy + geom_bar()
#summary(bar_sllvy)
ggplot(data, aes(x = sllvy)) +
  geom_bar(fill = "orange", width = 0.7) +
  xlab("Company group according its turnover") + ylab("Number of Observations")
```



###Analysis of family status **peas**.

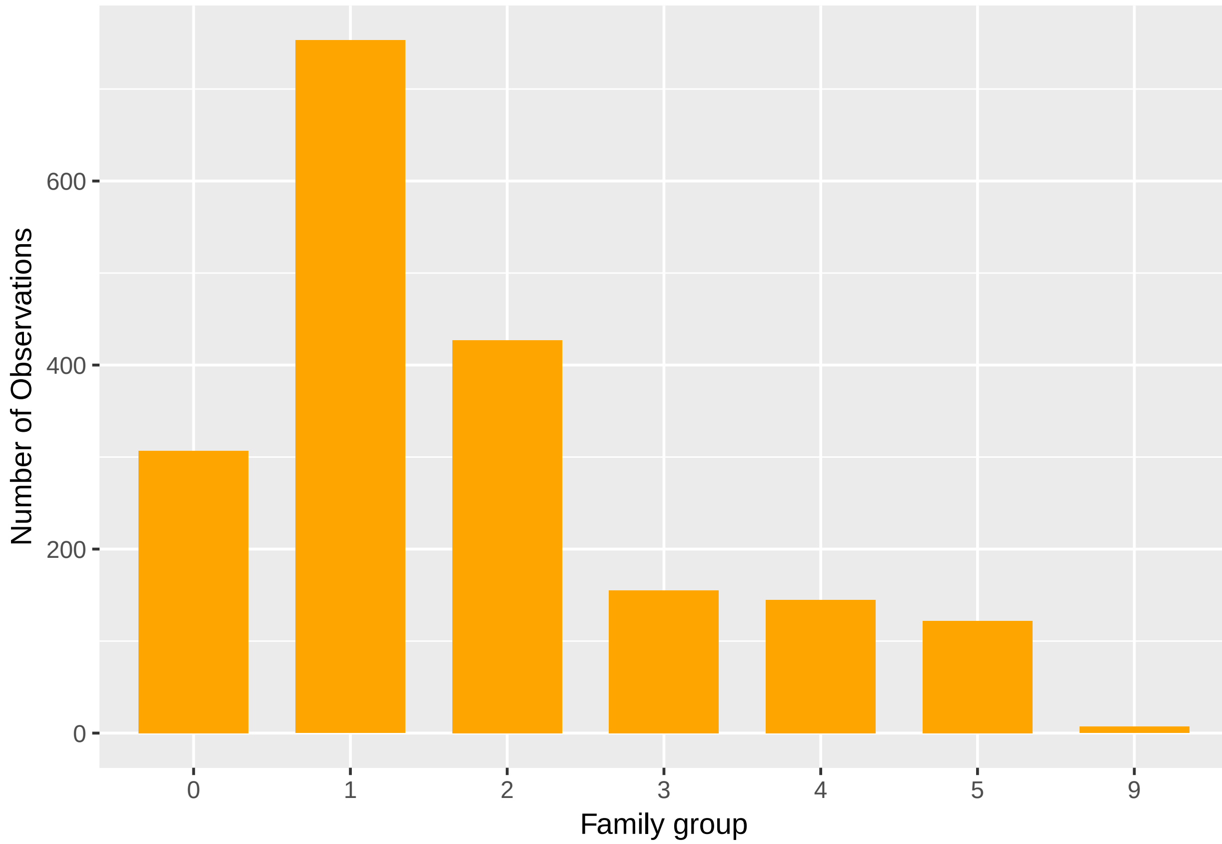
```
counts_peas <- table(peas)
proportions_peas<-counts_peas / sum(counts_peas)
proportions_peas
```

```
## peas
##           0           1           2           3           4           5
## 0.160229645 0.393006263 0.222860125 0.080897704 0.075678497 0.063674322
##           9
## 0.003653445
```

We see that the most common category (and also a sample mode) is group 1, which codes the man in the family who is also a had of the family.

The bar plot is visualized as:

```
bar_peas <- ggplot(data, aes(x = peas))
bar_peas <- bar_peas + geom_bar()
#summary(bar_peas)
ggplot(data, aes(x = peas)) +
  geom_bar(fill = "orange", width = 0.7) +
  xlab("Family group") + ylab("Number of Observations")
```



Bivariate data analysis: Total earned income vs. company turnover.

Exploratory data analysis.

Before to summarize and graph these data we first look them carefully and try to clarify which total earned incomes are associated with different groups of company turnover. For this purpose we first *sort* the observations and use *order* function to keep the indexes of sorted vector.

```
svatva[order(svatva)][1:100]
```

```
##   [1]    0    0    0    0 1000 1000 1000 1000 1000 1000 1000 1000 1000 2000
##  [15] 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 3000 3000 3000
##  [29] 3000 3000 3000 3000 3000 3000 3000 3000 3000 3000 3000 4000 4000 4000
##  [43] 4000 4000 4000 4000 4000 4000 4000 4000 4000 4000 5000 5000 5000 5000
##  [57] 5000 5000 5000 5000 5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
##  [71] 5000 5000 5000 5000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000
##  [85] 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000
##  [99] 7000 7000
```

```
sllvy[order(svatva)] [1:100]
```

```
##   [1] 3 3 4 2 6 1 7 4 3 2 2 7 5 6 8 1 8 9 2 3 3 5 2 5 1 5 4 5 9 6 6 2 6 4 4
##  [36] 1 1 8 1 4 2 4 4 5 9 9 9 2 4 2 5 4 4 9 8 2 4 2 6 1 1 4 9 2 4 1 7 2 2 2
##  [71] 3 2 7 5 4 4 8 6 7 4 2 8 4 6 2 9 4 1 2 4 4 2 3 1 2 3 5 9 6 9
## Levels: 1 2 3 4 5 6 7 8 9
```

It is difficult to make some conclusions only by looking these values. Next we use *by* method to compute some statistics for every level of the categorical variable. We compute also length of the range of income in every company turnover group.

```
by(svatva,sllvy,range)
```

```
## sllvy: 1
## [1]  1000 51000
## --------------------------------------------------------
## sllvy: 2
## [1]     0 63000
## --------------------------------------------------------
## sllvy: 3
## [1]     0 52000
## --------------------------------------------------------
## sllvy: 4
## [1]     0 68000
## --------------------------------------------------------
## sllvy: 5
## [1]  1000 99000
## --------------------------------------------------------
## sllvy: 6
## [1]  1000 74000
## --------------------------------------------------------
## sllvy: 7
## [1]  1000 64000
## --------------------------------------------------------
## sllvy: 8
## [1]  2000 53000
## --------------------------------------------------------
## sllvy: 9
## [1]  2000 85000
```

```
by(svatva,sllvy,function(x) max(x)-min(x))
```

```
## sllvy: 1
## [1] 50000
## --------------------------------------------------------
## sllvy: 2
## [1] 63000
## --------------------------------------------------------
## sllvy: 3
## [1] 52000
## --------------------------------------------------------
## sllvy: 4
## [1] 68000
## --------------------------------------------------------
## sllvy: 5
## [1] 98000
## --------------------------------------------------------
## sllvy: 6
## [1] 73000
## --------------------------------------------------------
## sllvy: 7
## [1] 63000
```

```
## ----------------------------------------------------------
## sllvy: 8
## [1] 51000
## ----------------------------------------------------------
## sllvy: 9
## [1] 83000
```

The number of observations in every group also can be obtained using *by* method:

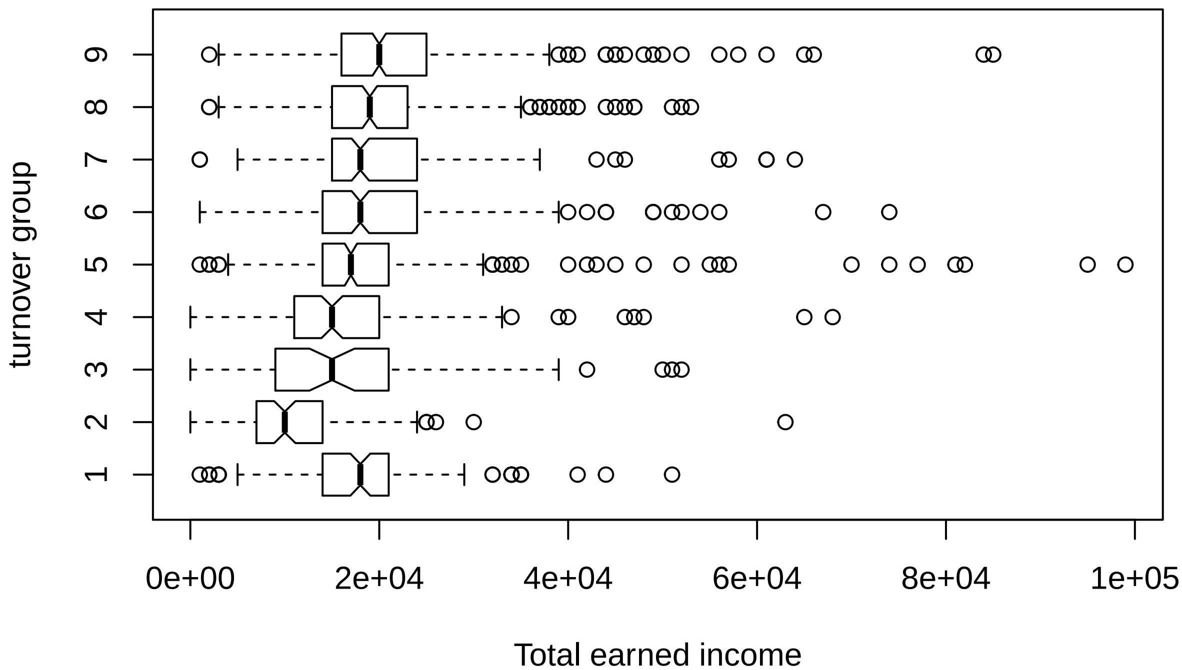
```r
by(svatva,sllvy,length)
```

```
## sllvy: 1
## [1] 110
## ----------------------------------------------------------
## sllvy: 2
## [1] 96
## ----------------------------------------------------------
## sllvy: 3
## [1] 63
## ----------------------------------------------------------
## sllvy: 4
## [1] 161
## ----------------------------------------------------------
## sllvy: 5
## [1] 285
## ----------------------------------------------------------
## sllvy: 6
## [1] 296
## ----------------------------------------------------------
## sllvy: 7
## [1] 239
## ----------------------------------------------------------
## sllvy: 8
## [1] 264
## ----------------------------------------------------------
## sllvy: 9
## [1] 402
```

This corresponds to the bar graph visualization during univariate analysis of **sllvy**.

Next we use boxplot visualization, where the incomes are divided by the different factors of companies turnover. The attribute *notch=T* on this plot shows if the class medians are significantly different.

```r
boxplot(svatva~sllvy, notch=T, horizontal=T, xlab="Total earned income", ylab="turnover group")
```



The results shown on this box plot are somehow surprising. It seems that there is not too big differences in the total earned income of different employees and the size of turnover of the company where they work. There are also not too large differences in the average income of employees in different groups. The strange thing is that in the lowest turnover group 1 the mean income is higher comparing to groups 2, 3 and 4. The mean income increases in groups 5, 6, 7, 8 and 9. When increase the group number (especially in group 5), the differences reflect in more outliers from right. When we compare this plot with the number of observations in every group shown above, we see that groups 1, 3 and 8 are significantly under-represented (about twice less then group 5). The groups 3, 6 and 9 have the widest ranges, while the groups 1 and 2 have the lowest ranges. In all groups the distributions are somehow skewed, negatively or positively. Only in group 3 the distribution seems to be symmetric.

A numerical conformation of the results shown on box plots could be obtained using summaries about different groups in *by* method:

```
by(svatva,sllvy,summary)
```

```
## sllvy: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000   14000   18000   18209   21000   51000
## --------------------------------------------------------
## sllvy: 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    7000   10000   11677   14000   63000
## --------------------------------------------------------
## sllvy: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    9000   15000   17016   21000   52000
```

```
## ------------------------------------------------------------
## sllvy: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   11000   15000   16453   20000   68000
## ------------------------------------------------------------
## sllvy: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000   14000   17000   19765   21000   99000
## ------------------------------------------------------------
## sllvy: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000   14000   18000   20101   24000   74000
## ------------------------------------------------------------
## sllvy: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000   15000   18000   20234   24000   64000
## ------------------------------------------------------------
## sllvy: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000   15000   19000   20462   23000   53000
## ------------------------------------------------------------
## sllvy: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000   16000   20000   21316   25000   85000
```

One-way analysis of variance (ANOVA)

The simplest ANOVA is one-way, where the total variance of the data is compared to the residual variance after each observation's value is adjusted for the mean for the one factor. Here the question is how big proportion of employees with their corresponding total income varies among the 9 companies group (divided by their turnover). In R the method, use for ANOVA, is *lm*. It is used also for linear regression. ANOVA is just another form of the same linear modeling, as it is shown in [4].

```r
lm_an<-lm(svatva~sllvy)
summary(lm_an)
```

```
##
## Call:
## lm(formula = svatva ~ sllvy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19316  -5462  -2101   2899  79235
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18209.1      965.8  18.853  < 2e-16 ***
## sllvy2       -6532.0     1414.8  -4.617 4.16e-06 ***
## sllvy3       -1193.2     1600.5  -0.746  0.45605
## sllvy4       -1755.7     1253.1  -1.401  0.16135
## sllvy5        1555.8     1137.1   1.368  0.17138
## sllvy6        1892.3     1131.2   1.673  0.09452 .
## sllvy7        2025.2     1167.1   1.735  0.08287 .
## sllvy8        2253.0     1149.6   1.960  0.05016 .
```

```
## sllvy9        3106.8     1090.0   2.850  0.00442 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10130 on 1907 degrees of freedom
## Multiple R-squared:  0.0475, Adjusted R-squared:  0.04351
## F-statistic: 11.89 on 8 and 1907 DF,  p-value: < 2.2e-16
```

The *Intersept* corresponds to the mean for Group 1 in **sllvy**. The mean estimates for other groups are shown in corresponding coefficients in column Estimate as difference the corresponding group and intercept (mean for group 1). This summary shows only first two groups as important, and group 9 as quite significant. The value of Adjusted R-squared is 0.043, which means that only 4.3% of total variation is explained by **sllvy**. Since p-value< 2.2e-16 this means that the probability this amount of variability to be explained by chance is practically 0.

Conclusion: There is no strong direct relationship between total earned income of employees and company turnover groups.

Bivariate data analysis: Total earned income vs. family status.

We repeat the same pipeline to investigate bivariate relationship between total earned income **svatva** and family status **peas**.

Exploratory data analysis.

```
svatva[order(svatva)][1:100]
```

```
##   [1]    0    0    0    0 1000 1000 1000 1000 1000 1000 1000 1000 1000 2000
##  [15] 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 3000 3000 3000
##  [29] 3000 3000 3000 3000 3000 3000 3000 3000 3000 3000 3000 4000 4000 4000
##  [43] 4000 4000 4000 4000 4000 4000 4000 4000 4000 4000 5000 5000 5000 5000
##  [57] 5000 5000 5000 5000 5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
##  [71] 5000 5000 5000 5000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000
##  [85] 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000
##  [99] 7000 7000
```

```
peas[order(svatva)] [1:100]
```

```
##   [1] 4 1 4 1 3 1 3 0 4 1 1 3 3 3 3 2 2 3 1 3 2 3 2 3 3 3 1 2 3 3 3 2 3 3 3
##  [36] 2 3 3 3 3 2 3 5 1 2 2 3 1 1 2 3 3 2 1 3 1 0 3 3 5 2 3 2 4 2 2 3 3 4 2
##  [71] 1 5 3 2 3 2 2 9 3 2 1 3 2 4 2 5 3 3 1 2 0 3 3 2 3 5 3 5 3 3
## Levels: 0 1 2 3 4 5 9
```

As before, here is also difficult to make some conclusions only by looking these values. Next apply *by* method.

```
by(svatva,peas,range)
```

```
## peas: 0
## [1]  1000 74000
## ---------------------------------------------------------
## peas: 1
## [1]     0 99000
## ---------------------------------------------------------
## peas: 2
```

```
## [1]  2000 81000
## ---------------------------------------------------------
## peas: 3
## [1]  1000 27000
## ---------------------------------------------------------
## peas: 4
## [1]     0 51000
## ---------------------------------------------------------
## peas: 5
## [1]  4000 37000
## ---------------------------------------------------------
## peas: 9
## [1]  6000 22000
```

```
by(svatva,peas,function(x) max(x)-min(x))
```

```
## peas: 0
## [1] 73000
## ---------------------------------------------------------
## peas: 1
## [1] 99000
## ---------------------------------------------------------
## peas: 2
## [1] 79000
## ---------------------------------------------------------
## peas: 3
## [1] 26000
## ---------------------------------------------------------
## peas: 4
## [1] 51000
## ---------------------------------------------------------
## peas: 5
## [1] 33000
## ---------------------------------------------------------
## peas: 9
## [1] 16000
```

Find out the number of observations, which fall in different family groups:

```
by(svatva,peas,length)
```

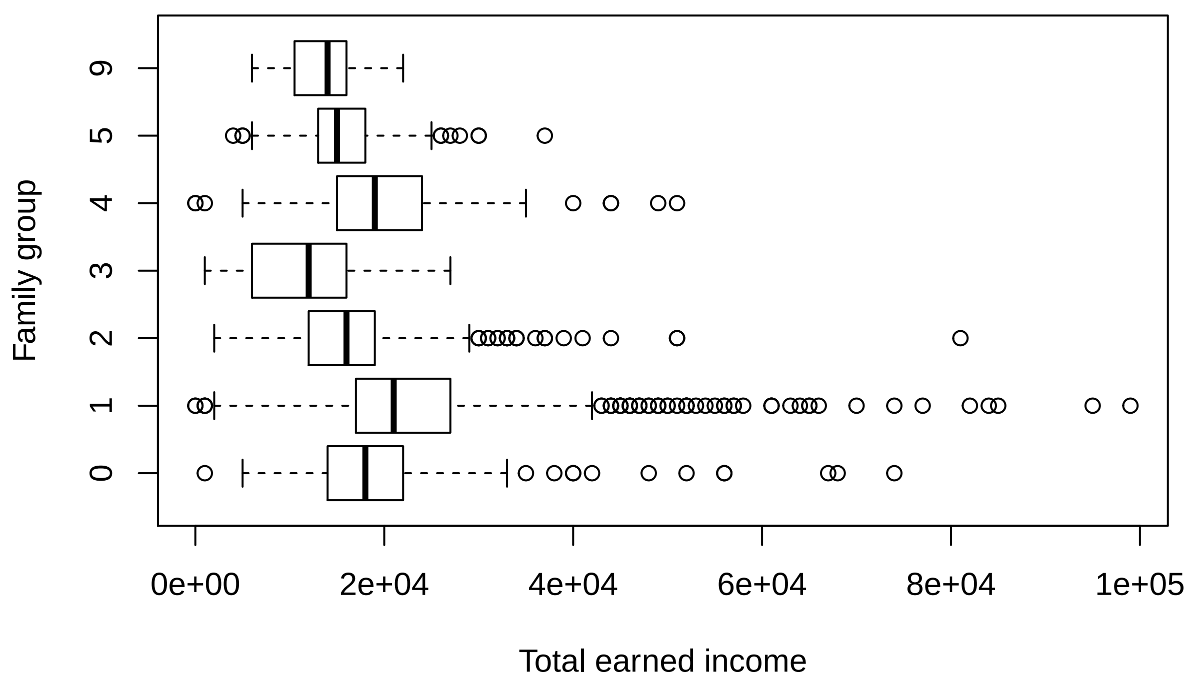
```
## peas: 0
## [1] 307
## ---------------------------------------------------------
## peas: 1
## [1] 753
## ---------------------------------------------------------
## peas: 2
## [1] 427
## ---------------------------------------------------------
## peas: 3
## [1] 155
## ---------------------------------------------------------
## peas: 4
## [1] 145
## ---------------------------------------------------------
```

```
## peas: 5
## [1] 122
## ------------------------------------------------------------
## peas: 9
## [1] 7
```

This corresponds to the bar graph visualization during univariate analysis of **peas**. Different groups are not equally represented. The biggest part fall in Group 1, 753 observations, followed by Group 2 (427 observations) and Group 0 (307 observations). The groups 3, 4 and 5 have comparable number of observations (155, 145 and 122), while the group 9 with unknown family status is presented only by 7 observations.

The box plot is shown below. Since some notches went outside hinges, the attribute *notch=T* is now set as *notch=F*.

```
boxplot(svatva~peas, notch=F, horizontal=T, xlab="Total earned income", ylab="Family group")
```



As it was expected, the highest average income in Group 1, for the heads of the families. It has the highest range, and the biggest number of outliers from right. The smallest range shows group 9. The range of group 5 is also small. All distributions are not symmetric.

Numerically the same box plot results are shown below:

```
by(svatva,peas,summary)
```

```
## peas: 0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000   14000   18000   18847   22000   74000
## ------------------------------------------------------------
## peas: 1
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   17000   21000   23578   27000   99000
## --------------------------------------------------------
## peas: 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000   12000   16000   16415   19000   81000
## --------------------------------------------------------
## peas: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    6000   12000   11871   16000   27000
## --------------------------------------------------------
## peas: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   15000   19000   19628   24000   51000
## --------------------------------------------------------
## peas: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4000   13000   15000   15615   18000   37000
## --------------------------------------------------------
## peas: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6000   10500   14000   13571   16000   22000
```

It seems that the family status has some influence on earned income. This statement is further examined using ANOVA.

One-way analysis of variance (ANOVA)

```r
lm_an_2<-lm(svatva~peas)
summary(lm_an_2)
```

```
##
## Call:
## lm(formula = svatva ~ peas)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23578  -5578  -1415   3385  75422
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18846.9      550.5  34.234  < 2e-16 ***
## peas1         4730.8      653.2   7.243 6.35e-13 ***
## peas2        -2432.4      721.8  -3.370 0.000767 ***
## peas3        -6975.9      950.5  -7.340 3.16e-13 ***
## peas4          780.7      972.0   0.803 0.421976
## peas5        -3232.2     1032.4  -3.131 0.001769 **
## peas9        -5275.5     3687.2  -1.431 0.152663
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9646 on 1909 degrees of freedom
## Multiple R-squared:  0.1354, Adjusted R-squared:  0.1327
```

```
## F-statistic: 49.83 on 6 and 1909 DF,  p-value: < 2.2e-16
```

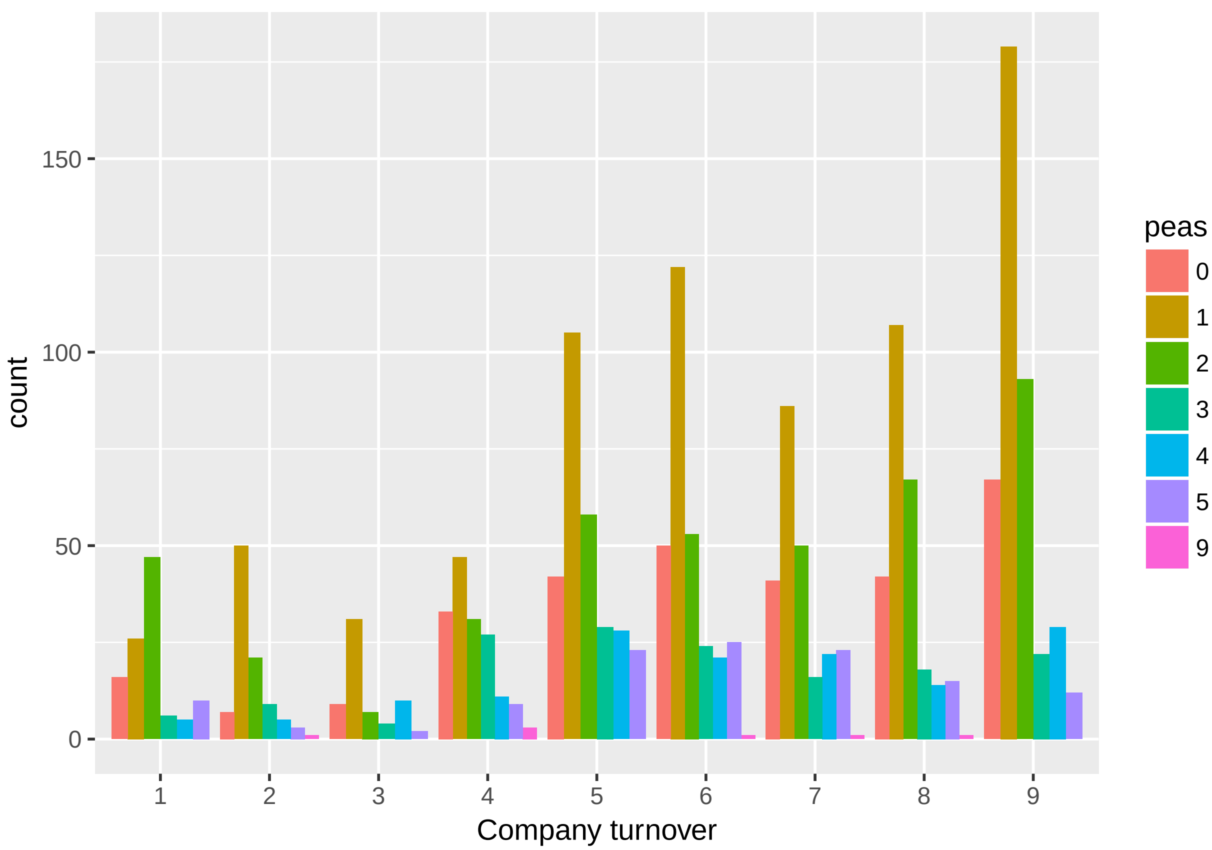
In this case the first 5 family groups (from 0 to 4) are shown to be significant, and also the family group 5 could be included in the model. The value of Adjusted R-squared now is 0.1327, which means that 13.27% of total variation is explained by **peas**. Since p-value< 2.2e-16 this means that the probability this amount of variability to be explained by chance is practically 0.

Conclusion: There is some relationship between total earned income of employees and family status. This relationship is not too strong since the proportion of explained variance is only 13.27%.

Bivariate analysis: visualization of the two categorical variables company turnover and family status

The *ggplot2* package in R allows visualization of two categorical variables using bar plots. We use this capability to visualize the company turnover and family status.

```
g0<-ggplot(data, aes(x = sllvy, fill = peas)) + geom_bar(position = "dodge")
g0 + xlab("Company turnover")
```



We see that except the company turnover group 1, in all other groups the biggest part of employees is form of family heads (family status 1). Next biggest proportion in most of the groups is the spouse (status 2). For the lowest turnover group 1 the proportion of spouses is highest. This most important family group is 3 = child.

Multivariate analysis.

Multivariate visualization

Box plot with multiple groups

The *ggplot2* package in R offers also visualization of the numerical predctor and two categorical variables on one plot. The total income (on y axes) for every company turnover group (on x axes) is shown as side-by-side box plots using different colors for different family statuses.
First generate frequency tables:

```r
table(sllvy, peas)
```

```
##      peas
## sllvy   0   1   2   3   4   5   9
##     1  16  26  47   6   5  10   0
##     2   7  50  21   9   5   3   1
##     3   9  31   7   4  10   2   0
##     4  33  47  31  27  11   9   3
##     5  42 105  58  29  28  23   0
##     6  50 122  53  24  21  25   1
##     7  41  86  50  16  22  23   1
##     8  42 107  67  18  14  15   1
##     9  67 179  93  22  29  12   0
```

Since the number of observations in every group is not equal, this is not balanced design.

At this stage, before to start actual ANOVA modeling, we compute some descriptive statiustics. Means of all groups:

```r
tapply(svatva,list(sllvy,peas),mean)
```

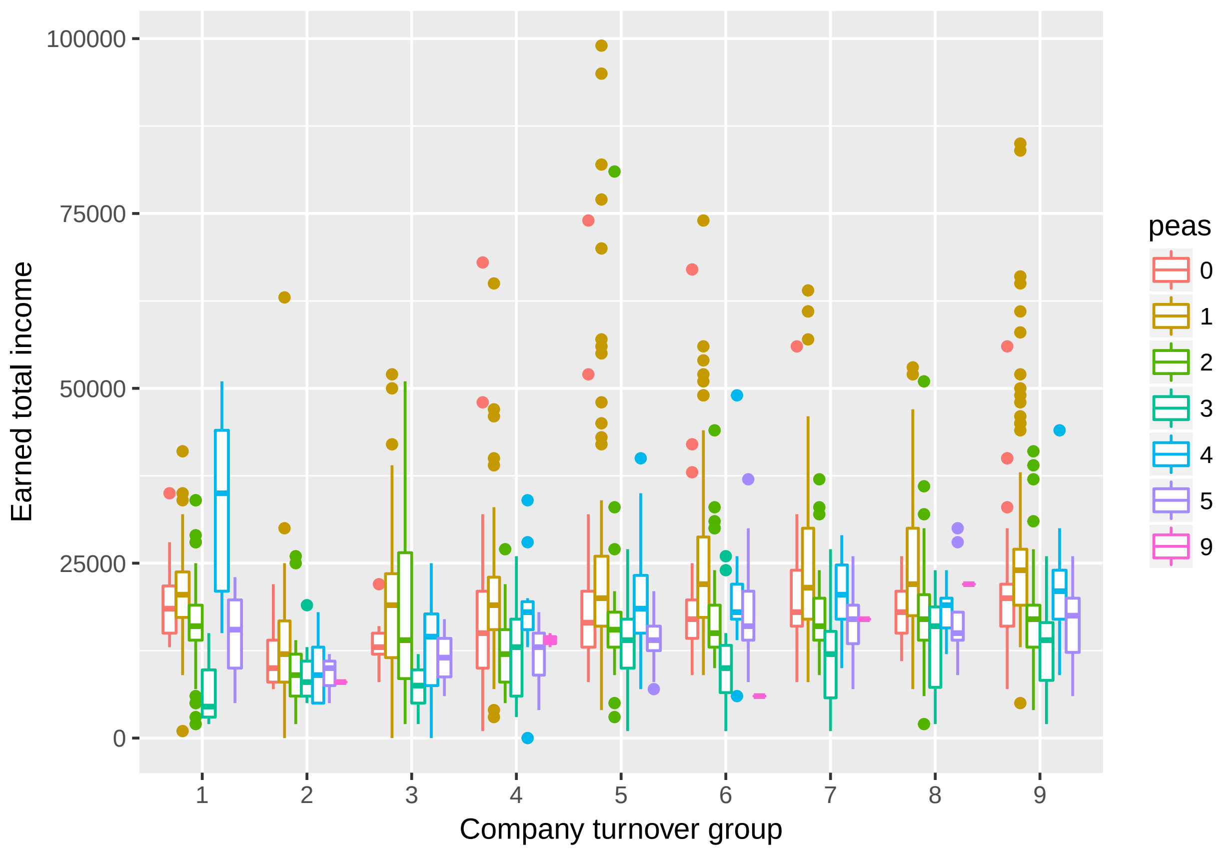
```
##          0        1        2         3        4        5     9
## 1 19687.50 21307.69 16595.74  6666.667 33200.00 14800.00    NA
## 2 11857.14 13160.00 10047.62  9333.333 10000.00  9000.00  8000
## 3 13444.44 20387.10 19571.43  7250.000 13000.00 11500.00    NA
## 4 17575.76 21255.32 12548.39 12444.444 18000.00 11666.67 14000
## 5 18976.19 24742.86 16637.93 13103.448 20142.86 14304.35    NA
## 6 18780.00 24418.03 17094.34 10291.667 20190.48 17960.00  6000
## 7 19975.61 24720.93 17280.00 12000.000 20000.00 16434.78 17000
## 8 18214.29 24785.05 17955.22 13944.444 18428.57 16733.33 22000
## 9 20402.99 25452.51 16709.68 12772.727 21241.38 16250.00    NA
```

Standard deviations:

```r
tapply(svatva,list(sllvy,peas),sd)
```

```
##           0         1         2        3         4        5    9
## 1  6139.693  8531.210  7033.026 5240.865 15139.353 6033.241   NA
## 2  5814.596  9792.459  6272.768 4663.690  5567.764 3605.551   NA
## 3  4126.473 12867.731 17232.306 4272.002  8666.667 7778.175   NA
## 4 12811.202 11333.805  5702.857 6710.115  8532.292 4387.482 1000
## 5 11595.811 16553.543  9809.945 6597.320  7998.677 3495.904   NA
## 6  9109.963 10589.906  6708.961 6300.305  8003.868 6711.185   NA
## 7  7900.911 11680.704  5671.411 7554.248  5209.881 5061.675   NA
## 8  3904.567  9659.318  6832.260 6999.767  3837.353 5535.169   NA
## 9  7563.989 11185.451  6233.815 6179.319  6550.110 6397.798   NA
```

```r
g1 <- ggplot(data, aes(x = sllvy, y = svatva, col=peas))
g1 + geom_boxplot() + xlab("Company turnover group")+ ylab("Earned total income")
```



We see that in the lowest turnover group 1 the highest mean income have employees with family status 4, heads of cohabiting family. It is interesting to observe that this is the hihgest mean income among all other turnover groups with different family statuses. ###Line plots with multiple groups

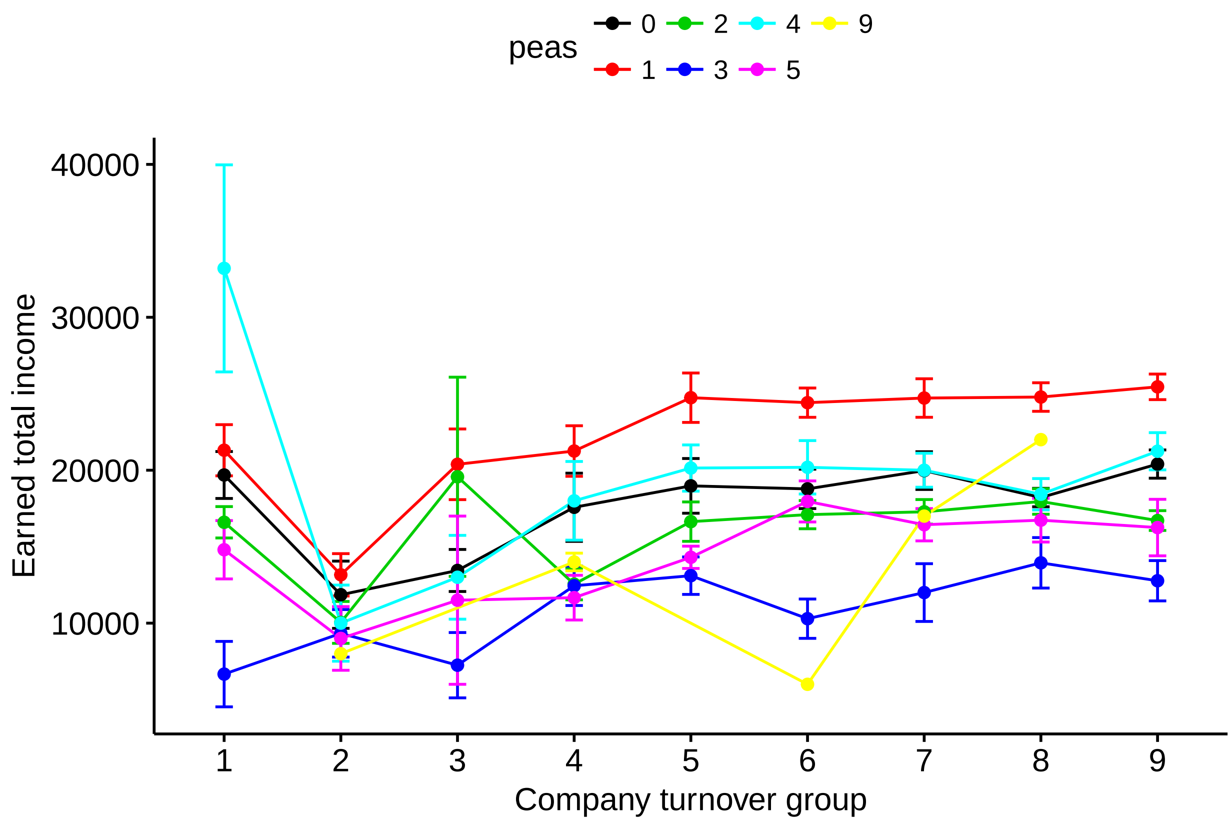
```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:GGally':
##
##     nasa

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggpubr)
```

```
## Loading required package: magrittr
```

```r
ggline(data, x = "sllvy", y = "svatva",
       xlab = "Company turnover group", ylab="Earned total income",
       color = "peas",
       add = "mean_se", palette = 1:7)
```



Here we see how the mean total income values are connected in different company turnover groups for different family statuses.

Two way ANOVA test to evaluate the effect of the two categorical variables sllvy and peas on a response variable svatva.

Since for these data we have the case of unbalanced design, there are 3 methods to apply two way ANOVA test in these data, namely Type I, Type II and Type III sum of squares [7].

We first examine the case without interactions, so called additive model. Here we also have to assume that the two categorical variables **sllvy** and **peas** are independent. For such additive models the Type II two-sided ANOVA method is recommended [7].

```r
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
anova_ind <- aov(svatva ~ sllvy + peas, data = data)
Anova(anova_ind, type = "II")
```

```
## Anova Table (Type II tests)
##
## Response: svatva
##               Sum Sq   Df F value    Pr(>F)
## sllvy     9.4266e+09    8  13.318 < 2.2e-16 ***
## peas      2.7487e+10    6  51.777 < 2.2e-16 ***
## Residuals 1.6820e+11 1901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From these results we can conclude that both company turnover group **sllvy** and family status **peas** are statistically significant with very small p-value (less than 2.2e-16). We expect that if there is no interaction between these variables, changing one of them will impact significantly the mean total income of employee.

Next we assume that there is a simultaneous effect of company turnover groups and family statuses and involve the interaction term in the model:

```r
#anova_inter <- aov(svatva ~ sllvy + peas + sllvy:peas, data = data)
#Anova(anova_inter, type = "III")
```

This code chunk produces the following error message:

"Error in Anova.III.lm(mod, error, singular.ok = singular.ok, ...) : there are aliased coefficients in the model"

This seems to be a warning about multicollinearity. Therefore we continue to analyze only the first model and compute some summary statistics.

Summary statistics

We compute the grand mean and the mean by groups:

```r
model.tables(anova_ind, type="means")
```

```
## Tables of means
## Grand mean
##
## 19433.72
##
##  sllvy
##         1     2     3     4     5     6     7     8     9
##     18209 11677 17016 16453 19765 20101 20234 20462 21316
## rep   110    96    63   161   285   296   239   264   402
##
##  peas
##         0     1     2     3     4     5     9
##     18666 23581 16394 12316 19572 15504 15600
## rep   307   753   427   155   145   122     7
```

These numbers are not much different from analogous ones during uni- ja bivariate analyses.

Multiple pairwise comparison between the means of the groups: Tukey Honest Significant Differences (THS)

A significant p-value in ANOVA two side tests means that some of the group means are different. Since we do not know which pairs of groups are different, we can clarify it by performing multiple pairwise comparison. In R this can be provided using the method TukeyHSD().

```r
TukeyHSD(anova_ind)
```

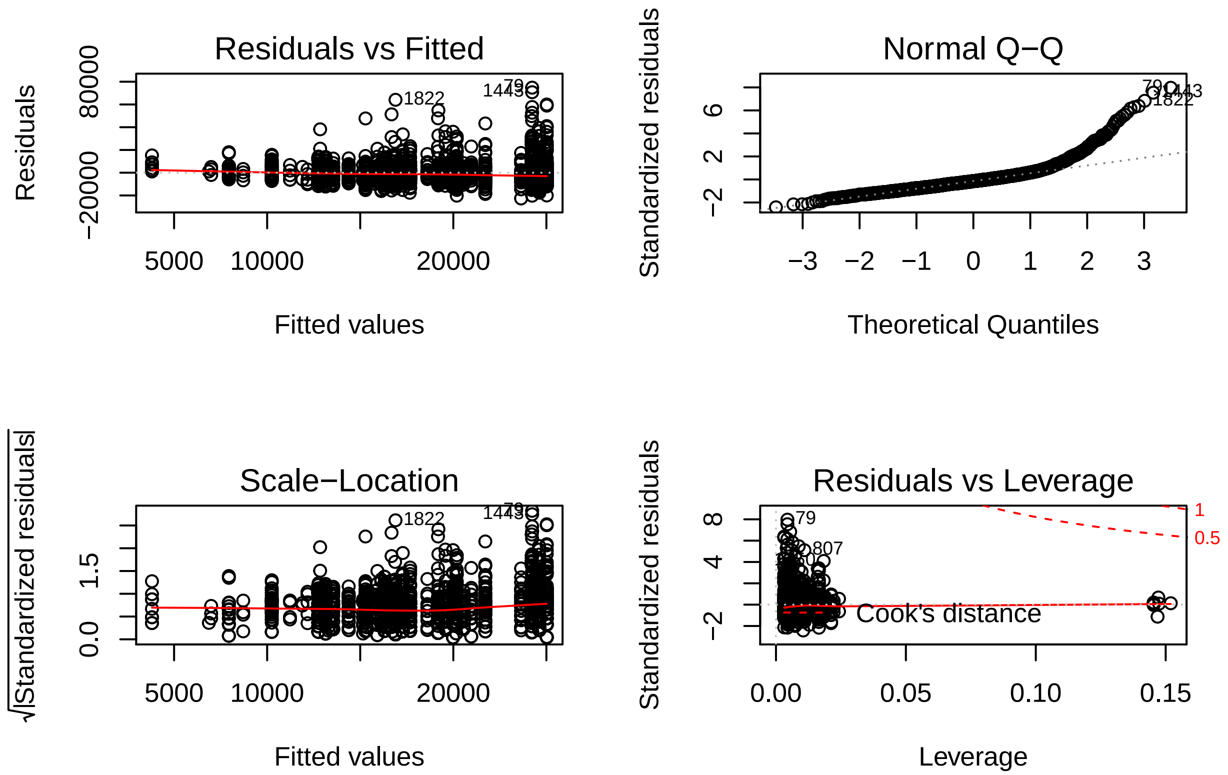
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = svatva ~ sllvy + peas, data = data)
##
## $sllvy
##           diff          lwr       upr     p adj
## 2-1 -6532.0076 -10611.71796 -2452.297 0.0000255
## 3-1 -1193.2179  -5808.34979  3421.914 0.9968070
## 4-1 -1755.6748  -5368.96666  1857.617 0.8516866
## 5-1  1555.8214  -1722.92225  4834.565 0.8678828
## 6-1  1892.2604  -1369.47326  5153.994 0.6812937
## 7-1  2025.2187  -1340.24648  5390.684 0.6355932
## 8-1  2253.0303  -1061.82960  5567.890 0.4662184
## 9-1  3106.8295    -36.23502  6249.894 0.0555466
## 3-2  5338.7897    602.69524 10074.884 0.0140182
## 4-2  4776.3328   1009.76588  8542.900 0.0027494
## 5-2  8087.8289   4640.90164 11534.756 0.0000000
## 6-2  8424.2680   4993.51666 11855.019 0.0000000
## 7-2  8557.2263   5027.70733 12086.745 0.0000000
## 8-2  8785.0379   5303.73865 12266.337 0.0000000
## 9-2  9638.8371   6320.70466 12956.969 0.0000000
## 4-3  -562.4569  -4903.24390  3778.330 0.9999813
## 5-3  2749.0393  -1317.49654  6815.575 0.4740587
## 6-3  3085.4783   -967.35530  7138.312 0.3045260
## 7-3  3218.4366   -918.33867  7355.212 0.2756386
## 8-3  3446.2482   -649.46296  7541.959 0.1817258
## 9-3  4300.0474    342.09233  8258.002 0.0215390
## 5-4  3311.4961    431.71326  6191.279 0.0109519
## 6-4  3647.9352    787.53375  6508.337 0.0025137
## 7-4  3780.8935    802.74881  6759.038 0.0027003
## 8-4  4008.7051   1087.86857  6929.542 0.0007152
## 9-4  4862.5042   2138.19807  7586.810 0.0000012
## 6-5   336.4391  -2087.64380  2760.522 0.9999683
## 7-5   469.3973  -2092.56030  3031.355 0.9997381
## 8-5   697.2089  -1797.89972  3192.318 0.9945378
## 9-5  1551.0081   -710.87727  3812.894 0.4533606
## 7-6   132.9583  -2407.19410  2673.111 1.0000000
## 8-6   360.7699  -2111.94408  2833.484 0.9999534
## 9-6  1214.5690  -1022.58821  3451.726 0.7553497
## 8-7   227.8116  -2380.20747  2835.831 0.9999991
## 9-7  1081.6108  -1304.24793  3467.469 0.8951240
## 9-8   853.7992  -1460.12863  3167.727 0.9670641
##
## $peas
```

```
##              diff        lwr        upr     p adj
## 1-0   4914.83416   3034.8788  6794.7895 0.0000000
## 2-0  -2272.51174  -4349.9407  -195.0828 0.0215216
## 3-0  -6350.65701  -9086.2232 -3615.0908 0.0000000
## 4-0    905.87564  -1891.6708  3703.4221 0.9631384
## 5-0  -3162.79075  -6134.0519  -191.5296 0.0283201
## 9-0  -3066.26861 -13678.5289  7545.9917 0.9791609
## 2-1  -7187.34590  -8869.2098 -5505.4820 0.0000000
## 3-1 -11265.49117 -13714.2210 -8816.7614 0.0000000
## 4-1  -4008.95852  -6526.7396 -1491.1775 0.0000573
## 5-1  -8077.62491 -10787.1171 -5368.1327 0.0000000
## 9-1  -7981.10278 -18523.0677  2560.8621 0.2774203
## 3-2  -4078.14527  -6681.5578 -1474.7328 0.0000823
## 4-2   3178.38738    509.9233  5846.8514 0.0081596
## 5-2   -890.27901  -3740.3363  1959.7783 0.9691287
## 9-2   -793.75687 -11372.7220  9785.2083 0.9999902
## 4-3   7256.53265   4048.9928 10464.0725 0.0000000
## 5-3   3187.86626   -172.2585  6547.9910 0.0760456
## 9-3   3284.38840  -7443.2448 14012.0216 0.9720702
## 5-4  -4068.66639  -7479.4408  -657.8920 0.0080076
## 9-4  -3972.14425 -14715.7497  6771.4612 0.9308300
## 9-5     96.52213 -10693.6208 10886.6651 1.0000000
```

These results show that the biggest part of pairs in company turnover groups show high p-values (3-1, 4-1, 5-1, 6-1, 7-1, 8-1, 4-3, 5-3, 6-3, 7-3, 8-3, 6-5, 7-5, 8-5, 9-5, 7-6, 8-6, 9-6, 8-7, 9-7, 9-8). Also quite many pairs in **peas** express higher p-values (4-0, 9-0, 9-1, 5-2, 9-2, 9-3, 9-4, 9-5).

Residual analysis: diagnostic plots to check the assumptions about normally distributed data and variance.

```
op <- par(mfrow = c(2, 2))
plot(anova_ind)
```




```r
par(op)
```

On the first plot the points 1822, 1443 and 79 are detected as outliers, which can affect normality and homogeneity of variance. It can be useful to remove them in order to match the test assumptions.

The second plot shows that the normality assumption is violated on the right upper part of the plot. As we have seen in the previous univariate analyses and plots, the total income data have long tail from right.

Because of the unbalanced design the leverages are not constant. On the fourth plot they are drawn in x-axis.

Conclusion: The provided analyses using two-side ANOVA with unbalanced design show that the dependences between categorical variables company turnover group sllvy and employees family status peas, and earned total employees income svatva as predictor are probably more complicated. For the investigated subset for year=2 even when quite small part of missing values are removed, the assumptions for normally distributed data and variance are not matched. The unbalanced design increases the complexity of this investigation.

Bivariate and multivariate analysis in the case of grouped levels of the categorical variables.

Next we observe that some of the levels of the categorical variables could be combined since they do not differ too much. For example the family status of people, who are officially married (family statuses 1 and 2) and others, who live together and probably have children, but are not officially married (family statuses 4 and 5), should not differ too much in the sense of their expenses and way of living. Since the children, who work, should not be too small, for me they look somehow similar (in the sense that financially they do not take care for the other members of family) to the singles. Also the group of unknowns is quite small and could be joined to the group of singles. Thus the new larger groups are: 0 - single/child/unknown; 1 - head, 2 - spouse. I also decided to group every 3 consequent (by their turnover) company groups in one bigger, so company group 1 has turnover LV <100 000, for group 2 the turnover LV is 100 000 <= LV < 10 000 000 , for group 3 the turnover 10 000 000 <= LV.

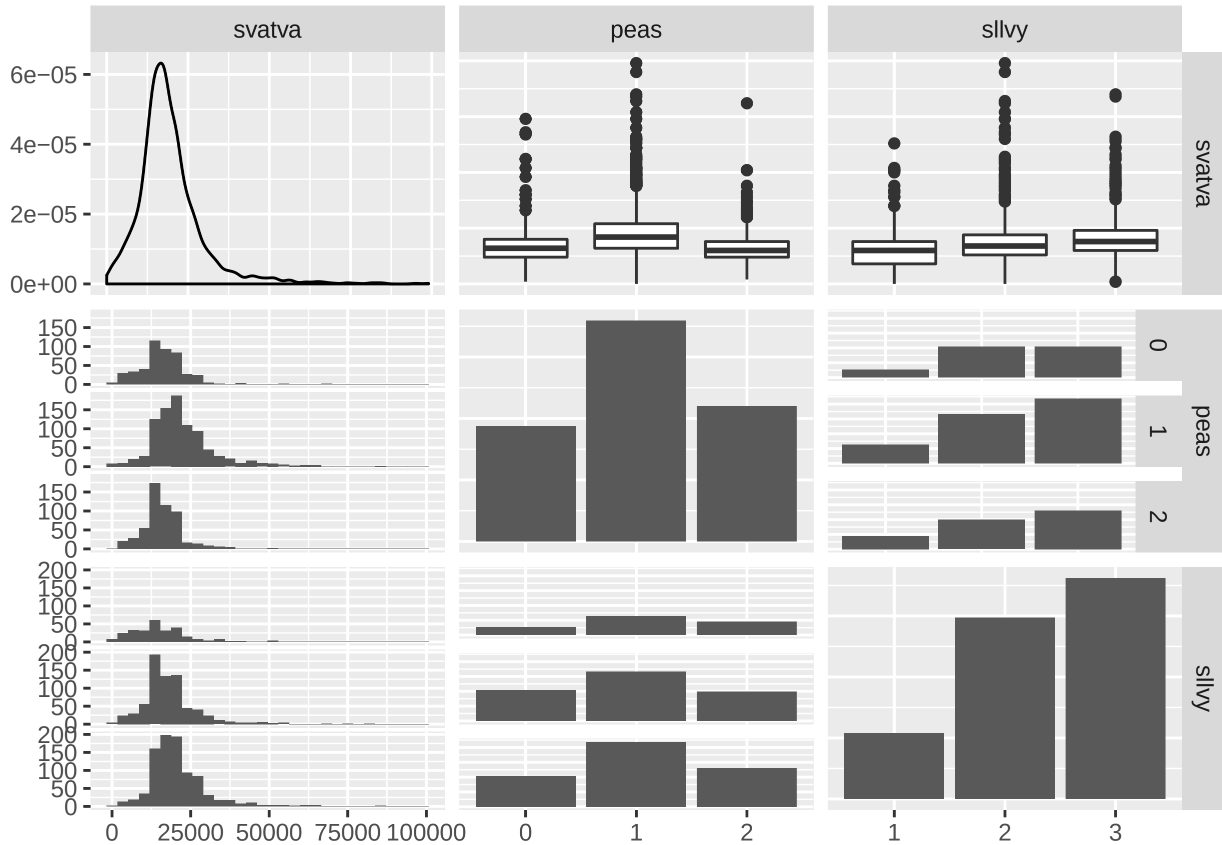
```
#5.1. Group the levels of peas and sllvy
#Group the levels in family status)
#0 - single/child/unknown
data$peas[data$peas=='3']<-'0'
data$peas[data$peas=='9']<-'0'
#1 - head
data$peas[data$peas=='4']<-'1'
#2 - spouse
data$peas[data$peas=='5']<-'2'
#drop levels
data$peas<-droplevels(data$peas)
#Group the levels in company turnover
#1 - new group formed from 1,2,3.
data$sllvy[data$sllvy=='2']<-'1'
data$sllvy[data$sllvy=='3']<-'1'
#4 - new group formed from 4,5,6.
data$sllvy[data$sllvy=='5']<-'4'
data$sllvy[data$sllvy=='6']<-'4'
#7 - new group formed from 7,8,9.
data$sllvy[data$sllvy=='8']<-'7'
data$sllvy[data$sllvy=='9']<-'7'
data$sllvy<-droplevels(data$sllvy)
levels(data$sllvy)<-c("1", "2", "3")
```

First look at the within correlation using ggpairs

To get a glimpse of these combined data we plot all investigated variables against each other.

```
var<-data[,c("svatva", "peas", "sllvy")]
assignInNamespace("ggally_cor", ggally_cor, "GGally")
ggpairs(var, upper = list(continuous = wrap("cor", size = 10)), lower = list(continuous = "smooth"))
```

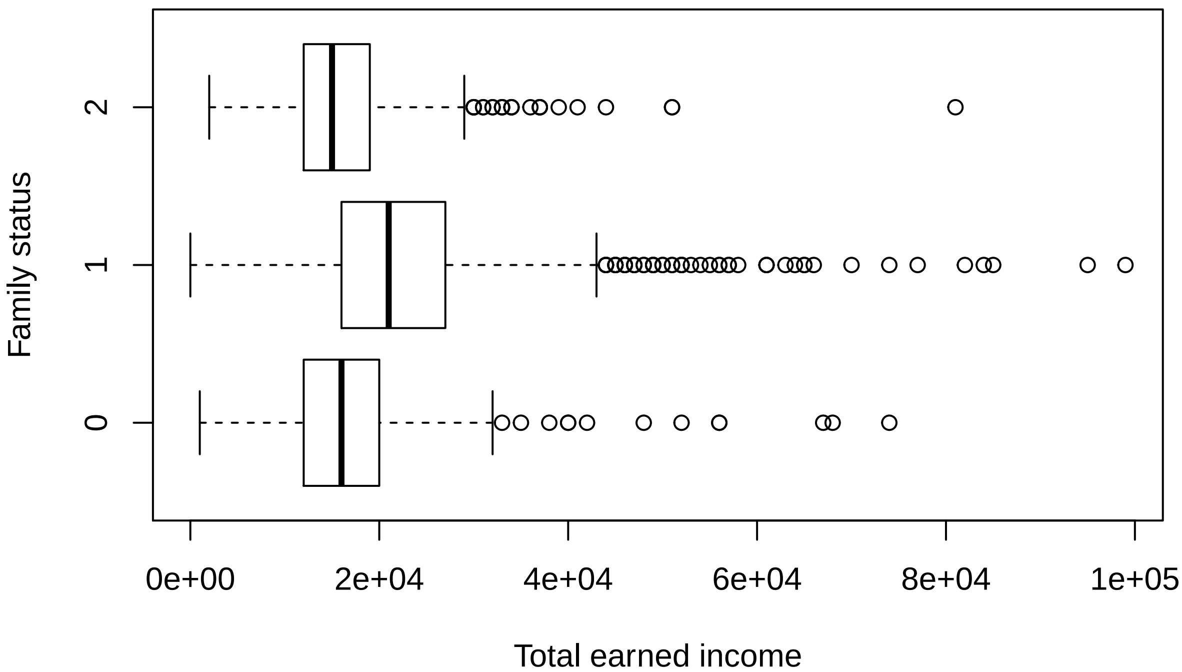
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We see that the distribution of total earned income for the highest group of heads of families deviates from normal distribution in similar way as the wholedistribution of the earned income. The distributions of earned income for highest two turnover company groups also have long right tails.

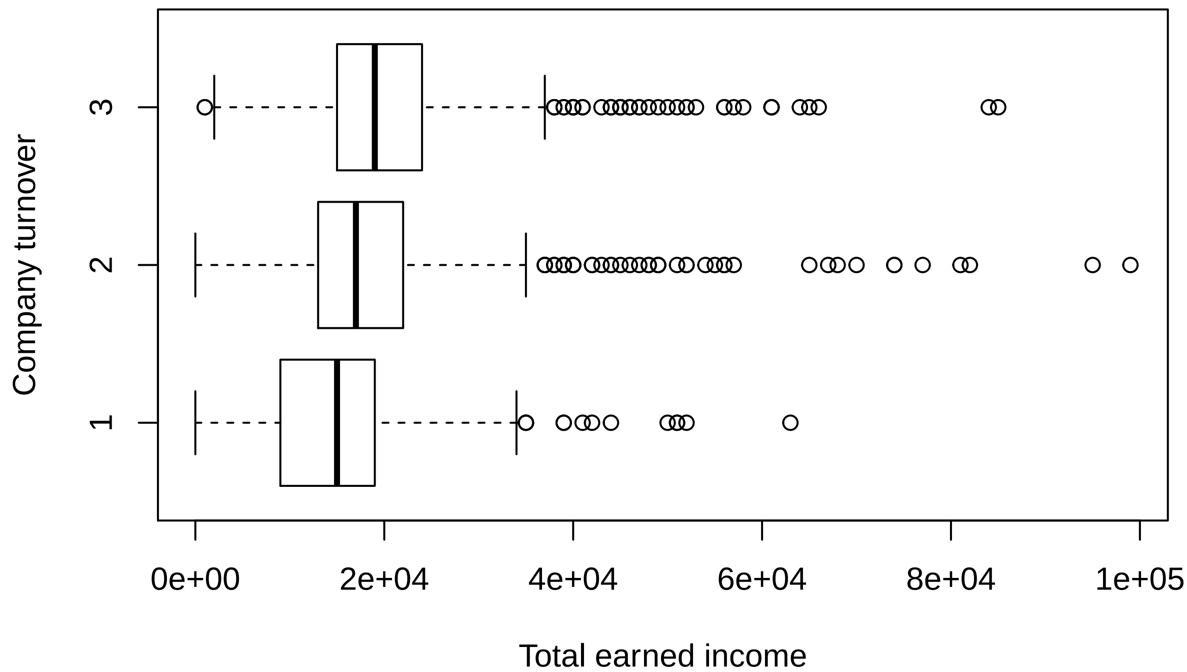
Next we repeat the bivariate and multivariate analyses following the same pipeline as before. ##Bivariate plots The corresponding bivariate box plots in this case are:

```
boxplot(data$svatva~data$peas, notch=F, horizontal=T, xlab="Total earned income", ylab="Family status")
```



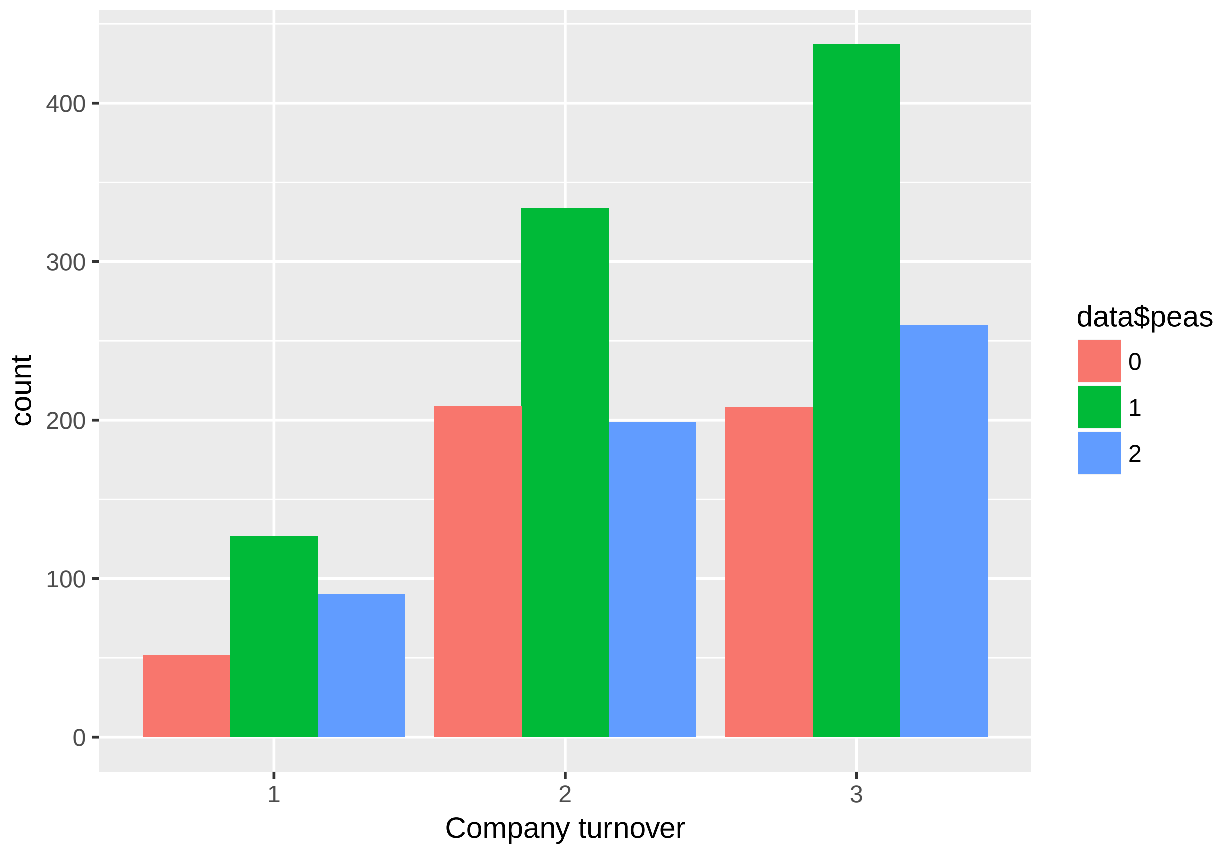
We observe that the mean values f both groups of singles and spouses are very similar, while heads of family have higher mean value of their total income in wider range.

```
boxplot(data$svatva~data$sllvy, notch=F, horizontal=T, xlab="Total earned income", ylab="Company turnov
```



Here we see clear tendency for increasing the mean total income in different company turnover groups. In this case there is clear difference between 3 different groups.

```
g0<-ggplot(data, aes(x = data$sllvy, fill = data$peas)) + geom_bar(position = "dodge")
g0 + xlab("Company turnover")
```



The observed dependences on the box plots are even better visualized on this bar plot. Only the family group 0 of singles/children/unknown does not show clear increasing tendency with increasing the company turnover.

Multivariate analysis

We follow the same steps as before. First generate frequency tables and make the same observation as before - the number of observations, which fall in different groups, is different, so again we observe that the design is not balanced.

We show some descriptive statistics anout the means

```
tapply(data$svatva,list(data$sllvy,data$peas),mean)
```

```
##          0        1        2
## 1 13076.92 17244.09 14733.33
## 2 15918.66 23239.52 15793.97
## 3 17855.77 24366.13 17096.15
```

and standard deviations:

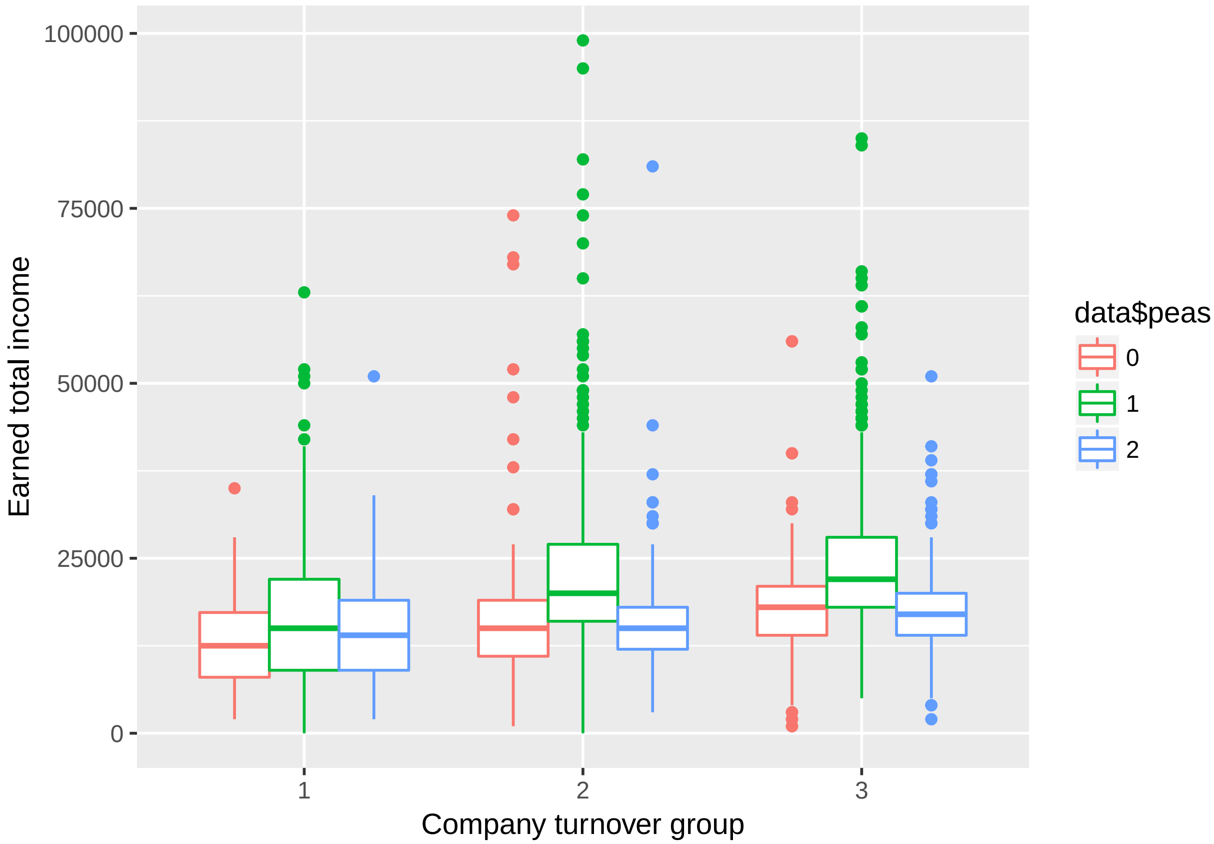
```
tapply(data$svatva,list(data$sllvy,data$peas),sd)
```

```
##          0        1        2
## 1 7023.340 11439.08 8323.245
## 2 9916.385 12624.26 7482.151
## 3 7428.784 10399.96 6138.868
```

Visualization of the total earned income and the two categorical variables company turnover and family status.

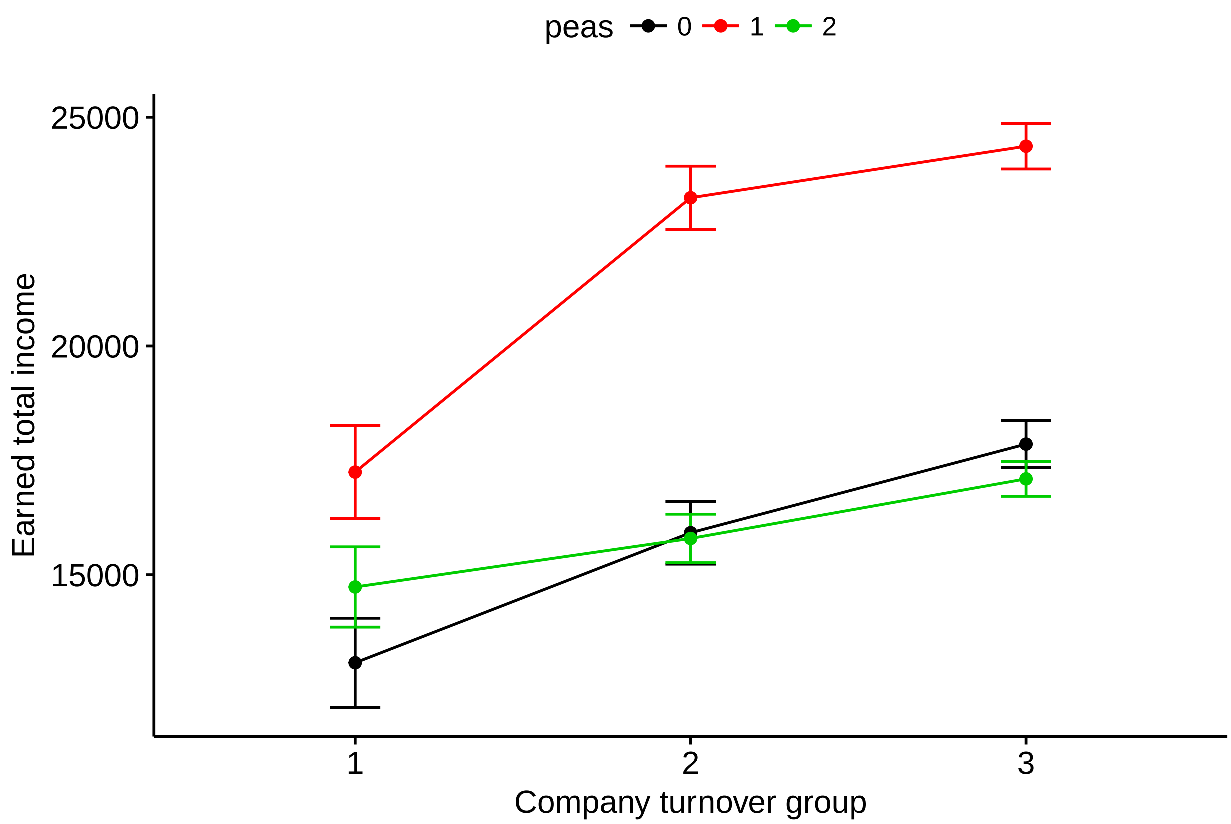
Box plot with multiple groups.

```
g1 <- ggplot(data, aes(x = data$sllvy, y = data$svatva, col=data$peas))
g1 + geom_boxplot() + xlab("Company turnover group")+ ylab("Earned total income")
```



Line plots with multiple groups

```
library(dplyr)
library(ggpubr)

ggline(data, x = "sllvy", y = "svatva",
       xlab = "Company turnover group", ylab="Earned total income",
       color = "peas",
       add = "mean_se", palette = 1:7)
```



Both graphs suggest that there is clear dependency between the earned total income and both categorical variables - company turnover and family status. ###Two way ANOVA test As before for the case of unbalanced design we apply type II sum of squares method to run ANOVA.
First we assume that both categorical variables are **independent**.

```
anova_ind <- aov(data$svatva ~ data$sllvy + data$peas, data = data)
Anova(anova_ind, type = "II")
```

```
## Anova Table (Type II tests)
##
## Response: data$svatva
##                Sum Sq   Df F value    Pr(>F)
## data$sllvy 5.4278e+09    2  28.937 4.162e-13 ***
## data$peas  2.0570e+10    2 109.666 < 2.2e-16 ***
## Residuals  1.7923e+11 1911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again both independent variables appear to be very significant because the very low p-values. We also check the model which involves the **simultaneous effect** of both predicates. Here we apply type III sum of squares method.

```
anova_inter <- aov(data$svatva ~ data$sllvy + data$peas + data$sllvy:data$peas, data = data)
Anova(anova_inter, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: data$svatva
##                          Sum Sq   Df F value    Pr(>F)
```

```
## (Intercept)          8.8923e+09    1 95.2037 < 2.2e-16 ***
## data$sllvy           1.0616e+09    2  5.6830  0.003461 **
## data$peas            7.4191e+08    2  3.9716  0.019000 *
## data$sllvy:data$peas 1.1073e+09    4  2.9638  0.018704 *
## Residuals            1.7812e+11 1907
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we observe that all predicates are significant. The intercept is the estimate of the predicate when all the independent variables are 0. By a rule the significance of the intercept is not of interest because its value can be changed by recoding the predictor, and this will not affect the meaning of the model. Since this model shows significant simultaneous effect of both predicates, we cannot assume that these are independent and continue its investigation.

We compute some **summary statistics**:

```
model.tables(anova_inter, type="means")
```

```
## Tables of means
## Grand mean
##
## 19433.72
##
##  data$sllvy
##         1     2     3
##     15599 19181 20781
## rep   269   742   905
##
##  data$peas
##         0     1     2
##     16403 22921 16319
## rep   469   898   549
##
##  data$sllvy:data$peas
##           data$peas
## data$sllvy 0     1     2
##        1   13077 17244 14733
##        rep    52   127    90
##        2   15919 23240 15794
##        rep   209   334   199
##        3   17856 24366 17096
##        rep   208   437   260
```

Next we obtain the **multiple pairwise-comparison** between the means of the groups by computing Tukey Honest Significant Differences (THS).

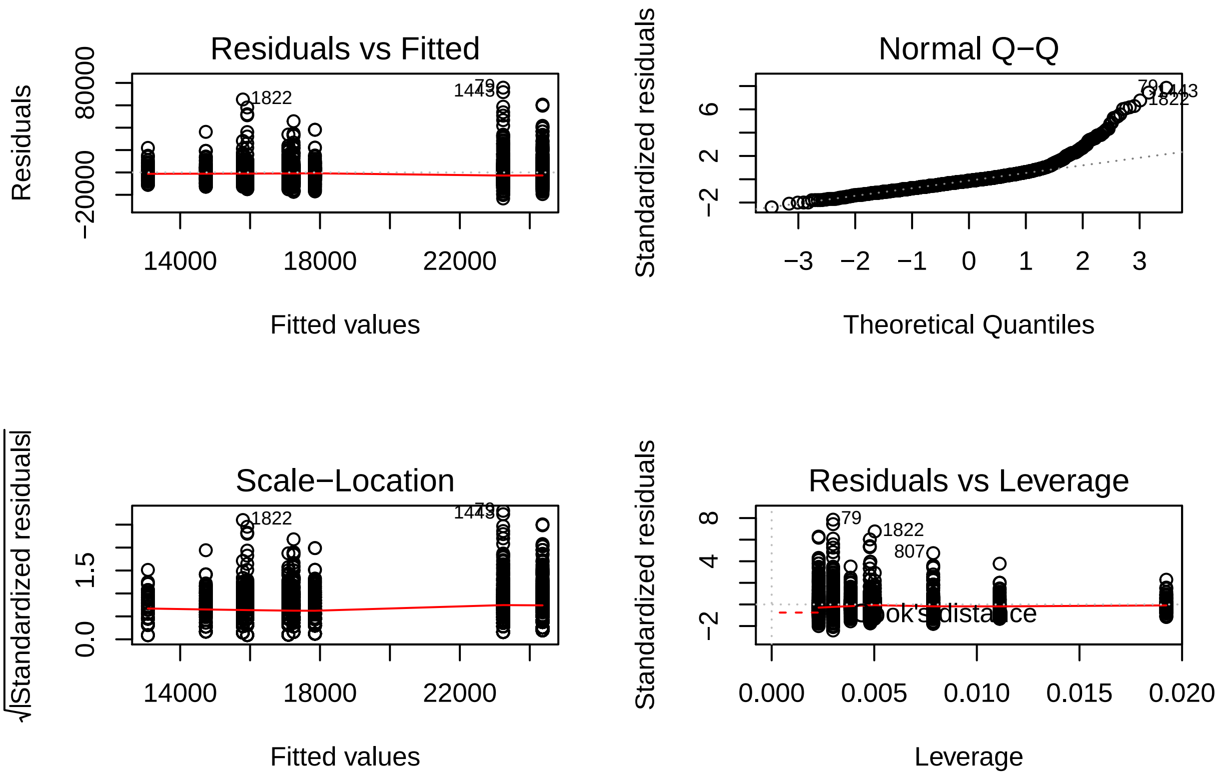
```
TukeyHSD(anova_inter)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data$svatva ~ data$sllvy + data$peas + data$sllvy:data$peas, data = data)
##
## $`data$sllvy`
##         diff       lwr      upr     p adj
## 2-1 3582.080 1968.7626 5195.397 0.0000006
## 3-1 5182.702 3608.5169 6756.888 0.0000000
```

```
## 3-2 1600.622  477.9764 2723.269 0.0024234
##
## $`data$peas`
##            diff       lwr       upr     p adj
## 1-0  6517.56548  5226.102  7809.029 0.0000000
## 2-0   -83.99162 -1509.348  1341.365 0.9895235
## 2-1 -6601.55710 -7829.653 -5373.461 0.0000000
##
## $`data$sllvy:data$peas`
##               diff         lwr        upr     p adj
## 2:0-1:0  2841.7372  -1809.1074  7492.5818 0.6156561
## 3:0-1:0  4778.8462    125.7747  9431.9176 0.0388103
## 1:1-1:0  4167.1714   -773.7662  9108.1090 0.1792236
## 2:1-1:0 10162.5979   5688.5036 14636.6922 0.0000000
## 3:1-1:0 11289.2096   6886.7191 15691.7001 0.0000000
## 1:2-1:0  1656.4103  -3571.2534  6884.0739 0.9873559
## 2:2-1:0  2717.0468  -1957.0214  7391.1149 0.6788812
## 3:2-1:0  4019.2308   -539.8296  8578.2911 0.1354280
## 3:0-2:0  1937.1089  -1002.2296  4876.4475 0.5108029
## 1:1-2:0  1325.4342  -2051.1767  4702.0451 0.9524928
## 2:1-2:0  7320.8607   4673.9461  9967.7753 0.0000000
## 3:1-2:0  8447.4724   5923.4757 10971.4692 0.0000000
## 1:2-2:0 -1185.3270  -4969.1218  2598.4679 0.9882630
## 2:2-2:0  -124.6904  -3097.1557  2847.7748 1.0000000
## 3:2-2:0  1177.4936  -1610.6363  3965.6234 0.9281598
## 1:1-3:0  -611.6747  -3991.3523  2768.0028 0.9997611
## 2:1-3:0  5383.7517   2732.9263  8034.5772 0.0000000
## 3:1-3:0  6510.3635   3982.2658  9038.4612 0.0000000
## 1:2-3:0 -3122.4359  -6908.9676   664.0958 0.2041206
## 2:2-3:0 -2061.7994  -5037.7477   914.1490 0.4383684
## 3:2-3:0  -759.6154  -3551.4583  2032.2275 0.9954501
## 2:1-1:1  5995.4265   2866.7421  9124.1109 0.0000001
## 3:1-1:1  7122.0382   4096.6343 10147.4421 0.0000000
## 1:2-1:1 -2510.7612  -6645.9290  1624.4067 0.6239620
## 2:2-1:1 -1450.1246  -4858.6520  1958.4028 0.9252049
## 3:2-1:1  -147.9406  -3396.9678  3101.0865 1.0000000
## 3:1-2:1  1126.6118  -1054.6090  3307.8325 0.8029284
## 1:2-2:1 -8506.1876 -12070.4924 -4941.8829 0.0000000
## 2:2-2:1 -7445.5511 -10133.0619 -4758.0403 0.0000000
## 3:2-2:1 -6143.3671  -8625.4696 -3661.2647 0.0000000
## 1:2-3:1 -9632.7994 -13106.7987 -6158.8001 0.0000000
## 2:2-3:1 -8572.1629 -11138.7008 -6005.6249 0.0000000
## 3:2-3:1 -7269.9789  -9620.5608 -4919.3970 0.0000000
## 2:2-1:2  1060.6365  -2751.6674  4872.9404 0.9946984
## 3:2-1:2  2362.8205  -1307.5718  6033.2128 0.5441775
## 3:2-2:2  1302.1840  -1524.5147  4128.8827 0.8861276
```

Again we observe some high p-values, indicating that there is not significant efect between corresponding pairs of variables. I am not very sure how strong the conclusions from this test for the case of unbalanced design are. **Residual analysis**: diagnostic plots to check the assumptions about normally distributed data and variance.

```
op1 <- par(mfrow = c(2, 2))
plot(anova_inter)
```




```
par(op1)
```

On the first plot we observe that the assumption about homogeneity of variances is probably valid. Again the points 1822, 1443 and 79 are detected as outliers and can affect the assumptions about normality and homogeneity of variance. In this model we are able to compute Levene's test to check the homogeneity of variances (leveneTest() in car package), but as it is stated in [10], it is not recommended for the case of unbalanced design because the significance level could be under- or overestimated. On the second plot we again observe that the normality assumption is violated on the right upper part of the plot. On the fourth plot the leverages are drawn in x-axis.

Conclusion: Grouping several similar levels in both categorical variables concerning family status and company turnover into bigger homogenous groups significantly improves the results in this particular task and data subset. Bivariate visualization revealed some clear tendencies like for example the increasing frequencies in almost all family groups with increasing turnover. Now we are able to observe the clear tendency for increasing the mean earned income in almost all family statuses when company turnover increases. The two way ANOVA analysis revealed the important fact of simultaneous effect of family status and company turnover together.

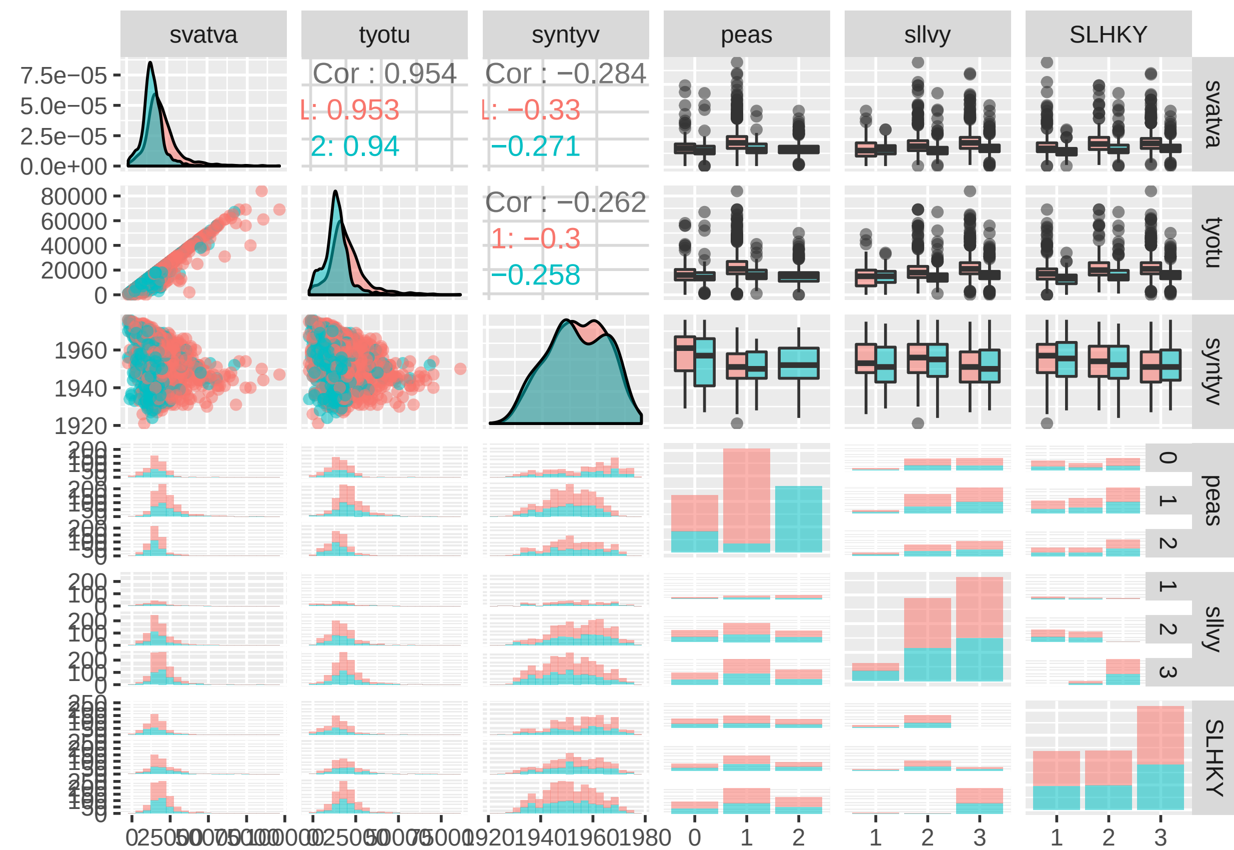
Some interesting relations.Possible future work.

After wrangling around all columns, I choose some of them which look interesting and plot them against each other. I added to the three already investigated variables the other numerical variable **tyotu** (income from salary), and **syntyv** (year of birth), and also one more categorical variable - **SLHKY** about the size of the company measured by number of employees. Since we have already observed that it is difficult to make some inference when the number of categories is too big, I have applied similar grouping of all 3 consequent categories to 1 bigger in **SLHKY** as well. There are two different colors on the plot, separating males (pink) and females (blue).

```
#6.2 plot chosen variables using the combined groups in peas, sllvy and SLHKY.
var1<-data_p[,c("svatva", "tyotu", "syntyv", "peas", "sllvy", "SLHKY")]
str(var1)
```

```
## 'data.frame':    1783 obs. of  6 variables:
##  $ svatva: int  24000 14000 25000 24000 21000 26000 42000 13000 23000 14000 ...
##  $ tyotu : int  24000 14000 25000 23000 21000 26000 41000 12000 21000 14000 ...
##  $ syntyv: int  1942 1951 1956 1947 1944 1952 1942 1950 1961 1954 ...
##  $ peas  : Factor w/ 3 levels "0","1","2": 2 2 2 2 2 2 1 3 2 2 ...
##  $ sllvy : Factor w/ 3 levels "1","2","3": 2 2 3 3 3 2 2 2 2 2 ...
##  $ SLHKY : Factor w/ 3 levels "1","2","3": 2 1 3 3 3 2 1 1 1 1 ...
```

```
p <- ggpairs(var1, mapping = aes(col=data_p$sukup, alpha=0.3), lower = list(combo = wrap("facethist", b
p
```



We observe that there is not so big difference between age distribution of males and females. This variable will probably become more informative if we transform it to age (current year - year of birth), and separate different age groups in different categories. We see that the companies with biggest turnover (group 3) are also the companies with bigger number of employees. The pairwise plot also show strong correlation (0.954) between the total earned income **svatva** and income from salary **tyotu**. All these findings should be investigated further.

Discussion

In this work we found that grouping some levels into bigger groups in two categorical variables significantly improved the modeling results and helped to clarify some important tendencies. This is valid only for this particular data subset and the modeling task defined as it is. The situation could be opposite in other situations, when dividing the existing groups will lead to discovering new dependences between the variables. It seems that we cannot say in advance what approach to apply; just have to explore different possibilities.

Here we have analyzed only a data subset for year=2, which became very small after omitting only a part of the missing values. Therefore we cannot be sure that the results are representative for the whole populations. Furthermore, we could just by chance to be able to observe dependences, which do not exist in other subsets. Therefore it is logical to analyze other samples for different years, providing some statistical analyses for comparing these samples. I have also checked some statistics for year *vuosi*=10. The data points there are a bit more, but the difference is not very significant.

```
summary(data$svatva)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   14000   18000   19434   23000   99000
```

```
summary(data_10$svatva)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   17000   25000   26490   33000  100000
```

Here we observe that all values (not only mean and median) for latter year=10 are higher, which is logical because there are 8 years difference and we expect increasing total earned income in time (at least because the inflation).

```
summary(sllvy)
```

```
##   1   2   3   4   5   6   7   8   9
## 110  96  63 161 285 296 239 264 402
```

```
summary(data_10$sllvy)
```

```
##   1   2   3   4   5   6   7   8   9
## 238  56  72 177 320 310 229 296 392
```

Here we observe slight decreases in the frequencies of company turnover in higher groups and increasing frequencies in group one, which also cannot be explained only by chance, but it is rather because the economical changes in Finland during this time period.

```
summary(peas)
```

```
##   0   1   2   3   4   5   9
## 307 753 427 155 145 122   7
```

```
summary(data_10$peas)
```

```
##   0   1   2   3   4   5   9
## 432 681 388 165 240 174  10
```

Here we observe decreasing the amount of officially married couples and increase of the proportion of couples who live together without official marriage. Also the proportion of singles is increased. This also is rather regular tendency in time than random event. It seems that if we would like to compare the modeling results, we have to choose some subsets quite near in time.

One of the simplified assumptions here is to **remove the missing data**. For this particular task we did not removed any values from both categorical variables concerning family status and company turnover because these data were complete. But if we were interested in some other categorical variables (like for example *toimiala* or *suuralue*), we have to probably remove some valuable information, which is involved in these missing values. For the categorical variable one of the possibilities is to form an additional category from the missing values, and then apply the multiple correspondance analysis to investigate the associations between different categories [2, 3].

Used and useful links

1. Sheldon Ross. Introductory Statistics, 4-th Edition, Elsevier, 2017, p.828.
2. Assignments Work during the Course on Multiple Correspondence Analysis (MCA): Theory and Practice, Spring 2017, University of Helsinki
3. Multiple Correspondence Analysis Essentials: Interpretation and application to investigate the associations between categories of multiple qualitative variables - R software and data mining
4. D G Rossite. Tutorial: An example of statistical data analysis using the R environment for statistical computing, Version 1.4; May 6, 2017.
5. Dylan Z. Childs. Exploring categorical variables, 2018.
6. Two-Way ANOVA Test in R.
7. Anova - Type I/II/III SS explained.
8. Raccoon | Ch 2.5 - Unbalanced and Nested Anova
9. Two-Way Factorial ANOVA with R
10. The Assumption of Homogeneity of Variance