Mandatory Assignment 1 - STK-IN4300

Alessia Sanfelici

September 21, 2023

The selected dataset is about Data Science Salaries and was found on the website Kaggle at the following link: <https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023>.

The dataset contains 3755 observations and 11 variables. Each row represents one employee, with all the informations described by the columns. There are no missing values, meaning that the whole dataset is full.

The variables are the following:

* : the year the salary was paid;
* : experience level of the employee during the year;
* : type of employee;
* : role of the employee during the year;
* : total salary expressed in the currency of the country of work;
* : currency of the country of work;
* : salary converted in USD;
* : primary country of residence of the employee during the working year;
* : overall amount of work done remotely;
* : country of the company where the employee works;
* : size (in terms of number of working people) of the company during the year;

The categorical variables are: , , , , , and .

The integer variables are: and .

The remaining variables, and , are integer variables but, since they are not continuous but discrete (they represent years and percentages respectively), it is better to consider them as categorical variables.

This dataset could be used for many purposes of analysis. For example, it could be used to make a prediction of the salary of an employee, based on his/hers characteristics about the country of residence, the currency, the company, and so on. Another application of this dataset could be a clustering of the data, applied with the aim of grouping similar observations together.

## Problem 1. Summary Statistics Table

For an easier understanding of the data, continuous and categorical variables have been summarized in two separated tables. That is due to the fact that these two types of variables can be described with different quantities and information. The summaries of the variables have been created by using the library **gtsummary**.

The following table reports information about the continuous variables: it includes the mean, the standard deviation, the median and the minimum and maximum values.

library("gtsummary")  
  
# Summary for continuous variables  
df%>%   
 tbl\_summary(type = c(salary ~ "continuous2", salary\_in\_usd ~ "continuous2"),   
 statistic = list(all\_continuous() ~ c("{mean} ({sd})",   
 "{median}", "{min}-{max}")),  
 digits = all\_continuous() ~ c(2, 2, 0, 0, 0),   
 include = c(salary, salary\_in\_usd),  
 label = list(salary ~ "Salary", salary\_in\_usd ~ "Salary in USD")) %>%  
 bold\_labels() %>%  
 modify\_caption("\*\*Continuous Variables\*\*")

**Continuous Variables**

| **Characteristic** | **N = 3,755** |
| --- | --- |
| **Salary** |  |
| Mean (SD) | 190,695.57 (671,676.50) |
| Median | 138,000 |
| Minimum-Maximum | 6,000-30,400,000 |
| **Salary in USD** |  |
| Mean (SD) | 137,570.39 (63,055.63) |
| Median | 135,000 |
| Minimum-Maximum | 5,132-450,000 |

For the remaining variables, a new table has been created. Since the variables are categorical, other information needs to be outlighted: the absolute and the relative frequency of each variable in the dataset.

For the creation of a clearer summary table for categorical variables, some columns have been removed, since they were characterized by many different values with very low frequency (otherwise the table would be infinite). These columns are: , and .

# Summary for categorical variables   
df %>%  
 tbl\_summary(type = everything() ~ "categorical",   
 digits = all\_categorical() ~ c(0, 2),   
 label = list(work\_year ~ "Work Year",   
 experience\_level ~ "Experience Level",  
 employment\_type ~ "Employment type",  
 salary\_currency ~ "Salary Currency",  
 remote\_ratio ~ "Remote Ratio",  
 company\_size ~ "company Size"),  
 include = c(work\_year, experience\_level, employment\_type,  
 salary\_currency, remote\_ratio, company\_size),  
 statistic = list(all\_categorical() ~ "{n} ({p}%)")) %>%  
 bold\_labels() %>%  
 modify\_caption("\*\*Categorical Variables\*\*")

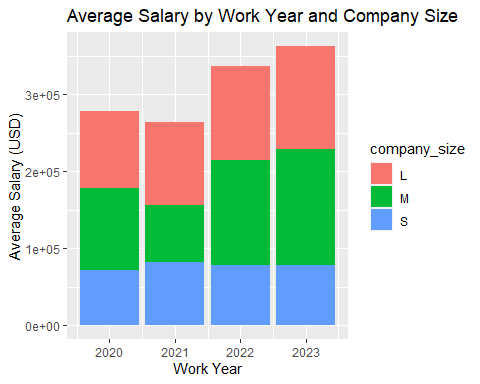
**Categorical Variables**

| **Characteristic** | **N = 3,755** |
| --- | --- |
| **Work Year** |  |
| 2020 | 76 (2.02%) |
| 2021 | 230 (6.13%) |
| 2022 | 1,664 (44.31%) |
| 2023 | 1,785 (47.54%) |
| **Experience Level** |  |
| EN | 320 (8.52%) |
| EX | 114 (3.04%) |
| MI | 805 (21.44%) |
| SE | 2,516 (67.00%) |
| **Employment type** |  |
| CT | 10 (0.27%) |
| FL | 10 (0.27%) |
| FT | 3,718 (99.01%) |
| PT | 17 (0.45%) |
| **Salary Currency** |  |
| AUD | 9 (0.24%) |
| BRL | 6 (0.16%) |
| CAD | 25 (0.67%) |
| CHF | 4 (0.11%) |
| CLP | 1 (0.03%) |
| CZK | 1 (0.03%) |
| DKK | 3 (0.08%) |
| EUR | 236 (6.28%) |
| GBP | 161 (4.29%) |
| HKD | 1 (0.03%) |
| HUF | 3 (0.08%) |
| ILS | 1 (0.03%) |
| INR | 60 (1.60%) |
| JPY | 3 (0.08%) |
| MXN | 1 (0.03%) |
| PLN | 5 (0.13%) |
| SGD | 6 (0.16%) |
| THB | 2 (0.05%) |
| TRY | 3 (0.08%) |
| USD | 3,224 (85.86%) |
| **Remote Ratio** |  |
| 0 | 1,923 (51.21%) |
| 50 | 189 (5.03%) |
| 100 | 1,643 (43.75%) |
| **company Size** |  |
| L | 454 (12.09%) |
| M | 3,153 (83.97%) |
| S | 148 (3.94%) |

## Problem 2. Bad Data Visualization

The first graph represents the average salary (in USD) of the employees with respect to the work year and also to the size of the company. Using a bar plot with stacked bars allows us to have an idea of the amount of average salary for each year, but it is difficult to make a comparison between the different values of the salary with respect to year and company size. That is, for each year, we can only see the total salary, obtained by the sum of the average salaries for small, medium and large companies, without any indication about the single values, which cannot be compared.

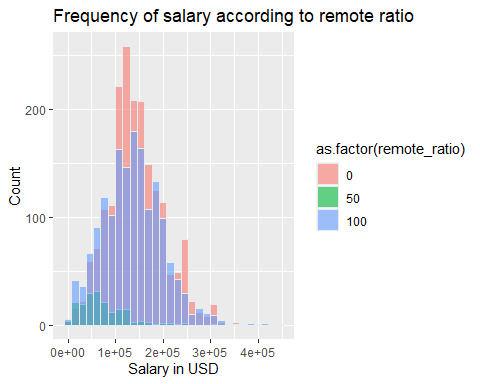
# Bad plot 1  
by\_year <- group\_by(df, work\_year, company\_size)  
new\_df <- summarize(by\_year, average\_salary = mean(salary\_in\_usd, na.rm = TRUE))  
  
ggplot(new\_df) +  
 geom\_bar(mapping = aes(x = work\_year, y = average\_salary, fill = company\_size), stat = "identity")+  
 labs(x = "Work Year", y = "Average Salary (USD)", title = "Average Salary by Work Year and Company Size")



The second “bad” plot shows the density of the salary according to the remote ratio. The graph is a barplot, containing bars of different colors, with respect to the amount of work done remotely (0, 50 or 100). We can identify three main problems in this plot:

* The bars overlap, making the trends confusing and difficult to understand;
* Since the 50 category has a very low number of cases, the corresponding bars are very small, making it very hard to catch the frequency and the behaviour of this category;
* The values on the X-axis are not so easy to interpret.

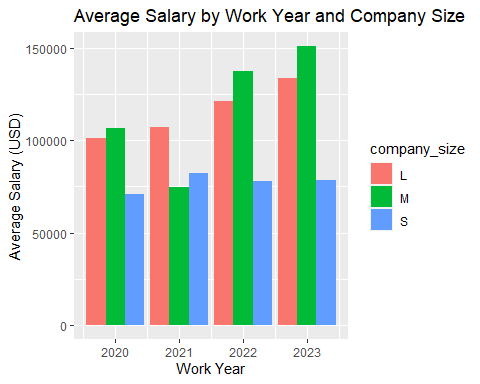
# Bad plot 2  
ggplot(df, aes(x = salary\_in\_usd, fill = as.factor(remote\_ratio)))+  
 geom\_histogram( color='#e9ecef', alpha=0.6, position='identity')+  
 labs(x = "Salary in USD", y = "Count", title = "Frequency of salary according to remote ratio")



## Problem 3. Good Data Visualization

The “good” version of the first graph uses a bar for each category of the company size. Thanks to this trick, it is possible to visualize data in a more organized way: for each year, we have 3 columns, each one representing the average salary for each company size. This allows a deeper analysis, with a particular interest on the behaviour of the average salary over time, according to the size of the company. For instance, it is possible to notice that, while small companies had more or less stable salaries over time, the average salary in large companies increased with the years, with a growth of more than 250000 USD from 2020 to 2023.

new\_df <- summarize(by\_year, average\_salary = mean(salary\_in\_usd, na.rm = TRUE))  
  
# Good plot 1  
ggplot(new\_df) +  
 geom\_bar(mapping = aes(x = work\_year, y = average\_salary, fill = company\_size),   
 stat = "summary", position = "dodge") +   
 labs(x = "Work Year", y = "Average Salary (USD)", title = "Average Salary by Work Year and Company Size")



The second plot have been improved by substituting the histogram with a density plot. This solution allows to obtain a clearer graph, composed of just three lines. Each line represents the density distribution of the salary, according to the remote ratio category. Moreover, since the data for all the categories have been transformed to densities, we don’t have small bars for a single category anymore. The problem of the values on the X-axis has been solved by changing the unit of measure of the axis (from USD to thousand of USD).

According to this graph, for example, it can be understood that employees with 50% of remote ratio tend to have a lower salary than the other categories (the pick of the density is under 100000 USD, probably around 80000 USD). On the contrary, the other two categories have similar densities, with a mean around 150000 USD.

#Good plot 2  
  
rem\_0 <- subset(df, df$remote\_ratio == 0, select = c(salary\_in\_usd))  
rem\_50 <- subset(df, df$remote\_ratio == 50, select = c(salary\_in\_usd))  
rem\_100 <- subset(df, df$remote\_ratio == 100, select = c(salary\_in\_usd))  
d0 <- density(as.vector(rem\_0[["salary\_in\_usd"]]))  
d50 <- density(as.vector(rem\_50[["salary\_in\_usd"]]))  
d100 <- density(as.vector(rem\_100[["salary\_in\_usd"]]))  
plot(d50, col = "red", ,main = "Density of salary according to remote ratio",  
 xlab = "Salary in thousands of USD", ylab = "Density", xaxt = "n")   
# Customize the x-axis labels to be in thousands  
axis(1, at = seq(1e5, 5e5, by = 1e5),  
 labels = scales::comma\_format(scale = 1e-3)(seq(1e5, 5e5, by = 1e5)))  
lines(d0, col = "blue")  
lines(d100, col = "green")  
legend("topright", legend = c("Remote ratio = 0", "Remote ratio = 50", "Remote ratio = 100"),  
 col = c("blue", "red", "green"), lty = 1)

