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: .

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 worked 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 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. Possible use of the data?????

## 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 . \ 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.

# 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",  
 job\_title ~ "Job Title",   
 salary\_currency ~ "Salary Currency",  
 remote\_ratio ~ "Remote Ratio",  
 company\_location ~ "Company Location",  
 company\_size ~ "company Size"),  
 include = c(work\_year, experience\_level, employment\_type, job\_title,  
 salary\_currency, remote\_ratio,  
 company\_location, company\_size),  
 statistic = list(all\_categorical() ~ "{n} ({p}%)")) %>%  
 bold\_labels() %>%  
 modify\_caption("\*\*Categorical Variables\*\*")

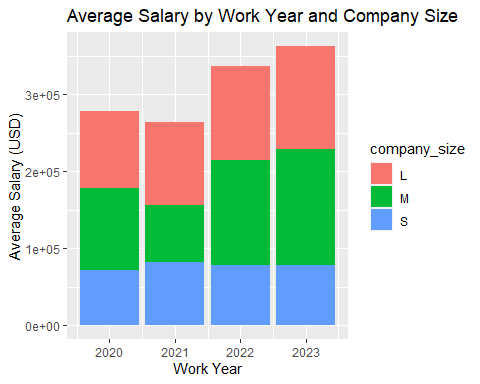
**Categorical Variables**

| **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%) |
| **Job Title** |  |
| 3D Computer Vision Researcher | 4 (0.11%) |
| AI Developer | 11 (0.29%) |
| AI Programmer | 2 (0.05%) |
| AI Scientist | 16 (0.43%) |
| Analytics Engineer | 103 (2.74%) |
| Applied Data Scientist | 10 (0.27%) |
| Applied Machine Learning Engineer | 2 (0.05%) |
| Applied Machine Learning Scientist | 12 (0.32%) |
| Applied Scientist | 58 (1.54%) |
| Autonomous Vehicle Technician | 2 (0.05%) |
| Azure Data Engineer | 1 (0.03%) |
| BI Analyst | 9 (0.24%) |
| BI Data Analyst | 15 (0.40%) |
| BI Data Engineer | 1 (0.03%) |
| BI Developer | 13 (0.35%) |
| Big Data Architect | 2 (0.05%) |
| Big Data Engineer | 11 (0.29%) |
| Business Data Analyst | 15 (0.40%) |
| Business Intelligence Engineer | 4 (0.11%) |
| Cloud Data Architect | 1 (0.03%) |
| Cloud Data Engineer | 3 (0.08%) |
| Cloud Database Engineer | 5 (0.13%) |
| Compliance Data Analyst | 1 (0.03%) |
| Computer Vision Engineer | 18 (0.48%) |
| Computer Vision Software Engineer | 5 (0.13%) |
| Data Analyst | 612 (16.30%) |
| Data Analytics Consultant | 2 (0.05%) |
| Data Analytics Engineer | 6 (0.16%) |
| Data Analytics Lead | 2 (0.05%) |
| Data Analytics Manager | 22 (0.59%) |
| Data Analytics Specialist | 2 (0.05%) |
| Data Architect | 101 (2.69%) |
| Data DevOps Engineer | 1 (0.03%) |
| Data Engineer | 1,040 (27.70%) |
| Data Infrastructure Engineer | 6 (0.16%) |
| Data Lead | 2 (0.05%) |
| Data Management Specialist | 1 (0.03%) |
| Data Manager | 29 (0.77%) |
| Data Modeler | 2 (0.05%) |
| Data Operations Analyst | 4 (0.11%) |
| Data Operations Engineer | 10 (0.27%) |
| Data Quality Analyst | 7 (0.19%) |
| Data Science Consultant | 24 (0.64%) |
| Data Science Engineer | 5 (0.13%) |
| Data Science Lead | 8 (0.21%) |
| Data Science Manager | 58 (1.54%) |
| Data Science Tech Lead | 1 (0.03%) |
| Data Scientist | 840 (22.37%) |
| Data Scientist Lead | 2 (0.05%) |
| Data Specialist | 14 (0.37%) |
| Data Strategist | 2 (0.05%) |
| Deep Learning Engineer | 6 (0.16%) |
| Deep Learning Researcher | 1 (0.03%) |
| Director of Data Science | 11 (0.29%) |
| ETL Developer | 10 (0.27%) |
| ETL Engineer | 2 (0.05%) |
| Finance Data Analyst | 1 (0.03%) |
| Financial Data Analyst | 3 (0.08%) |
| Head of Data | 10 (0.27%) |
| Head of Data Science | 9 (0.24%) |
| Head of Machine Learning | 1 (0.03%) |
| Insight Analyst | 2 (0.05%) |
| Lead Data Analyst | 5 (0.13%) |
| Lead Data Engineer | 6 (0.16%) |
| Lead Data Scientist | 9 (0.24%) |
| Lead Machine Learning Engineer | 3 (0.08%) |
| Machine Learning Developer | 7 (0.19%) |
| Machine Learning Engineer | 289 (7.70%) |
| Machine Learning Infrastructure Engineer | 11 (0.29%) |
| Machine Learning Manager | 3 (0.08%) |
| Machine Learning Research Engineer | 4 (0.11%) |
| Machine Learning Researcher | 6 (0.16%) |
| Machine Learning Scientist | 26 (0.69%) |
| Machine Learning Software Engineer | 10 (0.27%) |
| Manager Data Management | 1 (0.03%) |
| Marketing Data Analyst | 2 (0.05%) |
| Marketing Data Engineer | 1 (0.03%) |
| ML Engineer | 34 (0.91%) |
| MLOps Engineer | 4 (0.11%) |
| NLP Engineer | 7 (0.19%) |
| Power BI Developer | 1 (0.03%) |
| Principal Data Analyst | 2 (0.05%) |
| Principal Data Architect | 1 (0.03%) |
| Principal Data Engineer | 2 (0.05%) |
| Principal Data Scientist | 8 (0.21%) |
| Principal Machine Learning Engineer | 1 (0.03%) |
| Product Data Analyst | 5 (0.13%) |
| Product Data Scientist | 1 (0.03%) |
| Research Engineer | 37 (0.99%) |
| Research Scientist | 82 (2.18%) |
| Software Data Engineer | 2 (0.05%) |
| Staff Data Analyst | 1 (0.03%) |
| Staff Data Scientist | 1 (0.03%) |
| **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 Location** |  |
| AE | 3 (0.08%) |
| AL | 1 (0.03%) |
| AM | 1 (0.03%) |
| AR | 3 (0.08%) |
| AS | 3 (0.08%) |
| AT | 6 (0.16%) |
| AU | 14 (0.37%) |
| BA | 1 (0.03%) |
| BE | 4 (0.11%) |
| BO | 1 (0.03%) |
| BR | 15 (0.40%) |
| BS | 1 (0.03%) |
| CA | 87 (2.32%) |
| CF | 2 (0.05%) |
| CH | 5 (0.13%) |
| CL | 1 (0.03%) |
| CN | 1 (0.03%) |
| CO | 4 (0.11%) |
| CR | 1 (0.03%) |
| CZ | 3 (0.08%) |
| DE | 56 (1.49%) |
| DK | 4 (0.11%) |
| DZ | 1 (0.03%) |
| EE | 2 (0.05%) |
| EG | 1 (0.03%) |
| ES | 77 (2.05%) |
| FI | 3 (0.08%) |
| FR | 34 (0.91%) |
| GB | 172 (4.58%) |
| GH | 2 (0.05%) |
| GR | 14 (0.37%) |
| HK | 1 (0.03%) |
| HN | 1 (0.03%) |
| HR | 3 (0.08%) |
| HU | 2 (0.05%) |
| ID | 2 (0.05%) |
| IE | 7 (0.19%) |
| IL | 2 (0.05%) |
| IN | 58 (1.54%) |
| IQ | 1 (0.03%) |
| IR | 1 (0.03%) |
| IT | 4 (0.11%) |
| JP | 6 (0.16%) |
| KE | 2 (0.05%) |
| LT | 2 (0.05%) |
| LU | 3 (0.08%) |
| LV | 4 (0.11%) |
| MA | 1 (0.03%) |
| MD | 1 (0.03%) |
| MK | 1 (0.03%) |
| MT | 1 (0.03%) |
| MX | 10 (0.27%) |
| MY | 1 (0.03%) |
| NG | 5 (0.13%) |
| NL | 13 (0.35%) |
| NZ | 1 (0.03%) |
| PH | 1 (0.03%) |
| PK | 4 (0.11%) |
| PL | 5 (0.13%) |
| PR | 4 (0.11%) |
| PT | 14 (0.37%) |
| RO | 2 (0.05%) |
| RU | 3 (0.08%) |
| SE | 2 (0.05%) |
| SG | 6 (0.16%) |
| SI | 4 (0.11%) |
| SK | 1 (0.03%) |
| TH | 3 (0.08%) |
| TR | 5 (0.13%) |
| UA | 4 (0.11%) |
| US | 3,040 (80.96%) |
| VN | 1 (0.03%) |
| **company Size** |  |
| L | 454 (12.09%) |
| M | 3,153 (83.97%) |
| S | 148 (3.94%) |

## Problem 2. Bad Data Visualization

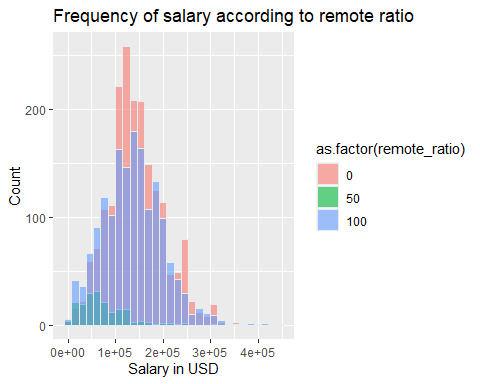
The first graph represents the average salary (in USD) of the employees with respect to the work year and also to the size of the company. Using a bar plot with stacked bars allows us to to have an idea of the amount of average salary for each year, but it is difficult to make a comparison between the 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:

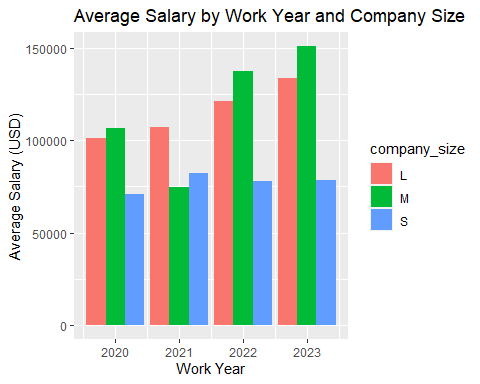
low\_salary <- subset(df, df$salary\_in\_usd < 150000 | df$salary\_in\_usd == 150000)  
medium\_salary <- subset(df, df$salary\_in\_usd > 100000 & df$salary\_in\_usd < 250000)  
high\_salary <- subset(df, df$salary\_in\_usd > 250000 | df$salary\_in\_usd == 250000)  
  
df <- mutate(df, salary\_category = 0)  
  
by\_year <- group\_by(df, work\_year, company\_size)  
new\_df <- summarize(by\_year, average\_salary = mean(salary\_in\_usd, na.rm = TRUE))  
  
# 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 each company size. 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.

# 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 subplot 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 too, 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 a value of 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)

