MAS8403 Formative Assignment: Data Science Salaries

This is a formative assignment for module MAS8403. **This assignment will not count towards your final grade for the module**, and is simply meant to provide an opportunity to receive feedback on your work. The task in this project is similar to that which you are required to do for the project for this module, so you are encouraged to take advantage of this opportunity.

Submit your report in PDF form to Canvas by **Friday 30th September at 16:30**.

Assignment Task

Write a 2 page summary of some interesting aspects of the `ds_salaries` dataset, including appropriate numerical and graphical summaries of the variables. All plots and tables should be included within the 2 pages. You do not need to submit any R code.

Things to Note

- Make sure the plots you include are readable, and have appropriate/understandable headings and axis labels
- Round your numerical summaries to a suitable number of decimal places (usually 2dp or 3dp is fine)
- Make sure all summaries/figures included in your report are discussed/commented on
- You won't be able to include/discuss every aspect of the dataset in the page limit. Be selective and choose what you think are the most important/interesting aspects to discuss.

Importing the Data

Download the `ds_salaries.csv` file from Canvas to your machine (and remember where you save it!

In RStudio, set your current working directory to be the folder you saved the `ds_salaries` file to by clicking Session -> Set Working Directory -> Choose Directory.

Alternatively you can set your working directory using the `setwd()` command

```
setwd("filepath")
```

and replace `filepath` with the path to the directory containing the `salaries` data.

We can import the data into R using the `read.csv` command

```
salaries = read.csv("ds_salaries.csv")
```

Data Summary

We can see the size of our dataset, either from looking in the `Environment` window in the top right corner of the RStudio display, or by using the `dim` command.

```
dim(salaries)
```

```
## [1] 607  11
```

From this we see we have 607 jobs included, and 11 variables observed on these jobs.

We can get a glimpse of what's contained in the data using the `head` command.

```
head(salaries)
```

```
##   work_year experience_level employment_type                 job_title salary
## 1      2020               MI              FT             Data Scientist  70000
## 2      2020               SE              FT Machine Learning Scientist 260000
## 3      2020               SE              FT          Big Data Engineer  85000
## 4      2020               MI              FT       Product Data Analyst  20000
## 5      2020               SE              FT  Machine Learning Engineer 150000
## 6      2020               EN              FT               Data Analyst  72000
##   salary_currency salary_in_usd employee_residence remote_ratio
## 1             EUR         79833                 DE            0
## 2             USD        260000                 JP            0
## 3             GBP        109024                 GB           50
## 4             USD         20000                 HN            0
## 5             USD        150000                 US           50
## 6             USD         72000                 US          100
##   company_location company_size
## 1               DE            L
## 2               JP            S
## 3               GB            M
## 4               HN            S
## 5               US            L
## 6               US            L
```

The variables in the data are:

- `work_year` – The year the salary was paid
- `experience_level` – The experience level in the job (**EN** = Entry Level/Junior, **MI** = Mid-level/Intermediate, **SL** = Senior level/Expert, **EX** = Expert level/Director)
- `employment_type` – The type of employment for the role (**PT** = Part time, **FT** = Full time, **CT** = Contract, **FL** = Freelance)
- `job_title` – The job role
- `salary` – The gross salary paid
- `salary_currency` – The currency the salary was paid in
- `salary_in_usd` – The salary converted into USD
- `employee_residence` – The employee's primary country of residence
- `remote_ratio` – The amount of work done remotely (**0** = No remote work, **50** = Partially remote, **100** = Fully remote)
- `company_location` – The country of the employer's headquarters
- `company_size` – The size of the company in terms of employee numbers (**S** = less than 50 employees, **M** = 50 - 250 employees, **L** = more than 250 employees)