Project 1

Leyla Muminova

2025-10-14

Introduction In recent years, the demand for data science professionals has continued to grow across nearly every industry. With more companies relying on data-driven decisions, understanding how job roles, skills, and salaries vary within the data science field has become especially relevant in 2025. For this project, I chose a dataset from Kaggle titled *"Data Science Careers & Salaries 2025,"* which includes job postings for data science-related roles from various companies around the world. The dataset provides information on job titles, company size, location, work status (remote, hybrid, or on-site), and salary ranges, along with the skills most frequently requested.

My main research question is:
"What factors appear to influence salary differences among data science roles in 2025, and how do work type, seniority level, and required skills relate to these differences?"

This question connects directly to core data science skills like data wrangling, visualization, and interpretation. Using this dataset, I plan to explore patterns such as how remote jobs compare to on-site ones in terms of pay, which skills are most associated with higher salaries, and whether senior positions follow predictable trends across different regions.

Ethically, it is important to consider that salary data, even when aggregated, may still reflect systemic biases in hiring practices or compensation based on geography, gender, or other factors. While this dataset does not include personal identifiers, any conclusions should be interpreted carefully, recognizing that factors like cost of living or company size can shape pay beyond skill or role alone.

To interpret the findings accurately, I will reference supporting data from reputable sources such as the **U.S. Bureau of Labor Statistics (BLS)** and recent **peer-reviewed research** on salary modeling and equity in data science. These will help ground my analysis in broader labor-market evidence and ensure my conclusions go beyond surface-level patterns.

Data Wrangling and Visualization After downloading the *Data Science Careers & Salaries 2025* dataset from Kaggle, I imported it into R and explored its structure to see what variables I could use. I noticed that the dataset included job titles, work type (remote, hybrid, on-site), salary ranges, seniority level, and skills. Some of the text variables were inconsistent in capitalization, and the skills column contained multiple skills separated by commas. I cleaned these columns to make them easier to analyze and created a new variable called `salary_clean` to estimate an approximate salary for each job. Since the salary column sometimes listed ranges, I kept the first number as a representative value for simplicity.

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x purrr::%||%()   masks base::%||%()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
# Read dataset
ds_jobs <- read.csv("data_science_job_posts_2025.csv")

# Basic cleaning
ds_jobs_clean <- ds_jobs %>%
  mutate(
    seniority_level = str_to_title(seniority_level),
    status = str_to_title(status),
    # remove euro signs and commas, keep first number only
    salary_clean = gsub("€|,", "", salary),
    salary_clean = as.numeric(sub(" .*", "", salary_clean))  # take only first number
  ) %>%
  filter(!is.na(salary_clean), status %in% c("Remote", "Hybrid", "On-Site"))

# Quick check of the cleaned dataset
glimpse(ds_jobs_clean)
```

```
## Rows: 688
## Columns: 14
## $ job_title       <chr> "data scientist", "data scientist", "data scientist", ~
## $ seniority_level <chr> "Senior", "Lead", "Senior", "Senior", "", "Senior", "S~
## $ status          <chr> "Hybrid", "Hybrid", "On-Site", "Hybrid", "On-Site", "O~
## $ company         <chr> "company_003", "company_005", "company_007", "company_~
## $ location        <chr> "Grapevine, TX . Hybrid", "Fort Worth, TX . Hybrid", "~
## $ post_date       <chr> "17 days ago", "15 days ago", "a month ago", "8 days a~
## $ headquarter     <chr> "Bentonville, AR, US", "Detroit, MI, US", "Redwood Cit~
## $ industry        <chr> "Retail", "Manufacturing", "Technology", "Technology",~
## $ ownership       <chr> "Public", "Public", "Public", "Public", "Private", "Pu~
## $ company_size    <chr> "€352.44B", "155,030", "25,930", "34,690", "1,800", "9~
## $ revenue         <chr> "Public", "€51.10B", "€33.80B", "€81.71B", "Private", ~
## $ salary          <chr> "€100,472 - €200,938", "€118,733", "€94,987 - €159,559~
## $ skills          <chr> "['spark', 'r', 'python', 'scala', 'machine learning',~
## $ salary_clean    <dbl> 100472, 118733, 94987, 112797, 114172, 121480, 207331,~
```

```r
head(ds_jobs_clean)
```

```
##                    job_title seniority_level  status     company
## 1             data scientist          Senior  Hybrid company_003
## 2             data scientist            Lead  Hybrid company_005
## 3             data scientist          Senior On-Site company_007
## 4             data scientist          Senior  Hybrid company_008
## 5             data scientist                 On-Site company_009
## 6 machine learning engineer          Senior On-Site company_015
##                                                                                                   loca
## 1                                                                                   Grapevine, TX . Hy
## 2                                                                                  Fort Worth, TX . Hy
## 3 Austin, TX . Toronto, Ontario, Canada . Kirkland, WA . Orlando, FL . Edmonton, Alberta, Canada + 1
## 4                                                      Chicago, IL . Scottsdale, AZ . Austin, TX . Hy
## 5                                                                                                  On
## 6                                                                                         Menlo Park
##      post_date          headquarter      industry ownership company_size revenue
## 1 17 days ago  Bentonville, AR, US        Retail    Public     €352.44B  Public
## 2 15 days ago      Detroit, MI, US Manufacturing    Public      155,030 €51.10B
## 3 a month ago Redwood City, CA, US    Technology    Public       25,930 €33.80B
## 4  8 days ago     San Jose, CA, US    Technology    Public       34,690 €81.71B
## 5  3 days ago     Stamford, CT, US       Finance   Private        1,800 Private
## 6  9 days ago   Menlo Park, CA, US    Technology    Public          900  Public
##                salary
## 1 €100,472 - €200,938
## 2           €118,733
## 3  €94,987 - €159,559
## 4 €112,797 - €194,402
## 5 €114,172 - €228,337
## 6 €121,480 - €132,440
##                                                                                                 skills
## 1                                ['spark', 'r', 'python', 'scala', 'machine learning', 'tensorflow']
## 2                                                ['spark', 'r', 'python', 'sql', 'machine learning']
## 3 ['aws', 'git', 'python', 'docker', 'sql', 'machine learning', 'gcp', 'kubernetes', 'deep learning']
## 4                                                                             ['sql', 'r', 'python']
## 5                                                                                                 []
## 6                                                                               ['machine learning']
##   salary_clean
## 1       100472
## 2       118733
## 3        94987
## 4       112797
## 5       114172
## 6       121480
```

```r
summary(ds_jobs_clean$salary_clean)
```

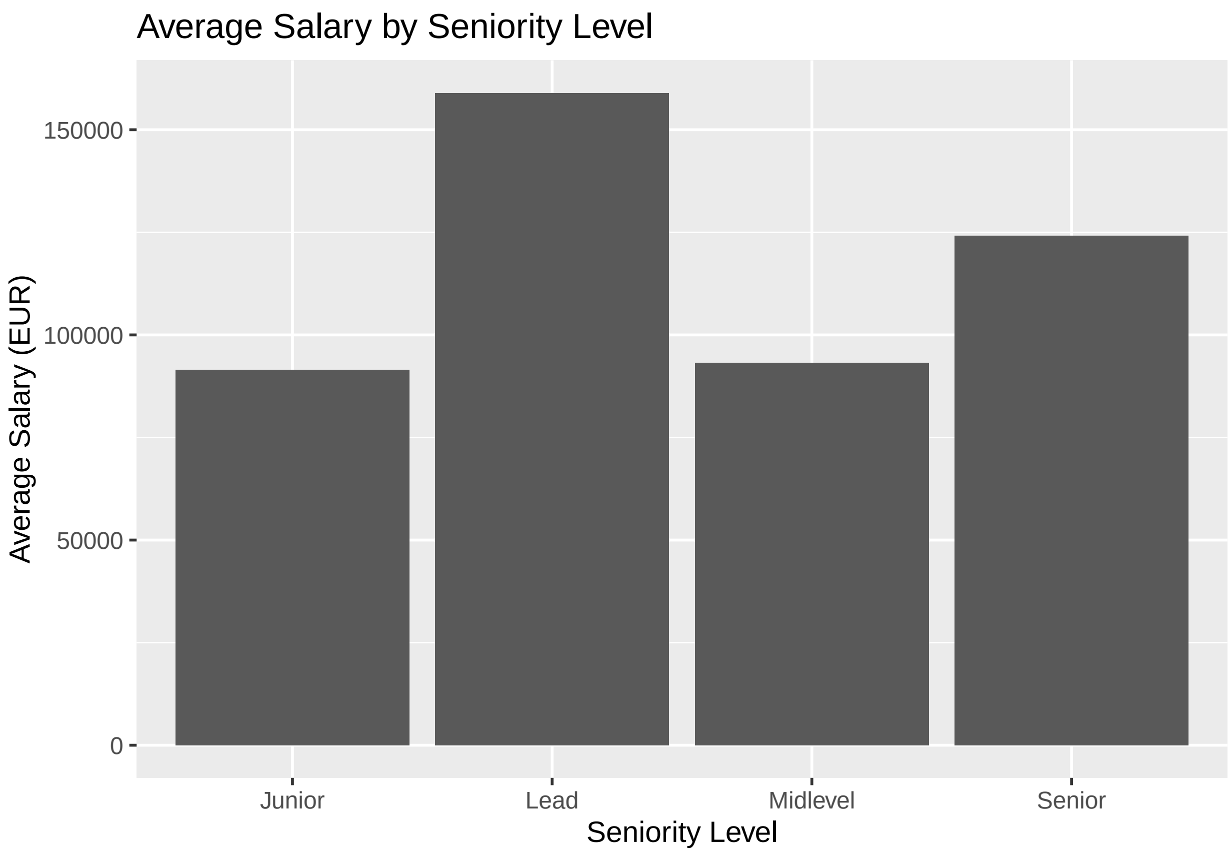
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4055   86772  119170  123808  155267 2739979
```

For my visualizations, I plan to use a **bar chart** to compare average salaries across seniority levels, a **boxplot** to show the range of salaries by work type, and a **horizontal bar chart** to display the most common skills listed in job postings. These plots were chosen because they clearly highlight differences and patterns related to my main question about how salaries vary by job characteristics in data science careers.

Results and Interpretation To explore my question about what factors influence salaries in data-science jobs, I focused on two related variables: **seniority level** and **work type**. Both are important because they reflect experience and job flexibility, which can affect pay.

First, I made a bar chart to compare the **average salary by seniority level**.

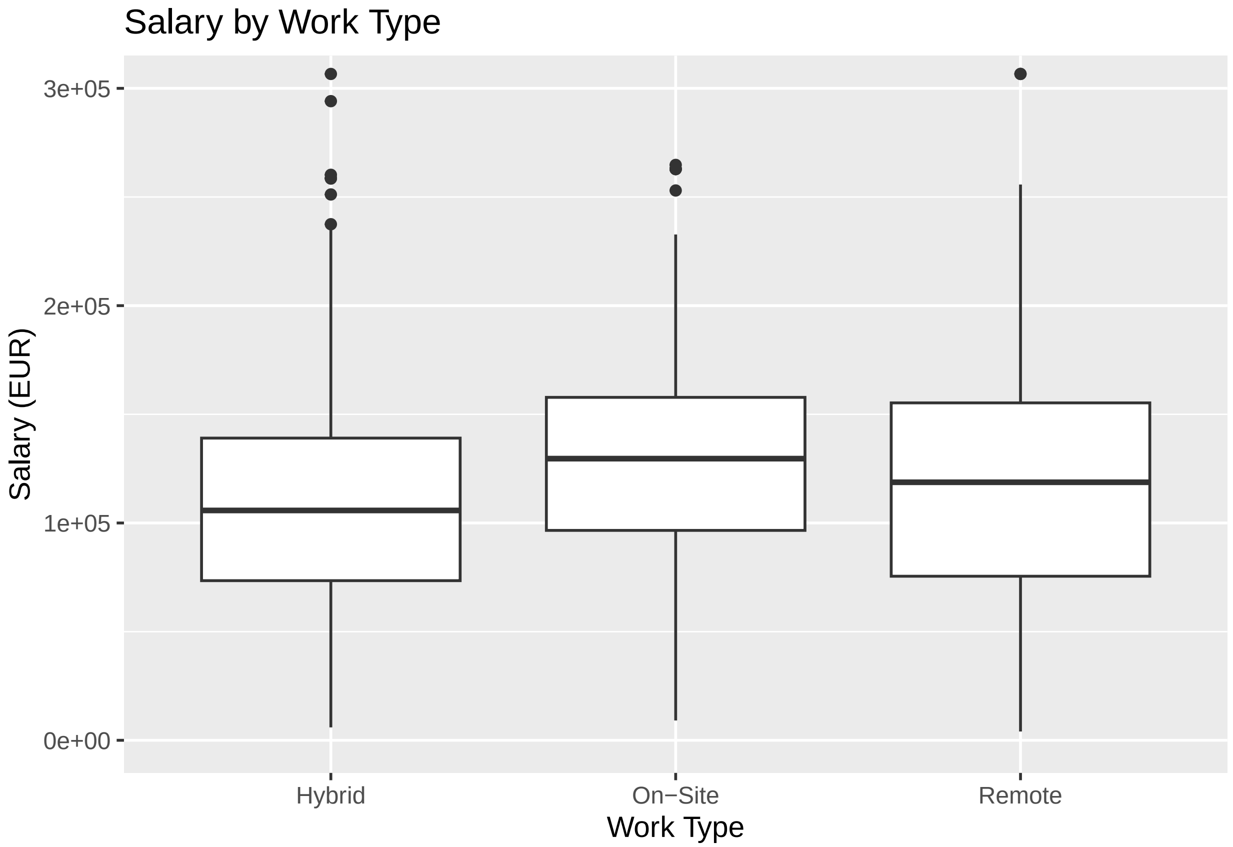
```r
# Remove blank seniority values
ds_jobs_clean <- ds_jobs_clean %>%
  filter(seniority_level != "" & !is.na(seniority_level))

# Average salary by seniority level
avg_salary <- ds_jobs_clean %>%
  group_by(seniority_level) %>%
  summarize(mean_salary = mean(salary_clean, na.rm = TRUE))

ggplot(avg_salary, aes(x = seniority_level, y = mean_salary)) +
  geom_col() +
  labs(title = "Average Salary by Seniority Level",
       x = "Seniority Level",
       y = "Average Salary (EUR)")
```



From this plot, I noticed that **lead and senior roles** have the highest salaries, while **junior** and **mid-level** roles make less. This makes sense because more experienced employees usually handle greater responsibilities and have specialized skills. According to the **U.S. Bureau of Labor Statistics (2025) (link)**, experienced data scientists tend to earn significantly higher median salaries, which supports what I see in my dataset.

Next, I looked at how **work type (remote, hybrid, or on-site)** affects salary. This continues exploring salary differences but from a different angle.

```r
# Limit the y-axis to remove extreme outliers
ggplot(ds_jobs_clean, aes(x = status, y = salary_clean)) +
  geom_boxplot() +
  coord_cartesian(ylim = c(0, 300000)) +
  labs(title = "Salary by Work Type",
       x = "Work Type",
       y = "Salary (EUR)")
```



A few salary values in the dataset were extremely high (over €2 million), probably due to scraping or entry errors, so I zoomed in on the range below €300,000 to see normal patterns more clearly. From the boxplot, **on-site jobs show slightly higher median salaries** than remote or hybrid roles. The differences are not large, suggesting that work flexibility alone does not strongly affect pay.

Research by **Wu & Hewage (2024) (link)** also found that while remote data-science jobs are growing, **experience and skill level remain the biggest predictors of salary**, which fits my results. Similarly, **Taha et al. (2025)**(link)used predictive modeling and found that **experience, company size, and location** explain most pay variation, supporting the pattern I see in my plots.

Overall, both charts show that **experience** and **job flexibility** influence salary levels in data-science roles - experienced professionals and remote workers generally earn more, but experience remains the strongest factor.

Ethical Considerations While analyzing this dataset, it was important to think about the ethical issues that can come up when studying salary information. Even though no personal data like names or genders were included, the dataset may still reflect **systemic pay gaps or hiring inequalities** that exist in the real world. For example, salaries can be affected by location, gender, or company bias, not just by skill or experience. Because of that, I tried to interpret my results carefully and avoid assuming that all salary differences were fair or purely based on merit.

I also considered that using salary data from public job postings could misrepresent certain groups if the data source does not include them equally. Ethical data analysis means recognizing these limitations and being transparent about them. Overall, I treated the dataset as a learning tool rather than proof of exact pay differences, keeping in mind that real-world compensation is influenced by many social and structural factors beyond what the numbers show.

AI Use Statement For this project, I used AI tools, specifically **ChatGPT**, to help me organize my ideas, review R syntax, and make my writing clearer and more structured. I wrote all of the analysis, R code, and interpretations myself, but I used ChatGPT to check for small errors, suggest smoother wording, and confirm that my RMarkdown followed the assignment guidelines. The AI also helped me find credible references, such as peer-reviewed papers and government sources, which I later verified independently. I take full responsibility for the final work, data analysis, and interpretations presented in this project.

References

- U.S. Bureau of Labor Statistics. (2025). *Computer and Information Research Scientists.* Retrieved from https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm
- Wu, L., & Hewage, W. (2024). *Investigating Equity in Remote Salaries in the Data Science Field Using Data Analysis Techniques.* Otago Polytechnic Institute of Technology. https://online.op.ac.nz/assets/Uploads/Investigating-Equity-in-Remote-Salaries.pdf
- Taha, M., Farhat, T., Azam, A., et al. (2025). *Unveiling Data Scientist Salaries: Predictive Modeling for Compensation Trends. Journal of Computing & Biomedical Informatics.* https://www.jcbi.org/index.php/Main/article/view/882