Data Science Salaries Project Report

# GitHub URL

<https://github.com/annie-maher/UCDPA_ann-marie_devlin.git>

# Abstract

This project combines two datasets to compare data science salaries over the vast range of data scientist positions using experience level and location in the world to look at correlations in data. I cleaned, investigated and analysed the data looking for relationships between experience, job titles and country specific salary range.

# Introduction

Data Science is of great interest to every industry to develop both their business and products. By using tools and techniques, data science can be a problem solving subject to find patterns, derive meaningful information, and make business decisions. Main target audience includes individuals searching for employment in the Data Science field. I found this dataset interesting as the **COVID-19 pandemic**has hugely appended the lives of millions of employed professionals, especially the ones working in the **data and analytics** industry. Not only has this crisis brought in the **new normal**of**remote working** but has also pushed analytics in the forefront, which in turn, has altered the way analytics professionals carry out their work.

The main research questions addressed are:

* What is the association between salary levels and experience?
* What are the top paying positions in the field of Data Science?
* Has Covid 19 impacted the remote working and increased demand in sector?
* Which country employs the most in Data Science sector?

# Datasets

Data Science Job Salaries Dataset contains 607 rows and 11 columns, each are:

1. work\_year: The year the salary was paid.
2. experience\_level: The experience level in the job during the year
3. employment\_type: The type of employment for the role
4. job\_title: The role worked in during the year.
5. salary: The total gross salary amount paid.
6. salary\_currency: The currency of the salary paid as an ISO 4217 currency code.
7. salary\_in\_usd: The salary in USD
8. employee\_residence: Employee's primary country of residence in during the work year as an ISO 3166 country code.
9. remote\_ratio: The overall amount of work done remotely.
10. company\_location: The country of the employer's main office or contracting branch
11. company\_size: The median number of people that worked for the company during the year.

World Data html table I web scrapped using pandas had 69 rows and 4 columns each are;

1. Rank – Highest paid countries with highest at 1 and lowest at 69
2. Country – name of country
3. Avg. Income per Year – average income per year 2021
4. Avg. Income per month – average monthly paid income for 2021

# Implementation Process

**Data Importing**

I imported python modules Numpy, Pandas, Matplotlib and Seaborn. I downloaded a csv dataset from Kaggle. I performed a web scrap using pandas to scrap html table from World Data website on country avg salaries for 2021. Reference to websites you will find at end of report.

**Data Cleaning and Preparation**

Before beginning exploratory data analysis, I prepared and cleaned the datasets to enhance analytics performance. Firstly, explored dataset ds\_salaries downloaded from Kaggle by creating a pandas dataframe and assigned it to a variable (Salaries\_df). Then I applied the info, shape and describe functions from the pandas library to familarise myself with the dataframe. I found unnecessary columns used and some that needed to be changed to understand the data when I present findings. I explored the web scrap table and explored headings for any missing items that needed correction. I web scrapped using the pandas read\_html function downloading the table and assigning a variable(df) and saving it as csv file. I created another pandas dataframe for this new csv using the read.csv function.

**Cleaning**

I replaced values using the function replace so we could easily understand information in the graph when presented. I changed the following columns to read;

Remote Ratio - ‘Remote’, ‘Hybrid’ & ‘On-Site’,

Experience level - ‘Entry’, ‘Mid’, ‘Senior’ & ‘Executive’

Company Size – ‘Small’, Medium’ & ‘Large’

I dropped some columns that were not needed using drop function from pandas. I dropped the ‘Salary’, ‘Salary Currency’ and ‘Unnamed’ columns in dataframe(Salaries-df)

I used the drop function again on the dataframe(df) to drop column Average Income per month as I will not need this when I merge both dataframes.

For the $ sign in column ‘average income per year’ I deleted this by using the replace function from pandas and used the same function to delete the \* sign in ‘country’ column in dataframe(df)

For the purpose of merging both dataframes it was important I had a column that was common in both dataframes so I could use a left join on merge. I renamed the ‘company location’ column in dataframe(Salaries\_df) to ‘Country’ using rename function. I also installed country\_converter from library and converted the country abbreviations to name of country, so I could make this column available for merging both datasets.

**Merging Datasets**

I validated how many matching countries were in each dataset by using the isin function along with value\_counts before I performed merge.

I then used pandas merge function to perform a left join so we could keep all the rows from the dataframe (Salaries\_df) and just bring in the columns from data frame(df) that was common.

When I had both dataframes merged into one(result\_df) I had to further cleaned columns to make comparisons. I converted ‘avg country income for 2021’ to a datatype integer using the astype function.

Renamed column to include country in avg salary 2021 so easier to read comparative analysis.

# Results and Visualisations

I first looked at plotting heat map correlation between numeric attributes in dataframe. We can see from below heat map there is not much correlation between the columns.Chart, bar chart

Description automatically generated

**Experience Level**

Senior experience level accounts for majority (280 or 46%) of experience level in our data whereas Executive level is the least (26 or 4%)

Chart, bar chart

Description automatically generated

The top 5 job titles in dataframe is mostly lead by mid to senior level staff with very low % at executive level.

Chart, bar chart

Description automatically generated

**Job Titles**

Various job titles in Data Science can be plotted on bar chart below. The most popular are;

1. Data Scientist 3. Data Analyst

2. Data Engineer 4. Machine Learning EngineerChart

Description automatically generated

**Average Salary**

I compared the average salary from a sample of job titles to the country average salary for year 2021. I had to extract data in relation to 2021 from dataframe and assigned new variable and then compared to the country average salary for 2021. You can see by having a data scientist qualification your earnings in comparison to country average salary are higher, with ML Engineer having the largest favorable gap.

Chart, bar chart

Description automatically generated

**Employment Type**

It is evident from below bar chart that most data scientist work full time.

Chart, histogram

Description automatically generated

**Number of Data Science Jobs Per Year**

Data Science job sector is growing every year from 2020 and the most number of employees occurred in 2022.

Chart, bar chart

Description automatically generated

**Remote Ratio**

The greatest number of employee work remotely

Chart, bar chart

Description automatically generated

We can further analyse this figure by looking are work year and see if there is any correalation with the global covid 19 pandemic. We can see remote working is increasing a lot from 2020 and this could be a factor of the pandemic but also the data science sector growith in the past two years.

Chart, bar chart

Description automatically generated

**Top 5 Locations in World**

* Unted states employs the most data science jobs with 355
* United Kingdom employs 46
* The remainder divided between Canada, Germany, India & France

Chart, bar chart

Description automatically generated

**Top Average Salary per Country**

* Russia pays the highest average salary among all of the countries at 157.5k USD
* United States is second with New Zealand and Israel third and fourth.

Chart, bar chart

Description automatically generated

**Insights**

* Most of the employees are Mid-level and Senior executives, so it is very rare to get a job in data science as entry level. You need to be qualified to enter this sector.
* The remote ratio chart by year shows the existence of a new culture that is most favored by Data Science workers, namely remote working. We can see this is a positive impact from Covid 19 pandemic where employers have trust in their employees and can see the benefits of work life balance by offering this flexibility to future employees.
* Most Data Science jobs are Full-time positions, and this as expected, the importance of Data Science roles during pandemic was very evident. Databases to collect covid data to produce reports for media coverage so countries could make informed decisions regarding lock downs, increase testing centers and share test results in an effective way.
* Most Data Science employees are resident in the United States. However, the location where the highest average salary is paid is Russia; followed closely by the United States.
* The most common job titles in the Data Science field are Data Scientist, Data Engineer, Data Analyst and Machine Learning Engineer.

# Machine Learning

# We have seen from the above dataframe the experience level, employment type, job title and company location impact salary in usd, although all of these could be classed as independent variables. The salary in usd is the dependent variable. I would use regression analysis to predict future salaries using any of the independent variables to try and predict future salary using linear regression of historic data.

# References

# Datasets

<https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries>. <https://www.worlddata.info/average-income.php>

Other resource Websites

[www.github.com](http://www.github.com)

[www.kaggle.com](http://www.kaggle.com)

[www.datacamp.com](http://www.datacamp.com)

[country-converter · PyPI](https://pypi.org/project/country-converter/)