

# DATA SCIENCE JOB SALARIES

Hello, this readme will show the aims for project and each steps i took into my Python project. Hope you have a nice read. 😊

## AIMS

The goal for my project is to identify the:

- Country with the highest earners.
- Company's size highest value counts.
- Highest earners with their experience level.
- Also highest earners using company size.
- Highest earners by employee residence with their employment type.
- Also highest job Branch earners.
- Average salaries by employment type.
- Most experienced level and their place of residence.

## STEPS

it will be divided into few sections:

- Imports
- Data Cleaning.
- Data Manipulations.
- Visualization.

## IMPORTS

I imported opendataset od for our Url downloads and named the link datasets\_url and download using od.download function. I made data\_dir to show the directory.

I imported os and used the function os.listdir of data\_dir to show us the csv file. Also,i imported Pandas as pd, and created a dataframe known as ds\_job\_salary\_df to read the csv file using pd.read\_csv function.

My final imports for visualization were seaborn sns,matplotlib.

## DATA CLEANING

I created a copy called ds\_job\_sal\_df.

### FUNCTIONS USED:

#### **.shape:**

Shows shape of the data.

#### **.column:**

Brings a list of the columns.

#### **.info:**

This provided the columns, Non-null count and Data type.

#### **.drop:**

I removed the column 'Unnamed: 0','salary'

#### **.isnull**

It shows the amount nulls within the data which was 0.

## DATA MANIPULATION

I created a new column called Job\_branch from Job\_title.

#### **.describe:**

This shows the counts,mean,std,min to max of the integer Datatype found the csv.

#### **.nunique**

To count the distinct number of company\_location,salary\_currency.

## value\_counts

This was used to show series counts of company\_location in a descending order. also the work\_year and company\_size column

## .replace

I used this function to divide the job\_branch into 3 categories.

```
ds_job_sal_df['job_branch'].replace(['Data Scientist',  
                                     'Research Scientist',  
                                     'Data Science Manager',  
                                     'Machine Learning Scientist',  
                                     'Principal Data Scientist',  
                                     'AI Scientist',  
                                     'Data Science Consultant',  
                                     'Director of Data Science',  
                                     'Applied Data Scientist',  
                                     'Applied Machine Learning Scientist',  
                                     'Head of Data Science',  
                                     'Lead Data Scientist',  
                                     'Data Specialist',  
                                     'Staff Data Scientist',  
                                     'Machine Learning Manager'], 'Data Science related jobs', inplace=True)
```

Python

\*Picture1: Data science Category

```
ds_job_sal_df['job_branch'].replace(['Data Analyst',  
                                     'Data Analytics Manager',  
                                     'BI Data Analyst',  
                                     'Head of Data',  
                                     'Business Data Analyst',  
                                     'Lead Data Analyst',  
                                     'Financial Data Analyst',  
                                     'Product Data Analyst',  
                                     'Principal Data Analyst',  
                                     'Marketing Data Analyst',  
                                     'Finance Data Analyst',  
                                     '3D Computer Vision Researcher', 'Data Analytics Lead'], 'Data Analysis re
```

Python

\*Picture2: Data analyst Category

```
ds_job_sal_df['job_branch'].replace(['Lead Machine Learning Engineer',
                                     'NLP Engineer',
                                     'Head of Machine Learning',
                                     'Big Data Architect',
                                     'Director of Data Engineering',
                                     'Cloud Data Engineer',
                                     'ETL Developer',
                                     'Principal Data Engineer',
                                     'Data Science Engineer',
                                     'Computer Vision Software Engineer',
                                     'Machine Learning Developer',
                                     'Machine Learning Infrastructure Engineer',
                                     'Data Analytics Engineer',
                                     'Analytics Engineer',
                                     'Data Engineering Manager',
                                     'Lead Data Engineer',
                                     'ML Engineer',
                                     'Computer Vision Engineer',
                                     'Big Data Engineer',
                                     'Data Architect',
                                     'Machine Learning Engineer',
                                     'Data Engineer'], 'Data Engineering related jobs', inplace=True)
```

Python

\*Picture3: Data engineering Category

## **.groupby**

I used this to group 2 columns together with their mean value.

# **VISUALIZATION**

I set the font size, figsize, facecolor and style.

## **sns.barplot:**

I created a df showing salary\_currency\* 100 /ds\_job\_sal\_df count to get the percentage. I used the created df as x and it's index as y for the bargraph.

## **plt.pie**

I created title called Size of Company, using company\_size column and it's index for laels with an autopct='%1.1f%'. It has startangle of 180.

## **sns.scatterplot**

It has figsize of 12,6 and a title

It also contains x-axis,y-axis and hue with s of 100.

Also, rotation can be 75,xlabel and ylabel can be included

## **sns.histplot**

It contains a title, x-axis and y-labels and a hue.

## **sns.pairplot**

Using the df and a hue showed my plots.

## **.plot**

I created df by Grouping 2 columns together and using the function value\_counts  
I gave it figsize,grid and used the descriptive column for my x-axis.

Also,you can determine the kind of Graph you want using 'kind='.