**Data Set Title**

**Exploratory Analysis**

Name, [bbrandon@bellarmine.edu](mailto:bbrandon@bellarmine.edu)

Name, [jpenaloza@bellarmine.edu](mailto:jpenaloza@bellarmine.edu)

1. **INTRODUCTION**

Short description of the data set including a reference to where it can be found and why you chose it.

1. **DATA SET DESCRIPTION**

Narrative summary of the data set: e.g. this data set contains 398 samples with 7 columns with various data types. A complete listing is shown in **Table 1**. For data types you want to indicate two things (nominal, ordinal, interval, or ratio) and the Pandas data type. For example, age might be ratio/int32. For missing data, indicate what percentage of data from that column are missing. Ensure you check to for NaN, NA, or any other indicators that actually mean missing data.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| work\_year | int | 0% |
| experience\_level | chr | 0% |
| employment\_type | chr | 0% |
| job\_title | chr | 0% |
| salary | int | 0% |
| salary\_currency | chr | 0% |
| salary\_in\_usd | int | 0% |
| employee\_residence | chr | 0% |
| remote\_ratio | int | 0% |
| company\_location | chr | 0% |
| company\_size | chr | 0% |

1. **Data Set Summary Statistics**

Narrative introduction to the section.

**Table 2: Summary Statistics for ds\_salaries\_2023:**

*A screenshot of a computer code

Description automatically generatedA screenshot of a computer

Description automatically generated*

There should be a table for **EACH** categorical variable.

Table 3: Proportions for data\_salaries\_2021 (n=yyy)

|  |  |  |  |
| --- | --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* | |
| experience\_level | *EN = 320 EX = 114 MI = 805 SE = 2516* |  | |
| *employment\_type* | *CT FL FT PT*  *10 10 3718 17* |  |
| *job\_title* | *3D Computer Vision Researcher AI Developer*  *4 11*  *AI Programmer AI Scientist*  *2 16*  *Analytics Engineer Applied Data Scientist*  *103 10*  *Applied Machine Learning Engineer Applied Machine Learning Scientist*  *2 12*  *Applied Scientist Autonomous Vehicle Technician*  *58 2*  *Azure Data Engineer BI Analyst*  *1 9*  *BI Data Analyst BI Data Engineer*  *15 1*  *BI Developer Big Data Architect*  *13 2*  *Big Data Engineer Business Data Analyst*  *11 15*  *Business Intelligence Engineer Cloud Data Architect*  *4 1*  *Cloud Data Engineer Cloud Database Engineer*  *3 5*  *Compliance Data Analyst Computer Vision Engineer*  *1 18*  *Computer Vision Software Engineer Data Analyst*  *5 612*  *Data Analytics Consultant Data Analytics Engineer*  *2 6*  *Data Analytics Lead Data Analytics Manager*  *2 22*  *Data Analytics Specialist Data Architect*  *2 101*  *Data DevOps Engineer Data Engineer*  *1 1040*  *Data Infrastructure Engineer Data Lead*  *6 2*  *Data Management Specialist Data Manager*  *1 29*  *Data Modeler Data Operations Analyst*  *2 4*  *Data Operations Engineer Data Quality Analyst*  *10 7*  *Data Science Consultant Data Science Engineer*  *24 5*  *Data Science Lead Data Science Manager*  *8 58*  *Data Science Tech Lead Data Scientist*  *1 840*  *Data Scientist Lead Data Specialist*  *2 14*  *Data Strategist Deep Learning Engineer*  *2 6*  *Deep Learning Researcher Director of Data Science*  *1 11*  *ETL Developer ETL Engineer*  *10 2*  *Finance Data Analyst Financial Data Analyst*  *1 3*  *Head of Data Head of Data Science*  *10 9*  *Head of Machine Learning Insight Analyst*  *1 2*  *Lead Data Analyst Lead Data Engineer*  *5 6*  *Lead Data Scientist Lead Machine Learning Engineer*  *9 3*  *Machine Learning Developer Machine Learning Engineer*  *7 289*  *Machine Learning Infrastructure Engineer Machine Learning Manager*  *11 3*  *Machine Learning Research Engineer Machine Learning Researcher*  *4 6*  *Machine Learning Scientist Machine Learning Software Engineer*  *26 10*  *Manager Data Management Marketing Data Analyst*  *1 2*  *Marketing Data Engineer ML Engineer*  *1 34*  *MLOps Engineer NLP Engineer*  *4 7*  *Power BI Developer Principal Data Analyst*  *1 2*  *Principal Data Architect Principal Data Engineer*  *1 2*  *Principal Data Scientist Principal Machine Learning Engineer*  *8 1*  *Product Data Analyst Product Data Scientist*  *5 1*  *Research Engineer Research Scientist*  *37 82*  *Software Data Engineer Staff Data Analyst*  *2 1*  *Staff Data Scientist*  *1* |  |
| *salary\_currency* | *AUD BRL CAD CHF CLP CZK*  *9 6 25 4 1 1*  *DKK EUR GBP HKD HUF ILS*  *3 236 161 1 3 1*  *INR JPY MXN PLN SGD THB*  *60 3 1 5 6 2*  *TRY USD*  *3 3224* |  |

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn’t make sense)

Table 4: Correlation Table/Tables

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

After the table with the raw data, include a heatmap of the correlation matrix as a figure.

1. **DATA SET GRAPHICAL EXPLORATION**

Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

* 1. *Distributions*
  2. *ScatterPlots / Pairwise Plots (continuous variables)*
  3. *Barcharts (categorical variables)*
  4. *Other Plots - don’t skimp – there are likely other plots that would be useful that I haven’t already specified. Include those in this section.*

All figures should be cited formatted like this and mentioned in the text.

*A graph of a graph

Description automatically generated*

**Figure 1: Comparison of X/Y from dataset (single plot) (8 pt)**

*A graph of employment type

Description automatically generated*

**Figure 2: (a) Function Output (b) A against B (multiple plots) (8 pt)**

*A graph of a company size

Description automatically generated*

*A graph of a graph showing a graph

Description automatically generated with medium confidence*

*A graph of salary

Description automatically generated*

*A graph with a bar graph

Description automatically generated*

1. **SUMMARY OF FINDINGS**

Finish up with a paragraph or two of summarizing your findings about this data set.

Through this dataset, we explored the trends of salaries for the different ramifications of data science and how each one of the ramifications has an impact on the salaries that everyone gets, in respect to their job title, years of experience, employment type, and company size. Generally, the more experience an individual has the higher the salary, and the higher the difficulty of the job title is the higher the salary will be. Also, there seem to be lots of outliers in the dataset that seem to not be represented accordingly because of the lack of workers in that job title.