**Data Set Title**

**Exploratory Analysis**

Brandon Boyle, [bboyle@bellarmine.edu](mailto:bboyle@bellarmine.edu)

Jonathan Penaloza Rumie, [jpenaloza@bellarmine.edu](mailto:jpenaloza@bellarmine.edu)

1. **INTRODUCTION**

The dataset that we chose is titled “Data Science Salaries 2023”. It consist of 11 variables with seven being strings and four being integers. We chose this topic as it has a big relevance to our major. We were interested in seeing just how much some data scientist make.

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 XXX (name of dataset)**

*A screenshot of a computer code

Description automatically generatedA screenshot of a computer

Description automatically generated*

1. **MLR & SLR**

We started by looking at the MLR for Salary. We were able to closely predict the salary.

*A screenshot of a computer

Description automatically generated*

Next we did a MLR on the remote ratio.

A screenshot of a computer

Description automatically generated

Then we did SLR between salary and work year. We were able to find a trend of an increase in salary every year with 2020 having the lowest predicted salary and 2023 having the highest predicted salary

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*

**Here we looked at the experience level of workers to find out how many entry level people there were compared to senior. And as shown, there are a bunch more seniors that we have data on**

*A graph of employment type

Description automatically generated*

**Here we have the employment type. We wanted to see what the main employment type was**

*A graph of a company size

Description automatically generated*

*Here we have company size. We were looking for the size of a majority of the companies on our data set*

*A graph of salary

Description automatically generated*

*Here we have a histogram of the salary of the workers and as we can see, most DS workers earn between 100,000 and 175,000 a year.*

*A graph with a bar graph

Description automatically generated*

*Here we have a boxplot of salary, backing up that a majority of the people earn between 100,000 and 175,000*

1. **SUMMARY OF FINDINGS**

In the end we learned that the salaries of DS workers are increasing year after year with workers earing between 100,000 and 175,000 a year. Most people have senior level experience and work full time in a medium sized companies.