1 Example Report: Analysis of Data Scientists Salaries

This example uses open data from Kaggel (https://www.kaggle.com/datasets/vladimirmijatovic/data-scientists-salaries-worldwide-annual-survey) currently only one data source but later on I will expand the data sources from other datasets as well this dataset is an open source and public in kaggle vladimirmijatovic is the author of dataset which is collected from Reddit opensource data.

1.1 Main Question

Dive into the various compensation possibilities - and use it to prepare to negotiate when the time is right! ## Description Every year there is an annual thread on Salaries in Data Science on Reddit. Data scientists from all over the world self-report their own salaries, bonuses and other compensation. Some of the salaries are insanely high!

1.2 About Data

data contains 14 columns and 55 rows including Index([‘date’, ‘title’, ‘location’, ‘salary’, ‘company_industry’, ‘education’, ‘prior_experience’, ‘bonus’, ‘stocks’, ‘total_comp’, ‘additional_benefits’, ‘tenure_length_period’, ‘tenure_length_period_units’, ‘survey_year’], dtype=‘object’) initially the data is dirty have to calculate the missing values first so here in picture of missing values we can see.

1.3 Data cleaning

1.3.1 Cleaning and Conversion Functions:

1.3.2 convert_salary:

This function takes a salary value as input, removes any non-numeric characters except for decimal points and ‘k’ (for thousands), and converts it to a numeric value. It also handles ‘k’ suffix by multiplying the numeric value by 1000 if present. ### Data Imputation: Missing values in columns like ‘salary’, ‘total_comp’, ‘bonus’, ‘stocks’, and ‘tenure_length_period’ are filled: ‘salary’, ‘total_comp’: Filled with the median value of their respective columns. ‘bonus’, ‘stocks’: Filled with 0. ‘title’, ‘location’, ‘company_industry’, ‘education’: Filled with the mode (most frequent value) of their respective columns. ‘prior_experience’: Filled with ‘Unknown’. ### Normalization:

The salary, bonus, stocks, and total_comp columns are normalized using StandardScaler from sklearn.preprocessing. This step standardizes the features by removing the mean and scaling to unit variance.

1.3.3 Handling Outliers:

Outliers are handled by capping the values at the 99th percentile using a lambda function.

1.3.4 Final Output:

The DataFrame df is displayed with the cleaned and preprocessed data.

1.4 Visualization Using histograms

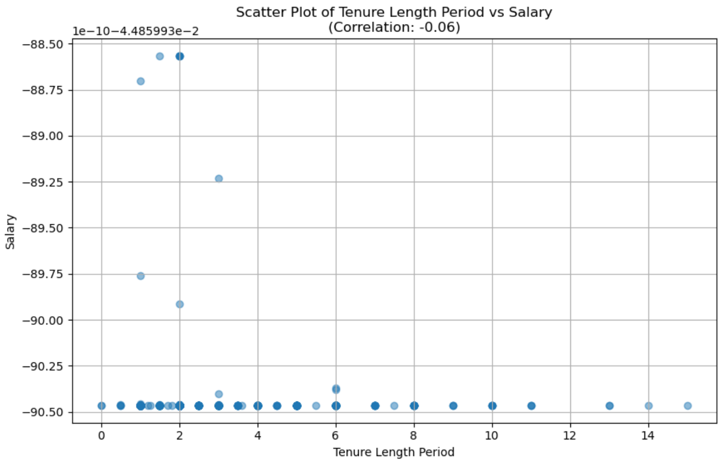
Histogram indicates The salary, bonus, stocks, and total compensation distributions are right-skewed, indicating that most data scientists earn lower amounts, with fewer receiving high compensation. Tenure length periods are varied, suggesting common employment durations. Recent survey years reflect current trends, and 'Data Scientist' is the predominant job title. Key hubs for data science jobs are highlighted in location data. The technology sector is the leading industry, and many data scientists hold advanced degrees. Diverse prior experiences are common pathways into data science. Companies offer a range of additional benefits, and different units measure tenure length, reflecting varied standards across industries

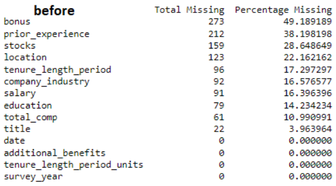
1.5 correlation using scater plot

Correlation between Tenure Length Period and Salary: -0.059131160687359255

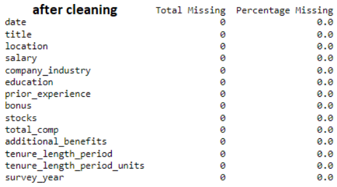
The plot shows that there is no significant linear relationship between tenure length period and salary. The correlation is weak and slightly negative, indicating that other variables likely have a more substantial impact on salary levels for data scientists.

This kind of analysis is crucial because it helps identify which factors truly influence salary, guiding more targeted investigations and decisions.

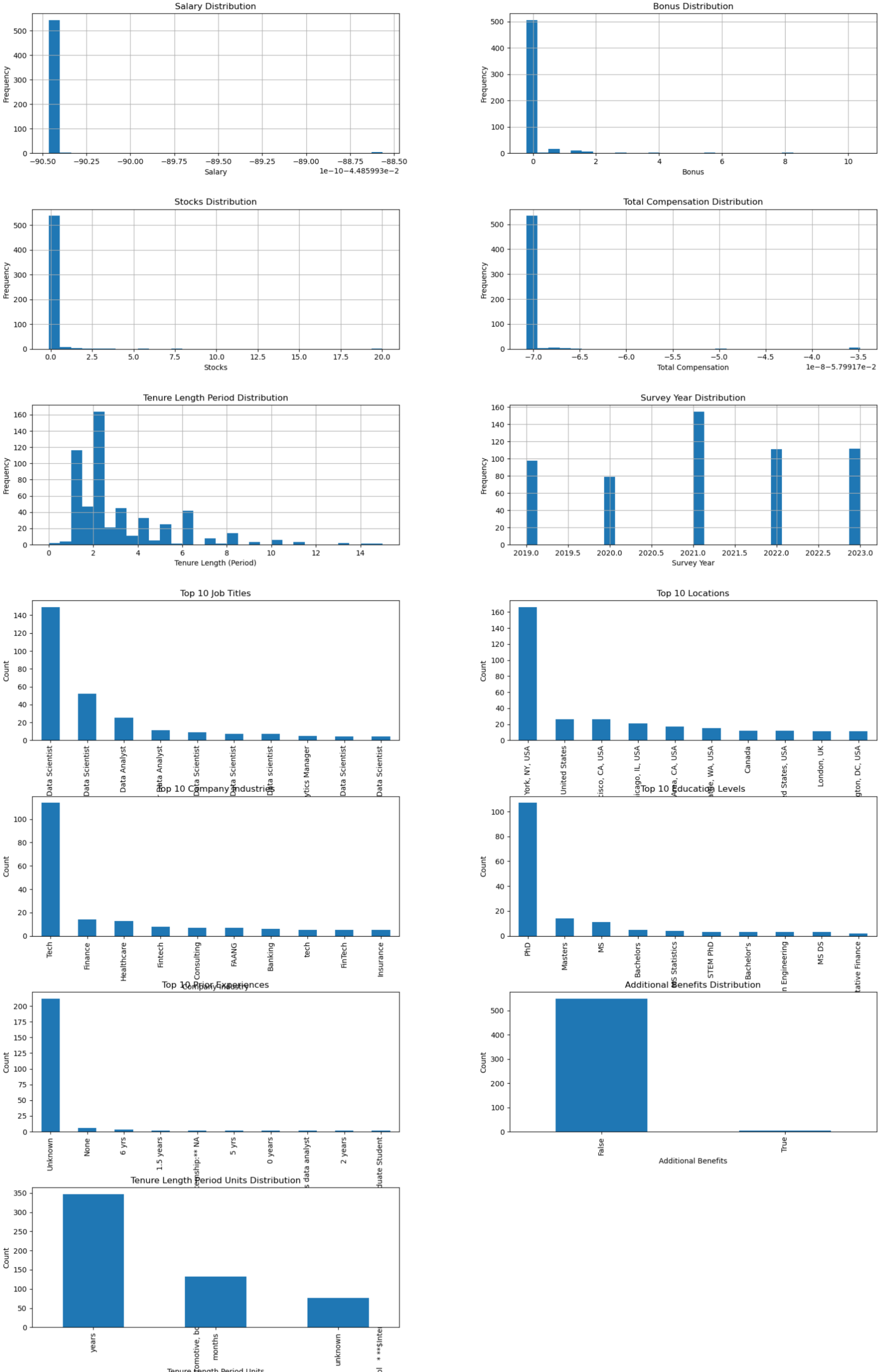




```
     before                   Total Missing  Percentage Missing
bonus                                   273           49.189189
prior_experience                        212           38.198198
stocks                                  159           28.648649
location                                123           22.162162
tenure_length_period                     96           17.297297
company_industry                         92           16.576577
salary                                   91           16.396396
education                                79           14.234234
total_comp                               61           10.990991
title                                    22            3.963964
date                                      0            0.000000
additional_benefits                       0            0.000000
tenure_length_period_units                0            0.000000
survey_year                               0            0.000000
```



```
     after cleaning           Total Missing  Percentage Missing
date                                      0                 0.0
title                                     0                 0.0
location                                  0                 0.0
salary                                    0                 0.0
company_industry                          0                 0.0
education                                 0                 0.0
prior_experience                          0                 0.0
bonus                                     0                 0.0
stocks                                    0                 0.0
total_comp                                0                 0.0
additional_benefits                       0                 0.0
tenure_length_period                      0                 0.0
tenure_length_period_units                0                 0.0
survey_year                               0                 0.0
```


Salary Distribution
Frequency
Salary
−90.50
−90.25
−90.00
−89.75
−89.50
−89.25
−89.00
−88.75
−88.50
1e−10−4.485993e−2
Bonus Distribution
Frequency
Bonus
Stocks Distribution
Frequency
Stocks
Total Compensation Distribution
Frequency
Total Compensation
1e−8−5.79917e−2
Tenure Length Period Distribution
Frequency
Tenure Length (Period)
Survey Year Distribution
Frequency
Survey Year
Top 10 Job Titles
Count
Data Scientist
Data Scientist
Data Analyst
r Data Analyst
Data Scientist
Data Scientist
Data scientist
ytics Manager
Data Scientist
Data Scientist
Top 10 Locations
Count
York, NY, USA
United States
cisco, CA, USA
icago, IL, USA
Area, CA, USA
ttle, WA, USA
Canada
d States, USA
London, UK
gton, DC, USA
Top 10 Company Industries
Count
Tech
Finance
Healthcare
Fintech
Consulting
FAANG
Banking
tech
FinTech
Insurance
Company Industry
Top 10 Education Levels
Count
PhD
Masters
MS
Bachelors
MS Statistics
STEM PhD
Bachelor's
n Engineering
MS DS
tative Finance
Top 10 Prior Experiences
Count
Unknown
None
6 yrs
1.5 years
nship:** NA
5 yrs
0 years
s data analyst
2 years
duate Student
Additional Benefits Distribution
Count
False
True
Additional Benefits
Tenure Length Period Units Distribution
Count
years
omotive, bo
months
unknown
ol * **$Inte
Tenure Length Period Units