NSBM GREEN UNIVERSITY TOWN

Faculty of Computing
Department of Data Science
Group Assignment

A Statistical Exploration Of The Data Science Job Market

INDEX NUMBER	**STUDENT NAME**		
24715	**N G D Nethmini**		
25648	**A S A Gunathilaka**		
25063	**S A D H M Samarathunga**		
GROUP NAME	**Group D**		
YEAR OF STUDY AND SEMESTER	**3rd year 1st semester**		
MODULE CODE	**DS304.3**	**MODULE NAME**	**Advanced Statistics for Data Science**
MODULE LECTURER	**Ms. Kavishka Rajapaksha**	**SUBMISSION DATE**	
For office purpose only:			
GRADE / MARK			

Abstract

The project aims to analyze the salary variable by considering the other factors which have to possibility of influencing the salary scale in the field of data science. To initial the project, identifying the scope of the project, project background and objectives are essential to lead. Studying about the factors such as company size, employee type, experience level, job titles and analyzing how those factors involve to effect salary of employees who are in the data science field is the identified project scope. Finding a dataset as secondary resource and starting off with exploratory data analysis to highlight the insights and patterns of data was the first task of the project. Moving towards the statistical techniques, the simple linear regression and hypothesis testing were conducted under experience level, company size and employee type variables to analyze the data based on salary which converted to USD dollars. After the analysis, the final interpretations and insights of the project were discussed.

Contents

Project Overview

Problem background

In obedience to the proposed project, the identified problem background is surrounded by the data science job market. In the transforming world, organizations and industries consist of employees who are capable of facing challenges in their relevant fields. In order to handle those challenges, nowadays, the data science job field filters out employees based on various type of factors such as salary scale, job title, company size, amount of work done, employment type, etc. The proposed project comprised extracting meaningful information and insights for people who are interested in the field of data science. Approaching the statistical techniques and methodologies to identify the mentioned aspects.

Problem Statement

The project statement is defined as examining the factors involved in the salary scale in the field of data science. Focusing on particular variables, such as work experience level, job title, and company characteristics determined to study the data distribution & how parameters impacted the field of data science.

Problem Objective

The primary objective of this project is to extract insights for employees and job seekers in the field of data science. To uplift the project objectives, the project proceeds by analyzing the factors that impacted the salary scale of the data science field.

Methodology

Data collection and preprocessing

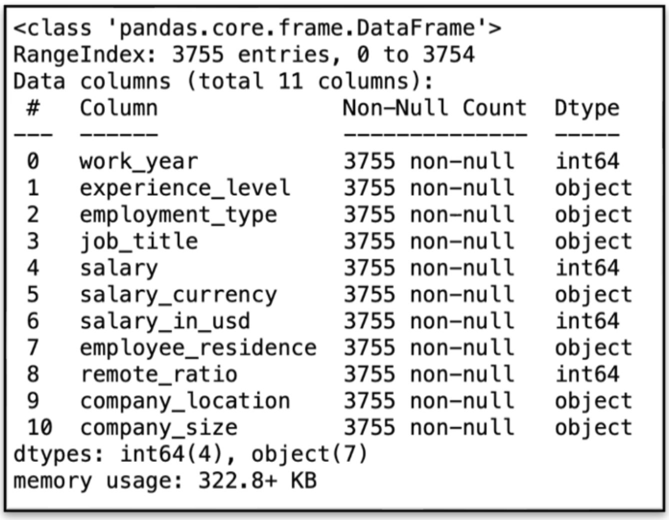
After identifying the problem statement and studying the problem background, the team members approached the data collection method. By considering the available resources, a secondary data collection method was selected. The dataset utilized for the project consists of salary information which is influenced by different types of factors. Pondering as a secondary source, the dataset was selected from the Kaggle website. It includes a considerable amount of variables that are involved in studying the impact of the salary scale in the field of data science.

Description of the dataset

The selected dataset provides insight into salary trends of data scientists, including a total of 11 variables, 4 numerical and 7 categorical variables. When conducting the data preprocessing several steps were taken. Encoding categorical variables into a numerical format to be used in the analysis process. Encoding the data is compatible with linear regression.

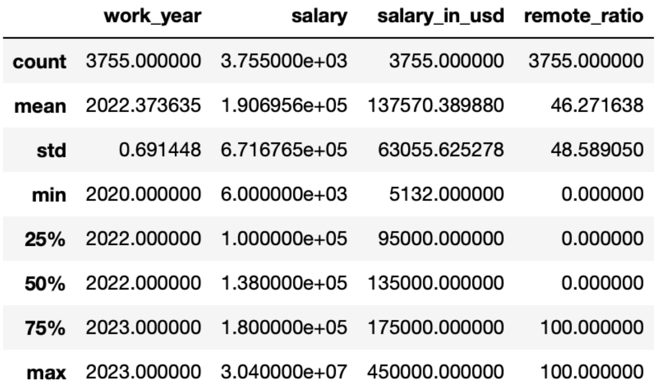
Summary Statistics

- The data frame consists of 3755 entries and each of the columns has 3755 non-null values which indicates the fact that there are no null values in the dataset. There are a total of 4 variables with data type ‘int64’, numerical variables, and a total of 7 with data type ‘object’, categorical variables. Since all variables have no null values, data imputation no removal of missing values needs to be done further.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3755 entries, 0 to 3754
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   work_year           3755 non-null   int64
 1   experience_level    3755 non-null   object
 2   employment_type     3755 non-null   object
 3   job_title           3755 non-null   object
 4   salary              3755 non-null   int64
 5   salary_currency     3755 non-null   object
 6   salary_in_usd       3755 non-null   int64
 7   employee_residence  3755 non-null   object
 8   remote_ratio        3755 non-null   int64
 9   company_location    3755 non-null   object
 10  company_size        3755 non-null   object
dtypes: int64(4), object(7)
memory usage: 322.8+ KB
```

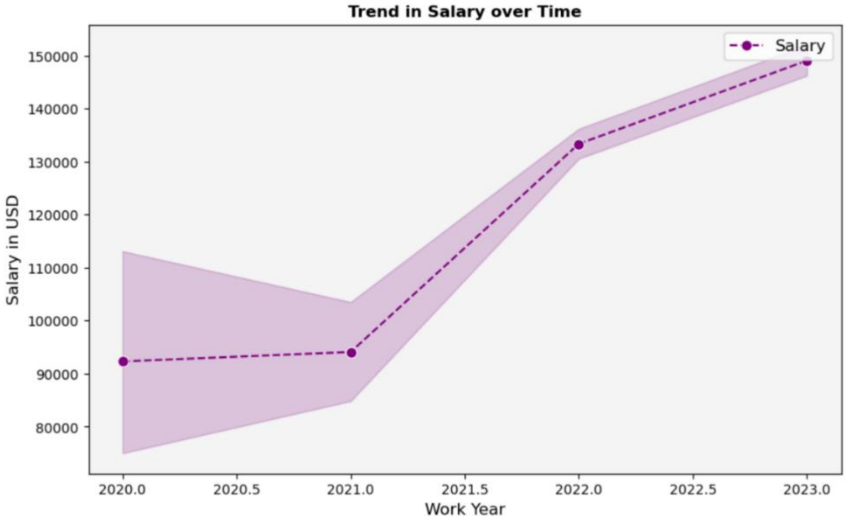
- Below is the descriptive summary statistics of the numerical variables which include the count, mean, standard deviation, minimum, maximum, and quartiles.



	work_year	salary	salary_in_usd	remote_ratio
count	3755.000000	3.755000e+03	3755.000000	3755.000000
mean	2022.373635	1.906956e+05	137570.389880	46.271638
std	0.691448	6.716765e+05	63055.625278	48.589050
min	2020.000000	6.000000e+03	5132.000000	0.000000
25%	2022.000000	1.000000e+05	95000.000000	0.000000
50%	2022.000000	1.380000e+05	135000.000000	0.000000
75%	2023.000000	1.800000e+05	175000.000000	100.000000
max	2023.000000	3.040000e+07	450000.000000	100.000000

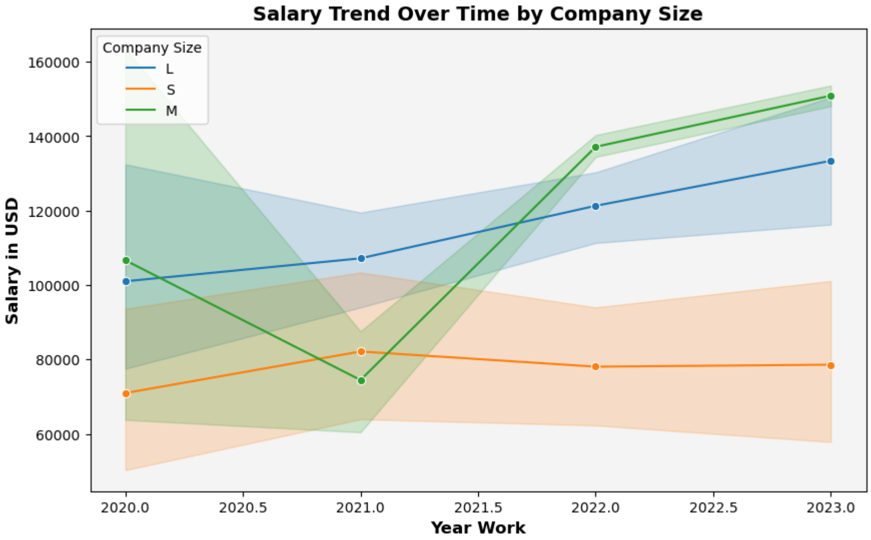
Exploratory Data Analysis

01. Salary trend over time



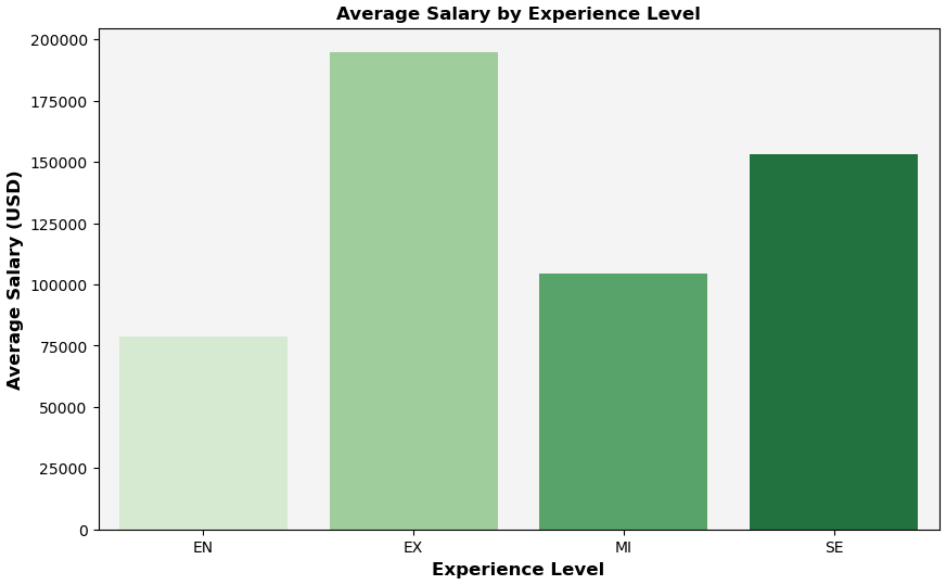
The above graph is a line plot that portrays the relationship between the two variables work_year and salary_in_usd. As shown in the above plot, from the year 2020 to 2021 there is very minimal fluctuation in the salary which depicts a stable income trend during this one year. However, from the year 2021 to 2022, there is a steep increase in salary. Furthermore, in the year 2022 to 2023, there is still a high change in salary which is however lower than the increase in salary that happens from the year 2021 to 2022.Overall this describes and represents the transition of salary in USD through the years from a time frame with a more stable salary to a time which has a major salary increase.

2. visualizing the salary scale by considering the company size



The above line graph was created using work year , company size and salary in usd dollars variables. The x axis represents the year of work while the y axis represents the salary range. The company size plotted using three different colors which can be refered using the legend in the left corner. By studying the line plot, from year 2020-2021, small size comapanies' and large size companies have slightly increamented meanwhile medium scale companies' salary has decremented at the same time period. However, from 2021 onwords medium companies salaries have incremented in significantly. Considering about small and large companies had slight increments and decrements throughout years compared to medium companies.

3. Visualizing the average salary based on the experience level

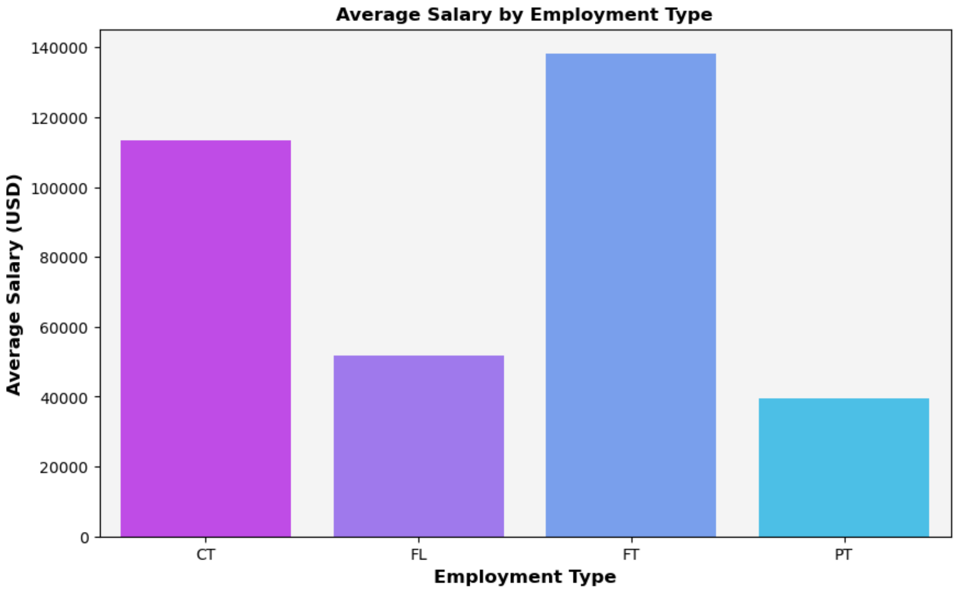


For the above bar plot the element declaration as follows,

- EN- entry level
- EX- experience level
- ML- mid level
- SE- senior level

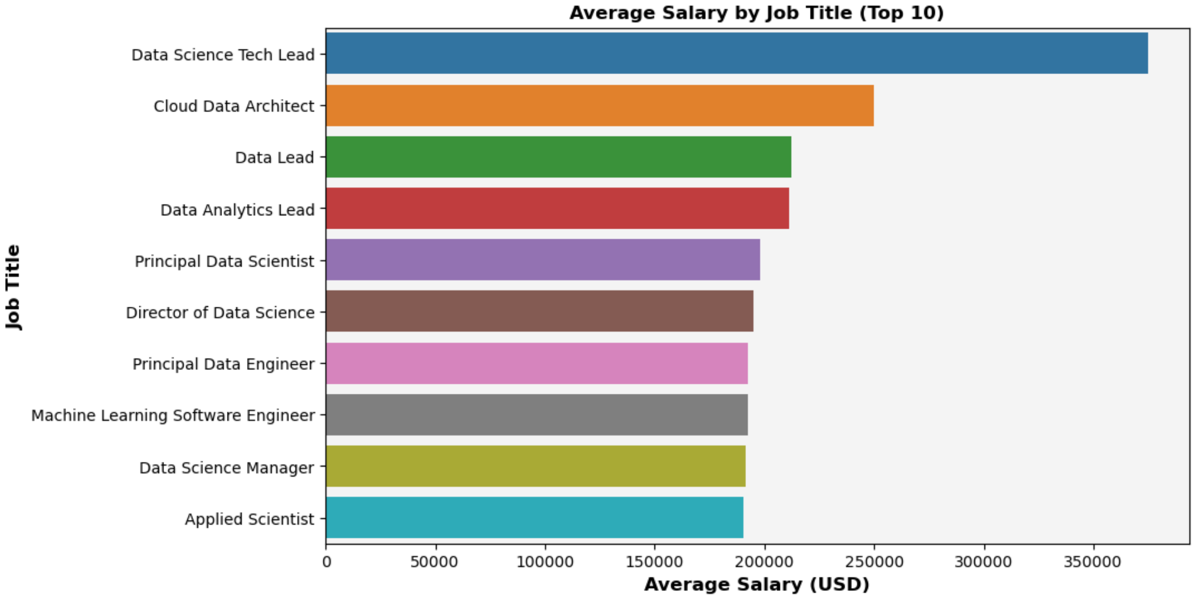
According with the bar plot, the highest average salary value scored for the experienced level employees and lowest value consists of the entry level employees. Senior level employees comprised comparative amount of average salary.

3. Visualizing how employee type engaging with the salary variable



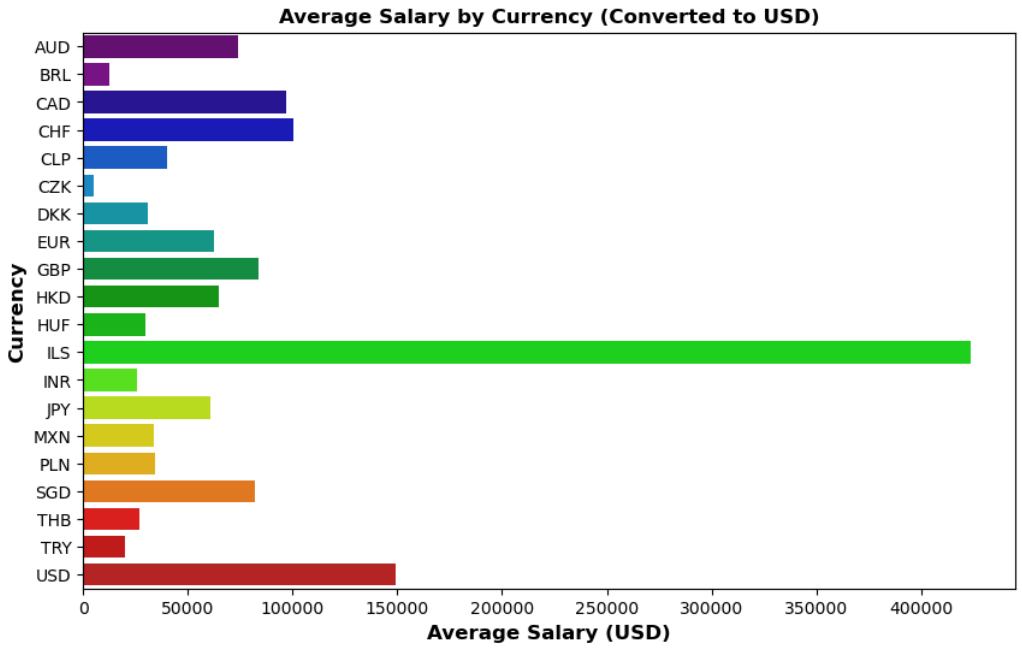
The above bar plot’s x axis represents the employee type (CT- contractor, FL- freelancer, FT- full time, PT- part time) and the y axis represents the average salary in USD dollars. The highest salary includes for the full-time employees and freelancers, part time employees consisting of low range of salary scale. Contractors is capable of competing with the full time employees.

4. visualizing how job title affects the salary scale



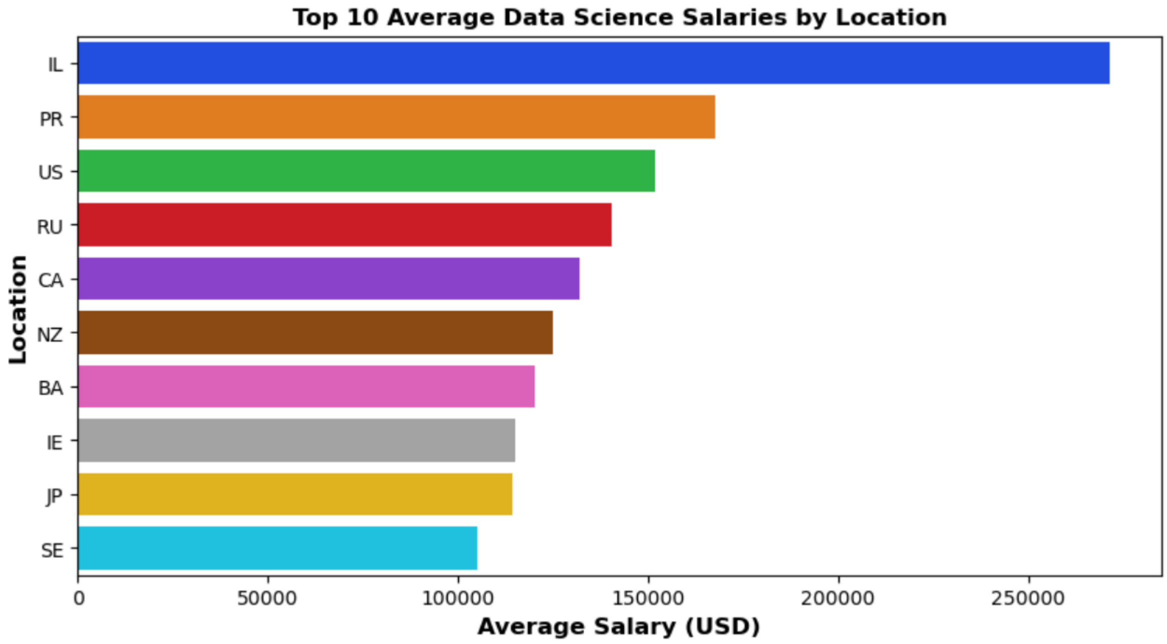
Align with the horizontal bar plot, the highest salary contains for the data science tech lead and second highest salary range goes for the cloud data architect job title. Data lead and data analytics titles are next leading job roles based on salary scale. Other job titles consist similar range of salary scale.

5. visualizing the different type of currency impacting the salary scale



The bar plot visualized using all currency types converted to USD dollars for the visualize perspectives. Based on the visualized data, Israeli Shekel is the highest salary value for the currency type. The second highest owns for the USD dollars. Other currency types include overall average salary values.

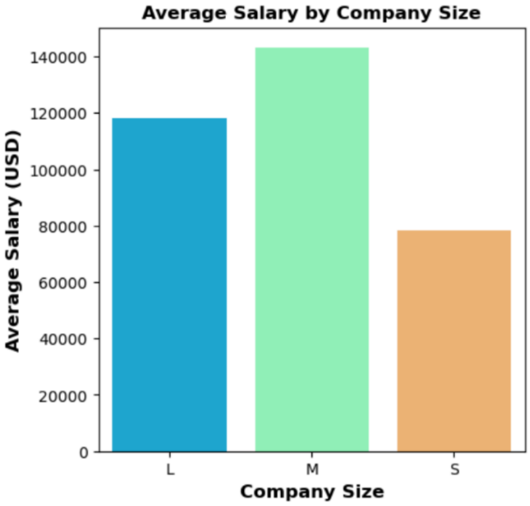
6. average data science salaries by location



Average salary and location variables are used to create the above bar graph. The x-axis represents the average salary of a data science professional in USD while the y-axis depicts the location. As shown in the above graph the average data science salary is notably high in

Illinois (IL). Puerto Rico (PR) and the United States (US) also have competitive average salaries. Additionally, Russia (RU), Canada (CA), New Zealand (NZ), Bosnia (BA), Ireland (IE), Japan (JP), and Sweden (SE) also offer average salaries above 100,000 USD for data science-related jobs.

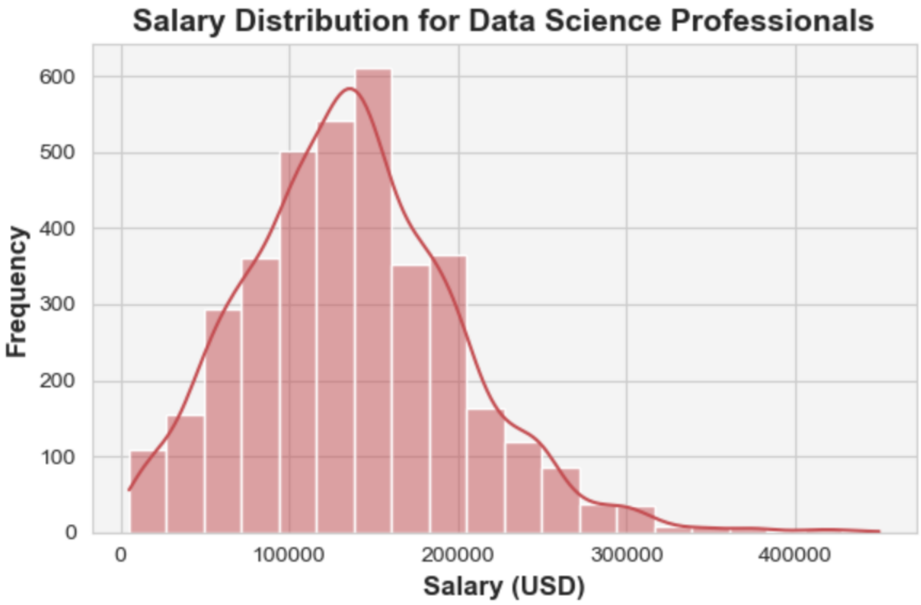
7. Average salary by company size



The above bar graph depicts the variation in average salary with the company size. The element declaration is as,

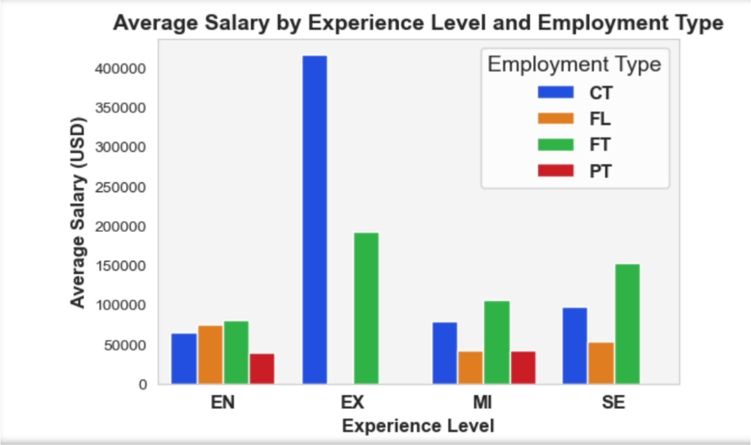
L- Large companies
M-Medium sized companies
S-small companies
As per the above visual medium-sized companies offer the highest average salary while small companies offer a lower average salary compared to large companies.

8. Salary distribution for Data Science Professionals



The above histogram represents the salary distribution for data science professionals. The distribution is negatively skewed, which explains that most professionals earn lower to mid-range salaries. There is a noticeable peak in the distribution suggesting a concentration of professionals within the 100,000 – 200,000 salary range.

9. Average salary by Experience level and Employment type



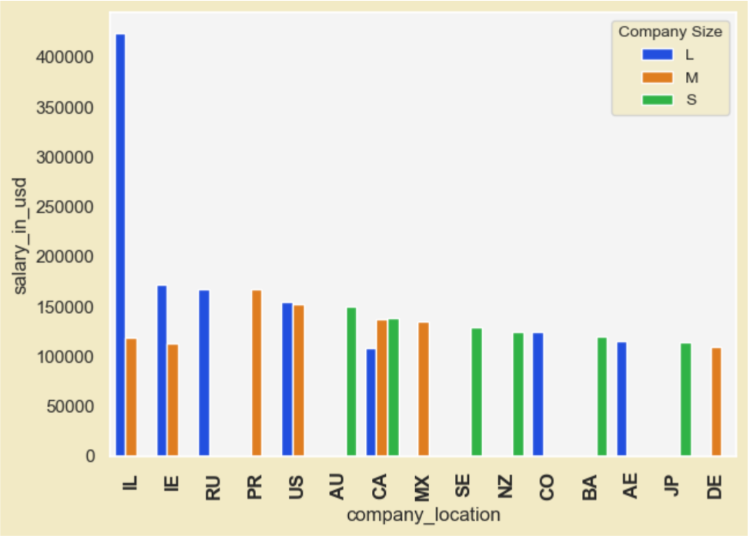
The above bar chart compares the average salary based on the experience level for different employment types. The element declaration is as follows:

CT – Contractor, FL- Freelancer, FT-Full-time, PT- Part-time

EN-Entry-Level, EX-Experienced, MI – Mid-Level, SE – Senior

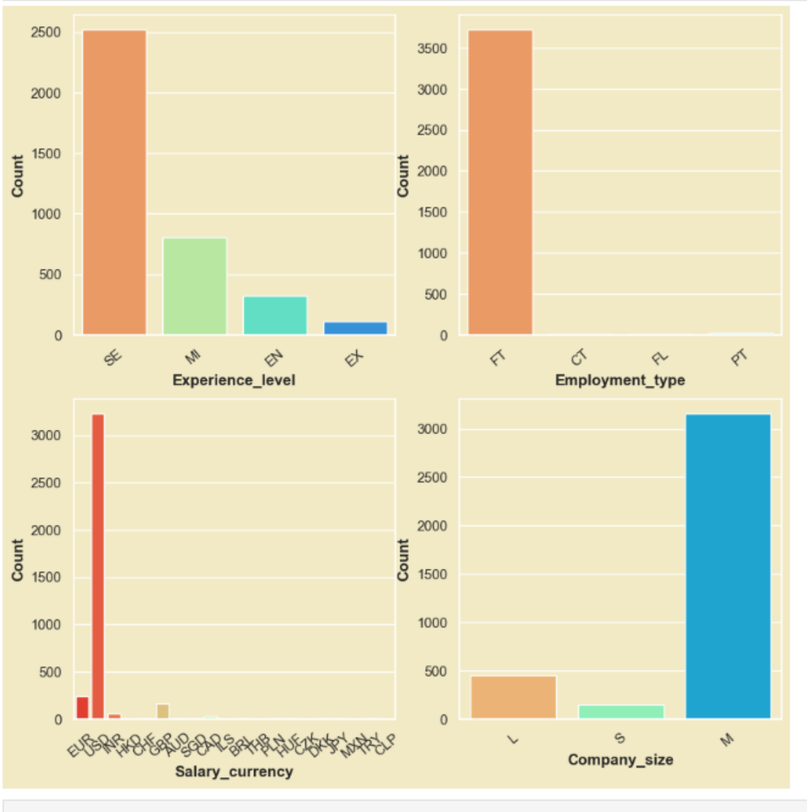
As per the above visualization experienced level contractors will have the highest average salary.

10. Average Salary by Company Location and Company Size



The above bar graph depicts a comparison of the average salary based on company size and location. According to the graph in Illinois (IL), large companies tend to offer comparatively an average salary.

11. Count plots for experience level, company size and salary scale



The above plots represent the count of each category in the respective variable. Element declaration for the variables are as follows,

Experience level- SE (Senior), MI (Mid-level), EN (Entry-Level), EX (Experienced)

Employment type- FT(Full-Time), CT (Part-Time), FL(Freelancer), PT(Part-Time)

Company size- L (Large Companies), S (Small companies), M (Medium companies)

Based on the above plots, the most common experienced level is “Senior” while the most common employment type is “Full-Time” Employment. Further, most salaries are offered in USD and the most prevalent company size is “Medium” scaled companies.

Conducting Statistical Techniques

Simple Linear regression analysis

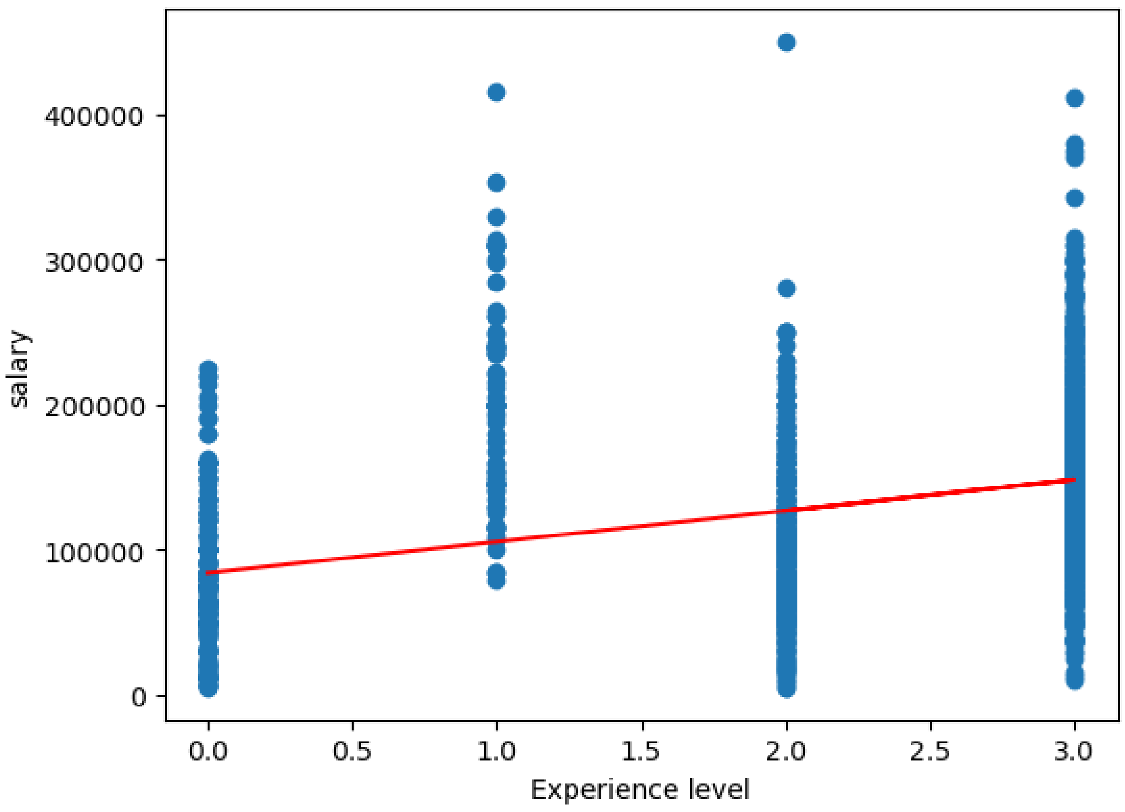
Simple linear regression on experience level vs salary

Referring to the data resource that was selected to do an analysis, the dataset includes categorical variables. In order to proceed a simple linear regression on experience level and salary. The first step was encoding the experience level variable, in this step, the label encoding process was applied as four different experience levels were aligned for 0-3 range numbers. Once the encoded process was done, to identify the correlation between variables the scatter plot was plotted. As for the simple linear regression, as the independence variable, the Experience level variable is considered and the salary in USD variable is considered as the dependent variable. Once the scatter plot was plotted, the linear regression line was plotted as below. Based on the plot data provided, the coefficient value returned as 21416.44 which means the mentioned value of average change has the ability to affect the salary variable by the level of experience that an employee obtains in the field of data science. the regression line plotted between 100,000 and 200,000 expounds the salary range can be affected by experience level in the mentioned range.

- c = 83901.70819792873
  array([21416.44663946])

  Regression Equation:

  $$\text{Salary\_in\_USD} = 21416.4 \times \text{Experience\_Level\_encoded} + 83901.7 + \in$$



Simple linear regression on employment type vs salary

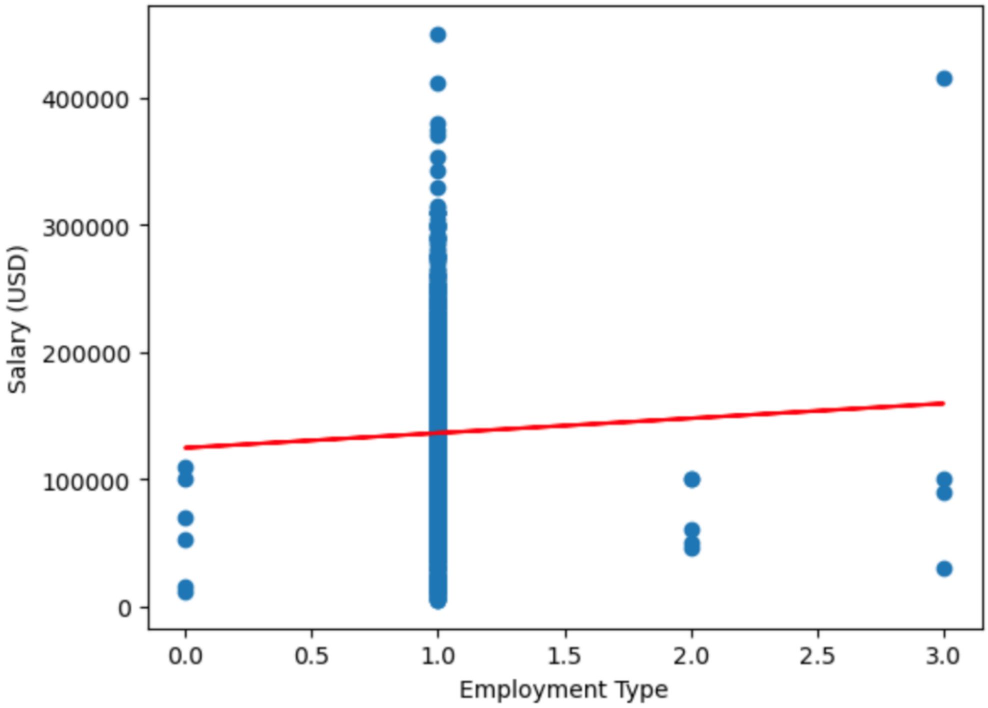
Data preprocessing-The employment_type variable has four categories which are PT (part-time), FT (full-time), CT (contract), and FL (freelance). These data was encoded and set in a new variable and assigned with numerical values as follows, CT = 3, FL = 2, FT = 1 ,PT = 0.

Simple Linear Regression Analysis-In the conducted linear regression using Python the considered dependent variable is salary_in_usd (Y) and the independent variable is the encoded employment type (X). Then the dataset was split into two, training and testing where 40% of the data was used for testing.The linear regression equation applied was Y=mX+c where the training parameters m is the coefficient and c is the intercept.

Using Python, the team found the training parameters to be as follows,

- c = 124638.74212760404
- m = array([11613.05309138])
  regression equeation :
  Salary_in_USD = 11613.0 × Encoded_Employment_Type + 124639.0 + ∈

The coefficient represents the average change in salary for a unit change in the employment type while other variables are constant.As visible in the regression plot generated, when plotting regression lines for categorical variables there will be disconnected lines of data points for each category.The linear regression model was generated for both the trained data as well as the test data.



Interpretation -the coefficient of this linear regression is positive which indicates and visually shows in the model that it is a positive linear regression. This implies that there is a positive linear relationship between the employment type and the salary.

- As the employment type goes higher the salary also increases. The increase in employment means, the encoded values for this variable increase from PT to FT to FL and then CT the corresponding salary in USD also according to the linear regression model.

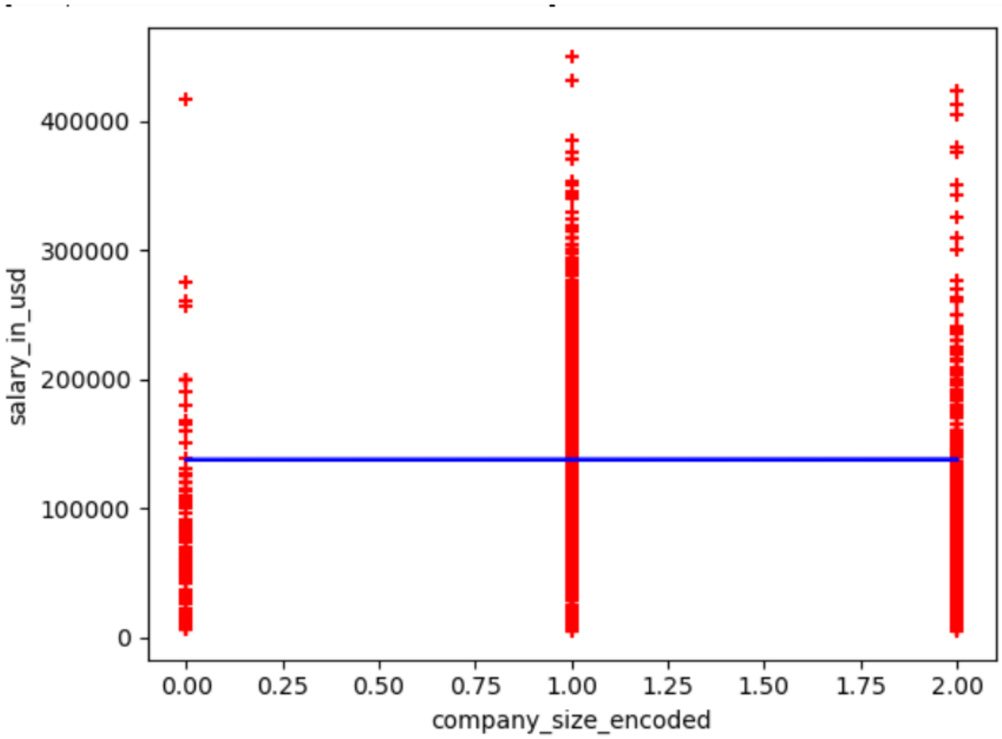
Conclusion

- The linear regression analysis reveals the positive linear relationship between the categorical variable employment type which was encoded to be compatible with linear regression and the variable salary in the data science field.
- The analysis reveals and suggests that advancement in the employment status increases the salary while other factors remain the same.

Simple linear regression on company size vs salary

When conducting a regression analysis to understand how company size impacts salary, since company size is a categorical variable it needed an appropriate encoding. In this case, label encoding was used, representing large-scale companies as 0.00, medium-scaled as 1.00, and small-scaled as 2.00. Upon plotting a scatter plot of company size against salary in USD, the visual representation showcased three horizontal lines, each corresponding to one of the encoded company size categories. This suggests that within each category, there was no apparent linear relationship between company size and salary.

Further, when the regression line was generated, it too appeared as a horizontal straight line. This outcome implies that there was no significant linear association between company size and salary.

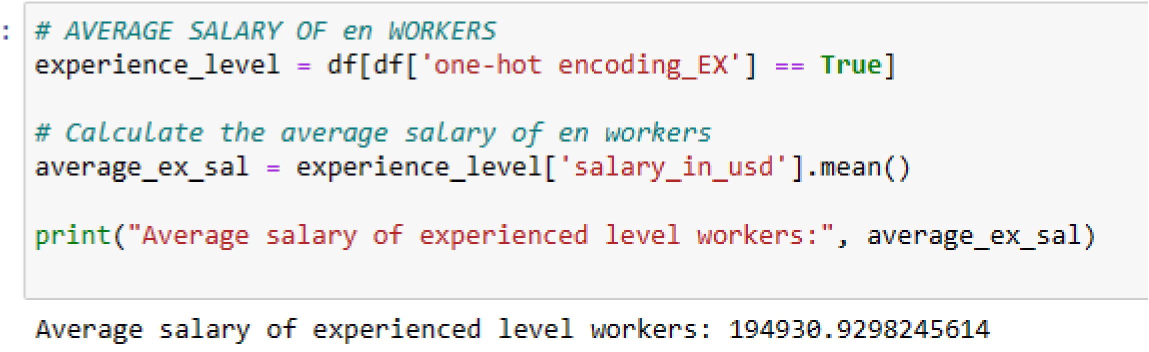


Hypothesis Testing

Hypothesis testing on experience level variable.

01. Considering the people who have experienced level qualification on working on the data science field for the salary range

Calculating the average salary:



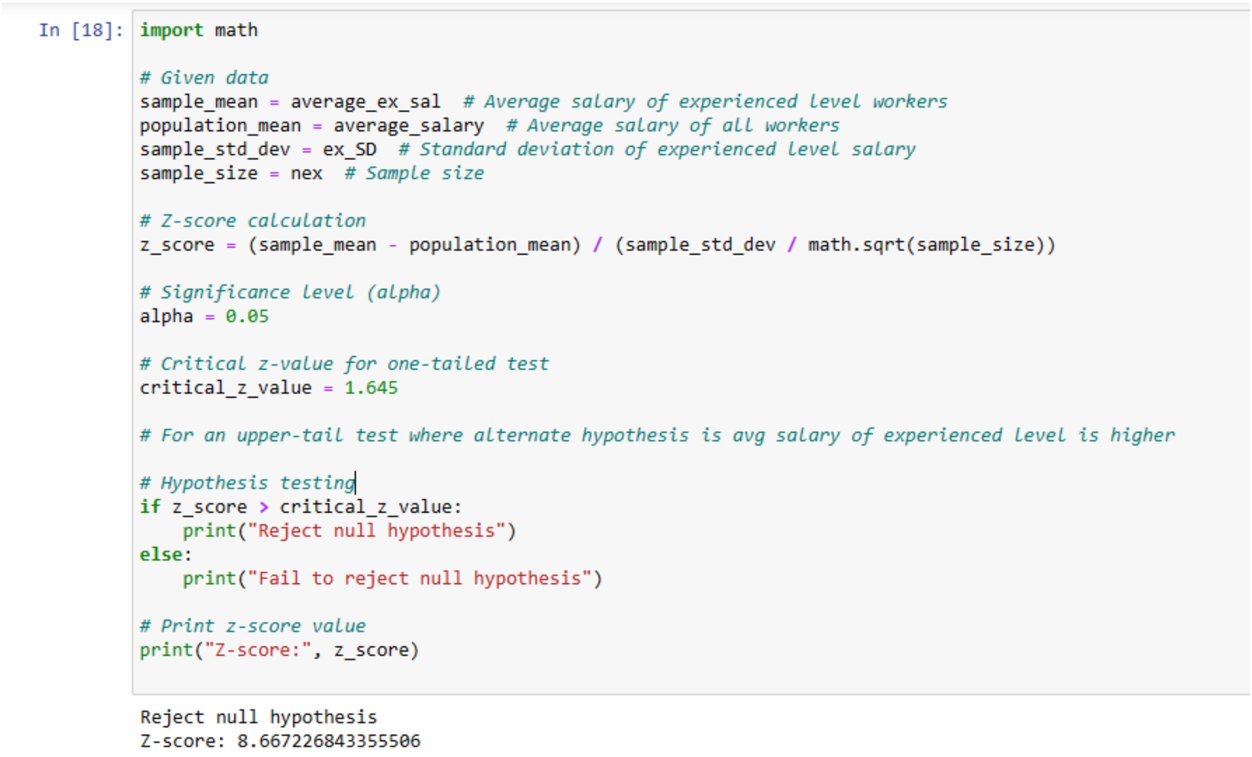
```
# AVERAGE SALARY OF en WORKERS
experience_level = df[df['one-hot encoding_EX'] == True]

# Calculate the average salary of en workers
average_ex_sal = experience_level['salary_in_usd'].mean()

print("Average salary of experienced level workers:", average_ex_sal)
```

```
Average salary of experienced level workers: 194930.9298245614
```

Conducting the hypothesis testing:



```
In [18]: import math

         # Given data
         sample_mean = average_ex_sal  # Average salary of experienced level workers
         population_mean = average_salary  # Average salary of all workers
         sample_std_dev = ex_SD  # Standard deviation of experienced level salary
         sample_size = nex  # Sample size

         # Z-score calculation
         z_score = (sample_mean - population_mean) / (sample_std_dev / math.sqrt(sample_size))

         # Significance level (alpha)
         alpha = 0.05

         # Critical z-value for one-tailed test
         critical_z_value = 1.645

         # For an upper-tail test where alternate hypothesis is avg salary of experienced level is higher

         # Hypothesis testing
         if z_score > critical_z_value:
             print("Reject null hypothesis")
         else:
             print("Fail to reject null hypothesis")

         # Print z-score value
         print("Z-score:", z_score)
```

```
Reject null hypothesis
Z-score: 8.667226843355506
```

Hypothesis :

μ : average salary of an experienced level employee in data science field

H_0 : μ < = 137570 vs H_1 : μ > 137570

Value of test statistic :

$Z_{cal} = 8.66$

Critical value:

Critical value = 1.645

Decision Rule:

Reject H_0 if $Z_{cal} > Z\alpha$

Decision :

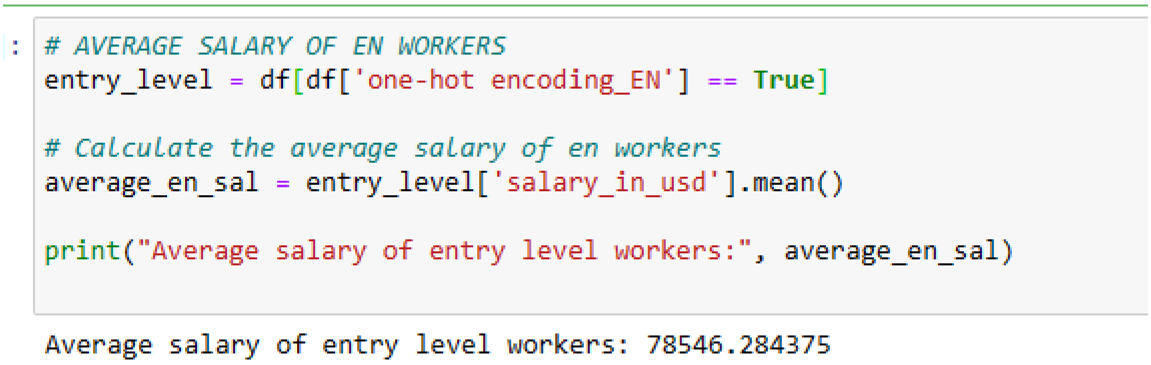
Reject null hypothesis

Interpretation:

According to the results of the testing, the null hypothesis have rejected. The interpretation is the employees who are working in the data science field which have experience level qualifications receive an average salary greater than 137570 dollars.

02. considering the people who entry level of experience in working data science field for the salary range.

Calculating the average salary:



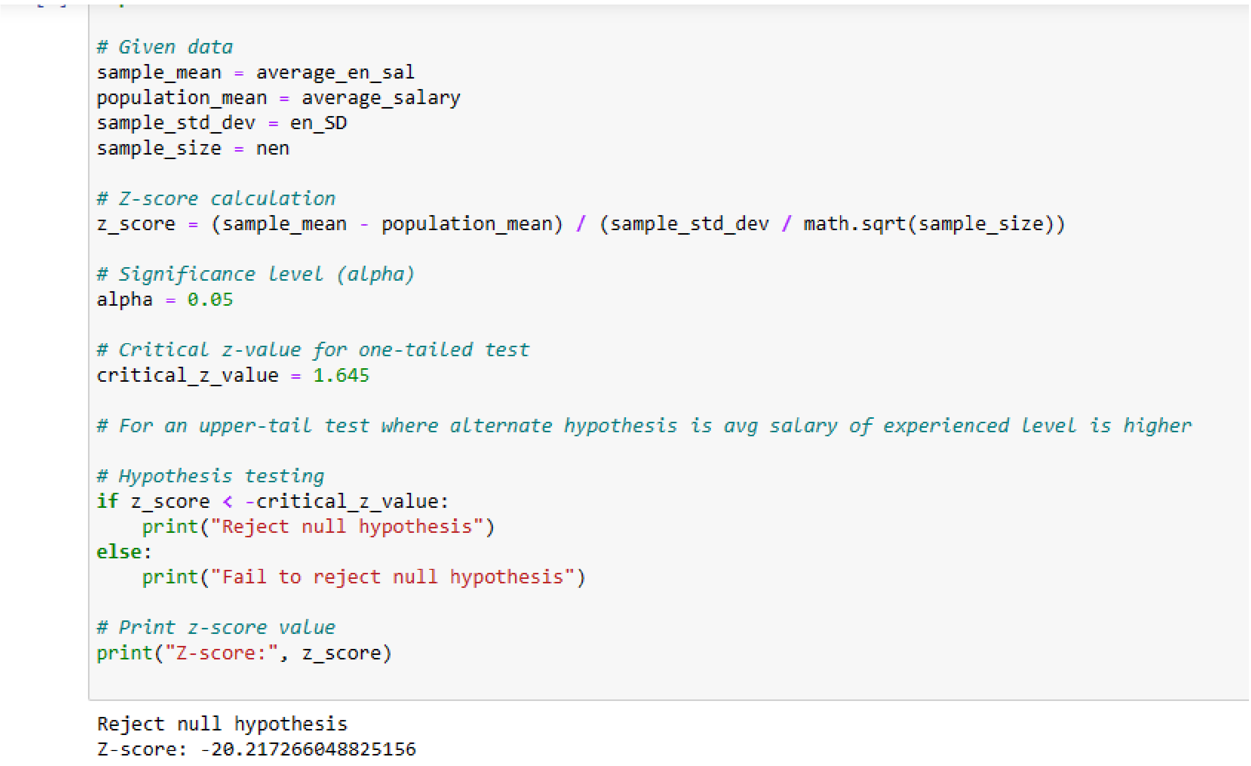
```
# AVERAGE SALARY OF EN WORKERS
entry_level = df[df['one-hot encoding_EN'] == True]

# Calculate the average salary of en workers
average_en_sal = entry_level['salary_in_usd'].mean()

print("Average salary of entry level workers:", average_en_sal)
```

```
Average salary of entry level workers: 78546.284375
```

Conducting the hypothesis testing:



```
# Given data
sample_mean = average_en_sal
population_mean = average_salary
sample_std_dev = en_SD
sample_size = nen

# Z-score calculation
z_score = (sample_mean - population_mean) / (sample_std_dev / math.sqrt(sample_size))

# Significance level (alpha)
alpha = 0.05

# Critical z-value for one-tailed test
critical_z_value = 1.645

# For an upper-tail test where alternate hypothesis is avg salary of experienced level is higher

# Hypothesis testing
if z_score < -critical_z_value:
    print("Reject null hypothesis")
else:
    print("Fail to reject null hypothesis")

# Print z-score value
print("Z-score:", z_score)
```

```
Reject null hypothesis
Z-score: -20.217266048825156
```

Hypothesis :

μ : average salary of an entry level employee in data science field

H_0 : $\mu >= 137570$ vs H_1 : $\mu < 137570$

Value of test statistic :

$Z_{cal} = -20.22$

Critical value:

Critical value = 1.645

Decision Rule:

Reject H_0 if $Z_{cal} < -Z\alpha$

Decision :

Reject null hypothesis

Interpretation:

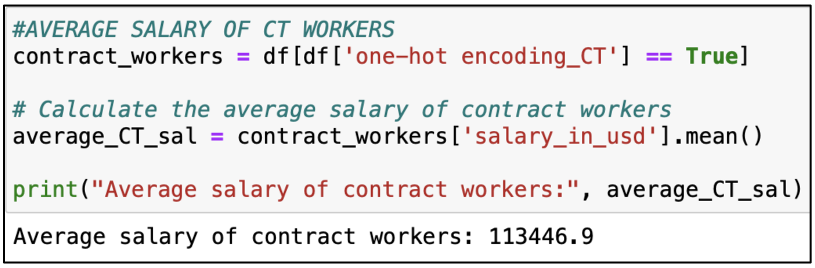
With the significance level of 0.05 , there’s sufficient evidence to conclude that, employees who have entry level of experience in the data science field receive an average salary less than 137570 dollars.

Hypothesis testing on employment type variable.

Upper tail testing

The following section contains the detailed descriptions of the statistical analysis conducted using a z-test to conduct the hypothesis testing about the effects that the variable employement_type has on the salary of data science jobs. This bariable is categorical and one hot encoding was performed therefore inorfer to encode the categorical variables into a numerical one in order to be compatible with the analysis.

1) The average salary of data science job employees is higher for contract-time workers than for other employment types.



```
#AVERAGE SALARY OF CT WORKERS
contract_workers = df[df['one-hot encoding_CT'] == True]

# Calculate the average salary of contract workers
average_CT_sal = contract_workers['salary_in_usd'].mean()

print("Average salary of contract workers:", average_CT_sal)
```

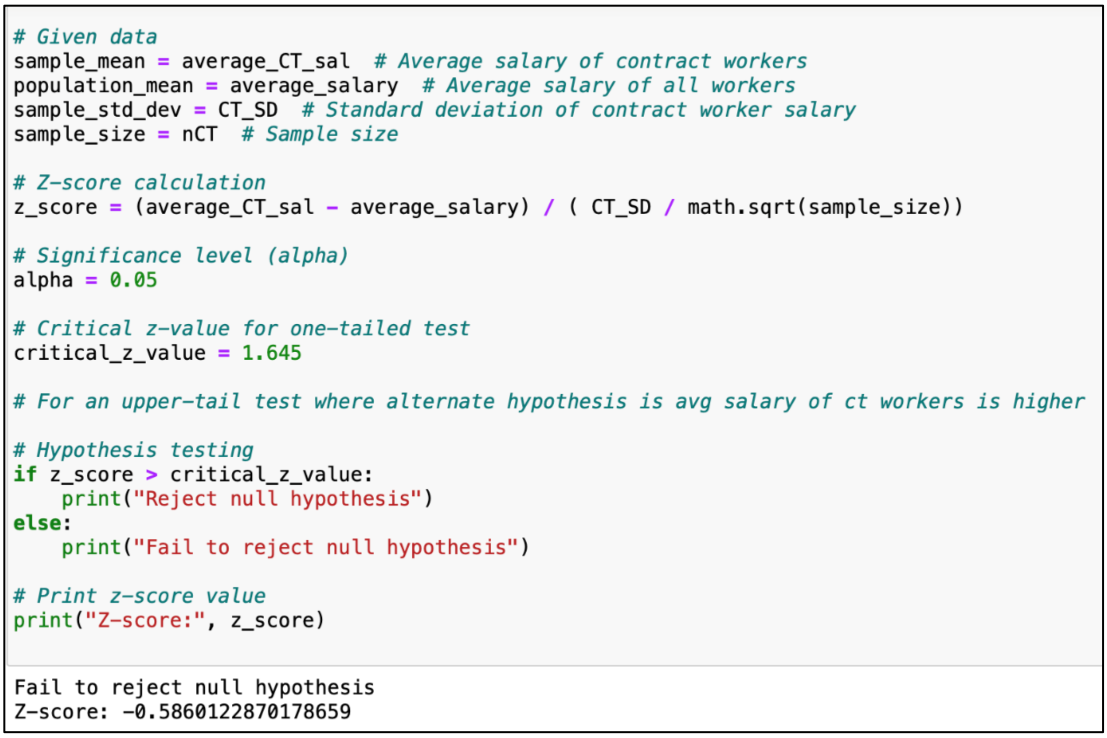
```
Average salary of contract workers: 113446.9
```

H_0 : $\mu \leq$ 113000 Vs H_1: $\mu > 113000$

This hypothesis testing aims to check if the average salary for contract-time employees (CT) is higher than the salary of other employment types.

Hypothesis testing was done initially to compare the salaries.

The results for this are below.



```
# Given data
sample_mean = average_CT_sal  # Average salary of contract workers
population_mean = average_salary  # Average salary of all workers
sample_std_dev = CT_SD  # Standard deviation of contract worker salary
sample_size = nCT  # Sample size

# Z-score calculation
z_score = (average_CT_sal - average_salary) / ( CT_SD / math.sqrt(sample_size))

# Significance level (alpha)
alpha = 0.05

# Critical z-value for one-tailed test
critical_z_value = 1.645

# For an upper-tail test where alternate hypothesis is avg salary of ct workers is higher

# Hypothesis testing
if z_score > critical_z_value:
    print("Reject null hypothesis")
else:
    print("Fail to reject null hypothesis")

# Print z-score value
print("Z-score:", z_score)
```

```
Fail to reject null hypothesis
Z-score: -0.5860122870178659
```

Hypothesis

$\mu = Average\ salary\ of\ CT\ workers$

$H_0 : \mu \leq 113000\ Vs$ $\qquad H_1: \quad \mu > 113000$

Value of test statistic:

Z_{cal} = -0.597

Critical value:

Critical value = 1.645

Decision rule:

Reject H_0 if $Z_{cal} > Z\alpha$

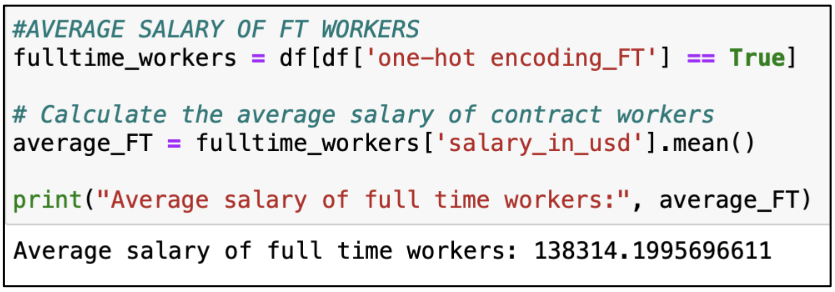
Decision:

Fail to reject null hypothesis

Interpretation :

The calculated z-score is -0.5968684646426328, and the critical z-value for an upper-tailed test with a significance level of 0.05 (alpha = 0.05) is 1.645. Since z_score < critical_z_value we fail to reject the null hypothesis. There is sufficient evidence to prove that the average salary for contract workers is lower than or equal to the average salary of all workers.

2) The average salary of data science job employees is higher among Full time employees (FT).

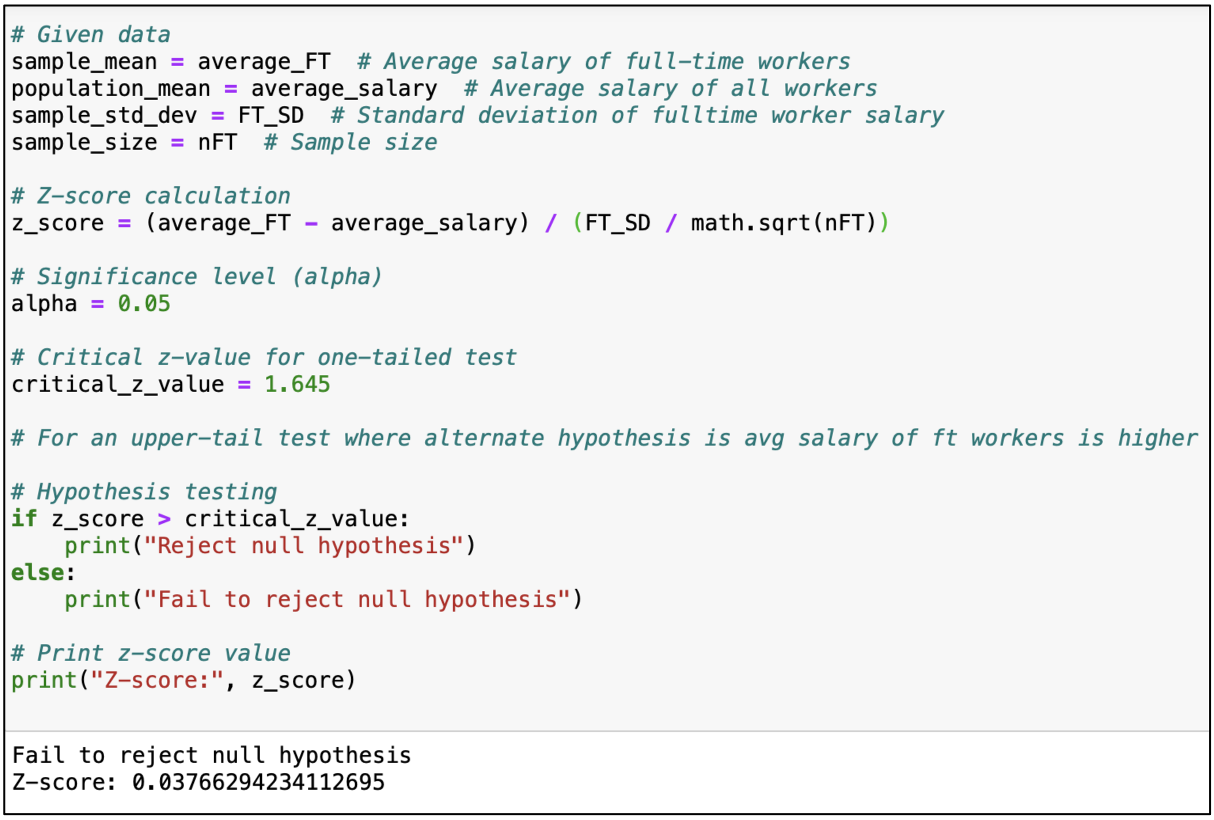


```
#AVERAGE SALARY OF FT WORKERS
fulltime_workers = df[df['one-hot encoding_FT'] == True]

# Calculate the average salary of contract workers
average_FT = fulltime_workers['salary_in_usd'].mean()

print("Average salary of full time workers:", average_FT)
```

```
Average salary of full time workers: 138314.1995696611
```



```
# Given data
sample_mean = average_FT  # Average salary of full-time workers
population_mean = average_salary  # Average salary of all workers
sample_std_dev = FT_SD  # Standard deviation of fulltime worker salary
sample_size = nFT  # Sample size

# Z-score calculation
z_score = (average_FT - average_salary) / (FT_SD / math.sqrt(nFT))

# Significance level (alpha)
alpha = 0.05

# Critical z-value for one-tailed test
critical_z_value = 1.645

# For an upper-tail test where alternate hypothesis is avg salary of ft workers is higher

# Hypothesis testing
if z_score > critical_z_value:
    print("Reject null hypothesis")
else:
    print("Fail to reject null hypothesis")

# Print z-score value
print("Z-score:", z_score)
```

```
Fail to reject null hypothesis
Z-score: 0.03766294234112695
```

Hypothesis:

$\mu = Average\ salary\ of\ FT\ workers$

$H_0 : \mu \leq 138000\ Vs$ $H_1 : \quad \mu > 138000$

Value of test statistic

$Z_{cal} = 0.03766$

Critical value:

Critical value = 1.645

Decision rule:

Reject H_0 if $Z_{cal} > Z\alpha$

Decision:

Fail to reject null hypothesis

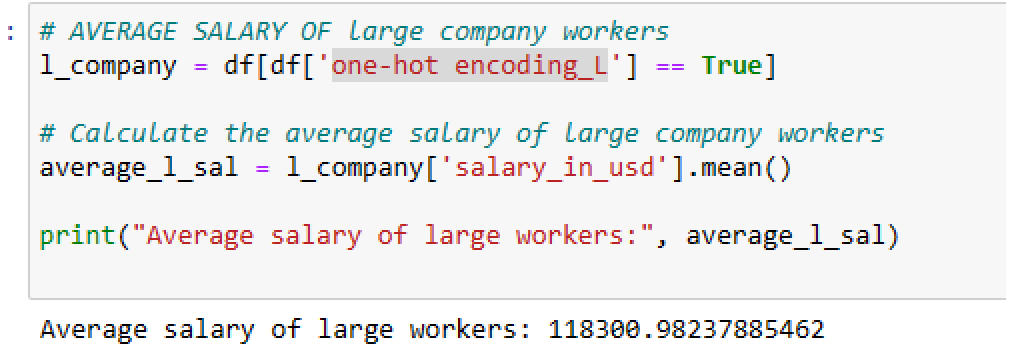
Interpretation :

The calculated z-score is 0.03766294234112695, and the critical z-value for an upper-tailed test with a significance level of 0.05 (alpha = 0.05) is 1.645. Since z_score < critical_z_value we fail to reject the null hypothesis. There is sufficient evidence to prove that the average salary for full-time workers is lower than or equal to the average salary of all workers.

Hypothesis testing on company size variable.

01. Considering the people who are working on large scale companies in the data science field for the salary range.

Calculating the average salary:



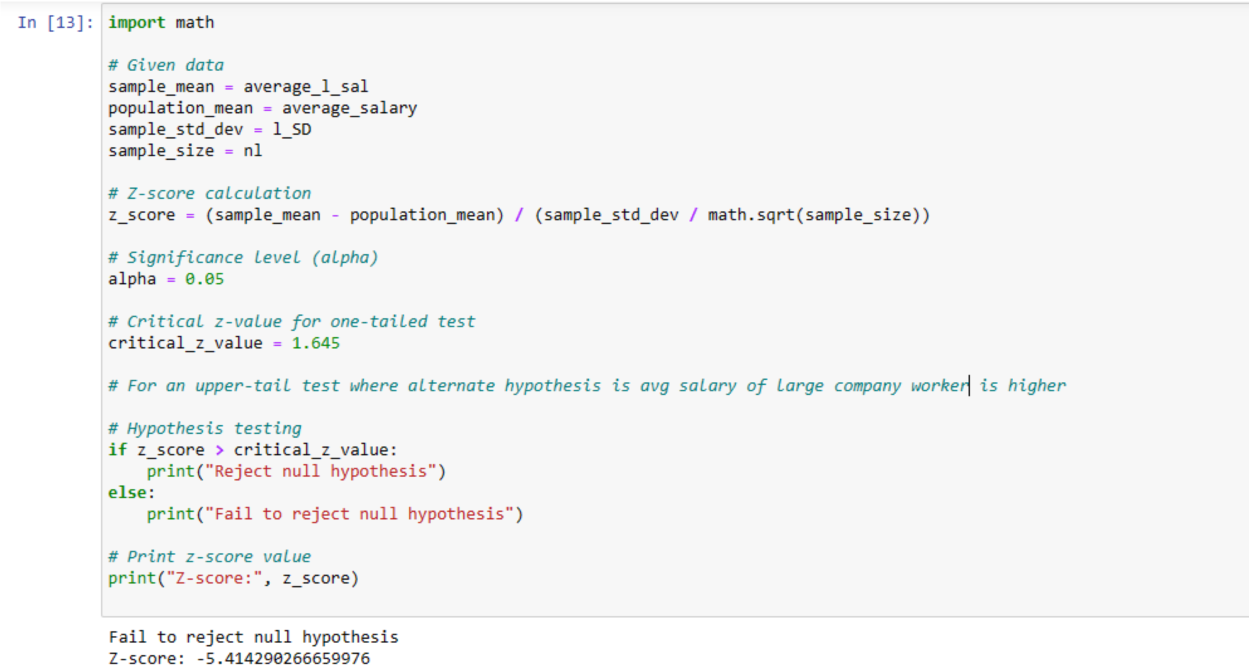
```
: # AVERAGE SALARY OF large company workers
l_company = df[df['one-hot encoding_L'] == True]

# Calculate the average salary of large company workers
average_l_sal = l_company['salary_in_usd'].mean()

print("Average salary of large workers:", average_l_sal)
```

```
Average salary of large workers: 118300.98237885462
```

Conducting hypothesis testing:



```
In [13]: import math

         # Given data
         sample_mean = average_l_sal
         population_mean = average_salary
         sample_std_dev = l_SD
         sample_size = nl

         # Z-score calculation
         z_score = (sample_mean - population_mean) / (sample_std_dev / math.sqrt(sample_size))

         # Significance level (alpha)
         alpha = 0.05

         # Critical z-value for one-tailed test
         critical_z_value = 1.645

         # For an upper-tail test where alternate hypothesis is avg salary of large company worker is higher

         # Hypothesis testing
         if z_score > critical_z_value:
             print("Reject null hypothesis")
         else:
             print("Fail to reject null hypothesis")

         # Print z-score value
         print("Z-score:", z_score)

         Fail to reject null hypothesis
         Z-score: -5.414290266659976
```

Hypothesis :

μ : average salary of a large company size employee in data science field

H_0 : $\mu <= 137570$ vs H_1 : $\mu > 137570$

Value of test statistic :

$Z_{cal} = -5.414$

Critical value:

Critical value = 1.645

Decision Rule:

Reject H_0 if $Z_{cal} > Z\alpha$

Decision :

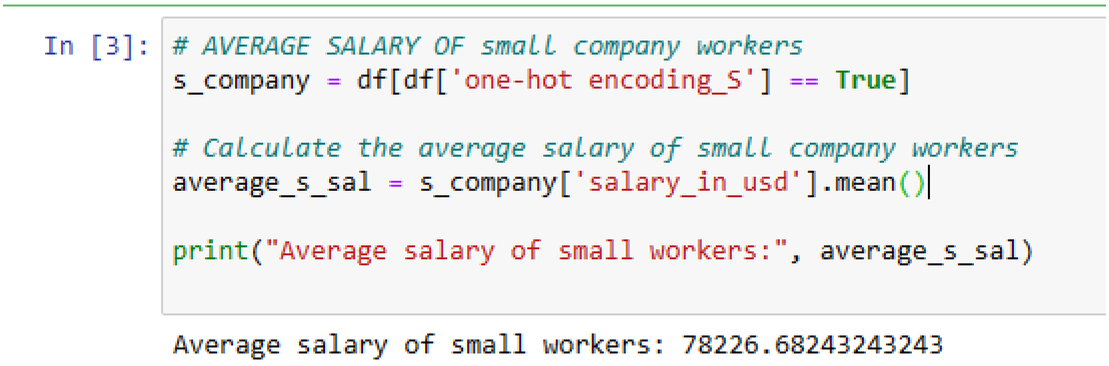
Fail to reject null hypothesis

Interpretation:

At 0.05 alpha level of significance, there is no evidence to reject the null hypothesis. Therefore, there is evidence to suggest that the average salary for data science professionals in large-scale companies is significantly smaller or equal to the average salary for employees in the field of data science.

02. Considering the people who are working on small scale companies in the data science field for the salary range.

Calculating the average salary:



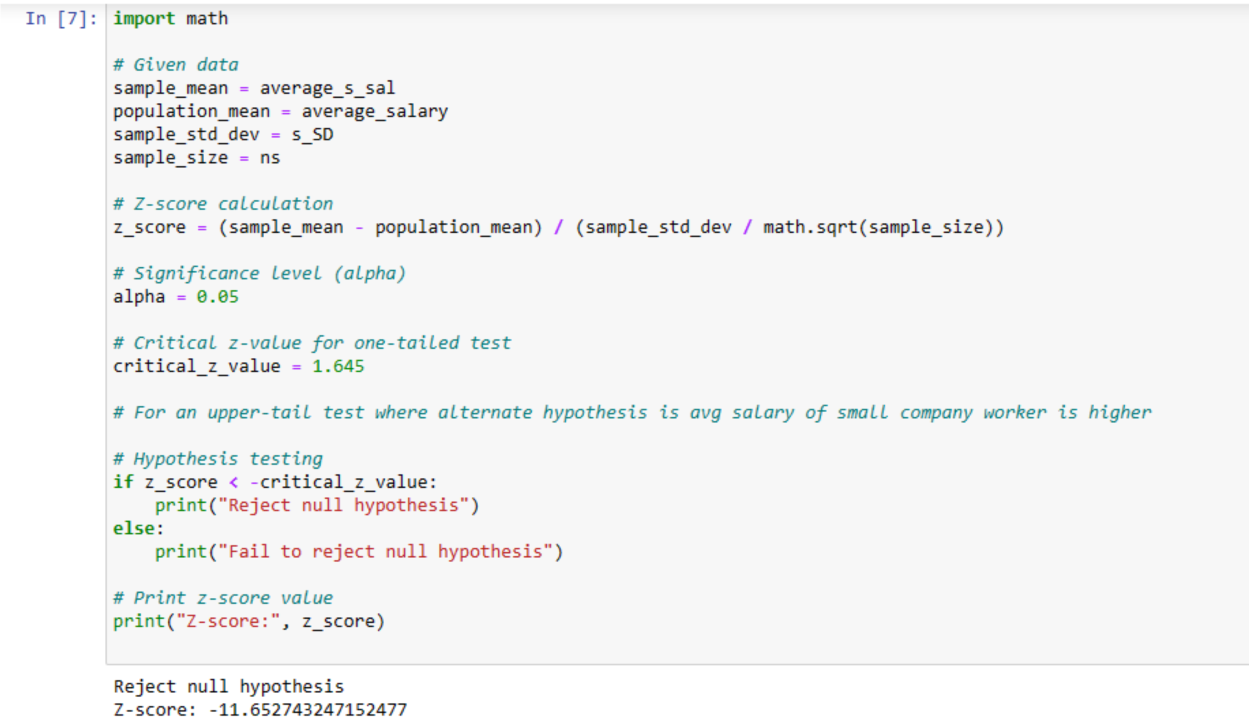
```
In [3]: # AVERAGE SALARY OF small company workers
        s_company = df[df['one-hot encoding_S'] == True]

        # Calculate the average salary of small company workers
        average_s_sal = s_company['salary_in_usd'].mean()

        print("Average salary of small workers:", average_s_sal)

        Average salary of small workers: 78226.68243243243
```

Conducting hypothesis testing:



```
In [7]: import math

        # Given data
        sample_mean = average_s_sal
        population_mean = average_salary
        sample_std_dev = s_SD
        sample_size = ns

        # Z-score calculation
        z_score = (sample_mean - population_mean) / (sample_std_dev / math.sqrt(sample_size))

        # Significance level (alpha)
        alpha = 0.05

        # Critical z-value for one-tailed test
        critical_z_value = 1.645

        # For an upper-tail test where alternate hypothesis is avg salary of small company worker is higher

        # Hypothesis testing
        if z_score < -critical_z_value:
            print("Reject null hypothesis")
        else:
            print("Fail to reject null hypothesis")

        # Print z-score value
        print("Z-score:", z_score)

        Reject null hypothesis
        Z-score: -11.652743247152477
```

Hypothesis :

µ : average salary of a small company size employee in data science field

H_0 : µ > = 137570 vs H_1 : µ < 137570

Value of test statistic :

Z_{cal} = -11.65

Critical value:

Critical value = 1.645

Decision Rule:

Reject H_0 if $Z_{cal} < - Z\alpha$

Decision :

Reject null hypothesis

Interpretation:

Based on the hypothesis results, there's sufficient evidence to conclude that the employees work in small size companies receive less that 137570 dollars average salary in the field of data science.

Analysis Results

Hypothesis Testing

Founded on the results of hypothesis testing conducted on three separate variables related to salary variable which known as employment type, company size and experience level which concerned as major factors that could effect to the salary scale in the field of data science. as the tests results came out overall manner, elements in the factors are involving to impact on the salary range in the field of data science. Expound through as example, medium scale companies obtain greater salary than or equal to other range of companies. People who have senior level or experienced qualification earn salary greater than people who have other different type of experience level of working as an employee. Contract employees have a higher average salary range compared to other types of employees. Based on analyzed hypothesis results, the findings can be considered as the mentioned factors are involving measuring the salaries of employees who have different types of job titles in the field of data science.

Linear Regression

The linear regression analysis conducted in this project aimed to understand the relationship between certain variables and the salary of jobs in data science field. The analysis was done using three key variables that are experience level, employement type and lastly company size. This was done by encoding all the three categorical variables using label encoding to make them compatible with the analysis process.

Briefly experience level and employement type linear regressions demonstrated positive correlations with the target variable salary while company size didn't depict a significant correlation with the salary. The company size was revealed to not have a big connection or affect to the salary of the jobs.

The final analysis of the results shows that employement type and experience level have a big influence in the salary in a positive way whilst the company size didn't have a significant effect or change. These insight empahasize and clearly potrary the factors affecting the salary in the industry of data science.

Conclusion

In conclusion, the project on a statistical exploration of the data science job market revealed insights into factors influencing salary scale within the field. Through exploratory data analysis and statistical techniques such as simple linear regression and hypothesis testing the team identified key attributes such as employment type, company size, and experience level as significant factors impacting salary ranges. The final results of the analysis highlight the importance of considering various factors when assessing salary scales within the data science industry, providing valuable guidance for both employers and job seekers in this rapidly evolving field.

Project Programming Files attachments

https://nsbm365-my.sharepoint.com/:f:/g/personal/ngdnethmini_students_nsbm_ac_lk/Eh1_Lky1MZVMupbCjxv0ohIBcEEIyNbRQldxg9ClASpyug?e=IK6QKH

Individual Contribution

Content	**N G D Nethmini**	**A S A Gunathilaka**	**S A D H M Samarathunga**
Project Overview	• Problem background • Problem Statement • Problem Objective		

Methodology			• Data collection and preprocessing
Exploratory Data Analysis	• visualizing the salary scale by considering the company size • Visualizing the average salary based on the experience level • Visualizing how employee type engaging with the salary variable • visualizing how job title affects the salary scale • visualizing the different type of currency impacting the salary scale	• average data science salaries by location • Average salary by company size • Salary distribution for Data Science Professionals • Average salary by Experience level and Employment type • Average Salary by Company Location and Company Size • Count plots for experience level, company size and salary scale	• salary over trend
Conducting Statistical Techniques	• Simple linear regression on experience level vs salary • Hypothesis testing on experience level variable.	• Simple linear regression on company size vs salary • Hypothesis testing – salary comparison based on company size	• Simple linear regression on employment type vs salary • Hypothesis testing – effects of employment type on the salary
Analysis Results	• Hypothesis Testing		• Linear Regression
Conclusion & Abstract	• abstract	• conclusion	