

# DSC630-T301 Predictive Analytics (2243-1)\_week1\_Samanta\_Rajib

December 10, 2023

## **0.1 Class : DSC630-T301 Predictive Analytics (2243-1)**

## **0.2 Name : Rajib Samanta**

### **0.2.1 Assignment 1.2 : Week 1**

This assignment is a refresher of data analysis and visualization using Python and/or R. Find a data set that interests you and has appropriate data to create some interesting visualizations. A few good sources for finding datasets include Kaggle, UCI ML Repository, and the US Bureau of Labor Statistics.

With the dataset that you choose, perform the following steps using Python and/or R:

1. Write a summary of your data and identify at least two questions to explore visually with your data.
2. Create a histogram or bar graph from your data.
3. Create a boxplot from your data.
4. Create a bivariate plot from your data.
5. Create any additional visualizations that will help to answer the question(s) you want to answer.
6. Summarize your results and make a conclusion. Explain how you arrived at this conclusion and how your visualizations support your conclusion.

### **0.2.2 Write a summary of your data and identify at least two questions to explore visually with your data.**

### **0.2.3 Data Set :**

Salary of Data Scientists

<https://www.kaggle.com/datasets/piyushborhade/salary-of-data-scientists/>

### **0.2.4 About Dataset:**

This dataset aims to shed light on the salary trends in the field of Data Science for the years 2021 to 2023. With a focus on various aspects of employment, including work experience, job titles, and company locations, this dataset provides valuable insights into salary distributions within the industry.

### 0.2.5 Data Set Fields:

1. **Work\_year:** Representing the specific year of salary data collection.
2. **Experience\_level:** The level of work experience of the employees, categorized as EN (Entry-Level), EX (Experienced), MI (Mid-Level), SE (Senior).
3. **Employment\_type:** The type of employment, labelled as FT (Full-Time), CT (Contractor), FL (Freelancer), PT (Part-Time).
4. **Job\_title:** The job titles of the employees, such as “Applied Scientist”, “Data Quality Analyst”
5. **Salary:** The salary figures in their respective currency formats.
6. **Salary\_currency:** The currency code representing the salary.
7. **Salary\_in\_usd:** The converted salary figures in USD for uniform comparison.
8. **Company\_location:** The location of the companies, specified as country codes (e.g., “US” for the United States)
9. **Company\_size:** The size of the companies, classified as “L” (Large), “M” (Medium), and “S” (Small).

### 0.2.6 Questions to explore:

1. **Q01: How the Data science Salary Trends over Time ?** Understanding the distribution of Data science Salary is fundamental as it provides insights into the Job market and the range of salary for each category. This knowledge helps in:
  - a) Market Understanding: Identifying the typical Data science Job Market.
  - b) Job Salary: Determining competitive salary for data science with different sub category.
  - c) Career path Decisions: This will help the computer science engineer to choose the career and future learning prospective.
2. **Q02: What is the best Job Title Recommendation ?**

Recommend suitable job titles for candidates based on their experience level and desired salary range. This will help to understand how the Job and salary depend on company size and domain.

Answering these questions through data analysis and visualization provides critical insights for data science professionals and individuals involved in studying the computer science and programming language. It helps them understand market dynamics, make informed decisions, and devise effective strategies based on observed trends within the dataset.

```
[23]: # Load the Libraries
import os
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

```
[24]: # 1. Load the dataset as a Pandas data frame.
# 2. Display the first ten rows of data.
# Read in the Video Game Sales with Ratings data file ('VData Science Salary_
↳2021 to 2023.csv') from local:
directory = '/Users/rajibsamanta/Documents/Rajib/College/Sem 7 Winter 2023/
↳Week1'
# Set the working directory
os.chdir(directory)
print(os.getcwd())
dataset1_csv = pd.read_csv("ds_salaries.csv")
dataset1_csv.head(10)
# Display the DataFrame 10 rows
```

/Users/rajibsamanta/Documents/Rajib/College/Sem 7 Winter 2023/Week1

```
[24]: work_year experience_level employment_type job_title \
0      2023                SE          FT Principal Data Scientist
1      2023                MI          CT          ML Engineer
2      2023                MI          CT          ML Engineer
3      2023                SE          FT          Data Scientist
4      2023                SE          FT          Data Scientist
5      2023                SE          FT    Applied Scientist
6      2023                SE          FT    Applied Scientist
7      2023                SE          FT          Data Scientist
8      2023                SE          FT          Data Scientist
9      2023                SE          FT          Data Scientist

      salary salary_currency salary_in_usd employee_residence remote_ratio \
0    80000          EUR          85847          ES          100
1    30000          USD          30000          US          100
2    25500          USD          25500          US          100
3   175000          USD         175000          CA          100
4   120000          USD         120000          CA          100
5   222200          USD         222200          US           0
6   136000          USD         136000          US           0
7   219000          USD         219000          CA           0
8   141000          USD         141000          CA           0
9   147100          USD         147100          US           0

      company_location company_size
0                ES          L
1                US          S
2                US          S
3                CA          M
4                CA          M
5                US          L
6                US          L
```

7	CA	M
8	CA	M
9	US	M

```
[25]: # describe the dataframe'
dataset1_csv.shape
## It has 3755 records with 11 columns
```

```
[25]: (3755, 11)
```

```
[26]: # describe the dataframe'
dataset1_csv.describe()
```

```
[26]:
```

	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

```
[27]: # missing values
dataset1_csv.isnull().sum()
#-- No null column
```

```
[27]: work_year          0
experience_level      0
employment_type      0
job_title            0
salary              0
salary_currency      0
salary_in_usd        0
employee_residence   0
remote_ratio         0
company_location     0
company_size         0
dtype: int64
```

```
[28]: # Calculate frequency of each job title
job_title_counts = dataset1_csv['job_title'].value_counts()
job_title_counts
```

```
[28]: job_title
Data Engineer          1040
Data Scientist          840
Data Analyst           612
```

Machine Learning Engineer	289
Analytics Engineer	103
...	
Principal Machine Learning Engineer	1
Azure Data Engineer	1
Manager Data Management	1
Marketing Data Engineer	1
Finance Data Analyst	1

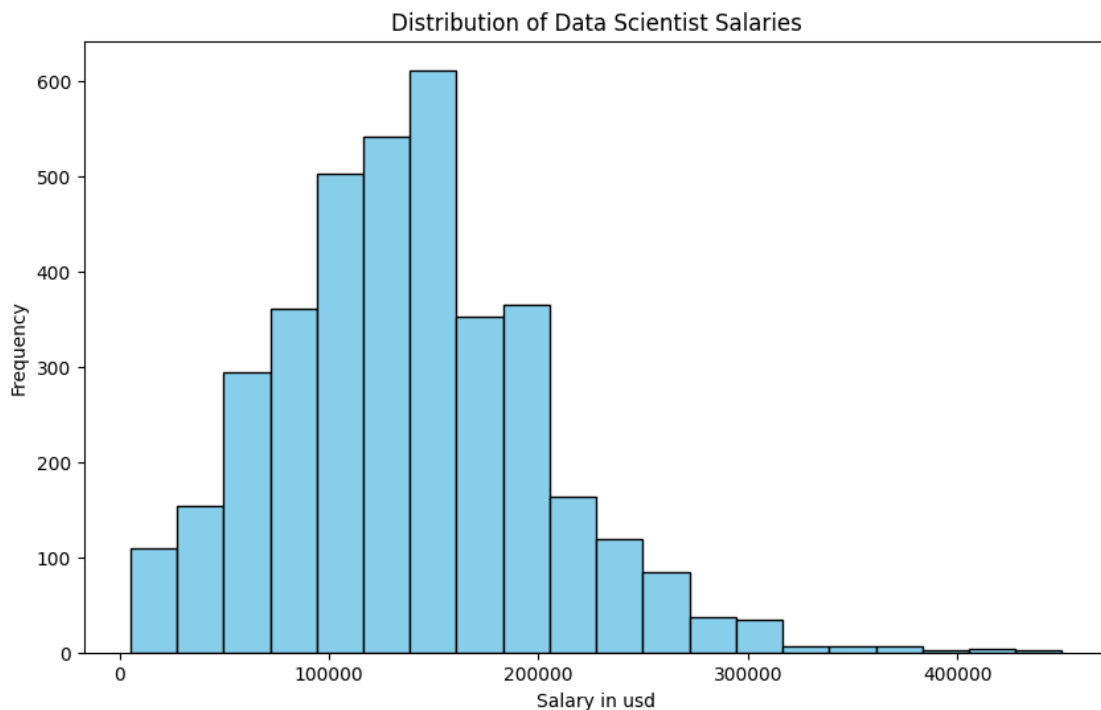
Name: count, Length: 93, dtype: int64

### 0.2.7 03: Create a histogram or bar graph from your data

```
[29]: # Create a histogram for Salary in usd
plt.figure(figsize=(10, 6))
plt.hist(dataset1_csv['salary_in_usd'], bins=20, color='skyblue',
         edgecolor='black')

# Customize the plot
plt.title('Distribution of Data Scientist Salaries')
plt.xlabel('Salary in usd')
plt.ylabel('Frequency')

# Show the plot
plt.show()
```

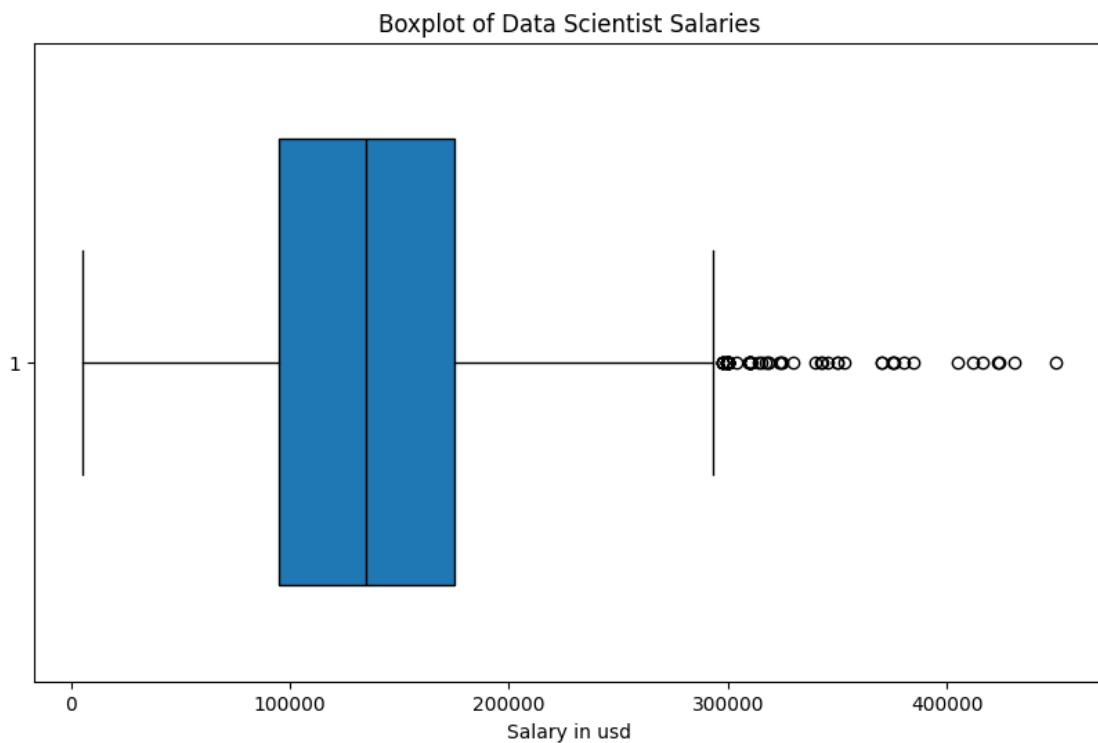


### 0.2.8 04: Create a boxplot from your data

```
[30]: # Create a boxplot
plt.figure(figsize=(10, 6))
plt.boxplot(dataset1_csv['salary_in_usd'], vert=False, widths=0.7, patch_artist=True, medianprops={'color': 'black'})

# Customize the plot
plt.title('Boxplot of Data Scientist Salaries')
plt.xlabel('Salary in usd')

# Show the plot
plt.show()
```

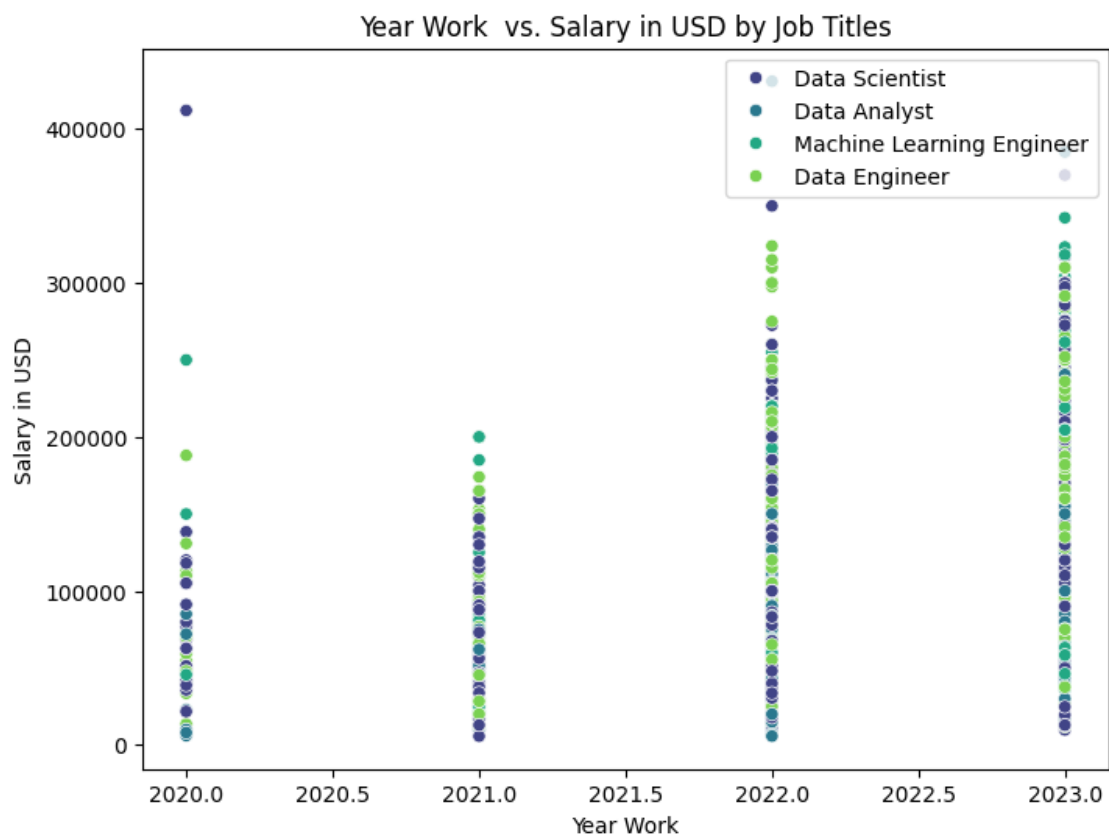


1. From the above histogram & boxplot the average data science job Salary is around 140K USD. Which is very good are per IT market.

### 0.2.9 05 : Create a bivariate plot from your data

```
[31]: # Bivariate Plot : Salry over years by Job titles
plt.figure(figsize=(8, 6))
# Limit the graph for most common job titles
job_titles_sal= ['Data Engineer', 'Data Scientist', 'Data Analyst', 'Machine_
↳ Learning Engineer']
```

```
# filter rows based on list values
dataset_mask = dataset1_csv['job_title'].isin(job_titles_sal)
dataset=dataset1_csv[dataset_mask]
#sns.scatterplot(x=dataset1_csv['Year Work'], y=df['salary_in_usd'],
#               ↪hue=df['job_title'],palette='viridis')
sns.scatterplot(x=dataset['work_year'], y=dataset['salary_in_usd'],
               ↪hue=dataset['job_title'],palette='viridis')
plt.xlabel('Year Work')
plt.ylabel('Salary in USD')
plt.title('Year Work vs. Salary in USD by Job Titles')
plt.legend(loc='upper right')
plt.show()
```



**0.2.10 06:** Create any additional visualizations that will help to answer the question(s) you want to answer.

```
[32]: # Determine titles below the threshold, e.g., less than N occurrences
N=50
low_frequency_titles = job_title_counts[job_title_counts < N].index
```

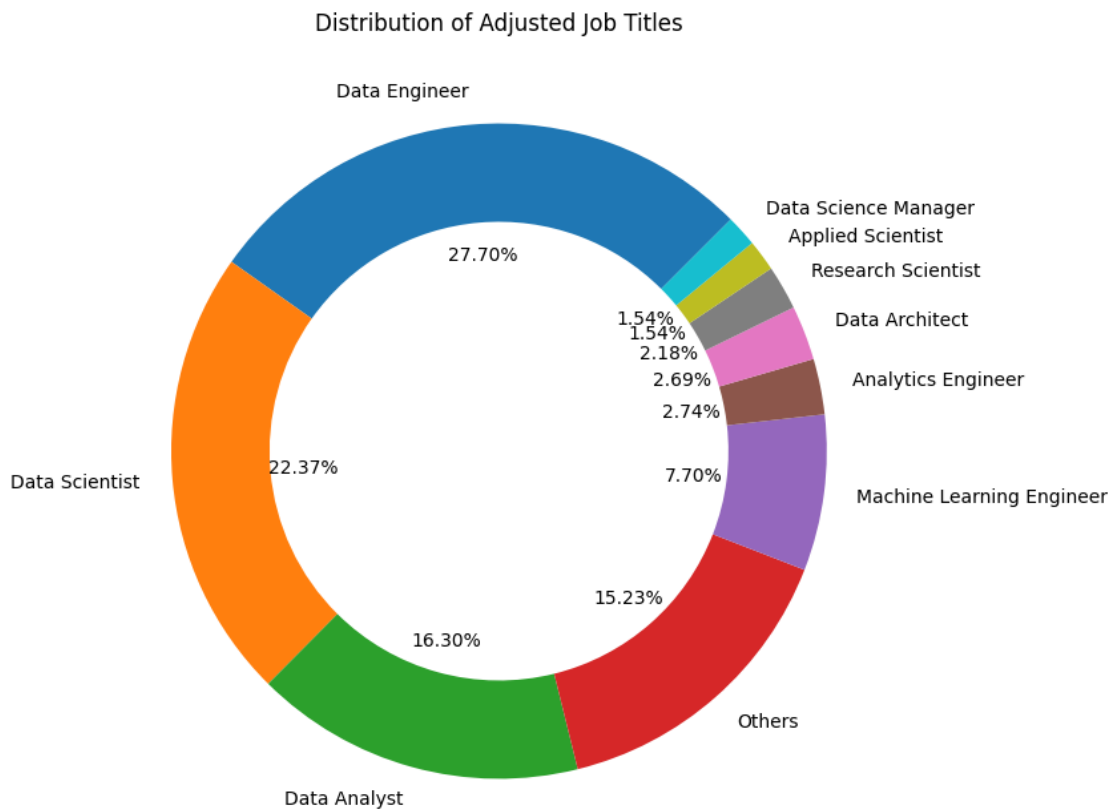
```

# Replace these titles in the dataframe with "Others"
dataset1_csv['adjusted_job_title'] = dataset1_csv['job_title'].apply(lambda x:
    ↪ "Others" if x in low_frequency_titles else x)

# Recalculate the frequency
adjusted_counts = dataset1_csv['adjusted_job_title'].value_counts()

# Plot
plt.figure(figsize=(10,8))
adjusted_counts.plot.pie(autopct='%.2f%%', startangle=45,
    ↪ wedgeprops=dict(width=0.3))
plt.title('Distribution of Adjusted Job Titles')
plt.ylabel('') # Hide the 'adjusted_job_title' y-label
plt.show()

```



### 0.2.11 Most frequent positions are:

1. Data Engineer
2. Data Scientist
3. Data Analyst



#### 4. Machine Learning Engineer

```
[33]: # Salary Trend Over Time by Job Title for
# a. Data Engineer          1040
# b. Data Scientist         840
# c. Data Analyst           614
# d. Machine Learning Engineer 291
# list of active subscription statuses
job_titles_sal= ['Data Engineer', 'Data Scientist', 'Data Analyst', 'Machine_
↳ Learning Engineer']
# filter rows based on list values
dataset_mask = dataset1_csv['job_title'].isin(job_titles_sal)
dataset=dataset1_csv[dataset_mask]

plt.figure(figsize=(10, 6))
p = sns.lineplot(data=dataset, x='work_year', y='salary_in_usd',
↳ hue='job_title', marker='o')

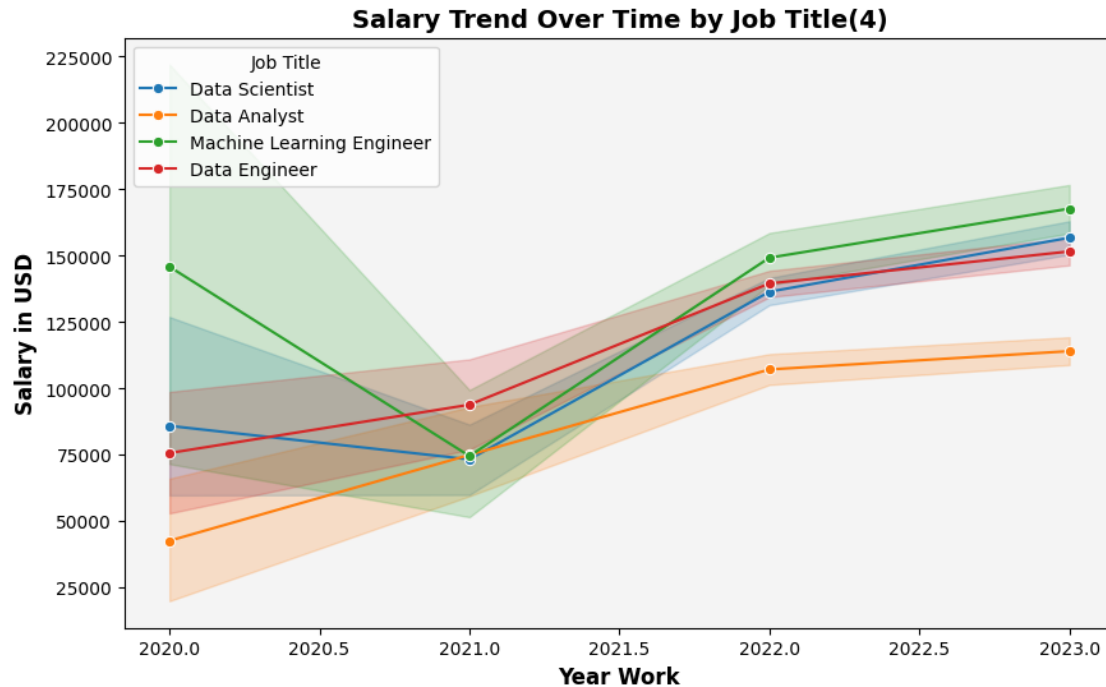
plt.xlabel('Year Work', fontsize=12, fontweight='bold')
plt.ylabel('Salary in USD', fontsize=12, fontweight='bold')

# Add a legend
plt.legend(title='Job Title', title_fontsize=10, fontsize=10, loc='upper left')

# Add a title
plt.title('Salary Trend Over Time by Job Title(4)', fontsize=14,
↳ fontweight='bold')

# Customize the background color
p.set_facecolor("#f4f4f4")

# Remove the grid lines
p.grid(False)
plt.show()
```



1. The salary trend in Machine learning engineer currently increasing better than data engineer/analyst/scientist
2. All the four job titles 'Data Engineer', 'Data Scientist', 'Data Analyst', 'Machine Learning Engineer' salary is in up trend.

```
[34]: # Average Data Science Salaries by Company Size
# Group the data by company_location and calculate the mean salary for Company_
↳Size
average_salaries_by_size = dataset1_csv.
↳groupby('company_size')['salary_in_usd'].mean().reset_index()

# Sort the Size by average salary in descending order
average_salaries_by_size = average_salaries_by_size.
↳sort_values(by='salary_in_usd', ascending=False)

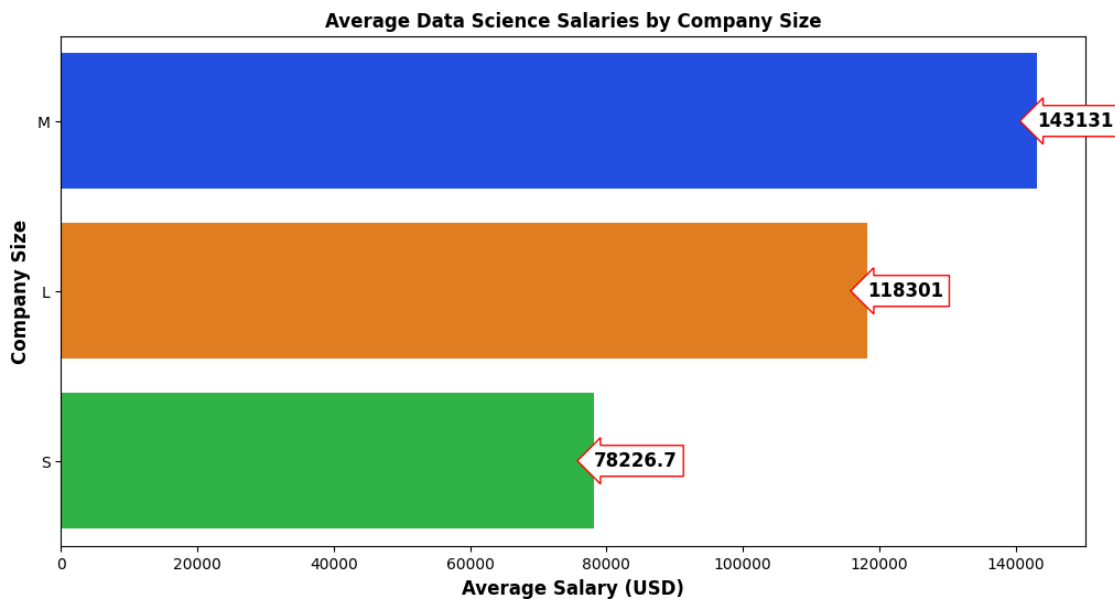
# Select the top N to plot
top_n_size = 10 # You can change this number as needed

# Create a bar chart to visualize average salaries by country
plt.figure(figsize=(12, 6))
p = sns.barplot(x='salary_in_usd', y='company_size',
↳data=average_salaries_by_size.head(top_n_locations), palette = 'bright')
plt.title('Average Data Science Salaries by Company Size', fontsize=12,
↳fontweight='bold' )
```

```
plt.xlabel('Average Salary (USD)', fontsize=12, fontweight='bold')
plt.ylabel('Company Size', fontsize=12, fontweight='bold')

for container in p.containers:
    p.bar_label(container,
                 fontsize = 12,
                 bbox = {'boxstyle': 'larrow', 'edgecolor': 'red', 'facecolor': 'white'},
                 label_type="edge",
                 fontweight = 'bold'
                )

# Customize the background color
#ax.set_facecolor("#f4f4f4")
plt.show()
```



**0.2.12 0.7 Summarize your results and make a conclusion. Explain how you arrived at this conclusion and how your visualizations support your conclusion.**

#### Histogram of Salary in USD :

The histogram showcases the distribution of Salary within the dataset. A histogram is a great way to understand the spread and frequency of different Salary ranges. In this case, it appears that the distribution of Salary is skewed to the right, indicating that most salary fall within a certain range , while there are a few jobs salary with significantly higher prices.

#### Boxplot of Salary in USD

From the above histogram & boxplot the average data science job salary is around 140K USD.

Which is very good are per IT market.

### **Scatterplot of year work vs. Salary in USD :**

This scatterplot illustrates the relationship between work year and salary in USD. The trend shows a general pattern where, as the work year increases, the salary tends to increase as well. This positive correlation suggests that in future data science job salary will increase. However, it's essential to note that while there's an overall trend, individual data science job type might have different salary despite having similar work experience. Other factors such as location or market conditions might influence these variations.

**Pie Chart of Different Job data science Job:** Pie Chart is a very useless visualization to represent the allocation or market share for multiple categories. In this plot we are able show the market job positions of each Job titles. The Pie chart show Most frequent positions are:

1. Data Engineer
2. Data Scientist
3. Data Analyst
4. Machine Learning Engineer

### **Bar chart for Salary depend on company size:**

With the above Bar chart, we can compare the Data Science job average salary that organization offer based on the organization size. Which will help the Job seeker to target the organization and set the salary expectation.

### **0.2.13 Conclusion:**

1. Average data science job Salary is around 140 K USD.
2. Most demanding data science job titles are a. Data Engineer b. Data Scientist c. Data Analyst d. Machine Learning Engineer
3. All the four job titles 'Data Engineer', 'Data Scientist', 'Data Analyst','Machine Learning Engineer' salary is in up trend.
4. Mid-Size organization comparatively pay more than big and small size organization
5. Data Engineer and Data Scientist job titles are more than other data science related job in the market.

[ ]: