C550-T301-Data_mining_2241_week2_Samanta_rajib

September 10, 2023

0.1 Class : C550-T301 Data Mining (2241-1)

0.2 Name : Rajib Samanta

0.2.1 Assignment : Week 2

1. Complete several of the Matplotlib tutorials at the following link until you feel comfortable: Matplotlib Tutorials.
2. Using a data set of your choice, write an introduction explaining the data set.
3. Identify a question or question(s) that you would like to explore in your data set.
4. Create at least three graphs that help answer these questions. Make sure your graphs are clearly readable and are labeled appropriately and professionally.
5. Explain what you have learned from each of your graphs.
6. Write a conclusion that summarizes your findings.

0.2.2 Data Set :

`Download the data file for Data Science Salary  2021 to 2023:`
`https://www.kaggle.com/datasets/harishkumardatalab/data-science-salary-2021-to-2023`

0.2.3 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.

0.2.4 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).

0.2.5 Data exploration:

1. **Optimal Hiring Decisions:** Analyze the dataset to determine the best employment type and experience level for hiring data science professionals for maximum cost-effectiveness.
2. **Salary Trends over Time:** Utilize the dataset to visualize and interpret data science salary trends from 2021 to 2023.
3. **Job Title Recommendation:** Recommend suitable job titles for candidates based on their experience level and desired salary range.

[40]:
```
# Load the Libraries
import os
import pandas as pd
import matplotlib.pyplot as plt
#%matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

[41]:
```
# 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/Sem6_fall_2023/week2'
# Set the working directory
os.chdir(directory)
print(os.getcwd())
dataset1_csv = pd.read_csv("Data Science Salary 2021 to 2023.csv")
dataset1_csv.head(10)
# Display the DataFrame 10 rows
```

```
/Users/rajibsamanta/Documents/Rajib/College/Sem6_fall_2023/week2
```

[41]:

	work_year	experience_level	employment_type	job_title	\
0	2023	EN	FT	Applied Scientist	
1	2023	EN	FT	Applied Scientist	
2	2023	EN	FT	Data Quality Analyst	
3	2023	EN	FT	Compliance Data Analyst	
4	2023	EN	FT	Applied Scientist	
5	2023	EN	FT	Applied Scientist	
6	2023	EN	FT	Machine Learning Engineer	
7	2023	EN	FT	Machine Learning Engineer	
8	2023	EN	FT	Research Scientist	
9	2023	EN	FT	Data Engineer	

	salary	salary_currency	salary_in_usd	company_location	company_size

```
0  213660              USD         213660              US            L
1  130760              USD         130760              US            L
2  100000              USD         100000              NG            L
3   30000              USD          30000              NG            L
4  204620              USD         204620              US            L
5  110680              USD         110680              US            L
6  163196              USD         163196              US            M
7  145885              USD         145885              US            M
8  220000              USD         220000              US            L
9   85000              USD          85000              US            M
```

[42]:
```
# describe the dataframe'
dataset1_csv.shape
## It has 3761 records with 9 columns
```

[42]:
```
(3761, 9)
```

[43]:
```
# describe the dataframe'
dataset1_csv.describe()
```

[43]:
```
         work_year        salary  salary_in_usd
count  3761.000000  3.761000e+03    3761.000000
mean   2022.374103  1.905999e+05  137555.178942
std       0.691252  6.711457e+05   63022.267974
min    2020.000000  6.000000e+03    5132.000000
25%    2022.000000  1.000000e+05   95000.000000
50%    2022.000000  1.375000e+05  135000.000000
75%    2023.000000  1.800000e+05  175000.000000
max    2023.000000  3.040000e+07  450000.000000
```

[44]:
```
# missing values
dataset1_csv.isnull().sum()
#-- No  null column
```

[44]:
```
work_year           0
experience_level    0
employment_type     0
job_title           0
salary              0
salary_currency     0
salary_in_usd       0
company_location    0
company_size        0
dtype: int64
```

[45]:
```
# Replace column values with more descriptive information
```

```
dataset1_csv['experience_level'] = dataset1_csv['experience_level'].
 ↪replace('EN', 'Entry-Level')
dataset1_csv['experience_level'] = dataset1_csv['experience_level'].
 ↪replace('EX', 'Experienced')
dataset1_csv['experience_level'] = dataset1_csv['experience_level'].
 ↪replace('MI', 'Mid-Level')
dataset1_csv['experience_level'] = dataset1_csv['experience_level'].
 ↪replace('SE', 'Senior')
dataset1_csv['employment_type'] = dataset1_csv['employment_type'].replace('FT',␣
 ↪'Full-Time')
dataset1_csv['employment_type'] = dataset1_csv['employment_type'].replace('CT',␣
 ↪'Contractor')
dataset1_csv['employment_type'] = dataset1_csv['employment_type'].replace('FL',␣
 ↪'Freelancer')
dataset1_csv['employment_type'] = dataset1_csv['employment_type'].replace('PT',␣
 ↪'Part-Time')
dataset1_csv['company_size'] = dataset1_csv['company_size'].replace('L',␣
 ↪"Large")
dataset1_csv['company_size'] = dataset1_csv['company_size'].replace('M',␣
 ↪"Medium")
dataset1_csv['company_size'] = dataset1_csv['company_size'].replace('S',␣
 ↪"Small")
dataset1_csv.head()
```

[45]:

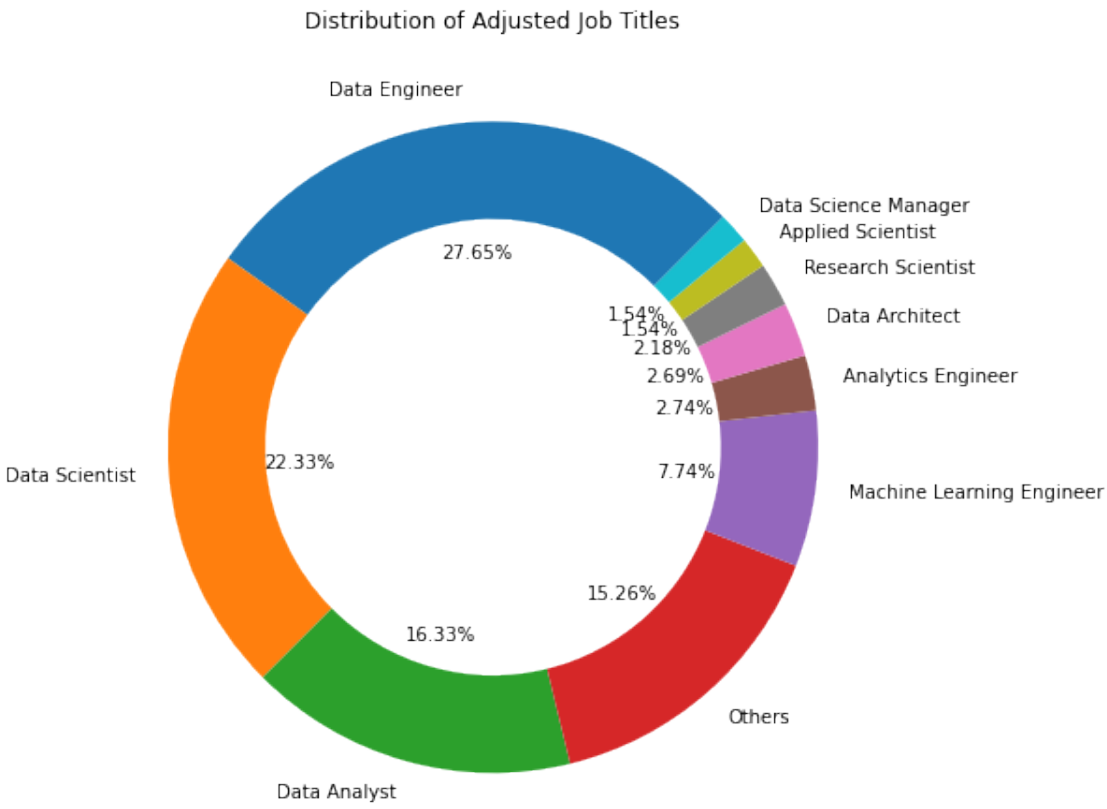
	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	company_location	company_size
0	2023	Entry-Level	Full-Time	Applied Scientist	213660	USD	213660	US	Large
1	2023	Entry-Level	Full-Time	Applied Scientist	130760	USD	130760	US	Large
2	2023	Entry-Level	Full-Time	Data Quality Analyst	100000	USD	100000	NG	Large
3	2023	Entry-Level	Full-Time	Compliance Data Analyst	30000	USD	30000	NG	Large
4	2023	Entry-Level	Full-Time	Applied Scientist	204620	USD	204620	US	Large

[46]:
```
# Calculate frequency of each job title
job_title_counts = dataset1_csv['job_title'].value_counts()
job_title_counts
```

[46]:
```
Data Engineer                1040
Data Scientist                840
Data Analyst                  614
Machine Learning Engineer     291
```

```
Analytics Engineer            103
                              …
Compliance Data Analyst         1
BI Data Engineer                1
Deep Learning Researcher        1
Head of Machine Learning        1
Staff Data Analyst              1
Name: job_title, Length: 93, dtype: int64
```

[47]:
```
# 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()
```



0.2.6 Most frequent positions are:

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

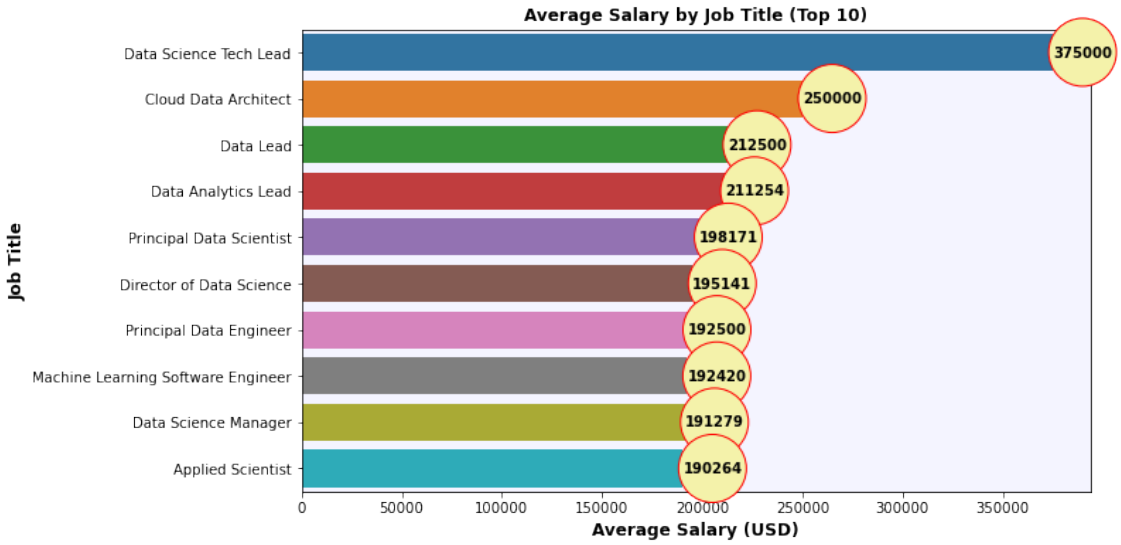
[48]:
```
dataset1_csv.head()
```

[48]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	company_location	company_size
0	2023	Entry-Level	Full-Time	Applied Scientist	213660	USD	213660	US	Large
1	2023	Entry-Level	Full-Time	Applied Scientist	130760	USD	130760	US	Large
2	2023	Entry-Level	Full-Time	Data Quality Analyst	100000	USD	100000	NG	Large
3	2023	Entry-Level	Full-Time	Compliance Data Analyst					
4	2023	Entry-Level	Full-Time	Applied Scientist					

```
3   30000             USD          30000               NG        Large
4  204620             USD         204620               US        Large

  adjusted_job_title
0  Applied Scientist
1  Applied Scientist
2             Others
3             Others
4  Applied Scientist
```

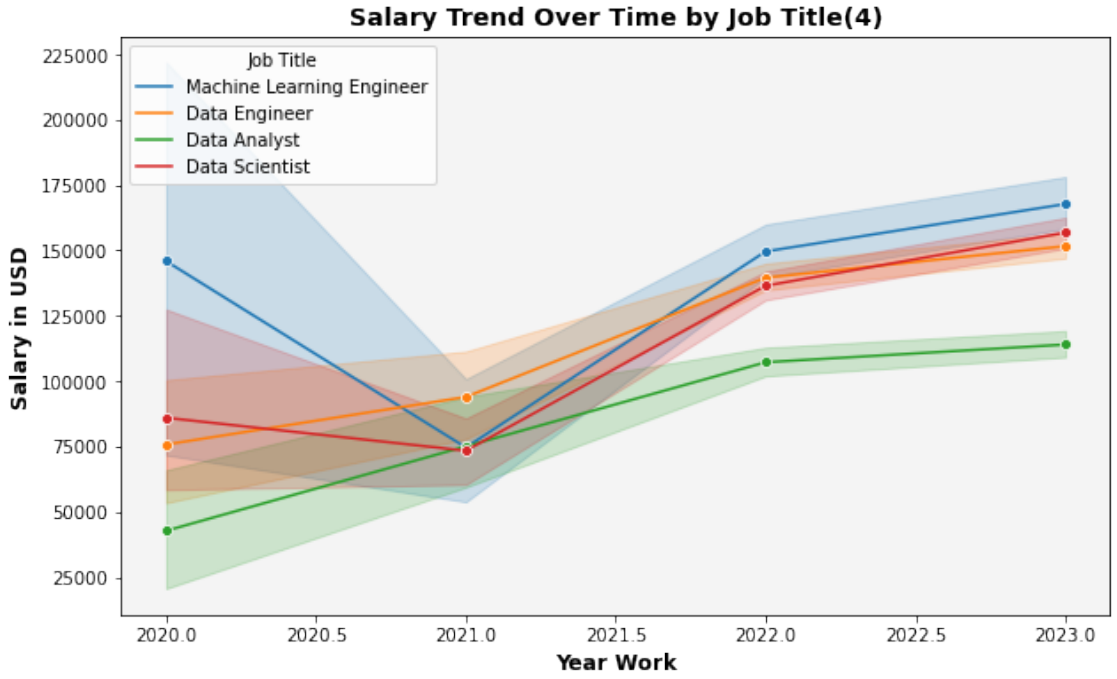
[49]:
```
# Determine Average Salary by Job Title
# Group data by 'adjusted_job_title' and calculate the average salary for each␣
 ↪title
job_title_salary= dataset1_csv.groupby('job_title')['salary_in_usd'].mean().
 ↪sort_values(ascending = False)

plt.figure(figsize = (10,6))
p = sns.barplot(x= job_title_salary.values[:10], y = job_title_salary.index[:
 ↪10])

plt.title('Average Salary by Job Title (Top 10)', fontsize=12,␣
 ↪fontweight='bold')
plt.xlabel('Average Salary (USD)', fontsize=12, fontweight='bold')
plt.ylabel('Job Title', fontsize=12, fontweight='bold')

for container in p.containers:
    p.bar_label(container,

                bbox = {'boxstyle': 'circle', 'facecolor': '#f4f2aa',␣
 ↪'edgecolor': 'red'},
                fontweight = 'bold'


               )
# Customize the background color
p.set_facecolor("#f4f4ff")

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



1. Data Science Tech Lead has highest average salary at 375,000 USD.
2. Cloud Data Architect & Data Lead also have notably high salaries.
3. The top 10 job titles exhibit strong earning potential in the data science field.

[50]:
```
# 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')
```

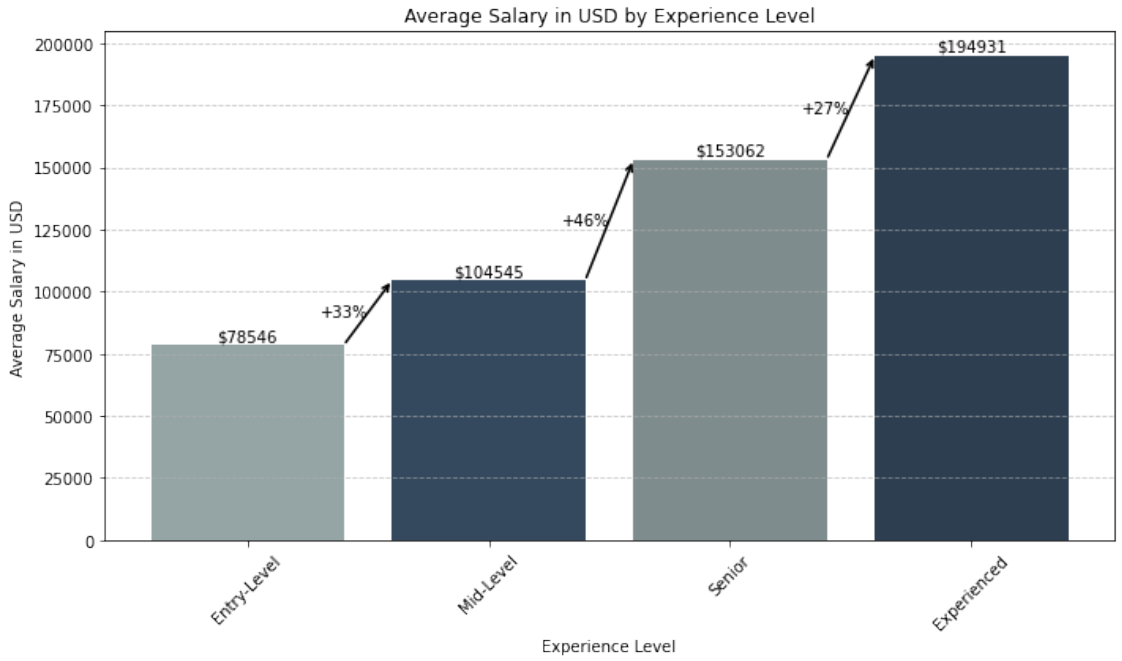
```
# Customize the background color
p.set_facecolor("#f4f4f4")

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



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

[51]:
```
# Average salary by experience level
# Calculate the average salary in USD for each experience level
avg_salaries = dataset1_csv.groupby('experience_level')['salary_in_usd'].mean().
 ↪reset_index()

# Sort values
avg_salaries = avg_salaries.sort_values(by='salary_in_usd', ascending=True)

# Neutral Colors
colors = ['#95a5a6', '#34495e', '#7f8c8d', '#2c3e50']
```

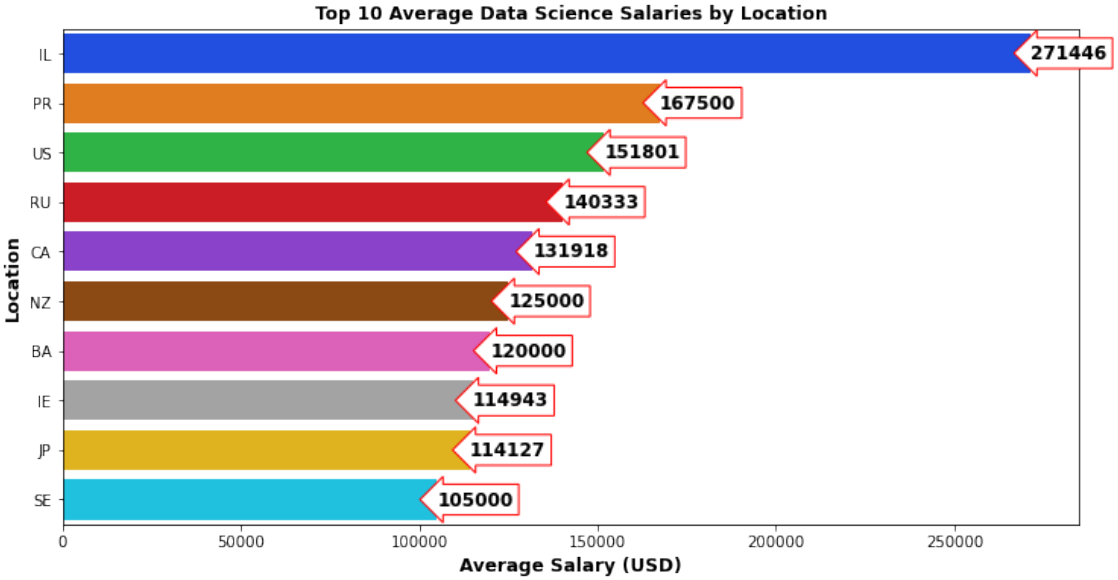
```
# Plotting
plt.figure(figsize=(10,6))
bars = plt.bar(avg_salaries['experience_level'], avg_salaries['salary_in_usd'],␣
 ↪color=colors)
plt.title('Average Salary in USD by Experience Level')
plt.xlabel('Experience Level')
plt.ylabel('Average Salary in USD')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.xticks(rotation=45)
# Display the average salary on top of each bar
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, yval + 500, f"${int(round(yval,␣
 ↪0))}", ha='center', va='bottom')

# Add arrows pointing diagonally to the next bar with percentage increase for␣
 ↪each
for i in range(len(bars)-1):
    start_height = bars[i].get_height()
    end_height = bars[i+1].get_height()

    percentage_increase = (end_height - start_height) / start_height * 100

    start_point = (bars[i].get_x() + bars[i].get_width(), start_height)
    end_point = (bars[i+1].get_x(), end_height)

    plt.annotate(
        '',
        xy=end_point,
        xytext=start_point,
        arrowprops=dict(facecolor='black', arrowstyle='->', lw=1.5),
    )
    plt.text((start_point[0] + end_point[0]) / 2, (start_point[1] +␣
 ↪end_point[1]) / 2, f"+{percentage_increase:.0f}%", ha='right', va='center')

plt.tight_layout()
plt.show()
```



1. There is a notable salary jump when transitioning from a Middle-level to a Senior-level position.
2. Experienced professionals earns the highest amount with an average of around 190K
3. Entry-level positions, on average, earn about half of what senior-level positions do.

[52]:
```
# Average Data Science Salaries by Location
# Group the data by company_location and calculate the mean salary for each␣
 ↪location
average_salaries_by_location = dataset1_csv.
 ↪groupby('company_location')['salary_in_usd'].mean().reset_index()

# Sort the locations by average salary in descending order
average_salaries_by_location = average_salaries_by_location.
 ↪sort_values(by='salary_in_usd', ascending=False)

# Select the top N locations to plot
top_n_locations = 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_location',␣
 ↪data=average_salaries_by_location.head(top_n_locations), palette = 'bright')
plt.title('Top {} Average Data Science Salaries by Location'.
 ↪format(top_n_locations), fontsize=12, fontweight='bold' )
plt.xlabel('Average Salary (USD)', fontsize=12, fontweight='bold')
```

```
plt.ylabel('Location', 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()
```



1. In Illinois (IL), the average data science salary is notably high, at approximately 271,447 USD.
2. Puerto Rico (PR) and the United States (US) also offer competitive average salaries, with approximately 167,500 USD 3. and 151,801 USD, respectively.
3. Russia (RU) and Canada (CA) have average data science salaries of around 140,333 USD and 131,918 USD, respectively.
4. New Zealand (NZ), Bosnia and Herzegovina (BA), Ireland (IE), Japan (JP), and Sweden (SE) round out the top 6. locations with varying average salaries.

0.2.7 Conclusion:

1. Experienced pros earn most. Seniors follow, then mid-level, and entry-level.
2. Data Science Tech Lead earns the most.

3. USD salaries are highest. ILS, GBP, and CHF follow.
4. There is an obvious correlation between the salary and the experience level for each job.
5. Company size doesn't appear to influence the income across job titles.