Analysis in Data Science Salaries in 2023

May 10, 2023

1 Analysis in Data Science Salaries in 2023

Data science is emerging career these days. Everybody wants to pursue their career in data science. To know whether its a growing opportunities or not analysis is performed in data science salaries in 2023. Data set is taken from the kaggle. In the dataset there are salaries of different data science fields in the data science domain.

1.1 Downloading the data

[2]:
```
!pip install jovian opendatasets --upgrade --quiet
```

[3]:
```
dataset_url = 'https://www.kaggle.com/datasets/arnabchaki/
↪data-science-salaries-2023'
```

[4]:
```
import opendatasets as od
od.download(dataset_url)
```

```
Please provide your Kaggle credentials to download this dataset. Learn more:
http://bit.ly/kaggle-creds
Your Kaggle username: kiranpandey98
Your Kaggle Key: ·····
Downloading data-science-salaries-2023.zip to ./data-science-salaries-2023
100%|      | 25.4k/25.4k [00:00<00:00, 18.9MB/s]
```

[5]:
```
data_dir = './data-science-salaries-2023'
```

[6]:
```
import os
os.listdir(data_dir)
```

[6]:
```
['ds_salaries.csv']
```

[7]:
```
project_name = "data-science-salaries-analysis"
```

[8]:
```
!pip install jovian --upgrade -q
```

[9]:
```
import jovian
```

[10]:
```
jovian.commit(project = project_name)
```

```
<IPython.core.display.Javascript object>
```

```
[jovian] Updating notebook "pandeykiran571/data-science-salaries-analysis" on
https://jovian.com
[jovian] Committed successfully! https://jovian.com/pandeykiran571/data-science-
salaries-analysis
```

[10]:
```
'https://jovian.com/pandeykiran571/data-science-salaries-analysis'
```

1.2 Data preparation and Cleaning

Loading the dataset

[11]:
```
import pandas as pd
```

[12]:
```
raw_data_df = pd.read_csv(data_dir + '/ds_salaries.csv')
```

[13]:
```
raw_data_df
```

[13]:

	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	
…	…	…	…	…	
3750	2020	SE	FT	Data Scientist	
3751	2021	MI	FT	Principal Data Scientist	
3752	2020	EN	FT	Data Scientist	
3753	2020	EN	CT	Business Data Analyst	
3754	2021	SE	FT	Data Science Manager	

	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	
…	…	…	…	…	…	
3750	412000	USD	412000	US	100	
3751	151000	USD	151000	US	100	
3752	105000	USD	105000	US	100	
3753	100000	USD	100000	US	100	
3754	7000000	INR	94665	IN	50	

```
     company_location company_size
0                  ES            L
1                  US            S
2                  US            S
3                  CA            M
4                  CA            M
…                 …           …
3750               US            L
3751               US            L
3752               US            S
3753               US            L
3754               IN            L

[3755 rows x 11 columns]
```

To explore the column name in dataset

[14]:
```
raw_data_df.info()
```

```
<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
```

To summarize the overall data

[15]:
```
raw_data_df.describe()
```

[15]:

	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
```

Cleaning the dataset by handling missing information and preparing dataset for analysis

[16]:
```
raw_data_df.isnull().sum()
```

[16]:
```
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
```

No missing values in data frame

[17]:
```
raw_data_df.dtypes
```

[17]:
```
work_year              int64
experience_level      object
employment_type       object
job_title             object
salary                 int64
salary_currency       object
salary_in_usd          int64
employee_residence    object
remote_ratio           int64
company_location      object
company_size          object
dtype: object
```

[18]:
```
new_df =␣
 ↪raw_data_df[['job_title','experience_level','employment_type','salary_in_usd','company_size
 ↪copy()
```

[19]:
```
new_df
```

[19]:
```
                    job_title experience_level employment_type  \
0     Principal Data Scientist               SE              FT
1                  ML Engineer               MI              CT
2                  ML Engineer               MI              CT
3               Data Scientist               SE              FT
4               Data Scientist               SE              FT
…                          …              …             …
3750            Data Scientist               SE              FT
3751  Principal Data Scientist               MI              FT
3752            Data Scientist               EN              FT
3753    Business Data Analyst                EN              CT
3754     Data Science Manager                SE              FT

      salary_in_usd company_size
0             85847            L
1             30000            S
2             25500            S
3            175000            M
4            120000            M
…               …          …
3750         412000            L
3751         151000            L
3752         105000            S
3753         100000            L
3754          94665            L

[3755 rows x 5 columns]
```

[20]:
```
job_types = new_df.pivot_table(index = ['job_title'], aggfunc ='size')
```

[21]:
```
job_types
```

[21]:
```
job_title
3D Computer Vision Researcher      4
AI Developer                      11
AI Programmer                      2
AI Scientist                      16
Analytics Engineer               103
                                  …
Research Engineer                 37
Research Scientist                82
Software Data Engineer             2
Staff Data Analyst                 1
Staff Data Scientist               1
Length: 93, dtype: int64
```

Selecting specific job title such as: Data Analyst, Data Scientist, Data Engineer, BI Analyst

[22]:
```
options = ['Data Analyst','Data Scientist', 'Data Engineer', 'BI Analyst']
db_df = new_df[new_df['job_title'].isin(options)]
db_df
```

[22]:
```
            job_title experience_level employment_type  salary_in_usd  \
3      Data Scientist               SE              FT         175000
4      Data Scientist               SE              FT         120000
7      Data Scientist               SE              FT         219000
8      Data Scientist               SE              FT         141000
9      Data Scientist               SE              FT         147100
…                 …                …               …              …
3743    Data Engineer               MI              FT         130800
3746   Data Scientist               MI              FT         119059
3748    Data Engineer               MI              FT          28369
3750   Data Scientist               SE              FT         412000
3752   Data Scientist               EN              FT         105000

     company_size
3               M
4               M
7               M
8               M
9               M
…             …
3743            M
3746            M
3748            L
3750            L
3752            S

[2501 rows x 5 columns]
```

[23]:
```
job_types = db_df.pivot_table(index = ['job_title'], aggfunc ='size')
```

[24]:
```
job_types
```

[24]:
```
job_title
BI Analyst           9
Data Analyst       612
Data Engineer     1040
Data Scientist     840
dtype: int64
```

As only specific job title is selected for the analysis

[25]:
```
data_df = db_df.sample(n=200)
```

[26]:
```
data_df
```

[26]:
```
           job_title experience_level employment_type  salary_in_usd  \
2063    Data Analyst               SE              FT         201000
2177    Data Analyst               SE              FT         100000
489   Data Scientist               SE              FT         170730
2226  Data Scientist               MI              FT         180000
2310   Data Engineer               SE              FT         135000
…                ...              ...             ...            ...
2223   Data Engineer               SE              FT         150000
750    Data Engineer               SE              FT          75000
1782  Data Scientist               SE              FT         140000
1717  Data Scientist               SE              FT         140000
457    Data Engineer               EN              FT          92700

     company_size
2063            M
2177            M
489             M
2226            M
2310            M
…             ...
2223            M
750             M
1782            M
1717            M
457             M

[200 rows x 5 columns]
```

For analysis 200 rows are selected using random sampling

[27]:
```
import jovian
```

[28]:
```
jovian.commit()
```

```
<IPython.core.display.Javascript object>
```

```
[jovian] Updating notebook "pandeykiran571/data-science-salaries-analysis" on
https://jovian.com
[jovian] Committed successfully! https://jovian.com/pandeykiran571/data-science-
salaries-analysis
```

[28]:
```
'https://jovian.com/pandeykiran571/data-science-salaries-analysis'
```

1.3 Exploratory Analysis and Visualization

[29]:
```
data_df
```

[29]:
```
           job_title experience_level employment_type  salary_in_usd  \
2063    Data Analyst               SE              FT         201000
2177    Data Analyst               SE              FT         100000
489   Data Scientist               SE              FT         170730
2226  Data Scientist               MI              FT         180000
2310   Data Engineer               SE              FT         135000
…                ...              ...             ...            ...
2223   Data Engineer               SE              FT         150000
750    Data Engineer               SE              FT          75000
1782  Data Scientist               SE              FT         140000
1717  Data Scientist               SE              FT         140000
457    Data Engineer               EN              FT          92700

     company_size
2063            M
2177            M
489             M
2226            M
2310            M
…             ...
2223            M
750             M
1782            M
1717            M
457             M

[200 rows x 5 columns]
```

[30]:
```
data_df.describe()
```

[30]:
```
       salary_in_usd
count     200.000000
mean   127795.365000
std     58974.837802
min      5679.000000
25%     85049.500000
50%    129300.000000
75%    162825.000000
max    300240.000000
```

importing libraries for visualization

[31]:
```
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Visualizing the salaries ranges of data science career

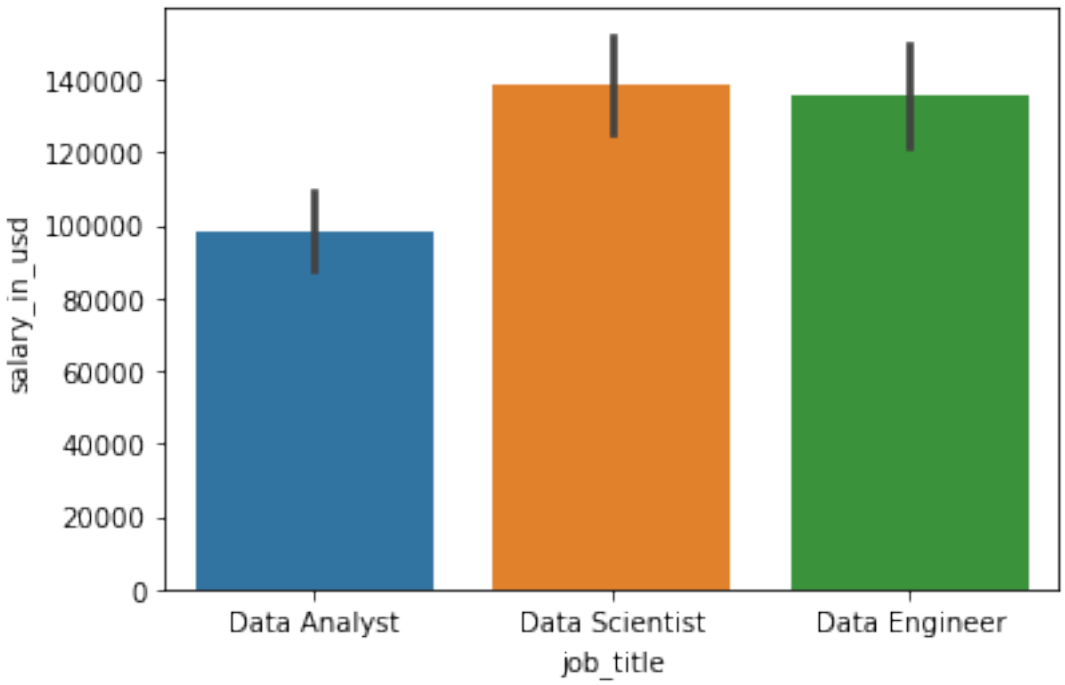
[32]:
```
data_df
```

[32]:
```
           job_title experience_level employment_type  salary_in_usd  \
2063    Data Analyst               SE              FT         201000
2177    Data Analyst               SE              FT         100000
489   Data Scientist               SE              FT         170730
2226  Data Scientist               MI              FT         180000
2310   Data Engineer               SE              FT         135000
…                …                …               …              …
2223   Data Engineer               SE              FT         150000
750    Data Engineer               SE              FT          75000
1782  Data Scientist               SE              FT         140000
1717  Data Scientist               SE              FT         140000
457    Data Engineer               EN              FT          92700

     company_size
2063            M
2177            M
489             M
2226            M
2310            M
…             …
2223            M
750             M
1782            M
1717            M
457             M

[200 rows x 5 columns]
```

[33]:
```
x= data_df.job_title
y= data_df.salary_in_usd
sns.barplot(x, y);

import warnings
warnings.filterwarnings('ignore')
```

```
/opt/conda/lib/python3.9/site-packages/seaborn/_decorators.py:36: FutureWarning:
Pass the following variables as keyword args: x, y. From version 0.12, the only
valid positional argument will be `data`, and passing other arguments without an
explicit keyword will result in an error or misinterpretation.
```

```
warnings.warn(
```



In above bar graph, highest salary received by each professional is shown

[34]:
```
d_a = data_df.groupby('job_title')
d_a.first
```

[34]:
```
<bound method GroupBy.first of <pandas.core.groupby.generic.DataFrameGroupBy
object at 0x7effccf4a430>>
```

[35]:
```
data_analyst = d_a.get_group('Data Analyst')
data_analyst
```

[35]:

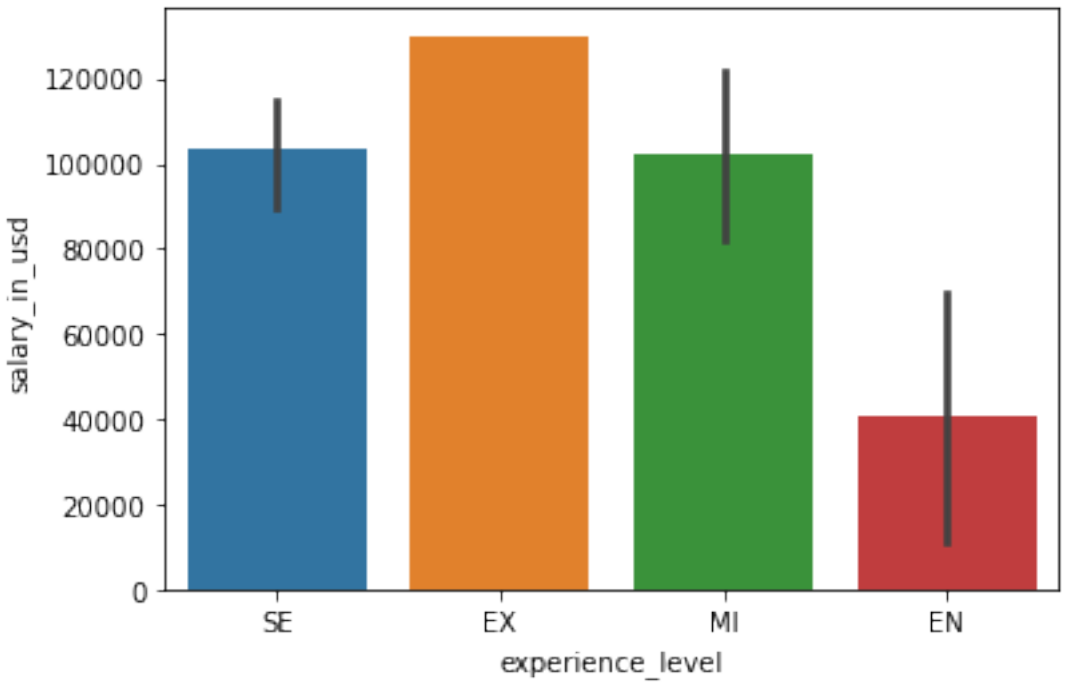
	job_title	experience_level	employment_type	salary_in_usd	\
2063	Data Analyst	SE	FT	201000	
2177	Data Analyst	SE	FT	100000	
3274	Data Analyst	EX	FT	130000	
3009	Data Analyst	SE	FT	99450	
2669	Data Analyst	SE	FT	81666	
1058	Data Analyst	SE	FT	153600	
1655	Data Analyst	MI	FT	116000	
1197	Data Analyst	EN	FT	75000	
2468	Data Analyst	SE	FT	120000	
2059	Data Analyst	SE	FT	95000	
1688	Data Analyst	MI	FT	150000	

1878	Data Analyst	SE	FT	149000
1052	Data Analyst	SE	FT	169000
2014	Data Analyst	SE	FT	39925
422	Data Analyst	SE	FT	70000
2840	Data Analyst	EN	FT	7799
1221	Data Analyst	SE	FT	55800
3149	Data Analyst	SE	FT	90320
1009	Data Analyst	SE	FT	152380
3255	Data Analyst	MI	FT	106260
3139	Data Analyst	EN	FT	15000
2538	Data Analyst	SE	FT	50432
1231	Data Analyst	SE	FT	80000
1732	Data Analyst	SE	FT	125000
1391	Data Analyst	MI	FT	100000
571	Data Analyst	SE	FT	100000
704	Data Analyst	MI	FT	85000
2939	Data Analyst	SE	FT	50000
544	Data Analyst	SE	FT	128500
2350	Data Analyst	MI	FT	150000
2169	Data Analyst	SE	FT	70186
2214	Data Analyst	SE	FT	110600
1656	Data Analyst	MI	FT	72000
282	Data Analyst	SE	FT	149500
2625	Data Analyst	SE	FT	85700
3072	Data Analyst	MI	FT	130000
1296	Data Analyst	SE	FT	93919
2389	Data Analyst	SE	FT	100000
1164	Data Analyst	MI	FT	72914
1455	Data Analyst	SE	FT	108000
2196	Data Analyst	MI	FT	100000
1841	Data Analyst	MI	FT	150000
3016	Data Analyst	SE	FT	61566
1148	Data Analyst	MI	FT	60000
1773	Data Analyst	SE	FT	64000
3197	Data Analyst	MI	FT	36940
144	Data Analyst	SE	FT	138900
425	Data Analyst	EN	FT	64200
2908	Data Analyst	SE	FT	115000

	company_size
2063	M
2177	M
3274	M
3009	M
2669	M
1058	M
1655	M

```
1197            M
2468            M
2059            M
1688            M
1878            M
1052            M
2014            M
422             M
2840            L
1221            M
3149            M
1009            M
3255            M
3139            L
2538            M
1231            M
1732            M
1391            M
571             M
704             M
2939            S
544             M
2350            M
2169            M
2214            M
1656            M
282             M
2625            M
3072            M
1296            M
2389            M
1164            M
1455            M
2196            M
1841            M
3016            M
1148            M
1773            M
3197            M
144             M
425             M
2908            L
```

[36]:
```
sns.barplot(data_analyst.experience_level,y)

warnings.filterwarnings('ignore')
```



In this bar graph, it is shown that how much does Data analyst can earn according to the experience. Such as senior, midlevel, Executive level

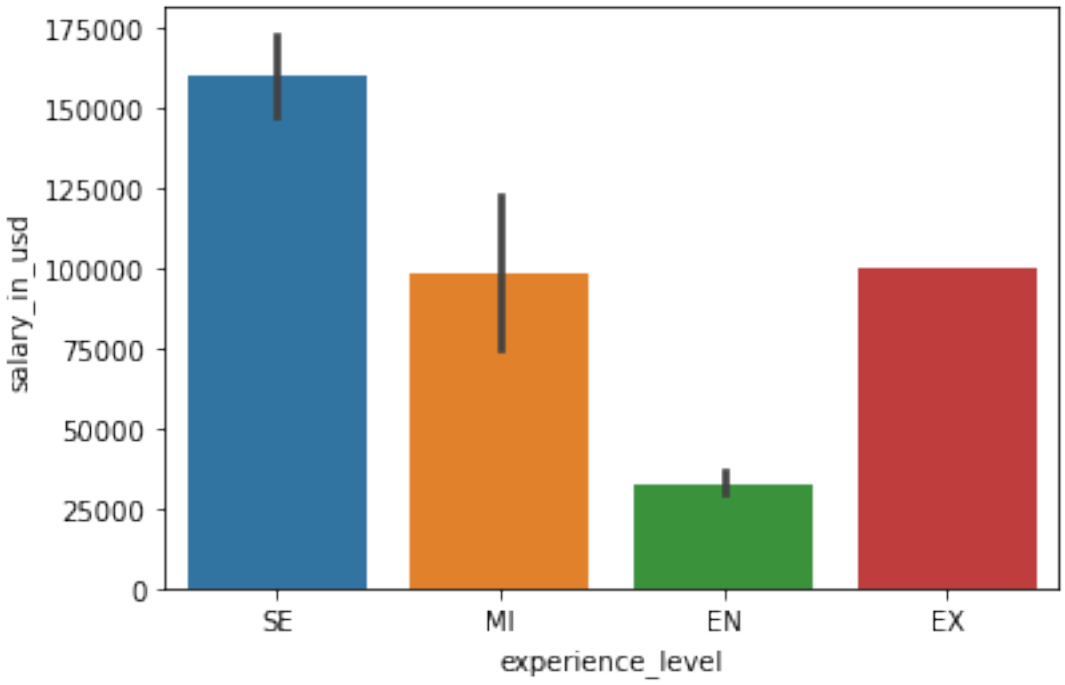
[37]:
```
data_scientist = d_a.get_group('Data Scientist')
data_scientist
```

[37]:

	job_title	experience_level	employment_type	salary_in_usd	company_size
489	Data Scientist	SE	FT	170730	M
2226	Data Scientist	MI	FT	180000	M
2909	Data Scientist	EN	FT	31520	M
2250	Data Scientist	SE	FT	37824	M
1050	Data Scientist	SE	FT	149076	M
…	…	…	…	…	
2101	Data Scientist	SE	FT	130000	
2756	Data Scientist	SE	FT	191475	
1826	Data Scientist	SE	FT	150000	
1782	Data Scientist	SE	FT	140000	
1717	Data Scientist	SE	FT	140000	

```
…            …
2101             M
2756             M
1826             M
1782             M
1717             M

[74 rows x 5 columns]
```

[38]:
```
sns.barplot(data_scientist.experience_level,y)

warnings.filterwarnings('ignore')
```



Similarly, data scientist mid level, entry level and senior level earning is shown in this bar graph

[39]:
```
c_l = data_df.groupby('company_size')
d_a.first
```

[39]:
```
<bound method GroupBy.first of <pandas.core.groupby.generic.DataFrameGroupBy
object at 0x7effccf4a430>>
```

[40]:
```
company_level = c_l.get_group('L')
```

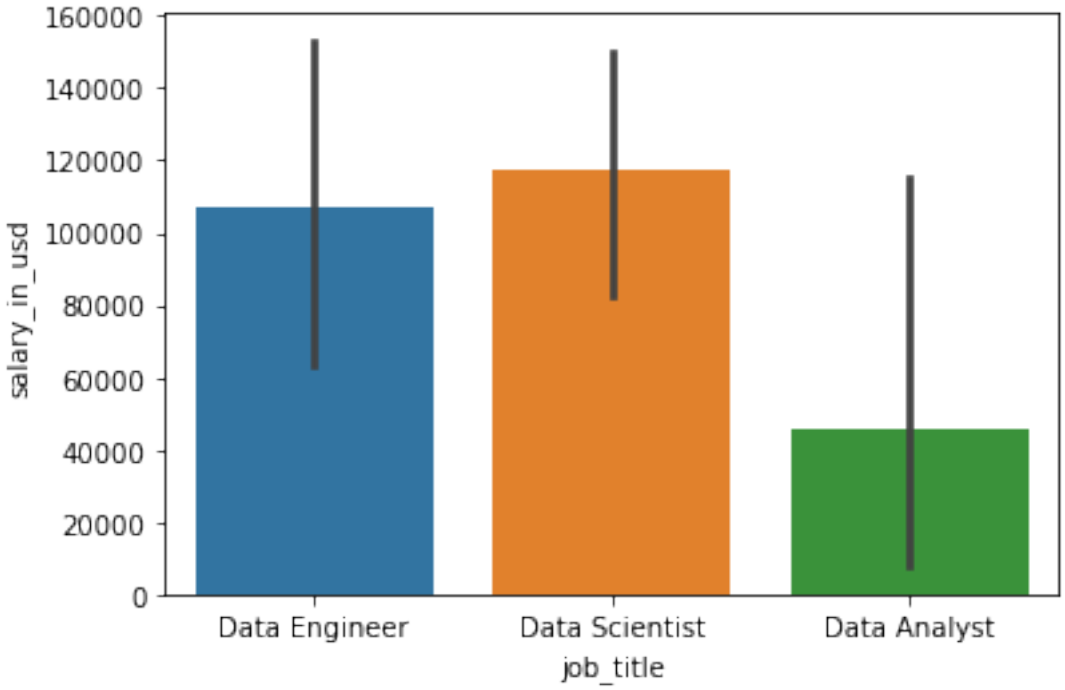
[41]:
```
company_level
```

[41]:

	job_title	experience_level	employment_type	salary_in_usd	company_size
3592	Data Engineer	MI	FT	70139	L
3598	Data Scientist	EN	FT	36643	L
3688	Data Scientist	MI	FT	88654	L
2715	Data Engineer	SE	FT	229998	L
2702	Data Engineer	SE	FT	132100	L
3604	Data Scientist	SE	FT	103691	L
3120	Data Engineer	EN	FT	13493	L
3581	Data Scientist	EN	FT	29751	L
1554	Data Engineer	SE	FT	205600	L
2840	Data Analyst	EN	FT	7799	L
2725	Data Scientist	SE	FT	236900	L
3212	Data Scientist	SE	FT	215300	L
3542	Data Engineer	MI	FT	28476	L
3090	Data Scientist	SE	FT	100000	L
3139	Data Analyst	EN	FT	15000	L
3453	Data Scientist	MI	FT	33609	L
3134	Data Scientist	SE	FT	160000	L
2662	Data Scientist	MI	FT	108000	L
1555	Data Engineer	SE	FT	105700	
3166	Data Scientist	SE	FT	215300	
3195	Data Engineer	SE	FT	100800	
1443	Data Scientist	MI	FT	155000	
3513	Data Engineer	EN	FT	80000	
2947	Data Scientist	SE	FT	119300	
997	Data Scientist	MI	FT	40000	
2908	Data Analyst	SE	FT	115000	

```
1555            L
3166            L
3195            L
1443            L
3513            L
2947            L
997             L
2908            L
```

[42]:
```
sns.barplot(company_level.job_title, y)

warnings.filterwarnings('ignore')
```



Large company pays more money to the data engineer in comparison with data scientist and data analyst. Least money goes to BI analyst

[43]:
```
import jovian
```

[44]:
```
jovian.commit()
```

```
<IPython.core.display.Javascript object>

[jovian] Updating notebook "pandeykiran571/data-science-salaries-analysis" on
https://jovian.com
[jovian] Committed successfully! https://jovian.com/pandeykiran571/data-science-
```

```
salaries-analysis
```

[44]:
```
'https://jovian.com/pandeykiran571/data-science-salaries-analysis'
```

1.4 Asking and Answering Question

1.4.1 Q1: Which job is the highest paid in all the category?

[80]:
```
df2= data_df.loc[data_df['salary_in_usd'].idxmax(), 'job_title']
print("highest paying job in all category is {}".format(df2))
```

```
highest paying job in all category is Data Scientist
```

1.4.2 Q2: Which job is most popular in data science?

[46]:
```
job = data_df.job_title.max()
print("Most popular job in data science is {}".format(job))
```

```
Most popular job in data science is Data Scientist
```

1.4.3 Q3: What post does Mid level company likes to offer?

[47]:
```
company_mid_level = c_l.get_group('M')
company_mid_level
offer = company_mid_level.job_title.max()
print("Mid Level company offer {} post more".format(offer))
```

```
Mid Level company offer Data Scientist post more
```

1.4.4 Q4: Which is the least popular job in data science?

[87]:
```
least = data_df.job_title.min()
print("Least popular post is {} ".format(least))
```

```
Least popular post is Data Analyst
```

1.4.5 Q5: What employment type do people perfer?

[90]:
```
df3 = data_df.employment_type.max()
print("People prefer {} the most ".format(df3))
```

```
People prefer FT the most
```

[91]:
```
import jovian
```

[92]:
```
jovian.commit()
```

```
<IPython.core.display.Javascript object>
```

```
[jovian] Updating notebook "pandeykiran571/data-science-salaries-analysis" on
https://jovian.com
[jovian] Committed successfully! https://jovian.com/pandeykiran571/data-science-
salaries-analysis
```

[92]: 'https://jovian.com/pandeykiran571/data-science-salaries-analysis'

1.5 Conclusion

Some results of analysis are: * Data Scientist is the most popular job in data science career * According to the sample data, data analyst is least popular job in data science * Data scientist are getting high salary * Most people likes to do full time job rather than doing freelancing or in contract basis.

[93]:
```
import jovian
```

[94]:
```
jovian.commit()
```

```
<IPython.core.display.Javascript object>
```

```
[jovian] Updating notebook "pandeykiran571/data-science-salaries-analysis" on
https://jovian.com
[jovian] Committed successfully! https://jovian.com/pandeykiran571/data-science-
salaries-analysis
```

[94]: 'https://jovian.com/pandeykiran571/data-science-salaries-analysis'

1.6 Future work

Small dataset was used in this analysis due to which result may vary while doing analysis in big dataset. Certain criteria was only focused while performing analysis.So, many other insights may be discovered in future by using big dataset with all the criteria.

[95]:
```
import jovian
```

[96]:
```
jovian.commit()
```

```
<IPython.core.display.Javascript object>
```

```
[jovian] Updating notebook "pandeykiran571/data-science-salaries-analysis" on
https://jovian.com
[jovian] Committed successfully! https://jovian.com/pandeykiran571/data-science-
salaries-analysis
```

[96]: 'https://jovian.com/pandeykiran571/data-science-salaries-analysis'

[ ]: