Step One: Set up your analysis and Jupyter Notebook

Import Libraries (Please ensure you have installed all the required dependencies)

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
#For presentable data
sns.set(color_codes=True)
import os
import operator
import matplotlib.pyplot as plt; plt.rcdefaults()
```

In [2]:
```
#List all the available files for this project
print(os.listdir())
```

```
['.ipynb_checkpoints', 'Data Analysis.ipynb', 'train_data.csv']
```

In [2]:
```
#List all the available files for this project
print(os.listdir())
```

```
['.ipynb_checkpoints', 'Data Analysis.ipynb', 'train_data.csv']
```

Insights: Read Data

In [3]:
```
career =pd.read_csv('train_data.csv')
```

In [4]:
```
career.head()
```

Out[4]:

	Year	Hobbyist	ConvertedComp	Country	EdLevel	Employment	JobSat	OrgSize	UndergradMajor	YearsCodePro	Data scientist or machine learning specialist	Database administrator	Data busin anal
0	2017	Yes, both	43750.00000	United Kingdom	Bachelor's degree	Employed full-time	4.0	2 to 9 employees	Computer science	2.0	1	1	N
1	2017	Yes, I program	51282.05128	Denmark	Some college/university study without	Employed part-time	10.0	100 to 499 employees	Computer science	3.0	1	0	N

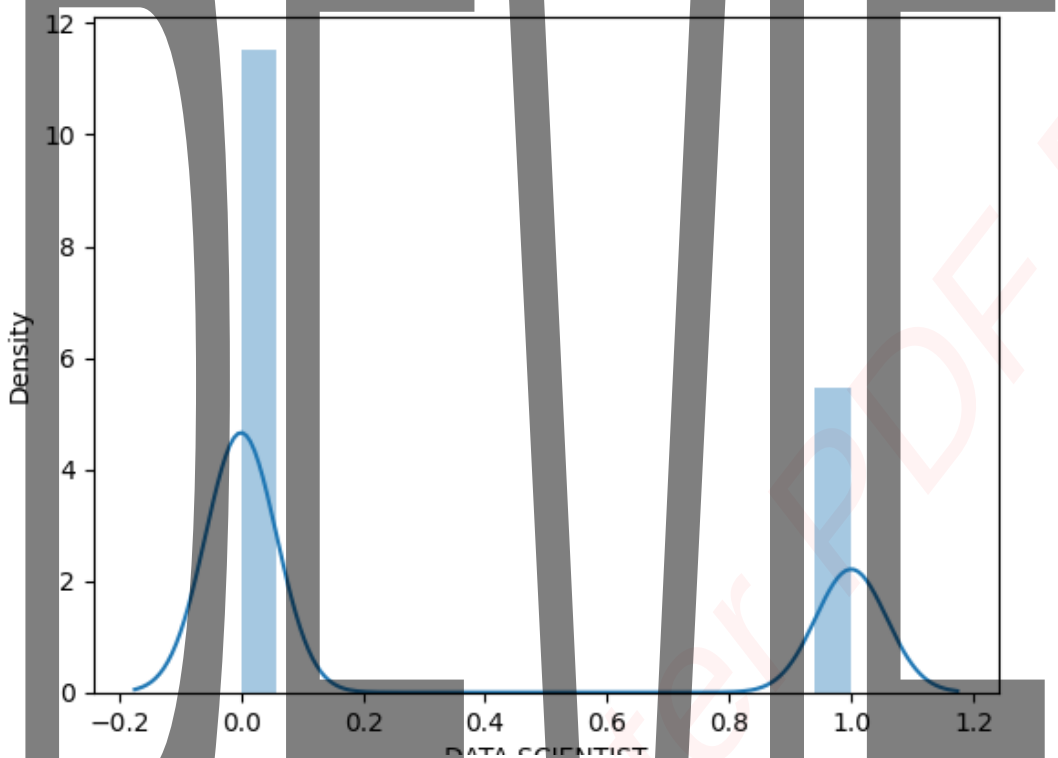
Step Three: Analyse your data (Exploratory Data Analysis)

Check all data size and data column

In [22]:
```
career_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33601 entries, 0 to 33600
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   YEAR            33601 non-null  int64
 1   HOBBY           33601 non-null  object
 2   SALARY          33601 non-null  float64
 3   COUNTRY         33601 non-null  object
 4   DATA SCIENTIST  33601 non-null  int64
 5   DB ADMIN        33601 non-null  int64
dtypes: float64(1), int64(3), object(2)
memory usage: 1.5+ MB
```

```
<AxesSubplot:xlabel='DATA SCIENTIST', ylabel='Density'>
```




```
Signature:
new_format.dropna(
    axis: 'Axis' = 0,
    how: 'str' = 'any',
    thresh=None,
    subset=None,
    inplace: 'bool' = False,
)
Docstring:
Remove missing values.

See the :ref:`User Guide <missing_data>` for more on which values are
considered missing, and how to work with missing data.

Parameters
----------
axis : {0 or 'index', 1 or 'columns'}, default 0
    Determine if rows or columns which contain missing values are
    removed.

    * 0, or 'index' : Drop rows which contain missing values.
    * 1, or 'columns' : Drop columns which contain missing value.

    .. versionchanged:: 1.0.0

       Pass tuple or list to drop on multiple axes.
       Only a single axis is allowed.

how : {'any', 'all'}, default 'any'
    Determine if row or column is removed from DataFrame, when we have
    at least one NA or all NA.

    * 'any' : If any NA values are present, drop that row or column.
    * 'all' : If all values are NA, drop that row or column.

thresh : int, optional
    Require that many non-NA values.
subset : array-like, optional
    Labels along other axis to consider, e.g. if you are dropping rows
    these would be a list of columns to include.
inplace : bool, default False
    If True, do operation inplace and return None.
```

In [24]:
```
career_data.head()
```

Out[24]:

	HOBBY	SALARY	COUNTRY	STATUS	DATA SCIENTIST	DB ADMIN
0	Yes, both	43750.00000	United Kingdom	Developed	1	1
1	Yes, I program as a hobby	51282.05128	Denmark	Developed	1	0
2	No	25000.00000	Israel	Developed	1	0
3	Yes, I program as a hobby	100000.00000	United States	Developed	0	1
4	Yes, both	27000.00000	Ukraine	Developed	0	1

In [25]:
```
cdi = career_data.groupby('STATUS')
```

In [ ]:
```
```

In [26]:
```
cdi
```

Out[26]:
```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x000001E7CD825B50>
```

In [27]:
```
for STATUS, STATUS_career_data in cdi:
    print(STATUS)
    print(STATUS_career_data)
```

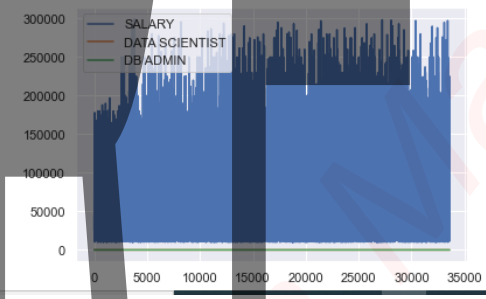
In [29]:
```
cdi.mean()
```

Out[29]:

STATUS	SALARY	DATA SCIENTIST	DB ADMIN
Developed	69387.730203	0.325577	0.529545
Underdeveloped	5292.665513	0.288602	0.629141

In [30]:
```
%matplotlib inline
cdi.plot()
```

Out[30]:
```
STATUS
Developed          AxesSubplot(0.125,0.125;0.775x0.755)
Underdeveloped     AxesSubplot(0.125,0.125;0.775x0.755)
dtype: object
```



In [32]:
```
career_data.head()
```

Out[32]:

	HOBBY	COUNTRY
0	Yes, both	United Kingdom
1	Yes, I program as a hobby	Denmark
2	No	Israel
3	Yes, I program as a hobby	United States
4	Yes, both	Ukraine

In [33]:
```
cdsat = career_data.groupby('HOBBY')
```

In [34]:
```
cdsat
```

Out[34]:
```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x000001E7D2ADD250>
```

In [35]:
```
for HOBBY, HOBBY_career_data in cdsat:
    print(HOBBY)
    print(HOBBY_career_data)
```

```
No
      HOBBY        COUNTRY
2        No         Israel
6        No         Canada
7        No         Canada
```

In [36]:
```
my_data = cdsat.count()
```

In [37]:
```
my_data
```

Out[37]:

	COUNTRY
HOBBY	
No	6218
Yes	25225
Yes, I contribute to open source projects	165
Yes, I program as a hobby	1210
Yes, both	783

In [38]:
```
#%matplotlib inline
my_data.plot.bar(rot=60)
```

Out[38]:
```
<AxesSubplot:xlabel='HOBBY'>
```

In [38]:
```
#%matplotlib inline
my_data.plot.bar(rot=60)
```

Out[38]:
```
<AxesSubplot:xlabel='HOBBY'>
```