Term Project Data Mining - Gender Pay Gap Analysis

Milestone 1 - Week 6

Create a Graphical Analysis creating a minium of four grouphs. Label your graphs appropriately and explain/analyze provided by each graph. Your analysis should begin to answer the question(s) you are addressing. Write a short overview/conclusion of the insights gained from your graphical anaylsis.

In [1]:

```
# import the data set using necesarry libraries
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt

# read csv Glassdoor Gender Pay Gap
df_pay = pd.read_csv("Glassdoor Gender Pay Gap.csv")
df_pay.head()
```

Out[1]:

	JobTitle	Gender	Age	PerfEval	Education	Dept	Seniority	BasePay	Bonus
0	Graphic Designer	Female	18	5	College	Operations	2	42363	9938
1	Software Engineer	Male	21	5	College	Management	5	108476	11128
2	Warehouse Associate	Female	19	4	PhD	Administration	5	90208	9268
3	Software Engineer	Male	20	5	Masters	Sales	4	108080	10154
4	Graphic Designer	Male	26	5	Masters	Engineering	5	99464	9319

In [2]:

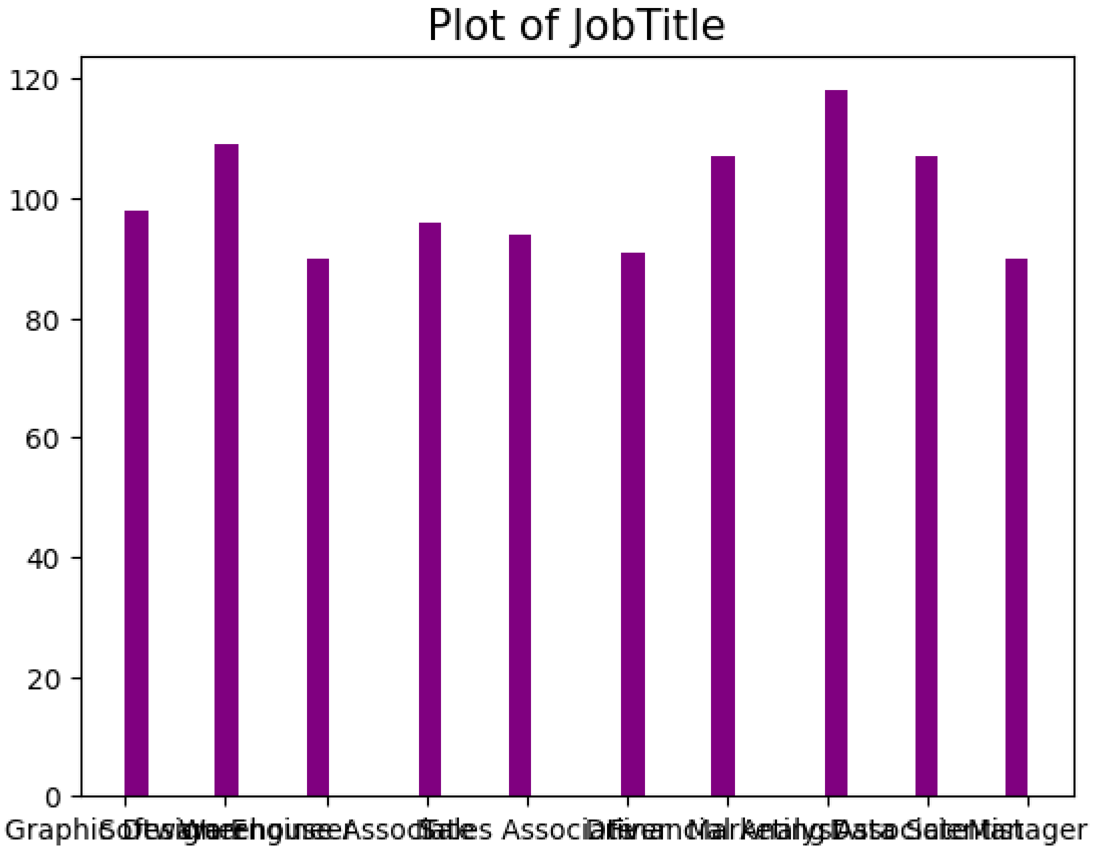
```
# find total number of records in csv by (rows, columns)
df_pay.shape
```

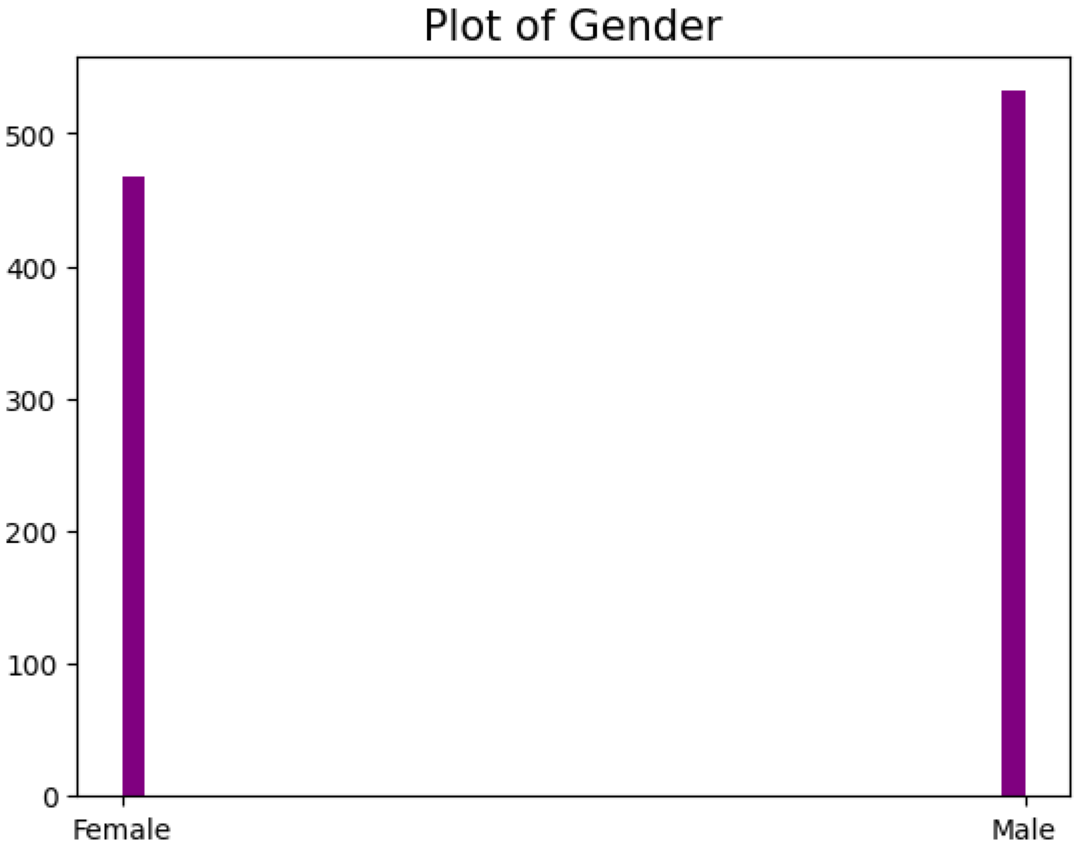
Out[2]:

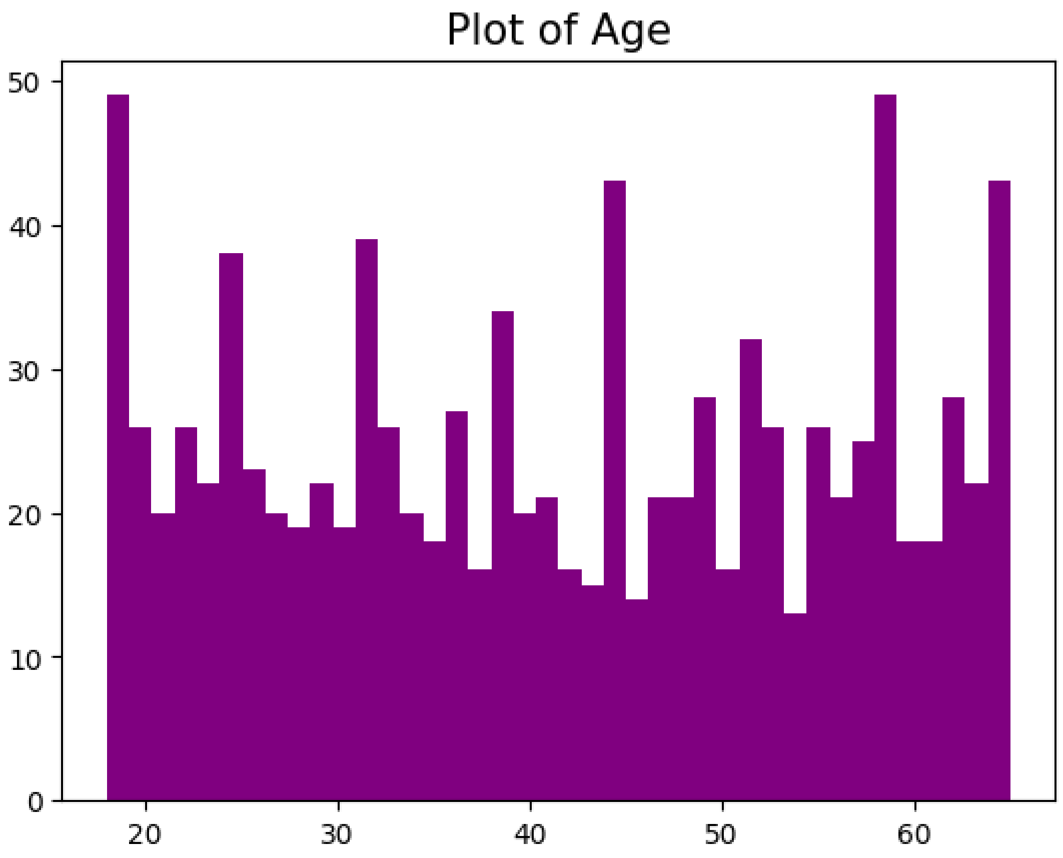
```
(1000, 9)
```

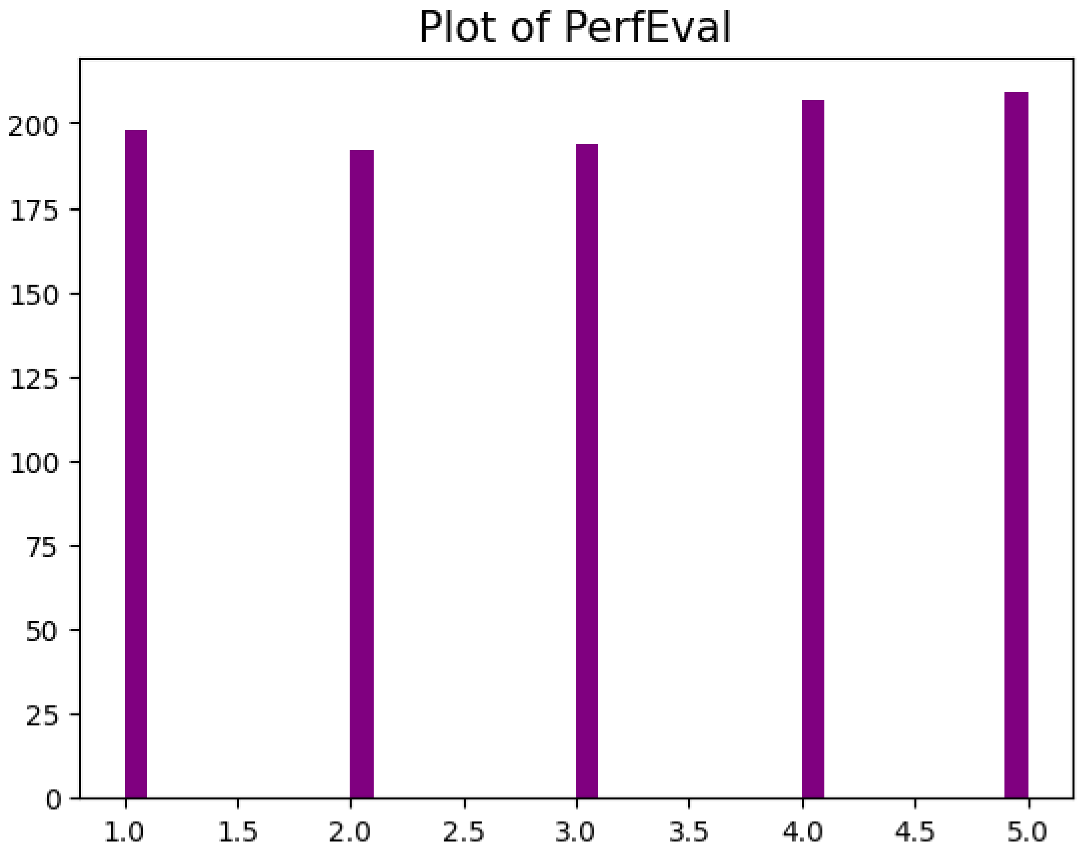
In [3]:

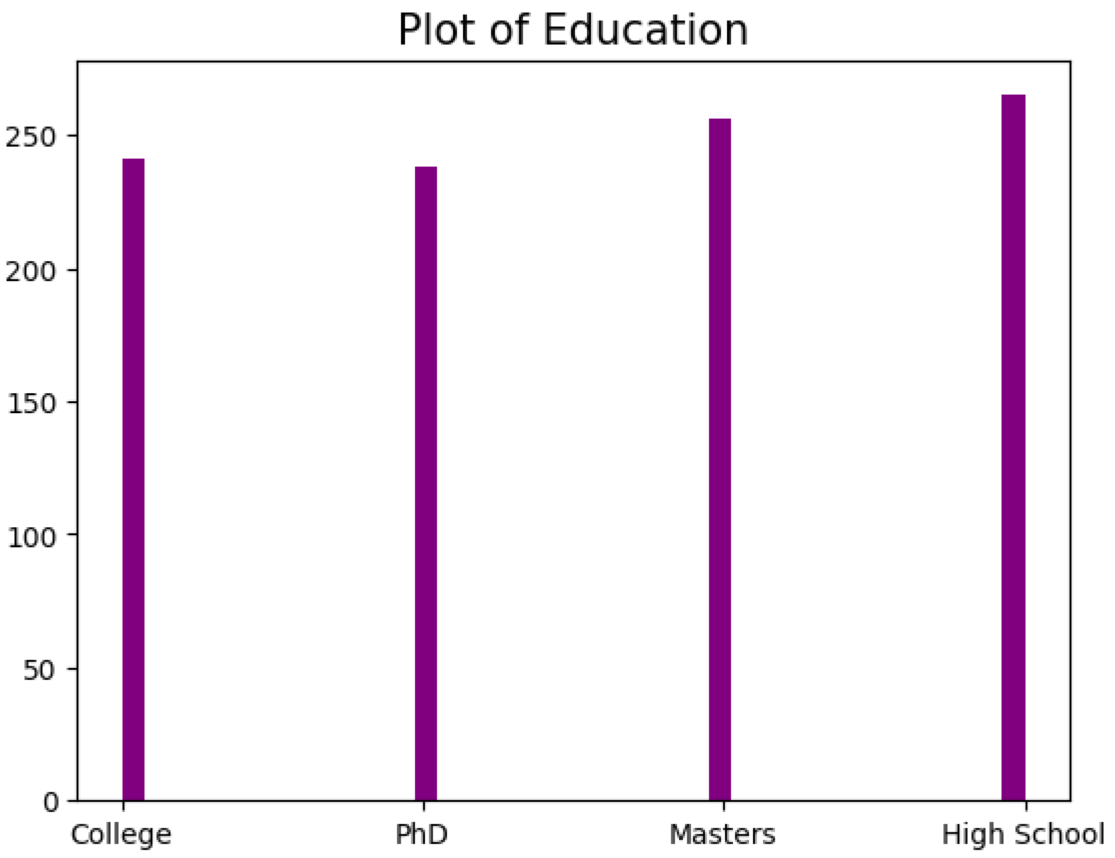
```
for c in df_pay.columns:
    plt.title("Plot of "+c,fontsize=15)
    plt.hist(df_pay[c],bins=40,color='purple')
    plt.show()
```

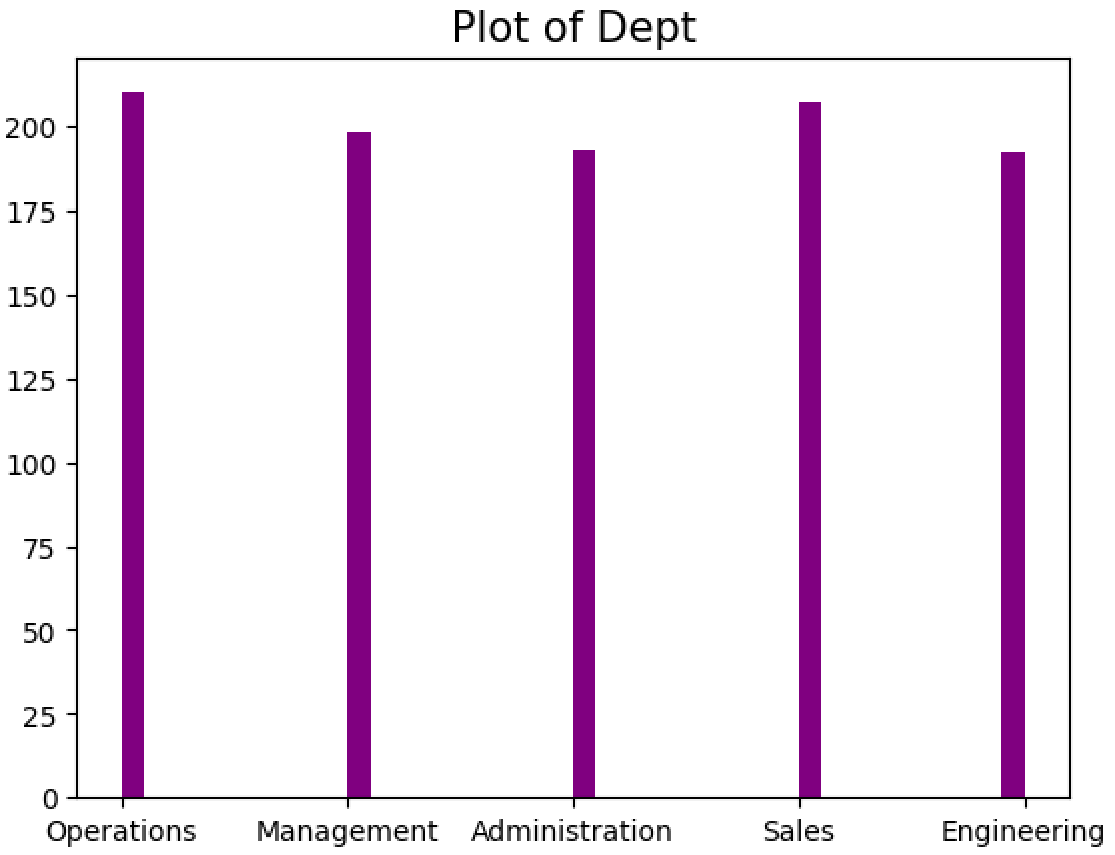

Plot of JobTitle

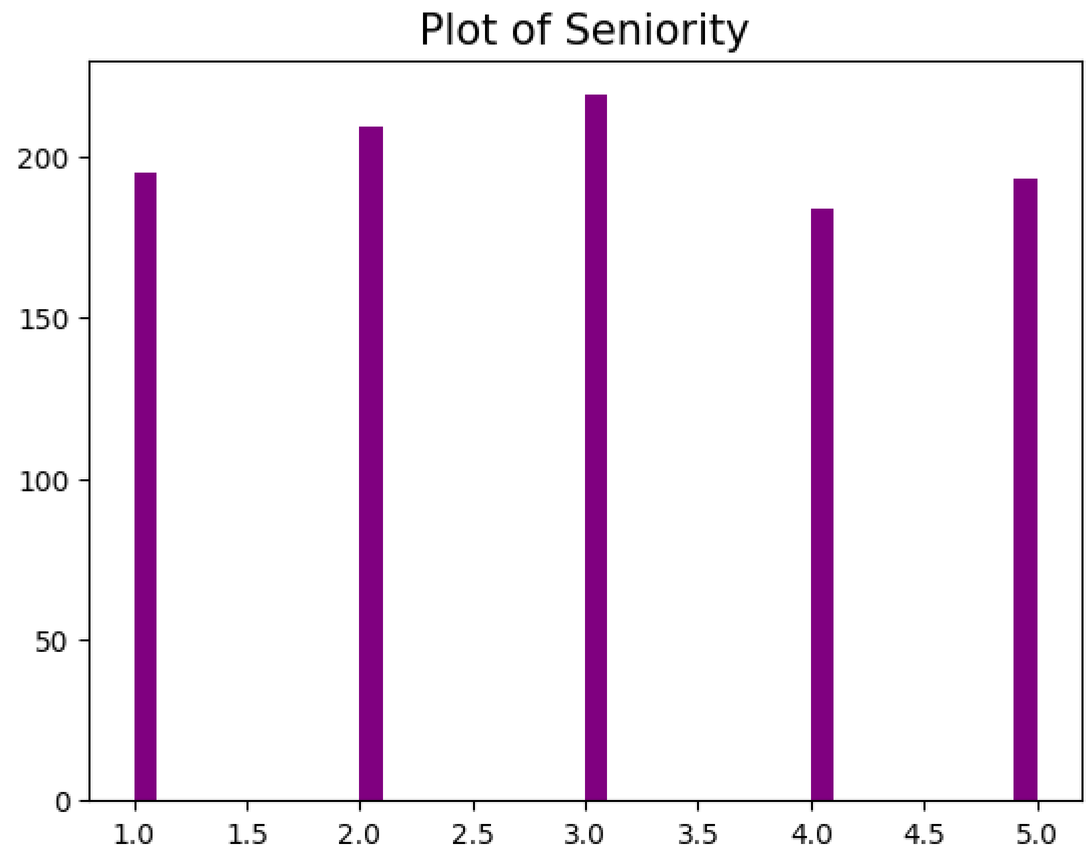

Plot of Gender

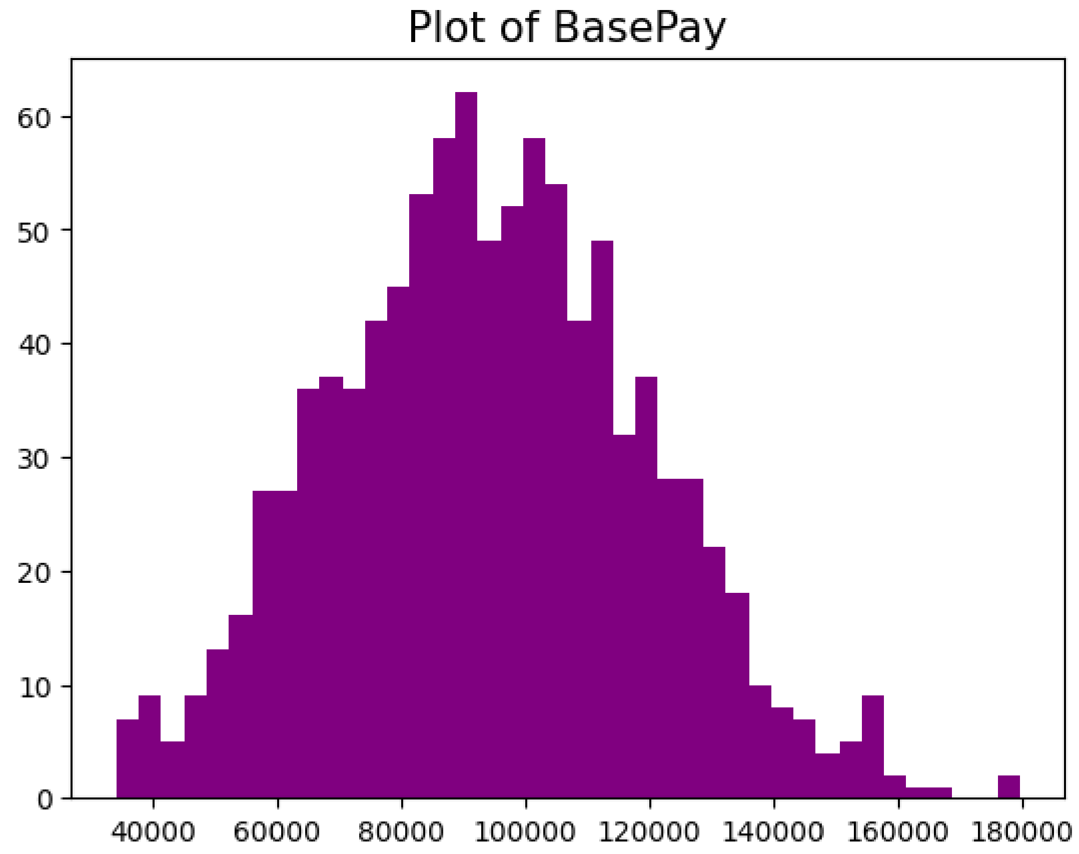

Plot of Age
0
10
20
30
40
50
20
30
40
50
60

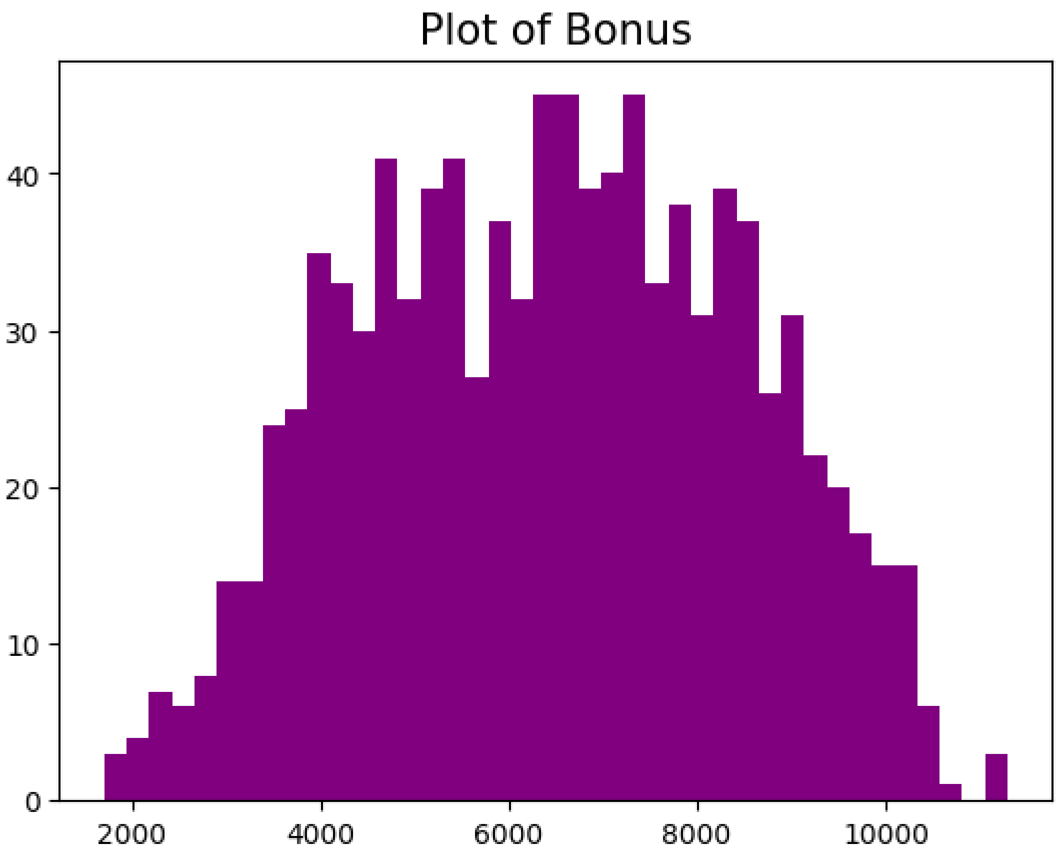

Plot of PerfEval
0
25
50
75
100
125
150
175
200
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Plot of Education
250
200
150
100
50
0
College
PhD
Masters
High School


Plot of Dept
200
175
150
125
100
75
50
25
0
Operations
Management
Administration
Sales
Engineering


Plot of Seniority
0
50
100
150
200
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

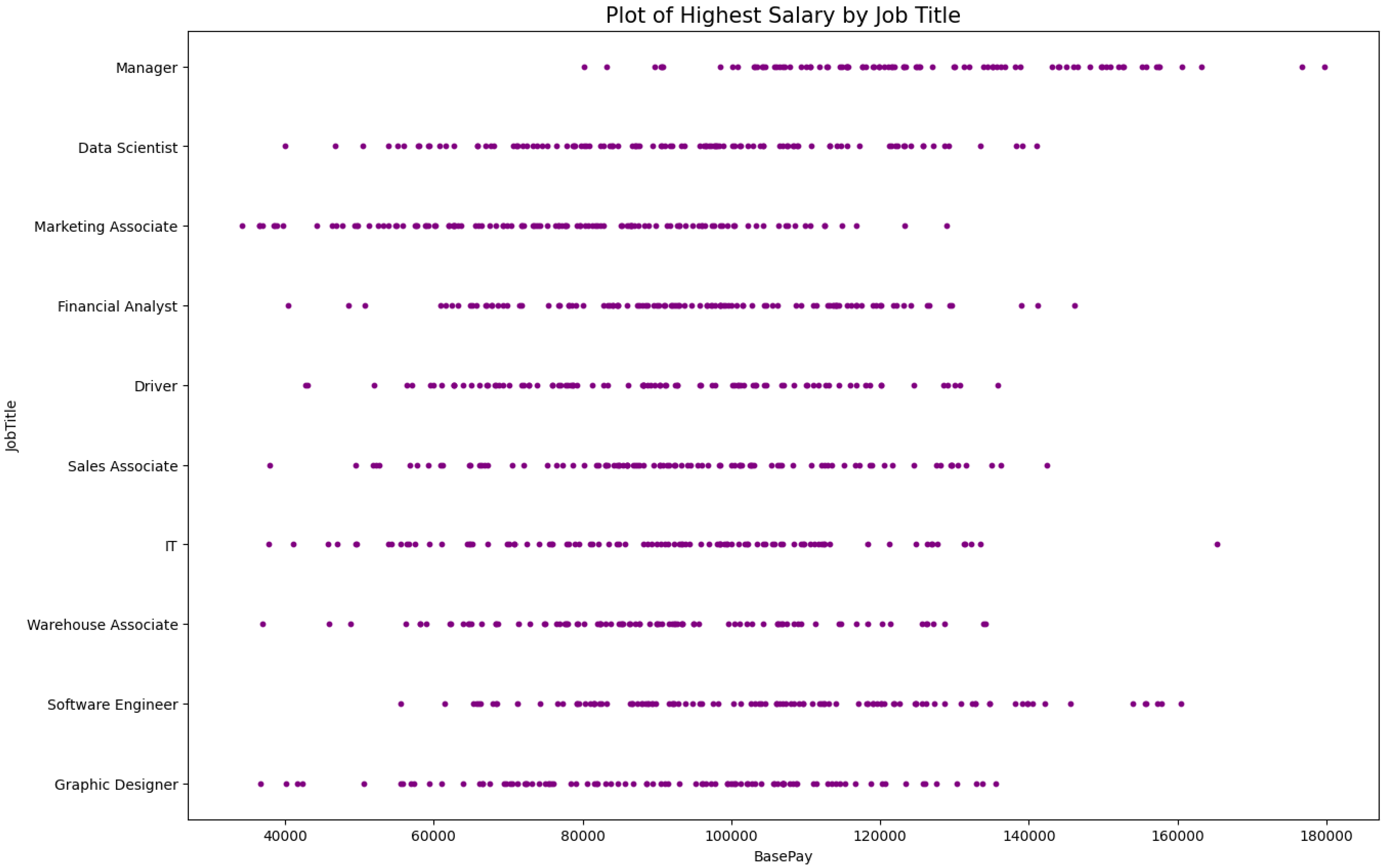

Plot of BasePay
0
10
20
30
40
50
60
40000
60000
80000
100000
120000
140000
160000
180000



Which job title had the highest salary?

In [4]:
```
plt.figure(figsize =(15,10))
plt.title('Plot of Highest Salary by Job Title',fontsize=15)
plt.scatter(df_pay['BasePay'], df_pay['JobTitle'], s=10, color ='purple')
plt.xlabel('BasePay')
plt.ylabel('JobTitle')
plt.plot()
```

Out[4]:
```
[]
```

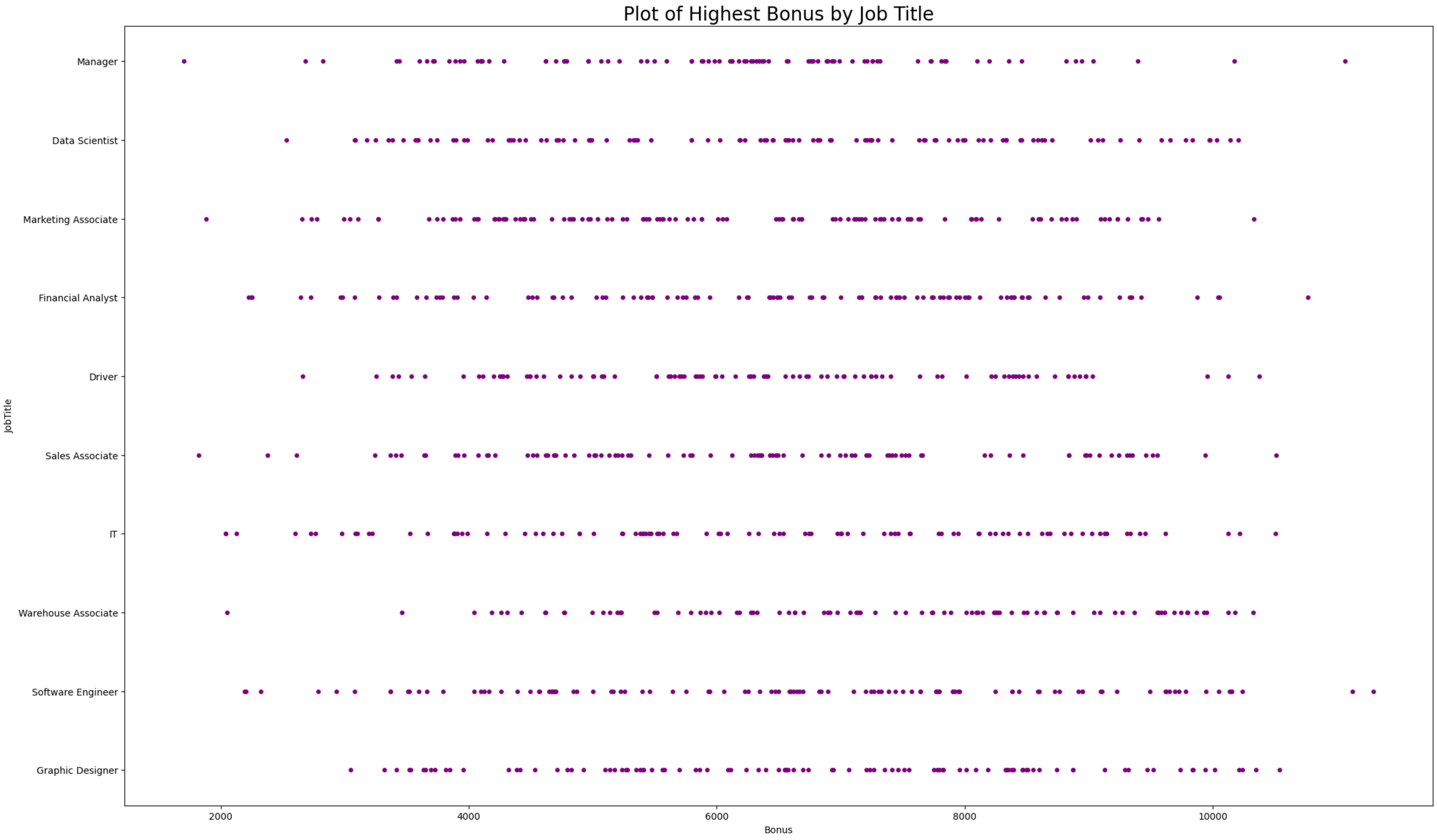


From the scatter plot above, I can see that the Highest Salary is of a Manager that is ranging around 180,000.

Which job had the highest bonus? Was it the same title as the highest salary?

In [5]:

```
# CORRECTED - changed from histogram to scatter plot to show visualization better.
plt.figure(figsize =(25,15))
plt.title('Plot of Highest Bonus by Job Title',fontsize=20)
plt.scatter(df_pay['Bonus'], df_pay['JobTitle'], s=15, color ='purple')
plt.xlabel('Bonus')
plt.ylabel('JobTitle')
plt.show()
```

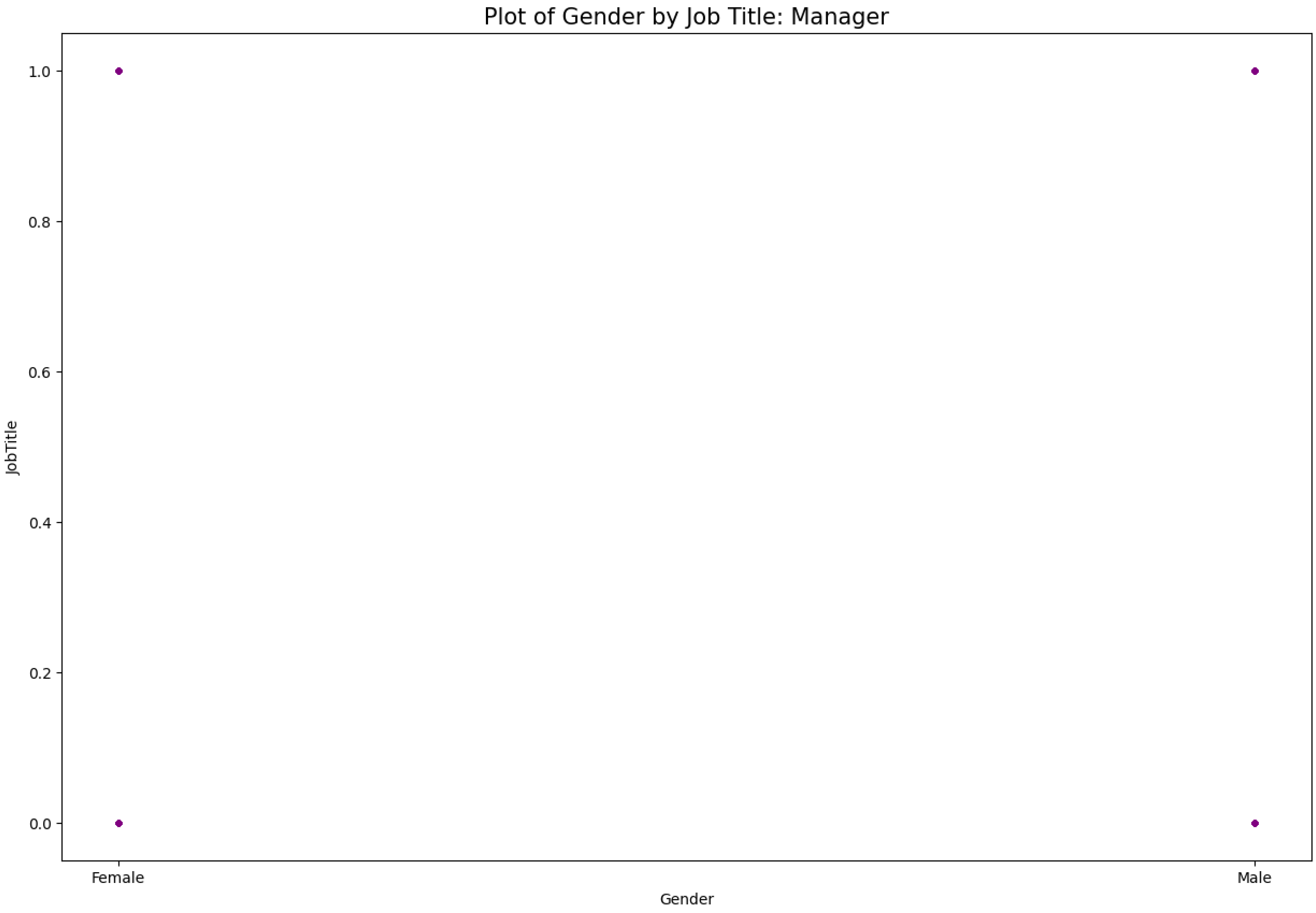


CORRECTED -From the scatterplot above, it looks as thought the title that has the highest bonus is the Software Engineer with a bonus of 11,000 that a Manager shows.

Out of the highest salary and bonus, which gender reflected that salary?

```
In [6]: # CORRECTED - changed from histogram to scatter plot to show visualization better.
plt.figure(figsize =(15,10))
plt.title('Plot of Gender by Job Title: Manager',fontsize=15)
plt.scatter(df_pay['Gender'], df_pay['JobTitle']=='Manager', s=10, color ='purple')
plt.xlabel('Gender')
plt.ylabel('JobTitle')
plt.plot()
```

```
Out[6]: []
```

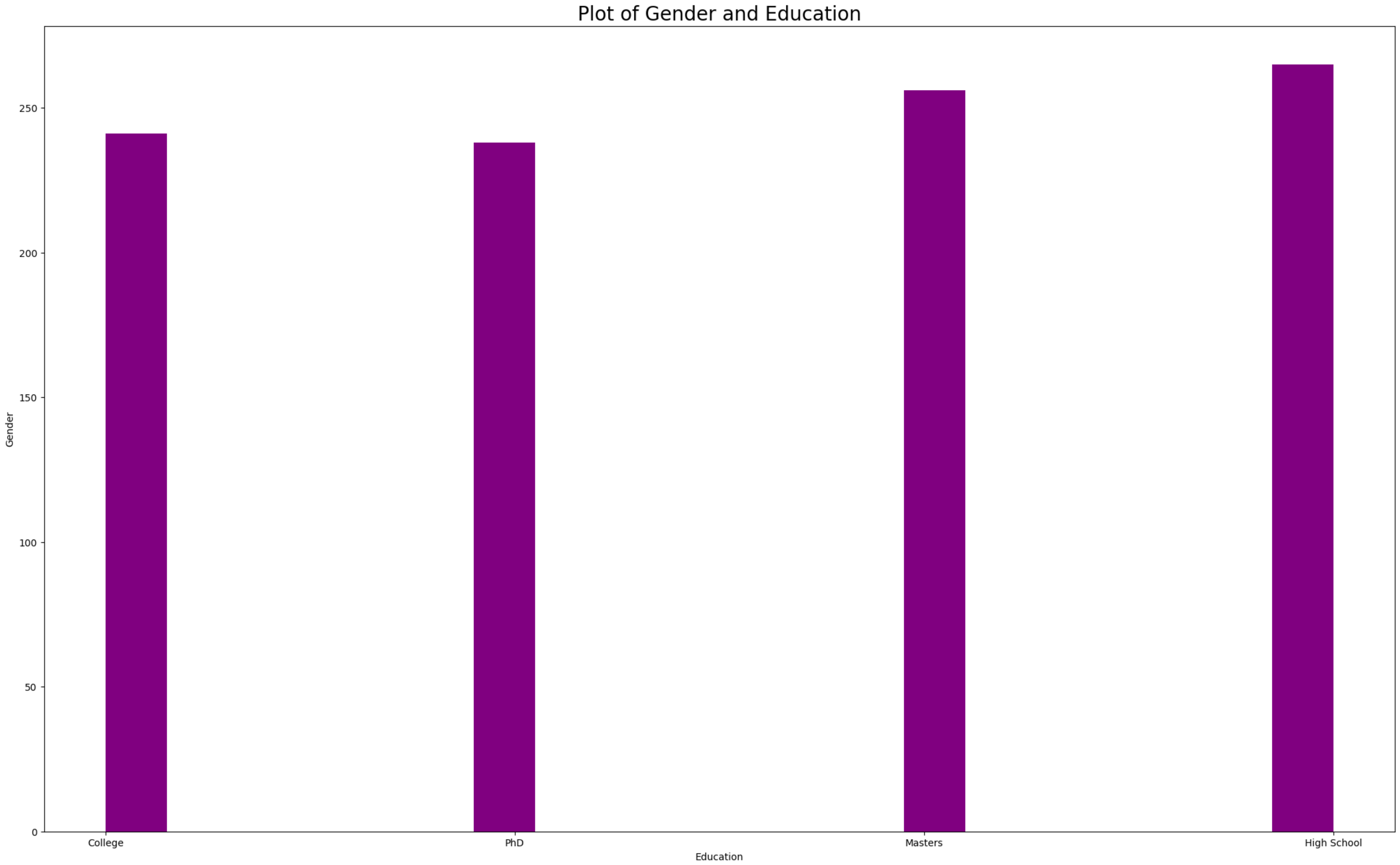


CORRECTED -From the scatterplot above, it looks as though since the Manager had the highest salary, and that both genders where considered for the Manager position. Though the graph doesn't show exactly what gender reflected the salary of 180,000.

Did the opposite gender have the same schooling as the gender that had the highest salary?

In [7]:

```
plt.figure(figsize =(25,15))
plt.title('Plot of Gender and Education',fontsize=20)
plt.hist(df_pay['Education'],bins=20,color='purple')
plt.xlabel('Education')
plt.ylabel('Gender')
plt.show()
```



CORRECTED- I was able to correct all of my graphs to show more of what I was looking for with my questions. However, With the last one, it does show the range for the education degrees, but still couldn't figure out how to pull which gender had which education using matplotlib as my visualization source.