Goal: Find out if there any evidence of systematic bias in how women vs men are paid in this dataset

In [1]:

```
import numpy as np
import pandas as pd
```

In [2]:

```
empdf = pd.read_csv('employee_data.csv')
```

In [3]:

```
empdf.head()
```

Out[3]:

	employee_id	role	gender	level	years_of_experience	compensation
0	1	Hardware Engineer	M	Junior	2.0	16.79
1	2	Hardware Engineer	F	Staff	12.5	30.00
2	3	Hardware Engineer	M	Junior	9.6	15.71
3	4	Hardware Engineer	F	Senior	11.6	21.54
4	5	Hardware Engineer	M	Principal	8.7	35.80

In [6]:

```
empdf['employee_id'].count()
```

Out[6]:

```
2000
```

In [7]:

```
empdf['gender'].value_counts()
```

Out[7]:

```
M    1495
F     505
Name: gender, dtype: int64
```

We have a total of 2000 employees - Out of which 1495 are Males and just 505 are Females.

Highlight 1: There seems to be a 3:1 ratio for Male to Female employees.

In [8]:

```
empdf[['years_of_experience','compensation']].mean()
```

Out[8]:

```
years_of_experience     6.41125
compensation           21.19294
dtype: float64
```

In [9]:

```
empdf.groupby(['gender'])[['years_of_experience','compensation']].mean()
```

Out[9]:

	years_of_experience	compensation
gender		
F	6.543960	19.825485
M	6.366421	21.654856

Next we see the comparision of average 'Compensation' and 'Years of experience' figures of Males and Females. In terms of Number of years of experiences there is no high deviation from the overall average of years of experience of the workforce.

Highlight 2: For females the average compensation numbers is less than the overall average.

The average female compensation is less than 8.41% (|19.83-21.65|/21.65) of average male compensation.

I am now testing the hypothesis that Males average Salary is greater than Females average salary.

Will be performing Two Sample - One Tail T-Test

Ho: Mean average salary for men <= Mean average salary of females

Ha: Mean average salary for men > Mean average salary of females

In [10]:

```
from scipy import stats

female_compensation = empdf[empdf['gender'] == 'F']['compensation']
male_compensation = empdf[empdf['gender'] == 'M']['compensation']
```

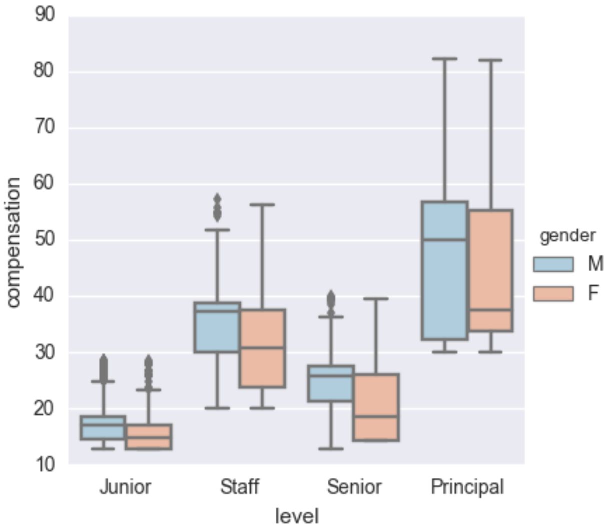
```
In [11]: stats.ttest_ind(male_compensation, female_compensation)

Out[11]: Ttest_indResult(statistic=3.6533400872963071, pvalue=0.00026552148611486777)
```

The T-Statistic value is 3.65. As the T-Statistic is greater than 1.645 (for 95% confidence interval), infact greater than 2.33 (for 99% confidence interval) we reject the Null Hypothesis (Male Salary is <= Female Salary)

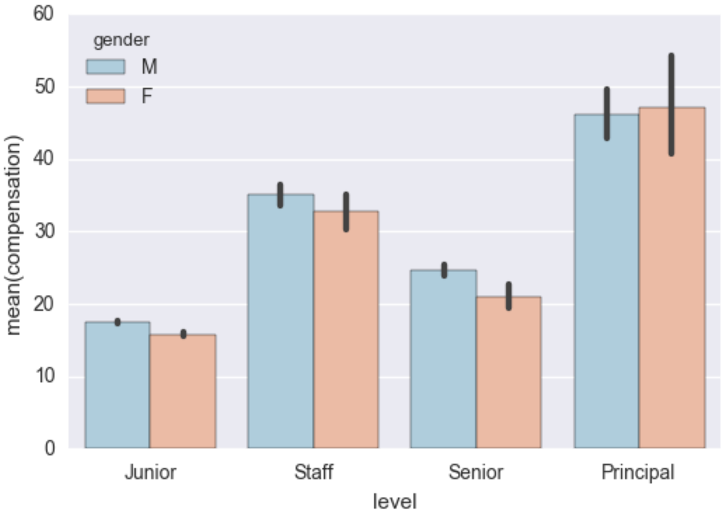
Highlight 3: With 99% confidence we can say that Mean Male Salary > Mean Female Salary

```
In [12]: from matplotlib import pyplot
         import seaborn as sns
         %matplotlib inline
```

```
In [13]: sns.factorplot(x="level", hue="gender", y="compensation", data=empdf, kind="box", palette= "RdBu_r")

Out[13]: <seaborn.axisgrid.FacetGrid at 0x2b49d508438>
```



```
In [14]: sns.barplot(x='level',hue='gender',y='compensation',data=empdf,estimator=np.mean,palette= "RdBu_r")

Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x2b49e6f8a58>
```



Highlight 4: If we look at boxplots and barplot of compensations broken down by gender and levels, we see that the interquartile range for women, for all the levels, is lower than that of men. In addition, we can see lower mean for females across all levels except Principal level - for which the mean is almost same for both males and females.

In [15]:
```
from sklearn.linear_model import LinearRegression

femaleempdf = empdf[empdf['gender']=='F']
maleempdf = empdf[empdf['gender']=='M']

X = femaleempdf['years_of_experience']
y = femaleempdf['compensation']
Female_lm = LinearRegression()
X=X.reshape(-1, 1)
Female_lm.fit(X,y)
X = maleempdf['years_of_experience']
y = maleempdf['compensation']
Male_lm = LinearRegression()
X=X.reshape(-1, 1)
Male_lm.fit(X,y)

diff=((Female_lm.coef_-Male_lm.coef_)/Male_lm.coef_)*100
print("For Females - with every year increase in experience the compensation increases by: {one} units of Salar
y".format(one=round(Female_lm.coef_[0],3)))
print("For Males   - with every year increase in experience the compensation increases by: {one} units of Salar
y".format(one=round(Male_lm.coef_[0],3)))
print("The difference between increase in Compensation for Females as compared to that of Males is: {one}%".form
at(one=round(diff[0],3)))
```

```
For Females - with every year increase in experience the compensation increases by: 1.022 units of Salary
For Males   - with every year increase in experience the compensation increases by: 1.048 units of Salary
The difference between increase in Compensation for Females as compared to that of Males is: -2.442%
```

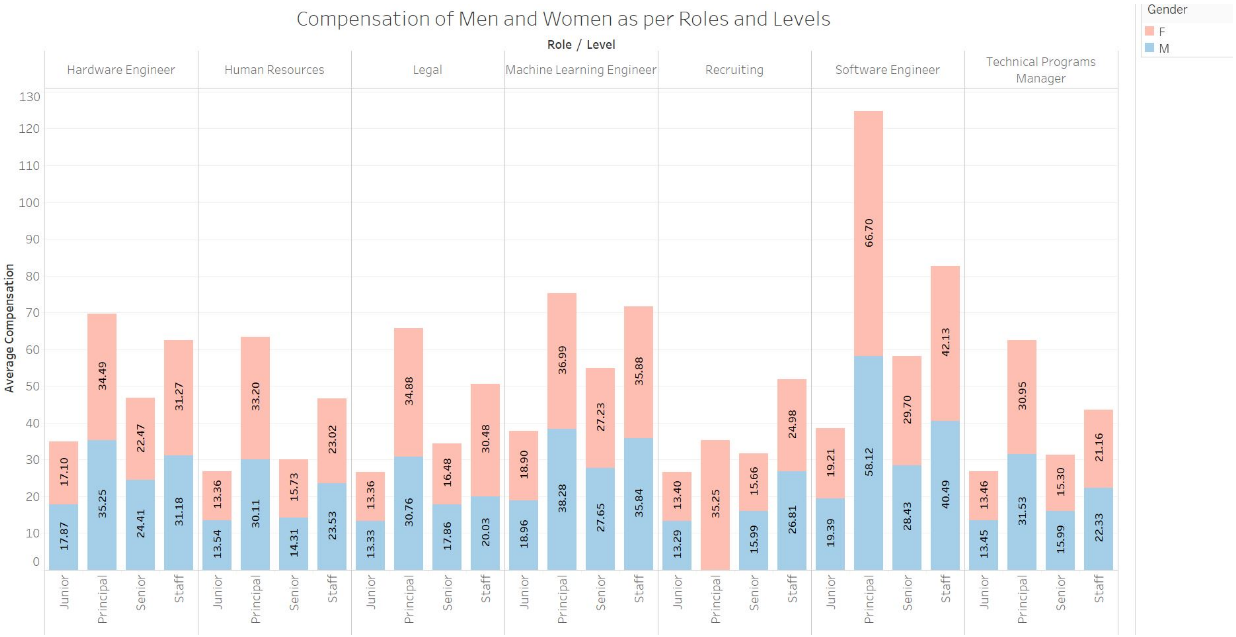
Highlight 5: With every year increase in experience Men are paid slightly higher than their Female counterparts, but the difference is not much. Note: For this analysis we have only considered years of experience as factor, there are other factors that we have not considered like effort hours, performance.

Now I am trying to find Gender pay differences as per Different Roles -

Goal is to check if there is any Role specific Pay bias.

In [16]:
```
from IPython.display import Image
Image(filename='CruiseAutomation-Gender-Role-Level-Compensation.JPG')
```

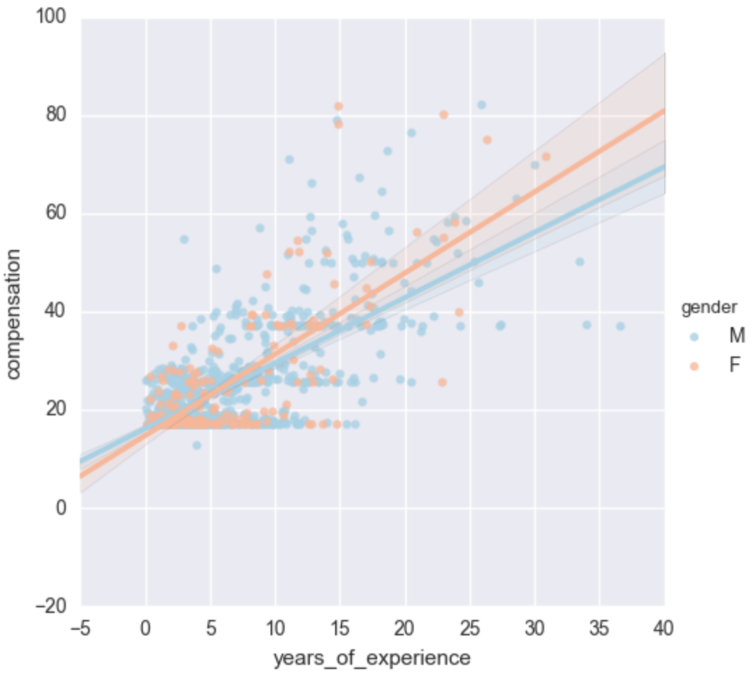
Out[16]:



The above graph was prepared in Tableau

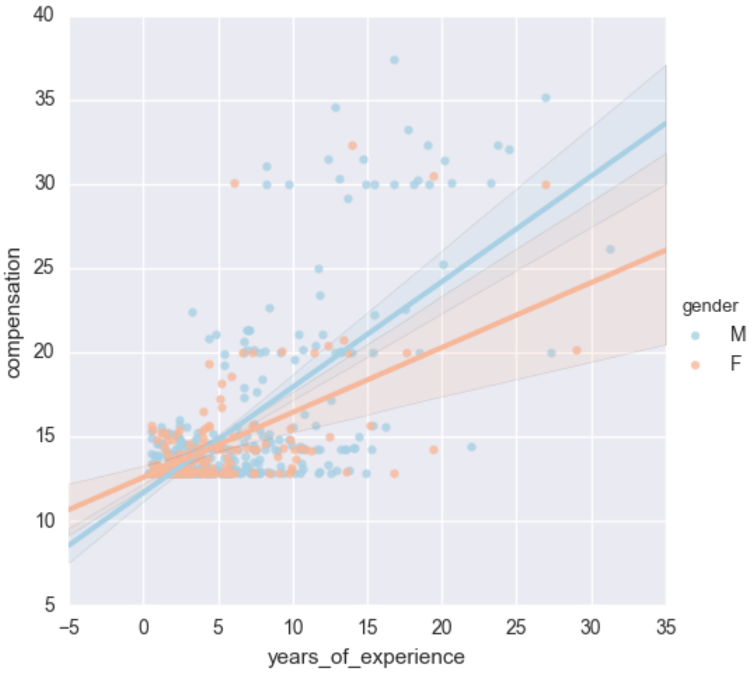
Highlight 6: The above graph shows us these important points: For Legal roles - Women at Staff level and at Principal levels are paid 52% and 13% higher than their Men counterparts. For Human Resources role - Women at Senior and Principal levels are paid 10% higher than their Men counterparts. For Software Engineer role - Women at Principal levels are paid 15% higher than their Men counterparts.

```
In [17]: sns.lmplot(x='years_of_experience',y='compensation',hue='gender',hue_order=['M','F'],data=empdf[empdf['role']=='Sof
         tware Engineer'],palette= "RdBu_r")

Out[17]: <seaborn.axisgrid.FacetGrid at 0x2b49e6fce10>
```



For Software Engineering roles - We see as experience increases Females are slightly paid higher than Males.

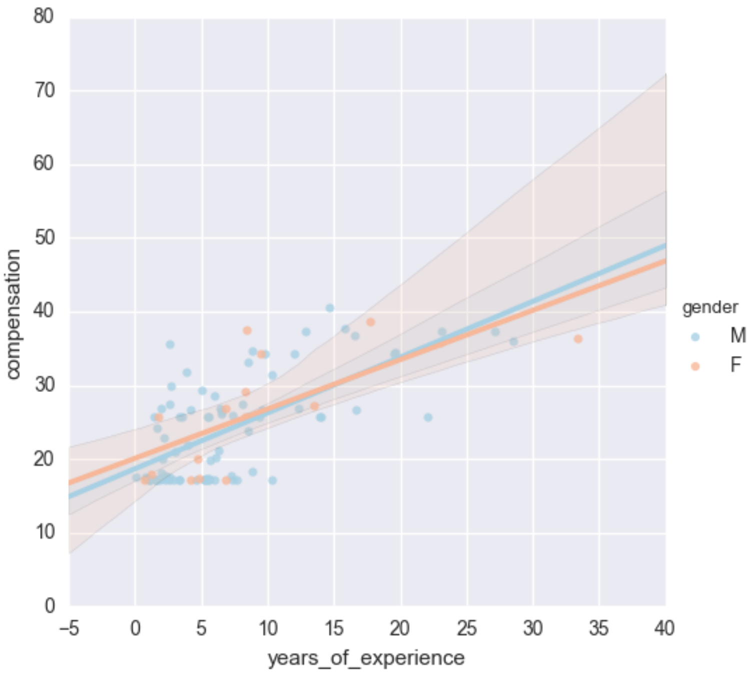
```
In [18]: sns.lmplot(x='years_of_experience',y='compensation',hue='gender',hue_order=['M','F'],data=empdf[empdf['role']=='Tec
         hnical Programs Manager'],palette= "RdBu_r")

Out[18]: <seaborn.axisgrid.FacetGrid at 0x2b49a2f9470>
```



For Technical Programs Manager roles - We see as experience increases Males are paid higher than Females and the difference between the pay increases with increase in experience

```
In [19]: sns.lmplot(x='years_of_experience',y='compensation',hue='gender',hue_order=['M','F'],data=empdf[empdf['role']=='Mac
         hine Learning Engineer'],palette= "RdBu_r")
```

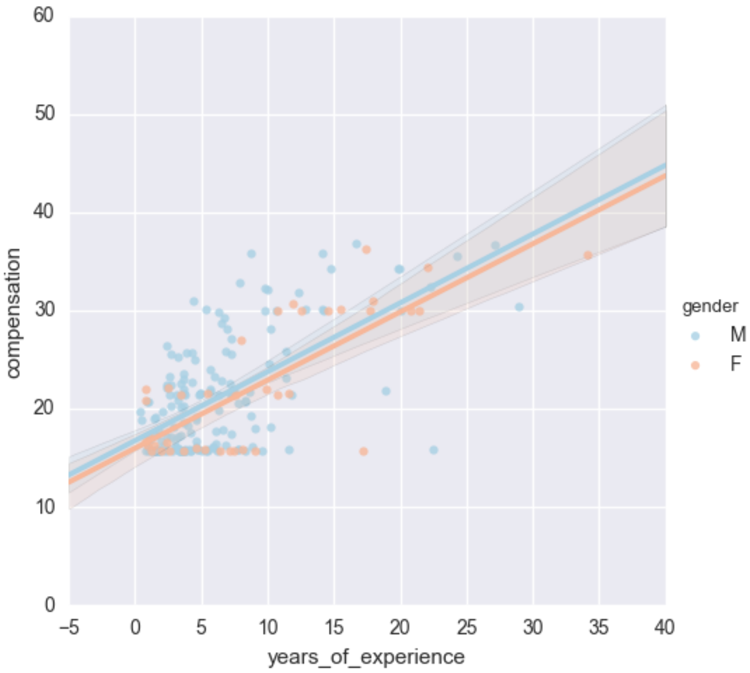
```
Out[19]: <seaborn.axisgrid.FacetGrid at 0x2b49ed51668>
```



For Machine Learning Engineer roles - We do not see a difference in compensations for Males and Females with increase in experience. Possible reason could be Machine Learning is highly sought skills and compensation is generally high irrespective of Gender

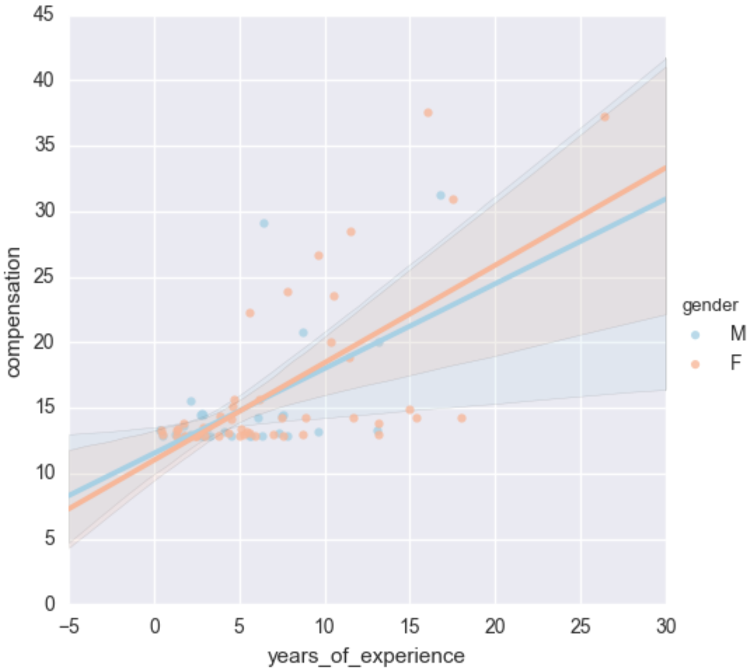
```
In [20]: sns.lmplot(x='years_of_experience',y='compensation',hue='gender',hue_order=['M','F'],data=empdf[empdf['role']=='Har
         dware Engineer'],palette= "RdBu_r")
```

```
Out[20]: <seaborn.axisgrid.FacetGrid at 0x2b49f4f12e8>
```



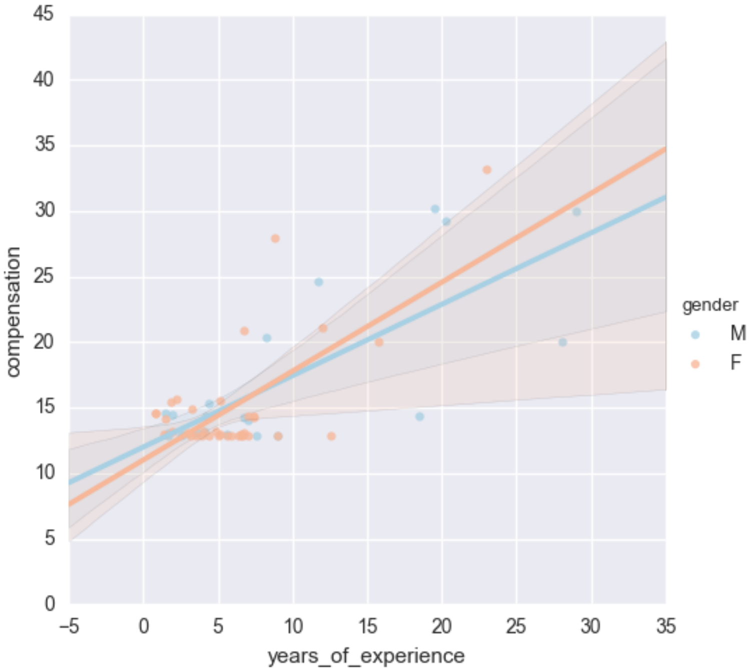
For Hardware Engineer roles - We do not see a difference in compensations for Males and Females with increase in experience.

```
In [21]: sns.lmplot(x='years_of_experience',y='compensation',hue='gender',hue_order=['M','F'],data=empdf[empdf['role']=='Rec
         ruiting'],palette= "RdBu_r")

Out[21]: <seaborn.axisgrid.FacetGrid at 0x2b49f217b38>
```



For Recruiting roles - We see after 12 years of experience, as experience increases Females are slightly paid higher than Males. Possible reason for this trend is we do not see lot of Males in this role for 10+ years of experience

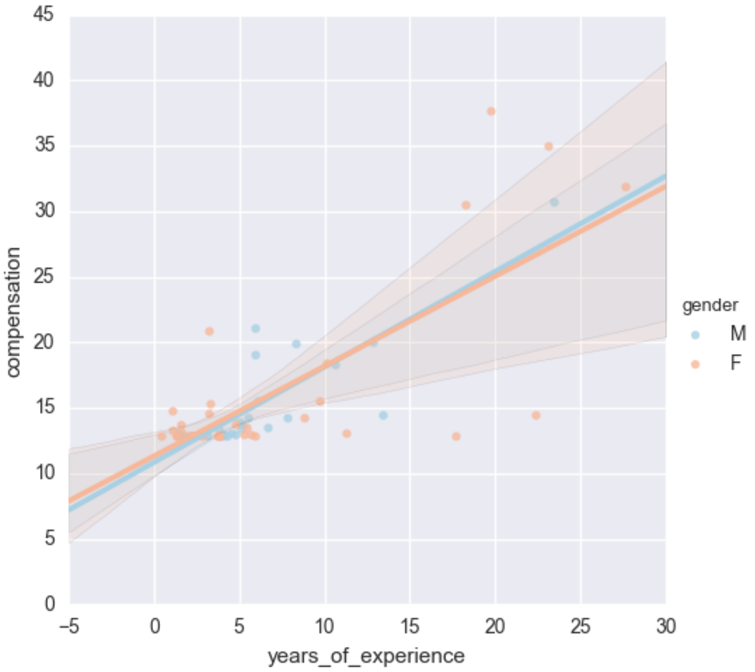
```
In [22]: sns.lmplot(x='years_of_experience',y='compensation',hue='gender',hue_order=['M','F'],data=empdf[empdf['role']=='Hum
         an Resources'],palette= "RdBu_r")

Out[22]: <seaborn.axisgrid.FacetGrid at 0x2b49f5292e8>
```



For Human Resources roles - We see after 12 years of experience, as experience increases Females are slightly paid higher than Males.

```
In [23]: sns.lmplot(x='years_of_experience',y='compensation',hue='gender',hue_order=['M','F'],data=empdf[empdf['role']=='Leg
         al'],palette= "RdBu_r")
```

```
Out[23]: <seaborn.axisgrid.FacetGrid at 0x2b49ed6eb00>
```



For Legal roles - We do not see a difference in compensations for Males and Females with increase in experience.

Conclusion

After analysis what I found is:

1. Number of Male employees are much higher than Female employees: In this data set we see a ratio of 3:1 in favour of Males. Need investigation on Hiring process to check if there is some sort of bias in Hiring process. It could also be a case where supply of Male employees is higher than females and thus this 3:1 could be justified.
2. Our hypothesis that Males are paid higher than Females stands justified - as with 99% confidence we can say that Mean Male Compensation is greater than Mean Female Compensation.
3. With every year increase in experience, Men are paid higher than Women, but the difference is just 2.5%
4. Further investigating to find out Bias in Compensation numbers as per different Roles, we do not see any specific trend in most of the roles. The only Role that stands-out is Technical Programs Manager, in which we see Males being paid more than Females with similar increase in experience.