CASE-STUDY-USING LINEAR REGRESSION

BANK GENDER DISCRIMINATION

The Following Analysis has been performed on the data:

1. Univariate Analysis:

Checked the count of the values using bar charts in the categorical columns and for numerical columns checked its distribution

Insights:

Got insights about the distribution of the data and count of values in the given categorical columns

2. Bivariate Analysis:

Conducted an analysis of each independent variable with the salary using bar charts and scatter plots (for two continuous numerical columns).

Insights:

1. The average salary of employees rose with the increase in education level

2. With the Increase in the grade of the employees there is an increase in the average salary. ie. The higher was the grader higher the average salary.

3. There is a positive correlation between years 1 and salary but not strong enough. correlation between salary and years1 is 0.61.

4. There is very less positive correlation between age and salary.

5. The average salary for males was higher than for females.

6. There isn't much difference is salary whether there is PC job or not

3. Multivariate Analysis:

Performed multivariate analysis using pivoted tables and also analyzed the data by segregating the data into male and female.

Insights:

1.The average salary for males and females for different grades. And the average salary for different grades.

	Salary	
Gender	Female	Male
Grade		
1	32649.0	31079.0
2	34865.0	34217.0
3	38536.0	39329.0
4	44351.0	43845.0
5	51100.0	49750.0
6	30000.0	70923.0

	Salary
Grade	
1	32335.0
2	34665.0
3	38665.0
4	44152.0
5	50329.0
6	68000.0

The salaries for male and female was the same for all levels when compared to the average salary as a whole except for level 6 where the average salary of men was higher when compared to women.

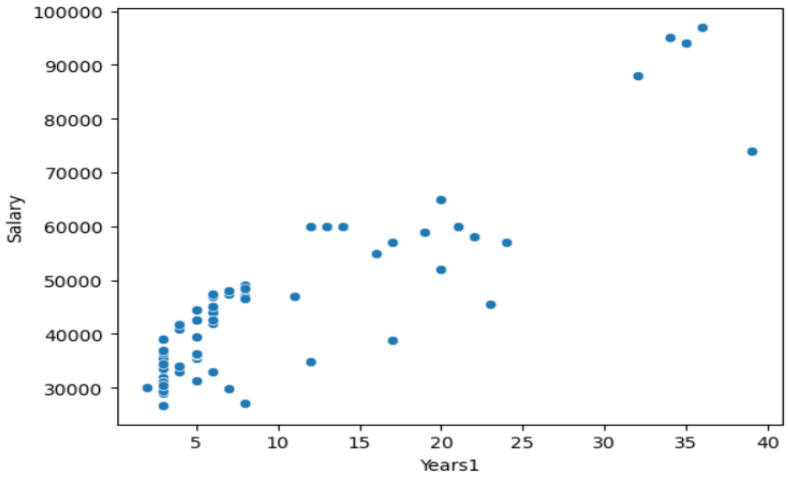
2.Comparing education level , salary and gender

	Salary	
Gender	Female	Male
Education		
1	35176.060606	39933.333333
2	35131.034483	31970.000000
3	36716.279070	39247.500000
4	39500.000000	44666.666667
5	41782.666667	51772.222222

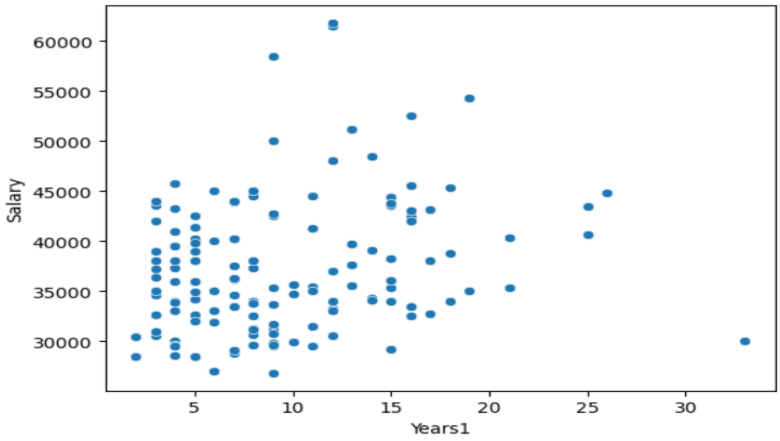
	Salary
Education	
1	35572.500000
2	34589.142857
3	37519.841270
4	41437.500000
5	47231.515152

Except for level 2, men had a higher salary than women at all education levels and their average were higher than the overall salary average for different education levels.

3.Scatter plot for male years1 and salary



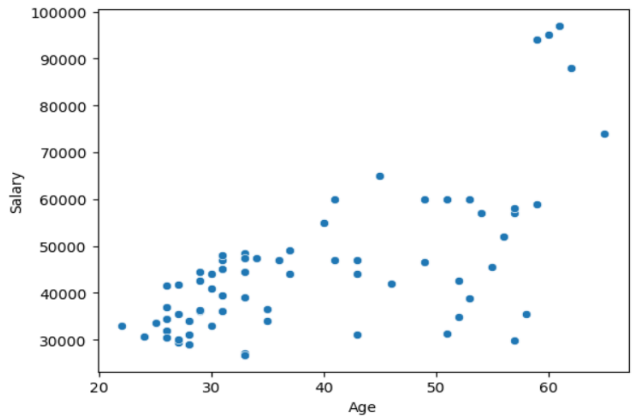
Previously we saw that years1 and salary had a positive correlation But when we check for only men there is a strong positive correlation between years 1 and salary. Correlation =0.88



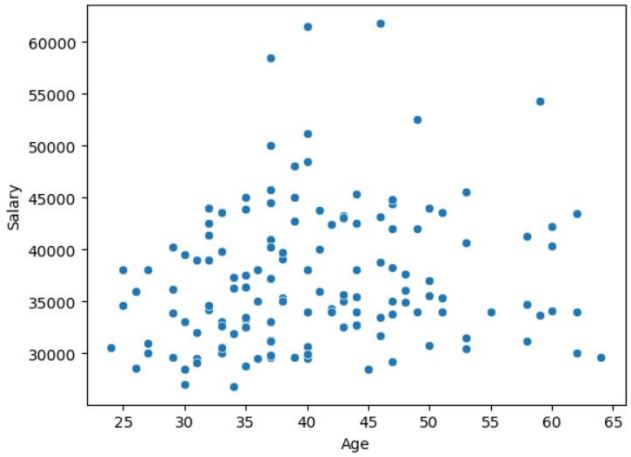
But for women, there isn't a strong correlation between age and salary. Correlation =0.23

4. Age, Salary and Gender:

Male Age and Salary:



Female Age and Salary:



Male Age and Salary have a better correlation than female age and salary.

5. Gender, Age, Salary, and Job Grade:

	Age		Salary	
Gender	Female	Male	Female	Male
Grade				
1	40.458333	35.833333	32649.166667	31079.166667
2	40.034483	32.538462	34865.172414	34216.923077
3	40.361111	33.571429	38535.555556	39328.571429
4	43.294118	36.000000	44350.588235	43845.454545
5	41.111111	39.666667	51100.000000	49750.000000
6	62.000000	55.230769	30000.000000	70923.076923

	Salary
Grade	
1	32335.166667
2	34664.523810
3	38664.651163
4	44152.142857
5	50328.571429
6	68000.000000

The average salary for women and men is similar to the overall average salary for men and women combined. But the age of women in each education level is higher when compared to that of men who are receiving similar levels of salaries.

Summary statistic for Linear Regression with gender as a feature and a constant:

For gender I have encoded Female as 1 and male as 0:

OLS Regression Results

Dep. Variable:	Salary	R-squared:	0.765
Model:	OLS	Adj. R-squared:	0.748
Method:	Least Squares	F-statistic:	44.94
Date:	Fri, 23 Jun 2023	Prob (F-statistic):	5.21e-53
Time:	22:08:47	Log-Likelihood:	-2084.3
No. Observations:	208	AIC:	4199.
Df Residuals:	193	BIC:	4249.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.227e+04	1281.001	17.383	0.000	1.97e+04	2.48e+04
Years1	515.5827	97.980	5.262	0.000	322.334	708.832
Years2	167.7270	140.442	1.194	0.234	-109.272	444.726
Age	-8.9621	57.699	-0.155	0.877	-122.763	104.839
Gender	-2554.4740	1011.974	-2.524	0.012	-4550.423	-558.525
Education_1	3849.9151	1166.514	3.300	0.001	1549.163	6150.668
Education_2	3364.3627	1117.904	3.010	0.003	1159.485	5569.241
Education_3	4377.8300	751.726	5.824	0.000	2895.177	5860.483
Education_4	4135.0909	1724.458	2.398	0.017	733.888	7536.294
Education_5	6540.7157	872.521	7.496	0.000	4819.815	8261.616
Grade_1	-5100.4288	1030.579	-4.949	0.000	-7133.073	-3067.785
Grade_2	-3535.9323	983.044	-3.597	0.000	-5474.821	-1597.043
Grade_3	118.9293	953.358	0.125	0.901	-1761.409	1999.268
Grade_4	3494.4040	1002.389	3.486	0.001	1517.361	5471.447
Grade_5	8558.9803	1243.702	6.882	0.000	6105.988	1.1e+04
Grade_6	1.873e+04	2097.609	8.930	0.000	1.46e+04	2.29e+04
PC Job_No	8672.5344	787.371	11.015	0.000	7119.578	1.02e+04
PC Job_Yes	1.36e+04	1134.288	11.986	0.000	1.14e+04	1.58e+04

Omnibus:	80.603	Durbin-Watson:	1.585
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2443.188
Skew:	-0.753	Prob(JB):	0.00
Kurtosis:	19.722	Cond. No.	1.03e+18

The coefficient for gender is -2554(approx.) which means when the gender is female the target variable reduces by -8295 whereas for males the effect is nullified as the gender becomes zero.

Fitted a linear regression model with Gender, Education Level, Age, Years1, and Grade(on basis of eda):

OLS Regression Results

Dep. Variable:	Salary	R-squared:	0.748
Model:	OLS	Adj. R-squared:	0.732
Method:	Least Squares	F-statistic:	48.21
Date:	Fri, 23 Jun 2023	Prob (F-statistic):	9.07e-52
Time:	22:09:56	Log-Likelihood:	-2091.7
No. Observations:	208	AIC:	4209.
Df Residuals:	195	BIC:	4253.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.828e+04	1687.072	16.764	0.000	2.5e+04	3.16e+04
Years1	453.8924	98.246	4.620	0.000	260.131	647.654
Age	24.3390	57.394	0.424	0.672	-88.853	137.531
Gender	-1757.3997	1013.021	-1.735	0.084	-3755.284	240.485
Education_1	4987.1516	1244.952	4.006	0.000	2531.853	7442.450
Education_2	4273.2388	1122.479	3.807	0.000	2059.481	6486.997
Education_3	5902.6512	778.889	7.578	0.000	4366.523	7438.779
Education_4	5344.0851	1791.243	2.983	0.003	1811.389	8876.781
Education_5	7774.6930	876.993	8.865	0.000	6045.084	9504.302
Grade_1	-4434.1568	1036.354	-4.279	0.000	-6478.058	-2390.255
Grade_2	-2199.6333	1003.926	-2.191	0.030	-4179.579	-219.687
Grade_3	1095.5800	980.822	1.117	0.265	-838.801	3029.960
Grade_4	4621.4528	1047.100	4.414	0.000	2556.357	6686.548
Grade_5	9391.8506	1276.795	7.356	0.000	6873.750	1.19e+04
Grade_6	1.981e+04	2196.640	9.017	0.000	1.55e+04	2.41e+04

Omnibus:	74.438	Durbin-Watson:	1.634
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1979.311
Skew:	-0.668	Prob(JB):	0.00
Kurtosis:	18.053	Cond. No.	6.33e+17

The coefficient for gender is -1757(approx.) which means when the gender is female the target variable reduces by -1757 whereas for males the effect is nullified as the gender becomes zero.

Observation and interpretation

* Yes, their is discrimination between gender, and we have evidence i.e thier difference in pay of salary

* Their is also gender discrimination on basis of promotion as men tend to have lower age compared women have higher age, comparatively the avg age gap btw men and women is 10 years

* With feature (Gender, Education-Level(1-5), Age, Years1, Grade-level(1-5), PC job)

The coefficient for gender is -2554(approx.) which means when the gender is female the target variable reduces by -2554 whereas for males the effect is nullified as the gender becomes zero

The coefficient for grade_1 is -5100(approx.) which means when the grade_1 is female the target variable reduces by -5100 whereas for males the effect is nullified as the gender becomes zero

The coefficient for grade_2 is -3535(approx.) which means when the gender is female the target variable reduces by -3535 whereas for males the effect is nullified as the gender becomes zero

* Best variation of model explain by above these set of of feature after removing insignificant feature
(['Salary','Years2',"Age",'Grade_3'])

* With some other in combination of interaction effect

Gender with combination job grades(promotion) are also discriminated, as the gender and grades coefficient remain negative

Gender with combination education, years1, years2 and age are discriminated. These are shown by negative Coeff in model after combining them in interaction effect

Conclusion:

The important insights covered from visuals and the regression model regarding the presence /absence of gender discrimination:

1. For grade six there the male salary was much higher than the female salary
2. Men with higher years1 of experience had a higher chance of getting a better salary. but the same can't be said for women as the years 1 and salary had less correlation.
3. Higher grades indicated a higher average salary, but the average age of women in each age is higher than the men but both were getting similar levels of salary.
4. One-way anova indicated that the average salary of women where significantly different from men. (but does not indicate which is higher than which one.)
5. While fitting a linear model with gender as a feature and salary as the label (where gender F was 1 and M was 0,) the value of the target variable decreased by a significant level indicating the decrease in salary in the presence of female)

From these above points, we can infer from the sample that there exists gender discrimination in the bank and that women are getting a lower salaries than men