UDACITY
Part of Accenture

Analysis of Salary Data

Data Scientist: Natalia María Bonilla Villalobos

Agenda

Findings of linear regression modeling with tech salary data

- Data Description
- Regression Results
- Interpretation and Next Steps

Data Description

The dataset has 375 rows and 6 columns with the below variables:

- Age = float
- Gender = String
- Education Level = String
- Job Title = String
- Years of Experience = float
- Salary = float

So, we find 3 quantitative variables and 3 categorical variables (qualitative), that eventually we'll transform due to the regression model.

Data Description

The data has 12 missing and 50 duplicate values, so that was handled using ‘dropna’ and also ‘drop_duplicates’ in order to clean the dataset.

Now, the dataset has 324 rows with 6 columns.

```
df.isnull().sum()
```

```
Age                    2
Gender                 2
Education Level        2
Job Title              2
Years of Experience    2
Salary                 2
dtype: int64
```

```
df.duplicated().sum()
```

```
49
```

```
df.shape
```

```
(324, 6)
```

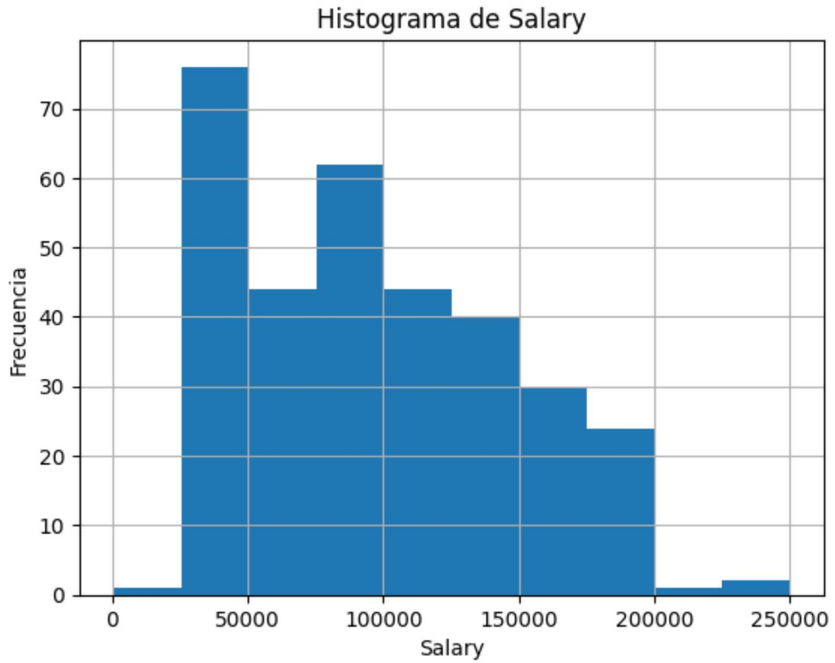
Data Description

Then going through the Salary variable, it has a range of $350.00 to $250,000.00. So, one employee earns $99,985.65 in average, but this one is affected for extreme variables, that means that not all of the earns it.

In other hand, we have $95,000.00 as a median, in comparison is not too far to the average. In statistical terms, the salaries stray more or less $48,652.27 from the average.

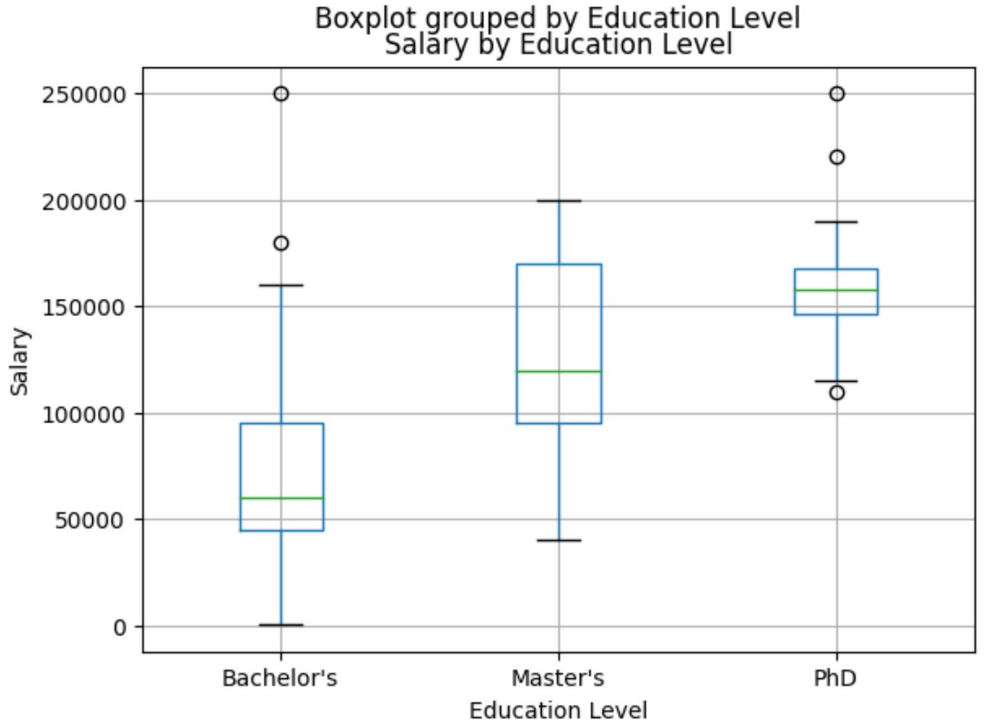
Data Description

Finally, the Salary does not have a normal distribution, it is a distribution skewed to the right. Consequently, we find extreme values far from the peak on the high end more frequently than on the low.



Data Description

Here we can see a Salary comparison between the Bachelor's, Master's and PhD and it outliers or extreme values.



Regression Results

Because of the categorical variables, a transformation technique was applied using dummy variables for “Gender”, “Educational Level” and “Job Title”.

Taking as a baseline:

- Gender: is_Not
- Education Level: is_director
- Job Title: Bachelor’s

Age	Education Level	Years of Experience	Salary	is_junior	is_senior	is_manager	is_analyst	is_engineer	is_Male	Education_Master's	Education_PhD
32.0	Bachelor's	5.0	90000.0	0	0	0	0	1	1	0	0
28.0	Master's	3.0	65000.0	0	0	0	1	0	0	1	0
45.0	PhD	15.0	150000.0	0	1	1	0	0	1	0	1
36.0	Bachelor's	7.0	60000.0	0	0	0	0	0	0	0	0
52.0	Master's	20.0	200000.0	0	0	0	0	0	1	1	0

Regression Results

Four variables were found to have high p-values that are not statistically significant. Therefore, they will be removed from the model.

Variables:

- is_junior 0.207 > p = 0.05
- is_manager 0.800 > p = 0.05
- is_analyst 0.170 > p = 0.05
- is_engineer 0.947 > p = 0.05

	coef	std err	t	P>|t|	[0.025	0.975]
is_Male	8652.8528	1819.677	4.755	0.000	5072.551	1.22e+04
Years of Experience	5640.4798	190.999	29.531	0.000	5264.680	6016.280
is_junior	-3609.6775	2855.816	-1.264	0.207	-9228.631	2009.277
is_senior	6458.6657	2151.545	3.002	0.003	2225.399	1.07e+04
is_manager	558.7344	2205.951	0.253	0.800	-3781.579	4899.048
is_analyst	-3586.7326	2607.052	-1.376	0.170	-8716.232	1542.767
is_engineer	-294.8188	4441.613	-0.066	0.947	-9033.904	8444.266
Education_Master's	1.78e+04	2322.145	7.665	0.000	1.32e+04	2.24e+04
Education_PhD	2.259e+04	3397.078	6.649	0.000	1.59e+04	2.93e+04
intercept	2.962e+04	2824.282	10.489	0.000	2.41e+04	3.52e+04

Regression Results

With a new model fitted, we can see that the rest of p-values are statistically significant related with salary

	coef	std err	t	P>|t|	[0.025	0.975]
is_Male	8382.5480	1778.387	4.714	0.000	4883.658	1.19e+04
Years of Experience	5803.5032	165.127	35.146	0.000	5478.624	6128.383
is_senior	7313.5528	1929.856	3.790	0.000	3516.654	1.11e+04
Education_Master's	1.814e+04	2272.966	7.980	0.000	1.37e+04	2.26e+04
Education_PhD	2.215e+04	3241.223	6.833	0.000	1.58e+04	2.85e+04
intercept	2.647e+04	1930.575	13.711	0.000	2.27e+04	3.03e+04

Interpretation and Next Steps

Having a PhD Education (compared to the baseline education level: Bachelor's) the salary increase by $22,150.00, holding all else constant.

We are confident that the salary increase for PhD Education oscillates between $15,800.00 and $28,500.00.

	coef	std err	t	P>|t|	[0.025	0.975]
is_Male	8382.5480	1778.387	4.714	0.000	4883.658	1.19e+04
Years of Experience	5803.5032	165.127	35.146	0.000	5478.624	6128.383
is_senior	7313.5528	1929.856	3.790	0.000	3516.654	1.11e+04
Education_Master's	1.814e+04	2272.966	7.980	0.000	1.37e+04	2.26e+04
Education_PhD	2.215e+04	3241.223	6.833	0.000	1.58e+04	2.85e+04
intercept	2.647e+04	1930.575	13.711	0.000	2.27e+04	3.03e+04

Interpretation and Next Steps

Similarly, having a Master's Education compared to the baseline the salary increase by $18,140.00, holding all else constant. With a 95% of confidence, the salary increase oscillates between $13,700.00 and $22,600.00.

	coef	std err	t	P>|t|	[0.025	0.975]
is_Male	8382.5480	1778.387	4.714	0.000	4883.658	1.19e+04
Years of Experience	5803.5032	165.127	35.146	0.000	5478.624	6128.383
is_senior	7313.5528	1929.856	3.790	0.000	3516.654	1.11e+04
Education_Master's	1.814e+04	2272.966	7.980	0.000	1.37e+04	2.26e+04
Education_PhD	2.215e+04	3241.223	6.833	0.000	1.58e+04	2.85e+04
intercept	2.647e+04	1930.575	13.711	0.000	2.27e+04	3.03e+04