UDACITY
Part of Accenture

Analysis of Salary Data

Data Scientist: Komgrich Srisak

Agenda

Findings of linear regression modeling with tech salary data

- Data Description
- Regression Results
- Interpretation and Next Steps

Data Description

Data Description

- Amount of Employee - include empty/null data : 375 persons

Data Description

- Amount of Employee - include empty/null data : 375 persons
- Amount of null/empty data : 2

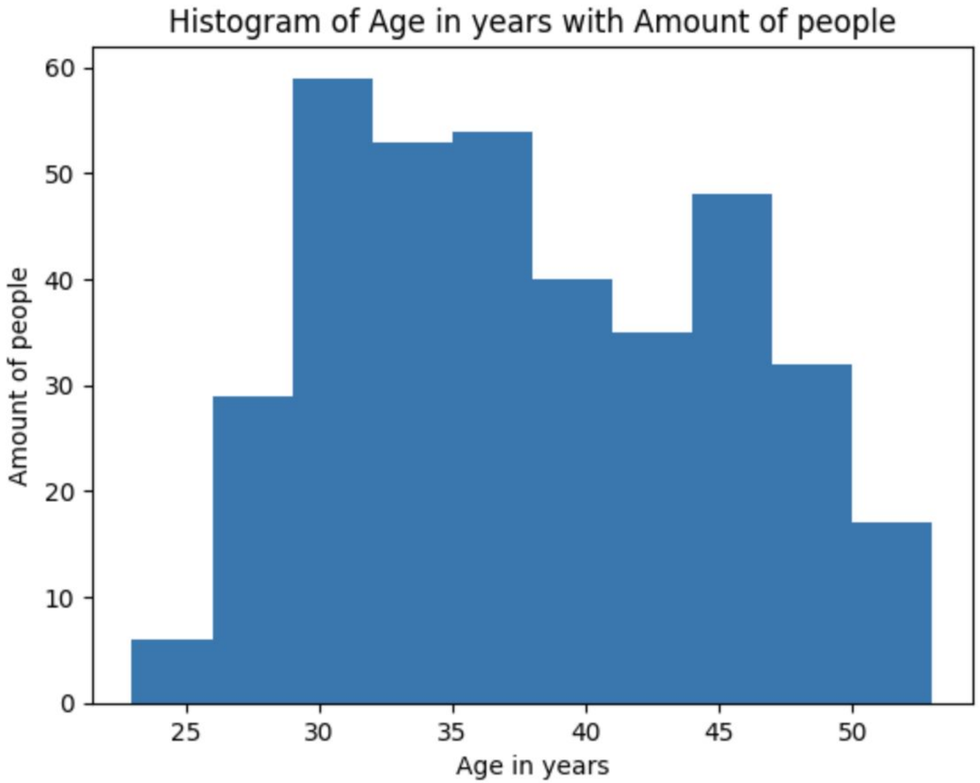
How to deal with null/empty data
: drop away 2 null/empty data

Data Description

- Amount of Employee - include empty/null data : 375 persons
- Amount of null/empty data : 2
- Amount of Employee - exclude empty/null data : 373 persons

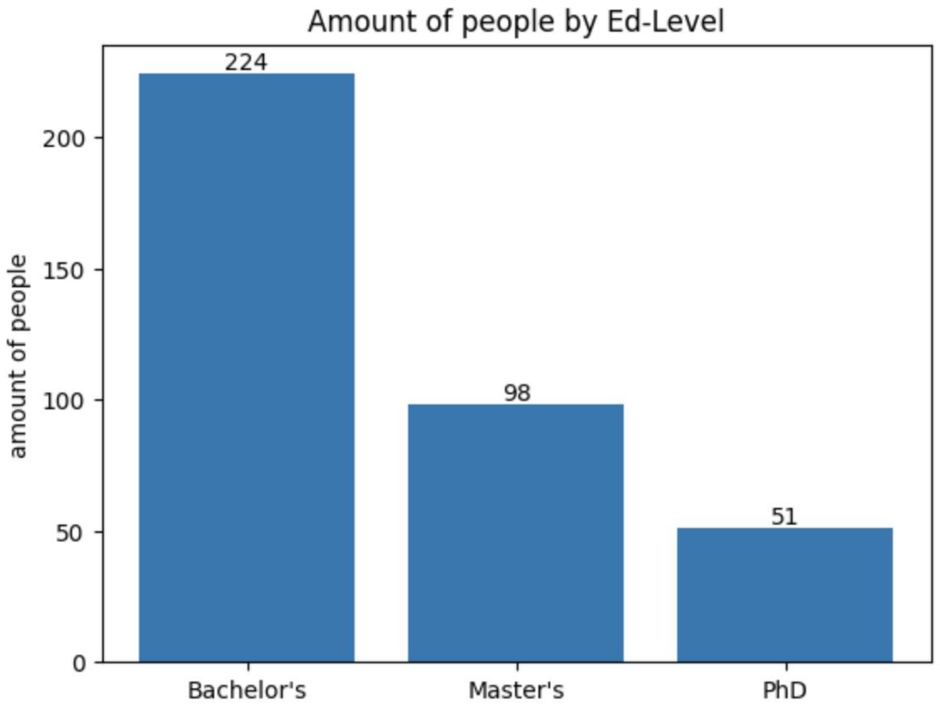
How to deal with null/empty data
: drop away 2 null/empty data

Data Description



- This is a histogram of Age in years with Amount of people (employee)
- Shape : Normally distributed
- Age range : 23 - 53 years old
- Lowest amount of people : 53 yrs
- Highest amount of people : 33 yrs

Data Description



- This is a histogram of Amount of people(employee) by Education level
- The highest amount of people belongs to who has only hold Bachelor’s degree with 224 persons
- The lowest amount of people belongs to who can reach at PhD level with just 51 persons.

Data Description

- This is a histogram of salary amount with amount of people (employee)
- The highest amount of people belongs to who get salary at 40000 dollar and amount of persons in this group is 31 persons.
- The lowest amount of people belongs to who get salary at 350 dollar and amount of persons in this group is just 1 person.

Regression Results

OLS Regression Results

Dep. Variable:	Salary	**R-squared:**	0.921
Model:	OLS	**Adj. R-squared:**	0.919
Method:	Least Squares	**F-statistic:**	382.7
Date:	Fri, 15 Aug 2025	**Prob (F-statistic):**	2.05e-191
Time:	09:15:18	**Log-Likelihood:**	-4077.7
No. Observations:	373	**AIC:**	8179.
Df Residuals:	361	**BIC:**	8227.
Df Model:	11		
Covariance Type:	nonrobust		

- { Before dropping some features which aren't statistical significant }
- Dependent variable : Salary
- R-squared score : 0.921

Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
Age	2624.6859	517.847	5.068	0.000	1606.310	3643.062
Years of Experience	2083.7700	594.583	3.505	0.001	914.489	3253.051
is_director	2.495e+04	3421.566	7.291	0.000	1.82e+04	3.17e+04
is_junior	-6064.7855	2398.876	-2.528	0.012	-1.08e+04	-1347.259
is_senior	1.125e+04	1945.966	5.781	0.000	7422.732	1.51e+04
is_manager	4722.5675	1897.589	2.489	0.013	990.850	8454.285
is_analyst	-674.2852	2037.975	-0.331	0.741	-4682.080	3333.510
is_engineer	1044.0553	3456.195	0.302	0.763	-5752.749	7840.859
Female	-9667.0949	3888.100	-2.486	0.013	-1.73e+04	-2020.924
Male	-1163.9769	3708.120	-0.314	0.754	-8456.207	6128.254
Bachelor's	-1.682e+04	2870.474	-5.861	0.000	-2.25e+04	-1.12e+04
Master's	-2260.8228	2742.452	-0.824	0.410	-7654.011	3132.366
PhD	8254.3988	2999.373	2.752	0.006	2355.960	1.42e+04
intercept	-1.083e+04	7453.881	-1.453	0.147	-2.55e+04	3827.410

Omnibus:	166.054	**Durbin-Watson:**	1.939
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1834.632
Skew:	1.563	**Prob(JB):**	0.00
Kurtosis:	13.405	**Cond. No.**	3.51e+17

- { Before dropping some features which aren't statistical significant }
- Criteria for dropping a feature
  - **(P>|t|) >= 0.05** ; 0.05 = Type I error rate,named "alpha rate".
- The features to drop away :
  - is_analyst
  - is_engineer
  - Male
  - Master's
  - intercept

Regression Results

Dep. Variable:	Salary	**R-squared (uncentered):**	0.985
Model:	OLS	**Adj. R-squared (uncentered):**	0.985
Method:	Least Squares	**F-statistic:**	2693.
Date:	Fri, 15 Aug 2025	**Prob (F-statistic):**	0.00
Time:	09:15:18	**Log-Likelihood:**	-4078.5
No. Observations:	373	**AIC:**	8175.
Df Residuals:	364	**BIC:**	8210.
Df Model:	9		
Covariance Type:	nonrobust		

- { After dropping some features which aren't statistical significant }
- R-squared is (**upraised**!) : 0.985 ; (0.921+0.064)

Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
Age	2093.9200	107.988	19.390	0.000	1881.561	2306.279
Years of Experience	2634.9036	277.819	9.484	0.000	2088.571	3181.236
is_director	2.503e+04	3381.219	7.402	0.000	1.84e+04	3.17e+04
is_junior	-6121.2512	2380.829	-2.571	0.011	-1.08e+04	-1439.344
is_senior	1.144e+04	1912.391	5.981	0.000	7677.770	1.52e+04
is_manager	4741.7958	1756.886	2.699	0.007	1286.875	8196.717
Female	-8475.0791	1436.933	-5.898	0.000	-1.13e+04	-5649.347
Bachelor's	-1.463e+04	1992.539	-7.344	0.000	-1.86e+04	-1.07e+04
PhD	1.053e+04	2656.770	3.963	0.000	5304.717	1.58e+04

Omnibus:	161.535	**Durbin-Watson:**	1.912
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1721.319
Skew:	1.520	**Prob(JB):**	0.00
Kurtosis:	13.076	**Cond. No.**	208.

- { After dropping some features which aren’t statistical significant }

Regression Results

The R-square of this model is 0.985 which means that approximately 98.5 percent of the variance in the dependent variable can be explained by the independent variable(s) in my model. This suggests a very strong relationship between the variables. According to the R-square score, it also indicate that there's only 1.5 percent of error of model prediction.

Shortly, this model is well-performed! but due to high score of it, we still need to consider some of the other factors like overfitting, or even a context for using this model.

(Phrased by me & AI)

Interpretation & Next steps

Interpretation and Next Steps

	coef	std err	t	P>|t|	[0.025	0.975]
Age	2093.9200	107.988	19.390	0.000	1881.561	2306.279
Years of Experience	2634.9036	277.819	9.484	0.000	2088.571	3181.236
is_director	2.503e+04	3381.219	7.402	0.000	1.84e+04	3.17e+04
is_junior	-6121.2512	2380.829	-2.571	0.011	-1.08e+04	-1439.344
is_senior	1.144e+04	1912.391	5.981	0.000	7677.770	1.52e+04
is_manager	4741.7958	1756.886	2.699	0.007	1286.875	8196.717
Female	-8475.0791	1436.933	-5.898	0.000	-1.13e+04	-5649.347
Bachelor's	-1.463e+04	1992.539	-7.344	0.000	-1.86e+04	-1.07e+04
PhD	1.053e+04	2656.770	3.963	0.000	5304.717	1.58e+04

Omnibus:	161.535	**Durbin-Watson:**	1.912
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1721.319
Skew:	1.520	**Prob(JB):**	0.00
Kurtosis:	13.076	**Cond. No.**	208.

- The chosen variable : “Years of Experience”
- Just an extra year of experience is associated with an average increase in salary of 2,634.90 dollar
- We’re 95 percent confident the true increase is between 2088.57 dollar and 3181.23 dollar

Interpretation and Next Steps

	coef	std err	t	P>|t|	[0.025	0.975]
Age	2093.9200	107.988	19.390	0.000	1881.561	2306.279
Years of Experience	2634.9036	277.819	9.484	0.000	2088.571	3181.236
is_director	2.503e+04	3381.219	7.402	0.000	1.84e+04	3.17e+04
is_junior	-6121.2512	2380.829	-2.571	0.011	-1.08e+04	-1439.344
is_senior	1.144e+04	1912.391	5.981	0.000	7677.770	1.52e+04
is_manager	4741.7958	1756.886	2.699	0.007	1286.875	8196.717
Female	-8475.0791	1436.933	-5.898	0.000	-1.13e+04	-5649.347
Bachelor's	-1.463e+04	1992.539	-7.344	0.000	-1.86e+04	-1.07e+04
PhD	1.053e+04	2656.770	3.963	0.000	5304.717	1.58e+04

Omnibus:	161.535	**Durbin-Watson:**	1.912
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1721.319
Skew:	1.520	**Prob(JB):**	0.00
Kurtosis:	13.076	**Cond. No.**	208.

- The chosen variable : “PhD”
- If an employee can reach at PhD level, his average increase in salary will be around 10530 dollar.
- We’re 95 percent confident the true increase is between 5304.71 dollar and 15800 dollar

FIN