UDACITY
Part of Accenture

Analysis of Salary Data

Data Scientist: [Your Name]

How to Use This Template

REFERENCE
REMOVE BEFORE SUBMITTING

- Make a copy of this slide deck.
- We have provided these slides as a guide to ensure that you submit all the required components to successfully complete your project.
- All red text should be replaced with the results of your analysis.
- When presenting your project, please only think of this as a guide. We encouraged you to use creative freedom when making changes as long as the required information is present.
- Don’t forget to delete this slide before you submit your project.

Agenda

Findings of linear regression modeling with tech salary data

- Data Description
- Regression Results
- Interpretation and Next Steps

Data Description

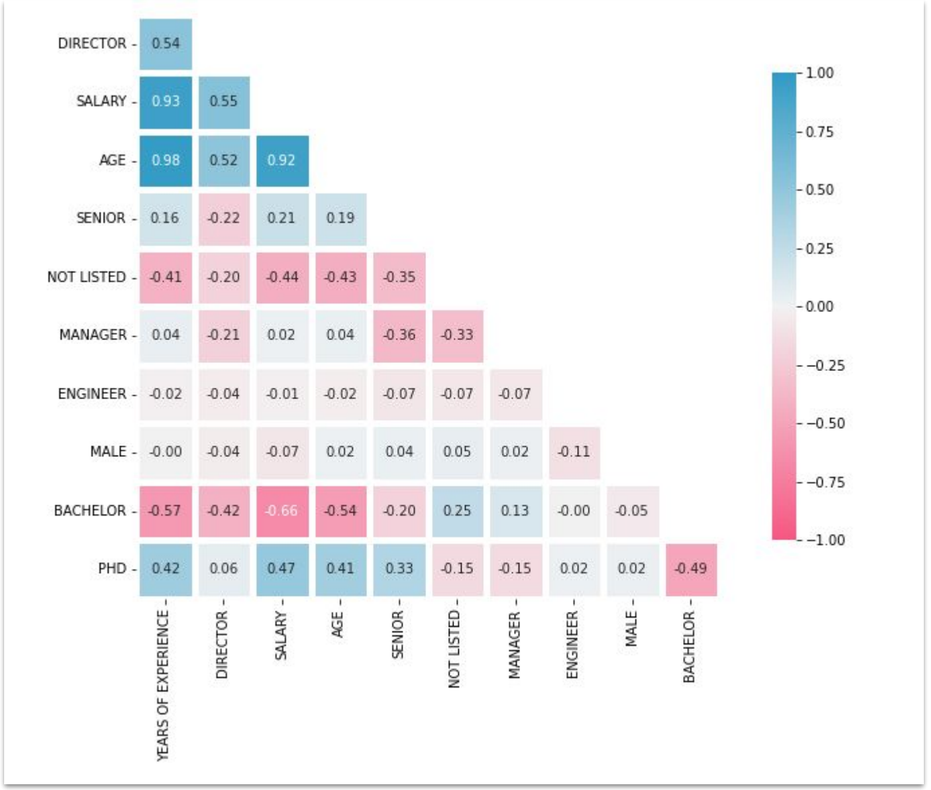
The number of employee = **373** after deducting the 2 empty rows

Missing data values = **2** rows/observations
Technique in handling missing values is: removed them because the "*NaN*" is across the columns I couldn't apply useful technique like imputation.

The range of the salaries is: **249650** (As the min = 350 & max = 250000)

Plots used regarding the salaries: **Histogram** to check the distribution and frequencies and **Box plot** to detect potential outliers.

Additionally I have used heatmap to display the correlation variables with salary, we can see that Years of Experience (YOE) has strong positive correlation with salary (~93) and Job Title "Director" with a confidence (~55), Age with strong positive correlation of (~92) and PhD with a correlation (~47).



Continue with Data Description

e) What are the possible values for `Salary`? What does the distribution of `Salary` look like?

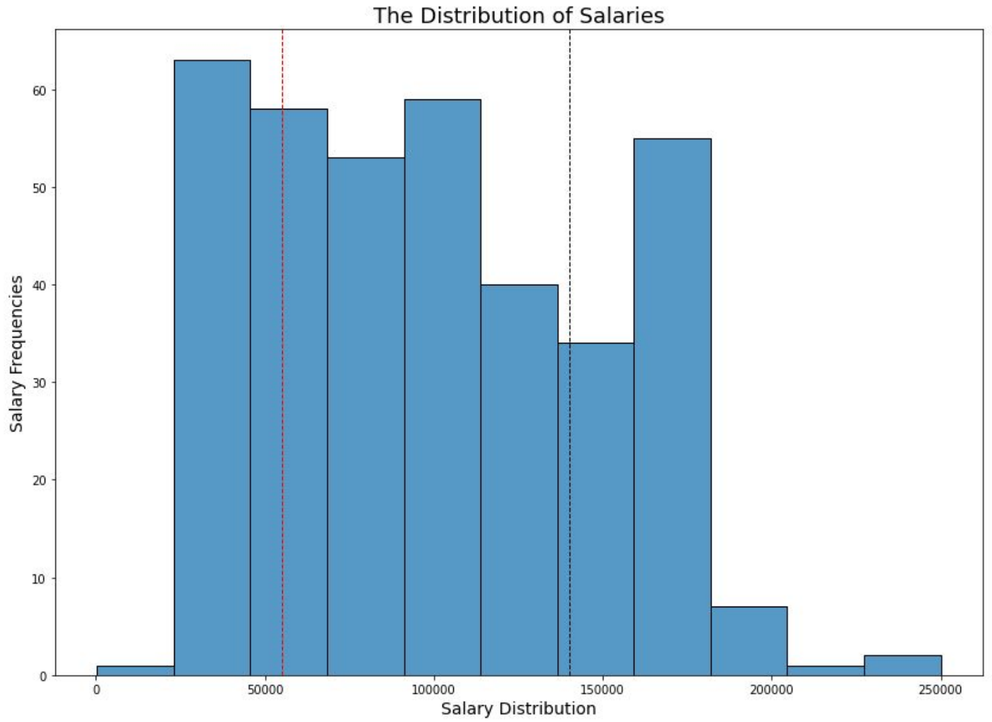
Salary distribution goes between around 50k to slightly beyond 200k, as you can see in the next two visuals (Histogram and Box plot) however I wanna emphasize that there're a potential of outliers regarding this variable and requires some attention in the future investigation.

The Salary looks like a right-skewed visual based on the histogram.

```
#Statistical calculations
Q1,Q3 = np.percentile(df['Salary'],[25,75])
IQR = Q3 - Q1
upper_bound = Q3 + 1.5*IQR
lower_bound = Q1 - 1.5*IQR
Range = (df['Salary'].max()-df['Salary'].min())
print("The 25% percentile = {}\nThe 75% percentile = {}\nThe IQR = {}\nThe Upper Bound = {}\nThe Lower Bound = {}\n
```

```
The 25% percentile = 55000.0
The 75% percentile = 140000.0
The IQR = 85000.0
The Upper Bound = -72500.0
The Lower Bound = 267500.0
The Range of Salary = 249650.0
```

```
import seaborn as sns
#using Histogram to see the distribution of "Salary"
plt.figure(figsize=(14,10),facecolor="w")
sns.histplot(df.Salary)
plt.title('The Distribution of Salaries',fontsize=18)
plt.axvline(Q1, color='r', linestyle='dashed', linewidth=1) #highlighting the 25% percentile
plt.axvline(Q3, color='k', linestyle='dashed', linewidth=1) #highlighting the 75% percentile
plt.ylabel('Salary Frequencies',fontsize=14)
plt.xlabel("Salary Distribution",fontsize=14)
plt.show()
```



Regression Results

Features correlated to the Salary our dependent response, that depends on the fluctuations of the independent variables such as (Years of Experience, Education Level, Job Title, Age, etc.)
In part II we zoom in to understand the features associated with the Salary "dependent response".

Variables with statistical significant with the Salary are:
1. Years of Experience (YOE)
2. Job Title - Director
3. Job Title - Senior
4. Job Title - Manager
5. Job Title - Engineer: It doesn't have significant impact on the Salary however we had to keep him because there's a question (d) in part III request to compute the predicted salary based on the Job Title - `Senior Engineer`.
6. Male
7. Education Level - Bachelor
8. Education Level - PhD

The model fits the data great. As I have computed the R2 score for this model and it scored around 91%, any % above 80 is considered a good score.

Interpretation and Next Steps

Our modified polish multiple linear regression model. Zooming in to the variables with significant impact on the Salary variable.

1st variable is **Years of Experience (YOE)**
Expected range (*4920.765, 5638.144*)

2nd variable is **PhD**
Expected Range (*3758.942, 14800*) with 95% confidence

Both of them have significant impact on the Salary. I find these two variables are practical and relevant to the market.

The expected salary = Intercept + YOE * Coefficients + PhD * Coefficients
= 49700 + 5279 * YOE + 9301 PhD