GENDER PAY GAP IN IT INDUSTRY

Subhashini Natarajan

OBJECTIVE

The objective of this project is to research and determine if gender pay gap exists in IT industry. Data source for the project is the salary survey conducted among IT employees in Europe. This dataset is available in the Kaggle link provided below.

https://www.kaggle.com/parulpandey/2020-it-salary-survey-for-eu-region

DATASET

There are three files of data with the surveys taken in 2018,2019 and 2020. The key variables observed in the dataset include,

- Gender - Identifies if the person is male , female or diverse.
- Experience Range – Number of years of work experience categorized in ranges 0-5, 6-10, 11-20, 20-25 and >25.
- Role- Provides the role, the person performs in the company as a Developer, Analyst, QA or Manager.
- Salary, Current & Prior – Provides the current annual salary and prior year annual salary in Euros.
- Age – Age of the employee. Mostly age will be on par with experience, however if the employee joined the IT industry late or is in with a break, this field will help understand that.

STEPS FOLLOWED TO PERFORM EDA ON THE DATASET

The following steps were followed in performing exploratory data analysis to find if gender pay gap exists in IT industry-

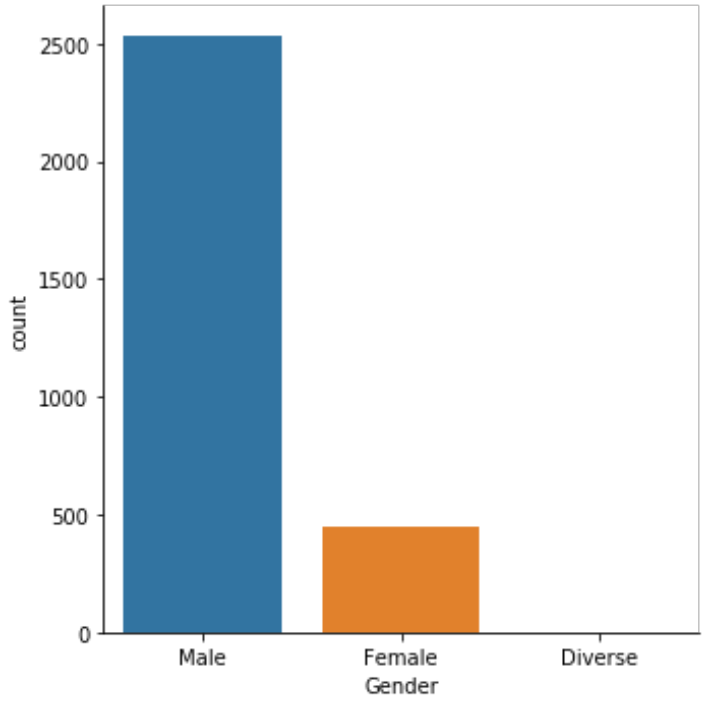
- Data Ingestion - The three files were loaded into three independent dataframes.
- Data profiling – Attributes in the dataframes were profiled to understand the list of unique values, presence of nulls etc.
- Data Transformation – For inconsistent data values, necessary transformations were performed to make the data consistent in the three dataframes and then were combined into one with the fields deemed necessary.
- Attributes were plotted to visually understand the data.

STEPS FOLLOWED TO PERFORM EDA ON THE DATASET (CONTINUED)

- Summary statistics were obtained to understand mean, median , mode, min, max and spread.
- PMF of salary for male and female datasets determined and plotted.
- CDF determined and plotted for male and female datasets to verify if female salary cdf is different than male salary cdf.
- Pareto distribution was determined and plotted for male and female datasets.
- Correlation analysis was performed between the variables.
- T-Test was performed to verify if null hypothesis is true.
- Simple and multiple linear regression performed on the dataset.

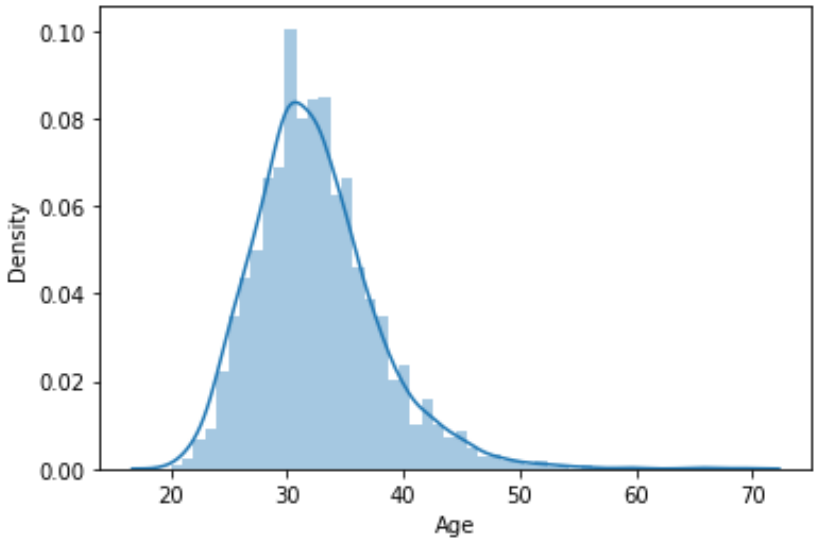
VISUAL OBSERVATION OF DATA

- From the distribution it can be observed that the number of women employees in the survey is only one fifth of the male employees. It is assumed to be a true representation of the women in IT workforce.



VISUAL OBSERVATION OF DATA

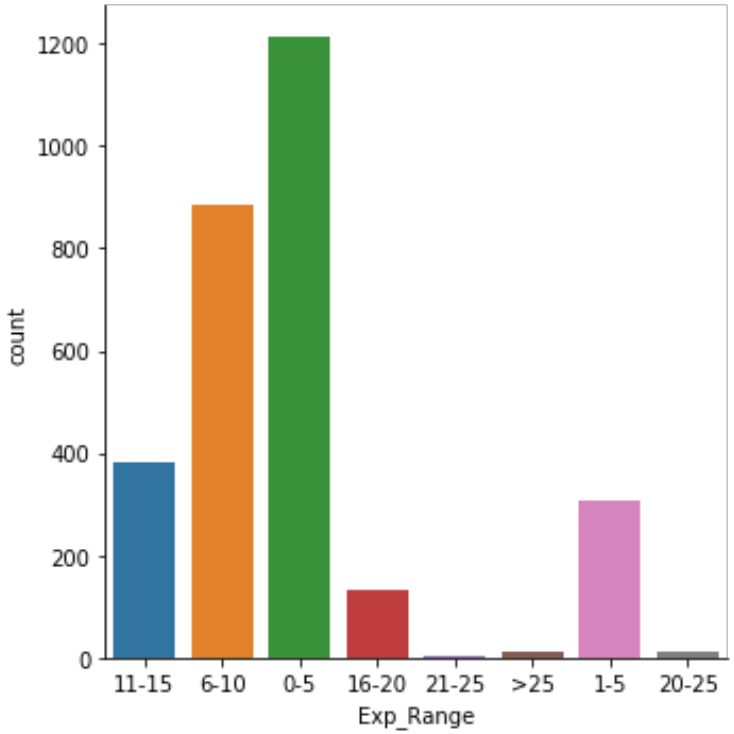
- The plot provides the distribution of age group in the salary survey.
- The plot is generated by distplot function from seaborn library.



VISUAL OBSERVATION OF DATA

- The plot provides the distribution of employees by the experience range. From the dataset it is observed that maximum number employees in the survey data is in the experience range 0-5. As compared with the prior chart where majority age group in the dataset was in 30s. This can indicate that they could have entered the IT workforce from a different career stream.

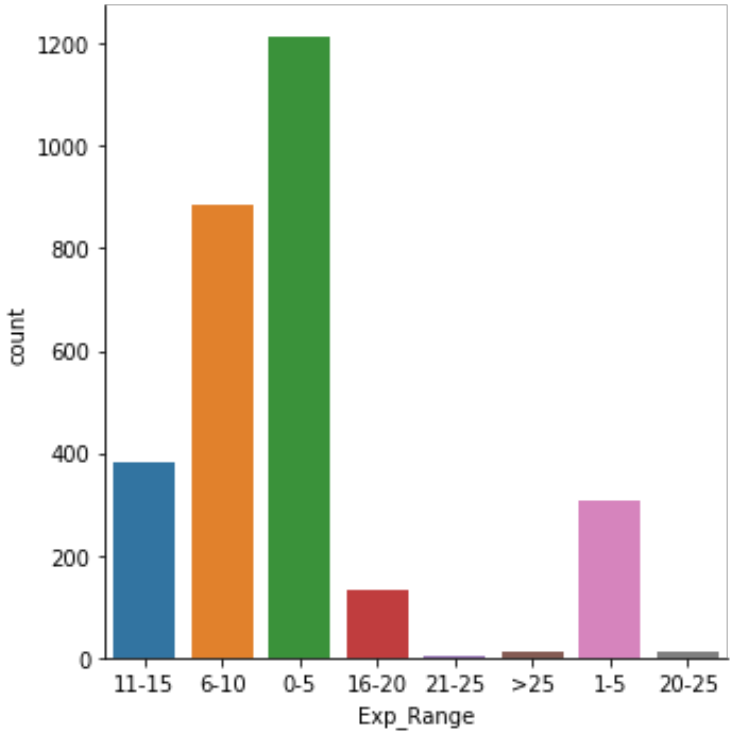
- The plot is generated by catplot function from seaborn library.



VISUAL OBSERVATION OF DATA

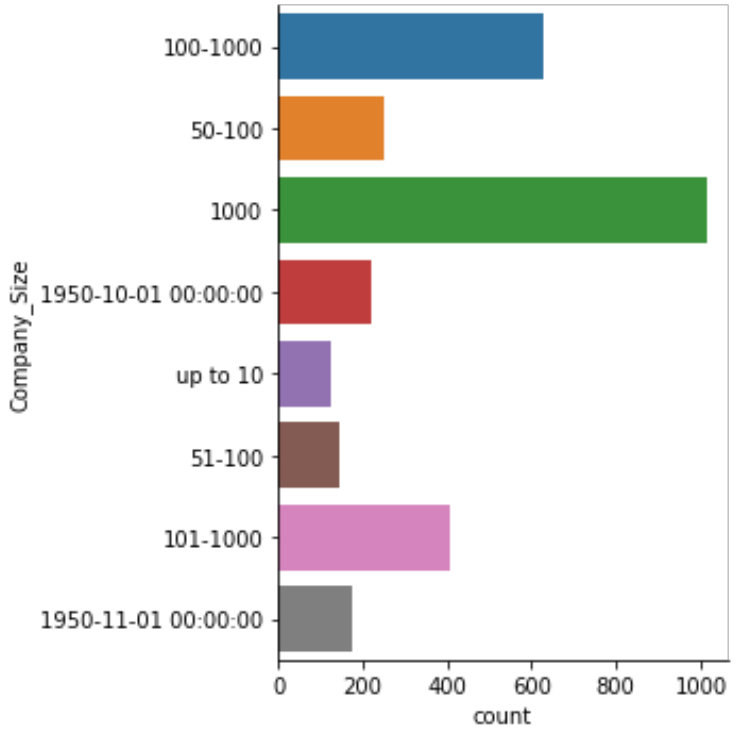
- The plot provides the distribution of employees by the experience range. From the dataset it is observed that maximum number employees in the survey data is in the experience range 0-5. As compared with the prior chart where majority age group in the dataset was in 30s. This can indicate that they could have entered the IT workforce from a different career stream.

- The plot is generated by catplot function from seaborn library.



VISUAL OBSERVATION OF DATA

- The plot provides the distribution of employees by the company size.
- The plot clearly indicates that the field requires clean up.
- The plot is generated by catplot function from seaborn library.



SUMMARY STATISTICS

Summary statistics for the numeric data in the dataset is as follows.

Age	Annual_Salary	Prior_Salary	
count	2780.00000	2993.00000	2084.00000
mean	32.40108	33649436.44007	305317.16123
std	5.40693	1827894112.95914	10951296.03375
min	20.00000	6000.00000	0.00000
25%	29.00000	58000.00000	55000.00000
50%	32.00000	68000.00000	65000.00000
75%	35.00000	80000.00000	75000.00000
max	69.00000	99999999999.00000	500000000.00000

SUMMARY STATISTICS – CONTD.

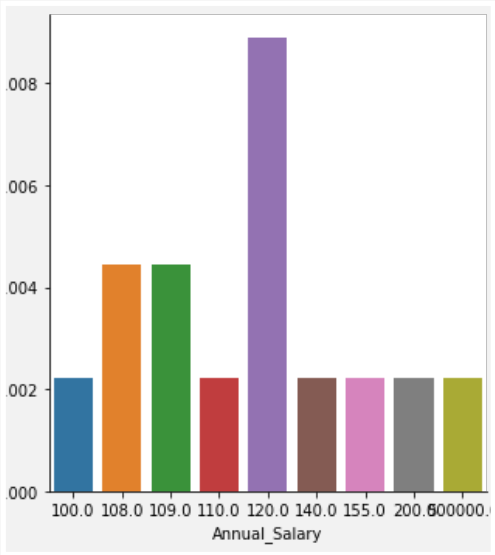
- The statistics indicate the presence of outliers. The outliers tend to skew the dataset. Upon comparing the current and prior salary for all the three years the outlier data (2 records with salary 5000000K and 800000 were removed).
- The summary statistics after the data clean up is to the right.

	Age	Annual Salary	Prior Salary
count	1891	1891	1891
mean	32.86462189	72192.35537	65439.54997
std	5.242043263	21844.81898	26054.63898
min	21	12000	0
0.25	29	60000	55000
0.5	32	70000	65000
0.75	36	80000	75000
max	66	250000	760000

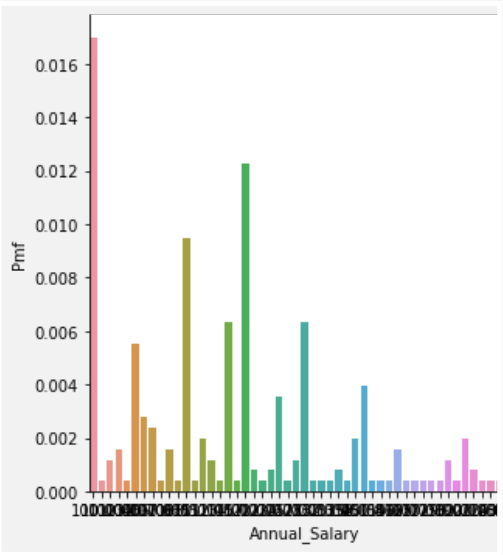
- PMF for male and female salary determined and plotted. Due to volume of dataset, considered salary above 100k for plotting. The plot for Female and Male employees are provided below –

- The plots indicate that the highest probability of female employees to earn is 0.010, however, for a male employee is 0.018.

PMF PLOT FOR HIGHEST SALARY



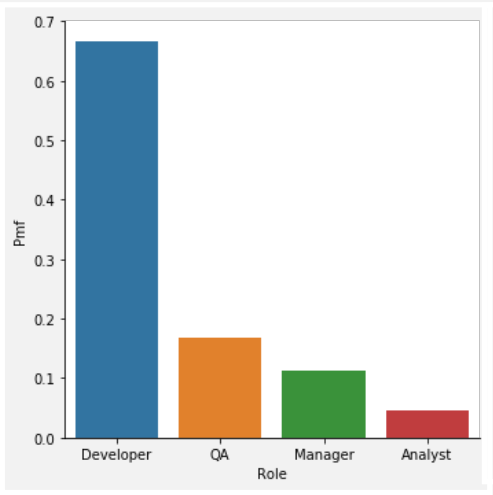
Female PMF plot



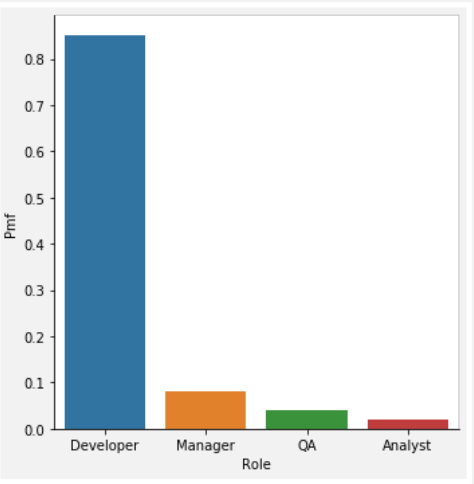
Male PMF plot

PMF PLOT FOR ROLE BETWEEN MALE/FEMALE EMPLOYEES

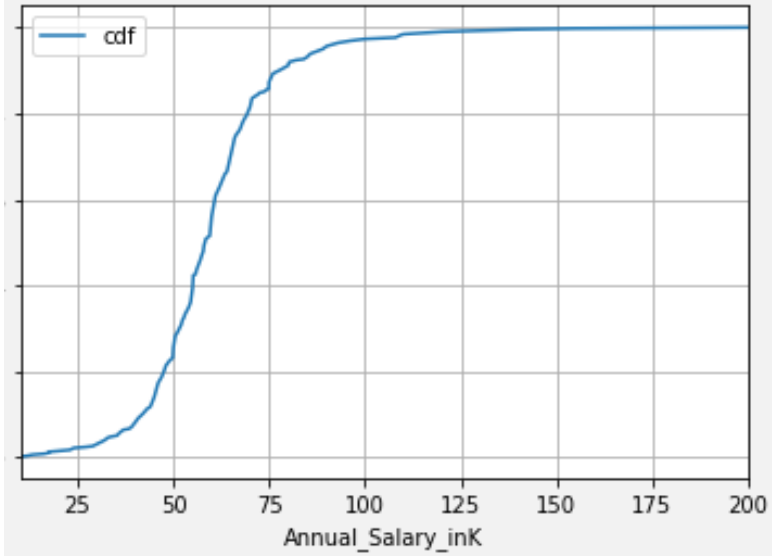
From the plot it is observed that the probability of women employees becoming QA is more than the probability of becoming Manager. Whereas for male employees the probability of becoming Manager is more than QA.



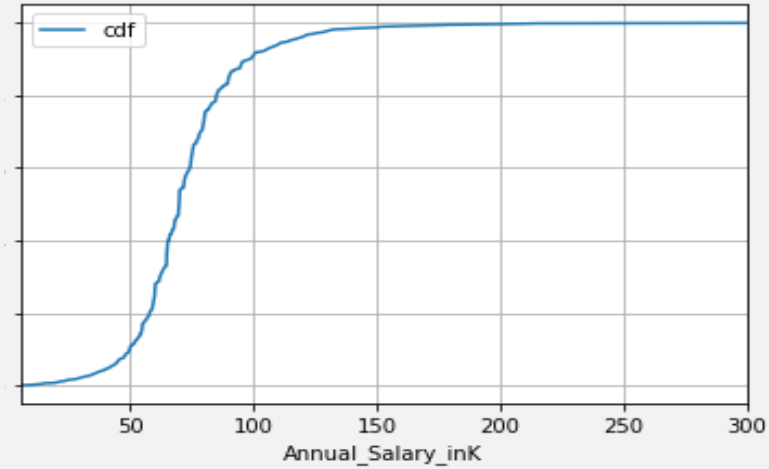
Female Role Distribution



Male Role Distribution



Female Salary CDF



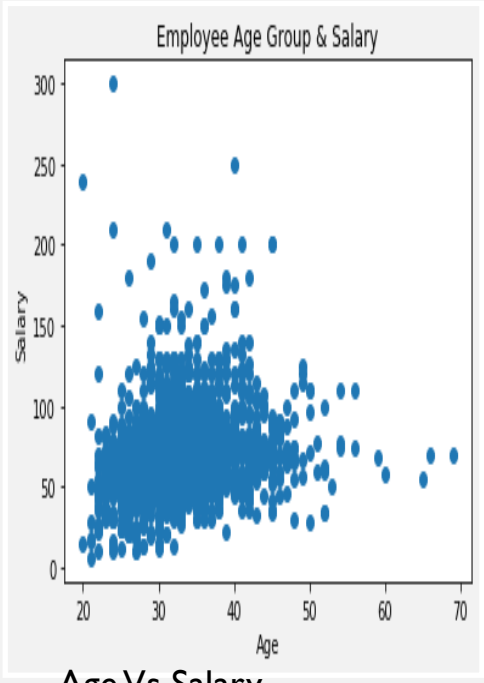
Male Salary CDF

SALARY CDF

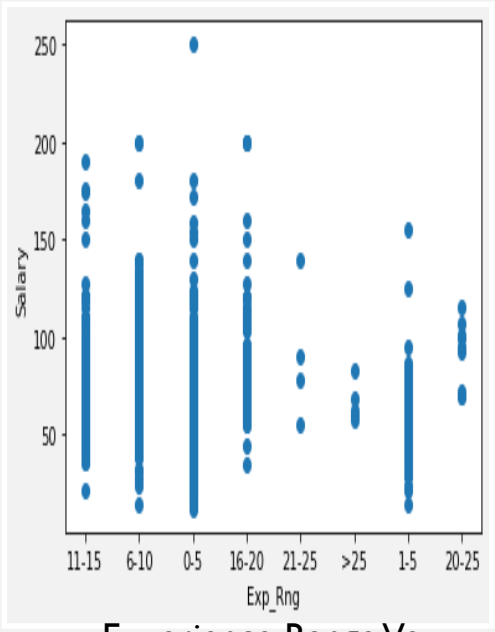
- Female salary CDF indicates that most of the pay is between 50K and 75K.
- Male salary CDF indicates that most of the pay is between 50 and 100K.

CORRELATION

- Scatterplot between age and salary. is represented. Plot indicates that higher salary is earned between 3—40 age group
- Employees with experience Range 16-20 earn the highest salary.



Age Vs Salary



Experience Range Vs Salary

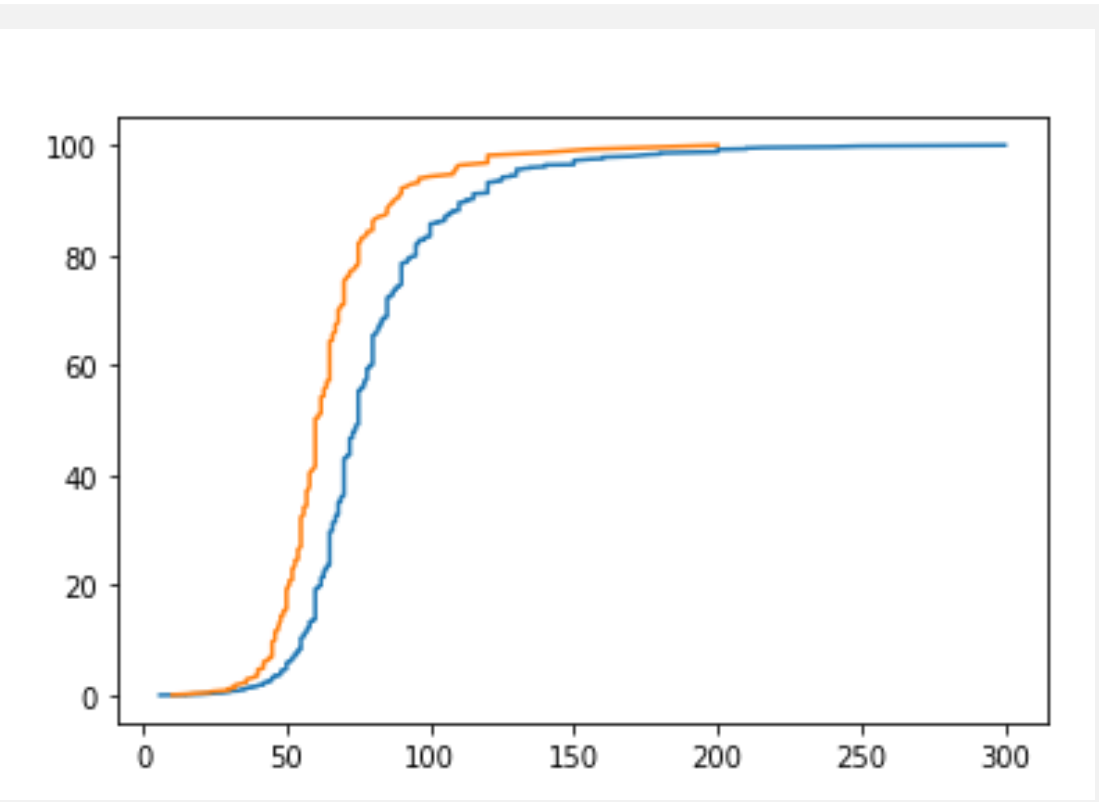
CORRELATION

- The objective of the project is to determine the relationship between Gender and Salary. Gender being categorical data, the correlation between Gender and Salary is determined using **Point Biserial correlation**.

	Gender_num	Annual_Salary_inK
Gender_num	1	0.191272
Annual_Salary_inK	0.191272	1

PARETO ANALYSIS

- Analytic distribution on salary data is performed using pareto analysis. Like CDF, pareto analysis indicate that majority female salary range is between 50K and 75K, whereas male salary range is between 50K and 100K.



REGRESSION ANALYSIS

- The simple regression analysis between Exp Range and Salary is represented. F statistic value of 26.02. The high F stat value indicate strong relationship between Exp Range and Salary.

OLS Regression Results						
Dep. Variable:	Annual_Salary_inK	R-squared:	0.088			
Model:	OLS	Adj. R-squared:	0.085			
Method:	Least Squares	F-statistic:	26.02			
Date:	Thu, 18 Nov 2021	Prob (F-statistic):	3.40E-34			
Time:	19:44:53	Log-Likelihood:	-8427.2			
No. Observations:	1891	AIC:	1.69E+04			
Df Residuals:	1883	BIC:	1.69E+04			
Df Model:	7					
Covariance Type:	nonrobust					
coef	std err	t	P>|t|	[0.025	0.975]	
Intercept	68.0602	0.747	91.074	0	66.595	69.526
Exp_Range[T.1-5]	-8.7129	1.89	-4.611	0	-12.419	-5.007
Exp_Range[T.11-15]	9.4342	1.492	6.324	0	6.509	12.36
Exp_Range[T.16-20]	19.7557	2.281	8.66	0	15.282	24.23
Exp_Range[T.20-25]	23.3842	7.006	3.338	0.001	9.644	37.124
Exp_Range[T.21-25]	22.6898	10.476	2.166	0.03	2.145	43.235
Exp_Range[T.6-10]	7.5765	1.14	6.645	0	5.34	9.813
Exp_Range[T.>25]	-2.5602	8.564	-0.299	0.765	-19.356	14.236
Omnibus:	779.1	Durbin-Watson:	2.002			
Prob(Omnibus):	0	Jarque-Bera (JB):	6252.755			
Skew:	1.73	Prob(JB):	0			
Kurtosis:	11.209	Cond. No.	23.3			

REGRESSION ANALYSIS

- When regression analysis is performed by adding Gender as an independent variable to the regression. The results with Gender female having a value -10 indicates that female salary is lesser than male salary.

OLS Regression Results

Dep. Variable:	Annual_Salary_inK	R-squared:	0.113
Model:	OLS	Adj. R-squared:	0.109
Method:	Least Squares	F-statistic:	29.94
Date:	Thu, 18 Nov 2021	Prob (F-statistic):	2.27e-44
Time:	19:59:57	Log-Likelihood:	-8401.2
No. Observations:	1891	AIC:	1.682e+04
Df Residuals:	1882	BIC:	1.687e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	69.7500	0.773	90.187	0.000	68.233	71.267
Exp_Range[T.1-5]	-7.7792	1.869	-4.163	0.000	-11.444	-4.114
Exp_Range[T.11-15]	8.6233	1.476	5.842	0.000	5.728	11.518
Exp_Range[T.16-20]	18.7050	2.255	8.293	0.000	14.281	23.128
Exp_Range[T.20-25]	21.6945	6.916	3.137	0.002	8.130	35.259
Exp_Range[T.21-25]	21.0000	10.338	2.031	0.042	0.725	41.276
Exp_Range[T.6-10]	6.9745	1.128	6.183	0.000	4.762	9.187
Exp_Range[T.>25]	-2.5816	8.450	-0.306	0.760	-19.153	13.990
Female	-10.0105	1.383	-7.238	0.000	-12.723	-7.298

Omnibus:	804.630	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6909.986
Skew:	1.776	Prob(JB):	0.00
Kurtosis:	11.665	Cond. No.	23.5

T-TEST

Hypothesis : Gender Pay gap exists in IT industry

To test the hypothesis, performed T-Test.

Summary statistics between Male and Female datasets.

	Annual_Salary_inK							
count	mean	std	min	25%	50%	75%	max	
Gender								
Female	264	61.477	18.92	12	50	60	68	200
Male	1626	73.879	21.70	14	62	71.625	82	250

ASSUMPTION VALIDATION

- T-Test requires the assumption of normality and homogeneity to be true.
- Assumption of Normality is verified using shapiro test.
  - 0.8340311646461487, 3.9997917502840688e-16)
- Assumption of homogeneity is verified using Levene's test.
  - LeveneResult(statistic=5.146387144375454, pvalue=0.023407260166392O3)

T-TEST RESULTS

- Ttest_indResult(statistic=8.759985992836686,
  - pvalue=4.255770985606472e-18)

Lower P value indicates that null hypothesis is not true and Alternative hypothesis is true. Hence it is concluded that gender pay gap exists between male and female employees and regression results indicate that female employee salary is lesser than the male employee salary.