STATISTICS

GROUP REGRESSION PROJECT

REPORT BY

Balaji Venkatesh | Dhananjai Hariharan | Goudam Muralitharan

Sabari Nathan Masilamani | Sameer Raghuram

INTRODUCTION

ITPro is a large national systems company that specializes in programming and system development for client companies. Your consulting firm has been approached by the HR Director to perform an independent analysis to develop a model to explain the firm's salary structure and to provide an opinion to ITPro concerning the adherence to state and federal equal employment opportunity (EEO) practices relating to factors such as age and gender.

A sample of 40 analyst's annual salaries ($K) and the following factors believed to impact salary is included in your Course Data File:

Tenure: Years employed by ITPro

Gender: Male or Female

Age: Age in years of the employee.

Degree: Highest degree obtained as follows:

H: High School

B: Bachelors

M: Masters (or higher)

Position: Two levels defined as follows:

Jr: Junior

Sr: Senior – reserved for those who have demonstrated project leadership capabilities and the ability to be mentors for the Junior Analysts.

PROBLEM STATEMENT

PART I: Descriptive Analysis and Difference in Means.

- Are there any outliers in the data of Salaries?
- Perform a descriptive analysis on the data. The analysis should include descriptive summaries on the key attributes.
- The descriptive analysis should help you identify issues such as differences in means of Salaries across the independent variables of Gender, Degree and Position.
- Did your analysis of difference in means identify any potential EEO concerns?

PART II: Regression Modeling

Using EXCEL, build a regression model to explain Salary. In your report I will be looking for such items as:

- A multiple regression model that you recommend or one you found to be the "best" from all the models you considered.
- The logic you used in selecting the model.
- A clear explanation of the model and its elements.
- A detailed statistical analysis of the model selected.
- Concern for any violations of the model assumptions.

ANALYSIS & SOLUTION

PART 1: DESCRIPTIVE ANALYSIS

Salary ($K)		Salary ($K) (MALE)		Salary ($K)(FEMALE)	
Mean	82.23	Mean	93.89047619	Mean	69.33157895
Standard Error	4.44	Standard Error	6.542330678	Standard Error	4.451585253
Median	72.60	Median	91.5	Median	69.4
Mode	#N/A	Mode	#N/A	Mode	#N/A
Standard Deviation	28.09	Standard Devia	29.98072555	Standard Devia	19.40401026
Sample Variance	788.99	Sample Varianc	898.8439048	Sample Varianc	376.515614
Kurtosis	-0.45	Kurtosis	-1.22774734	Kurtosis	-0.04181609
Skewness	0.61	Skewness	0.192325304	Skewness	0.384486149
Range	109.80	Range	102.2	Range	73
Minimum	36.10	Minimum	43.7	Minimum	36.1
Maximum	145.90	Maximum	145.9	Maximum	109.1
Sum	3289.00	Sum	1971.7	Sum	1317.3
Count	40.00	Count	21	Count	19
Confidence Level(95.0%)	8.98	Confidence Lev	13.64706265	Confidence Lev	9.352433572

Table 1: Descriptive statistics of the Salaries data

1. OUTLIERS

No outliers were observed in the data of Salaries. The minimum (36.1), and maximum (145.9) salaries were well within the outlier limits (-1.775 – 166.225)

2. DESCRIPTIVE STATISTICS

As observed from Table 1, the sample mean salary for the 40 analysts' was around 82.23 K $ with a sample standard deviation of 28.09 K $. The maximum paid employee had an annual take home of 145.9 K $, while the minimum was at 36.1 K $.

From Table 1, it can also be inferred that the sample mean salaries of male analysts (93.89 K $) was higher than the same for female analysts (69.33 K $).

Salary ($K) - Bachelors Degree		*Salary ($K) High School*		*Salary ($K) Masters Degree*	
Mean	77.63333333	Mean	77.67	Mean	98
Standard Error	5.915855746	Standard Error	10.1836803	Standard Error	7.31060873
Median	71	Median	69.6	Median	97.8
Mode	#N/A	Mode	#N/A	Mode	#N/A
Standard Deviation	27.10985676	Standard Deviation	32.20362471	Standard Deviation	21.93182619
Sample Variance	734.9443333	Sample Variance	1037.073444	Sample Variance	481.005
Kurtosis	1.007849521	Kurtosis	-0.327829545	Kurtosis	-0.40103364
Skewness	1.275282778	Skewness	0.613395633	Skewness	0.105044651
Range	102.2	Range	98.9	Range	68.3
Minimum	43.7	Minimum	36.1	Minimum	62.2
Maximum	145.9	Maximum	135	Maximum	130.5
Sum	1630.3	Sum	776.7	Sum	882
Count	21	Count	10	Count	9
Confidence Level(95.	12.34025885	Confidence Level(95.0%)	23.03708533	Confidence Level(95.0%)	16.85829396

Figure: Descriptive Statistics by Degree

As we can observe from the above figure, salaries for High School and bachelors degrees are approximately the same at 77.67K$ and 77.63K$ respectively. Individuals with Masters degrees have a greater mean salary of about 98K$.

Salaries - Senior Position		*Salaries - Junior Position*	
Mean	109.0375	Mean	64.35
Standard Error	5.654569502	Standard Error	2.666804797
Median	105.35	Median	67.15
Mode	#N/A	Mode	#N/A
Standard Deviation	22.61827801	Standard Deviation	13.06462199
Sample Variance	511.5865	Sample Variance	170.6843478
Kurtosis	-0.001459725	Kurtosis	0.484484304
Skewness	-0.392634322	Skewness	-0.421953385
Range	86.9	Range	56.5
Minimum	59	Minimum	36.1
Maximum	145.9	Maximum	92.6
Sum	1744.6	Sum	1544.4
Count	16	Count	24
Confidence Level(95.0%	12.05242959	Confidence Level(95.0%)	5.516706039

Figure: Descriptive Statistics by Position

As we can observe from the figure above, the mean salaries of the *senior* position and the *junior* position greatly differ.

3. DIFFERENCES IN MEANS

(1) INDEPENDENT VARIABLE – GENDER

Null Hypothesis: Population mean of Salaries for Male = Population mean of Salaries for Female

Alternate Hypothesis: Population means are not equal

Difference in employee salaries based on gender			
t-Test: Two-Sample Assuming Unequal Variances			
	Men	*Women*	
Mean	93.89	69.33	
Variance	898.84	376.51	
Observations	21	19	
Hypothesized Mean Di	0		
df	35		
t Stat	3.104		
P(T<=t) one-tail	0.0019		
t Critical one-tail	1.69		
P(T<=t) two-tail	0.0038		
t Critical two-tail	2.03		

Table 2: t-Test for Difference in means of Salaries between Male, and Female analysts

From Table 2, it is evident that t statistic value is greater than t critical value (two-tail). Hence, we reject the Null Hypothesis. We can say with 95% confidence that there is a difference in the populations' means of Salaries for male analysts, and the populations' means of Salaries for female analysts.

(2) INDEPENDENT VARIABLE - POSITION

Null Hypothesis: Population mean of Salaries for senior analysts = Population mean of Salaries for junior analysts

Alternate Hypothesis: Population means are not equal.

t-Test: Two-Sample Assuming Unequal Variances		
	Senior *Salary ($K)*	Junior *Salary ($K)*
Mean	64.35	109.03
Variance	170.68	511.58
Observations	24	16
Hypothesized Mean Difference	0	
df	22	
t Stat	-7.147	
P(T<=t) one-tail	1.82E-07	
t Critical one-tail	1.717	
P(T<=t) two-tail	3.63E-07	
t Critical two-tail	2.073	

Table 3: t-Test for Difference in means of Salaries between Senior, and Junior analysts

From Table 3, it is evident that t statistic value is lesser than - t critical value (two-tail). Hence, we reject the Null Hypothesis. We can say with 95% confidence that there is a difference in the populations' means of Salaries for Senior analysts, and the populations means of Salaries for Junior analysts.

(3) INDEPENDENT VARIABLE – DEGREE (Bachelors vs High School)

Null Hypothesis: Population mean of Salaries for analysts who are Bachelor's degree holders = Population mean of Salaries for analysts who are High school passouts

Alternate Hypothesis: Population means are not equal

t-Test: Two-Sample Assuming Unequal Variances		
	Salary ($K)(B)	*Salary ($K)(H)*
Mean	77.63	77.67
Variance	734.94	1037.07
Observations	21	10
Hypothesized Mean Difference	0	
df	15	
t Stat	-0.0031	
P(T<=t) one-tail	0.498	
t Critical one-tail	1.753	
P(T<=t) two-tail	0.997	
t Critical two-tail	2.131	

Table 4: t-Test for Difference in means of Salaries between analysts who are Bachelor's degree holders, and High school pass outs

From Table 4, it is evident that t statistic value is not in the rejection region (two-tail). Hence, we cannot reject the Null Hypothesis. We can say with 95% confidence that there is no difference in the populations' means of Salaries for analysts who are Bachelor's degree holders, and the populations means of Salaries for analysts who are High school passouts.

(4) INDEPENDENT VARIABLE – DEGREE (Masters vs High School)

Null Hypothesis: Population mean of Salaries for analysts who are Master's degree holders = Population mean of Salaries for analysts who are High school passouts

Alternate Hypothesis: Population means are not equal

t-Test: Two-Sample Assuming Unequal Variances		
	Salary ($K)(H)	*Salary ($K)(M)*
Mean	77.67	98
Variance	1037.073	481.005
Observations	10	9
Hypothesized Mean Difference	0	
df	16	
t Stat	-1.621	
P(T<=t) one-tail	0.062	
t Critical one-tail	1.745	
P(T<=t) two-tail	0.124	
t Critical two-tail	2.119	

Table 5: t-Test for Difference in means of Salaries between analysts who are Master's degree holders, and High school pass outs

From Table 5, it is evident that t statistic value is not in the rejection region (two-tail). Hence, we cannot reject the Null Hypothesis. We can say with 95% confidence that there is no difference in the populations means of Salaries for analysts who are Master's degree holders, and the populations means of Salaries for analysts who are High school pass outs.

(5) INDEPENDENT VARIABLE – DEGREE (Bachelors vs Masters)

Null Hypothesis: Population mean of Salaries for analysts who are Bachelor's degree holders = Population mean of Salaries for analysts who are Master's degree holders

Alternate Hypothesis: Population means are not equal

t-Test: Two-Sample Assuming Unequal Variances		
	Salary ($K)(B)	*Salary ($K)(M)*
Mean	77.63	98
Variance	734.94	481.005
Observations	21	9
Hypothesized Mean Difference	0	
df	19	
t Stat	-2.16	
P(T<=t) one-tail	0.021	
t Critical one-tail	1.72	
P(T<=t) two-tail	0.043	
t Critical two-tail	2.093	

Table 6: t-Test for Difference in means of Salaries between analysts who are Bachelor's degree holders, and Master's degree holders

From Table 6, it is evident that t statistic value is lesser than - t critical value (two-tail). Hence, we reject the Null Hypothesis. We can say with 95% confidence that there is a difference in the populations' means of Salaries for analysts who are Bachelor's degree holders, and the populations' means of Salaries for analysts who are Master's degree holders.

4. EEO Concerns

From descriptive analysis, and analysis of differences in means, we have identified potential Equal Employment Opportunity concerns. We opine that the salary of an analyst varies depending on the following variables:

1. Gender
2. Position in the company
3. Degree - Bachelor's versus Master's

PART II: REGRESSION ANALYSIS

1. Logic used to arrive at Final Model.

To arrive at the final model of regression, we need to follow a procedure of variable selection, wherein we shall determine which independent variables can be used to properly describe/predict the dependent Salary variable. There are two methods, which we used, and both of them led to the same conclusion (our final model).

I. Forward Selection Method.

In this method, we start the model off by selecting one variable, and check for significance. If the model is significant, we add another variable, perform regression again and perform another check for significance. If the newly added variable is significant, we keep it in the model. If it isn't, we delete it. We continue this process until there are no more independent variables left.

It is always good practice to perform a test of correlation before starting off, as shown below.

Co-Relation

	Tenure	*Age*	*Gender*	*Degree*	*Position*	*Salary ($K)*
Tenure	1					
Age	**0.9540818**	1				
Gender	0.512278203	0.529926	1			
Degree	0.054934862	0.250357	0.03816	1		
Position	0.622475512	0.593902	0.265694	0.326006	1	
Salary ($K)	0.898555139	0.865427	0.442179	0.245415	0.789323	1

From the table above, it is observable that there is a high degree of co-relation between the variable Age and Tenure, which is quite logical when one thinks about it, a higher tenure is directly reflective of the age of the employee. So, in our final model we wouldn't want both Tenure and Age together.

Step 1: Add variable Tenure and run regression.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	**0.898555**							
R Square	0.807401							
Adjusted R Square	0.802333							
Standard Error	12.48826							
Observations	40							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	24844.12	24844.12	159.3015	3.64338E-15			
Residual	38	5926.352	155.9566					
Total	39	30770.48						
	Coefficients	*Standard Err*	*t Stat*	***P-value***	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	51.26231	3.149125	16.27827	1.01E-18	44.88724174	57.637383	44.88724174	57.6373832
Tenure	2.251832	0.178413	12.62147	**0%**	1.890653951	2.6130097	1.890653951	2.61300969

Step 2: Since Tenure is significant (**p value< 0.05**), add another variable.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.945389							
R Square	0.893761							
Adjusted R Square	0.888018							
Standard Error	9.399582							
Observations	40							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	27501.45	13750.72	155.6354	9.6887E-19			
Residual	37	3269.03	88.35215					
Total	39	30770.48						
	Coefficients	*Standard Err*	*t Stat*	***P-value***	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	50.81303	2.371678	21.42492	1.89E-22	46.0075491	55.618503	46.00754909	55.6185031
Tenure	1.666089	0.171582	9.710178	1.02E-11	1.31843149	2.0137467	1.318431486	2.0137467
Position	21.25812	3.876246	5.484204	3.14E-06	13.4041021	29.112142	13.40410215	29.1121421

Step 3: Since both the IV's are significant, we will add a new variable – Degree.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.949988536							
R Square	0.902478219							
Adjusted R Square	0.894351404							
Standard Error	9.129913476							
Observations	40							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	27769.68	9256.561	111.0494345	2.9631E-18			
Residual	36	3000.792	83.35532					
Total	39	30770.48						
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	47.10362236	3.095574	15.21644	2.95691E-17	40.8255076	53.3817371	40.8255076	53.38173714
Tenure	1.7271239	0.170097	10.15378	4.12576E-12	1.38215176	2.07209604	1.38215176	2.072096044
Position	18.53937513	4.058628	4.567893	5.58059E-05	10.3080967	26.7706536	10.3080967	26.77065359
Degree	4.059152786	2.262778	1.793881	**8%**	-0.529973	8.64827859	-0.52997302	8.648278593

Step 4: Since, Degree is not significant (p value>5%), we reject it and add Gender.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.945392							
R Square	0.893765							
Adjusted R Square	0.884913							
Standard Error	9.529031							
Observations	40							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	27501.59	9167.196	100.9576058	1.37633E-17			
Residual	36	3268.887	90.80243					
Total	39	30770.48						
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	50.78388	2.514675	20.19501	3.17347E-21	45.68388561	55.88387856	45.68388561	55.88387856
Tenure	1.662527	0.195874	8.487747	4.09281E-10	1.265276236	2.059776874	1.265276236	2.059776874
Position	21.27046	3.941989	5.39587	4.46179E-06	13.27573734	29.26518675	13.27573734	29.26518675
Gender	0.139415	3.524149	0.03956	**97%**	-7.007891027	7.286721371	-7.00789103	7.286721371

The Gender variable is highly insignificant as seen above. We reject it and settle for our final model of regression containing the two independent variables of Tenure and Position

II. Backwards Elimination Method of Variable Selection.

In this method, we start with a model having all independent variables. We run regression, and delete the most insignificant variable of the bunch. The following series of pictures show the procedure. The field highlighted in red is the most insignificant.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.951694995
R Square	0.905723364
Adjusted R Square	0.891859152
Standard Error	9.236973156
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	27869.53812	5573.907623	65.32815663	1.87189E-16
Residual	34	2900.936885	85.32167308		
Total	39	30770.475			

	Coefficients	*Standard Error*	*t Stat*	***P-value***	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	59.074735	11.50269207	5.13573124	1.14525E-05	35.69845221	82.4510178	35.69845221	82.4510178
Tenure	2.410810073	0.65836707	3.661802335	0%	1.072847211	3.748772935	1.072847211	3.748772935
Gender	0.560744799	3.478866372	0.16118607	**87%**	-6.509162283	7.630651882	-6.509162283	7.630651882
Age	-0.651057455	0.602194771	-1.081140997	0.287243517	-1.874864471	0.572749561	-1.874864471	0.572749561
Degree	6.605885905	3.283777558	2.011672773	0.052233901	-0.067553007	13.27932482	-0.067553007	13.27932482
Position	16.88348599	4.393690856	3.842665891	0%	7.954431871	25.81254011	7.954431871	25.81254011

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.951694995
R Square	0.905723364
Adjusted R Square	0.891859152
Standard Error	9.236973156
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	27869.53812	5573.907623	65.32815663	1.87189E-16
Residual	34	2900.936885	85.32167308		
Total	39	30770.475			

	Coefficients	*Standard Error*	*t Stat*	***P-value***	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	59.074735	11.50269207	5.13573124	1.14525E-05	35.69845221	82.4510178	35.69845221	82.4510178
Tenure	2.410810073	0.65836707	3.661802335	0%	1.072847211	3.748772935	1.072847211	3.748772935
Gender	0.560744799	3.478866372	0.16118607	**87%**	-6.509162283	7.630651882	-6.509162283	7.630651882
Age	-0.651057455	0.602194771	-1.081140997	0.287243517	-1.874864471	0.572749561	-1.874864471	0.572749561
Degree	6.605885905	3.283777558	2.011672773	0.052233901	-0.067553007	13.27932482	-0.067553007	13.27932482
Position	16.88348599	4.393690856	3.842665891	0%	7.954431871	25.81254011	7.954431871	25.81254011

SUMMARY OUTPUT

Regression Statistics	
Multiple R	**0.949989**
R Square	0.902478
Adjusted R Square	0.894351
Standard Error	9.129913
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	27769.68	9256.561	111.0494	2.96314E-18
Residual	36	3000.792	83.35532		
Total	39	30770.48			

	Coefficients	*Standard Erro*	*t Stat*	***P-value***	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	47.10362	3.095574	15.21644	2.96E-17	40.82550757	53.3817371	40.82550757	53.3817371
Tenure	1.727124	0.170097	10.15378	0%	1.382151756	2.07209604	1.382151756	2.07209604
Degree	4.059153	2.262778	1.793881	**8%**	-0.52997302	8.64827859	-0.529973022	8.64827859
Position	18.53938	4.058628	4.567893	0%	10.30809666	26.7706536	10.30809666	26.7706536

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.945389
R Square	0.893761
Adjusted R Square	0.888018
Standard Error	9.399582
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27501.45	13750.72	155.63541	9.68871E-19
Residual	37	3269.03	88.35215		
Total	39	30770.48			

	Coefficients	*Standard Erro*	*t Stat*	***P-value***	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	50.81303	2.371678	21.42492	1.887E-22	46.00754909	55.6185031	46.00754909	55.61850314
Tenure	1.666089	0.171582	9.710178	1.017E-11	1.318431486	2.0137467	1.318431486	2.013746701
Position	21.25812	3.876246	5.484204	3.137E-06	13.40410215	29.1121421	13.40410215	29.1121421

As we can observe, both methods of variable selection lead to the same final regression model. This is how we came to a conclusion.

2. Final Regression Model

SOLUTION: Below shows the final model we arrived at to describe the predictors of the salary variable.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.945389251
R Square	0.893760836
Adjusted R Square	0.888018178
Standard Error	9.399582444
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27501.44545	13750.72	155.6354	9.68871E-19
Residual	37	3269.029554	88.35215		
Total	39	30770.475			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	50.81302612	2.371678462	21.42492	1.89E-22	46.00754909	55.6185031	46.00754909	55.61850314
Tenure	1.666089093	0.17158173	9.710178	1.02E-11	1.318431486	2.0137467	1.318431486	2.013746701
Position	21.25812212	3.876245775	5.484204	3.14E-06	13.40410215	29.1121421	13.40410215	29.1121421

As is observable, we have chosen Tenure and Position as predictors for salary.

i. **Estimated Regression Equation**

From the third table, we can obtain the Estimated Regression Equation.

SALARY = 50.8130 + 1.666(Tenure) + 21.2581(Position)

ii. **Interpreting the co-efficients**

The Coefficients in the regression equation can be interpreted as follows:

b_1 = 1.666

For every year in a tenure, expect an increase in Salary of 1.666 K .

b_2= 21.2581

A senior can expect an increase in salary of 21.2581 K $

iii. ANOVA Output

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	**SSR = 27501.44544**	13750.72	155.6354	9.68871E-19
Residual	37	**SSE = 3269.0295544**	88.35215		
Total	39	**SST = 30770.475**			

SST = SSR + SSE

Where,

SST = Sum of Squares (Total)

SSR = Sum of Squares (Regression)

SSE = Sum of Squares (Error)

iv. Regression Statistics

Regression Statistics	
Multiple R	0.945389251
R Square	0.893760836
Adjusted R Square	0.888018178
Standard Error	9.399582444
Observations	40

v. Tests for Significance

a. **F-test for overall significance.**

Hypotheses:

H_0: $B_1 = B_2 = 0$

H_a: Coefficients are non-zero.

Test Statistic:

F = MSR/MSE

Rejection Rule:

Reject H_0 if $F>F_\alpha$ or if p-value<α (0.05)

ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	**SSR = 27501.44544**	13750.72	155.6354	9.68871E-19			
Residual	37	**SSE = 3269.0295544**	88.35215					
Total	39	**SST = 30770.475**						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	50.81302612	2.371678462	21.42492	1.89E-22	46.00754909	55.6185031	46.00754909	55.61850314
Tenure	1.666089093	0.17158173	9.710178	1.02E-11	1.318431486	2.0137467	1.318431486	2.013746701
Position	21.25812212	3.876245775	5.484204	3.14E-06	13.40410215	29.1121421	13.40410215	29.1121421

Conclusion

F = 155.635 | F_α = 3.59 | P-value = 9.68E-19

Since, $F>F_\alpha$ and p-value < α , we reject the null hypothesis. Therefore, our regression model is statistically significant overall.

b. T-test for individual significance

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	50.81302612	2.371678462	21.42492	1.89E-22
Tenure	1.666089093	0.17158173	9.710178	1.02E-11
Position	21.25812212	3.876245775	5.484204	3.14E-06

t-tests are used to determine whether are the independent variables are statistically significant or not.

Hypotheses

H_0: B_i = 0

H_α: Coefficient of variable is non-zero.

Rejection Rule

p-value < 0.05 (α) | t-stat > $t_{0.025}$ ($t_{\alpha/2}$ = 2.11)

Conclusion

Tenure: We can reject the null hypothesis because p-value<0.05 and t-stat >2.11.

Position: We can reject the null hypothesis because p-value<0.05 and t-stat >2.11.

3.Multi-Colinearity

Multi-colinearity refers to the degree of correlation between independent variables. If two variables are highly correlated, we shouldn't use them together to describe the dependent variable. Multi-colinearity is usually not a problem in cases of prediction. In Excel, we can express multi-colinearity using the correlation matrix. It is always good practice to perform a test of correlation before starting off, as show below.

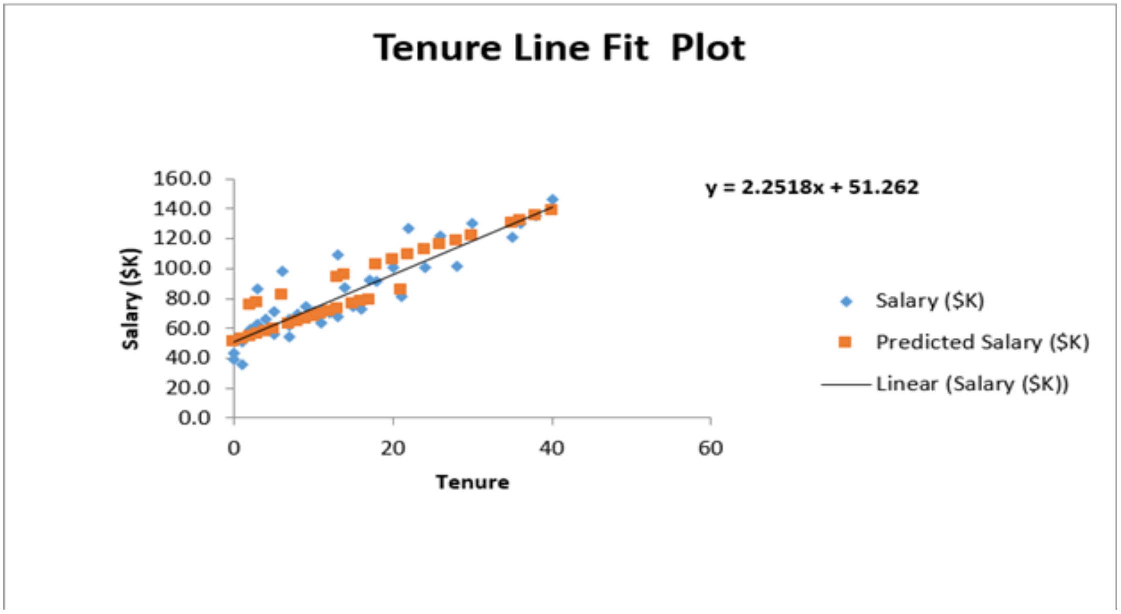
Co-Relation

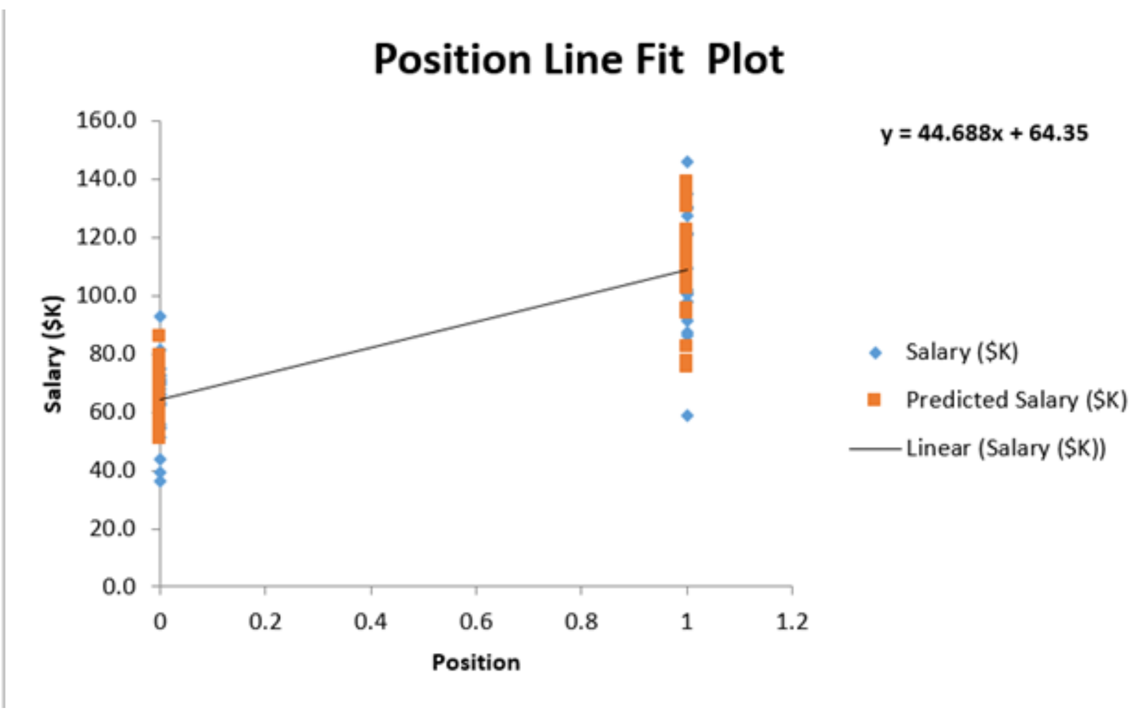
	Tenure	*Age*	*Gender*	*Degree*	*Position*	*Salary ($K)*
Tenure	1					
Age	**0.9540818**	1				
Gender	0.512278203	0.529926	1			
Degree	0.054934862	0.250357	0.03816	1		
Position	0.622475512	0.593902	0.265694	0.326006	1	
Salary ($K)	0.898555139	0.865427	0.442179	0.245415	0.789323	1

From the table above, it is observable that there is a high degree of co-relation between the variable Age and Tenure, which is quite logical when one thinks about it, a higher tenure is directly reflective of the age of the employee. So, in our final model we wouldn't want both Tenure and Age together.

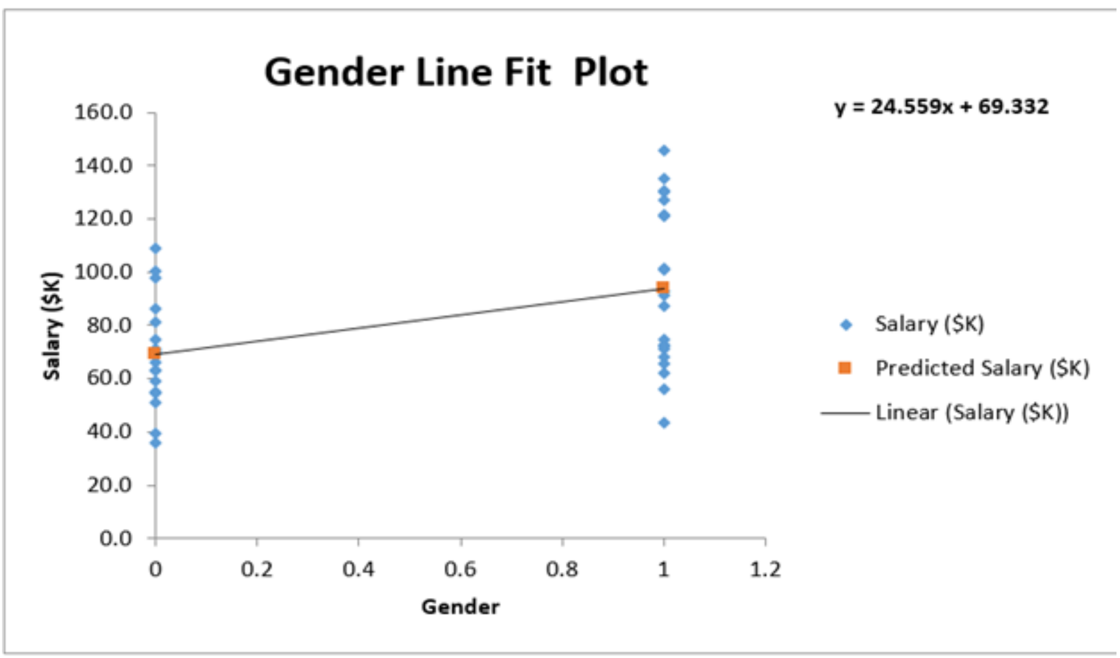
4. Regression Assumption

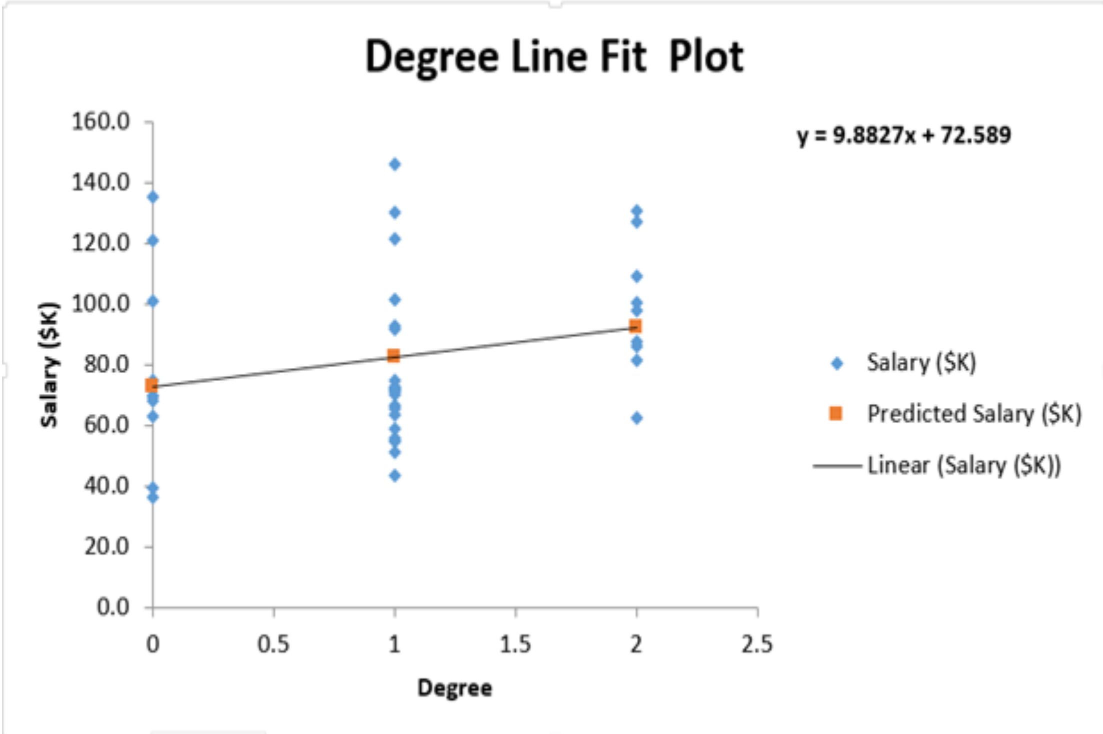
I. Linear Scatter Plots.






Age-Salary
y = 2.042x + 8.5589
Salary
Age
160.0
140.0
120.0
100.0
80.0
60.0
40.0
20.0
0.0
10
20
30
40
50
60
70


Gender Line Fit Plot
y = 24.559x + 69.332
Salary ($K)
Gender
160.0
140.0
120.0
100.0
80.0
60.0
40.0
20.0
0.0
0
0.2
0.4
0.6
0.8
1
1.2
Salary ($K)
Predicted Salary ($K)
Linear (Salary ($K))



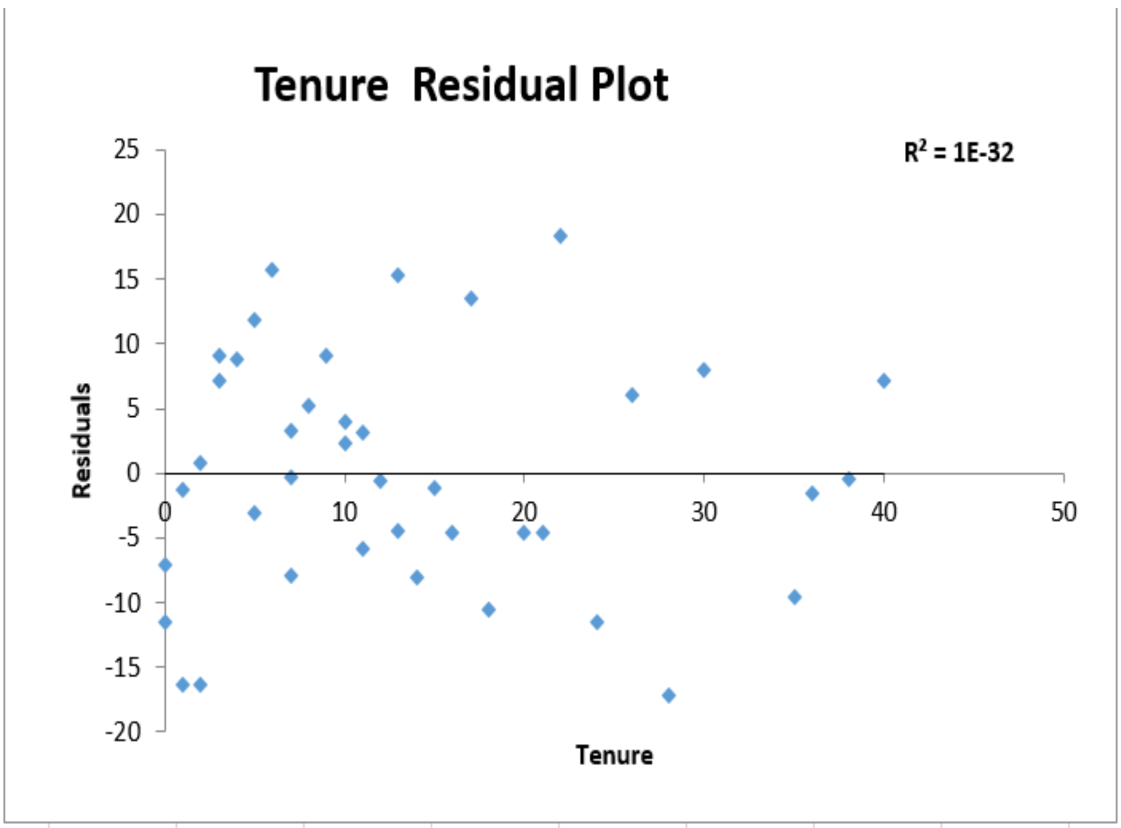
II. Residual Means = 0

69.14000614	-5.840006144	-0.637875885
108.7251083	18.37489171	2.007001367
SUM	-4.26326E-13	
OBSERVATIONS	40	
MEAN	0	

III. Residuals Constant

The two figures below illustrate that the residuals of both the independent variables (Tenure and Position) have the same spread on the Y-axis. This indicates that the residuals are constant.

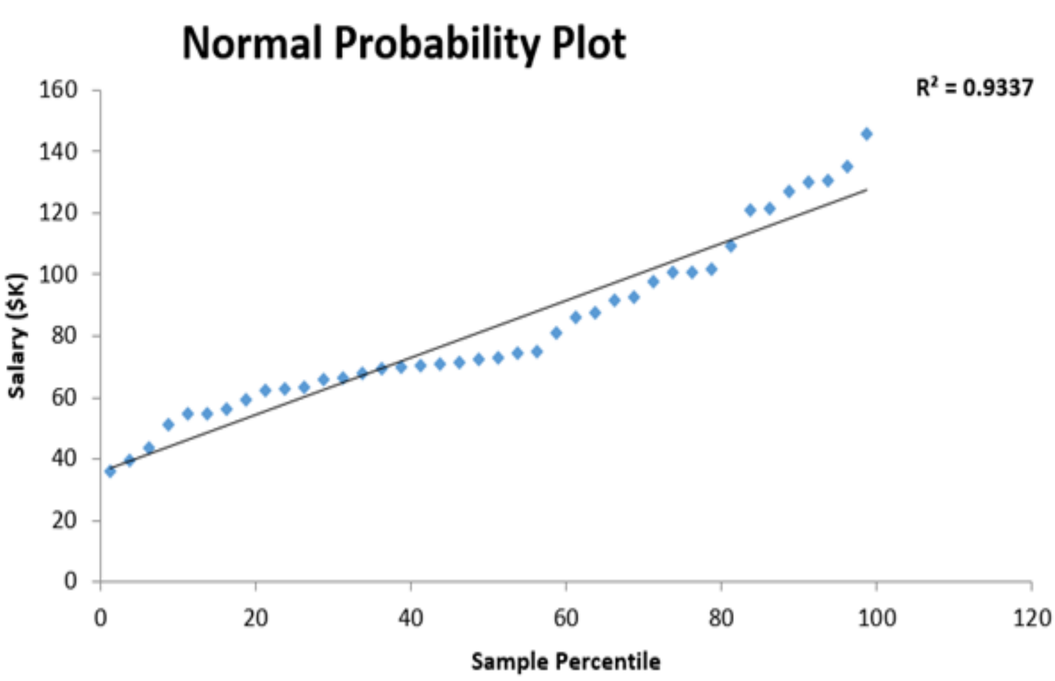

Position Residual Plot
R² = 6E-33
Residuals
Position
25
20
15
10
5
0
-5
-10
-15
-20
0
0.2
0.4
0.6
0.8
1
1.2


Tenure Residual Plot
R² = 1E-32
Residuals
Tenure
25
20
15
10
5
0
-5
-10
-15
-20
0
10
20
30
40
50

IV. Residuals Independent

As seen above in the figures of the residuals we can observe that the distribution of the residuals occurs randomly about the trend line drawn across them. It shows that they are independent

V. Residuals Normal



5. How to Use?

The estimated regression equation we have obtained as a result of our final model is as follows:

SALARY = 50.8130 + 1.666(Tenure) + 21.2581(Position)

Using the above formula, we can obtain estimates of salary for different values of position and tenure. For example, the estimated difference between two individuals having the same position but differing tenures of 10 and 15 years is **1.666(15-10) = $ 8.33k.**

Similarly, the estimated difference in salaries of two individuals having the same tenure but of different positions is **21.2581(1-0) = $ 21.2581k.**

6. Conflict between Part I and Part II

On further analysis of the results of Part I and Part II, we can conclude that they contradict each other. The analysis of descriptive statistics done in Part I by comparing the means of salary, position and degree, suggest that there may be a potential EEO concern. We found that the difference between the population means of salary in case of men and women are not equal. This however does not completely indicate that there is an EEO concern. Equal Employment Opportunity ensures that two individuals with the same level of qualifications and level of seniority earn the same amount of money, regardless of the individual's race, sex, origin etc. We need to take note the specific conditions of level of qualification and seniority.

In Part II of our project after conducting a regression analysis, we found out that the independent variable of gender was very insignificant and couldn't be selected to describe the Salary. After following two methods of variable selection, we arrived at our final model, which consisted of two of our five independent variables - Tenure and Position. This indicates that the only two variables that can successfully describe the salary of an individual working in the company are the tenure and position. This is suitable, since experienced people with senior job roles do demand a higher pay than people with not as much experience and simpler job roles.

A look at the table will reveal that the average tenure for females is 7.79 and that for males is 19.14 (The difference is 11.35). The average position for females is 0.26 and for males it is 0.52 (the difference is 0.26). Therefore there is an estimated difference in salaries of males and females using our regression equation of - **(11.35)*1.666 + 21.2586*(0.26) = $24.50 k**

7. Conclusion

To conclude the overall analysis of various regression methods, we have analyzed the salary data provided to us using two methods - descriptive statistics and regression. Our objective was to investigate the relationship between salary and other variables like gender, position and degree. The result of these investigations will enable us to determine whether there are any pressing EEO concerns or not.

In the first part we analyzed the data and obtained descriptive statistics for male versus female, junior versus senior and for degree. We concluded that there were no outliers in the data set, and that the differences in mean salaries were significant for male versus female, position, masters versus bachelors and masters versus high school.

A multi-colinearity test revealed that there is a high correlation between the age and tenure variables. As a result, we determined that the final regression model should not contain these two variables together.

In part two, we first derived our final multiple regression model by using variable selection techniques that includes forward selection and backward elimination. Both these methods led to a model with the estimated regression equation of:

$$\text{SALARY} = 50.8130 + 1.666(\text{Tenure}) + 21.2581(\text{Position})$$

As the above equation illustrates, the salary of an individual can only be described using *Tenure* and *Position*, and not *Gender*. This relieves our EEO concerns in the example, as we realized that the reason for the conflict in results from the two parts is because of the relative difference between the number of male senior roles and female senior roles.