“SAS FINAL PROJECT”

My dataset is “HR-ANALYTICS”, it basically contains variables related to all the employees in a company related to their Education Field, Job Role, Work Experience and their overall satisfaction related to different aspects in the company.

```
PROC IMPORT OUT=WORK.HR_NK
DATAFILE='\\adm.suffolk.edu\uem\STD-RedirectedFolders\rnk03410\
Documents\ISOM631 FILE\HR ANALYTICS.xlsx'
      DBMS=XLSX REPLACE;
      SHEET='Data';
      GETNAMES=yes;
RUN;
```

- After Importing my file, I have firstly separated variables related to Ratings given by employees in a different table named “HR_NK1”

“Descriptive Statistics”

```
DATA WORK.HR_NK1 (KEEP = JobInvolvement WorkLifeBalance
EnvironmentSatisfaction JobSatisfaction RelationshipSatisfaction);
Set WORK.HR_NK;
RUN;
PROC FREQ DATA= WORK.HR_NK1;
      Table JobInvolvement WorkLifeBalance EnvironmentSatisfaction
JobSatisfaction RelationshipSatisfaction;
RUN;

Data WORK.HR_NK2 (Keep=Satisfaction_Factor Rating);
Set WORK.HR_NK1;
Satisfaction_Factor = 'EnvironmentSatisfaction';
Rating = EnvironmentSatisfaction;
OUTPUT;
Satisfaction_Factor = 'JobSatisfaction';
Rating = JobSatisfaction;
OUTPUT;
Satisfaction_Factor = 'RelationshipSatisfaction';
Rating = RelationshipSatisfaction;
OUTPUT;
Satisfaction_Factor = 'WorkLifeBalance';
Rating = WorkLifeBalance;
OUTPUT;
Satisfaction_Factor = 'JobInvolvement';
Rating = EnvironmentSatisfaction;
OUTPUT;
RUN;
```

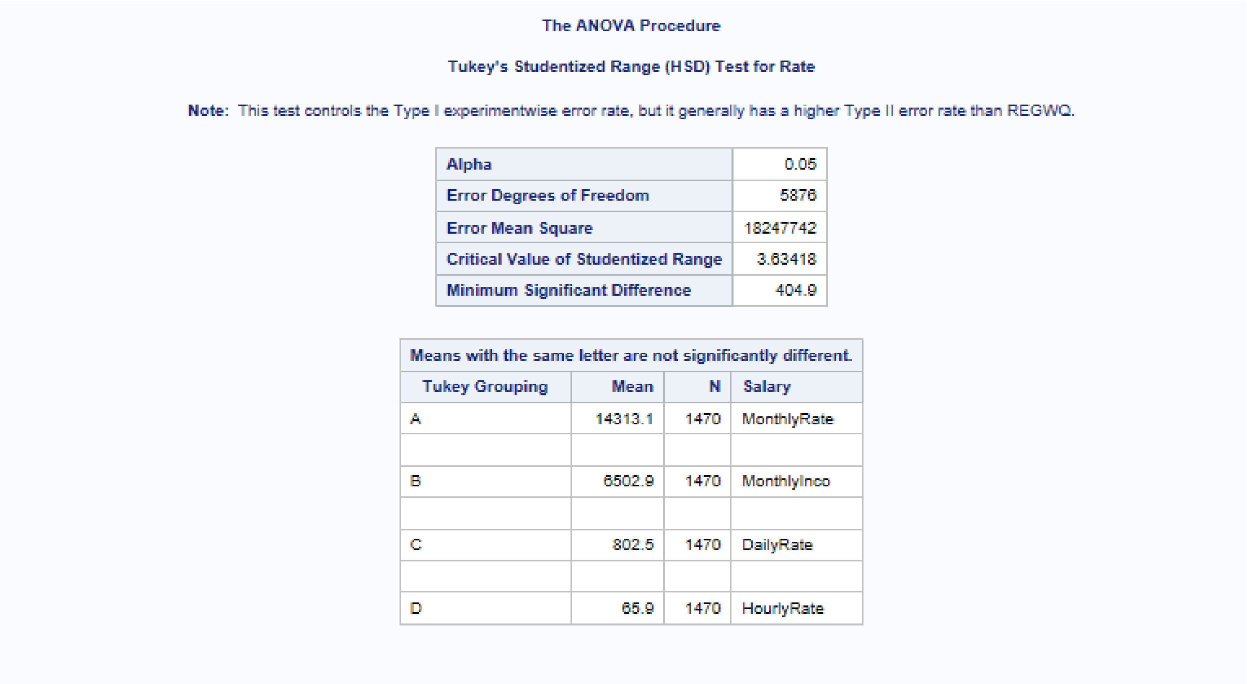
```
PROC UNIVARIATE DATA=WORK.HR_NK2;
      HIstogram Rating /  midpoints=1 to 4 by 1;
RUN;
PROC CORR DATA=WORK.HR_NK1;
      var JobInvolvement WorkLifeBalance RelationshipSatisfaction
JobSatisfaction EnvironmentSatisfaction;
RUN;
```

- After creating a different table for Ratings, I ran Proc Frequency on them just to get an idea on whether the employees are satisfied or not and which ratings score has the highest frequency.
- The results I got from running Proc Freq were positive in which the highest Frequency for all lied on either rating 3 or 4 telling us that majority of the employees are satisfied with their job.
- After running Proc Freq, I pivoted my table of Ratings and named it “HR_NK2” on which I ran Proc Univariate from which I got a “Normally Distributed graph” and from which I learned that its means is 2.72 provided me with other information as well.
- After which I also ran Proc Corr to get a table containing the Co-Relation between all the variables.

“ANOVA TEST”

```
PROC IMPORT OUT=WORK.HR_NK
DATAFILE='\\adm.suffolk.edu\uem\STD-RedirectedFolders\rnk03410\
Documents\ISOM631 FILE\HR ANALYTICS.xlsx'
      DBMS=XLSX REPLACE;
      SHEET='Data';
      GETNAMES=yes;
RUN;

Data work.HR_NK1 (Keep=Salary Rate);
Set WORK.HR_NK;
Salary = 'DailyRate  ';
Rate = DailyRate;
OUTPUT;
Salary='HourlyRate ';
Rate = HourlyRate;
OUTPUT;
Salary='MonthlyIncome   ';
Rate=MonthlyIncome;
Salary='MonthlyRate    ';
Rate=MonthlyRate;
OUTPUT;
RUN;
PROC ANOVA DATA=WORK.HR_NK1;
     class Salary;
     Model Rate = Salary;
     means Salary / hovtest welch tukey;
RUN;
```



The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for Rate

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

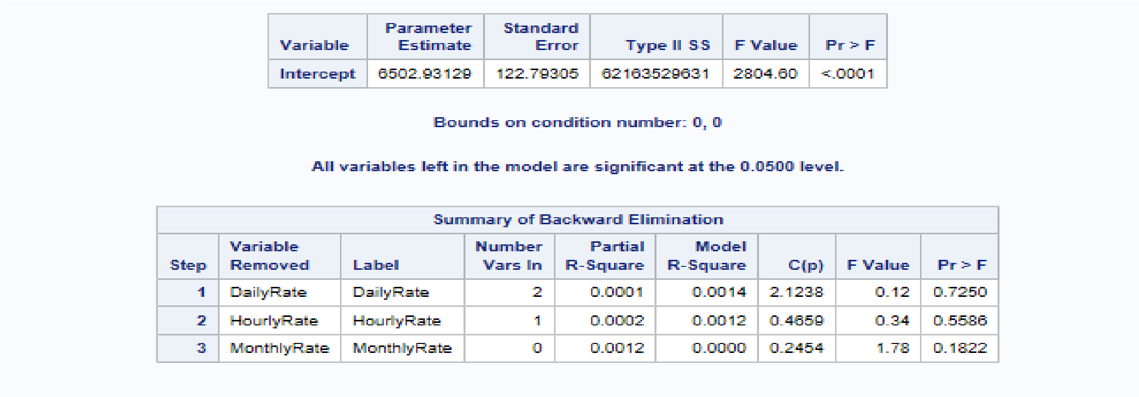
Alpha	0.05
Error Degrees of Freedom	5876
Error Mean Square	18247742
Critical Value of Studentized Range	3.63418
Minimum Significant Difference	404.9

Means with the same letter are not significantly different.			
Tukey Grouping	Mean	N	Salary
A	14313.1	1470	MonthlyRate
B	6502.9	1470	MonthlyInco
C	802.5	1470	DailyRate
D	65.9	1470	HourlyRate

- I separated variables related to Salary and decided to run a Proc Anova on those to see if they are Homogenous or not and if they fall under same groupings or not an have any impact on each other or not.
- H0: Means of all factors related to salary are same
- H1: At least one of the means of the Salary factors is different.
- We are using ANOVA because we have more than 2 variables in our dataset for which we have to compare their means so basically, it's helpful in comparing the means among three or more groups.
- After running the hovtest test and welch test, I saw that the p-value for both of them is <0.0001 which is less than 0.05 which shows us that the means are not homogeneous which basically means we are rejecting the null hypothesis and we are accepting the alternative hypothesis meaning that the means for at least one of them is different.
- After running the Tukey test, the results showed that all the factors related to Salary fall under different groups and none of them are similar.

“Regression Model”

```
PROC REG DATA=WORK.HR_NK;
MODEL MonthlyIncome = DailyRate HourlyRate MonthlyRate / vif;
RUN;
PROC REG DATA=WORK.HR_NK;
MODEL MonthlyIncome = DailyRate HourlyRate MonthlyRate / SELECTION =
Backward SLS=0.05 ADJRSQ;
RUN;
```



Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6502.93129	122.79305	62163529631	2804.60	<.0001

Bounds on condition number: 0, 0

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	DailyRate	DailyRate	2	0.0001	0.0014	2.1238	0.12	0.7250
2	HourlyRate	HourlyRate	1	0.0002	0.0012	0.4659	0.34	0.5586
3	MonthlyRate	MonthlyRate	0	0.0012	0.0000	0.2454	1.78	0.1822

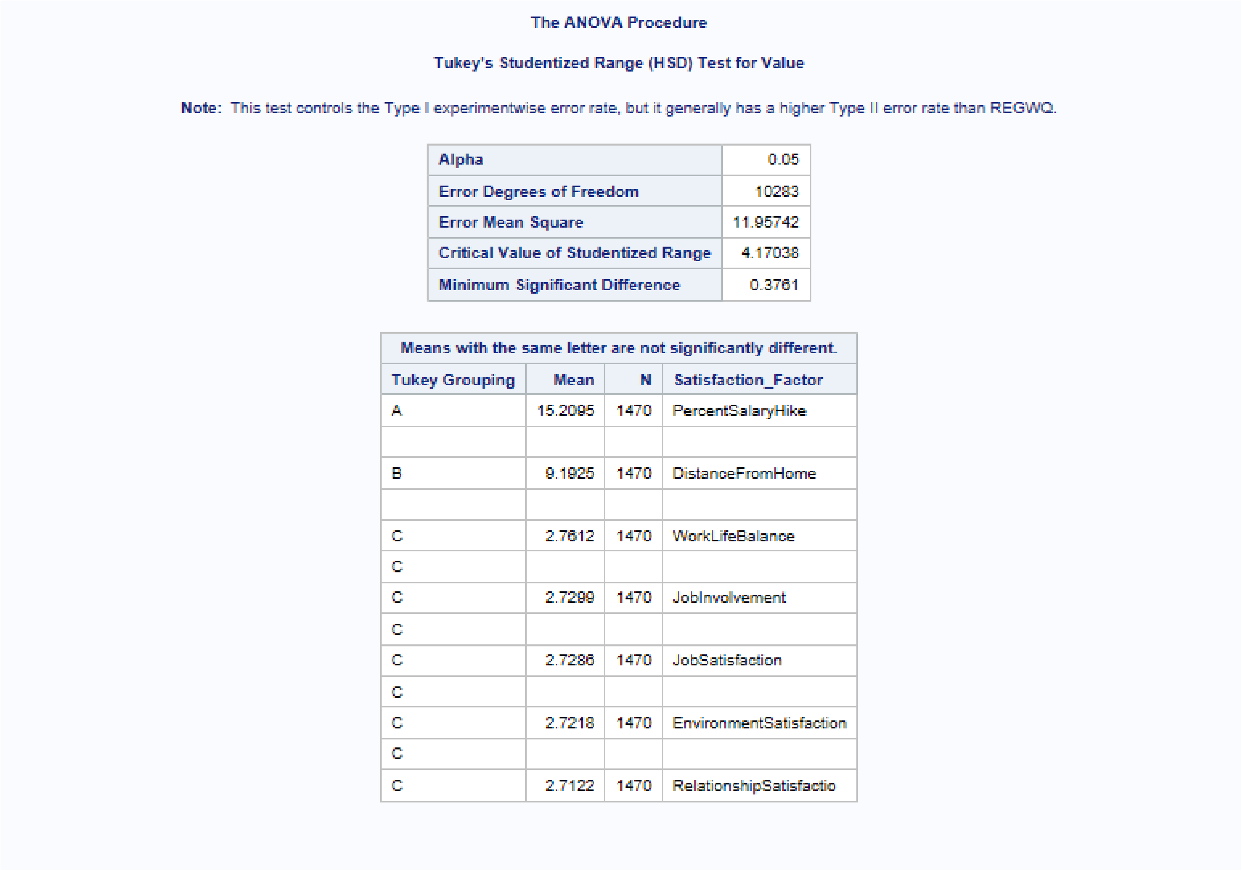
- Linear Regression in SAS is the best way to identify the relationship between one or more independent variables or a dependent variable. The model of relationship is first proposed, and then the estimation of the parameter values is made to develop a regression equation (estimated).
- After running a multiple linear regression model on the variables that are based on an employee’s salary, my output tells me that all the VIFs are less than 8, meaning that there is multicollinearity between the variables. The overall p-value is <0.0001, which is less than 0.05 alpha level, meaning we can reject the null hypothesis and conclude that the slope is 0.
- My Conclusion is that Monthly Income of any employee is not dependent on any of the other independent variables.

“ANOVA TEST”

```
Data work.HR_NK2 (Keep=Satisfaction_Factor Value);
Set WORK.HR_NK;
Satisfaction_Factor = 'DistanceFromHome       ';
Value = DistanceFromHome;
OUTPUT;
Satisfaction_Factor='EnvironmentSatisfaction       ';
Value = EnvironmentSatisfaction;
OUTPUT;
Satisfaction_Factor='JobInvolvement      ';
Value=JobInvolvement;
OUTPUT;
Satisfaction_Factor='JobSatisfaction       ';
```

```
Value=JobSatisfaction;
OUTPUT;
Satisfaction_Factor='PercentSalaryHike          ';
Value=PercentSalaryHike;
OUTPUT;
Satisfaction_Factor='RelationshipSatisfaction           ';
Value=RelationshipSatisfaction;
OUTPUT;
Satisfaction_Factor='WorkLifeBalance         ';
Value=WorkLifeBalance;
OUTPUT;
RUN;

PROC ANOVA DATA=WORK.HR_NK2;
     class Satisfaction_Factor;
     Model Value = Satisfaction_Factor;
     means Satisfaction_Factor / hovtest welch tukey;
RUN;
```



The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for Value

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	10283
Error Mean Square	11.95742
Critical Value of Studentized Range	4.17038
Minimum Significant Difference	0.3761

Means with the same letter are not significantly different.

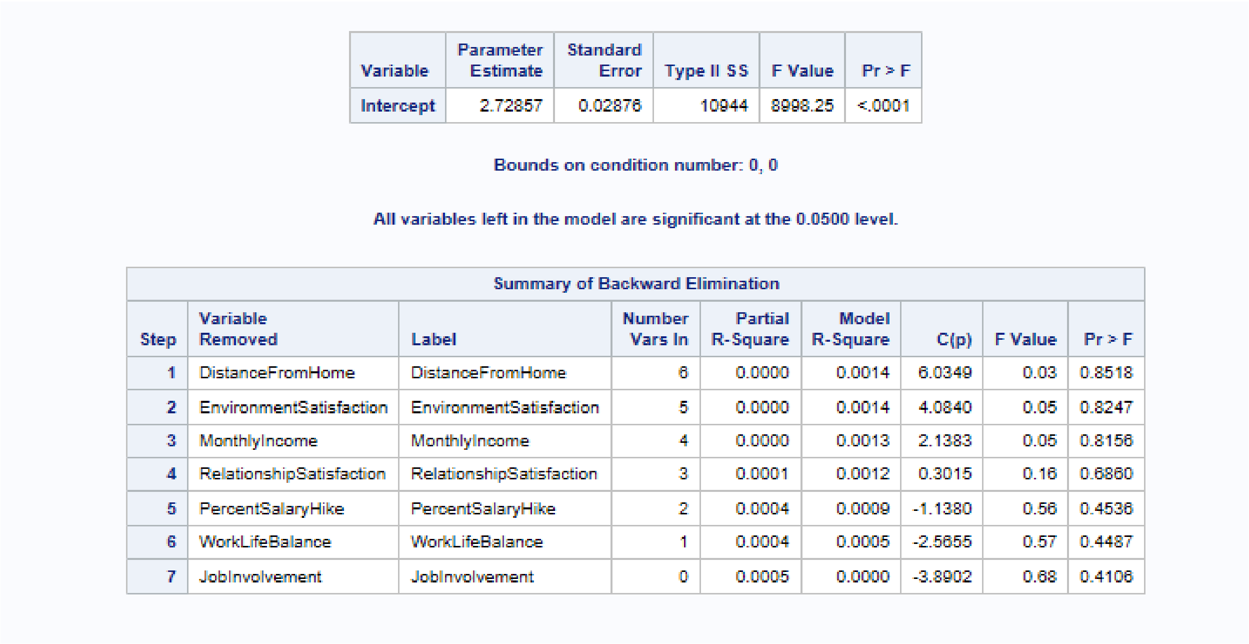
Tukey Grouping	Mean	N	Satisfaction_Factor
A	15.2095	1470	PercentSalaryHike
B	9.1925	1470	DistanceFromHome
C	2.7612	1470	WorkLifeBalance
C			
C	2.7299	1470	JobInvolvement
C			
C	2.7286	1470	JobSatisfaction
C			
C	2.7218	1470	EnvironmentSatisfaction
C			
C	2.7122	1470	RelationshipSatisfactio

- I separated variables related to Satisfaction of the employees and decided to run a Proc Anova on those to see if they are Homogenous or not and if they fall under same groupings or not and impact each other in any way.
- H0: Means of all factors related to Job Satisfaction are same

- H1: At least one of the means of the Variables is different.
- After running the Hovtest test and welch test, I saw that the p-value for both of them is <0.0001 which is less than 0.05 which shows us that the means are not homogeneous which basically means we are rejecting the null hypothesis and we are accepting the alternative hypothesis meaning that the means for at least one of them is different.
- After running the Tukey test, the results showed Percent Salary Hike, Distance from Home are the different ones out of all and Work Life Balance, Environment Satisfaction, Relationship Satisfaction and Job Involvement fall under same group “C” meaning they all have same means.

“Regression Model”

```
PROC REG DATA=WORK.HR_NK;
MODEL JobSatisfaction = MonthlyIncome DistanceFromHome WorkLifeBalance
EnvironmentSatisfaction RelationshipSatisfaction JobInvolvement
PercentSalaryHike / vif;
RUN;
PROC REG DATA=WORK.HR_NK;
MODEL JobSatisfaction = MonthlyIncome DistanceFromHome WorkLifeBalance
EnvironmentSatisfaction RelationshipSatisfaction JobInvolvement
PercentSalaryHike / SELECTION = Backward SLS=0.05 ADJRSQ;
RUN;
```



Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.72857	0.02876	10944	8998.25	<.0001

Bounds on condition number: 0, 0

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	DistanceFromHome	DistanceFromHome	6	0.0000	0.0014	6.0349	0.03	0.8518
2	EnvironmentSatisfaction	EnvironmentSatisfaction	5	0.0000	0.0014	4.0840	0.05	0.8247
3	MonthlyIncome	MonthlyIncome	4	0.0000	0.0013	2.1383	0.05	0.8156
4	RelationshipSatisfaction	RelationshipSatisfaction	3	0.0001	0.0012	0.3015	0.16	0.6860
5	PercentSalaryHike	PercentSalaryHike	2	0.0004	0.0009	-1.1380	0.56	0.4536
6	WorkLifeBalance	WorkLifeBalance	1	0.0004	0.0005	-2.5655	0.57	0.4487
7	JobInvolvement	JobInvolvement	0	0.0005	0.0000	-3.8902	0.68	0.4106

- After running a multiple linear regression model on variables that revolve around an employee’s satisfaction level in the company, my output tells me that all the VIFs are less than 8, meaning that there is multicollinearity between the variables. The overall p-value is <0.0001, which is less than 0.05 alpha level, meaning we can reject the null hypothesis and conclude that the slope is 0.

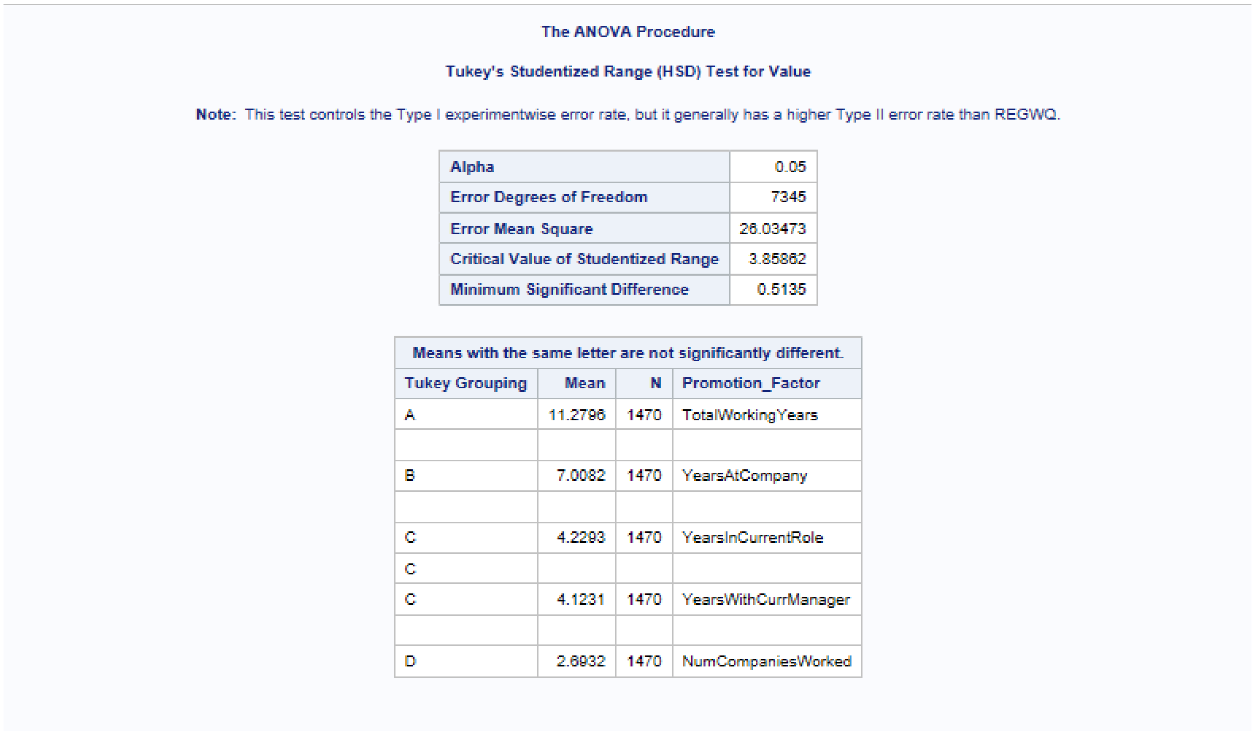
- My Conclusion is that Job Satisfaction of any employee is not dependent on any of the other independent variables.

“ANOVA TEST”

```
Data work.HR_NK3 (Keep=Promotion_Factor Value);
Set WORK.HR_NK;
Promotion_Factor = 'NumCompaniesWorked  ';
Value = NumCompaniesWorked;
OUTPUT;
Promotion_Factor='TotalWorkingYears ';
Value = TotalWorkingYears;
OUTPUT;
Promotion_Factor='YearsAtCompany   ';
Value=YearsAtCompany;
OUTPUT;
Promotion_Factor='YearsInCurrentRole    ';
Value=YearsInCurrentRole;
OUTPUT;
Promotion_Factor='YearsWithCurrManager    ';
Value=YearsWithCurrManager;
OUTPUT;
RUN;



PROC ANOVA DATA=WORK.HR_NK3;
     class Promotion_Factor;
     Model Value = Promotion_Factor;
     means Promotion_Factor / hovtest welch tukey;
RUN;
```



The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for Value

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	7345
Error Mean Square	26.03473
Critical Value of Studentized Range	3.85862
Minimum Significant Difference	0.5135

Means with the same letter are not significantly different.

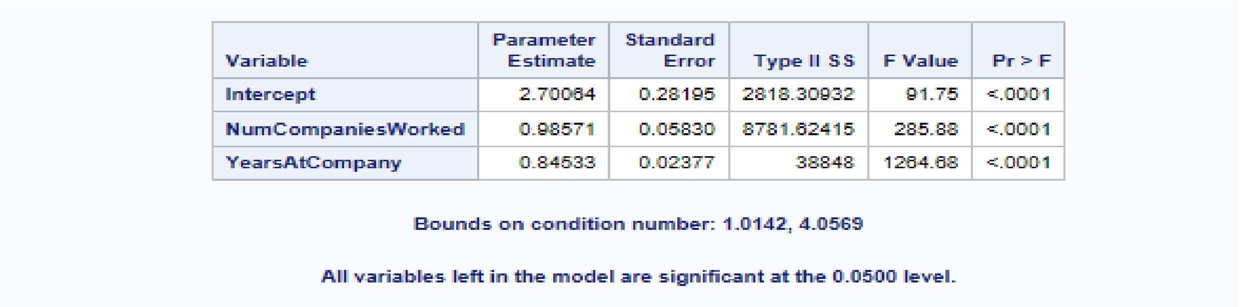
Tukey Grouping	Mean	N	Promotion_Factor
A	11.2796	1470	TotalWorkingYears
B	7.0082	1470	YearsAtCompany
C	4.2293	1470	YearsInCurrentRole
C			
C	4.1231	1470	YearsWithCurrManager
D	2.6932	1470	NumCompaniesWorked

- I separated variables that are somehow related to chances of Promotion of employees and decided to run a Proc Anova on those to see if they are Homogenous or not and if they fall under same groupings or not and impact each other in any way.
- H0: Means of all factors related to salary are same
- H1: At least one of the means of the Salary factors is different.
- After running the Hovtest test and welch test, I saw that the p-value for both of them is <0.0001 which is less than 0.05 which shows us that the means are not homogeneous which basically means we are rejecting the null hypothesis and we are accepting the alternative hypothesis meaning that the means for at least one of them is different.

- After running the Tukey test, the results showed that Total Working Years and Years at Company fall under different group then others whereas Num Companies Worked, Years in Current Role and Years With Curr Manager falls under same group "C" meaning they have same means.

"Regression Model"

```
PROC REG DATA=WORK.HR_NK;
MODEL TotalWorkingYears = NumCompaniesWorked YearsInCurrentRole
YearsAtCompany YearsWithCurrManager    / vif;
RUN;
PROC REG DATA=WORK.HR_NK;
MODEL TotalWorkingYears = NumCompaniesWorked YearsInCurrentRole
YearsAtCompany YearsWithCurrManager    / SELECTION = Backward SLS=0.05
ADJRSQ;
RUN;
```



Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.70064	0.28195	2818.30932	91.75	<.0001
NumCompaniesWorked	0.98571	0.05830	8781.62415	285.88	<.0001
YearsAtCompany	0.84533	0.02377	38848	1264.68	<.0001

Bounds on condition number: 1.0142, 4.0569

All variables left in the model are significant at the 0.0500 level.

- After running a multiple linear regression model on variables that are somehow responsible for promotion of employees, my output tells me that all the VIFs are less than 8, meaning that there is multicollinearity between the variables. The overall p-value is <0.0001, which is less than 0.05 alpha level, meaning we can reject the null hypothesis and conclude that the slope is 0.
- My Conclusion is that with the perceptive of Promotion for which Total working years was our dependent variables for which Num Companies Worked and Years at Company proved to be significant for our model.