# Descriptive statistics 1

* **Set up the purpose of analysis**:

I selected the data set – “BankLoan”, variable Income in thousands dollar. I want to know more about income of household, to understand client more deeply and better serve.

* **Analysis**:

Since data are whole information of bank, so I decided to see it as population.





* **Interpretation**:
* There are 850 clients taking loan service in the bank. The mean income is $46,700. The median income is $35,000. The minimum income is $13,000 and the maximum is $446,000. The range is $433,000. It means very high income range exist in this bank client.
* The standard deviation is $38,520. About 90% of household incomes are within one standard deviation. We can say 90% of household incomes are between $8,000 and $85,000.
* The skewness is 3.7; it is highly positively skewed (more than 2), so I would suggest to use the median ($35,000) as typical income.
* From the BoxPlot, I will suggest find out all the high outliers (35) and extremes (18). These are all high valuable clients, provide them better service, or consider offering them other kind of bank services. The clients whose income are higher than $100,000, can be considered as high value customers.

# Descriptive statistics 2

* **Set up the purpose of analysis**:

I selected the data set – “Diabetes\_cost”, and chosen the variable - COST. I want to see how much people paying for diabetes disease.

* **Analysis**:

I considered data as sample.





* **Interpretation**:
* There are 250 sample data in this data set. The mean diabetes treatment cost is $10,221. The median treatment cost is $8,029. The minimum treatment cost is $314 and the maximum is $53,359. The range is $53,045; it is very high range.
* The sample standard deviation is $8,555. The skewness is 1.25. It is about symmetric. It almost match Empirical Rule. 96% of diabetes treatment cost are within two standard deviation. We can say 96% of diabetes treatment cost are less than $27,330.
* The skewness is 1.25; it is almost symmetric, so I would suggest to use the mean ($10,221) as typical diabetes treatment cost.
* From the BoxPlot, there are 4 high and extreme outliers. These people may have higher age or influencing by other disease.

# Hypothesis test 1

* **Set up the purpose of analysis**:

I selected the data set – “EXEC SALARY”. I want to know if there is a salary difference between executive in board members and not board members. I am expecting no difference. The management and shareholder need to be separated.

* **Analysis**:

1. Generate the data:



1. State the Hypothesis:

*H*0: there is no salary difference.

*H*1: there is a salary difference if executives are board member.

1. Set the significance level : α =.05
2. Run the hypothesis test: compare two independent groups



**P-Value = 0.6336**. It is greater than significance level α =.05. So we fails to reject *H*0.

* **Conclusion**: there are no salary difference between the executive who is board member and not board member.

# Hypothesis test 2

* **Set up the purpose of analysis**:

I selected the data set – “WAGES”. I want to know if there is an income difference between the married and the unmarried. I think there should be difference; the unmarried can concentrate more on work, make more income. The married have more distraction to family, relatively less time on work, so less wage.

* **Analysis**:

1. Generate the data:



1. State the Hypothesis:

*H*0: there is no income difference.

*H*1: there is an income difference between the married and the unmarried.

1. Set the significance level : α =.05
2. Run the hypothesis test: compare two independent groups



**P-Value = 0.0127**. It is less than significance level α =.05. So we reject *H*0.

From **comparing the mean value above**, we can conclude that the married has higher wage than the unmarried.

* **Conclusion**: at the .05 significance level, there is an income difference between the married and the unmarried. The married workers have higher wages than the unmarried one. It is totally different than what I guessed in the beginning.

# ANOVA test 1

* **Set up the purpose of analysis**:

I selected the data set – “Car Insurance Claims”. I want to know if there are different number of claims within different age groups.

* **Analysis**:

1. Generate the data:



1. State the Hypothesis:

*H*0: within all age group, number of claims are same.

*H*1: at least two age groups have different number of claims.

1. Set the significance level : α =.05
2. Run the ANOVA test: one factor ANOVA



**ANOVA P-Value = 0.0012**. It is less than significance level α =.05. So we reject *H*0. **At least two age groups have different number of claims.**

From **Tukey output**, we can get:

* Age group ’17-20’ and Age group ’60+’ P-VALUE is .0292, less than significance level α =.05. We conclude Age group ’17-20’ and Age group ’60+’ have different number of claims. The Age group ‘60+’ has greater number of claims than Age group ’17-20’.
* Age group ’17-20’ and Age group ’35-39’ P-VALUE is .0303, less than significance level α =.05. We conclude Age group ’17-20’ and Age group ’35-39’ have different number of claims. The Age group ’35-39’ has greater number of claims than Age group ’17-20’.
* Age group ’17-20’ and Age group ’50-59’ P-VALUE is .0014, less than significance level α =.05. We conclude Age group ’17-20’ and Age group ’50-59’ have different number of claims. The Age group ’50-59’ has greater number of claims than Age group ’17-20’.
* Age group ’17-20’ and Age group ’40-49’ P-VALUE is .0001, less than significance level α =.05. We conclude Age group ’17-20’ and Age group ’40-49’ have different number of claims. The Age group ’40-49’ has greater number of claims than Age group ’17-20’.
* Age group ’21-24’ and Age group ’50-59’ P-VALUE is .0057, less than significance level α =.05. We conclude Age group ’21-24’ and Age group ’50-59’ have different number of claims. The Age group ’50-59’ has greater number of claims than Age group ’21-24’.
* Age group ’21-24’ and Age group ’40-49’ P-VALUE is .0003, less than significance level α =.05. We conclude Age group ’21-24’ and Age group ’40-49’ have different number of claims. The Age group ’40-49’ has greater number of claims than Age group ’21-24’.
* Age group ’25-29’ and Age group ’40-49’ P-VALUE is .0084, less than significance level α =.05. We conclude Age group ’25-29’ and Age group ’40-49’ have different number of claims. The Age group ’40-49’ has greater number of claims than Age group ’25-29’.
* Age group ’30-34’ and Age group ’40-49’ P-VALUE is .0214, less than significance level α =.05. We conclude Age group ’30-34’ and Age group ’40-49’ have different number of claims. The Age group ’40-49’ has greater number of claims than Age group ’30-34’.
* Age group ’60+’ and Age group ’40-49’ P-VALUE is .0441, less than significance level α =.05. We conclude Age group ’60+’ and Age group ’40-49’ have different number of claims. The Age group ’40-49’ has greater number of claims than Age group ’60+’.
* Age group ’35-39’ and Age group ’40-49’ P-VALUE is .0497, less than significance level α =.05. We conclude Age group ’35-39’ and Age group ’40-49’ have different number of claims. The Age group ’40-49’ has greater number of claims than Age group ’35-39’.
* **Conclusion**: at the .05 significance level, there are different number of claims within different age groups.

# ANOVA test 2

* **Set up the purpose of analysis**:

I selected the data set – “GOLFRBD”. I want to know if there are different driving distance between four different brands of golf balls and golf players.

* **Analysis**:

1. State the Hypothesis:

For the treatments:

*H*0: all driving distance of four brand balls are same.

*H*1: at least two brand of golf balls have different driving distance.

For the blocks:

*H*0: all golfers hit the golf balls to same distance.

*H*1: at least two golfers hit the golf ball in different distances.

1. Set the significance level : α =.05
2. Run the ANOVA test: Randomized blocks ANOVA



**Treatment ANOVA P-Value = 1.45E-11**. It is less than significance level α =.05. So we reject *H*0. **At least two brand of golf balls have different driving distance.**

**Blocks ANOVA P-Value = 4.50E-16**. It is less than significance level α =.05. So we reject *H*0. A**t least two golfers hit different driving distance.**

* **Conclusion**: Different brands of golf balls have different driving distance. Different golfers have different driving distance.

# Regression Analysis 1

**[Purpose]:** I chosen the data set – ‘EXEC SALARY’. I want to see how the executive’s salary is related to experience, education, number of employees supervised, assets and profits. In my guess:

* More experience more salary
* More education provide more knowledge to manage company, the more salary are given
* More employees more work to supervise, the more salary
* More assets to manage, the more salary
* Make more profits to company, the more salary are given

**[Analysis]:**

1. Define independent variable and dependent variable.

Salary: dependent variable

Experience, education, number of employees supervised, assets and profits: independent variables

1. Run the multiple regression model.



* Global test of the model P-Value is less than α =.05; it means at least one out of five independent variables is related to dependent variable – Salary.
* Profits’ P-Value is 0.9431, more than α =.05; we need to drop this independent variable.

1. Drop the independent variable - “Profits”, and run the regression model.



* Global test of the model P-Value is less than α =.05; it means at least one out of four independent variables is related to dependent variable – Salary.
* All four independent variables’ P-Value are less than α =.05; it means each independent variable is related to Salary.

1. Interpretation

* Global test of the model P-Value 2.36E-28 is less than α =.05; it means at least one out of four independent variables are related to dependent variable – Salary.
* Adjusted R2 = 0.747: **75%** of Salary (dependent variable) can be explained by experience, education, Empl Super and assets (independent variables).
* **For every unit increase** in years of work experience, **holding** the other three independent variables **constant**, the executive’s salary **is expected to increase by** $2,696.
* **For every one year increase** in education, **holding** the other three independent variables **constant**, the executive’s salary **is expected to increase by** $2,656.
* **For every unit increase** in number of employees supervised, **holding** the other three independent variables **constant**, the executive’s salary **is expected to increase by** $41.
* **For every million dollars increase** in assets, **holding** the other three independent variables **constant**, the executive’s salary **is expected to increase by** $245.

1. Test the assumptions.

* From bellow list, VIF <10, so there is no Multicollinearity problem.



* Residuals by predicted



* Normal Probability plot of residuals



**[Conclusion]:** the executive’s salary is related to experience, education, number of employees supervised and company assets. 75% of salary can be explained by experience, education, number of employees supervised and company assets. All these variables are positively related to the salary. Years of education and experience distribute more to increase salary.

# Regression Analysis 2

**[Purpose]:** I selected the data set – ‘WAGES’. I want to see how annual wage is related to years of education, years of work experience, south and Union. In my guess:

* More education more wage
* More work experience more wage
* Working in South part, has more opportunities, more wage
* Working in Union, gives more wage

**[Analysis]:**

1. Define independent variable and dependent variable.

Wage: dependent variable

Years of education, years of work experience, south and Union: independent variables

1. Run the multiple regression model.



* Global test of the model P-Value is less than α =.05; it means at least one out of four independent variables is related to dependent variable – wage.
* Union’s P-Value is 0.4543, more than α =.05; we need to drop this independent variable.

1. Drop the independent variable - “Union”, and run the regression model.



* Global test of the model P-Value is less than α =.05; it means at least one out of three independent variables is related to dependent variable – wage.
* South’s P-Value is 0.3781, more than α =.05; we need to drop this independent variable.

1. Drop the independent variable - “South”, and run the regression model.



* Global test of the model P-Value is less than α =.05; it means at least one out of two independent variables is related to dependent variable – wage.
* Two independent variable’s P-Value is all less than α =.05; it means these two are all related to dependent variable - Wage.

1. Interpretation

* Global test of the model P-Value 1.26E-06 is less than α =.05; it means at least one out of two independent variables is related to dependent variable – wage.
* Adjusted R2 = 0.229: 23% of wages (dependent variable) can be explained by years of education, years of work experience.
* **For every unit increase** in years of education, **holding** the experience **constant**, the annual wage **is expected to increase by** $3,307.
* **For every unit increase** in years of work experience, **holding** the variable - education **constant**, the annual wage **is expected to increase by** $388.

1. Test the assumptions.

* From bellow list, VIF <10, so there is no Multicollinearity problem.



* Residuals by predicted



* Normal Probability plot of residuals



**[Conclusion]:** the annual wage is related to years of education and years of work experience. It is not strong relation though.About 23% percent of wages can be explained by years of education and years of work experience. Years of education contribute more to the annual wage.

# Nonparametric Test 1 – Chi-Square

* **Set up the purpose of analysis**:

I selected the data set – “news”. And at α =.05 significance level, I want to know if there is relationship between two variables Location and TV Show. There are 3 kinds of TV Shows: national news, sports and comics. There are also 3 types of location provided: city, suburb and rural.

* **State the hypothesis**:

*H*0: there is no relationship between location and TV show.

*H*1: there is a relationship between location and TV show.

* **Analysis**:



**P-Value = 0.1190**. It is greater than significance level α =.05. So we fails to reject *H*0.

* **Conclusion**: at the .05 significance level there is no relationship location and TV show.

# Nonparametric Test 2

* **Set up the purpose of analysis**:

I selected the data set – “judges”. There are 10 judges scoring the softness of product A and product B. At α =.05 significance level, I want to know if there is difference between the softness of product A and product B.

* **State the hypothesis**:

*H*0: the softness of product A and product B is same.

*H*1: the softness of product A and product B is different.

* **Analysis**:

Since the 2 samples are **dependent**, so we will use Wilcoxon Signed Rank Test.



**P-Value = 0.0593**. It is greater than significance level α =.05. So we fails to reject *H*0.

* **Conclusion**: at the .05 significance level, the softness of product A and product B is same in given samples.

# Extra Credit 1 - Regression Model Test

**[Purpose]:** I have chosen the data set – ‘CALIRAIN’. I want to see how the average rain fall is related to the variables - Altitude, Latitude and Distance Coast. I don’t know how the altitude and latitude influence rain falls. But I know the more close to coast the more rain falls.

**[Analysis]:**

1. Define independent variable and dependent variable.

Average Rain Fall: dependent variable

Altitude, Latitude and Distance Coast: independent variables

1. Run the multiple regression model.



* Global test of the model P-Value is less than α =.05; it means at least one out of three independent variables is related to dependent variable – Rain Fall.
* All three independent variables’ P-Value are less than α =.05; it means each independent variable is related to Rain Fall.

1. Interpretation

* Global test of the model P-Value 2.21E-05 is less than α =.05; it means at least one out of three independent variables are related to dependent variable – Rainfall.
* Adjusted R2 = 0.554: **55%** of Rainfall (dependent variable) can be explained by Altitude, Latitude and Distance Coast (independent variables).
* **For every unit increase** in Altitude, **holding** the other two independent variables **constant**, the Average Rainfall **is expected to increase by** 0.0041.
* **For every unit increase** in Latitude, **holding** the other two independent variables **constant**, the Average Rainfall **is expected to increase by** **3.4511**.
* **For every unit increase** in Distance Coast, **holding** the other two independent variables **constant**, the Average Rainfall **is expected to decrease by** 0.1429.

1. Test the assumptions.

* From bellow list, VIF <10, so there is no Multicollinearity problem.



* Residuals by predicted



* Normal Probability plot of residuals



**[Conclusion]:** The average Rainfall is related to Altitude, Latitude and Distance Coast. 55% of Average Rainfall can be explained by Altitude, Latitude and Distance Coast. The Altitude and Latitude are positively related to the Average Rainfall. The Distance Coast is negatively related to the Average Rainfall.

# Extra Credit 2 – ANOVA Test

* **Set up the purpose of analysis**:

I selected the data set – “GOLFRBD”. I want to know if there are different driving distance between four difference brands of golf balls.

* **Analysis**:

1. State the Hypothesis:

*H*0: all driving distance of four brand balls are same.

*H*1: at least two brand of golf balls have difference driving distance.

1. Set the significance level : α =.05
2. Run the ANOVA test: one factor ANOVA



**ANOVA P-Value = 0.0372**. It is less than significance level α =.05. So we reject *H*0. **At least two brand of golf balls have difference driving distance.**

From **Tukey output**, we can get:

* Brand D and Brand C P-VALUE is .0057, less than significance level α =.05. We conclude Brand D and Brand C have difference driving distance.
* Brand A and Brand C P-VALUE is .0356, less than significance level α =.05. We conclude Brand A and Brand C have difference driving distance.
* **Conclusion**: at the .05 significance level, there are different driving distance between four difference brands of golf balls. Brand A ball and Brand C ball has different driving distance. Brand D ball and Brand C ball has different driving distance.