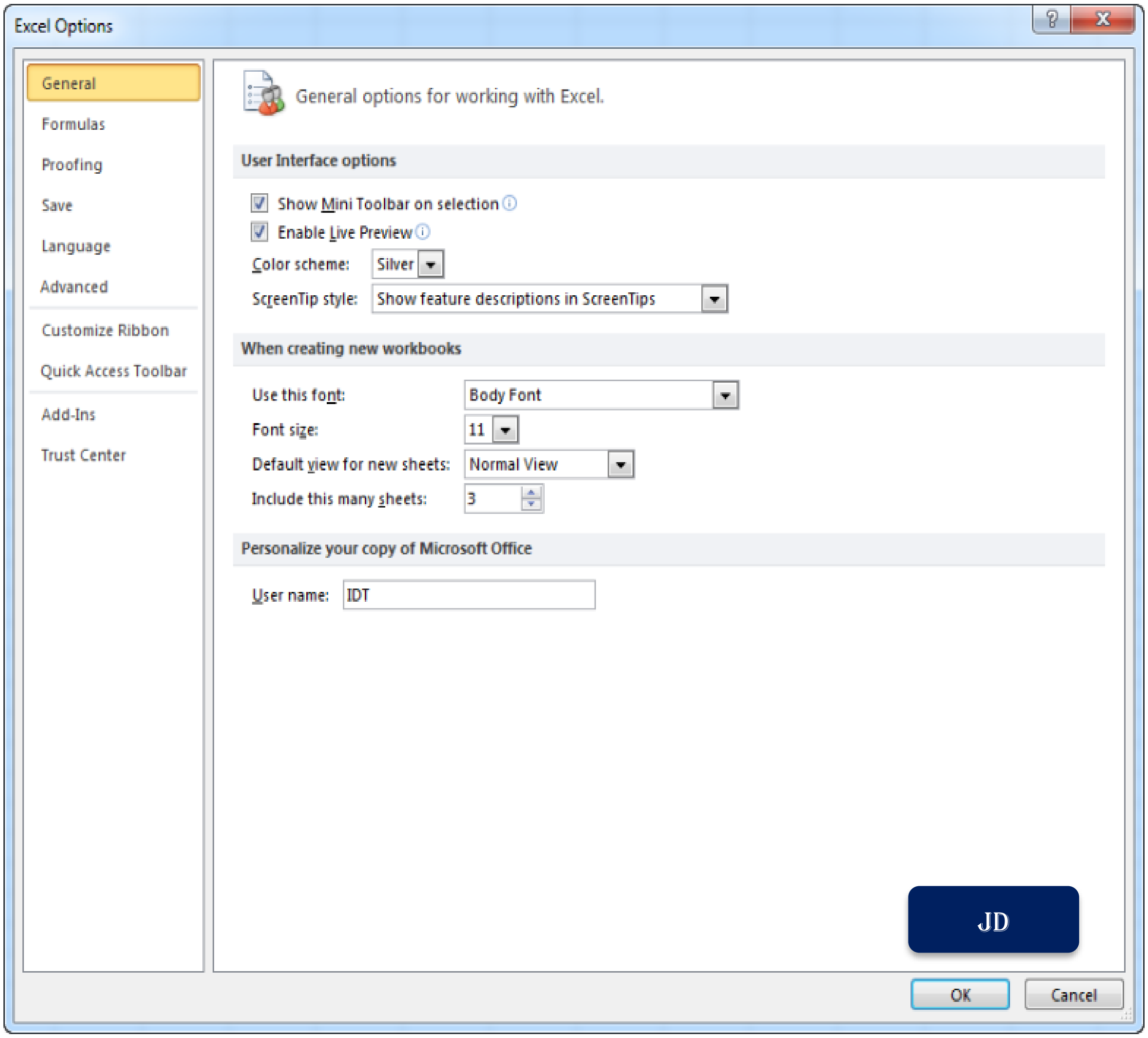
LINEAR REGRESSION

JD

To use Excel for Statistical Analysis, we will be using the Excel Data Analysis ToolPak Add-in. To use the Data Analysis Add-in, click on Data in the ribbon bar, then you will see the Data Analysis tool. But if you can’t find it there, **don’t worry** because we will be activating it right away.

To activate the Data Analysis Tool, go to File □
Then Click on **Options** □ **Add-ins**. An Excel Options window will appear as shown here



General options for working with Excel.

User Interface options

☑ Show Mini Toolbar on selection

☑ Enable Live Preview

Color scheme: Silver

ScreenTip style: Show feature descriptions in ScreenTips

When creating new workbooks

Use this font: Body Font

Font size: 11

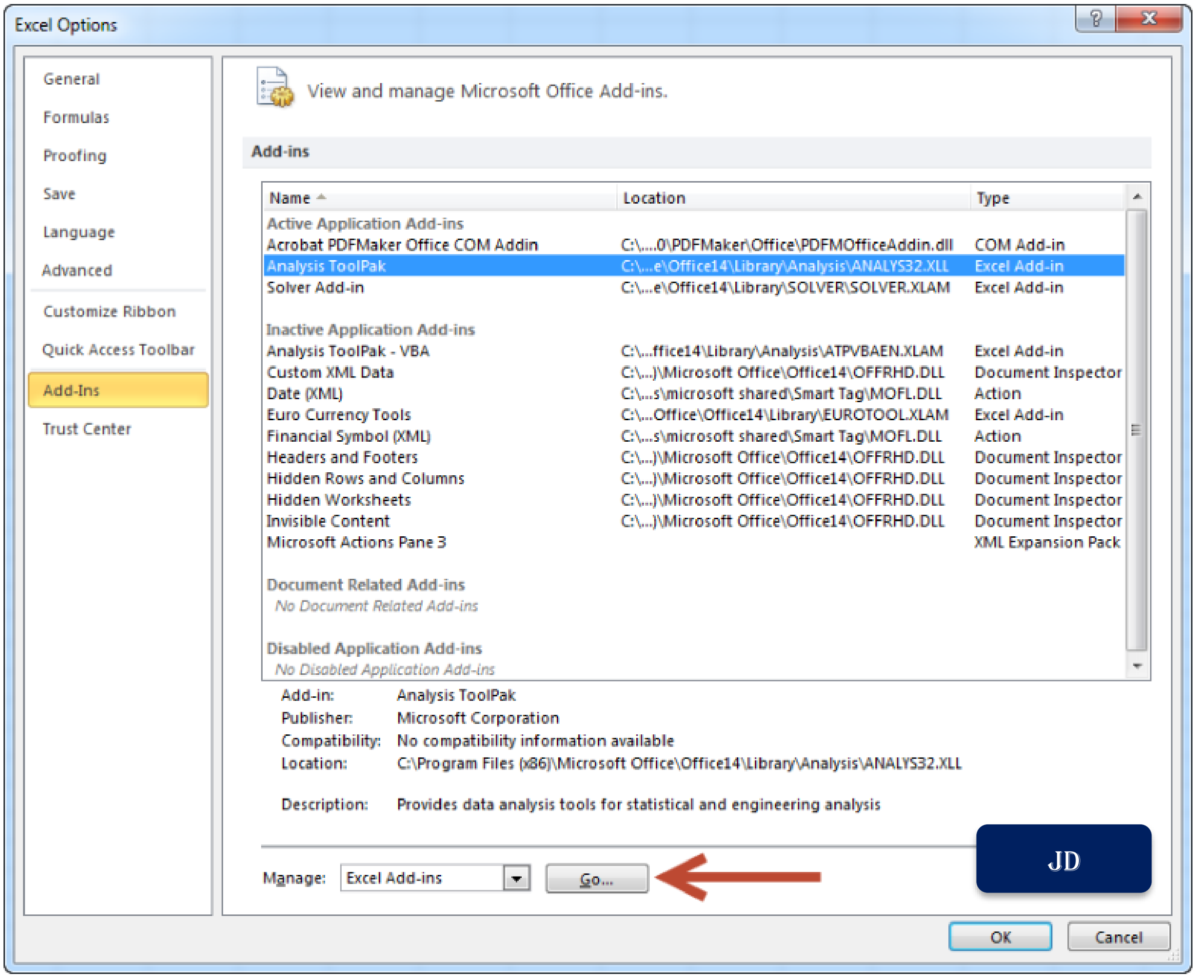
Default view for new sheets: Normal View

Include this many sheets: 3

Personalize your copy of Microsoft Office

User name: IDT

You will see **Add-ins** at the lower part of the Excel Options, Click the Add-ins. Then, be sure to click on the **Analysis ToolPak** and click on **Go**.



View and manage Microsoft Office Add-ins.

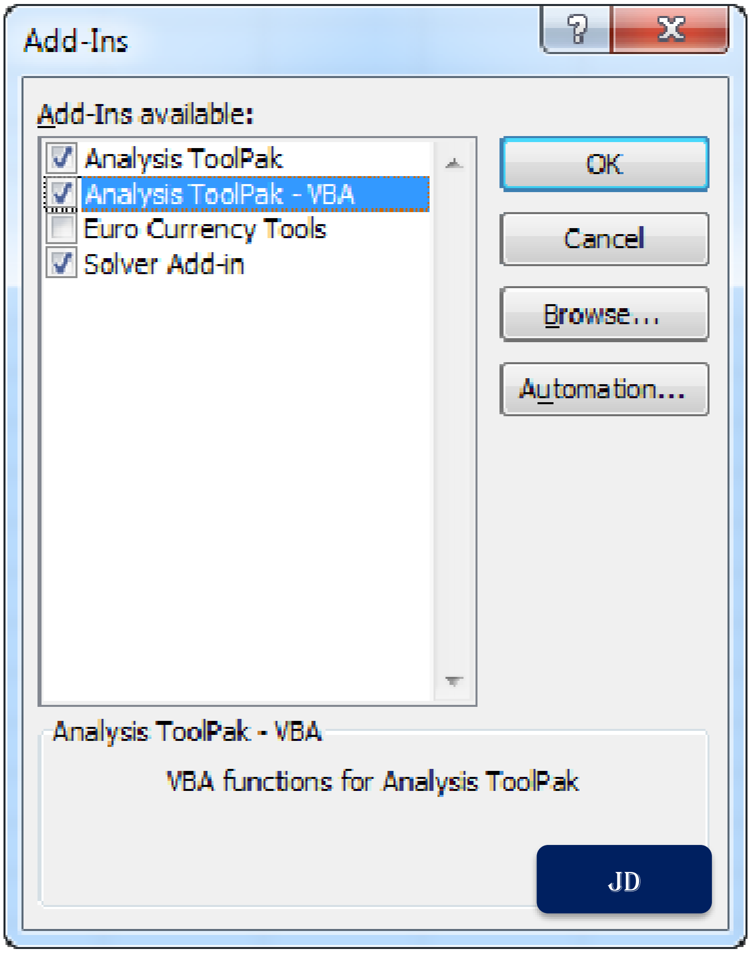
Add-ins

Name	Location	Type
Active Application Add-ins		
Acrobat PDFMaker Office COM Addin	C:\...0\PDFMaker\Office\PDFMOfficeAddin.dll	COM Add-in
Analysis ToolPak	C:\...e\Office14\Library\Analysis\ANALYS32.XLL	Excel Add-in
Solver Add-in	C:\...e\Office14\Library\SOLVER\SOLVER.XLAM	Excel Add-in
Inactive Application Add-ins		
Analysis ToolPak - VBA	C:\...ffice14\Library\Analysis\ATPVBAEN.XLAM	Excel Add-in
Custom XML Data	C:\...)\Microsoft Office\Office14\OFFRHD.DLL	Document Inspector
Date (XML)	C:\...s\microsoft shared\Smart Tag\MOFL.DLL	Action
Euro Currency Tools	C:\...Office\Office14\Library\EUROTOOL.XLAM	Excel Add-in
Financial Symbol (XML)	C:\...s\microsoft shared\Smart Tag\MOFL.DLL	Action
Headers and Footers	C:\...)\Microsoft Office\Office14\OFFRHD.DLL	Document Inspector
Hidden Rows and Columns	C:\...)\Microsoft Office\Office14\OFFRHD.DLL	Document Inspector
Hidden Worksheets	C:\...)\Microsoft Office\Office14\OFFRHD.DLL	Document Inspector
Invisible Content	C:\...)\Microsoft Office\Office14\OFFRHD.DLL	Document Inspector
Microsoft Actions Pane 3		XML Expansion Pack
Document Related Add-ins		
No Document Related Add-ins		
Disabled Application Add-ins		
No Disabled Application Add-ins		

Add-in:	Analysis ToolPak
Publisher:	Microsoft Corporation
Compatibility:	No compatibility information available
Location:	C:\Program Files (x86)\Microsoft Office\Office14\Library\Analysis\ANALYS32.XLL
Description:	Provides data analysis tools for statistical and engineering analysis

Manage: Excel Add-ins ▾ Go...

After clicking on **Go,** you will see this pop-up, then you must make sure to check Analysis ToolPak, Analysis ToolPak VBA, and the Solver Add-in and click on **OK.**



Add-Ins available:

☑ Analysis ToolPak
☑ Analysis ToolPak - VBA
□ Euro Currency Tools
☑ Solver Add-in

Analysis ToolPak - VBA

VBA functions for Analysis ToolPak

Once you have the Add-ins in place, you are ready to get started. You will then see the Data Analysis Tab on the data in the Ribbon bar.

Regression Analysis

In statistical modeling, regression analysis is used to determine the relationships between two or more variables:

- Dependent variable (y-axis variable or predicted variable) is the main factor you are trying to understand and predict.
- Independent variables (x-axis variable or explanatory variables, or predictors) are the factors that might affect or influence the dependent variable (y-axis).

Regression analysis helps us understand the relationship between two variables i.e. how the dependent variable (y-axis) changes when one of the independent variables (x-axis) varies or increases/decreases and allows you to mathematically determine which of those variables really has an impact.

Technically, a regression analysis model is based on the sum of squares (R^2), which is a mathematical way to find the dispersion of data points. The goal of a model is to get the smallest possible sum of squares (R^2) and draw a line that comes closest to the data.

Types of Linear Regression

In statistics, there are two different types of Linear Regression

- Simple linear regression
- Multiple linear regression.

Simple linear regression deals with and models the relationship between a dependent variable and one independent variable using a linear function. But If you use two or more explanatory variables (x1, x2, x3… xn) to predict the dependent variable (y), you are dealin**g** with **multiple linear regression**.

Furthermore, If the dependent variable (y) is modeled as a nonlinear function because the data relationships do not follow a straight line, then it is nonlinear regression and we can use other regression models like polynomial regression, etc.

The focus of will be on a simple linear regression.

To begin, lets add our data in the first two columns of the Excel Worksheet (X variable first). We can decide to remove the headers if we like, but for proper arrangement, lets add the headers.

Here is an example,

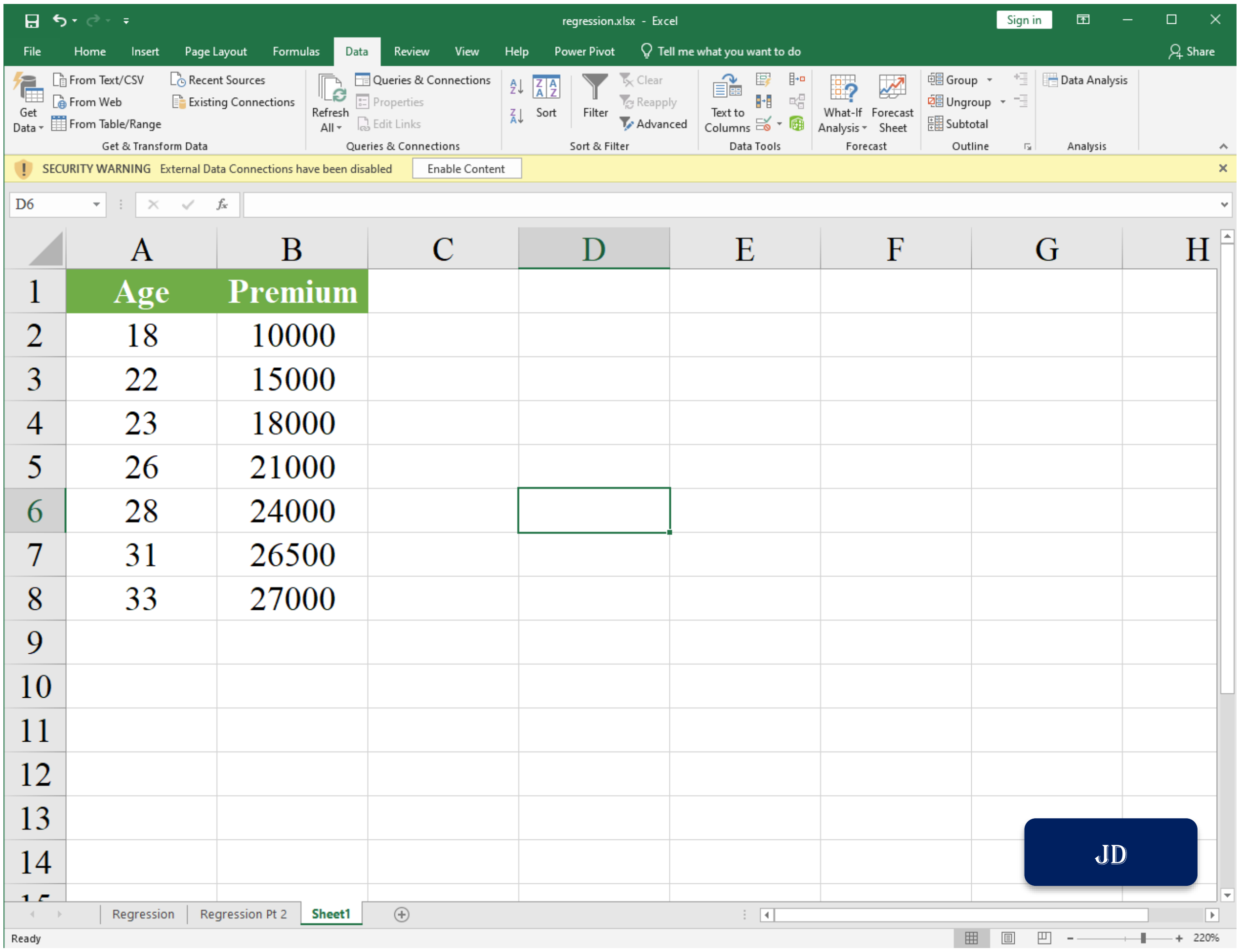
There is a dataset that consists of the age of staffs (x variable) and their Salary range (y variable). So, we will fit a Regression model that will help us predict the salary (y variable or outcome) vs the age (x variable or explanatory).

Don’t get confused

X variable – Independent/explanatory variable/predictors i.e. It explains our y variable and they are independent and also, they are what we use to predict the y variable.

Y variable – Dependent/Outcome variable i.e., this is dependent on our x variable. Without the X-variable, there is no y variable.

So, if I’m using all of this term in this book, that’s what I’m referring to, very simple, right?



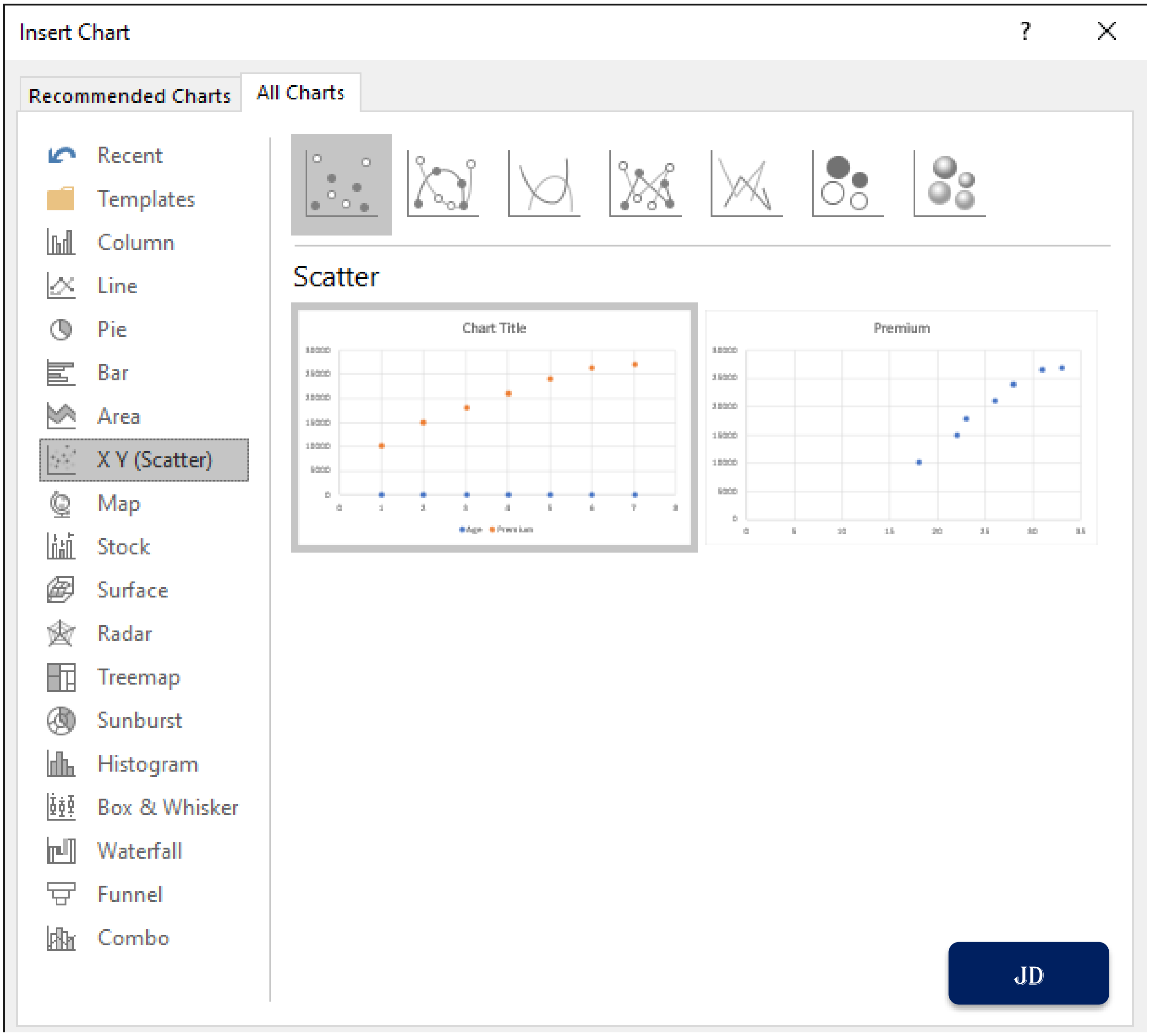
SECURITY WARNING External Data Connections have been disabled

Age	Premium
18	10000
22	15000
23	18000
26	21000
28	24000
31	26500
33	27000

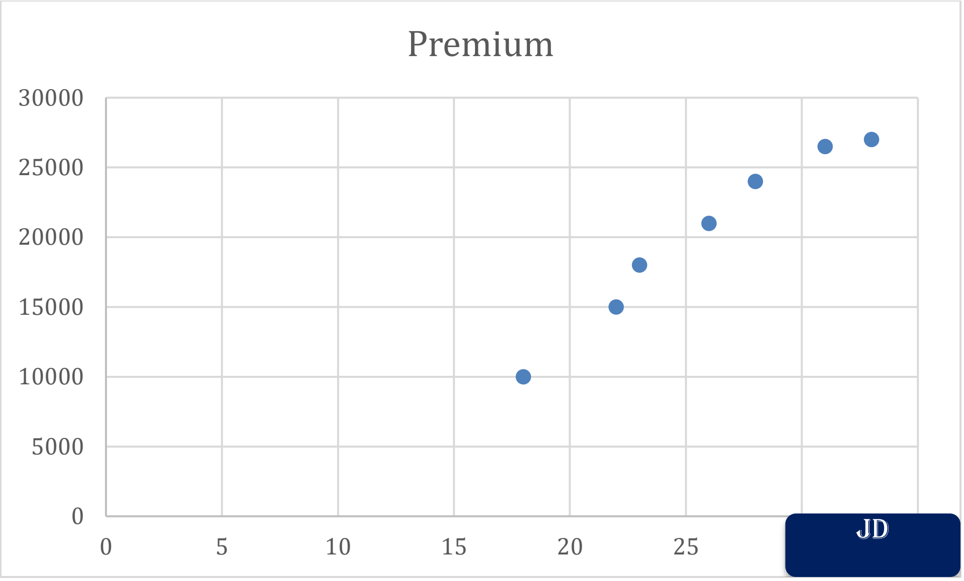
Looking at the above data, our dependent variable (y variable) is the Premium column while the independent variable (x variable) is the Age column.

Once you have the data imported, it is a good idea to create a Scatterplot when working with a Simple Linear Regression i.e. **one x variable** and **one y variable**.

Select/Highlight all of your data and click on **Insert tab**. Go to the recommended charts and click on All charts to choose a Scatterplot from the Insert tab.



After selecting the scatterplot, you will have a chart like this below. OK!



We can do some formatting to the chart:

- Give your plot a title.
- Click chart elements (the green plus sign) at the right upper part of your chart in excel, to add axis titles.
- Right click on an Axis to choose specific values.
- Right click on the points to change their color.

Let's move straight to Linear Regression. To add the regression line, never forget the formula which is y = mx + c where y is the outcome (dependent variable), m is the coefficient, x is the predictor (independent variable) and c is the intercept.

So, what does all of these things mean, we will get to it very shortly. OK!
Now, to add the regression line, we click on the chart element (green plus sign) at the right upper part of the chart and add the trendline. You will also see an arrow in front of the trend line, click on it and this will show some options. Click on the more options to have the full view.

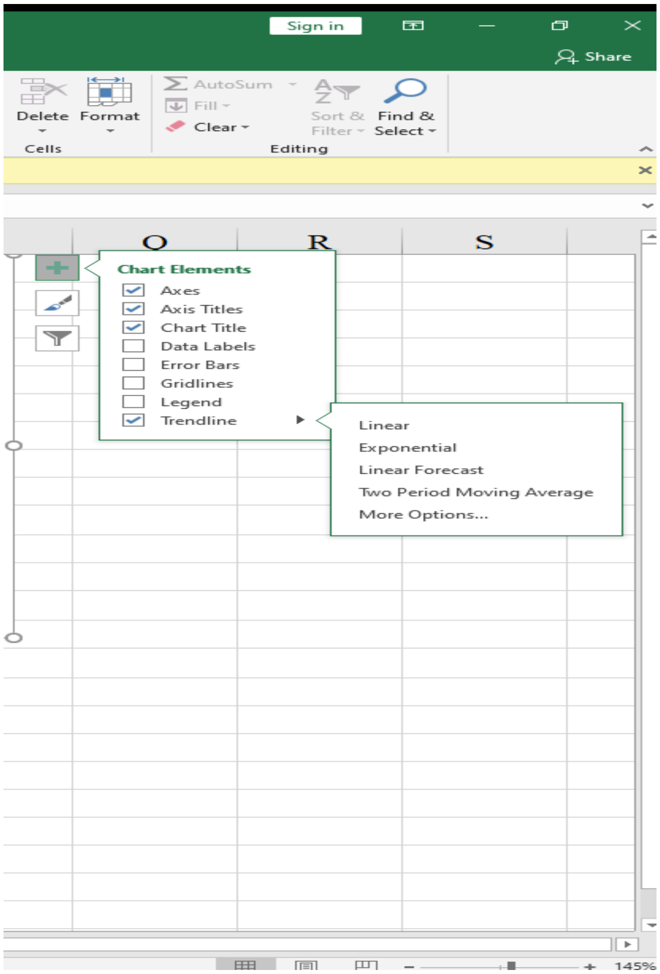
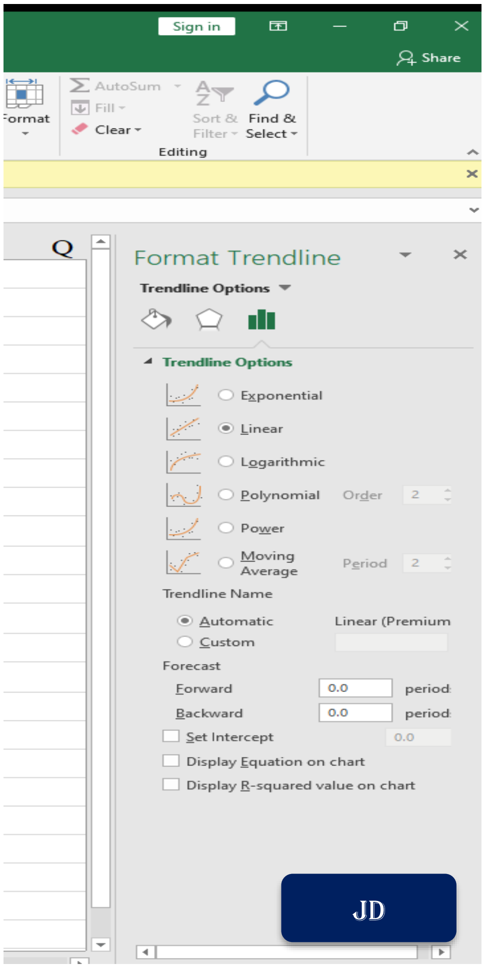


Chart Elements

- ☑ Axes
- ☑ Axis Titles
- ☑ Chart Title
- ☐ Data Labels
- ☐ Error Bars
- ☐ Gridlines
- ☐ Legend
- ☑ Trendline



Format Trendline

Trendline Options

Trendline Options

- ○ Exponential
- ● Linear
- ○ Logarithmic
- ○ Polynomial Order 2
- ○ Power
- ○ Moving Average Period 2

Trendline Name

- ● Automatic Linear (Premium
- ○ Custom

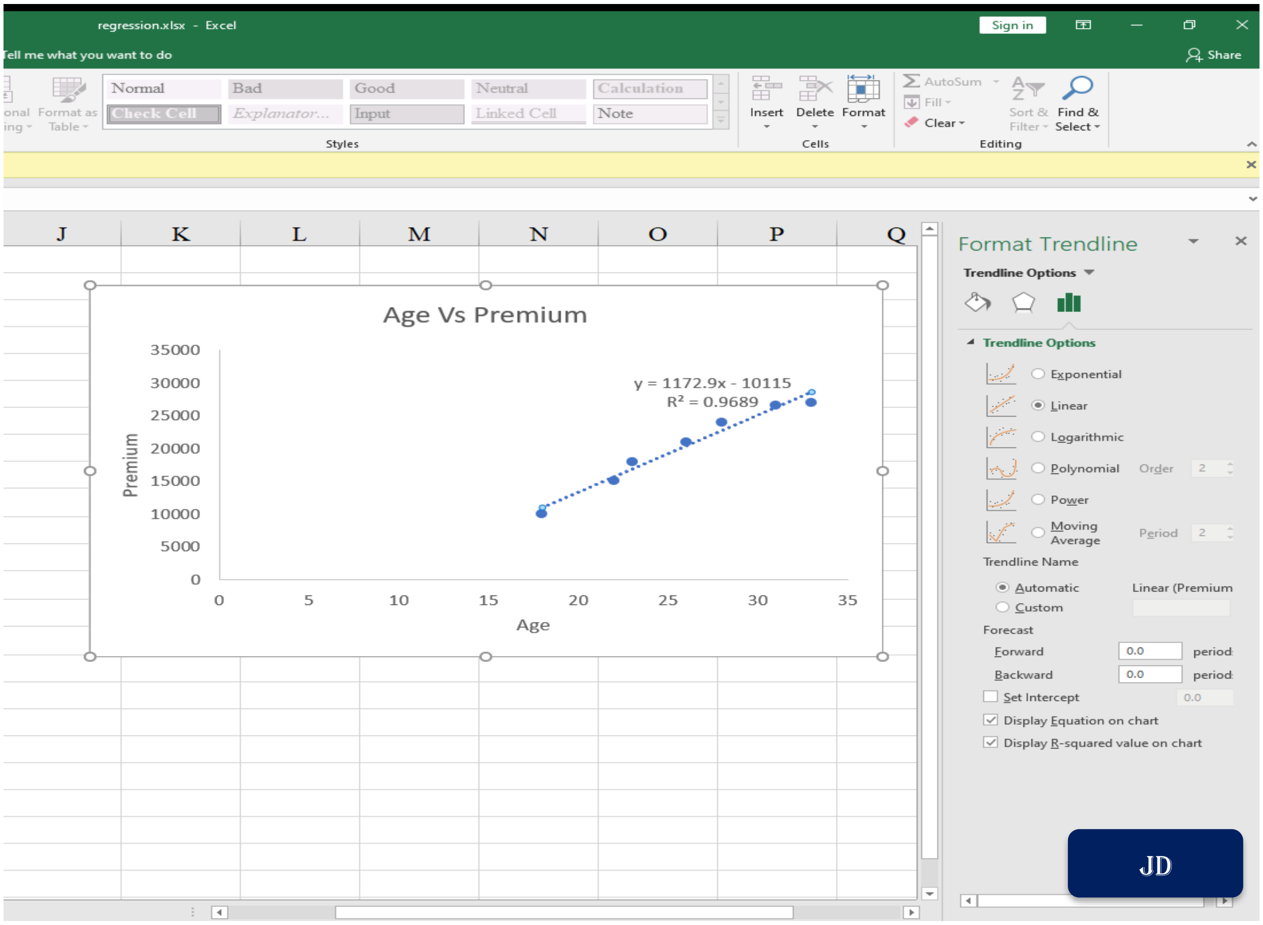
Forecast

Forward 0.0 period

Backward 0.0 period

- ☐ Set Intercept 0.0
- ☐ Display Equation on chart
- ☐ Display R-squared value on chart

Here, we will want to set the equation to our chart to know the relationship between the two variables. Check the **Display Equation on chart** and **Display R-squared value on chart.**
Once you check the two, you will have them on your chart as below.



Format Trendline

Trendline Options

Trendline Options

○ Exponential

● Linear

○ Logarithmic

○ Polynomial Order 2

○ Power

○ Moving Average Period 2

Trendline Name

● Automatic Linear (Premium

○ Custom

Forecast

Forward 0.0 period

Backward 0.0 period

□ Set Intercept 0.0

☑ Display Equation on chart

☑ Display R-squared value on chart

We now have the equation on the chart which is **y = 1172.9x – 10115** and **R squared value = 0.9689**

Explanation:

- y = 1172.9x – 10115. This means that at x = 0, the value of y will be -10115. This can help us to calculate a new value of x.
- R Square. It is the Coefficient of Determination, which is used as an indicator of the goodness of fit. It shows how many points fall on the regression line. The R2 value is calculated from the total sum of squares, more precisely, it is the sum of the squared deviations of the original data from the mean.

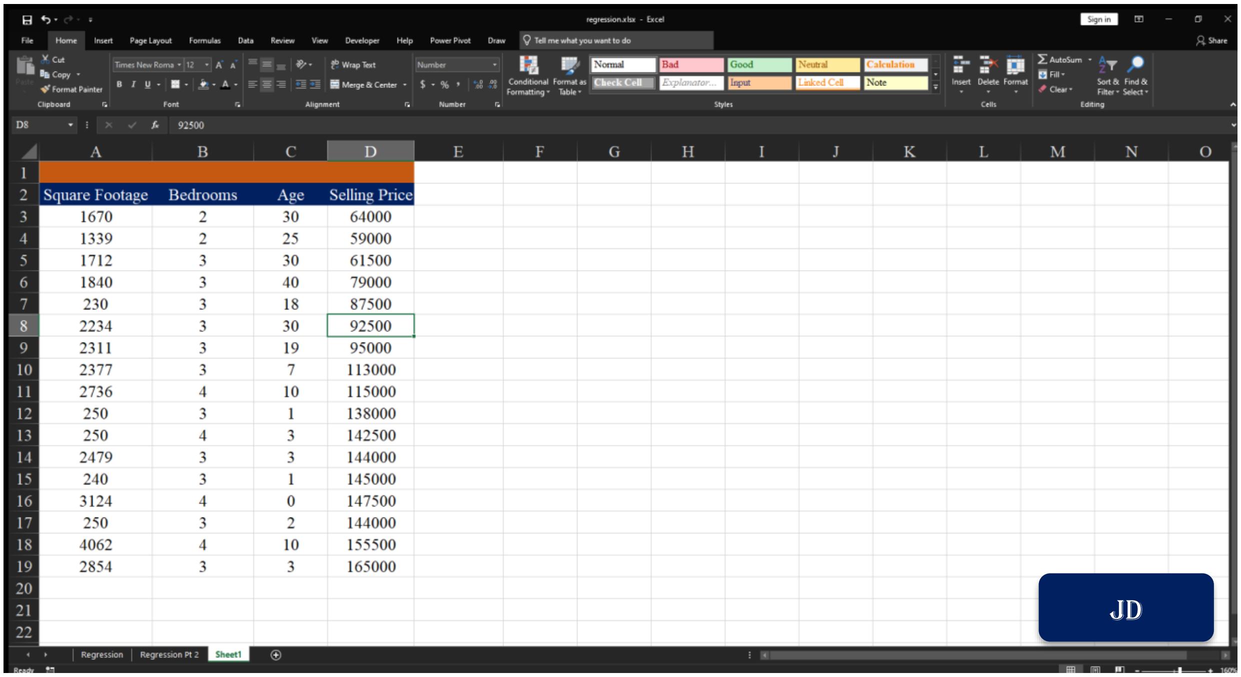
In our example, R2 is 0.97 (rounded to 2 digits), which is perfectly good. It means that 97% of our values fit the regression analysis model. In other words, 97% of the dependent variables (y-values) are explained by the independent variables (x-values). Generally, R Squared of 97% or more is considered a good fit.

Multiple Regression

In this part, we will be covering the basics of multiple linear regression. The following data gives us the selling price, square footage, number of bedrooms, and age of house (in years) that have sold in a neighborhood, let's say probably in the past four/five months.

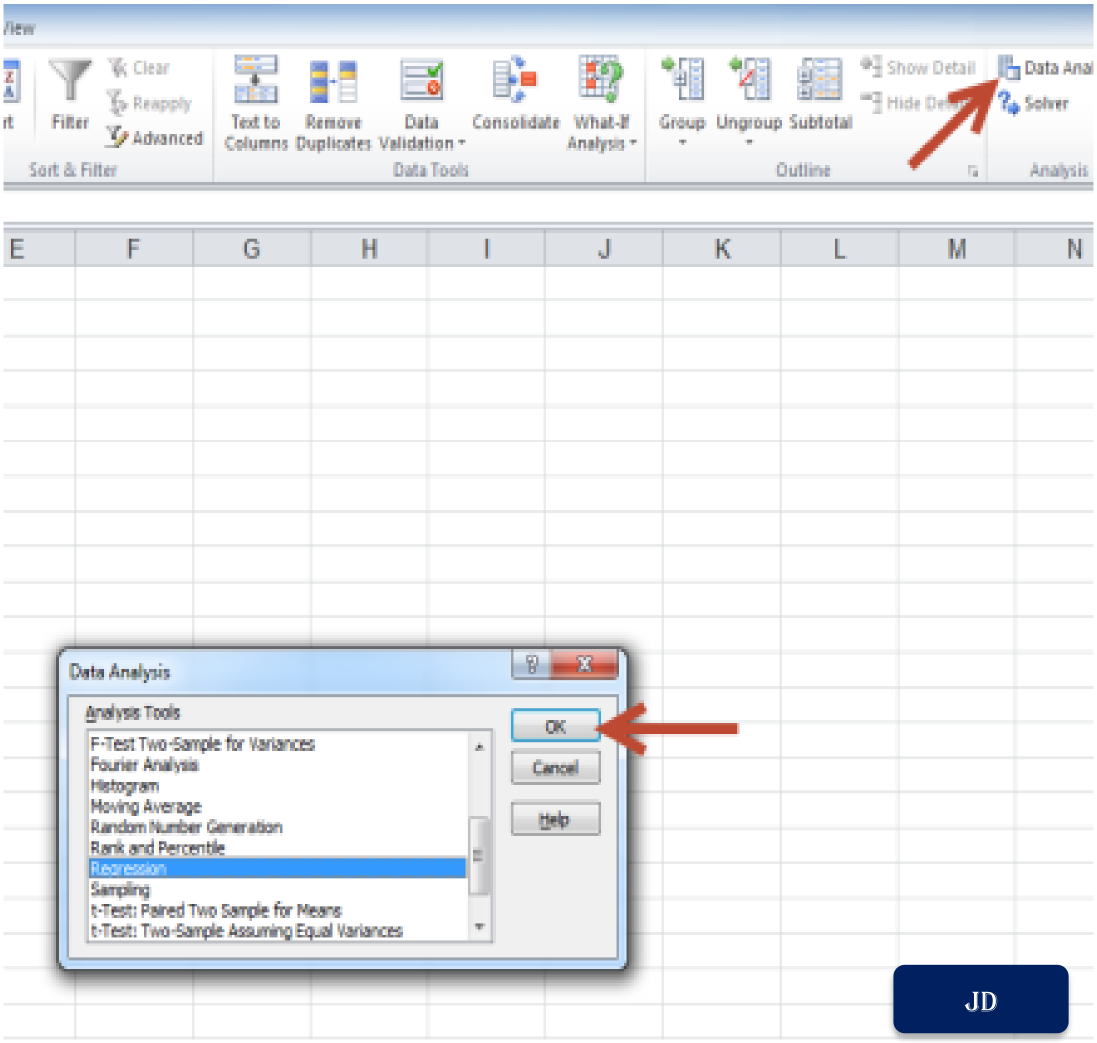
We need to develop a multiple regression model to predict the selling price based on each of the individual factors and determine which one is the best model. Then, we will develop a model that will help us predict the selling price of a house based on the square footage, number of bedrooms, and age and will discuss if all three variables should be included and if it is a better model.

Enter or copy the data into a blank Excel spreadsheet as shown here.

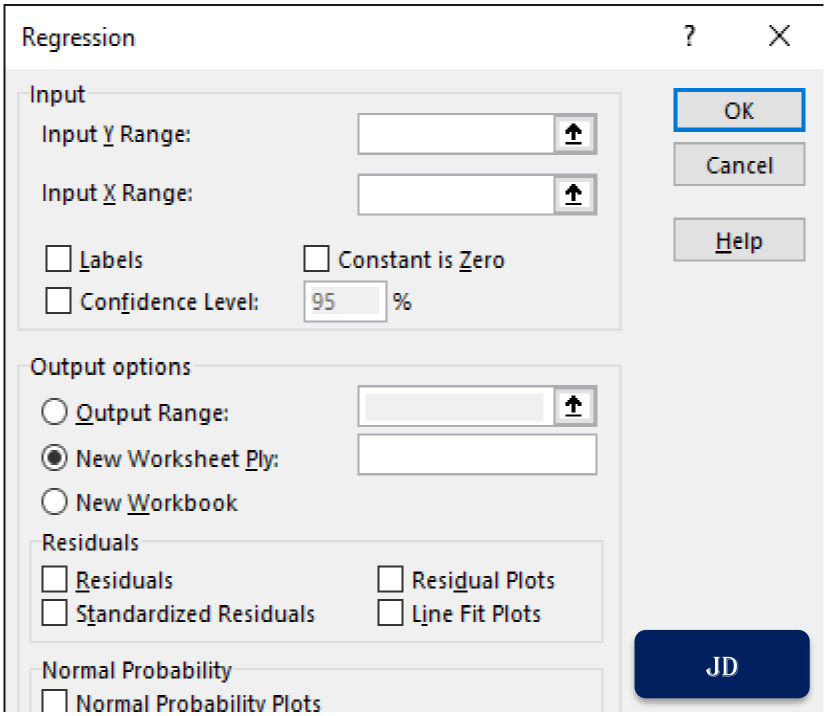


Square Footage	Bedrooms	Age	Selling Price
1670	2	30	64000
1339	2	25	59000
1712	3	30	61500
1840	3	40	79000
230	3	18	87500
2234	3	30	92500
2311	3	19	95000
2377	3	7	113000
2736	4	10	115000
250	3	1	138000
250	4	3	142500
2479	3	3	144000
240	3	1	145000
3124	4	0	147500
250	3	2	144000
4062	4	10	155500
2854	3	3	165000

Click on the Data tab and Data Analysis tool and, in the Data Analysis pop-up window, scroll down and select Regression and click OK.



Click in the box for Input Y Range and this is going to be our dependent variable, or in this case, the selling price, so highlight cells E2-E19.



Input

Input Y Range:

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

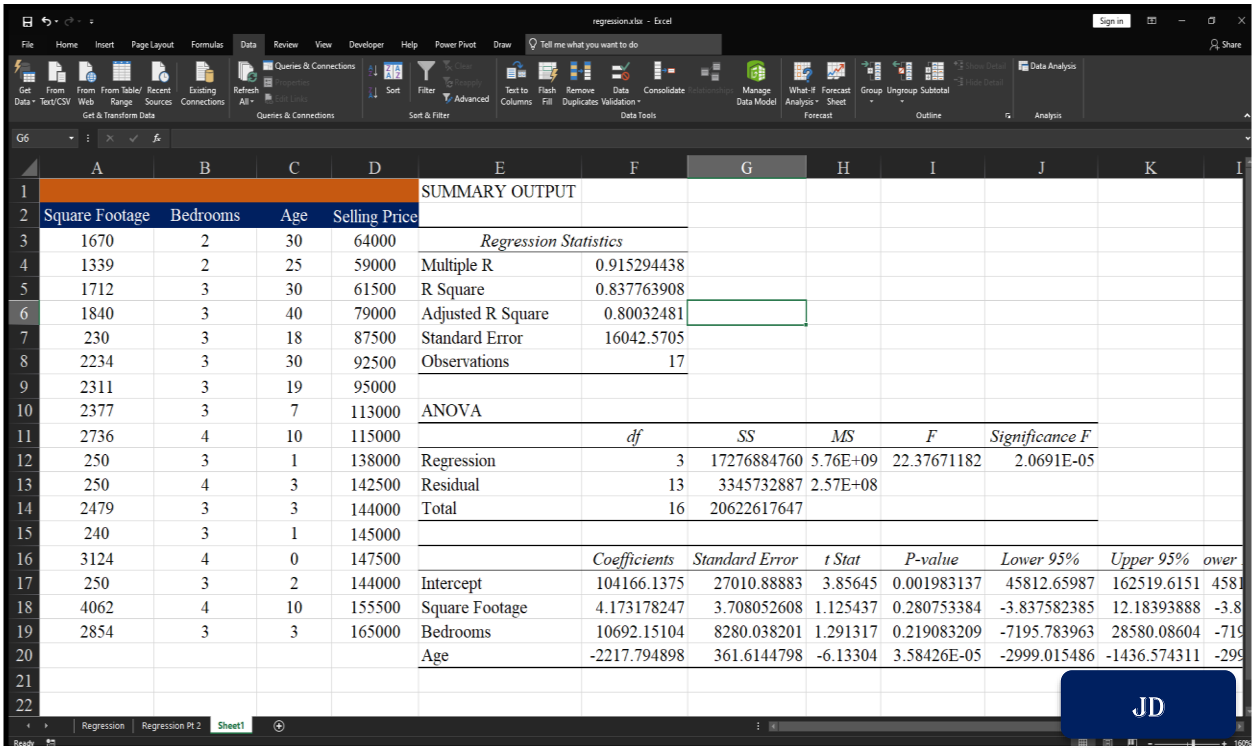
□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

- Our independent variable will be square footage, so click in the box for Input X Range and select cells A2-19. Be sure that the box is ticked next to Labels and select the Output Range as E1.
- Click OK. This will put the summary output next to our data table.



	A	B	C	D	E	F	G	H	I	J	K	L
1					SUMMARY OUTPUT							
2	Square Footage	Bedrooms	Age	Selling Price								
3	1670	2	30	64000	*Regression Statistics*							
4	1339	2	25	59000	Multiple R	0.915294438						
5	1712	3	30	61500	R Square	0.837763908						
6	1840	3	40	79000	Adjusted R Square	0.80032481						
7	230	3	18	87500	Standard Error	16042.5705						
8	2234	3	30	92500	Observations	17						
9	2311	3	19	95000								
10	2377	3	7	113000	ANOVA							
11	2736	4	10	115000		*df*	*SS*	*MS*	*F*	*Significance F*		
12	250	3	1	138000	Regression	3	17276884760	5.76E+09	22.37671182	2.0691E-05		
13	250	4	3	142500	Residual	13	3345732887	2.57E+08				
14	2479	3	3	144000	Total	16	20622617647					
15	240	3	1	145000								
16	3124	4	0	147500		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower*
17	250	3	2	144000	Intercept	104166.1375	27010.88883	3.85645	0.001983137	45812.65987	162519.6151	4581
18	4062	4	10	155500	Square Footage	4.173178247	3.708052608	1.125437	0.280753384	-3.837582385	12.18393888	-3.8
19	2854	3	3	165000	Bedrooms	10692.15104	8280.038201	1.291317	0.219083209	-7195.783963	28580.08604	-719
20					Age	-2217.794898	361.6144798	-6.13304	3.58426E-05	-2999.015486	-1436.574311	-299
21												
22												

Under the Regression Statistics

- **Multiple R** – the correlation coefficient – notes the strength of the relationship – in this case, 0.91529 – a pretty strong positive relationship.
- **R squared** – the amount of variability in the dependent variable explained by the independent variable(s). In this case, 0.8377 – again, a pretty strong number – almost 84% of the variability in purchase price is explained by the independent variables.
- **Adjusted R squared** – We use this when you have more than one independent variable and have adjusted the R squared value for the number of independent variables.

Under the ANOVA Tables

- **Significance F** – this test the significance of the overall model. We look for this to be less than 0.05. If it is less than 0.05, we can reject the null hypothesis and determine that the model is statistically valid. In this case, it's 0.000206, so we have a valid model.
- **Intercept Coefficient** – this is the intercept for our line if we were to plot it out. With X as zero (0), this is where the line crosses the Y axis. The value is 104,116.
- **X Coefficient** – this is the coefficient for our independent variable for the linear equation. It is the slope of our line or the amount that our dependent variable changes for every $1 change in our independent variable. For every increase in square footage by one, our price will change by this amount, 4.1731, For every increase in bedroom by one, our price will change by this amount, 10,692 and age by -2,222.
- **X P-Value** – this test the significance of the variable. We look for this to be less than 0.05. If it less than 0.05, we can reject the null hypothesis and determine that the variable is statistically significant.

So, again we can use the multiple linear regression on each of the columns i.e. square footage vs price, bedrooms vs price and age vs price and then compare to select which one is the best predictor for the house price. Checking the highest Multiple R, highest R-Squared.

This will be the end of Linear Regression.