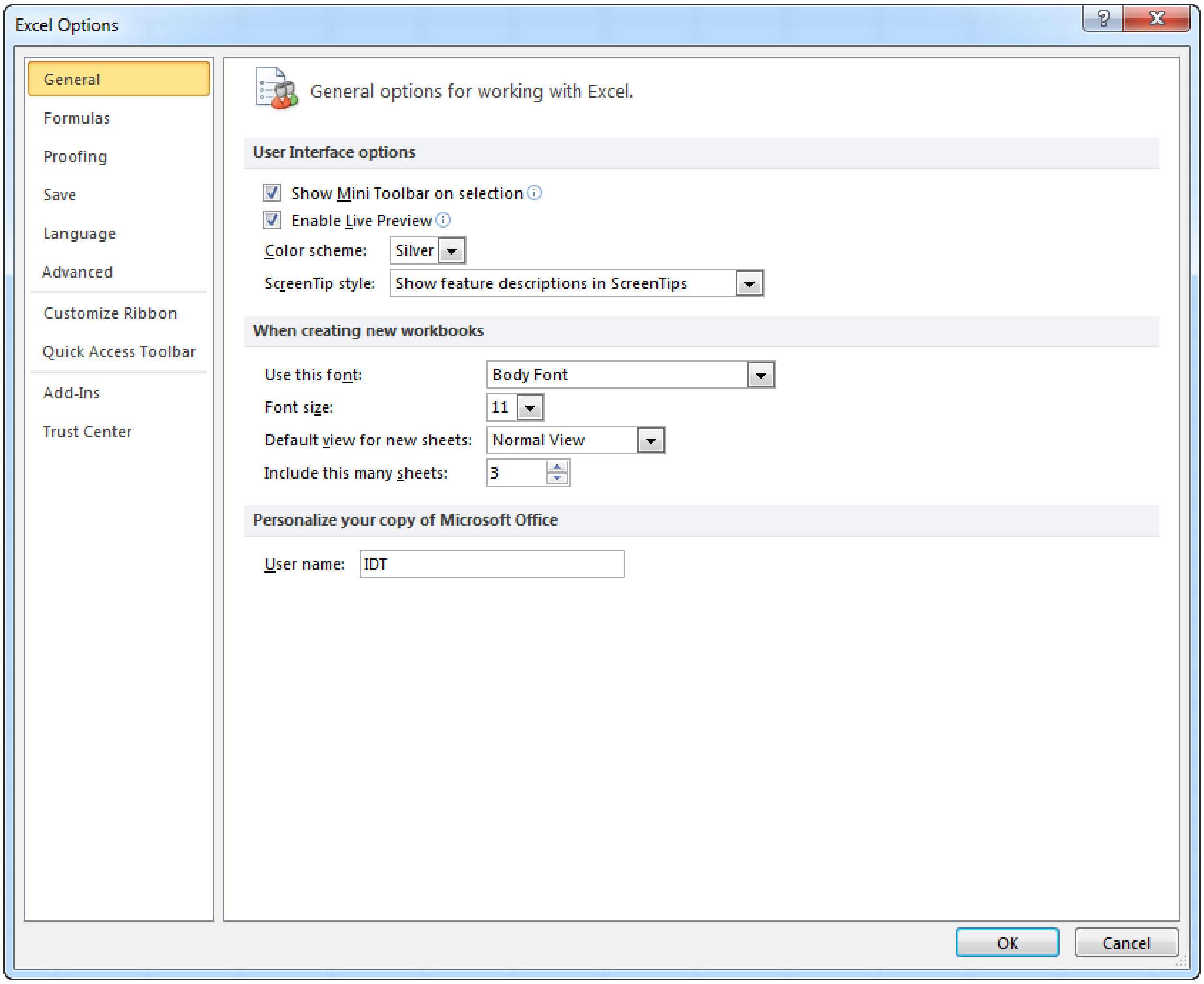
Simple Linear and Multiple Regression

In this tutorial, we will be covering the basics of linear regression, doing both simple and multiple regression models. 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 in the past six months.

Selling Price	Square Footage	Bedrooms	Age
64000	1670	2	30
59000	1339	2	25
61500	1712	3	30
79000	1840	3	40
87500	2300	3	18
92500	2234	3	30
95000	2311	3	19
113000	2377	3	7
115000	2736	4	10
138000	2500	3	1
142500	2500	4	3
144000	2479	3	3
145000	2400	3	1
147500	3124	4	0
144000	2500	3	2
155500	4062	4	10
165000	2854	3	3

We need to develop three simple regression models to predict the selling price based on each of the individual factors and determine which one is the best model. Next, we will develop a model to 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 than just the three simple regression models.

To use Excel for regression, we do not want to use the Excel QM module, but rather will be using the data analysis add-in. To check and be sure that it is activated, go to **File → 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

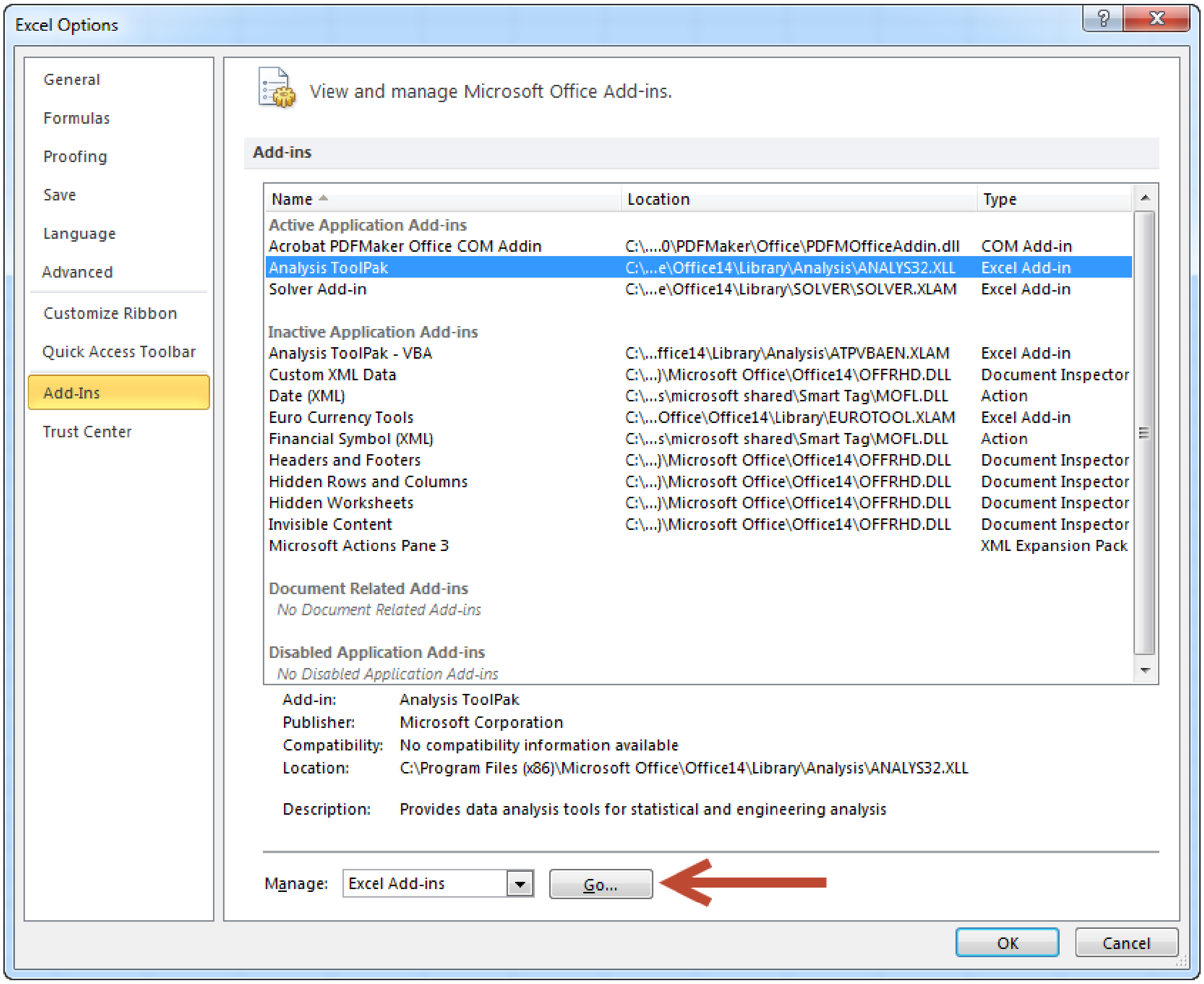
Default view for new sheets: Normal View

Include this many sheets: 3

Personalize your copy of Microsoft Office

User name: IDT

Under **Active Application Add-ins** be sure that **Analysis ToolPak** is there.



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

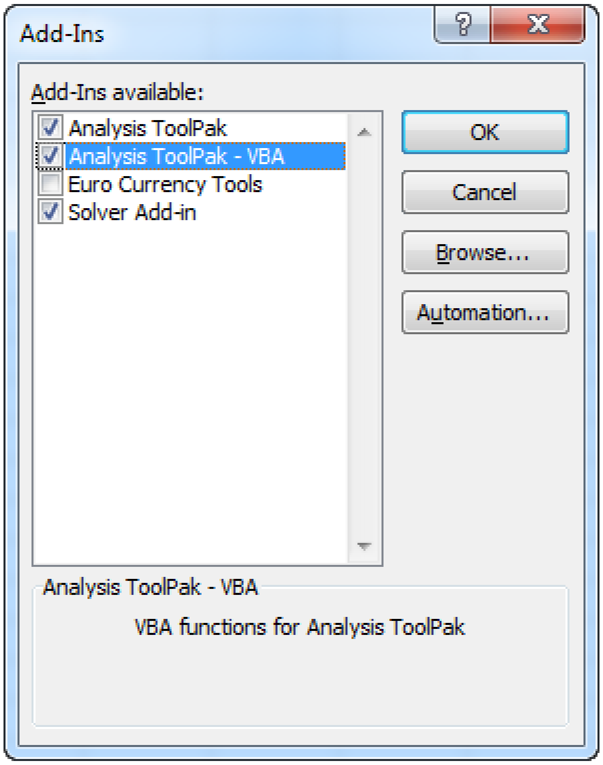
Compatibility: No compatibility information available

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

If not, click the “Go” button at the bottom of the window next to “Manage Excel Add-Ins” and simply tick the box next to **Analysis ToolPak** and **Analysis ToolPak VBA** then click **OK**.



Add-Ins available:

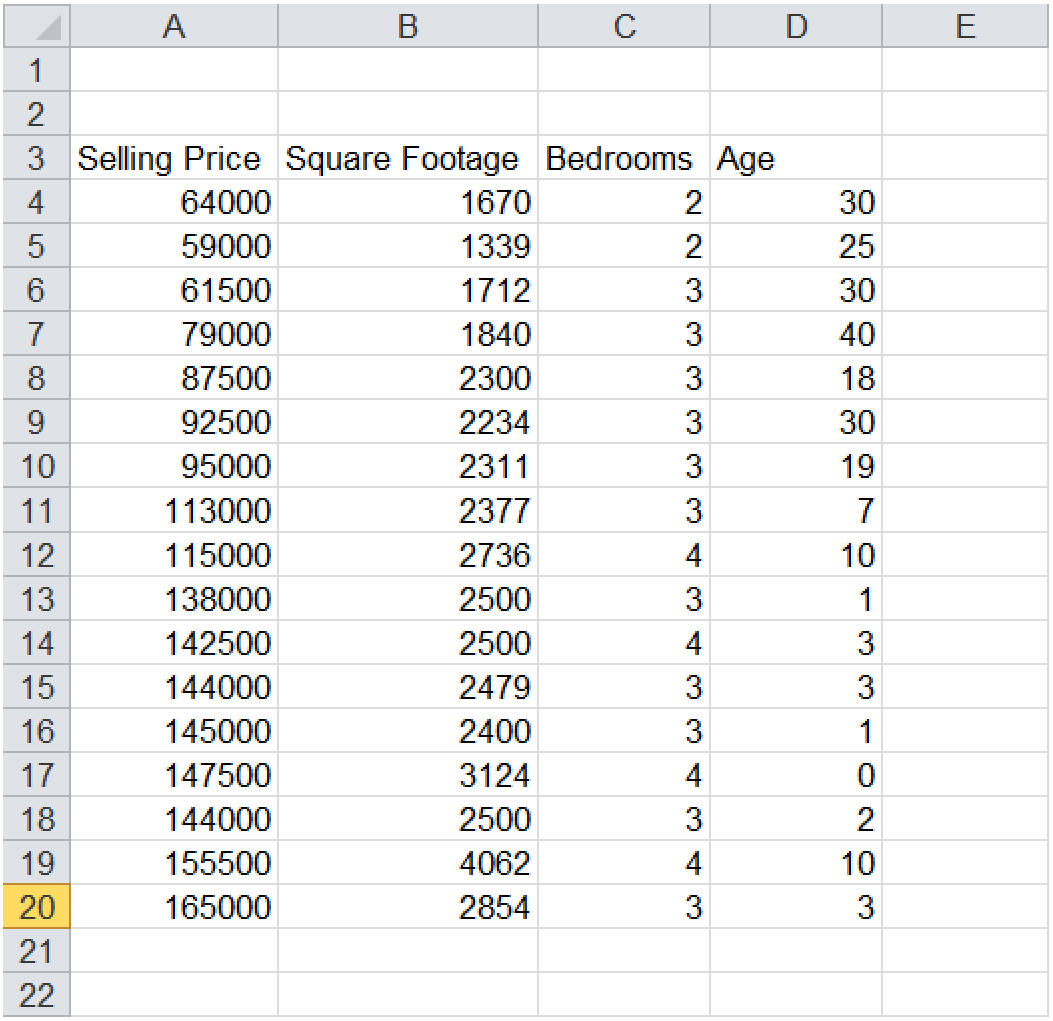
☑ Analysis ToolPak
☑ Analysis ToolPak - VBA
□ Euro Currency Tools
☑ Solver Add-in

Analysis ToolPak - VBA

VBA functions for Analysis ToolPak

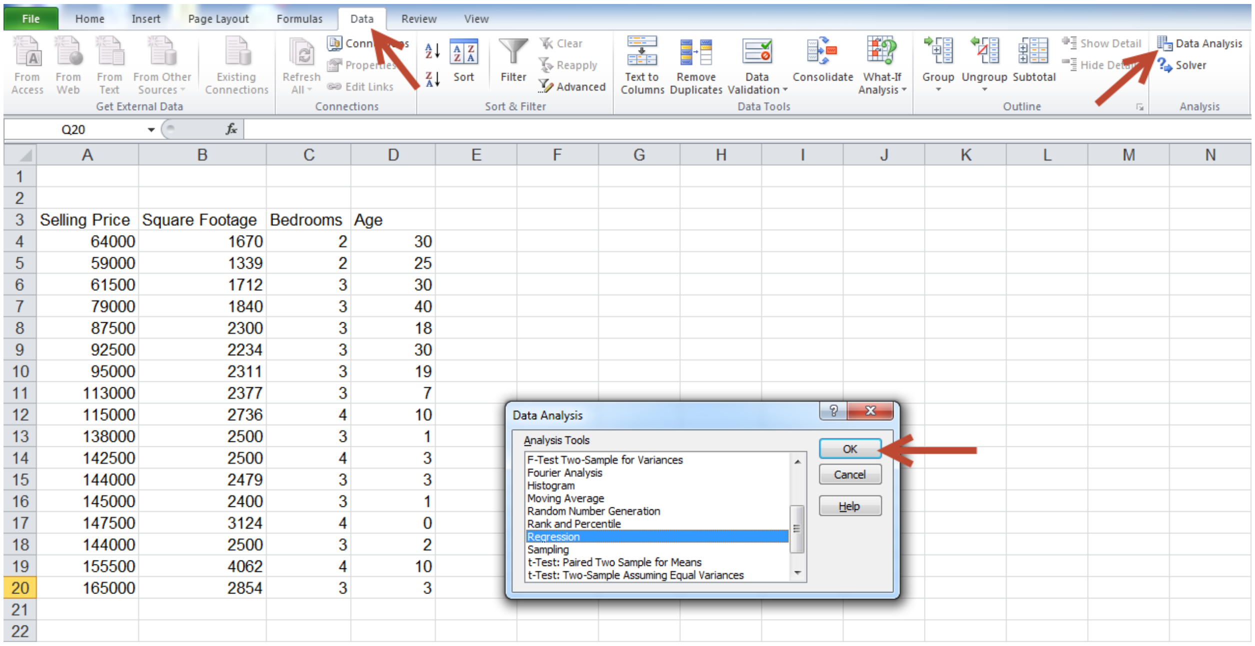
Once you have the Add-ins in place, you are ready to get started.

1. Enter or copy the data from the table above into a blank Excel spreadsheet as shown here.



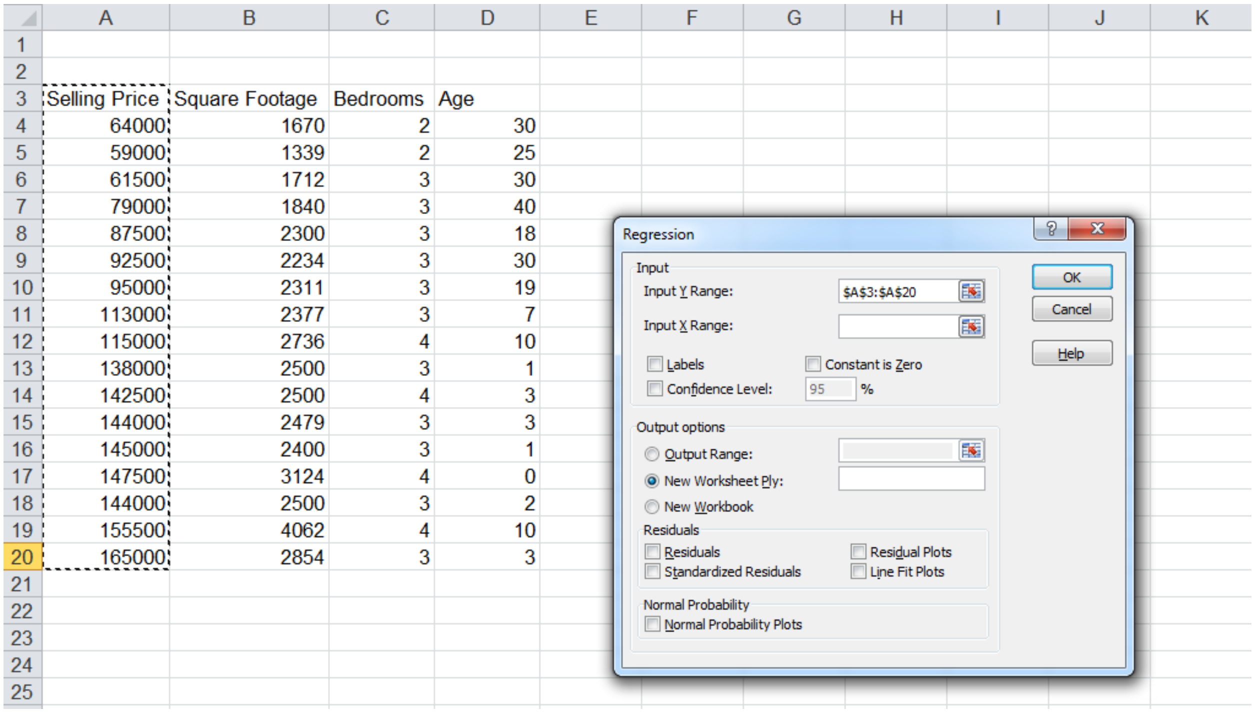
	A	B	C	D	E
1					
2					
3	Selling Price	Square Footage	Bedrooms	Age	
4	64000	1670	2	30	
5	59000	1339	2	25	
6	61500	1712	3	30	
7	79000	1840	3	40	
8	87500	2300	3	18	
9	92500	2234	3	30	
10	95000	2311	3	19	
11	113000	2377	3	7	
12	115000	2736	4	10	
13	138000	2500	3	1	
14	142500	2500	4	3	
15	144000	2479	3	3	
16	145000	2400	3	1	
17	147500	3124	4	0	
18	144000	2500	3	2	
19	155500	4062	4	10	
20	165000	2854	3	3	
21					
22					

2. Click on **Data→ Data Analysis** and, in the Data Analysis pop-up window, scroll down and select **Regression** and click **OK**.



	A	B	C	D
1				
2				
3	Selling Price	Square Footage	Bedrooms	Age
4	64000	1670	2	30
5	59000	1339	2	25
6	61500	1712	3	30
7	79000	1840	3	40
8	87500	2300	3	18
9	92500	2234	3	30
10	95000	2311	3	19
11	113000	2377	3	7
12	115000	2736	4	10
13	138000	2500	3	1
14	142500	2500	4	3
15	144000	2479	3	3
16	145000	2400	3	1
17	147500	3124	4	0
18	144000	2500	3	2
19	155500	4062	4	10
20	165000	2854	3	3
21				
22				

3. 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 A3-A20.



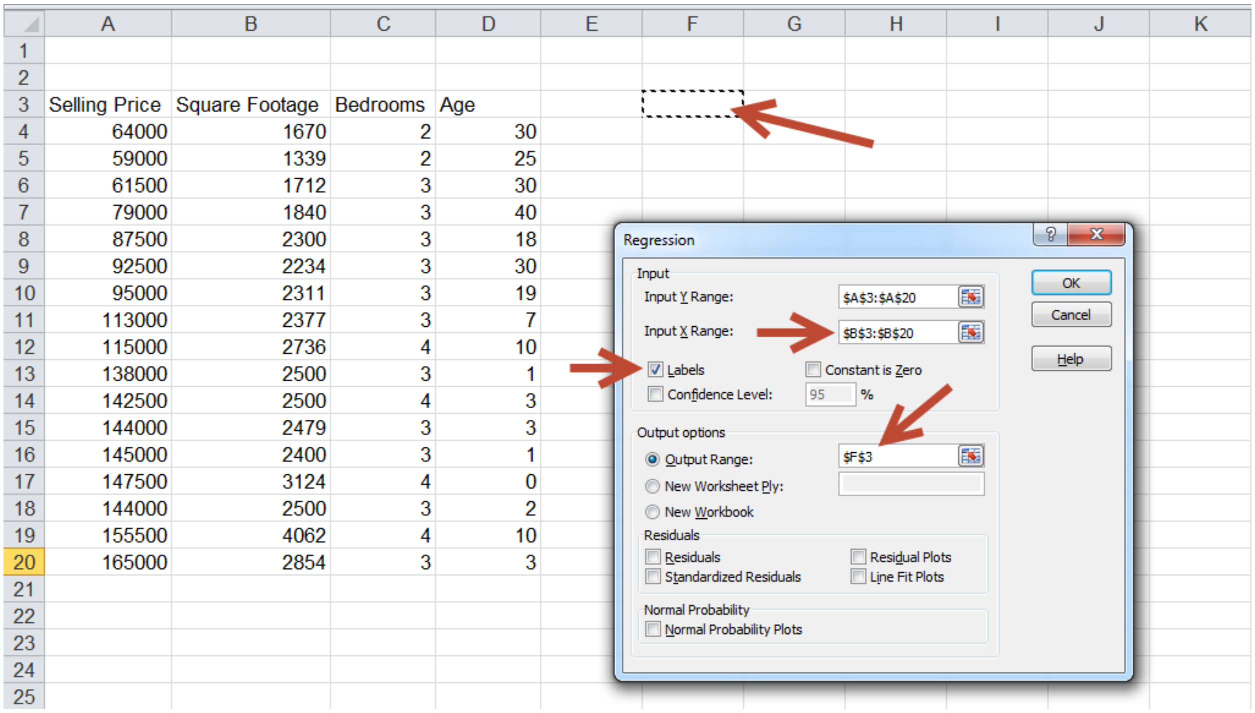
Input Y Range: $A$3:$A$20

□ Confidence Level: 95 %

● New Worksheet Ply:

	A	B	C	D
1				
2				
3	Selling Price	Square Footage	Bedrooms	Age
4	64000	1670	2	30
5	59000	1339	2	25
6	61500	1712	3	30
7	79000	1840	3	40
8	87500	2300	3	18
9	92500	2234	3	30
10	95000	2311	3	19
11	113000	2377	3	7
12	115000	2736	4	10
13	138000	2500	3	1
14	142500	2500	4	3
15	144000	2479	3	3
16	145000	2400	3	1
17	147500	3124	4	0
18	144000	2500	3	2
19	155500	4062	4	10
20	165000	2854	3	3
21				
22				
23				
24				
25				

4. Our first independent variable will be square footage, so click in the box for **Input X Range** and select cells B3-B20. Be sure that the box is ticked next to **Labels** and select the **Output Range** as F3.



	A	B	C	D
1				
2				
3	Selling Price	Square Footage	Bedrooms	Age
4	64000	1670	2	30
5	59000	1339	2	25
6	61500	1712	3	30
7	79000	1840	3	40
8	87500	2300	3	18
9	92500	2234	3	30
10	95000	2311	3	19
11	113000	2377	3	7
12	115000	2736	4	10
13	138000	2500	3	1
14	142500	2500	4	3
15	144000	2479	3	3
16	145000	2400	3	1
17	147500	3124	4	0
18	144000	2500	3	2
19	155500	4062	4	10
20	165000	2854	3	3

Regression

Input

Input Y Range: $A$3:$A$20

Input X Range: $B$3:$B$20

☑ Labels　□ Constant is Zero

□ Confidence Level: 95 %

Output options

● Output Range: $F$3

○ New Worksheet Ply:

○ New Workbook

Residuals

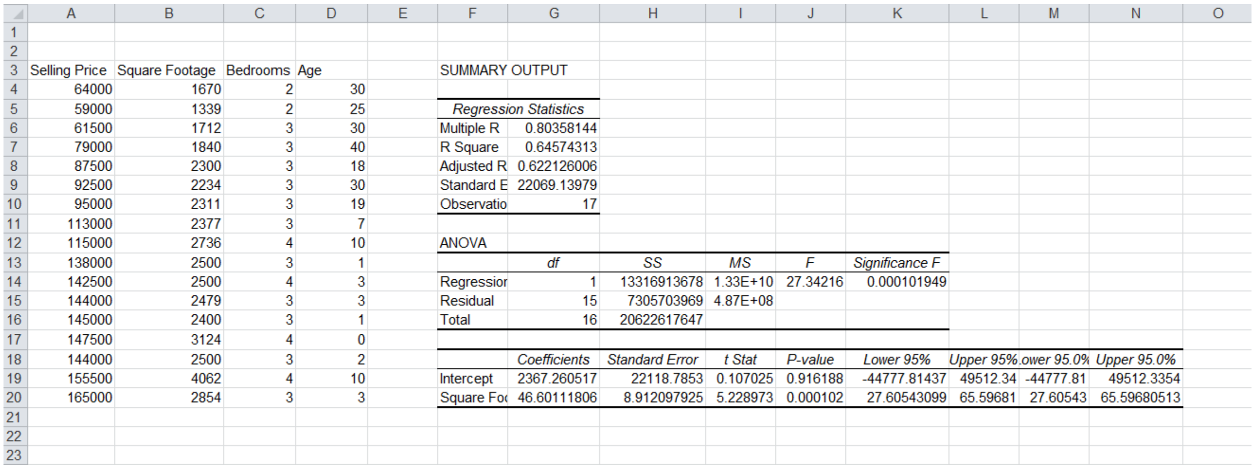
□ Residuals　□ Residual Plots

□ Standardized Residuals　□ Line Fit Plots

Normal Probability

□ Normal Probability Plots

5. Click **OK**. This will put the regression output next to our data table.



	A	B	C	D
3	Selling Price	Square Footage	Bedrooms	Age
4	64000	1670	2	30
5	59000	1339	2	25
6	61500	1712	3	30
7	79000	1840	3	40
8	87500	2300	3	18
9	92500	2234	3	30
10	95000	2311	3	19
11	113000	2377	3	7
12	115000	2736	4	10
13	138000	2500	3	1
14	142500	2500	4	3
15	144000	2479	3	3
16	145000	2400	3	1
17	147500	3124	4	0
18	144000	2500	3	2
19	155500	4062	4	10
20	165000	2854	3	3

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.80358144
R Square	0.64574313
Adjusted R	0.622126006
Standard E	22069.13979
Observatio	17

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	1	13316913678	1.33E+10	27.34216	0.000101949
Residual	15	7305703969	4.87E+08		
Total	16	20622617647			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*.ower 95.0%*	*Upper 95.0%*
Intercept	2367.260517	22118.7853	0.107025	0.916188	-44777.81437	49512.34	-44777.81	49512.3354
Square Foo	46.60111806	8.912097925	5.228973	0.000102	27.60543099	65.59681	27.60543	65.59680513

Repeat steps 2-5, but select C3-C20 for the number of bedrooms and put the **Output Range** as F23, then, repeat steps 2-5 again but select D3-D20 for Age, and put the **Output Range** as F43.

You should now have all three simple regression models. Click here to download the completed sample spreadsheet so you can compare it to yours.

The key parts of this output are as follows (using the square footage example):

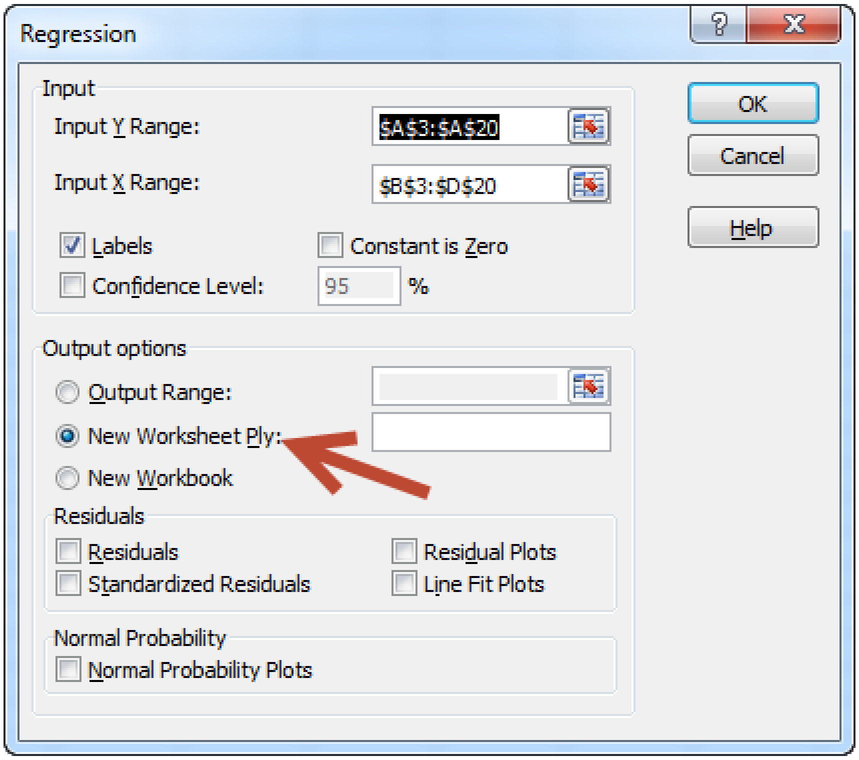
Under the "Regression Statistics"

- Multiple R – the correlation coefficient – notes the strength of the relationship – in this case, 0.80358 – a pretty strong positive relationship.
- R squared – the amount of variability in the dependent variable explained by the independent variable(s). In this case, 0.6457 – again, a pretty strong number – almost 65% of the variability in purchase price is explained by square footage.
- Adjusted R squared – this is when you have more than one independent variable and have adjusted the R squared value for the number of independent variables. Use this when looking at a multiple regression model.

Under the ANOVA Tables

- Significance F – this tests 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 0.000102, 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, this is where the line crosses the Y axis. Here its 2367. So a house with zero square feet will sell for $2,367.
- 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, or $46.6.
- X P-Value – this tests 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. It's 0.000102, so we have a significant variable.

Running a multiple regression is the same as a simple regression, the only difference being that we will select all three independent variables as our 'X variables' – our **Input Y Range** is A3-A20 while our **Input X Range** is now B3-D20. Again, be sure to tick the box for **Labels** and this time select **New Worksheet Ply** as your Output option.



Input

Input Y Range: $A$3:$A$20

Input X Range: $B$3:$D$20

☑ 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

Click here to download the completed sample spreadsheet so you can compare it to yours.

If we look at those statistics for all three simple models and our multiple regression model, we get the following:

Model	Significance (model/variable)	Multiple R	R-Square	Adjusted R Square
Square Footage	0.000102	.803581	.645743	.622126
Bedrooms	0.010206	.604196	.365053	.322723
Age	0.00000295	.881338	.776757	.761874
Multiple model: SF Bedrooms Age	0.00000212 .017955 .811196 .000162	.941348	.886137	.859861

Comparing the three simple models, we can see that the model using **age** as the predictor of price is the best. It has the highest Multiple R (i.e., strongest relationship) and highest R-Square (explains most of the variability in the dependent variable).

Looking at the multiple model, this is even better. Both Multiple R and R-Square are higher, even when adjusting for the number of dependent variables. What is interesting here is that the number of bedrooms is not significant in this model, so that should not be included in the final model.

This concludes the tutorial on both simple and multiple regression models.