Working with multiple groups: demo notes

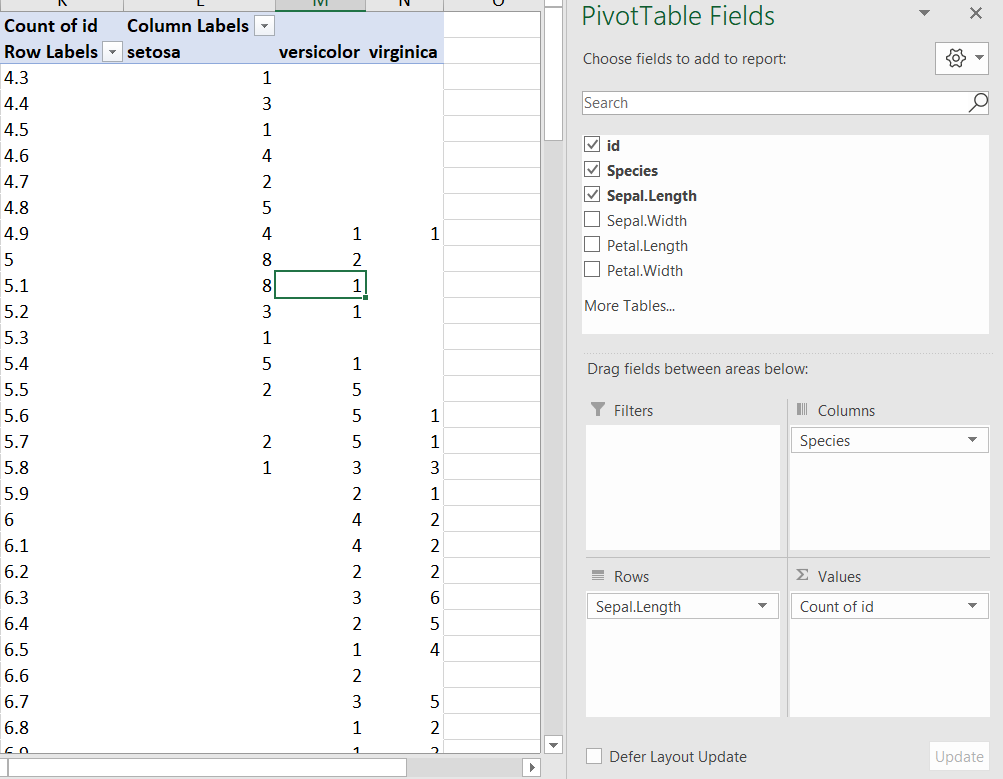
**Visualizing distributions: histograms and box plots**

File: iris-viz.xlsx

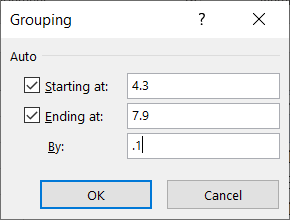
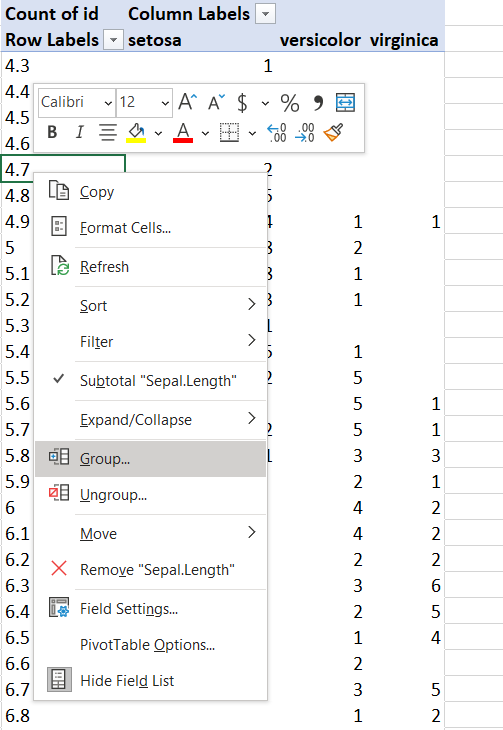
*Histograms*

We will visualize the distribution of sepal length for each species.

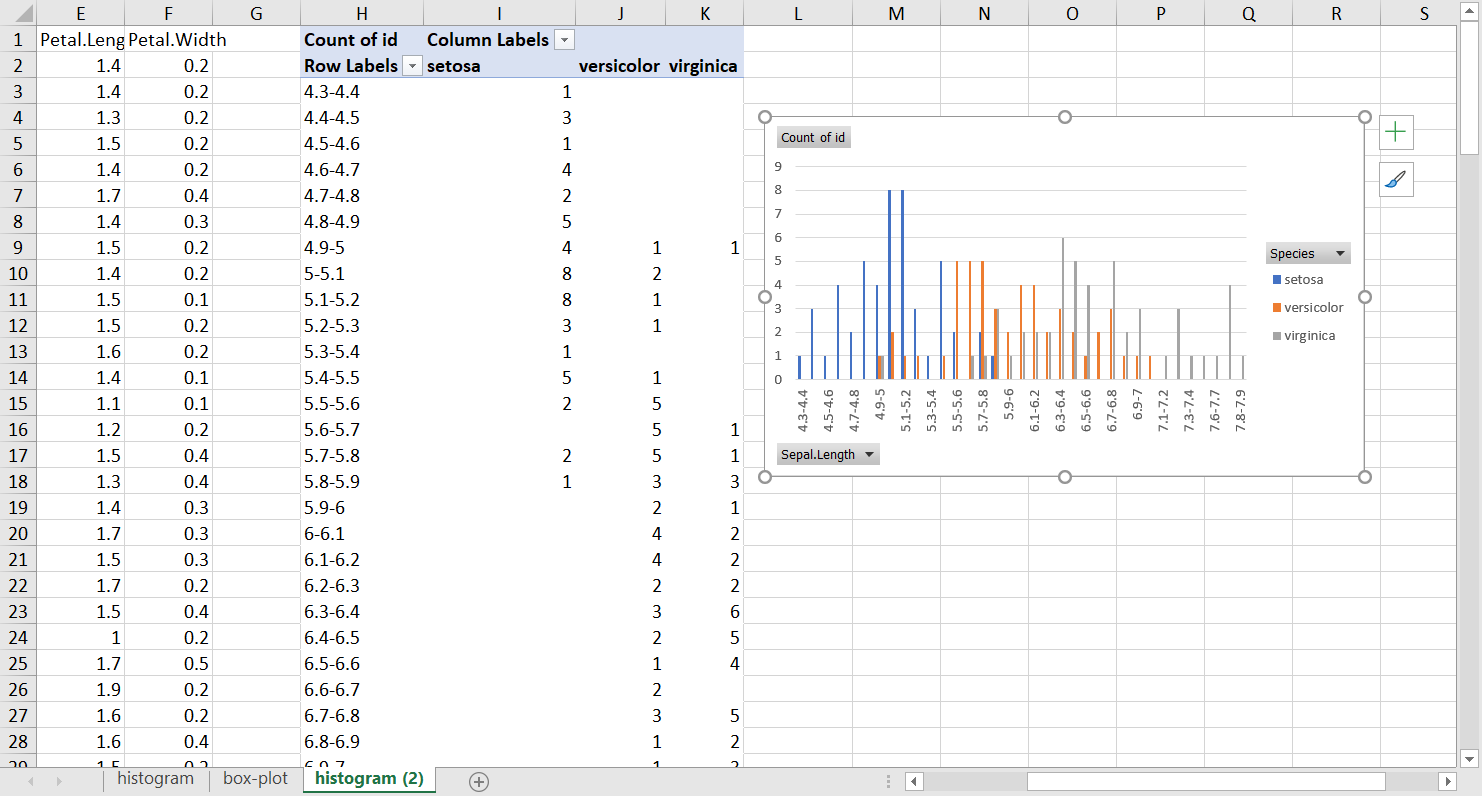
1. Insert a PivotTable from the source data. Place Species in the Columns section, Sepal.Length in the Rows section and Count of id in the Values section.



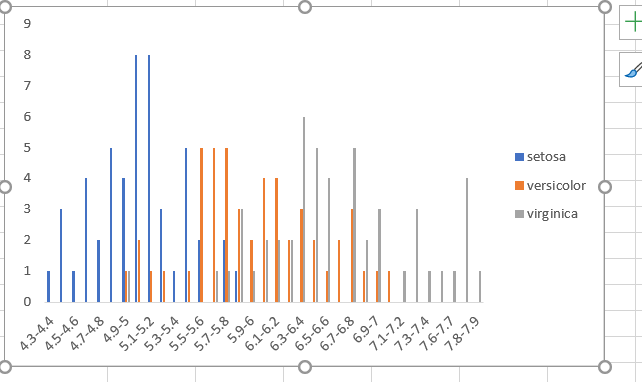
1. Right-click on the Row Labels and select Group. Group the variable at intervals of .1.



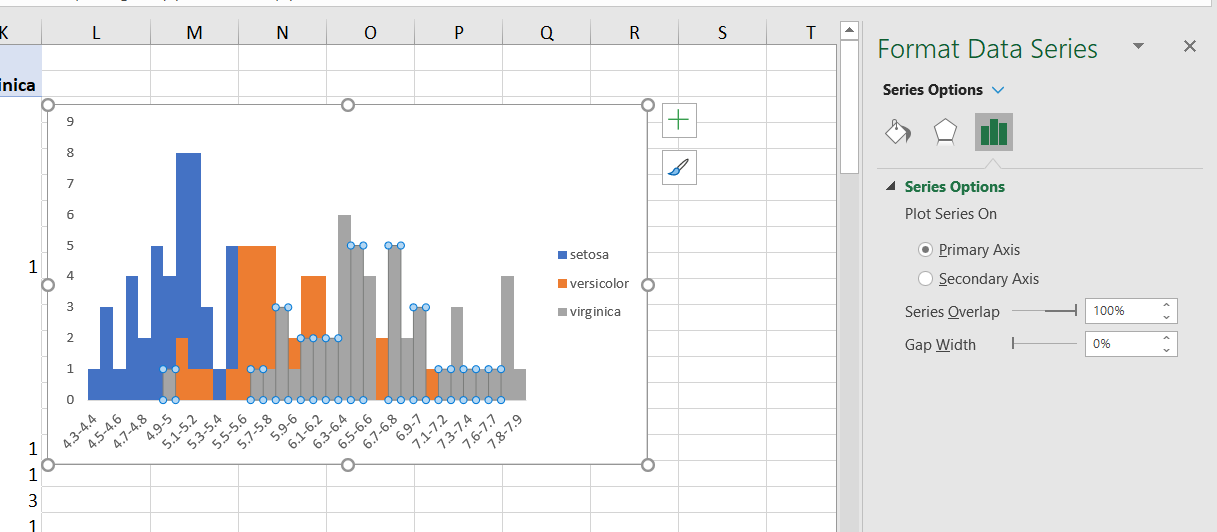
1. Insert the recommended chart: clustered column.



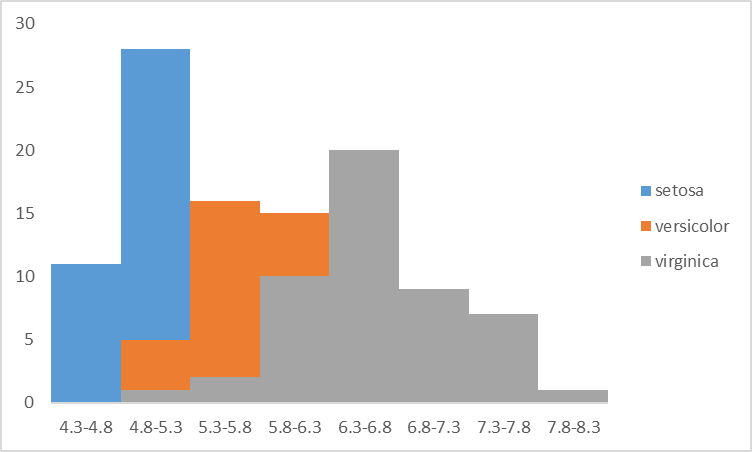
1. Clean up this chart by right-clicking on any of the labels and selecting “Hide All Field Buttons on Chart.” You can also remove the chart gridlines by clicking on any of them and pressing the Delete key.



1. Right-click on any of the bars and select “Format Data Series.” A menu will appear to the right. Set Series Overlap to 100% and Gap Width to 0%.



1. You can resize the bins of the histogram by right-clicking back on the Row Labels of the PivotTable and selecting Group. What happens if we put it in intervals of .5?

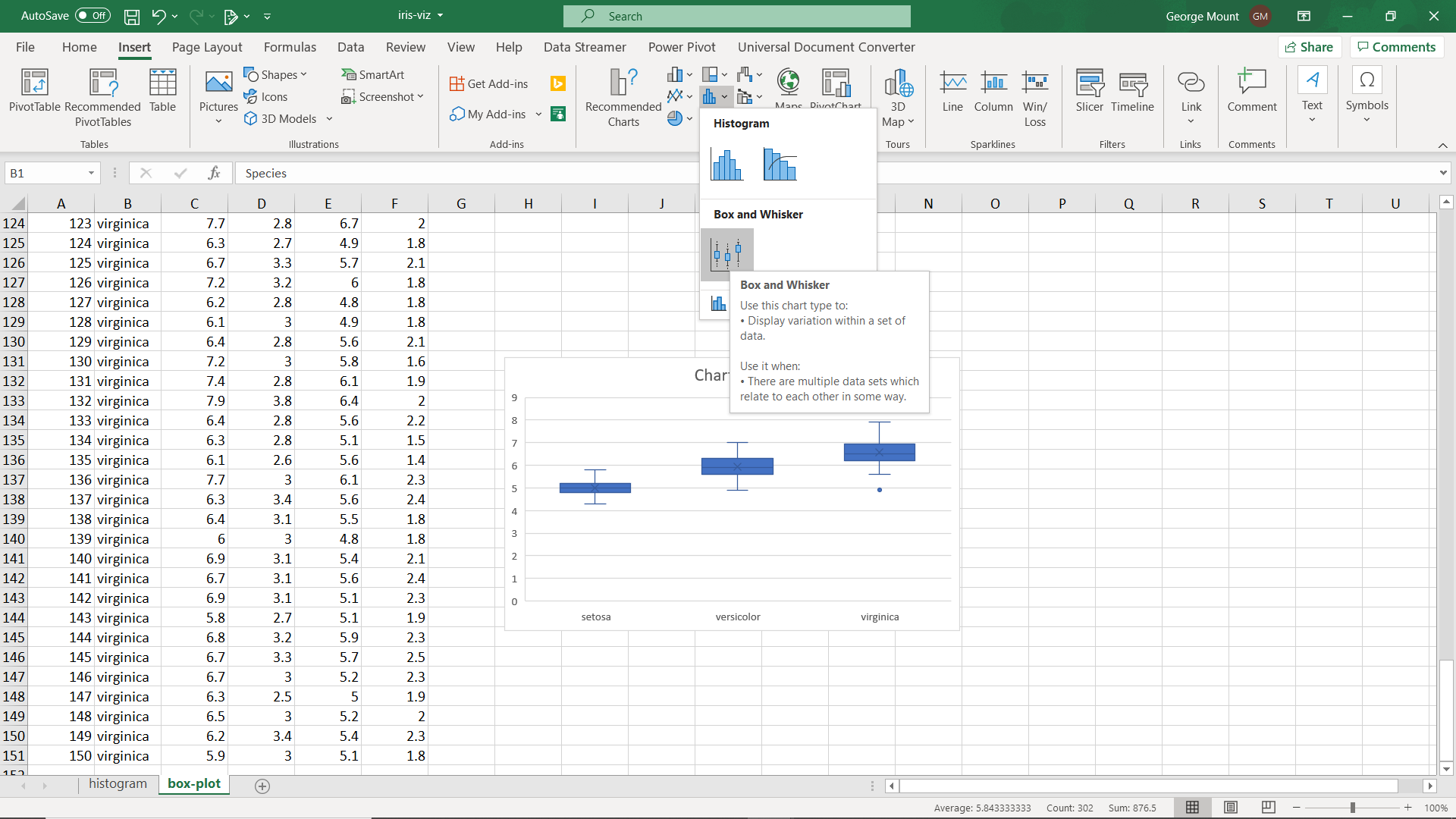


*Box plots*

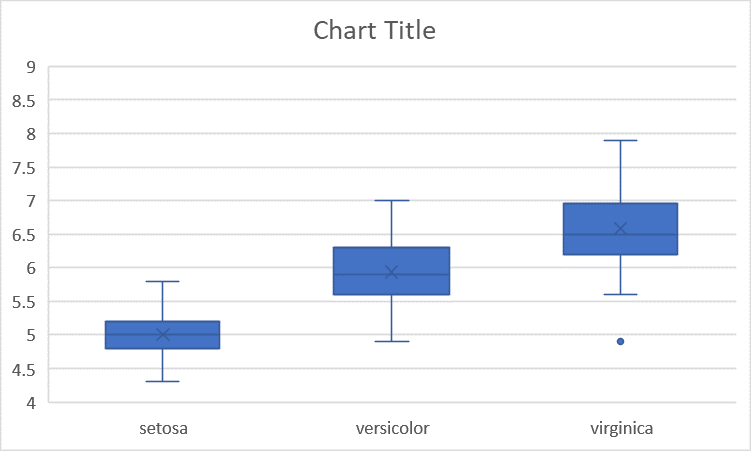
Multiple histograms on the same chart can get messy. Let’s try a different visualization: the box plot.

We will again plot the distributions of sepal length by species.

1. Select columns B-C and head to Insert > Chart. Under Histogram there will be an option, Box and Whisker.



1. Fortunately there is not too much more prep needed for this chart. We could re-set the y-axis to start at a value besides 0 (controversial, but sometimes useful).
   1. Right-click on the y-axis and select “Format Axis.” You can now set the minimum bound to 4.



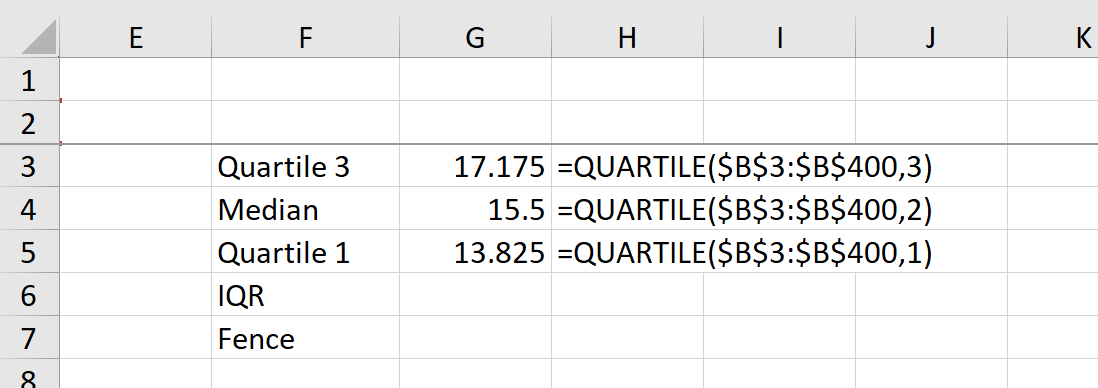
1. Take a look at the example box-and-whisker chart in the file to make sense of these distributions. What is the point under virginica doing there?
   1. Any datapoint that is 1.5 times the IQR is an outlier and excluded from the box and whisker plot.

**Outlier detection**

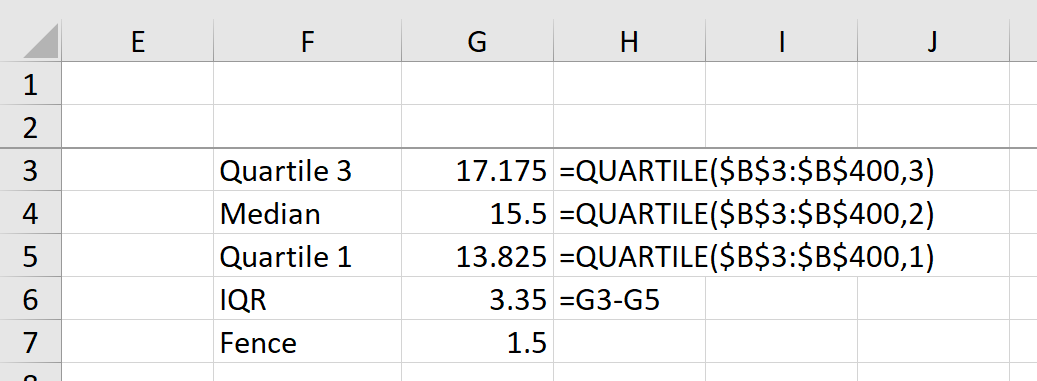
File: outliers.xlsx

Let’s calculate for ourselves the outlier range, and remove those datapoints from the box plot to see what happens.

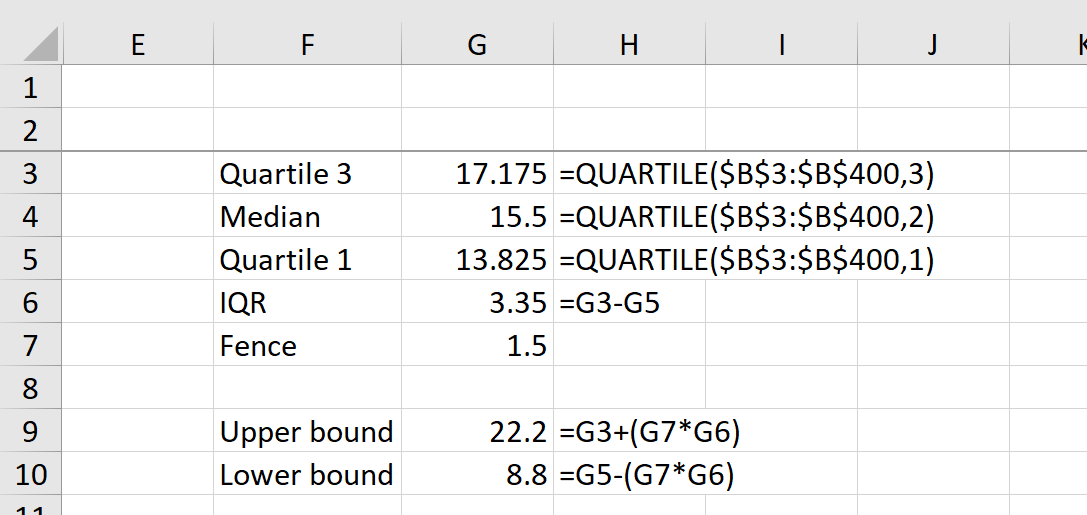
1. To do this, we will calculate the 3rd, 2nd or *median*, and 1st quartiles of our acceleration data using the QUARTILE() function. We will pass in our data range, and what number quartile we want:



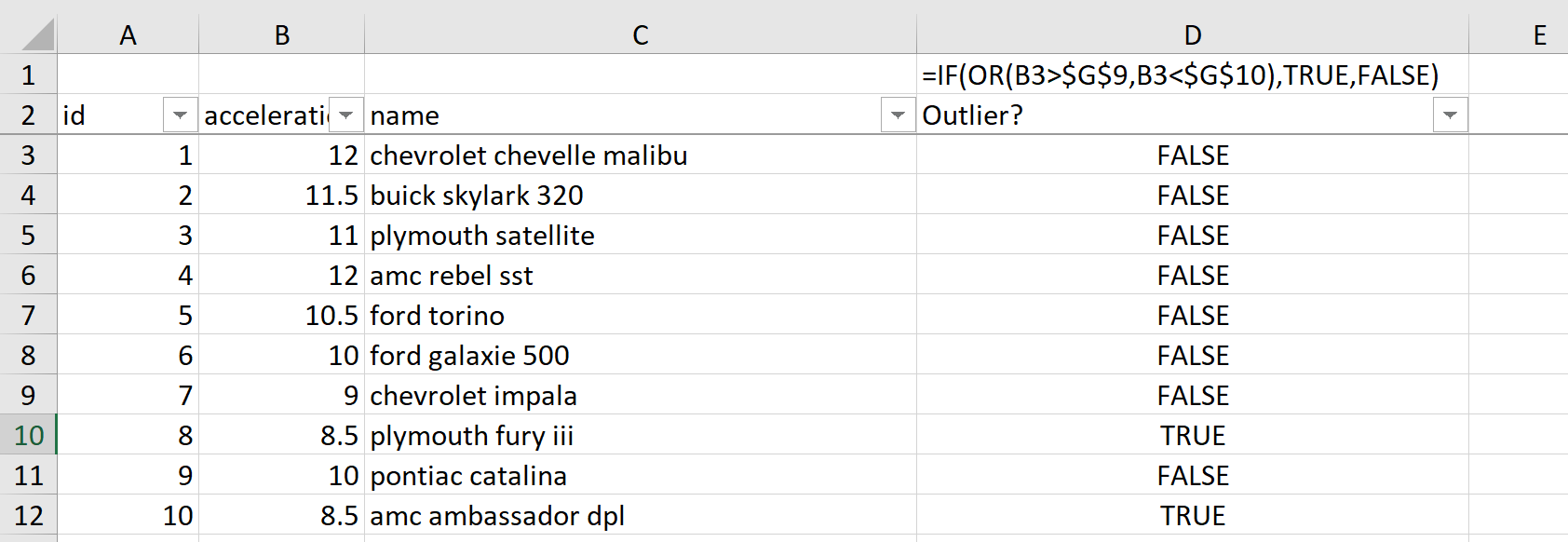
1. We will now calculate the interquartile range (IQR) as the difference between the third and first quartiles.
   1. The fence will be 1.5. This is a hard-coded value that we will use to set the threshold for being an outlier.



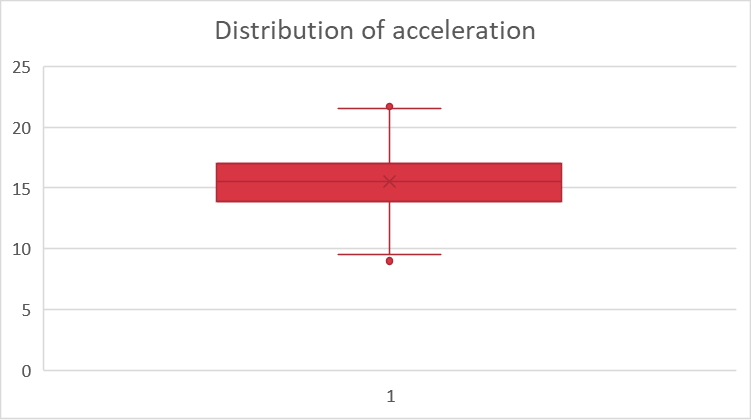
1. We will now calculate the outlier thresholds as 1.5 times the IQR, plus our upper bound and minus our lower bound respectively.



1. We now know that any datapoint less than 8.8 or 22.2 is considered an outlier. We can use conditional logic to flag each value as TRUE or FALSE as being an outlier.



1. We can now filter our source data to exclude datapoints where Outlier? equals TRUE.
   1. The outliers have been removed from our box plot, with the exceptions of some datapoints that are right on the cusp of being outliers.

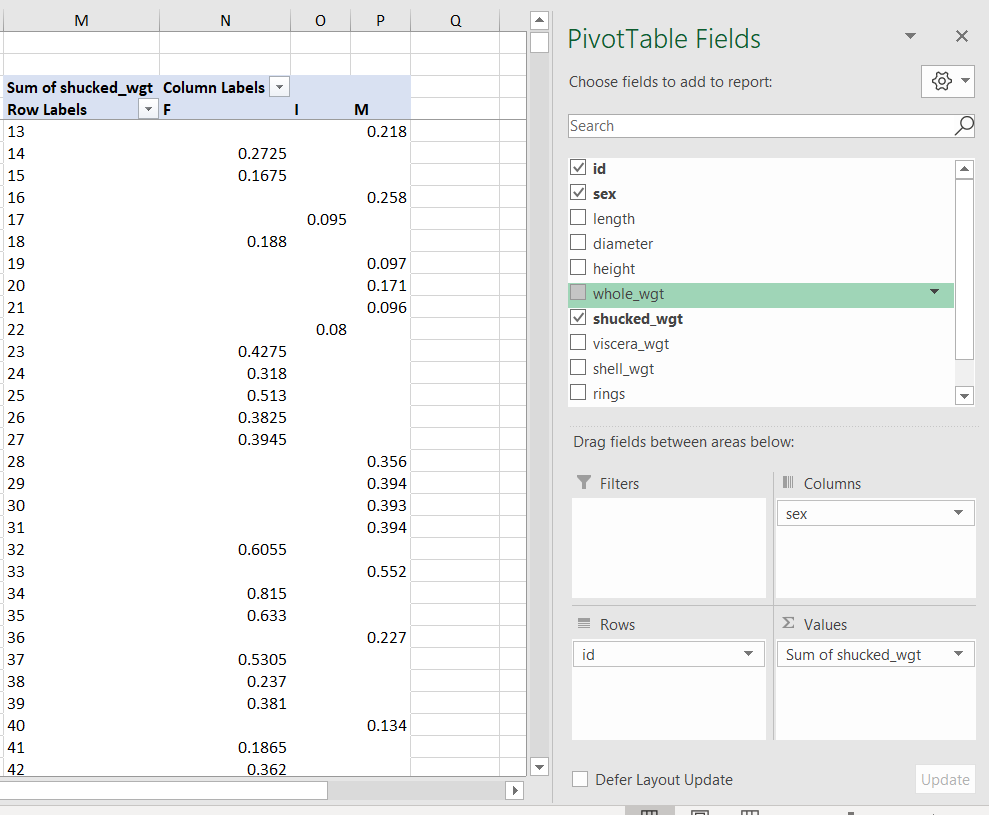


**Analysis of variance (ANOVA)**

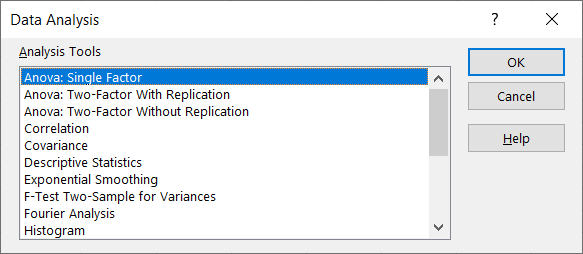
File: abalone-anova.xlsx

Let’s check for a significant difference in shucked weights across male, female and infant snails.

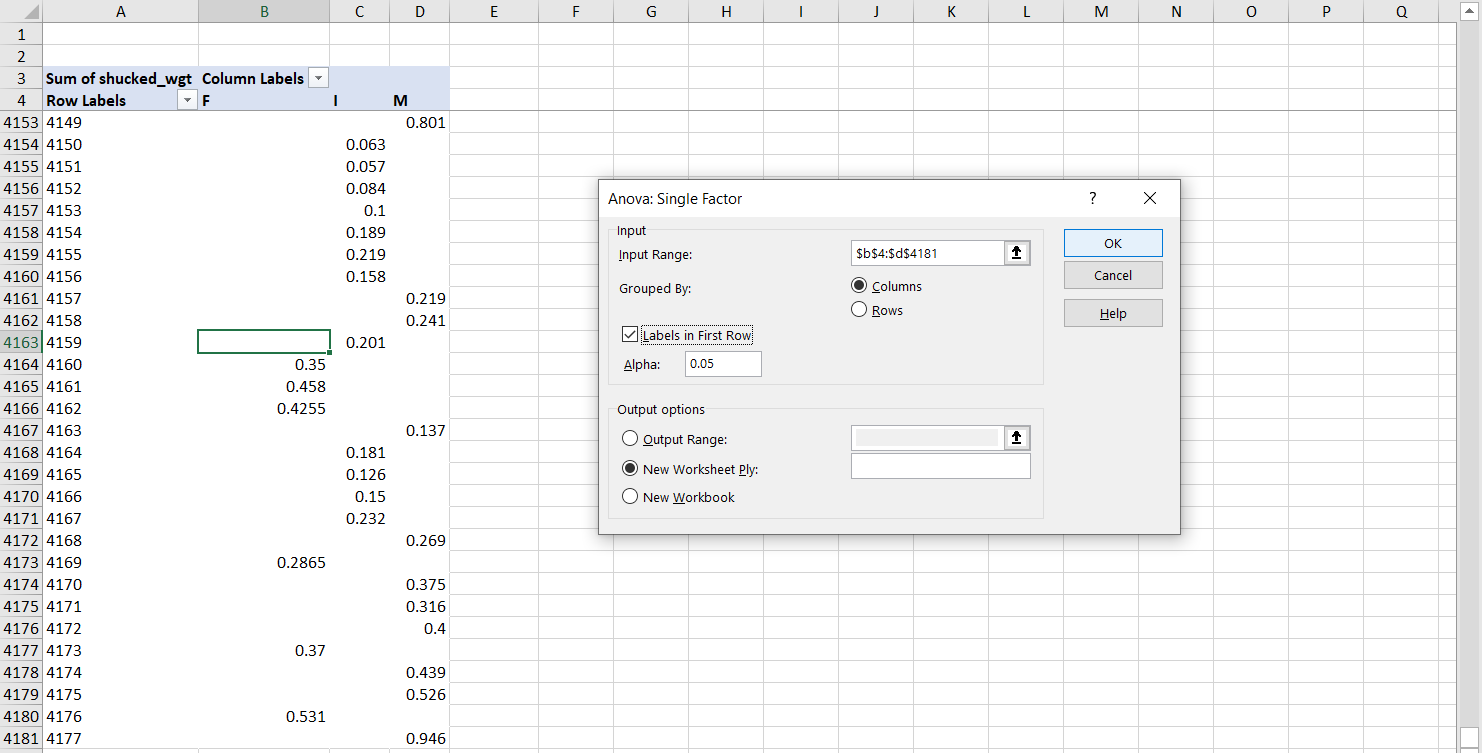
1. Insert a PivotTable. Put id in the Rows section, sex in the Columns section and Sum of shucked\_wgt in the Values section.
   1. Turn off the totals by clicking inside the PivotTable and selecting Design > Grand Totals > Off for Rows and Columns.



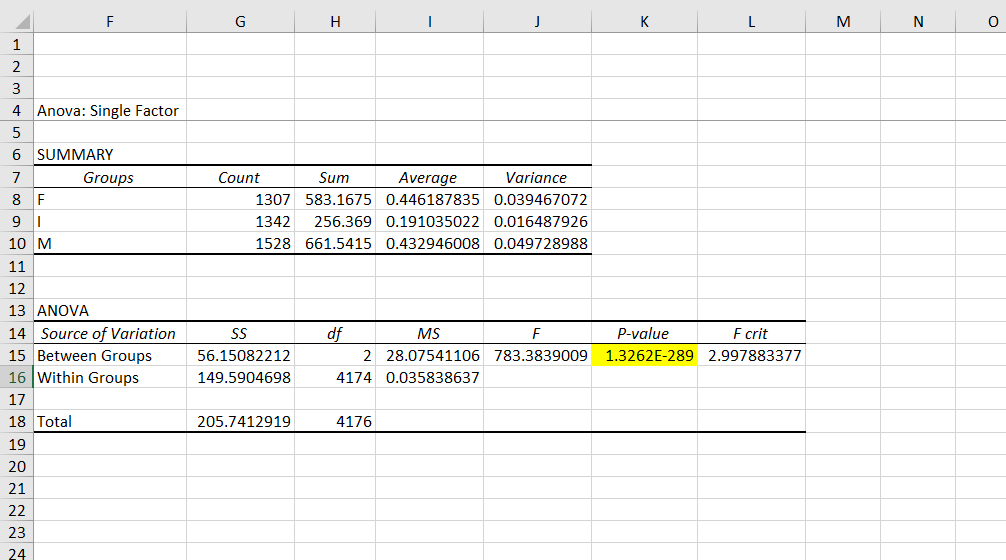
1. In the Analysis ToolPak, select Anova: Single Factor.



* 1. The input range is the three columns for each category: F, I and M.



1. The results of the ANOVA are available in the second box of outputs. The p-value for between-groups variation tells us if there is a significant difference across group means.

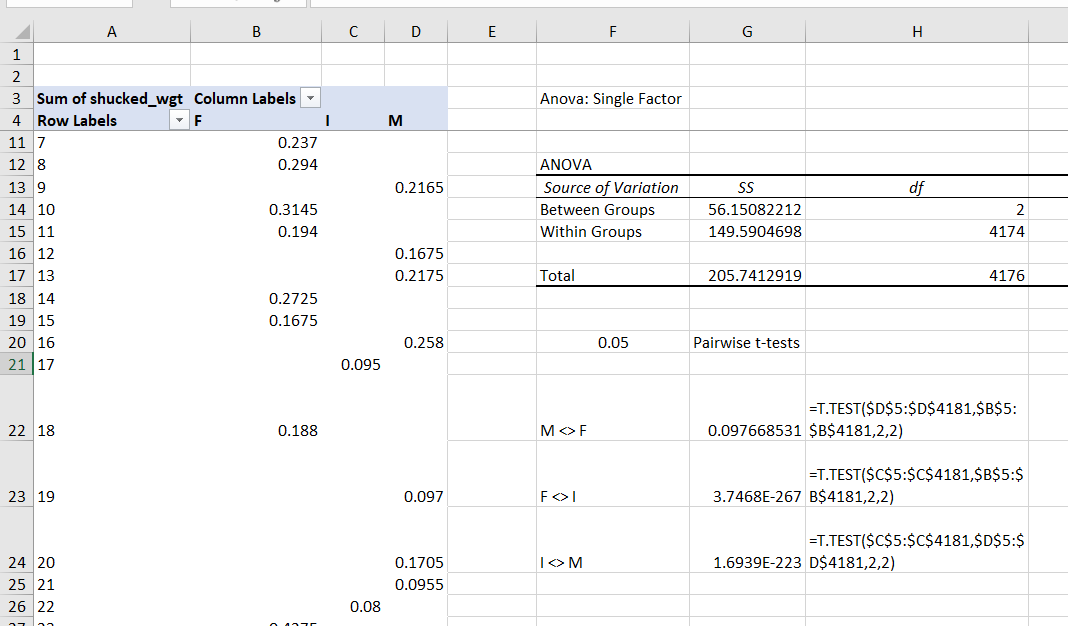


**ANOVA post-hoc tests: pairwise comparisons with Bonferroni correction**

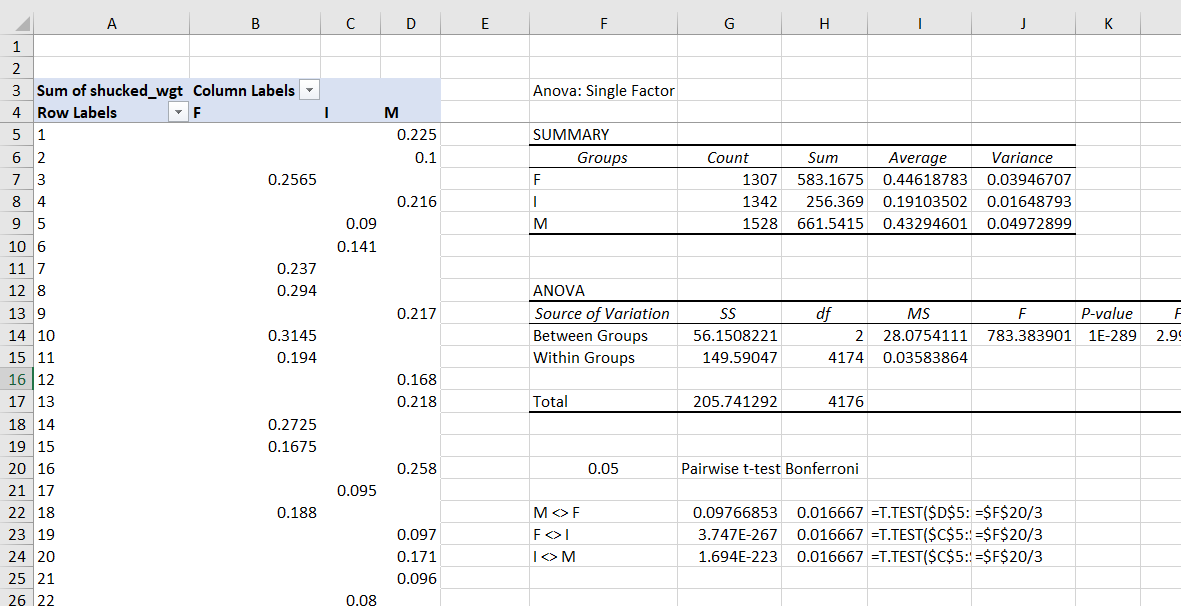
File: abalone-post-hoc.xlsx

The ANOVA in itself does *not* tell us *which* groups in particular are significantly higher/lower than the others. To do that, we will run *post-hoc* tests while adjusting for *experimentwise error*.

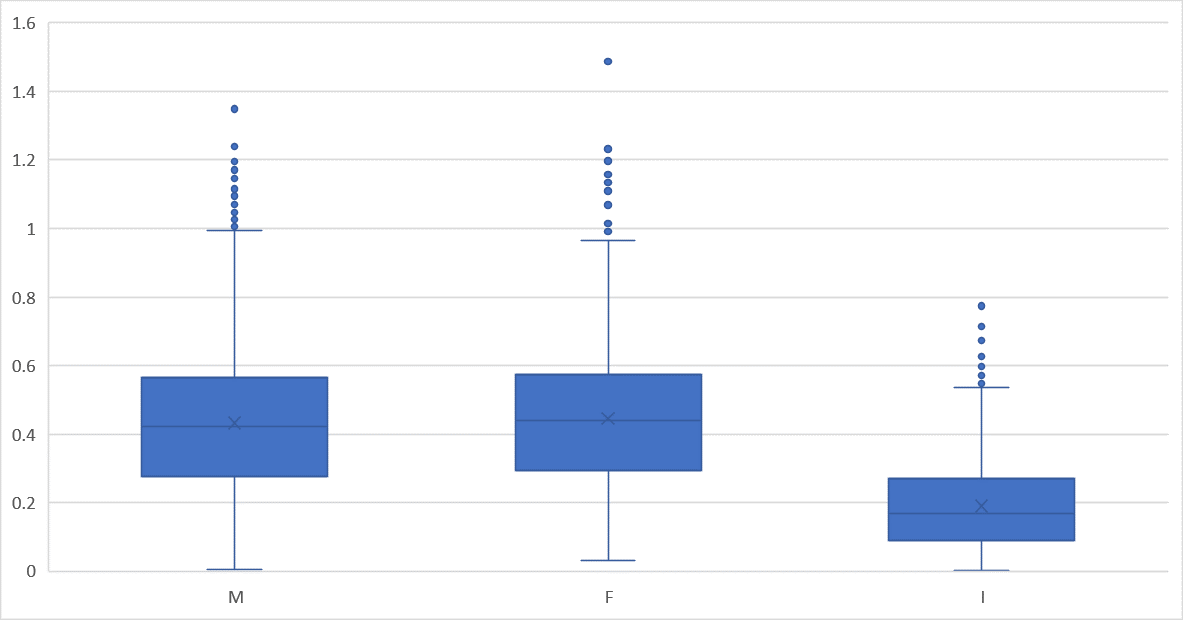
1. Conduct a *pairwise t-test* for to compare each pair of categories using the T.TEST() function.
   1. This will take four arguments:
      1. The range containing the first category to compare
      2. The range containing the second category to compare
      3. Whether this is a one- or two-tail test. We are using two-tail tests, so the argument is 2.
      4. The type of t-test. Since these are independent samples, this is not a paired t-test. We will assume equal variances as that is an assumption of the ANOVA. So the argument here is also 2.
      5. The result of T.TEST() is the test’s p-value. We will compare it against the adjusted alpha next.



1. We will now compare these p-values to a Bonferroni-adjusted alpha. This number will be our original alpha (.05) divided by the number of groups we are comparing (3). This makes our new alpha .0017.



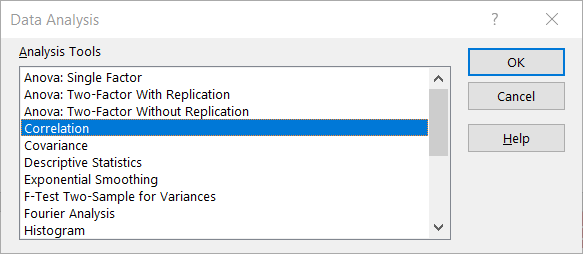
* 1. Based on these results, there is no significant difference in weights between male and female snails, but there is a significant difference between female and infant snails, and a significant difference between male and infant snails.
     1. This is a good time to refer back to the box plots for a visual understanding of the analysis.



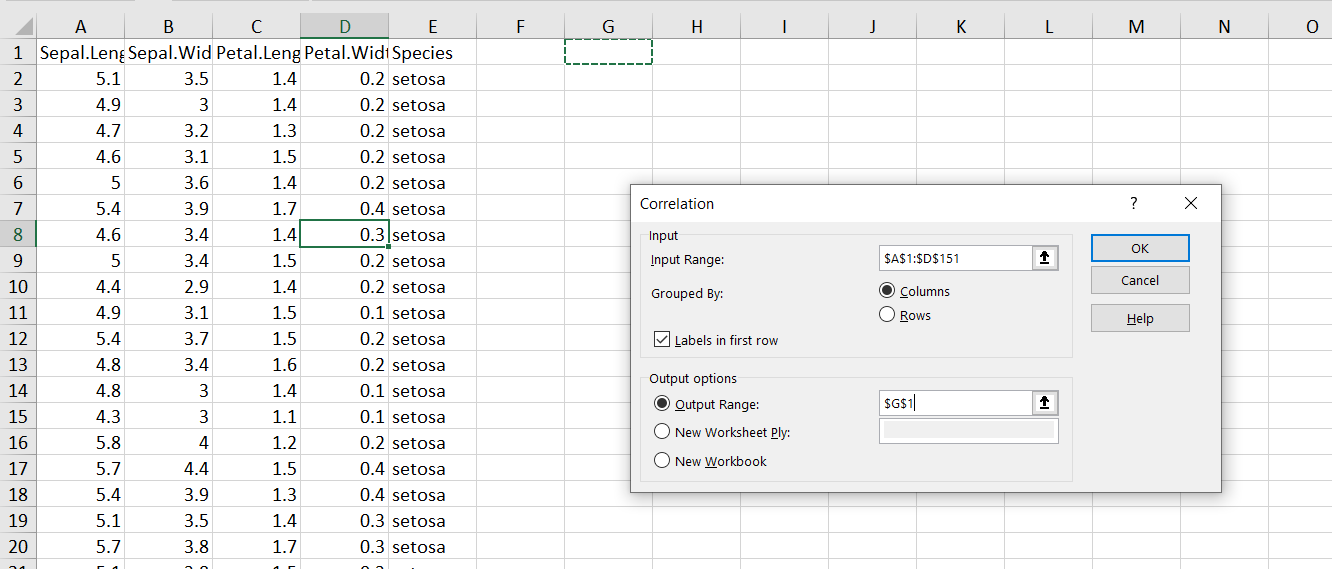
**Pearson correlations**

File: iris-corr.xlsx

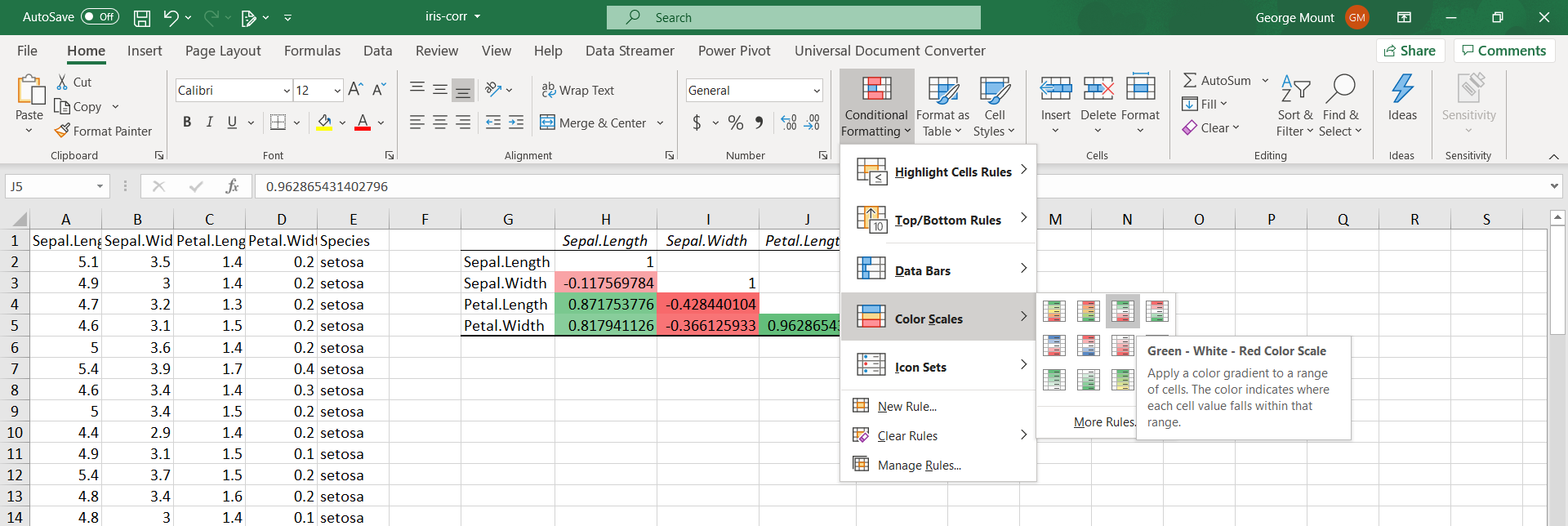
1. To insert a correlation matrix, go to the Analysis ToolPak and select Correlation.



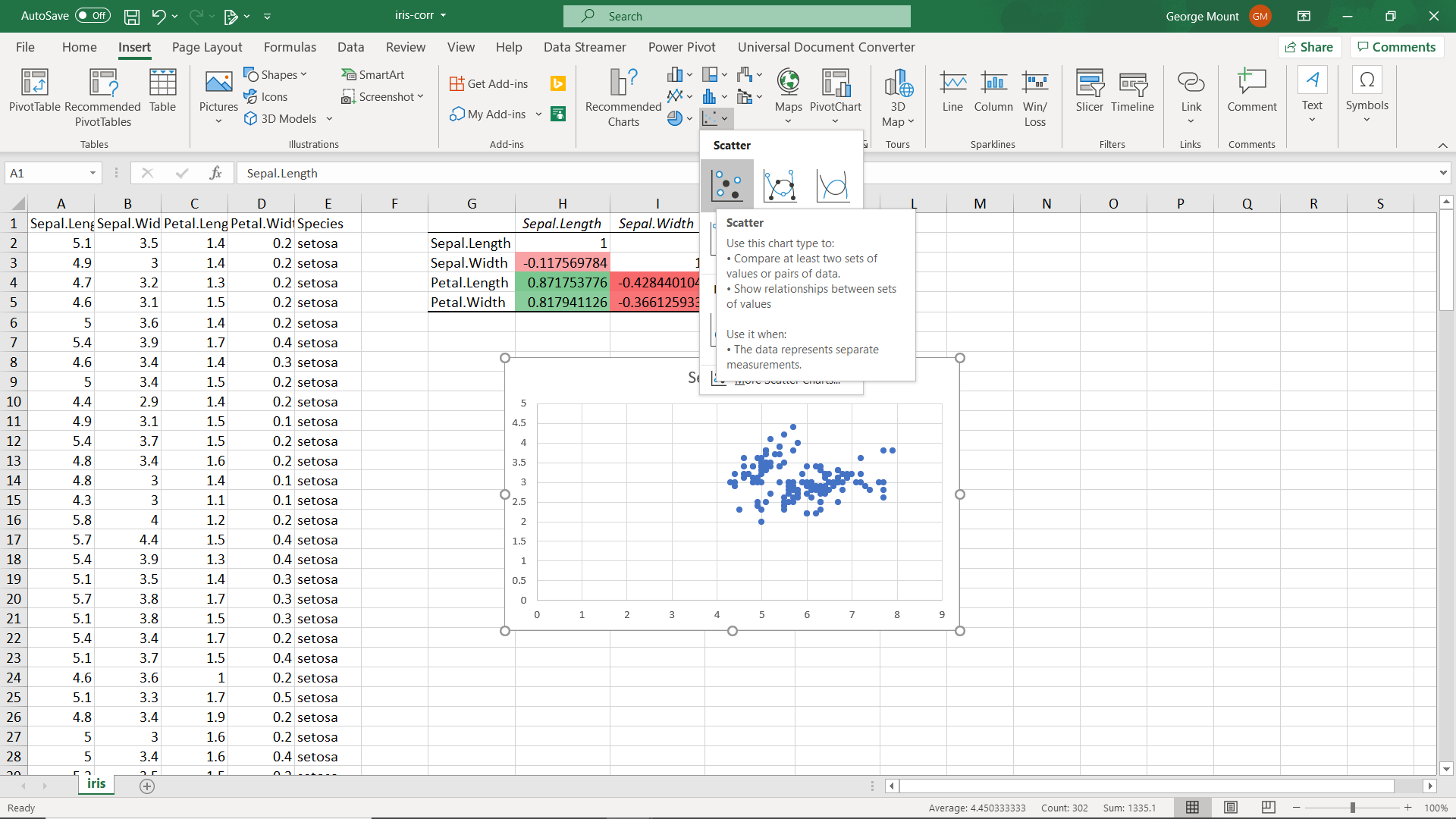
1. The input range will be all the *numeric* fields. String fields cannot be included in the correlation analysis.



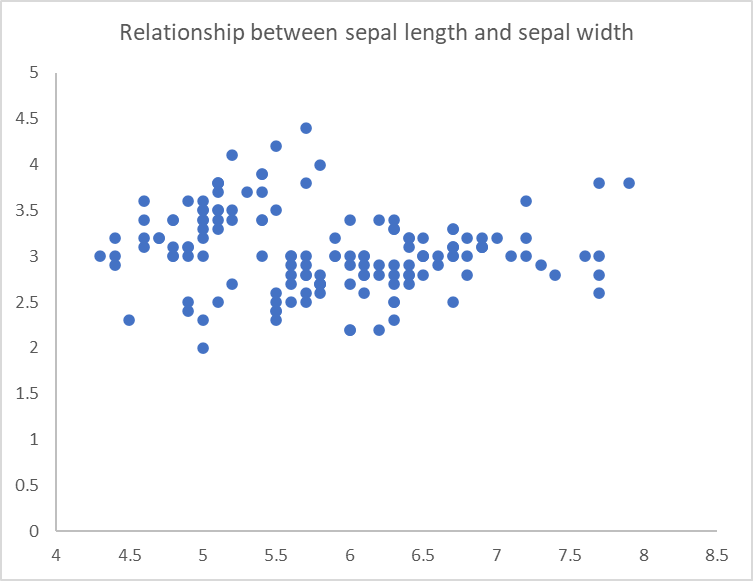
1. For ease of interpretation, select all the correlation values and select Home > Conditional Formatting > Color Scales > Green – White – Red Color Scales.



1. To plot the relationship between two variables (in this case, sepal length and sepal width), highlight the data and select Insert > Scatter.



1. To clean up this scatter chart, set the X-axis to 4 and remove the gridlines. It’s also a good idea to label this chart more clearly.

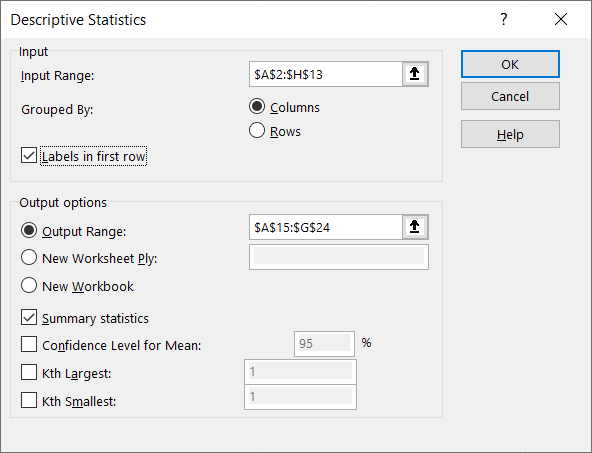


1. What does the scatterplot of sepal length and petal length look like?
2. By custom, you want to put the *independent* variable on the X-axis, and the *dependent* on the Y-axis.

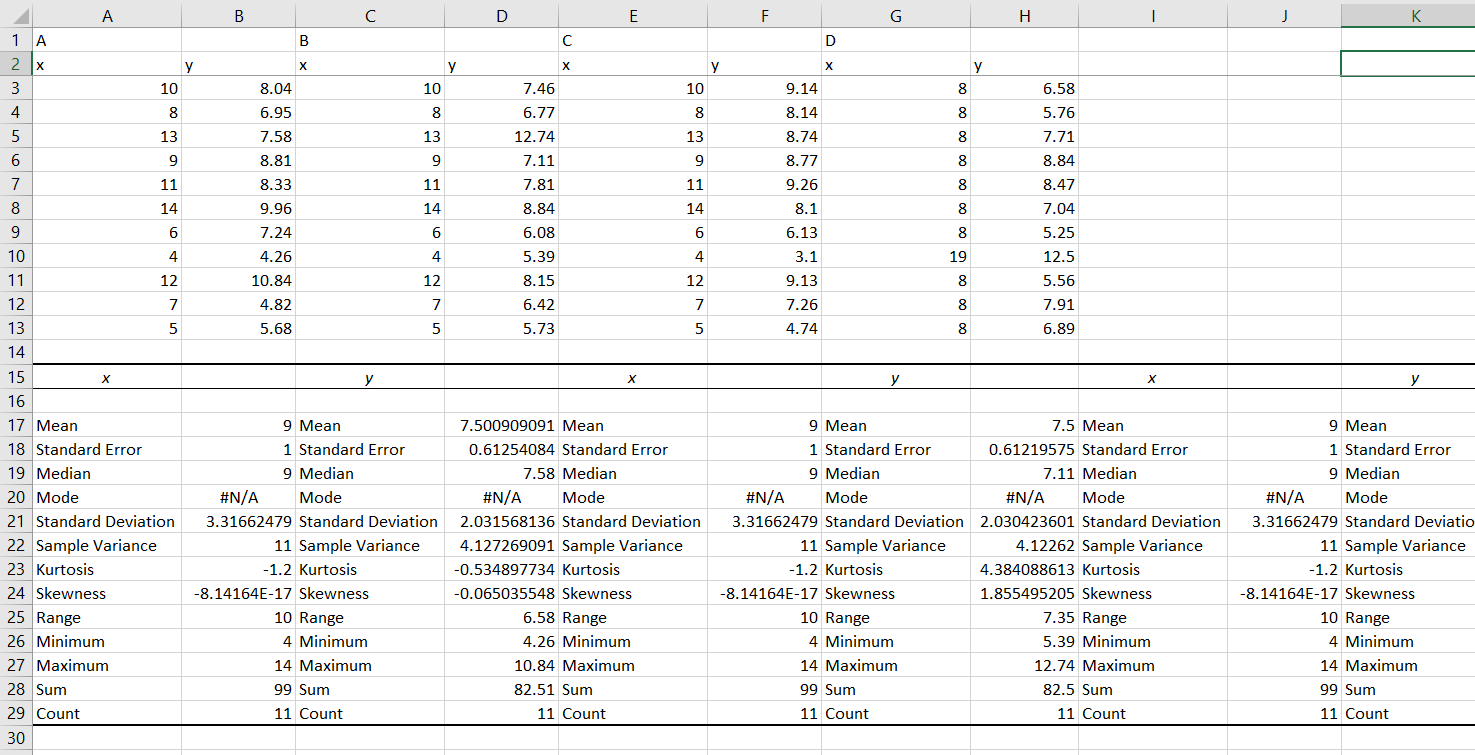
**Careful with correlations!**

File: anscombe.xlsx

1. Perform descriptive statistics on all variables by selecting Descriptive Statistics from the ToolPak. Make sure you select “Summary statistics” from the Output Options.



1. From the results of these descriptive statistics, each X-Y pair is *very* similar.



1. What happens if we graph this data using a scatter plot? Highlight the range and select Insert > Recommended Charts > Scatter.
   1. Only from the visualization can we easily see that these datasets are quite different!

