Exercise 1:

No Google Account?
Go Here to sign up
Make a copy of our DataSet Google Sheet

1. Lets explore our dataset
2. Determining the Domain of a data set
    - Open this tab
    - Using the Unique function determine the domain of the various columns
    - A full list of Google Sheet functions can be found HERE

Exercise 2

Let's build a Histogram Chart to review our data distribution

- Open the first tab of our dataset HERE.
- Go to the last column and select the entire SalesPrice column
- On the menu select ‘Insert’ and then ‘Chart’
- On the right you will see a Chart Editor and a Setup tab
- Under Setup select ‘Chart Type’ and find the ‘Histogram Chart’
- Now select the ‘Customize’ Tab in Chart Editor
- Find the Histogram section and expand it
- Let's select a bucket size of $50,000 or 50000

The bucket size is the range of values that are grouped together to form a single bar. It is also known as bin width or bin size.

We are grouping all prices that fall within each $50,000 range together. Each bar represents the number of houses that fall within the specified price range. For example, the bar on the far left represents the number of houses that are priced between $0 and $50,000. The height of each bar represents the frequency or count of houses in that price range. By adjusting the bucket size, we can control the level of detail in the histogram. A smaller bucket size would result in a more detailed histogram, while a larger bucket size would result in a more generalized histogram.

Exercise 3: Collect Semi-Structured Data

1. Lets Collect some structured Data using Google Sheets
    - Google Sheet list of functions can be found here
        - https://support.google.com/docs/table/25273?hl=en
    - Open the spreadsheet
    - =IMPORTHTML("https://finance.yahoo.com/quote/TSLA/", "table", 0)
2. Lets collect some unstructured data
    - The demographics of the USA
    - Go to https://en.wikipedia.org/wiki/Demographics_of_the_United_States
    - Use Google Sheets Function IMPORTHTML to extract the data
        - https://support.google.com/docs/answer/3093339
3. Let's use a more advanced Tool - DataMiner
    - https://chrome.google.com/webstore/detail/data-scraper-easy-web-scr/nndknepjnldbdbepjfgmncbggmopgden?hl=en
    - Lets scrape this site
        - https://dataminer.io/sandbox/index

Exercise 4:

1. Lets determined the key characteristics of our dataset
    - Open this tab
    - Use the following formulas to ‘describe’ our dataset

=COUNT(A:A)
=sum(A:A)
=MIN(A:A)
=max(A:A)
Range is MAX - MIN
=AVERAGE(A:A)
=MEDIAN(A:A)
=MODE(A:A)

=STDEV(A:A)
=VAR(A:A)
=QUARTILE(A:A,1)
=QUARTILE(A:A,2)
=QUARTILE(A:A,3)
=QUARTILE(A:A,4)
=SKEW(B:B)
=KURT(B:B)

Exercise 5:

1. Lets calculate the Standard Deviation of our sales prices
   - Open this tab
   - Recall the formula:

$$\text{Standard Deviation} = \sqrt{(\Sigma(x_i - \mu)^2 / n)}$$

   - Use the following steps to calculate

i. Calculate the Difference versus the mean, ($180,921)
ii. Square the differences, COLUMN^2
iii. Sum the squared differences, =SUM(COLU<N)
iv. Divid by the number of data entries, (1460)
v. Get the Square Root of the result, =SQRT(COLUMN)

Exercise 6:

1. Lets generate the correlations for some of our data
    - Open this tab
    - Use the following steps to calculate

        i. Select the Living Area and Sales Price Columns
        ii. In the Menu look for ‘Insert’ and Select ‘Chart’
        iii. In the Chart Editor select Setup
            - You will then see ‘Chart Type’. Select Scatter Chart
        iv. Lets Insert a trend line
            - In the Chart Editor select ‘Customize’
            - Move down to the ‘Series’ section in the Chart Editor
            - Select ‘Trend Line’ Check box near the bottom

Exercise 7:

1. Lets see what happens when we remove outliers
    - Open this tab
    - Use the following steps to find and remove

        i. Look for the yellow rows
        ii. Copy your previous scatter chart to that area
        iii. Remove one row and then the next and observe the trend line
        iv. Observe how easy it is to ‘manipulate data’

Exercise 8:

Let’s prepare and transform some data

- Open this tab

1. using the function COUNTIF to find N/A in the column.
   a. The syntax is =COUNTIF(B3:B8,"*") which counts any cells with characters in it. That’s what this wildcard character * represents
   b. Use =COUNTBLANK(A:A) to find blank cells
   c. Use =SUBSTITUTE(A:A,"N/A",0) to replace “N/A” with value 0
   d. Lets replace blank cells with the average size
      i. Calculate the average basement size
      ii. For Each cell use =ISBLANK(CELL) to determine if blank or not
      iii. Use IF Function to replace TRUE with Average value
      IF(logical_expression, value_if_true, value_if_false)

Exercise 9:

1. We can use Pivot Tables to answer questions such as:
    - What is the total sale by house price
    - how many house are there in the dataset by house type?
    - What's the average cost of the houses by type?
    - What is the highest price for each house type?

Lets Try it:

- Open this tab
- In the Sheet menu to to "Insert" and "Pivot Table"
- In the 'Create Pivot' popup
    i. For 'Data Range' select both columns
    ii. For 'Insert To', select the existing sheet and pick a cell to place the pivot table
    iii. You'll now see the 'Pivot Table Editor'
        - In Rows select 'Add' and select 'BldgType'
        - In Values select 'Add' and select 'SalesPrice'

- Now you can select SUM, AVERAGE, and MAX

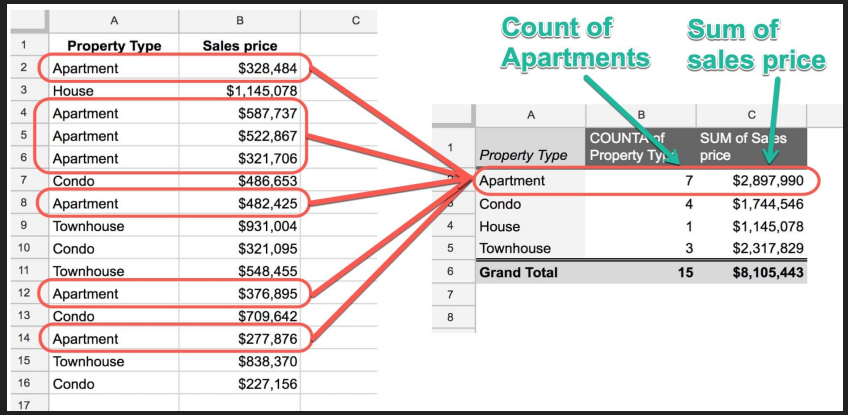

Image credit: benlcollins.com