Data Preparation

Summary: Data comes in all different types and formats. It needs to be cleaned, formatted, and blended properly so that you can conduct your analysis

STEP 1: Gathering
When gathering data, you may need to collect data from multiple sources. This can involve doing SQL queries, scraping data off the internet, or gathering csvs and excel sheets.

Useful Alteryx tool: Input Data

STEP 2: Cleansing
The data set you are working with may have issues that you want to resolve prior to your analysis. This can be in the form of incorrect, duplicate, outlier, or missing data. The data will need to be corrected, deleted, or imputed based on the analysis being conducted. Look for misspelled data, extra characters, and null values then take the appropriate action to adjust for each issue.

Useful Alteryx tool: Data Cleansing, Text to Columns, Formula, Field Summary

STEP 3: Formatting
You may need to format the data by renaming fields, changing variable types, reformatting date variables, summarizing, or pivoting the data. Transposing the data creates records for column names, and cross tabulation creates columns out of record values. Aggregation provides summary data for various groupings in the data.

Here is a quick overview of the most common types of data:

- Strings - combination of characters, can be alpha-numeric including symbols
- Numeric - numbers which can be whole numbers such as integers or numbers with decimal places.
- Date/Time - data can contain a specific date or a combination of both date and time.
- Boolean - also called the Logical type and is a conditional flag representing either true or false.

Useful Alteryx tools: Select, Auto Field, Transpose, Cross Tabulation, Summarize

STEP 4: Blending
You may want to blend, or combine, your data with other datasets to enrich it with additional variables.

- Union - add more records to data that has the same fields
- Join - add more columns to data from another data set (need column to join on)
- Fuzzy Matching - join data on records that are similar but spelled a little differently
- Spatial Matching - join data based on geolocation fields

Useful Alteryx tool: Join, Union, Fuzzy Matching, Spatial Matching

STEP 5: Sampling
Lastly, you may want to sample the dataset and work with a more manageable number of records.This will help your workflows run faster while you are still testing the workflow out.

Useful Alteryx tool: Sample