datagenomix™

Best Practices for Data Quality Assurance

- **Ensure Consistent Formatting in Data Entry**
  - Make sure all data entries have consistent formatting within their respective variables
  - Note the difference in formatting between the data for Customers 1 and 2 below:

	A	B	C	D	E	F	G	H	I
1	**Customer ID**	**First Name**	**Last Name**	**Date of Birth**	**Phone Number**	**Mailing Address**	**City**	**State**	**ZIP**
2	1514070332	John	Smith	5/7/80	(216)407-9932	256 W 3rd St	St. Paul	MN	58271
3	5638662901	jane	Doe	12/9/83	4409983746	9882 South Fifth St	Saint Paul	Minnesota	58273-8443

  - The data below has been formatted consistently:

	A	B	C	D	E	F	G	H	I
1	**Customer ID**	**First Name**	**Last Name**	**Date of Birth**	**Phone Number**	**Mailing Address**	**City**	**State**	**ZIP**
2	1514070332	John	Smith	5/7/80	(216)407-9932	256 W 3rd St	St. Paul	MN	58271
3	5638662901	Jane	Doe	12/9/83	(440)998-3746	9882 S 5th St	St. Paul	MN	58273

  - When sending multiple data files for analysis, ensure that all variables containing the same information have the same naming convention (ex: the columns containing ZIP code info in both files are both labeled ‘ZIP’)

- **Ensure Data Is Free of Duplicate Entries (Within Rows and Columns)**
  - Duplicate entries can skew the results of targeted advertisements, as advertising algorithms will treat the duplicates as two separate, unique entries, while both entries contain the same data
  - Some targeting algorithms (such as Facebook) limit the number of individuals that can be targeted per data set. Duplicate entries within a data set will take up unnecessary ‘slots’ within the targeting algorithms that could otherwise be used to target another individual
  - Eliminating duplicate entries will ensure maximum efficiency of the targeting algorithms

- **Ensure All Data Is Up-to-Date (Or As Close To Up-to-Date As Possible)**
    - Targeted advertising quickly loses accuracy the older the data is, as old data will contain outdated information that has since changed
    - Consumer/customer data is especially susceptible to inaccuracy since many variables (ex: age, mailing address, job information, etc) can change rapidly and without notice
    - Make sure all data is as up-to-date as possible to improve efficiency and accuracy of targeting algorithms

- **Ensure Data Set Is Sufficiently Large and Fields Are As Complete As Possible**
    - Data sets must be sufficiently large in order to run analyses and gain accurate insights (good rule of thumb: at least 5000 data entries)
    - Blank or poorly-populated fields within a data file make analysis difficult
    - Blank fields will either need to be imputated using values from other, similar entries within the data set (a time-consuming and inexact process), or be disregarded altogether
    - Insights on data sets will be easier to conduct and will be more accurate as a data set becomes more complete
    - Data is often hard to comprehend in a vacuum; explanation of the variable names in a data set and what these variables represent is helpful, especially for variables that can't be easily interpreted

- **Eliminate All Irrelevant Fields From Data File**
    - Including irrelevant data fields greatly increases the size of data files, leading to slower loading/processing times and higher costs for storing the data.
    - Irrelevant fields within the data will also detract from the ultimate goal of the data set; overcrowded, messy data sets make for more difficult analysis
    - It is still better to have too much data than not enough data; however, it is important to not pass the point of diminishing returns

- **Ensure At Least One Variable Includes a Unique Identifier**
    - Including a unique identifier for each data entry allows for quick elimination of potential duplicate entries
    - Unique identifiers also allow data files to be easily merged with other data files if more/newer data is to be analyzed in the future
    - Unique identifiers also help distinguish between data entries that may have similar attributes within other variables
    - Examples of unique identifiers: Voter ID numbers, Customer ID numbers, Account ID numbers, Social Security numbers, etc.
    - Ex: two customers, both named John Smith, were born on the same day and both live in St. Paul, MN. The unique identifier, Customer ID, helps distinguish between these two different people that might otherwise be impossible (see table below):

	A	B	C	D	E	F	G	H	I
1	**Customer ID**	**First Name**	**Last Name**	**Date of Birth**	**Phone Number**	**Mailing Address**	**City**	**State**	**ZIP**
2	1514070332	John	Smith	5/7/80	(216)407-9932	256 W 3rd St	St. Paul	MN	58271
3	1586458989	John	Smith	5/7/80	(216)523-0177	7496 N High St	St. Paul	MN	58271

- **Ensure Data is Error-Free**
    - Errors within data sets can rapidly decrease the quality of even the most complete and up-to-date data.
    - Spelling errors are the most frequent type of error within data sets, followed by formatting errors
    - Formatting errors are usually able to be corrected by a data scientist, whereas spelling errors are much more difficult to correct
    - Spelling errors can often go unnoticed and are tough to correct without knowing the correct value for the data point
    - Try to avoid manually typing in data whenever possible, instead opting to download data directly from the original source, or utilizing copy/paste functions from the original source to the data set to be analyzed
    - Ex: in the data below, the customer's first name is misspelled as 'Joh'. Without any previous information on the data, it is impossible to know if the customer's first name should be corrected to John, Joe, Jon, etc.)

	A	B	C	D	E	F	G	H	I
1	**Customer ID**	**First Name**	**Last Name**	**Date of Birth**	**Phone Number**	**Mailing Address**	**City**	**State**	**ZIP**
2	1514070332	Joh	Smith	5/7/80	(216)407-9932	256 W 3rd St	St. Paul	MN	58271