# Capstone Project Data Wrangling

Daniel Merino

Foundations of Data Science

1. Data Wrangling the Census Data (in MS Excel)
   1. Deleted the first row which contained metadata.
   2. Deleted irrelevant columns.
   3. Combined male age columns 0-14 as a sum; named the column male\_csum (children). Combined male age columns 15-85+ as a sum; named the column male\_asum (adults). Deleted other age columns and kept the sums. Repeated this for the female columns.
   4. Renamed other remaining columns to ZIP for ZIP Code, and TotalPop for Total Population.
2. Data Wrangling the Sites Data (in MS Excel)
   1. Deleted all columns except for SITE and ZIP.
   2. Renamed SITE to Site to easily merge with other data files.
3. Data Wrangling the Men’s Apparel Data (in MS Excel & R)
   1. Deleted columns A-C, F, H-N, P, S-AD, AF-AQ (irrelevant data).
   2. Reformatted column names to not include spaces or extra lines.
   3. Deleted discontinued items (starting with \*).
   4. Deleted items with no cost (generic items which stores use – can’t do anything with that data).
   5. Imported to R.
   6. Reformatted all columns to the correct types (numeric, factor, etc).
4. Data Wrangling the Ladies Apparel Data (in MS Excel & R)
   1. Deleted columns A-C, F, H-N, P, S-AD, AF-AQ (irrelevant data).
   2. Reformatted column names to not include spaces or extra lines.
   3. Deleted items with no cost (generic items which stores use – can’t do anything with that data).
   4. Imported to R.
   5. Reformatted all columns to the correct types (numeric, factor, etc).
5. Data Wrangling the Children’s Apparel Data (in MS Excel & R)
   1. Deleted columns A-C, F, H-N, P, S-AD, AF-AQ (irrelevant data).
   2. Reformatted column names to not include spaces or extra lines.
   3. Deleted items with no cost (generic items which stores use – can’t do anything with that data).
   4. Imported to R.
   5. Reformatted all columns to the correct types (numeric, factor, etc).
6. MERGING ALL FILES (in R)
   1. Merged Mens dataframe with Sites dataframe.
   2. Merged again with Census dataframe.
   3. Merged again with ZIP Coordinates dataframe (contains Latitude & Longitude; to be used in Tableau geospatial analysis).
   4. Repeated for Ladies and Childrens data.
7. Exported the 3 dataframes to Excel files.