Capstone Project 1: Data Wrangling

1. **What kind of cleaning steps did I perform?**

<u>**For genotypic data:**</u>

Step 1: Checked the dimension of the two genotypic data sets and found that there were mismatches.

Step 2: Merge two genotypic dataset with 'inner' option based on SNP id.

Step 3: Found out common lines that have both genotypic and phenotypic data.

Step 4: Got a subset of genotypic data that match phenotypic data.

Step 5: Imputed missing values for genotypic data.

Step 6: Drop monomorphic SNP data.

<u>**For phenotypic data:**</u>

Step 1: Checked the dimension of phenotypic data and found the mismatch with genotypic data

Step 2: Found out common lines that have both phenotypic and genotypic data.

Step 3: Found duplicates in lines' Id and dropped those duplicates.

Step 4: Using Z-score (-3 and 3) as thresholds to find out outliers.

Step 5: Wrote out a subset of phenotypic data that match genotypic data and without duplicates and outliers.

2. **How did I deal with missing values, if any?**

<u>**For genotypic data:**</u> Using 'Middle' genotypic code (0) replaced missing values [in 240].

<u>**For phenotypic data:**</u> Found low proportion missing value for protein and oil (<0.001) trait. Missing values were kept. Relative high proportions of missing values were found in yield data. All missing values were discarded.

3. **Were there outliers, and how did I handle them?**

<u>**For genotypic data:**</u> SNP data with missing data more than 20% were considered as outliers. I discarded those markers.

<u>**For phenotypic data**</u>: There were outliers for all three traits. I used Z-score , namely >-3 and < 3, as thresholds to find out outliers.