Capstone Project 1: Data Wrangling

1. What kind of cleaning steps did I perform?

For genotypic data:

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: Created a subset of genotypic data that match phenotypic data.

Step 5: Replace the missing values with the major value for genotypic data.

Step 6: Unify the heterozygous genotype code among two genotypic data sets

Step 7: Drop monomorphic SNP data.

Step 8: Drop SNP with minor allele frequency (MAF) less than 5%.

Step 9: Drop SNP with high heterozygous ration (large than 10%).

Step 10: Wrote out a cleaned genotypic data set.

For phenotypic data:

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 1.5 IQR (<-1.5 IQR and >1.5 IQR) 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?

For genotypic data: Using 'major' genotypic code replaced missing values.

For phenotypic data: 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?

For genotypic data: SNP data with monomorphic SNP, missing data more than 20% and high heterozygous(large than 10%) ratio were considered as outliers. I discarded those markers.

For phenotypic data: There were outliers for all three traits. I used 1.5 IQR, namely <-1.5 IQR and >1.5 IQR, as thresholds to find out outliers.