1. Code.xlsx

a. Description

This dataset consists of 2 sheets and the first sheet contains 2815 observations with 8 variables and the second sheet is empty (so we will ignore this). There are few duplicated values in "Inbred line" and "Accesion N." We may need to remove them before performing "join" (merge) two datasets. Currently, I did not drop any observations and perform "join." Please see below to see detail.

Variable Description

Inbred line	The homozygous genotypes.
Accesion N (Accession name)	A unique identifier given to a DNA or protein sequence record to allow for tracking of different versions of that sequence record.
N GBS samples	Number of Genotyping By Sequencing (GBS) Sample (simply number of samples).
N Plants	Number of plants
Avg. IBS (Identical By State)	Average IBS value for all the samples. (IBS: measurement to describe how similar two sequences of DNA between 0 and 1) * NA if N GBS samples = N plants = 1.
% missing	Percentage of missing data
Breeding program	Breeding program
Pop structure (Population structure)	Pedigree group; the organization of genetic variation.

b. Duplicated Values

In the "Inbred Line," "PHW86" has 2 observations (2304th and 2305th).

```
# A tibble: 2 x 8
  `Inbred line` `Accesion N` `N GBS samples` `N Plants` `Avg. IBS` `% missing` `Breeding program` `Pop structure`
  <chr>         <chr>                  <dbl>      <dbl> <chr>           <dbl> <chr>              <chr>
1 PHW86         PI543850                   2          2 0.997            0.75 ExPVP              unclassified
2 PHW86         PI543850                   3          3 0.998            0.28 Other              unclassified
```

In the "Accesion N," "Landraces" has 4 observations, "Ames27101" has 2 observations, "Ames27260" has 2 observations, and "PI543850" has 2 observations (their inbred lines are "PHW86").

Landraces

```
# A tibble: 4 x 8
  `Inbred line`           `Accesion N` `N GBS samples` `N Plants` `Avg. IBS`          `% missing` `Breeding program` `Pop structure`
  <chr>                   <chr>                  <dbl>      <dbl> <chr>                     <dbl> <chr>              <chr>
1 MR06ChapaloteS6         Landraces                  1          1 NA                         0.53 Other              landraces
2 MR13HickoryKing         Landraces                  2          2 0.85499999999999998        0.6  Other              landraces
3 MR15PalomerodeJaliscoS6 Landraces                  2          2 0.85799999999999998        0.44 Other              landraces
4 MR26PolloS5             Landraces                  1          1 NA                         0.66 Other              landraces
```

Ames27101

```
# A tibble: 2 x 8
  `Inbred line` `Accesion N` `N GBS samples` `N Plants` `Avg. IBS` `% missing` `Breeding program` `Pop structure`
  <chr>         <chr>                  <dbl>      <dbl> <chr>            <dbl> <chr>             <chr>
1 CML333        Ames27101                  9          3 0.995             0.23 Mexico            tropical
2 CO255         Ames27101                  3          2 0.998             0.31 Ontario           unclassified
```

Ames27260

```
# A tibble: 2 x 8
  `Inbred line` `Accesion N` `N GBS samples` `N Plants` `Avg. IBS` `% missing` `Breeding program` `Pop structure`
  <chr>         <chr>                  <dbl>      <dbl> <chr>            <dbl> <chr>             <chr>
1 Ki44          Ames27260                  2          1 0.998             0.54 Thailand          tropical
2 KUI44         Ames27260                  1          1 NA                0.6  Thailand          tropical
```

PI543850

See above “Inbred Line” = “PHW86”.

2. Kernel_Color_Data.xlsx

a. Description

This dataset consists of 4 sheets and
1) the first sheet contains 1547 observations with 26 variables – Final_Product
2) the second sheet contains 1595 observations with 26 variables – Heavy_Lfiting
3) the third sheet contains 4476 observations with 25 variables – Genotype_Data
4) the fourth sheet contains 2648 observations with 4 variables – Phenotype_Data.

We can see “Complete_name” has three parts divided by “:” (colon). We assume the first part corresponds to the “Accesion N” in Code.xlsx.

b. Duplicated Values

Final_Product, Heavy_Lfiting, and Phenotype_data do not have duplicated value.
For Genotype_data, out of 4476 observations, 1864 observations are duplicated.
E.g. Top 10 duplicated Accesion Name in Genotype_data:

```
B73      SA24 Ames28291      Ki11       P39  PI601573     Tx303 Ames19311      B103       B97
 35        30        11         8         8         8         8         7         7         7
```

(Continue on the next page)

3. Join Process

a. Duplicated value from Code.xlsx in each sheet from Kernel_Color_Data.xlsx

	Landraces	Ames27101	Ames27260	PI543850
Final_Proudct	NO	NO	YES	YES
Heavy_Lfiting	NO	NO	YES	YES
Genotype	NO	YES	YES	YES
Phenotype	NO	NO	YES	YES

Since Genotype and Code have duplicated, we need a rule for join (i.e. how we will handle duplicated values).
For Final_Product, Heavy_Lfiting, and Phenotype_Data, we can perform a join in this way:
E.g. Joining Code and Final_Product:

Code.xlsx				
Inbred line	Accesion N	Column 1	Column 2	…
Ki44	Ames27260	A	C	…
KUI44	Ames27260	B	D	…

+ (Inner Join)

Final_Product sheet			
Complete_name	Column 1.1	Column 2.1	…
Ames27260	a	b	…

=

Code_Final_Product.xlsx						
Inbred line	Accesion N	Column 1	Column 2	Column 1.1	Column 2.1	…
Ki44	Ames27260	A	C	a	b	…
KUI44	Ames27260	B	D	a	b	…

* Based on “Accesion N” in Code.xlsx, performed inner join. That is, if there is a observation presents in Final_Product sheet but not in the Code.xlsx, this observaition will be removed in the combined dataset.

4. Combined data

Combined_Final_Product.xlsx: 1547 x 34
Combined_Heavy_Lfiting.xlsx: 1595 x 34
Combined_Phenotype_Data.xlsx: 2316 x 12