In Samples\_2004: missed one row where Origin = 3; I replaced it with Unknown. Please make a note in data cleaning doc that "3" = Unknown Origin, and was replaced with such)

Additional data cleaning: as Mayuresh mentioned previously, "Variety" is a free form text field, so there are various versions of variety names entered that had to be unified. A new column named "Grade" is manually added to store the fancy descriptive grade information in the unification process.

"Organic" in name

"Extra Fancy" in name

"Premium" in name

Renamed the following and stored the fancy grade info manually in a separate column:

Red Del Xfcy 88 Ct

Red Delicious Fancy

Red Delicious X Fanc

Red Del Xfcy

Extra Fancy Red Deli

Red Del Waex Fcy

Wash. Red Delicious

Wa Red Delicious

Sweet Tango Red Deli

To: "Red Delicious"

Wa Golden Delicious

Washington Golden De

Gold Delicious

To: Golden Delicious

Premium Braburn

To: "Braeburn"

Fuji Apple

Fuji Extra Fancy

Extra Fancy Fuji

Aztec Fuji

To: "Fuji"

Organic Gala

Fancy Gala

Royal Gala

Premium Gala

To: "Gala"

Honey Crisp

To: "Honeycrisp"

Ex Fancy Granny Smit

Granny Smith Premium

To: "Granny Smith"

Mcintosh

Macintosh

To: McIntosh

Gingergold

To: Ginger Gold

Empire Crist-Aire

To: "Empire"

Law Rome

Red Rome

To: Rome

Cameo Washington Ex

To: Cameo

Washington Ex Fancy

Washington Extra Fan

To: Unknown (Note: Because this is not a variety but a grade level)

None

Not Stated

Unknown

NA

To: "Unknown"