Establishing Data

- NAME: Chloe Wang
- ID: Qtr7bs

In [3]:
```
import pandas as pd
```

In [4]:
```
#https://www.kaggle.com/datasets/fatihb/coffee-quality-data-cqi
```

In [5]:
```
pd.set_option('display.max_columns', None)
df = pd.read_csv("df_arabica_clean.csv")
df.head()
```

Out[5]:

	Unnamed: 0	ID	Country of Origin	Farm Name	Lot Number	Mill	ICO Number	Company	A
0	0	0	Colombia	Finca El Paraiso	CQU2022015	Finca El Paraiso	NaN	Coffee Quality Union	
1	1	1	Taiwan	Royal Bean Geisha Estate	The 2022 Pacific Rim Coffee Summit,T037	Royal Bean Geisha Estate	NaN	Taiwan Coffee Laboratory	
2	2	2	Laos	OKLAO coffee farms	The 2022 Pacific Rim Coffee Summit,LA01	oklao coffee processing plant	NaN	Taiwan Coffee Laboratory	
3	3	3	Costa Rica	La Cumbre	CQU2022017	La Montana Tarrazu MIll	NaN	Coffee Quality Union	
4	4	4	Colombia	Finca Santuario	CQU2023002	Finca Santuario	NaN	Coffee Quality Union	

2) I acquired this dataset through Kaggle. My teammate and I wanted something to do with coffee; therefore, we tried looking for datasets that were not only about coffee but also had a good amount of data entries. Here, the amount of data (in rows) is about 207, which is one of the higher numbers within Kaggle's coffee datasets. After that, I went into the 'input' tab and hit download on the 'df_arabica_clean.csv'.

3) This dataset is available on Kaggle through a user named 'pannmie'; however, the original dataset was provided by the Coffee Quality Union in May 2023. Specifically, the dataset came from the Coffee Quality Institute (CQI) branch, which educates on coffee quality. The data itself is submitted by either the producer or the certifiers. Evaluations come from the certifier. For example, CQI outlines a new metric, "Total Cup Points", which is based on an amalgamation of sensory evaluations, namely aroma, flavor, aftertaste, acidity, body, balance, uniformity, clean cup, sweetness, and overall. Total Cup Points is the sum of all the subjective scores given to the above attributes, subtracting any defects noted.

In [16]:
```
cols = pd.DataFrame({
    "Column Name": df.columns,
    "Type": df.dtypes.astype(str)
})
cols = pd.DataFrame({
    "Column Name": df.columns,
    "Type": df.dtypes.astype(str),
    "Description": ""  # fill this in by hand in your writeup
})

cols.loc[cols["Column Name"] == "ID", "Description"] = "Unique identifier fo
cols.loc[cols["Column Name"] == "Country of Origin", "Description"] = "Count
cols.loc[cols["Column Name"] == "Farm Name", "Description"] = "Name of the f
cols.loc[cols["Column Name"] == "Lot Number", "Description"] = "Producer or
cols.loc[cols["Column Name"] == "Mill", "Description"] = "Wet or dry mill wh
cols.loc[cols["Column Name"] == "ICO Number", "Description"] = "Internationa
cols.loc[cols["Column Name"] == "Company", "Description"] = "Company or orga
cols.loc[cols["Column Name"] == "Altitude", "Description"] = "Growing altitu
cols.loc[cols["Column Name"] == "Region", "Description"] = "Region where the
cols.loc[cols["Column Name"] == "Producer", "Description"] = "Producer, coop
cols.loc[cols["Column Name"] == "Number of Bags", "Description"] = "Number o
cols.loc[cols["Column Name"] == "Bag Weight", "Description"] = "Weight of ea
cols.loc[cols["Column Name"] == "In-Country Partner", "Description"] = "CQI-
cols.loc[cols["Column Name"] == "Harvest Year", "Description"] = "Year or ha
cols.loc[cols["Column Name"] == "Grading Date", "Description"] = "Date on wh
cols.loc[cols["Column Name"] == "Owner", "Description"] = "Owner of the coff
cols.loc[cols["Column Name"] == "Variety", "Description"] = "Botanical coffe
cols.loc[cols["Column Name"] == "Status", "Description"] = "Status of the sa
cols.loc[cols["Column Name"] == "Processing Method", "Description"] = "Post-
cols.loc[cols["Column Name"] == "Aroma", "Description"] = "Cupping score for
cols.loc[cols["Column Name"] == "Flavor", "Description"] = "Cupping score fo
cols.loc[cols["Column Name"] == "Aftertaste", "Description"] = "Cupping scor
cols.loc[cols["Column Name"] == "Acidity", "Description"] = "Cupping score f
cols.loc[cols["Column Name"] == "Body", "Description"] = "Cupping score for
cols.loc[cols["Column Name"] == "Balance", "Description"] = "Cupping score f
cols.loc[cols["Column Name"] == "Uniformity", "Description"] = "Score repres
```

```
cols.loc[cols["Column Name"] == "Clean Cup", "Description"] = "Score for abs
cols.loc[cols["Column Name"] == "Sweetness", "Description"] = "Score for swe
cols.loc[cols["Column Name"] == "Overall", "Description"] = "Overall impress
cols.loc[cols["Column Name"] == "Defects", "Description"] = "Total defect-re
cols.loc[cols["Column Name"] == "Total Cup Points", "Description"] = "Total
cols.loc[cols["Column Name"] == "Moisture Percentage", "Description"] = "Mea
cols.loc[cols["Column Name"] == "Category One Defects", "Description"] = "Co
cols.loc[cols["Column Name"] == "Quakers", "Description"] = "Number of quake
cols.loc[cols["Column Name"] == "Color", "Description"] = "Color or appearar
cols.loc[cols["Column Name"] == "Category Two Defects", "Description"] = "Co
cols.loc[cols["Column Name"] == "Expiration", "Description"] = "Expiration d
cols.loc[cols["Column Name"] == "Certification Body", "Description"] = "Name
cols.loc[cols["Column Name"] == "Certification Address", "Description"] = "A
cols.loc[cols["Column Name"] == "Certification Contact", "Description"] = "C

cols
```

Out[16]:

	Column Name	Type	Description
Unnamed: 0	Unnamed: 0	int64	
ID	ID	int64	Unique identifier for each coffee record in th...
Country of Origin	Country of Origin	object	Country where the coffee was grown
Farm Name	Farm Name	object	Name of the farm where the coffee was grown
Lot Number	Lot Number	object	Producer or exporter batch identifier for this...
Mill	Mill	object	Wet or dry mill where the coffee was processed
ICO Number	ICO Number	object	International Coffee Organization identifier a...
Company	Company	object	Company or organization that submitted or owns...
Altitude	Altitude	object	Growing altitude of the coffee
Region	Region	object	Region where the coffee was produced
Producer	Producer	object	Producer, cooperative, or farmer responsible f...
Number of Bags	Number of Bags	int64	Number of bags represented by this row of data
Bag Weight	Bag Weight	object	Weight of each bag
In-Country Partner	In-Country Partner	object	CQI-affiliated partner organization that coord...
Harvest Year	Harvest Year	object	Year or harvest season when the coffee was har...
Grading Date	Grading Date	object	Date on which this coffee batch was graded
Owner	Owner	object	Owner of the coffee batch
Variety	Variety	object	Botanical coffee variety
Status	Status	object	Status of the sample or certificate (all compl...
Processing Method	Processing Method	object	Post-harvest processing method used (e.g., was...
Aroma	Aroma	float64	Cupping score for aroma of the coffee
Flavor	Flavor	float64	Cupping score for flavor of the coffee
Aftertaste	Aftertaste	float64	Cupping score for aftertaste of the coffee
Acidity	Acidity	float64	Cupping score for acidity of the coffee

	Column Name	Type	Description
Body	Body	float64	Cupping score for body or mouthfeel
Balance	Balance	float64	Cupping score for how well the different attri...
Uniformity	Uniformity	float64	Score representing consistency across multiple...
Clean Cup	Clean Cup	float64	Score for absence of off-flavors and presence ...
Sweetness	Sweetness	float64	Score for sweetness in the cup
Overall	Overall	float64	Overall impression score given by the cupper
Defects	Defects	float64	Total defect-related deductions or notes assoc...
Total Cup Points	Total Cup Points	float64	Total cupping score (sum of sensory attributes...
Moisture Percentage	Moisture Percentage	float64	Measured moisture content of the green coffee ...
Category One Defects	Category One Defects	int64	Count of primary (Category 1) green coffee def...
Quakers	Quakers	int64	Number of quakers (underdeveloped beans) detec...
Color	Color	object	Color or appearance description of the green c...
Category Two Defects	Category Two Defects	int64	Count of secondary (Category 2) green coffee d...
Expiration	Expiration	object	Expiration date of the evaluation or Q certifi...
Certification Body	Certification Body	object	Name of the organization or laboratory that ce...
Certification Address	Certification Address	object	Address of the certifying organization or labo...
Certification Contact	Certification Contact	object	Contact person or contact details for the cert...

In [8]:
```
len(df)
```

Out[8]: 207