In [1]:
```
import pandas as pd
import numpy as np
```

Step 1 - Data Engineering

Clint Goodman

In [2]:
```
# load CSV file into a variable
measurements_file = "Resources/hawaii_measurements.csv"
stations_file = "Resources/hawaii_stations.csv"

#load CSV file data in a dataframe
dfMeasurements = pd.read_csv(measurements_file)
dfStations = pd.read_csv(stations_file)
# dfMeasurements.head()
# dfStations.head()
```

In [3]:
```
# Examine data to find missing values
# dfMeasurements.describe()
dfMeasurements.isnull().sum() # - none
# dfStations.describe()
dfStations.isna().sum() # - 1447 missing from prcp
```

Out[3]:
```
station      0
name         0
latitude     0
longitude    0
elevation    0
dtype: int64
```

In [4]:
```
# I chose to replaced all Null/NaN values in the prcp column with
the arithmetic mean of the other columns that had a value.
# The other option was to simply drop the rows with Null/NaN in t
he prcp column.
# Using the average prcp value allows us to keep the other data p
oints for the other columns with little or no impact to the prcp
column
dfMeasurementsMean = dfMeasurements.fillna(dfMeasurements.mean())
# dfMeasurementsDrop = dfMeasurements.dropna()
```