Analysis of meteorites data using pandas\_profiling

Observations

Number of Variables: 14

Number of Observations: 45,726

Number of “cells”: 45726 \* 14 = 640,164

Missing Cells: (is this the same thing as a missing variable in a row?) 29,682

Percentage Missing Cells: 26,682/640164 =4.64% -- Matches worksheet

Variable Types:

7 categorical

Name

nametype

recclass

fall

GeoLocation – missing tide to latitude and longitude

source – All values are the same, so this was rejected

mixed

5 numerical

id – a numerical ID. All distinct

mass

reclat –16% missing, 14.1 % zero (this is the equator)

reclong – 16% missing, 13.6% zero (this is the prime meridian)

reclat\_city – 16% missing, tied to location

1 date

year

1 boolean

boolean – none are missing

Some duplicates exist but the name includes the word “copy” in the description

Mass is fairly skewed – but this may be because the bin size isn’t set correctly and because three seems to be an “Other values) category that is causing a problem – what is this other values? Small number of missing masses.