# Link to results notebook

https://github.com/vkdvamshi/DataCleanup/blob/master/DataCleanup.ipynb

# What kind of cleaning steps did you perform?

Given data of IDS various steps were performed resulting in cleaning data

a. Identifying junk characters and removing them like smiley, space , kg, kgs , cm, cms etc...

b. Converting all data into one metric like from tonns to kg conversion mm to cm conversion

c. Making data into one data type for complete column

d. converting data into datetime type

# How did you deal with missing values, if any?

e. use ffill & bfill methods to fill missing string values

f. use intrapolate to fill missing numbering values

# Were there outliers, and how did you handle them?

2 ways to identify outliers

univariant analysis

multi variant analysis

if we have categorical values with continous data then multi variant analysis is used

Ways to identify & handle outliers

a. box plot method ( visual method )

box-wisker plot helps easily identify outliers based on datapoints

import seaborn as sns

sns.boxplot(x=boston\_df['DIS'])

b. scattered plot ( visual method )

fig, ax = plt.subplots(figsize=(16,8))

ax.scatter(boston\_df['INDUS'], boston\_df['TAX'])

ax.set\_xlabel('Proportion of non-retail business acres per town')

ax.set\_ylabel('Full-value property-tax rate per $10,000')

plt.show()

c. z-score ( mathematical method )

The intuition behind Z-score is to describe any data point by finding their relationship with the Standard Deviation and Mean of the group of data points. Z-score is finding the distribution of data where mean is 0 and standard deviation is 1 i.e. normal distribution.

while calculating the Z-score we re-scale and center the data and look for data points which are too far from zero. These data points which are way too far from zero will be treated as the outliers

In most of the cases a threshold of 3 or -3 is used i.e if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers.

Detect Outliers:

from scipy import stats

import numpy as np

z = np.abs(stats.zscore(boston\_df))

print(z)

threshold = 3

print(np.where(z > 3))

print(z[55][1])

3.375038763517309

remove or filter the outliers and get the clean data.

boston\_df\_o = boston\_df\_o[(z < 3).all(axis=1)]

d. IQR Score ( mathematical method )

The interquartile range (IQR), also called the midspread or middle 50%, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1.

Q1 = boston\_df\_o1.quantile(0.25)

Q3 = boston\_df\_o1.quantile(0.75)

IQR = Q3 - Q1

print(IQR)

print(boston\_df\_o1 < (Q1 - 1.5 \* IQR)) |(boston\_df\_o1 > (Q3 + 1.5 \* IQR))

Fileter & remove outliers:

boston\_df\_out = boston\_df\_o1[~((boston\_df\_o1 < (Q1 - 1.5 \* IQR)) |(boston\_df\_o1 > (Q3 + 1.5 \* IQR))).any(axis=1)]

boston\_df\_out.shape

# Submit a link to the document.

Discuss it with your mentor at the next call.

Revise and resubmit if needed.

Convert the final document to a .pdf and add it to your GitHub repository for this project. This document will eventually become part of your milestone report.