Python Pandas Reference for Preprocessing

1. To read data from csv to pandas df

df=pd.read\_csv('carbon-monoxide-results-20181007-145932.csv')

df.dtypes to check types of each column

1. To check if cell contains numeric true or false val

df.isnull().sum()

will return columns with number of missing values in each

1. How to check percentage of values missing in each column

columns = df.columns

percent\_missing = df.isnull().sum() \* 100 / len(df)

missing\_value\_df = pd.DataFrame({'column\_name': columns,

'percent\_missing': percent\_missing})

missing\_value\_df.sort\_values('percent\_missing', inplace=True)

1. Impute missing values

from sklearn.preprocessing import Imputer

imp=Imputer(missing\_values='NaN', strategy='mean')

df["aqi"]=imp.fit\_transform(df[["aqi"]]).ravel()

df["method\_code"]=imp.fit\_transform(df[["method\_code"]]).ravel()

1. Fill missing with zeroes

## Fill missing values with zero

df["aqi"].fillna(0, inplace=True)

df["method\_code"].fillna(0, inplace=True)

SELECT \* FROM `bigquery-public-data.epa\_historical\_air\_quality.co\_daily\_summary`

where state\_name ='California' AND EXTRACT(YEAR FROM date\_local) = 2016

https://www.kaggle.com/epa/epa-historical-air-quality/kernels?sortBy=voteCount&group=everyone&pageSize=20&datasetId=5779