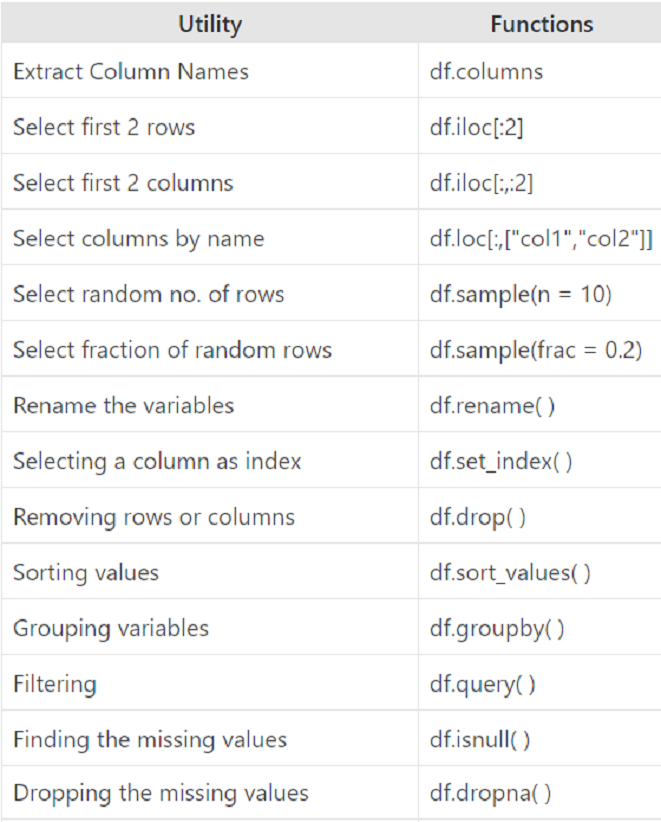
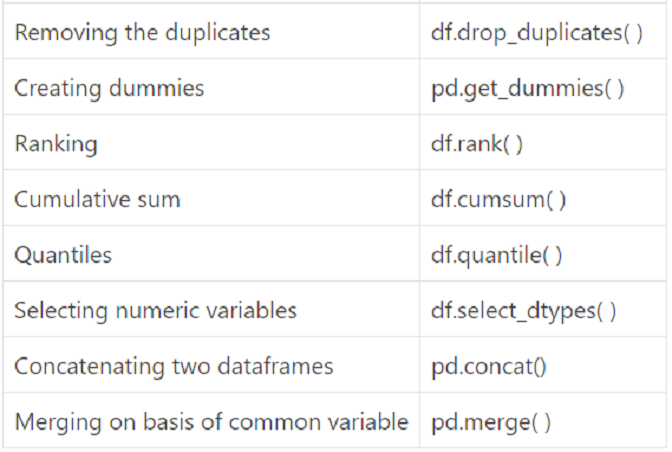
Pandas: popular package in Python for data manipulation

1. Create a structured data set (data frame) and Excel spreadsheet.
2. Reading data from various sources such as CSV, TXT, XLSX, SQL database, R etc.
3. Selecting particular rows or columns from data set
4. Arranging data in ascending or descending order
5. Filtering data based on some conditions
6. Summarizing data by classification variable
7. Reshape data into wide or long format
8. Time series analysis
9. Merging and concatenating two datasets
10. Iterate over the rows of dataset
11. Writing or Exporting data in CSV or Excel format

**Important pandas functions to remember**

****

****

**Importing pandas library**

loading it into the memory and then use it **import pandas as pd**

**Importing Dataset**

read\_csv() function

**dataFrameName= pd.read\_csv("C:\\Users\\FILENAME.csv")**

**Get Variable Names**

**dataFrameName.columns**

**dataFrameName.columns[0:2] :** returns first two column names

**Knowing the Variable types**

**dataFrameName.dtypes**

**Changing the data types**

**dataFrameName.ColName=dataFrameName.ColName.astype(datatype)**

**To view the dimensions or shape of the data dataFrameName.shape**

**dataFrameName.shape[0] : no. of rows**

**dataFrameName.shape[1] : no of columns**

**dataFrameName.head( ) shows first 5 rows dataFrameName[0:5]**

**dataFrameName.tail( ) shows last 5 rows**

**dataFrameName.head(2)**

**dataFrameName.ColName.unique() // shows the unique levels or categories in the dataset**

**nunique( ) shows the number of unique values**

**dataFrameName["colname"]**

**dataFrameName.colname**

**#Renaming all the variables.**

**dataFrameName.columns = ['NewName1','NewName2']**

**#Renaming only some of the variables.**

**dataFrameName.rename(columns = {"Names":"Cust\_Name"},inplace = True)**

By default in pandas inplace = False which means that no changes are made in the original dataset. Thus if we wish to alter the original dataset we need to define inplace = True

**Removing columns and rows**

**dataFrameName.drop('ColName',axis = 1)**

By default axis = 0 which row wise. To remove a column set axis = 1

**dataFrameName.drop("ColName",axis = "columns")**

**dataFrameName.drop(0,axis = 0)**

**Sorting Data**

By default inplace = False and ascending = True.

**dataFrameName.sort\_values("ColName",ascending = False)**

**dataFrameName.sort\_values("ColName",ascending = False,inplace = True)**

**dataFrameName.ColName.sort\_values()**

**Finding Descriptive Statistics**

**dataFrameName.describe() #for numeric variables mean,minimum, quartiles etc**

**dataFrameName.describe(include = ['object']) #Only for strings / objects total count, maximum occurring string and its frequency**

**To find out specific descriptive statistics of each column of data frame**

**dataFrameName.mean()**

**dataFrameName.median()**

**dataFrameName.agg(["mean","median"]) //performs aggregation with summary functions like sum, mean, median, min, max etc**

**dataFrameName.ColName.mean()**

**GroupBy function**

**dataFrameName.groupby("ColNameCat")["ColName1","ColName2"].min()**

**income.groupby("Index")["Y2002","Y2003"].agg(["min","max","mean"])**

**income.groupby("Index").agg({"Y2002": ["min","max"],"Y2003" : "mean"})**

**Filtering**

**dataFrameName[dataFrameName.ColNameCat== "Fruit"]**