7 8 5.0 3.4 1.5 0.2 Iris-setosa In [72]: #3.Display bottom rows in the datset dataset.tail() Out[72]: Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm **Species 145** 146 2.3 Iris-virginica 6.7 3.0 5.2 1.9 Iris-virginica **146** 147 6.3 2.5 **147** 148 2.0 Iris-virginica 6.5 3.0 5.2 **148** 149 6.2 3.4 2.3 Iris-virginica **149** 150 1.8 Iris-virginica In [73]: dataset.tail(3) #Displays bottom 3 rows in the dataset Out[73]: Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm **Species 147** 148 2.0 Iris-virginica **148** 149 3.4 2.3 Iris-virginica **149** 150 1.8 Iris-virginica In [74]: #4.Display count of rows and columns in the dataset dataset.shape Out[74]: (150, 6) In [75]: print(dataset.shape[0]) #Display rows count print(dataset.shape[1]) #Display columns count 150 6 In [76]: #5.Display summary of the Dataset dataset.info() <class 'pandas.core.frame.DataFrame'> RangeIndex: 150 entries, 0 to 149 Data columns (total 6 columns): 150 non-null int64 SepalLengthCm 150 non-null float64 SepalWidthCm 150 non-null float64 PetalLengthCm 150 non-null float64 PetalWidthCm 150 non-null float64 150 non-null object dtypes: float64(4), int64(1), object(1)memory usage: 7.1+ KB In [77]: #6.Display statistical summary of all columns dataset.describe() #statistical description of all numerical data columns Out[77]: Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm **count** 150.000000 150.000000 150.000000 150.000000 150.000000 mean 75.500000 5.843333 3.758667 3.054000 1.198667 **std** 43.445368 0.828066 0.433594 1.764420 0.763161 1.000000 min 1.000000 4.300000 2.000000 0.100000 25% 38.250000 5.100000 2.800000 1.600000 0.300000 4.350000 75.500000 5.800000 3.000000 1.300000 **75**% 112.750000 6.400000 3.300000 5.100000 1.800000 max 150.000000 7.900000 4.400000 6.900000 2.500000 In [78]: print(dataset.describe(include = ['O'])) #statistical description of categorical columns Species 150 count 3 unique Iris-setosa top freq In [79]: print(dataset.describe(include = 'all')) #statistical description of all columns Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm \ 150.000000 150.000000 150.000000 150.000000 150.000000 count unique NaN NaN NaN NaN NaN NaN NaN NaN NaN top freq NaN NaN NaN NaN NaN 75.500000 5.843333 3.054000 3.758667 1.198667 mean 0.433594 0.763161 43.445368 0.828066 1.764420 0.100000 min 1.000000 4.300000 2.000000 1.000000 25% 38.250000 5.100000 2.800000 1.600000 0.300000 50% 75.500000 5.800000 3.000000 4.350000 1.300000 75% 112.750000 6.400000 3.300000 5.100000 7.900000 max 150.000000 4.400000 6.900000 2.500000 Species 150 count unique 3 top Iris-setosa freq 50 NaN mean std NaN min NaN 25% NaN 50% NaN 75% NaN max NaN In [80]: #7.Display datatypes in datset dataset.dtypes #Datatypes of all columns in the dataset Out[80]: Id int64 SepalLengthCm float64 float64 SepalWidthCm PetalLengthCm float64 PetalWidthCm float64 Species object dtype: object In [81]: dataset.Id.dtype #Datatype of particular column in the dataset dataset['Id'].dtype Out[81]: dtype('int64') In [82]: #8.no of uique rows in each column dataset.nunique() #count of unique rows in all columns Out[82]: Id SepalLengthCm 35 SepalWidthCm 23 43 PetalLengthCm PetalWidthCm 22 Species 3 dtype: int64 In [83]: dataset.Id.nunique() #count of unique rows of a particular column dataset['Id'].nunique() Out[83]: 150 In [84]: #9.select row with particular name or row no dataset.loc[:,'Id'] #select all rows in column Id Out[84]: 0 2 3 3 4 5 6 9 10 10 11 11 12 12 13 13 14 14 15 16 17 18 19 19 20 20 21 21 22 22 24 25 25 26 26 27 27 28 28 29 29 30 121 120 121 122 122 123 123 124 125 124 125 126 126 127 127 128 128 129 130 131 130 131 132 132 133 133 134 134 135 135 136 136 137 137 138 138 139 139 140 140 141 141 142 143 142 143 144 145 147 146 147 148 148 149 150 Name: Id, Length: 150, dtype: int64 In [85]: dataset.iloc[:,1] #select all rows in column no 1 Out[85]: 0 5.1 4.7 4.6 5.4 4.6 5.0 4.9 5.4 11 4.8 13 4.3 14 5.8 15 5.7 5.4 5.1 18 5.7 5.1 19 20 4.6 23 5.1 4.8 25 5.0 26 5.0 27 5.2 28 5.2 4.7 120 6.9 121 5.6 122 7.7 123 6.3 124 6.7 125 7.2 126 127 6.1 128 6.4 129 7.2 130 7.4 7.9 131 132 6.4 133 6.3 134 6.1 135 7.7 136 6.3 137 138 6.0 139 6.9 140 6.7 141 6.9 142 143 144 6.7 145 6.7 146 6.3 147 6.5 148 6.2 149 5.9 Name: SepalLengthCm, Length: 150, dtype: float64 In [86]: #10.find rows with duplicate data sum(dataset.duplicated()) #count of rows with duplicated data Out[86]: 0 In [87]: dataset[dataset.duplicated()] #displays the rows with duplicate data. Out[87]: ld SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species In [88]: # 11. drop a column dataset.drop('Id',axis = 1) Out[88]: SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm **Species** 3.5 1.4 5.1 0.2 Iris-setosa 1 4.9 3.0 1.4 Iris-setosa 4.7 3.2 1.3 Iris-setosa 3 4.6 3.1 1.5 Iris-setosa 5.0 3.6 1.4 Iris-setosa 1.7 5.4 3.9 Iris-setosa 4.6 3.4 1.4 Iris-setosa 5.0 3.4 1.5 Iris-setosa 2.9 1.4 0.2 Iris-setosa 9 4.9 3.1 1.5 Iris-setosa 10 5.4 3.7 1.5 0.2 Iris-setosa 11 4.8 3.4 1.6 0.2 Iris-setosa 3.0 12 4.8 1.4 0.1 Iris-setosa 13 4.3 3.0 1.1 0.1 Iris-setosa 14 5.8 4.0 1.2 0.2 Iris-setosa 15 4.4 1.5 5.7 Iris-setosa 16 5.4 3.9 1.3 0.4 Iris-setosa 17 5.1 3.5 1.4 0.3 Iris-setosa 18 5.7 3.8 1.7 0.3 Iris-setosa 19 5.1 3.8 1.5 0.3 Iris-setosa 20 5.4 3.4 1.7 0.2 Iris-setosa 21 5.1 3.7 1.5 Iris-setosa 22 4.6 3.6 1.0 0.2 Iris-setosa 23 5.1 3.3 1.7 0.5 Iris-setosa 24 4.8 3.4 1.9 0.2 Iris-setosa 25 5.0 1.6 Iris-setosa 3.0 0.2 26 5.0 3.4 1.6 0.4 Iris-setosa 27 5.2 3.5 1.5 0.2 Iris-setosa 28 5.2 3.4 1.4 0.2 Iris-setosa 29 4.7 3.2 1.6 0.2 Iris-setosa 2.3 Iris-virginica 120 6.9 3.2 5.7 121 5.6 2.8 4.9 2.0 Iris-virginica 122 7.7 2.8 6.7 2.0 Iris-virginica 123 6.3 2.7 4.9 1.8 Iris-virginica 124 6.7 3.3 5.7 2.1 Iris-virginica 7.2 125 3.2 6.0 1.8 Iris-virginica 126 6.2 2.8 4.8 1.8 Iris-virginica 127 6.1 3.0 4.9 1.8 Iris-virginica 128 6.4 2.1 Iris-virginica 2.8 5.6 129 7.2 3.0 5.8 1.6 Iris-virginica 130 7.4 6.1 1.9 Iris-virginica 2.8 131 7.9 3.8 6.4 2.0 Iris-virginica 132 6.4 2.8 5.6 2.2 Iris-virginica 133 6.3 2.8 5.1 1.5 Iris-virginica 134 6.1 1.4 Iris-virginica 2.6 5.6 135 7.7 3.0 6.1 2.3 Iris-virginica 3.4 2.4 Iris-virginica 136 6.3 5.6 137 6.4 3.1 5.5 1.8 Iris-virginica 138 6.0 3.0 4.8 1.8 Iris-virginica 139 6.9 3.1 5.4 2.1 Iris-virginica 140 3.1 5.6 6.7 2.4 Iris-virginica 141 6.9 3.1 5.1 2.3 Iris-virginica 142 5.8 2.7 5.1 1.9 Iris-virginica 143 6.8 3.2 5.9 2.3 Iris-virginica 144 6.7 3.3 2.5 Iris-virginica 145 6.7 3.0 5.2 2.3 Iris-virginica 146 6.3 2.5 5.0 1.9 Iris-virginica 147 6.5 3.0 5.2 2.0 Iris-virginica 5.4 148 6.2 3.4 2.3 Iris-virginica 149 5.9 3.0 5.1 1.8 Iris-virginica 150 rows × 5 columns In [89]: #12.count of unique values in the column dataset['Species'].value_counts() Out[89]: Iris-setosa 50 50 Iris-versicolor Iris-virginica Name: Species, dtype: int64 **Advanced Commands** In [90]: #1. query command dataset.query('SepalLengthCm < 5') #fetch the rows in dataset which vae speallength less than 5</pre> Out[90]: Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm **Species** 1 4.9 1.4 0.2 Iris-setosa 2 3 4.7 3.2 1.3 0.2 Iris-setosa 4.6 3.1 1.5 0.2 Iris-setosa 7 6 4.6 3.4 1.4 0.3 Iris-setosa 2.9 1.4 Iris-setosa 9 10 4.9 3.1 1.5 0.1 Iris-setosa **11** 12 4.8 3.4 1.6 Iris-setosa **12** 13 4.8 3.0 0.1 1.4 Iris-setosa **13** 14 4.3 3.0 1.1 Iris-setosa **22** 23 4.6 3.6 1.0 0.2 Iris-setosa **24** 25 4.8 3.4 1.9 Iris-setosa **29** 30 4.7 3.2 1.6 0.2 Iris-setosa 31 4.8 3.1 1.6 30 Iris-setosa 34 35 4.9 3.1 1.5 0.1 Iris-setosa 37 38 4.9 3.1 1.5 Iris-setosa 38 39 4.4 3.0 1.3 0.2 Iris-setosa **41** 42 2.3 1.3 4.5 Iris-setosa **42** 43 4.4 3.2 1.3 0.2 Iris-setosa 46 4.8 3.0 1.4 0.3 Iris-setosa **47** 48 4.6 3.2 1.4 0.2 Iris-setosa 4.9 2.4 3.3 **57** 58 1.0 Iris-versicolor **106** 107 4.9 2.5 4.5 1.7 Iris-virginica dataset.query('PetalLengthCm < @length') # fetch rows with lentgh lessthan 1.2</pre> Out[91]: $Id \quad SepalLengthCm \quad SepalWidthCm \quad PetalLengthCm \quad PetalWidthCm \\$ **13** 14 4.3 1.1 3.0 0.1 Iris-setosa **22** 23 4.6 1.0 3.6 0.2 Iris-setosa In [92]: #2.Get dummie variables dataset[['Iris-setosa', 'Iris-versicolor', 'Iris-virginica']] = pd.get_dummies(dataset['Species']) dataset.head() Out[92]: Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species Iris-setosa Iris-versicolor Iris-virginica 0 1 5.1 3.5 0.2 Iris-setosa **1** 2 4.9 3.0 1.4 0.2 Iris-setosa 0 **2** 3 4.7 0.2 Iris-setosa 3.2 1.3 **3** 4 4.6 3.1 1.5 1 0 0.2 Iris-setosa 3.6 **4** 5 5.0 0.2 Iris-setosa In [93]: #3.group by dataset.groupby('Species').count() Out[93]: ld SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Iris-setosa Iris-versicolor Iris-virginica **Species** Iris-setosa 50 50 50 50 50 50 50 50 50 50 Iris-versicolor 50 50 50 50 50 Iris-virginica 50 50 50 50 In [94]: #4. order the datset dataset.sort_values(by = 'Species', inplace = True) dataset.head() Out[94]: Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species Iris-setosa Iris-versicolor Iris-virginica 0 1 5.1 1.4 0 0 3.5 0.2 Iris-setosa **27** 28 5.2 3.5 1.5 0.2 Iris-setosa **28** 29 5.2 3.4 1.4 0.2 Iris-setosa **29** 30 4.7 3.2 1.6 0.2 Iris-setosa 0 0 **30** 31 4.8 3.1 1.6 0.2 Iris-setosa #5.rename the column dataset.rename(columns = {"Id":"id", "Species":"species"}, inplace = True) dataset.head() Out[95]: id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm species Iris-setosa Iris-versicolor Iris-virginica 0 1 5.1 1.4 3.5 0.2 Iris-setosa **27** 28 5.2 3.5 1.5 0.2 Iris-setosa 0 0 5.2 **28** 29 3.4 1.4 0.2 Iris-setosa **29** 30 4.7 3.2 1.6 0 0 0.2 Iris-setosa 1 4.8 3.1 1.6 **30** 31 0.2 Iris-setosa In [96]: dataset.rename(str.lower,axis = columns', inplace = True) dataset.head() Out[96]: id sepallengthcm sepalwidthcm petallengthcm petalwidthcm species iris-setosa iris-versicolor iris-virginica 0 1 0.2 Iris-setosa **27** 28 3.5 1.5 0.2 Iris-setosa 5.2 **28** 29 0.2 Iris-setosa 4.7 **29** 30 3.2 1.6 0.2 Iris-setosa 0 **30** 31 0.2 Iris-setosa In [97]: #6. merge datsets df1 = dataset[['id','sepallengthcm','sepalwidthcm']] print('df1') print(df1.columns) df2 = dataset[['id','petallengthcm','petalwidthcm','species']] print('df2') print(df2.columns) new_data = pd.merge(df1,df2,on = 'id') print(new_data.head()) print('') print(new_data.shape) Index(['id', 'sepallengthcm', 'sepalwidthcm'], dtype='object') Index(['id', 'petallengthcm', 'petalwidthcm', 'species'], dtype='object') id sepallengthcm sepalwidthcm petallengthcm petalwidthcm species

 0
 1
 5.1
 3.5
 1.4
 0.2
 Iris-setosa

 1
 28
 5.2
 3.5
 1.5
 0.2
 Iris-setosa

 2
 29
 5.2
 3.4
 1.4
 0.2
 Iris-setosa

 3
 30
 4.7
 3.2
 1.6
 0.2
 Iris-setosa

 4
 31
 4.8
 3.1
 1.6
 0.2
 Iris-setosa

 (150, 6)In [98]: #7. concat datsets df1 = dataset.iloc[0:75,:] print('df1') print(df1.head(3)) print('----') df2 = dataset.iloc[75:150,:] print('df2') print(df2.head(3)) df1 id sepallengthcm sepalwidthcm petallengthcm petalwidthcm species \

 0
 1
 5.1
 3.5
 1.4
 0.2
 Iris-setosa

 27
 28
 5.2
 3.5
 1.5
 0.2
 Iris-setosa

 28
 29
 5.2
 3.4
 1.4
 0.2
 Iris-setosa

 iris-setosa iris-versicolor iris-virginica 1 0 0 0 1 0 ______ id sepallengthcm sepalwidthcm petallengthcm petalwidthcm \ 74 75 6.4 2.9 4.3 1.3 73 74 6.1 2.8 4.7 1.2 51 52 6.4 3.2 1.5 4.5 species iris-setosa iris-versicolor iris-virginica 74 Iris-versicolor 0 1 0 73 Iris-versicolor In [99]: new data = pd.concat([df1,df2]) print(new_data.shape) print('') new_data.head(3) (150, 9)Out[99]: id sepallengthcm sepalwidthcm petallengthcm petalwidthcm species iris-setosa iris-versicolor iris-virginica 0 1 5.1 3.5 1.4 0.2 Iris-setosa **27** 28 5.2 3.5 1.5 0.2 Iris-setosa 0 0 **28** 29 5.2 3.4 0.2 Iris-setosa **Plotting with pandas** In [100]: #1. Scatter plot dataset.plot.scatter('sepallengthcm','petallengthcm'); 7.5 5.0 6.0 6.5 7.0 4.5 sepallengthcm In [101]: #2. Histogram plot dataset['sepalwidthcm'].plot.hist(); 35 30 25 20 15 10 3.0 3.5 2.0 2.5 In [102]: #3. Bar plot dataset['species'].value_counts().plot.bar(); 50 40 30 20 10 In [103]: #4. Horizontal bar plot dataset['species'].value_counts().plot.barh(); Iris-virginica Iris-versicolor Iris-setosa In [104]: #5. box plots dataset[['sepallengthcm','petallengthcm','sepalwidthcm','petalwidthcm']].plot.box();

sepallengthcm petallengthcm sepalwidthcm petalwidthcm

33.33

dataset['species'].value_counts().plot.pie(autopct='%.2f',figsize=(6,6));

33.33

33.33

find all kinds of pandas visualiations in the link

Iris-setosa

Iris-virginica

In [105]: #6 pie plots

lris-verឡ៊ីcolor

we will explore useful pandas commands on iris dataset

In [68]: import pandas as pd

Basic Commands

In [70]: #2.Displaying top 5 rows in dataset

4.9

4.7

4.6

5.0

5.1

4.9

4.6

5.0

5.4

4.6

In [71]: dataset.head(8) #Displays top 8 rows in the dataset

dataset.head()

0 1

1 2

2 3

3 4

4 5

0 1

1 2

3 4

5 6

Out[70]:

Out[71]:

In [69]: # 1.Reading or Loading a datset into a variable
dataset = pd.read csv('Iris.csv');

Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm

3.5

3.0

3.2

3.1

3.6

Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm

3.5

3.0

3.1

3.6

3.9

1.5

1.5

1.7

Species

0.2 Iris-setosa

0.2 Iris-setosa

0.2 Iris-setosa0.2 Iris-setosa

0.2 Iris-setosa

Species

0.2 Iris-setosa

0.2 Iris-setosa0.2 Iris-setosa

0.2 Iris-setosa

0.2 Iris-setosa

0.4 Iris-setosa

0.3 Iris-setosa