Part 1: Cleaning,wrangling data

1) By using shape function we can get the no.of rows and columns in the data.
By ".head()" we get the first five rows in the data.

2)By using "dtypes" fuction we can get the type of each variable where we can observe int,float and object datatypes in the above data.
By "info" function we get information of:
1.The number of non-null values in each column.
2.The data type of each column.
3.The memory usage of the DataFrame.

.

3)By using dropna function we can drop the null values
By axis=1 specifies to drop only columns

By how='all' specifies drop only columns which has all null values assigning new data to df1

4)By using dropna function we can drop the null values

axis=1 specifies to drop only columns

how='any' specifies drop only columns which has alteast one null value

assigning new data to df2

By ".head()" we get the first five rows in the data.

5)By using df2.rename(old_name=new_name) function we can rename the columns

By df3.head() function we get the top five rows of the data

6)By usig df3['date'].str.split(' ',1,expand=True) function we can split date column into two new columns as dates and time

By.split(' ') function we split data by space

By "1" we split data at first space

By "expand = True" tells that the split result should be expanded into separate columns

7)By df3['atd_loc_id'].str.split('-',1, expand=True) function we can split atd_loc_id column into two new columns as loc and code

By".split('-')" function we split data by -

By "1" we split data at first - symbol

By "expand = True" tells that the split result should be expanded into separate columns

By ".head()" we get the first five rows in the data.

8)By using df3["loc_type"].unique() we get unique values in the loc_type data

By ".head()" we get the first five rows in the data.

9) By df3['loc_type'].replace({'ROADWAY': '0', 'BUILDING': '1'}) function we can replace the values ROADWAY and BUILDING with o and 1 in loc_type

By ".head()" we get the first five rows in the data.

10) By "df3['loc_name'].str.split('/',1, expand=True)" function we can split loc_name column into two new columns as corner1 and corner2

By".split('-') "function we split data by /

By "1" we split data at first / symbol

By "expand = True" tells that the split result should be expanded into separate columns

By ".head()" we get the first five rows in the data.

Exploratory Data Analysis

Name: Nuthan Reddy Vaddi Reddy

UTA ID:1002175133

Step **1:**

- In The step 1 we are finding the complete information about the data like the number of null values in the data,datatype of each column,mean,quartile etc.

- By using "info" function we get the information about:
  - 1.The number of non-null values in each column.
  - 2.The data type of each column.
  - 3.The memory usage of the DataFrame.
  - In the above we have six float datatype and 1 integer datatype.

- By "describe" we get the small statical summary of data:
  - 1.mean and median of the data.
  - 2.standard deviation of the data.
  - 3.minimum and maximum vaule of the data.
  - 4.count of the values.

Step 2:

- In the second step we are finding the missing values(null values) and duplicates, Removing the null values and duplicates form the data and making the data more efficient this is called as data cleaning.
- Where are finding the number values count in the each class it is important for the predictions.

- By "isna().sum()" we get the sum of null values in each columns
  - 1"isna()" gives the null values
  - 2."sum()" the sum function will add the sum of all the null values in each column

- By using "drop_duplicates" we can identify duplicate values and drop the duplicate values from the data.

- By using "unique" function we get all unique values in the species column that only in the data, In the above code we are finding unique values in the species column.

- By "value_counts" we get the value count of each class in the species column.
  - In the above data we can see that value count of each class is not same and it varies excessively and it is not good for prediction.
  - For prediction the same value count in each class is very important
  - The accuracy of prediction is high when we have same value count in each class

Step3:

- In the third step we are finding the shape of the data in the plotting (visualization) by that we can see the data in the plotting, By ploting we can easily give judication on the data by seeing the plots.

Count plot:

- The count plot is used to see the values of class in the column and for the ploting we need to import seaborn.
- The x-axis is classes of species and y-axis is the number of values in the class.
- In the above count plot we can see that the count of each class:
- The perch has highest number of values with more than 50 values and whitefish has the lowest number of values below 10 values

STRIPPLOT:

- The stripplot is used specify that Weight is on the y-axis and 'Species' is on the x-axis.

- In The stripplot we can see observe that the pike has the highest Weight with above 1500 and the smelt has the lowest weight lies between 0 to 250.
- pernch has more weight than whitefish,roach and Parkki.

- The stripplot is used specify that Length3 is on the y-axis and 'Species' is on the x-axis.
- In The above stripplot we can see observe that the pike has the highest length3 with 7 and the smelt has the lowest length3 lies between 10 to 20.
- All Bream Length 3 lies between 30 and 50.
- whitefish has more length3 than roach and Parkki.

- The stripplot is used specify that Length1 is on the y-axis and 'Species' is on the x-axis.
- In The above stripplot we can see observe that the pike has the highest length1 with 60 and the smelt has the lowest length1 lies between 5 to 20.
- The pernch is equally distrubuted length form below 0 to 50.
- The Bream Length1 lies between 20 and 40.
- whitefish has more length1 than roach and Parkki.

- The stripplot is used specify that Length2 is on the y-axis and 'Species' is on the x-axis.
- In The above stripplot we can see observe that the pike has the highest length2 with above 60 and the smelt has the lowest length2 lies between 10 to 20.
- the pernch is equally distrubuted length2 form below 10 to 50.
- All Bream Length 2 lies between 30 and 45.
- whitefish has more length2 than roach and Parkki.

- The stripplot is used specify that Width is on the y-axis and 'Species' is on the x-axis.
  In The above stripplot we can see observe that the perch has the highest Width with above 8 and the smelt has the lowest width lies between 1 to 2.
  - The pernch is equally distrubuted width form 1 to above 8.
  - All Bream width lies between 4 and 7.
  - whitefish has more width than roach and Parkki.
  - The pike has the second heightest width.

- The stripplot is used specify that Height is on the y-axis and 'Species' is on the x-axis.
  - In The above stripplot we can see observe that the Bream has the highest height with above 17.5 and the smelt has the lowest below 5 .
  - ->the pernch is equally distrubuted he form bottom to 12.5.
  - ->whitefish has more height than roach,pike and Parkki.
  - ->The Perch has the second heightest height.

- The stripplot is used specify that Length3 is on the y-axis and 'Species' is on the x-axis.
  - In The above stripplot we can see observe that the pike has the highest length3 with 80 to 70 and the smelt has the lowest length3 lies between 10 to 20.
  - ->whitefish has more length3 than roach,bream and Parkki.

- The stripplot is used specify that Length3 is on the y-axis and 'Species' is on the x-axis.
  - In The above stripplot we can see observe that the pike has the highest length3 with 80 to 70 and the smelt has the lowest length3 lies between 10 to 20.
  - ->whitefish has more length3 than roach,bream and Parkki.
  - ->"size =8" adjusts the width of the violin plots.

Violinplot:

- The violinplot is used specify that weight is on the y-axis and 'Species' is on the x-axis.
  - In The above violinplot we can see observe that the pike has the highest weight with above 2000 and the smelt has the lowest weight below 500 .
  - ->whitefish has more highest than roach,perch and Parkki.
  - ->"size =8" adjusts the width of the violin plots.

- The violinplot is used specify that Height is on the y-axis and 'Species' is on the x-axis.
  - In The above violinplot we can see observe that the Bream has the highest height with above 20 and the smelt has the lowest heighest below 5 .
  - ->whitefish has more highest than roach,pike and Parkki.
  - ->The Perch has the second heightest height above 15.
  - ->"size =8" adjusts the width of the violin plots.

- The violinplot is used specify that Width is on the y-axis and 'Species' is on the x-axis.
  - In The above violinplot we can see observe that the perch has the highest Width with above 8 and the smelt has the lowest width lies between 1 to 2.
  - ->whitefish has more width than roach and Parkki.
  - ->The pike has the second heightest width.
  - ->"size =8" adjusts the width of the violin plots.

- The violinplot is used specify that Length1 is on the y-axis and 'Species' is on the x-axis.
  - In The above violinplot we can see observe that the pike has the highest length1 with above 60 and the smelt has the lowest length1 lies between 5 to 20.
  - ->whitefish has more length1 than roach,bream and Parkki.
  - ->"size =8" adjusts the width of the violin plots.

- The violinplot is used specify that Length2 is on the y-axis and 'Species' is on the x-axis.
  - In The above violinplot we can see observe that the pike has the highest length2 with above 70 and the smelt has the lowest length2 lies between 10 to 20.
  - ->whitefish has more length2 than roach,bream and Parkki.
  - ->"size =8" adjusts the width of the violin plots.

Histograms:

- The highest frequency of the Fish_width is between 30 and 35 which is between 3 and 4
  - ->bins=10 specifies the number of bins (intervals) for the histogram
  - ->color='Brown' sets the color of the bars
  - ->edgecolor='black' sets the color of the edges of the bars.

- The highest frequency of the Fish_height is between 40 and 35 which is between 5 and 7.5
  - ->bins=10 specifies the number of bins (intervals) for the histogram
  - ->color='Brown' sets the color of the bars
  - ->edgecolor='black' sets the color of the edges of the bars.

- The highest frequency of the Fish_Length1 is between 35 and 40 which is between 15 and 25
  - ->bins=10 specifies the number of bins (intervals) for the histogram
  - ->color='Brown' sets the color of the bars
  - ->edgecolor='black' sets the color of the edges of the bars.

- The highest frequency of the Fish_Length2 is between 35 and 40 which is between 15 and 25
  - ->bins=10 specifies the number of bins (intervals) for the histogram
  - ->color='Brown' sets the color of the bars
  - ->edgecolor='black' sets the color of the edges of the bars

- The highest frequency of the Fish_Length3 is between 35 and 40 which is between 20 and 25
  - ->bins=10 specifies the number of bins (intervals) for the histogram
  - ->color='Brown' sets the color of the bars
  - ->edgecolor='black' sets the color of the edges of the bars
- The highest frequency of the Fish_Weight is between 60 and 50 which is between 0 and 250
  - ->bins=10 specifies the number of bins (intervals) for the histogram
  - ->color='Brown' sets the color of the bars
  - ->edgecolor='black' sets the color of the edges of the bars.

Step4 :

- In the fourth step we can correlations between the colums :
- If correlation is positive one then the correlations is highly positive
- If correlation is negative one then the correlations is highly negative.
- If correlation is zero then there is no correlation.
- we can correlations between the colums :
- If correlation is positive one then the correlations is highly positive
- If correlation is negative one then the correlations is highly negative.
- If correlation is zero then there is no correlation.
- In the above correlation table there is no negative correlation between the columns .
- there is length1 and length2 highly positive correlation.
- In the above table all columns has positive correlation between them.

Step5:

- Spot/deal with outliers in the data.

Box plot:

- species on x-axis and width on y-axis.
- Species Smelt has the smallest features and less distributed with some outliers.
- Species Whitefish has the highest features.
- Species parkki,pike,perch,roach and bream has the average features.

- species on x-axis and Length3 on y-axis.
- Species Smelt has the smallest features and less distributed with some outliers.
- Species pike has the highest features.
- Species Whitefish,pike,perch,roach and bream has the average features.

- species on x-axis and Length2 on y-axis.
- Species Smelt has the smallest features and less distributed with some outliers.
- Species pike has the highest features.
- Species Whitefish,pike,perch,roach and bream has the average features.

- species on x-axis and Length1 on y-axis.
- Species Smelt has the smallest features and less distributed with some outliers.
- Species pike has the highest features.
- Species Whitefish,pike,perch,roach and bream has the average features.

- species on x-axis and Height on y-axis.
- Species Smelt has the smallest features and less distributed with some outliers.
- Species pike has the highest features.
- Species Whitefish,pike,perch,roach and bream has the average features.
- species on x-axis and Weight on y-axis.
- Species Smelt has the smallest features and less distributed with some outliers.
- Species pike has the highest features.
- Species Whitefish,pike,perch,roach and bream has the average features.

For the outliers:

- We calculate the interquartile range as the difference between Q3 and Q1.
- We calculate the lower and upper bounds for outliers based on the IQR.
- We identify outliers by comparing each data point to the calculated bounds.