Assignment NO-7.

Page No.
Date

Dr.D.Y.PATIL EDUCATIONAL FEDERATION Varale Campus

Title- Text Analytics.

objective- Students should able to do the text analytics using TFIDE algorithm.

prerequisite - 1) Basic of python
2) concept of TFIDF and Text analysis.

Theory -

1. Basic Concepts of Text Analytics.

one of the most frequent types of day-to-day conversion is text communication, In our everyday routine, we that, message, tweet, share status, email create blogs, and offer opinions and criticism.

2. Text Analysis operations using natural Language Toolkit.

NLTK (natural Language Toolkit) is a leading platform for building python program to work with human Language data. It provides easy-to-use interfaces and lexical resources such as WordNet, along with a suite of text processing libraries for classification. ~~Reg~~

Tokenization

Tokenization is the first step in text analytics. The process of breaking down a text paragraph into smaller chunks such as words or sentences is called Tokenization.

Page No.
Date

Dr.D.Y.PATIL EDUCATIONAL FEDERATION Varale Campus

Sentence tokenization : split the paragraph into list of sentences using sent - tokenize() method.

- word tokenization - split a sentence into list of words - tokenize() method.

Stop words removal -

stopwords considered as noise in the text. Text may contain stop words such as is, am, are this, a, an, the, etc. In NLTK for removing stopwords, you need to create a list of stopwords, ~~you need to create~~. and filter out your list of token from these words.

Stemming and Lemmatization.

stemming is a normalization technique where list of tokenized words are converted into shortend root words to remove redundancy. stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form.

1) Lemmatization vs stemming

Stemming algorithm works by cutting the suffix from the word. In a brodder sense cuts either the begining of end of the word.

On the country, Lemmatization is a more powerfed operation, and it takes into considering morphological analysis of the words. It returns the lemma which is the base from of all its inflectional forms.

In-deapth linguistic knowledge is required to create dictionaries and look for the proper form of the word.

Pos Tagging:

Pos (parts of speech) tell us about grammatical information of words of the sentence by assigning specific token (Determiner, noun, adjective, adverb, verb, Personal pronoun etc.) as tag (DT, NN, JJ, RB, PRP etc) to each words.

Text Analysis Model using TF-IDF.

Term Frequency - inverse document frequency (TF-IDF), is a numerical ~~stastic~~ statistic that is intended to reflect how important a word is to a document in a collection or corpus.

- Term Frequency (TF)

It is measure of the frequency of word (w) in a document (d). TF is defined as the ratio of a word's occurance in a document to the total number of words in a document. The denominator term in the formula is to normalize since all the corpus documents are of different lengths.

$$TF(w,d) = \frac{\text{occurances of } w \text{ document } d}{\text{total no. of words in document } d}$$

ex.

Documents	Text	Total no. of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun.	8

The initial steps is to make a vocabulary of unique words and calculate TF for each document.

Inverse Document Frequency (IDF)

It is the measure of the a word Term frequency (TF) does not consider the importance of words such as of, 'and', etc. can be most frequently present but are of little significance.

$$IDF(w,D) = \ln\left(\frac{\text{Total no. of documents (N) in corpus D}}{\text{no. of documents containing w}}\right)$$

In our example, since we have two documents in the corpus, N = 2.

Words	TF (For A)	TF (For B)	IDF
Jupiter	1/5	0	$\ln(2/1) = 0.69$
Is	1/5	1/8	$\ln(2/2) = 0$
the	1/5	2/8	$\ln(2/2) = 0$
largest	1/5	0	$\ln(2/1) = 0.69$
From	0	1/8	$\ln(2/1) = 0.69$
Sun	0	1/8	$\ln(2/1) = 0.69$

Term Frequency - Inverse Document Frequency (TFIDF)

It is the product of TF and IDF.

TFIDF gives more weightage to the word that is rare in the corpus (all the documents).

TFIDF provides more importance to the word that is more frequent in the document.

$$TFIDF(w, d, D) = TF(w, d) * IDF(w, D)$$

Words	TF (For A)	TF (For B)	IDF	TFIDF(A)	TFIDF(B)
Jupiter	1/5	0	$\ln(2/1) = 0.69$	0.138	0
Is	1/5	1/8	$\ln(2/2) = 0$	0	0
The	1/5	2/8	$\ln(2/1) = 0.69$	0.138	0
planet	1/5	0	$(\ln 2/1) = 0$	0.138	0
From	0	1/8	$(\ln 2/2) = 0$	0.138	0
sun	0	1/8	$(\ln 2/1) = 0.69$	0	0.086

After applying TFIDE, text in A and B documents can be represented as a TFIDF vector of dimension equal to the vocabulary words. The value corresponding to each word represents.

the importance of that word in particular document.

Disadvantage of TFIDF

It is unable to capture the semantics. For example. Funny and humorous as synonyms, but TFIDF does not capture that. TFIDF can be computationally expensive if the vocabulary is vast.

Bag of words (BOW)

Machine learning algorithms cannot works with raw Text directly. Rather, the text must be converted into vectors of Numbers. In natural language processing, a common technique for extracting features from text is to place all of the words that occur in the text in a bucket.

Algorithm for Tokenization, pos Tagging, Stop words removal, Stemming and Lemmatization.

Conclusion :-

In this way we have done text data analysis using TF IDF algorithm.

Assignment No - 8.

Page No.
Date

Dr. D. Y. PATIL EDUCATIONAL FEDERATION Varale Campus

Title - Data visualization.

1. use the inbuilt dataset 'titanic' The datasets contains 891 rows and contains information about.

objective - Students should able to do the Data visualization for Titanic dataset using seaborn library.

Prequisite - 1) Basics of python
2) Concepts of seaborn library.

Contents for Theory.

1) Seaborn library Basics.
2) know your data.
3) Finding patterns of Data.
4) cheaking how the price of ticket for each passenger is distributed by plotting histogram.

Theory - Data Visualization plays a very important role in data mining. To accelerate this process we need to have a well document - ation of all plots.

1) Seaborn library Basics.

Seaborn is a python data visualization library based on matplotlib It provides

high level interprice for drawing attractive and informative statistical graphics.

2) Know your data

The dataset that we are going to are to draw our plots will be titanic dataset which is downloaded by default with the seaborn library.

let's see what the titanic dataset looks like. Execate the following scripts.

```
import pandas as pd.
import numpy as np.
import matplotlib.pyplot as plt.
import seaborn as sns.

dataset = sns.load_dataset("titanic")
dataset.head().
```

3. Finding patterns of Data.

patterns of data can be find out with the help of different types of plots

Types of plots are :.

A. Distribution plots.

a. Dist-plot
b. Joint plot
c. Rug plot

Page No.
Date

Dr.D.Y.PATIL EDUCATIONAL FEDERATION Varale Campus

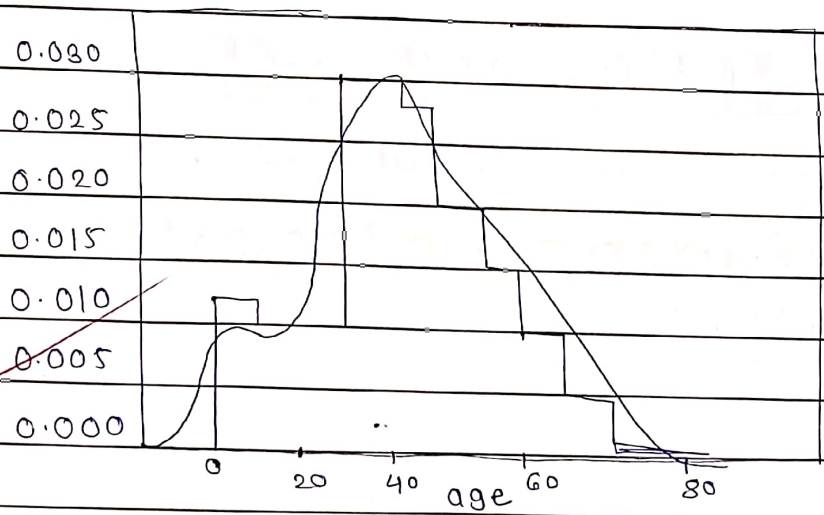
A. Distribution plots :.

These plots help us to visualize the distribution of data. we can use these plots to understand the mean, median, range, variance, deviation, etc of the data.

a. Displot.

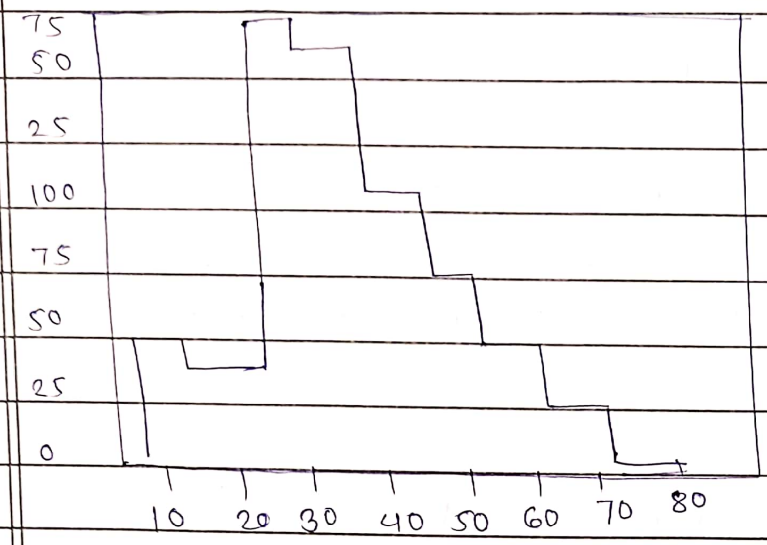
- Dist plot gives us the histogram of the selected continuous variable.
- It is an example of a univariate analysis.
- we can change the number of bins, i.e. no of vertical bars in a histogram.

```
import seaborn as sns.
sns.displot(x = dataset['age'], bins = 10).
```



The line that you see represents the kernal density estimation, you can remove this line by passing False as the parameter for the kde Attribute as shown below.

```
sns.distplot(dataset['age'], bins=10, kde=False).
```

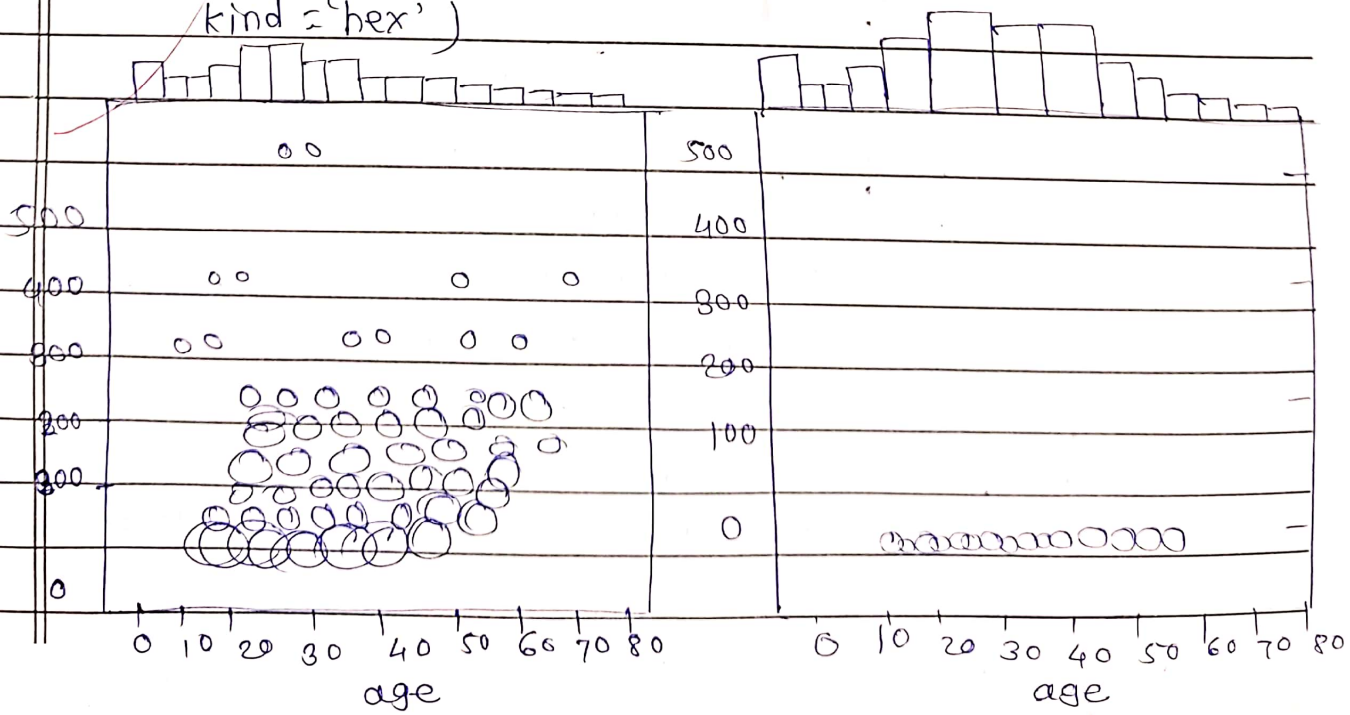


i.b.

Joint plot

- It is the combination of the distplot of Two variables.
- It is an example of bivariate analysis.

```
import seaborn as sns.
    # For plot 1
sns.jointplot (x = dataset ['age'], y = dataset ['Fare']
kind = 'scatter')
    # For plot 2.
Sns.jointplot (x = dataset ['age'], = y = dataset ['Fare']
    kind = 'hex')
```

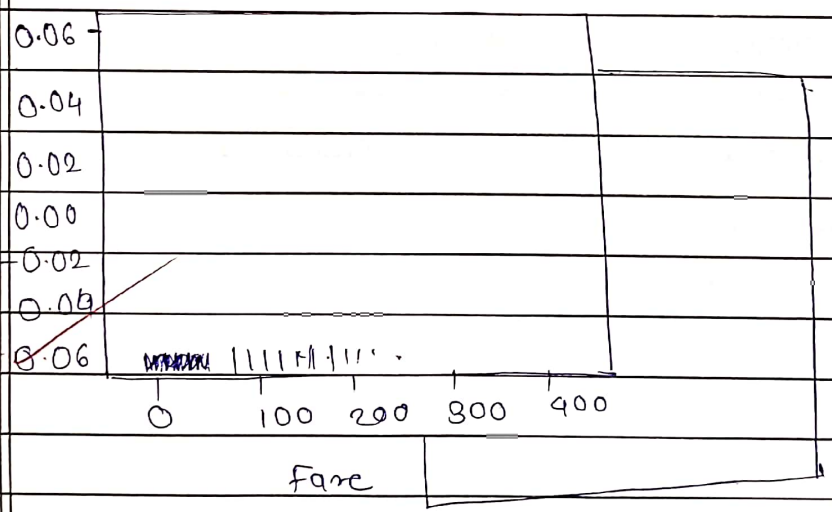


Page No.
Date

- From the output, you can see that a joint plot has three parts. A distribution plot at the Top for the column on the x-axis, a distribution plot on the right for the column on the y-axis & a Scatter plot in betn that shows the mutual distribution of data.

c. The Rug plot

b. The rugplot() is used to draw small bars along the x-axis for each points in the dataset. To plot a rug plot, you need to pass the ~~most of the~~ name of the column. Let's plot a rug plot for fare.



From the output, you can see that most of the instances for the fares have values betn 0 and 100.

conclusion-

Seaborn is an advanced data visualization library built on top of matplotlib library. In this assignment, we looked at how we can drew distributed & categorical plots using the Seaborn library.

TS	PR	VC	VA	RH	Total marks (10)	Sign
(2)	(2)	(2)	(2)	(2)		

Page No.
Date

Dr.D.Y.PATIL EDUCATIONAL FEDERATION Varale Campus

Assignment No-9.

Title - Data visualization.

1) Use the inbuit dataset 'titanic' The dataset used in the problem plot box plot for distribution of age with respect to each gender
2) white observations on the inference from the above statistics.

Objectives- students should able to do the data visualization for titanic dataset using seaborn library.

~~Prere~~ prerequisite - 1) Basic of python
2) Concept of seaborn library.

Contents of Theory
1) know you data.
2) Finding pattern of Data
3) cheaking how price of Ticket.

Theory -

Categorical plots;

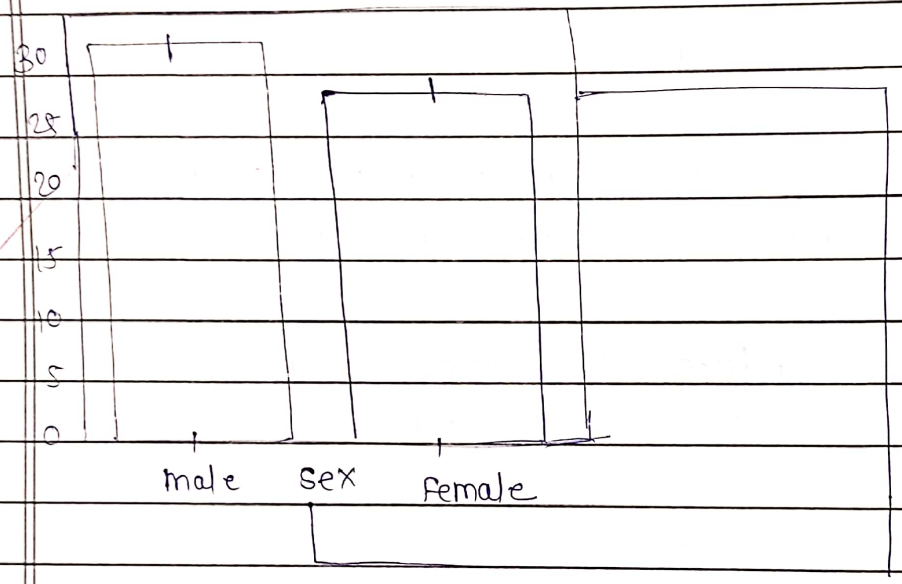
Categorical plots, as the name suggests, are normally used to plot categorial Data. The categorial plots the values in the categorial columns.

b. The Bar plot.

The barplot () is used to display the mean value in a categorial column, against a numeric column against a while the Third parameter is the dataset.

For instance, if you want to know the mean value the age of the male and female passenger. you can use the bar plot as follows.

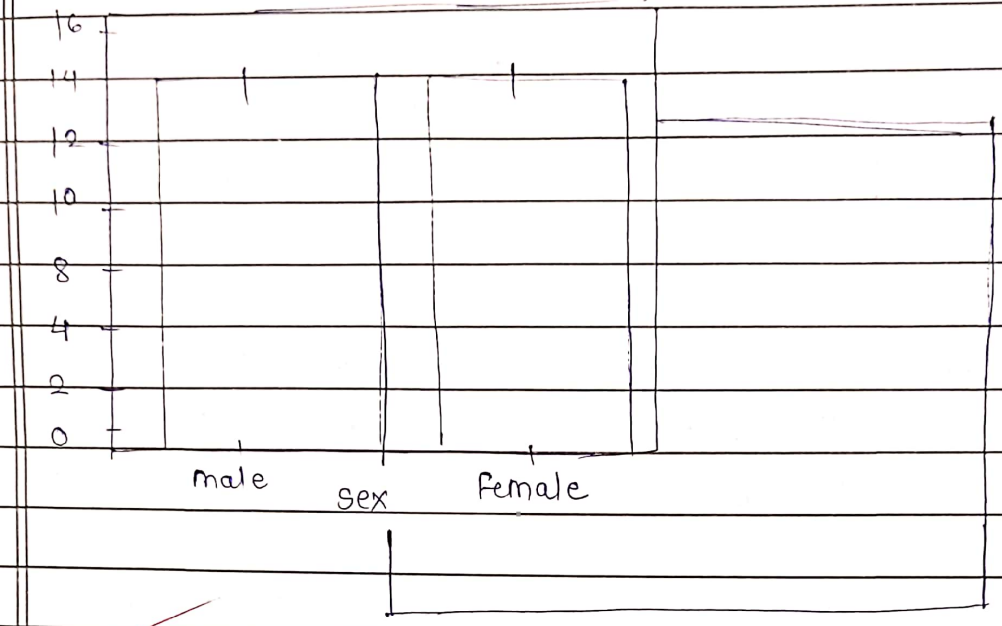
sns.barplot(x='sex', y='age', data = dataset)



From the output, you can clearly see that the average age of male passengers is just less than 40 while the average age of female passengers is around 33.

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns.
sns.barplot(x='sex', y='age', data - dataset, estimator
    =np.std)
```

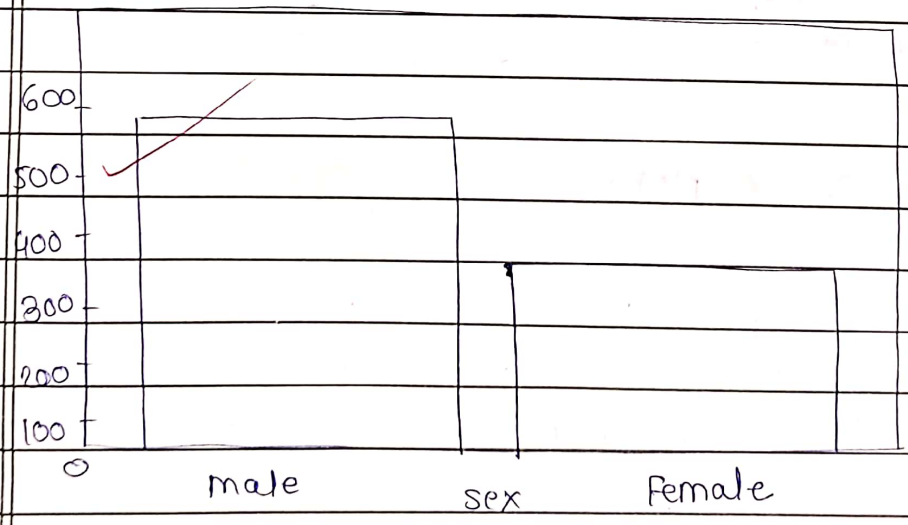
c.



c. The count plot.

The count plot is similar to the bar plot, however is displays the count of the categories in the specific column.

sns.countplot(x='sex', data=dataset)



d. The Box plot.

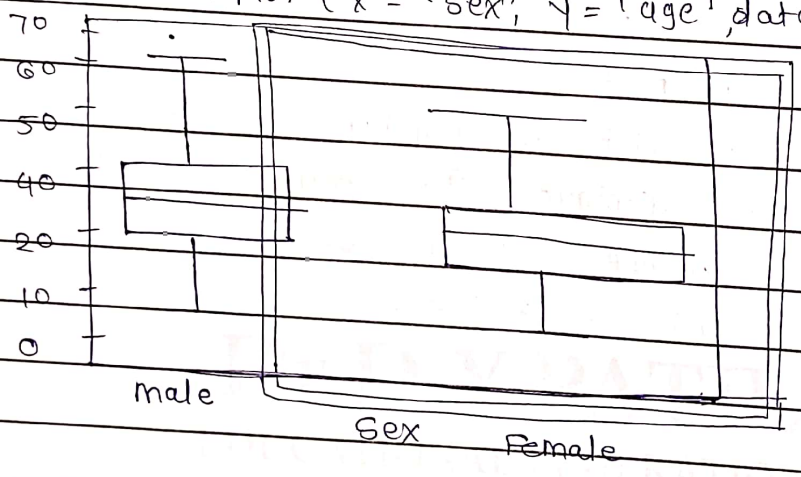
The Box plot is used to display the distribution of the categorial data in the

Form of quartiles. The centre of the box shows the median value.

The value from the lower whisker to the bottom of the box shows the first quartile.

Now let's The Box plot that displays, the distribution for the age with respect to each gender.

Sns. boxplot (x = 'sex', y = 'age', data = dataset)

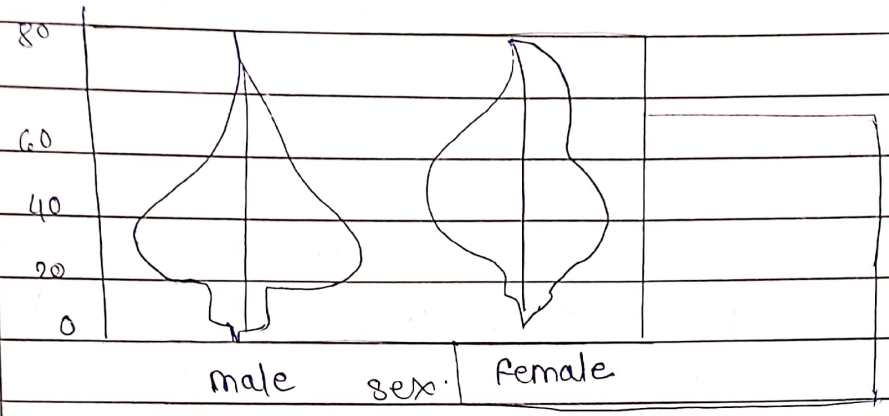


e. The violin plot.

The violin plot is similar to the box plot however, The violin plot allows us to display all the component that actually corresponds to the data point.

let's plot in violin plot.

Sns. violinplot (x = 'sex', y = 'age', data - dataset



You can see from the figure below the violin plots provide much more information about the data as compared to the box plot. Instead of plotting the quartile, the violin plot allows us to see all the components the actually correspond to the data.

b. The swarm plot.

The swarm plot is a combination of strip and violin plot.

In the swarm plot, the points are adjusted in such way that they dont overlaps.

c. A strip plot

A strip plot draws a scatter plot where one of the variables is categorical. We have seen scatter plots is the point plot and the pair plot sections where we had two numeric variables.

The strip plot is different in a way that one of the variables is categorical in this case.

sns. stripplot (x = 'sex', y = 'age', data = dataset, jitter = False)

conclusion -

Seaborn is an advanced data visualisation library built on top of Matplotlib library. In this assignment, we Looked at how we can draw distributional and categorical plots using the seaborn library. we have seen how to plot matrix plots in Seaborn. we also saw how to change plot styles and use grid functions to manipate subplots.

TS	PR	VC	VA	RN	Total marks (10)	Sign
(2)	(2)	(2)	(2)	(2)		
02	02	01	02	02	09	

Page No.
Date

Assignment No-10.

Title - Data visualization.

1. Download the Iris Flower dataset or any other dataset into a Dataframe (e.g. https://archeive.ics.uci.edu/ml/datasets/Iris)
Scan the dataset and give the inference as.
1. List down the features and their types.
2. Create a boxplot for each features in the dataset.

objective - students should able to do the data visualization for Titanic dataset using seaborn library.

Theory :-

Matrix plots are the type of plots that show data in the form of rows and columns. Heat maps are the prime examples of matrix plots.

a. Heat Maps.

Heat maps are normally used to plot correlation betn numeric column in the form of matrix. It is important to mention here that to draw matrix plots, you need to have meaningful information on rows as well as columns.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
important Seaborn as sns
dataset = sns.load_dataset('titanic')
```

dataset.head()

	Survied	pclass	Sex	age	sibsp	parch	embarked	class	who	adult-male	deck
0	0	3	male	22.0	1	0	S	Third	man	True	NAN
1	1	1	female	38.0	1	0	C	First	woman	False	C
2	1	3	female	26.0	0	0	S	Third	women	False	NAN
3	1	1	female	35.0	1	0	S	First	cooman	False	C
4	0	3	male	35.0	0	0	S	Third	man	True	NAN

Both all the numeric columns of the dataset Execute the following script.

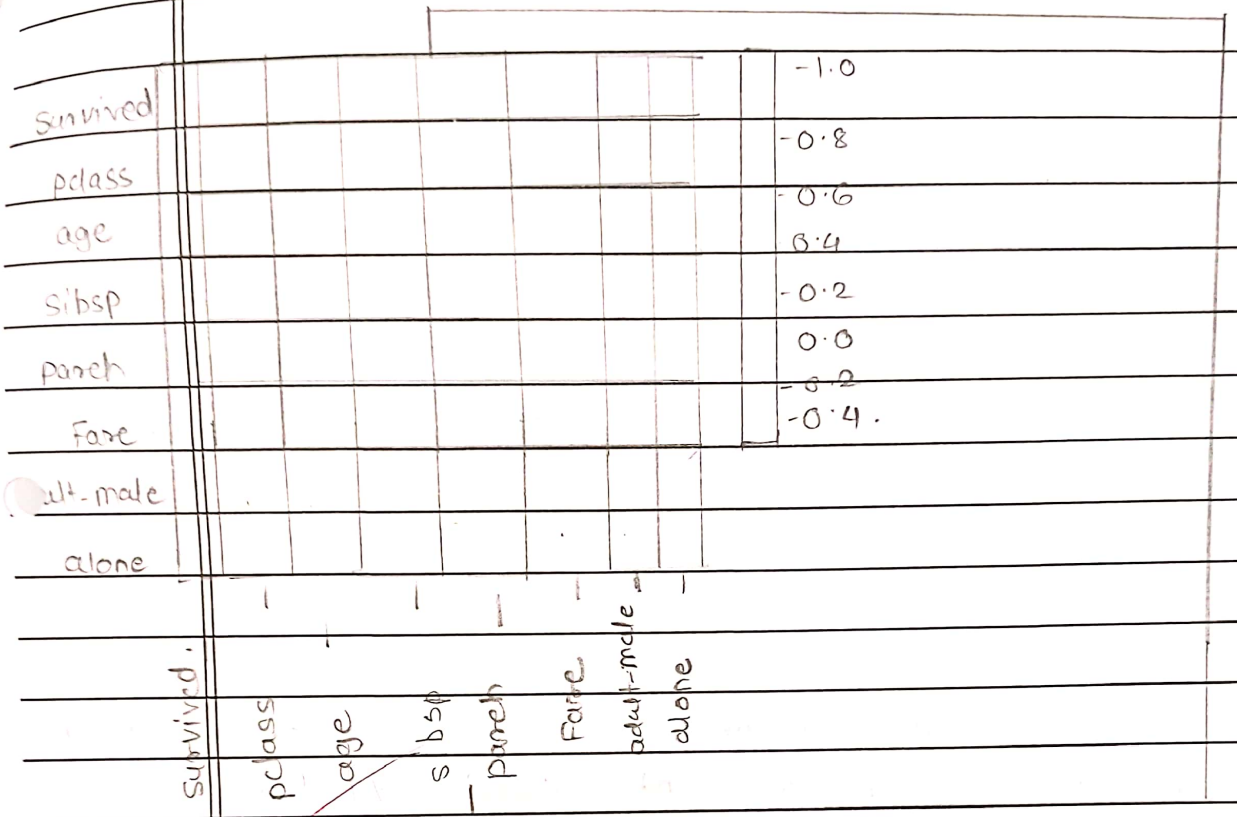
dataset.corr().

In the output, you will see that both the columns and the rows have meaningful header information. as shown below.

	Survived	Pclass	age	Sibsp	Parch	alone
Survived	1.000000	-0.338481	-0.077221	-0.035322	0.081629	-0.203367
Pclass	-0.338481	1.000000	-0.369226	0.083081	0.018443	0.135207
age	-0.035322	-0.369226	1.00000	-0.308247	-0.189119	0.198270
Sibsp	-0.081629	-0.083081	-0.308247	1.00000	-0.414838	0.584471
Parch	0.257307	0.018443	-0.189119	0.414838	1.00000	-0.583398
Fare	0.257307	-0.549500	0.96067	0.159657	0.216225	-0.271832
adult_male	0.557080	0.094035	0.280328	-0.349943	-0.349943	0.404744
alone	-0.203367	0.135207	0.19270	-0.583398	-0.271832	1.000000

Now create a heat map with these correlation values, you need to call the heatmap() function and pass it your dataframe. look at the following Script

corr = dataset.corr()



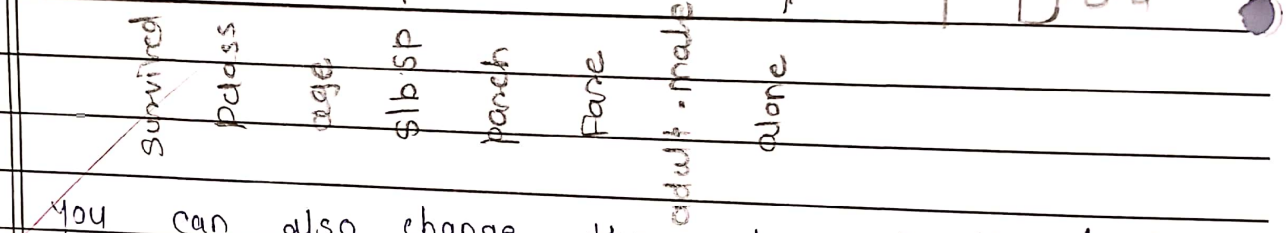
From the output, it can be seen that what heatmap essentially does is that is plots a box for every combination of rows and column value. The color the box depends upon the gradient.

The correlation values can also be plotted on the heatmap by passing True for the annot parameter.

execute the following script to see this in action.

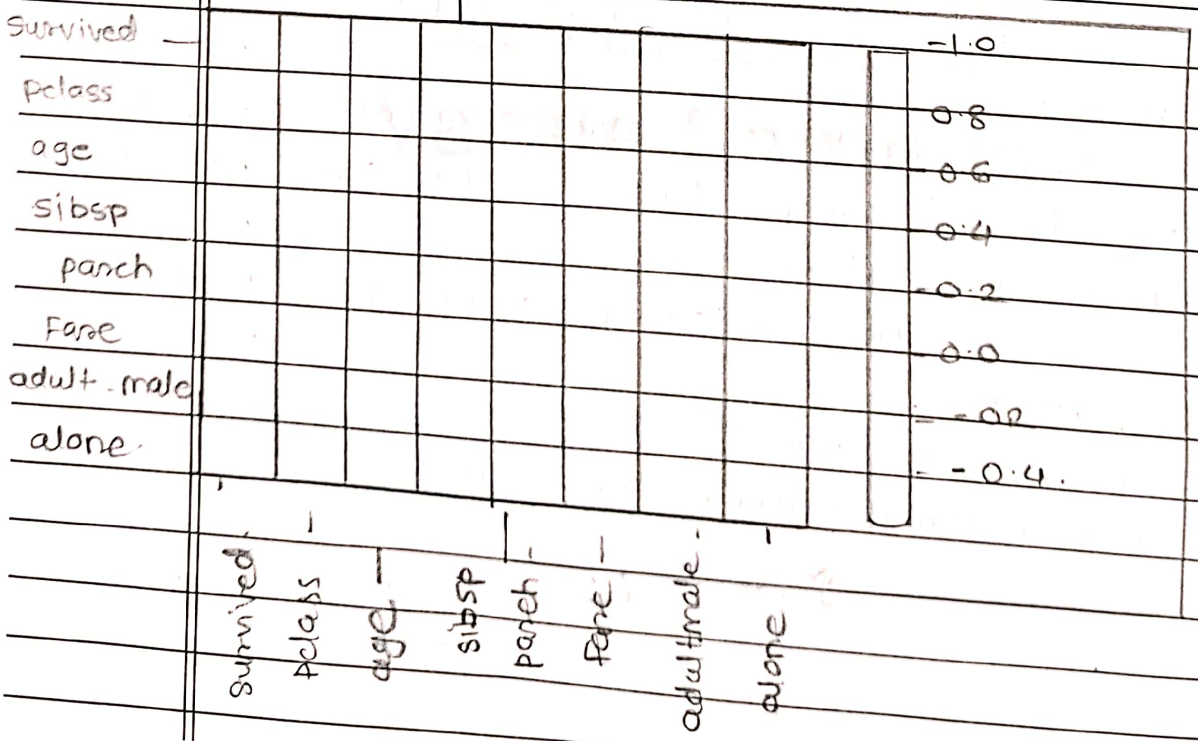
```
corr = dataset.corr()
sns.heatmap(corr, annot=True)
```

	Survived	Pclass	age	sibsp	parch	Fare	adult_male	alone
Survived	1	-0.34	-0.077	-0.035	0.082	0.26	-0.56	-0.2
Pclass	-0.34	1	-0.37	0.083	0.018	-0.55	0.094	0.14
age	-0.077	-0.37	1	-0.31	-0.19	0.096	0.28	0.2
sibsp	-0.035	0.083	-0.31	1	0.41	0.16	-0.25	-0.58
parch	0.082	0.018	-0.19	0.16	1	0.2	0.22	-0.38
Fare	0.26	-0.55	0.096	-0.25	0.22	1	-0.38	-0.27
adult_male	-0.56	0.094	0.28	-0.58	-0.38	0.27	1	1
alone	-0.2	0.14	0.2	-0.58	-0.38	0.27	1	1



You can also change the color of the heatmap by passing an argument for the cmap parameter for now, just look at the following script.

```
Corr = dataset.corr()
Sns = heatmap (corr)
```



b. cluster map.

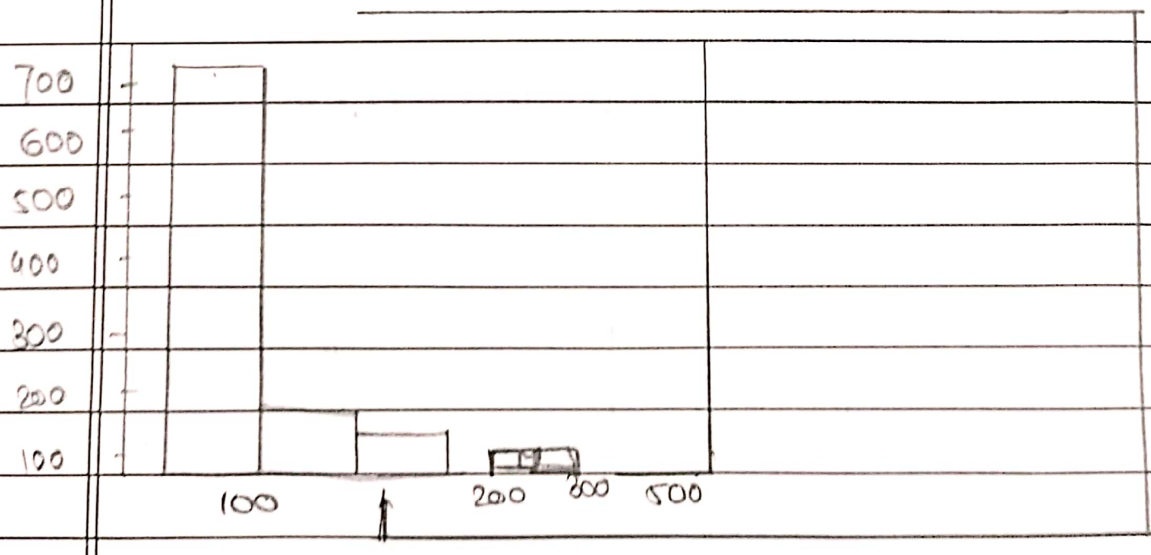
In addition to the heat map, another commonly used matrix plot is cluster map. The cluster map basically uses Hierarchical clustering to cluster the rows and columns of the matrix.

b. clustermap

In addition to the heat map, another commonly used matrix plot is the cluster map. The cluster map basically uses Hierarchical clustering to cluster the rows and columns of the matrix.

4. cheaking how the price of the ticket (column name : 'fare') for each passenger is distributed by plotting a histogram.

```
import seaborn as sns.
dataset = sns.load_dataset('titanic')
sns.histplot(dataset['fare'], kde=False, bins=10)
```



Page No.
Date

From the histogram, it is seen that for around 730 passengers the price of the ticket is 50. for 100 passengers the price of the ticket is 100 and so on.

conclusion -

Seaborn is an advanced data visualisation library built on top of matplotlib library. In this assignment, we looked at how we can draw distributional and categorial plots using the seaborn library.

TS	PR	UC	VA	RN	Total marks	sign
02	02	01	02	02	09	