We at The Data Monk hold the vision to make sure everyone in the IT industry has an equal stand to work in an open domain such as analytics. Analytics is one domain where there is no formal under-graduation degree and which is achievable to anyone and everyone in the World.

We are a team of 30+ mentors who have worked in various product-based companies in India and abroad, and we have come up with this idea to provide study materials directed to help you crack any analytics interview.

Every one of us has been interviewing for at least the last 6 to 8 years for different positions like Data Scientist, Data Analysts, Business Analysts, Product Analysts, Data Engineers, and other senior roles. We understand the gap between having good knowledge and converting an interview to a top product-based company.

Rest assured that if you follow our different mediums like our blog cum questions-answer portal www.TheDataMonk.com , our youtube channel - The Data Monk, and our e-books, then you will have a very strong candidature in whichever interview you participate in.

There are many blogs that provide free study materials or questions on different analytical tools and technologies, but we concentrate mostly on the questions which are asked in an interview. We have a set of 100+ books which are available both on Amazon and on The Data Monk e-shop page

We would recommend you to explore our website, youtube channel, and e-books to understand the type of questions covered in our articles. We went for the question-answer approach both on our website as well as our e-books just because we feel that the best way to go from beginner to advance level is by practicing a lot of questions on the topic.

We have launched a series of 50 e-books on our website on all the popular as well as niche topics. Our range of material ranges from SQL, Python, and Machine Learning algorithms to ANN, CNN, PCA, etc.

We are constantly working on our product and will keep on updating it. It is very necessary to go through all the questions present in this book.

Give a rating to the book on Amazon, do provide your feedback and if you want to help us grow then please subscribe to our Youtube channel.

Data Pre-processing

Question.1 What is Data??

Ans: - Data is a representation of facts stored in digital form. Data may be the clean or not. This representation of the facts may or may not be valid and accurate.

Question.2 What is the difference between data and information?

Ans: - The terms information and data are often used interchangeably. There is a difference between these two. Data can be any sequence of values, numbers, text, picture, files and so on. All of these things do not necessarily have to be informative to a consumer of that data. In most cases, data needs to be processed and put into context to make it informative for the consumer.

Question.3 Why is data important?

Ans: - Data is information stored in digital form. Information has always been important. When information was stored exclusively in analogue form, the information storage capacity of human mankind was extremely limited. By storing information as data in digital form, we have decoupled the growth of information from these limitations. Thanks to computers, hard drives, our smartphones and other technological innovations we are able to create, store and process data at staggering levels. Today, data growth is following an exponential path. Data is the oil of the 21st century. Most importantly, data is often regarded as the fuel of the 21st century. Much like oil and electricity have powered innovations and economies in the past, data will be the (not so natural) resource that fuels these in the present and future.

Question.4 What is Data Pre-processing?

Ans: - Data pre-processing is a process of preparing and transforming the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

Question.5 Why data pre-processing is so important?

Ans: - Mistakes, redundancies, missing values, and inconsistencies all effect the quality of the dataset, we need to fix all those issues for a more accurate outcome. That is why data pre processing is very important. Imagine you are training a Machine Learning algorithm to deal with the share market with a faulty dataset. Chances are that the system will develop biases and deviations that will produce a poor user experience. Thus, before using the data for the purpose you want, you need it to be clean as possible.

Question.6 What is raw data?

Ans: - Raw data refers to any data that hasn't gone thorough any processing, either manually or through automated computer software. Raw data may be gathered from various processes, manual data entry or IT resources. Raw data is also known as source data or primary data.

Question.7 What are the types of raw data?

Ans: -

- Missing data: Missing data often appears when there's a problem in the collection phase, such as a glitch that caused a system's downtime, mistakes in data entry, or issues with biometrics use, among others. This is common in pretty much in any data available
- Noisy data: This group encompasses outliers that you can find in the data set but that is just meaningless information. Here you can see noise made of human mistakes, rare exceptions, mislabels, and other issues during data gathering.
- Inconsistent data: Duplicates in different formats, mistakes in codes of names, or the absence of data constraints often lead to inconsistent data, that introduces deviations that you have to deal with before analysis.

Question.8 What is types of data?

Ans: - Data can be categorized into three types: -

1. Structured Data
2. Unstructured Data
3. Semi- Structured data

Structured Data: - Data with a high degree of organization, typically stored in a spreadsheet-like manner. Think of a spreadsheet or data in a tabular format. Data is structured in a spreadsheet-like manner Within that table, entries have the same format and a predefined length and follow the same order. It Is easily machine-readable and can therefore be analysed without major pre processing of the data. It is commonly said that around 20% of the world's data is structured.

E.g.

- Excel spreadsheets
- Comma-separated value file (.csv)
- Relational database tables

Semi-structured Data: - Data with some degree of organization. Think of a TXT file with text that has some structure (headers, paragraphs, etc.)

E.g.: -

- Hypertext Markup Language (HTML) files
- JavaScript Object Notation (JSON) files
- Extensible Markup Language (XML) files

Data is stored in files that have some degree of organization and structure. Tags or other markers separate elements and enforce hierarchies, but the size of elements can vary and their order is not important. Needs some pre-processing before it can be analysed by a computer. Has gained importance with the emergence of the World Wide Web

Unstructured Data: - Data with no predefined organizational form and no specific format. Essentially anything that is not structured or semi-structured data (which is a lot). Data that can take any form and thus be stored as any kind of file (formless) . Within that file, there is no structure of content. Typically needs major pre-processing before it can be analysed by a computer, but often easily consumable for humans (e.g., pictures, videos, plain texts. Most of the data that is created today is unstructured.

E.g.

- Images such as .jpeg or .png files
- Videos such as .mp4 or m4a files
- Sound files such as .mp3 or .wav files
- Plain text files
- Word files
- PDF files

Question.9 What are the 5 Major Steps in Data Pre-processing?

Ans: -

1. Import the libraries.

Importing the libraries which you'll need to work with mainly pandas ,NumPy,seaborn

2. Import the data-set

Importing the dataset which you want to do your work

3. Check out the missing values.

Analysis and imputing the missing values

4. Transforming the Categorical Values.

Converting the Categorical values into numerical ones

5. Splitting the data-set into Training and Test Set.

Cutting the dataset into train and test so we'll train the dataset on train and test its results on test.

6. Feature Scaling.

Scaling the features so every feature has the same level of importance for the model.

Question.10 What type of dataset does the machine learning algorithms work?

Ans: - A machine learning model completely works on data. Each dataset is different from another dataset. To use the dataset in our code, we usually put it into a CSV file. However, sometimes, we may also need to use an HTML, Json or xlsx file.
Question.11 What are Libraries and why do we need them?

Ans: - In order to perform data pre-processing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs.

Question.12 What are the popular libraries in Python we use for data pre-processing?

Ans: - Pandas: The last library is the Pandas library, which is one of the most famous Python libraries and used for importing and managing the datasets. It is an open-source data manipulation and analysis library.

import pandas as pd

Matplotlib: The second library is matplotlib, which is a Python 2D plotting library, and with this library, we need to import a sub-library pyplot.

Import matplotlib.pyplot as plt

Numpy: Numpy Python library is used for including any type of mathematical operation in the code. It is the fundamental package for scientific calculation in Python.

Import numpy as np

Here we use alias words for efficient coding.

Question.13 How to set the working directory in Spyder?

Ans: - To set a working directory in Spyder IDE:

- Save your Python file in the directory which contains dataset.
- Click on to the File explorer option in Spyder IDE, and select the required directory. • Execute the file.

Question.14 How to import the dataset which we want to work on?

Ans: - We use read_csv () function of pandas library, which is used to read a csv file and performs various operations on it. Using this function, we can read a csv file locally as well as through an URL.

Syntax: - df= pd.read_csv('Dataset.csv')

Here, df is a name of the variable to store our dataset, and inside the function, we have passed the name of our dataset.

Question.15: - How to pass the address of the dataset on the function?

Ans: - We can the pass the address of the dataset by using the read_csv() function

Syntax: - df = pd.read_csv(r”Address of the dataset”)

Question.16 What are the additional parameters of the read_csv()

function? Ans: - read_csv() has multiple paramters:

- Index_col : It is used to set which columns to be used as the index of the dataframe. The default value is None, and pandas will add a new column start from 0 to specify the index column. It can be set as a column name or column index, which will be used as the index column.
  Syntax :- df = pd.read_csv(“FIlename”, index_col = 0)

- Header : Header parameter is used to specify you have the names of columns in the first row in the file and if you don't you will have to specify header=None.
  Syntax: df = pd.read_csv(“Filename”,header= None)

- Sep : The sep parameter is used to specify by which element are the columns separated so that the pandas library treats the data that way
  Syntax:- df = pd.read_csv(“filename” ,sep = “,”)

- Skiprows : The skiprows parameter of the read_csv() function is used to the rows from csv at specified indices in the list
  Syntax:- df = pd.read_csv(“filename”, , skiprows=[0,2,5])

Question.17 How to have a brief look at the data ?

Ans:-

- Info() function :You can use the info() function to have a brief look at the data
  Syntax:- df.info()

- Head() function :The head( ) Function is used to look at the first 5 rows of the dataset.
  Syntax: df.head()

- Tail() function :The tail() function is used to look at the first 5 rows of the dataset.
  Syntax: df.tail()

Question.18 Which commands will give you a descriptive look at the data?

Ans:-

- Describe() function: describe() is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values.

- Shape() Function: The shape attribute returns a tuple of the number of rows and the number of columns in the DataFrame.
  Syntax:- df.shape

Question.19 What is missing data and why it is a big problem?

Ans:- Missing data presents various problems.

- The absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false.
- The lost data can cause bias in the estimation of parameters.
- It can reduce the representativeness of the samples.

Question.20 When should be consider deleting the missing data?

Ans:- If data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.

Question.21 How can we impute the categorical missing values?

Ans:-

1. Ignore the observation and let the algorithm handle it
2. Replace by most frequent value.

Question 22. How can we impute the continuous missing variable?

Ans:-

1. Ignore the observation and let the algorithm handle it.
2. Replace by the mean

Question 23. How to find out the missing numbers present in the data?

Ans:- The isnull() function is used to show where missing values are present in the data.

The output is in Boolean format.

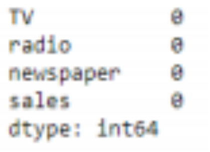
Question.24 How to find out the total number of missing values in the data?

Ans:- The isnull().sum() is used to return the total number of missing values in the

data.

Syntax:- df.isnull().sum()

OP:-



```
TV           0
radio        0
newspaper    0
sales        0
dtype: int64
```

Question.25 What are Outliers??

Ans: - Outliers are considered to be extreme values. They are defined as samples that are significantly different from the remaining data. Those are points that lie outside the overall pattern of the distribution. Statistical measures such as mean, variance, and correlation are very susceptible to outliers.

Example: - Suppose you are handling a dataset of cricketers of the Indian cricket team. In the variable of total runs, you will encounter that Sachin Tendulkar and Virat Kohli would be considered as an outlier since their total runs scored is much higher than other cricketers, this doesn't mean that the data entered in their field is wrong, it’s just that they are better than other cricketers that much.

Question.26 How should you handle Outliers in the dataset?

Ans: - If you encounter outliers in the data, you should look at the quantity of the outliers. If the quantity is high then you should consider leaving them as they are and let the algorithm handle it. If the quantity is less then consider them imputing them.

Question.27 How can Outliers occur in the dataset?

Ans: - Outliers can occur in the dataset due to one of the following reasons: -

1. Genuine extreme high and low values in the dataset
2. Introduced due to human or mechanical error
3. Introduced by replacing missing values.

Question.28 How to Detect Outliers??

Ans: - Outliers can be detected by the following ways: -

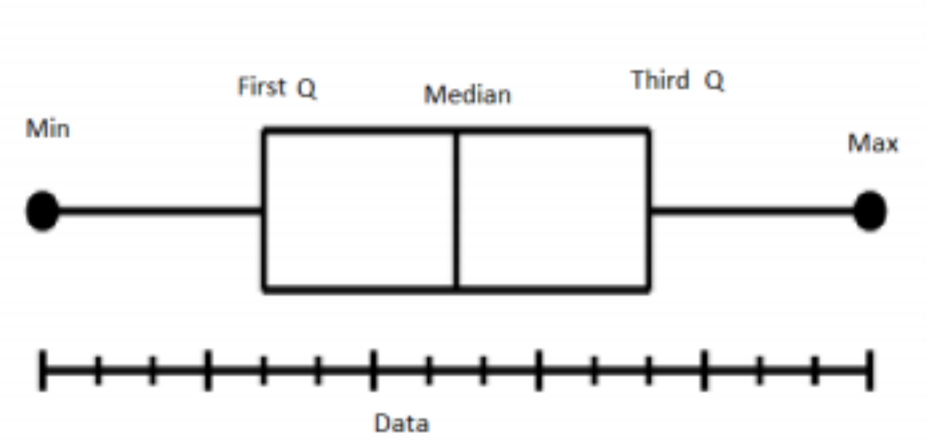
1. Extreme Value Analysis by Box Plot

2. Visualizing the data

Question.29 What is BoxPlot?

Ans: - A box and whisker plot (box plot) summarizes the data in 5 numbers.
The five-numbers are the minimum, first quartile, median, third quartile, and maximum.

In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.



The "interquartile range", abbreviated "IQR", is just the width of the box in the box-and-whisker plot. That is, IQR = Q3 – Q1 . The IQR can be used as a measure of how spread-out the values are.

Outliers are the values which fall outside these margins.

Question.30 How to Treat Outliers?

Ans:-

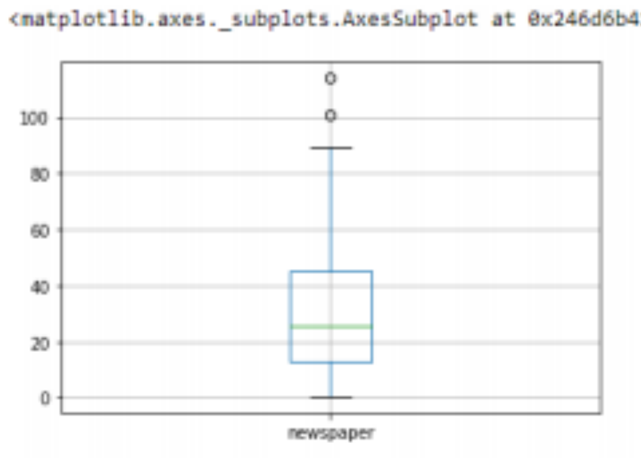
1.Mean/Median or random Imputation

2.Trimming

3.Discretization

Question.31 What is the syntax of the boxplot for visualizing the outliers?

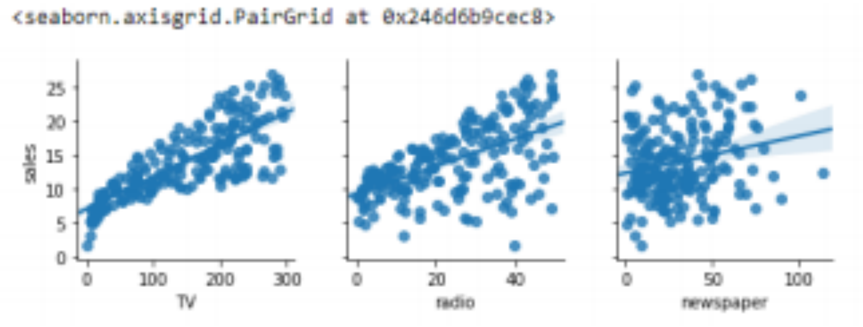
Ans:- Syntax:- df.boxplot(column='TV')



```
<matplotlib.axes._subplots.AxesSubplot at 0x246d6b41c88>
```

Question.32 How to generate a pairplot ?

Ans:- Syntax:- sns.pairplot(df,x_vars = ['TV','radio','newspaper'],y_vars = 'sales',kind = 'reg') OP:-



```
<seaborn.axisgrid.PairGrid at 0x246d6b9cec8>
```

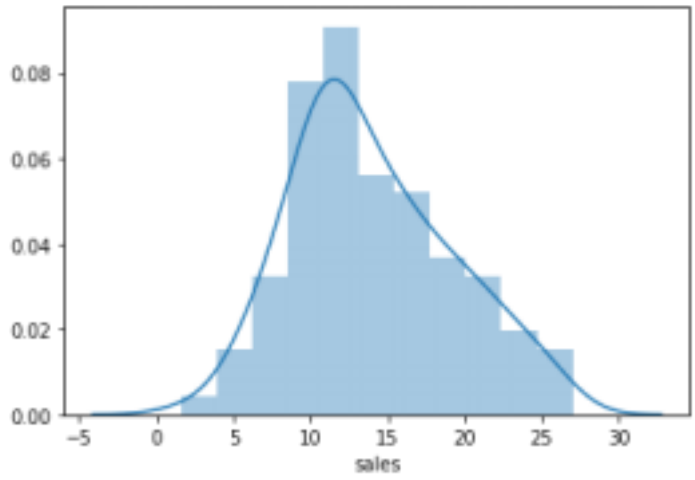
Question.33 What does the 'kind' parameter stand for?

Ans:- The kind parameter is used to show the regression line if needed.

Question.34 What is Histogram and how to generate the histogram of the variables?

Ans:- A histogram is an approximate representation of the distribution of numerical data.

Syntax:- sns.distplot(Y,hist = True)



Question.35 What is Normal Distribution?

Ans:- Normal distribution also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In normal distribution mean = mode = median In graph form, normal distribution will appear as a bell curve.

Question.36 Why do we need Log Transformation?

Ans:- Logarithmic transformation is a convenient means of transforming a highly skewed variable into a more normalized dataset. When modelling variables with non-linear relationships, the chances of producing errors may also be skewed negatively.

Question.37 What is Skewness?

Ans: - Skewness is a measure of the asymmetry of a distribution. This value can be positive or negative.

- A negative skew indicates that the tail is on the left side of the distribution, which extends towards more negative values.
- A positive skew indicates that the tail is on the right side of the distribution, which extends towards more positive values.
- A value of zero indicates that there is no skewness in the distribution at all, meaning the distribution is perfectly symmetrical.

Question.38 What is Kurtosis?

Ans: - Kurtosis is a measure of whether or not a distribution is heavy-tailed or light-tailed relative to a normal distribution.

- The kurtosis of a normal distribution is 3.
- If a given distribution has a kurtosis less than 3, it is said to be platykurtic, which means it

tends to produce fewer and less extreme outliers than the normal distribution. • If a given distribution has a kurtosis greater than 3, it is said to be leptokurtic, which means it tends to produce more outliers than the normal distribution.

Question.39 How can we convert the variables data type ?

Ans:- In machine learning there might be situations when you need to convert the data type of the variable to your liking and hence there are function for that.

Example :-

```
pincode = 400070

mystring = str(pincode) # '400070'
```

Now the value 400070 is not integer but string

In this scenario the pin code variable is not a continuous variable ,it has a categorical essence to it and hence it should be treated as such ..

Question.39 What is correlation?

Ans: - Correlation is usually defined as a measure of the linear relationship between

two quantitative variables.

Question.40 How are covariance and correlation different from one another?

Ans: - Covariance measures how two variables are related to each other and how one would vary with respect to changes in the other variable. If the value is positive, it means there is a direct relationship between the variables and one would increase or decrease with an increase or decrease in the base variable respectively, given that all other conditions remain constant.

Correlation quantifies the relationship between two random variables and has only three specific values, i.e., 1, 0, and -1.

1 denotes a positive relationship, -1 denotes a negative relationship, and 0 denotes that the two variables are independent of each other.

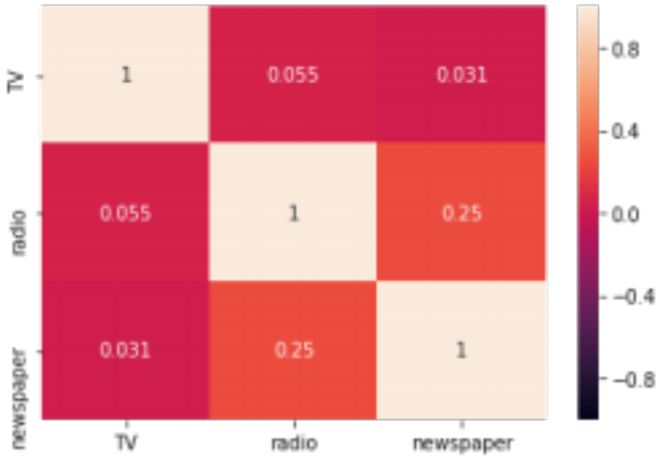
Question.41 What is Multicollinearity and how we find it in Python?

Ans: - Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model.

Syntax: -

```
a = sns.heatmap(corr_df,vmax = 1.0, vmin = -1.0, annot = True)
```

```
b, t = a.get_ylim()

a.set_ylim(b+0.5, t-0.5)
```



If any variables show high score, then that feature should be eliminated.

Question.42 What is VIF?

Ans: - Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. In general, a VIF above 5 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above and it depends on the situation.

Question.43 What is a confusion matrix and why do you need it?

Ans: - Confusion matrix is a table that is frequently used to illustrate the performance of a classification model i.e., classifier on a set of test data for which the true values are well-known. It allows us to visualize the performance of an algorithm/model. It allows us to easily identify the confusion between different classes. It is used as a performance measure of a model/algorithm. It is summary of predictions on a classification model.

Question.44 How do we check the normality of a data set or a feature?

Ans: - There is a list of Normality checks, they are as follow:

- Shapiro-Wilk W Test
- Anderson-Darling Test
- Martinez-Iglewicz Test
- Kolmogorov-Smirnov Test
- D’Agostino Skewness Test

Question.45 What is the idea behind Splitting the data?

Ans: - In Machine Learning, we split the data into 2 parts, training and testing parts. We train the model on training data and compare its results with the test data.

Question.46 What is the threshold for splitting the data?

Ans: - Usually we follow the threshold of 70:30 of the data i.e., 70 % of the data to the training and 30% of the data. In depends on the situation whether you need more data for your model if its not giving you the accuracy.

Syntax: -

```
from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3, random_state=101)
```

Question.47 What is Scaling?

Ans: - Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. ... If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Machine learning is like making a mixed fruit juice. If we want to get the best-mixed juice, we need to mix all fruit not by their size but based on their right proportion. We just need to remember apple and strawberry are not the same unless we make them similar in some context to compare their attribute. Similarly, in many machine learning algorithms, to bring all features in the same standing, we need to do scaling so that one significant.

The two major techniques for Feature Scaling are:

- Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1].
- Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

Question.48 What are the different types of Scalers available?

Ans:-

1) Min Max Scaler

2) Standard Scaler

3) Max Abs Scaler

4) Robust Scaler
5) Quantile Transformer Scaler

6) Power Transformer Scaler

7) Unit Vector Scaler

Question.49 Explain in Min Max Scaler

Ans: - An alternative approach to Z-score normalization (or standardization) is the so-called Min-Max scaling (often also simply called "normalization" - a common cause for ambiguities). In this approach, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers.

A Min-Max scaling is typically done via the following equation:

X(sc) = X – X(min)/X(max) – X(min)

Question.50 Explain Max abs Scaler

Ans: - Scale each feature by its maximum absolute value.

This estimator scales and translates each feature individually such that the maximal absolute value of each feature in the training set will be 1.0. It does not shift/centre the data, and thus does not destroy any sparsity.

Attributes:

- scale_
  Per feature relative scaling of the data.
  New in version 0.17: scale_ attribute
- max_abs_
  Per feature maximum absolute value.
- n_samples_seen_
  The number of samples processed by the estimator. Will be reset on new calls to fit, but increments across partial_fit calls.

Syntax: - class node_preprocessing.MaxAbsScaler

Question.51 Explain Robust Scaler

Ans: - Robust Scaler algorithms scale features that are robust to outliers. It uses the interquartile range. The median and scales of the data are removed by this scaling algorithm according to the quantile range.

It, thus, follows the following formula:

$$\frac{X(i) - Q1(x)}{Q3(x) - Q1(x)}$$

Where Q1 is the 1st quartile, and Q3 is the third quartile.

EG:-

```
data = [[0,5],[2,13],[-3,7],[1,-4],[6,0]]

from sklearn.preprocessing import RobustScaler

rs = RobustScaler().fit(data)

print(rs.transform(data))
```

OP:-

```
[[-0.5 0. ]

[ 0.5 1.14285714]

[-2. 0.28571429]

[ 0. -1.28571429]

[ 2.5 -0.71428571]]
```

Question.52 Explain Standard Scaler

Ans: - Standard Scaler assumes a normal distribution for data within each feature. The scaling makes the distribution centred around 0, with a standard deviation of 1 and the mean removed.

Formula:-

$$\frac{x(i) - mean(x)}{Sd(x)}$$

Where sd is the standard deviation of x.

Syntax:-

```
from sklearn.preprocessing import StandardScaler

ss = StandardScaler().fit(data)

print(ss.transform(data))
```

Question.53 Why do we need to convert the categorical variables into numerical ones?

Ans: - The machines we develop only understand categorical data, they only understand numerical data that is why we need to convert every categorical variable into numerical one.

Question.54 What to do when you have a variable with no missing values but has no variance?

Ans: - In such situations where the variable has passed the other parameters into conducted into the model but has no variance. Then you should consider removing that variable because that variable is not contributing anything useful to the model.

Question.55 If your dataset is suffering from high variance, how would you handle it?

Ans: - For datasets with high variance, we could use the bagging algorithm to handle it. Bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use polling technique to combine all the predicted outcomes of the model.

Question.56 What do you mean by feature engineering?

Ans: - Features are the core characteristics of any prediction that impact the results. Feature engineering is the process of creating a new feature, transforming a feature, and encoding a feature. Sometimes we also use the domain knowledge to generate new features.

For e.g. Using the selling price variable and the cost price variable to calculate the

profit. It prepares the data that easily input to the model and improves model

performance.

Question.57 What do you mean by feature splitting?

Ans: - A feature splitting is a technique to generate a few other features from the existing one to improve the model performance. for example, splitting names into first and last names.

Question.58 How do you select the important features in your data?

Ans: - We can select the important features using random forest, or remove redundant features using recursive feature elimination. Let's all the categories of such methods.

1. Filter Methods: Pearson Correlation, Chi-Square, Anova, Information gain, and LDA.
2. Wrapper Methods: Recursive feature elimination.
3. Embedded Methods: Ridge and Lasso Regression

Question.59 What are the different ways by Which you can convert the categorical variables into numerical ones.?

Ans: -

1. Label Encoder
2. Manually Mapping
3. Dummy Variables
4. One hot label Encoding

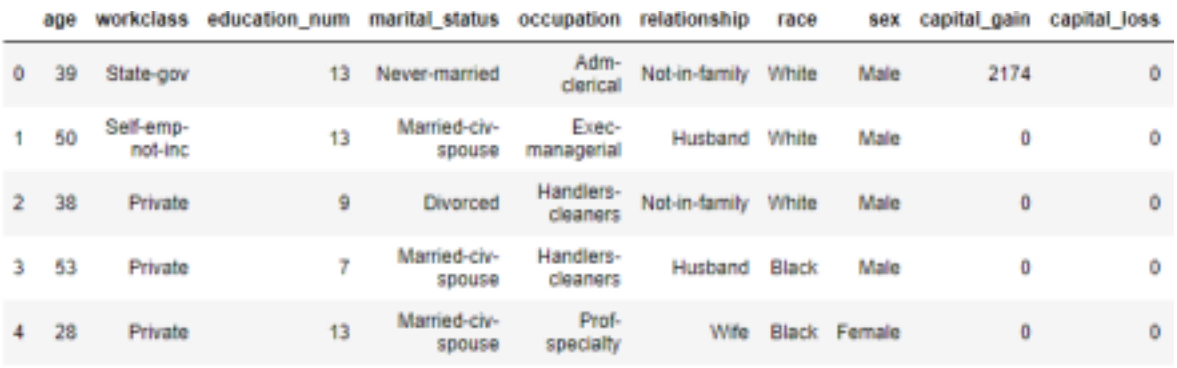
Question.60 Explain Label Encoder

Ans:- One hot encoding is used to encode the categorical column. It replaces a categorical column with its labels and fills values either 0 or 1. For example, you can see the "color" column, there are 3 categories such as red, yellow, and green. 3 categories labeled with binary values.

Syntax:-

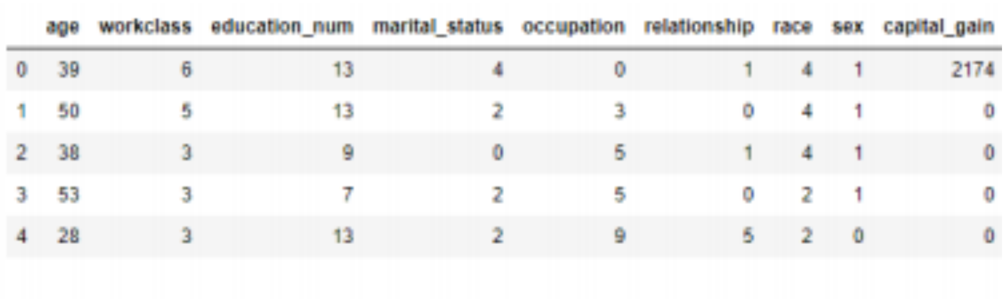
```
from sklearn import preprocessing

le=preprocessing.LabelEncoder()

for x in colname:

 adult_df_rev[x]=le.fit_transform(adult_df_rev[x])
```

Before transforming



	age	workclass	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss
0	39	State-gov	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0
1	50	Self-emp-not-inc	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0
2	38	Private	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0
3	53	Private	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0
4	28	Private	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0

After transforming



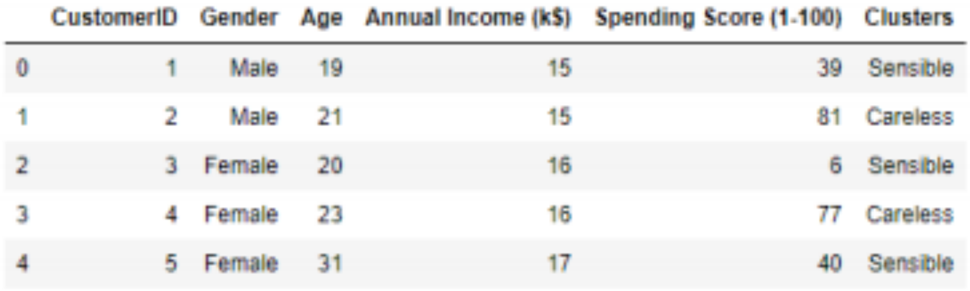
	age	workclass	education_num	marital_status	occupation	relationship	race	sex	capital_gain
0	39	6	13	4	0	1	4	1	2174
1	50	5	13	2	3	0	4	1	0
2	38	3	9	0	5	1	4	1	0
3	53	3	7	2	5	0	2	1	0
4	28	3	13	2	9	5	2	0	0

Question.61 Explain Manual Mapping

Ans:- Manual mapping is a technique where we individually take one by one element and assign them a value. This is done where you need to convert specific values and the number of these values in less.

Example : -

df["Clusters"]=df.Clusters.map({0:"Careless",1:"Standard",2:"Target",3:"Sensible",4:"Careful"}



	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	Clusters
0	1	Male	19	15	39	Sensible
1	2	Male	21	15	81	Careless
2	3	Female	20	16	6	Sensible
3	4	Female	23	16	77	Careless
4	5	Female	31	17	40	Sensible

)

Question.62 What are Dummy Variables?

Ans:- A Dummy variable is an artificial variable created to represent an attribute with two or more distinct categories/levels. Its requires less computational power compared to other techniques. However the coding length is more compared to other techniques.

Syntax:-

import pandas as pd

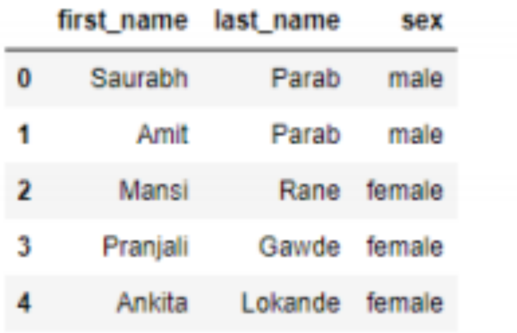
raw_data = {'first_name': ['Saurabh', 'Amit', 'Mansi', 'Pranjali', 'Ankita'],

'last_name': ['Parab', 'Parab', 'Rane', 'Gawde', 'Lokande'],

'sex': ['male', 'male', 'female', 'female', 'female']}

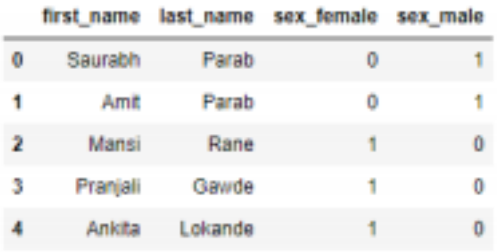
df = pd.DataFrame(raw_data, columns = ['first_name', 'last_name', 'sex'])

df



	first_name	last_name	sex
0	Saurabh	Parab	male
1	Amit	Parab	male
2	Mansi	Rane	female
3	Pranjali	Gawde	female
4	Ankita	Lokande	female

pd.get_dummies(df, columns=['sex'])



	first_name	last_name	sex_female	sex_male
0	Saurabh	Parab	0	1
1	Amit	Parab	0	1
2	Mansi	Rane	1	0
3	Pranjali	Gawde	1	0
4	Ankita	Lokande	1	0

So after creating these variables the parent variable is of no use and hence must be eliminated.

Question.63 Explain One Hot Label Encoding

Ans:- It is a process that converts categorical data to integers or a vector of ones and zeros. The length of vector is determined by number of expected classes or categories. Each element in the vector represents a class. Therefore, a one is used to indicate which class it is and everything else will be zero.

Code:-

```
 from sklearn.preprocessing import OneHotEncoder

type_one_hot = OneHotEncoder(sparse=False).fit_transform(

train_new.array.to_numpy().reshape(-1,1))
```

If we have categorical data that we think may be important, we want to be able to use this in the model. This is because regression algorithms and classification algorithms won't be able to process it. This is when one-hot encoding is useful.

Question.64 What are the Different Types of Feature Selection Techniques?

Ans:- Its is not possible that all the variables will be useful to the model and hence in machine learning you have to apply some feature selection techniques to make your model the best. Using all the features to the model reduces the overall accuracy of a classifier.

The goal of feature selection in machine learning is to find the best set of features to build useful models.

The techniques for feature selection in machine learning can be broadly classified into the following categories:

- Filter methods
- Wrapper methods
- Embedded methods
- Hybrid methods

Filter Methods: - Filter methods use the properties of the features measured via univariate statistics. These methods are faster and less computationally expensive than wrapper methods. When dealing with high-dimensional data, it is computationally cheaper to use filter methods.

Some of the filter method Techniques: -

1. Information Gain: - Information gain calculates the reduction in entropy from the transformation of a dataset. It can be used for feature selection by evaluating the Information gain of each variable in the context of the target variable.
2. Chi-square Test: - The Chi-square test is used for categorical features in a dataset. We calculate Chi-square between each feature and the target and select the desired number of features with the best Chi-square scores.
3. Correlation coefficient: - Correlation is a measure of the linear relationship of 2 or more variables. Through correlation, we can predict one variable from the other. The logic behind using correlation for feature selection is that the good variables are highly correlated with the target. Furthermore, variables should be correlated with the target but should be uncorrelated among themselves. If two variables are correlated, we can predict one from the other. Therefore, if two features are correlated, the model only really needs one of them, as the second one does not add additional information.

Wrapper Methods: - Wrappers require some method to search the space of all possible subsets of features, assessing their quality by learning and evaluating a classifier with that feature subset. The feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given data.

Some of the Wrapper method techniques are: -

1. Forward Feature Selection: - This is an iterative method wherein we start with the best performing variable against the target. Next, we select another variable that gives the best performance in combination with the first selected variable. This process continues until the target is achieved.
2. Recursive Feature Elimination :- Recursive feature elimination (RFE) selects features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coef_ attribute or through a feature_importances_ attribute. Then, the least important features are pruned from the current set of features. That procedure is recursively repeated on the data set until the desired number of features to select is eventually reached.