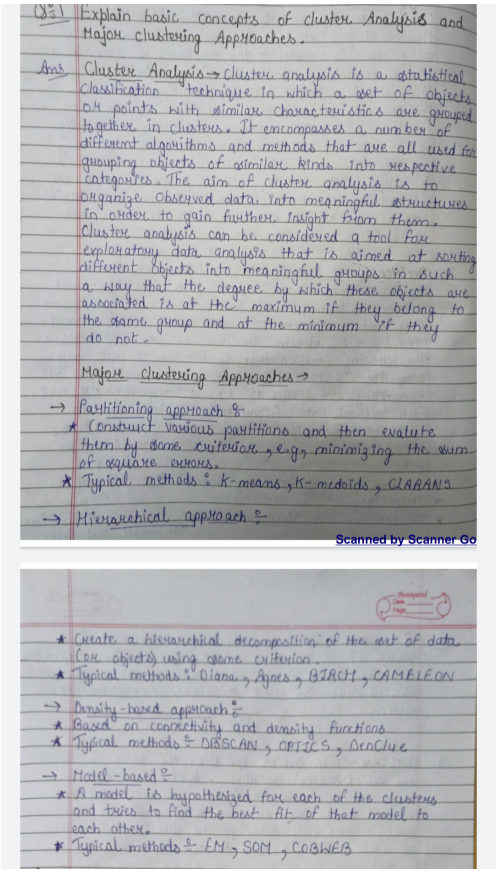
Cluster Analysis:-



Q.5.1 Explain basic concepts of cluster Analysis and Major clustering Approaches.

Ans Cluster Analysis → Cluster analysis is a statistical classification technique in which a set of objects or points with similar characteristics are grouped together in clusters. It encompasses a number of different algorithms and methods that are all used for grouping objects of similar kinds into respective categories. The aim of cluster analysis is to organize observed data into meaningful structures in order to gain further insight from them. Cluster analysis can be considered a tool for exploratory data analysis that is aimed at sorting different objects into meaningful groups in such a way that the degree by which these objects are associated is at the maximum if they belong to the same group and at the minimum if they do not.

Major Clustering Approaches →

→ Partitioning approach :-
* Construct various partitions and then evalute them by some criterion, e.g., minimizing the sum of square errors.
* Typical methods : k-means, k-medoids, CLARANS

→ Hierarchical approach :-
* Create a hierarchical decomposition of the set of data (or objects) using some criterion.
* Typical methods : Diana, Agnes, BIRCH, CAMELEON

→ Density-based approach :-
* Based on connectivity and density functions
* Typical methods : DBSCAN, OPTICS, DenClue

→ Model-based :-
* A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other.
* Typical methods : EM, SOM, COBWEB

Applications Of Cluster Analysis:

- It is widely used in image processing, data analysis, and pattern recognition.
- It helps marketers to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.
- It can be used in the field of biology, by deriving animal and plant taxonomies and identifying genes with the same capabilities.
- It also helps in information discovery by classifying documents on the web.

Advantages of Cluster Analysis:

- It can help identify patterns and relationships within a dataset that may not be immediately obvious.
- It can be used for exploratory data analysis and can help with feature selection.
- It can be used to reduce the dimensionality of the data.
- It can be used for anomaly detection and outlier identification.
- It can be used for market segmentation and customer profiling.

Disadvantages of Cluster Analysis:

- It can be sensitive to the choice of initial conditions and the number of clusters.
- It can be sensitive to the presence of noise or outliers in the data.
- It can be difficult to interpret the results of the analysis if the clusters are not well-defined.
- It can be computationally expensive for large datasets.
- The results of the analysis can be affected by the choice of clustering algorithm used.
- It is important to note that the success of cluster analysis depends on the data, the goals of the analysis, and the ability of the analyst to interpret the results.

 # Clustering Structures :-

Hierarchical clustering results in a clustering structure consisting of nested partitions. In an agglomerative clustering algorithm, the clustering begins with singleton sets of each point. That is, each data point is its own cluster. At each time step, the most similar cluster pairs are combined according to the chosen similarity measure, and this step is repeated either until all data points are included in a single cluster or until some predetermined criteria are met. If the nesting occurs in the other direction, that is, the clustering begins with one large cluster and breaks down into smaller clusters, it is known as a divisive clustering algorithm.

 # Why Outlier analysis:-

Outlier analysis is the process of identifying outliers, or abnormal observations, in a dataset. Also known as outlier detection, it's an important step in data analysis, as it removes erroneous or inaccurate observations which might otherwise skew conclusions.

There are a wide range of techniques and tools used in outlier analysis. However, as we'll see later, it's often very easy to spot outlying data points. As a result, there's really no excuse not to perform outlier analysis on any and all datasets.

–Outlier Analysis Techniques
There are a wide variety of techniques that can be used to identify outliers in datasets. In this section, we'll look at just a few of these techniques, including both straightforward and sophisticated ones.

Sorting
For an amateur data analyst, sorting is by far the easiest technique for outlier analysis. The premise is simple: load your dataset into any kind of data manipulation tool (such as a spreadsheet), and sort the values by their magnitude. Then, look at the range of values of various data points. If any data points are significantly higher or lower than others in the dataset, they may be treated as outliers.

Graphing

An equally forgiving tool for outlier analysis is graphing. Once again, the premise is straightforward: plot all of the data points on a graph, and see which points stand out from the rest. The advantage of using a graphing approach over a sorting approach is that it visualizes the magnitude of the data points, which makes it much easier to spot outliers.

Z-score
A more statistical technique that can be used to identify outliers is the Z-score. The Z-score measures how far a data point is from the average, as measured in standard deviations. By calculating the Z-score for each data point, it's easy to see which data points are placed far from the average. Unfortunately, like sorting, this doesn't take into account the influence of a second variable.

Why ?

Outlier is a data point in the dataset that differs significantly from the other data or observations. Just look at the picture above, there are a series of bottles, but one is colored differently. This one bottle is what we called an outlier.

The outlier is inherently different than Noise. While Outlier is a data that significantly different compared to the other data, Noise is a random error or variance. The outlier is part of the data, but Noise is just a random error (could be mislabeled or mistake or even missing data).

Many parametric statistics, like mean, correlations, and every statistic based on these is sensitive to outliers. Since the assumptions of standard statistical procedures or models, such as linear regression and ANOVA also based on the parametric statistic, outliers can mess up your analysis.

 # Identifying and Handling of outliers :-

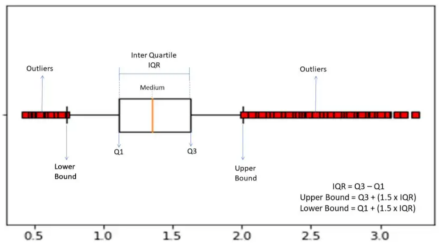
Outlier detection is one of the analysis and cleansing data. In several cases, these outliers disturb our model as in the regression model. If we don't handle it well, it will make our model will not perform better. So, here I want to tell you how to detect outliers and handle them well. By handling properly, it makes our model will perform better. You can try it!

Here are the outlines of what we're going to cover:

- Outlier Detection with IQR
- Handling Outliers
- Practice Detect and Handling
- Conclusion

1. Outlier Detection with IQR

First, Before we detect outliers, we have to know what outliers are. Please see the picture below!



Boxplot Explanation By Author

The picture above shows that outliers are located outside the upper bound and the lower bound. So, it simply understands. But, how can we know the sum of outliers in each variable? To solve it, we have to implement the formula like the picture above. First, we have to calculate the IQR by dividing Q3 with Q1. Second, calculating *Upper Bound* with Q3 plus the result of 1.5 times IQR. Third, calculating *Lower Bound* with Q1 plus the result of 1.5 times IQR.

When we have found the IQR, the upper bound and the lower bound, we can see which values exceed the upper bound and which values are smaller than the lower bound. Then we can calculate that value as the number of outliers.

So that we can understand more, let's practice in the third session.

2. Handling Outliers

Ok, here the second session, talking about how can we handle outliers properly. There several techniques for handling outliers. We only give 3 techniques:

a. Dropping the outliers data: You omit the outliers values.

b. Caping the outliers data: You replace the outliers values with upper bound and lower bound. outliers that are located at more upper bound be replaced by upper bound values. Otherwise, outliers that are located at more the lower bound can be replaced with lower bound.

c. Replacing with new values: You replace outliers value with mean, median, or mode.

If you have another method you can give us a suggestion in a comment. Ok, let's practice detect and handle outliers in the third session.

3. Practicing to Detect and Handling Outliers

In the third session, we practising detects and handling outliers. In this case, we can use this dataset. Please download if you want to practice now. When you process the dataset, there are 2 columns, *date* and *average_price* columns. Ok, here we want to see the outliers of the *average_price* column and handle it. Please follow this path:

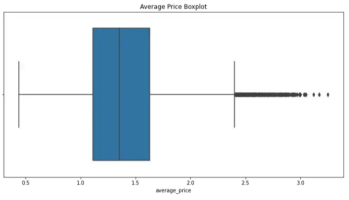
1. First, we have to analyze the outliers of the average_price variable which shows the boxplot. here's the code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
data = 'avocado_price.csv'
df = pd.read_csv(data)
fig,ax = plt.subplots(figsize=(12,6))
fig = sns.boxplot(df.average_price).set_title
fig.figure.savefig('boxplot_origin.png')
```

analyze outliers hosted with ❤ by GitHub view raw

Boxplot Visualization Code By Author

after you write and run the code, it will show this result:



Resul of Boxplot Visualization

The picture shows that average_price has many outliers values. When you want to know how's the summary of outliers. Like how many sums of outliers, what's the value of IQR, upper bound, and lower bound. You can follow this code:

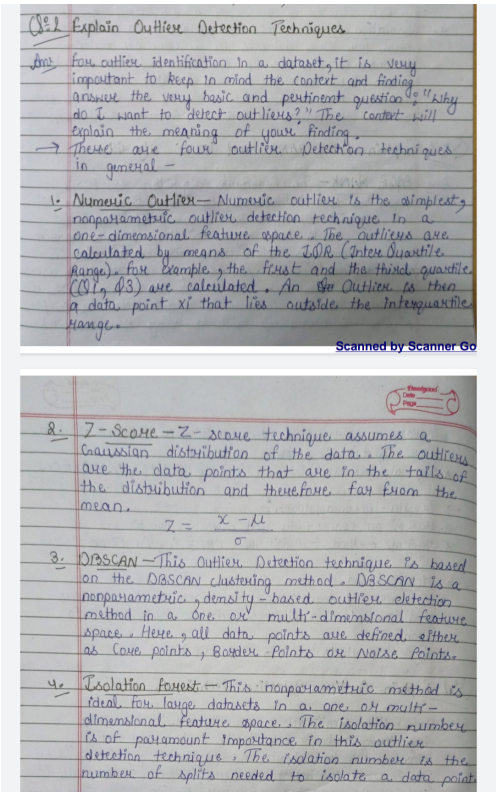
```
import numpy as np
import pandas as pd
def outliers(s):
    iqr = (np.quantile(s, 0.75))-(np.quantil
    upper_bound = np.quantile(s, 0.75)+(1.5*
    lower_bound = np.quantile(s, 0.25)-(1.5*
    f = []
    for i in s:
        if i > upper_bound:
            f.append(i)
        elif i < lower_bound:
            f.append(i)
    sums = len(f)
    pros = len(f)/len(s)*100
    d = {'IQR':iqr,
         'Upper Bound':upper_bound,
        'Lower Bound':lower_bound,
        'Sum outliers': sums,'percentage out
    d = pd.DataFrame(d.items(),columns = ['s
    return(d)

outliers(df.average_price)
```

Conclusion

So, in this case, we can conclude that the technique which handles outliers are cap and drop outliers. Which one is the best? You can make an experiment and choose which give us better accuracy.

If you have several suggestions to make better or has other technique, please write in comment.

Outlier Detection Techniques :-


Q.2 Explain Outlier Detection Techniques

Ans For outlier identification in a dataset, it is very important to keep in mind the context and finding answers the very basic and pertinent question, "Why do I want to detect outliers?" The context will explain the meaning of your finding.

→ There are four outlier detection techniques in general -

1. Numeric Outlier - Numeric outlier is the simplest, nonparametric outlier detection technique in a one-dimensional feature space. The outliers are calculated by means of the IQR (Inter Quartile Range). For example, the first and the third quartile (Q1, Q3) are calculated. An Outlier is then a data point xi that lies outside the Interquartile range.

Scanned by Scanner Go

2. Z-Score - Z-score technique assumes a Gaussian distribution of the data. The outliers are the data points that are in the tails of the distribution and therefore far from the mean.

$$Z = \frac{x - \mu}{\sigma}$$

3. DBSCAN - This Outlier Detection technique is based on the DBSCAN clustering method. DBSCAN is a nonparametric, density-based outlier detection method in a one or multi-dimensional feature space. Here, all data points are defined, either as Core points, Border points or Noise points.

4. Isolation Forest - This nonparametric method is ideal for large datasets in a one or multi-dimensional feature space. The isolation number is of paramount importance to this outlier detection technique. The isolation number is the number of splits needed to isolate a data point

Web Mining :-

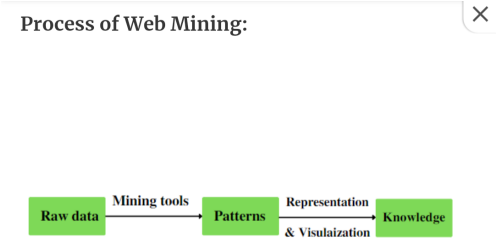
Web Mining

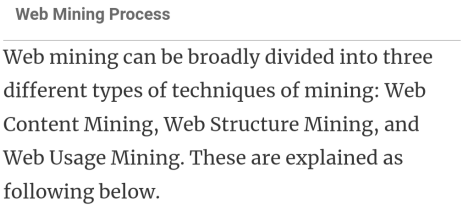
Web Mining is the process of Data Mining techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.

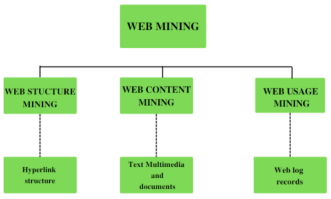
Applications of Web Mining:

Web mining helps to improve the power of web search engines by classifying web documents and identifying web pages.
It is used for Web Searching e.g., Google, Yahoo, etc, and Vertical Searching e.g., FatLens, Become, etc.

Web mining is used to predict user behavior.
Web mining is very useful for a particular Website and e-service e.g., landing page optimization.

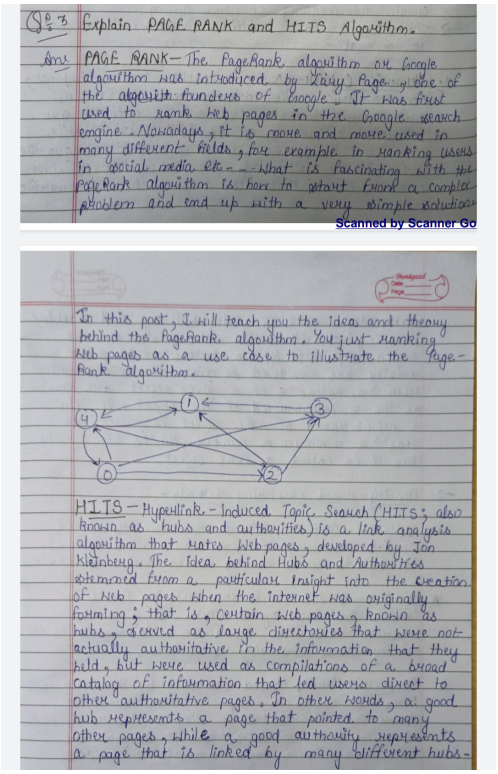






1. **Web Content Mining:** Web content mining is the application of extracting useful information from the content of the web documents. Web content consist of several types of data – text, image, audio, video etc. Content data is the group of facts that a web page is designed. It can provide effective and interesting patterns about user needs. Text documents are related to text mining, machine learning and natural language processing. This mining is also known as text mining. This type of mining performs scanning and mining of the text, images and groups of web pages according to the content of the input.
2. **Web Structure Mining:** Web structure mining is the application of discovering structure information from the web. The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Structure mining basically shows the structured summary of a particular website. It identifies relationship between web pages linked by information or direct link connection. To determine the connection between two commercial websites, Web structure mining can be very useful.

3. **Web Usage Mining:** Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets. And these patterns enable you to understand the user behaviors or something like that. In web usage mining, user access data on the web and collect data in form of logs. So, Web usage mining is also called log mining.



Q.3 Explain PAGE RANK and HITS Algorithm.

Ans PAGE RANK – The PageRank algorithm or Google algorithm was introduced by Larry Page, one of the algorithm founders of Google. It was first used to rank web pages in the Google search engine. Nowadays, it is more and more used in many different fields, for example in ranking users in social media etc. What is fascinating with the PageRank algorithm is how to start from a complex problem and end up with a very simple solution.

Scanned by Scanner Go

In this post, I will teach you the idea and theory behind the PageRank algorithm. You just ranking web pages as a use case to illustrate the Page-Rank algorithm.

HITS – Hyperlink - Induced Topic Search (HITS; also known as hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that they held, but were used as compilations of a broad catalog of information that led users direct to other authoritative pages. In other words, a good hub represents a page that pointed to many other pages, while a good authority represents a page that is linked by many different hubs.

Q.24 Explain Density-Based Methods and Model-Based Clustering.

Ans Density-Based Methods - Density-Based clustering refers to one of the most popular unsupervised learning

methodologies used in model building and machine learning algorithms. The data points in the region separated by two clusters of low point density are considered as noise. The surroundings with a radius ε of a given object are known as the ε neighborhood of the object. If the ε neighborhood of the object comprises at least a minimum no number, MinPts of objects, then it is called a core object.

→ Density-Based clustering Methods -

→★ DBSCAN - DBSCAN stands for Density-Based Spatial clustering of Applications with Noise. It depends on a density-based notion of cluster

★ OPTICS - OPTICS stands for ordering Points to Identify the clustering structure. It gives a significant order of database with respect to its density-based clustering structure.

★ DENCLUE - Density-based clustering by Hinneburg and Keim. It enables a compact mathematical description of arbitrarily shaped clusters in high dimension state of data, and it is good for data sets with a huge amount of noise.

Model Based Clustering - Model-based clustering is a statistical approach to data clustering. The observed (multivariate) data is considered to have been created from a finite combination of component models. Each component

model is a probability distribution, generally a parametric multivariate distribution. For instance, in a multivariate Gaussian mixture model, each component is a multivariate Gaussian distribution. The component responsible for generating a particular observation determines the cluster to which the observation belongs.