Aniket kumar
200801

Name :- Aniket Kumar
Enroll no :- AJU/200801
Sub :- Exploratory Data Analysis.

Assignment no: 01

1. What is Explanatory Data Analysis?

Exploratory data analytics (EDA) is a form of Data analytics. This field is involved with analyzing and exploring datasets in order to summarize their dominant characteristics.

The American Mathematician John Tukey developed EDA in 1977, and in the time since, it has continued to play an integral role in the data discovery process.

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns to spot anomalies to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

EDA is all about making sense of data in hand set before getting them dirty with it.

2 What is the importance of EDA in real world data analysis?

importance of EDA in real world data analysis -

Healthcare - EDA is helpful for spotting natural patterns embedded in large stores of medical data. In addition, healthcare networks, healthcare department, and hospital store large amounts of data in electronic medical records.

Retail - EDA can be used by business managers to spots weak areas in a store or franchise in order to suggest areas that can be targeted for increased revenue.

Fraud detection - When EDA mining techniques are used on medicare datasets, it's possible to evaluate the risk of a given individual for fraudulent activity

Auditing - EDA can be applied to several stages of auditing, for both internal and external audit cycles.

2 The food industry

3. Compare Exploratory Data Analysis with classical and Bayesian Analysis

Classical data Analysis	Exploratory data analysis	Bayesian Data Analysis
Problem Definition	Problem Definition	Problem Definition
Data collection	Data collection	Data collection
Model development	Data analysis	Model Development
Data Analysis	Model Development	Prior Distribution
Results communication	results communication	Data Analysis
		Result communication

4 What are the software tools available for EDA

Some of the most common tools used to create an EDA are :-

R - An open-source programming language and free software environment for statistical computing and graphics supported by the R foundation for statistical computing.

Python pandas and NumPy are quite useful together to spot missing values in the data set [illegible] in [illegible] so we decide the way to handle missing value for machine learning.

5 Write down the goal of Exploratory Data Analysis?

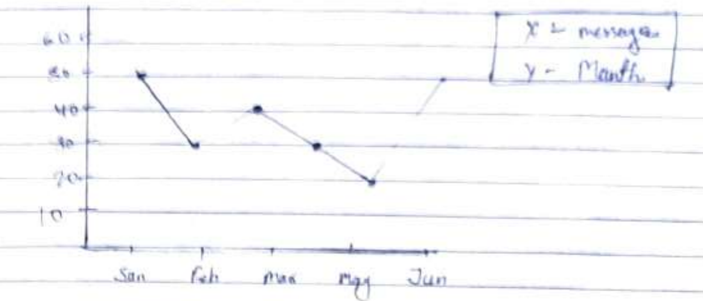
The primary goal of EDA is to maximize the analyst insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as

- a good-fitting, parsimonious model
- a list of outliers
- a sense of robustness of conclusion
- estimate for parameters
- uncertainties for those estimates
- a ranked list of important factors
- conclusion as to whether individual factors are statistically significant.
- optimal settings

Q with the suitable example, Explain

1) Line Chart - Line chart are used to represen quantitative data collected over a specific subje and a specific time interval, all the data points are connected by a line. Data points represent the observations that are collected on a survey or research.

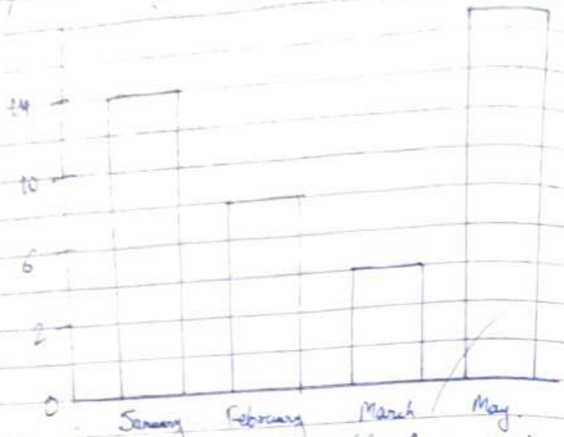
The line graph has an x-axis and a y-axis.



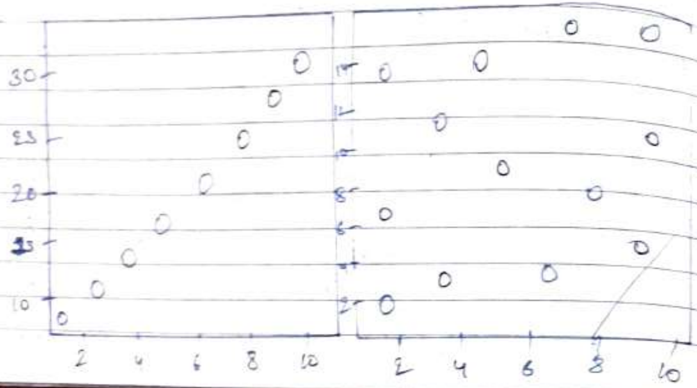
Bar Chart - A bar chart represents categorical data with rectangular bars having lengths proportio-al to the values that they represent.

This is one of the most common types of visualiza ion that almost everyone must have encountered.

Bar can be drawn horizontally or vertically to represent categorical variables.



Scatter plot - A scatter plot is a series of points that show how two variables are related to each other.



Area plot - An area plot displays graphically quan data. It is based on the line chart.
The area between axis and line are commonly emphasised with color, textures and hatching

Area charts are commonly used to showcase d that depits a time - series relationship.

$20K
$15K
$10K
$5K
$0
Q1
Q2
Q3
Q4

Pie chart - A pie chart is a circular statistical graphical, which is divided into slices to illustrate numerical proportion.

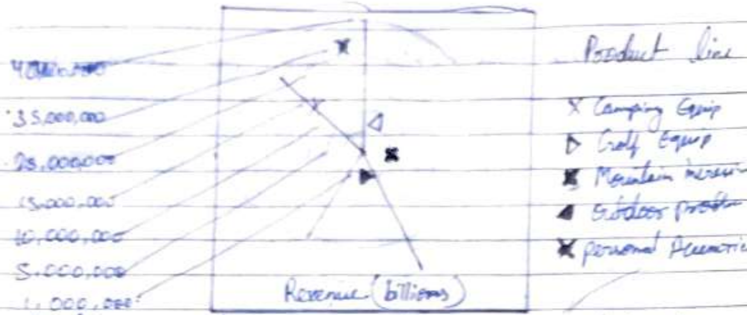
In a pie chart, they are length of each slice is proportional to the quality it represents.
Pie chart are very widely used in the business world and the mass media

Table chart - A table chart is a means of arranging data in rows and columns. The use of table is pervasive throughout all communication, research and data analysis.

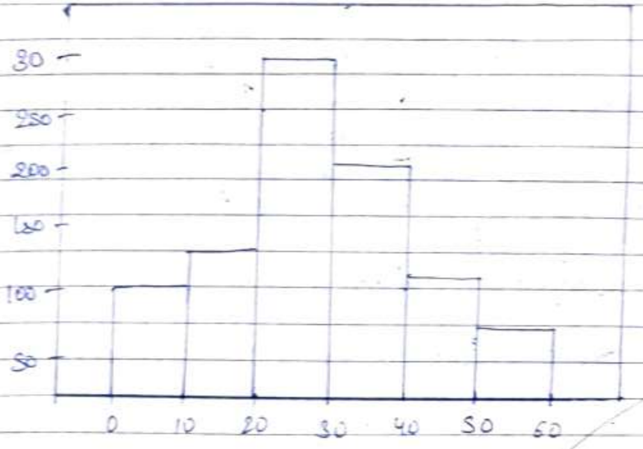
Tables appear in print media, handwritten notes, computer software, architectural ornamentation, traffic signs and many other places.

	A	B	C
X	$40	240	48
Y	$50	200	59
Z	$60	310	29

Polar chart - polar chart are circular charts that use value and angle to show information as polar coordinates. polar charts are useful for showing scientific data



Histogram - A histogram is a value distribution plot of numerical columns. It basically creates lines in various ranges in values and plots it where we can visualize how values are distributed.



7 Write a short note on dplyr. package

The dplyr package in R programming language is a structure of data manipulation that provides a uniform set of verbs, helping to resolve the most frequent data manipulation hurdles.

dplyr package provides various important functions that can be used for Data manipulation.

8. Write a short note on Data Transformation

Data Transformation is having a sense of how Data is distributed, both from using visual or quantitatives summaries. we can consider transformation of variables to ease both interpretation of data analysis and the application statistical and machine learning models to a dataset.

9 with the suitable example explain :-

select () and rename (): for choosing variable and using their names as a base for doing so.

example of select ()

```
# Create a dataframe with missing data
d <- data.frame(name = c("Abhi", "Bhavya", "Ben", "Raju"),
                age = c(7, 5, 9, 16),
                ht = c(46, NA, NA, 69))

# starts_with() function to input only ht data
select (d, starts_with ("ht"))

# -starts_with() function to print everything except ht data
Select (d, -starts_with ("ht"))

# printing data of column heading containing 1 to 2
select (d, 1:2)

# printing data to column heading containing 'a'
select (d, contains ("a"))

# printing data of column heading which matches 'na'

select (d, matches ("na"))
```

filter () function - for choosing cases and using their values as a base for doing so.

example :-

```
# Create a data frame with missing data

   d <- data.frame (name = c ("Abhi", "Ani", "Piya", "Sonu")
                    age = c ( 7, 5, 9, 16)
                    ht = c (46, NA, NA, 69)
                    School = c ("yes", "yes", "no", "no"))

   d
# Finding rows with NA value

   d %>% filter (is.na (ht))

# Finding rows with no NA value.

   d %>% Filter (! is.na (ht))
```

arrange () function - It is used for re-ordering of the classes

```
# Example

# Create a data frame with missing data

   d <- data.frames (name = c ("Ani", "Sonu", "Piya"),
                     age = c (7, 5, 9)
                     ht = c (46, NA, NA)
                     school = c ("yes", "No", "No"))

# Arranging name according to the age

   d.name <- arrange (d, age)
   print (d.name)
```

Mutate () function — mutate () and transmute () is used for Addition of new variables which are the functions of preceding variables.

example of mutate

```
# Create a data frame with missing data
  d <- data.frame (name = c ("Ani", "Piya", "Sonu"),
                   age = c (7, 5, 9)
                   school = c ("yes", "yes", "No")
                   ht = c (46, NA, NA))
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

Histogram - Histogram is the plot of values of data vs. their frequency in the dataset

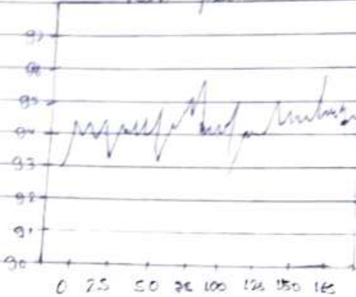
The histogram is used to know the distribution of the process i.e. whether it is uniform, normal etc.

Vertical axis : counts / frequency / probability
Horizontal axis : Y

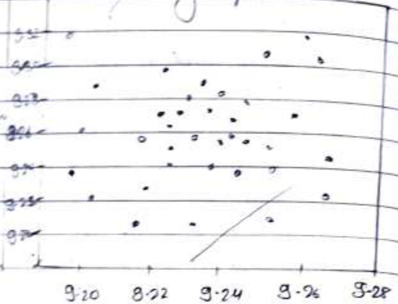
Normal probability - Normal probability is used to know how close the process distribution to normal distribution

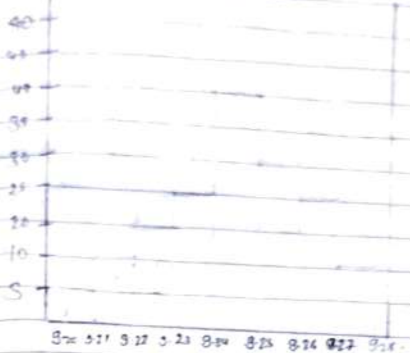
Vertical axis - Ordered Y:
Horizontal axis - The theoretical values from the normal distribution N(0,1)

Run plot

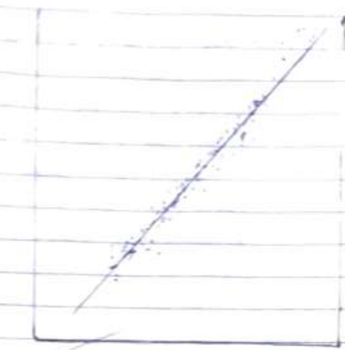


lag plot





Histogram



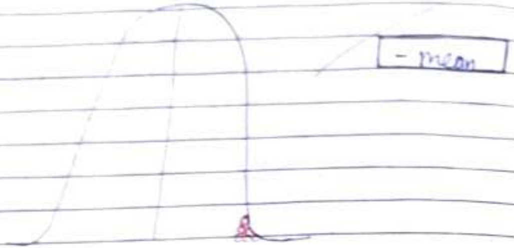
Normal probability

2) What does it mean if a normal probability plot is linear ?

if a normal probability plot is linear or the data follow a normal distribution with mean and variance, then a plot of the theoretical percentiles of the normal distribution versus the observed sample percentil should be approximately linear.

3) What is p value in probability plot ?

The P value means the probability under a given Statistical model that when the null hypothesis is true the Statistical Summary would be equal to or more extreme than the actual Observed results.



P. value corresponding to the real point tells us about the total probability of getting any value to the right hand side of the red point, when the values are picked randomly from population distribution

P- value does not hold any value by itself. A large p value implies that sample Scores are more aligned or similar to the population score.

4) What is the importance of 4plot ?

4 plot is important for testing underlying Assumptio Helps ensures the validity of the final Scientific and engineering conclusion.

The 4 plot sequence plot, lag plot, histogram and normal probability is seen as a simple efficient and powerful way of carrying out routine" Checking.

5) Define consequences ?

Consequences of inappropriate data analysis afte reanalysis

To determine the effects of the incorrect Data analysis procedures observes in the literatu survey. it would be necessary to reanalyze the data correctly and compare difference in the results. This is not possible without access to the raw data.

5) What is the difference among Univariate, Bivariate and multivariate analysis?
Univariate data - This type of data consists of only one variable.

The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes.

It does not deal with causes or relationships and the main deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

Bivariate data - This type of data involves two different variables.

The analysis of this type of data deals with causes with cause and relationships and the analysis is done to find out the relationship among the two variables.

Multivariate data - When the data involves three or more variables.

It is similar to bivariate but contains more than one dependent variable.

The ways to perform analysis on this data depends on the goals to be achieved.

7) During the data pre-processing step, how should one treat missing / null values? How will you deal with them through R programming?

Missing values are practical in life for eg- some cells in spread-sheet are empty. If an insensible or impossible arithmetic operations is tried than NAS occur.

Dealing missing values in R :-

missing values in R are handled with the use of some predefined function.

is na() function for finding Missing values :-

A logical vector is returned by the function that indicates all the NA values present. It return a Boolean value. If NA is present in a vector i return TRUE else else FALSE

```
X <- c(NA , 3, 4, NA , NA, NA )
is. na (x)
```

Output :

```
[1] TRUE FALSE FALSE TRUE TRUE TRUE
```

is.nan function for finding Missing values :-

```
x<c (NA, 3. 4, NA. NA , 0/, 0/0 )
is. nan (x)
```

Output :-

```
[1] FALSE FALSE FALSE FALSE FALSE TRUE TRUE
```

25/08/22

Assignment no :- 03

1) Explain EDA Techniques

There are four explanatory data analysis techniques that data experts use, which include :-

Univariate non-Graphical - This is the simplest type of EDA, where data has a single variable. Since there is only one variable, data professionals do not have to deal with relationships

Univariate Graphical - Non graphical technique do not present the complete picture of data. Therefore, for comprehensive EDA, data specialists implement graphical method, such as stem-and-leaf plot, box plot and histograms.

Multivariate Graphical This EDA technique makes use of graphics to show relationships between 2 or more datasets.

Multivariate Non Graphical.- Multivariate data consist of several variables. Non-graphical multivariate EDA methods illustrate relationships between 2 or more data variable using statistic or cross tabulation

2) What is Alphabetical Graphical techniques ?

This section provides a galley of some useful graphical techniques

The techniques are orders alphabetically so this section is not intended be read in a sequential fashion.

The use of most of these graphical technique is demostrated in Graphical technique.

3) Explain EDA using probability density function ?

A function that defines the relationship between a random variable and its probability, such that we can find the probability of the variable using the function, is called a probability Density function

The probability Density function is called as the probability mass function when deal with Discreate Data.

4) What is a quantitative exploratory analysis ?

Exploratory data analysis is the essential fi step of any quantitative data analysis.

It provides you with an overview of the d and allow to select variable of interest verify your first inteitions about the data and explane possible relationships

5) Explain EDA using Quantitative distribution function

The EDA Types of techniques are either graphic or quantitative. While the graphical method invo summarising the data in a diagramatic or vis way the quantitative method on the other ha involves the calculation of summary statistics

These two type of method are further divided into univariate and multivariate meth

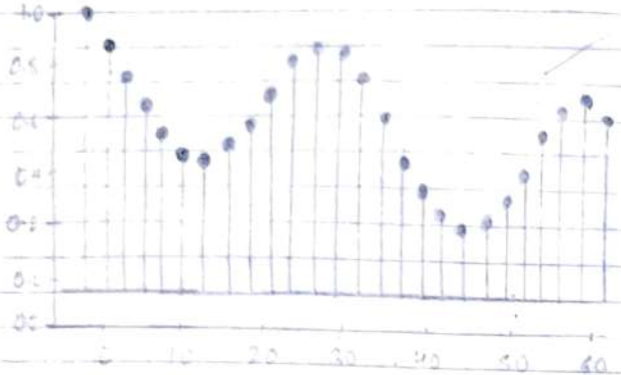
—— x —— x ——

1/08/22

Assignment no : 04

1) What is Auto correlation Structure in Random Walk ?

Autocorrelation involves finding the correlation between a time series and a lagged version of itself.



The XAxis is the lag k, and the XAxis is the pearson's correlation coefficient at each lag. The red shaded region is a confidence interval. If the height of the bars is outside this region it means the correlation is statistically significant.

2. Explain the credit risk analysis with EDA

Credit risk analysis is a form of analysis performed by a Credit analyst to determine a borrower's ability to meet their deft obligations.

The purpose of credit analysis is to determine the creditworthiness of borrowers by quanlifying the risk of loss that the lender is exposed to.

3) Write down the steps of the ceramic strength analysis.

To test the hypothesis of the steps of the ceramic strength analysis is effective to predict the reliability of an alumina based ceramic.

The ceramic strength analysis divided into 3 groups

1) Step-Stress accelerating test
2) Flexual strength - control
3) Flexural Strength - mechanical aging.

4) Write the goal of the case study of Heat Flow meter

The heat flow meter determines the thermal conductivity and the thermal transmittance, which is also known as U-value, of materials with a low thermal conductivity.

GHP measurements have a higher accuracy since the HFM measurement is comparative method of data analysis.

5) Write the steps of Beam Deflection analysis?

The beam can be bent or moved away from its original position. There are mainly 4 steps which can determine the magnitude of beam deflection Analysis These includes -

- How much loading is on the structure
- The length of the unsupported member
- The material ; specifically the young's modulus.
- The cross section size, Specifically the moment of Inertia

7/09/22

—— x —— x ——

Assignment NO:- 05

1. What is the advantages and benefits of good data visualization

Advantages and the benefits of good data visualization

visualization allows visual access to huge amount of data in easily digestible visuals

Faster Decision making

Making sense of complicated data

Identifies errors and inaccuracies in data quickly

Access real time information and assist in management functions

It provides storytelling and conveys the exact message to the audience.

Solve data inefficiencies and absorb vast amount of data presented in visual format.

2) How do you visualize website data ?

We can visualise website data through

Google chart - Google Charts is a completely free tool that ~~affect~~ offers a number of default models to visualize data.

Tableau - Tableau is an enterpoize level data analytics platform that can drill into different kinds of data to make sense of them.

3) Write a short note on - Content based document clustering.

The idea of content based document clustering algorithm is to cluster the document by using the concept that present in sufficient number of documents. Our approach doesn't consider the documents as bag of woods but as a set of semantically related words for example - Content : based document clustering created a cluster for the frequent concept announce made up all the documents that contain words which are either to identical or related

4 How do you visualize time series data in R ?

Time series in R is used to see how an object behaves over a period of time.

xts() function is the most useful tool in the R time series data visualization artillery. It is fairly similar to general plotting but its x axis contains a time scale.

plot() and plot.xts() function are generally use to visualize the time series data in R.

5 What is a ggplot 2 in R ?

ggplot2 in R is the latest version of the famous open data visualization tool ggplot for the statistical programming language R.

The term ggplot-2 relates to the package's name. we use the function ggplot() to produce the plots when using the package. Therefore, ggplot() is the command, and the whole package is called ggplot2.