DAV SAMPLE QUESTIONS

Module 1 – Data Analytics Lifecycle

1. What are the roles of key stakeholders of an analytics project? 2B2P3D

Ans)Key roles for a successful analytics project

- **Business User:** Someone who understands the domain area and usually benefits from the results. This person can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized. Usually a business analyst, line manager, or a deep subject matter expert in the project domain fulfills this role.
- **Project Sponsor:** Responsible for the genesis of the project. Provides the impetus and requirements for the project and defines the core business problem. Generally provides the funding and gauges the degree of value from the final outputs of the working team. This person sets the priorities for the project and clarifies the desired outputs.
- **Project Manager:** Ensures that key milestones and objectives are met on time and at the expected quality.
- **Business Intelligence Analyst :** Provides business domain expertise based on a deep understanding of the data, key performance indicators (KPis), key metrics, and business intelligence from a reporting perspective. Business Intelligence analysts generally create dashboards and reports and have knowledge of the data feeds and sources.

- **Database Administrator (DBA):** Provisions and configures the database environment to support the analytics needs of the working team. These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data Repositories.
- **Data Engineer:** Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox. The data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics. The data engineer works closely with the data scientist to help shape data in the right ways for analyses.

- **Data Scientist:** Provides subject matter expertise for analytical techniques, data modeling, and applying valid analytical techniques to given business problems. Ensures overall analytics objectives are met. Designs and executes analytical methods and approaches with the data available to the project.

2. What are the five main activities performed during *'identifying potential data sources sub-phase* under Discovery phase?

Ans) The team should perform five main activities during this step of the discovery phase:

o **Identify data sources:** Make a list of candidate data sources the team may need to test the initial hypotheses outlined in this phase. Make an inventory of the datasets currently available and those

that can be purchased or otherwise acquired for the tests the team wants to perform.
o **Capture aggregate data sources:** This is for previewing the data and providing high-level understanding. It enables the team to gain a quick overview of the data and perform further exploration on specific areas. It also points the team to possible areas of interest within the data.
o **Review the raw data:** Obtain preliminary data from initial data feeds. Begin understanding the interdependencies among the data attributes, and become familiar with the content of the data, its quality, and its limitations.
o **Evaluate the data structures and tools needed:** The data type and structure dictate which tools the team can use it to analyze the data. This evaluation gets the team thinking about which technologies may be good candidates for the project and how to start getting access to these tools.
• **Scope the sort of data infrastructure needed for this type of problem:** In addition to the tools needed, the data influences the kind of infrastructure that's required, such as disk storage and network Capacity.

For example,

Let's say a team is working on a project to optimize the production process for a manufacturing plant. They want to use data analytics to identify bottlenecks in the production line, and develop strategies to increase efficiency.

Here's how they might approach the discovery phase using the five main activities:

1. **Identify data sources:** The team may start by identifying candidate data sources such as production logs, equipment sensors, and maintenance records.
2. **Capture aggregate data sources:** The team might use visualization tools to create charts and graphs of production rates, equipment downtime, and maintenance schedules.
3. **Review the raw data:** They might examine equipment sensor data to identify patterns in temperature, pressure, and vibration.
4. **Evaluate the data structures and tools needed:** They might use statistical analysis to identify correlations between equipment performance and production rates.
5. **Scope the data infrastructure needed:** This might include investing in additional disk storage, upgrading network capacity to handle large volumes of data, or using cloud-based services to scale up as needed.

3. What is Data Conditioning?

Ans)

Data conditioning refers to the process of cleaning data, normalizing datasets, and performing transformations on the data. A critical step within the Data Analytics Lifecycle, data conditioning can involve many complex steps to join or merge data sets or otherwise get datasets into a state that enables analysis in further phases. It is viewed as a preprocessing step. It involves many operations on the dataset before developing models to process or analyze the data. The data-conditioning step is performed only by IT, the data owners, a DBA, or a data engineer. It is also important to involve the data scientist in this step because many decisions are made in the data conditioning phase that affects the subsequent analysis.

Suppose you have a dataset containing customer information, including their names, ages, addresses, and purchase history. Here's how data conditioning can be applied to this dataset:

1. Data Cleaning: For example, you may remove duplicate entries, correct misspelled names, and fill in missing age values with reasonable estimates.
2. Data Normalization: For instance, you may convert all addresses to a standardized format or transform age values into a standardized range, such as normalizing ages between 0 and 100.
3. Data Transformation: Depending on the analysis goals, you may need to transform certain variables. For example, you might convert categorical variables like purchase history into numerical values or apply mathematical transformations, such as taking the logarithm of a variable, to achieve a better distribution for analysis purposes.
4. Data Integration: For instance, you could merge customer data with sales data from different stores or combine it with demographic data to gain additional insights.
5. Feature Engineering: For example, you might derive a new feature like "total purchases" by summing up the purchase amounts from the purchase history.

By performing these data conditioning steps, the dataset becomes cleaner, consistent, and well-prepared for analysis.

4. In which phase would the team expect to invest most of the project time? Why?

Ans)Data prep part if for 5/10 marks

Understanding the data in detail is critical to the success of the project. The team also must decide how to condition and transform data to get it into a format to facilitate subsequent analysis. The team may perform data visualizations to help team members understand the data, including its trends, outliers,etc. Each of these are steps of the data preparation phase. Data preparation tends to be the most labor-intensive step in the analytics lifecycle. So, it is common for teams to spend at least 50% of a data science project's time in this critical phase.

5. What are the common questions that are helpful to ask during the Discovery phase when interviewing the project sponsor?

Ans) Following is a brief list of common questions that are helpful to ask during the discovery phase when interviewing the project sponsor. The responses will begin to shape the scope of the project and give the team an idea of the goals and objectives of the project.

- What business problem is the team trying to solve?
- What is the desired outcome of the project?
- What data sources are available?
- What industry issues may impact the analysis?
- What timelines need to be considered?
- Who could provide insight into the project?
- Who has final decision-making authority on the project?
- How will the focus and scope of the problem change if the following dimensions change:
- Time: Analyzing 1 year or 10 years' worth of data?
- People: Assess impact of changes in resources on project timeline.
- Risk: Conservative to aggressive
- Resources: None to unlimited (tools, technology, systems)
- Size and attributes of data: Including internal and external data sources

6. Mention the list of activities in Phase 1. (loa)

Ans) **L**ong **F**ruity **I**s **I**nterviewing **D**eveloper in **I**ndia

1. **Learning the Business Domain:** This activity involves researching and understanding the industry, market, and specific business domain that the data analytics project will address. The goal is to gain a comprehensive understanding of the context in which the data is being collected and analyzed.
2. **Framing the Problem:** This activity involves clearly defining the business problem or question that the data analytics project aims to solve or answer. This involves defining the scope of the project, identifying key metrics that will be used to measure success, and setting expectations.
3. **Identifying Key Stakeholders:** This activity involves identifying the individuals or groups who will be impacted by the project's outcomes, and ensuring that their needs and expectations are taken into account throughout the project. This includes defining roles and responsibilities for stakeholders, and establishing communication channels.
4. **Interviewing the Analytics Sponsor:** This activity involves meeting with the person responsible for overseeing the project, to understand their goals, expectations, and available resources. This helps to align the project team with the sponsor's expectations, and ensures that the project is feasible and aligned with the organization's strategic objectives.
5. **Developing Initial Hypotheses:** This activity involves developing initial hypotheses or assumptions about the business problem or question that the data analytics project aims to solve or answer. This helps to guide the analysis and ensure that the team is focusing on the most relevant factors.
6. **Identifying Potential Data Sources:** This activity involves identifying the sources of data that will be used to test the hypotheses or answer the business question. This includes understanding the quality and reliability of the data, as well as any potential limitations or biases.

7. How to prepare the Analytic Sandbox?

Ans) Data preparation definition and importance

The first subphase of data preparation requires the team to obtain an analytic sandbox (also commonly referred to as a workspace), in which the team can explore the data without interfering with live production databases. Consider an **example** in which the team needs to work with a company's financial data. The team should access a copy of the financial data from the analytic sandbox rather than interacting with the production version of the organization's main database, because that will be tightly controlled and needed for financial reporting.

When developing the analytic sandbox, it is a best practice to collect all kinds of data there, as team members need access to high volumes and varieties of data for a Big Data analytics project. This can include everything from summary-level aggregated data, structured data, raw data feeds, and unstructured text data from call logs or web logs, depending on the kind of analysis the team plans to undertake. This expansive approach for attracting data of all kind differs considerably from the approach advocated by many information technology (IT) organizations. Because of these differing views on data access and use, it is critical for the data science team to collaborate with IT, make clear what it is trying to accomplish, and align goals. The analytic sandbox enables organizations to undertake more ambitious data science projects and move beyond doing traditional data analysis and Business Intelligence to perform more robust and advanced predictive analytics.

8. List of questions to consider after phase 4.

Ans) Explain the below questions for 10 marks..

Creating robust models that are suitable to a specific situation requires thoughtful consideration to ensure the models being developed ultimately meet the objectives outlined in Phase 1

Questions to consider include these:

• Does the model appear valid and accurate on the test data?

• Does the model output/behavior make sense to the domain experts? That is, does it appear as if the model is giving answers that make sense in this context?

• Do the parameter values of the fitted model make sense in the context of the domain?

• Is the model sufficiently accurate to meet the goal?

• Does the model avoid intolerable mistakes? Depending on context, false positives may be more serious or less serious than false negatives, for instance.

• Are more data or more inputs needed? Do any of the inputs need to be transformed or eliminated?

• Will the kind of model chosen support the runtime requirements?

• Is a different form of the model required to address the business problem? If so, go back to the model planning phase and revise the modeling approach.

1. Give a brief overview of the main phases of the Data Analytics Lifecycle (with diagram)

Ans)

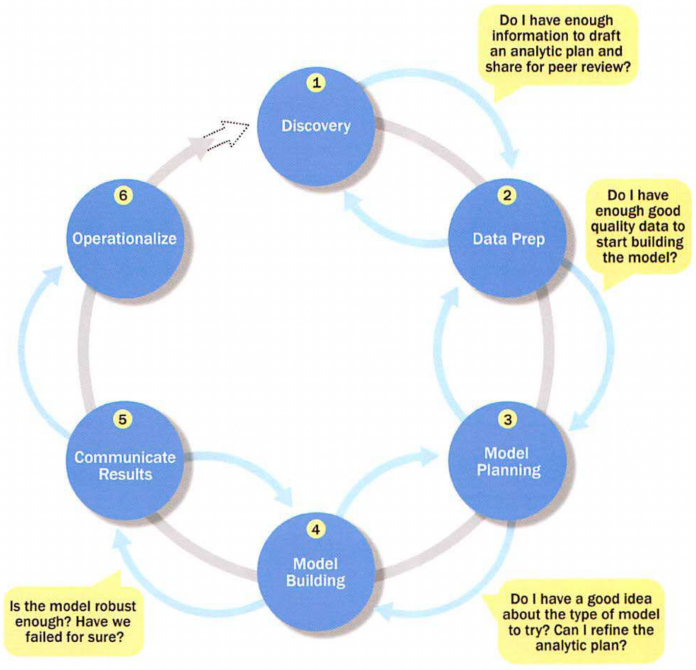


FIGURE 2-2 *Overview of Data Analytics Lifecycle*

Here is a brief overview of the main phases of the Data Analytics Lifecycle:

• **Phase 1- Discovery:** In Phase 1, the team learns the business domain, including relevant history

such as whether the organization or business unit has attempted similar projects in the past from which they can learn. The team assesses the resources available to support the project in terms of people, technology, time, and data. Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.

• **Phase 2- Data preparation:** Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT. Data should be transformed in the ETLT process so the team can work with it and analyze it. In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data.

• **Phase 3-Model planning:** Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

• **Phase 4-Model building:** In Phase 4, the team develops datasets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing, if applicable).

• **Phase 5-Communicate results:** In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

• **Phase 6-Operationalize:** In Phase 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production Environment.

Once team members have run models and produced findings, it is critical to frame these results in a way that is tailored to the audience that engaged the team. Moreover, it is critical to frame the results of the work in a manner that demonstrates clear value. If the team performs a technically accurate analysis but fails to translate the results into a language that resonates with the audience, people will not see the value, and much of the time and effort on the project will have been wasted.

2. What are the common tools used in Phase 2, 3 and 4?

Ans)

Common Tools for the Data Preparation Phase (Phase 2) (HODA)

Several tools are commonly used for this phase:

• **Hadoop:** can perform massively parallel ingest and custom analysis for web traffic parsing, GPS location analytics, genomic analysis, and combining of massive unstructured data feeds from multiple sources.

• **Alpine Miner :** provides a graphical user interface (GUI) for creating analytic workflows, including data manipulations and a series of analytic events such as staged data-mining techniques on Postgresql and other Big Data sources.

• **Open Refine (formerly called Google Refine):** It is "a free, open source, powerful tool for working with messy data." It is a popular GUI-based tool for performing data transformations, and it's one of the most robust free tools currently available.

• **Data Wrangler:** It is an interactive tool for data cleaning and transformation. It can be used to perform many transformations on a given dataset.

Common Tools for the Model Planning Phase (Phase 3)

Many tools are available to assist in this phase. Here are several of the more common ones:

o **R:** has a complete set of modeling capabilities and provides a good environment for building interpretive models with high-quality code.

o **SQL Analysis services:** can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models.

• **SAS/ACCESS :** provides integration between SAS and the analytics sandbox via multiple data connectors such as OBDC, JDBC, etc.

Common Tools for the Model Building Phase (Phase 4)

There are many tools available to assist in this phase, focused primarily on statistical analysis or data mining
software. Common tools in this space include, but are not limited to, the following:

• **Commercial Tools:**

• **SAS Enterprise Miner**: allows users to run predictive and descriptive models based on large volumes of data from across the enterprise. It interoperates with other large data stores, has many partnerships, and is built for enterprise-level computing and analytics.

• **SPSS Modeler:** offers methods to explore and analyze data through a GUI.

• **Matlab:** provides a high-level language for performing a variety of data analytics, algorithms, and data exploration.

• **Free or Open Source tools:**

• **Rand PL/R [14)** R was described earlier in the model planning phase, and PL!R is a procedural language for PostgreSQL with R. Using this approach means that R commands can be executed in database. This technique

provides higher performance and is more scalable than running R in memory.
• **Octave [22),** a free software programming language for computational modeling, has some of the functionality of Matlab. Because it is freely available, Octave is used in major universities when teaching machine learning.
• **WEKA [23)** is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.
• **Python** is a programming language that provides toolkits for machine learning and analysis, such as scikit-learn, numpy, scipy, pandas, and related data visualization using matplotlib, seaborn,etc.
• **SQL** in-database implementations, such as MADlib. Provide an alternative to in-memory desktop analytical tools.
MADlib provides an open-source machine learning library of algorithms that can be executed in-database, for PostgreSQL or Greenplum.

3. Explain: Why Phase 5 – Communicate Results is critical.
Ans)
Phase 5, "Communicate Results", is critical in the data analytics lifecycle because it involves presenting the analysis results to stakeholders in a way that is easy to understand and actionable. Effective communication of the analysis results is essential for making informed decisions based on the insights gained from the data.

Here are some reasons why Phase 5 is critical:
Making decisions: The primary objective of data analytics is to help stakeholders make informed decisions. Without effective communication of the analysis results, stakeholders may not be able to understand the insights gained from the data or how they can be used to inform decisions.

Trust and credibility: Effective communication of the analysis results builds trust and credibility with stakeholders. If the analysis results are not communicated effectively, stakeholders may question the validity of the analysis or the expertise of the analytics team.

Actionable insights: Effective communication of the analysis results ensures that stakeholders can understand the insights gained from the data and how they can be used to inform decisions. If the analysis results are not communicated effectively, stakeholders may not be able to take action based on the insights gained from the data.

Continued engagement: Effective communication of the analysis results keeps stakeholders engaged in the data analytics process. By presenting the results in an understandable and actionable way, stakeholders are more likely to continue to support the data analytics efforts and invest in future projects.

Eg: If accurate insights are communicated → Stakeholders will work on their limitations and will gain profits.
If accurate insights are not communicated → Stakeholders will work on other things and will not gain profits.

4. What are the benefits of doing a pilot program before a full-scale rollout of a new analytical methodology?

Ans)

A pilot program is a small-scale implementation of a new analytical methodology, typically involving a limited set of data or a specific subset of users. The purpose of a pilot program is to test and refine the methodology in a controlled environment, before deploying it on a larger scale.

There are several benefits to doing a pilot program before a full-scale rollout of a new analytical methodology. Here are some of the most important ones:

Identify and resolve issues: A pilot program provides an opportunity to identify and resolve issues before the methodology is implemented on a larger scale. This can help to prevent potential problems that could negatively impact the effectiveness of the methodology and the accuracy of its results.

Test the methodology in a controlled environment: A pilot program allows you to test the methodology in a controlled environment, which can help to identify areas where improvements can be made. This can help to ensure that the methodology is effective and reliable when implemented on a larger scale.

Evaluate the feasibility of the methodology: A pilot program provides an opportunity to evaluate the feasibility of the methodology in terms of resources, costs, and time. This can help to identify any limitations or constraints that may need to be addressed before the methodology is implemented on a larger scale.

Gather feedback from stakeholders: A pilot program provides an opportunity to gather feedback from stakeholders, including users, management, and other stakeholders. This feedback can be used to improve the methodology and ensure that it meets the needs of all stakeholders.

Build support and buy-in: A successful pilot program can build support and buy-in for the methodology, making it easier to implement on a larger scale. This can help to ensure that the methodology is widely adopted and effectively used throughout the organization.

Overall, a pilot program can help to ensure that analytical methodology is effective, reliable, and feasible before it is implemented on a larger scale. This can help to reduce risks and increase the chances of success when rolling out the methodology more broadly.

5. Explain the sub-phase – Performing ETLT under Data Preparation phase.

Ans)

- In ETL, users perform extract, transform, load processes to extract data from a datastore, perform data transformations, and load the data back into the datastore.
- However, the analytic sandbox approach differs slightly; it advocates extract, load, and then transform. ln this case, the data is extracted in its raw form and loaded into the datastore, where analysts can choose to

transform the data into a new state or leave it in its original, raw condition.

- The reason for this approach is that there is significant value in preserving the raw data and including it in the sandbox before any transformations take place.
- Following the ELT approach gives the team access to clean data to analyze after the data has been loaded into the database and gives access to the data in its original form for finding hidden nuances in the data.
- This approach is part of the reason that the analytic sandbox can quickly grow large. The team may want clean data and aggregated data and may need to keep a copy of the original data to compare against or look for hidden patterns that may have existed in the data before the cleaning stage.
- This process can be summarized as ETLT to reflect the fact that a team may choose to perform ETL in one case and ELT in another.

MOD - 2 Regression Analysis

9. Differentiate Correlation and Regression analysis

Ans)

	CORRELATION	REGRESSION ANALYSIS
Purpose	Measures the strength and direction of the linear relationship between two variables	Models and analyzes the relationship between a dependent variable and one or more independent variables
Focus	Association between variables	Prediction and understanding causal relationships
Variables	Two or more variables, typically continuous	Dependent variable (response variable) and one or more independent variables (predictor variables)
Causality	Does not assume causality	Assumes causality: independent variables are believed to have a causal impact on the dependent variable
Coefficients	Calculates correlation coefficient (r)	Estimates coefficients to quantify the impact of independent variables
Range	Correlation coefficient ranges from -1 to +1	Coefficients can be any real number
Interpretation	Measures strength and direction of linear relationship	Describes how changes in independent variables relate to changes in the dependent variable
Prediction	Does not make predictions or infer	Generates predictive models to estimate or predict

	values	the dependent variable
Analysis type	Descriptive analysis	Inferential analysis
Assumptions	No assumptions about variable roles	Assumes linearity, independence, normality, homoscedasticity, etc.

Eg: Regression :- If student studies more no of hours → Scores good marks (i.e marks dependent on no of hours)

Correlation:- There exists a positive relationship between the sales of ice cream and the climate temperature. This implies that the sales of ice cream are higher in hotter weather conditions. Obviously one tends to crave the ice cream in summers more than in winters.

Extra points:

- Regression establishes how *x* causes *y* to change, and the results will change if *x* and *y* are swapped. With correlation, *x* and *y* are variables that can be interchanged and get the same result.
- Correlation is a single statistic, or data point, whereas regression is the entire equation with all of the data points that are represented with a line.
- Correlation shows the relationship between the two variables, while regression allows us to see how one affects the other.
- The data shown with regression establishes a cause and effect, when one changes, so does the other, and not always in the same direction. With correlation, the variables move together.

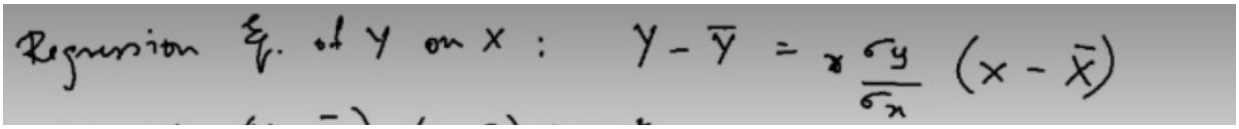
Extra question

Definition of simple linear regression

Ans) Simple linear regression provides a model of the relationship between the magnitude of one variable and that of a second—for example, as X increases, Y also increases. Or as X increases, Y decreases. Regression quantifies the nature of the relationship.

10. What is the regression equation of Y on X

Ans)



Where,

$y - \bar{y}$ = standard deviation from mean of y

$r \frac{\sigma_y}{\sigma_x}$ = regression coefficient of y on x

R = pearson's coefficient

Sigma x = SD of x

11. Explain the concepts: Fitted Values and Residuals.

Ans)

A fitted value is a statistical model's prediction of the mean response value when you input the values of the predictors, factor levels, or components into the model. Suppose you have the following regression equation: y = 3X + 5. If you enter a value of 5 for the predictor, the fitted value is 20. Fitted values are also called predicted values.

The *fitted* values, also referred to as the *predicted* values, are typically denoted by $\hat{Y}_i$ (Y-hat). These are given by:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

Residuals in a statistical or machine learning model are the differences between observed and predicted values of data. They are a diagnostic measure used when assessing the quality of a model. They are also known as errors.

We compute the residuals $\hat{e}_i$ by subtracting the *predicted* values from the original data:

$$\hat{e}_i = Y_i - \hat{Y}_i$$

Least Square:

The regression line is the estimate that minimizes the sum of squared residual values, also called the residual sum of squares or RSS:

$$RSS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$= \sum_{i=1}^{n} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2$$

The method of minimizing the sum of the squared residuals is termed least squares regression, or ordinary least squares (OLS) regression. Least squares regression can be computed quickly and easily with any standard statistical software. Computational convenience is one reason for the widespread use of least squares in regression. With the advent of big data, computational speed is still an important factor. Least squares, like the mean, are sensitive to outliers, although this tends to be a significant problem only in small or moderate-sized data sets.

Q. Multiple Linear Regression

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. You can use multiple linear regression when you want to know:

1. How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
2. The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

When there are multiple predictors, the equation is simply extended to accommodate them:
$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p + e$

Instead of a line, we now have a linear model. The relationship between each coeffi-cient and its variable (feature) is linear.

Q. Assessing the model

The model assessment phase starts when we create a holdout set which consists of examples the learning algorithm didn't see during training. If our model performs well on the holdout set we can say that our model generalizes well and is of good quality. The most common way to assess whether a model is good or not is to compute a performance metric on the holdout data.

The most important performance metric from a data science perspective is root mean squared error, or RMSE. RMSE is the square root of the average squared error in the predicted yi values:

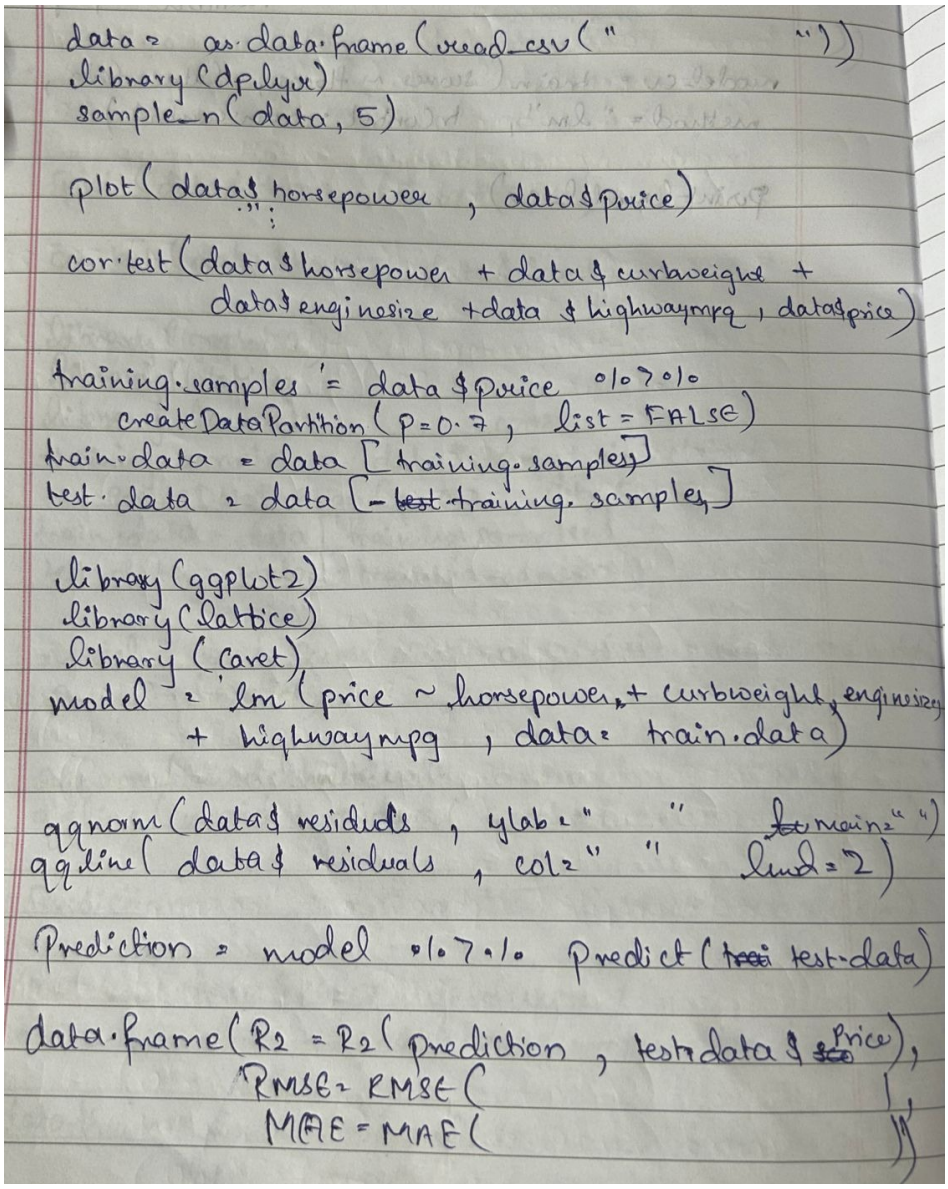
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n}}$$

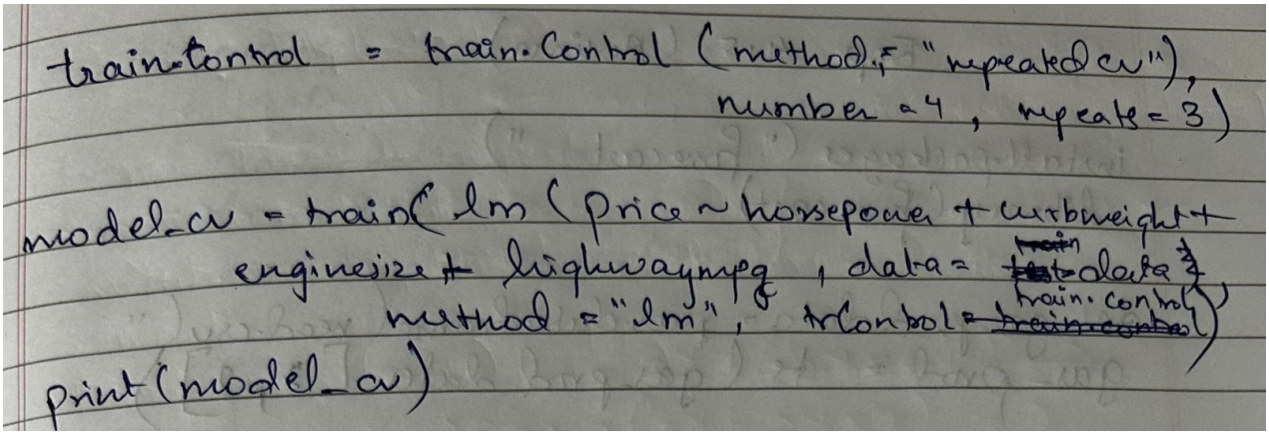
This measures the overall accuracy of the model and is a basis for comparing it to

other models (including models fit using machine learning techniques). Similar to RMSE is the residual standard error, or RSE. In this case we have p predictors, and the RSE is given by:

$$RSE = \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-p-1)}}$$

The only difference is that the denominator is the degrees of freedom, as opposed to number of records.In practice, for linear
regression, the difference between RMSE and RSE is very small, particularly for big data applications.
The summary function in R computes RSE as well as other metrics for a regression Model:





Q.Cross Validation

Classic statistical regression metrics (R2, F-statistics, and p-values) are all “in-sample” metrics—they are applied to the same data that was used to fit the model. Intuitively, you can see that it would make a lot of sense to set aside some of the original data, not use it to fit the model, and then apply the model to the set-aside (holdout) data to see how well it does.

Cross-validation extends the idea of a holdout sample to multiple sequential holdout samples.

The algorithm for basic k-fold cross-validation is as follows:

1. Set aside 1/k of the data as a holdout sample.
2. Train the model on the remaining data.
3. Apply (score) the model to the 1/k holdout, and record needed model assessment metrics.
4. Restore the first 1/k of the data, and set aside the next 1/k (excluding any records that got picked the first time).
5. Repeat steps 2 and 3.
6. Repeat until each record has been used in the holdout portion.
7. Average or otherwise combine the model assessment metrics.

The division of the data into the training sample and the holdout sample is also called a fold.

Q. Model Selection and Stepwise Regression

Model Selection and Stepwise Regression are two approaches used in regression analysis for selecting the most appropriate variables to include in a predictive model. However, they differ in their methodology and complexity. Here are the key differences:

Model Selection:

Model selection involves evaluating and comparing multiple models with different sets of variables to identify the best-fitting model.
It considers various criteria, such as goodness-of-fit measures (e.g., R-squared, adjusted R-squared), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or cross-validation techniques.
Model selection techniques include forward selection, backward elimination, and stepwise selection.
Model selection is often performed by evaluating a subset of possible models and selecting the one that achieves the best balance between model complexity and predictive accuracy.
It may require domain expertise and careful consideration of the research question, as it involves manually deciding which variables to include or exclude.

Stepwise Regression:
Stepwise regression is a specific approach to model selection that automatically selects variables based on a combination of forward selection and backward elimination.
It starts with an initial model (empty or with a few variables) and iteratively adds or removes variables based on statistical criteria, usually using p-values or a predetermined significance level.
Stepwise regression typically involves two steps: forward selection (adding variables one by one) and backward elimination (removing non-significant variables).
The algorithm stops when certain stopping criteria are met (e.g., a variable fails to meet the significance threshold or all remaining variables have significant p-values).
Stepwise regression can be performed in a forward or backward direction, resulting in different models.
It is an automated procedure that aims to identify a subset of variables that best explain the variation in the dependent variable while balancing model complexity.

Q. Prediction using Regression
Prediction using regression refers to the process of using a regression model to estimate or forecast values of the dependent variable based on the values of the independent variables. Regression analysis is a statistical technique used to model the relationship between a dependent variable and one or more independent variables, and it can be utilized for predictive purposes.

Q.Logistic response function and logit
In statistics, the logistic response function (also known as the sigmoid function) is a mathematical function that maps any input value to a value between 0 and 1. The function is typically used to model binary outcomes or probabilities, such as whether a customer will buy a product or not.
The logistic response function is defined as follows:
$f(x) = 1 / (1 + e^{(-x)})$
where x is the input variable and e is the mathematical constant approximately equal to 2.71828.
The function has an S-shaped curve and is symmetrical around the point where x equals zero. The output of the function approaches 0 as x approaches negative infinity, and approaches 1 as x approaches positive infinity. The inflection point of the function is at x = 0. At this point, the output of the function is 0.5.
The inverse of the logistic response function is the logit function, which is used to transform probabilities or proportions to a linear scale. The logit function is defined as:

logit(p) = ln(p / (1 - p))
where p is the probability or proportion of a binary outcome.
The logit function maps probabilities between 0 and 1 to values between negative infinity and positive infinity. A probability of 0 maps to negative infinity, a probability of 0.5 maps to 0, and a probability of 1 maps to positive infinity. The logit function is used in logistic regression models to estimate the relationship between predictor variables and a binary outcome variable.

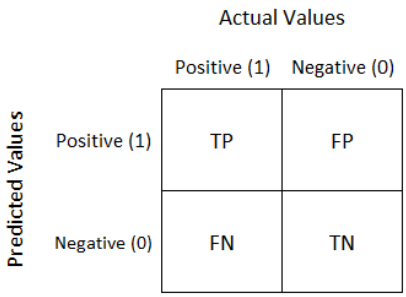
Q. Generalized Linear model

A Generalized Linear Model (GLM) is a flexible and powerful statistical framework used to model the relationship between a response variable and one or more predictor variables. GLMs extend the linear regression model to handle a wide range of response variables that may not necessarily follow a normal distribution, allowing for greater flexibility in modeling various types of data.

1. Explain any two metrics that measure the overall accuracy of the model. 2
Ans)
Accuracy: Accuracy is a basic metric that measures the overall correctness of a model's predictions. It is calculated as the ratio of the number of correct predictions to the total number of predictions made by the model. Mathematically, it can be expressed as:

$$\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Confusion
Matrix: A matrix that summarizes the number of true positives, false positives, true negatives, and false negatives. Confusion matrix is used to calculate other performance measures such as accuracy, precision, recall, and F1-score.



2. Explain lower t-statistics indicate a predictor should be dropped.
Ans)
In a linear regression model, each predictor variable has an associated coefficient estimate and a corresponding t-statistic. The t-statistic is calculated by dividing the estimated coefficient by its standard error. It measures the number of standard errors that the coefficient estimate is away from zero. A higher absolute value of the t-statistic indicates a more significant predictor variable, while a lower absolute value suggests a less significant predictor variable.

A common practice is to use a threshold, such as a significance level (e.g., 0.05 or 0.01), to determine the statistical significance of a predictor. If the absolute value of the t-statistic for a predictor is lower than the threshold, it may indicate that the predictor is not statistically significant in explaining the variation in the dependent variable. In such cases, it may be considered for removal from the model to simplify the model and potentially improve its interpretability and predictive accuracy.

3. What is logistic regression?

Ans) Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted or not to a particular college. These binary outcomes allow straightforward decisions between two alternatives.

A logistic regression model can take into consideration multiple input criteria. In the case of college acceptance, the logistic function could consider factors such as the student's grade point average, SAT score and number of extracurricular activities. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into one of two outcome categories.

4. Explain how logistic regression differs from linear regression. 2

Ans)

similarities:

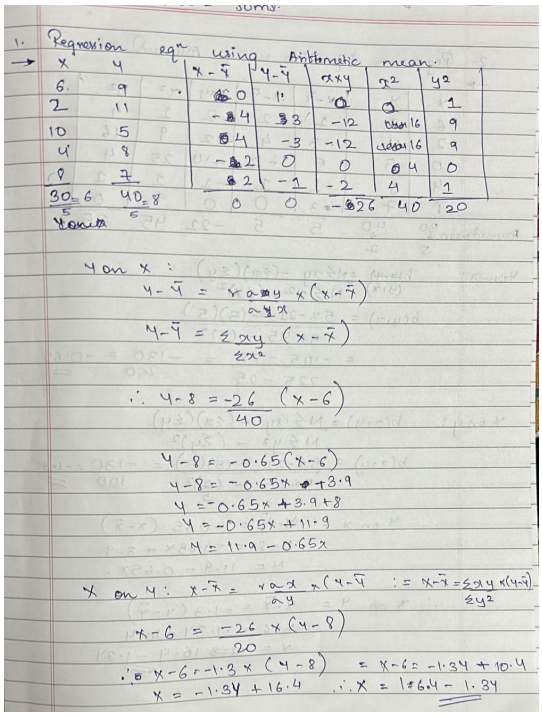
Linear regression and logistic regression are both widely used statistical techniques for modeling and analyzing relationships between variables.

Similarities:

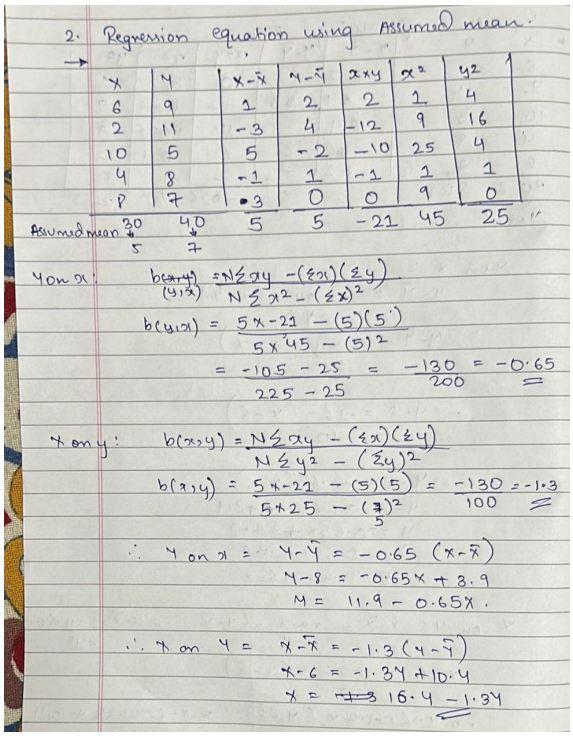
1. Both are types of regression analysis techniques used for modeling the relationship between a dependent variable and one or more independent variables.
2. Both techniques aim to understand the relationship and predict the value of the dependent variable based on the values of the independent variables.
3. Both regression models involve estimating regression coefficients that quantify the relationship between the variables.

Linear Regression	Logistic Regression
Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear Regression is used for solving Regression problem.	Logistic regression is used for solving Classification problems.
In Linear regression, we predict the value of continuous variables.	In logistic Regression, we predict the values of categorical variables.
In linear regression, we find the best fit line, by which we can easily predict the output.	In Logistic Regression, we find the S-curve by which we can classify the samples.
Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.
The output for Linear Regression must be a continuous value, such as price, age, etc.	The output of Logistic Regression must be a Categorical value such as 0 or 1, Yes or No, etc.
In Linear regression, it is required that relationship between dependent variable and independent variable must be linear.	In Logistic regression, it is not required to have the linear relationship between the dependent and independent variable.
In linear regression, there may be collinearity between the independent variables.	In logistic regression, there should not be collinearity between the independent variable.

6. Compute the regression equations for the given data using the Arithmetic mean.



7. Compute the regression equations for the given data using the Assumed mean.



2. Regression equation using Assumed mean.

→

X	Y	$x-\bar{x}$	$y-\bar{y}$	xy	x^2	y^2
6	9	1	2	2	1	4
2	11	-3	4	-12	9	16
10	5	5	-2	-10	25	4
4	8	-1	1	-1	1	1
P	7	3	0	0	9	0
Assumed mean	30/5	40/7	5	5	-21	45	25

Y on X:

$$b_{(y,x)} = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{N\Sigma x^2 - (\Sigma x)^2}$$

$$b_{(y,x)} = \frac{5(-21) - (5)(5)}{5(45) - (5)^2}$$

$$= \frac{-105 - 25}{225 - 25} = \frac{-130}{200} = -0.65$$

X on Y:

$$b_{(x,y)} = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{N\Sigma y^2 - (\Sigma y)^2}$$

$$b_{(x,y)} = \frac{5(-21) - (5)(5)}{5(25) - (7)^2} = \frac{-130}{100} = -1.3$$

$\therefore$ Y on X: $y - \bar{y} = -0.65(x - \bar{x})$

$y - 8 = -0.65x + 8.9$

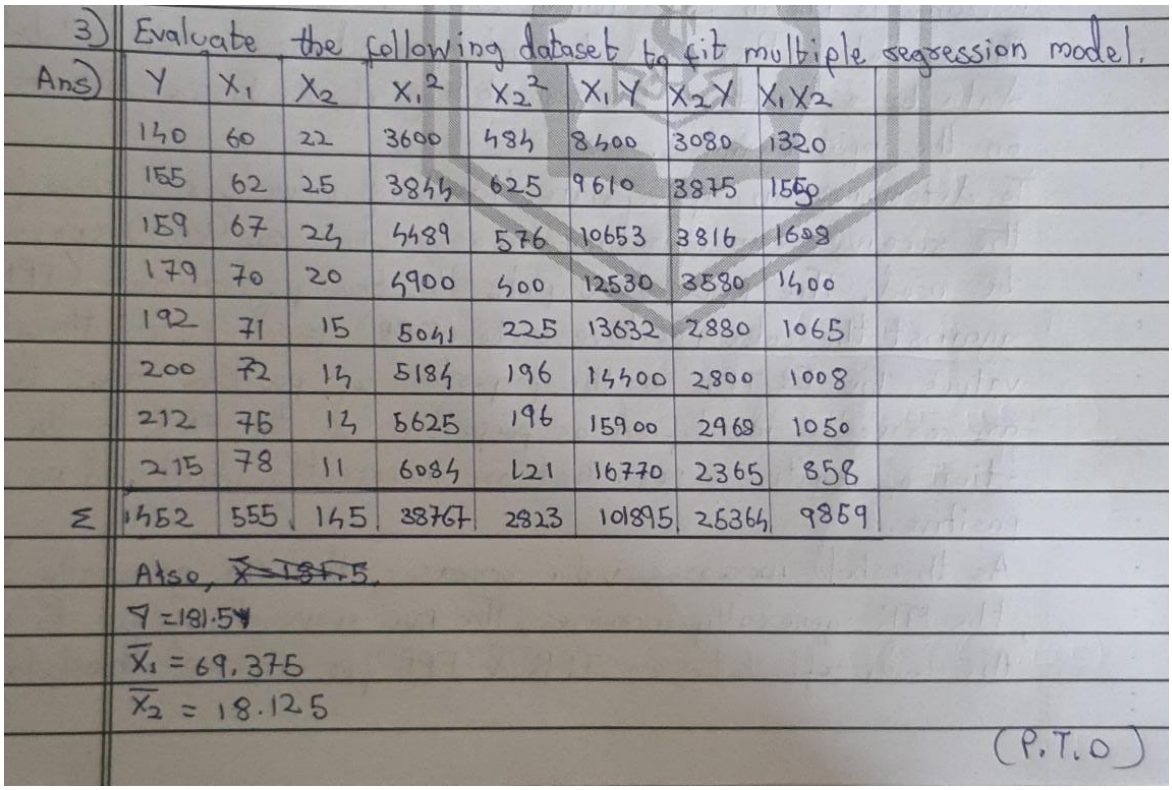
$y = 11.9 - 0.65x$.

$\therefore$ X on Y: $x - \bar{x} = -1.3(y - \bar{y})$

$x - 6 = -1.3y + 10.4$

$x = 16.4 - 1.3y$

8. Multiple Regression Sum



3) Evaluate the following dataset to fit multiple regression model.

Ans)

Y	X_1	X_2	X_1^2	X_2^2	X_1Y	X_2Y	X_1X_2
140	60	22	3600	484	8400	3080	1320
155	62	25	3844	625	9610	3875	1550
159	67	24	4489	576	10653	3816	1608
179	70	20	4900	400	12530	3580	1400
192	71	15	5041	225	13632	2880	1065
200	72	14	5184	196	14400	2800	1008
212	75	14	5625	196	15900	2968	1050
215	78	11	6084	121	16770	2365	858
Σ 1452	555	145	38767	2823	101895	25364	9859

Also, ~~$\bar{X} = 181.5$~~.

$\bar{Y} = 181.5$

$\bar{X_1} = 69.375$

$\bar{X_2} = 18.125$

(P.T.O)

Now,

$$\Sigma x_1^2 = \Sigma X_1^2 - \frac{(\Sigma X_1)(\Sigma X_1)}{N} = 38767 - \frac{(555)^2}{8} = 263.875$$

$$\Sigma x_2^2 = \Sigma X_2^2 - \frac{(\Sigma X_2)^2}{N} = 2823 - \frac{(145)^2}{8} = 194.875$$

$$\Sigma x_1 y = \Sigma X_1 Y - \frac{(\Sigma X_1)(\Sigma Y)}{N} = 101895 - \frac{(555)(1452)}{8} = 1162.5$$

$$\Sigma x_2 y = \Sigma X_2 Y - \frac{(\Sigma X_2)(\Sigma Y)}{N} = 25364 - \frac{(145)(1452)}{8} = -953.5$$

$$\Sigma x_1 x_2 = \Sigma X_1 X_2 - \frac{(\Sigma X_1)(\Sigma X_2)}{N} = 9859 - \frac{(555)(145)}{8} = -200.375$$

We know that,

$$b_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$\therefore b_1 = \frac{(194.875)(1162.5) - (-200.375)(-953.5)}{(263.875)(194.875) - (-200.375)^2}$$

$$\therefore b_1 = 3.1478 \approx 3.148$$

$$b_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$\therefore b_2 = \frac{(263.875)(-953.5) - (-200.375)(1162.5)}{(263.875)(194.875) - (-200.375)^2}$$

$$\therefore b_2 = -1.656$$

Now, since -

$$a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

$$\therefore a = 181.5 - (3.148)(69.375) - (-1.656)(18.125)$$

$$\therefore a = -6.87$$

(P.T.O)

$\therefore$ Regression eq^n will be -

$$Y = a + b_1 X_1 + b_2 X_2$$

$$\therefore Y = -6.87 + 3.148 X_1 - 1.656 X_2 \; //$$

Module 3 – Time Series Analysis

1. What are the components of time series?

Ans)

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. Data collected irregularly or only once are not time series. The components of the time series are as follows:

1. The **trend** refers to the long-term movement in a time series. It indicates whether the observation values are increasing or decreasing over time. Examples of trends are a steady increase in sales month over month or an annual decline in fatalities due to car accidents.
2. The **seasonality** component describes the fixed, periodic fluctuation in the observations over time. As the name suggests, the seasonality component is often related to the calendar. For example, monthly retail sales can fluctuate over the year due to the weather and holidays.
3. A **cyclic** component also refers to a periodic fluctuation, but one that is not as fixed as in the case of a seasonality component. For example, retail sales are influenced by the general state of the economy. Thus, a retail sales time series can often follow the lengthy boom-bust cycles of the economy.
4. There is another kind of movement that can be seen in the case of time series. It is pure **Irregular and Random** Movement. As the name suggests, no hypothesis or trend can be used to suggest irregular or random movements in a time series. These outcomes are unforeseen, erratic, unpredictable, and uncontrollable in nature. Eg COVID

2. Write the steps to perform box-jenkins methodology. 2

Ans)

This approach starts with the assumption that the process that generated the time series can be approximated using an ARMA model if it is stationary or an ARIMA model if it is non-stationary.

Box-Jenkins methodology is an iterative approach that consists of the following 3 steps:

1. **Identification:** Use the data and all related information to help select a subclass of model that may best summarize the data.
2. **Estimation:** Use the data to train the parameters of the model (i.e. the coefficients).
3. **Diagnostic Checking:** Evaluate the fitted model in the context of the available data and check for areas where the model may be improved.

It is an iterative process, so that as new information is gained during diagnostics, you can circle back to step 1 and incorporate that into new model classes.

3. What is the result of the absolute value of ACF(h), when it is closer to 1?

Ans) The difficulty is that the plot does not provide insight into the covariance of the variables in the time series and its underlying structure. The plot of the autocorrelation function (ACF) provides this insight. For a stationary time series, the ACF is defined as shown,

$$ACF(h) = \frac{cov(y_t, y_{t+h})}{\sqrt{cov(y_t, y_t)\, cov(y_{t+h}, y_{t+h})}} = \frac{cov(h)}{cov(0)}$$

By convention, the quantity h in the ACF is referred to as the lag, the difference between the time points t and t +h. At lag 0, the ACF provides the correlation of every point with itself. So ACF(O) is always equals 1 and it goes on decreasing as the lag increases.

When ACF(h) is closer to 1, it suggests a strong linear relationship between the values of the time series at the current time step and the values at the lag h time step. This indicates that the values of the time series are highly correlated with their past values at the specific lag h.

A value of 1 indicates a perfect positive autocorrelation, -1 indicates a perfect negative autocorrelation, and 0 indicates no autocorrelation.

4. **What are the three conditions for stationary time series?**
Ans)
As stated in the first step of the Box-Jenkins methodology, it is necessary to remove any trends or seasonality in the time series. This step is necessary to achieve a time series with certain properties to which autoregressive and moving average models can be applied. Such a time series is known as a stationary time series. A time series, Y_t for t = 1,2,3, ...,, is a stationary time series if the following three conditions are met:

(a) The expected value (mean) of Y_t is constant for all values of t.

(b) The variance of Y_t is finite.

(c)The covariance of Y_t and Y_{t+h}, depends only on the value of h = 0,1,2,... for all t.

1. Explain the application of time series in the following sectors. Finance, Economic, Engineering, Retail and Manufacturing.
Ans)
Finance: Time series analysis has become an intrinsic part of financial analysis and can be used in predicting interest rates, foreign currency risk, volatility in stock markets and many more. Policymakers and business experts use financial forecasting to make decisions about production, purchases, market sustainability, allocation of resources, etc. In investment, this analysis is employed to track the price fluctuations and price of a security over time. For instance, the price of a security can be recorded.

Economics: Time series analysis is used in economics to study macroeconomic trends, forecast economic indicators, and measure the effectiveness of economic policies. For example, economists may use time series analysis to analyze trends in GDP, inflation, and unemployment. They can then use this information to identify economic cycles and assess the impact of economic policies.

Engineering: In engineering, time series analysis can be used to analyze data collected over time to identify patterns and trends, and to make predictions about future behavior. For example, time series analysis can be used to analyze sensor data from equipment to detect signs of wear and tear, and to schedule maintenance accordingly. It can also be used to monitor energy consumption and optimize energy usage in buildings or industrial processes.

Retail sales: For various product lines, a clothing retailer is looking to forecast future monthly sales. These forecasts need to account for the seasonal aspects of the customer's purchasing decisions. For example, in the northern hemisphere, sweater sales are typically brisk in the fall season, and swimsuit sales are the highest during the late

spring and early summer. Thus, an appropriate time series model needs to account for fluctuating demand over the calendar year.

Manufacturing: Time series analysis is used in manufacturing to monitor production processes, predict maintenance needs, and optimize product quality. For example, manufacturers may use time series analysis to monitor machine performance, predict maintenance needs, and identify trends in product quality. They can then use this information to optimize production processes, reduce downtime, and improve product quality.

2. Which are the models used for forecasting?
Ans)

Autoregressive Moving Average (ARMA):

The autoregressive moving average (ARMA) model is used on a stationary time series and is a combination of the autoregressive and moving average models. The ARMA model is defined as a regression model in which the dependent/response variable is a linear function of past values of both the dependent/response variable and the error term. The order of an ARMA model is represented by 'p' for the autoregressive part and 'q' for the moving average part.

Autoregressive Integrated Moving Average (ARIMA):

The autoregressive integrated moving average (ARIMA) model is a generalization of the ARMA model and is applied on non-stationary models. The ARIMA model is defined as a regression model in which the dependent/response variable is a linear function of past values of both the dependent/response variable and the error term, where the error term has been differentiated 'd' times. The order of an ARIMA model is represented by 'p' for the autoregressive part, 'q' for the moving average part, and 'd' for the differencing part.

Seasonal Autoregressive Integrated Moving-Average (SARIMA):

SARIMA is a type of time-series forecasting model that takes into account both seasonality and autocorrelation. SARIMA models are based on a combination of differencing, autoregression, and moving average processes. These models can be used to forecast short-term or long-term trends in data. SARIMA models are generally considered to be more accurate than other types of time-series forecasting models, such as ARIMA models. SARIMA models are also relatively easy to interpret and use. The SARIMA model can be used to forecast demand for a product or service over the course of a year.

Vector Autoregression (VAR):

VAR is a multivariate time series model that can forecast multiple variables simultaneously. It models the dependencies and interactions between multiple time series variables, making it useful for forecasting in situations where multiple variables influence each other.

3. Explain Auto-correlation function and partial auto-correlation function. 2
Ans)
Auto-correlation Function:

The difficulty is that the simple plot does not provide insight into the covariance of the variables in the time series and its underlying structure. The plot of the autocorrelation function (ACF) provides this insight. For a stationary time series, the ACF is defined as shown,

$$ACF(h) = \frac{cov(y_t, y_{t+h})}{\sqrt{cov(y_t, y_t)\, cov(y_{t+h}, y_{t+h})}} = \frac{cov(h)}{cov(0)}$$

By convention, the quantity h in the ACF is referred to as the lag, the difference between the time points t and t+h. At lag 0, the ACF provides the correlation of every point with itself. So ACF(O) always equals 1 and it goes on decreasing as the lag increases.

A value of 1 indicates a perfect positive autocorrelation, -1 indicates a perfect negative autocorrelation, and 0 indicates no autocorrelation.

Partial Auto-correlation Function:

A measure of the autocorrelation between Y_t and Y_{t+h} for h = 1, 2, 3 ... with the effect of the

Y_{t+1} and Y_{t+h-1} values excluded from the measure. The partial autocorrelation function (PACF) provides such a measure.
The PACF measures the correlation between a time series and its lagged values, while accounting for the correlations that may exist at shorter lags. By removing the influence of intermediate lags, the PACF allows us to identify the "pure" or "direct" relationship between the time series and its lagged values.
The PACF can be interpreted in a similar way to the autocorrelation function (ACF), but it measures the correlation between a time series and its lagged values after controlling for the influence of shorter lags.
Because the ACF and PACF are based on correlations, negative and positive values are possible. Thus, the magnitudes of the functions at the various lags should be considered in terms of absolute values.

4. What are the other methods of Time Series Analysis? 2
Ans)
Additional time series methods include the following:

• **Autoregressive Moving Average with Exogenous inputs** (ARM AX) is used to analyze a time series that is dependent on another time series. For example, retail demand for products can be modeled based on the previous demand combined with a weather-related time series such as temperature or rainfall.
• **Spectral analysis** is commonly used for signal processing and other engineering applications. Speech recognition software uses such techniques to separate the signal for the spoken words from the overall signal that may include some noise.
• **Generalized Autoregressive Conditionally Heteroscedastic** (GARCH) is a useful model for addressing time series with nonconstant variance or volatility. GARCH is used for modeling stock market activity and price fluctuations.
• **Kalman filtering** is useful for analyzing real-time inputs about a system that can exist in certain states. Typically, there is an underlying model of how the various components of the system interact and affect each other. A Kalman filter processes the various inputs, attempts to identify the errors in the input, and predicts the current state. For example, a Kalman filter in a vehicle navigation system can process various inputs, such as speed and direction, and update the estimate of the current location.

• **Vector ARIMA (VARIMA):** Multivariate time series analysis examines multiple time series and their effect on each other. Vector ARIMA (VARIMA**)** extends ARIMA by considering a vector of several time series at a particular time, t. VARIMA can be used in marketing analyses that examine the time series related to a company's price and sales volume as well as related time series for the competitors.

EXTRA

Q. AutoRegressive Model (p)

Ans)

Autoregressive models, often referred to as AR models, are a class of time series models that use past values of a variable to predict its future values. These models assume that the current value of a variable depends linearly on its previous values.

Autoregressive models are widely used in various fields, including finance, economics, signal processing, and climate science. They can capture temporal dependencies and patterns in time series data, making them valuable for prediction, understanding underlying dynamics, and generating insights from time-dependent variables.

Q. Moving Average Model (q)

Ans)

Moving Average (MA) models are another class of time series models used to analyze and forecast time-dependent data. Unlike autoregressive models that focus on past values of the variable, MA models consider the residual errors or the discrepancy between the predicted values and the actual values. Moving average models are commonly used in time series analysis and forecasting. They capture the short-term dependencies and random fluctuations in the data, which can be valuable for understanding and predicting time-dependent phenomena. MA models are often combined with autoregressive models to form more powerful and comprehensive models such as the Autoregressive Moving Average (ARMA) or the Seasonal Autoregressive Moving Average (SARMA) models.

Q. Building and Evaluating ARIMA model.

Ans) Building and evaluating an ARIMA (Autoregressive Integrated Moving Average) model involves several steps. Here's a general outline of the process:

Data Preparation: Ensure your data is in a suitable format for ARIMA modeling. It should be a time series with evenly spaced observations. If necessary, preprocess the data by addressing missing values, outliers, or other anomalies.

Stationarity Check: ARIMA models assume stationarity, which means the mean, variance, and autocorrelation structure of the data remain constant over time. Conduct a stationarity check on your time series using statistical tests or visual inspection. If the data is non-stationary, apply differencing to make it stationary.

Model Identification: Determine the appropriate order of the ARIMA model by examining the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. The ACF plot shows the correlation between the current observation and lagged observations, while the PACF plot reveals the direct relationship between the current

observation and specific lags. The plots can help identify the orders of autoregressive (p), integrated (d), and moving average (q) components of the model.

Model Estimation: Estimate the parameters of the ARIMA model based on the identified orders. This can be done using maximum likelihood estimation or other optimization techniques. Software packages like Python's statsmodels or R's forecast package provide functions to estimate ARIMA models.

Model Diagnostic Checking: Evaluate the model's goodness of fit and check for any remaining issues. Examine the residuals of the model for stationarity, normality, and lack of autocorrelation.

Model Refinement: If the diagnostic checks reveal any problems, adjust the model accordingly. You can try different orders of the ARIMA components or consider alternative models like seasonal ARIMA (SARIMA) or ARIMA with exogenous variables (ARIMAX). Iterate this step until the model adequately captures the data's patterns.

Model Forecasting: Once a satisfactory ARIMA model is obtained, use it to generate forecasts for future time points. The forecast horizon can vary depending on the specific application and requirements.

Model Evaluation: Compare the model's forecasts with the actual values to evaluate its performance. Use evaluation metrics such as mean squared error (MSE), mean absolute error (MAE), or others to quantify the accuracy of the forecasts. Additionally, visual inspection of the forecasts against the observed data can provide insights into the model's performance

Q.Reason to choose and Cautions in ARIMA model.
Ans) Reasons to Choose ARIMA:

Time Series Analysis: ARIMA models are specifically designed for analyzing and modeling time series data, which exhibit temporal dependencies and patterns. If you are working with a time-dependent dataset and want to understand its underlying dynamics, ARIMA can be a suitable choice.

Flexibility: ARIMA models can handle various types of time series data, including those with trend, seasonality, and irregular fluctuations. By incorporating autoregressive (AR), moving average (MA), and differencing components, ARIMA models offer flexibility in capturing different patterns and behaviors exhibited by the data.

Forecasting: ARIMA models are widely used for forecasting future values of time series variables. They utilize historical data and the identified patterns to make predictions for future time points. ARIMA models can be valuable in applications such as demand forecasting, stock market analysis, or predicting economic indicators.

Established Methodology: ARIMA models have a solid theoretical foundation and are supported by well-developed methodologies and statistical techniques. The principles and procedures for building ARIMA models have been extensively studied and documented, providing a reliable framework for analysis.

Cautions of ARIMA:

Stationarity Requirement: ARIMA models assume that the time series being analyzed is stationary, meaning its mean, variance, and autocorrelation structure remain constant over time. If the data is non-stationary, it may require

differencing or other transformations to achieve stationarity. Failing to meet the stationarity assumption can lead to unreliable results and inaccurate forecasts.

Model Identification Challenges: Determining the appropriate order of the ARIMA model (i.e., the values of p, d, and q) can be challenging. It often requires analyzing autocorrelation and partial autocorrelation plots, conducting hypothesis tests, and considering information criteria. Selecting the wrong order can result in poor model performance and inaccurate predictions.

Data Quality and Outliers: ARIMA models are sensitive to outliers and extreme values in the data. Outliers can have a disproportionate influence on the model estimation process, potentially distorting the parameter estimates and leading to unreliable forecasts. It is important to carefully examine and handle outliers before fitting an ARIMA model.

Limited Handling of Seasonality: While ARIMA models can capture trends and some forms of seasonality, they may struggle with more complex seasonal patterns or long-term dependencies. In such cases, alternative models like seasonal ARIMA (SARIMA) or other advanced time series models may be more appropriate.

External Factors: ARIMA models focus primarily on the time series data itself and may not explicitly incorporate external factors or predictors. If there are known external variables that influence the time series, it may be necessary to consider additional modeling techniques, such as ARIMAX (ARIMA with exogenous variables), to incorporate those factors effectively.

Adequate Data Availability: ARIMA models typically require a sufficient amount of historical data to estimate the model parameters accurately and generate reliable forecasts. If the time series dataset is very short or contains gaps, the model's performance may be limited.

MOD 4 – Text Analytics

12. Explain how you can Summarize Text.

Ans)
One of the use cases of text mining is the extraction of meaning when the goal is to quickly summarize one or a few very large documents.

There are two types of text summarizations.

- One type summarizes themes across the chapters or paragraphs of the text, in which case the individual paragraphs or chapters can be considered different documents of a larger corpus (the entire text). The goal of this type is to identify the different themes across the various documents (e.g., as just described) or to identify common dimensions or relationships among individuals, events, and so on.
- The second type summarizes the contents of a large text document into a meaningful narrative which cannot be accomplished effectively (yet) using automatic text mining methods and algorithms. In short, it is not realistic to expect that present computer algorithms are capable of summarizing the "essence" of a very large book into a single paragraph. At present, this can be done only in other highly subjective ways.

13. What do you mean by topic in Documents by Topics? How is it useful?

Ans) A topic consists of a cluster of words that frequently occur together and share the same theme. The topics of a document are not as straightforward as they might initially appear.
For example,
ACME wants to categorize the reviews by topics.
We Consider the following review:
1. While I love ACME's bPhone series, I've been quite disappointed by the bEbook. The text is illegible, and it makes even my old NBook look blazingly fast.

For machines, it is difficult to answer whether the review is about bPhone series, bEbook or NBook.

Since a document typically consists of multiple themes running through the text in different proportions, document grouping is required. It can be achieved with various clustering or classification methods. However, a more feasible and prevalent approach is to use topic modeling. Topic models are statistical models that examine words from a set of documents, determine the themes over the text, and discover how the themes are associated or change over time. Thus, it is very useful as it provides tools to automatically organize, search, understand, and summarize from vast amounts of information.

14. What is a caveat of IDF? How does TFIDF address the problem?

Ans)

The inverse document frequency (IDF) is a widely used metric in natural language processing and information retrieval for measuring the importance of a word in a corpus of documents. However, there are some caveats to be aware of when using IDF.

- Bias towards rare words.
- Inability to capture word context.
- Sensitivity to document size.
- Difficulty in comparing IDF values across different corpora.
- Vulnerability to noise

In tf-idf, the frequency of each word in a document is weighted by its inverse document frequency (IDF) score, which gives more weight to words that are rare in the corpus and less weight to words that are common. This helps to address the bias towards rare words that can occur with IDF alone.

In addition, tf-idf also takes into account the term frequency (TF) of each word in a document. This helps to address the limitation of IDF in capturing word context. By considering both the TF and IDF of each word, tf-idf can give higher weight to words that are both rare and important in a specific document.

15. Explain IDF and TF IDF. How does TFIDF overcome the issue of using IDF?

Ans)

IDF measures the rarity of a word in a corpus of documents. The intuition behind IDF is that words that appear in many documents are less informative than words that appear in few documents.

The TFIDF (or TF-IDF) is a measure that considers both the prevalence(context) of a term within a document (TF) and the scarcity(rarity) of the term over the entire corpus (IDF).

Tf-idf overcomes some of the limitations of using IDF alone. Here are some of the ways in which tf-idf addresses the issues with using IDF:

- Addresses the bias towards rare words.
- Captures word context.
- Less sensitive to document size.
- Can be used to compare scores across different corpora.
- Less vulnerable to noise.

16. Explain Precision and Recall. Calculate precision and recall from above given confusion matrix.

Ans)

Precision and Recall

- **Precision**: exactness – what % of tuples that the classifier labeled as positive are actually positive

$$precision = \frac{TP}{TP + FP}$$

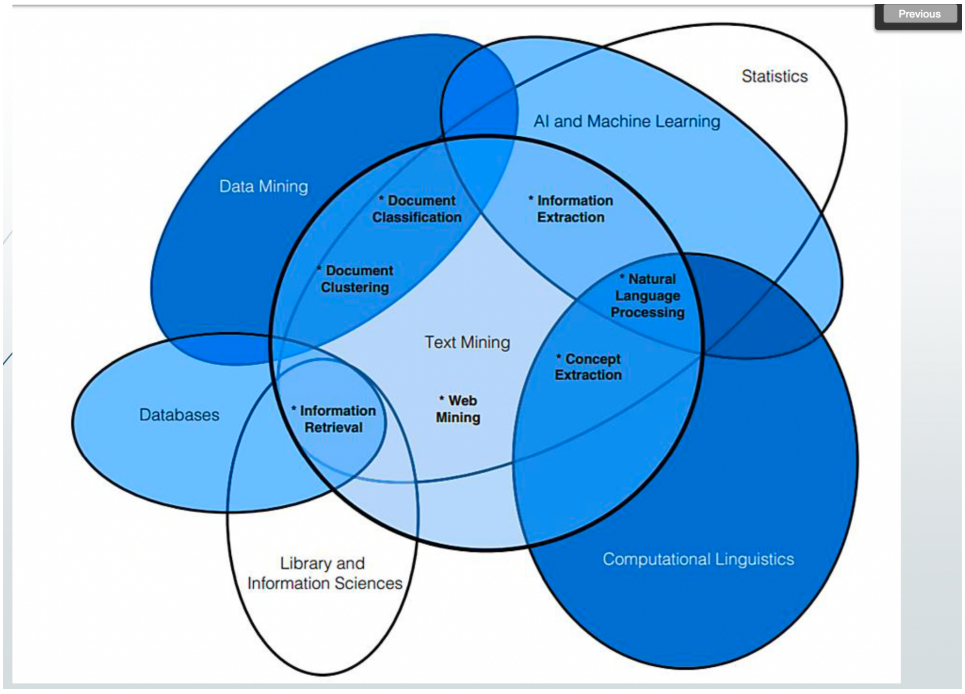
- **Recall**: completeness – what % of positive tuples did the classifier label as positive?

$$recall = \frac{TP}{TP + FN}$$

- Perfect score is 1.0
- Inverse relationship between precision & recall

8. Explain seven practices of text analytics. Give an example application in that area.

Ans)



- Search and information retrieval (IR): Storage and retrieval of text documents, including search engines and keyword search.
- Document clustering: Grouping and categorizing terms, snippets, paragraphs, or documents, using data mining clustering methods.
- Document classification: Grouping and categorizing snippets, paragraphs, or documents, using data mining classification methods, based on models trained on labeled examples.
- Web mining: Data and text mining on the Internet, with a specific focus on the scale and interconnectedness of the web.

- Information extraction (IE): Identification and extraction of relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semi structured text.
- Natural language processing (NLP): Low-level language processing and understanding tasks (e.g., tagging part of speech); often used synonymously with computational linguistics.
- Concept extraction: Grouping of words and phrases into semantically similar groups.

Examples:

IR: one of the real time applications of IR is that it is used in search engines like google,bing,etc

Document Clustering: A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories.

Document Classification: It can be used in Gmail as a spam classifier making it easy for users and keeping their inbox clear.

Web Mining: It can be used to collect data from the web i.e retrieve relevant information using mining.
IE: One of the example of IE is as follows-
“Shanaya lives in Mumbai city.”
Through information extraction, the following basic facts can be pulled out of the free-flowing text and organized in a structured, machine-readable form:

Person: Shanaya

Location: Mumbai

NLP: For example, it is used in Chatbot like ChatGPT.

Concept Extraction: For example, when searching Google images for “fluffy cat”, I have defined a set of features *F={fluffy, cat}*, and the response Google gives—a collection of fluffy cat images—is the concept *Y* extracted for the feature set *F*

9. Explain three important steps for text analysis problems in detail.

Ans)

Text Analysis Steps

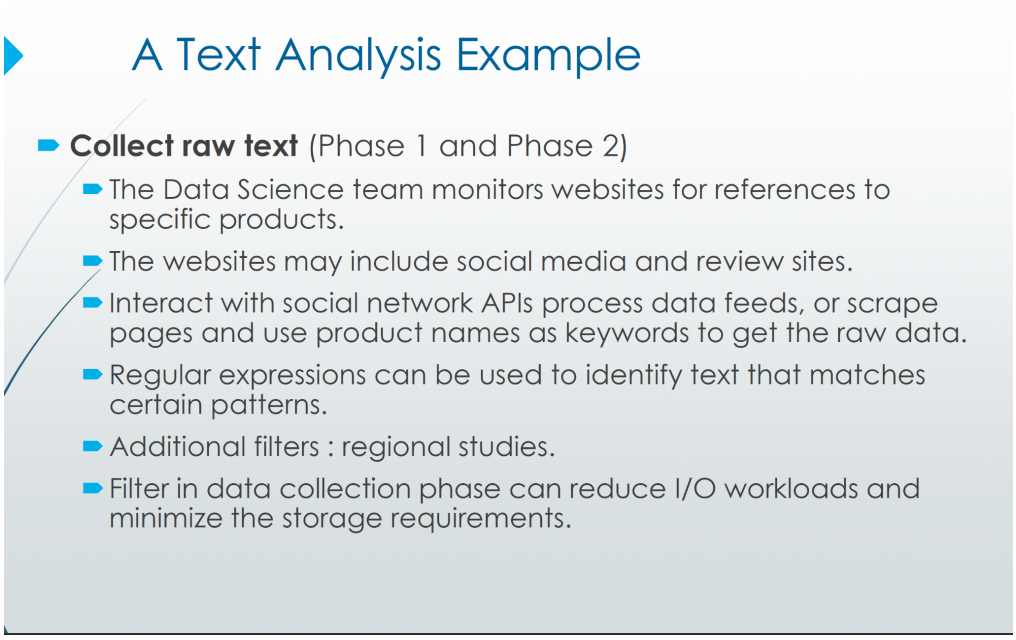
A text analysis problem usually consists of three important steps: parsing, search and retrieval, and text mining.

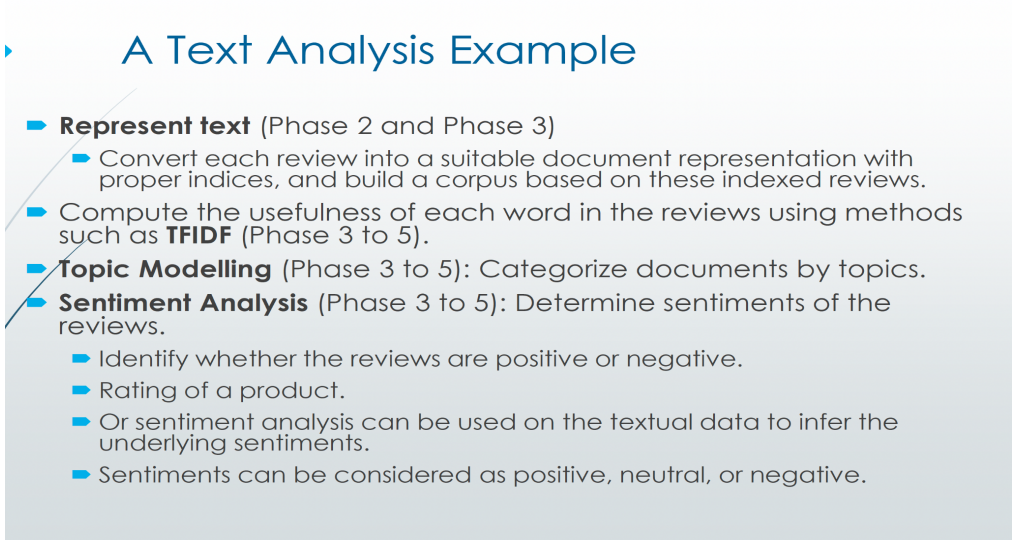
1. **PARSING** is the process that takes unstructured text and imposes a structure for further analysis. The unstructured text could be a plain text file, a weblog, an Extensible Markup Language (XML) file, a HyperText Markup Language (HTML) file, or a Word document. Parsing deconstructs the provided text and renders it in a more structured way for the subsequent steps.
2. **SEARCH AND RETRIEVAL** is the identification of the documents in a corpus that contain search items such as specific words, phrases, topics, or entities like people or organizations. These search items are generally called key terms. Search and retrieval originated from the field of library science and is now used extensively by web search engines.
3. **TEXT MINING** uses the terms and indexes produced by the prior two steps to discover meaningful insights pertaining to domains or problems of interest. With the proper representation of the text, many of the techniques, such as clustering and classification, can be adapted to text mining. K-means can be modified to cluster text documents into groups, where each group represents a collection of documents with a similar topic. The distance of a document to a centroid represents how closely the document talks about that topic. Classification tasks such as sentiment analysis and spam filtering are prominent use cases for the naïve Bayes classifier.

All three steps may not be included in all projects. Also they may have any sequence.

10. Explain how you can perform Collection Raw Text and Represent that Text using any example application.

Ans)





Let's take the example of an application that collects and represents news articles. This application could collect news articles from various sources using web scraping tools or APIs and store them in a database. Once the news articles are collected, the application could represent them using a bag-of-words approach. The bag-of-words approach involves creating a matrix of word frequencies in the news articles. For example, the application could use the bag-of-words matrix to perform sentiment analysis on the news articles to determine the overall sentiment of each article. The application could also use the matrix to perform topic modeling to identify the main topics discussed in the news articles. The application could then use this information to provide users with personalized news recommendations based on their interests.

EXTRA QUESTIONS

Q Finding the right practice area

FIVE QUESTIONS FOR FINDING THE RIGHT PRACTICE AREA

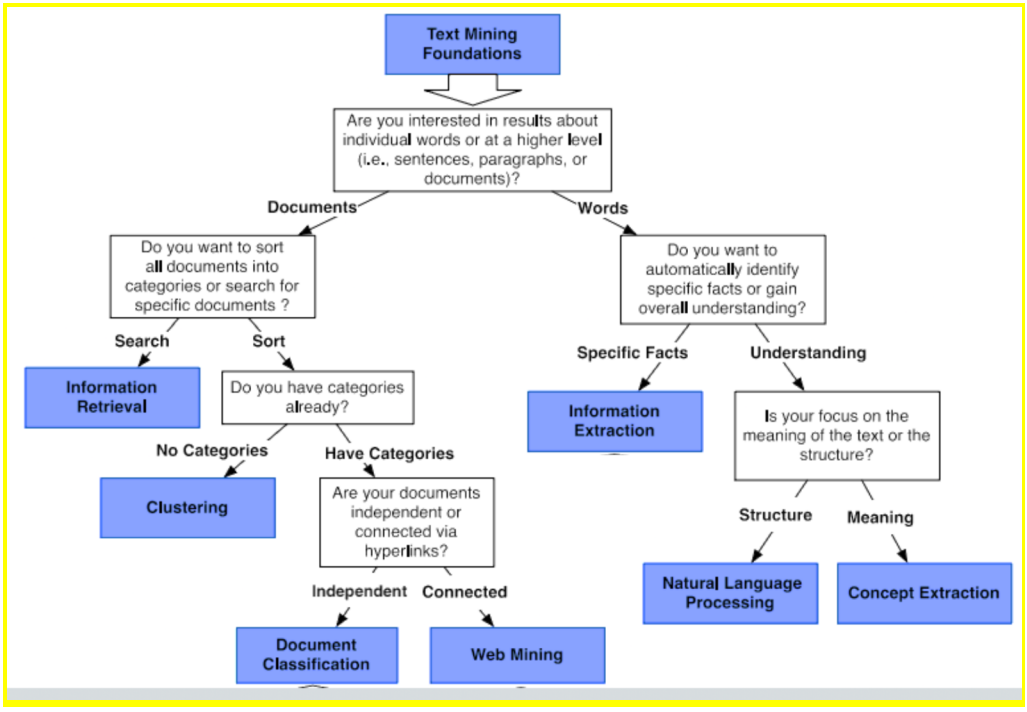
- **Question 1: Granularity**
- This question finds the desired granularity (level of detail of focus) of the text mining task.
- While documents and words are both integral to successful text mining, an algorithm virtually always emphasizes one or the other.
- To determine the granularity of your text mining problem, ask yourself about the desired outcome: Is it about characterizing or grouping together words or documents?
- This is the biggest division between classes of text mining algorithms.

- **Question 2: Focus**
- The focus of the algorithm: Are you interested in finding specific words and documents or characterizing the entire set?
- The two practice areas separated by this question - search and information extraction - both concentrate on identifying specific pieces of information within a document database, whereas the other solutions attempt to cluster or partition the space.

- **Question 3: Available Information**
- If you are interested in documents, the next question regards the available information at the time of analysis.
- This is equivalent to the supervised/unsupervised question from data mining.
- A supervised algorithm requires training data with an answer (outcome label) for positive and negative examples of the classes you're trying to model (such as distinguishing "interesting versus not interesting" articles for an analyst studying a specialized topic).
- An unsupervised algorithm does not require any labeled data, and it can be applied to any data set without any available information at analysis time. Supervised learning is much more powerful when possible to used that is, when enough example cases with tar outcomes are known.

- **Question 4: Syntax or Semantics**
- If you are interested in words, the major question is about syntax or semantics.
- Syntax is about what the words "say," while semantics is about what the words "mean."
- Because natural language is so fluid and complex, semantics is the harder problem. However, there are text mining algorithms to address both areas.

- **Question 5: Web or Traditional Text**
- The rise of the Internet (including blogs, Twitter, and Facebook) is largely responsible for the prominence that text mining holds today by making available a vast number of previously unreachable text documents. The structure and style of web documents provide both unique opportunities and challenges when compared to non web documents. Though many of the algorithms are theoretically the same for web and traditional text, the scale of the web and its unique structural characteristics justify defining two different categories.



Module 5 – Data analytics and visualization with R

5. Explain Kernel density plot in r with proper example.

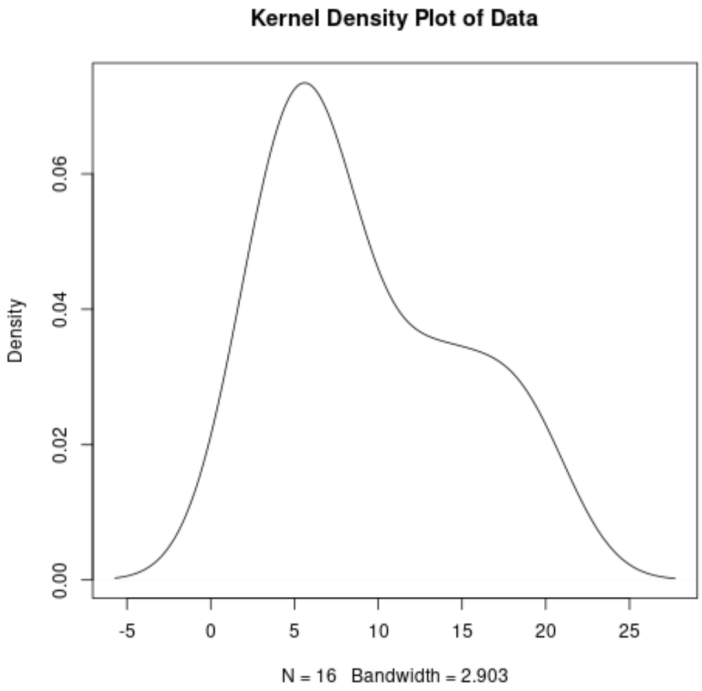
Ans)

A kernel density plot is a type of plot that displays the distribution of values in a dataset using one continuous curve. A kernel density plot is similar to a histogram, but it's even better at displaying the shape of a distribution since it isn't affected by the number of bins used in the histogram

- Create One Kernel Density Plot

The following code shows how to create a kernel density plot for one dataset in R:

```
#create data

data <- c(3, 3, 4, 4, 5, 6, 7, 7, 7, 8, 12, 13, 14, 17, 19, 19)

#define kernel density

kd <- density(data)

#create kernel density plot

plot(kd, main='Kernel Density Plot of Data')
```



The x-axis shows the values of the dataset and the y-axis shows the relative frequency of each value. The highest point in the plot shows where the values occur most often.

6. What is the main idea for exploratory data analysis and why do you need visualization before analysis?
Ans)
Exploratory data analysis is a data analysis approach to reveal the important characteristics of a dataset, mainly through visualization. The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, and find interesting relations among the variables. Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals.

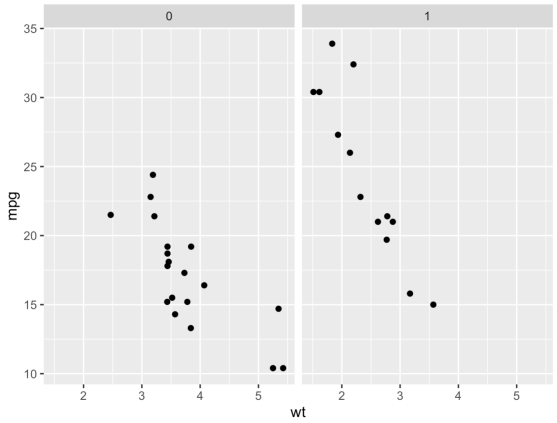
Visualization is an important part of EDA because it allows us to quickly and easily explore the data and identify patterns or trends that may not be immediately apparent from numerical summaries or statistics. By visualizing the data, we can gain insight into its distribution, identify outliers or unusual observations, and see how different variables are related to each other.

For example, suppose we have a dataset of sales data for a retail store, including variables such as date, product, price, and quantity sold. Before analyzing the data, we may want to visualize it to gain a better understanding of its structure and identify any patterns or trends. We might create a line chart of sales over time to see if there are any seasonal patterns, a scatterplot of price versus quantity sold to see if there is a relationship between these variables, or a histogram of sales by product to see which products are selling the most. By visualizing the data in this way, we can identify interesting features or patterns and generate hypotheses about the data, which we can then test using statistical techniques.

1. How would you use facet wrap and facet grid methods of visualization with R and give proper examples. 3
Ans)
In R, the ggplot2 package provides the facet_wrap() and facet_grid() functions for creating multi-panel plots that allow you to visualize subsets of your data. Both functions are useful for exploring relationships between variables or comparing groups of data.

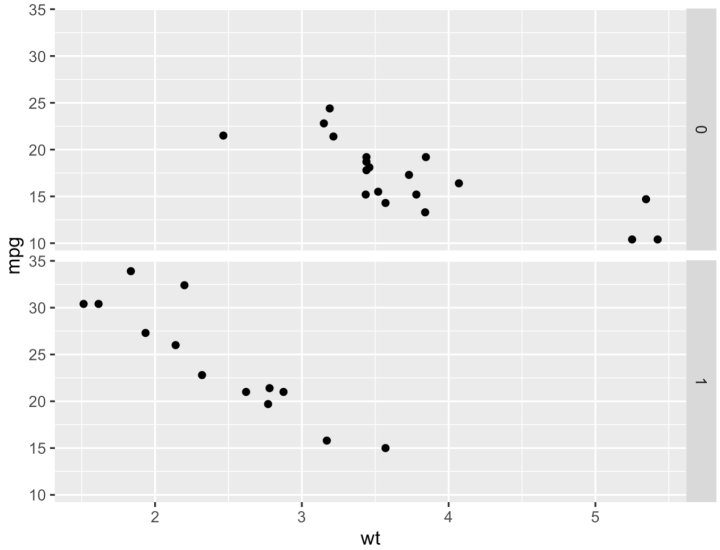
Example using the mtcars dataset that comes with R:
Suppose we want to compare the relationship between miles per gallon (mpg) and weight (wt) for different types of transmission (am):

```
library(ggplot2)
ggplot(mtcars, aes(x = wt, y = mpg)) +
geom_point() +
facet_wrap(~ am)
```



This code creates a scatterplot of mpg versus wt using the ggplot2 package. The facet_wrap() function is used to create separate panels for each type of transmission (am). The ~ symbol in the facet_wrap() function indicates that am is the variable used to create the subsets.
Alternatively, we can use facet_grid() to create a 2x1 grid with separate panels for automatic (am = 0) and manual (am = 1) transmissions:

```
ggplot(mtcars, aes(x = wt, y = mpg)) +
geom_point() +
facet_grid(am ~ .)
```



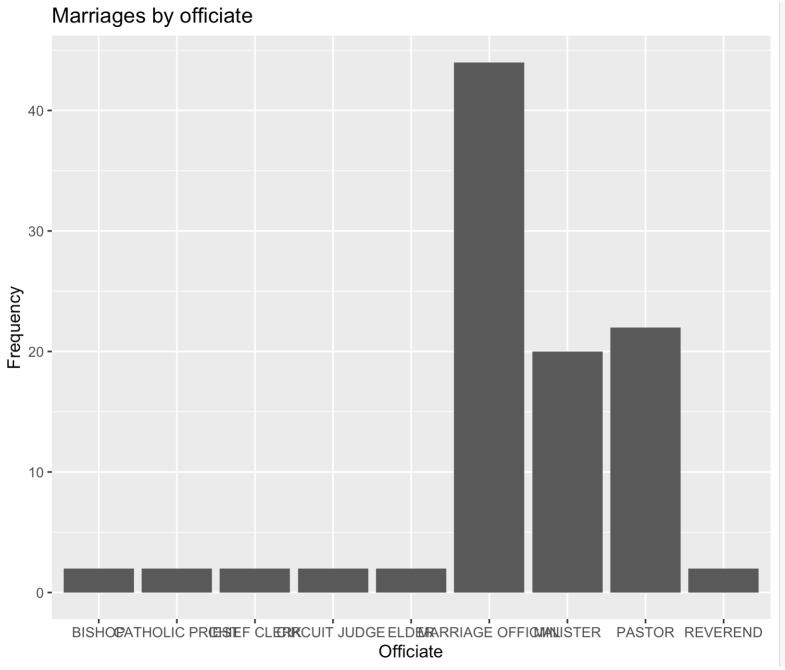
This code creates the same scatterplot, but with facet_grid() instead of facet_wrap(). The am ~ . argument tells R to use am for the rows and to put all remaining variables (in this case, none) in the columns.

Both facet_wrap() and facet_grid() are useful for comparing subsets of your data and identifying patterns or trends within those subsets.

2. How will you enhance the following R code to display horizontal bar charts and avoid axis labels to overlap?
Ans) (TAKING A CODE FOR EXAMPLE ANY CODE MIGHT COME IN EXAM)

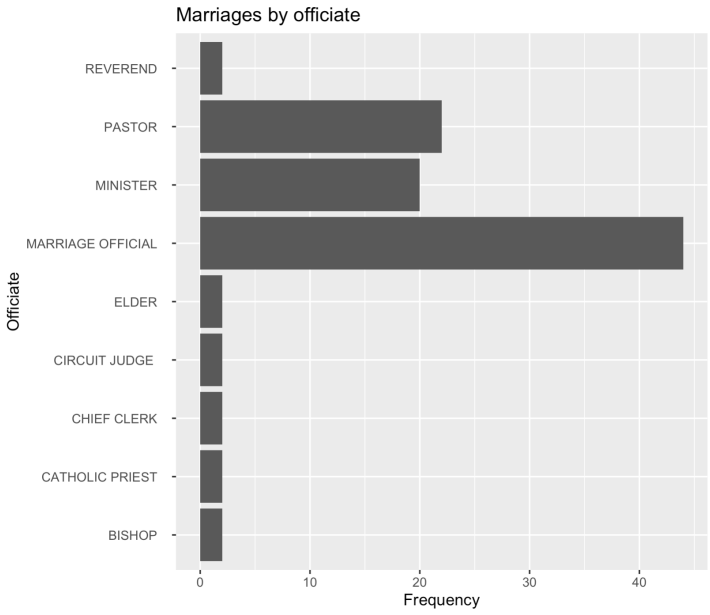
In this code, we first load the Marriage dataset from the mosaicData package. We then create a bar chart using ggplot() and specify the officialTitle variable as the x-axis variable. We use geom_bar() to create the bar chart and labs() to add axis labels and a chart title.

```
data(Marriage, package = "mosaicData")
library(ggplot2)
p=ggplot(Marriage,aes(x=officialTitle))+
geom_bar()+labs(x="Officiate",y="Frequency",title="Marriages by officiate")
p
```



As we can see from above output the officiate labels are overlapping making it difficult to read the labels. Thus we modify the code to avoid overlapping of data and better get a better understanding of the bar chart.

```
p + coord_flip() + theme(axis.text.y = element_text(margin = margin(r = 10)))
```



We can see that by increasing the space between the right edge of the plot and the y-axis labels, we ensure that the labels do not overlap with each other or with the plot title, making it easier to read and interpret the chart. Additionally, we flipped the coordinate system of the plot using coord_flip() to create a horizontal bar chart. This allows for longer axis labels to be displayed without being truncated, as the horizontal orientation provides more space for the labels to be displayed.

EXTRA

Single and multiple variable

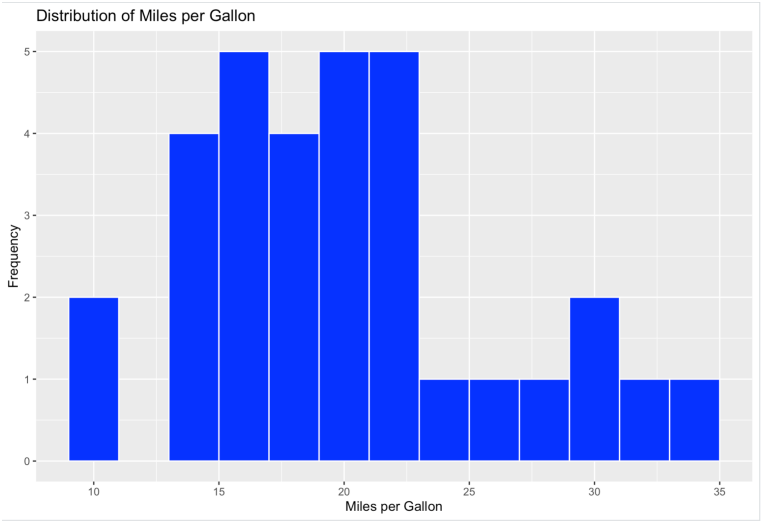
The visualization for mtcars dataset in R for single as well as multiple variables is as follows:

```
> library(ggplot2)
> data(mtcars)
```

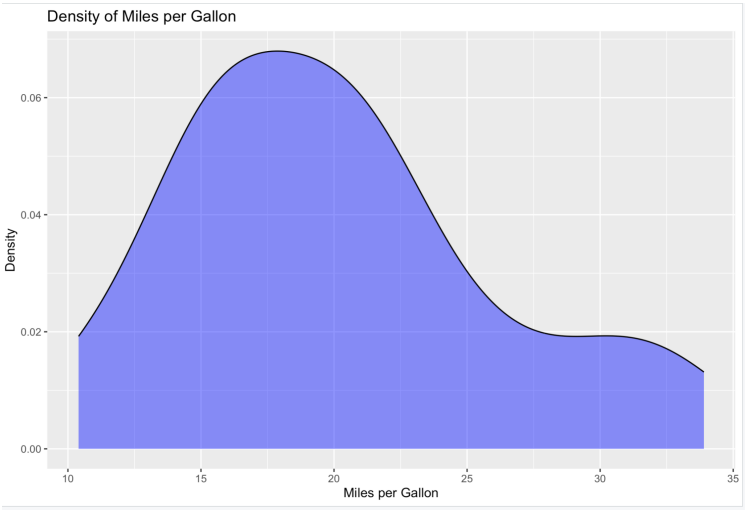
row names	mpg	cyl	disp	hp	drat	wt	qsec	vs	am
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0

Showing 1 to 11 of 32 entries, 11 total columns

```
> # visualize single variable
> #histogram
> ggplot(mtcars, aes(x = mpg)) +
+   geom_histogram(binwidth = 2, fill = "blue", color = "white") +
+   labs(x = "Miles per Gallon", y = "Frequency"
+        , title = "Distribution of Miles per Gallon")
```

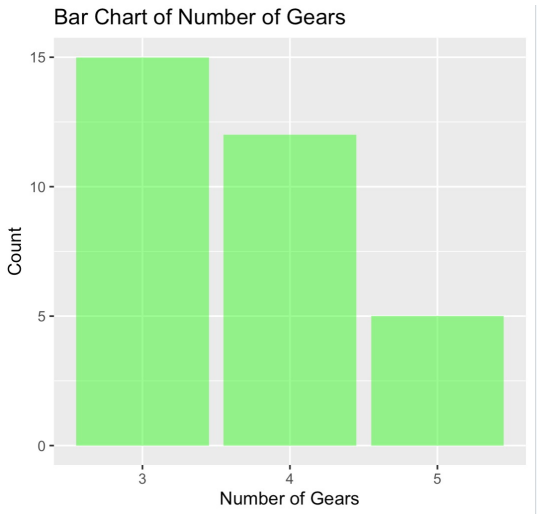


```
> #density plot
> ggplot(mtcars, aes(x = mpg)) +
+   geom_density(fill = "blue", alpha = 0.5) +
+   labs(x = "Miles per Gallon", y = "Density"
+        , title = "Density of Miles per Gallon")
>
```



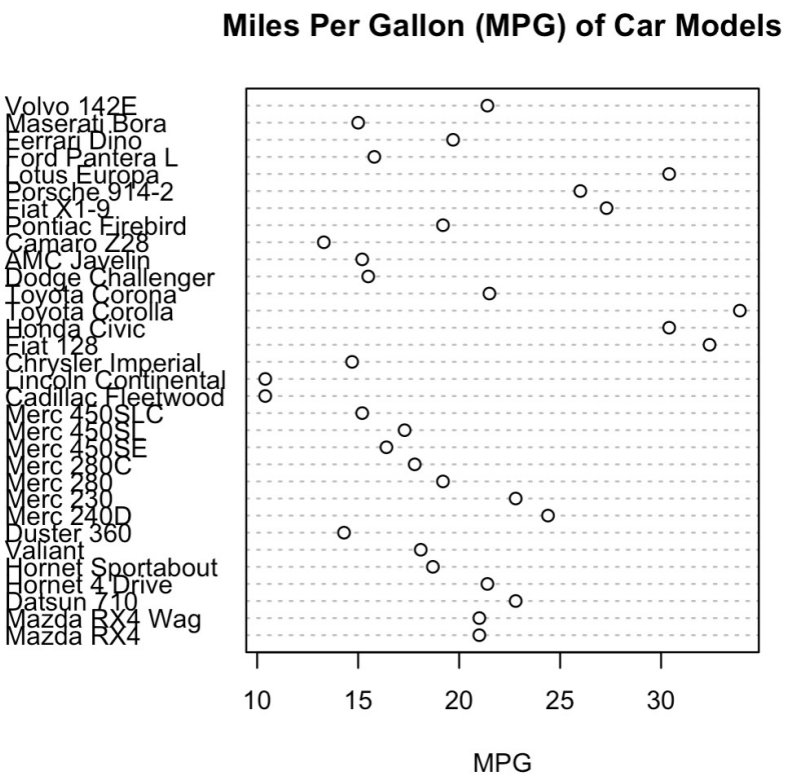
Bar plot

```
ggplot(mtcars, aes(x = factor(gear))) +
  geom_bar(fill = "green", alpha = 0.5) +
  labs(x = "Number of Gears", y = "Count",
       title = "Bar Chart of Number of Gears")
```



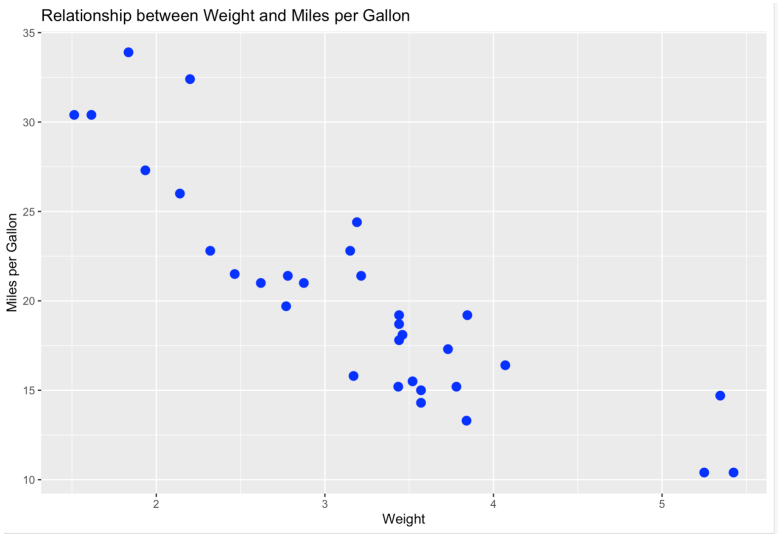
Dot Chart

```
dotchart(mtcars$mpg,labels=row.names(mtcars),cex=.8,
        main="Miles Per Gallon (MPG) of Car Models",
        xlab="MPG")
```

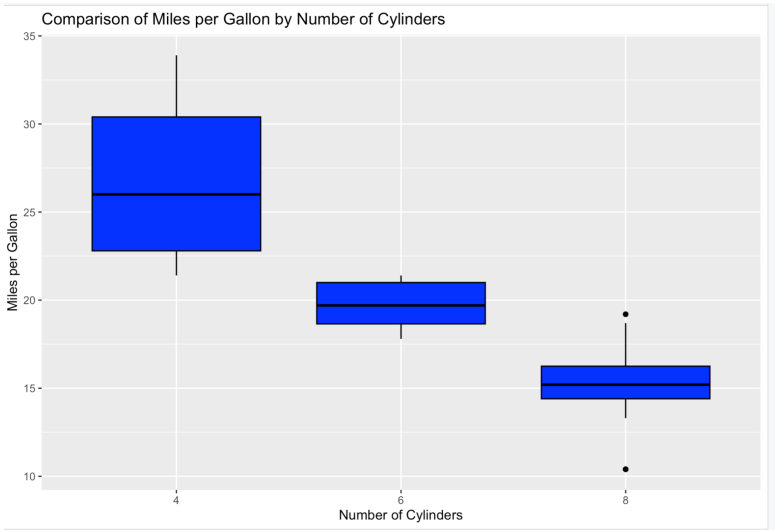


Scatter Plot

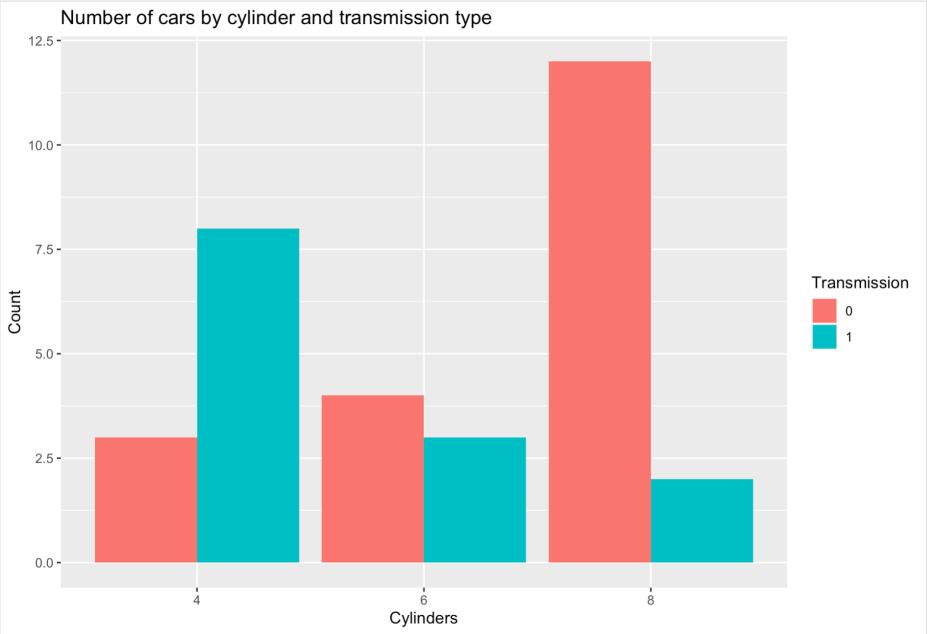
```
> #visualization of multiple variables
> #scatter plot
> ggplot(mtcars, aes(x = wt, y = mpg)) +
+   geom_point(color = "blue", size = 3) +
+   labs(x = "Weight", y = "Miles per Gallon"
+        , title = "Relationship between Weight and Miles per Gallon")
```




```
> #box plot
> ggplot(mtcars, aes(x = factor(cyl), y = mpg)) +
+   geom_boxplot(fill = "blue", color = "black") +
+   labs(x = "Number of Cylinders", y = "Miles per Gallon"
+        , title = "Comparison of Miles per Gallon by Number of Cylinders")
```

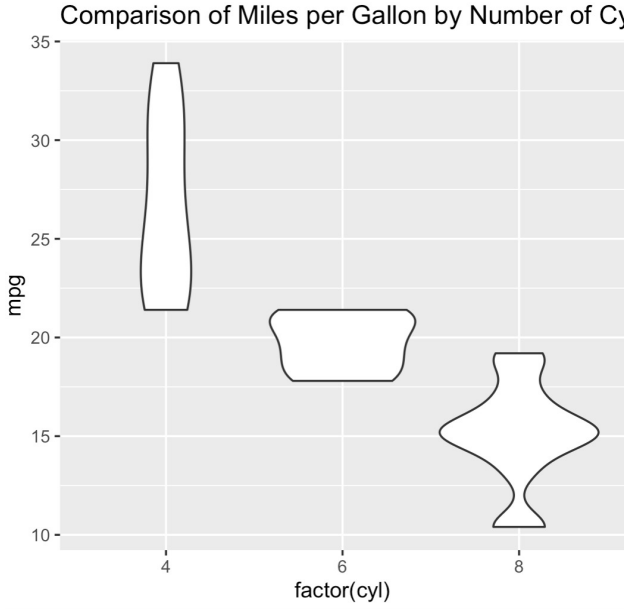


```
> #bar plot
> library(ggplot2)
> ggplot(mtcars, aes(x = factor(cyl), fill = factor(am))) +
+   geom_bar(position = "dodge") +
+   labs(title = "Number of cars by cylinder and transmission type",
+        x = "Cylinders",
+        y = "Count",
+        fill = "Transmission")
```



Violin Plot

```
ggplot (mtcars, aes (x = factor(cyl), y = mpg)) +
  geom_violin () +
  labs (title = "Comparison of Miles per Gallon by Number of Cylinders")
```



Module 6 – Data analytics and visualization with Python

1. How would you apply str.cat() to following python code to concatenate the address column with the name column?

Ans) **Pandas:**

- Pandas is a powerful library for data manipulation and analysis. It provides data structures like DataFrame and Series, which are highly efficient for handling structured data.
- Key features of Pandas include data cleaning, data wrangling, reshaping, merging, slicing, and indexing.
- It enables operations such as filtering, grouping, sorting, and aggregation on datasets.
- Pandas integrates well with other libraries, making it a primary tool for data preprocessing and exploratory data analysis.

Pandas is primarily used for data manipulation and analysis

Example:

Assuming you have a DataFrame with columns "name" and "address" and you want to concatenate these two columns together into a new column named "full_name_address", you can use the str.cat() method in the following way:

```
import pandas as pd
# create example DataFrame
df = pd.DataFrame({ 'name': ['John', 'Jane', 'Bob'], 'address': ['123 Main St', '456 Oak Ave', '789 Elm Blvd'] })
# concatenate name and address columns into a new column named full_name_address
df['full_name_address'] = df['name'].str.cat(df['address'], sep=', ')
print(df)
```

```
   name         address         full_name_address
0  John     123 Main St         John, 123 Main St
1  Jane     456 Oak Ave         Jane, 456 Oak Ave
2  Bob      789 Elm Blvd        Bob, 789 Elm Blvd
```

2. How would you find the determinant and rank for the following matrix using python code?

Ans)**NumPy:**

- NumPy (Numerical Python) is a fundamental library for numerical computing in Python.
- It provides a multidimensional array object called ndarray, which allows efficient handling of large datasets.
- NumPy offers a wide range of mathematical functions for array manipulation, element-wise operations, linear algebra, Fourier transforms, random number generation, and more.
- It is known for its speed and efficiency, as many NumPy operations are implemented in C or Fortran for performance optimization.
- NumPy serves as the foundation for many other scientific computing libraries in Python.

NumPy provides efficient numerical computing capabilities.

Example:

To find the determinant and rank of a matrix in Python, you can use the NumPy library. Here's an example code for a matrix A:

```
import numpy as np

# define the matrix A

A = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])

# calculate the determinant of A

det_A = np.linalg.det(A)

print("Determinant of A:", det_A)

# calculate the rank of A

rank_A = np.linalg.matrix_rank(A)

print("Rank of A:", rank_A)
```

Output:

```
Determinant of A: 0.0

Rank of A: 2
```

In this example, we first define the matrix A using NumPy's array() function. Then, we use the linalg.det() function to calculate the determinant of A and store it in the variable det_A. We use the linalg.matrix_rank() function to calculate the rank of A and store it in the variable rank_A. Finally, we print out the values of det_A and rank_A.

1. How would you use CSR and CSC in scipy to handle spare data, explain the methods with examples?

Ans)

SciPy:

- SciPy (Scientific Python) is a library built on top of NumPy and provides additional functionality for scientific computing.
- It offers a comprehensive collection of algorithms and tools for optimization, interpolation, integration, linear algebra, signal and image processing, statistics, and more.
- SciPy includes submodules such as scipy.optimize, scipy.integrate, scipy.stats, scipy.signal, and scipy.linalg, each focusing on specific scientific computing tasks.
- It is widely used in fields such as physics, engineering, biology, finance, and machine learning for advanced scientific computations and data analysis.

SciPy extends the functionality of NumPy with additional scientific computing tools and algorithms.

Example:

CSR and CSC are two storage formats used in the Scipy library to efficiently handle sparse matrices. CSR stands for "Compressed Sparse Row" and it represents a sparse matrix using three arrays: one for the non-zero values, one for the column indices of the non-zero values, and one for the row pointers that indicate the start and end of each row in the first two arrays. Here's an example of how to create a CSR matrix using Scipy:

```
import numpy as np
from scipy.sparse import csr_matrix
# define a dense matrix
A_dense = np.array([[1, 0, 2], [0, 3, 0], [4, 0, 5]])
# create a CSR matrix from the dense
matrix A_csr = csr_matrix(A_dense)
# print the CSR matrix
print(A_csr)
```

Output:

```
(0, 0) 1
(0, 2) 2
(1, 1) 3
(2, 0) 4
(2, 2) 5
```

In this example, we first define a dense matrix A_dense. Then, we use the csr_matrix() function to create a CSR matrix A_csr from the dense matrix. Finally, we print out the CSR matrix using the print() function. As you can see, the CSR matrix only stores the non-zero values and their indices, and it uses the row pointers to reconstruct the original matrix.

CSC stands for "Compressed Sparse Column" and it represents a sparse matrix using the same three arrays as CSR, but with the column indices and row pointers swapped. Here's an example of how to create a CSC matrix using Scipy:

```
import numpy as np
from scipy.sparse import csc_matrix
# define a dense matrix
A_dense = np.array([[1, 0, 2], [0, 3, 0], [4, 0, 5]])
# create a CSC matrix from the dense
matrix A_csc = csc_matrix(A_dense)
# print the CSC matrix
print(A_csc)
```

Output:

```
(0, 0)  1
(2, 0)  4
(1, 1)  3
(0, 2)  2
(2, 2)  5
```

In this example, we first define a dense matrix A_dense. Then, we use the csc_matrix() function to create a CSC matrix A_csc from the dense matrix. Finally, we print out the CSC matrix using the print() function. As you can see, the CSC matrix stores the same non-zero values and their indices as the CSR matrix, but it uses the column indices and row pointers to reconstruct the original matrix.

Both CSR and CSC formats are useful for handling sparse matrices because they can save a lot of memory and computation time compared to dense matrices. The choice between CSR and CSC depends on the specific problem and the available algorithms for each format.

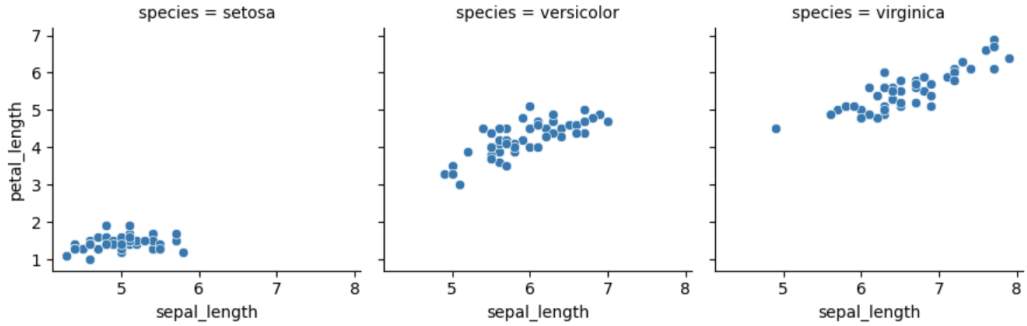
2. How would you use facet grid and pair grid methods of visualization with python seaborn and give proper examples. 3

Ans)

Facet Grid

Suppose we want to compare the relationship between sepal length (sepal_length) and petal length (petal_length) for each species (species) in the iris dataset. We can use FacetGrid() to create separate panels for each species:

```
import seaborn as sns
# Load iris dataset
iris = sns.load_dataset("iris")
# Create FacetGrid
g = sns.FacetGrid(iris, col="species")
g.map(sns.scatterplot, "sepal_length", "petal_length")
```

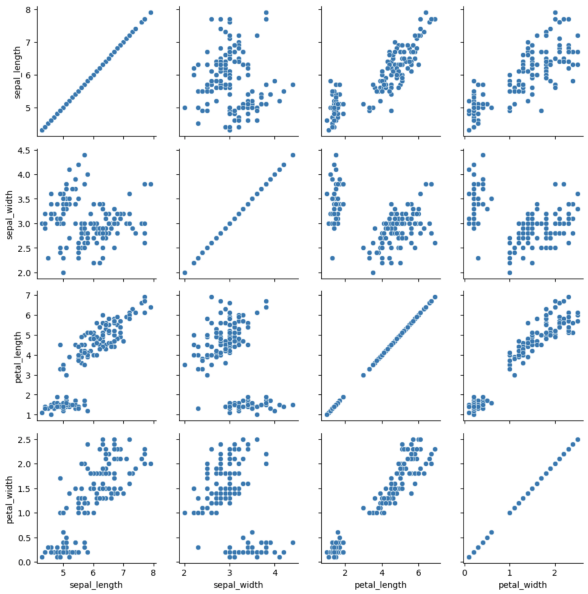


This code creates a FacetGrid with col="species", which tells Seaborn to create separate panels for each species in the iris dataset. We then use map() to apply a scatterplot to each panel, plotting sepal_length on the X-axis and petal_length on the Y-axis.

Pair Grid

Suppose we want to compare the relationship between all pairs of variables in the iris dataset. We can use PairGrid() to create a grid of scatter plots, with each variable plotted against all other variables:

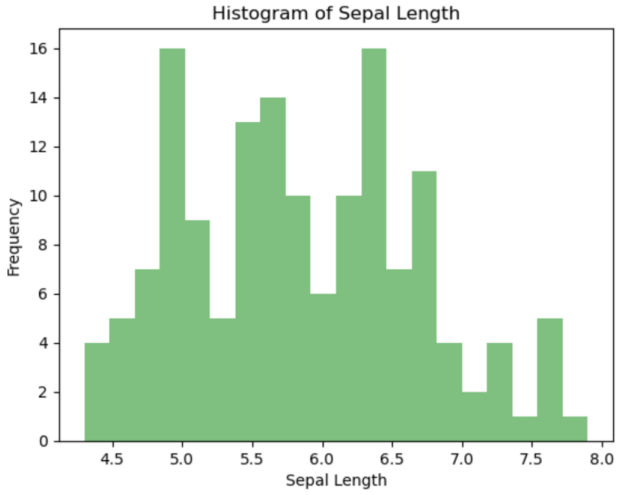
```
import seaborn as sns
# Load iris dataset
iris = sns.load_dataset("iris")
# Create PairGrid
g = sns.PairGrid(iris)
g.map(sns.scatterplot)
```



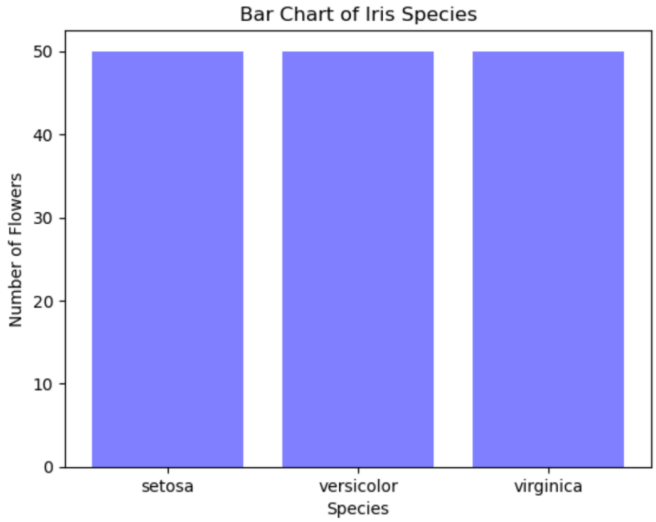
This code creates a PairGrid with the iris dataset. We then use map() to apply a scatterplot to each cell in the grid, plotting one variable on the X-axis and another variable on the Y-axis.
Both FacetGrid and PairGrid are useful for comparing subsets of your data and identifying patterns or trends within those subsets.

EXTRA
Matplotlib
Q.Create Histogram, BarChart, Pie chart, Box Plot, violin plot using Matplotlib.

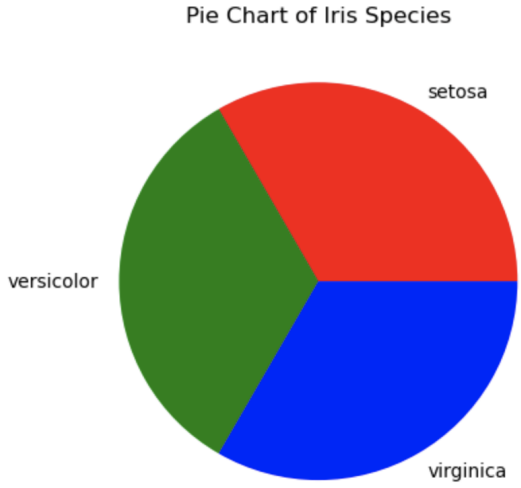
In [1]:
```
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

# load iris dataset
iris = sns.load_dataset("iris")

# plot the histogram
plt.hist(iris['sepal_length'], bins=20, color='green', alpha=0.5)
plt.title('Histogram of Sepal Length')
plt.xlabel('Sepal Length')
plt.ylabel('Frequency')
plt.show()
```



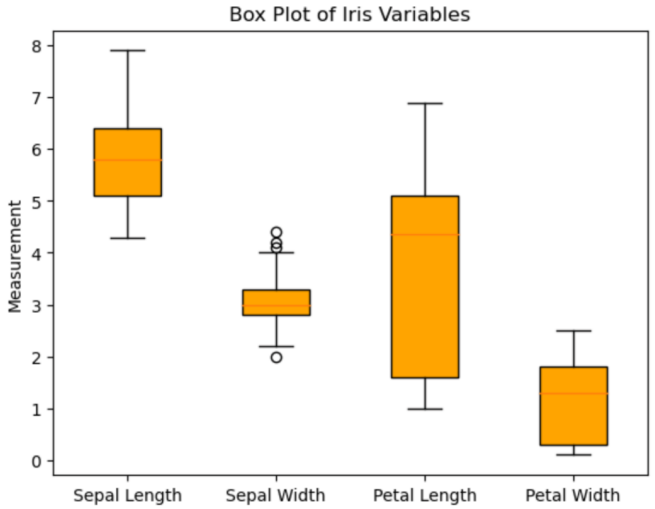
In [2]:
```
# plot the bar chart
class_counts = iris['species'].value_counts()
plt.bar(class_counts.index, class_counts.values, color='blue', alpha=0.5)
plt.title('Bar Chart of Iris Species')
plt.xlabel('Species')
plt.ylabel('Number of Flowers')
plt.show()
```



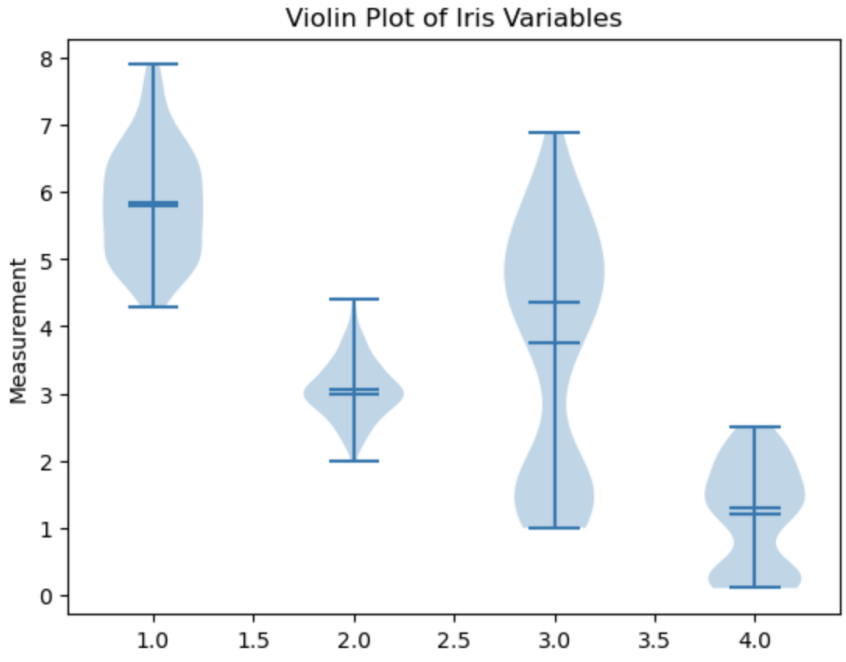
In [9]:
```
# plot the pie chart
species_counts = iris['species'].value_counts()
plt.pie(species_counts, labels=species_counts.index, colors=['red', 'green', 'blue'])
plt.title('Pie Chart of Iris Species')
plt.show()
```



In [12]:
```
# create box plot
fig, ax = plt.subplots()
ax.boxplot([iris['sepal_length'], iris['sepal_width'], iris['petal_length'],
            iris['petal_width']], labels=['Sepal Length', 'Sepal Width', 'Petal Length', 'Petal Width'],
           patch_artist=True, boxprops=dict(facecolor='orange', color='black'))
ax.set_title('Box Plot of Iris Variables')
ax.set_ylabel('Measurement')
plt.show()
```



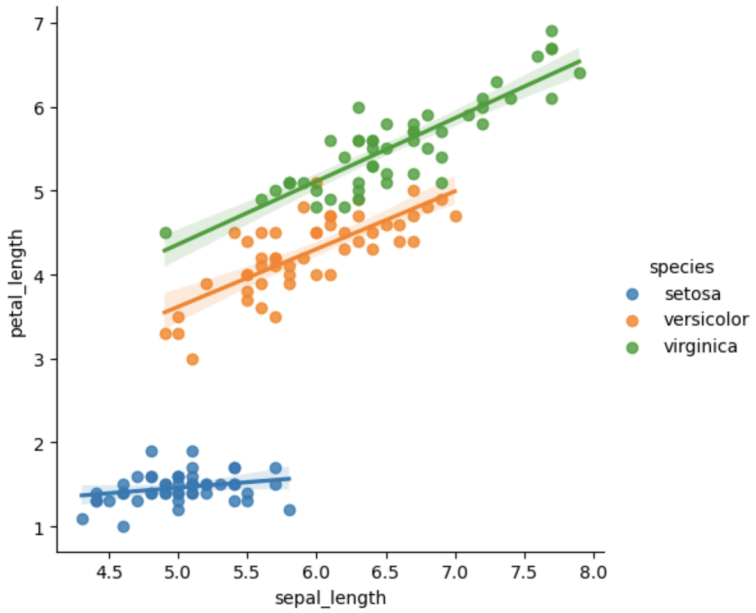
In [7]:
```
# create violin plot
fig, ax = plt.subplots()
ax.violinplot([iris['sepal_length'], iris['sepal_width'], iris['petal_length'],
               iris['petal_width']], showmeans=True, showmedians=True)
ax.set_title('Violin Plot of Iris Variables')
ax.set_ylabel('Measurement')
plt.show()
```



Seaborn:

Q.Regression Plot

In [15]:
```
# create lm plot
sns.lmplot(x='sepal_length', y='petal_length', hue='species', data=iris);
```



Reg plot:

In [18]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Create a scatter plot with regression line
sns.regplot(x=iris['sepal_length'], y=iris['petal_length'])

# Set plot title and axis labels
plt.title('Scatter Plot with Regression Line')


# Display the plot
plt.show()
```