**ASSIGNMENT 4**

**1.list the tasks performed in data analysis and demonstrate five core components of analytics.**

Tasks Performed in Data Analysis:

1. Data Collection(<https://datacatalog.worldbank.org/dataset/data-collection-methods>) : This is the initial step where relevant data is acquired from diverse sources, including databases, files, surveys, sensors, or APIs. The collected data can be structured (in databases) or unstructured (like text documents).

2. Data Cleaning(<https://towardsdatascience.com/data-cleaning-and-preprocessing-307584497936> ): Identifying and addressing data quality issues is critical to ensure data integrity. This involves handling missing values, dealing with outliers, and resolving inconsistencies to make the data suitable for analysis.

3. Exploratory Data Analysis (EDA)( <https://www.datacamp.com/community/tutorials/tutorial-python-exploratory-data-analysis> ): EDA is a crucial phase that entails in-depth examination of the data's characteristics. It utilizes statistical and visualization techniques to uncover hidden patterns, trends, and outliers in the data, providing insights into its overall structure.

4. Data Transformation (<https://www.tutorialspoint.com/data_mining/dm_data_transformation.html> ): This step involves preparing the data for analysis. It encompasses tasks like data normalization, standardization, or feature engineering. The objective is to make the data suitable for the chosen analysis methods.

5. Data Modeling(<https://www.coursera.org/learn/machine-learning> ): Building statistical, machine learning, or predictive models is the core of data analysis. Models are used to answer specific questions, make predictions, or identify patterns in the data. This phase can range from linear regression to complex neural networks.

6. Hypothesis Testing(<https://stattrek.com/statistics/hypothesis-testing.aspx> ): Hypothesis testing is fundamental to determine whether the observed patterns are statistically significant. It involves formulating hypotheses, collecting data, and using statistical tests to make informed decisions.

7. Data Visualization( <https://serialmentor.com/dataviz/> ): Visual representations such as charts, graphs, and dashboards are created to effectively communicate the data's insights. Visualization aids in conveying complex information in an easily understandable format.

8. Descriptive Statistics(<https://stattrek.com/statistics/descriptive-statistics.aspx> ): Calculating summary statistics like mean, median, variance, and standard deviation helps in summarizing and characterizing the data's central tendencies and variability.

9. Inferential Statistics(<https://www.statistics.com/inferential-statistics/> ): Inferential statistics enables drawing conclusions about populations based on sample data. Techniques like confidence intervals and hypothesis testing are employed to make generalizations.

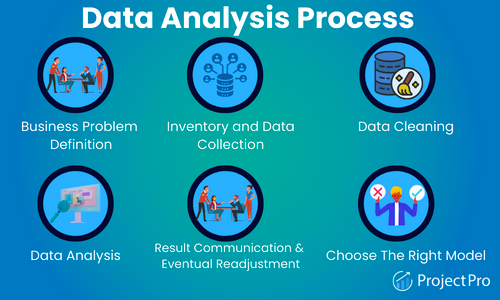
10. Time Series Analysis (<https://www.datacamp.com/community/tutorials/time-series-analysis-tutorial> ): This specialized analysis is used when data is collected over time. Time series analysis helps identify temporal patterns and make forecasts, which is invaluable in fields like finance and climate science.

11. Text Analysis(<https://nlp.stanford.edu/IR-book/html/htmledition/text-analysis-1.html> ): Text analysis involves processing and extracting insights from unstructured text data. Natural language processing (NLP) techniques are employed to extract information from documents, social media, or customer reviews.

12. Cluster Analysis(<https://online.stat.psu.edu/stat505/lesson/14> ): Cluster analysis groups data points with similar characteristics to discover underlying structures within the data. It is widely used in customer segmentation and image processing.

13. Anomaly Detection(<https://dl.acm.org/doi/10.1145/2670529> ): Anomaly detection identifies unusual data points that deviate from the expected patterns. It has applications in fraud detection, network security, and quality control.

14. Data Interpretation(<https://www.coursera.org/learn/data-analysis> ): This is the final stage where analysis results are interpreted and translated into actionable recommendations or decisions. Clear and meaningful insights are drawn from the analysis to inform stakeholders.



Five Core Components of Analytics:

1. Data(<https://towardsdatascience.com/what-is-data-a9cf9c79f664> ): Data is the foundational component of analytics. It includes the raw information collected from various sources. This data can be structured, residing in databases, or unstructured, such as text or multimedia content.

2. Tools and Technologies( <https://www.g2.com/categories/data-analysis> ): Tools and technologies encompass the software, platforms, and hardware used for data processing, analysis, and visualization. Examples include programming languages like Python and R, data analytics platforms, databases, and data visualization tools.

3. Methods and Models( <https://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/lecture-notes/MIT15_097S12_lec02.pdf> ): The heart of data analysis lies in the methods and models. This component covers the diverse statistical, mathematical, and machine learning techniques used to analyze data, make predictions, and derive insights. Selecting the appropriate method or model is critical for the success of the analysis.

4. Domain Knowledge(<https://www.kdnuggets.com/2020/03/role-domain-knowledge-data-science.html> ): Domain knowledge involves expertise in the specific subject area related to the data. It's essential for understanding the context of the data, asking relevant questions, and interpreting the results accurately. This expertise ensures that the analysis is meaningful and aligns with the domain's requirements.

5. Business Impact(<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century> ): Ultimately, the purpose of data analysis is to drive business decisions and create value. The business impact component assesses how the insights derived from the analysis can be applied to achieve specific goals, improve business outcomes, enhance decision-making, and gain a competitive edge in the market. It's the bridge that connects data analysis to real-world applications and results.

**2.Describe five job roles in the overall spectrum of data discovery and understanding. (**[**https://www.simplilearn.com/data-analytics-job-roles-and-responsibilities-article**](https://www.simplilearn.com/data-analytics-job-roles-and-responsibilities-article) **)**

**1. Data Analyst:**

* Data analysts are responsible for collecting, cleaning, and analyzing data to extract actionable insights. They use statistical and analytical tools to understand trends, patterns, and relationships within the data. Data analysts are often involved in creating reports and visualizations to communicate their findings.

**2. Business Intelligence (BI) Analyst:**

* BI analysts focus on translating data into insights that drive business decisions. They design and maintain data dashboards, perform ad-hoc analysis, and collaborate with stakeholders to provide data-driven recommendations. BI analysts play a crucial role in helping organizations understand their data in the context of business objectives.

**3. Data Scientist:**

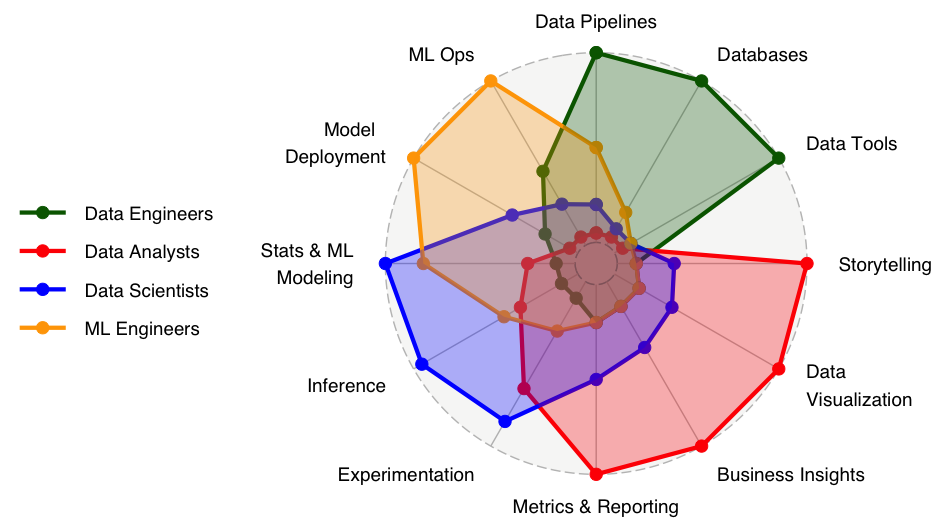
* Data scientists are skilled in advanced analytics, machine learning, and predictive modeling. They dive deep into data to discover hidden insights, build predictive models, and develop algorithms. Data scientists are responsible for uncovering complex patterns and solving challenging problems with data.

**4. Data Engineer:**

* Data engineers are responsible for the architecture, construction, and maintenance of data systems and pipelines. They ensure data is collected, stored, and made accessible to analysts and data scientists. Data engineers play a crucial role in the data discovery process by providing clean and reliable data for analysis.

**5. Data Architect:**

* Data architects design and manage the overall structure of data systems, including databases and data warehouses. They create the blueprint for how data should be organized and stored. Data architects help organizations understand the data's structure and relationships, which is essential for effective data discovery.



**3.Describe five key areas in the data analysis process.(** [**https://towardsdatascience.com/**](https://towardsdatascience.com/) **)**

**. Data Collection:**

* Data collection is the foundation of data analysis. It involves gathering raw data from various sources, such as databases, files, surveys, sensors, web scraping, or APIs. The collected data can be structured (in relational databases) or unstructured (text, images, videos), and it can come in different formats and scales. Data collection often requires careful planning to ensure that the data collected is relevant and of high quality. It's essential to consider data privacy and ethics when collecting data, especially in the context of personal or sensitive information.

**2. Data Cleaning and Preprocessing:**

* Once data is collected, it usually needs cleaning and preprocessing. This step involves identifying and addressing data quality issues. Common tasks include dealing with missing values, handling outliers, resolving inconsistencies, and standardizing data formats. Data preprocessing is critical to ensure that the data is accurate and consistent before analysis. The quality of the analysis results heavily depends on the quality of the data, making this step crucial in the process.

**3. Exploratory Data Analysis (EDA):**

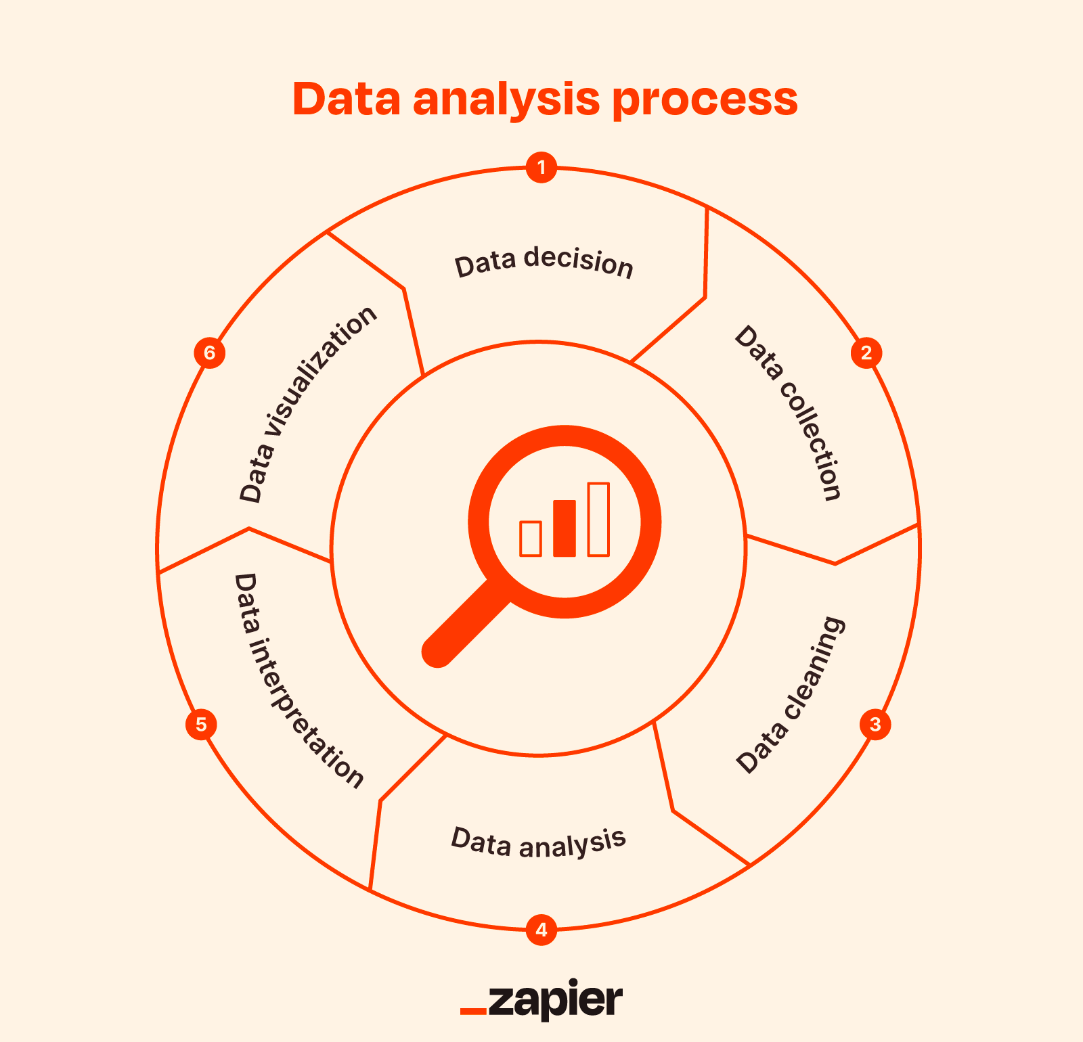
* Exploratory Data Analysis is a crucial phase where the data analyst aims to understand the data's characteristics and identify initial patterns or trends. Techniques in EDA include summary statistics, data visualization, and data mining. Data visualization tools like histograms, scatter plots, and box plots can reveal distribution, correlations, and anomalies in the data. EDA helps generate hypotheses, uncover insights, and guides the selection of appropriate analysis methods.

**4. Data Modeling and Analysis:**

* Data modeling is the heart of data analysis. This step involves building and applying statistical, mathematical, or machine learning models to the data. It is where you answer specific questions, make predictions, identify patterns, or test hypotheses. The choice of the modeling technique depends on the nature of the data and the objectives of the analysis. Regression analysis, clustering, classification, and time series forecasting are examples of modeling techniques used in data analysis.

**5. Data Visualization and Communication:**

* Effective data visualization is essential to communicate the findings and insights from the analysis. This includes creating clear and meaningful visual representations such as charts, graphs, and dashboards. Data visualization is not only about making the analysis results understandable but also about making them actionable. Effective communication is a key skill in data analysis, as it allows stakeholders to make informed decisions based on the analysis results. Dashboards and interactive visualizations are often used to provide real-time insights to decision-makers.



**4.Discuss ten difference between fact and dimension table.**

Fact and dimension tables are essential components of a star or snowflake schema in data warehousing and are used to organize and store data for business intelligence and reporting. Here are ten differences between fact and dimension tables, elaborated with sources where you can find further information:

1. Definition:

- Fact Table: A fact table contains numerical data (facts) that can be analyzed or measured, such as sales revenue, quantities, or profit.

- Dimension Table: Dimension tables contain descriptive textual or categorical data that provide context to the facts, like customer names, product categories, or dates.

Source: [What are Fact and Dimension Tables?](<https://www.vertabelo.com/blog/technical-articles/data-warehousing-dictionary-what-are-fact-and-dimension-tables> )

2. Data Type:

- Fact Table: Contains quantitative data typically stored as numbers.

- Dimension Table: Contains qualitative data stored as text.

Source: [Data Warehousing Concepts](<https://www.talend.com/resources/what-is-data-warehousing/> )

3. Size:

- Fact Table: Fact tables are usually larger in size since they store numeric data that can accumulate over time.

- Dimension Table: Dimension tables are relatively smaller as they contain categorical and textual data.

Source: [Fact Tables vs Dimension Tables](<https://www.sqlchick.com/entries/2011/8/14/fact-tables-vs-dimension-tables> )

4. Granularity:

- Fact Table: Fact tables have a finer level of granularity and are often associated with specific events or transactions.

- Dimension Table: Dimension tables have a coarser level of granularity, typically containing attributes that apply to a group of facts.

Source: [Data Warehouse Fact Tables and Dimension Tables](<https://docs.oracle.com/en/database/oracle/oracle-database/21/dwhsg/data-warehouse-fact-tables-and-dimension-tables.html> )

5. Primary Key:

- Fact Table: Fact tables have a composite primary key that consists of foreign keys from dimension tables.

- Dimension Table: Dimension tables have a single primary key, typically a surrogate key.

Source: [Fact and Dimension Tables](<https://learn.microsoft.com/en-us/sql/analysis-services/multidimensional-models/multidimensional-models-lesson-1-4-creating-a-fact-table-and-relationship-lesson?view=sql-server-ver15> )

6. Foreign Keys:

- Fact Table: Contains foreign keys that reference dimension tables, creating relationships between facts and dimensions.

- Dimension Table: Contains no foreign keys, as it serves as the primary reference source for other tables.

Source: [Understanding Fact and Dimension Tables](<https://www.kimballgroup.com/1997/07/28/fact-tables-and-dimension-tables/> )

7. Aggregations:

- Fact Table: Aggregations, like sums or averages, are often applied to fact table data to provide summarized views.

- Dimension Table: Aggregations are not applied to dimension table data; they provide context for facts.

Source: [Aggregation Design](<https://www.kimballgroup.com/1996/02/our-early-years-part-i-our-first-data-warehouse/> )

8. Volatility:

- Fact Table: Fact tables are more volatile as they can change frequently due to the addition of new facts.

- Dimension Table: Dimension tables are relatively stable and change infrequently.

Source: [Star Schema Design for Data Warehousing](<https://docs.sas.com/documentation/casml/8.2/casml_dgen_star_schema_overview.htm> )

9. Measures:

- Fact Table: Contains measures, which are numeric values that can be aggregated (e.g., sales amount, profit, or quantity sold).

- Dimension Table: Contains attributes, which provide context for measures (e.g., customer name, product category, or date).

Source: [Fact and Dimension Tables in Data Warehousing](<https://intellipaat.com/blog/tutorial/data-warehousing/fact-and-dimension-tables/> )

10. Usage in Queries:

- Fact Table: Fact tables are primarily used in aggregation queries, where numerical data is summarized.

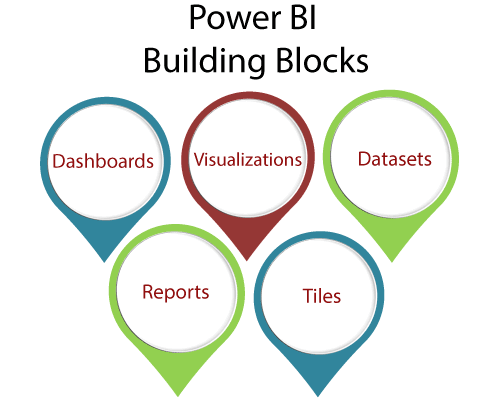
- Dimension Table: Dimension tables are used to filter, group, or slice and dice data in queries.

Source: [Star Schema vs. Snowflake Schema in Data Warehousing](<https://www.talend.com/resources/star-schema-vs-snowflake-schema/> )

**5. Describe Building blocks of Power BI. (**[**https://docs.microsoft.com/en-us/power-bi/**](https://docs.microsoft.com/en-us/power-bi/) **)**

Power BI is a powerful business intelligence and data visualization tool developed by Microsoft. It provides a suite of building blocks and components that work together to create interactive reports and dashboards. Here's an elaboration of the building blocks of Power BI:

1. **Power Query**: Power Query is the data transformation and preparation component of Power BI. It allows you to connect to various data sources, perform data cleansing, transformation, and shaping operations. Power Query enables you to extract, transform, and load (ETL) data from multiple sources into a format suitable for analysis.
2. **Data Model**: The data model is at the core of Power BI. It's a powerful and flexible data structure that holds your data and relationships between tables. You can create relationships, calculations, and hierarchies within the data model. The model allows for data integration from multiple sources and provides a foundation for building visualizations.
3. **DAX (Data Analysis Expressions)**: DAX is a formula language used for creating custom calculations and aggregations in Power BI. It allows you to define calculated columns, measures, and calculated tables. DAX functions enable complex calculations and analysis within your data model.
4. **Power BI Desktop**: Power BI Desktop is the authoring tool that enables you to design reports and dashboards. It's where you import data, create data models, design visuals, and define interactions. Power BI Desktop provides a user-friendly, drag-and-drop interface for report development.
5. **Power BI Service**: The Power BI Service is the cloud-based platform for publishing, sharing, and collaborating on Power BI reports and dashboards. Once you've created a report in Power BI Desktop, you can publish it to the Power BI Service, where you can access it from any device with an internet connection. The service also allows you to share reports with others and collaborate on them.
6. **Visualizations**: Power BI offers a wide range of visualizations, such as bar charts, line charts, pie charts, tables, maps, and custom visuals. These visualizations allow you to represent your data in a meaningful and interactive way. You can customize visuals, add drill-through actions, and create cross-filtering interactions.
7. **Dashboards**: Dashboards are a collection of related visualizations and reports in a single canvas. You can pin specific visuals and report pages to a dashboard to create a unified view of your data. Dashboards provide a high-level overview and interactivity with your data.
8. **Gateways**: Power BI Gateways are used to connect on-premises data sources to the Power BI Service. They facilitate data refresh from sources located within your organization's network, ensuring your reports stay up to date.
9. **Power Automate (formerly Microsoft Flow)**: Power BI integrates with Power Automate to automate processes and trigger actions based on data-driven events. You can use Power Automate to set up notifications, alerts, or workflows based on changes in your Power BI data.
10. **Power BI Mobile**: Power BI offers mobile apps for various platforms (iOS, Android, and Windows) to access your reports and dashboards on mobile devices. The mobile app provides an optimized experience for interacting with your data on the go.

****