**Part A**

**Task 1: Data Analytics Foundation**

**Data Analytics:** Data Analytics is the process of collecting, organizing, analyzing, and interpreting various forms of data to gain useful insights for making informed decisions. As I am an urban planner, in the context of urban and regional planning, data analytics helps planners understand spatial patterns, population trends, transportation needs, land use changes, and environmental impacts.

**Two historical milestones:**

1. Rise of Relational Databases (1970s): Edgar F. Codd introduced the relational database model (SQL-based). It allowed structured storage, retrieval, and querying of data, which is essential for organized data analysis.
2. Rise of Machine Learning & AI (2010s): Advances in algorithms and computing power enabled machine learning models to analyze data automatically. It led to predictive analytics and automation in fields like healthcare, finance, and marketing.

**Three current trend in data analysis:**

1. **AI-powered Data Analytics:** AI and machine learning are becoming integral to data analysis, enabling automation of tasks, faster insights, and the development of more sophisticated models. Augmented analytics is a type which combines AI and ML with natural language processing (NLP), is automating data preparation, visualization, and analysis, making it easier for non-technical users to gain insights. Generative AI is also emerging in this sector. AI is also driving the evolution of predictive analytics, moving beyond simple statistical methods to incorporate causal inference and prescriptive analytics.
2. **Real-time Analytics:** The demand for real-time insights is growing, driven by the need for faster decision-making and improved customer experiences.
3. **Natural Language Processing (NLP):** NLP is enhancing data analysis by enabling users to interact with data using natural language, making it easier for non-technical users to ask questions and get insights. NLP is also being used to automate tasks like data extraction, summarization, and sentiment analysis.

**Task 2: Analytics Process Mapping**

End -to-end data analytics process:

1. **Data Collection:** This is first step in data analytics process.Data is gathered from both primary and secondary sources like surveys, sensors, databases, web scraping, APIs, or files (CSV, Excel). The quality of analysis depends on the quality and relevance of collected data.
2. **Data Cleaning (Preprocessing):** Data cleaning means Removing duplicates, handling missing values, fixing errors, converting formats (e.g., dates), and filtering outliers. Dirty and inconvenient data leads to misleading insights. Cleaning ensures data is accurate, consistent, and usable.
3. **Data Exploration (EDA – Exploratory Data Analysis):** Use statistics and visualizations to understand data patterns, distributions, and relationships. Helps identify trends, spot anomalies, and guide analysis strategy.
4. **Data Analysis / Modeling:** Apply statistical tests, machine learning, or other analytical methods to extract insights or make predictions. This is where real value is created—turning raw data into useful knowledge or forecasts.
5. **Reporting & Decision Making:** Create dashboards, charts, or written reports to share findings with stakeholders. Makes insights actionable. Helps organizations make data-driven decisions and solve problems.

**Task 3: Terminologies and Tools**

| **Terms** | **Definitions** | **Importance** |
| --- | --- | --- |
| Data types | Categories that define the nature of data (e.g., integer, float, string). | Helps in understanding and processing data correctly during analysis. |
| Variables | Named storage in a program that holds data that can change during execution. | Essential for storing and manipulating data in programming and analysis. |
| Metrices | Quantitative measurements used to track performance or progress. | Provides concrete numbers to evaluate and compare results. |
| KPI’s | Key Performance Indicators – strategic metrics that reflect business success. | Crucial for measuring goal achievement and guiding decisions. |
| Structured Data | Data organized in rows and columns, like in databases or spreadsheets. | Easy to store, query, and analyze using tools like SQL or Pandas. |
| API | Application Programming Interface – a tool to connect different software apps. | Enables automated data access and integration from external sources. |
| SQL | Structured Query Language – used to manage and query relational databases. | Fundamental tool for accessing, modifying, and analyzing structured data. |
| Data Warhouse | Centralized storage system for large-scale, structured data. | Supports complex queries and business intelligence reporting. |
| Pandas | A Python library for data manipulation and analysis. | Widely used in data cleaning, transformation, and EDA in Python. |
| Tidyverse | A collection of R packages for data science (e.g., dplyr, ggplot2). | Simplifies and standardizes data analysis workflows in R. |
| Matplotlib | A Python library for data visualization. | Helps in creating charts and graphs to understand data visually. |
| Jupyter Notebook | An interactive coding environment for Python and data science. | Allows combining code, output, and notes in one place for reproducible work. |

**Part B: Practical Tasks**

**Task 4: Data collection exploration**

**Data source and format:** The dataset is derived from Kaggle. <https://www.kaggle.com/datasets/prevek18/ames-housing-dataset?resource=download>

The file is in CSV (Comma-Separated Values) format, which is a structured text file. Each row represents a house sale record, and each column represents a feature such as sale price, number of rooms, year built, etc.

**Data structure:** The dataset is structured as it contains well-defined rows and columns, with headers identifying each variable (e.g., LotArea, YearBuilt, SalePrice, etc.). The data types are a mix of numerical, categorical, and ordinal variables, but overall it conforms to a tabular, structured schema.

**Ethical consideration:** While the dataset doesn’t contain names or direct identifiers, combining it with other data could risk re-identification (e.g., location-based info like neighborhoods).

**Task 5 &6: Data profiling and quality check & Data cleaning**

**Task 7: Tool Evaluation and Reflection**

**Why I chose Jupyter notebook:** I chose Jupyter Notebook because it provides an interactive environment for data exploration, cleaning, and visualization. It allows me to combine code, visual output, and commentary in a single document, which makes the data cleaning process more transparent and easier to document and debug.

**Strengths:**

* Codes are easy and simple
* I could see results immediately after running each code block, which helped in checking progress step by step.
* Handling null values for big database is easy to handle

**Limitations:**

* Jupyter becomes slow for very large datasets.

**Task 8: Mini Case Study – Application of Data Analytics in Sports**

**Real-World Application:** Player Performance Optimization in Football

One of the most impactful applications of data analytics in sports is the optimization of player performance and team strategy in professional football (soccer). Football teams actively use analytics to gain a competitive edge in matches, training, and transfers.

Using tracking devices, video analysis systems (like Opta or StatsBomb), and GPS-enabled wearables, analysts collect vast amounts of data during training sessions and matches. This data helps teams make data-driven decisions that improve player fitness, tactical awareness, and overall team performance.

**What Kind of Data Is Used?**

**Player Tracking Data:**

* Position, speed, acceleration, distance covered
* Heatmaps showing zones covered during the match
* Collected using GPS devices or camera systems

**Event Data:**

* Passes, shots, tackles, interceptions, fouls
* Time and location of each event on the pitch
* Used to evaluate tactical movements and decisions

**Biometric and Fitness Data:**

* Heart rate, VO2 max, muscle strain, recovery time
* Used to monitor player fatigue and injury risk
* Collected using wearables and medical assessments

**Match Statistics:**

* Possession percentage, expected goals (xG), pass accuracy
* Aggregated to assess overall team or individual efficiency

**What Decisions Are Made From the Data:**

**Training Adjustments:** Coaches personalize training plans based on player fatigue, workload, and performance metrics to avoid injuries and maximize fitness.

**Tactical Strategy:** Managers use heatmaps and pass networks to adjust team formation, pressing intensity, and attacking/defending strategies before or during matches.

**Recruitment and Transfers:** Data is used to identify undervalued players or rising talents by comparing key metrics like pass accuracy, defensive duels, or chance creation across leagues.

**Injury Prevention:** Biometric monitoring helps the medical team detect early signs of overtraining or injury risk, enabling timely rest or treatment.

**Opponent Analysis:** Teams analyze opponent patterns (e.g., where they lose possession, where they attack from) to prepare counter-strategies.