Assignment 1: Data Analysis Fundamentals

This assignment covers the first three weeks of the course — reading and inspecting data, exploratory data analysis (EDA), and data wrangling. Each idea below provides a unique dataset, a set of EDA questions, and data cleaning/wrangling tasks.

1. Olympic Athletes Performance Analysis

Dataset: 120 Years of Olympic History – Athletes and Results (https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results)

Instructions:

- Load and inspect the dataset using Pandas functions such as .head(), .info(), .describe(), and .shape.

- Explore athlete demographics, medal distribution, and performance trends over time.

Exploratory Data Analysis (EDA) Questions:

- Which countries have won the most medals overall and by sport?

- How has athlete participation changed over the years?

- Is there a relationship between athlete height, weight, and winning medals?

- Which sports have the most male vs female participation?

- Which Olympic year saw the highest number of athletes or countries?

Data Wrangling & Cleaning Tasks:

- Handle missing values in 'Age', 'Height', and 'Weight' columns.

- Convert categorical columns such as 'Sex' and 'Medal' to appropriate data types.

- Remove duplicates or invalid entries (e.g., athletes listed multiple times per event).

- Fix inconsistencies in 'Sport' or 'NOC' naming formats.

- Create a new column for 'BMI' to explore physical attributes of athletes.

---

2. Netflix Movies & TV Shows

Dataset: Netflix Titles Dataset (https://www.kaggle.com/shivamb/netflix-shows)

Instructions:

- Load and inspect the dataset to understand the structure and data types.

- Analyze trends in release years, genres, and countries.

Exploratory Data Analysis (EDA) Questions:

- Which country produces the most Netflix content?

- How many movies vs TV shows are there?

- What are the most common genres?

- What is the trend of content production over the years?

- Which directors have the most shows/movies on Netflix?

Data Wrangling & Cleaning Tasks:

- Clean the 'date_added' column by converting it to datetime format.

- Fill missing values in 'director' and 'cast' with 'Unknown'.

- Split 'listed_in' (genres) into separate columns or lists.

- Standardize country names to consistent format.

- Remove any duplicate titles.

---

3. World Happiness Report

Dataset: World Happiness 2021 Dataset
(https://www.kaggle.com/datasets/mathurinache/world-happiness-report)

Instructions:

- Load the dataset and identify the key variables related to happiness.

- Explore the relationship between happiness score and contributing factors.

Exploratory Data Analysis (EDA) Questions:

- Which regions have the highest and lowest happiness scores?

- Which factors (GDP, freedom, social support) most correlate with happiness?

- How does happiness vary by continent?

- What is the trend of happiness over the years?

Data Wrangling & Cleaning Tasks:

- Handle missing values in numeric columns such as GDP and Healthy Life Expectancy.

- Standardize region names to be consistent.

- Detect and remove outliers in GDP per capita or Score using boxplots.

- Fix data types (ensure numeric columns are properly formatted).

---

4. Students Performance in Exams

Dataset: Students Performance Dataset (https://www.kaggle.com/spscientist/students-performance-in-exams)

Instructions:

- Load and inspect the dataset to identify columns related to performance.

- Analyze relationships between demographics and exam scores.

Exploratory Data Analysis (EDA) Questions:

- Which gender performs better in each subject?

- Does test preparation improve performance?

- What is the relationship between parental education level and student performance?

- Are there correlations between math, reading, and writing scores?

Data Wrangling & Cleaning Tasks:

- Check for and remove duplicate rows.

- Handle any missing values in the score columns.

- Encode categorical columns ('gender', 'parental level of education', etc.).

- Remove outliers with impossible scores (e.g., >100 or <0).

---

5. Airbnb Listings (NYC 2019)

Dataset: Airbnb NYC 2019 (https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data)

Instructions:

- Inspect the dataset and understand the columns related to listings, prices, and reviews.

- Analyze pricing patterns across neighborhoods.

Exploratory Data Analysis (EDA) Questions:

- Which neighborhoods have the highest average price?

- How do prices vary by room type?

- Is there a relationship between the number of reviews and price?

- What is the distribution of listings by host?

Data Wrangling & Cleaning Tasks:

- Convert 'last_review' to datetime format.

- Fill missing values in 'reviews_per_month' with 0.

- Remove listings with missing critical data (e.g., price or neighborhood).

- Detect and remove extreme outliers in 'price' using IQR.

---

6. COVID-19 World Data

Dataset: (https://www.kaggle.com/datasets/imdevskp/corona-virus-report?select=worldometer_data.csv)

Instructions:

- Load and inspect the dataset to understand columns such as Country/Region, TotalCases, TotalDeaths, TotalRecovered, ActiveCases, and TotalTests- Analyze global and regional patterns.

- Perform exploratory data analysis (EDA) to uncover global and regional COVID-19 patterns.

- Identify relationships between variables such as cases, deaths, recovery rate, and testing levels.

Exploratory Data Analysis (EDA) Questions:

- Which countries have the highest total cases, deaths, and recoveries?

- What is the recovery and fatality rate distribution across countries?

- Is there a correlation between total tests conducted and total confirmed cases?

- Which continents show the highest average number of cases and deaths per million population?

- What trends can be observed between population size and infection rate?

- Which countries have managed to control the spread most effectively (low active cases vs. high recovery rate)?

Data Wrangling & Cleaning Tasks:

- □ Handle missing values in numeric columns such as TotalCases, TotalDeaths, and TotalTests using median or mean imputation.

□ Convert percentage or string-formatted columns (like “1,234,567”) to numeric using str.replace() and pd.to_numeric().

 Standardize country names and ensure consistent capitalization.

 Detect and handle outliers in cases and deaths using the IQR or Z-score method.

Remove duplicate or invalid rows (e.g., aggregate entries like “World” or “Total:”).

Ensure all numeric columns have correct data types

---

7. Global CO_2 Emissions

Dataset: Global CO_2 Emissions Dataset
(https://www.kaggle.com/datasets/yoannboyere/co2-ghg-emissionsdata)

Instructions:

- Analyze emission patterns across countries and years.

- Visualize changes in emissions over time.

Exploratory Data Analysis (EDA) Questions:

- Which countries have the highest and lowest CO_2 emissions?

- What is the global trend of CO_2 emissions?

- How does GDP relate to emissions?

- Which sectors contribute most to emissions?

Data Wrangling & Cleaning Tasks:

- Fix missing or inconsistent country names.

- Handle missing numeric values for emission columns.

- Remove unrealistic zero or negative emission entries.

- Convert year columns to integer format.

---

8. YouTube Channel Statistics Analysis

Dataset: YouTube Statistics Dataset
(https://www.kaggle.com/datasets/datasnaek/youtube-new)

Instructions:

- Load and inspect the dataset for trending YouTube videos.

- Explore the relationships between views, likes, dislikes, and comment counts.

Exploratory Data Analysis (EDA) Questions:

- Which videos have the highest number of views and likes?

- Is there a correlation between likes and dislikes?

- Which categories receive the most engagement (views/comments)?

- How does video publishing time affect performance?

- What are the most common tags among trending videos?

Data Wrangling & Cleaning Tasks:

- Handle missing values in columns such as 'tags' or 'description'.

- Convert 'publish_time' to datetime and extract date and hour.

- Remove duplicate video entries across multiple days.

- Standardize text columns (convert to lowercase, remove special characters).

- Detect and handle outliers in views, likes, and comments.

---

9. E-commerce Sales Data Analysis

Dataset: Unlock Profits with E-commerce Sales Data (https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data)

Instructions:

- Load the CSV dataset and inspect its structure and columns.

- Perform sales analysis to identify key revenue drivers and profitability trends.

Exploratory Data Analysis (EDA) Questions:

- Which product categories generate the most revenue and profit?

- How do sales and profit vary by region and customer segment?

- Is there a relationship between discount rate and profit margin?

- Which month or quarter records the highest sales volume?

- Who are the top-performing sales representatives or cities?

Data Wrangling & Cleaning Tasks:

- Handle missing or zero values in 'Sales', 'Profit', and 'Discount' columns.

- Convert date columns to datetime and extract month and year.

- Detect and remove duplicate transaction records.

- Handle negative profit or sales entries if found.

- Standardize categorical columns such as 'Category' and 'Region'.

---

10. Spotify Songs Dataset

Dataset: Spotify Tracks Dataset
(https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db)

Instructions:

- Analyze features influencing song popularity.

- Explore distributions of tempo, energy, and danceability.

Exploratory Data Analysis (EDA) Questions:

- Which features most correlate with popularity?

- How does popularity vary by genre or release year?

- What is the distribution of tempo and energy?

- Which artists appear most frequently?

Data Wrangling & Cleaning Tasks:

- Handle missing values in 'tempo' and 'duration_ms'.

- Convert 'release_date' to datetime and extract year.

- Remove duplicate tracks.

- Detect and remove popularity outliers.

---

11. Global Temperature and Climate Change

Dataset: Global Temperature Time Series

Instructions:

- Load and inspect global average temperature records from 1850 onward.

- Analyze long-term temperature trends and variations by region.

Exploratory Data Analysis (EDA) Questions:

- How has the global temperature changed over the past 150 years?

- Which regions show the fastest warming trends?

- What are the annual and seasonal patterns of temperature change?

Data Wrangling & Cleaning Tasks:

- Parse date column and extract year/month.
- Handle missing temperature values using interpolation.
- Remove duplicates and invalid readings.
- Aggregate data by decade for trend visualization.

---

12. Heart Disease Prediction Dataset

Dataset: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

Instructions:

- Explore risk factors (age, cholesterol, blood pressure) influencing heart disease.

Exploratory Data Analysis (EDA) Questions:

- What is the age distribution of patients with heart disease?
- Which variables correlate most strongly with heart disease presence?
- Does gender or chest-pain type affect diagnosis rate?

Data Wrangling & Cleaning Tasks:

- Handle missing or inconsistent numeric values.
- Encode categorical columns (`sex`, `cp`, `thal`).
- Detect and remove outliers using IQR.
- Normalize numeric features for analysis.

---

13. Employee Attrition and Satisfaction

Dataset: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Instructions:

- Analyze why employees leave the company and what factors influence attrition.

Exploratory Data Analysis (EDA) Questions:

- What is the attrition rate by age, gender, and department?
- How does monthly income relate to attrition?
- Which job roles have the highest turnover?

Data Wrangling & Cleaning Tasks:

- Handle missing categorical data (e.g., Department).
- Convert binary columns (`Attrition`) to numeric.
- Standardize column names (snake_case).
- Remove outliers in `MonthlyIncome`.

---

14. Global Education Statistics

Dataset: https://www.kaggle.com/datasets/andrewmvd/global-education-statistics

Instructions:

- Explore literacy rates, school enrollment, and education spending.

Exploratory Data Analysis (EDA) Questions:

- Which countries invest most in education?
- How do literacy rates vary by region?
- What is the relationship between education spending and GDP?

Data Wrangling & Cleaning Tasks:

- Fill missing values with median or regional averages.
- Convert columns to numeric where necessary.
- Remove duplicate country records.
- Standardize region names.

---

15. Startups Investment Data

Dataset: https://www.kaggle.com/datasets/arindam235/startup-investments-crunchbase

Instructions:

- Analyze funding trends across industries and countries.

Exploratory Data Analysis (EDA) Questions:

- Which industries receive the most investment?
- What are the top countries by funding volume?
- How does funding change over years?

Data Wrangling & Cleaning Tasks:

- Handle missing funding amounts and dates.
- Convert funding columns to numeric (remove $ symbols).
- Remove duplicate company entries.
- Extract year from date for temporal analysis.

---

16. FIFA Players Statistics

Dataset: https://www.kaggle.com/datasets/stefanoleone992/fifa-21-complete-player-dataset

Instructions:

- Explore attributes (overall, potential, wages) and performance trends.

Exploratory Data Analysis (EDA) Questions:

- Which countries produce the best players (highest overall)?
- What is the relationship between age, potential, and market value?
- Which clubs have the most valuable squads?

Data Wrangling & Cleaning Tasks:

- Handle missing `Value` or `Wage` values (convert from string to float).

- Remove outliers in `Age` and `Overall`.
- Standardize positions and nations.
- Detect duplicate player records.

---

17. Health Insurance Cost Analysis

Dataset: https://www.kaggle.com/datasets/mirichoi0218/insurance

Instructions:

- Load and inspect the dataset to understand key variables like `age`, `sex`, `bmi`, `children`, `smoker`, `region`, and `charges`.
- Perform exploratory data analysis to understand which factors influence insurance costs.
- Investigate the relationships between demographic and lifestyle variables and the total charges.

Exploratory Data Analysis (EDA) Questions:

- What is the average insurance charge, and how does it vary by region?
- How does smoking status affect medical charges?
- What is the relationship between BMI and charges — are higher BMI values associated with higher costs?
- How do charges differ by gender and number of children?

- Are there any correlations among age, BMI, and charges?

Data Wrangling & Cleaning Tasks:

- Check for and handle missing values in all numeric and categorical columns.
- Detect and handle outliers in the `charges` and `bmi` columns using IQR or z-score methods.
- Convert categorical variables (`sex`, `smoker`, `region`) into numeric format using label or one-hot encoding.

- Standardize column names and ensure consistent casing.

---

18. Supermarket Sales Analysis

Dataset: https://www.kaggle.com/datasets/faresashraf1001/supermarket-sales

Instructions:

- Study branch-wise performance, payment types, and customer behavior.

Exploratory Data Analysis (EDA) Questions:

- Which branch has the highest average sales?
- What is the most common payment method?
- How do ratings vary by customer type and gender?

Data Wrangling & Cleaning Tasks:

- Convert `Date` to datetime and extract month.
- Handle missing numeric values in `gross_income`.
- Standardize categorical labels.
- Remove duplicate invoices.

---

19. World Population Data

Dataset: https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset

Instructions:

- Explore global population growth and distribution.

Exploratory Data Analysis (EDA) Questions:

- Which countries have the largest and smallest populations?
- How does population growth vary by continent?
- What is the relationship between density and urban population?

Data Wrangling & Cleaning Tasks:

- Convert year columns to numeric.
- Handle missing values in population metrics.
- Detect and remove outliers in `Density`.
- Standardize continent/country names.

---

20. Road Accidents Data Analysis

Dataset: https://www.kaggle.com/datasets/devansodariya/road-accident-united-kingdom-uk-dataset

Instructions:

- Analyze accident frequency, location, and contributing factors.

Exploratory Data Analysis (EDA) Questions:

- Which regions report the most accidents?
- What time of day sees the highest number of accidents?
- How does weather condition affect accident severity?

Data Wrangling & Cleaning Tasks:

- Convert `Date` and `Time` columns to datetime.
- Handle missing values in `Weather_Conditions` and `Road_Surface`.
- Remove duplicate or incomplete records.
- Create new column for `Hour` to analyze hourly trends.