Pandas Practice Tasks Based on customers-10000.csv

Simple, Intermediate, and Hard Tasks

Simple Tasks

S1. Load the CSV file and display the first 5 rows.

S2. Show the number of rows and columns in the dataset.

S3. Print all column names.

S4. Show summary info using `df.info()` and statistical summary using `df.describe()`.

S5. Select and display only the `First Name` column.

S6. Select and display the `Country` and `Email` columns together.

S7. Filter and show all customers from the United States.

S8. Filter and show all customers whose email ends with ".org".

S9. Create a new column called `Full Name` combining first and last names.

S10. Convert the `Subscription Date` column to datetime.

S11. Find the earliest and latest subscription dates.

S12. Count the number of missing values in each column.

Intermediate Tasks

I1. Clean phone numbers by removing dots, parentheses, dashes, spaces, and extensions.

I2. Extract domain names from emails and count the top 10 most frequent domains.

I3. Extract top-level domains (TLDs) such as "com", "org", "net" and count their frequency.

I4. Group customers by country and show customer count and percentage distribution.

I5. Extract the year from subscription dates and count customers per year.

I6. Detect duplicate customers using email, phone number, and (First Name + Last Name + Phone).

I7. Calculate the length of each company name and find the longest and shortest names.

I8. Count how many customers do not have a Phone 2.

I9. Find all customers whose phone numbers include an extension (like "x123" or "ext456").

I10. Group customers by month (YYYY–MM) and analyze monthly signup trends.

I11. Detect outlier subscription dates using the IQR method.

Hard Tasks

H1. Standardize phone numbers into an international format:

- Extract country code if present.
- Extract main number.
- Extract extension separately.

H2. Identify inconsistent or misspelled country names and standardize them.

H3. Perform customer segmentation (clustering) using encoded features such as country, city, company name length, and signup month.

H4. Build a small EDA report including:

- Customer count by country
- Monthly signup trend
- Top email domains
- Outlier detection summary

H5. Create a correlation-style analysis by encoding categorical features and examining relationships between email domain, country, company size, etc.

H6. Build a simple model to predict the customer's country based on email domain, company name length, and phone number patterns.

H7. Construct a mini CRM-style dashboard (matplotlib/plotly) showing:

- Total customers
- Monthly new customers
- Top 10 countries
- Top email domains
- Duplicate count
- Outlier count

H8. Parse phone numbers into structured fields and analyze common phone number patterns by country.

H9. Build an anomaly detection model using Isolation Forest for unusual phone formats, strange company names, or abnormal dates.

H10. Create a full data-cleaning pipeline:

- Loading
- Missing value handling

- Data type conversion
- Normalization of text fields
- Feature extraction
- Exporting cleaned dataset