Exercise 1:

Data analysis is the process of looking at data carefully to find useful information, draw conclusions, and make decisions. It turns raw data into insights that can help in making better choices. This process is essential for businesses and researchers today.

Data analysis is important because it helps organizations improve their efficiency, satisfy customers, and innovate. It is used in many industries. For example, in healthcare, data analysis helps predict disease outbreaks and create personalized treatments. Financial institutions use it to detect fraud and make smart investments, while marketers use it to understand consumer behavior and create targeted campaigns.

Data analysis is also crucial in scientific research and public policy. It helps researchers prove their ideas and policymakers create informed decisions. With the rise of big data and advanced technologies, data analysis can now provide deeper insights and faster results. As the amount of data grows, good data analysis becomes even more important, making it a key part of modern life and work.

Exercise 2, Exercise 3:

Dataset Description: Average Hours Spent Sleeping per Day.

Columns:

1. **Index:** Sequential identifier for each row.
2. **Year:** The year in which the data was collected.
3. **Period:** The time period for the data collection, specified as annual.
4. **Avg hrs per day sleeping:** The average number of hours spent sleeping per day.
5. **Standard Error:** The standard error associated with the average hours of sleep.
6. **Type of Days:** Specifies that the data includes all days of the week.
7. **Age Group:** The age group of the individuals included in the dataset, specified as 15 years and over.
8. **Activity:** The activity being measured, which is sleeping.
9. **Sex:** The dataset includes data for both sexes combined.

Categorization of Columns as Quantitative or Qualitative:

1. **Index:** Quantitative
    - This column contains numerical values that identify each row uniquely. It represents discrete numerical data.
2. **Year:** Quantitative
    - This column contains numerical values indicating the year of data collection. It represents time in a numerical format, making it quantitative.
3. **Period:** Qualitative

    - This column contains descriptive data indicating the time period, specified as "Annual." It categorizes the data rather than measuring it numerically.
4. **Avg hrs per day sleeping:** Quantitative
    - This column contains numerical values representing the average number of hours spent sleeping per day. It is a continuous measurement.
5. **Standard Error:** Quantitative
    - This column contains numerical values representing the standard error associated with the average hours of sleep. It is a continuous measurement.
6. **Type of Days:** Qualitative
    - This column contains descriptive data indicating the type of days included in the dataset (e.g., "All days"). It categorizes the data rather than measuring it numerically.
7. **Age Group:** Qualitative
    - This column contains descriptive data indicating the age group of individuals included in the dataset (e.g., "15 years and over"). It categorizes the data rather than measuring it numerically.
8. **Activity:** Qualitative
    - This column contains descriptive data indicating the activity being measured (e.g., "Sleeping"). It categorizes the data rather than measuring it numerically.
9. **Sex:** Qualitative
    - This column contains descriptive data indicating the sex of the individuals included in the dataset (e.g., "Both"). It categorizes the data rather than measuring it numerically.

Dataset Description: Mental health Depression disorder Data.

Columns:

1. **Index:** Sequential identifier for each row.
2. **Entity:** The name of the country or region (Afghanistan in this case).
3. **Code:** The three-letter country code (AFG for Afghanistan).
4. **Year:** The year in which the data was collected.
5. **Schizophrenia (%):** The percentage of the population affected by schizophrenia.
6. **Bipolar disorder (%):** The percentage of the population affected by bipolar disorder.
7. **Eating disorders (%):** The percentage of the population affected by eating disorders.
8. **Anxiety disorders (%):** The percentage of the population affected by anxiety disorders.
9. **Drug use disorders (%):** The percentage of the population affected by drug use disorders.
10. **Depression (%):** The percentage of the population affected by depression.
11. **Alcohol use disorders (%):** The percentage of the population affected by alcohol use disorders.

Categorization of Columns as Quantitative or Qualitative:

1. **Index:** Quantitative

    - This column contains numerical values that identify each row uniquely. It represents discrete numerical data.
2. **Entity:** Qualitative
    - This column contains the name of the country, which is a descriptive label.
3. **Code:** Qualitative
    - This column contains the three-letter country code, which is a categorical identifier.
4. **Year:** Quantitative
    - This column contains numerical values indicating the year of data collection. It represents time in a numerical format, making it quantitative.
5. **Schizophrenia (%):** Quantitative
    - This column contains numerical values representing the percentage of the population affected by schizophrenia. It is a continuous measurement.
6. **Bipolar disorder (%):** Quantitative
    - This column contains numerical values representing the percentage of the population affected by bipolar disorder. It is a continuous measurement.
7. **Eating disorders (%):** Quantitative
    - This column contains numerical values representing the percentage of the population affected by eating disorders. It is a continuous measurement.
8. **Anxiety disorders (%):** Quantitative
    - This column contains numerical values representing the percentage of the population affected by anxiety disorders. It is a continuous measurement.
9. **Drug use disorders (%):** Quantitative
    - This column contains numerical values representing the percentage of the population affected by drug use disorders. It is a continuous measurement.
10. **Depression (%):** Quantitative
    - This column contains numerical values representing the percentage of the population affected by depression. It is a continuous measurement.
11. **Alcohol use disorders (%):** Quantitative
    - This column contains numerical values representing the percentage of the population affected by alcohol use disorders. It is a continuous measurement.

<u>Dataset Description: Credit Card Approvals.</u>

Columns:

1. **kCustomer:** Sequential identifier for each customer.
2. **Industry:** Numeric code representing the customer's industry.
3. **Ethnicity:** Numeric code representing the customer's ethnicity.
4. **YearsEmployed:** Number of years the customer has been employed.
5. **PriorDefault:** Indicates whether the customer has a prior default (1 = Yes, 0 = No).
6. **Employed:** Indicates whether the customer is currently employed (1 = Yes, 0 = No).
7. **CreditScore:** Numeric credit score of the customer.
8. **DriversLicense:** Indicates whether the customer has a driver's license (1 = Yes, 0 = No).

9. **Citizen:** Indicates the customer's citizenship status.
10. **ZipCode:** Numeric code representing the customer's postal area.
11. **Income:** Annual income of the customer.
12. **Approved:** Indicates whether the customer's credit card was approved (1 = Yes, 0 = No).

Categorization of Columns as Quantitative or Qualitative:

1. **nkCustomer:** Quantitative
    - This column contains numerical values that identify each customer uniquely. It represents discrete numerical data.
2. **Industry:** Quantitative
    - This column contains numerical codes representing different industries. It is categorized numerically.
3. **Ethnicity:** Quantitative
    - This column contains numerical codes representing different ethnicities. It is categorized numerically.
4. **YearsEmployed:** Quantitative
    - This column contains numerical values representing the number of years the customer has been employed. It is a continuous measurement.
5. **PriorDefault:** Quantitative
    - This column contains binary numerical values indicating whether the customer has a prior default (1 = Yes, 0 = No).
6. **Employed:** Quantitative
    - This column contains binary numerical values indicating whether the customer is currently employed (1 = Yes, 0 = No).
7. **CreditScore:** Quantitative
    - This column contains numerical values representing the customer's credit score. It is a continuous measurement.
8. **DriversLicense:** Quantitative
    - This column contains binary numerical values indicating whether the customer has a driver's license (1 = Yes, 0 = No).
9. **Citizen:** Qualitative
    - This column contains categorical data indicating the customer's citizenship status (e.g., "ByBirth," "ByOtherMeans").
10. **ZipCode:** Quantitative
    - This column contains numerical codes representing the customer's postal area. It is categorized numerically.
11. **Income:** Quantitative
    - This column contains numerical values representing the customer's annual income. It is a continuous measurement.
12. **Approved:** Quantitative
    - This column contains binary numerical values indicating whether the customer's loan was approved (1 = Yes, 0 = No).

Exercise 4:

Dataset Description: Iris Dataset

Columns Classification:

1. **Id:** Quantitative
    - **Reasoning:** This column contains numerical values that uniquely identify each row. It represents discrete numerical data and serves as an identifier for each sample.
2. **SepalLengthCm:** Quantitative
    - **Reasoning:** This column contains numerical values representing the length of the sepal in centimeters. It is a continuous measurement.
3. **SepalWidthCm:** Quantitative
    - **Reasoning:** This column contains numerical values representing the width of the sepal in centimeters. It is a continuous measurement.
4. **PetalLengthCm:** Quantitative
    - **Reasoning:** This column contains numerical values representing the length of the petal in centimeters. It is a continuous measurement.
5. **PetalWidthCm:** Quantitative
    - **Reasoning:** This column contains numerical values representing the width of the petal in centimeters. It is a continuous measurement.
6. **Species:** Qualitative
    - **Reasoning:** This column contains categorical data representing the species of the Iris flower (e.g., "Iris-setosa," "Iris-versicolor," "Iris-virginica"). It categorizes the data rather than measuring it numerically.

Exercise 6:

Columns of Interest for Specific Analyses:

1. **Trend Analysis:**
    - **Columns:** Year, Avg hrs per day sleeping
    - **Explanation:** For trend analysis, the 'Year' and 'Avg hrs per day sleeping' columns are crucial. By analyzing these columns, we can identify trends over time, such as whether the average hours of sleep are increasing, decreasing, or remaining stable from 2003 to 2007.
2. **Group Comparison:**
    - **Columns:** Type of Days, Age Group, Avg hrs per day sleeping
    - **Explanation:** For group comparison, it would be interesting to have data that includes different age groups and types of days (e.g., weekdays vs. weekends). However, since the dataset includes only one age group (15 years and over) and all days, it limits the scope for detailed group comparisons within this dataset. If additional data on different age groups or days were available, it would enable comparisons of average sleep hours across different demographics.
3. **Variance and Uncertainty Analysis:**
    - **Columns:** Avg hrs per day sleeping, Standard Error

- **Explanation:** To understand the variance and uncertainty in the average hours of sleep data, we can use the 'Avg hrs per day sleeping' and 'Standard Error' columns. The standard error provides insight into the reliability of the average sleep hours reported for each year, helping to assess the precision of the measurements.