**Sorting Problems:**

1. **Sort by State Name:**
   * Problem: Sort the dataset by the state column in ascending order.
   * Questions:
     1. Which state comes first in alphabetical order?
     2. Which state comes last in alphabetical order?
     3. What is the 10th state in alphabetical order?
     4. How many states start with the letter 'A'?
2. **Sort by Number of Individuals:**
   * Problem: Sort the dataset by the individuals column in descending order.
   * Questions:
     1. Which state has the highest number of homeless individuals?
     2. What is the number of homeless individuals in the state with the second-highest count?
     3. Which state has the lowest number of homeless individuals?
     4. How many states have more than 5,000 homeless individuals?
3. **Sort by Family Members:**
   * Problem: Sort the dataset by the family\_members column in ascending order.
   * Questions:
     1. Identify the state with the fewest family members who are homeless.
     2. How many states have fewer than 500 homeless family members?
     3. Which state has the highest number of homeless family members?
     4. What is the median number of homeless family members across all states?
4. **Sort by State Population:**
   * Problem: Sort the dataset by the state\_pop column in descending order.
   * Questions:
     1. What are the top three states with the largest populations?
     2. Which state has the smallest population?
     3. How many states have a population greater than 10 million?
     4. What is the total population of the top five states?
5. **Sort by Region and State:**
   * Problem: First, sort by the region column in ascending order, then by the state column within each region.
   * Questions:
     1. What is the first state listed in the East South Central region?
     2. Which region has the most states listed?
     3. What is the last state listed in the Mountain region?
     4. How many states are in the West South Central region?
6. **Sort by Individuals per Capita:**
   * Problem: Add a new column for homeless individuals per capita (individuals/state\_pop), then sort by this new column in descending order.
   * Questions:
     1. Which state has the highest number of homeless individuals per capita?
     2. What is the per capita rate for the state with the second-highest number?
     3. Which state has the lowest number of homeless individuals per capita?
     4. How many states have a per capita rate higher than 0.001?
7. **Sort by Family Members per Capita:**
   * Problem: Add a new column for homeless family members per capita (family\_members/state\_pop), then sort by this new column in ascending order.
   * Questions:
     1. Which state has the fewest homeless family members per capita?
     2. What is the per capita rate for the state with the highest number?
     3. Which states have a per capita rate higher than 0.0005?
     4. How many states have a per capita rate lower than 0.0001?
8. **Sort by Total Homeless:**
   * Problem: Add a new column for the total number of homeless people (individuals + family\_members), then sort by this new column in descending order.
   * Questions:
     1. Which state has the highest total number of homeless people?
     2. What is the total number of homeless people in the state with the second-highest total?
     3. Which state has the lowest total number of homeless people?
     4. How many states have a total number of homeless people greater than 10,000?
9. **Sort by Percentage of Family Members:**
   * Problem: Add a new column for the percentage of family members among the homeless (family\_members / (individuals + family\_members)), then sort by this new column in ascending order.
   * Questions:
     1. Which state has the lowest percentage of family members among the homeless?
     2. What is the percentage for the state with the highest number?
     3. Which states have a percentage higher than 50%?
     4. How many states have a percentage lower than 10%?
10. **Sort by Region and Total Homeless:**
    * Problem: First, sort by the region column in ascending order, then by the total number of homeless people within each region in descending order.
    * Questions:
      1. Which state has the highest total number of homeless people in the Pacific region?
      2. What is the total number of homeless people in the state with the highest count in the Mountain region?
      3. Which state has the lowest total number of homeless people in the East North Central region?
      4. How many states in the South Atlantic region have a total number of homeless people greater than 5,000?

**Subsetting Problems:**

1. **Subset by Region:**
   * Problem: Create a subset of the dataset containing only the states from the "Pacific" region.
   * Questions:
     1. How many states are in the Pacific region?
     2. Which state in the Pacific region has the highest number of homeless individuals?
     3. What is the total population of the Pacific region?
     4. How many family members are homeless in the Pacific region in total?
2. **Subset by Homeless Individuals:**
   * Problem: Subset the dataset to include only states with more than 10,000 homeless individuals.
   * Questions:
     1. How many states have more than 10,000 homeless individuals?
     2. Which state has the most homeless individuals among this subset?
     3. What is the average number of homeless individuals in this subset?
     4. How does the total population of this subset compare to the total population of all states?
3. **Subset by Family Members:**
   * Problem: Create a subset of the dataset with states where the number of family members is between 500 and 2000.
   * Questions:
     1. How many states fall within this range?
     2. Which state has the highest number of family members within this range?
     3. What is the average number of family members in this subset?
     4. How many individuals are homeless in these states in total?
4. **Subset by State Population:**
   * Problem: Subset the dataset to include states with a population of less than 1 million.
   * Questions:
     1. Which states have a population of less than 1 million?
     2. What is the total population of these states?
     3. How many homeless individuals are there in total in these states?
     4. Which state has the highest number of homeless individuals in this subset?
5. **Subset by Region and Homeless Individuals:**
   * Problem: Create a subset of states from the "Mountain" region with more than 2,000 homeless individuals.
   * Questions:
     1. How many states in the Mountain region have more than 2,000 homeless individuals?
     2. Which state in this subset has the highest number of homeless individuals?
     3. What is the total number of homeless individuals in this subset?
     4. How does the average number of homeless individuals in this subset compare to the overall average?
6. **Subset by Multiple Conditions:**
   * Problem: Create a subset where the number of homeless individuals is greater than 5,000 and the number of family members is greater than 1,000.
   * Questions:
     1. Which states meet both of these criteria?
     2. What is the total number of homeless individuals in these states?
     3. How many family members are homeless in total in these states?
     4. What is the average state population in this subset?
7. **Subset by High Population and Low Homelessness:**
   * Problem: Subset the dataset to include states with a population greater than 5 million but with fewer than 5,000 homeless individuals.
   * Questions:
     1. Which states meet these criteria?
     2. What is the total population of these states?
     3. How many homeless individuals are there in total in these states?
     4. What is the average number of homeless individuals in this subset?
8. **Subset by Region and Family Members:**
   * Problem: Create a subset of states from the "West South Central" region with fewer than 1,000 homeless family members.
   * Questions:
     1. How many states in the West South Central region have fewer than 1,000 homeless family members?
     2. Which state in this subset has the lowest number of family members?
     3. What is the total number of family members in this subset?
     4. How does the total number of homeless individuals in this subset compare to the total number of family members?
9. **Subset by High Homelessness per Capita:**
   * Problem: Subset the dataset to include states with homeless individuals per capita (individuals/state\_pop) greater than 0.001.
   * Questions:
     1. Which states have a per capita rate greater than 0.001?
     2. What is the highest per capita rate in this subset?
     3. How many total homeless individuals are there in this subset?
     4. What is the average number of homeless individuals in this subset?
10. **Subset by Low Homelessness per Capita:**
    * Problem: Subset the dataset to include states with homeless individuals per capita (individuals/state\_pop) less than 0.0005.
    * Questions:
      1. Which states have a per capita rate less than 0.0005?
      2. What is the lowest per capita rate in this subset?
      3. How many total homeless individuals are there in this subset?
      4. What is the average number of homeless individuals in this subset?

**Creating New Columns Problems:**

1. **Homeless Individuals per Capita:**
   * Problem: Add a new column for homeless individuals per capita (individuals/state\_pop).
   * Questions:
     1. Which state has the highest number of homeless individuals per capita?
     2. What is the per capita rate for the state with the second-highest number?
     3. Which state has the lowest number of homeless individuals per capita?
     4. How many states have a per capita rate higher than 0.001?
2. **Family Members per Capita:**
   * Problem: Add a new column for homeless family members per capita (family\_members/state\_pop).
   * Questions:
     1. Which state has the fewest homeless family members per capita?
     2. What is the per capita rate for the state with the highest number?
     3. Which states have a per capita rate higher than 0.0005?
     4. How many states have a per capita rate lower than 0.0001?
3. **Total Homeless:**
   * Problem: Add a new column for the total number of homeless people (individuals + family\_members).
   * Questions:
     1. Which state has the highest total number of homeless people?
     2. What is the total number of homeless people in the state with the second-highest total?
     3. Which state has the lowest total number of homeless people?
     4. How many states have a total number of homeless people greater than 10,000?
4. **Percentage of Family Members:**
   * Problem: Add a new column for the percentage of family members among the homeless (family\_members / (individuals + family\_members)).
   * Questions:
     1. Which state has the lowest percentage of family members among the homeless?
     2. What is the percentage for the state with the highest number?
     3. Which states have a percentage higher than 50%?
     4. How many states have a percentage lower than 10%?
5. **Homeless per Thousand Residents:**
   * Problem: Add a new column for the number of homeless individuals per thousand residents (individuals / state\_pop \* 1000).
   * Questions:
     1. Which state has the highest number of homeless individuals per thousand residents?
     2. What is the rate for the state with the second-highest number?
     3. Which state has the lowest number of homeless individuals per thousand residents?
     4. How many states have a rate higher than 1?
6. **Family Members per Thousand Residents:**
   * Problem: Add a new column for the number of homeless family members per thousand residents (family\_members / state\_pop \* 1000).
   * Questions:
     1. Which state has the fewest homeless family members per thousand residents?
     2. What is the rate for the state with the highest number?
     3. Which states have a rate higher than 0.5?
     4. How many states have a rate lower than 0.1?
7. **Total Homeless per Thousand Residents:**
   * Problem: Add a new column for the total number of homeless people per thousand residents ((individuals + family\_members) / state\_pop \* 1000).
   * Questions:
     1. Which state has the highest total number of homeless people per thousand residents?
     2. What is the rate for the state with the second-highest number?
     3. Which state has the lowest total number of homeless people per thousand residents?
     4. How many states have a rate higher than 1?
8. **Percentage of Homeless Individuals:**
   * Problem: Add a new column for the percentage of homeless individuals among the total homeless population (individuals / (individuals + family\_members) \* 100).
   * Questions:
     1. Which state has the highest percentage of homeless individuals among the total homeless population?
     2. What is the percentage for the state with the lowest number?
     3. Which states have a percentage higher than 90%?
     4. How many states have a percentage lower than 50%?
9. **Percentage of Homeless Family Members:**
   * Problem: Add a new column for the percentage of homeless family members among the total homeless population (family\_members / (individuals + family\_members) \* 100).
   * Questions:
     1. Which state has the highest percentage of homeless family members among the total homeless population?
     2. What is the percentage for the state with the lowest number?
     3. Which states have a percentage higher than 50%?
     4. How many states have a percentage lower than 10%?
10. **Homeless Rate Change:**
    * Problem: Assuming you have a dataset with previous year’s homeless numbers, add a new column for the percentage change in the number of homeless individuals from the previous year to the current year.
    * Questions:
      1. Which state had the highest percentage increase in homeless individuals?
      2. Which state had the highest percentage decrease in homeless individuals?
      3. What is the average percentage change across all states?
      4. How many states had a positive percentage change?

**Combining Sorting, Subsetting, and Creating New Columns Problems:**

1. **Subset by Region and Sort by State Name:**
   * Problem: Subset the dataset to include only states from the "Pacific" region and sort by state in ascending order.
   * Questions:
     1. Which state comes first in the Pacific region?
     2. How many states are in the Pacific region?
     3. What is the total population of the Pacific region?
     4. How many homeless individuals are there in the Pacific region in total?
2. **Subset by High Homelessness and Sort by Total Homeless:**
   * Problem: Subset the dataset to include states with more than 10,000 homeless individuals, then add a column for the total number of homeless people and sort by this column in descending order.
   * Questions:
     1. Which state has the highest total number of homeless people among those with more than 10,000 homeless individuals?
     2. How many states have more than 10,000 homeless individuals?
     3. What is the average total number of homeless people in this subset?
     4. How does the total population of this subset compare to the total population of all states?
3. **Subset by Family Members Range and Sort by Family Members per Capita:**
   * Problem: Subset the dataset to include states with family members between 500 and 2000, add a column for family members per capita, then sort by this new column in ascending order.
   * Questions:
     1. Which state has the fewest family members per capita within this subset?
     2. How many states fall within this range?
     3. What is the average number of family members per capita in this subset?
     4. How many total homeless individuals are there in these states?
4. **Subset by Low Population and Sort by Homeless Individuals per Capita:**
   * Problem: Subset the dataset to include states with a population of less than 1 million, add a column for homeless individuals per capita, then sort by this column in descending order.
   * Questions:
     1. Which state has the highest number of homeless individuals per capita within this subset?
     2. How many states have a population of less than 1 million?
     3. What is the total number of homeless individuals in these states?
     4. How does the average number of homeless individuals per capita compare to the overall average?
5. **Subset by Region and High Homelessness, then Sort by Total Homeless:**
   * Problem: Subset the dataset to include states from the "Mountain" region with more than 2,000 homeless individuals, add a column for the total number of homeless people, then sort by this column in descending order.
   * Questions:
     1. Which state has the highest total number of homeless people in this subset?
     2. How many states in the Mountain region have more than 2,000 homeless individuals?
     3. What is the total number of homeless individuals in this subset?
     4. How does the average number of homeless individuals in this subset compare to the overall average?
6. **Subset by Multiple Conditions and Sort by Individuals per Capita:**
   * Problem: Subset the dataset where the number of homeless individuals is greater than 5,000 and the number of family members is greater than 1,000, add a column for homeless individuals per capita, then sort by this column in descending order.
   * Questions:
     1. Which state has the highest number of homeless individuals per capita in this subset?
     2. How many states meet both criteria?
     3. What is the total number of homeless individuals in these states?
     4. How many family members are homeless in total in these states?
7. **Subset by High Population and Low Homelessness, then Sort by Family Members per Capita:**
   * Problem: Subset the dataset to include states with a population greater than 5 million but with fewer than 5,000 homeless individuals, add a column for family members per capita, then sort by this column in ascending order.
   * Questions:
     1. Which state has the fewest homeless family members per capita in this subset?
     2. How many states meet these criteria?
     3. What is the total population of these states?
     4. How many homeless individuals are there in total in these states?
8. **Subset by Region and Low Family Members, then Sort by Total Homeless per Capita:**
   * Problem: Subset the dataset to include states from the "West South Central" region with fewer than 1,000 homeless family members, add a column for the total number of homeless people per capita, then sort by this column in descending order.
   * Questions:
     1. Which state has the highest total number of homeless people per capita in this subset?
     2. How many states in the West South Central region have fewer than 1,000 homeless family members?
     3. What is the total number of homeless family members in this subset?
     4. How does the total number of homeless individuals in this subset compare to the total number of family members?
9. **Subset by High Homelessness per Capita, then Sort by Percentage of Homeless Family Members:**
   * Problem: Subset the dataset to include states with homeless individuals per capita greater than 0.001, add a column for the percentage of homeless family members, then sort by this column in ascending order.
   * Questions:
     1. Which state has the lowest percentage of homeless family members among the total homeless population in this subset?
     2. How many states have a per capita rate greater than 0.001?
     3. What is the highest per capita rate in this subset?
     4. How many total homeless individuals are there in this subset?
10. **Subset by Low Homelessness per Capita, then Sort by Percentage Change in Homelessness:**
    * Problem: Subset the dataset to include states with homeless individuals per capita less than 0.0005, add a column for the percentage change in the number of homeless individuals from the previous year, then sort by this column in descending order.
    * Questions:
      1. Which state had the highest percentage increase in homeless individuals in this subset?
      2. Which state had the highest percentage decrease in homeless individuals in this subset?
      3. What is the average percentage change across all states in this subset?
      4. How many states in this subset had a positive percentage change?

**Problem 1: Summary Statistics, Sorting, and Subsetting**

1. **Calculate Summary Statistics and Subset the Data**
   * **Calculate the mean, median, and standard deviation of the weekly\_sales column.**
   * **Sort the dataset by weekly\_sales in descending order.**
   * **Create a subset of transactions that occurred in the year 2010.**
   * **Questions:**
     1. **What are the mean, median, and standard deviation of the weekly\_sales?**
     2. **Which transaction has the highest weekly\_sales?**
     3. **How many transactions occurred in the year 2010?**
     4. **What are the earliest and latest transaction dates in this subset?**

**Problem 2: Creating New Columns, Aggregating, and Cumulative Statistics**

1. **Create a New Column and Aggregate Statistics**
   * **Create a new column adjusted\_sales by dividing weekly\_sales by fuel\_price\_usd\_per\_l.**
   * **Use the .agg() method to calculate the mean, median, and sum of the adjusted\_sales column.**
   * **Calculate the cumulative sum and cumulative maximum of the adjusted\_sales column.**
   * **Questions:**
     1. **What are the mean, median, and sum of the adjusted\_sales?**
     2. **What is the cumulative sum of adjusted\_sales up to the 10th transaction?**
     3. **What is the cumulative maximum adjusted\_sales up to the 10th transaction?**

**Problem 3: Subsetting, Sorting, and Multiple Summaries**

1. **Subset, Sort, and Multiple Summaries**
   * **Create a subset of transactions where the temperature\_c is greater than 20.**
   * **Sort this subset by weekly\_sales in descending order.**
   * **Use the .agg() method to calculate the mean, median, minimum, and maximum for the weekly\_sales column in this subset.**
   * **Questions:**
     1. **Which transaction has the highest weekly\_sales in this subset?**
     2. **What are the mean, median, minimum, and maximum of weekly\_sales in this subset?**
     3. **How many transactions have temperature\_c greater than 20?**

**Problem 4: Multiple Summaries and Dates**

1. **Multiple Summaries and Summarizing Dates**
   * **Use the .agg() method to calculate the mean and standard deviation of both weekly\_sales and temperature\_c.**
   * **Calculate the earliest and latest transaction dates in the dataset.**
   * **Questions:**
     1. **What are the mean and standard deviation of weekly\_sales and temperature\_c?**
     2. **What is the earliest transaction date?**
     3. **What is the latest transaction date?**

**Problem 5: Cumulative Statistics and Sorting**

1. **Cumulative Statistics and Sorting**
   * **Calculate the cumulative sum of weekly\_sales and the cumulative product of temperature\_c.**
   * **Sort the dataset by the cumulative sum of weekly\_sales.**
   * **Questions:**
     1. **What is the cumulative sum of weekly\_sales up to the 15th transaction?**
     2. **What is the cumulative product of temperature\_c up to the 15th transaction?**
     3. **Which transaction has the highest cumulative sum of weekly\_sales?**

**Aggregating DataFrames: Counting**

**Dropping Duplicates**

1. **Unique Stores**
   * How many unique stores are there in the dataset?
2. **Unique Departments**
   * How many unique departments are there in the dataset?

**Dropping Duplicate Pairs**

1. **Unique Store-Department Pairs**
   * How many unique store-department pairs are there in the dataset?
2. **Unique Store-Date Pairs**
   * How many unique store-date pairs are there?

**Value Counts**

1. **Department Sales Count**
   * What is the count of sales entries for each department?
2. **Store Sales Count**
   * What is the count of sales entries for each store?
3. **Holiday Sales Count**
   * What is the count of sales entries for holidays versus non-holidays?
4. **Department Sales Count (Sorted)**
   * What is the count of sales entries for each department, sorted in descending order?

**Proportions**

1. **Proportion of Sales per Department**
   * What proportion of the total sales entries does each department represent?
2. **Proportion of Sales per Store**
   * What proportion of the total sales entries does each store represent?
3. **Proportion of Holiday Sales**
   * What proportion of the sales entries are holidays versus non-holidays?

**GROUPED SUMMARY STATISTICS**

**1. Grouped Summaries**

* What is the average weekly sales for each store?

**2. Multiple Grouped Summaries**

* What are the minimum, maximum, and total weekly sales for each department?

**3. Grouping by Multiple Variables**

* What is the average temperature for each store in each department?

**4. Many Variables, Many Summaries**

* What are the average weekly sales and average temperature for each store and department?

**5. Additional Grouped Summaries**

* What is the total number of holiday and non-holiday sales for each store?

**6. Combining Aggregations**

* What are the minimum, average, and maximum unemployment rates for each store type?

**7. Grouping by Multiple Variables with Multiple Summaries**

* For each store and department, what are the minimum and maximum fuel prices, and the total weekly sales?

**8. Summarizing Weekly Sales by Holiday Status**

* What is the total and average weekly sales for holidays versus non-holidays for each department?

**9. Store Performance Over Time**

* What is the average weekly sales and average unemployment rate for each store for each year?

**10. Department-wise Sales During Holidays**

* For each department, what are the total and average weekly sales during holidays?

**PIVOT TABLES**

 **Group By .pivot\_table()**

* How would you create a pivot table to calculate the average weekly sales for each department?
* Create a pivot table that shows the mean weekly sales for each store.

 **Different Statistics**

* Create a pivot table to calculate the median weekly sales for each department.
* How can you modify the pivot table to calculate the median weekly sales for each store?

 **Multiple Statistics**

* Generate a pivot table that calculates both the mean and median weekly sales for each department.
* Create a pivot table showing both the mean and sum of weekly sales for each store.

 **Pivot on Two Variables**

* How can you create a pivot table to calculate the average weekly sales for each combination of department and holiday status?
* Create a pivot table to show the median weekly sales for each combination of store and holiday status.

 **Fill Missing Values with .fill\_value**

* Create a pivot table that calculates the sum of weekly sales for each department, filling any missing values with 0.
* How would you modify the pivot table to show the mean weekly sales for each store, filling missing values with a specified fill value?

 **Summing with .pivot\_table BY margins=True**

* Generate a pivot table that calculates the total weekly sales for each department and includes the overall total using margins=True.
* Create a pivot table to display the sum of weekly sales for each store and overall total, including the overall mean.

**AGGREGATING DATAFRAME PROBLEMS:**

 What are the total and average weekly sales for each store and department combination?

 Calculate the median, minimum, and maximum weekly sales for each department, segmented by holiday and non-holiday status.

 Identify the top 5 stores with the highest average weekly sales and the bottom 5 stores with the lowest average weekly sales.

 How many unique stores and departments have recorded sales during holiday weeks? What is the total sales amount for these combinations?

 Determine the count of sales entries and the proportion of total entries for each department, both during holidays and non-holidays.

 Compare the total weekly sales and average weekly sales for stores that have recorded sales both during holidays and non-holidays.

 Create a summary that shows the total weekly sales for each store, including a row for the overall total sales across all stores.

 Find the store and department combinations with the highest and lowest median weekly sales, and display the top 5 and bottom 5 combinations.

 Calculate the total, mean, and count of weekly sales for each store and department, replacing any missing values with zero, and provide a summary including overall totals.

 Generate a pivot table to display the average weekly sales for each department, segmented by store, and include margins to show overall totals and averages.

**EXPLICIT INDEXES**

1. **Setting a column as index**:
   * How would you set the date column as the index of the temperature DataFrame?
2. **Removing index**:
   * After setting date as the index, how would you reset it back to a regular column?
3. **Dropping index**:
   * How can you reset the index to a default integer index and drop the date column completely from the DataFrame?

**INDEX SUBSETTINGS**

1. **Index subsettings**:
   * Suppose you set city as the index. How would you subset the DataFrame to include only data for 'Abidjan' and 'New York'?
2. **Subsetting on duplicated index values**:
   * If city is set as the index, how can you subset all rows where the city is 'Abidjan'?

**MULTI-LEVEL INDEXES AKA HIERARCHICAL INDEXES**

1. **Multi-level indexes**:
   * How would you set a multi-level index with country as the first level and city as the second level?
2. **Subset outer level with a list**:
   * Given the multi-level index from the previous question, how would you subset the DataFrame to include only data for 'Côte D'Ivoire' and 'India'?
3. **Subset inner level with a list of tuples**:
   * How would you subset the DataFrame to include only data for ('Côte D'Ivoire', 'Abidjan') and ('India', 'New Delhi')?

**SORTING BY INDEX VALUES**

1. **Sorting by index values**:
   * With the multi-level index set as country and city, how would you sort the DataFrame by these index values?
2. **Controlling sorting**:
   * How can you sort the DataFrame by country in ascending order and city in descending order?