Principal Component Analysis (PCA) to a biomedical dataset

Name: Masixole Boya
Student number: 1869204

Section A: The data I would use to apply PCA

This is the link to the dataset:

https://datasetsearch.research.google.com/search?src=0&query=hospital&docid=L2cvMTFzczRwM3JkNQ%3D%3D

This dataset provides information about hospital bed capacity in U.S. hospital markets, along with data related to various models of COVID-19 infection scenarios. The dataset was compiled by a team of researchers at the Harvard Global Data Institute. It includes data for nine different models of COVID-19 infection scenarios. The scenarios, in which 20%, 40%, and 60% of the adult population would be infected with the novel coronavirus, many of whom would have no or few symptoms, and examined whether hospitals had the capacity to handle them if the cases came in over six months, 12 months, and 18 months. Hospital bed statistics were sourced from recent surveys conducted by the American Hospital Association and comprehensive data compiled by the American Hospital Directory. The dataset is organized into approximately 300 regions, referred to as hospital referral regions, for analysis and insights.

Columns and their description:

- **HRR**: Hospital Referral Region (HRR), specifying a market within which people generally go to the same hospitals.
- **Total Hospital Beds**: Count of all hospitable beds within an HRR that are set up and staffed.
- **Total ICU Beds**: Count of all ICU beds within an HRR that are set up and staffed.
- **Available Hospital Beds**: How many hospital beds are unoccupied at any given time, on average.
- **Potentially Available Hospital Beds***: How many beds could be available if the occupancy rate was reduced by 50% for non-COVID patients.
- **Available ICU Beds**: How many ICU beds are unoccupied on average.
- **Potentially Available ICU Beds***: How many beds could be available if the occupancy rate was reduced by 50% for non-COVID patients.

- **Adult Population**: How many people over the age of 18 are living within the HRR.
- **Population 65+**: How many people over the age of 65 are living within the HRR.
- **Projected Infected Individuals**: How many individuals over the age of 18 are expected to get infected with COVID-19 over the entire course of the pandemic.
- **Projected Hospitalized Individuals**: How many individuals over the age of 18 are expected to need hospitalization due to COVID-19 over the entire course of the pandemic.
- **Projected Individuals Needing ICU Care**: How many individuals over the age of 18 are expected to need ICU care due to COVID-19 over the entire course of the pandemic.
- **Hospital Beds Needed, Six Months**: How many hospital/ICU beds would have to be available to care for all patients requiring hospital care within six months.
- **Percentage of Available Beds Needed, Six Months**: What percentage of available hospital/ICU beds would need to be committed to COVID patients to care for all patients in six months.
- **Percentage of Potentially Available Beds Needed, Six Months**: What percentage of potentially available hospital/ICU beds would need to be committed to COVID patients to care for all patients in six months.
- **Percentage of Total Beds Needed, Six Months**: What percentage of all hospital/ICU beds would need to be committed to COVID patients to care for all patients in six months.
- **Hospital Beds Needed, Twelve Months**: How many hospital/ICU beds would have to be available to care for all patients requiring hospital care within twelve months.
- **Percentage of Available Beds Needed, Twelve Months**: What percentage of available hospital/ICU beds would need to be committed to COVID patients to care for all patients in twelve months.
- **Percentage of Potentially Available Beds Needed, Twelve Months**: What percentage of potentially available hospital/ICU beds would need to be committed to COVID patients to care for all patients in twelve months.
- **Percentage of Total Beds Needed, Twelve Months**: What percentage of all hospital/ICU beds would need to be committed to COVID patients to care for all patients in twelve months.
- **Hospital Beds Needed, Eighteen Months**: How many hospital/ICU beds would have to be available to care for all patients requiring hospital care within eighteen months.
- **Percentage of Available Beds Needed, Eighteen Months**: What percentage of available hospital/ICU beds would need to be committed to COVID patients to care for all patients in eighteen months.

- **Percentage of Potentially Available Beds Needed, Eighteen Months**: What percentage of potentially available hospital/ICU beds would need to be committed to COVID patients to care for all patients in eighteen months.
- **Percentage of Total Beds Needed, Eighteen Months**: What percentage of all hospital/ICU beds would need to be committed to COVID patients to care for all patients in eighteen months.
- **ICU Beds Needed, Six Months**: ICU beds within an HRR that are set up and staffed.
- **Percentage of Available ICU Beds Needed, Six Months**: What percentage of available hospital/ICU beds would need to be committed to COVID patients to care for all patients in six months.
- **Percentage of Potentially Available ICU Beds Needed, Six Months**: What percentage of potentially available hospital/ICU beds would need to be committed to COVID patients to care for all patients in six months.
- **Percentage of Total ICU Beds Needed, Six Months**: What percentage of all hospital/ICU beds would need to be committed to COVID patients to care for all patients in six months.
- **ICU Beds Needed, Twelve Months**: ICU beds within an HRR that are set up and staffed.
- **Percentage of Available ICU Beds Needed, Twelve Months**: What percentage of available hospital/ICU beds would need to be committed to COVID patients to care for all patients in twelve months.
- **Percentage of Potentially Available ICU Beds Needed, Twelve Months**: What percentage of potentially available hospital/ICU beds would need to be committed to COVID patients to care for all patients in twelve months.
- **Percentage of Total ICU Beds Needed, Twelve Months**: What percentage of all hospital/ICU beds would need to be committed to COVID patients to care for all patients in twelve months.
- **ICU Beds Needed, Eighteen Months**: ICU beds within an HRR that are set up and staffed.
- **Percentage of Available ICU Beds Needed, Eighteen Months**: What percentage of available hospital/ICU beds would need to be committed to COVID patients to care for all patients in eighteen months.
- **Percentage of Potentially Available ICU Beds Needed, Eighteen Months**: What percentage of potentially available hospital/ICU beds would need to be committed to COVID patients to care for all patients in eighteen months.

- **Percentage of Total ICU Beds Needed, Eighteen Months**: What percentage of all hospital/ICU beds would need to be committed to COVID patients to care for all patients in eighteen months.

Section B: How I would do the steps in the order given in the instructions

0. **Tools and Libraries:**

- **Python:** The entire project would be done in python as this is a versatile programming language suitable for data preprocessing, analysis, and visualization.
- **Libraries:**
    - **Pandas:** Used for data manipulation and preprocessing.
    - **scikit-learn:** Offers various machine learning tools, including PCA for dimensionality reduction.
    - **Matplotlib and Seaborn:** Popular for creating visualizations in Python.

The following are bullet points that summarize my proposed workflow in the form of numbers for the main steps, as given in the instructions, and indented bullet points for the sub-steps within the main steps. I say what library I would use and from that library what function or class would I use. I took this approach as I believed it is much better than writing it in an essay format as this is much easier to read and it enables me to imagine the workflow even better.

1. **Preprocess the Biomedical Dataset:**
    - Handle missing values and normalize the data:
        - Library: Pandas
        - Functions: **fillna()** for filling missing values, and **normalize()** for data normalization.
2. **Use PCA to Simplify the Data:**
    - Apply PCA algorithm to reduce dimensionality:
        - Library: scikit-learn
        - Function: **PCA()** from scikit-learn, specifically **fit_transform()** to fit the model and transform the dataset into the new reduced-dimensional space.
3. **Analyze the PCA Results:**

- Interpret variance explained by each principal component and assess feature contribution:
    - This step is about analyzing the output of the PCA, such as explained variance ratios. Therefore, I would not use libraries but rather I would interpret these myself.

4. **Visualize the PCA Results:**
    - Utilize appropriate plots and charts for visualization:
        - Library: Matplotlib, Seaborn
        - Functions: **scatter()** for scatter plots, **biplot()** for biplots, or another function based on the exact graph I want for visualization
5. **Summarize and Understand:**
    - Summarize insights and findings from PCA analysis:
        - This step calls for understanding the PCA results and summarizing the important insights in a clear and concise manner.