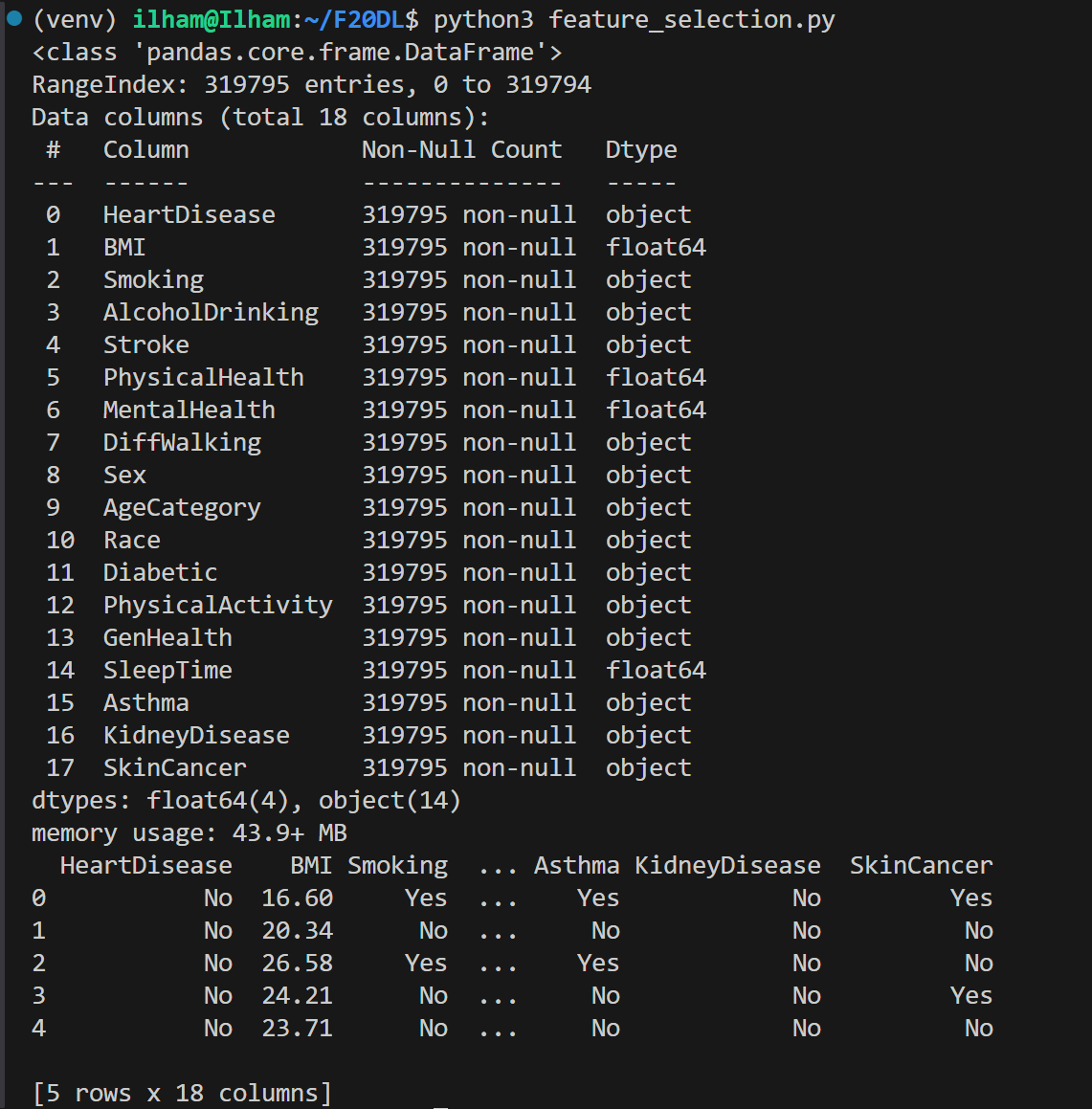
**F21DL LAB 2: Data Exploration and Visualization**

**Data Inspection**

* Understand the structure, types, and any missing values in the dataset.



- This initial output from the dataset inspection step gives us a clear view of the structure, types, and memory usage of the data.

**1. DataFrame Structure and Summary**

* **RangeIndex**: The dataset has **319,795 entries**
* **Data Columns**: There are **18 columns** in total. (features)

**2. Columns and Data Types**

Each feature is listed with its name, the count of non-null values, and its data type:

* **float64**: Continuous, numerical features (e.g., BMI, PhysicalHealth, MentalHealth, SleepTime).
* **object**: Categorical/text features (e.g., HeartDisease, Smoking, Stroke, etc.).

Since all columns show **319,795 non-null counts**, this dataset is complete, with no missing values in any column.

**3. Memory Usage**

* The dataset uses around **43.9 MB** of memory, which provides a sense of the dataset's size relative to our computational resources.

**4. Sample Data**

The preview shows the first five rows with sample values.

* **HeartDisease**: The target variable (Yes/No) indicates whether a person has heart disease.
* **BMI**: Numerical body mass index value for each individual.
* **Smoking, AlcoholDrinking, Stroke, etc.**: Lifestyle and health indicators that may correlate with heart disease risk.
* We have a mix of categorical and continuous variables. This data inspection confirms that all columns are fully populated, and now, we focus on exploring these features further, particularly the relationships between these lifestyle/health variables and the target variable HeartDisease.

**Correlation of features with HeartDisease.**  
This output shows the correlation values between the target variable (HeartDisease) and each feature in the dataset:

A screenshot of a computer

Description automatically generated

**Interpreting the Correlation Values**

* **Correlation Coefficient**: The values range from -1 to 1.
  + **Positive correlation** (closer to 1): As one variable increases, the other tends to increase as well.
  + **Negative correlation** (closer to -1): As one variable increases, the other tends to decrease.
  + **Near zero correlation**: No linear relationship between the variables.

**For HeartDisease:**

* **AgeCategory (0.23)**: This has a positive correlation, indicating that older individuals are more likely to have heart disease.
* **DiffWalking (0.20)**: There is a positive correlation, showing that difficulty in walking is associated with heart disease.
* **Stroke (0.19)** and **Diabetic\_Yes (0.18)**: Both are positively correlated, indicating that individuals who had a stroke or are diabetic are more likely to have heart disease.
* **GenHealth (-0.24)**: This negative correlation suggests that as general health improves, the likelihood of heart disease decreases.
* The lower correlation values (closer to zero) for some features mean that they have less influence or no linear relationship with HeartDisease.

**Insights from the Evaluation**

From this correlation matrix, we identify which features have a stronger relationship with HeartDisease:

1. **AgeCategory**, **DiffWalking**, **Stroke**, **Diabetic\_Yes**, and **GenHealth** have notable correlations with HeartDisease. These might be important features to focus on for further analysis or model training, as they appear more predictive of heart disease risk.
2. **PhysicalActivity** shows a negative correlation, indicating that physically active individuals are less likely to have heart disease, which aligns with general health knowledge.

**Next Steps**

1. **Feature Selection**:
   * Based on the correlation values, we choose to select features with higher absolute correlations with HeartDisease for model training.
   * We set a threshold (i.e. selecting features with an absolute correlation above 0.1) to filter out features with weaker correlations.
2. **Prepare Datasets for Model Training**:
   * Create separate datasets with selected features (e.g., top 5, 10, 15, etc.) as mentioned in our lab requirements.
   * These different feature sets can be used to train models and evaluate how the number of features affects performance.
3. **Model Training**:
   * Use machine learning models (e.g., logistic regression, decision trees) to predict HeartDisease based on the selected features.
   * Evaluate the models’ accuracy and performance metrics to determine which features are most predictive.
4. **Answering Lab Questions**:
   * **Identify features with the strongest correlation to HeartDisease**: We have identified them (e.g., AgeCategory, DiffWalking, Stroke, etc.).
   * **Prepare feature subsets**: For instance, we create datasets with only the top 2, top 5, top 10 features, etc., based on their correlation with HeartDisease.
   * **Analyze performance**: After training models with each feature set, we assess whether using more features improves accuracy or if only a few top features are sufficient.

**Feature Selection:**

* **Top 2 features**: The two features with the highest absolute correlation (e.g., GenHealth and AgeCategory).
* **Top 5 features**: The top five based on absolute correlation (e.g., GenHealth, AgeCategory, DiffWalking, Stroke, Diabetic\_Yes).
* **Top 10 features**: Expanding further to include features like PhysicalHealth, KidneyDisease, Smoking, etc.
* **All features**: Using all features to understand if the extra information improves performance or if it adds noise.

**Suggested Feature Counts**

* **Top 2 features** (highest correlation with HeartDisease).
* **Top 5 features** (best balance between simplicity and predictive power).
* **Top 10 features** (captures most of the higher correlations).

**All features with a correlation > |0.1|**

**The features include:**  
HeartDisease 1.000000

AgeCategory 0.234493

DiffWalking 0.201258

Stroke 0.196835

Diabetic\_Yes 0.183072

PhysicalHealth 0.170721

KidneyDisease 0.145197

Smoking 0.107764

PhysicalActivity -0.100030

GenHealth -0.243182