Milestone 1

Topic : Heart Health Indicator data preparation and visualization

In this project, the datasets of interest were the Heart health datasets.

I am using three different datasets which has some relationship between them by type of heart decease, age groups & region columns. It means that the dataset can be joined together by the standard columns and explored together.

In other milestones of the project, I would be like the option to pull the data from Kaggle API/ chronic data.cdc.gov and apply data wrangling techniques that I have learned throughout the course using pandas. And as a part of data visualization, I would be using matplotlib and ggplot2.

DataSets & Sources:

Data Source 1: Flat file data source, Kaggle

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Data Source 2: Data Pull from API, Kaggle

API download from https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

Data Source 3: Website Data, chronic data.cdc.gov

https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Rates-and-Trends-in-Coronary-Heart-Disease-and-Str/9cr5-2tt7

Relationship exist between the datasets or relationship to be created:

The grain of the dataset is identified as Heart Disease, Age Group & Region.

Interpretation of data and next steps for upcoming milestones:

Dataset1: Source:Kaggle

The first dataset was personal key indicator of heart health which is available at https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease and had 18 variables. The variables in the dataset were both continuous and categorical.

Data Dictionary:

	Variable	Description
1	HeartDisease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
2	BMI	Body Mass Index
3	Smoking	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
4	AlcoholDrinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
5	Stroke	(Ever told) (you had) a stroke?
6	PhysicalHealth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days)
8	MentalHealth	Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days)
9	DiffWalking	Do you have serious difficulty walking or climbing stairs?
10	Sex	Are you male or female?
11	AgeCategory	Fourteen-level age category
12	Race	Imputed race/ethnicity value
13	Diabetic	(Ever told) (you had) diabetes?
14	PhysicalActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
15	SleepTime	On average, how many hours of sleep do you get in a 24-hour period?
16	Asthma	(Ever told) (you had) asthma?
17	KidneyDisease	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
18	SkinCancer	(Ever told) (you had) skin cancer?

Data Source 2: API Data Source:

The second dataset was https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset and had 23 columns. Similarly, the dataset had both numerical and categorical variables.

Variable
HeartDiseaseorAttack
HighBP
HighChol
CholCheck

BMI
Smoker
Stroke
Diabetes
PhysActivity
Fruits
Veggies
HvyAlcoholConsump
AnyHealthcare
NoDocbcCost
GenHlth
MentHlth
PhysHlth
DiffWalk
Sex
Age
Education
Income

Data Source 3: HTML source from cdc.gov website

URL: https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Rates-and-Trends-in-Coronary-Heart-Disease-and-Str/9cr5-2tt7

I would be scraping the url to pull details

Variable	Description
Name	HeartDiseaseorAttack
Year	Year
Reason	Types
Region	States

Overall, I like to clean the data as needed by dropping off columns which may not be required in further data analysis / visualizations, adding new columns / updating the data elements in selected columns to maintain consistent relationships between various data sources using the country code and then create the planned visualizations.