Data Dictionary & Project Plan // SYDE 780 - Benjamin Luo (20890448)

Data: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
Full Data (2022): https://www.cdc.gov/brfss/annual_data/annual_2022.html

---

Data dictionary and preliminary preprocessing

The Kaggle dataset is actually an excerpt of data from the CDC BRFSS annual survey with only features relevant to heart disease (n=44/328). The original data is made available through SAS files which can be read through Python's Pandas library for preprocessing. The Kaggle dataset renames the original feature names without documenting them so I instead use the original data with 445132 data points and 328 features. The methodology used to preprocess is as follows:

<table>
<tr><td>1</td><td>

```
df22 = pd.read_sas('./data/LLCP2022.XPT', format='xport')
df22
```

✓ 0.0s

	_STATE	FMONTH	IDATE	IMONTH	IDAY	IYEAR	DISPCODE	SEQNO	_PSU	CTELENM1
0	1.0	1.0	b'02032022'	b'02'	b'03'	b'2022'	1100.0	b'2022000001'	2.022000e+09	1.0
1	1.0	1.0	b'02042022'	b'02'	b'04'	b'2022'	1100.0	b'2022000002'	2.022000e+09	1.0
2	1.0	1.0	b'02022022'	b'02'	b'02'	b'2022'	1100.0	b'2022000003'	2.022000e+09	1.0
3	1.0	1.0	b'02032022'	b'02'	b'03'	b'2022'	1100.0	b'2022000004'	2.022000e+09	1.0
4	1.0	1.0	b'02022022'	b'02'	b'02'	b'2022'	1100.0	b'2022000005'	2.022000e+09	1.0
...	...	...	...	...	...	...	...	...	...	...
445127	78.0	11.0	b'12192022'	b'12'	b'19'	b'2022'	1100.0	b'2022001527'	2.022002e+09	NaN
445128	78.0	11.0	b'12212022'	b'12'	b'21'	b'2022'	1100.0	b'2022001528'	2.022002e+09	NaN
445129	78.0	11.0	b'11292022'	b'11'	b'29'	b'2022'	1100.0	b'2022001529'	2.022002e+09	NaN
445130	78.0	11.0	b'12082022'	b'12'	b'08'	b'2022'	1100.0	b'2022001530'	2.022002e+09	NaN
445131	78.0	11.0	b'12142022'	b'12'	b'14'	b'2022'	1100.0	b'2022001531'	2.022002e+09	NaN

</td><td>Download the 2022 dataset (.XPT) from the CDC and read it into a Pandas dataframe</td></tr>
<tr><td>2</td><td>

Label: Cellular Telephone
Section Name: Land Line Introduction
Section Number: 0
Question Number: 5
Column: 67
Type of Variable: Num
SAS Variable Name: CELPHON1
Question Prologue: Variable only on the land line survey
Question: Is this a cell telephone?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes, it is a cell phone—Terminate Phone Call	2	0.00	0.00
2	Not a cell phone—Go to LL.06, LADULT1	96,050	100.00	100.00
BLANK	Missing or Cell Phone Interview Notes: QSTVER >= 20	349,080	.	.

</td><td>Download the 2022 codebook (.HTM) that explains each of the 328 variables. Crawl this HTM file in Python using BeautifulSoup to form a dataframe of variables, data types, descriptions, and possible values. Use string processing to remove unnecessary data (ex. 'Notes')</td></tr>
<tr><td>3</td><td>

```
Independent variables
    CVDINFR4: (Ever told) you had a heart attack, also called a myocardial infarction?
    CVDCRHD4: (Ever told) (you had) angina or coronary heart disease?
    _MICHD: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).

Demographic and socioeconomic
    _STATE: State of residence (geographic location can influence access to healthcare and environmental factors).
    SEXVAR: Sex of the respondent (cardiovascular risk varies by sex).
    _AGEG5YR: Age group (cardiovascular risk increases with age).
    _RACEGR4: Race/ethnicity (certain racial/ethnic groups have higher cardiovascular risk).
    _EDUCAG: Education level (lower education levels are associated with higher cardiovascular risk).
    _INCOMG1: Income level (lower income is associated with higher cardiovascular risk).
    MARITAL: Marital status (social support can influence cardiovascular health).
    EMPLOY1: Employment status (stress and job type can impact cardiovascular health).
```

</td><td>Use an LLM to identify and cluster relevant variables (n=53). Manually validate results</td></tr>
<tr><td>4</td><td>

```
df22c = df22c.dropna(axis=1, thresh=int(0.75 * len(df22)) + 1)
removed_cols = df22.columns.difference(df22c.columns)
df22c
```

✓ 0.0s

	CVDINFR4	CVDCRHD4	_MICHD	_STATE	SEXVAR	_AGEG5YR	_RACEGR4	_EDUCAG	_INCOMG1	MARITAL	...	HEIGHT3	_BMI5	_BMI5CAT
0	2.0	2.0	2.0	1.0	2.0	13.0	1.0	4.0	9.0	1.0	...	9999.0	NaN	NaN
1	2.0	2.0	2.0	1.0	2.0	13.0	1.0	2.0	3.0	3.0	...	503.0	2657.0	3.0
2	2.0	2.0	2.0	1.0	2.0	8.0	1.0	4.0	6.0	1.0	...	502.0	2561.0	3.0
3	2.0	2.0	2.0	1.0	2.0	14.0	1.0	2.0	9.0	1.0	...	505.0	2330.0	2.0
4	2.0	2.0	2.0	1.0	2.0	5.0	1.0	3.0	3.0	1.0	...	502.0	2177.0	2.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
445127	2.0	2.0	2.0	78.0	2.0	1.0	2.0	2.0	1.0	5.0	...	505.0	2563.0	3.0
445128	2.0	2.0	2.0	78.0	2.0	7.0	2.0	4.0	5.0	1.0	...	507.0	2866.0	3.0
445129	2.0	2.0	2.0	78.0	2.0	10.0	9.0	2.0	9.0	1.0	...	507.0	1723.0	1.0
445130	1.0	2.0	1.0	78.0	1.0	11.0	2.0	3.0	5.0	1.0	...	600.0	3255.0	4.0
445131	2.0	2.0	2.0	78.0	1.0	5.0	2.0	1.0	2.0	3.0	...	506.0	2260.0	2.0

</td><td>Remove features where more than 75% of the responses are NaN. In retrospect, I could have first removed users where the majority of responses were NaN before removing features.</td></tr>
<tr><td>5</td><td>

```
dd22c = dd22c.loc[~dd22c.index.isin(removed_cols)]
dd22c.head(3)
```

✓ 0.0s

Data Element Name	Format	Description	Survey Question	Possible values	Example Num
CVDINFR4	Num	Ever Diagnosed with Heart Attack	(Ever told) you had a heart attack, also calle...	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not s...	1
CVDCRHD4	Num	Ever Diagnosed with Angina or Coronary Heart D...	(Ever told) (you had) angina or coronary heart...	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not s...	1
_MICHD	Num	Ever had CHD or MI	Respondents that have ever reported having cor...	{'1': 'Reported having MI or CHD', '2': 'Did n...	1

</td><td>Prune the data dictionary dataframe to match the dataset's remaining features. 36 features remain</td></tr>
</table>

The final dataset contains all numeric values, with most of the values being identifiers that map to a natural language description using the 'Possible values' header in the data dictionary dataframe.

Snapshot of the data dictionary

Data Element Name	Format	Description	Possible values	Example Char	Example	Renamed Variable
CVDINFR4	Char	Ever Diagnosed with Heart Attack	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not sure', '9': 'Refused', 'BLANK': 'Not asked	Yes	2	heart_attack
CVDCRHD4	Char	Ever Diagnosed with Angina or Coronary Heart Disease	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not sure', '9': 'Refused', 'BLANK': 'Not asked	Yes	2	angina_or_chd
_MICHD	Char	Ever had CHD or MI	{'1': 'Reported having MI or CHD', '2': 'Did not report having MI or CHD', 'BLANK':	Reported having MI or CHD	2	chd_or_mi
_STATE	Char	State FIPS Code	{'1': 'Alabama', '2': 'Alaska', '4': 'Arizona', '5': 'Arkansas', '6': 'California', '8': 'Col	Alabama	1	state
SEXVAR	Char	Sex of Respondent	{'1': 'Male', '2': 'Female'}	Male	2	sex
_AGEG5YR	Char	Reported age in five-year age categories calculated variab	{'1': 'Age 18 to 24', '2': 'Age 25 to 29', '3': 'Age 30 to 34', '4': 'Age 35 to 39', '5': 'Ag	Age 18 to 24	13	age
_RACEGR4	Char	Computed Five level race/ethnicity category.	{'1': 'White only, Non-Hispanic', '2': 'Black only, Non-Hispanic', '3': 'Other race o	White only, Non-Hispanic	1	ethnicity
_EDUCAG	Char	Computed level of education completed categories	{'1': 'Did not graduate High School', '2': 'Graduated High School', '3': 'Attended (	Did not graduate High School	2	education
_INCOMG1	Char	Computed income categories	{'1': 'Less than $15,000', '2': '$15,000 to < $25,000', '3': '$25,000 to < $35,000	Less than $15,000	3	income
MARITAL	Char	Marital Status	{'1': 'Married', '2': 'Divorced', '3': 'Widowed', '4': 'Separated', '5': 'Never married',	Married	3	marital
EMPLOY1	Char	Employment Status	{'1': 'Employed for wages', '2': 'Self-employed', '3': 'Out of work for 1 year or mor	Employed for wages	2	employment
SMOKE100	Char	Smoked at Least 100 Cigarettes	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not Sure', '9': 'Refused', 'BLANK': 'Not asked	Yes	2	smoker_past
USENOW3	Char	Use of Smokeless Tobacco Products	{'1': 'Every day', '2': 'Some days', '3': 'Not at all', '7': 'Don't know/Not Sure', '9': 'R	Every day	3	smoker_now
ECIGNOW2	Char	Do you now use e-cigarettes, or vaping products every da	{'1': 'Never used e-cigarettes in your entire life', '2': 'Use them every day', '3': 'Use	Never used e-cigarettes in your entire life	1	vaper
ALCDAY4	Char	Days in past 30 had alcoholic beverage	{'101 - 199': 'Days per week', '201 - 299': 'Days in past 30 days', '777': 'Don't kn	Days per week	888	alcohol
EXERANY2	Char	Exercise in Past 30 Days	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not Sure', '9': 'Refused', 'BLANK': 'Not asked	Yes	2	exercise
SLEPTIM1	Num	How Much Time Do You Sleep	{'1 - 24': 'Number of hours [1-24]', '77': 'Don't know/Not Sure', '99': 'Refused', 'I		6	sleep
GENHLTH	Char	General Health	{'1': 'Excellent', '2': 'Very good', '3': 'Good', '4': 'Fair', '5': 'Poor', '7': 'Don't know/I	Excellent	1	general_health
PHYSHLTH	Num	Number of Days Physical Health Not Good	{'1 - 30': 'Number of days', '88': 'None', '77': 'Don't know/Not sure', '99': 'Refuse		88	physical_health_days
MENTHLTH	Num	Number of Days Mental Health Not Good	{'1 - 30': 'Number of days', '88': 'None', '77': 'Don't know/Not sure', '99': 'Refuse		88	mental_health_days
DIABETE4	Char	(Ever told) you had diabetes	{'1': 'Yes', '2': 'Yes, but female told only during pregnancy', '3': 'No', '4': 'No, pre-c	Yes	3	diabetes
HAVARTH4	Char	Told Had Arthritis	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not Sure', '9': 'Refused', 'BLANK': 'Not asked	Yes	2	arthritis
CHCCOPD3	Char	Ever told you had C.O.P.D. emphysema or chronic bronch	{'1': 'Yes', '2': 'No', '7': 'Don't know / Not sure', '9': 'Refused', 'BLANK': 'Not asked	Yes	2	copd
ADDEPEV3	Char	(Ever told) you had a depressive disorder	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not sure', '9': 'Refused', 'BLANK': 'Not asked	Yes	2	depression
ASTHMA3	Char	Ever Told Had Asthma	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not Sure', '9': 'Refused', 'BLANK': 'Not asked	Yes	2	asthma
WEIGHT2	Num	Reported Weight in Pounds	{'50 - 0776': 'Weight (pounds)', '7777': 'Don't know/Not sure', '9023 - 9352': 'W		150	weight
HEIGHT3	Num	Reported Height in Feet and Inches	{'200 - 711': 'Height (ft/inches)', '7777': 'Don't know/Not sure', '9061 - 9998': 'H		503	height
_BMI5	Num	Computed body mass index	{'1 - 9999': '1 or greater', 'BLANK': 'Don't know/Refused/Missing'}	1 or greater	2657	bmi
_BMI5CAT	Char	Computed body mass index categories	{'1': 'Underweight', '2': 'Normal Weight', '3': 'Overweight', '4': 'Obese', 'BLANK': '	Underweight	3	bmi_category
PRIMINSR	Char	What is Primary Source of Health Insurance?	{'1': 'A plan purchased through an employer or union (including plans purchase	A plan purchased through an employer or	3	health_insurance
MEDCOST1	Char	Could Not Afford To See Doctor	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not sure', '9': 'Refused', 'BLANK': 'Not asked	Yes	2	low_ses
CHECKUP1	Num	Length of time since last routine checkup	{'1': 'Within past year (anytime less than 12 months ago)', '2': 'Within past 2 yea		8	checkup_days
COVIDPOS	Char	Have you ever been told you tested positive for COVID 19	{'1': 'Yes', '2': 'No', '3': 'Tested positive using home test without health professic	Yes	2	covid
FLUSHOT7	Char	Adult flu shot/spray past 12 mos	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not Sure', '9': 'Refused', 'BLANK': ''}	Yes	2	flu_shot
PNEUVAC4	Char	Pneumonia shot ever	{'1': 'Yes', '2': 'No', '7': 'Don't know/Not Sure', '9': 'Refused', 'BLANK': 'Not asked	Yes	2	pneumonia_shot
TETANUS1	Char	Received Tetanus Shot Since 2005?	{'1': 'Yes, received Tdap', '2': 'Yes, received tetanus shot, but not Tdap', '3': 'Yes, r	Yes, received Tdap	4	tetanus_shot

Project Overview

Objective	To examine the prevalence and distribution of cardiovascular diseases (CVD) in the United States across demographic groups and evaluate the influence of socioeconomic factors and lifestyle choices on CVD outcomes. **RQ1 (Demographics)**: What is the prevalence of CVD across different age, sex, and ethnic groups? **RQ2 (Socioeconomics)**: Are there significant differences in CVD prevalence based on socioeconomic factors such as income, education, or insurance status? **RQ3 (Lifestyle)**: How do behavioral and lifestyle choices such as smoking, alcohol use, and physical activity influence the prevalence of CVD?
Hypothesis	**RQ1 (Demographics)**: - CVD prevalence increases with age as the body naturally weakens over its lifetime. - CVD prevalence is higher for females with children as the heart undergoes physiological changes during pregnancy that may be detrimental to heart health. - CVD prevalence is higher for the Black and Hispanic populations compared to the White population due to increased socioeconomic disparities that deter access to healthcare as well as potential implicit bias in medical care. **RQ2 (Socioeconomics)**: CVD prevalence is inversely correlated with socioeconomic status as it limits access to healthcare. **RQ3 (Lifestyle)**: Increased exercise and decreased substance use leads to decreased CVD prevalence. Exercise is known to improve heart physiology while substance use has adverse effects on the brain which negatively affects the heart through hormone and neurotransmitter release.
Motivation	CVD comprises some of the world's leading causes of death. By studying cardiovascular diseases, we can develop a deeper understanding of how demographic, socioeconomic, and lifestyle factors impact heart health. This knowledge can help: **(RQ1)** Identify high-risk groups in healthcare settings **(RQ2)** Inform policies that promote health equity and access to care **(RQ3)** Inspire individuals to make healthier lifestyle choices
Approach	Exploratory analysis will be done to determine variable inclusion, variable distributions, and causes of missing data. All approaches will involve controlling for confounding variables through preprocessing. Final results will report the p-values, effect size, and confidence intervals. For the modeling, I want to cover the broad topics taught in this course from statistics (descriptive, inferential) to machine learning (decision trees) **RQ1 (Demographics)**: Use descriptive statistics to describe CVD prevalence rates across each demographic group (age, sex, and ethnicity) **RQ2 (Socioeconomics)**: Train tree-based model(s) to rank socioeconomic factors by their impact on CVD prevalence. Emphasize the explainability of decision trees for describing model outputs through examples and visualization **RQ3 (Lifestyle)**: Use multiple regression to model the relationship between lifestyle factors and CVD prevalence