Project Information

Data

Data Sourcing & Collection

The Dataset was sourced from Kaggle, a community driven platform for publishing datasets. The Data itself was originally sourced from the 2020 Annual CDC survey. A survey was done on 400k individuals regarding their health status. The dataset can be considered as good quality as it is from a trusted source.

Heart Disease Dataset

Data Content

The dataset contains information regarding the health of individuals based off of key indicators of heart disease, as well as some demographic information.

The Original CDC dataset consisted of 279 columns and 401958 rows. The Heart Disease indicators dataset was reduced to 18 columns and 319795 rows. See Data dictionary for detailed view of each column (Data Dictionary frequencies calculated pre-cleaning)

Data Selection

I've studied and worked in the health sector for the last 5.5 years, and cardiovascular conditions have always been an interest of mine. As a first open project I wanted to choose a topic I have substantial knowledge in, in order to improve my general understanding of the project and the data itself. CDC data is reliable and easy to understand, therefore working with this dataset is preferred.

Data Dictionary

HeartDisease

Lable: Ever had coronary heart disease (CHD) or myocardial infarction(MI)
Column: 1
Type of Variable: Categorical
Column name: HeartDisease

Value	Value Lable	Frequency	Percentage
Yes	Reported having MI or CHD	27373	8.56
No	Did not report having MI or CHD	292422	91.44

BMI

Lable: Computed body mass index
Column: 2
Type of Variable: Continuous
Column name: BMI

Value	Value Lable	Frequency	Percentage
1-9999	1 or greater	319795	100

Smoking

Lable: Current smoking calculated variable
Column: 3
Type of Variable: Categorical
Column name: Smoking

Value	Value Lable	Frequency	Percentage
Yes	Reported as a current smoker	131908	41.25
No	Did not report as a current smoker	187887	58.75

AlcoholDrinking

Lable: Heavy Alcohol Consumption Calculated Variable
Column: 4
Type of Variable: Categorical
Column name: AlcoholDrinking
Question: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)

Value	Value Lable	Frequency	Percentage
Yes	Yes	13415	4.19
No	No	118495	95.81

Stroke

Lable: Ever Diagnosed with a stroke
Column: 5
Type of Variable: Categorical
Column name: Stroke

Value	Value Lable	Frequency	Percentage
Yes	Yes	6411	2.00
No	No	112082	98.00

PhysicalHealth

Lable: Number of Days Physical Health Not Good
Column: 6
Type of Variable: Continuous
Column name: PhysicalHealth
Question: 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?

Value	Value Lable	Frequency	Percentage
0-30	Number of days	319795	100

MentalHealth

Lable: Number of Days Mental Health Not Good
Column: 7
Type of Variable: Continuous
Column name: MentalHealth
Question: Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?

Value	Value Lable	Frequency	Percentage
0-30	Number of days	319795	100

DiffWalking

Lable: Difficulty Walking or Climbing Stairs
Column: 8
Type of Variable: Categorical
Column name: DiffWalking

Value	Value Lable	Frequency	Percentage
Yes	Do have difficulty walking or climbing stairs	44410	13.89
No	Do not have difficulty walking or climbing stairs	275385	86.11

Sex

Lable: Sex
Column: 9
Type of Variable: Categorical
Column name: Sex

Value	Value Lable	Frequency	Percentage
Male	Male	151990	47.53
Female	Female	167805	52.47

AgeCategory

Lable: Age category
Column: 10
Type of Variable: Categorical
Column name: AgeCategory

Value	Value Lable	Frequency	Percentage
18-24	18-24	21064	6.59
25-29	25-29	16955	5.30
30-34	30-34	18753	5.86
35-39	35-39	20550	6.43
40-44	40-44	21006	6.57
45-49	45-49	21791	6.81
50-54	50-54	25382	7.94
55-59	55-59	29757	9.31
60-64	60-64	33686	10.53
65-69	65-69	34151	10.68
70-74	70-74	31065	9.71
75-79	75-79	21482	6.72
80 or older	80 or older	24153	7.55

Race

Lable: Race
Column: 11
Type of Variable: Categorical
Column name: Race

Value	Value Lable	Frequency	Percentage
American Indian/Alaskan Native	American Indian/Alaskan Native	5202	1.63
Asian	Asian	8068	2.52
Black	Black	22939	7.17
Hispanic	Hispanic	27446	8.58
Other	Other	10928	3.42
White	White	245212	76.68

Diabetic

Lable: Diabetic or not
Column: 12
Type of Variable: Categorical
Column name: Sex

Value	Value Lable	Frequency	Percentage
Yes	Yes	40802	12.76
Yes (during pregnancy)	Yes (during pregnancy)	2559	0.80
No	No	269653	84.32
No, borderline diabetes	No, borderline diabetes	6781	2.12

PhysicalActivity

Lable: Leisure Time Physical Activity Calculated Variable
Column: 13
Type of Variable: Categorical
Column name: PhysicalActivity
Question: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job

Value	Value Lable	Frequency	Percentage
Yes	Yes	247957	77.54
No	No	71838	22.46

GenHealth

Lable: General Health
Column: 14
Type of Variable: Categorical
Column name: GenHealth

Value	Value Lable	Frequency	Percentage
Excellent	Excellent	66842	20.90
Very Good	Very Good	113858	35.60
Good	Good	93129	29.12
Fair	Fair	34677	10.84
Poor	Poor	11289	3.53

SleepTime

Lable: How much time do you sleep
Column: 15
Type of Variable: Continuous
Column name: SleepTime

Value	Value Lable	Frequency	Percentage
1-24	Number of Hours	319795	100

Asthma

Lable: Do you suffer from Asthma?
Column: 16
Type of Variable: Categorical
Column name: Asthma

Value	Value Lable	Frequency	Percentage
Yes	Do suffer from asthma	42872	13.41
No	Do not suffer from asthma	276923	86.59

KidneyDisease

Lable: Do you suffer from Kidney Disease?
Column: 17
Type of Variable: Categorical
Column name: KidneyDisease

Value	Value Lable	Frequency	Percentage
Yes	Do suffer from Kidney Disease	11779	3.68
No	Do not suffer from Kidney Disease	308016	96.32

SkinCancer

Lable: Do you suffer from Skin Cancer?
Column: 18
Type of Variable: Categorical
Column name: SkinCancer

Value	Value Lable	Frequency	Percentage
Yes	Do suffer from skin cancer	29819	9.32
No	Do not suffer from skin cancer	289976	90.68

Data Cleaning

Data Types

There were no mixed data types

Missing values

There were no missing values

Duplicates

There were 18078 duplicate rows, duplicates were removed. The dataset now has 301717 rows.

Columns & Renaming

No columns were dropped or renamed

Descriptive Statistics

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	301717.000000	301717.000000	301717.000000	301717.000000
mean	28.441970	3.572298	4.121475	7.084559
std	6.468134	8.140656	8.128288	1.467122
min	12.020000	0.000000	0.000000	1.000000
25%	24.030000	0.000000	0.000000	6.000000
50%	27.410000	0.000000	0.000000	7.000000
75%	31.650000	2.000000	4.000000	8.000000
max	94.850000	30.000000	30.000000	24.000000

Manual checks were done on the statistics to look for odd values that should be investigated, non were found.

Data Limitations and Ethics

No ethical concerns with PII data, as the dataset does not have any personal information, all the data is anonymous.

The data is limited to individuals above 18, therefor the analysis would not be applicable to early signs of heart disease in children.

As visible in the data dictionary, even before cleaning, the majority of the data (70%) is based on White people, therefor finding variations between races would be more difficult. However, races with a similar amount of data can be compared with one another. Ideally, similar race statistics should be used.

The data only covers information in 2020, so statistics will only be valid for 2020 findings, so analysis would have to be repeated once newer data is available.

Data Questions

Which health factors contribute the most to heart disease?

What are the differences in heart disease cases between white males and white females?

Is there a difference in heart disease cases in races besides white?

What ages are the most susceptible to heart disease?