

CMP9139 Research Methods Assessment

Analysis of Risk Factors for Heart Disease Using CDC Dataset: An Inferential Statistical Approach

Niracha Chaiwong 26683869

Dataset: Heart Disease recorded by CDC

- CDC is the Centers for Disease Control and Prevention of the United States
- Collects and analyses data on health-related topics
- NHANES: a survey that evaluates through interviews, physical examinations, and laboratory testing
- BRFSS is a state-based telephone which provides essential data to monitor health behaviours trends

heart_disease_health_indicators_BRFSS2015.csv (22.74 MB)

Detail Compact Column 22 of 22 columns

About this file

Remote source: https://www.kaggle.com/code/alexteboul/heart-disease-health-indicators-dataset-notebook

1 binary target variable: HeartDiseaseorAttack and 21 feature variables that are either binary or ordinal. No missing values. Column names are changed from the originals to be more readable/understandable See source notebook for how I cleaned the BRFSS2015 dataset.

# HeartDiseaseorAt...	# HighBP	# HighChol	# CholCheck	# BMI	# Smoker	# Stroke	# Diabetes	# PhysActivity
0 1	0 1	0 1	0 1	12 98	0 1	0 1	0 2	0
0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0
0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0
0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0
0.0	1.0	0.0	1.0	30.0	1.0	0.0	0.0	0.0
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0
1.0	1.0	1.0	1.0	30.0	1.0	0.0	2.0	0.0
0.0	0.0	0.0	1.0	24.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	25.0	1.0	0.0	2.0	1.0
0.0	1.0	1.0	1.0	34.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	26.0	1.0	0.0	0.0	0.0
0.0	1.0	1.0	1.0	28.0	0.0	0.0	2.0	0.0
0.0	0.0	1.0	1.0	33.0	1.0	1.0	0.0	1.0

The dataset contains 253,680 samples and 22 risk factors

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

Research Question

1. There is a significant association between cholesterol, smoking status, alcohol drinking, diabetes, general health, age, gender and the risk of heart disease.

2. There is a significant difference in the mean BMI, and mental health between individuals with and without heart disease.

3. There is a significant correlation between smoking status and alcohol drinking among individuals with heart disease

Methodology

1. Chi-square test: is used to find the difference between the factor's frequency, which can be used in nominal and ordinal data
2. Independent-sample t-test: is used to compare the mean difference between two factors which analyse quantity data
3. Correlation: is used to find the direction of the relationship between factors.

IBM

Power BI

Analysis Result

1. Chi-square test

Crosstab

		HeartDisease					
		No		Yes		Total	
		N	%	N	%	N	%
Sex	Female	131769	57.3%	10205	42.7%	141974	56.0%
	Male	98018	42.7%	13688	57.3%	111706	44.0%
Total		229787	100.0%	23893	100.0%	253680	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1880.387[a]	1	<.001		
Continuity Correction[b]	1879.793	1	<.001		

Pearson's Chi-Square value is less than a significant value of 0.05. So, it can be concluded that the H0 is rejected, or **heart disease has a relationship with gender**

1. Independent sample t-test

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
BMI	Equal variances assumed	49.440	<.001	-26.683	253678
	Equal variances not assumed			-26.181	28838.927
MentHlth	Equal variances assumed	3157.785	<.001	-32.616	253678
	Equal variances not assumed			-26.740	27012.310

The P-value is less than a significant value of 0.05. So, it can be concluded that the null hypothesis is rejected **or people with heart disease have a different BMI and mental health**.

Analysis Result

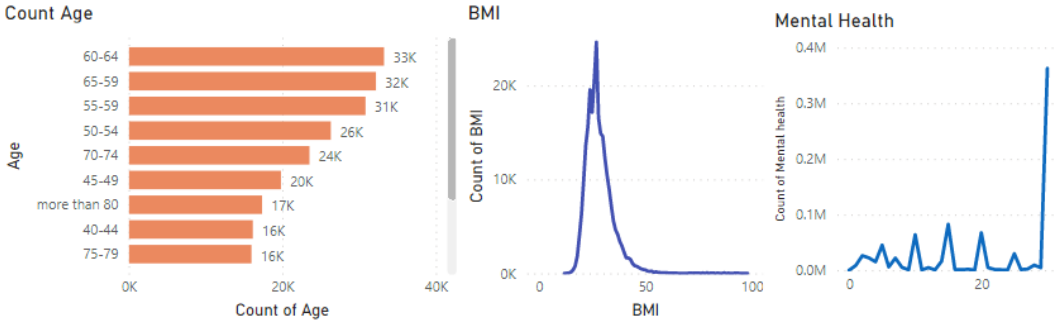
1. Correlation

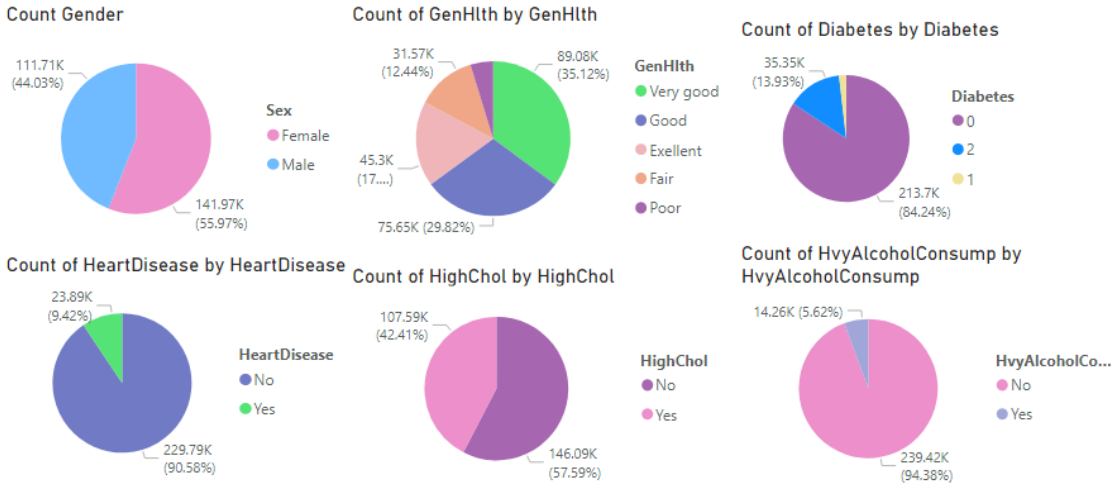
Correlations

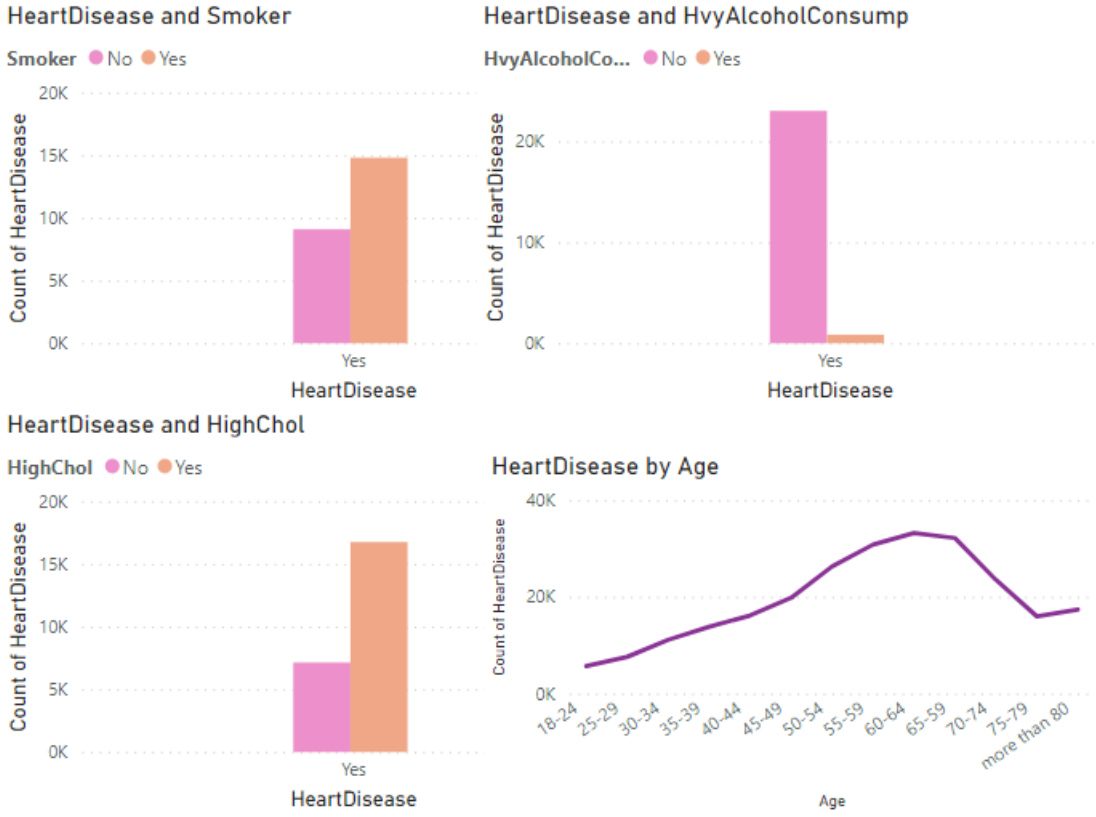
		HeartDisease	Diabetes	Smoker	HighChol	GenHlth
HeartDisease	Pearson Correlation	1	.180**	.114**	.181**	.258**
	Sig. (2-tailed)		<.001	<.001	<.001	<.001
	N	253680	253680	253680	253680	253680
Diabetes	Pearson Correlation	.180**	1	.063**	.209**	.303**
	Sig. (2-tailed)	<.001		<.001	<.001	<.001
	N	253680	253680	253680	253680	253680
Smoker	Pearson Correlation	.114**	.063**	1	.091**	.163**
	Sig. (2-tailed)	<.001	<.001		<.001	<.001
	N	253680	253680	253680	253680	253680
HighChol	Pearson Correlation	.181**	.209**	.091**	1	.208**
	Sig. (2-tailed)	<.001	<.001	<.001		<.001
	N	253680	253680	253680	253680	253680
GenHlth	Pearson Correlation	.258**	.303**	.163**	.208**	1
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	
	N	253680	253680	253680	253680	253680
MentHlth	Pearson Correlation	.065**	.074**	.092**	.062**	.302**
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001
	N	253680	253680	253680	253680	253680
BMI	Pearson Correlation	.053**	.224**	.014**	.107**	.239**
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001
	N	253680	253680	253680	253680	253680
HvyAlcoholConsump	Pearson Correlation	-.029**	-.058**	.102**	-.012**	-.037**
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001
	N	253680	253680	253680	253680	253680

All factors have a positive direction on heart disease except for alcohol consumption which has a negative value, indicating that **alcohol consumption increases, heart disease tends to decrease**

Analysis Result








Research Presentation End

Thank you