Exploratory Data Analysis

Diabetes Health Indicators Dataset

INFO 648 - Data Science Assignment
Allison Schneider
November 6, 2022

In [19]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
df = pd.read_csv('Diabetes_Dataset.csv')
```

In [4]:
```
df.head()
```

Out[4]:

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	Phys
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	18.0	
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0	0.0	
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0	30.0	
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0	
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	3.0	

5 rows × 22 columns

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```

In [21]:
```
df.describe()
```

Out[21]:

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	A
count	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	...	2
mean	0.296921	0.429001	0.424121	0.962670	28.382364	0.443169	0.040571	0.094186	0.756544	0.634256	...	
std	0.698160	0.494934	0.494210	0.189571	6.608694	0.496761	0.197294	0.292087	0.429169	0.481639	...	
min	0.000000	0.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	
25%	0.000000	0.000000	0.000000	1.000000	24.000000	0.000000	0.000000	0.000000	1.000000	0.000000	...	
50%	0.000000	0.000000	0.000000	1.000000	27.000000	0.000000	0.000000	0.000000	1.000000	1.000000	...	
75%	0.000000	1.000000	1.000000	1.000000	31.000000	1.000000	0.000000	0.000000	1.000000	1.000000	...	
max	2.000000	1.000000	1.000000	1.000000	98.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	

8 rows × 22 columns

In [22]:
```
1 df.dtypes
```

Out[22]:
```
Diabetes_012            float64
HighBP                  float64
HighChol                float64
CholCheck               float64
BMI                     float64
Smoker                  float64
Stroke                  float64
HeartDiseaseorAttack    float64
PhysActivity            float64
Fruits                  float64
Veggies                 float64
HvyAlcoholConsump       float64
AnyHealthcare           float64
NoDocbcCost             float64
GenHlth                 float64
MentHlth                float64
PhysHlth                float64
DiffWalk                float64
Sex                     float64
Age                     float64
Education               float64
Income                  float64
dtype: object
```

In [6]:
```
1 df.shape
```

Out[6]:
```
(253680, 22)
```

In [7]:
```
1 df.isnull().sum()
```

Out[7]:
```
Diabetes_012            0
HighBP                  0
HighChol                0
CholCheck               0
BMI                     0
Smoker                  0
Stroke                  0
HeartDiseaseorAttack    0
PhysActivity            0
Fruits                  0
Veggies                 0
HvyAlcoholConsump       0
AnyHealthcare           0
NoDocbcCost             0
GenHlth                 0
MentHlth                0
PhysHlth                0
DiffWalk                0
Sex                     0
Age                     0
Education               0
Income                  0
dtype: int64
```

In [8]:
```
1 df.nunique()
```

Out[8]:
```
Diabetes_012             3
HighBP                   2
HighChol                 2
CholCheck                2
BMI                     84
Smoker                   2
Stroke                   2
HeartDiseaseorAttack     2
PhysActivity             2
Fruits                   2
Veggies                  2
HvyAlcoholConsump        2
AnyHealthcare            2
NoDocbcCost              2
GenHlth                  5
MentHlth                31
PhysHlth                31
DiffWalk                 2
Sex                      2
Age                     13
Education                6
Income                   8
dtype: int64
```

In [9]:
```
1 df['Diabetes_012'].value_counts()
```

Out[9]:
```
0.0    213703
2.0     35346
1.0      4631
Name: Diabetes_012, dtype: int64
```

In [15]:
```
dfcorr = df.corr()
dfcorr
```

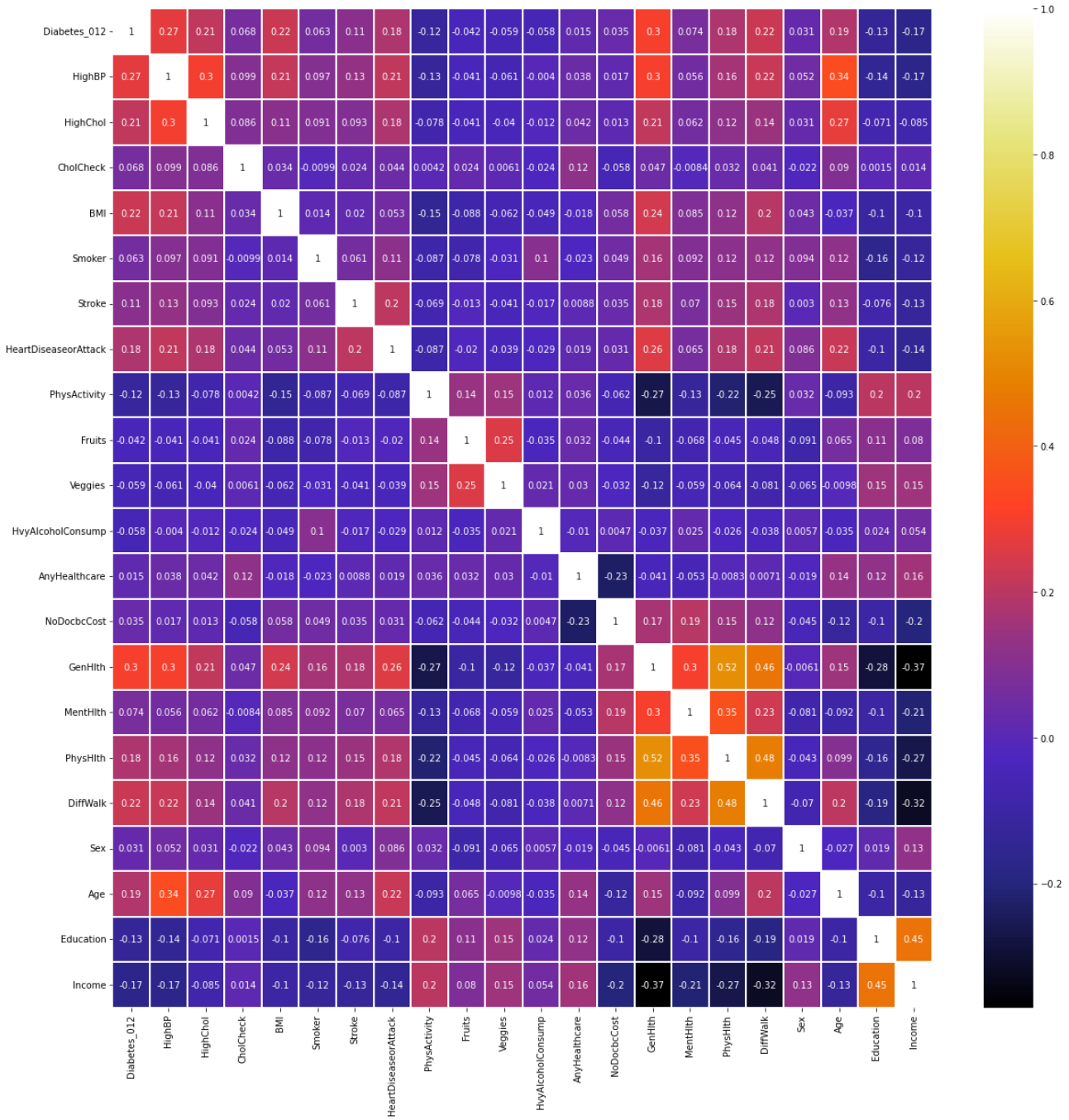
Out[15]:

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDo
Diabetes_012	1.000000	0.271596	0.209085	0.067546	0.224379	0.062914	0.107179	0.180272	-0.121947	-0.042192	...	0.015410	0
HighBP	0.271596	1.000000	0.298199	0.098508	0.213748	0.096991	0.129575	0.209361	-0.125267	-0.040555	...	0.038425	0
HighChol	0.209085	0.298199	1.000000	0.085642	0.106722	0.091299	0.092620	0.180765	-0.078046	-0.040859	...	0.042230	0
CholCheck	0.067546	0.098508	0.085642	1.000000	0.034495	-0.009929	0.024158	0.044206	0.004190	0.023849	...	0.117626	-0
BMI	0.224379	0.213748	0.106722	0.034495	1.000000	0.013804	0.020153	0.052904	-0.147294	-0.087518	...	-0.018471	0
Smoker	0.062914	0.096991	0.091299	-0.009929	0.013804	1.000000	0.061173	0.114441	-0.087401	-0.077666	...	-0.023251	0
Stroke	0.107179	0.129575	0.092620	0.024158	0.020153	0.061173	1.000000	0.203002	-0.069151	-0.013389	...	0.008776	0
HeartDiseaseorAttack	0.180272	0.209361	0.180765	0.044206	0.052904	0.114441	0.203002	1.000000	-0.087299	-0.019790	...	0.018734	0
PhysActivity	-0.121947	-0.125267	-0.078046	0.004190	-0.147294	-0.087401	-0.069151	-0.087299	1.000000	0.142756	...	0.035505	-0
Fruits	-0.042192	-0.040555	-0.040859	0.023849	-0.087518	-0.077666	-0.013389	-0.019790	0.142756	1.000000	...	0.031544	-0
Veggies	-0.058972	-0.061266	-0.039874	0.006121	-0.062275	-0.030678	-0.041124	-0.039167	0.153150	0.254342	...	0.029584	-0
HvyAlcoholConsump	-0.057882	-0.003972	-0.011543	-0.023730	-0.048736	0.101619	-0.016950	-0.028991	0.012392	-0.035288	...	-0.010488	0
AnyHealthcare	0.015410	0.038425	0.042230	0.117626	-0.018471	-0.023251	0.008776	0.018734	0.035505	0.031544	...	1.000000	-0
NoDocbcCost	0.035436	0.017358	0.013310	-0.058255	0.058206	0.048946	0.034804	0.031000	-0.061638	-0.044243	...	-0.232532	1
GenHlth	0.302587	0.300530	0.208426	0.046589	0.239185	0.163143	0.177942	0.258383	-0.266186	-0.103854	...	-0.040817	0
MentHlth	0.073507	0.056456	0.062069	-0.008366	0.085310	0.092196	0.070172	0.064621	-0.125587	-0.068217	...	-0.052707	0
PhysHlth	0.176287	0.161212	0.121751	0.031775	0.121141	0.116460	0.148944	0.181698	-0.219230	-0.044633	...	-0.008276	0
DiffWalk	0.224239	0.223618	0.144672	0.040585	0.197078	0.122463	0.176567	0.212709	-0.253174	-0.048352	...	0.007074	0
Sex	0.031040	0.052207	0.031205	-0.022115	0.042950	0.093662	0.002978	0.086096	0.032482	-0.091175	...	-0.019405	-0
Age	0.185026	0.344452	0.272318	0.090321	-0.036618	0.120641	0.126974	0.221618	-0.092511	0.064547	...	0.138046	-0
Education	-0.130517	-0.141358	-0.070802	0.001510	-0.103932	-0.161955	-0.076009	-0.099600	0.199658	0.110187	...	0.122514	-0
Income	-0.171483	-0.171235	-0.085459	0.014259	-0.100069	-0.123937	-0.128599	-0.141011	0.198539	0.079929	...	0.157999	-0

22 rows × 22 columns

In [20]:

```
plt.figure(figsize = (20,20))
sns.heatmap(dfcorr,linewidth=0.1, annot = True, cmap='CMRmap')
plt.yticks(rotation=0)
plt.show()
```



Answering Questions from The Data

1. Are patients with high blood pressure and high cholesterol more likely to also have diabetes?

In [47]:
```
df_bp_chol = df[(df.HighBP == 1) & (df.HighChol == 1)]
diabetes_count = df_bp_chol[(df_bp_chol['Diabetes_012'] == 1) | (df_bp_chol['Diabetes_012'] == 2)].shape[0]
no_diabetes_count = df_bp_chol[(df_bp_chol['Diabetes_012'] == 0)].shape[0]
total_count = df_bp_chol.shape[0]

diabetes_percent = (diabetes_count / total_count) * 100
no_diabetes_percent = (no_diabetes_count / total_count) * 100

print(round(diabetes_percent,2), '% of patients with high blood pressure and high cholestorol also have diabetes.')
print(round(no_diabetes_percent,2),
      '% of patients with high blood pressure and high cholestorol do not also have diabetes.')
```

```
32.84 % of patients with high blood pressure and high cholestorol also have diabetes.
67.16 % of patients with high blood pressure and high cholestorol do not also have diabetes.
```

2. Does Diabetes affect more women or men?

In [43]:
```
df_diabetes = df[(df['Diabetes_012'] == 1) | (df['Diabetes_012'] == 2)]
female_count = df_diabetes[(df_diabetes['Sex'] == 0)].shape[0]
male_count = df_diabetes[(df_diabetes['Sex'] == 1)].shape[0]
total_count = df_diabetes.shape[0]

female_percent = (female_count / total_count) * 100
male_percent = (male_count / total_count) * 100

print(round(female_percent,2), '% of patients with diabetes are female.')
print(round(male_percent,2), '% of patients with diabetes are male.')
```

```
52.57 % of patients with diabetes are female.
47.43 % of patients with diabetes are male.
```

3. Does diabetes affect more patients with poor health?

In [37]:
```
df_poor_health = df[(df['PhysActivity'] == 0) & # no physical activity in the last 30 days
                    (df['HvyAlcoholConsump'] == 1) & # heavy alcohol consumption
                    ((df['GenHlth'] == 4) | (df['GenHlth'] == 5)) & # general health of fair or poor
                    (df['PhysHlth'] > 10) & # physical health was not good for more than 10 days in the month
                    (df['DiffWalk'] == 1)] # serious difficulty walking or climbing stairs

diabetes_count = df_poor_health[(df_poor_health['Diabetes_012'] == 1) | (df_poor_health['Diabetes_012'] == 2)].shape[0]
total_count = df_poor_health.shape[0]

diabetes_percent = (diabetes_count / total_count) * 100

print(round(diabetes_percent,2), '% of patients with poor health also have diabetes.')
```

```
23.53 % of patients with poor health also have diabetes.
```

In [48]:
```
df_good_health = df[(df['PhysActivity'] == 1) & # physical activity in the last 30 days
                    (df['HvyAlcoholConsump'] == 0) & # no heavy alcohol consumption
                    ((df['GenHlth'] == 1) | (df['GenHlth'] == 2) | (df['GenHlth'] == 3)) & # excellent/very good/good
                    (df['PhysHlth'] < 10) & # physical health was not good for less than 10 days in the month
                    (df['DiffWalk'] == 0)] # no serious difficulty walking or climbing stairs

diabetes_count = df_good_health[(df_good_health['Diabetes_012'] == 1) | (df_good_health['Diabetes_012'] == 2)].shape[0]
total_count = df_good_health.shape[0]

diabetes_percent = (diabetes_count / total_count) * 100

print(round(diabetes_percent,2), '% of patients with good health have diabetes.')
```

```
9.4 % of patients with good health have diabetes.
```

4. Do paitents with lower incomes have a higher chance of diabetes?

In [45]:
```
df_low_income = df[(df['AnyHealthcare'] == 0) & # no healthcare coverage
                    (df['NoDocbcCost'] == 1) & # needed to see a doc but couldnt because of cost
                    (df['Income'] < 8) & # income less than $75K
                    (df['Education'] < 5)] # no college education

diabetes_count = df_low_income[(df_low_income['Diabetes_012'] == 1) | (df_low_income['Diabetes_012'] == 2)].shape[0]
total_count = df_low_income.shape[0]

diabetes_percent = (diabetes_count / total_count) * 100

print(round(diabetes_percent,2), '% of low income patients have diabetes.')
```

```
19.27 % of low income patients have diabetes.
```

In [46]:
```
df_high_income = df[(df['AnyHealthcare'] == 1) & # healthcare coverage
                    (df['NoDocbcCost'] == 0) & # never needed to see a doc but couldnt because of cost
                    (df['Income'] >= 8) & # income greater than $75K
                    (df['Education'] > 4)] # college educated

diabetes_count = df_high_income[(df_high_income['Diabetes_012'] == 1) | (df_high_income['Diabetes_012'] == 2)].shape[0]
total_count = df_high_income.shape[0]

diabetes_percent = (diabetes_count / total_count) * 100

print(round(diabetes_percent,2), '% of high income patients have diabetes.')
```

```
8.78 % of high income patients have diabetes.
```