Data Analysis – Diabetes

Variables:

A short overview of the variables present in the dataset.

Year

Measure	Value	Null
Count	10000.000	Not null
Mean	2018.360	Not null
Standard deviation	1.345	Not null
Min	2015	Not Null
25% Quantile	2015	Not null
Median	2019	Not null
75% Quantile	2019	Not null
Max	2022	Not null

Gender

	Male	Female
Count	41430	58552

The Male population is less represented in comparison to the female population. To see whether the race plays a role in this representation:

	African American	Asian	Caucasian	Hispanic	Other
Male Count of Dataset	8414	8297	8144	8284	8291
Female Count of Dataset	11807	11716	11723	11601	11705

There is a clear bias towards the Female population with respect to representation in the population. When split up by population, the differences between the races are not significantly different.

Age

Measure	Value	Null
Count	10000.000	Not null
Mean	41.885	Not null
Standard deviation	22.516	Not null
Min	0.000	Not null
25% Quantile	24.000	Not null
Median	43.000	Not null
75% Quantile	60.000	Not null
Max	80.000	Not null

Location

Place	Count
Kentucky	2038
Iowa	2038
Hawaii	2038
Nebraska	2038
Florida	2037
Minnesota	2037
Arkansas	2037
New Jersey	2037
Massachusetts	2036
Kansas	2036
Louisiana	2036
District of Columbia	2036
Maine	2036
Delaware	2036
Georgia	2036
Michigan	2036
Illinois	2036
Pennsylvania	2036
Oregon	2036
Alabama	2036
Connecticut	2035
Maryland	2035
Alaska	2035
North Dakota	2035
New York	2035
North Carolina	2035
Mississippi	2035
Rhode Island	2035
Colorado	2035
Missouri	2035
New Hampshire	2035
New Mexico	2033
South Dakota	2033
Montana	2033
Idaho	1988
South Carolina	1987
Indiana	1987
Arizona	1986
California	1986
Nevada	1986
Oklahoma	1986
Ohio	1986

Tennessee	1574
United States	1401
Washington	1363
Utah	1359
Virginia	1350
Vermont	1338
Texas	1337
Puerto Rico	1295
Guam	1204
West Virginia	1132
Virgin Islands	763
Wisconsin	388
Wyoming	388

Race

	African American	Asian	Caucasian	Hispanic	Other
In % of Dataset	0.202230	0.200150	0.198760	0.19888	0.199980

The Race categories are binary – either 0 or 1. The data is balanced and well defined.

Hypertension

Measure	Value	Null
Count	10000	Not null
Mean	0.074	Not null
Standard deviation	0.263	Not null
Min	0.000	Not null
25% Quantile	0.000	Not null
Median	0.000	Not null
75% Quantile	0.000	Not null
Max	1.000	Not null

There are only 7485 subjects with the hypertension condition – with following distribution per race:

	African American	Asian	Caucasian	Hispanic	Other
Count Hypertension	1501	1540	1493	1503	1448

There are few differences between the races with respect to Hypertension. Should be used carefully as diabetes indicator due to small sample size.

Heart_disease

Measure	Value	Null
Count	10000.000	Not null
Mean	0.039	Not null
Standard deviation	0.194	Not null

Min	0.000	Not null
25% Quantile	0.000	Not null
Median	0.000	Not null
75% Quantile	0.000	Not null
Max	1.000	Not null

There are only 3942 subjects with the heart_disease condition – with following distribution per race:

	African American	Asian	Caucasian	Hispanic	Other
Count heart disease	792	837	774	778	761

There are few differences between the races with respect to heart disease. Should be used carefully as diabetes indicator due to small sample size.

Smoking_history

Value	Count
No Info	35816
Never	35095
Former	9352
Current	9286
Not current	6447
Ever	4004

Male distribution of smokers:

	African American	Asian	Caucasian	Hispanic	Other
No Info	3241	3260	3148	3261	3200
Never	2494	2427	2402	2460	2440
Former	980	906	858	894	940
Current	809	819	855	881	864
Not current	532	518	501	471	504
Ever	358	819	855	881	864

Female distribution of smokers:

	African American	Asian	Caucasian	Hispanic	Other
No Info	3918	4005	3879	3891	4007
Never	4682	4574	4601	4434	4578
Former	926	955	973	975	945
Current	1018	965	809	776	760
Not current	706	772	809	776	760
Ever	567	445	431	463	432

BMI

Measure	Value	Null
Count	10000	Not null
Mean	27.320767	Not null
Standard deviation	6.636783	Not null
Min	10.01	Not null
25% Quantile	23.63	Not null
Median	27.32	Not null
75% Quantile	29.58	Not null
Max	95.69	Not null

Mean of BMI across race:

	African American	Asian	Caucasian	Hispanic	Other
Mean BMI	27.304	27.390	27.292	27.352	27.264

The mean BMI is consistent across the races – there are no severe outliers.

hbA1c_level

Measure	Value	Null
Count	10000	Not null
Mean	5.527	Not null
Standard deviation	1.070	Not null
Min	3.500	Not null
25% Quantile	4.800	Not null
Median	5.800	Not null
75% Quantile	6.200	Not null
Max	9.000	Not null

Mean of hbA1c_level across race:

	African American	Asian	Caucasian	Hispanic	Other
Mean hbA1c_level	5.530	5.526	5.518	5.528	5.533

The mean of hbA1c_level across the different races is consistent at around 5.5.

blood_glucose_level

Measure	Value	Null
Count	10000	Not null
Mean	138.058	Not null
Standard deviation	40.708	Not null
Min	80.000	Not null
25% Quantile	100.000	Not null
Median	140.000	Not null
75% Quantile	159.000	Not null

Max	300.000	Not null

Mean of blood_glucose_level across race:

	African American	Asian	Caucasian	Hispanic	Other
Mean blood_glucose_level	138.243	138.071	138.394	137.838	137.740

The mean of blood_glucose_level across the different races is consistent at around 137/138.

diabetes

Measure	Value	Null
Count	10000.000	Not null
Mean	0.085	Not null
Standard deviation	0.278	Not null
Min	0.000	Not null
25% Quantile	0.000	Not null
Median	0.000	Not null
75% Quantile	0.000	Not null
Max	1.000	Not null

Sum of diabetes per race

	African American	Asian	Caucasian	Hispanic	Other
Sum diabetes per race	1768	1743	1670	1676	1643

Results

Using Chi^2, the resulting variables chosen are the following: age, hypertension, bmi, hbA1c_level and blood_glucose_level.
These variables will now be used in a prediction model.