Project Documentation - Information Visualization

Hello, Healthcare Data Sketch!

Elena Hrincescu, Emma Ionescu

Group 511

Data Overview & Preprocessing

Table 1: **Initial DataFrame Details**

Property	**Value**
DataFrame Shape	(445132, 40)
Number of Columns	40
Number of rows with NaN values	199110
Columns with NaN values	True
Has Duplicates	True
Columns	State, Sex, GeneralHealth, PhysicalHealthDays, MentalHealthDays, LastCheckupTime, PhysicalActivities, SleepHours, RemovedTeeth, HadHeartAttack, HadAngina, HadStroke, HadAsthma, HadSkinCancer, HadCOPD, HadDepressiveDisorder, HadKidneyDisease, HadArthritis, HadDiabetes, DeafOrHardOfHearing, BlindOrVisionDifficulty, DifficultyConcentrating, DifficultyWalking, DifficultyDressingBathing, DifficultyErrands, SmokerStatus, ECigaretteUsage, ChestScan, RaceEthnicityCategory, AgeCategory, HeightInMeters, WeightInKilograms, BMI, AlcoholDrinkers, HIVTesting, FluVaxLast12, PneumoVaxEver, TetanusLast10Tdap, HighRiskLastYear, CovidPos

Usually, duplicates are dropped from a dataframe, but we cannot guarantee that there are not 2 individuals with the same health problems, therefore we decided to keep them.

First, we decided to remove the following columns:

```
data = data.drop(columns=['RemovedTeeth', 'LastCheckupTime',
                          'HIVTesting', 'FluVaxLast12',
                          'PneumoVaxEver', 'TetanusLast10Tdap',
                          'ECigaretteUsage'])
```

Reasoning:

Columns	**Values**	**Reasoning**
RemovedTeeth	'1 to 5', '6 or more, but not all', 'all', 'nan', 'none of them'	It is a bit vague, some people might lose their teeth because of an accident, some due to pregnancy, some due to old age, it does not necessarly indicate if someone has or not health problems
LastCheckupTime	'5 or more years ago', 'nan', 'within past 2 years (1 year but less than 2 years ago)', 'within past 5 years (2 years but less than 5 years ago)', 'within past year (anytime less than 12 months ago	It does not indicate if someone has or not health problems, some people are more hypochondriac than others and get tested for various reasons
HIVTesting	Yes/ No	It does not indicate if someone has or not the disease, some people are more hypochondriac than others and get tested for various reasons
FluVaxLast12	'nan', 'no', 'yes'	It does not indicate if someone has or not health problems, some people are more hypochondriac than others and get vaccines for various reasons
PneumoVaxEver	'nan', 'no', 'yes	It does not indicate if someone has or not health problems, some people are more hypochondriac than others and get vaccines for various reasons
TetanusLast10Tdap	'nan', 'no, did not receive any tetanus shot in the past 10 years', 'yes, received tdap', 'yes, received tetanus shot but not sure what type', 'yes, received tetanus shot, but not tdap'	Too vague, it does not indicate if someone has or not health problems
ECigaretteUsage	'nan', 'never used e-cigarettes in my entire life', 'not at all (right now)', 'use them every day', 'use them some days'	We already have a SmokerStatus column , also high number of NaNs

NaN handling - Assumptions

In order to fill some of the missing values we made the following assumptions and put the following conditions:

1. For the column **BMI**:

- we filled NaNs using the data available, in ’HeightInMeters’ and ’WeightInKilograms’.

! Following, **we also dropped these 2 columns** considering that we have the BMI column and the have a lot of unique values.

2. For the column **ChestScan**:

- if the columns:

- HadHeartAttack == ’yes’
- HadAngina == ’yes’
- HadStroke == ’yes’
- HadCOPD == ’yes’

- it is obvious that a chest scan was performed. So in those cases, the NaN values were filled with ’yes’.

3. For the columns **DifficultyWalking**, **DifficultyDressingBathing**, **DifficultyErrands**:

- if the column:

- HadArthritis == ’yes’

- it is obvious that having arthritis is affecting the other conditions. So in this case, the NaN values were filled with ’yes’.

4. For the columns **HadDepressiveDisorder**:

- if the value in ’MentalHealthDays’ is greater than 15, we decided that the person probably has a level of depressive disorder.

5. For the column **GeneralHealth**:

- **Excellent Condition**: If all the health-related conditions are ’no’, the individual never smoked, and they are physically active, it is assumed their general health is excellent. Thus, NaN values in these cases are filled with ’excellent’.
  - Conditions checked include HadHeartAttack, HadAngina, HadStroke, HadAsthma, HadSkinCancer, HadCOPD, HadDepressiveDisorder, HadKidneyDisease, HadArthritis, HadDiabetes, DeafOrHardOfHearing, BlindOrVisionDifficulty, DifficultyConcentrating, DifficultyWalking, DifficultyDressingBathing, DifficultyErrands, HighRiskLastYear.
  - Additional checks: SmokerStatus is ’never smoked’, PhysicalActivities is ’yes’.

- **Good Condition**: If all the health-related conditions are ’no’ and the individual is either a non-smoker or smokes occasionally, and they are physically active and drink alcohol either yes or no, then their general health is considered good. Hence, NaN values in these cases are filled with ’good’.
  - Same health conditions as for excellent.
  - Additional checks: SmokerStatus includes ’never smoked’, ’former smoker’, ’current smoker - now smokes some days’, and AlcoholDrinkers includes ’yes’, ’no’.
- **Poor Condition**: If three or more of the health-related conditions are ’yes’, it indicates poor general health. Therefore, NaN values here are filled with ’poor’.

NaN Handling - Chi-Square Test - Verifying Assumptions

1. ChestScan

HadHeartAttack

Chi-Squared statistic: 14519.83, P-value: 0.0

HadHeartAttack	**no**	**yes**
no	217384	153615
yes	4921	20187

Interpretation: The p-value indicates a significant relationship between experiencing hearth attack and the results of the chest scan.

HadAngina

Chi-Squared statistic: 16372.62, P-value: 0.0

HadAngina	**no**	**yes**
no	217286	150955
yes	4958	21593

Interpretation: The p-value indicates a significant relationship between experiencing angina and the results of the chest scan.

HadStroke

Chi-Squared statistic: 9952.33, P-value: 0.0

HadStroke	**no**	**yes**
no	218752	159395
yes	4089	15150

Interpretation: The p-value indicates a significant relationship between experiencing stroke and the results of the chest scan.

HadCOPD

Chi-Squared statistic: 20238.36, P-value: 0.0

HadCOPD	no	yes
no	215335	145921
yes	7279	28377

Interpretation: The p-value indicates a significant relationship between experiencing CODP and the results of the chest scan.

2. DifficultyWalking, DifficultyDressingBathing, DifficultyErrands

DifficultyWalking

Chi-Squared statistic: 55825.68, P-value: 0.0

DifficultyWalking	no	yes
no	254492	96808
yes	19847	54340

Interpretation: The p-value indicates a significant relationship between having arthritis and difficulty in walking.

DifficultyDressingBathing

Chi-Squared statistic: 19498.79, P-value: 0.0

DifficultyDressingBathing	no	yes
no	269155	133131
yes	4953	18017

Interpretation: The p-value suggests a significant relationship between having arthritis and difficulty in dressing or bathing.

DifficultyErrands

Chi-Squared statistic: 18410.39, P-value: 0.0

DifficultyErrands	no	yes
no	260096	124968
yes	12928	26180

Interpretation: The p-value indicates a significant relationship between having arthritis and difficulty in managing errands.

3. Mental Health Days

Chi-Squared statistic: 92535.21, P-value: 0.0

Mental Health Days	No Depressive Disorder	With Depressive Disorder
0.0	243211	21219
1.0	11890	2438
2.0	18323	5320
3.0	10988	4269
4.0	5475	2418
5.0	13360	6444
6.0	1413	867
7.0	4850	2909
8.0	1056	679
9.0	176	143
10.0	8696	6572
11.0	61	58
12.0	655	583
13.0	81	86
14.0	1548	1301
15.0	6814	7538
16.0	116	164
17.0	90	154
18.0	134	185
19.0	20	27
20.0	3648	5502
21.0	244	305
22.0	88	105
23.0	31	66
24.0	43	81
25.0	1069	2009
26.0	40	66
27.0	88	153
28.0	343	567
29.0	215	287
30.0	9792	17198

Interpretation: The statistic and the p-value of zero strongly suggest that there is a significant association between having a depressive disorder and the number of mental health days reported. Notably, as the number of days increases, especially beyond 10 days per month, there is an increase in the proportion of individuals with a depressive disorder compared to those without.

!! It can be noted that our initial assumptions were supported by the Chi-Square Test

!! The rest of the rows that contained NaN values were dropped, considering that we have a big enough dataframe

Data Binning

Binned Columns	**Values Before Binning**	**Values After Binning**
PhysicalHealthDays, MentalHealthDays	Various numerical values (0-30 days)	none (0 days), low (1-5 days), moderate (6-15 days), high (16-30 days)
SleepHours	Various numerical values (0-24 hours)	very low (0-5 hours), low (6-7 hours), normal (8-9 hours), high (10 hours), very high (11-24 hours)
SmokerStatus	current smoker - now smokes every day, current smoker - now smokes some days, former smoker, never smoked	current smoker, former smoker, non-smoker
CovidPos	no, yes, tested positive using home test without a health professional	no, yes
AgeCategory	age 18 to 24, age 25 to 29, age 30 to 34, age 35 to 39, age 40 to 44, age 45 to 49, age 50 to 54, age 55 to 59, age 60 to 64, age 65 to 69, age 70 to 74, age 75 to 79, age 80 or older	18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+
RaceEthnicityCategory	black only, non-hispanic, hispanic, multiracial, non-hispanic, other race only, non-hispanic, white only, non-hispanic	black, hispanic, multiracial, other, white

Final Data Details

Table 2: **Final DataFrame Details**

Property	**Value**
DataFrame Shape	(312266, 31)
Number of Columns	31
Number of rows with NaN values	0
Columns with NaN values	False
Columns	State, Sex, GeneralHealth, PhysicalHealthDays, MentalHealthDays, PhysicalActivities, SleepHours, HadHeartAttack, HadAngina, HadStroke, HadAsthma, HadSkinCancer, HadCOPD, HadDepressiveDisorder, HadKidneyDisease, HadArthritis, HadDiabetes, DeafOrHardOfHearing, BlindOrVisionDifficulty, DifficultyConcentrating, DifficultyWalking, DifficultyDressingBathing, DifficultyErrands, SmokerStatus, ChestScan, RaceEthnicityCategory, AgeCategory, BMI, AlcoholDrinkers, HighRiskLastYear, CovidPos

Data Visualization Plan

Columns Categorization

Lifestyle	Medical Stuff	Characteristics
PhysicalHealthDays	GeneralHealth	State
MentalHealthDays	HadHeartAttack	Sex
PhysicalActivities	HadAngina	RaceEthnicityCategory
SleepHours	HadStroke	AgeCategory
SmokerStatus	HadAsthma	BMI
AlcoholDrinkers	HadSkinCancer	
	HadCOPD	
	HadDepressiveDisorder	
	HadKidneyDisease	
	HadArthritis	
	HadDiabetes	
	DeafOrHardOfHearing	
	BlindOrVisionDifficulty	
	DifficultyConcentrating	
	DifficultyWalking	
	DifficultyDressingBathing	
	DifficultyErrands	
	HighRiskLastYear	
	CovidPos	
	ChestScan	

Data Visualization

1. Lifestyle Data Associations

1 vs 1 Relationships

Variables	$\rightarrow$
PhysicalActivities	General Health
SmokerStatus	General Health
SleepHours	General Health
MentalHealthDays	PhysicalActivities

2 vs 1 Relationships

Variable Pairs	→
PhysicalHealthDays, MentalHealthDays	General Health
SmokerStatus, AlcoholDrinkers	General Health
SmokerStatus, AlcoholDrinkers	SleepHours
PhysicalHealthDays, MentalHealthDays	HadDepression

n vs 1 Relationships

Variable Group	→
PhysicalHealthDays, MentalHealthDays, PhysicalActivities, SleepHours, SmokerStatus, AlcoholDrinkers	General Health
PhysicalHealthDays, MentalHealthDays, PhysicalActivities, SleepHours, SmokerStatus, AlcoholDrinkers	HighRiskLastYear
PhysicalHealthDays, MentalHealthDays, PhysicalActivities, SleepHours, SmokerStatus, AlcoholDrinkers	HasKidney

n vs n Relationships

Variable Group	→
PhysicalActivities, SleepHours, SmokerStatus, AlcoholDrinkers	HadHeartAttack, HadAngina, HadStroke, HadCOPD
PhysicalActivities, SleepHours, SmokerStatus, AlcoholDrinkers	HadAsthma, DifficultyConcentrating, HadKidneyDisease
DeafOrHardOfHearing, BlindOrVisionDifficulty, DifficultyConcentrating, DifficultyWalking, DifficultyDressingBathing, DifficultyErrands	PhysicalHealthDays, MentalHealthDays, PhysicalActivities, SleepHours, SmokerStatus, AlcoholDrinkers

2. Medical Stuff Data Associations

1 vs 1 Relationships

Variables	→
HadAsthma	Covid
HadAsthma	DifficultyErrands

2 vs 1 Relationships

Variable Pairs	→
HadKidneyDisease, HadDiabetes	HighRiskLastYear
HadAsthma, CovidPos	ChestScan

n vs 1 Relationships

Variable Group	→
HadHeartAttack, HadAngina, HadStroke	General Health
HadAsthma, HadAsthma, Covid	General Health

n vs n Relationships

Variable Group	→
HadHeartAttack, HadAngina, HadStroke, HadAsthma, HadSkinCancer, HadCOPD, HadDepressiveDisorder, HadKidneyDisease, HadArthritis, HadDiabetes	DeafOrHardOfHearing, BlindOrVisionDifficulty, DifficultyConcentrating, DifficultyWalking, DifficultyDressingBathing, DifficultyErrands
DeafOrHardOfHearing, BlindOrVisionDifficulty, DifficultyConcentrating, DifficultyWalking, DifficultyDressingBathing, DifficultyErrands	PhysicalHealthDays, MentalHealthDays, PhysicalActivities, SleepHours, SmokerStatus, AlcoholDrinkers

3. Characteristics Data Associations

1 vs 1 Relationships

Variables	→
Sex	General Health
State	Sex / Age
Age	General Health
Race	General Health
Race	Had Skin Cancer

2 vs 1 Relationships

Variable Pairs	→
Age, BMI	General Health
State, Race	General Health

n vs 1 Relationships

Variable Group	→
State, Sex, Age, Race	CovidPos
State, Sex, Age, Race	MentalHealthDays
State, Sex, Age, Race	PhysicalHealthDays

n vs n Relationships

Variable Group	→
State, Sex, Age, Race	HasDiabetes, Alcohol, Smoke, HasDepression
State, Sex, Age, Race	HadHeartAttack, HadAngina, HadStroke
Sex, Age, Race	DeafOrHardOfHearing, BlindOrVisionDifficulty, DifficultyConcentrating, DifficultyWalking, DifficultyDressingBathing, DifficultyErrands