



## Table of Contents

- About Our Project: Data Source, Questions, Technologies
- Database & Connections
- Analysis
- Machine Learning
- Results
- Future Analysis Recommendations





## About Our Project













Our group had a shared interest in and/or background in the health industry.











# About Our Project





#### **Data Source**



Health-related telephone survey in the US.

>400k interviews each year.

(dataset from the CDC <u>via Kaggle</u>)







## About Our Project







#### Questions:

Which health choices & common diseases are most correlated with mental health?

Could you predict mental health outcomes using physical health indicators?

Do other indicators have an impact on mental health?







## Technologies We Used









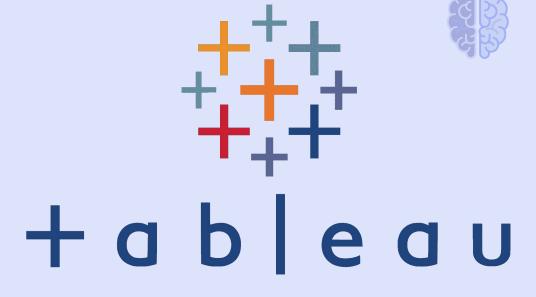












## Database and Connections



#### 2015 Behavioral Risks Factors Survey



#### Codebook

Mapped columns in the dataset to descriptions, frequency, percentages, etc.

### S3 + PySpark

Loaded data into an S3 bucket.

Extracted data using pyspark.

Transformed data and Loaded Tables to SQL DataBase

#### Postgres+RDS

Used SQL Database on RDS Server

Used psycopg2 to connect Database to ML model and visualization tools





# Analysis

### Data Exploration ...



#### **Understanding Data:**

++		<b></b> -	+	<b>+</b>	<b></b>	<del></del>	++		H		+	+-
PHYSHLTH	POORHLTH	GENHLTH	MENTHLTH	BPHIGH4	TOLDHI2	CVDINFR4	CVDCRHD4	CVDSTRK3	CHCSCNCR	CHCOCNCR	CHCCOPD1	HAVARTH3   C
++	·	+	+	+	<del></del>	<del></del>	++		tt		++	+-
15	10	5	18	1	1	2	2	2	2	2	1	1
88	null	3	88	3	2	2	2	2	2	2	2	2
15	88	4	88	3	1	7	2	1	2	1	2	1
30	30	5	30	1	1	2	2	2	2	1	2	1
20	30	5	88	3	2	2	2	2	2	2	2	1
88	null	2	88	1	2	2	2	2	2	2	2	1
88	88	2	3	1	1	2	2	2	2	2	2	2
8	8	5	88	1	1	7	2	2	2	2	2	1
77	77	5	88	3	null	2	2	2	2	2	2	1
2	2	2	88	1	1	2	2	2	2	2	2	2
++			+	+			++		+ <del>-</del>		+	+-

	BEHAVIORAL RISK FACTOR SURVEILLANCE S CODEBOOK REPORT, 2015 Land-Line and Cell-Phone data	SYSTEM		
ıld Not See I	Octor Because of Cost			
Section:	3.3 Health Care Access		Type:	Num
Column:	99	SAS V	riable Name:	MEDCOST
Prologue:				
Description:	Was there a time in the past 12 months when you needed to see a	doctor but could not	because of co	st?
Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	43,514	9.86	13.22
2	No	396,748	89.87	86.46
7	Don't know/Not sure	945	0.21	0.25
9	Refused	248	0.06	0.07
BLANK	Not asked or Missing	1		
•	since last routine checkup			
Section:	3.4 Health Care Access		Type:	Num
Column:	100	SAS V	riable Name:	CHECKUP1
Prologue:				
Description:	About how long has it been since you last visited a doctor for a rout physical exam, not an exam for a specific injury, illness, or condition		outine checkup	is a general
				Weighted
Value	Value Label	Frequency	Percentage	Percentage
	Value Label Within past year (anytime less than 12 months ago)	Frequency 325,432	Percentage 73.72	Percentage 69.13
Value				-
Value	Within past year (anytime less than 12 months ago)	325,432	73.72	69.13
Value	Within past year (anytime less than 12 months ago) Within past 2 years (1 year but less than 2 years ago)	325,432 50,529	73.72 11.45	69.13 13.10
Value 1 2 3	Within past year (anytime less than 12 months ago) Within past 2 years (1 year but less than 2 years ago) Within past 5 years (2 years but less than 5 years ago)	325,432 50,529 28,570	73.72 11.45 6.47	69.13 13.10 8.02
Value 1 2 3 4	Within past year (anytime less than 12 months ago) Within past 2 years (1 year but less than 2 years ago) Within past 5 years (2 years but less than 5 years ago) 5 or more years ago	325,432 50,529 28,570 26,967	73.72 11.45 6.47 6.11	69.13 13.10 8.02 7.28
Value 1 2 3 4 7	Within past year (anytime less than 12 months ago) Within past 2 years (1 year but less than 2 years ago) Within past 5 years (2 years but less than 5 years ago) 5 or more years ago Don't know/Not sure	325,432 50,529 28,570 26,967 5,453	73.72 11.45 6.47 6.11 1.24	69.13 13.10 8.02 7.28 1.18

	BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM CODEBOOK REPORT, 2015 Land-Line and Cell-Phone data	8		
ver Told Blood	I Pressure High			
Section:	4.1 Hypertension Awareness		Type:	Num
Column:	101	SAS Va	riable Name:	BPHIGH4
Prologue:				
	Have you EVER been told by a doctor, nurse or other health professional t "Yes" and respondent is female, ask "Was this only when you were pregna	hat you have int?".)	high blood pre	ssure? (If
Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	178,188	40.36	31.90
2	Yes, but female told only during pregnancy—Go to Section 05.01 $\ensuremath{BLOODCHO}$	3,271	0.74	0.93
3	No-Go to Section 05.01 BLCODCHO	254,318	57.61	66.02
4	Told borderline high or pre-hypertensive—Go to Section 05.01 $_{\rm BLOODCHO}$	4,312	0.98	0.83
7	Don't know/Not Sure-Go to Section 05.01 BLOODCHO	862	0.20	0.19
9	Refused—Go to Section 05.01 BLOODCHO	504	0.11	0.12
BLANK	Not asked or Missing	1		
urrently Takin	g Blood Pressure Medication			
Section:	4.2 Hypertension Awareness		Type:	Num
Column:	102	SAS Va	riable Name:	BPMEDS
Prologue:				
Description:	Are you currently taking medicine for your high blood pressure?			
Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	149,034	83.64	77.20
2	No	28,833	16.18	22.68
7	Don't know/Not Sure	258	0.14	0.10
9	Refused	63	0.04	0.03
BLANK	Not asked or Missing Notes: Section 04.01, BPHIGH4, is coded 2, 3, 4, 7, 9, or Missing	263,268		

	CODEBOOK REPORT, 2015 Land-Line and Cell-Phone data			
uld Not See I	Doctor Because of Cost			
Section:	3.3 Health Care Access		Type:	Num
Column:	99	SAS Va	riable Name:	MEDCOST
Prologue:				
Description:	Was there a time in the past 12 months when you needed to se	e a doctor but could not	because of co	st?
Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	43,514	9.86	13.22
2	No.	396,748	89.87	86.46
7	Don't know/Not sure	945	0.21	0.25
9	Refused	248	0.06	0.07
BLANK	Not asked or Missing	1		
ngth of time : Section:	since last routine checkup  3.4 Health Care Access		Туре:	
Section: Column:	3.4 Health Care Access	SAS Va	Type:	
Section: Column: Prologue:	3.4 Health Care Access	routine checkup? [A ro	riable Name:	CHECKUP1
Section: Column: Prologue:	Health Care Access     Mout how long has it been since you last visited a doctor for a	routine checkup? [A ro	riable Name:	CHECKUP1
Section: Column: Prologue: Description:	3.4 Health Care Access 100  About how long has it been since you last visited a doctor for a physical exam, not an exam for a specific injury, illness, or concentrations.	routine checkup? [A re lition.]	riable Name: outine checkup	CHECKUP1 is a general Weighted
Section: Column: Prologue: Description:	3.4 Health Care Access 100  About how long has it been since you last visited a doctor for a physical exam, not an exam for a specific injury, illness, or con-	routine checkup? [A ra fition.]	riable Name: outine checkup Percentage	CHECKUP1 is a general Weighted Percentage
Section: Column: Prologue: Description:  Value	3.4 Health Care Access 100  About how long has it been since you last visited a doctor for a physical exam, not an exam for a specific injury, limess, or conclude Label  Whith past year (anytime less than 12 months ago)	routine checkup? [A redition.]  Frequency 325,432	putine checkup  Percentage  73.72	CHECKUP1 is a general Weighted Percentage 69.13
Section: Column: Prologue: Description:  Value 1 2	3.4 Health Care Access 100  About how long has it been since you last visited a doctor for a plysical exam, not an exam for a specific injury, illness, or conc Value Label  Nithin past year (asystime less than 12 months ago)  Nithin past years (1 years to Less than 2 years ago)	routine checkup? [A ra fition.]  Frequency 325,432 50,529	Percentage 73.72 11.45	CHECKUP1 is a general Weighted Percentage 69.13 13.10
Section: Column: Prologue: Description:  Value  1 2 3	3.4 Health Care Access 100  About how long has it been since you last visited a doctor for a physical exam, not an exam for a specific injury, illness, or con Value Label  Within past year (anytime less than 12 months ago)  Within past 2 years (1 years but less than 2 years ago)  Within past 2 years (2 years but less than 5 years ago)	routine checkup? [A radition.]  Frequency 325,432 50,529 28,570	Percentage 73.72 11.45 6.47	checkup1 is a general  Weighted Percentage 69.13 13.10 8.02
Section: Column: Prologue: Description:  Value  1 2 3 4	3.4 Health Care Access 100  About how long has it been since you last visited a doctor for a pyrightical exam, not an exam for a specific injury, illness, or conc Value Label  Nithin past year (anytime less than 12 months ago)  Nithin past 2 years (1 year but less than 2 years ago)  Nithin past 5 years (2 years but less than 3 years ago) 5 or more years ago	routine checkup? [A ra fition.]  Frequency 325, 432 50, 529 28, 570 26, 967	Percentage 73.72 11.45 6.47 6.11	CHECKUP1 is a general  Weighted Percentage 69.13 13.10 8.02 7.28
Section: Column: Prologue: Description:  Value  1 2 3 4 7	3.4 Health Care Access 100  About how long has it been since you lest visited a doctor for a physical exam, not an exam for a specific injury, illness, or conc Value Label  Within past year (asyrtime less than 12 months ago) Within past 5 years (1 years but less than 2 years ago) Soft more years ago Don't know/Niko surse	routine checkup? [A rd fition.]  Frequency 325,432 50,529 28,570 26,967 5,453	Percentage 73.72 11.45 6.47 6.11 1.24	CHECKUP1  is a general  Weighted Percentage 69.13 13.10 8.02 7.28 1.18

		BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM CODEBOOK REPORT, 2015 Land-Line and Cell-Phone data			
Ever Told Bloom	d Pres	sure High			
Section:	4.1	Hypertension Awareness		Type:	Num
Column:	101		SAS Va	riable Name:	BPHIGH4
Prologue:					
Description:		you EVER been told by a doctor, nurse or other health professional and respondent is female, ask "Was this only when you were pregni		high blood pre	ssure? (If
Value	Value	Label	Frequency	Percentage	Weighted
1	Yes		178,188	40.36	31.90
2	Yes, BLOOD	but female told only during pregnancy—Go to Section 05.01 CHO	3,271	0.74	0.93
3	No-G	o to Section 05.01 BLOODCHO	254,318	57.61	66.02
4	Told BLOOD	borderline high or pre-hypertensive—Go to Section 05.01 CHO	4,312	0.98	0.83
7	Don't	know/Not Sure-Go to Section 05.01 BLOODCHO	862	0.20	0.19
9	Refus	ed-Go to Section 05.01 BLOODCHO	504	0.11	0.12
BLANK	Not a	sked or Missing	1		
Currently Takin	g Bloc	d Pressure Medication			
Section:	4.2	Hypertension Awareness		Type:	Num
Column:	102		SAS Va	riable Name:	BPMEDS
Prologue:					
Description:	Are y	ou currently taking medicine for your high blood pressure?			
Value	Value	Label	Frequency	Percentage	Weighted
1	Yes		149.034	83.64	77.20
2	No		28,833	16.18	22.68
7	Don't	know/Not Sure	258	0.14	0.10
9	Refus	ed	63	0.04	0.03
BLANK		sked or Missing : Section 04.01, BPHIGH4, is coded 2, 3, 4, 7, 9, or Missing	263,268		



	BEHAVIORAL RISK FACTOR SURVEILLIANCE SYST CODEBOOK REPORT, 2015 Land-Line and Cel-Phone data	ЕМ				BEHAVIORAL RISK FACTOR SURVEILLANCE S' COOEBOOK REPORT, 2015 Land-Une and Cal-Phone data	'STEM		
Could Not See	Doctor Because of Cost				Ever Told Bloo	d Pressure High			
Section:	3.3 Health Care Access		Type:	Num	Section:	4.1 Hypertension Awareness		Type:	Num
Column:	99	SAS V	riable Name:	MEDCOST	Column:	101	SAS Variable Name: BPH		
Prologue:					Prologue:				
Description:	Was there a time in the past 12 months when you needed to see a doct	or but could not	because of co	weighted	Description:	Have you EVER been told by a doctor, nurse or other health profess "Yes" and respondent is female, ask "Was this only when you were p		high blood pre	issure? (If
Value	Value Label	Frequency	Percentage	Percentage					Weighted
1	Yes	43,514	9.86	13.22	Value	Value Label	Frequency	Percentage	Percentage
2	No	396,748	89.87	86.46	1	Yes	178,188	40.36	31.90
7	Don't know/Not sure	945	0.21	0.25	2	Yes, but female told only during pregnancy—Go to Section 05 BLOODCHO	.01 3,271	0.74	0.93
9	Refused	248	0.06	0.07	3	No-Go to Section 05.01 BLOODCHO	254,318	57.61	66.02
BLANK	Not asked or Missing	1			4	Told borderline high or pre-hypertensive—Go to Section 05.0 BLOODCHO	1 4,312	0.98	0.83
enath of time	since last routine checkup				7	Don't know/Not Sure-Go to Section 05.01 BLOODCHO	862	0.20	0.19
Section:	•		Type:	Num	9	Refused-Go to Section 05.01 BLOODCHO	504	0.11	0.12
Column:		CACV	riable Name:		BLANK	Not asked or Missing	1		
	100	SAS V	mable Name:	CHECKUPI					
Prologue:					Currently Takin	ng Blood Pressure Medication			
Description:	About how long has it been since you last visited a doctor for a routine of physical exam, not an exam for a specific injury, illness, or condition.]	heckup? [A n	outine checkup	o is a general	Section:			Type:	
Value	Value Label	Frequency	Percentage	Weighted Percentage			SAS VI	iriable Name:	BPIMEUS
1	Within past year (anytime less than 12 months ago)	325, 432	73.72	69.13	Prologue:				
2	Within past 2 years (1 year but less than 2 years ago)	50,529	11.45	13.10	Description:	Are you currently taking medicine for your high blood pressure?			
1	Within past 5 years (2 years but less than 5 years ago)	28,570	6.47	8.02	Value	Value Label	Frequency	Percentage	Weighted Percentage
4	5 or more years ago	26,967	6.11	7.28	value	Yes	149,034	83.64	77.20
7	Don't know/Not sure	5,453	1,24	1.18	2	No.	28,833	16.18	22.68
8	Never	3,843	0.87	1.14	2	Don't know/Not Sure	258	0.14	0.10
9	Refused	661	0.15	0.15		Refused	63	0.04	0.03
BLANK	Not asked or Missing	1			BLANK	Not asked or Missing Notes: Section 04.01, BPHIGH4, is coded 2, 3, 4, 7, 9, or Mis	263,268 sing		
					l <del></del>		The second		

	BEHAVIORAL RISK FACTOR SURVEILLANCE SY CODEBOOK REPORT, 2015 Land-Line and Cell-Phone data	YSTEM		
ld Not See	Doctor Because of Cost			
Section:	3.3 Health Care Access		Type:	Num
Column:	99	SAS Va	riable Name:	MEDCOST
Prologue:				
escription:	Was there a time in the past 12 months when you needed to see a d	octor but could not	because of co	st?
	-27272			Weighted
Value	Value Label	Frequency	Percentage	Percentage
1	Yes	43,514	9.86	13.22
2	No	396,748	89.87	86.46
7	Don't know/Not sure	945	0.21	0.25
9 BLANK	Refused Not asked or Missing	248	0.06	0.07
gth of time	since last routine checkup			
gth of time : Section: Column:	3.4 Health Care Access	SAS Va	Type:	
Section: Column: Prologue:	3.4 Health Care Access		riable Name:	CHECKUP1
Section: Column: Prologue:	3.4 Health Care Access	ne checkup? [A re	riable Name:	CHECKUP1
Section: Column: Prologue:	Health Care Access     Hour Access     About how long has it been since you last visited a doctor for a routing the second	ne checkup? [A re	riable Name:	CHECKUP1
Section: Column: Prologue: escription:	3.4 Health Care Access 100  About how long has it been since you last visited a doctor for a routin physical exam, not an exam for a specific injury, illness, or condition.	ne checkup? [A ro	priable Name:	CHECKUP1  is a general  Weighted
Section: Column: Prologue: escription:	3.4 Health Care Access 100  About how long has it been since you last visited a doctor for a routin physical exam, not an exam for a specific injury, illness, or condition. Value Label	ne checkup? [A ro	putine checkup	CHECKUP1  is a general  Weighted Percentage
Section: Column: Prologue: escription: Value	3.4 Health Care Access 100  About how long has it been since you last visited a doctor for a routin physical examt, not an exam for a specific injury, liness, or condition.  Value Label  Within past year (anytime less than 12 months age)	re checkup? [A re] Frequency	Percentage	CHECKUP1  D is a general  Weighted Percentage  69.13
Section: Column: Prologue: escription: Value	3.4 Health Care Access 100  About how long has it been since you last visited a doctor for a routily physical exam, not an exam for a specific injury, illness, or condition.  Value Label  Within past year (anytime less than 12 months app)  Within past years (i year but less than 2 years app)	Frequency 325,432 50,529	Percentage 73.72 11.45	CHECKUP1  D is a general  Weighted Percentage 69.13 13.10
Section: Column: Prologue: escription:  Value 1 2 3	3.4 Health Care Access 100  About how long has it been since you last visited a doctor for a routin physical exam, not an exam for a specific injury, literes, or condition.  Value Label  Within past year (anytime less than 12 months app)  Within past 2 years (1) years but less than 2 years app)  Within past 2 years (2) years but less than 5 years app)	Frequency 325,432 50,529 28,570	Percentage 73.72 11.45 6.47	CHECKUP1  Dis a general  Weighted Percentage 69.13 13.10 8.02
Section: Column: Prologue: escription:  Value 1 2 3 4	3.4 Meath Care Access 100  About how long has it been since you last visited a doctor for a routine hypical exam, not an exam for a specific injury, illness, or condition.  Value Label  Within past year (anytime I see than 12 months app)  Within past 2 years (1 years but less than 2 years app)  Within past 5 years (2 years but less than 5 years app)  So core years area.	Frequency 325, 432 50, 529 28, 570 26, 967	Percentage 73.72 11.45 6.47 6.11	CHECKUP1  Dis a general  Weighted Percentage 69.13 13.10 8.02 7.28
Section: Column: Prologue: escription:  Value  1 2 3 4 7	3.4 Health Care Access 100  About how long has it been since you lest visited a doctor for a routil physical exam, not an exam for a specific injury, illness, or condition.  Value Label  Hithin past year (asytime less than 12 months app)  Within past 5 years (1 years but less than 2 years app)  Within past 5 years (2 years but less than 3 years app)  S or note years app  Don't know/host sure	Frequency 325,432 50,529 28,570 26,967 5,453	Percentage 73.72 11.45 6.47 6.11 1.24	CHECKUP1  Dis a general  Weighted Percentage 69.13 13.10 8.02 7.28 1.18

	BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM CODEBOOK REPORT, 2015 Land-Line and Cell-Phone data	4		
Ever Told Blood	Pressure High			
Section:	4.1 Hypertension Awareness		Type:	Num
Column:	101	SAS Va	riable Name:	BPHIGH4
Prologue:				
Description:	Have you EVER been told by a doctor, nurse or other health professional "Yes" and respondent is female, ask "Was this only when you were pregn		high blood pre	essure? (If
Value	Value Label	Frequency	Percentage	Weighte
1	Yes	178,188	40.36	31.90
2	Yes, but female told only during pregnancy—Go to Section 05.01 $_{\rm BLOODCHO}$	3,271	0.74	0.93
3	No-Go to Section 05.01 BLOODCHO	254,318	57.61	66.02
4	Told borderline high or pre-hypertensive—Go to Section 05.01 $_{\mbox{\footnotesize{BLOODCHO}}}$	4,312	0.98	0.83
7	Don't know/Not Sure-Go to Section 05.01 BL00DCH0	862	0.20	0.19
9	Refused-Go to Section 05.01 BLOODCHO	504	0.11	0.12
BLANK	Not asked or Missing	1		
Currently Takin Section:	g Blood Pressure Medication 4.2 Hypertension Awareness		Type:	Norm
	.,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,			
Column:	102	SAS Va	riable Name:	BPMEDS
Prologue:				
Description:	Are you currently taking medicine for your high blood pressure?			
Value	Value Label	Frequency	Percentage	Weighte Percenta
1	Yes	149.034	83.64	77.20

# Analysis Data Clean Up ...

#### Reviewed DataSet and Classified Relevant Columns

А	В	С	D	E	
ColumnCod	Topic	Question		Relevant -Y	
DISPCODE	Disposition Code		1100 completed	X	only completed Ir
GENHLTH	Health Status/Healthy Days	Would you say that in general your health is	1=Excellent 2=Very good 3=Good 4=Fair 5=Poor 7=DK/NS 9=Refused	ML?	Remove 7,9
PHYSHLTH	Health Status/Healthy Days	how many days during the past 30 days was your physical health not good? (Moved to Healthy Days in 2004)	=Number of days 88=None 77=DK/NS 99=Refused	ML	Remove 88, 77,99
MENTHLTH	Health Status/Healthy Days	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? (Moved to Healthy Days in 2004)	=Number of days 88=None 77=DK/NS 99=Refused	Target	132,972 3 btwn 1 but we could also threshold, or con answers see yello
POORHLTH	Health Status/Healthy Days	you from doing your usual activities, such as self-care, work, or recreation? (Moved to Healthy Days in 2004)	=Number of days 88=None 77=DK/NS 99=Refused	ML	
BPHIGH4	Hypertension (Awareness)	high blood pressure? (If "Yes" and respondent is female, ask "Was this only when you were pregnant?").	1=Yes 2=Yes, but female told only during pregnancy 3=No 4=Told borderline high or pre-hypertensive 7=DK/NS 9=Refused	ML	Combine Respons
TOLDHI2	Cholesterol (Awareness)	Have you ever been told by a doctor, nurse or other health professional that your blood cholesterol is high? (Change in variable name.)	1=Yes 2=No 7=DK/NS 9=Refused	ML	Remove 7,9
		[Prologue: Now I would like to ask you some questions about cardiovascular disease]. Has a doctor, nurse, or other health professional EVER told you that you had any of the following? For each. tell me "Yes". "No". or vou're "Not sure". (Ever told) you had a			

## Cleaned Data, removed NA, renamed Columns for Machine Learning Table

ENHLTH   ME	NTHLTH	PHYSHLTH	SEX	MARITAL_STATUS	EDUCATION	HOME_STATUS	VETERAN	EMPLOYMENT_SITUATION	INTERNET DI
+ ۱ د	+   88		tt   3		+ <del>-</del>		tt	٠	+   1
3		88	4	2	0	1	4	3	1
4	88	88	4	3	3	1		2	4
2	3	88	2	3	5	1	2	7	2
2	88	2	1	1	6	1	2	7	1
3	88	88	2	3	4	1	2	5	1
3	88	14	2	3	3	1	2	7	2
3	88	88	1	1	6	1	1	7	1
5	99	99	1	2	4	2	1	99	2
4	30	28	2	1	6	1	2	8	1
2	5	88	2	1	6	1	2	1	1
+	+		++		+	<del></del>	++		+



#### Optimized List Creation For Transformations

SU	M 🛊	$\times$ $\checkmark$ $f_x$	= 19&"'"&F	<mark>20</mark> &"'"&","					
	E	F	Н	1	J	К	L	М	N
17	list string	Renamed		List String					
18		STATE							
19	'FMONTH',	FMONTH		'FMONTH',					
20	'FMONTH','I	DATE		F20&"'"&","					
21	'FMONTH','I	MONTH		'FMONTH','	ATE','MONT	Ή',			
22	'FMONTH','I	DAY		'FMONTH','E	DATE','MONT	H','DAY',			
23	'FMONTH','I	YEAR		'FMONTH','E	DATE','MONT	H','DAY','YE	AR',		
24	'FMONTH','I	DISPCODE		'FMONTH','	DATE','MONT	H','DAY','YE	AR','DISPCOD	Ε',	
25	'FMONTH','I	PHYSHLTH		'FMONTH','	DATE','MONT	H','DAY','YE	AR','DISPCOD	E','PHYSHLTH',	
26	'FMONTH','I	POORHLTH		'FMONTH','	DATE','MONT	H','DAY','YE	AR','DISPCOD	E','PHYSHLTH',	'POORI
27	'FMONTH','I	GENHLTH		'FMONTH','	DATE','MONT	H','DAY','YEA	AR','DISPCOD	E','PHYSHLTH',	'POORI

### Transform to categorical Values for DashBoard Table

+		+	t			+	+		
EDUCATI	AGE_2LEVEL	AGE_14LEVEL	METRO_CODE	RACE	MARITAL_STATUS	PHYSHLTH	SEX	MENTHLTH	GENHLTH
+   Graduated	+   18-64	+   50-54	+   Not in MSA	White	Divorced	+   88	+  Female	 88	+   Good
Did not Grad	65-Older	70-74	Not in MSA	White	Widowed	88	Female	88	Very Good
Attended	65-Older	70-74	Center of MSA	White	Widowed	88	Female	3	Very Good
Graduated	65-Older	65-69	Center of MSA	White	Married	2	Male	88	Very Good
Graduated Hi	65-Older	70-74	Inside a Suburban	White	Widowed	88	Female	88	Good
Did not Grad	65-01der	80-Older	Inside a Suburban	White	Widowed	14	Female	88	Good
Graduated	65-01der	80-Older	Center of MSA	White	Married	88	Male	88	Good
Graduated Hi	18-64	Dont know	Center of MSA	White	Divorced	99	Male	99	Poor
Graduated	18-64	35-39	Inside a Suburban	White	Married	28	Female	30	Fair
Graduated	18-64	45-49	Not in MSA	White	Married	88	Female	5	Very Good
Graduated Hi	65-01der	65-69	Not in MSA	White	Married	88	Female	88	Good
Attended	18-64	50-54	Center of MSA	White	Married	88	Male	88	Very Good
Did not Grad	65-01der	80-Older	Center of MSA	White	Widowed	88	Male	99	/ery Good
Did not Grad	65-01der	80-Older	Not in MSA	White	Widowed	30	Female	88	Fair
Attended	65-01der	80-Older	Not in MSA	African American	Widowed	88	Female	88	Excellent
Graduated	18-64	40-44	Center of MSA	White	Married	88	Female	88	Very Good
Graduated	65-01der	75-79	Outside Center MSA	White	Married	88	Male	88	Excellent
Attended	18-64	Dont know	Outside Center MSA	African American	Never Married	7	Female	88	Fair
Did not Grad	65-01der	80-Older	Inside a Suburban	White	Widowed	88	Male	88	Excellent
Graduated Hi	18-64	45-49	Not in MSA	White	Married	88	Female	2	Excellent

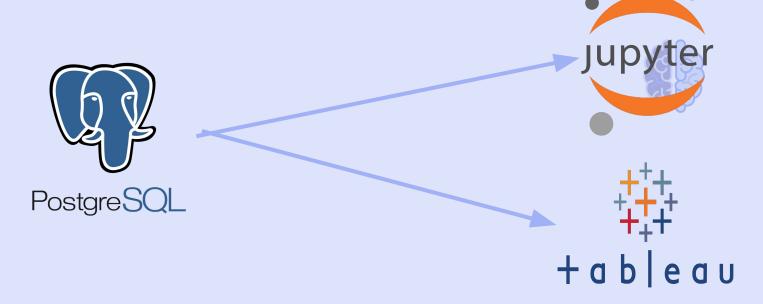
## Load and Connect

#### Load Tables to SQL RDS Server

#### Connect

```
import psycopg2
import pandas as pd
conn = psycopg2.connect(database = 'postgres',
                                    'postgres',
                        password = 'password',
                                    'bootcampproject.cs8v5ggqsbn0.us-west-1.rds.amazonaws.com',
                        host =
                                   '5432')
                        port =
cur = conn.cursor()
cur.execute('''
            SELECT *
            FROM ml table;
data = cur.fetchall()
# Create a dataframe
for elt in cur.description:
    cols.append(elt[0])
health df = pd.DataFrame (data=data,columns=cols)
pd.set_option('display.max_columns', None)
health df.head(10)
```





# Machine Learning Feature Engineering



Changed non-answer codes
Removed outliers
Binned target variable and features
with over 12 unique values





Section: 1.1 Health Status Type: Num

Column: 90 SAS Variable Name: GENHLTH

Prologue:

**Description:** Would you say that in general your health is:

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Excellent	76,032	17.22	18.68
2	Very good	145,065	32.86	31.78
3	Good	136,975	31.03	31.59
4	Fair	58,962	13.36	13.06
5	Poor	23,175	5.25	4.60
7	Don't know/Not Sure	799	0.18	0.18
9	Refused	446	0.10	0.11
BLANK	Not asked or Missing	2		

# GENHLTH: 7/9 (don't know/refused/missing) changed to the rounded median
GENHLTH\_median = round(health\_df[(health\_df.GENHLTH != 9) & (health\_df.GENHLTH != 7)].GENHLTH.median())
health\_df['GENHLTH'] = health\_df['GENHLTH'].replace({7:GENHLTH\_median, 9:GENHLTH\_median})

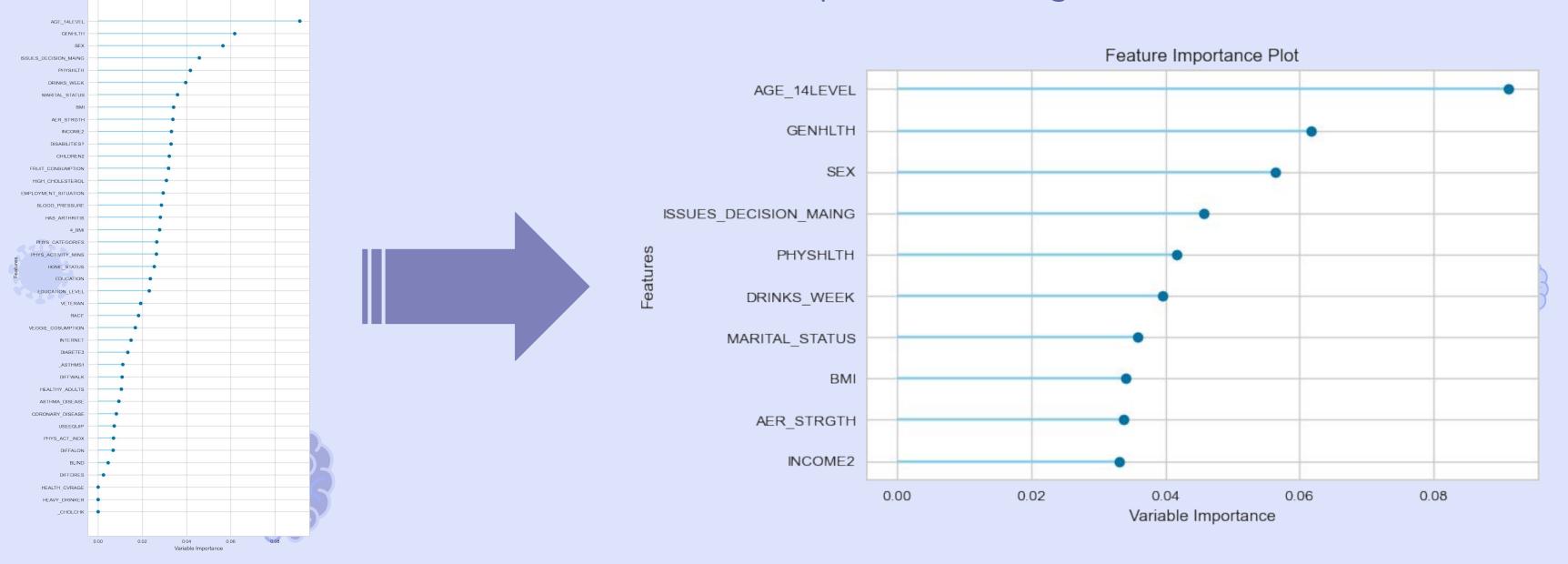


# Machine Learning Feature Selection

Created feature importance chart

Set a cutoff of 0.02 variable feature importance due to drop-off

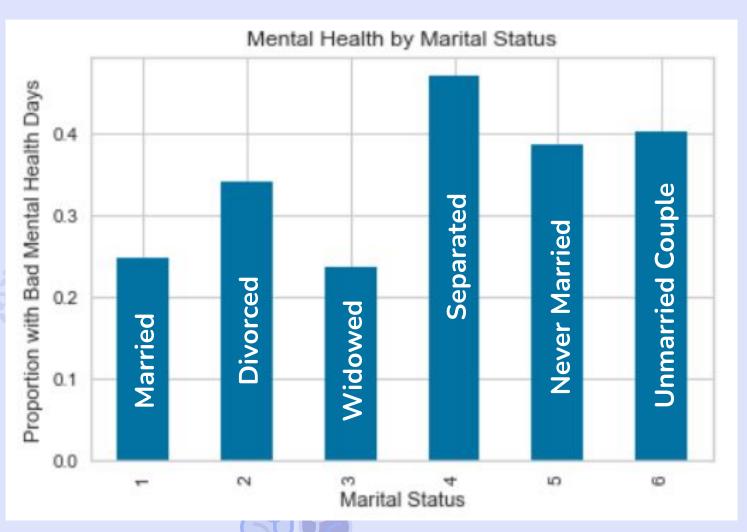
23/41 features were kept for modeling

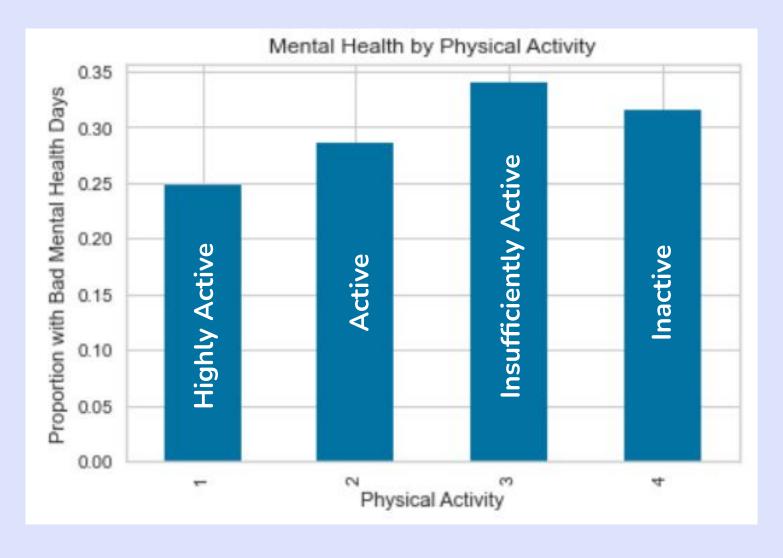


# Machine Learning Feature Selection



Interesting Correlations Discovered through Feature Importance Plot:







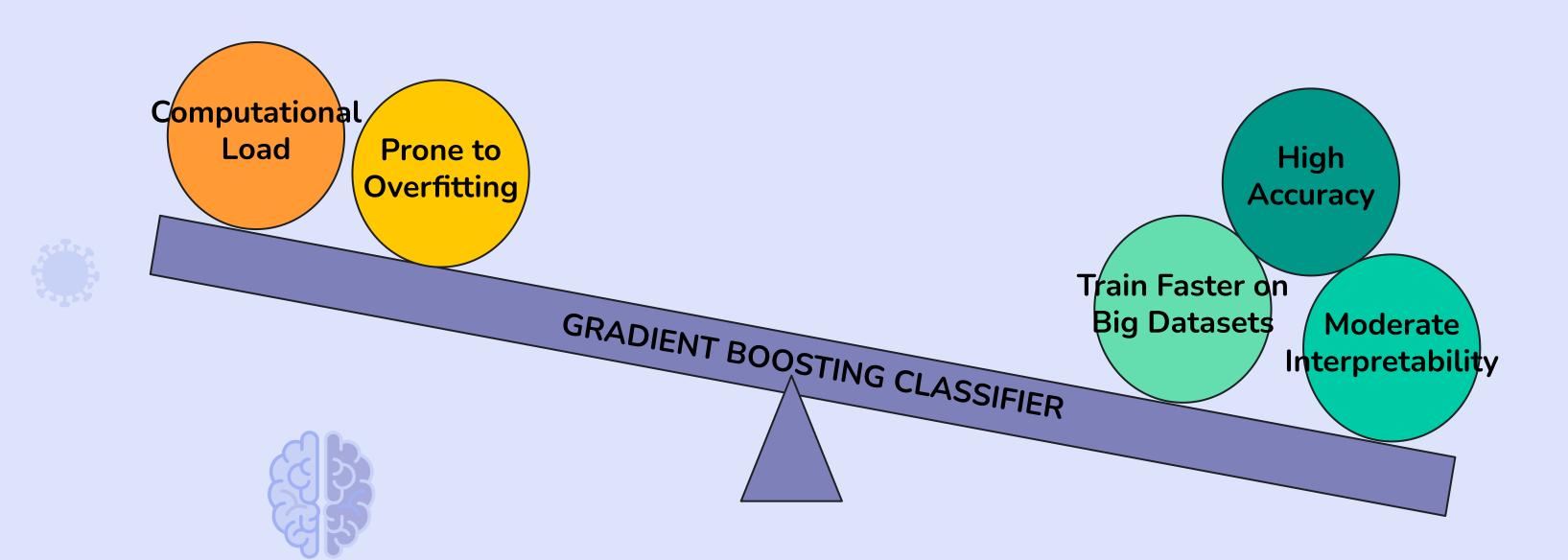


# Machine Learning Model Choice

Used PyCaret's function compare\_models(), which runs 12 ML models

A gradient boosting classifier model was chosen to balance accuracy and interpretability

Accuracy: 75.46%



# Machine Learning Optimization

SMOTE oversampling to balance the uneven target variable (0: 70%, 1: 30%) Used PyCaret's tune\_model() function to automatically tune hyperparameters

# Automatically tune model parameters using PyCaret (runtime ~50min)
tuned\_gbc = tune\_model(gbc)

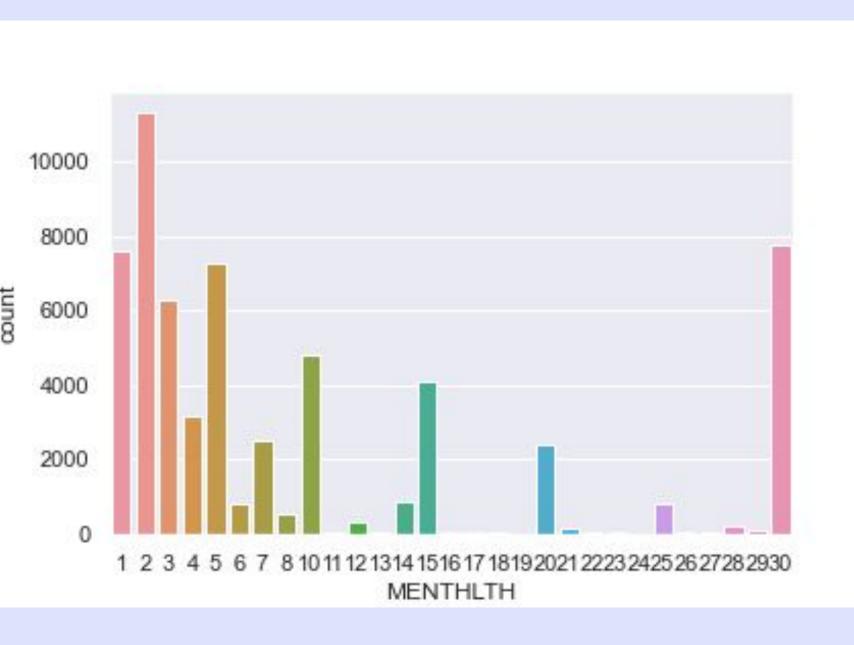
		AUG	Decall	Duna	F4	V	MCC	
	Accuracy	AUC	Kecali	Prec.	-	карра	WICC	
Fold								
0	0.7569	0.7353	0.3153	0.6370	0.4218	0.2895	0.3182	
1	0.7532	0.7258	0.3031	0.6269	0.4087	0.2759	0.3050	
2	0.7527	0.7282	0.3060	0.6232	0.4104	0.2764	0.3044	
3	0.7547	0.7333	0.3023	0.6346	0.4096	0.2786	0.3091	
4	0.7574	0.7349	0.3106	0.6425	0.4188	0.2883	0.3187	
5	0.7526	0.7374	0.3050	0.6231	0.4095	0.2756	0.3037	
6	0.7562	0.7352	0.2997	0.6432	0.4088	0.2801	0.3125	
7	0.7548	0.7394	0.3094	0.6311	0.4152	0.2824	0.3111	
8	0.7534	0.7315	0.3017	0.6283	0.4076	0.2754	0.3050	
9	0.7537	0.7235	0.3100	0.6256	0.4145	0.2804	0.3081	
Mean	0.7546	0.7324	0.3063	0.6315	0.4125	0.2802	0.3096	
Std	0.0017	0.0049	0.0046	0.0071	0.0046	0.0049	0.0052	
Processing: 0%   0/7 [00:00 , ?it/s] Fitting 10 folds for each of 10 candidates, totalling 100 fits</td								

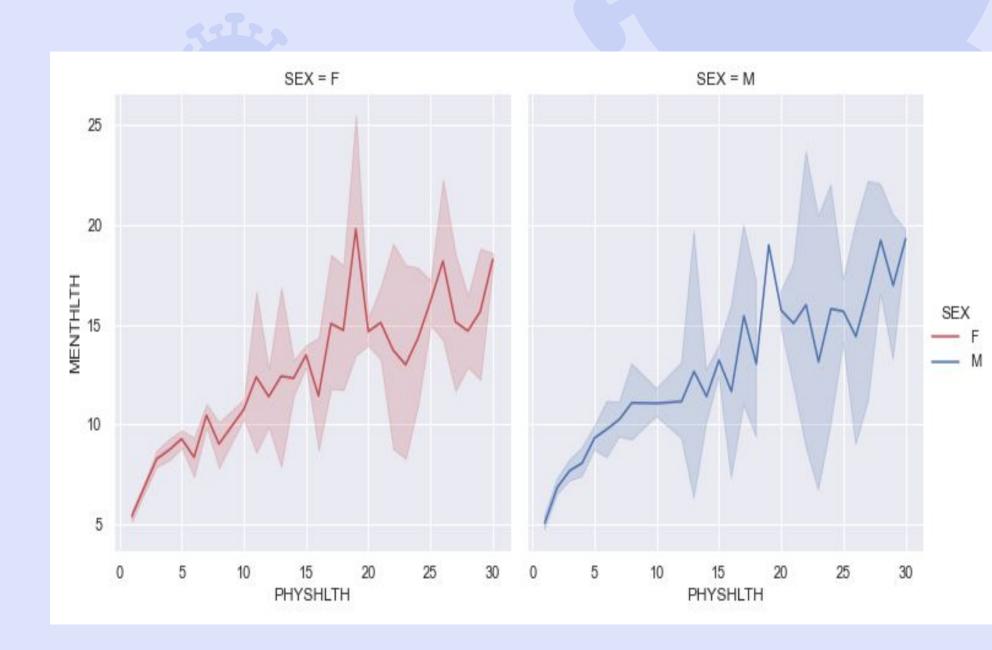






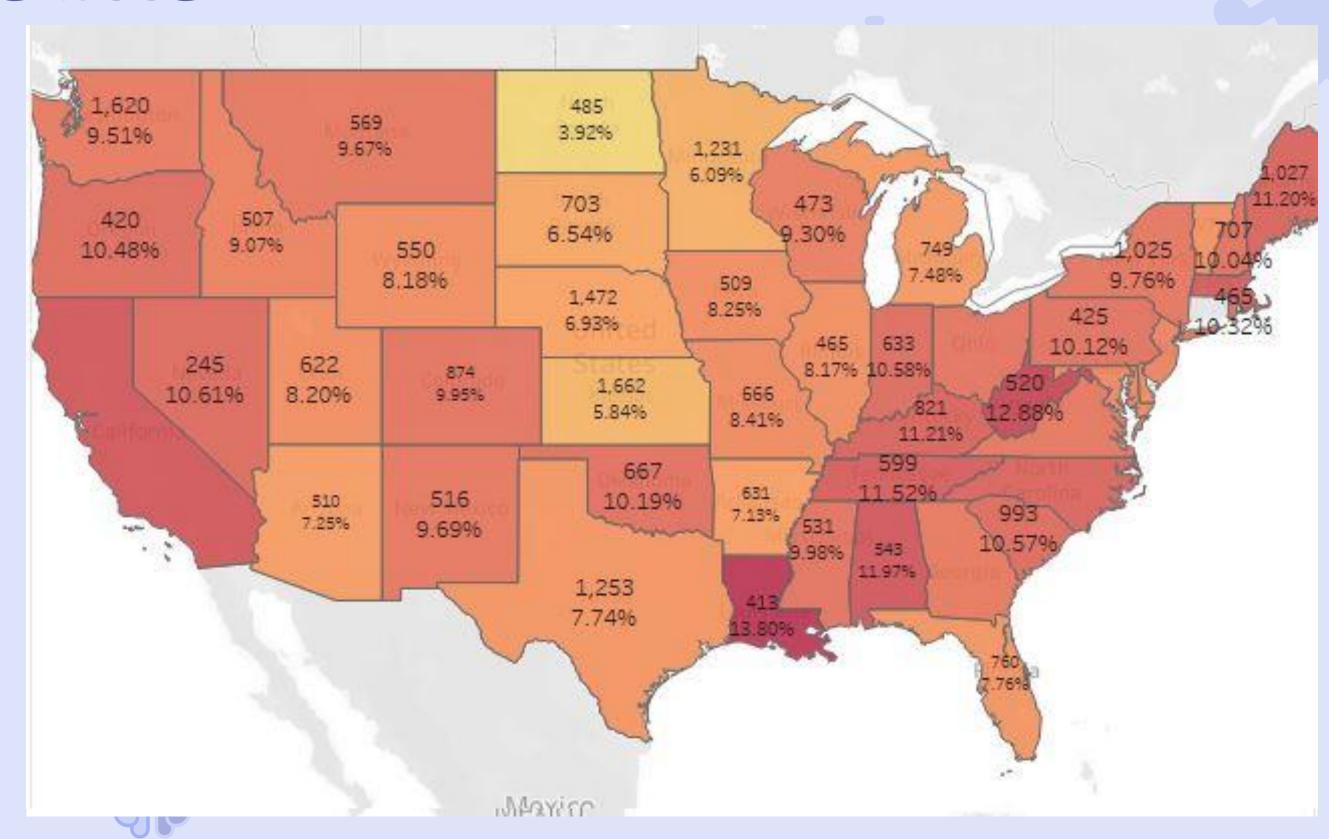
## Results

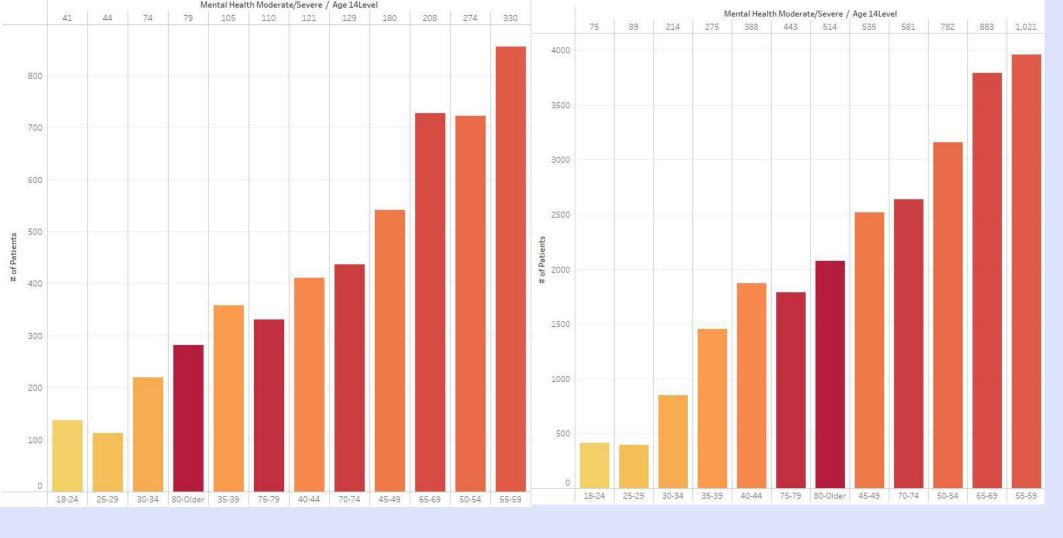






## Results





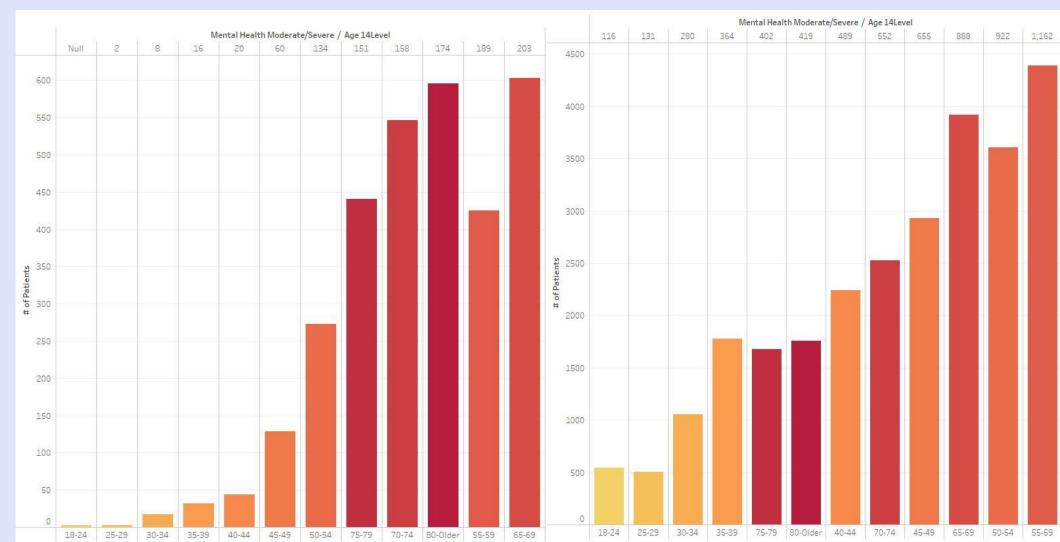


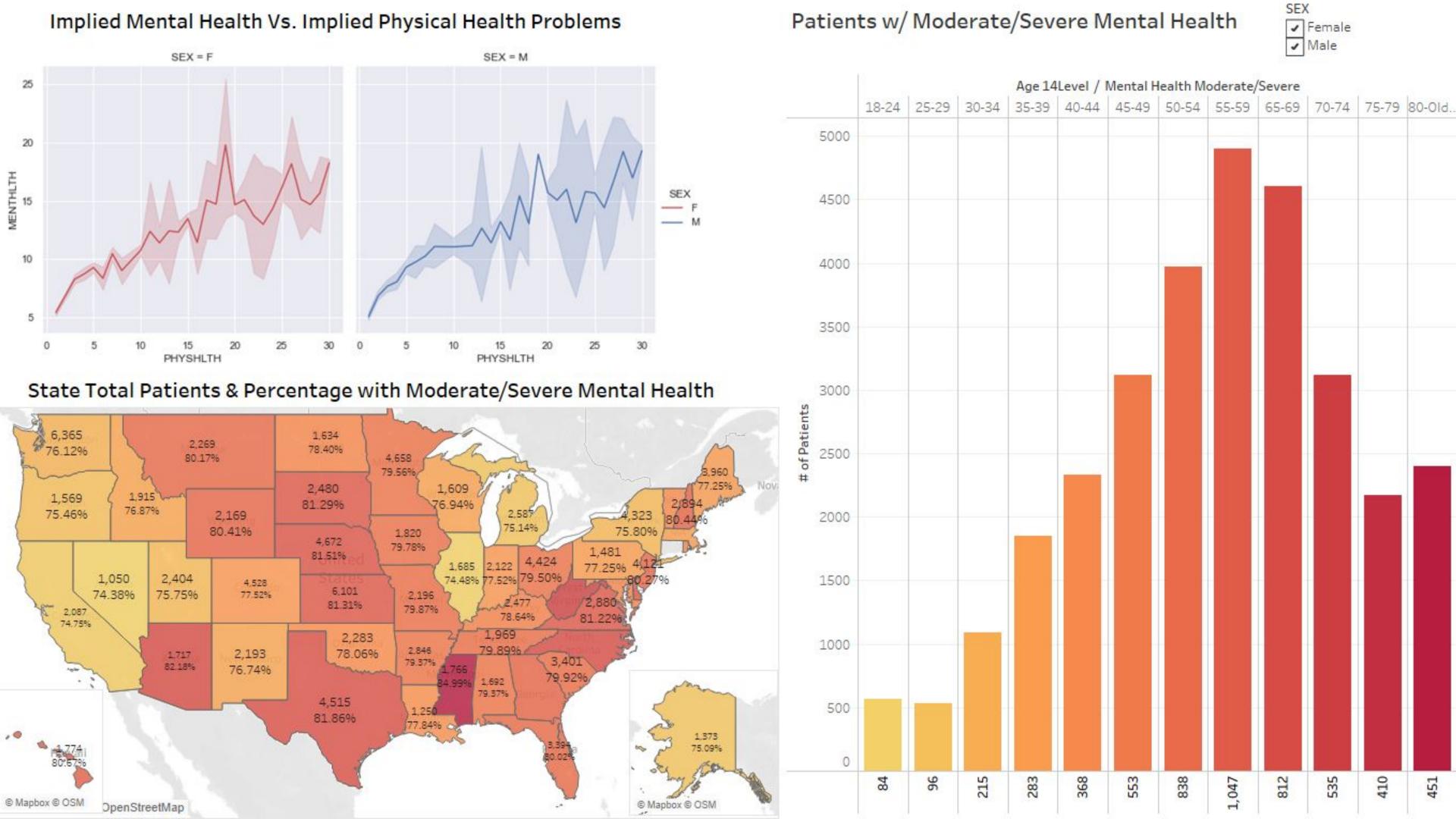


### **Coronary Disease**

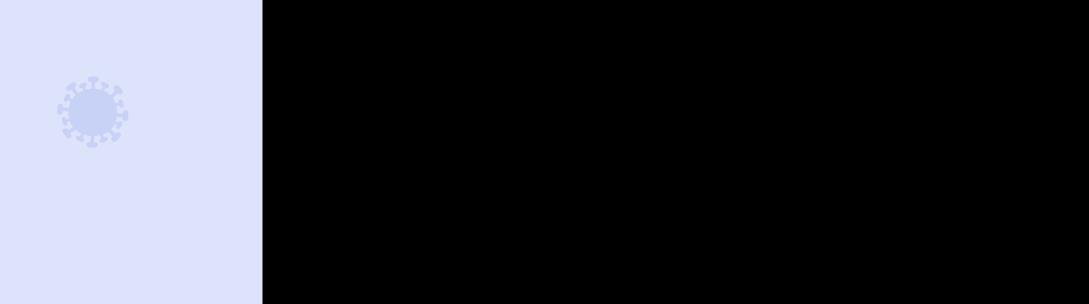
5x more likely







# Interactivity







## Future Analysis Recommendations



#### **Additional Datasets**

Our data was limited to 2015. For future analysis loading additional years would allow for more trend analysis and high relevant-low data columns to play a bigger role.

### Clear & Complete Data

Some fields that were deemed of higher importance had odd calculations leading to unclear results with visualization. Refusal to answer certain important questions caused problems in visualizations.



