

# Correlation Between Physical & Mental Health

Aaron Miller, Adriana Garcia, Kylie Hefner, & Lukas Van de Velde



#### Table of Contents

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





### About Our Project





#### Our Topic













Exploring correlations between physical health indicators and mental health outcomes.



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







### About Our Project









Health-related telephone survey in the US.

>400k interviews each year – largest continuously conducted health survey in the world.

(dataset from the CDC via Kaggle)







### 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:**

++	+	+		<del>-</del>			<del>-</del>		<del>-</del>		+·	++-
PHYSHLTH	POORHLTH	GENHLTH	MENTHLTH	BPHIGH4	TOLDHI2	CVDINFR4	CVDCRHD4	CVDSTRK3	CHCSCNCR	CHCOCNCR	CHCCOPD1	HAVARTH3 C
++	+	+		·+	+		·		++		٠·	++-
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>					<del>-</del>			+

Section: 3.3 Health Care Access	YSTEM		
		Type:	Num
Column: 99	SAS Va	riable Name:	MEDCOST
Prologue:			
Description: Was there a time in the past 12 months when you needed to see a d	loctor but could not	because of co	st?
The state of the s	bould not		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 Refused	248	0.06	0.07
BLANK Not asked or Missing	1		
ngth of time since last routine checkup			
Section: 3.4 Health Care Access		Type:	Num
Column: 100	SAS Va	riable Name:	CHECKUP1
Proloque:			
Description: 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	outine checkup	
Value Value Label	Frequency	Percentage	Weighted Percentage
1 Within past year (anytime less than 12 months ago)	325,432	73.72	69.13
Within past 2 years (1 year but less than 2 years ago)	50,529	11.45	13.10
3 Within past 5 years (2 years but less than 5 years ago)	28,570	6.47	8.02
4 5 or more years ago	26,967	6.11	7.28
	5,453	1.24	1.18
7 Don't know/Not sure	5,453		
	3,843	0.87	1.14

	REHAVIORAL RISK FACTOR SURVEILLANCE S			
	BEHAVIORAL RISK FACTOR SURVEILLANCE S' CODEBOOK REPORT, 2015 Land-Line and Cell-Phone data	YSTEM		
er 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 professi "Yes" and respondent is female, ask "Was this only when you were p		high blood pre	
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 BLOODCHO	.01 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 05.0 $_{\rm BLOODCHO}$	1 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		
	g Blood Pressure Medication			
Section:			Type:	
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

ould Not See	Doctor Be	cause of Cost			
Section:	3.3 He	ealth Care Access		Type:	Num
Column:	99		SAS Va	riable Name:	MEDCOST
Prologue:					
Description:	Was then	e a time in the past 12 months when you needed to see a do	octor but could not	because of cos	st?
					Weighted
Value	Value Lab	el .	Frequency	Percentage 9.86	Percentage 13.22
2	Yes		43,514 396,748	9.86	13.22 86.46
7		ow/Not sure	396,748	0.21	0.25
9	Befused	ow/Not sure	248	0.21	0.25
BLANK		d or Missing	248	0.06	0.07
				Type:	
Column:	100		SAS Va	riable Name:	
Column: Prologue:			SAS Va	riable Name:	
Prologue:	About ho	w long has it been since you last visited a doctor for a routin exam, not an exam for a specific injury, illness, or condition.)	e checkup? [A ro		CHECKUP1
Prologue:	About ho	exam, not an exam for a specific injury, illness, or condition.]	e checkup? [A ro		CHECKUP1
Prologue: Description:	About hor physical of Value Laborate	exam, not an exam for a specific injury, illness, or condition.]	e checkup? [A ro	outine checkup	CHECKUP1 is a general Weighted
Prologue: Description: Value	About hor physical of Value Laboration within p	exam, not an exam for a specific injury, illness, or condition.]	e checkup? [A ro	outine checkup	CHECKUP1 is a general Weighted Percentage
Prologue: Description:  Value  1 2 3	About hor physical of Value Labout Within p	exam, not an exam for a specific injury, illness, or condition.]  el  ast year (anytime less than 12 months ago)	e checkup? [A ro 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
Prologue: Description:  Value  1 2 3 4	About hor physical of Value Lab Within p Within p Within p 5 or more	<pre>axxam, not an exam for a specific injury, illness, or condition.)  st st year (anytime less than 12 months ago) ast 2 years (1 year but less than 2 years ago) ast 5 years (2 years but less than 5 years ago) years ago) years ago) years</pre>	e checkup? [A rc]  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
Prologue: Description:  Value  1 2 3 4 7	About hor physical of Value Lab Within p Within p Within p 5 or mor Don't kn	exam, not an exam for a specific injury, illness, or condition.]  st year (anytime less than 12 months ago)  sat 2 years (1 year but less than 2 years ago)  sat 5 years (2 years but less than 5 years ago)	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
Prologue: Description:  Value  1 2 3 4 7 8	About hor physical of Value Lab Within p Within p Within p 5 or mor Don't kn Never	<pre>axxam, not an exam for a specific injury, illness, or condition.)  st st year (anytime less than 12 months ago) ast 2 years (1 year but less than 2 years ago) ast 5 years (2 years but less than 5 years ago) years ago) years ago) years</pre>	Frequency 325,432 50,529 28,570 26,967 5,453 3,843	Percentage 73.72 11.45 6.47 6.11 1.24 0.87	CHECKUP1  is a general  Weighted Percentage 69.13 13.10 8.02 7.28 1.18 1.14
Prologue: Description:  Value  1 2 3 4 7	About hor physical of Value Labout North part of Within	<pre>axxam, not an exam for a specific injury, illness, or condition.)  st st year (anytime less than 12 months ago) ast 2 years (1 year but less than 2 years ago) ast 5 years (2 years but less than 5 years ago) years ago) years ago) years</pre>	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	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 pregna		high blood pre	issure? (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		



	15 01 157			rugues au, au tu			14 01 107			magast Eu, Eu
	BEHAVIORAL RISK FACTOR SURVEILLANCE SYST CODEBOOK REPORT; 2015 Land-Line and Cell-Phone data	ЕМ					BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM CODEBOOK REPORT 2015 Land-Line and Cell-Phone data	(		
Could Not See I	Doctor Because of Cost				Ever Told Bloom	d Pressu	re High			
Section:	3.3 Health Care Access		Type:	Num	Section:	4.1 H	Hypertension Awareness		Type:	Num
Column:	99	SAS Va	riable Name:	MEDCOST	Column:	101		SAS Va	riable Name:	BPHIGH4
Prologue:					Prologue:					
	Was there a time in the past 12 months when you needed to see a doctor	or but could not	because of co	st? Weighted		Have yo	ou EVER been told by a doctor, nurse or other health professional and respondent is female, ask "Was this only when you were pregna	hat you have ant?".)	high blood pres	ssure? (If
Value	Value Label	Frequency	Percentage	Percentage						Weighted
1	Yes	43,514	9.86	13.22	Value	Value La	ibel	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, bu BLOODCH	it female told only during pregnancy—Go to Section 05.01	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 bo	orderline high or pre-hypertensive—Go to Section 05.01	4,312	0.98	0.83
enath of time	since last routine checkup				7	Don't k	now/Not Sure-Go to Section 05.01 BLOODCHO	862	0.20	0.19
	1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		-		9	Refused	-Go to Section 05.01 BLOODCHO	504	0.11	0.12
Section:		N 100-01.	Type:		BLANK	Not ask	ted or Missing	1		
Column:	100	SAS Va	riable Name:	CHECKUP1						
Prologue:					Currently Takin	ng Blood	Pressure Medication			
Description:	About how long has it been since you last visited a doctor for a routine or physical exam, not an exam for a specific injury, illness, or condition.]	heckup? [A re	outine checkup	is a general	Section:		Hypertension Awareness		Type:	
				Weighted	Column:	102		SAS Va	riable Name:	BPMEDS
Value	Value Label	Frequency	Percentage	Percentage	Prologue:					
1	Within past year (anytime less than 12 months ago)	325,432	73.72	69.13	Description:	Are you	currently taking medicine for your high blood pressure?			
2	Within past 2 years (1 year but less than 2 years ago)	50,529	11.45	13.10	200.500					Weighted
3	Within past 5 years (2 years but less than 5 years ago)	28,570	6.47	8.02	Value	Value La	ibel	Frequency	Percentage	Percentage
4	5 or more years ago	26,967	6.11	7.28	1	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	7		now/Not Sure	258	0.14	0.10
9	Refused	661	0.15	0.15	9	Refused		63	0.04	0.03
BLANK	Not asked or Missing	1			BLANK	Not ask Notes:	med or Missing Section 04.01, BPHIGH4, is coded 2, 3, 4, 7, 9, or Missing	263,268		
					l <del></del>		-0 (	b		

	BEHAVIORAL RISK FACTOR SURVEILLANCE SYS COOEBOOK REPORT, 2015 Land-line and Gel-Phone data	тем				BEHAVIORAL RISK FACTOR SI CODEBOOK REPC Land-Line and Cal-
uld Not See [	Octor Because of Cost				Ever Told Bloom	d Pressure High
Section:	3.3 Health Care Access		Type:	Num	Section:	4.1 Hypertension Awareness
Column:	99	SAS Va	riable Name:	MEDCOST	Column:	101
Prologue:					Prologue:	
	Was there a time in the past 12 months when you needed to see a do	ctor but could not	herause of co	st?		Have you EVER been told by a doctor, nurse or other
	The trade of the first the past 12 months when you needed to dee a de	olor but could not	000000000000000000000000000000000000000	Weighted		"Yes" and respondent is female, ask "Was this only wi
Value	Value Label	Frequency	Percentage	Percentage		
1	Yes	43,514	9.86	13.22	Value	Value Label
2	No	396,748	89.87	86.46	1	Yes
7	Don't know/Not sure	945	0.21	0.25	2	Yes, but female told only during pregnancy-Go BLOODCHO
9	Refused	248	0.06	0.07	3	No-Go to Section 05.01 BLOODCHO
BLANK	Not asked or Missing	1			4	Told borderline high or pre-hypertensive—Go to ${\tt BLOODCHO}$
nath of time s	since last routine checkup				7	Don't know/Not Sure-Go to Section 05.01 BLOODC
	25.0 -0.000 -0.000 -0.000		_		9	Refused-Go to Section 05.01 BLOODCHO
Section:			Type:		BLANK	Not asked or Missing
Column:	100	SAS Va	riable Name:	CHECKUP1		
Prologue:					Currently Takin	g Blood Pressure Medication
Description:	About how long has it been since you last visited a doctor for a routine physical exam, not an exam for a specific injury, illness, or condition.]	checkup? [A ro	outine checkup	is a general	Section:	4.2 Hypertension Awareness
				Weighted	Column:	102
Value	Value Label	Frequency	Percentage	Percentage	Prologue:	
1	Within past year (anytime less than 12 months ago)	325,432	73.72	69.13	Description:	Are you currently taking medicine for your high blood a
2	Within past 2 years (1 year but less than 2 years ago)	50,529	11.45	13.10		,,,
3	Within past 5 years (2 years but less than 5 years ago)	28,570	6.47	8.02	Value	Value Label
4	5 or more years ago	26,967	6.11	7.28	1	Yes
7	Don't know/Not sure	5,453	1.24	1.18	2	No
8	Never	3,843	0.87	1.14	7	Don't know/Not Sure
9	Refused	661	0.15	0.15	9	Refused
BLANK	Not asked or Missing	1			BLANK	Not asked or Missing Notes: Section 04.01, BPHIGH4, is coded 2, 3, 4, $\frac{1}{2}$

# Analysis Data Clean Up ...

#### Reviewed DataSet and Classified Relevant Columns

А	В	С	D	E	
ColumnCod	Topic	Question	Score	Relevant -T	Comments
DISPCODE	Disposition Code		1100 completed	x	only completed I
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		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 Respon
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 you're "Not sure". (Ever told) you had a			

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

ENHLTH	MENTHLTH	PHYSHLTH	SEX	MARITAL_STATUS	EDUCATION	HOME_STATUS	VETERAN	EMPLOYMENT_SITUATION	INTERNET	DISAE
+	+		++		+	+	+	<u></u>	+	
3	88	88	2	2	!  6	1	2	3	1	
2	88	88	2	3	3	1	2	2	2	
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	e L
4	30	28	2	1	.  6	1	2	8	1	
2	5	88	2	1	.  6	1	2	1	1	
+			++		+	+	++		+	



#### Optimized List Creation For Transformations

SU	M 🛊	$\mathbf{X}$ $\checkmark$ $f_x$	= 19&""&F	20&"""&","					
1	Е	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','	DATE','MONT	H','DAY',			
23	'FMONTH','I	YEAR		'FMONTH','	DATE','MONT	H','DAY','YEA	AR',		
24	'FMONTH','I	DISPCODE		'FMONTH','	DATE','MONT	H','DAY','YEA	AR','DISPCODI	Ε',	
25	'FMONTH','I	PHYSHLTH		'FMONTH','	DATE','MONT	H','DAY','YEA	AR','DISPCODI	E','PHYSHLTH',	,
26	'FMONTH','I	POORHLTH		'FMONTH','	DATE','MONT	H','DAY','YEA	AR','DISPCODI	E','PHYSHLTH',	'POORI
27	'FMONTH','I	GENHLTH		'FMONTH','	DATE','MONT	'H','DAY','YEA	AR','DISPCODE	E','PHYSHLTH',	'POORI

#### Transform to categorical Values for DashBoard Table

EDUCATI	AGE_2LEVEL	AGE_14LEVEL	O_CODE	METRO	RACE	MARITAL_STATUS	PHYSHLTH	SEX	MENTHLTH	NHLTH	GEI
Graduated	18-64	50-54	in MSA	Not :	White	Divorced	88	Female	88	Good	
Did not Grad	65-Older	70-74	in MSA	Not i	White	Widowed	88	Female	88	Good	Very
Attended	65-01der	70-74	of MSA	Center o	White	Widowed	88	Female	3	Good	Very
Graduated	65-01der	65-69	of MSA	Center o	White	Married	2	Male	88	Good	Very
Graduated Hi	65-01der	70-74	ban	Inside a Suburk	White	Widowed	88	Female	88	Good	
Did not Grad	65-Older	80-Older	ban	Inside a Suburb	White	Widowed	14	Female	88	Good	
Graduated	65-01der	80-Older	of MSA	Center o	White	Married	88	Male	88	Good	
Graduated Hi	18-64	Dont know	of MSA	Center o	White	Divorced	99	Male	99	Poor	
Graduated	18-64	35-39	ban	Inside a Suburb	White	Married	28	Female	30	Fair	
Graduated	18-64	45-49	in MSA	Not i	White	Married	88	Female	5	Good	Very
Graduated Hi	65-01der	65-69	in MSA	Not i	White	Married	88	Female	88	Good	
Attended	18-64	50-54	of MSA	Center o	White	Married	88	Male	88	Good	Very
Did not Grad	65-Older	80-Older	of MSA	Center of	White	Widowed	88	Male	99	Good	Very
Did not Grad	65-01der	80-Older	in MSA	Not i	White	Widowed	30	Female	88	Fair	
Attended	65-Older	80-Older	in MSA	Not i	ican American	Widowed	88	Female	88	llent	Exce:
Graduated	18-64	40-44	of MSA	Center o	White	Married	88	Female	88	Good	Very
Graduated	65-Older	75-79	er MSA	Outside Cente	White	Married	88	Male	88	llent	Exce:
Attended	18-64	Dont know	er MSA	Outside Cente	ican American	Never Married	7	Female	88	Fair	
Did not Grad	65-01der	80-Older	ban	Inside a Suburk	White	Widowed	88	Male	88	llent	Exce:
Graduated Hi	18-64	45-49	in MSA	Not i	White	Married	88	Female	2	llent	Exce

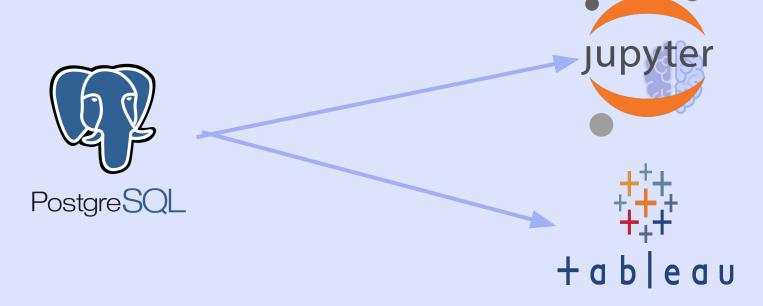
#### Load and Connect

#### 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



#### **General Health**

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})

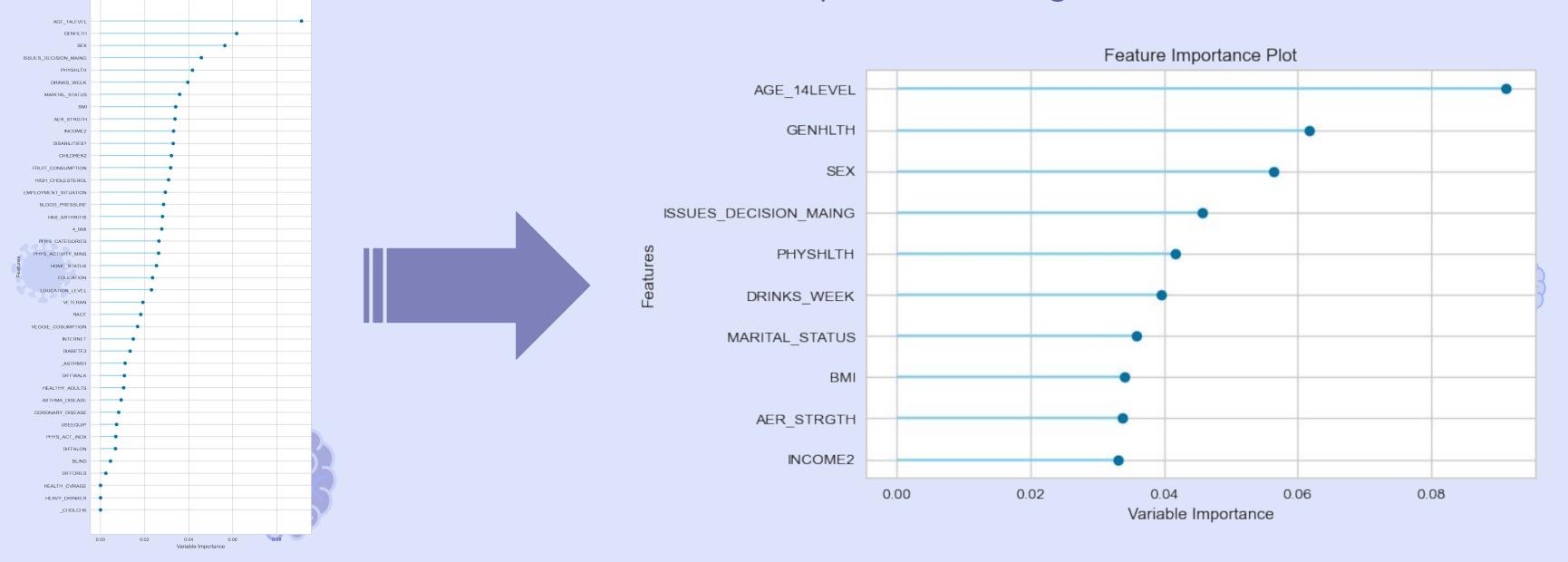


# Machine Learning Feature Selection

Created feature importance chart

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

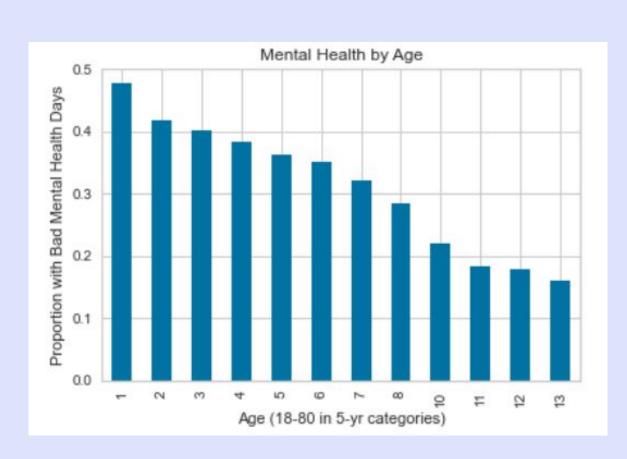
23/41 features were kept for modeling

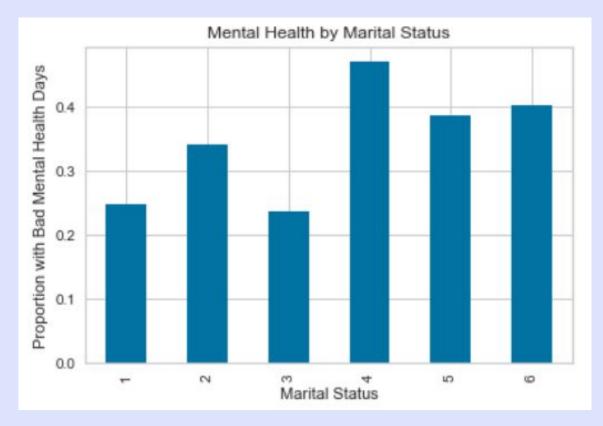


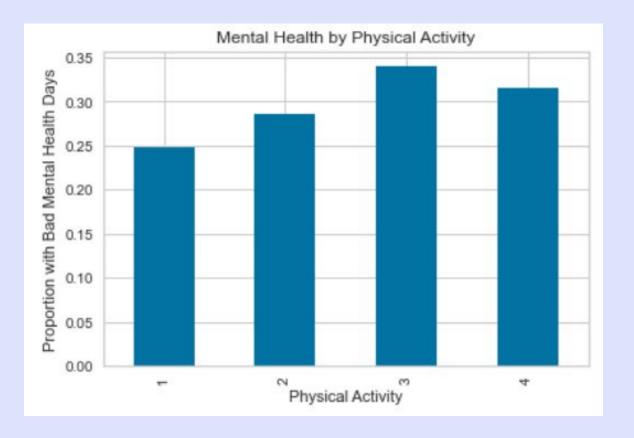
# Machine Learning Feature Selection



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

Initial Accuracy: 74.63% Initial F1: 0.7248

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.7524	0.7409	0.7524	0.7291	0.7211	0.2982	0.3253	4.1350
gbc	Gradient Boosting Classifier	0.7463	0.7365	0.7463	0.7230	0.7248	0.3132	0.3262	77.0060
rf	Random Forest Classifier	0.7332	0.7053	0.7332	0.7024	0.7075	0.2673	0.2817	21.5730
et	Extra Trees Classifier	0.7193	0.6895	0.7193	0.6904	0.6986	0.2491	0.2572	22.5210
ada	Ada Boost Classifier	0.6833	0.6934	0.6833	0.7032	0.6895	0.2873	0.2913	6.5870
dt	Decision Tree Classifier	0.6340	0.5798	0.6340	0.6457	0.6395	0.1500	0.1503	1.4990
qda	Quadratic Discriminant Analysis	0.6247	0.6874	0.6247	0.7065	0.6607	0.2262	0.2336	0.5660
nb	Naive Bayes	0.5885	0.6771	0.5885	0.7019	0.6350	0.1856	0.1976	3.9590
Ir	Logistic Regression	0.5583	0.6979	0.5583	0.7151	0.6203	0.2064	0.2262	24.3200
lda	Linear Discriminant Analysis	0.5562	0.6965	0.5562	0.7140	0.6180	0.2047	0.2247	0.6530
ridge	Ridge Classifier	0.5548	0.0000	0.5548	0.7147	0.6181	0.2033	0.2235	0.2500
svm	SVM - Linear Kernel	0.5457	0.0000	0.5457	0.7121	0.6104	0.1891	0.2095	1.5120
dummy	Dummy Classifier	0.2775	0.5000	0.2775	0.0770	0.1206	0.0000	0.0000	0.1770



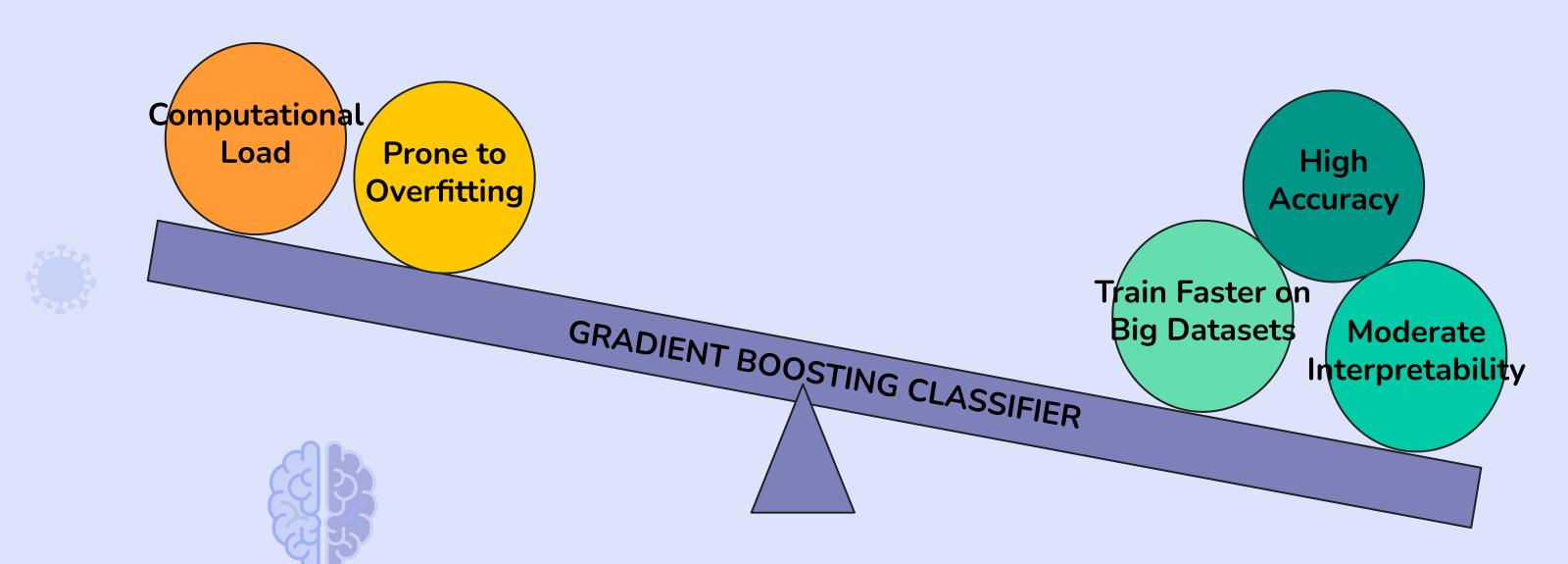


# Machine Learning Model Choice

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

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

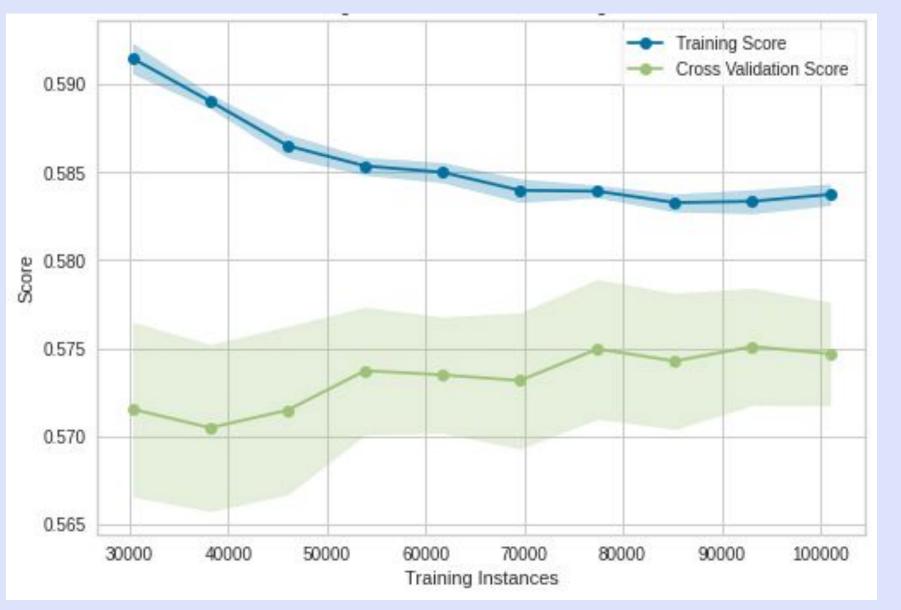
Initial Accuracy: 74.63%





# 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 Final accuracy: 75.46% Final F1: 74.00%

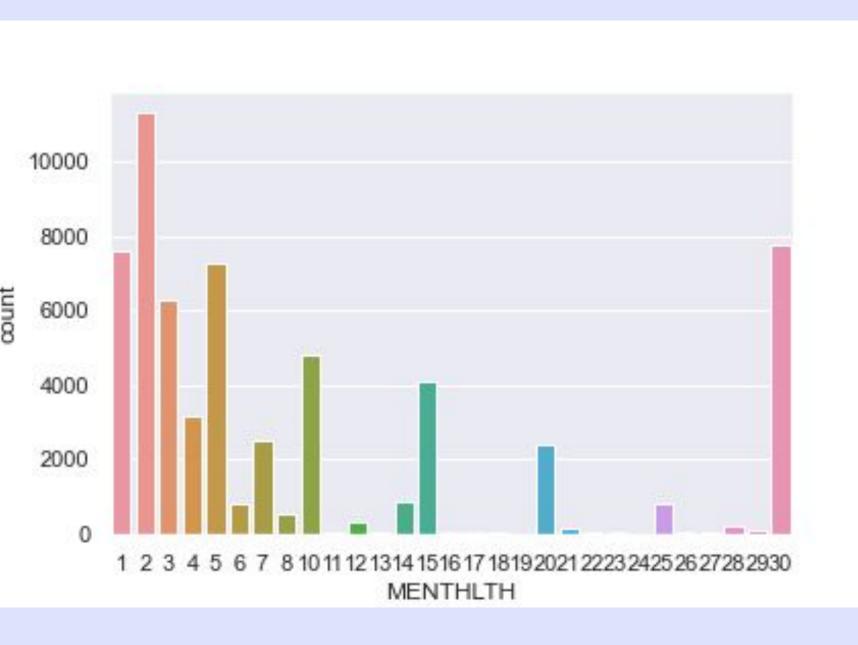


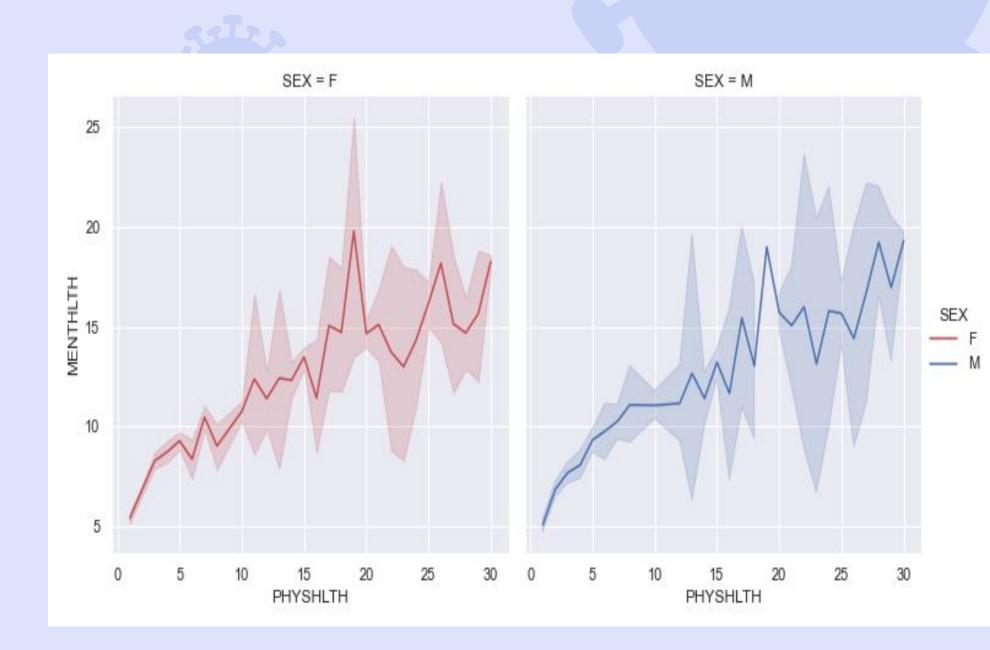






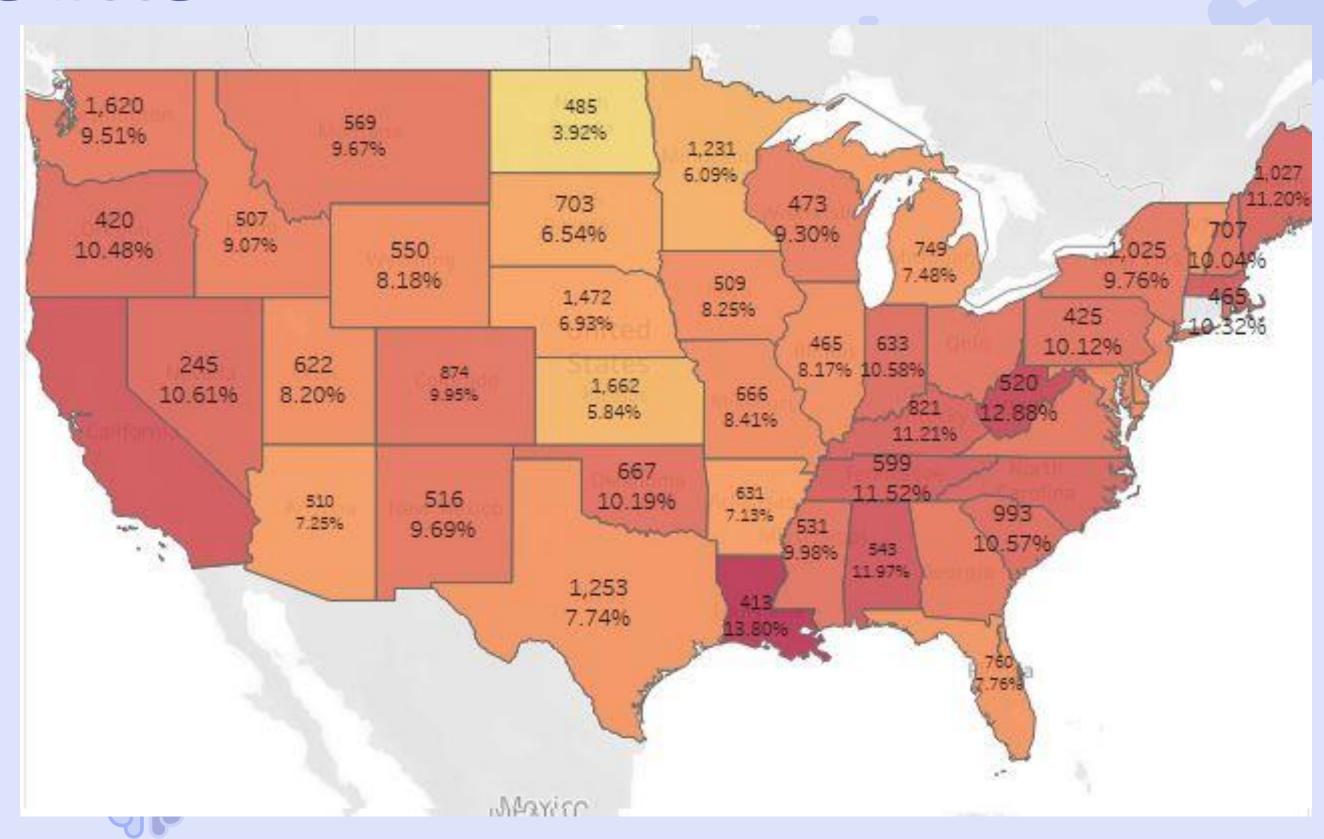
### Results



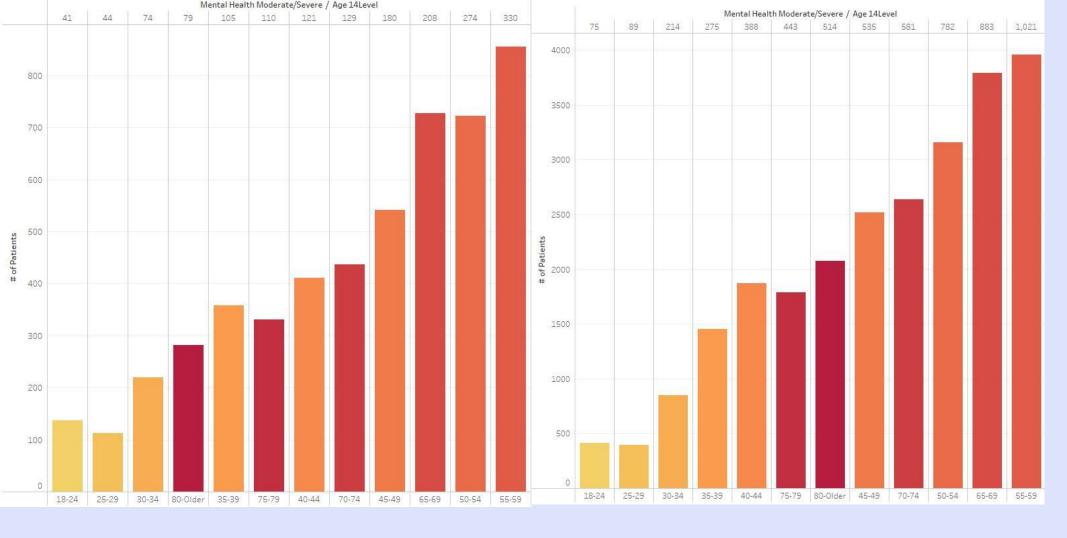




### Results







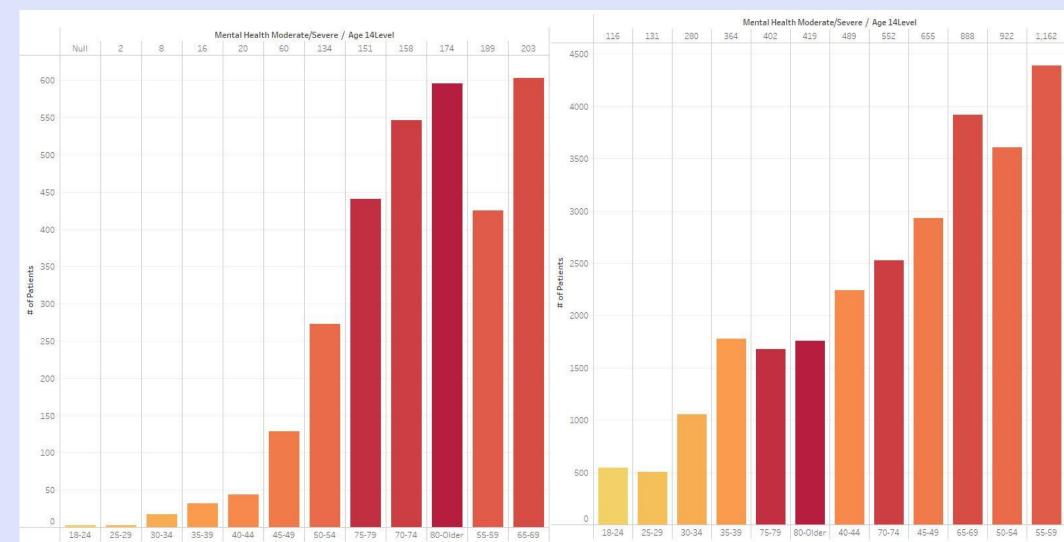




#### **Coronary Disease**

5x more likely





# Results







### 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.

#### More Feature Cleaning Before Machine Learning

Machine learning was performed on 42 columns. Some codes for features included a very small percentage of the responses. The accuracy of the model may be improved if we devoted time to manually go through all 42 columns, removing or recategorizing these very small response categories (such as "refused to answer" or "didn't know")

## Future Analysis Recommendations



#### 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.





