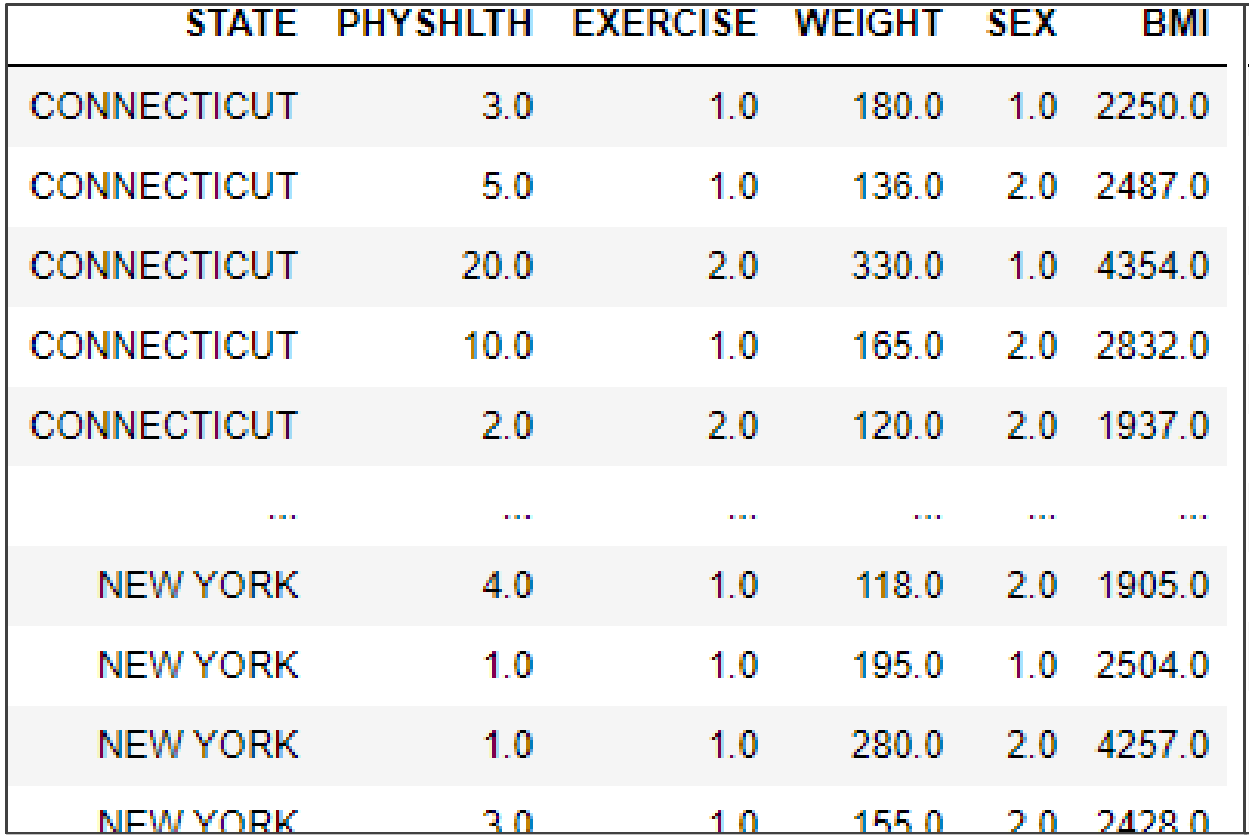
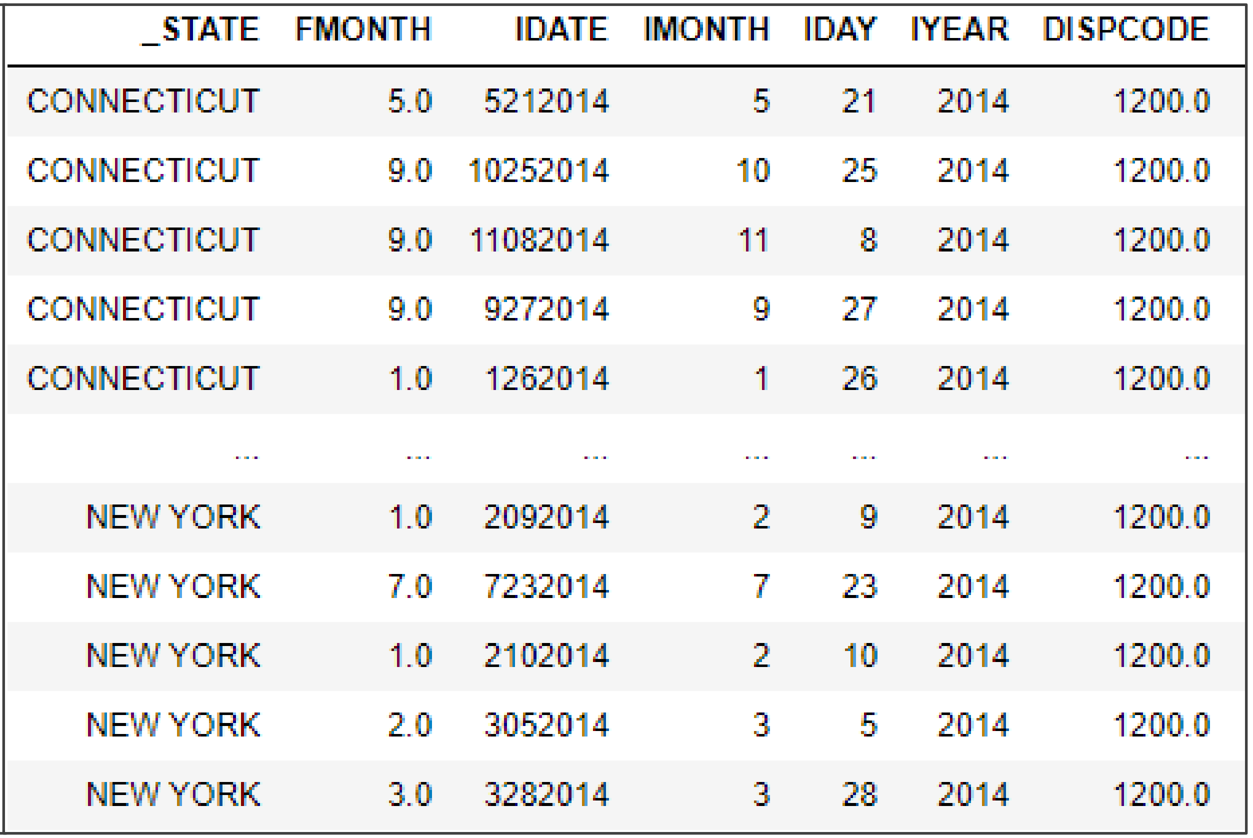
Problem Statement

- Mansourian Insurance is a growing private healthcare insurance provider. The company was founded on the premise that advanced analytics can improve the health insurance industry.
- Based on the yearly CDC data, analyze the 2014 published survey results for three States.
- Understand the demographics, health, and fitness trends affecting Physical Health for the residents of these States.

Data Gathering & Cleansing



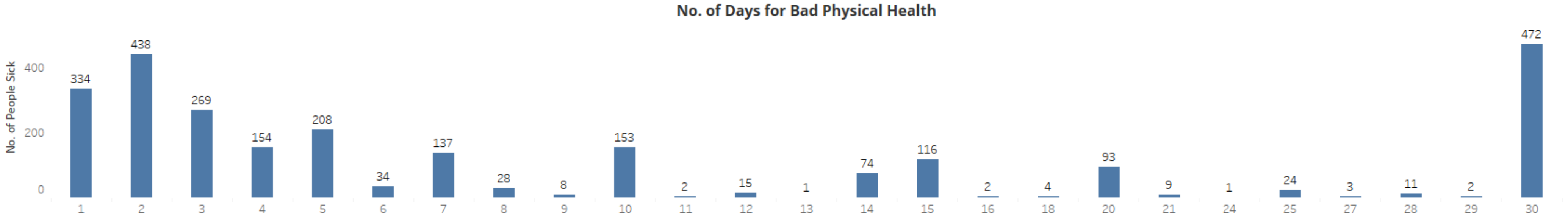
STATE	PHYSHLTH	EXERCISE	WEIGHT	SEX	BMI
CONNECTICUT	3.0	1.0	180.0	1.0	2250.0
CONNECTICUT	5.0	1.0	136.0	2.0	2487.0
CONNECTICUT	20.0	2.0	330.0	1.0	4354.0
CONNECTICUT	10.0	1.0	165.0	2.0	2832.0
CONNECTICUT	2.0	2.0	120.0	2.0	1937.0
...	...	...	...	...	...
NEW YORK	4.0	1.0	118.0	2.0	1905.0
NEW YORK	1.0	1.0	195.0	1.0	2504.0
NEW YORK	1.0	1.0	280.0	2.0	4257.0
NEW YORK	3.0	1.0	155.0	2.0	2428.0



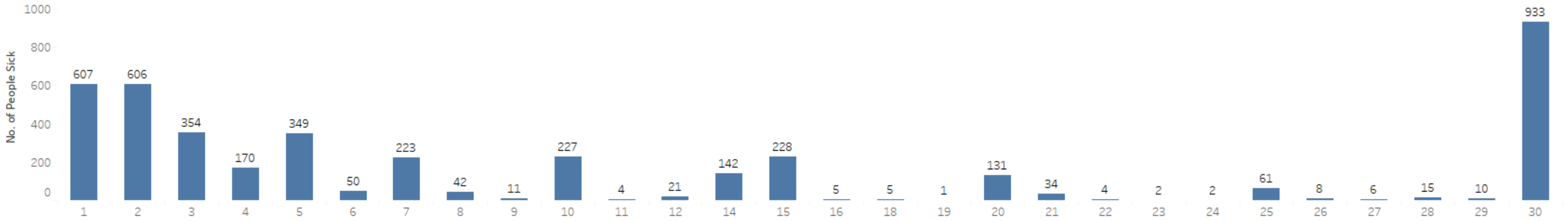
_STATE	FMONTH	IDATE	IMONTH	IDAY	IYEAR	DISPCODE
CONNECTICUT	5.0	5212014	5	21	2014	1200.0
CONNECTICUT	9.0	10252014	10	25	2014	1200.0
CONNECTICUT	9.0	11082014	11	8	2014	1200.0
CONNECTICUT	9.0	9272014	9	27	2014	1200.0
CONNECTICUT	1.0	1262014	1	26	2014	1200.0
...	...	...	...	...	...	...
NEW YORK	1.0	2092014	2	9	2014	1200.0
NEW YORK	7.0	7232014	7	23	2014	1200.0
NEW YORK	1.0	2102014	2	10	2014	1200.0
NEW YORK	2.0	3052014	3	5	2014	1200.0
NEW YORK	3.0	3282014	3	28	2014	1200.0

Tri-State Physical Health Histogram

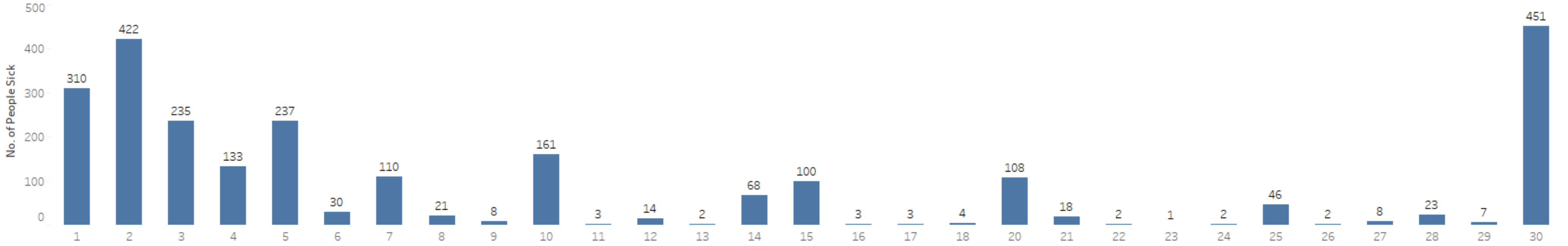
Connecticut



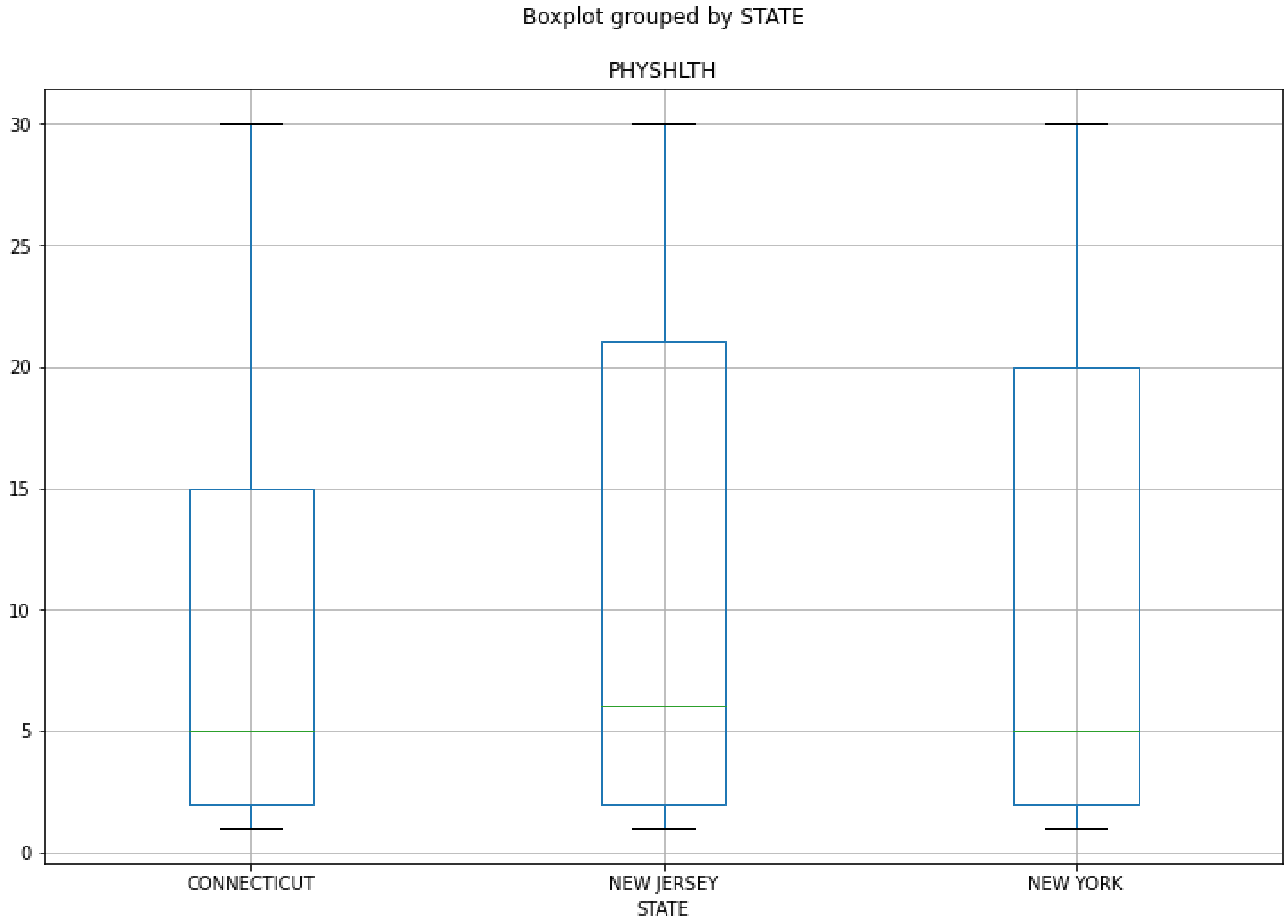
New Jersey



New York



Tri-State Physical Health Boxplot



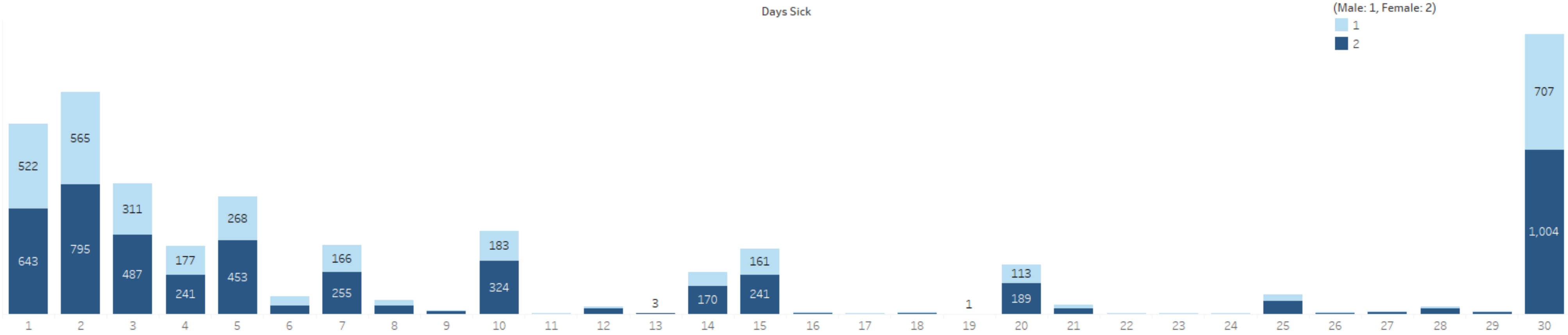
Tri-State Physical Health Statistics

State	Mean (days)	Median (days)	Standard Deviation (days)
CONNECTICUT	10.43	5.0	10.67
NEW JERSEY	11.73	6.0	11.23
NEW YORK	10.93	5.0	10.80
ANOVA, Statistic	12.09		
ANOVA, P-Value	5.72×10^{-6}		

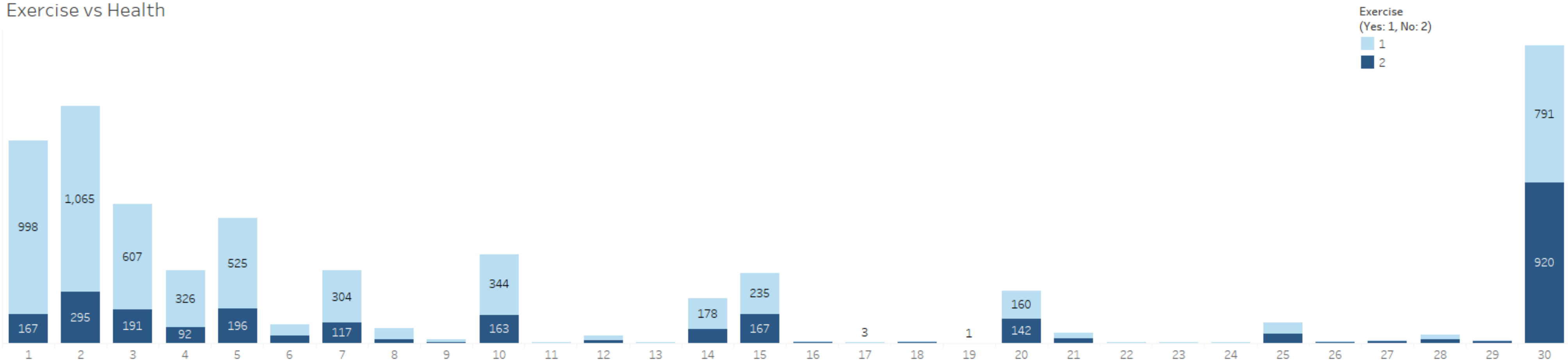
Impact of Exercise & Sex on Physical Health

State
All

SEX vs Health



Exercise vs Health



Correlations Between Dependent Variables & Physical Health

	PHYSHLTH	EXERCISE	WEIGHT	SEX	BMI
PHYSHLTH	**1**	**- 0.285597**	**0.076465**	**- 0.012965**	**0.118236**
EXERCISE	- 0.285597	- 1	- 0.090831	- 0.05867	- 0.143793
WEIGHT	0.076465	0.090831	1	- 0.37463	0.866612
SEX	0.012965	- 0.05867	- 0.37463	1	- 0.05174
BMI	0.118236	0.143793	0.866612	- 0.05174	1

Multiple Linear Regression Model

Dep. Variable	Model	Method	Prob (F-statistic)	Log-Likelihood
PHYSHLTH	OLS	Least Squares	5.06×10^{-174}	-32549
R-squared	Adj. R-squared	F-statistic	No. Observations	Df Model
0.09	0.089	212.3	8636	4

	Coefficients	Standard Error	t-value	P > |t|
Y-Intercept	12.542	0.597	20.999	0.000
EXERCISE	- 6.432	0.245	- 26.21	0.000
WEIGHT	- 0.027	0.006	- 4.276	0.000
SEX	- 0.847	0.308	- 2.754	0.006
BMI	0.003	0.000	6.989	0.000

Taking it Further

- Finding where the significance lies between other variables
- Understand how physical health is different amongst other races
- Investigate physical health data based on disease trends in the region. Such as how seasonal flu is impacting physical health.
- Compare healthcare expenditure per 100k residents versus number of bad physical health days suffered by citizens
- Compare health amongst all SES (Socio-Economic Status) levels
- Conduct Post-hoc analyses to run comparative studies:
  - Duncan's Multiple Range Test (MRT)
  - Tukey Honest Significance Difference (HSD)
- Use the regression model to create supervised machine learning algorithm to predict physical health status based on given features.