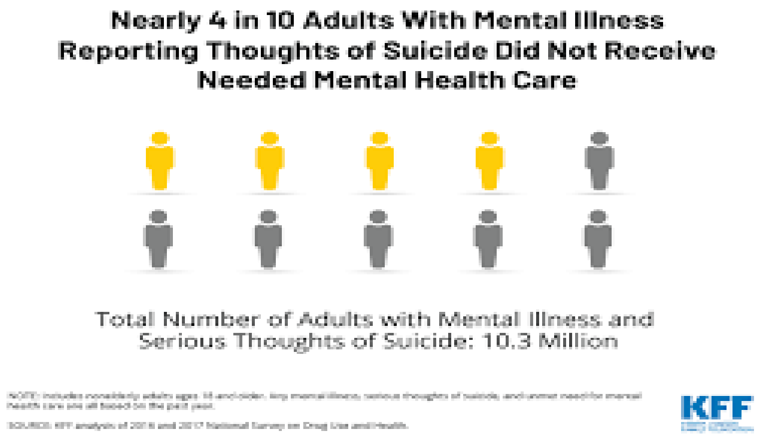
BIG DATA PROJECT

By: Team NSMR

Introduction



Picture Credits: [Kaiser Family Foundation]

Picture Credits: [Tower MSA Partners]

- Important to measure how common mental illness is, so we can understand its physical, social and financial impact.
- Our goal: Analysing the effect of various lifestyle choices/conditions on mental health and producing actionable countermeasures to mitigate mental health issues for at risk demographic.

Background

1. Machine-learning boosted regression analysis found some biomarkers related to depression in the National Health and Nutrition Examination Survey dataset.[1, 7]
2. The effectiveness of predicting a set of common mental conditions based on textual description of patient's history of present illness was discussed by Tung, 2014. [2]
3. There are works identifying depression from social media that use natural language processing and quantify activities on social media platforms, identifying positive or negative emotions, social isolation and other platform specific statistics.[3, 4]
4. There is also growing interest in using sensing data from wearables to infer human mobility and mental health. [5,6].

Picture Credits: Lynda.com

Dataset

Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

National Center for Health Statistics

Mental Depression Questionnaire

- SEQN - Respondent sequence number
- DPQ010 - Have little interest in doing things
- DPQ020 - Feeling down, depressed, or hopeless
- DPQ030 - Trouble sleeping or sleeping too much
- DPQ040 - Feeling tired or having little energy
- DPQ050 - Poor appetite or overeating
- DPQ060 - Feeling bad about yourself
- DPQ070 - Trouble concentrating on things
- DPQ080 - Moving or speaking slowly or too fast
- DPQ090 - Thought you would be better off dead
- DPQ100 - Difficulty these problems have caused

Income Questionnaire

- SEQN - Respondent sequence number
- INQ020 - Income from wages/salaries
- INQ012 - Income from self employment
- INQ030 - Income from Social Security or RR
- INQ060 - Income from other disability pension
- INQ080 - Income from retirement/survivor pension
- INQ090 - Income from Supplemental Security Income
- INQ132 - Income from state/county cash assistance
- INQ140 - Income from interest/dividends or rental
- INQ150 - Income from other sources
- IND235 - Monthly family income
- INDFMMPI - Family monthly poverty level index
- INDFMMPC - Family monthly poverty level category
- INQBOX1 - CHECK ITEM
- INQ244 - Family has savings more than $5000
- IND247 - Total savings/cash assets for the family

Alcohol Consumption Questionnaire

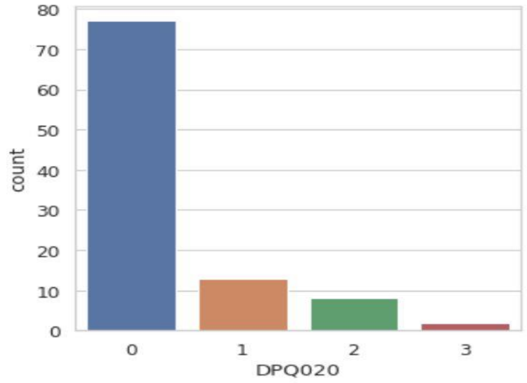
- SEQN - Respondent sequence number
- ALQ101 - Had at least 12 alcohol drinks/1 yr?
- ALQ110 - Had at least 12 alcohol drinks/lifetime?
- ALQ120Q - How often drink alcohol over past 12 mos
- ALQ120U - # days drink alcohol per wk, mo, yr
- ALQ130 - Avg # alcoholic drinks/day -past 12 mos
- ALQ140Q - #days have 5 or more drinks/past 12 mos
- ALQ140U - # days per week, month, year?
- ALQ150 - Ever have 5 or more drinks every day?

Analysis

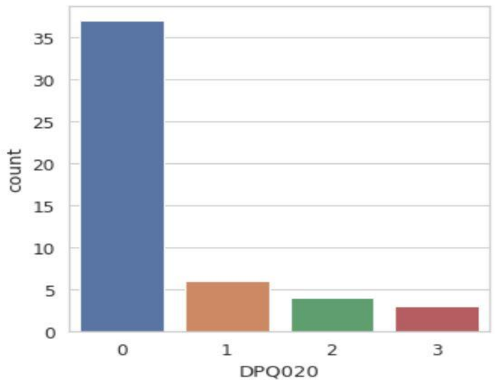
Feeling down, depressed or hopeless?

Value	Meaning
0	0: Not at all
1	1: Several days
2	2: More than half the days
3	3: Nearly everyday

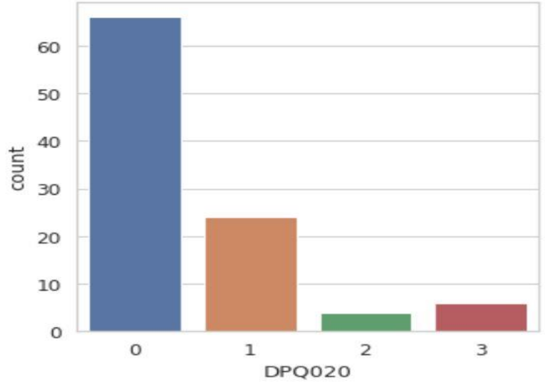
Avg # drinks per day <= 5



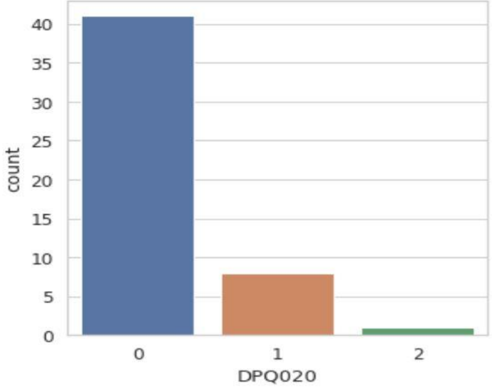
Poverty Index <= 1.3

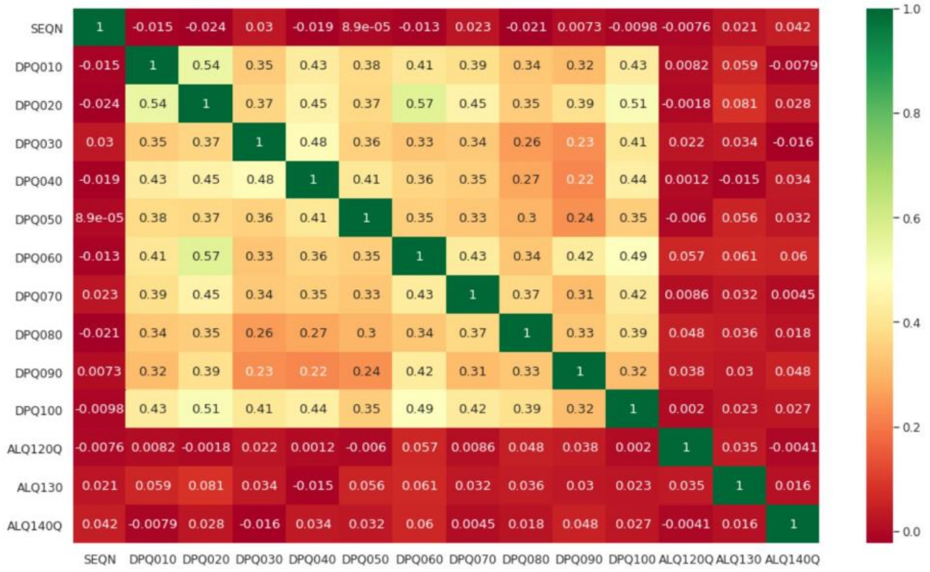


Avg # drinks per day > 5

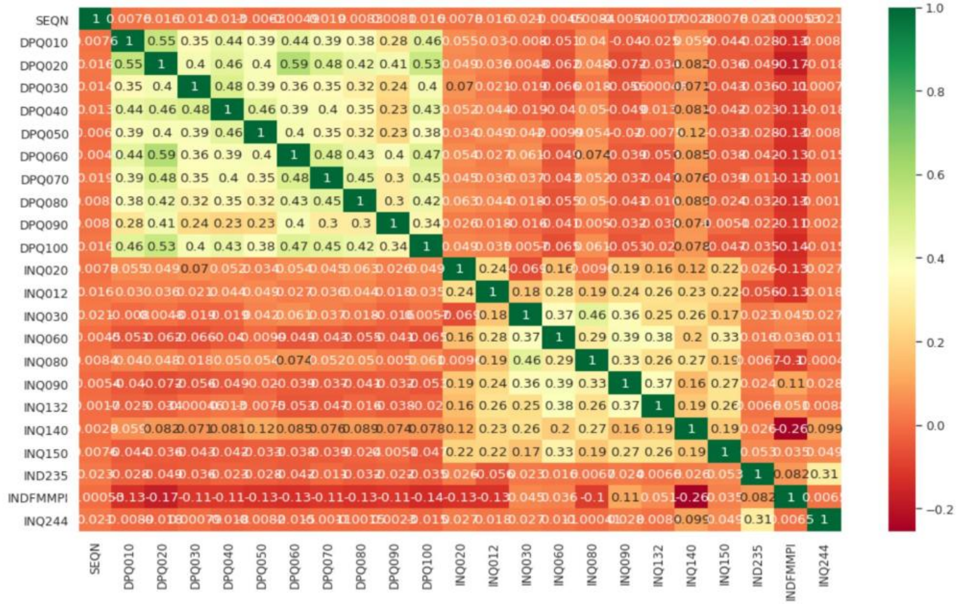


Poverty Index > 1.3





Correlation between Mental Health and Alcohol Consumption



Correlation between Mental Health and Income

Method

Picture Credits: Towards Data Science

- The plan is to analyze the dataset using a multifaceted approach.
- Since the data has been conveniently divided across the years and by topics its easier to manage data in chunks.
- Working on a random sample of the data will help us find interesting insights in the dataset. Repeating the process significant number of times will help us decide about the nature of our hypothesis.
- Since the dataset includes significant number of missing values in important columns because some participants refusing to answer some queries, we plan to use collaborative filtering to fill in the values.

Data pipeline 1 (Recommendation)

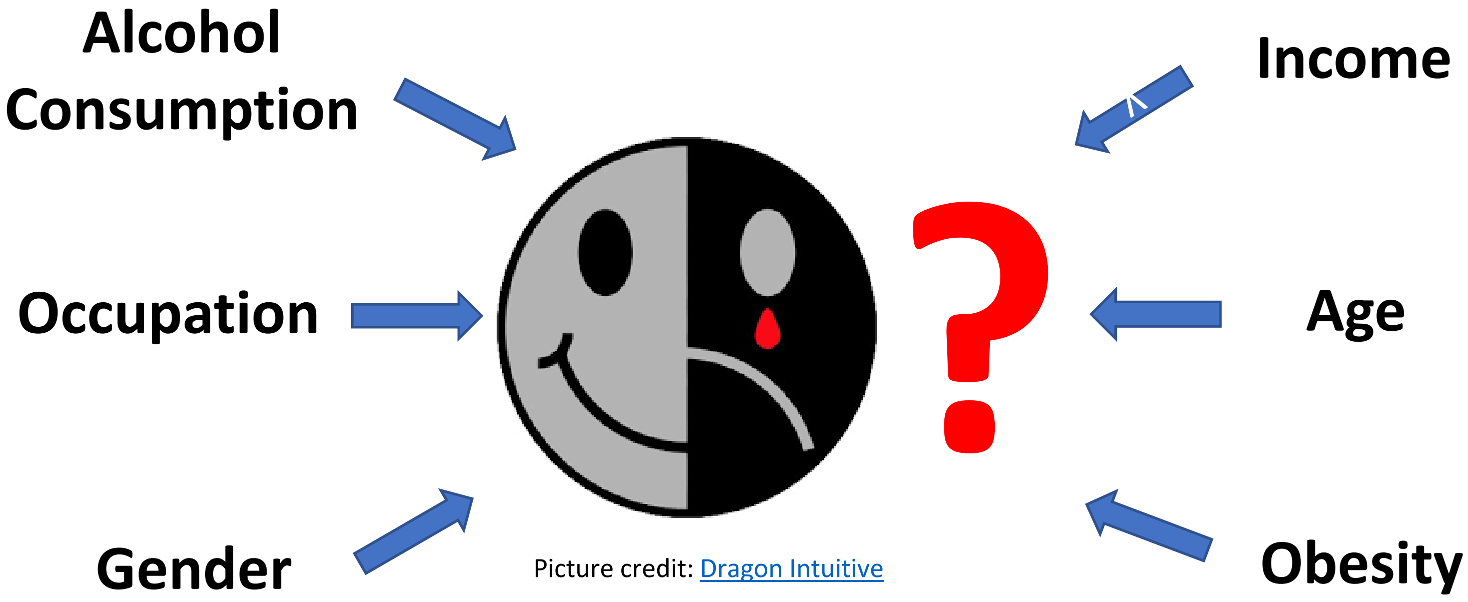
- Access data from HDFS
- Filter a subset of data based on predetermined conditions. (Spark)
- Find nearest neighbors. (Cosine similarity, Spark)
- Run collaborative filtering.(Collaborative Filtering, Spark)
- Impute missing values. (Linear Regression, Spark)
- Predict at risk subjects and output the results.

Data pipeline 2 (Hypothesis testing)

- State a null hypothesis.
- Access data from HDFS
- Sample a subset of data(Random Sampling)
- Calculate the t-value.(Hypothesis testing)
- Apply Bonferroni correction.(Bonferroni Principle)
- Check if its true for $\alpha = \alpha_0$
- Repeat the process N times.
- Output the result.

Mock Results

Goal 1: Predict tendency of depression



Picture credit: Dragon Intuitive

Respondent	Tendency of being depressed?
#1	Very Low
#2	Low
#3	Average
#4	High
#5	Very High

Mock Results

Goal 2: Hypothesis Testing

Picture credit: TheLawOfAttraction.com

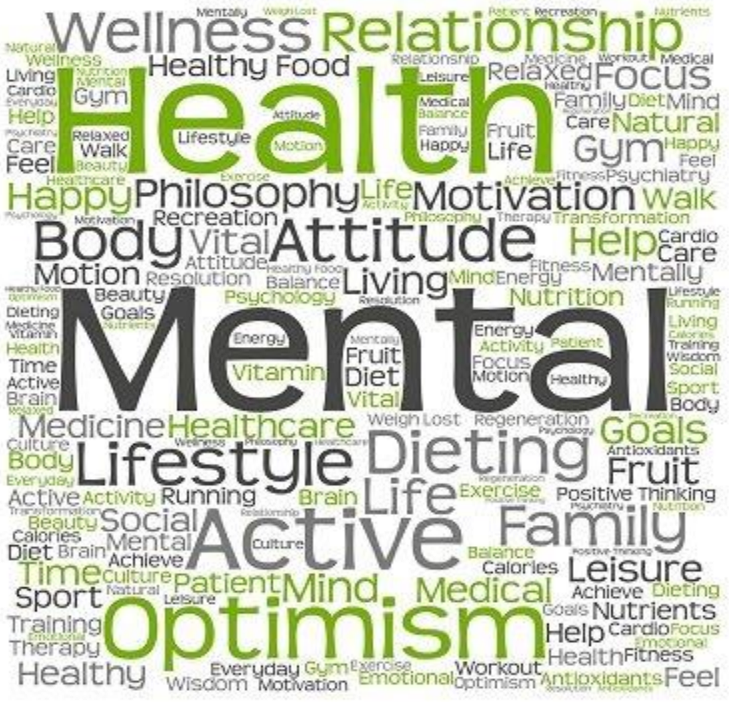
Picture credit: The Calorie Ninja

Picture credit: Deccan Herald

Hypothesis	Sample Result
Money can buy happiness	Failed to reject hypothesis
Obesity causes depression	Hypothesis rejected
Domestically abused males are more likely to commit suicide as compared to females	Failed to reject hypothesis

Conclusion

- With this project we aim to provide multiple data-backed actionable insights which will help the world move towards sustainable development.
- By testing hypothesis surrounding various UN sustainable development goals we aim to find ways in which the goals are intertwined. This specifically will help bring awareness to the fact that sustainable development needs more of a holistic approach instead of an isolated one.



Picture Credits: [Globest]

References

1. Dipnall, Joanna F et al. “Into the Bowels of Depression: Unravelling Medical Symptoms Associated with Depression by Applying Machine-Learning Techniques to a Community Based Population Sample.” PloS one vol. 11,12 e0167055. 9 Dec. 2016, doi:10.1371/journal.pone.0167055
2. Tran, Tung, and Ramakanth Kavuluru. “Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks.” Journal of biomedical informatics. doi:10.1016/j.jbi.2017.06.010
3. Coppersmith, Glen, Mark Dredze, and Craig Harman. "Quantifying mental health signals in Twitter." Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality. 2014.
4. Resnik, Philip, et al. "Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter." Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2015.
5. Miranda Olff. 2015. Mobile mental health: A challenging research agenda. European journal of psychotraumatology 6 (2015).
6. Mohr, David C., Mi Zhang, and Stephen M. Schueller. "Personal sensing: understanding mental health using ubiquitous sensors and machine learning." Annual review of clinical psychology 13 (2017): 23-47.
7. Dipnall, Joanna F., et al. "Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression." PloS one 11.2 (2016).