Analyzing Depression dataset for California State

Prakruti Rajendra Kothari
George Mason University
pkothari@gmu.edu

Abstract— **Depression is one of the common causes of mental illness that has been linked to an increased risk of death among individuals. Furthermore, Depression is a leading cause of suicide and significantly impairs everyday functioning of the individual. Depression is also one of the leading contributors to global disability. Hence, it is important to spread awareness about depression among people. In this paper, I analyzed data from California Behavioral Risk Factor Surveillance Survey (BRFSS) from 2012 to 2017 data. The dataset consists of number of people who were told by doctor or nurse that they have depression, and are labelled in different Strata names such as Sex, Education, Income, Age etc. The results show some interesting insights about the data. The results have been presented using R and Python software.**

Keywords—Depression, BRFSS, Strata, Linear Regression.

I. INTRODUCTION

Depression is the emotional expression of a state which has certain strong narcissistic aspirations [1]. Depression causes unhappiness and a loss of interest in previously appreciated activities [1]. It can cause a slew of mental and physical issues, as well as it affects your capacity to operate in your daily life. Nowadays, due to increase in stress millions of people worldwide suffer from depression. Its diagnosis is made if at least five of the below symptoms occur almost every day for at least 2 weeks [2]:

- Depressed Mood
- Loss on interest in activities
- Suicidal thoughts
- Feeling of worthlessness or hopelessness
- Worsened ability to think and concentrate

Depression can be caused by a variety of factors, including age; older persons are more susceptible to depression. Furthermore, genes can be one of the key causes of depression, and the family history of depression might raise the risk of person getting depression. Furthermore, big life events and other personal issues such as isolation or exclusion from one's family or social group can contribute to depression. As the severity of depression increases, the subjective quality of life decreases. At the worst, depression can lead to suicide. WHO estimates that, in the year 2015, 788,000 people have died by suicide and that it was the second most common cause of death for people between 15 and 29 years old worldwide [5]. One of the reasons for persevering with active treatment for depression is that even if the outlook for survival is poor, quality of life may still be improved [3]. Good health is not limited to physical health [8]. Mental health is also important for well-being of an individual; Therefore, it is important to spread awareness about depression among people so that the person who is depressed doesn't feel insecure in the society. Previous studies have shown that cognitive coping strategies such as ruminating, self-blame and catastrophizing are positively related to depression and/or other measures of mental ill-health, while strategies such as positive reappraisal are negatively related to depression [9].

This paper focuses on California State Behavioral Risk Factor Surveillance Survey (BRFSS) which will help us to answer following research questions about California State:

- Which Age group is mostly affected by Depression?
- Are females more affected by depression compared to Men?
- Is depression getting more common among people day by day?

By analyzing this data California State people can have a interesting insights of their state. The main aim of this paper is to analyze California State residents' mental health condition and their general demographic category under which responses have been stratified. Age, Education, Race/Ethnicity, Sex, Income, Total Population.

II. RELATED WORK

There are various researchers have done research on depression. Some of them which are related to this study are listed in this section. Depression reaches its lowest level in the middle aged, at about age 45 [4]. Young adults seem to benefit psychologically from getting older [4]. Middle-aged adults are the least depressed of all. Average levels of depression rise with age over 60 [4]. We can see this from this dataset also as the frequency count for people with 65+ age is the highest for all the years between 2012-2017.

It is widely acknowledged that women are about twice as likely as men to suffer from clinically relevant symptoms of depression [6]. However, the processes that underline these widespread disparities are still poorly understood. Women's higher propensity to depression might be due to biological differences between men and women, such as hormonal or genetic predispositions [7]. The study results shows that

although differences exist in the extent to which certain cognitive strategies are used by men and women, they play an important role in the reporting of symptoms of depression in both groups. First, the strongest significant differences between men and women were found in the cognitive emotion regulation strategies Rumination and Catastrophizing: women reported to ruminate as well as to catastrophize more often than men [7]. According to the Hispanic Health and Nutrition Ex- amination Survey (HHANES), low educational achievement and low income are factors associated with increased risk for higher levels of depressive symptoms [10]. The National Health and Nutrition Examination Survey III findings indicate that prevalence of depression differs significantly by race/ethnicity but that comparative rates depend on the *type* of depression. African American and Mexican American individuals have higher lifetime prevalence rates of dysthymic disorder, whereas White individuals have higher lifetime prevalence rates of major depressive disorder [11]. Previous research on depression detection has also revealed a set of machine learning models that can be used to detect whether the person has depression or not.

III. DATASET

The dataset for this research was taken from the California Behavioral Risk Factor Surveillance Survey (BRFSS) [6]. The California BRFSS is an annual cross-sectional health-related telephone survey that collects data about California residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. The BRFSS is conducted by Sacramento State under contract from CDPH. This indicator is based on the question: "“Has a doctor, nurse or other health professional EVER told you that you have a depressive disorder [6]. The dataset consists of data from year 2012 to 2017 for California State only.

adult-depression-lghc-indicator-24

Year	Strata	Strata Name	Frequency	Weighted Frequency	Percent	Lower 95% CL	Upper 95% CL
2012	Total	Total	1920		11.74	11.11	12.37
2012	Sex	Male	561	1116664	8.12	7.32	8.92
2012	Sex	Female	1359	2163108	15.25	14.3	16.2
2012	Race-Ethnicity	White	1314	1806371	14.57	13.67	15.46
2012	Race-Ethnicity	Black	97	222022	13.54	10.44	16.65
2012	Race-Ethnicity	Hispanic	412	923174	9.98	8.91	11.05
2012	Race-Ethnicity	Asian/Pacific Islander	61	220418	5.48	3.92	7.03
2012	Race-Ethnicity	Other	36	107786	17.34	11.09	23.6
2012	Education	No High School Diploma	282	579047	14.43	12.6	16.26
2012	Education	High School Graduate or GED Certificate	337	665672	11.44	9.98	12.89
2012	Education	Some College or Tech School	563	947473	13.25	11.95	14.55
2012	Education	College Graduate or Post Grad	717	1040822	10	9.11	10.9
2012	Income	< $20,000	642	1118292	16.97	15.42	18.51
2012	Income	$20,000 - $34,999	295	490532	12.39	10.67	14.1
2012	Income	$35,000 - $49,999	187	296225	9.88	8.16	11.6
2012	Income	$50,000 - $74,999	250	440798	12.65	10.82	14.49
2012	Income	$75,000 - $99,999	160	290492	10.17	8.26	12.07
2012	Income	$100,000.00	270	454444	8.31	7.17	9.46
2012	Age	18 to 34	219	705000	7.63	6.55	8.72
2012	Age	35 to 44	240	576639	11.13	9.57	12.69
2012	Age	45 to 54	409	770238	14.67	13.14	16.2
2012	Age	55 to 64	511	692056	17.19	15.55	18.82
2012	Age	65+ years	541	535838	12.63	11.38	13.87
2013	Total	Total	1689		13.08	12.33	13.82
2013	Sex	Male	539	1307668	9.53	8.53	10.52
2013	Sex	Female	1150	2337817	16.52	15.42	17.62
2013	Race-Ethnicity	White	1103	1979888	15.97	14.91	17.02
2013	Race-Ethnicity	Black	93	252871	15.46	11.91	19.02
2013	Race-Ethnicity	Hispanic	403	1011594	10.96	9.72	12.2

Fig. 1. Data File

The dataset consists of 138 rows and 8 columns with each year’s total California residents count who were told by a doctor or nurse that they have depression. Fig 1 shows a snippet of the data set. The dataset consists of columns Strata, Strata name, frequency, weighted frequency, Percent, Lower 95% CL, and Upper 95%CL. Strata is basically general demographic category under which responses have been stratified. Age, Education, Race/Ethnicity, Sex, Income, Total Population. These five Strata are classified into subcategories into the strata name column. Sex Strata is classified into Male and Female. Race-Ethnicity strata is classified into White, Black, Hispanic and Asian/Pacific Islander. Education strata is classified into No High School Diploma, High School Graduate or GED Certificate, Some College or Tech School and College Graduate or Post Grad. Income strata is classified into <$20k, $20k-$35k, $35k-$50k, $50k-$75k and $75k-100k and Age strata is classified into 18-34yrs ,35-44yrs,45-54yrs,55-64yrs and 65+years.

IV. METHODOLOGY

The dataset from the California Behavioral Risk Factor Surveillance Survey (BRFSS) was in CSV format which made it easier to do various statistical analysis on the data. The data had various variables such as Strata, Strata Name, frequency, weighted frequency percentage etc. The data was analyzed for each Strata from the year 2012 to 2017.

The first step in methodology was to clean the dataset which was done using python. The data consists of some null values for weighted frequency column for the total Strata which was to be removed. Also, the lower 95%CL and Upper 95% CL column from dataset was removed as these columns were irrelevant for this study. Then the data was filtered into different csv files based on each Strata Name one by one and rows which are relevant to only one specific Strata was appended to the csv files. For example, the data for sex Strata consists of two sub parts that is Male and Female, so all the data related to sex Strata was appended to sex csv file.

After the cleaning step the next step was to do some visual analysis of the data. For this R software was used to do some visual and statistical analysis of the data. The Frequency column of the dataset is Nominal, and the Strata Names are of Categorical data types so for this bar plot and line plot was used to do the analysis. Then Linear regression model was also applied to predict the frequency of people who will be diagnosed with depression for the next two decades using the R software. But since that data is too small the model will not give accurate results. The results are shown in Fig 2. In addition, SQL was also used to do some analysis and get some information about the dataset.

```
Call:
lm(formula = Year ~ Frequency, data = Year)

Residuals:
     1      2      3      4      5      6
-1.717 -1.413 -1.559  1.066  1.454  2.168

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.020e+03  6.690e+00 301.848 7.23e-10 ***
Frequency   -3.015e-03  4.001e-03  -0.754    0.493
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.957 on 4 degrees of freedom
Multiple R-squared:  0.1243,    Adjusted R-squared:  -0.09462
F-statistic: 0.5678 on 1 and 4 DF,  p-value: 0.4931
```

Fig. 2. Linear regression Summary

V. RESULTS AND ANALYSIS

A. *Age-wise Analysis*

Fig 3 shows Age wise analysis of number of people who where told by the doctor or nurse that they have depression in the California State. For this the age of people was categorized into five age groups that is from 18-34 years, 35-44 years, 45- 54 years, 55-64 years and 65+ years. The graph shows that as the age increases the number of people who are detected with depression in California State also increases for all the years from 2012 to 2017. And also there is a drop in frequency for 2014 year for all the five age groups.

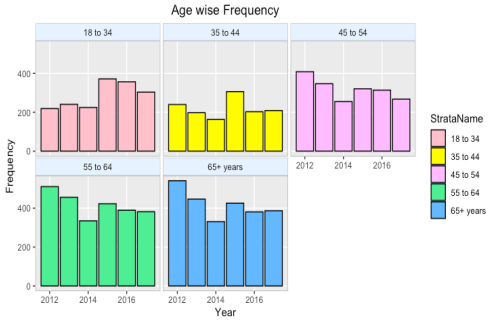

Fig. 3. Age-wise Frequency Plot.

B. *Gender wise analysis*

Fig 4 shows Gender wise analysis of number of people who where told by the doctor or the nurse that they have depression in the California State. From the graph we can see that Female's have higher frequency count compared to Male's for all the years from 2012 to 2017. Also, from the graph we can see that the count is reduced for the year 2014 and then the count has increase for the year 2015 and then the frequency decreases for 2016 and 2017.

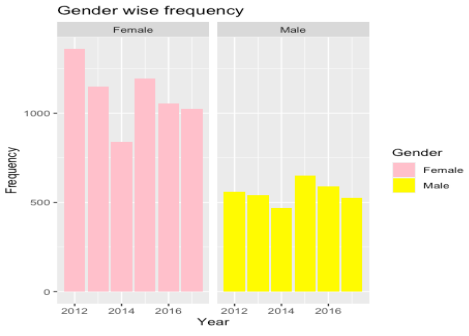

Fig. 4. Gender wise- Frequency Plot.

C. *Year wise total Frequency count for all the Stratas*

Fig 5 shows Total count for all the Strata's in total that is sex, Education, Income, Age and Race-Ethnicity.The graph shows that there is a drastic change in the total count. The total count decreases from 2012 to 2014 and then it suddenly increases with large number in the year 2015 and then again decreases for the year 2016 to 2017. To predict the Total count for the next two decades Linear Regression was performed on this data.

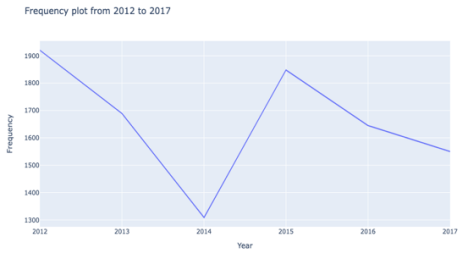

Fig. 5. Total Frequency plot for year from 2012 to 2017

D. *Income wise Analysis*

Fig 6 shows frequency count for various income categorized people in the California Sate. We can see from the graph that people with <$20,000 income are highest affected with depression in the California State and people with all the other income categories are almost equal in count for all the years from 2012-2017.According to the Hispanic Health and Nutrition Ex- amination Survey (HHANES), low educational achievement and low income are factors associated with increased risk for higher levels of depressive symptoms [10].

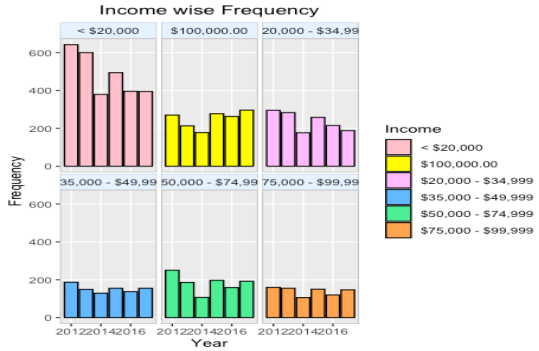

Fig. 6. Income wise frequency plot.

E. *Race-Ethnicity wise Analysis*

Fig 7 shows Frequency count for various subcategories for the Race-Ethnicity Strata of the California State. From the graph we can see that White has the highest count then Hispanic and then Asian/Pacific Islander, Black and others have almost equal count for all the years then the Hispanic Race. One reason of White being the largest would be of the distribution of Race-Ethnicity in the California State.

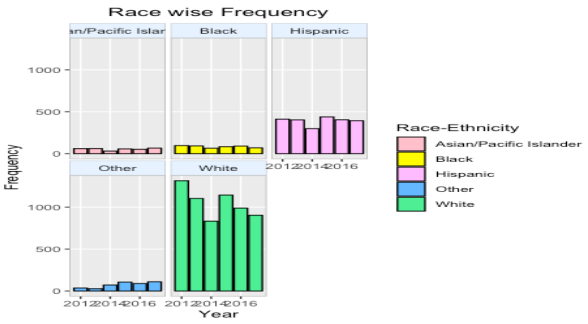

Fig. 7. Year wise frequency for Gender Strata

F. *Education wise Analysis*

Fig 8 shows frequency count for various subcategories for the Education Strata of the California State. From the graph we can be seen that people with Higher education have higher count compared to all the No high school Diploma resident for all the years from 2012-2017.

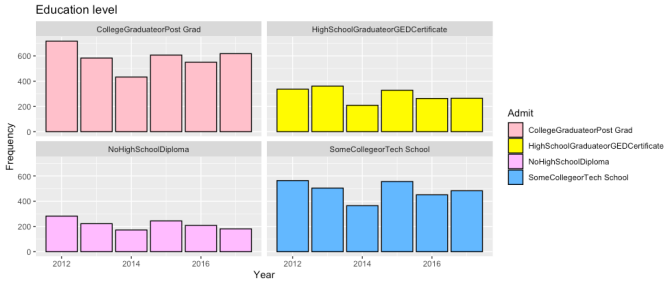

Fig. 8. Year wise frequency for Education Strata

VI. Limitations and Future Work

One limitations of this dataset are that the data provided by the residents of California State may not be 100% accurate. As this, data is collected on Telephone by the California State Government so there is no proof if the residents are being honest to the survey or not. Second, limitation of this dataset is that this dataset may not be equally distributed among all the Strata categories. As we can see in the Race-Ethnicity category it shows highest count for White Race it is because there may be more White's in California compared to other categories.
Third, as this study consists of data from 2012-2017, there can be further improvements in this research which makes it difficult to lay down any conclusions from this dataset. As, the data is too small it is difficult to fit a good model with high accuracy to analyze this data. Future work that can be done in this dataset is that the dataset can be increase which would help to have better analysis of the data. Future work that can be done on this dataset. Also, as in this study the data is collected on telephone which makes it less accurate, we can bud various models using Artificial Intelligence to detect whether the person has depression or not.

VII. Conclusion

The data analysis undertaken in this study has highlighted some of the interesting insights about depression in the California Sate which was gathered to answer the research topics mentioned previously. From this study we can conclude that People above the age of 65+ years are more likely to be diagnosed as depressed by the doctor or the nurse in the California State. From this study we can also conclude that Females are more likely to diagnosed by depression compared to males. Also, we can see that low-income people whose income is less than $20,0000 have high frequency compared to other classes. This study has insufficient data to answer the question whether depression is getting common day by day.

References

[1] Bibring, E. (1953). The mechanism of depression. In P. Greenacre (Ed.), Affective disorders; psychoanalytic contributions to their study (pp. 13–48). International Universities Press.

[2] M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence", IEEE ICISS Palladam, 2017, ISBN:978-1-5386-1959-9

[3] Goldberg D. (2010). The detection and treatment of depression in the physically ill. World psychiatry : official journal of the World Psychiatric Association (WPA), 9(1), 16–20. https://doi.org/10.1002/j.2051-5545.2010.tb00256.x

[4] Mirowsky, J., & Ross, C. E. (1992). Age and Depression. Journal of Health and Social Behavior, 33(3), 187–205. https://doi.org/10.2307/2137349.

[5] Depression and Other Common Mental Disorders: Global Health Estimates. World Health Organization 2017, https://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/

[6] Nolen-Hoeksema, S. (1991). Responses to depression and their effects on the duration of depressive episodes. Journal of Abnormal Psychology, 100, 569–582

[7] Garnefski, N., Teerds, J., Kraaij, V., Legerstee, J., & van Den Kommer, T. (2004). Cognitive emotion regulation strategies and depressive symptoms: Differences between males and females. *Personality and individual differences*, *36*(2), 267-276

[8] Office of Health Equity. (August 2015). Portrait of Promise: California Statewide Plan to Promote Health Equity and Mental Health Equity. California Department of Public Health., http://www.cdph.ca.gov/programs/Documents/CDPHOHEDisparityReportAug2015.pdf

[9] Anderson, C. A., Miller, R. S., Riger, A. L., Dill, J. C., & Sedikides, C. (1994). Behavioral and characterological styles as predictors of depression and loneliness: review, refinement, and test. Journal of Personality and Social Psychology, 66, 549–558

[10] Moscicki, E. K., Locke, B. Z., Donald, S. R., & Boyd, J. H. (1989). Depressive symptoms among Mexican Americans: The Hispanic Health and Nutrition Examination Survey.*American Journal of Epiderminology,130(2),348-360*

[11] Riolo, S. A., Nguyen, T. A., Greden, J. F., & King, C. A. (2005). Prevalence of depression by race/ethnicity: findings from the National Health and Nutrition Examination Survey III. *American journal of public health*, *95*(6), 998-1000