UNIVERSITY of BRADFORD

School of Electrical Engineering & Computer Science

COS7046-B

# Big Data Visualization

**What category or class of people have most instances of Depression diagnosis in the state of California**

[Author name]

# ABSTRACT

Visualisation for Health care has become quite a challenge over the past few years as the target audience has not been limited to health professionals anymore. The awareness to gain knowledge for several health disorders has gained momentum over the years among the common masses. This has significantly added to the target audience that views health data and its visualisation, by quite a margin in the past two decades.

The presence of internet on the fingertips has added a rather significant amount of responsibility over the shoulders of stakeholders that compile data to draw information from it.

In this respect, it is quite integral to use most effective methods to analyse the data and utilise most suitable techniques for presenting it visually, bearing in mind who the target audience is (Munzner 2009).

This report draws observation from a Dataset that has been collected from an online source (health.gov 2017) and aims to see whether the dataset answers the questions that have been proposed, by means of data visualisation techniques.

# Table of Contents

# INTRODUCTION TO DATASET

The sudden surge in the size of data, and the challenges that follow as a result have raised several questions on the data's effective management and usability. It is not just the enormity of the Big Data that makes the task of handling it challenging, it is Big Data's heterogeneous growth as well, that makes it even more hard to manipulate (Dasgupta 2018). Furthermore, the pace at which the data continues to grow, and change is an even bigger challenge, and there is a dire need to come up with systematic and effective approaches to harness this data and put it to most favourable use (Madsen 2014).

What is equally challenging is presenting this data to general public, that may have limited knowledge of information that data is trying to convey. Hence, use of techniques that tell a story that depicts the statistics to a close effect is integral.

The data set that has been selected for this coursework comprises of key indicators with regards to depression in adults in the state of California. It has been collected by Let's get Healthy California (LGHC) initiative. This data set measures depression in adults from ages 18 to 65 and over, who have ever been told that they had depression by a clinician. It is measured against gender, race, education, and income per annum to identify most vulnerable groups. It is a relatively concise data set and comprises of records between 2012 and 2017 and has been published by the healthdata.gov website.

This data requires to be visually presented to a wider audience, and not just health professionals. It needs to have employed use of techniques that are most suitably understood by a wide target audience.

The objective of this report is to state the techniques that have been utilised, and to highlight why they were appropriate for the dataset. The report also makes recommendations on which other techniques could have been more useful.

## QUESTIONS DERIVED FROM THE DATASET

The dataset was initially published as a means of awareness amongst the masses on the significant increase in mental illnesses in the state of California. This data can be used to draw several important conclusions with regards to clinical depression and its possible implications as well as the shifting trends in the diagnosis. While the visualisations done for this project draw light on the increasing trends, they also shift considerable focus on whether the trends remain steady or not, or to identify any changes that might be occurring in these trends. The dataset information will be retrieved to see how expectations are a lot different than reality.

The question that this dataset can answer is: **What category or class of people have most instances of Depression diagnosis in the state of California.**

When deciding on the visualization techniques, it is important to understand that the human brain finds it quite easy to accept complicated statistics, when presented in form of visuals that best depict and translate conciseness of a dataset (Agarwal, et al. 2016).

Since the report aims at showing trends in the diagnosis of depression in California, the visualisation methods and techniques that depict the increase or decrease in trends most effectively have been utilised for this purpose. Furthermore, the visualisations also aim to highlight most vulnerable groups in this regard.

## ANALYSIS METHODOLOGY

Although the data set is relatively small, Munzner model of visualisation design has been employed to make the visualisation process more systematic and enhance the learning experience for future projects.

For the purpose of making the data easy to understand and visualise, the dataset has been treated. A second form of Normalisation has been conducted so that it is visually presented to show similar measurements in all plots. The figures for the number of populations in the state of California that have been diagnosed with Depression have been converted to percentage of the diagnosed groups against total population. This keeps a uniformity and parallelism of measure in all the tools that have been used in this report.

The raw data that was originally obtained from health.gov (2017) has been treated to make it more categorical and tabular (Munzner 2009). This has provided the author with significant clarity on how to use and visually present the data.

Several mock-up and paper designs were created to ascertain which methods and techniques were most suited for the given dataset. Also, in order to see how various platforms would output the figures, mock data was used (Backonja et al. 2018).

Linear Interpolation was conducted by employing the use of MATLAB to see the varying trends in the provided data.  Several software platforms were used along with MATLAB, Power Bi, and Tableau to create other visualizations and mock-ups. Microsoft Excel was extensively used in mock-up as well as final plots.

## VISUALISATIONS and RESULTS

The dataset in this case has been normalised and updated using MS Excel and Matlab. Three formats of visualization techniques have been used to maintain uniformity, considering the use of similar parameters for representation of ascending trends in all plots. The data is relatively small; hence it doesn't crowd the plots in any instance.

During the experimentation phase, several plots were generated in varying formats to test which is best representation. The use of Radar, Horizontal line bar graph, or scatter plots were experimented with. However, it was decided to use Line, column, and scatter graphs.

The first graph is a representation of the trends in diagnosis of depression across the board, without any categorisation in years 2012 to 2017. A simple line graph is used to show a relative increase in this trend over the years the data has been collected for.

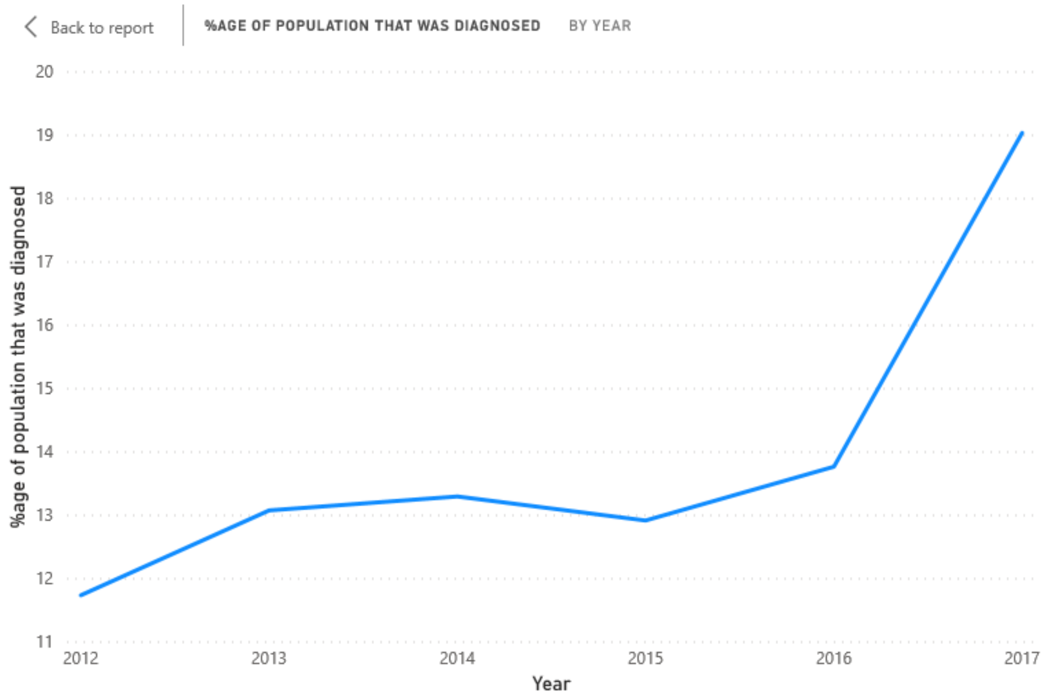

*Figure 1: %age of Total populations that have been diagnosed with depression in years 2012 to 2017 in the state of California*

The second graph represents the diagnosis, categorised in genders. The use of line graph, with points highlighted for each year makes it easier to observe the trend for any particular year. The use of different colours also helps in categorising the two gender classes.

Rest of the graphs, with the exception of one have been plotted in column graph format that help in seeing a steady increase quite effectively. They also provide to compare several parameters within one graph.

Scatter plot technique was put to use to show trends that peaked in all categories, against each other to identify the most vulnerable groups that could be targeted by the health care provision systems dedicated to California.

The first graph that depicts the diagnosis in depression amongst the population in California quite clearly points towards an ascend, particularly between 2016 and 2017.

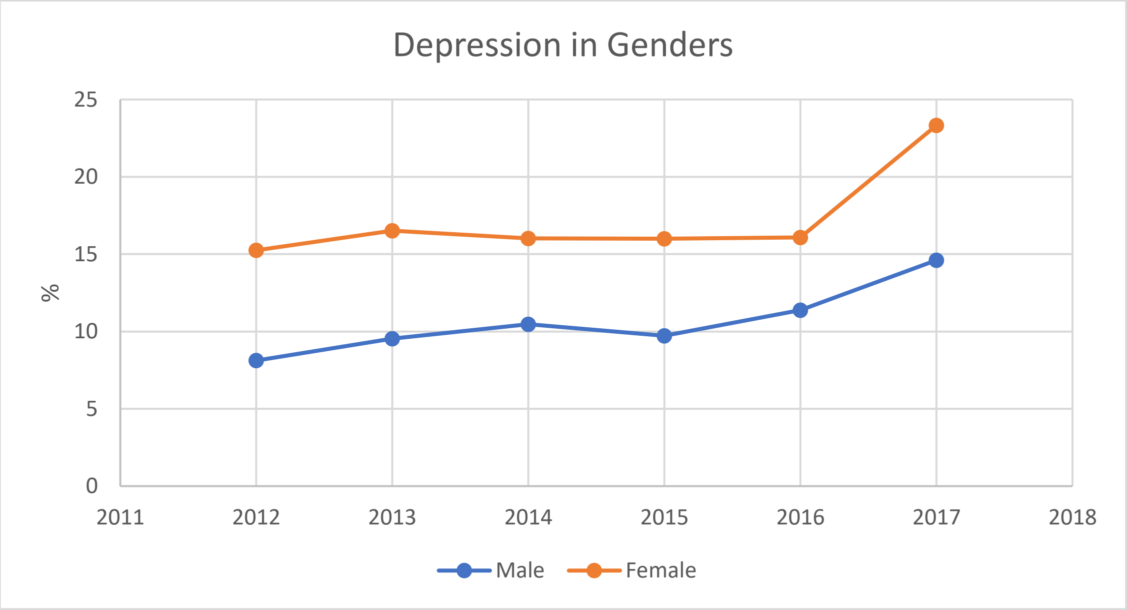

*Figure 2: Varying trends among the Genders*

The graph that represents trends in gender classes is quite clearly pointing towards not only increase through the years, but also how females form a more vulnerable group.

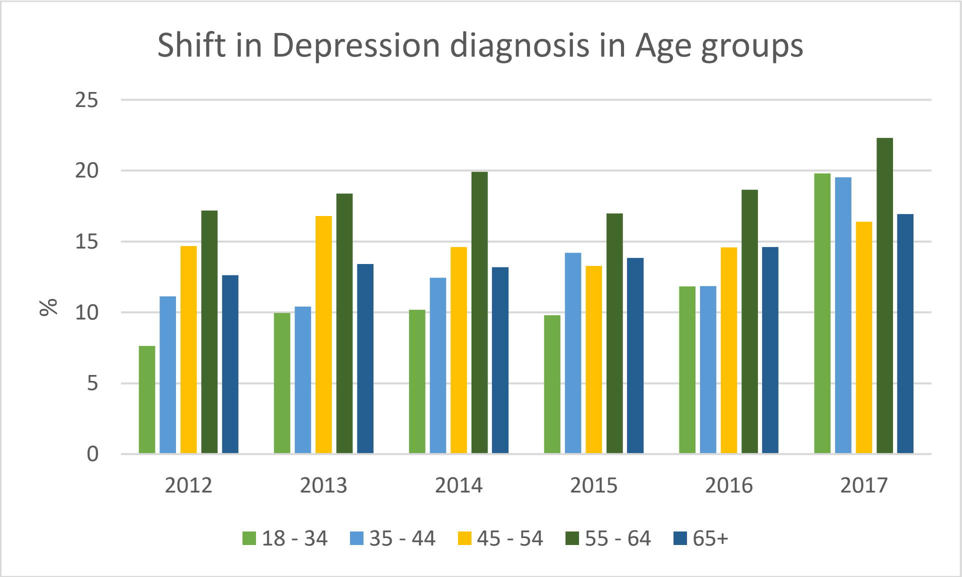

*Figure 3: Visualisation of Shift in depression diagnosis in varying age groups*

The other class factors that have been highlighted employ the use of column graphs, each sub classification presented in a different colour.

The readings are taken against percentages against the total population. The graphs clearly indicate the attributes that are prominent affected.

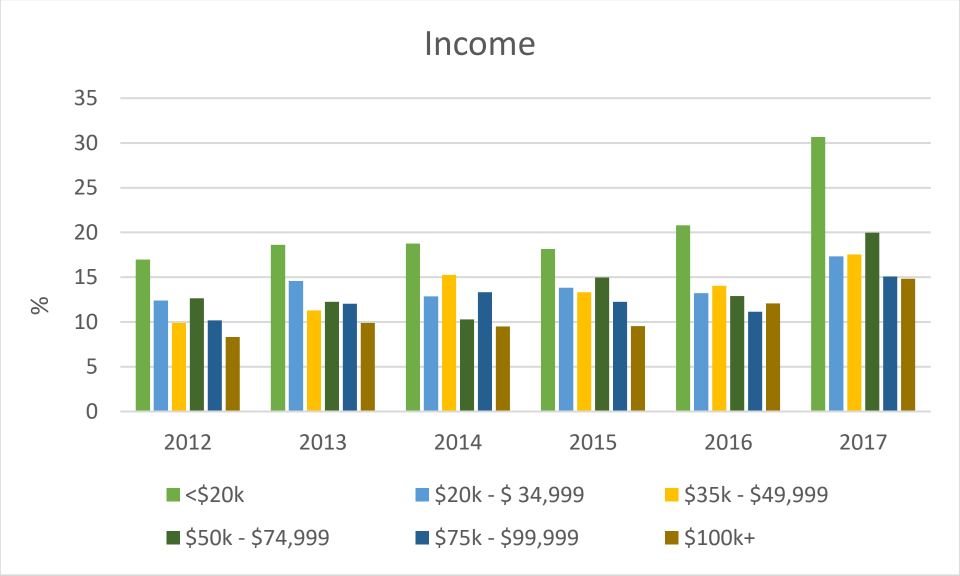

*Figure 4: Visualisation of Shift in depression diagnosis in varying income groups*

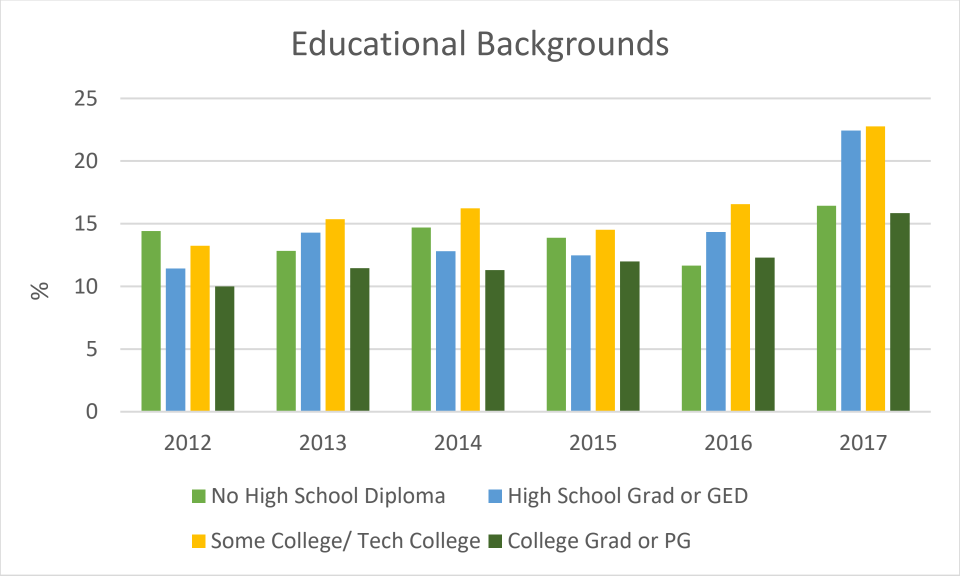

*Figure 5: Visualisation of Shift in depression diagnosis in varying qualification groups*

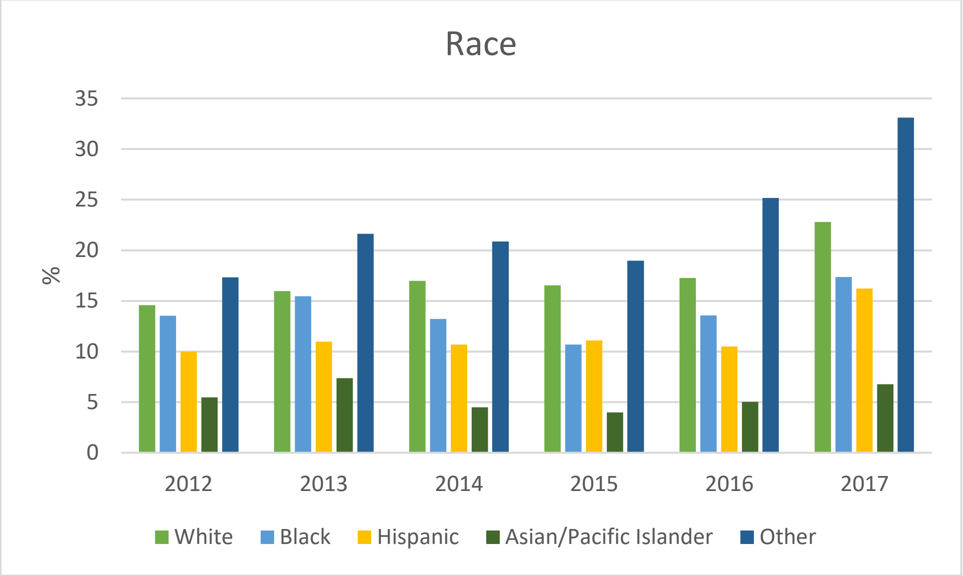

*Figure 6: Visualisation of Shift in depression diagnosis in ethnic groups*

After observation taken from the graphs that have been potted, it was decided to visually present the data to indicate most common cause, by means of a comparison of the most vulnerable groups in all cases over the years.

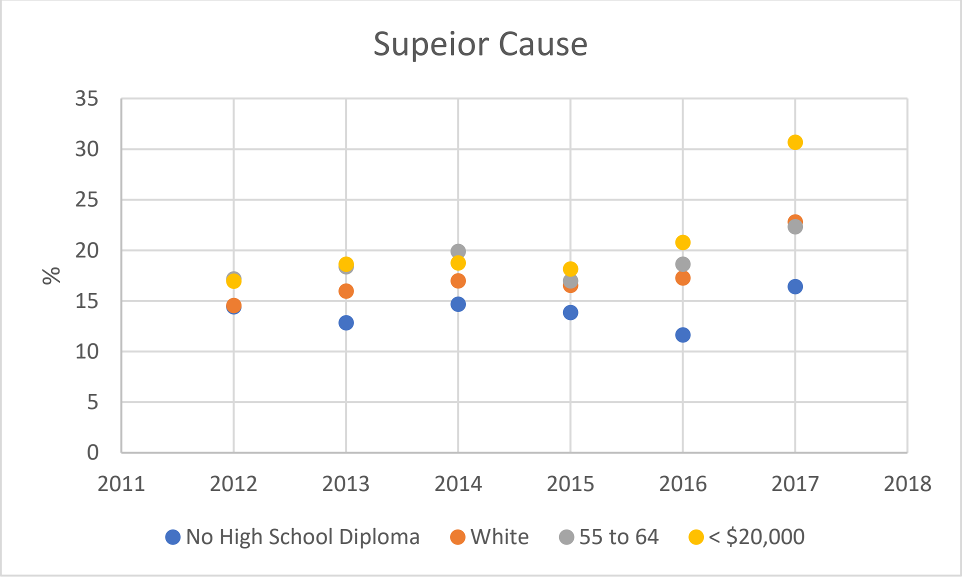

*Figure 7: A plot to see most vulnerable group*

While the graph clearly displays some sort of alignment in these trends, the year 2017 clearly points out the cause of this mental disorder amongst the masses in California to be low income group, followed by race and age factors.

## DISCUSSION

The dataset that was chosen for this project is quite an integral one and could play an important role in identifying where special focus needs to be rendered in efforts to treat depression. While it will be cumbersome for the concerned bodies that might be interested in using this data, the visual representation aims to show the statistics by means of effective story telling.

The use of a line graph clearly indicates increasing trends, but line graphs are quite limited or incomplete in their scope, and may show the changing trend, but reading the varying slopes can be cumbersome or misleading (Shah, Carpenter 1995). The gaps that the process of linear interpolation plots based on prediction may not be error free, or accurate (Caruso, Quarta 1998).

The use of column bar graphs in this respect quite clearly takes the lead, however, visually presenting this data by using contour maps or heat maps could have proved a better technique to this effect, as these plots represent varying trends for locations on maps (Kirk 2016). This technique could have been a visual reminder of where the data is from. Perhaps an added attribute of the locations or neighbourhood where residents have been diagnosed with depression most, could have given a more pronounced depth to the dataset, as well as its visualisations.

To show a comparison amongst most vulnerable groups, a scatter plot has been utilised. This technique seems to show the class of depression that falls in low income group tends to have larger figures in depression diagnosis. However, a tree map could have effectively indicated the size of the effect (Kirk 2016). A Matrix plot can also be experimented with and might have generated interesting results.

## CONCLUSION

As the world gallops towards the digitalisation of all processes, health care industry is quite effectively taking a big chunk of the pie. Healthcare data in various capacities is huge and can be used for improvement of current procedures or research purposes, effectively.

Various organisations aim to apply use of visualisation techniques to present big data, and as visuals tell a better story than statistics, it is important for that story to be told as accurately as the numbers depict. Here, decision making on which method to apply is quite important in differentiating a visually well represented data from the one that is not (Few 2007).

The use of line graph, column graphs, and scatter plot for the dataset are helpful in indicating the varying trends in depression diagnosis in California. The visualisation can deliver a lot more to the target audience, should the data itself supply information that add more depth to the plots.

The database however, successfully answers the question and highlights the five main vulnerable groups in this regard, in addition to showing ascending trends, which is quite alarming.

This data and visualisations can be used to predict future trends and remedies for current scenarios.

# REFERENCES

Dasgupta, N (2018) *Practical Big Data Analytics*. 1st Edition. Birmingham: Packt.

T. Munzner (2009) A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics* 15 (6), 921-928.

Madsen, L (2014) What does data mean to you? In John Wiley and sons (editors). *Data driven healthcare*.

Nitin Goyal, Mayank Dave, Anil K. Verma (2019) Data aggregation in underwater wireless sensor network. *Recent approaches and issues Journal of King Saud University - Computer and Information Sciences* 31 (3), 275-286.

H. B. Sankaranarayanan, G. Agarwal and V. Rathod (2016) An exploratory data analysis of airport wait times using big data visualisation techniques. *International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, Bangalore. 324-329.

Kirk, A (2016) *Data Visualisation*. Glasgow: Sage.

Shah, P., Carpenter, P. (1995) Conceptual Limitations in Comprehending Line Graphs. *Journal of Experimental Psychology* 124 (1), 43 – 61.

Caruso, C., Quarta, F. (1998) Interpolation Methods Comparison. *Computer Math Applic* 35 (12), 109-126.

Few, S. (2007) Data Visualisation: Past, Present and Future. *Perceptual Edge.*