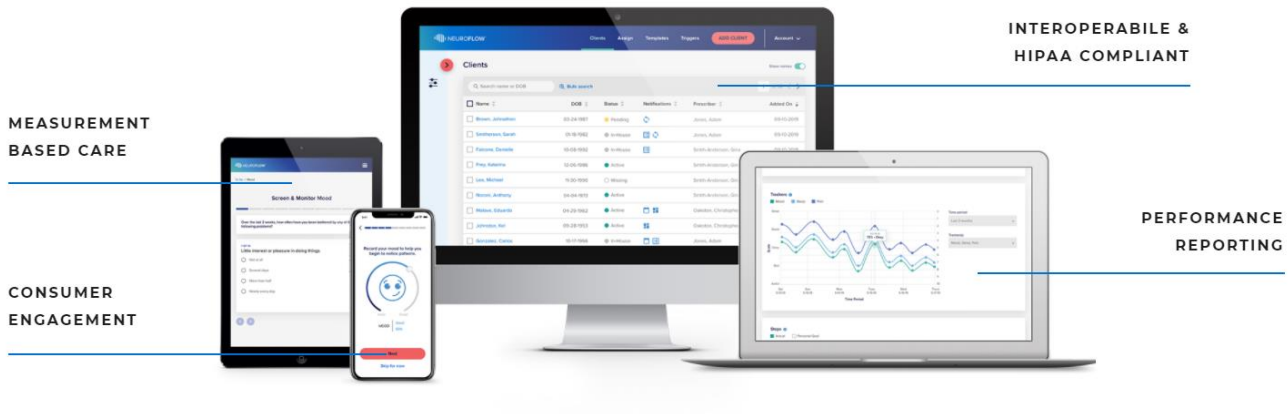
Ayan Kumar Das

NEUROFLOW DATA TEAM TAKE-HOME PROJECT

GAD-7 Assessment Analysis

May 2021



CONTENTS

1. INTRODUCTION

1.1. INTRODUCTION

For this project, we are considering data from GAD-7 validated assessments, which is used to aid in the diagnosis of generalized anxiety disorder as well as screen for panic, social anxiety, and post-traumatic stress disorder.

The GAD-7 is a 7-question assessment that asks how often one has been bothered by the problems represented in each question over the last two weeks.

The person then scores each question from 0 to 3 using the following scale:

Question Score	Question Label
0	Not at all sure
1	Several days
2	Over half of the days
3	Nearly every day

To score the GAD-7, the values in each question are summed to get the final score.

The final scoring scale is as follows:

GAD7 Score	Severity Label
0-5	Low to Minimal
6-10	Mild
11-15	Moderate
16-21	Severe

When screening for anxiety disorders, a recommended threshold for further clinical evaluation is a score of 10 or greater.

1.2. BUSINESS PROBLEM

During treatment, patients are asked by their care provider to complete these validated assessments. The cadence at which these are given are typically monthly or twice a

month, but it can vary. The clinical purpose of these assessments is to help support clinicians in making a diagnosis, to quantify anxiety symptoms, and to monitor changes over time to see if therapy is making a difference. The provider can see this data too, building the basis of a conversation they can have together.

We currently have the problem of not being able to visualize progress well for this assessment to mental health providers and their patients.

1.3. OBJECTIVE

Given the information and data we have, we will be implementing EDA and descriptive statistics to draw insights and inference. Our main aim is to check whether the treatment is helping the patient to overcome their anxiety problems and translate our outcome into user-friendly visualizations.

1.4. ASSUMPTIONS

There are few assumptions made prior to analysis as mentioned below.

- The data is reliable and true depiction of patient's behavior.
- The parameters of GAD-7 score are uncorrelated or else they can create a bias in our analysis.
- All the assessments of a patient are taken during the same daytime to emulate same mood and feelings.

1.5. ANALYSIS PROCESS

R is an open-source software environment for statistical computing and graphics. We will be using it to perform EDA and descriptive analysis. Further, we will be using dplyr package for data manipulation and ggplot2 package to produce visualizations.

2. DATA

2.1. DESCRIPTION OF DATA SAMPLE

- **Data Structure** – There are in total 5 columns and 53698 rows in the dataset.
- **Dimensionality** – There are total 5 dimensions in the dataset namely *date*, *patient_id*, *type*, *patient_date_created*, and *score* out of which only 3 are significant (*date*, *patient_id*, and *score*). We can skip *type* column as it has low cardinality (zero-variance) and will not contribute to our analysis.
- **Quality** – Data is consistent and reliable. The combination of Patient ID, date, and patient_date_created uniquely identifies each row. There are no duplicates, missing values, or junk data in the dataset.

2.2. DESCRIPTION OF VARIABLES

- **Ordinal variable** – score.
- **Nominal Variable** – patient_id, date, type, patient_date_created
- **Target (Y) Variable** – score is our target variable with discrete continuous set of values. It is the summation of all GAD-7 assessment scores.
- **Variable of interest** – Our focus will be on patients with a GAD-7 score of 10 or more. We will look if there are significant no. of patients who are in *Moderate* or *Severe* conditions.

2.3. DESCRIPTIVE STATISTICS & EDA

- **Basic Descriptive Statistics**

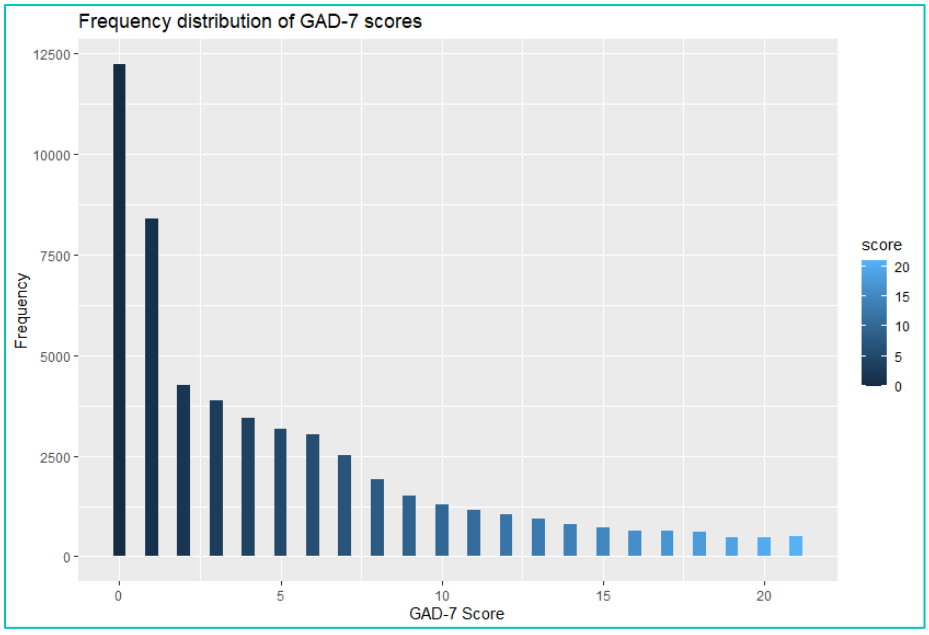
As we can see the mean of patient_id is almost equal to its median which implies the data is normally distributed. However, the mean of our target variable (score) is slightly greater than its median. This means the distribution of score data is slightly skewed towards right.

```
     date              patient_id       type              patient_date_created      score
Length:53698       Min.   :    0   Length:53698       Length:53698         Min.   : 0.000
Class :character   1st Qu.: 4697   Class :character   Class :character     1st Qu.: 1.000
Mode  :character   Median : 9228   Mode  :character   Mode  :character     Median : 3.000
                   Mean   : 9223                                           Mean   : 4.785
                   3rd Qu.:13809                                           3rd Qu.: 7.000
                   Max.   :18400                                           Max.   :21.000
```

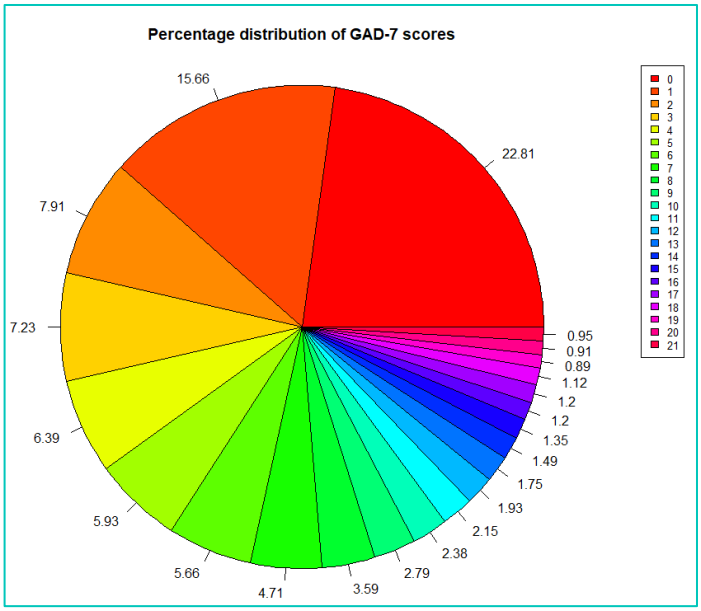
- **Variable Association**

Comment – Due to low number of numerical variables, we cannot interpret variable association.

- **Understand Target variable (score)**



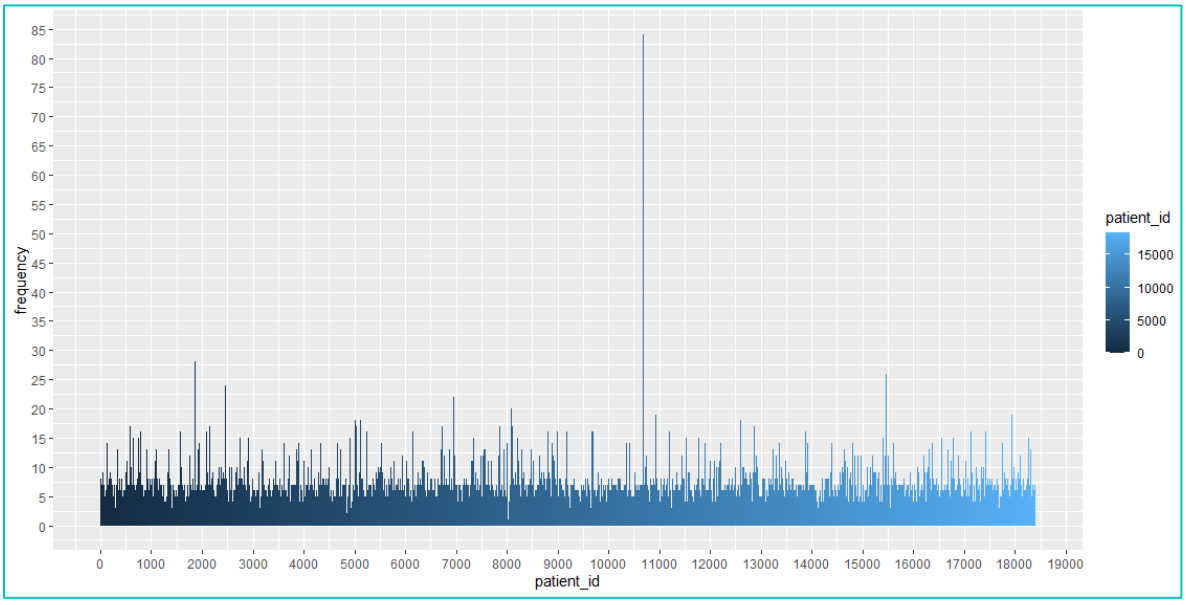
Comment - From the barplot, it looks like majority of patients have 0 or 1 GAD-7 score. Also, no. of patients decreases as the GAD-7 score increases.



Comment – 0 GAD-7 score has the highest percentage proportion with 22.81% and 19 GAD-7 score is the least with 0.89% share.

- **A holistic view on patient_id variable**

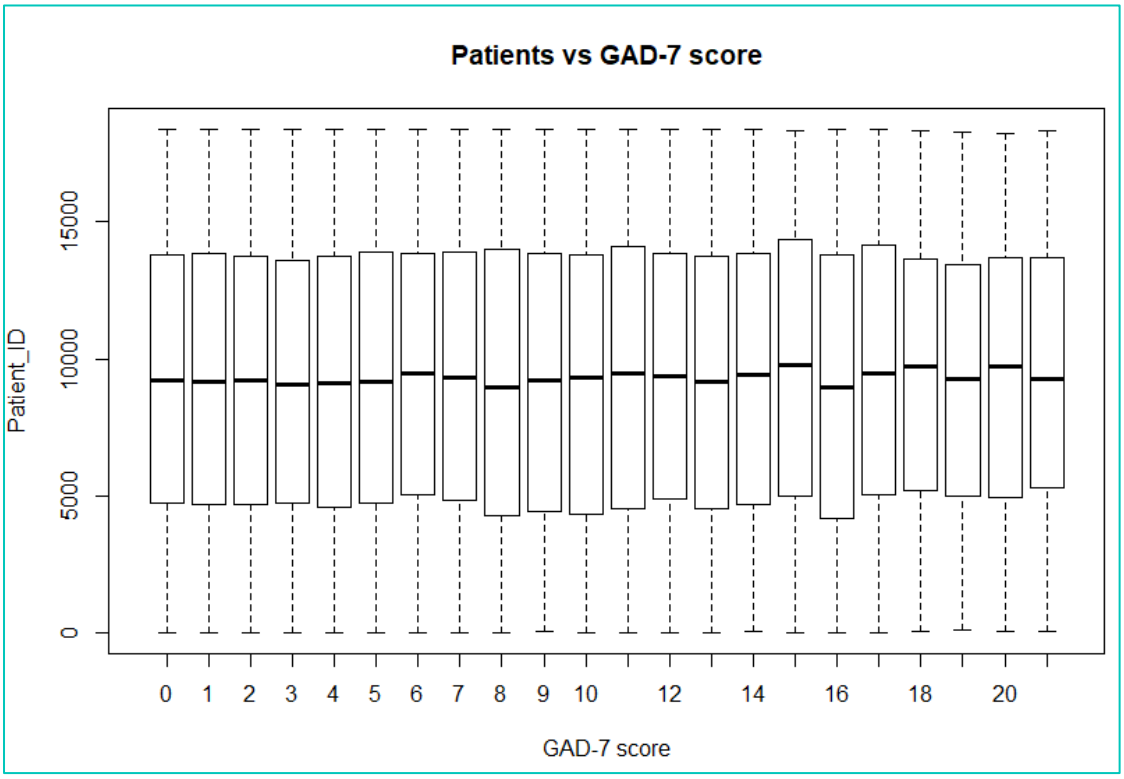
Frequency distribution of patients wrt no. of assessments



patient_id	frequency
10687	84
6574	42
12307	38
13855	34
13874	34
14779	33
11538	32
1939	31
1867	28
5670	28
15464	26
2462	24
4576	22
6958	22

Comment – We can see patient no. 10687 has the highest no. of assessments followed by 6574, 12307,13855, 13874, 14779, etc.

Boxplot of patient_id vs GAD-7 score



Comment – From the boxplot, we can see all GAD-7 scores are similarly distributed among the patients. This also means GAD-7 scores are uniform among patients with almost same range.

- **Monthly creation and assessments**

Monthly Creation Frequency

created_date	patient_count
2019-06	1616
2019-07	1686
2019-08	5100
2019-09	3739
2019-10	4273
2019-11	4870
2019-12	5505
2020-01	6152
2020-02	7323
2020-03	5844
2020-04	3843
2020-05	1508
2020-06	1446
2020-07	793

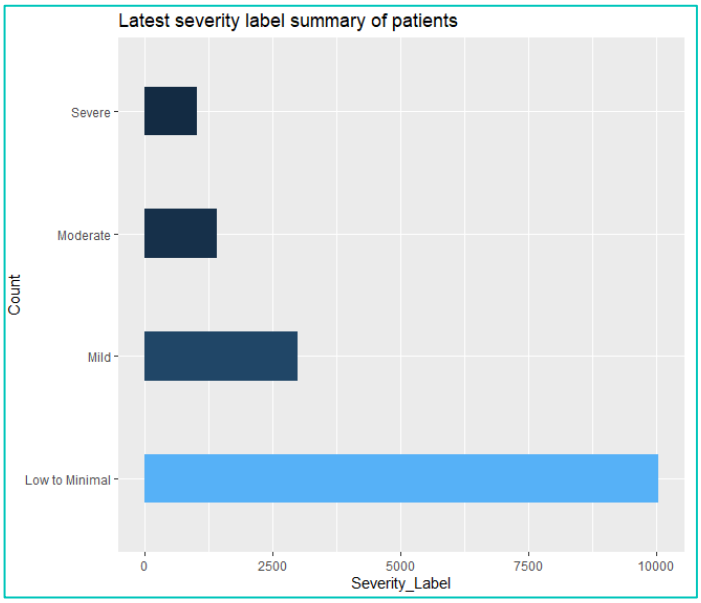
Comment – We cannot see a relevant pattern in the monthly creation frequency which makes it to hard infer anything from it.

Monthly Assessment Frequency

meas_date	patient_count
2019-06	177
2019-07	471
2019-08	1067
2019-09	1950
2019-10	2744
2019-11	3234
2019-12	3814
2020-01	4953
2020-02	5304
2020-03	6158
2020-04	7072
2020-05	5797
2020-06	5420
2020-07	5537

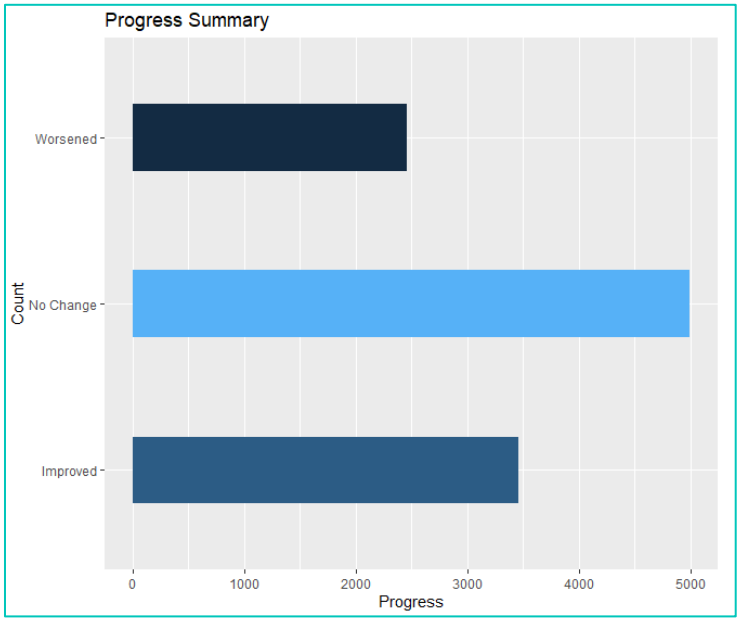
Comment – We can see significant increase in the no. of assessments with time. This is a good sign as it shows people are more open to assessments now as they were before.

- **Summary view of latest severity label**



Comment – As per the latest assessment of each patient, majority of them are in *low to minimal* status. There are almost 1000 patients which are in *severe* condition and need immediate care. Almost 1500 patients are in *moderate* status and are potential risk to severe condition. They need necessary therapies.

- **Summary view of Progress report**



Comment – As per each patient's latest assessment, almost 5000 patients have no impact of therapy, the condition of almost 2500 patients worsened, and almost 3500 patients have positive impact of therapy. Also, almost 4500 patients are redundant for this analysis as their both initial and latest assessment score is 0.

3. ANALYSIS RESULTS

3.1.1. ADDRESSING BUSINESS PROBLEM

Our analysis shows promising result and exposes few important inferences.

- GAD-7 scores are uniform among patients with almost same range.
- There is a significant increase in the no. of assessments with time. This is a good sign as it shows people are more open to assessments now as they were before.
- Majority of people are in *low to minimal* status yet there is still significant no. of people who need immediate care.
- Almost 5000 patients have no impact of therapy. These patients need considerate consultation and therapies.
- The condition of almost 2500 patients worsened. These patients need extensive therapies and frequent help.
- Almost 4500 patients are redundant for this analysis as their both initial and latest assessment score is 0. There is a possibility that these patients are not still open to revealing their anxiety issues. They need considerate consultation from therapists.

3.1.2. RECOMMENDATIONS

Some additional pieces of information which could have been important to collect are listed below:

- ✓ The data is quite generic and needs more important parameters like age, gender, location, relationship status, etc. which could have helped us in detailed analysis like decision tree and clustering.
- ✓ The date and time of the therapy sessions should also be recorded.

4. REFERENCES

Online Sources

Generalised Anxiety Disorder Assessment (GAD-7), https://patient.info/doctor/generalised-anxiety-disorder-assessment-gad-7

Robert L. Spitzer, MD; Kurt Kroenke, MD; Janet B. W. Williams, DSW; et al, A Brief Measure for Assessing Generalized Anxiety Disorder The GAD-7, 22 May 2006, https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/410326

Images Source

https://www.neuroflow.com/digital-behavioral-health/