MENTAL HEALTH ANALYSIS

Data Visualization
(CSE3020)

E2 Slot Fall Semester 2020-21

J Component Project Report

Submitted by

TRIPTI MISHRA (18BCE0888)

Under the guidance of

RAMANI S

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

October 2020

INTRODUCTION

1.1.Introduction to the area of study

Today the need of the medical services has grown exponentially. As many technologies have come into existence the level of pressure has been increased on the people. Everyone has to be better than other only then he can survive in this competing world. Due to this everyone is in great stress, the stress to prove oneself, the stress to become better in its own field, the stress of growing better with each day. So, wherever we can see we have seen people fighting to become better. Due to this fighting spirit every individual is going through a lot of stress and tension. Hence, this tension ultimately ruins the mental health of the person. Mental health is an important aspect of everyone's life. Suicidal behaviour is a leading cause of injury and death worldwide. Information about the epidemiology of such behaviour is important for policy-making and prevention [4]. Suicide is the third-leading cause of death in young people aged 10–19 years in the U.S. and represents a worldwide public health problem [6]. But this field goes unnoticed by the people. Therefore, I have chosen this field to study.

1.2.Relevance to practical field

Mental health has become the matter of concern in today's world. Everyone is running a race and this race to achieve the best is mentally exhausting. Mental health includes our emotional, psychological, and social well-being. It affects how we think, feel, and act. It also helps determine how we handle stress, relate to others, and make choices. Mental health is important at every stage of life, from childhood and adolescence through adulthood. Loneliness is increasingly becoming a major concern in modern Western societies. Severe loneliness (reporting feeling lonely 'almost all of the time' or 'most of the time') affects 6% of the adults in the UK [3]. The number of suicides has increased significantly in the past few years. More than 1.39 lakh Indians have committed suicide in the year 2019 and 67% of them are from the age group of 18-45. The most common reason for the suicides were drug abuse, mental illness and family problems. Hence, I have tried to visualize this thing in python as well as tableau for better understanding and what are the major reasons for the suicide. People don't see mental health as the medical issue hence, efforts are not even made to make sure the person is being taken care of. Mental health must be taken care in the same way as any other disease s being diagnosed.

1.3.Importance of the study proposed

As the number of suicide rates are increasing it is very essential to know what is the reason for it. Also, with the data visualization everything can be understood as easily as possible. Studies have shown human brain understands easily with graphs and pictures as compared to text. Our eyes are tempted to visualization. Many people have mental health concerns from time to time. But a mental health concern becomes a mental illness when ongoing signs and symptoms cause frequent stress and affect your ability to function. In today's world problems related to mental health are increasing each day because everyone has something which bothers them and this affects them mentally and ultimately leads to poor mental condition of the person. The days when people used to talk and share their problems have been lost long back and now a days there is only an atmosphere of sadness and stress everywhere which depreciates the mental condition of an individual. Mental health issues are increasingly prevalent among North American post-secondary students and often impede academic progress. However, students appear reluctant to seek help and access mental health services due to stigma associated with mental health issues [2]. In this project I have used two tools for the visualization. One is static which is done using python and other one is dynamic which is done using tableau. If we know what cost the mental health and what are the reasons for the increasing suicide rate and we can stop it from happening. Once the cause is known then efforts can be made to stop it. People must pay close attention to the mental health.

BRIEF BACKGROUND

2.1. Brief background of the problem

According to WHO Mental health is a state of well-being in which an individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively, and is able to make a contribution to his or her community. The WHO stress that mental health is "more than just the absence of mental disorders or disabilities." Peak mental health is about not only avoiding active conditions but also looking after ongoing wellness and happiness. In the United States, the National Alliance on Mental Illness estimate that almost 1 in 5 adults experience mental health problems each year. As the mental health gets deteriorated and the person is not able to suffer the pain or if he/she finds that he cannot do anything or nothing is going right in his life he decides to take his life. So ultimately poor mental health results in suicides. These two concepts are interrelated but people don't take the mental health so seriously. India reported an average 381 deaths by suicide daily in 2019, totalling 1,39,123 fatalities over the year, according to the latest National Crime Records Bureau (NCRB) data. The rate of suicide (incidents per 1 lakh population) rose by 0.2 per cent in 2019 over 2018, as per the data. According to the statistics by the NCRB, which functions under the Union Home Ministry, the suicide rate in cities (13.9 per cent) was higher as compared to all-India suicide rate (10.4 per cent) in 2019. Suicide by "hanging" (53.6 per cent), "consuming poison" (25.8 per cent), "drowning" (5.2 per cent) and "self-immolation" (3.8 per cent) were the prominent means of committing suicides during the year, the data showed. As we can see first the person's mental health is affected in the above scenario and then the person attempts for the suicide. Hence it is very necessary to know that the person really needs treatment or not. Because sometime due to some trivial tension the person tells that he is not having a good mental health. So first we must educate the people what exactly is mental health and what are the things to be done when people are going through a difficult phase. Also, other well-known countries have started taking mental health as a serious issue but in India still the person see the mental health as if the person has gone mad. Right steps must be taken to educate the people regarding this issue and the people who really needs support must be extended the hands. More than 450 million people suffer from mental disorders. The suicide rate in India is 10.3. In the last three decades, the suicide rate has increased by 43% but the male female ratio has been stable at 1.4: 1. Majority (71%) of suicide in India are by persons below the age of 44 years which imposes a huge social, emotional and economic burden [8]. According to WHO, by the year 2020, depression will constitute the second largest disease burden worldwide (Murray & Lopez, 1996). Global burden of mental health will be well beyond the treatment capacities of developed and developing countries. The social and economic costs associated with growing burden of mental ill health focused the possibilities for promoting mental health as well as preventing and treating mental illness. Thus, the Mental Health is linked to behaviour and seen as fundamental to physical health and quality of life. Most studies agree that suicide is closely linked to mental disorders. About 90% of people who commit suicide have suffered from at least one mental disorder. Mental disorders are found to contribute between 47 and 74% of suicide risk. Affective disorder is the disorder most frequently found in this context. Criteria for depression were found in 50–65% of suicide cases, more often among females than males. Substance abuse, and more specifically alcohol misuse, is also strongly associated with suicide risk, especially in older adolescents and males [9]. Each year over 30,000 people in the United States and approximately 1 million people worldwide die by suicide, making it one of the leading causes of death. A recent report from the Institute of Medicine (National Academy of Sciences) estimated that in the United States the value of lost productivity due to suicide is $11.8 billion per year [4].

1. Physical health and mental health are closely associated and it is proved beyond doubt that depression leads to heart and vascular diseases
2. Mental disorders also affect persons health behaviour like eating sensibly, regular exercise, adequate sleep, engaging in safe sexual practices, alcohol and tobacco use, adhering to medical therapies thus increasing the risk of physical illness.
3. Mental ill health also leads to social problems like unemployment, broken families, poverty, drug abuse and related crime.
4. Poor mental health plays a significant role in diminished immune functioning.
5. Medically ill patients with depression have worse outcome than those without.
6. Chronic illnesses like diabetes, cancer, heart disease increases the risk of depression.

WHO supports governments in the goal of strengthening and promoting mental health. WHO has evaluated evidence for promoting mental health and is working with governments to disseminate this information and to integrate the effective strategies into policies and plans. Early childhood interventions (e.g. home visits for pregnant women, pre-school psycho-social activities, combined nutritional and psycho-social help for disadvantaged populations)

1. support to children (e.g. skills building programmes, child and youth development programmes)
2. socio-economic empowerment of women (e.g. improving access to education and micro credit schemes)
3. social support for elderly populations (e.g. befriending initiatives, community and day centres for the aged)
4. programmes targeted at vulnerable groups, including minorities, indigenous people, migrants and people affected by conflicts and disasters (e.g. psycho-social interventions after disasters)
5. mental health promotional activities in schools (e.g. programmes supporting ecological changes in schools and child-friendly schools)
6. mental health interventions at work (e.g. stress prevention programmes)
7. housing policies (e.g. housing improvement)
8. Violence prevention programmes (e.g. community policing initiatives); and community development programmes (e.g. 'Communities That Care' initiatives, integrated rural development)

2.2. Earlier works / studies on this type of problem and the lead points for taking the present work

We have an organization named as National Alliance on Mental Illness (NAMI) which is helping people in their difficult phase of time where they are fighting the battle with themselves. They have various centres all over the world. As we can see there are very few organizations which are specifically operating on the mental health campaign. The population especially in India are not much aware of the mental health fact. The existing methodology fails in creating and understanding of the issues of mental health and only aims at predicting whether a person is mentally fit or not. The model is extremely useful in case a person wishes to know about his individual mental health but the project fails at producing the output for the mental health analysis which is the main concern of the data visualization field where the analysis of anything is extremely necessary. This not only throws light on the cause of the problem but also deals with the cause, the effect and the solution. In general, mental illness stigma was described as a barrier to building friendships and advancing friendships beyond a “surface-level,” leading students to feel isolated, disconnected, alienated, and excluded. Consequently, there was no place for students with mental illness to belong because they were unable to be authentic [2]. The dataset will be analysed thoroughly and the final outcome will be presented in this project.

Current issues in the mental health field are:

- Lack of support from the people
- Less number of people talking about these issues
- Lack of awareness among people regarding this problem
- Still considered as a taboo by a lot of people

The above points were the leading points for seeing whether people can be made educated by using visuals. Data visualization field helps to visualize even very complicated things to laymen’s language. As India is considered to be a Developing country many people must be educated about the mental health. Hence by using graphs, pictures message can be easily conveyed. My model will detect whether a person needs a mental health care and what are the reasons for the poor mental health.

2.3. Objectives

According to the World Health Organization and World Economic Forum, mental illness is the largest economic burden the world has today. Only about half the people in developed countries struggling with mental illness get the help they need, while in developing countries 90% go without any form of treatment. The stigma against mental illness and lack of mental health awareness worsens already existing mental health issues. With increases mental health awareness comes more support and care for those that need it. Currently, the public’s opinion on mental issues have a negative impact on those dealing with them, but the public has the power to make a positive impact on the society. When people are educated on mental illness, their effect and how often they occur, it overall lessens the bad reputation mental illness issues already have. Mental health is just as important as physical health and, in some aspects, it is more important because we can’t keep our physical health good without a healthy mental capacity. Objectives of any project implies the thought process of the person who is working on the project and what he wants to bring out of it. I wanted to study and analyse the cause of poor mental health in today’s generation where everyone is in such a hurry to move ahead in life. The results which have been presented by different institutes and health organizations will be taken and then an analysis will be made up because data visualization is all about how a person presents his idea to a lot of people in a different yet interesting way such that the idea reaches the people in the best possible and true way without being misinterpreted and misunderstood. Mental health and physical health are intertwined. One can rarely go without the other or in other words we can say they feed each other. Especially for children, mental health decided how we can move throughout the world. If affects how we think, act and feel about the world around us. If a child has some mental health issues that goes untreated then it can shape the child lives for the rest of their life. A person’s mental health is going to choose whether a person lives a healthy or an unhealthy life. For example, we can take that a person is having a lot stress and his mental health is worsening each day because of the negative thoughts and actions. Now we can think of it as the people around him notice that he is not behaving like a mentally fir person should and they take him to the hospital and get him treated. In this case the person’s chances of getting better increases but if we consider the scenario where people ignore the person’s change in behaviour and do nothing about it then in this case his mental condition worsens and there is a huge chance that he might not become normal or might lose his mental health more. Often mental health is related to madness which is an entirely wrong concept and, in this project, we aim to normalize the talking about mental health and its issues so that people come forward to talk about it more and more and people tend to become more fit mentally.

We can thus summarize the objectives of the project into the following points:

- Normalize the talking of mental health issues among people
- Study the statistics of mental health condition in India
- Come up with ways how mental health conditions can be improved
- Visualize the pattern of mental health in the past years.

2.4. Identification and exact definition of the problem

The project title is Mental Health Analysis. The project basically aims to create a better visualization of the mental health causes and thereby focus on what leads to poor mental health. This issue has long been not talked about in our society and we must normalize its thinking and therefore this project will help us learn more about the declining mental health condition in the society. Mental illness, also called mental health disorders, refers to a wide range of mental health conditions — disorders that affect our mood, thinking and behaviour. Examples of mental illness include depression, anxiety disorders, schizophrenia, eating disorders and addictive behaviours. Many people have mental health concerns from time to time. But a mental health concern becomes a mental illness when ongoing signs and symptoms cause frequent stress and affect your ability to function. A mental illness can make you miserable and can cause problems in your daily life, such as at school or work or in relationships. In most cases, symptoms can be managed with a combination of medications and talk therapy (psychotherapy). In today’s world problems related to mental health are increasing each day because everyone has something which bothers them and this affects them mentally and ultimately leads to poor mental condition of the person. The days when people used to talk and share their problems have been lost long back and now a days there is only an atmosphere of sadness and stress everywhere which depreciates the mental condition of an individual. The project deals with the analysis of mental health because in these days of extreme tension and stress cases of depression and poor mental health has become a very common thing. The analysis performed by us aims to throw light on some of the major causes of mental illness and also the statistics of how many people are under this light.

METHODOLOGY

3.1. Methodology Adopted

The issue of mental health is still considered as a taboo in our society and people are not keen to talk more about it. The fact is that mental health for any person is as important as physical health because if a person is mentally unwell, he will not be able to take care of own physical self. In this fast-moving world where everyone wants to succeed and gain something in the least span of time, poor mental health has become a common thing. This is a major reason that mental health is still not included and talked about freely in today’s society and people don’t want to talk more about it and learn the causes and result of it. Due to this reason mental health analysis is not done as much as other physical issues are dealt with. This is a huge loophole which must be filled out in order to build a better society where we live in. The idea we as a group are aiming at is to use the concept of mental health and normalize the way people are thinking of it. Existing methodology is the simple method of studying the mental health charts and coming out with some conclusion. This is almost the same technique which we will try to incorporate into in project but I have tried to use different types of dataset and thereby create a variety of results in my project which will focus on a lot of properties. Existing methodologies deal with the fact that we can solve this rising issue with a little part of our contribution and trying to build a healthy lifestyle but what it fails to talk about is the necessity of talking to people and trying to build a healthy mindset where stress and worries don't pop up immediately. Visualization techniques as far as observed by our team, we came up to the conclusion that there are several mental health predictor and analyser available in the internet which has the aim of telling how likely a person is of having a bad mental health issue. This result is driven by several factors which will be discussed further in this project and also dealt with in detail. The fault in this model is that it does not analyse the mental health of a particular age group or a particular region wherein the person can say this is the major cause of his poor mental health. Studying of different factors is equally important and must be considered while making any analysis project where the topic is dealt with in detail and a detailed analysis of the mental health issue is presented and discussed with in great detail. The field of data visualization has made great progress and with advancing time new methods need to be put into practise to come up with better results. I will be using the static as well as the dynamic visualization. The static visualization will be done using python and the dynamic one will be done using tableau. The existing system just talks that everyone must talk about the mental health but the part where they lack is that what are the major cause for the deteriorating mental health and how can we tackle that. Now a days no one is interested in reading long paragraphs and seeing what is the situation in the country. Instead they require visualizes for better understanding of the topic. Hence data visualization is a major tool for changing how the people think. And with the advancements in the technology now people can see how other things have been affected over the years. With the help of machine learning algorithms, the data will be trained and accordingly we will get a csv file which will state whether a person needs to have a treatment or not. So, the dynamic as well as static visualization is used in this project for better understanding.

3.2. Data Collection

In this project I have used two datasets. One dataset focuses on the data related to the mental health. The dataset consists of 27 attributes and 1259 tuples. The attributes are timestamp, age, gender, country, state, self_employed, family_history, treatment, work_interfere, no_employees, remote_work, tech_company, benefits, care_options, wellness_program, seek_help, anonymity, leave, mental_health_consequence, phys_health_consequence, coworkers, supervisor, mental_health_interview, phys_health_interview, mental_vs_physical, obs_consequence and comments.

- Timestamp is the time the survey was submitted. It indicates the time in the format of date followed by time.
- Age field tells about the age of the person
- Gender field tells about the gender of the person.
- Country field tells about the country the person belongs to.
- State field tells about the state the person lives in. •
- Self_employed tells whether the person is self-employed or not. They can fill this column value with either yes or no.
- Family_history tells whether the person has any family history of mental illness or not.
- Treatment field fills the value of whether the person has sought any treatment for mental health condition or not.

- Work_interfere deals with the query if in case the person is having any sort of mental illness then do they feel that it interferes with their work or not.
- No_employees seek the input of the number of employees the organization have in which the person works.
- Remote_work focuses on the query that whether the person works outside an office for atleast 50% of the time.
- Tech_company field takes in the information whether the person works in a tec company or not.
- Benefits deals with the query that whether the employer provides mental health facilities or not.
- Care_options field works when in case the employer provies mental health facilities then is the person aware of the fact or not.
- Wellness_program tells whether the employer has ever discussed mental health as part of employer wellness program.
- Seek_help is the field which asks whether the employer provides resources to learn more about mental health issues and how to seek help.
- Anonymity deals with the query of whether the anonymity of the person is protected if they choose to take advantage of mental health or substance abuse treatment.
- Leave field asks the person that how easy it is for him to take a medical leave for a mental health condition.
- Mental_health_consequence asks the person that if they discuss their mental health issues with their employer then will it have any type of negative consequences or not.
- Phys_health_consequence asks the person that if they discuss their physical health issues with their employer then will it have any type of negative consequences or not.
- Coworkers asks the person if they are willing to discuss a mental health issue with their coworkers or not
- Supervisor asks the person if they are willing to discuss a mental health issue with their supervisor directly or not
- Mental_health_interview asks the person if they are willing to bring up a mental health issue with a potential employer in an interview or not
- Phys_health_interview asks the person if they are willing to bring up a physical health issue with a potential employer in an interview or not
- Mental_vs_physical field asks the person if their employers take mental health as seriously as physical health or not
- Obs_consequence asks the person if they have heard of or observed any negative consequences for coworkers with mental health conditons in their working environment.
- Comments field is for any additional note or field.

TABLE 3.1-ATTRIBUTES OF DATASET OF MENTAL HEALTH

TimeStamp	Age	Gender	Country	State	Self_employed
Work_interface	No_employees	Remote_work	Benefits	Anonymity	Care_options
Family_history	Tech_company	Mental_health_conseque	Leave	Coworkers	Supervisor
Men Vs Phy	Physhealthinter	Mental_health_interview	Obs_consequence		

The next dataset which I have taken is regarding the suicides in India. Suicide is the poor mental health. Hence, by studying this we can see what are the major reasons for the suicide. This dataset contains 7 attributes and 237520 tuples. The attributes are State, Year, Type_code, Type, Gender, Age_group, Total.

- State mentions all the states in India along with the Union territories.
- Year tells us about when the incident took place.
- Type_code tells what is the type of the case. These includes causes, education_status, Means_adopted, Professional_profile, Sociall_status.
- Types include different ways in which the suicide was performed such as by hanging or by taking drugs, etc.
- Gender tells us about the gender of the person.
- Age_group field includes the age group of the person. The age group mentioned in the dataset are 0-14,15-49,30-44,45-59,60+.
- Total includes the sum of the total suicide cases in the respective states and also the total cases all over the country.

TABLE 3.2- ATTRIBUTES OF DATASET OF SUICIDES IN INDIA

State	Year	Type_code	Tyoe	Gender
Age_group	Total			

This is the descriptive view of the dataset which we will be using for my project and present an analysis on this dataset information. The information provided is quite enough with different types of information related to working environment of an individual which will aim at studying in a better way how organizations and companies are trying to boost up the working experience of the employees and make a better environment in the company for the people to work in. All of this analysis will be made in this project and presented using data visualization techniques and methods.

3.3. Experiments / Analytical Computations

In python all the computations were done. Tableau is user-friendly and hence it can be easily used. In the following images we can see all the python computations. Here I have used machine-learning algorithms. Firstly, I have compared most of the known ml algorithms and once giving the most accuracy was chosen was the model. The machine-learning algorithms which I have compared are Logistic regression with an accuracy of 80%, Treeclassifier with an accuracy of 81%, k-nearest neighbours with an accuracy of 80.4%, Boosting with an accuracy of 82%. As it is clearly visible that Boosting has the highest accuracy hence, I have used this for predicting the results.

3.3.1. Code:

```
import numpy as np # linear algebra

import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

import matplotlib.pyplot as plt

from matplotlib.patches import Polygon

import seaborn as sns
```

```
from scipy import stats
from scipy.stats import randint
# prep
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
from sklearn.datasets import make_classification
from sklearn.preprocessing import binarize, LabelEncoder, MinMaxScaler
# models
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import RandomizedSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, ExtraTreesClassifier
# Validation libraries
from sklearn import metrics
from sklearn.metrics import accuracy_score, mean_squared_error, precision_recall_curve
from sklearn.model_selection import cross_val_score
#Neural Network
from sklearn.neural_network import MLPClassifier
#from sklearn import grid_search
#from grid_search import RandomizedSearchCV
#Bagging
from sklearn.ensemble import BaggingClassifier, AdaBoostClassifier
from sklearn.neighbors import KNeighborsClassifier
#Naive bayes
from sklearn.naive_bayes import GaussianNB
#Stacking
from mlxtend.classifier import StackingClassifier
# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list the files in the input directory
# Any results you write to the current directory are saved as output.
```

```
#reading in CSV's from a file path
train_df = pd.read_csv('survey.csv')
#Pandas: whats the data row count?
print(train_df.shape)
#Pandas: whats the distribution of the data?
print(train_df.describe())
#Pandas: What types of data do i have?
print(train_df.info())
#dealing with missing data
#Let’s get rid of the variables "Timestamp",“comments”, “state” just to make our lives easier.
train_df = train_df.drop(['comments'], axis= 1)
train_df = train_df.drop(['state'], axis= 1)
train_df = train_df.drop(['Timestamp'], axis= 1)
train_df.isnull().sum().max() #just checking that there's no missing data missing...
train_df.head(5)
defaultInt = 0
defaultString = 'NaN'
defaultFloat = 0.0
# Create lists by data tpe
intFeatures = ['Age']
stringFeatures = ['Gender', 'Country', 'self_employed', 'family_history', 'treatment', 'work_interfere',
                 'no_employees', 'remote_work', 'tech_company', 'anonymity', 'leave', 'mental_health_consequence',
                 'phys_health_consequence', 'coworkers', 'supervisor', 'mental_health_interview', 'phys_health_interview',
                 'mental_vs_physical', 'obs_consequence', 'benefits', 'care_options', 'wellness_program',
                 'seek_help']
floatFeatures = []
# Clean the NaN's
for feature in train_df:
    if feature in intFeatures:
        train_df[feature] = train_df[feature].fillna(defaultInt)
    elif feature in stringFeatures:
        train_df[feature] = train_df[feature].fillna(defaultString)
    elif feature in floatFeatures:
        train_df[feature] = train_df[feature].fillna(defaultFloat)
    else:
        print('Error: Feature %s not recognized.' % feature)
train_df.head(5)
#clean 'Gender'
#Slower case all columm's elements
gender = train_df['Gender'].str.lower()
#print(gender)
#Select unique elements
gender = train_df['Gender'].unique()
#Made gender groups
male_str = ["male", "m", "male-ish", "maile", "mal", "male (cis)", "make", "male ", "man","msle", "mail", "malr","cis man", "Cis Male", "cis male"]
trans_str = ["trans-female", "something kinda male?", "queer/she/they", "non-binary","nah", "all", "enby", "fluid", "genderqueer", "androgyne", "agender", "male leaning androgynous", "guy (-ish) ^_^", "trans woman", "neuter", "female (trans)", "queer", "ostensibly male, unsure what that really means"]
female_str = ["cis female", "f", "female", "woman",  "femake", "female ","cis-female/femme", "female (cis)", "femail"]
for (row, col) in train_df.iterrows():


    if str.lower(col.Gender) in male_str:
        train_df['Gender'].replace(to_replace=col.Gender, value='male', inplace=True)


    if str.lower(col.Gender) in female_str:
        train_df['Gender'].replace(to_replace=col.Gender, value='female', inplace=True)


    if str.lower(col.Gender) in trans_str:
         train_df['Gender'].replace(to_replace=col.Gender, value='trans', inplace=True)
stk_list = ['A little about you', 'p']
```

```
train_df = train_df[~train_df['Gender'].isin(stk_list)]
print(train_df['Gender'].unique())
#complete missing age with mean
train_df['Age'].fillna(train_df['Age'].median(), inplace = True)

# Fill with media() values < 18 and > 120
s = pd.Series(train_df['Age'])
s[s<18] = train_df['Age'].median()
train_df['Age'] = s
s = pd.Series(train_df['Age'])
s[s>120] = train_df['Age'].median()
train_df['Age'] = s
#Ranges of Age
train_df['age_range'] = pd.cut(train_df['Age'], [0,20,30,65,100], labels=["0-20", "21-30", "31-65", "66-100"], include_lowest=True)
#There are only 0.014% of self employed so let's change NaN to NOT self_employed
#Replace "NaN" string from defaultString
train_df['self_employed'] = train_df['self_employed'].replace([defaultString], 'No')
print(train_df['self_employed'].unique())
#Replace "NaN" string from defaultString
train_df['work_interfere'] = train_df['work_interfere'].replace([defaultString], 'Don\'t know' )
print(train_df['work_interfere'].unique())
#Encoding data
labelDict = {}
for feature in train_df:
    le = preprocessing.LabelEncoder()
    le.fit(train_df[feature])
    le_name_mapping = dict(zip(le.classes_, le.transform(le.classes_)))
    train_df[feature] = le.transform(train_df[feature])
    # Get labels
    labelKey = 'label_' + feature
    labelValue = [*le_name_mapping]
    labelDict[labelKey] =labelValue
for key, value in labelDict.items():
    print(key, value)
#Get rid of 'Country'
train_df = train_df.drop(['Country'], axis= 1)
train_df.head()
#missing data
total = train_df.isnull().sum().sort_values(ascending=False)
percent = (train_df.isnull().sum()/train_df.isnull().count()).sort_values(ascending=False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_data.head(20)
print(missing_data)
#correlation matrix
corrmat = train_df.corr()
f, ax = plt.subplots(figsize=(12, 9))
sns.heatmap(corrmat, vmax=.8, square=True);
plt.show()
#treatment correlation matrix
k = 10 #number of variables for heatmap
cols = corrmat.nlargest(k, 'treatment')['treatment'].index
cm = np.corrcoef(train_df[cols].values.T)
sns.set(font_scale=1.25)
hm = sns.heatmap(cm, cbar=True, annot=True, square=True, fmt='.2f', annot_kws={'size': 10},
yticklabels=cols.values, xticklabels=cols.values)
plt.show()
# Distribiution and density by Age
plt.figure(figsize=(12,8))
sns.histplot(train_df["Age"], bins=24)
plt.title("Distribuition and density by Age")
plt.xlabel("Age")
#Separate by treatment or not
g = sns.FacetGrid(train_df, col='treatment', height=5)
```

```
g = g.map(sns.histplot, "Age")
# Let see how many people has been treated
plt.figure(figsize=(12,8))
labels = labelDict['label_Gender']
g = sns.countplot(x="treatment", data=train_df)
g.set_xticklabels(['Female','Male'])
plt.title('Total Distribuition by treated or not')
o = labelDict['label_age_range']

g = sns.catplot(x="age_range", y="treatment", hue="Gender", data=train_df, kind="bar",  ci=None, height=5, aspect=2, legend_out = True)
g.set_xticklabels(o)
plt.title('Probability of mental health condition')
plt.ylabel('Probability x 100')
plt.xlabel('Age')
# replace legend labels
new_labels = labelDict['label_Gender']
for t, l in zip(g._legend.texts, new_labels): t.set_text(l)
# Positioning the legend
g.fig.subplots_adjust(top=0.9,right=0.8)
plt.show()
o = labelDict['label_family_history']
g = sns.catplot(x="family_history", y="treatment", hue="Gender", data=train_df, kind="bar", ci=None, height=5, aspect=2, legend_out = True)
g.set_xticklabels(o)
plt.title('Probability of mental health condition')
plt.ylabel('Probability x 100')
plt.xlabel('Family History')
# replace legend labels
new_labels = labelDict['label_Gender']
for t, l in zip(g._legend.texts, new_labels): t.set_text(l)
# Positioning the legend
```

```
g.fig.subplots_adjust(top=0.9,right=0.8)
plt.show()
o = labelDict['label_care_options']
g = sns.catplot(x="care_options", y="treatment", hue="Gender", data=train_df, kind="bar", ci=None, height=5, aspect=2, legend_out = True)
g.set_xticklabels(o)
plt.title('Probability of mental health condition')
plt.ylabel('Probability x 100')
plt.xlabel('Care options')
# replace legend labels
new_labels = labelDict['label_Gender']
for t, l in zip(g._legend.texts, new_labels): t.set_text(l)

# Positioning the legend
g.fig.subplots_adjust(top=0.9,right=0.8)
plt.show()
o = labelDict['label_benefits']
g = sns.catplot(x="care_options", y="treatment", hue="Gender", data=train_df, kind="bar", ci=None, height=5, aspect=2, legend_out = True)
g.set_xticklabels(o)
plt.title('Probability of mental health condition')
plt.ylabel('Probability x 100')
plt.xlabel('Benefits')
# replace legend labels
new_labels = labelDict['label_Gender']
for t, l in zip(g._legend.texts, new_labels): t.set_text(l)
# Positioning the legend
g.fig.subplots_adjust(top=0.9,right=0.8)
plt.show()
o = labelDict['label_work_interfere']
g = sns.catplot(x="work_interfere", y="treatment", hue="Gender", data=train_df, kind="bar", ci=None, height=5, aspect=2, legend_out = True)
```

```
g.set_xticklabels(o)
plt.title('Probability of mental health condition')
plt.ylabel('Probability x 100')
plt.xlabel('Work interfere')
# replace legend labels
new_labels = labelDict['label_Gender']
for t, l in zip(g._legend.texts, new_labels): t.set_text(l)
# Positioning the legend
g.fig.subplots_adjust(top=0.9,right=0.8)
plt.show()
scaler = MinMaxScaler()
train_df['Age'] = scaler.fit_transform(train_df[['Age']])
train_df.head()
feature_cols = ['Age', 'Gender', 'family_history', 'benefits', 'care_options', 'anonymity', 'leave', 
'work_interfere']
X = train_df[feature_cols]
y = train_df.treatment
# split X and y into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=0)
# Create dictionaries for final graph
# Use: methodDict['Stacking'] = accuracy_score
methodDict = {}
rmseDict = ()
# Build a forest and compute the feature importances
forest = ExtraTreesClassifier(n_estimators=250,
                              random_state=0)

forest.fit(X, y)
importances = forest.feature_importances_
std = np.std([tree.feature_importances_ for tree in forest.estimators_],
             axis=0)
indices = np.argsort(importances)[::-1]
```

```
labels = []
for f in range(X.shape[1]):
    labels.append(feature_cols[f])
# Plot the feature importances of the forest
plt.figure(figsize=(12,8))
plt.title("Feature importances")
plt.bar(range(X.shape[1]), importances[indices],
       color="r", yerr=std[indices], align="center")
plt.xticks(range(X.shape[1]), labels, rotation='vertical')
plt.xlim([-1, X.shape[1]])
plt.show()
def evalClassModel(model, y_test, y_pred_class, plot=False):
    #Classification accuracy: percentage of correct predictions
    # calculate accuracy
    print('Accuracy:', metrics.accuracy_score(y_test, y_pred_class))

    #Null accuracy: accuracy that could be achieved by always predicting the most frequent class
    # examine the class distribution of the testing set (using a Pandas Series method)
    print('Null accuracy:\n', y_test.value_counts())
    # calculate the percentage of ones
    print('Percentage of ones:', y_test.mean())
    # calculate the percentage of zeros
    print('Percentage of zeros:',1 - y_test.mean())
    #Comparing the true and predicted response values
    print('True:', y_test.values[0:25])
    print('Pred:', y_pred_class[0:25])
    #Confusion matrix
    # save confusion matrix and slice into four pieces
    confusion = metrics.confusion_matrix(y_test, y_pred_class)
    #[row, column]
    TP = confusion[1, 1]
    TN = confusion[0, 0]
    FP = confusion[0, 1]
    FN = confusion[1, 0]
    # visualize Confusion Matrix
    sns.heatmap(confusion,annot=True,fmt="d")
    plt.title('Confusion Matrix')
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.show()
    #Metrics computed from a confusion matrix
    #Classification Accuracy: Overall, how often is the classifier correct?
    accuracy = metrics.accuracy_score(y_test, y_pred_class)
    print('Classification Accuracy:', accuracy)
    #Classification Error: Overall, how often is the classifier incorrect?
    print('Classification Error:', 1 - metrics.accuracy_score(y_test, y_pred_class))
    #False Positive Rate: When the actual value is negative, how often is the prediction incorrect?
    false_positive_rate = FP / float(TN + FP)
    print('False Positive Rate:', false_positive_rate)
    #Precision: When a positive value is predicted, how often is the prediction correct?
    print('Precision:', metrics.precision_score(y_test, y_pred_class))
    plt.show()
def logisticRegression():
    # train a logistic regression model on the training set
    logreg = LogisticRegression()
    logreg.fit(X_train, y_train)
    # make class predictions for the testing set
    y_pred_class = logreg.predict(X_test)
    print('########### Logistic Regression ###############'
    accuracy_score = evalClassModel(logreg, y_test, y_pred_class, True)
logisticRegression()
def treeClassifier():
    # Calculating the best parameters
    tree = DecisionTreeClassifier()
    # train a decision tree model on the training set
    tree = DecisionTreeClassifier(max_depth=3, min_samples_split=8, max_features=6, criterion='entropy', min_samples_leaf=7)
    tree.fit(X_train, y_train)

    # make class predictions for the testing set
    y_pred_class = tree.predict(X_test)

    print('########### Tree classifier ###############')

    accuracy_score = evalClassModel(tree, y_test, y_pred_class, True)
treeClassifier()
def Knn():
    # Calculating the best parameters
    knn = KNeighborsClassifier(n_neighbors=5)
    # define the parameter values that should be searched
    k_range = list(range(1, 31))
    weight_options = ['uniform', 'distance']

    # specify "parameter distributions" rather than a "parameter grid"
    param_dist = dict(n_neighbors=k_range, weights=weight_options)
    tuningRandomizedSearchCV(knn, param_dist)

    # train a KNeighborsClassifier model on the training set
    knn = KNeighborsClassifier(n_neighbors=27, weights='uniform')
    knn.fit(X_train, y_train)
    # make class predictions for the testing set
    y_pred_class = knn.predict(X_test)

    print('########### KNeighborsClassifier ###############')
    accuracy_score = evalClassModel(knn, y_test, y_pred_class, True)
    Knn()
    # Building and fitting
    clf = DecisionTreeClassifier(criterion='entropy', max_depth=1)
    boost = AdaBoostClassifier(base_estimator=clf, n_estimators=500)
    boost.fit(X_train, y_train)
    # make class predictions for the testing set
    y_pred_class = boost.predict(X_test
    print('########### Boosting ###############')
    accuracy_score = evalClassModel(boost, y_test, y_pred_class, True)
    #Data for final graph
    methodDict['Boosting'] = accuracy_score * 100
def plotSuccess():
    s = pd.Series(methodDict)
    s = s.sort_values(ascending=False)
    plt.figure(figsize=(12,8))
    #Colors
    ax = s.plot(kind='bar')
    for p in ax.patches:
        ax.annotate(str(round(p.get_height(),2)), (p.get_x() * 1.005, p.get_height() * 1.005))
    plt.ylim([70.0, 90.0])
    plt.xlabel('Method')
    plt.ylabel('Percentage')
    plt.title('Success of methods')
    plt.show()
# Generate predictions with the best method
clf = AdaBoostClassifier()
clf.fit(X, y)
dfTestPredictions = clf.predict(X_test)
# Write predictions to csv file
# We don't have any significative field so we save the index
```

```
results = pd.DataFrame({'Index': X_test.index, 'Treatment': dfTestPredictions})
# Save to file
# This file will be visible after publishing in the output section
results.to_csv('results1.csv', index=False)
results.head()


df = pd.read_csv("Suicides in India 2001-2012.csv")
df_suicide = df.loc[(df['Total'] != 0) & (~df['State'].isin(["Total (All India)", "Total (States)", "Total (Uts)"]))]
df_no_suicide = df.loc[(df['Total'] == 0) & (~df['State'].isin(["Total (All India)", "Total (States)", "Total (Uts)"]))]
df_suicide_state = df_suicide.loc[~df_suicide['State'].isin(["A & N Islands", "Chandigarh", "D & N Haveli", "Daman & Diu", "Delhi (Ut)", "Lakshadweep", "Puducherry"])]
df_suicide_state.info()
plt.subplots(figsize=(10,14))
sns.countplot(y= "State",data = df_suicide_state).set_title("Suicide Distribution by State")
plt.subplots(figsize=(10,14))
sns.countplot(y= "State", hue= "Gender" ,data = df_suicide_state, palette= "autumn").set_title("Gender Suicide Distribution by State")
plt.subplots(figsize=(20,20))
sns.countplot(y= "State", hue= "Age_group" ,data = df_suicide_state).set_title("State Suicide Distribution by Age-Group")
print(df_suicide_state.Gender.value_counts())
df_suicide_state.Gender.value_counts().plot(kind= "pie", autopct='%1.1f%%', shadow=True, figsize=(5,5))
plt.title("Gender Suicide Distribution of State")
sns.countplot(x = "Age_group", hue= "Gender", data= df_suicide_state, palette= "spring").set_title("Gender Suicide Distribution of State by Age-Group")
plt.subplots(figsize=(9,5))
sns.countplot(x = "Year", data= df_suicide_state, palette= "spring").set_title("Year Suicide Distribution of State")
print(df_suicide_state.Type_code.value_counts())
df_suicide_state.Type_code.value_counts().plot(kind= "pie", autopct='%1.1f%%', shadow=True, figsize=(7,7))
plt.title("Type-Code Distribution of State")
```

```
df_suicide_state_causes = df_suicide_state.loc[df_suicide_state["Type_code"]=='Causes']
plt.subplots(figsize=(15,12))
sns.countplot(y = "Type", data= df_suicide_state_causes).set_title("Type:'Causes' for Suicide in State")
df_suicide_state_causes = df_suicide_state.loc[df_suicide_state["Type_code"]=='Means_adopted']
plt.subplots(figsize=(15,8))
sns.countplot(y = "Type", data= df_suicide_state_causes).set_title("Type:'Means_adopted' for Suicide in State")
df_suicide_state_causes = df_suicide_state.loc[df_suicide_state["Type_code"]=='Social_Status']
plt.subplots(figsize=(10,2))
sns.countplot(y = "Type", data= df_suicide_state_causes).set_title("Type: 'Social_Status' for Suicide in State")
df_suicide_state_causes = df_suicide_state.loc[df_suicide_state["Type_code"]=='Education_Status']
plt.subplots(figsize=(10,2))
sns.countplot(y = "Type", data= df_suicide_state_causes).set_title("Type: 'Education_Status' for Suicide in State")
df_suicide_state_causes = df_suicide_state.loc[df_suicide_state["Type_code"]=='Professional_Profile']
plt.subplots(figsize=(15,5))
sns.countplot(y = "Type", data= df_suicide_state_causes).set_title("Type:'Professional_Profile' for Suicide in State")
df_suicide_ut = df_suicide.loc[df_suicide['State'].isin(["A & N Islands", "Chandigarh", "D & N Haveli", "Daman & Diu", "Delhi (Ut)", "Lakshadweep", "Puducherry"])]
sns.countplot(y= "State",data = df_suicide_ut).set_title("Suicide Distribution by UT")
sns.countplot(y= "State", hue= "Gender" ,data = df_suicide_ut).set_title("Gender Suicide Distribution by UT")
plt.subplots(figsize=(15,10))
sns.countplot(y= "State", hue= "Age_group" ,data = df_suicide_ut).set_title("UT Suicide Distribution by Age-Group")
print(df_suicide_ut.Gender.value_counts())
df_suicide_ut.Gender.value_counts().plot(kind= "pie", autopct='%1.1f%%', shadow=True, figsize=(5,5))
plt.title("Gender Suicide Distribution of UT")
sns.countplot(x = "Age_group", hue= "Gender", data= df_suicide_ut, palette= "spring").set_title("Gender Suicide Distribution of UT by Age-Group")
```

```
plt.subplots(figsize=(9,5))

sns.countplot(x = "Year", data= df_suicide_ut, palette= "spring").set_title("Year Suicide Distribution of UT")

print(df_suicide_ut.Type_code.value_counts())

df_suicide_ut.Type_code.value_counts().plot(kind= "pie", autopct='%1.1f%%', shadow=True, figsize=(7,7))

plt.title("Type-Code Distribution of UT")

df_suicide_ut_causes = df_suicide_ut.loc[df_suicide_ut["Type_code"]=='Causes']

plt.subplots(figsize=(15,12))

sns.countplot(y = "Type", data= df_suicide_ut_causes).set_title("Type:'Causes' for Suicide in UT")

df_suicide_ut_causes = df_suicide_ut.loc[df_suicide_ut["Type_code"]=='Means_adopted']

plt.subplots(figsize=(15,8))

sns.countplot(y = "Type", data= df_suicide_ut_causes).set_title("Type:'Means_adopted' for Suicide in UT")

df_suicide_ut_causes = df_suicide_ut.loc[df_suicide_ut["Type_code"]=='Social_Status']

plt.subplots(figsize=(10,2))

sns.countplot(y = "Type", data= df_suicide_ut_causes).set_title("Type: 'Social_Status' for Suicide in UT")

df_suicide_ut_causes = df_suicide_ut.loc[df_suicide_ut["Type_code"]=='Education_Status']

plt.subplots(figsize=(10,2))

sns.countplot(y = "Type", data= df_suicide_ut_causes).set_title("Type: 'Education_Status' for Suicide in UT")

df_suicide_ut_causes = df_suicide_ut.loc[df_suicide_ut["Type_code"]=='Professional_Profile']

plt.subplots(figsize=(15,5))

sns.countplot(y = "Type", data= df_suicide_ut_causes).set_title("Type:'Professional_Profile' for Suicide in UT")
```

3.4. Tools Used

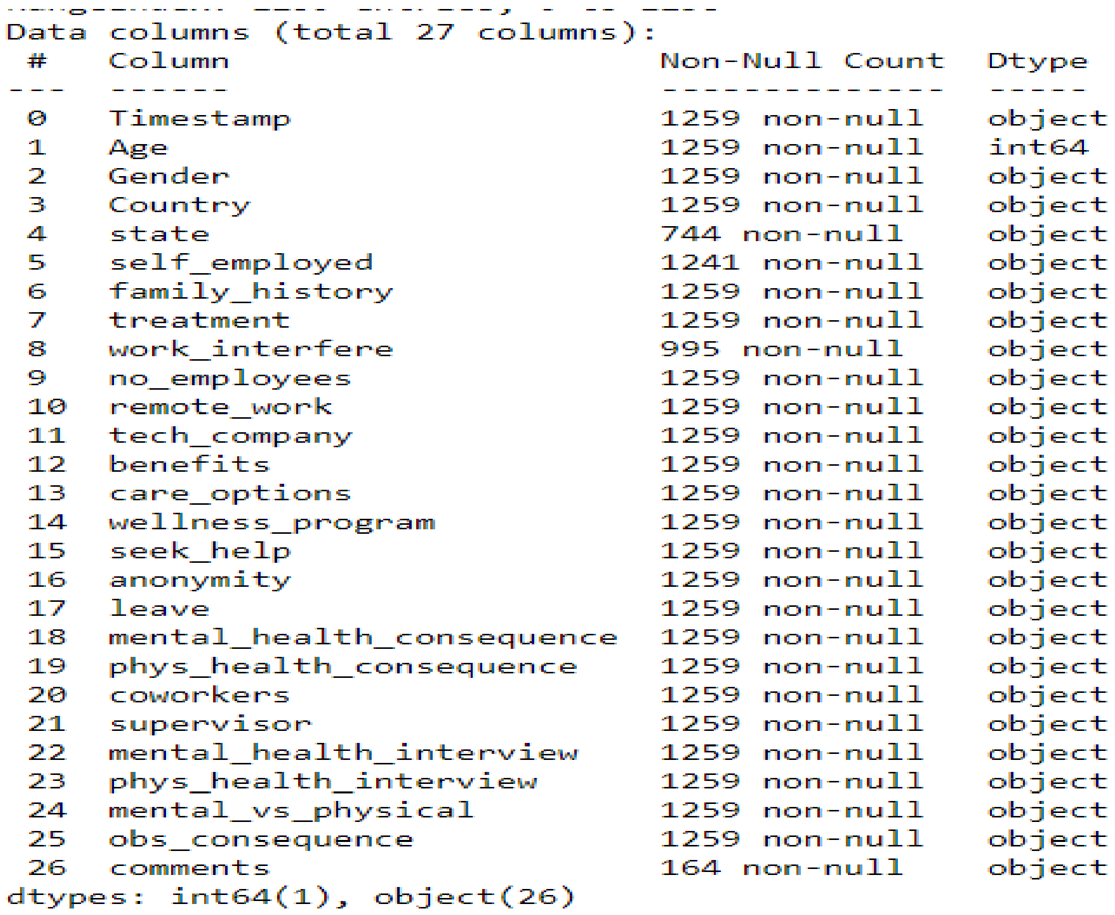
In this project I will be using both dynamic as well as static visualization. Hence, I have used two tools here namely jupyter notebook and tableau. For the dynamic visualization tableau is used and for the static one python is used. Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data into the very easily understandable format. Python is very much useful in plotting static visualizes. There are many libraries in python with is solely for this purpose. Hence, following tools are used in the project:

- Tableau
- Jupyter notebook

RESULTS

4.1. Results

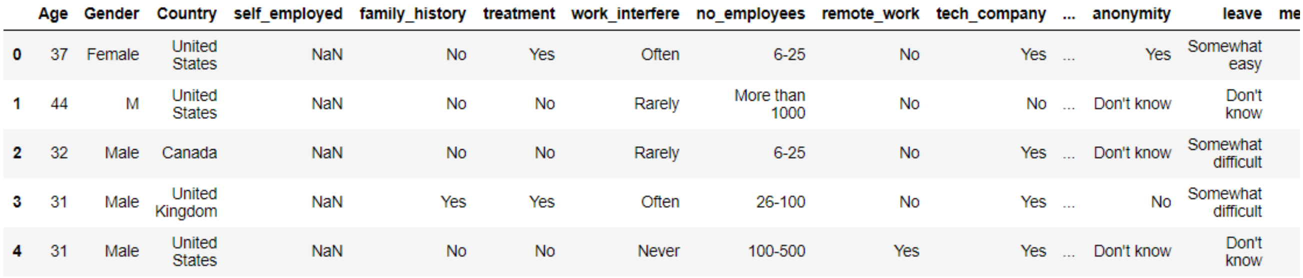
Following visualizations were made with the help of python and tableau. Various graphs were plotted which will be useful in getting useful conclusions.



```
Data columns (total 27 columns):
 #   Column                     Non-Null Count  Dtype 
---  ------                     --------------  ----- 
 0   Timestamp                  1259 non-null   object
 1   Age                        1259 non-null   int64 
 2   Gender                     1259 non-null   object
 3   Country                    1259 non-null   object
 4   state                      744 non-null    object
 5   self_employed              1241 non-null   object
 6   family_history             1259 non-null   object
 7   treatment                  1259 non-null   object
 8   work_interfere             995 non-null    object
 9   no_employees               1259 non-null   object
 10  remote_work                1259 non-null   object
 11  tech_company               1259 non-null   object
 12  benefits                   1259 non-null   object
 13  care_options               1259 non-null   object
 14  wellness_program           1259 non-null   object
 15  seek_help                  1259 non-null   object
 16  anonymity                  1259 non-null   object
 17  leave                      1259 non-null   object
 18  mental_health_consequence  1259 non-null   object
 19  phys_health_consequence    1259 non-null   object
 20  coworkers                  1259 non-null   object
 21  supervisor                 1259 non-null   object
 22  mental_health_interview    1259 non-null   object
 23  phys_health_interview      1259 non-null   object
 24  mental_vs_physical         1259 non-null   object
 25  obs_consequence            1259 non-null   object
 26  comments                   164 non-null    object
dtypes: int64(1), object(26)
```

Fig. 4.1 Study Area

Here we can find the summary about the dataset which we have considered for study. This dataset is about mental health analysis. It consists of 27 columns and 1024 tuples of data.



	Age	Gender	Country	self_employed	family_history	treatment	work_interfere	no_employees	remote_work	tech_company	...	anonymity	leave	me
0	37	Female	United States	NaN	No	Yes	Often	6-25	No	Yes	...	Yes	Somewhat easy	
1	44	M	United States	NaN	No	No	Rarely	More than 1000	No	No	...	Don't know	Don't know	
2	32	Male	Canada	NaN	No	No	Rarely	6-25	No	Yes	...	Don't know	Somewhat difficult	
3	31	Male	United Kingdom	NaN	Yes	Yes	Often	26-100	No	Yes	...	No	Somewhat difficult	
4	31	Male	United States	NaN	No	No	Never	100-500	Yes	Yes	...	Don't know	Don't know	

Fig. 4.2 Final dataset

After refining the dataset, we will get the above dataset. Here all the missing values of integer type are filled with the average of the respective columns and if the datatype is string then we can fill the values such as no or sometimes.

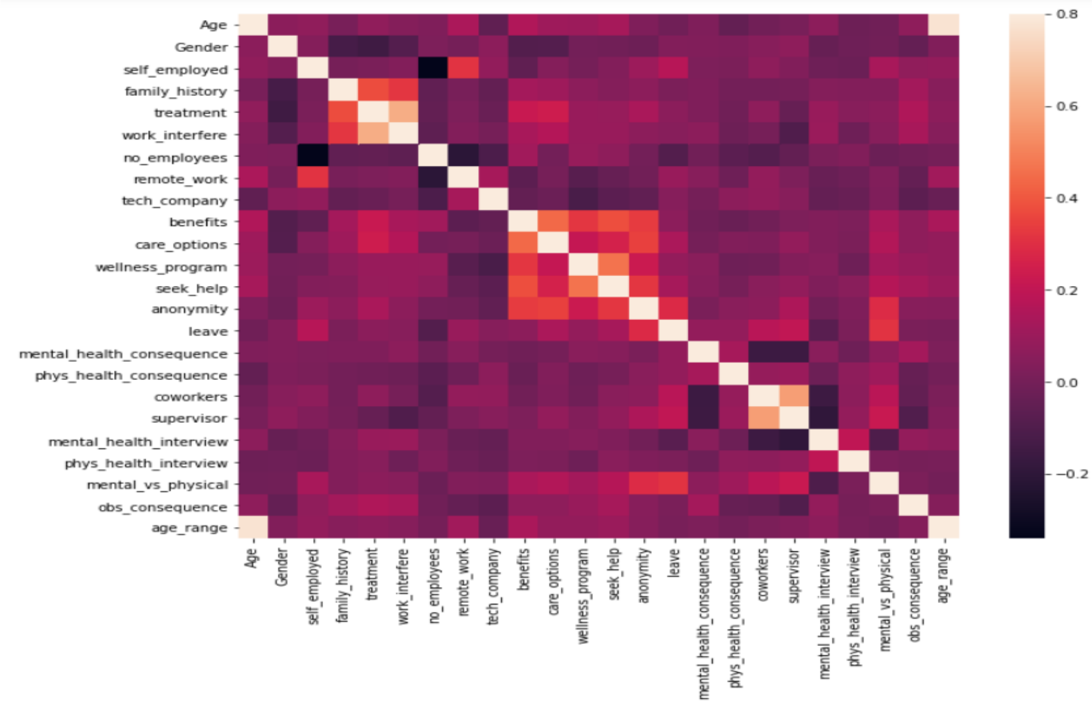


Fig. 4.3 Correlation Matrix

This is the correlation matrix. It defines the relationship between all the columns which is present in the dataset. As we can see the attribute of the same name will have the highest correlation. From the index it is clearly visible lighter the colour greater the relation and darker the colour weaker the relationship.

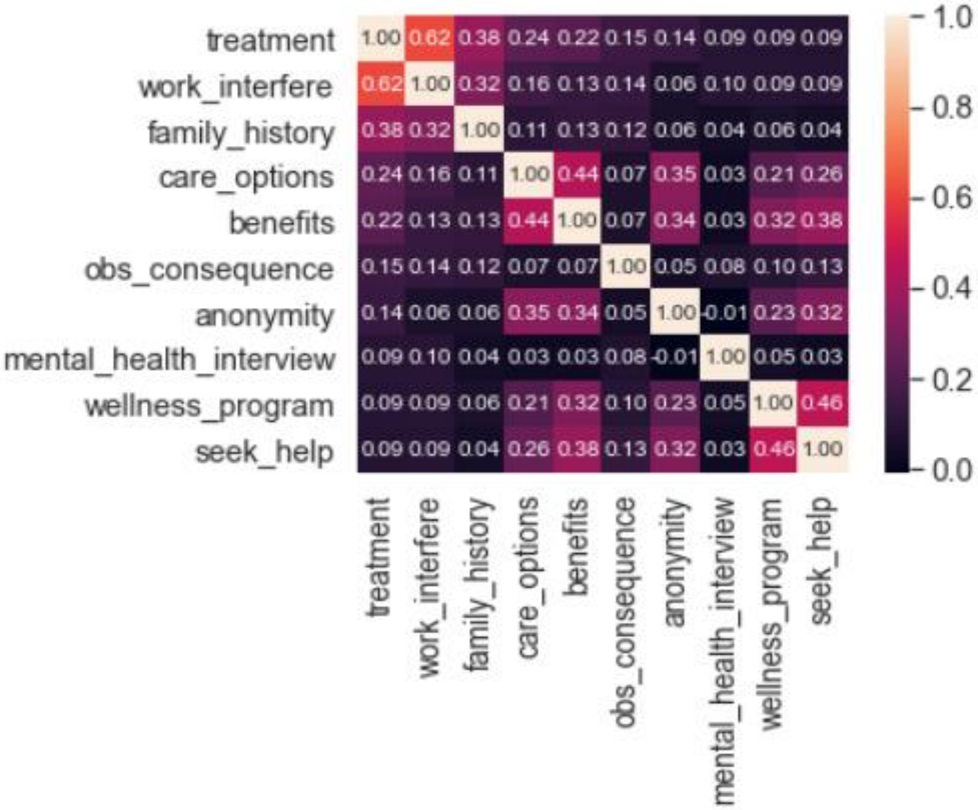


Fig. 4.4 Treatment Correlation Matrix

The above correlation matrix is specifically for the treatment attributes. All the attributes related to treatment are covered here. Here also the same rules apply darker the colour between the attributes weaker is the relationship and vice-versa.

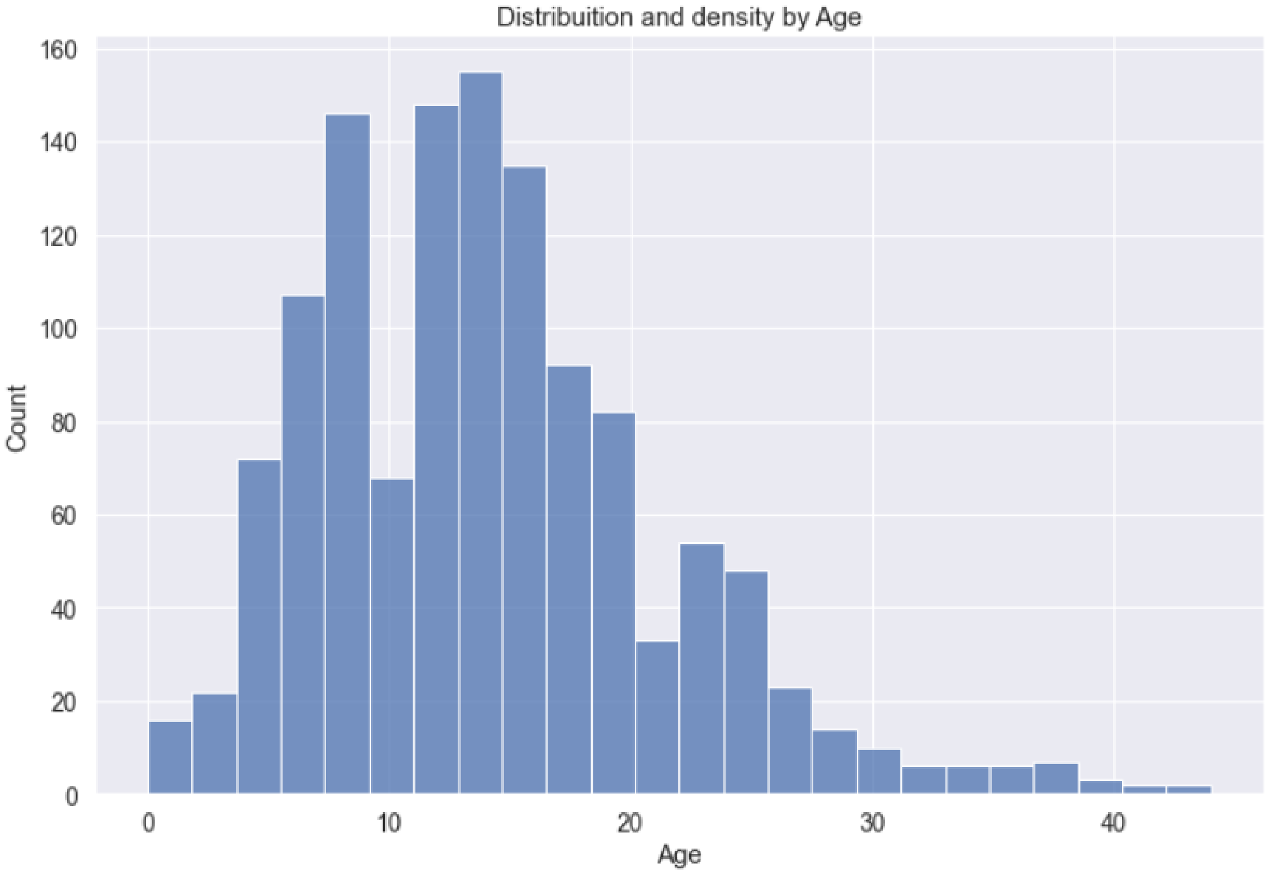


Fig. 4.5 Distribution and Density by Age

In the above histogram age is plotted on x axes and count is plotted on y axes. From it is visible that the age ranges from 10-20 are more prone to weaker mental health condition.

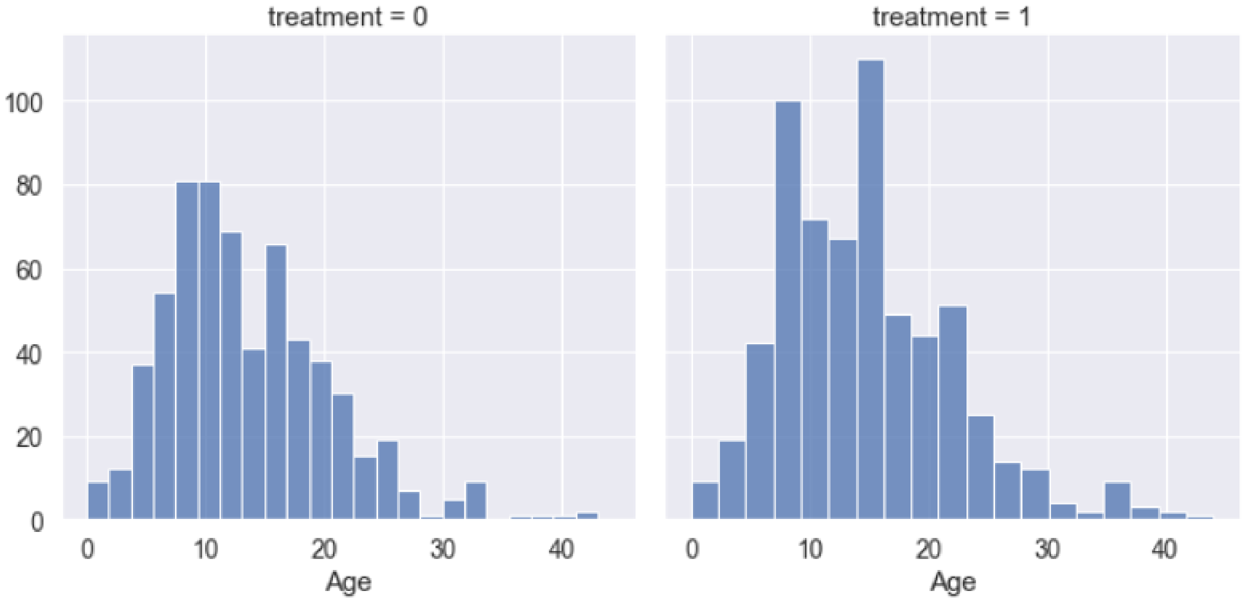


Fig. 4.6 Separated by Treatment or Not

Here age is plotted on the x axes with percentage on y axes. The people who have undergone treatment for mental health are more as compared to the people who haven't taken any treatment. Also, the age ranges from 10-20 are seen more responsible for taking proper treatment.

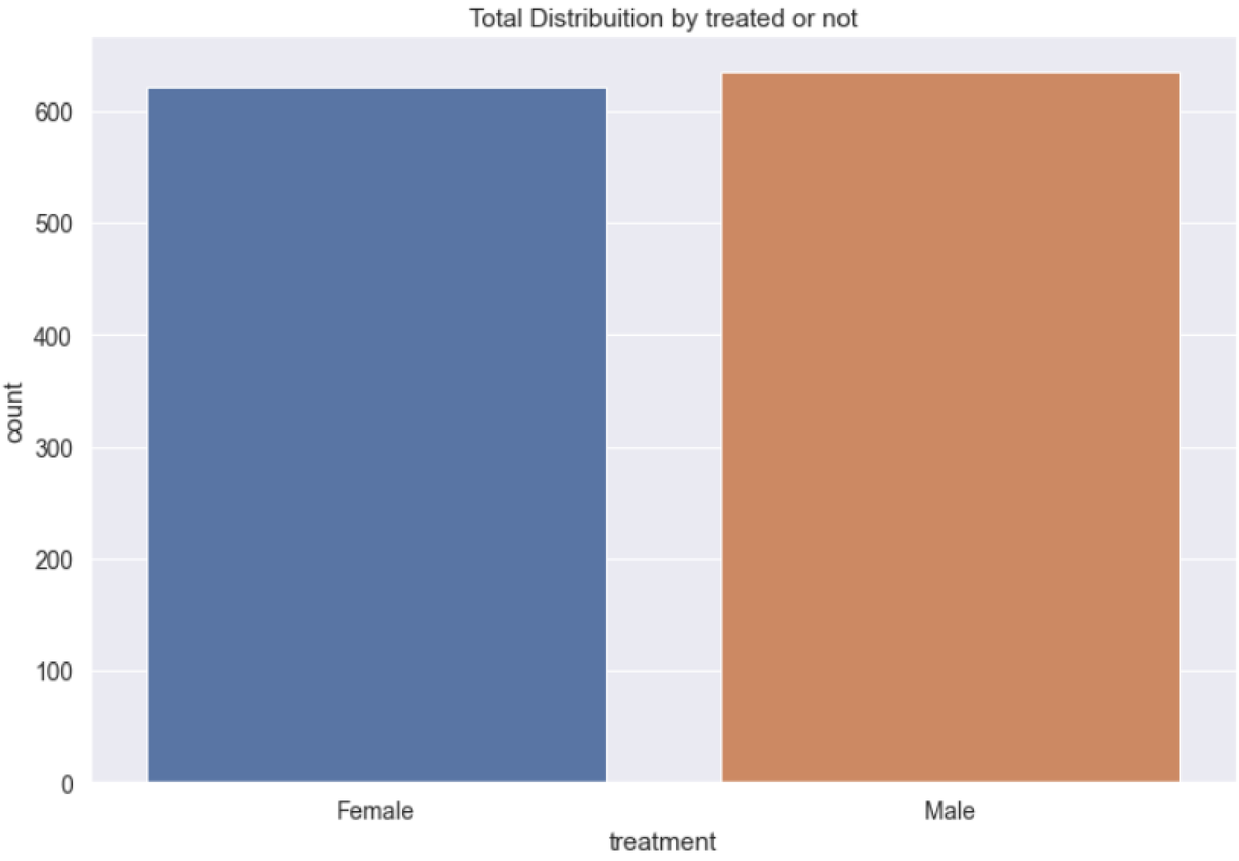


Fig. 4.7 Total Distribution by Treated or Not

The treatment is labelled on x axes with count on y axes. Most of the gender which have undergone treatment are male. Hence, we can conclude that males are more responsible when compared to female in terms of undergoing treatment.

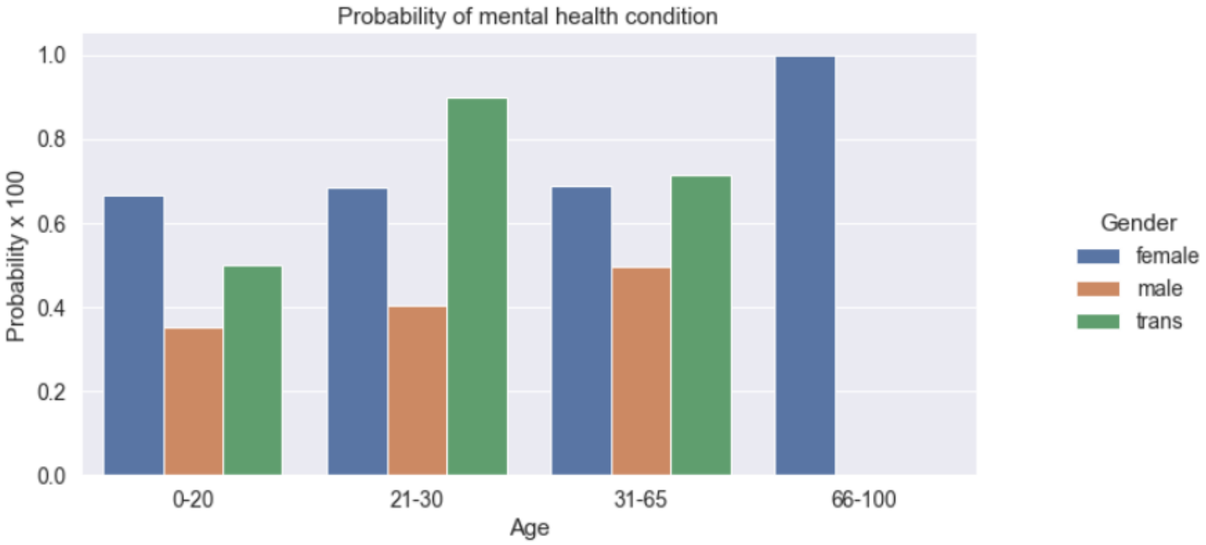


Fig. 4.8 Probability of Mental Health Condition Based on Age

Here age range is plotted on x axes and probability is marked on y axes. Form here we can conclude the age range from 21-30 are more prone to the bad mental health condition. The adulthood of the nation is prone to weaker mental health conditions.

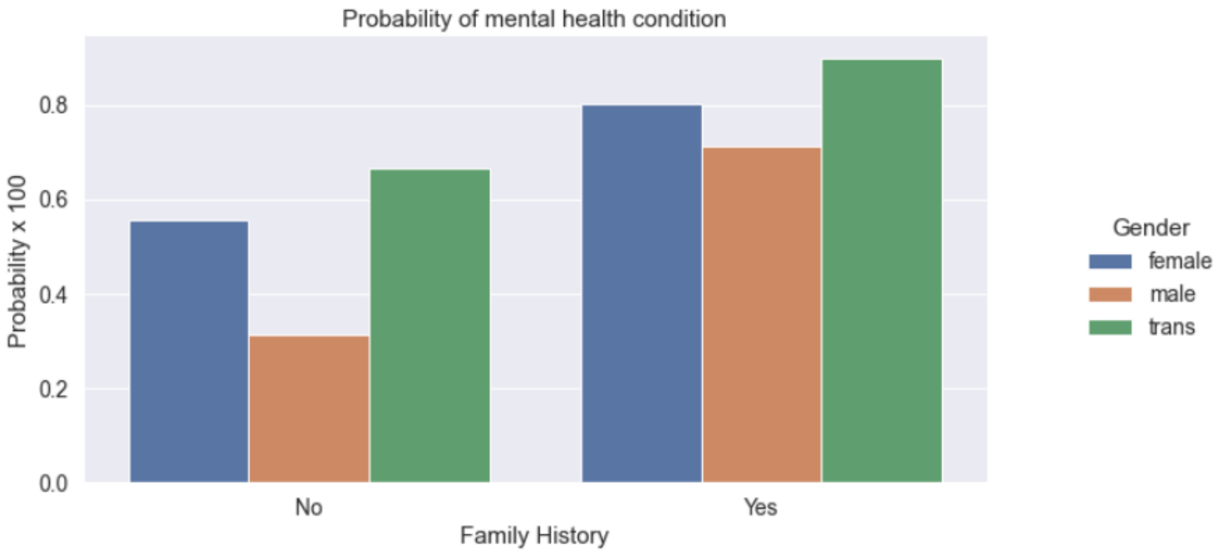


Fig. 4.9 Probability of Mental Health Condition Based on Family History

The above graph tells us that whether family history of mental health plays a major role in one’s life. And it is quite inevitable that most of the people have voted for yes.

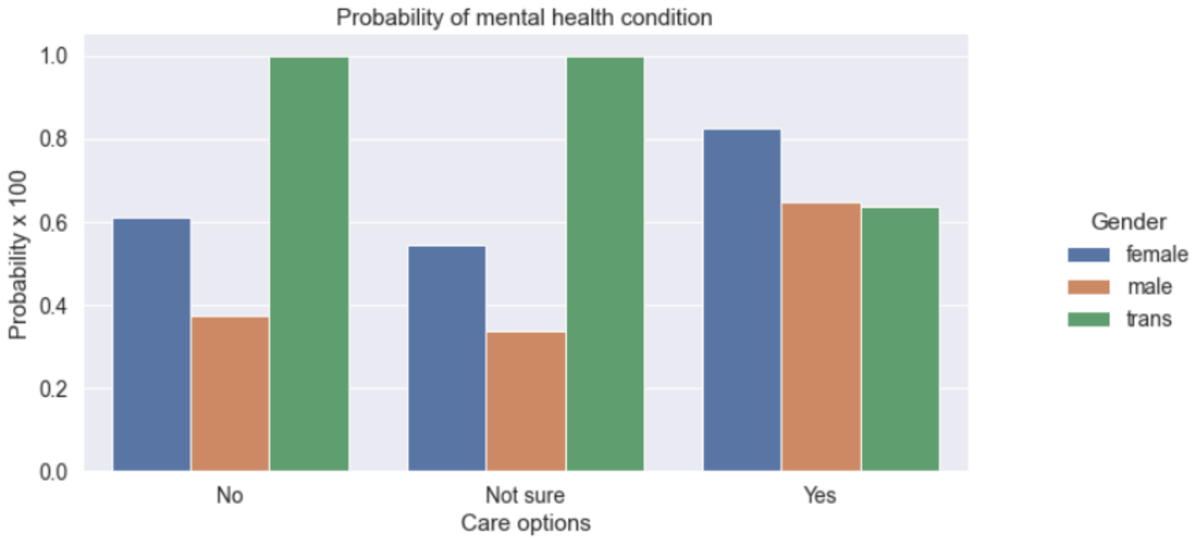


Fig. 4.10 Probability of Mental Health Condition Based on Care Options

Here care options are plotted on x axes and probability is plotted on y axes. Here people were asked whether they have taken any special care for their mental health and most of them have said no. From here we come to know mental health is not given much attention.

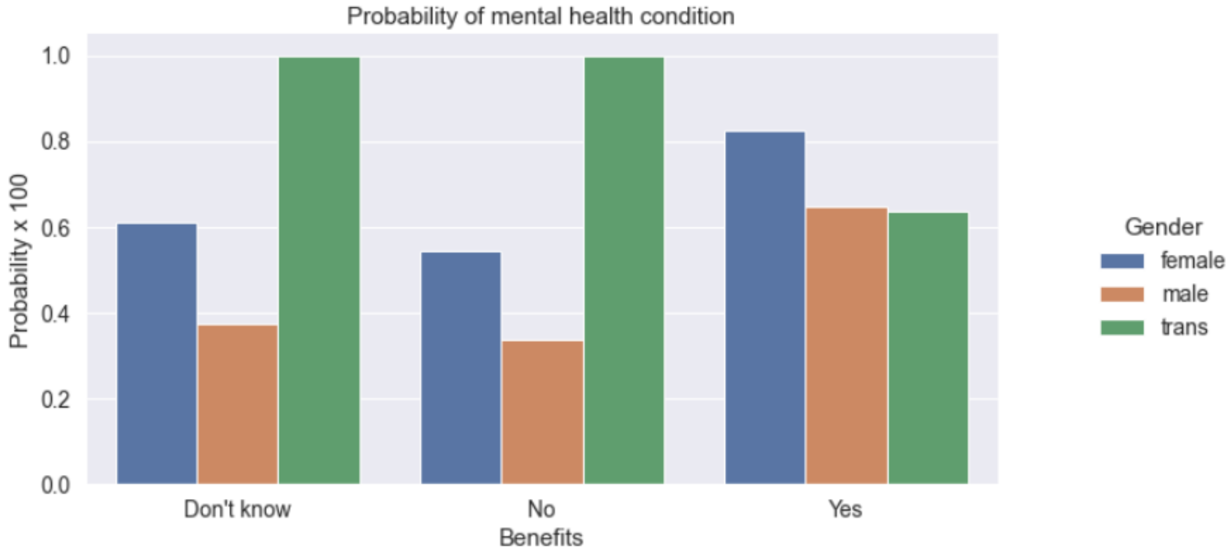


Fig. 4.11 Probability of Mental Health Condition Based on Benefits

On the x axes benefits are plotted and on the y axes probability is plotted. When the people were asked whether they can find any significant benefits after going through the treatment most of the males and females have answered yes and the trans have answered no.

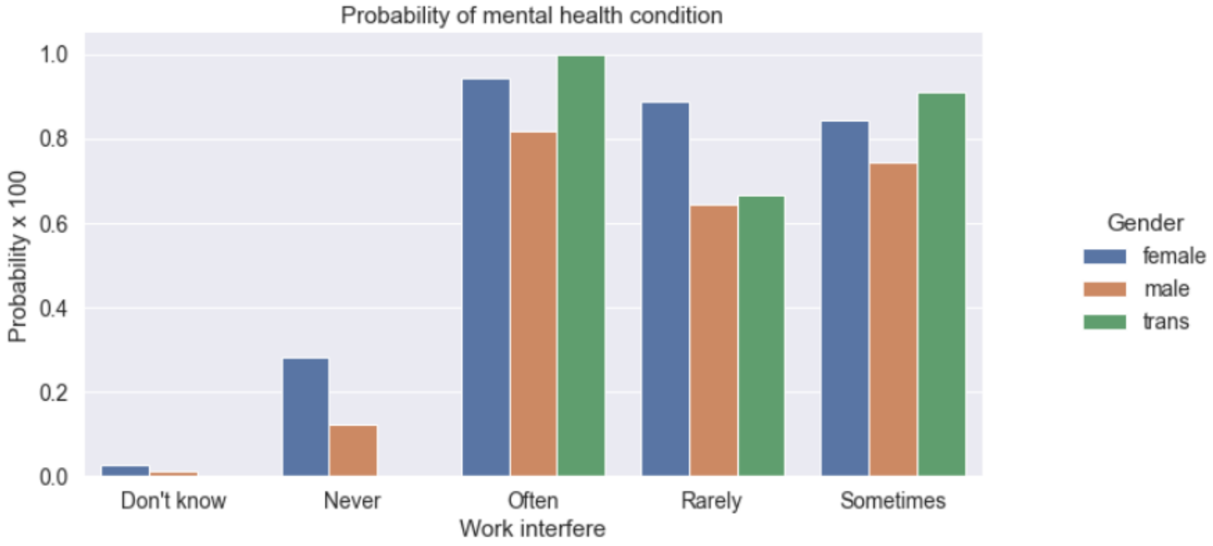


Fig. 4.12 Probability of Mental Health Condition Based on Work Interface

Work interfere was marked on the x axes and probability was marked on the y axes. People were asked whether work environment plays a major role in mental health. Most of them have answered often to this question.

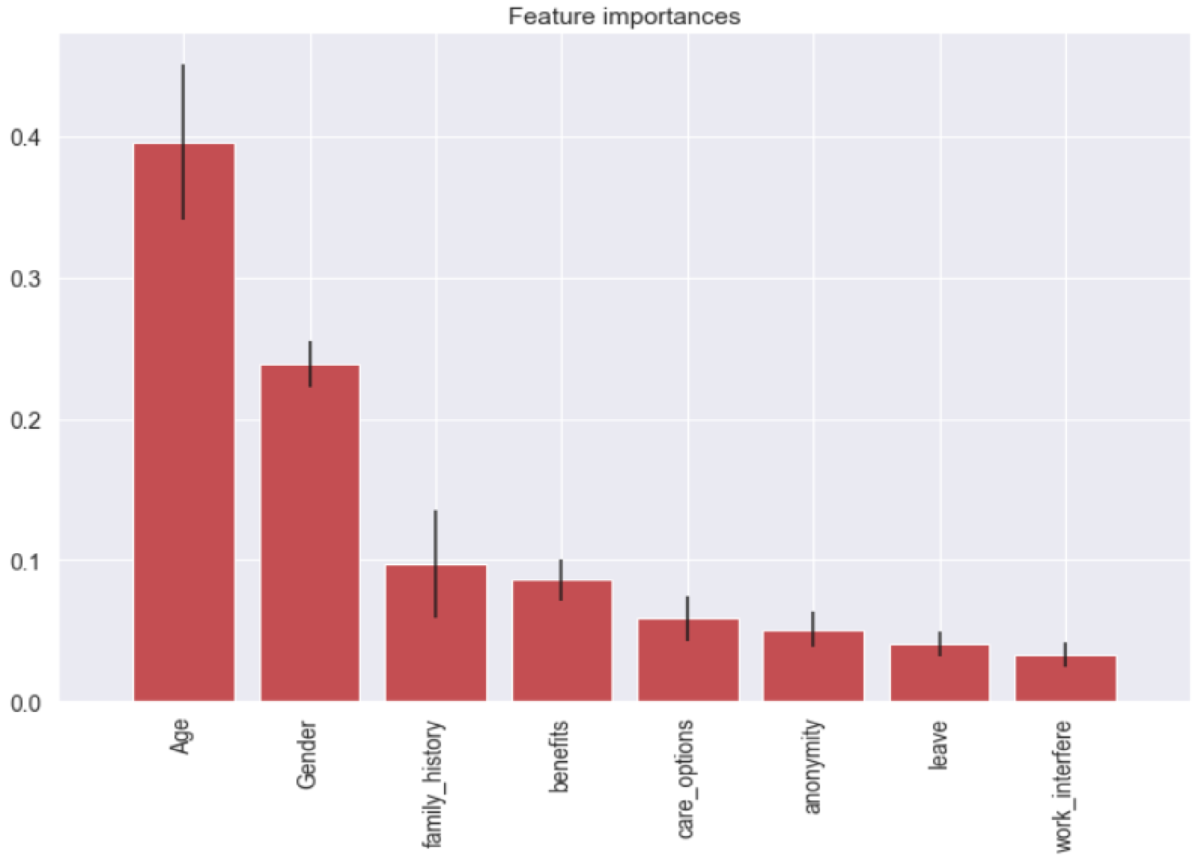
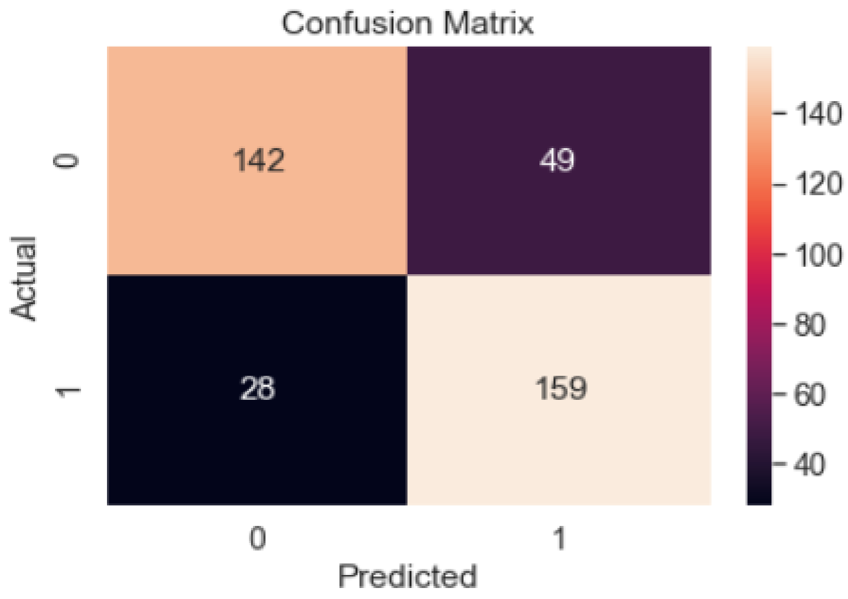


Fig. 4.13 Importance of Features

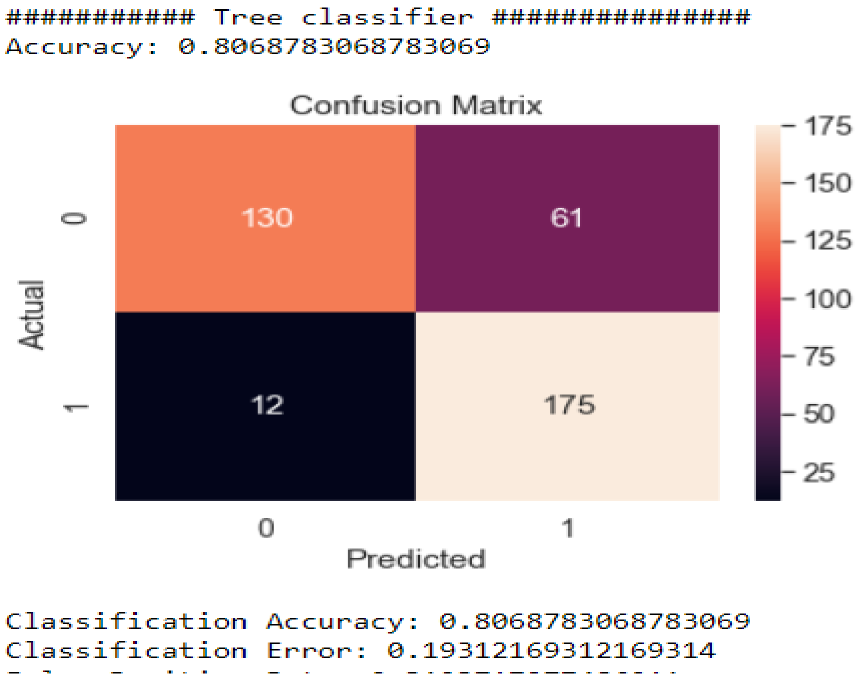
From the above graph we can see age plays an important role in the poor mental health conditions. Mostly the adulthood is affected by this. This is followed by gender, family-history, benefits and care-options.



```
Classification Accuracy: 0.7962962962962963
Classification Error: 0.20370370370370372
```

Fig. 4.14 Confusion Matrix for Logistic Regression

The above matrix is known as confusion matrix. This is plotted for logistic regression. When we predict the value as 0 and the result is also 0 is having the number 142. This can be applied to all the boxes of the matrix. Lighter the colour better is the accuracy.

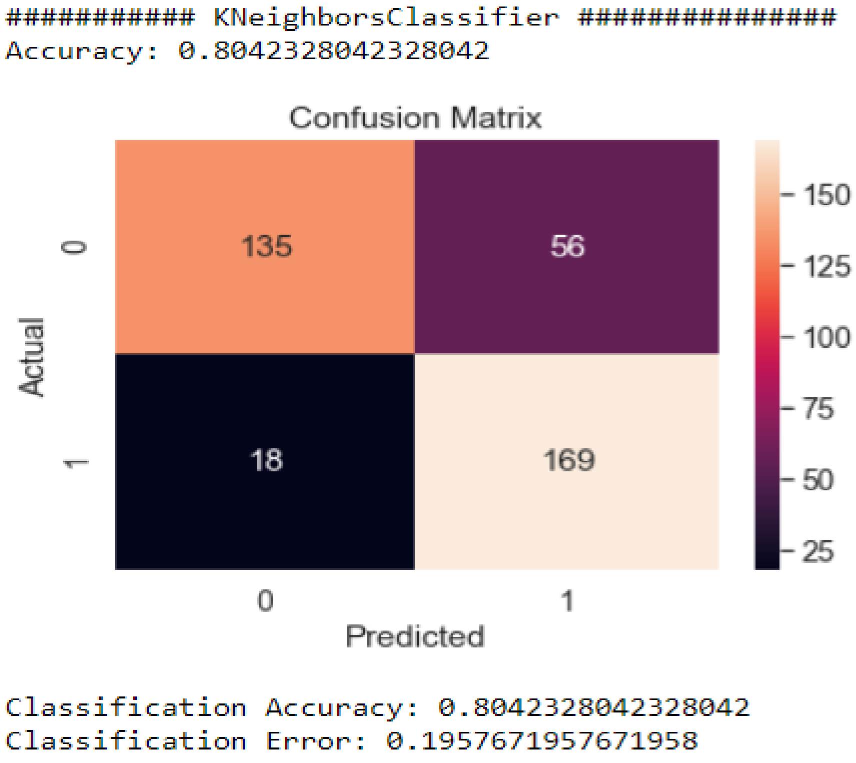


```
########### Tree classifier ###############
Accuracy: 0.8068783068783069
```

```
Classification Accuracy: 0.8068783068783069
Classification Error: 0.19312169312169314
```

Fig. 4.15 Confusion Matrix for Tree Classifier

The above matrix is known as confusion matrix. This is plotted for tree classifier. When we predict the value as 0 and the result is also 0 is having the number 130. This can be applied to all the boxes of the matrix. Lighter the colour better is the accuracy.

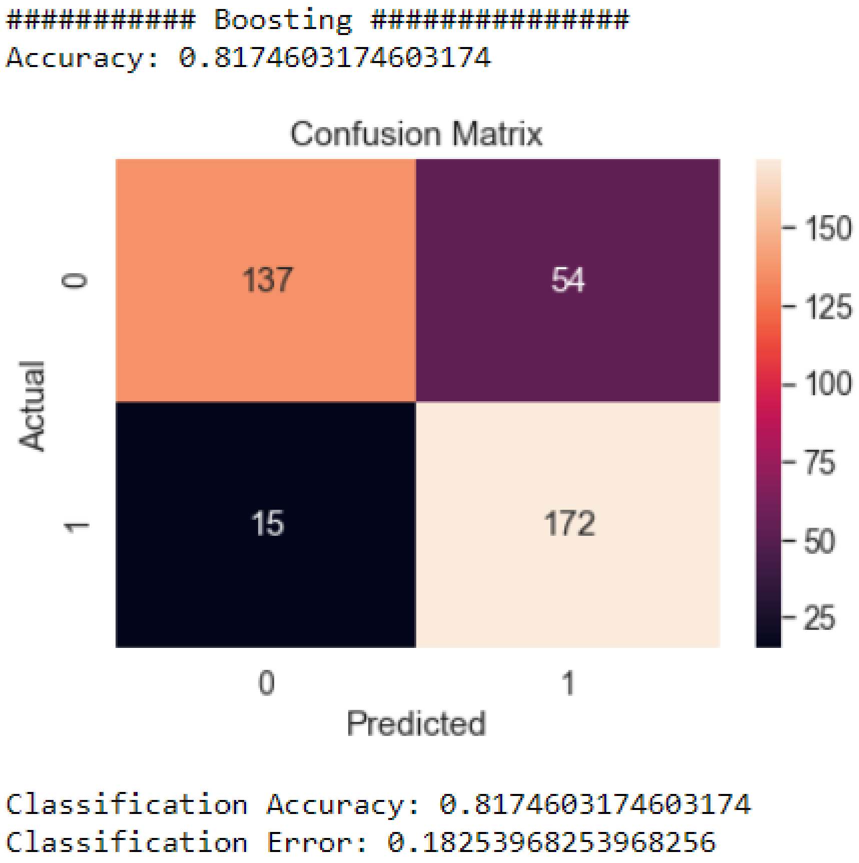


```
########### KNeighborsClassifier ###############
Accuracy: 0.8042328042328042
```

```
Classification Accuracy: 0.8042328042328042
Classification Error: 0.1957671957671958
```

Fig. 4.16 Confusion Matrix for KNN

The above matrix is known as confusion matrix. This is plotted for KNN. When we predict the value as 0 and the result is also 0 is having the number 135. This can be applied to all the boxes of the matrix. Lighter the colour better is the accuracy.

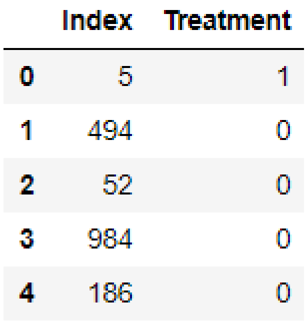


```
########### Boosting ###############
Accuracy: 0.8174603174603174
```

```
Classification Accuracy: 0.8174603174603174
Classification Error: 0.18253968253968256
```

Fig. 4.17 Confusion Matrix for Boosting

The above matrix is known as confusion matrix. This is plotted for logistic regression. When we predict the value as 0 and the result is also 0 is having the number 137. This can be applied to all the boxes of the matrix. Lighter the colour better is the accuracy.



	Index	Treatment
0	5	1
1	494	0
2	52	0
3	984	0
4	186	0

Fig. 4.18 Result CSV

The above table is the snippet of the result we have obtained by training the dataset. The index number indicates the row number in the dataset which we have taken. Treatment labelled as 1 is that the corresponding index number needs treatment for his mental health.

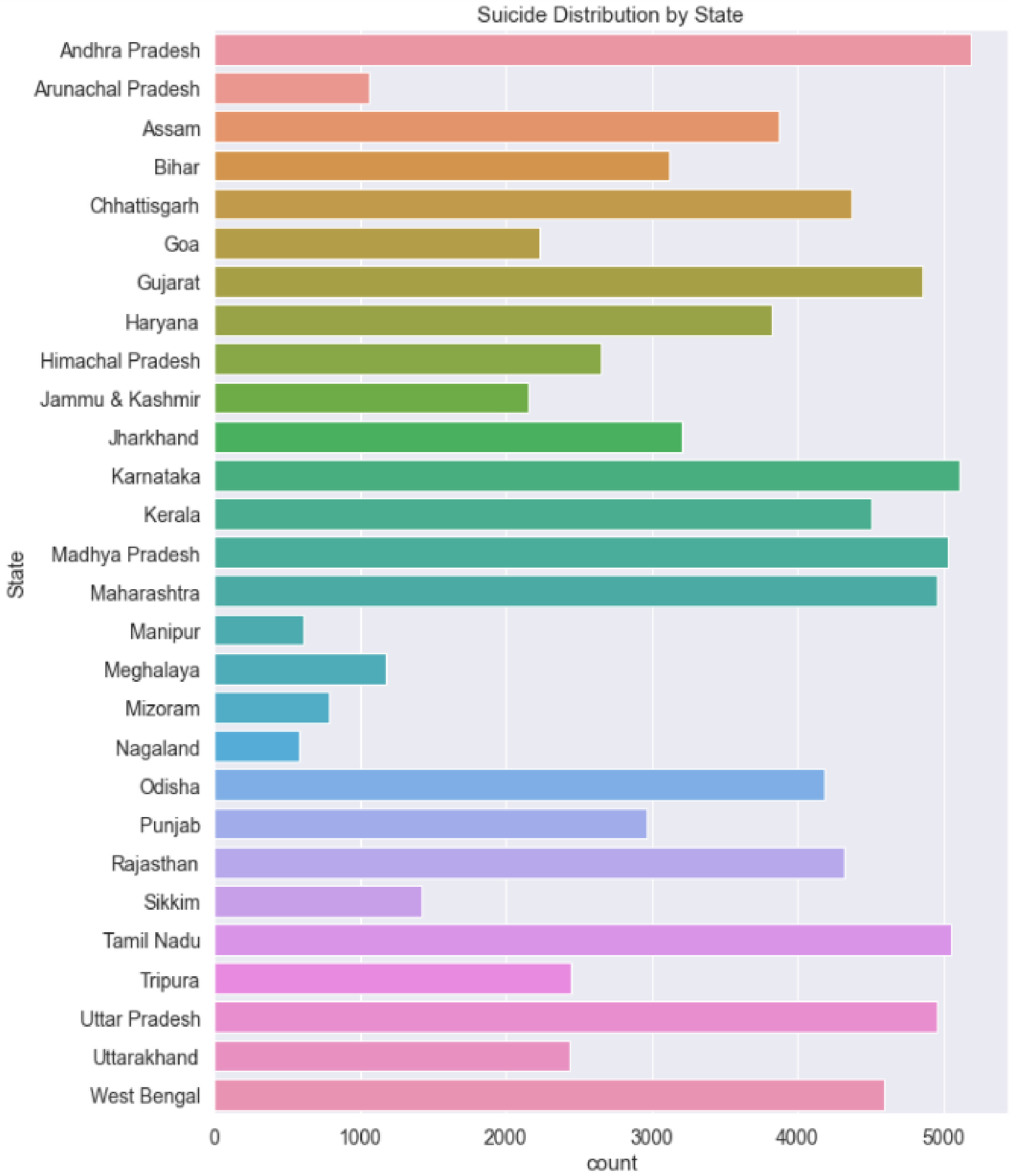


Fig. 4.19 Suicide Distribution by State

On the x axes count is plotted along with state on y axes. From the above bar graph we can see that most of the suicide cases are found in Andhra Pradesh followed by Karnataka and Madhya Pradesh.

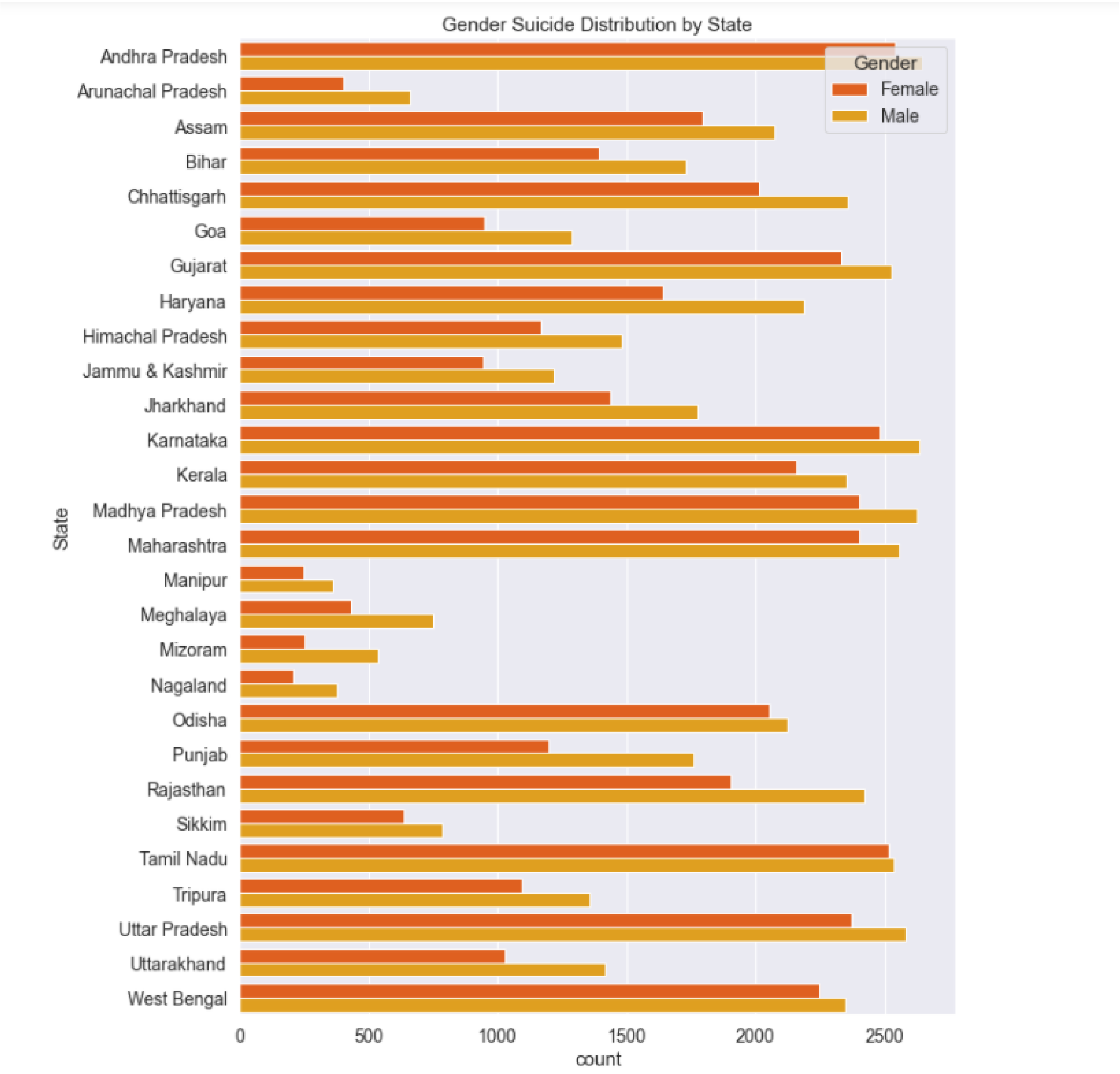


Fig. 4.20 Suicide Distribution by Gender

On the x axes count is plotted and on the y axes state is plotted. In this difference is made on the gender. In Andhra Pradesh most of the males have committed suicides. In all the states it has been observed that most of the males are committing suicide when compared to females.

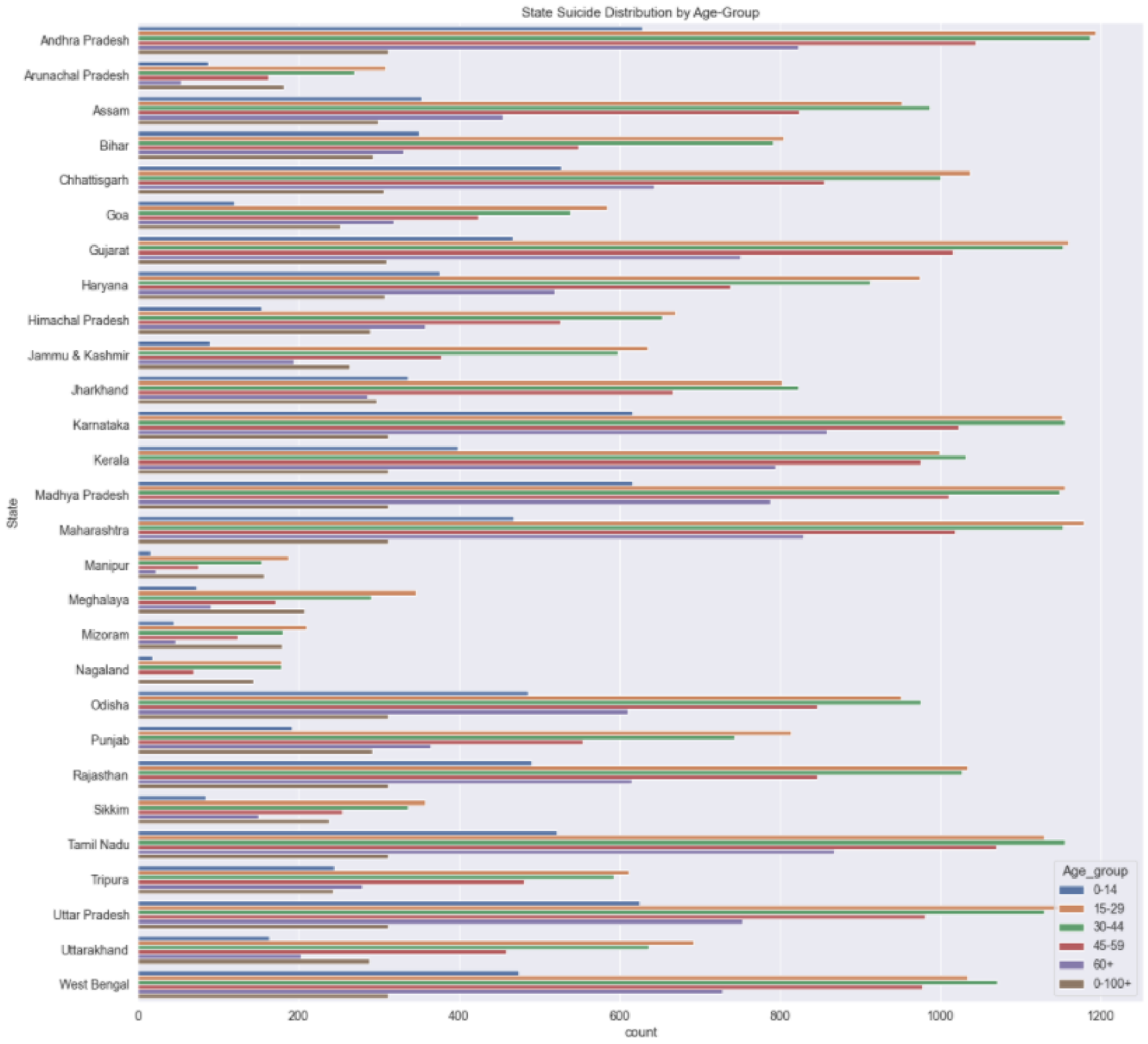


Fig. 4.21 State Suicide Distribution by Age-Group

On the x axes count is plotted and on the y axes state is plotted. In this difference is made on the age-range. The age-range which has committed the greatest number of suicides are 15-30. It is evident that most of the adulthood people are committing suicides.

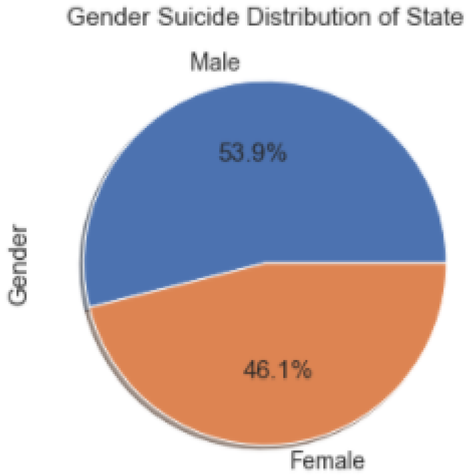


Fig. 4.22 Gender Suicide Distribution of State

The pie-chat describes the distribution of suicide based on the gender. Here, males have occupied 53.9% of the chart while females contribute to 46.1% of the total.

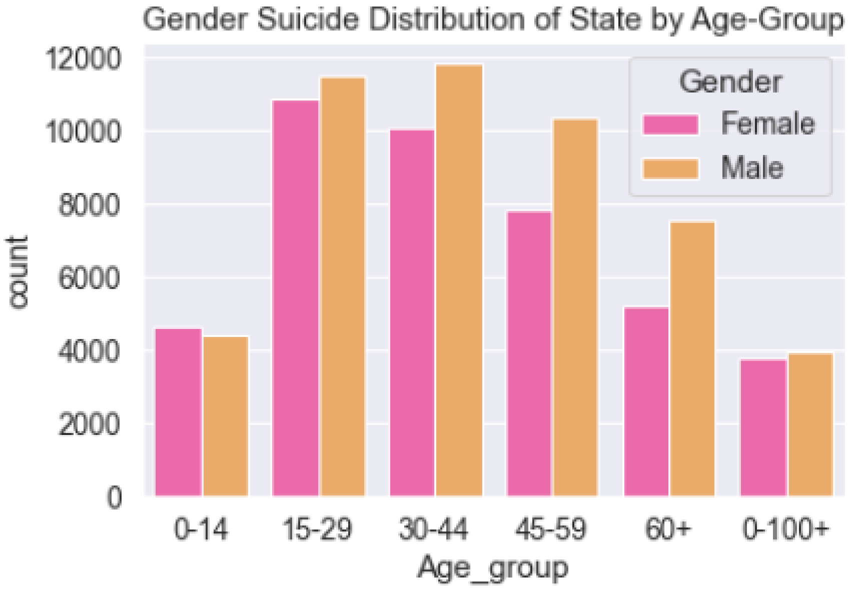


Fig. 4.23 Gender Suicide Distribution of State by Age-Group

Age-group is plotted on the x axes while count is plotted on the y axes. The distinction is based on gender. In the age-group 30-44 most of the males have committed suicides. Similarly, in the age-range of 15-29 most of the females have committed suicides.

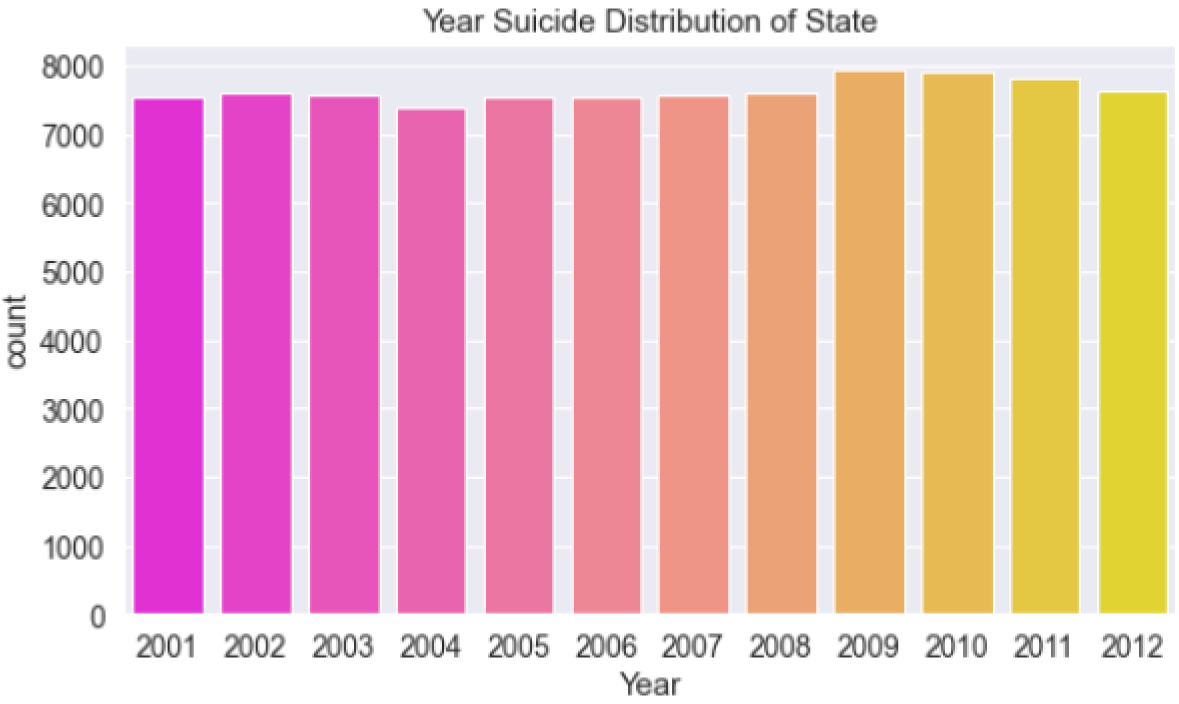


Fig. 4.24 Year Suicide Distribution of State

The above bar-graph tells us how the count of suicides has been continuously increasing over the years. It is clearly visible that people are not paying much attention to their mental health and hence the death rates are increasing.

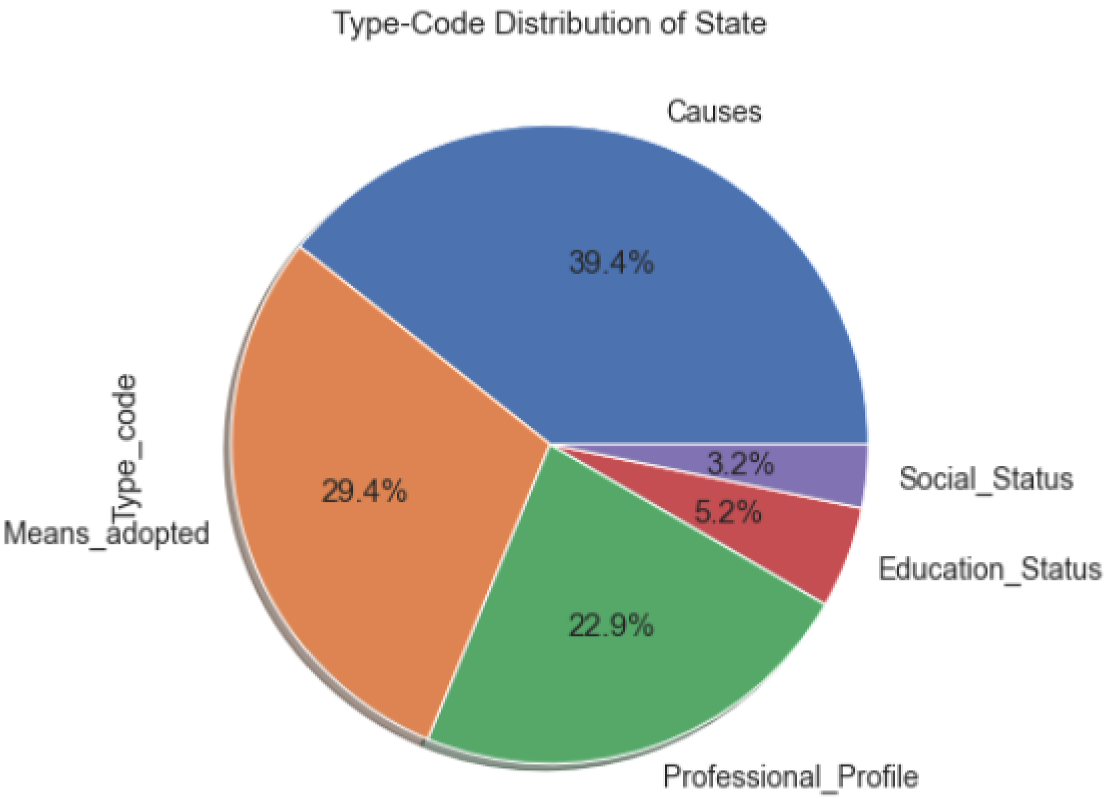


Fig. 4.25 Distribution Based on Type-Code

This pie-chart tells us about the various components such as the causes for the suicide, mean-adopted while the person has attempted suicide, education-status of the person who has committed suicide and the profession-profile of the person who has committed suicide.

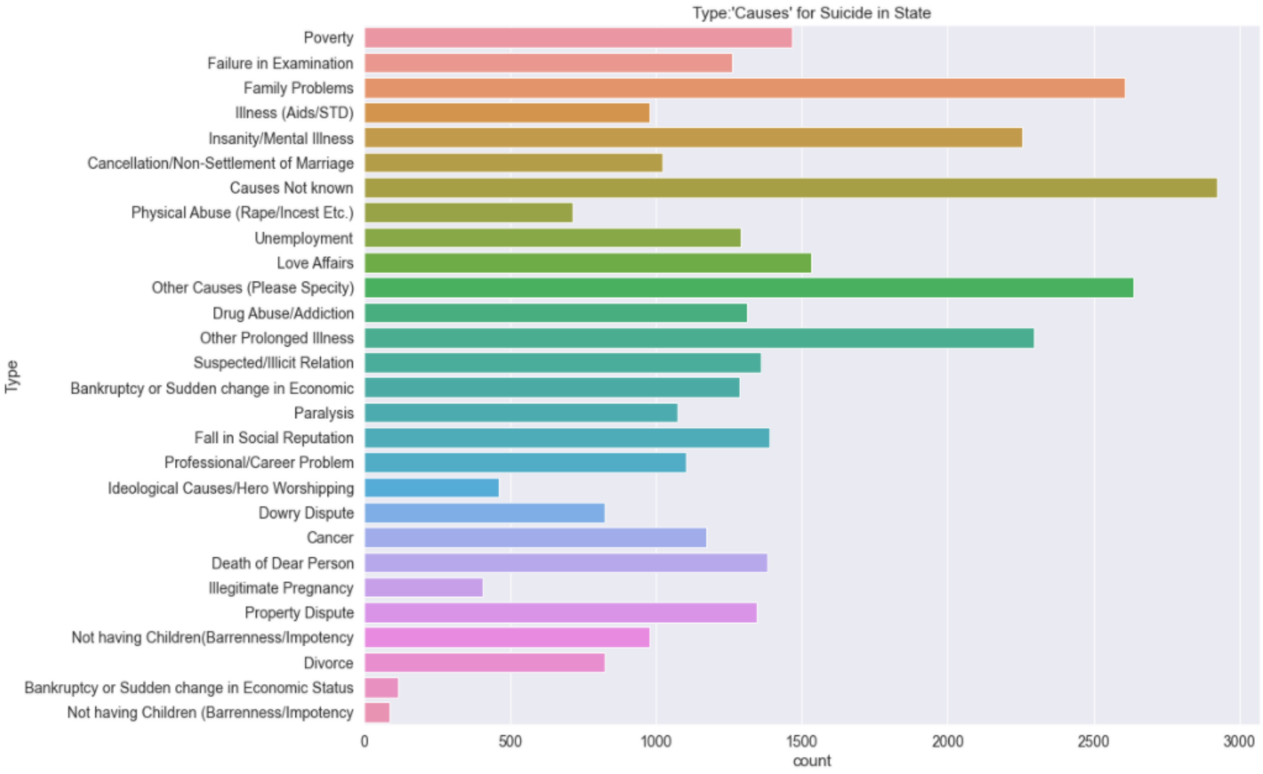


Fig. 4.26 Causes of Suicide

On x axes count is plotted and y axes causes of suicide is plotted. The main reason which can be seen from the graph is family problems and disease.

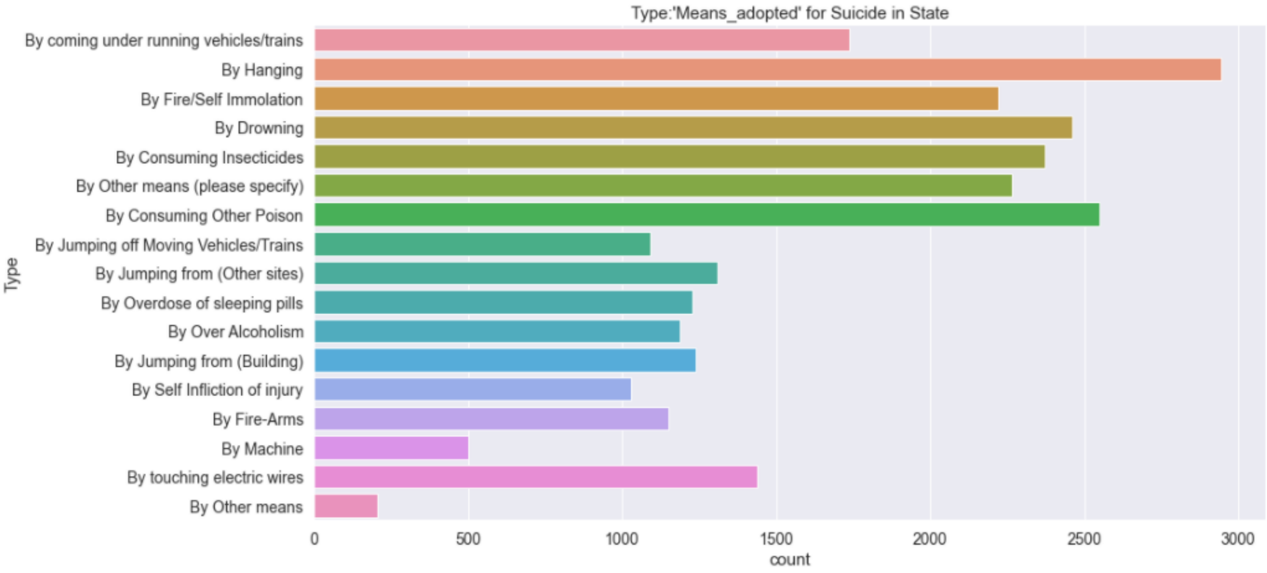


Fig. 4.27 Means Adopted for Suicide

On x axes count is plotted and y axes means-adopted for suicide is plotted. The main means were by hanging or by consuming poison.

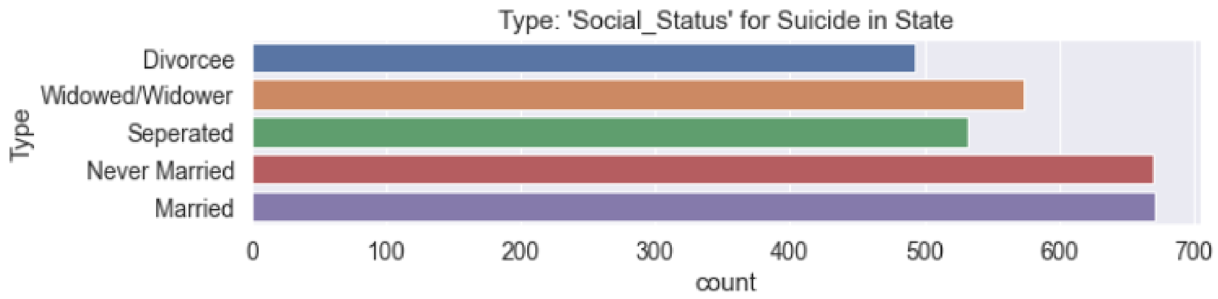


Fig. 4.28 Social-Status for Suicide

On x axes count is plotted and on y axes Social-status is plotted. Mostly married people were found to have committed suicides.

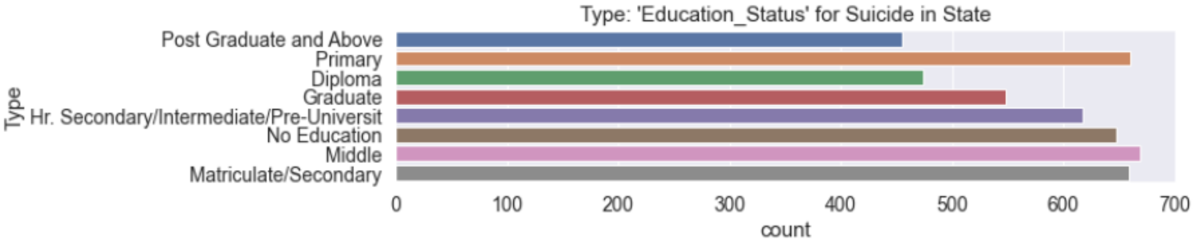


Fig. 4.29 Educational Status of the People who Committed suicide

On x axes count is plotted and on y axes education-status is plotted. Most of the people who have committed suicide are either ill literate or were having primary education.

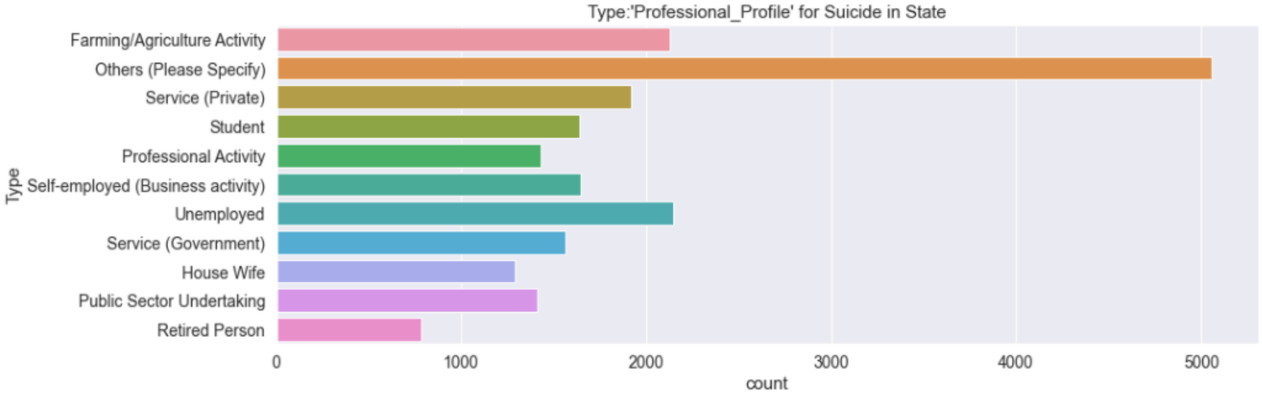


Fig. 4.30 Professional-Profile of the People who Committed suicide

On x axes count is plotted and y axes professional-profile is plotted. Most of the people who have committed were either farmers or were unemployed.

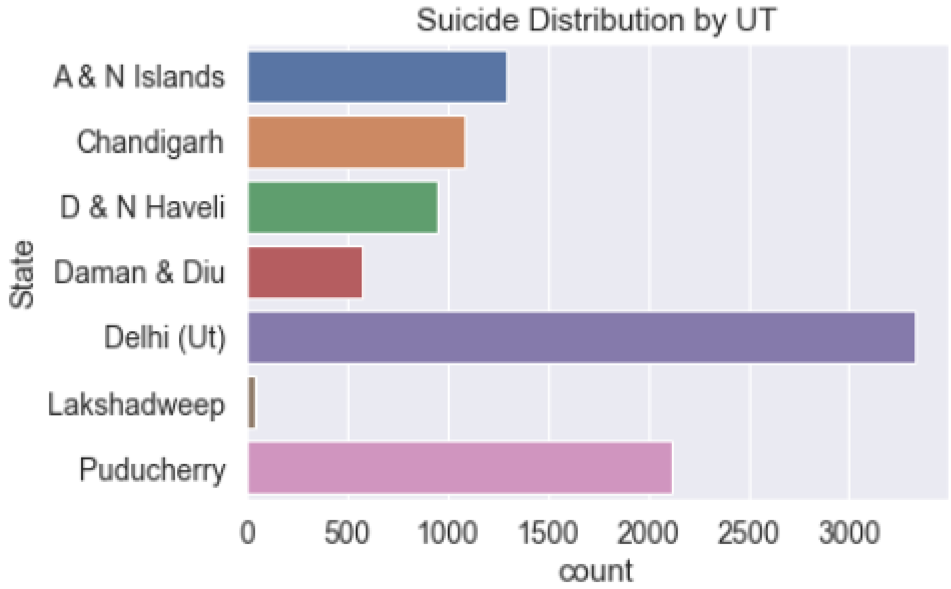


Fig. 4.31 Suicide Distribution Based on UT

On x axes count is plotted and y axes union territory is plotted. Delhi was having the highest number of suicides in the UT.

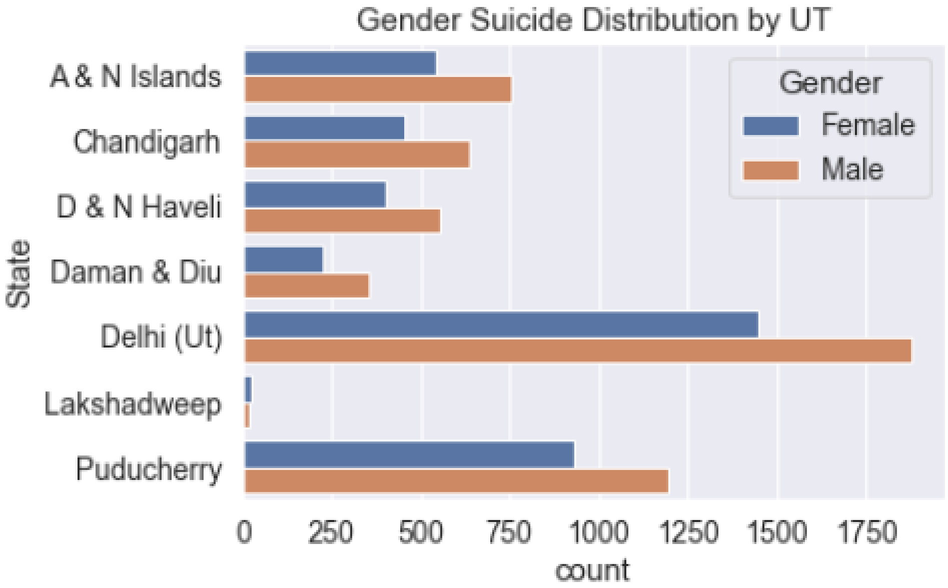


Fig. 4.32 Gender Suicide Distribution by UT

On x axes count is plotted and on y axes state was plotted. Here the difference was made based on the gender. Most of the males have committed suicides in all the union territories.

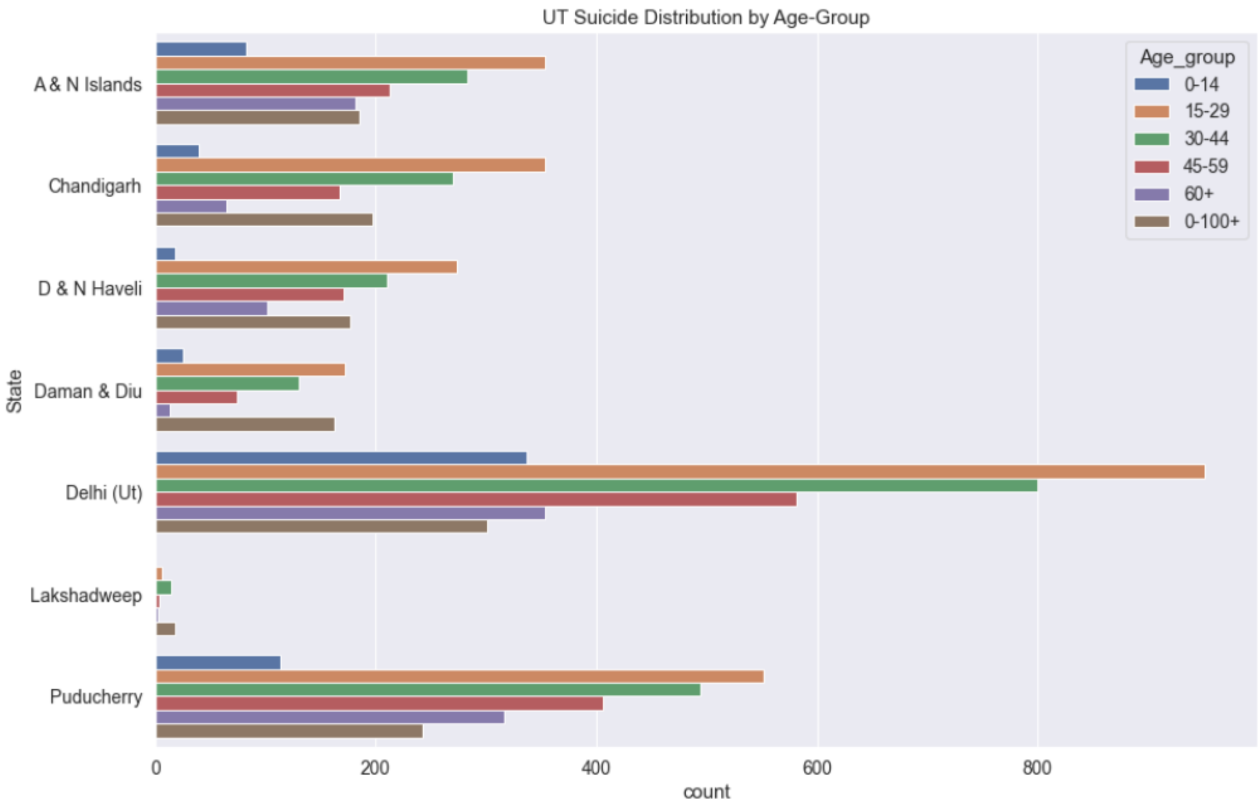


Fig. 4.33 UT Suicide Distribution by Age-Group

On the x axes we can see the count is plotted and on the y axes state is plotted. Here, the distinction is made based on the age-range. The age-range 15-29 are more prone to the suicides. This is visible from the graph above.

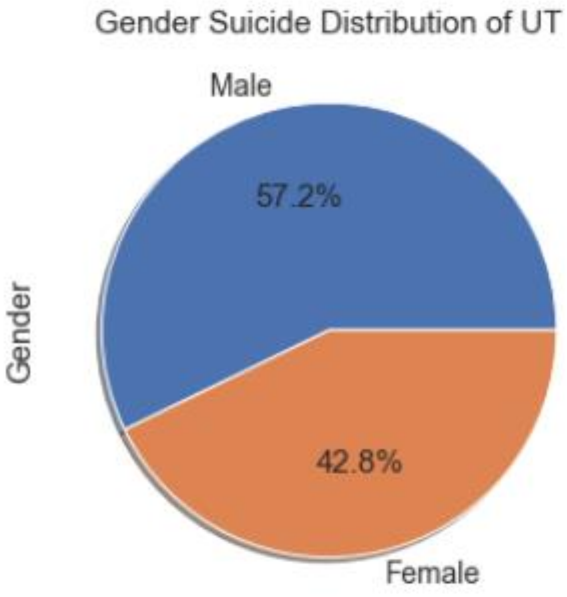


Fig. 4.34 Gender Suicide Distribution of UT

The pie-chart shows the distribution based on the gender. Here the males contribute to 57.2 % of the total suicides and female contributes to 42.8%.

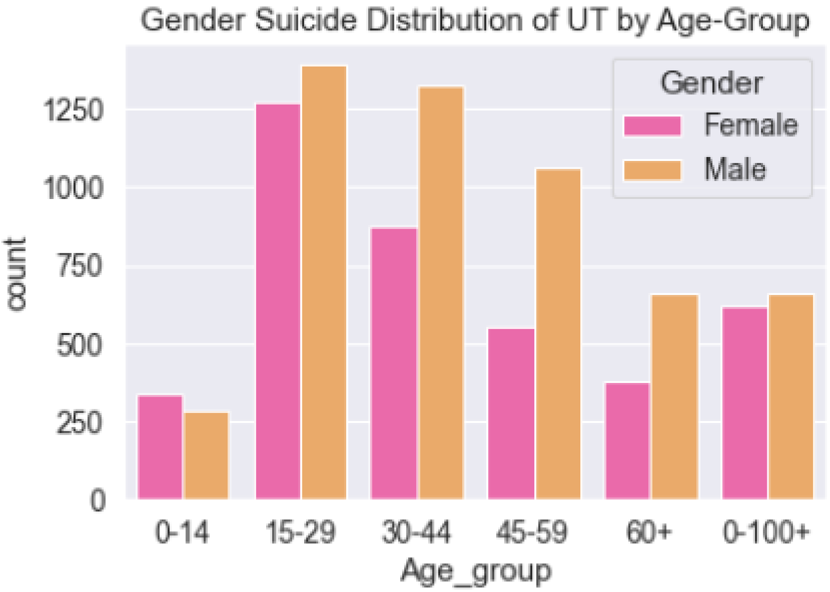


Fig. 4.35 Gender Suicide Distribution of UT by Age-Group

Age-group is plotted on the x axes while count is plotted on the y axes. The distinction is based on gender. In the age-group 15-29 most of the males have committed suicides. Similarly, in the age-range of 15-29 most of the females have committed suicides.

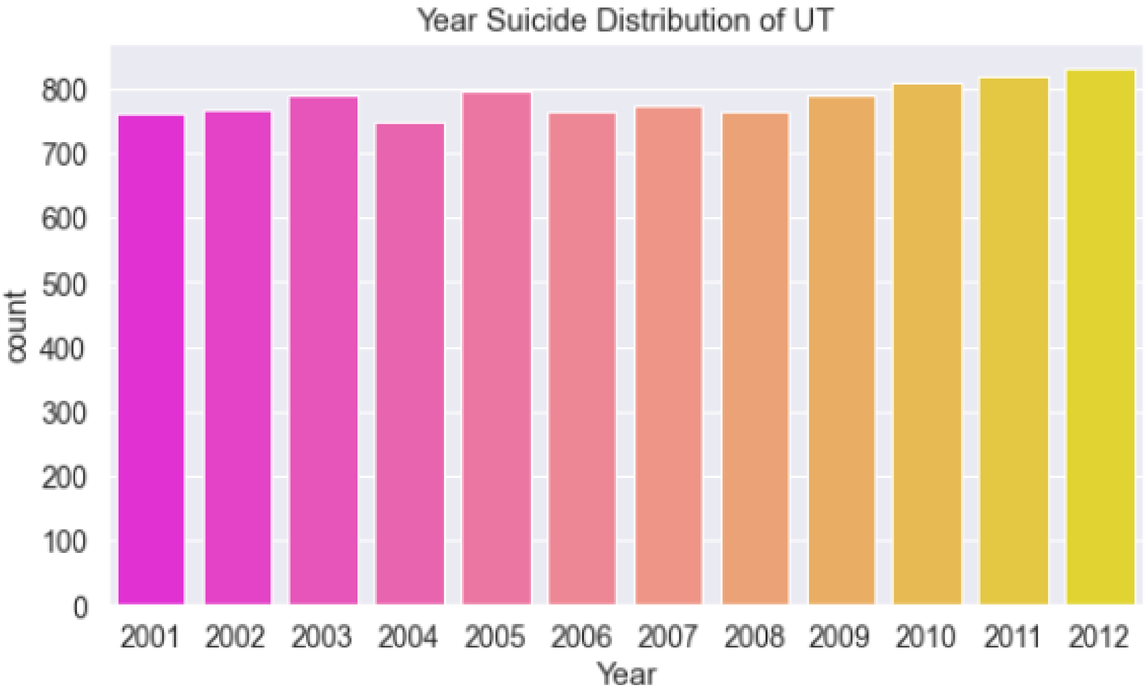


Fig. 4.36 Year Suicide Distribution of UT

The above bar-graph tells us how the count of suicides has been continuously increasing over the years. It is clearly visible that people are not paying much attention to their mental health and hence the death rates are increasing.

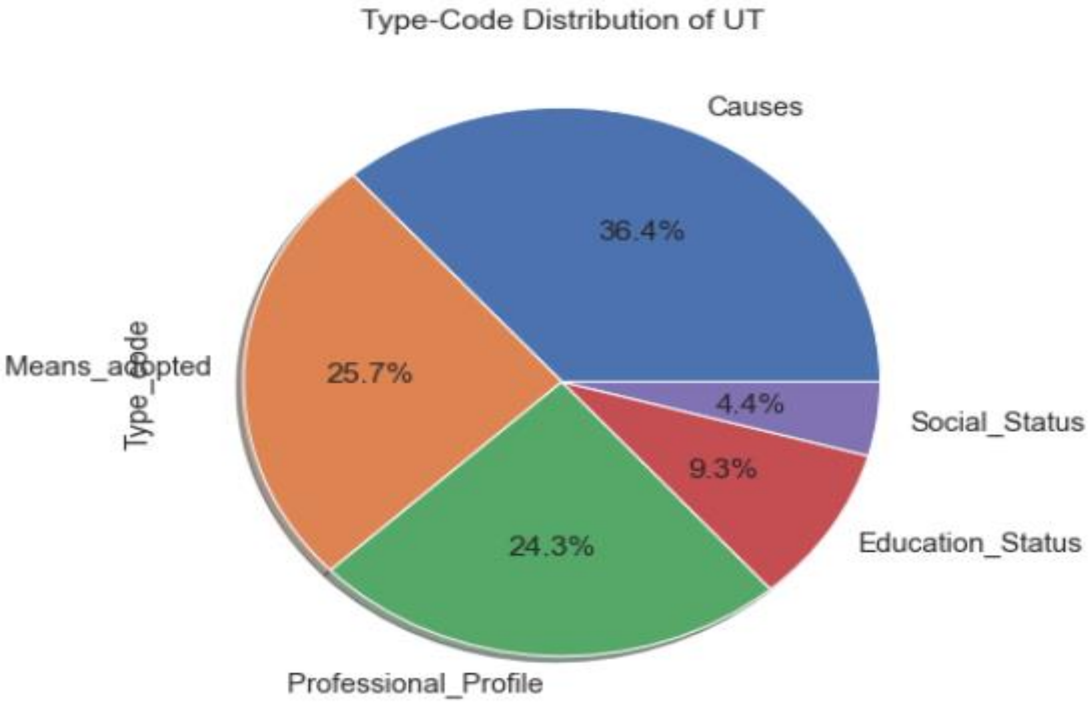


Fig. 4.37 distribution Based on Type-Code

This pie-chart tells us about the various components such as the causes for the suicide, mean-adopted while the person has attempted suicide, education-status of the person who has committed suicide and the profession-profile of the person who has committed suicide.

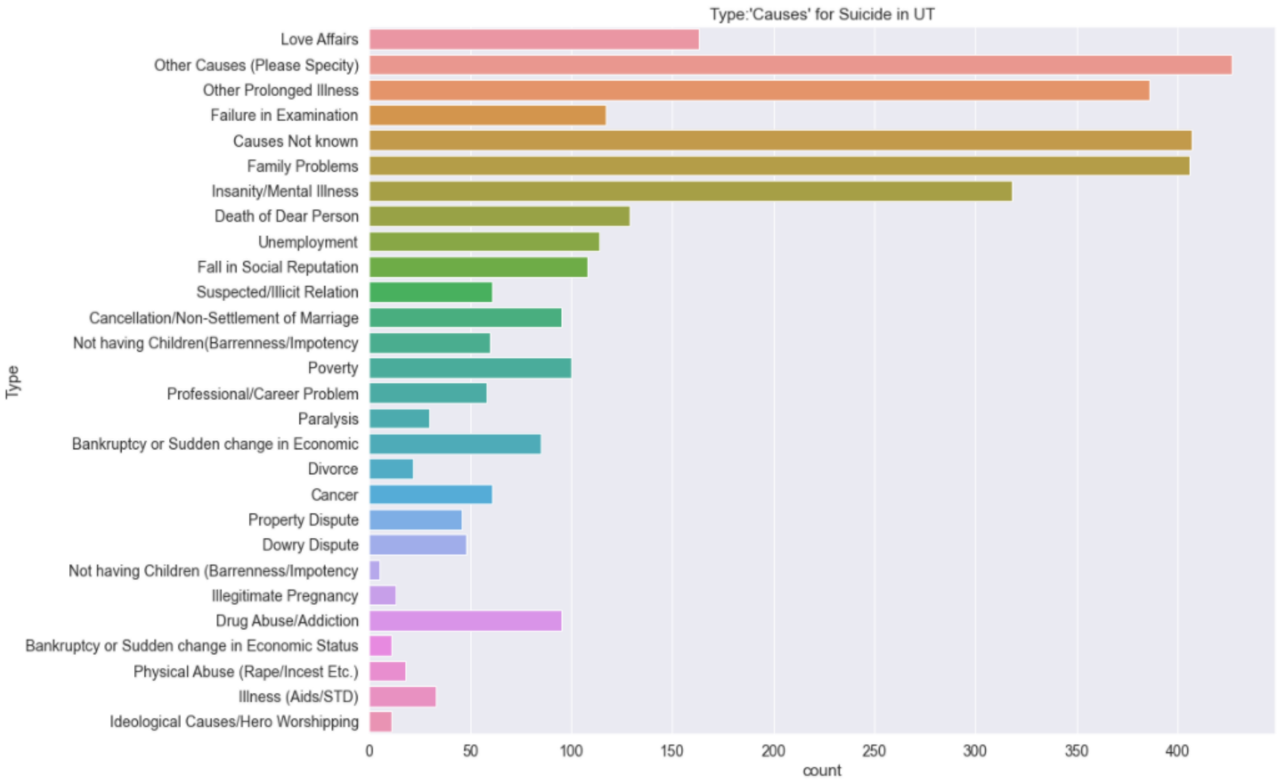


Fig. 4.38 Causes for Suicide in UT

On x axes count is plotted and y axes causes of suicide is plotted. The main reason which can be seen from the graph is family problems and disease.

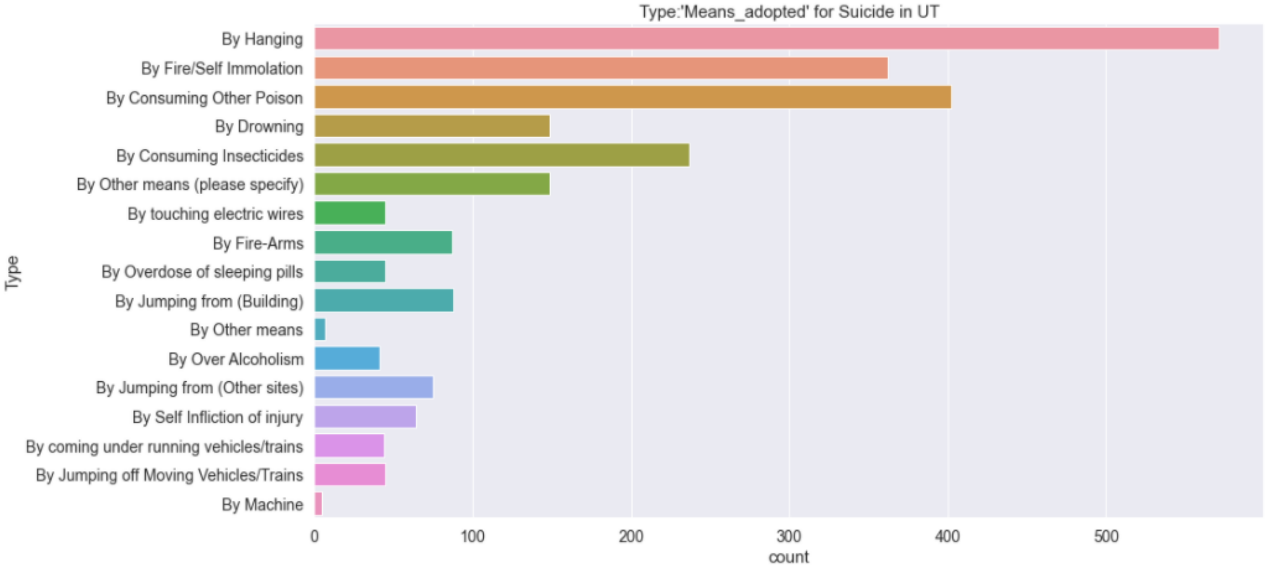


Fig. 4.39 Means-Adopted for Suicide in UT

On x axes count is plotted and y axes means-adopted for suicide is plotted. The main means were by hanging or by consuming poison.

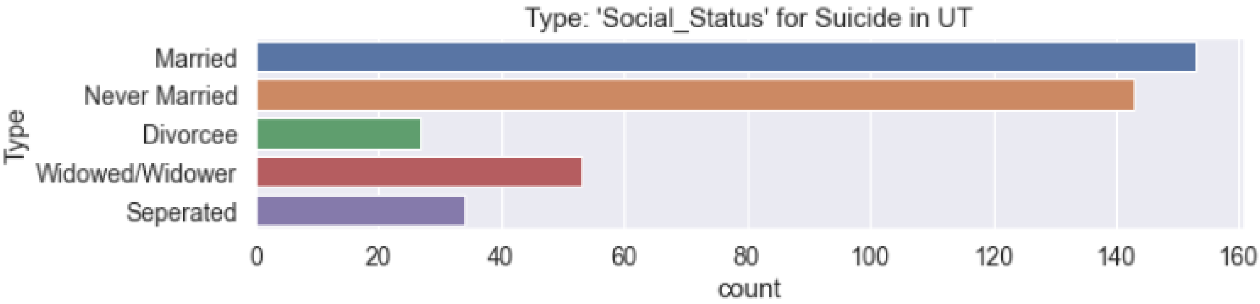


Fig. 4.40 Social-Status for Suicide in UT

On x axes count is plotted and on y axes Social-status is plotted. Mostly married people were found to have committed suicides.

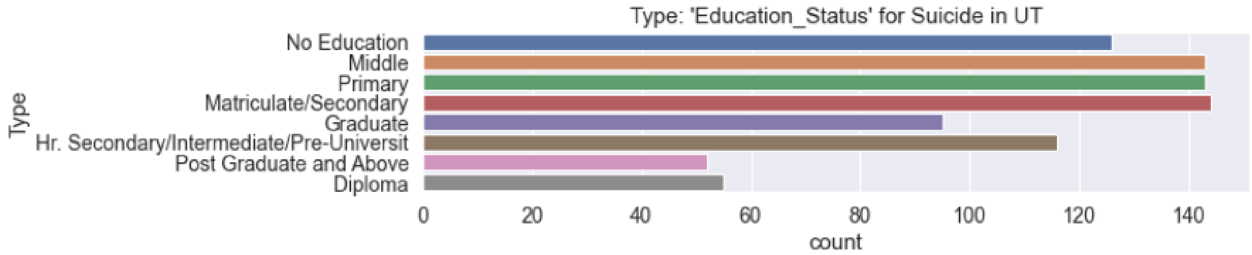


Fig. 4.41 Education-Status for Suicide in UT

On x axes count is plotted and on y axes education-status is plotted. Most of the people who have committed suicide are either ill literate or were having primary education or were having middle class education.

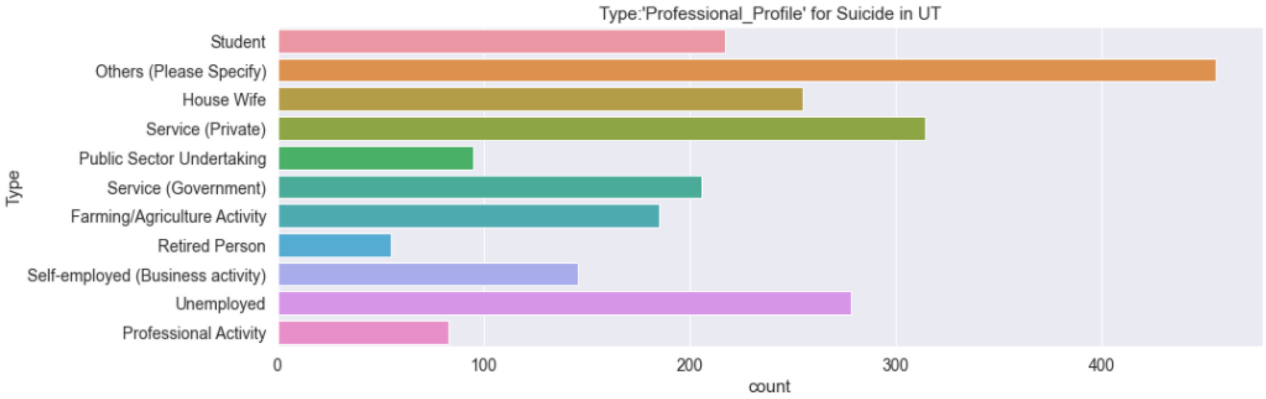


Fig. 4.42 Professional-Profile for Suicide in UT

On x axes count is plotted and y axes professional-profile is plotted. Most of the people who have committed were either students or were unemployed.

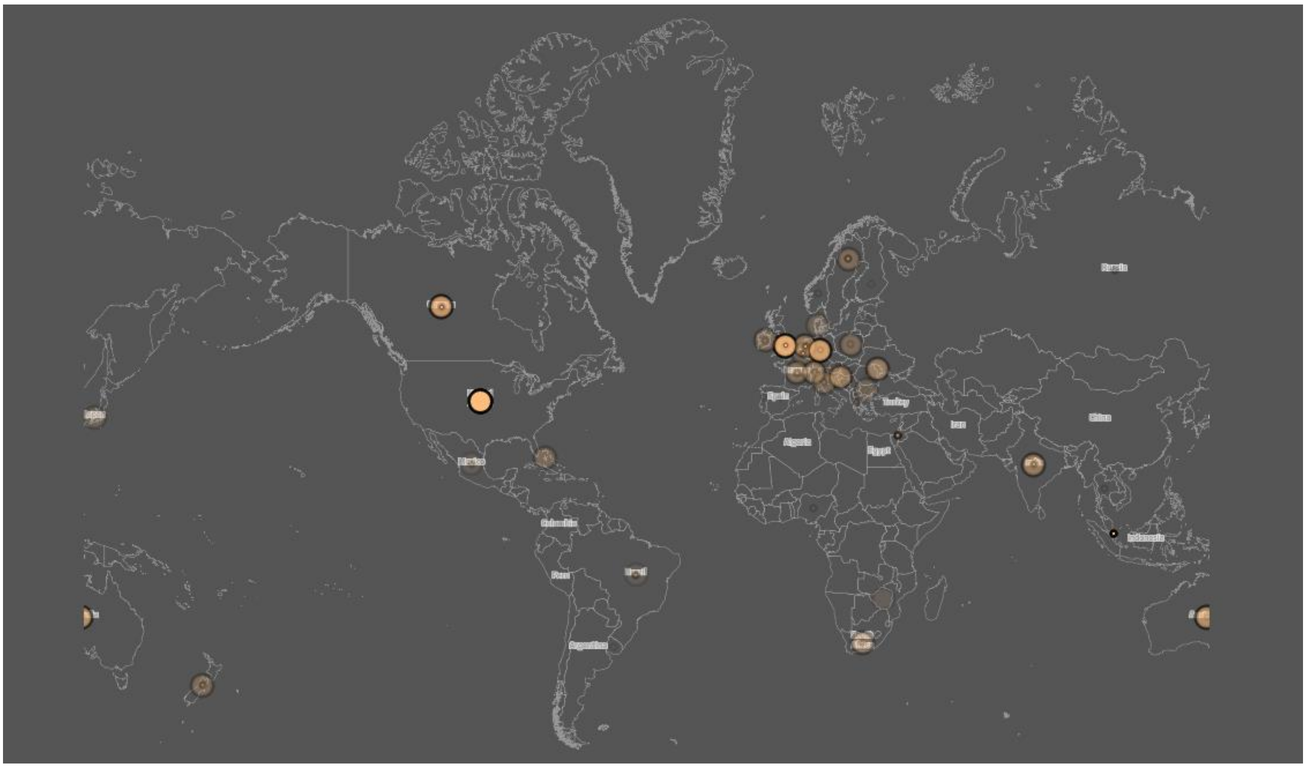
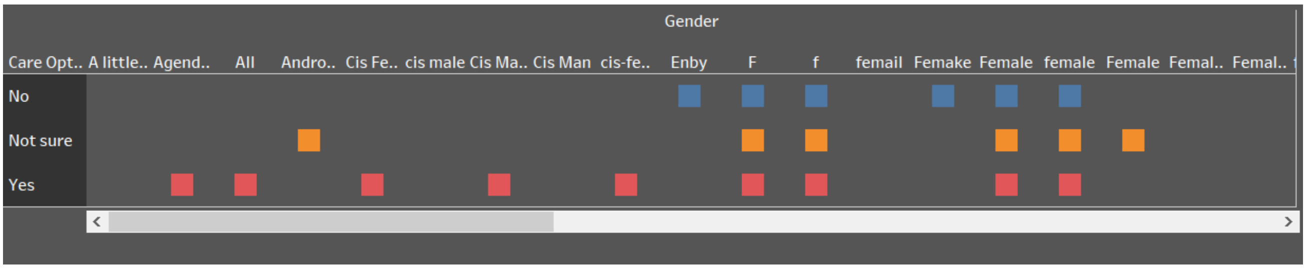


Fig. 4.43 How Many People Have considered Treatment

This map represents the number people who have considered treatment from year 2014-2016. The circles represent the number of people or we can see the population of the people.



	Gender																		
Care Opt..	A little..	Agend..	All	Andro..	Cis Fe..	cis male	Cis Ma..	Cis Man	cis-fe..	Enby	F	f	femail	Femake	Female	female	Female	Femal..	Femal..
No										■	■	■		■	■	■			
Not sure				■							■	■			■	■	■		
Yes		■	■		■		■		■		■	■			■	■			

Fig. 4.44 Care-Option Distribution Based on Gender

This visualization represents the gender v/s care options. The gender which think that care options make any significant difference in one's health is listed above.

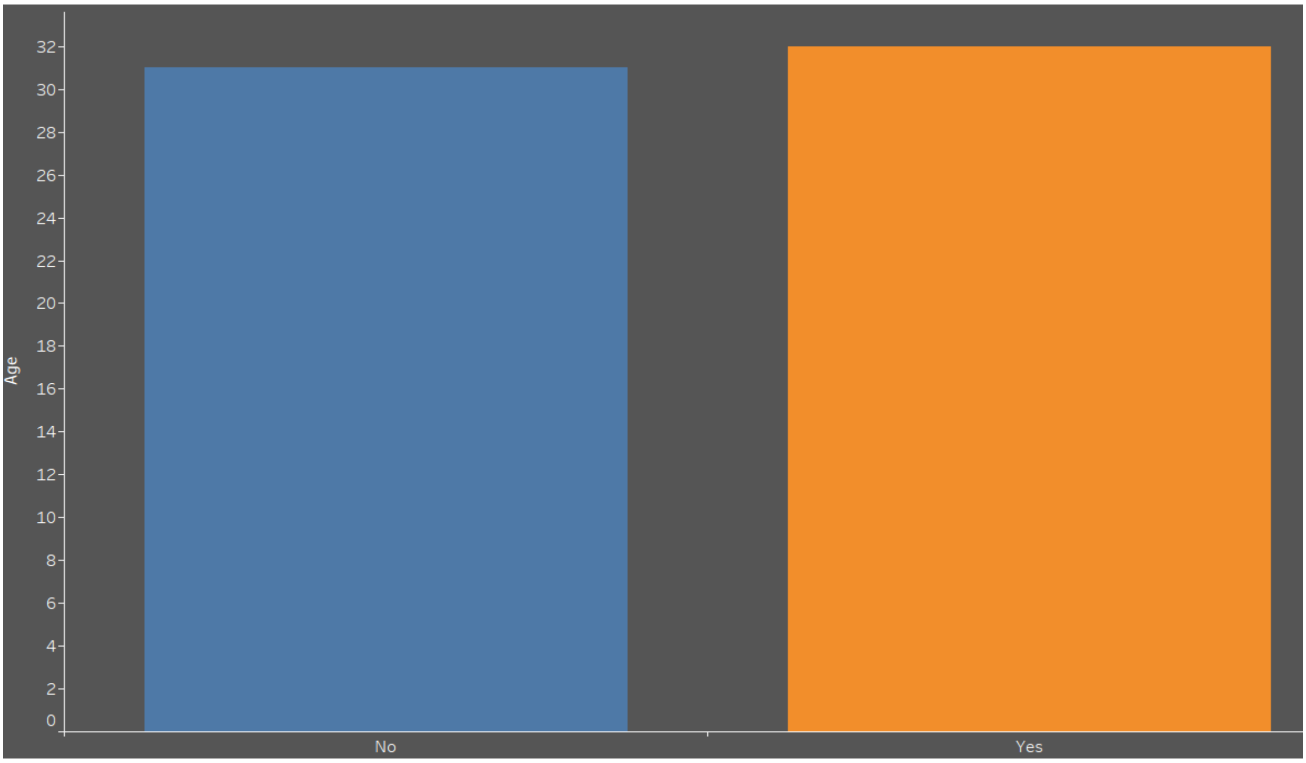


Fig. 4.45 Treatment Based on the Age

Here age is plotted on the y axes and treatment performed is plotted on the x axes. As we can see most of the people have opted for the treatments.

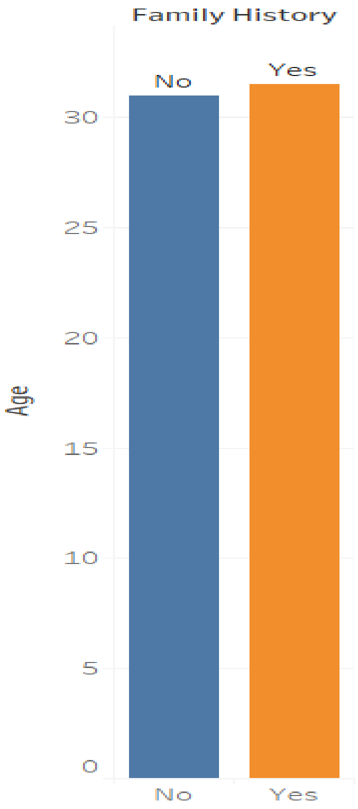


Fig. 4.46 Comparison Based on Family History and Age

On the x axes family history is plotted whereas on the y axes age is plotted. Here it is visible family history of mental health is also the major reason for a person suffering of poor mental health condition.

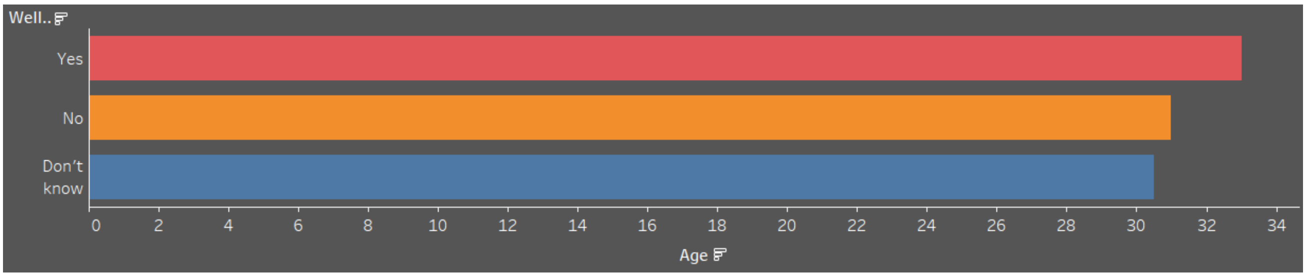


Fig. 4.47 Comparison Based on Age and Well-Being Treatment

The above bar-graph shows the comparison based on well-being and age. Most of the people who have undergone treatment says that treatment was really helpful for making their mental health better.

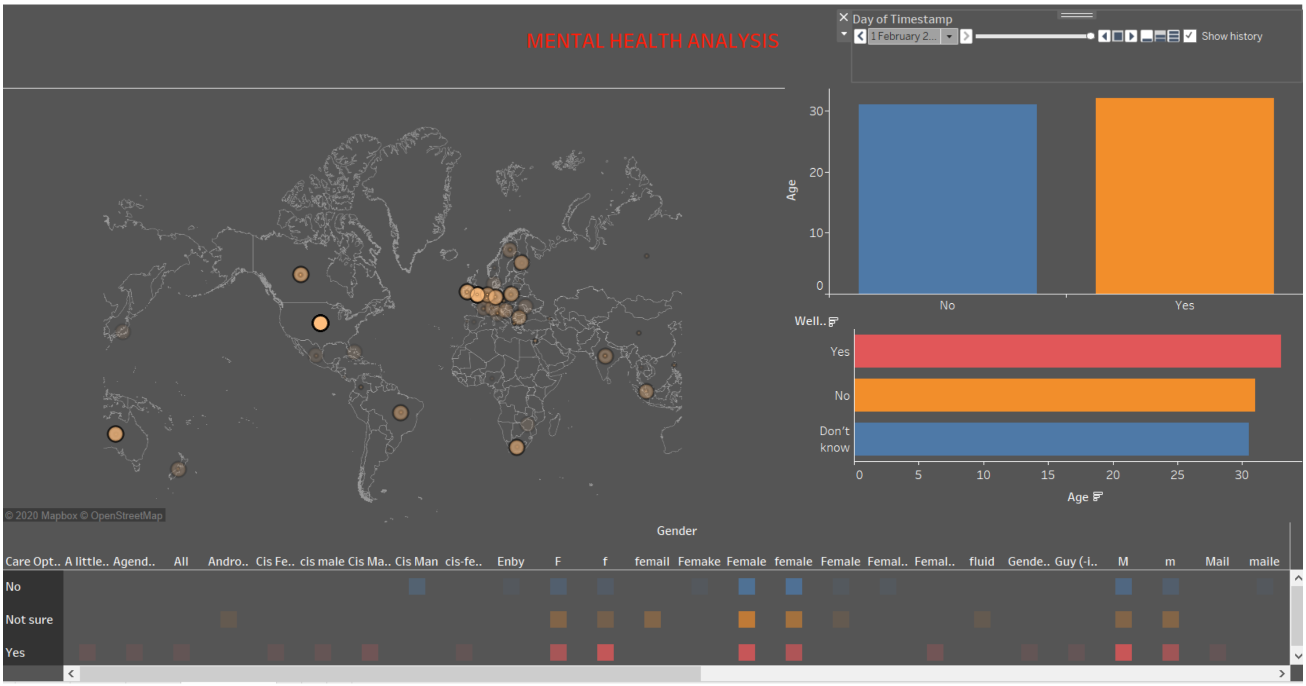


Fig. 4.48 Dashboard of the Mental Health Analysis

By combining all the above visualization, we will get the above dashboard.

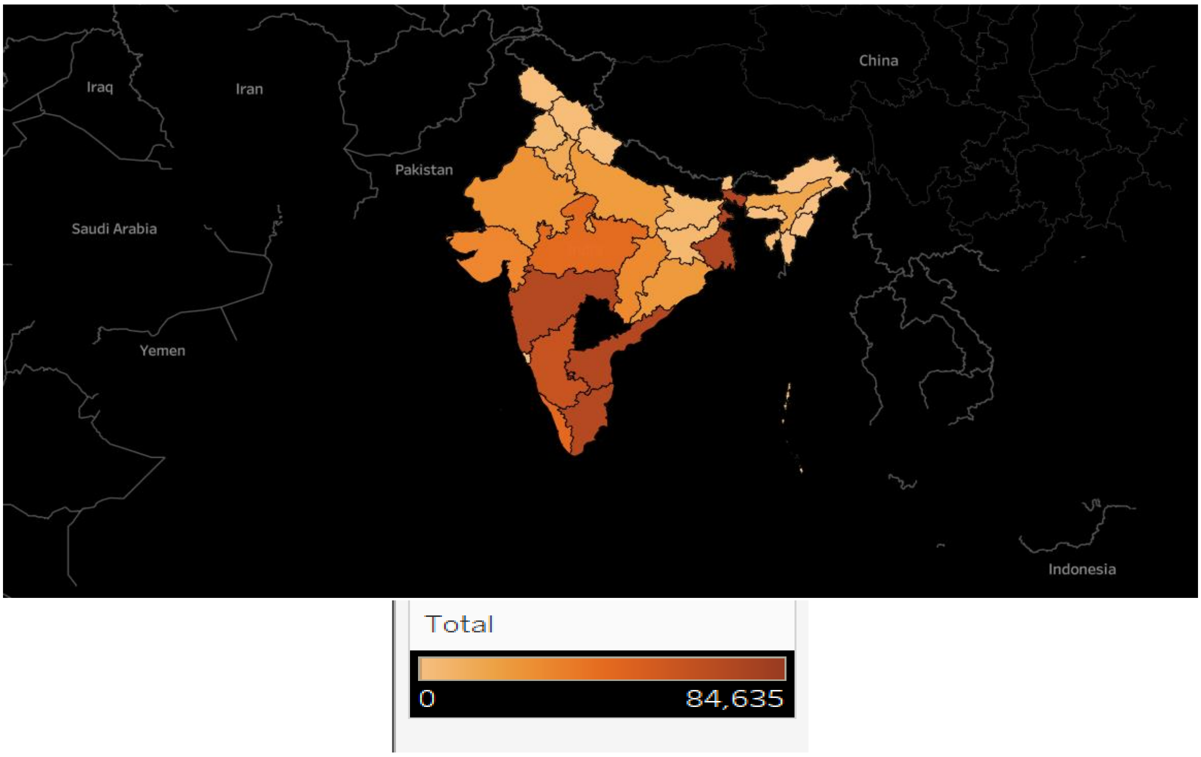


Fig. 4.49 Plotting Total Cases on the Map

The above map represents the total suicide cases in India from the year 2001 to 2012. Also the index is specified below the map from which we can see darker the colour results in more suicide rate.

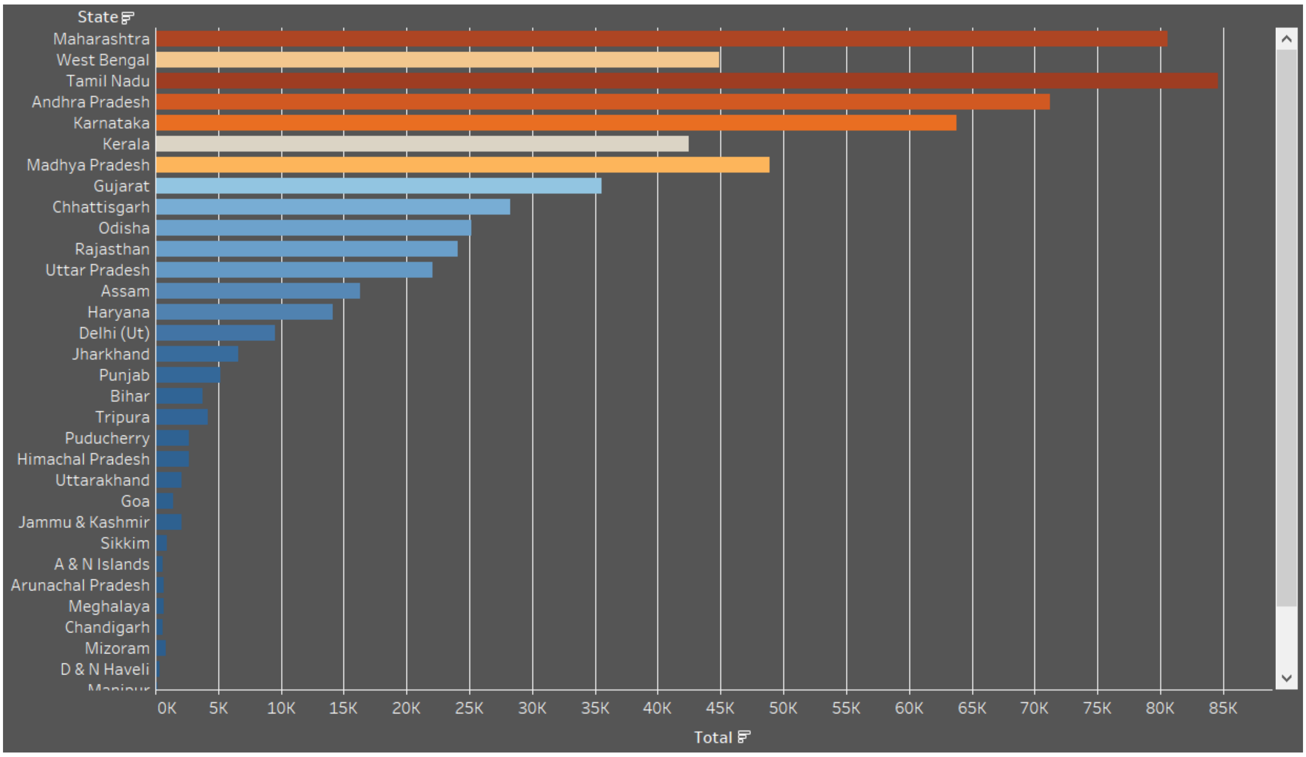


Fig. 4.50 Distribution Based on the States

On the x axes count is plotted along with state on y axes. From the above bar graph, we can see that most of the suicide cases are found in Andhra Pradesh followed by Karnataka and Madhya Pradesh.

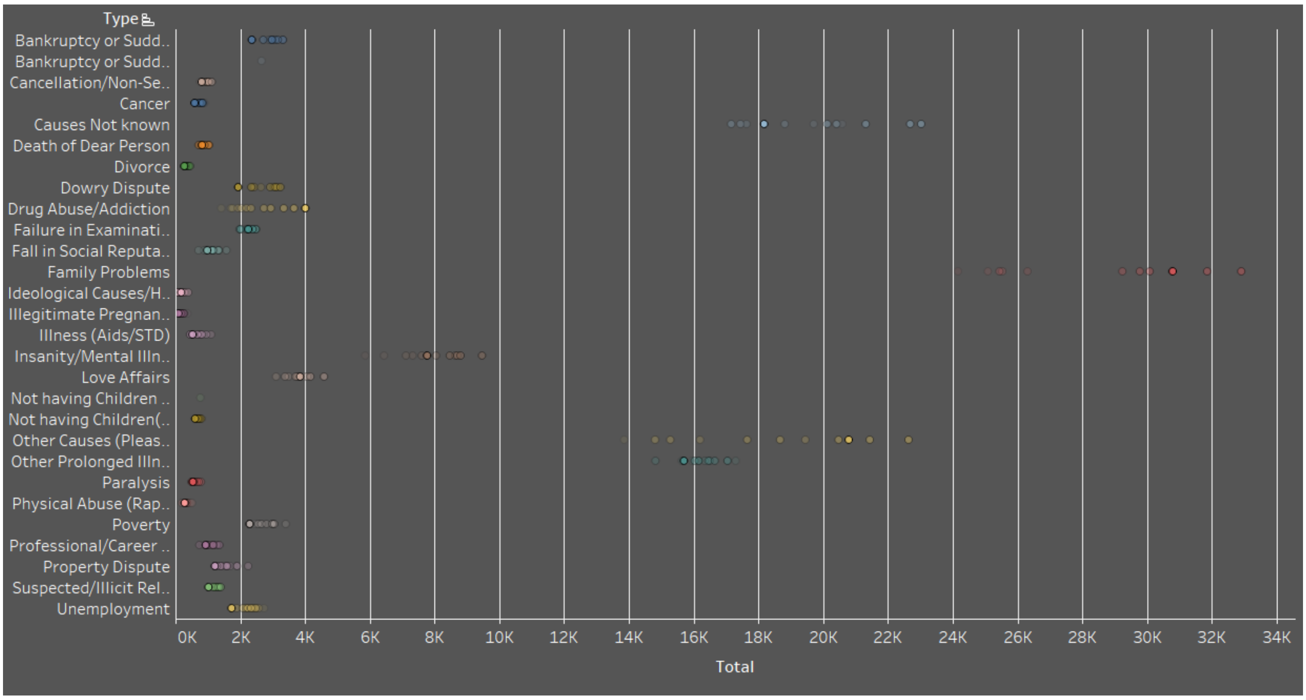


Fig. 4.51 Distribution Based on the Cause for the Suicide

On x axes count is plotted and y axes causes of suicide is plotted. The main reason which can be seen from the graph is family problems and disease.

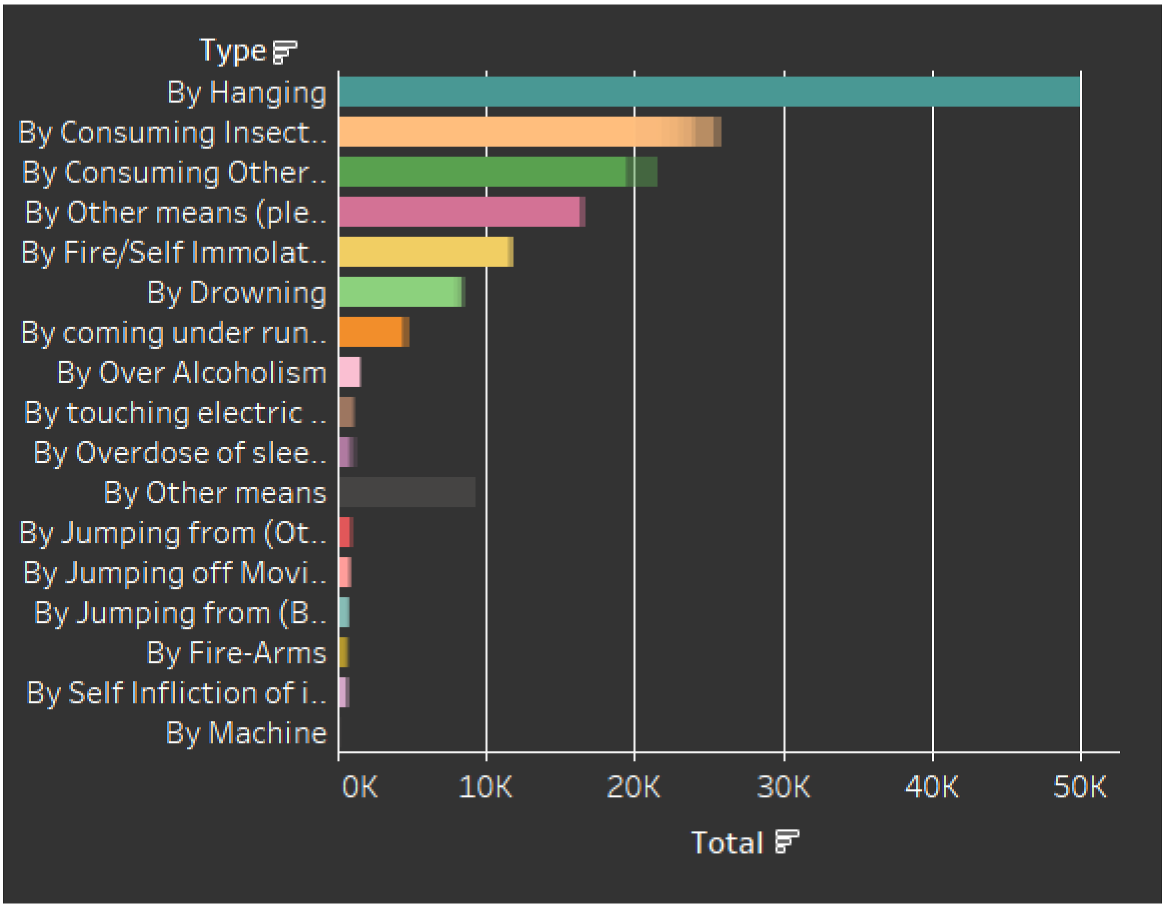


Fig. 4.52 Means Adopted for the Suicide

On x axes count is plotted and y axes means-adopted for suicide is plotted. The main means were by hanging or by consuming poison.

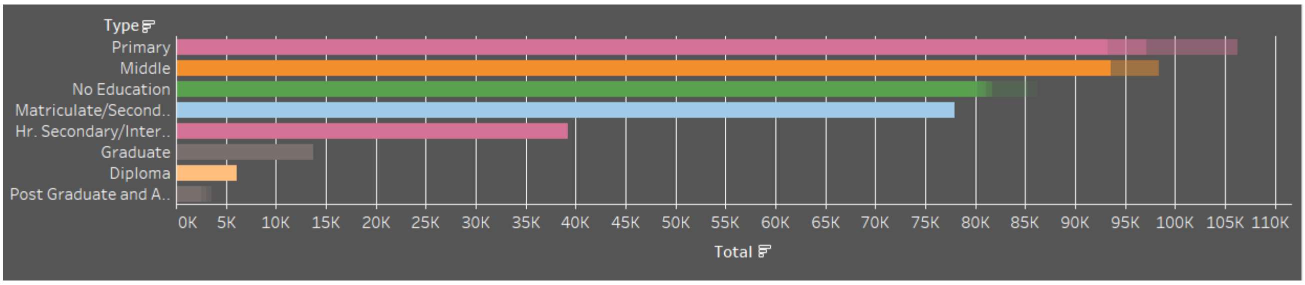


Fig. 4.53 Education-status of the People who Committed Suicide

On x axes count is plotted and on y axes education-status is plotted. Most of the people who have committed suicide are either ill literate or were having primary education or were having middle class education.

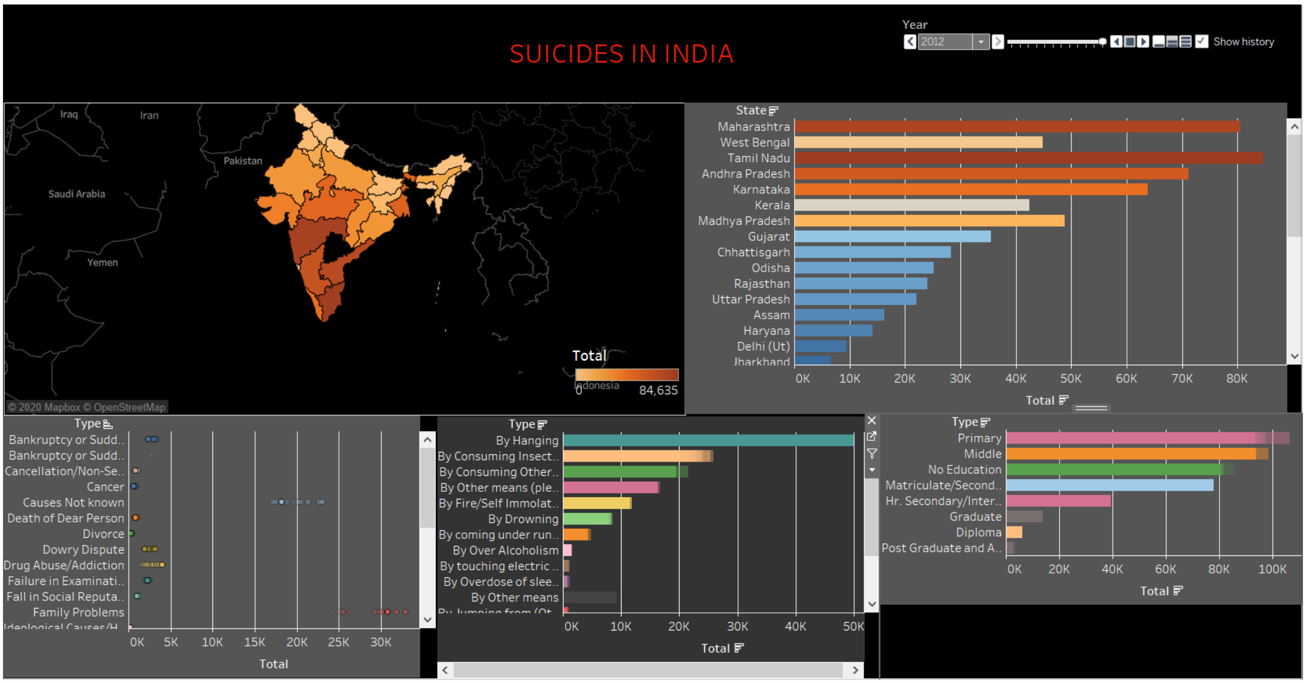


Fig. 4.54 Dashboard for the Suicides in India

By combining all the above visualization, we will get the above dashboard.

4.2. Interpretation of result

- Firstly, a correlation matrix was generated which relates every attribute with all the attributes present in the dataset.
- A bar plot with age on the y axes and count on the x axes was plotted. Here, it was observed that the age-group from 10-20 are more prone to weaker mental health.
- A histogram plot showed that the people from the age group of 10-20 has also gone under treatment for better mental health.
- A comparison was made which gender has gone under treatment. Studies showed that Male has undergone treatment more than that of female.

- A stacked histogram was plotted with age range and gender on x axes and probability on y axes. It showed from 0-20 age group, female with a percentage of 65 are more prone to weaker mental health. In the age group from 21-30, transgender with a percentage of 85 are more prone to weaker mental health.
- Th next histogram showed the relationship between the family problems and mental health. And, to the surprise family problems plays and important role in the mental health.
- The next plot shows whether people where aware that they are having a weaker mental health. Most of the female were sure that what they are going through on the other hand man were neutral in these cases and transgender were mainly not sure what they were going through.
- The next survey shows whether people think that the work is a major reason was their depriving mental health. To this most of the people have answered that sometimes working can be too exhausting and hence it can affect their mental health.
- A bar graph signified what are the major reasons for the poor mental health and the age was in the top of the list followed by gender, family history and care-options.
- A csv file was generated which contains the index number and whether the person needs treatment or not.
- Another dataset was studied which showed the suicides in India.
- When the bar graph was plotted with states on the y axes and count on the x axes it was observed that Andhra Pradesh has the maximum number of suicides in India.
- When a comparative study was done that which gender has the max suicide rate. It was observed that the males are more prone to these.
- A histogram was also plotted to see what age range is more prone to the suicides.
- A pie chart for better understanding of the dataset was used. Males occupied 53.9% of the pie chart while the female occupied 46.1%.
- With years on the x axes and count on the y axes it was observed that the number of suicides has been continuously increasing.
- Most of the suicide cases were observed to be due to family problems and the second reason for this was illness.
- Most of the suicides were performed by means of hanging followed by consuming poison.
- Educational status of the people committing suicides were mostly illiterate.
- The above case study was also done for the union territories.
- Delhi was most affected by this among the union territories.
- Here also the same set of observations were recorded as that of the states of India.

4.3. Inferences from the results and analysis

As we can see from the interpretation that mostly the middle-aged people are prone to weaker mental health. The major reason for this is family issues and the work environment. As the competition is increasing the people have to work harder to be in the race. Due to this their mental health is affected. Also, it was observed that many people were not aware that they are having a poor mental condition. It is quite visible that this is happening since people are not educated. Suicide can be described as the poor health condition. Most of the people who commit suicide are illiterate. They just have their primary degree and the major reason for their suicide was either family problems or due to unemployment. As a person cannot earn living for his family it becomes very difficult to sustain in this world. Hence, to overcome this problem he/she decides to commit suicide. To overcome this firstly the people must be given education in a proper way. Employment opportunities must be made available to decrease this hunger. Various camps must be organized in order to educate the people regarding the mental health.

SUMMARY

5.1. Summary

Here, I have taken two datasets. One which focuses on the mental health and second one was the number of suicides in India. In the first dataset I tried to link various factors which were contributing to the weak mental health. Among these the age was playing the major role. The people from the adulthood were mostly prone to weaker mental health. After that gender was followed. Mostly male were more prone to weaker mental health. Many people were not even aware that they were having poor mental health. The people who have enrolled into wellness program were feeling much relaxed and calm than before. The second dataset focused on the increasing suicide rates in India. Among the states Andhra Pradesh and among the union territories Delhi were the most affected zones. Here, also the males have committed more suicides when compared to females. Most of the people who have committed suicide were either ill literate or were having education till primary section. It is quite evident that education plays a important role in making people aware of these things. Most of them were either unemployed or were farmers. Since people have not completed their education, they don’t get the jobs and then they have to suffer a lot. To end this suffering, they try to end their life. On the other hand, farmers are working day and night to grow the crops but then too he is not getting the proper price for his crops. Due to which he has to suffer a huge loss. In India reason for suicide is poverty. India is still considered to be a developing nation and many people are below the poverty line. Steps are taken by the government to reduce the poverty but since the population so huge all the reformation cannot be made easily. Hence, if we want to eradicate the mental health problem from our world firstly employment must be provided to everyone. This is help to some extent.

5.2. Conclusion

Through this project I tried to make people aware how much important is the mental health. Previously people were not much concerned about it. But slowly they are understanding the importance of mental health. Today, a person doesn’t have enough time to at least it with family and have a quality time. Due to this the distance is increased between them and further they aren’t able to share problems with each other. Hence, in this way people commit suicide for their family problems. Suicide is a leading cause of death and a complex clinical outcome [5]. Older adolescents are more likely to die by suicide than children and younger adolescents [5]. In order to optimize recommendations for engagement in physical exercise so the largest preventive effects can be reached, it is crucial to further test the assumption that exercising outdoors results in more beneficial effects than exercising indoors [1].

5.3. Scope for Further Study

In this project I haven’t covered all the countries of the world. Depending on the country the statistics will change. So, a thorough study be made on every country. By studying this pattern every country can try to reform their policies and make efforts for providing a better mental health to people. Also, by studying theses trends the suicides can also be decreased. In the countries which are having more suicide cases first the research must be started from them. Also, various camps can be organized which can aware the people how important their life as well as mental health is. Policies can be made to support economically weak people as we have seen earlier farmers are more prone to suicide in India. They can be given more money for their crops. And in this way by studying small-small details change can be made.

5.4. References

1. Klaperski, S., Koch, E., Hewel, D., Schempp, A., & Müller, J. (2019). Optimizing mental health benefits of exercise: The influence of the exercise environment on acute stress levels and wellbeing. *Mental Health & Prevention*, *15*, 200173.

2. Wada, M., Suto, M. J., Lee, M., Sanders, D., Sun, C., Le, T. N., ... & Chauhan, S. (2019). University students’ perspectives on mental illness stigma. *Mental Health & Prevention*, *14*, 200159.

3. Vasileiou, K., Barnett, J., Barreto, M., Vines, J., Atkinson, M., Long, K., ... & Wilson, M. (2019). Coping with loneliness at University: A qualitative interview study with students in the UK. *Mental Health & Prevention*, *13*, 21-30.

4. Pumariega, A. J., Rothe, E., & Pumariega, J. B. (2005). Mental health of immigrants and refugees. *Community mental health journal*, *41*(5), 581-597.

5. Cha, C. B., Franz, P. J., M. Guzmán, E., Glenn, C. R., Kleiman, E. M., & Nock, M. K. (2018). Annual Research Review: Suicide among youth–epidemiology,(potential) etiology, and treatment. *Journal of Child Psychology and psychiatry*, *59*(4), 460-482.

6. Masango, S. M., Rataemane, S. T., & Motojesi, A. A. (2008). Suicide and suicide risk factors: A literature review. *South African Family Practice*, *50*(6), 25-29.

7. Bridge, J. A., Horowitz, L. M., Fontanella, C. A., Grupp-Phelan, J., & Campo, J. V. (2014). Prioritizing research to reduce youth suicide and suicidal behavior. *American journal of preventive medicine*, *47*(3), S229-S234.

8. Vijayakumar, L. (2010). Indian research on suicide. *Indian journal of psychiatry*, *52*(Suppl1), S291.

9. Bilsen, J. (2018). Suicide and youth: risk factors. *Frontiers in psychiatry*, *9*, 540.