**Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR Nagar, MSRIT post, Bangalore-54

A Dissertation Report on

Mental Health Survey in Technical Workplace

Submitted by

Aishwarya.K 1MS15CS007

Shalini Sarode 1MS15CS111

Varsha D Vadone 1MS15CS137

Varsha C.M 1MS15CS138

*Bachelor of Engineering in Computer Science & Engineering*

Under the guidance of

Name of the guide Sowmya B J

Designation Assistant Professor

Dept name Computer Science

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

www.msrit.edu, **2017**

Abstract

In this method KNN algorithm is used as a classification method. A survey on mental health in technical workplace is used as a dataset to study the algorithm. We aim to measure the attitudes towards mental health in the IT sphere, and try to examine the frequency of mental health disorders among tech workers. We also wish to determine the major predictors which affect mental health and hence raise awareness and solve the problem based on the result. Using ‘treatment’ attribute as the target variable we’ll try to predict whether a person would undergo treatment or not using our assigned algorithm-KNN.

Introduction

Data Mining is used to gain knowledge based on identifying the patterns and dependencies derived from various data source to solve real world problems.

In supervised learning all the attributes with similar characteristics are classified into a class label which is used to predict the class label of unseen data. Unsupervised training tries to find a hidden pattern in the group of unlabelled attributes.

Data mining methods use classification and regression for predicting the attributes and class labels. In case of classification we group an unknown sample into a known class label. Regression measures the change in dependent variables with respect to independent variables. It is widely used in Machine learning for forecasting and predicting.

Data mining techniques is used in various fields. It is mainly used in Market basket analysis, education, health care, fraud detection, financial banking, corporate surveillance and research analysis.

Dataset Description

The dataset is from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace.

Attribute Description

This dataset contains the following data:

* **timestamp:Time at which the survey was obtained.**
* **Age: Age of an individual**
* **gender: Gender of an individual**
* **Country:Country of an individual**
* **State**: If you live in the United States, which state or territory do you live in?
* **self\_employed**: Are you self-employed?
* **family\_history**: Do you have a family history of mental illness?
* **treatment**: Have you sought treatment for a mental health condition?
* **work\_interfere**: If you have a mental health condition, do you feel that it interferes with your work?
* **no\_employees**: How many employees does your company or organization have?
* **remote\_work**: Do you work remotely (outside of an office) at least 50% of the time?
* **tech\_company**: Is your employer primarily a tech company/organization?
* **benefits**: Does your employer provide mental health benefits?
* **care\_options**: Do you know the options for mental health care your employer provides?
* **wellness\_program**: Has your employer ever discussed mental health as part of an employee wellness program?
* **seek\_help**: Does your employer provide resources to learn more about mental health issues and how to seek help?
* **anonymity**: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
* **leave**: How easy is it for you to take medical leave for a mental health condition?
* **mental\_health\_consequence**: Do you think that discussing a mental health issue with your employer would have negative consequences?
* **phys\_health\_consequence**: Do you think that discussing a physical health issue with your employer would have negative consequences?
* **coworkers**: Would you be willing to discuss a mental health issue with your coworkers?
* **supervisor**: Would you be willing to discuss a mental health issue with your direct supervisor(s)?
* **mental\_health\_interview**: Would you bring up a mental health issue with a potential employer in an interview?
* **phys\_health\_interview**: Would you bring up a physical health issue with a potential employer in an interview?
* **mental\_vs\_physical**: Do you feel that your employer takes mental health as seriously as physical health?
* **obs\_consequence**: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
* **comments**: Any additional notes or comments

Source of Dataset

Kaggle

Mental Health in Tech Survey

Survey on Mental Health in the Tech Workplace in 2014

<https://www.kaggle.com/osmi/mental-health-in-tech-survey>

Data Set size in terms of Bytes and Number of Tuples

Data set size:297 KB

Number of tuples:1260

KNN Algorithm

for all the unknown samples UnSample(i)

for all the known samples Sample(j)

compute the distance between

UnSamples(i) and Sample(j)

end for

find the k smallest distances

locate the corresponding samples

Sample(j1),..,Sample(jk)

assign UnSample(i) to the class which

appears more frequently

end for

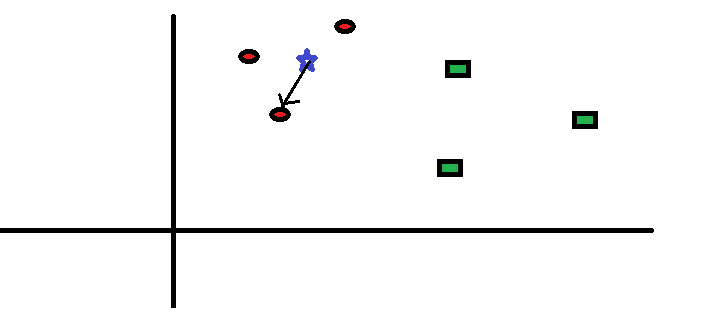
Algorithm Description

KNN algorithm is a lazy learner which is used to classify the data based on nearest training samples. Knn algorithm or K nearest neighbour algorithm is an algorithm used for classification and regression. Generally KNN is used when we have less or no knowledge about the data tuples.

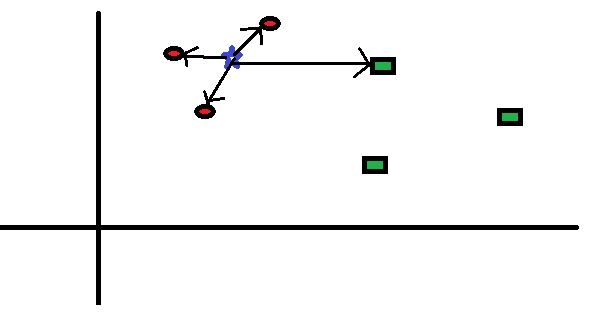
The tuples will be classified based upon its similarity with the surrounding tuples. For an unknown tuple its distance with the surrounding tuples is calculated. Depending upon the value of K it is classified to the respective class labels.

Formulas to calculate distance

K is the number of neighbors for the unknown tuple. The value of K must be carefully chosen. If the K value is too large then it may over estimate the data and it may include outliers and a very small value of K may result in weak estimation of data and may also misclassify the tuples.



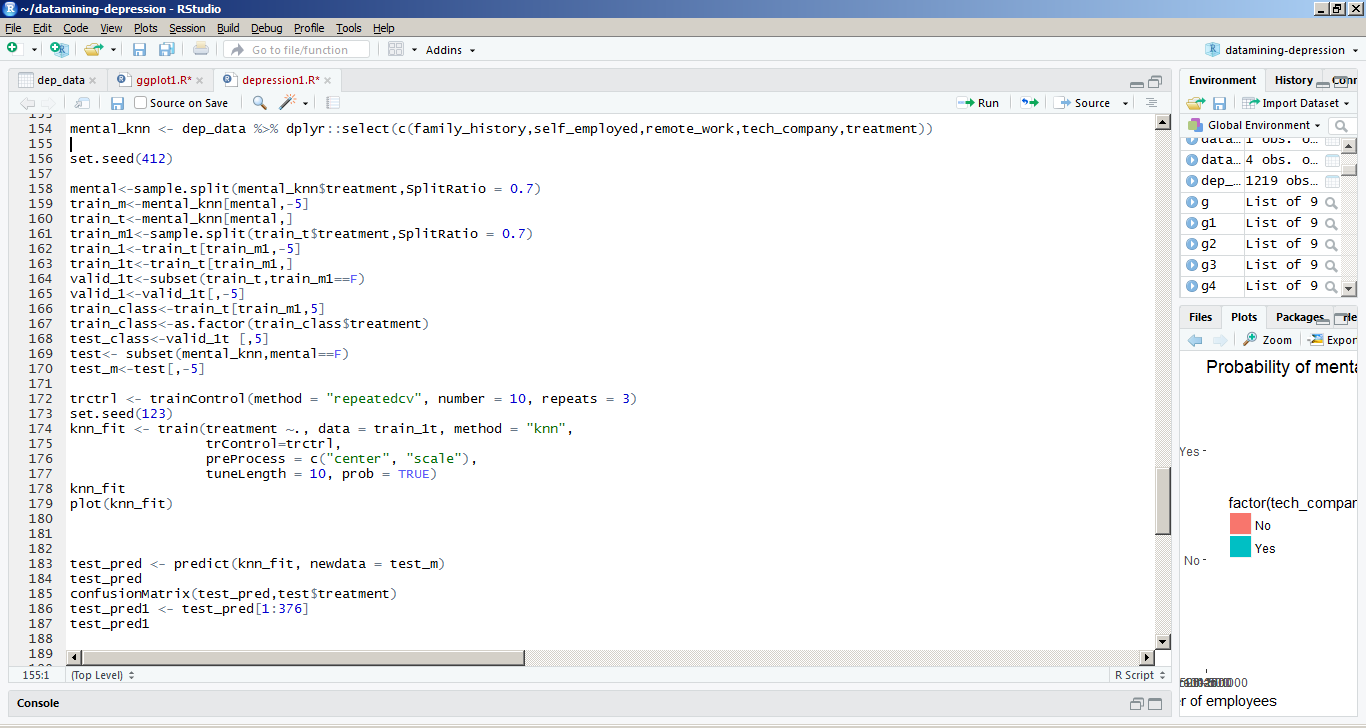
In this example K=1, nearest one sample is used to classify the unknown tuple.

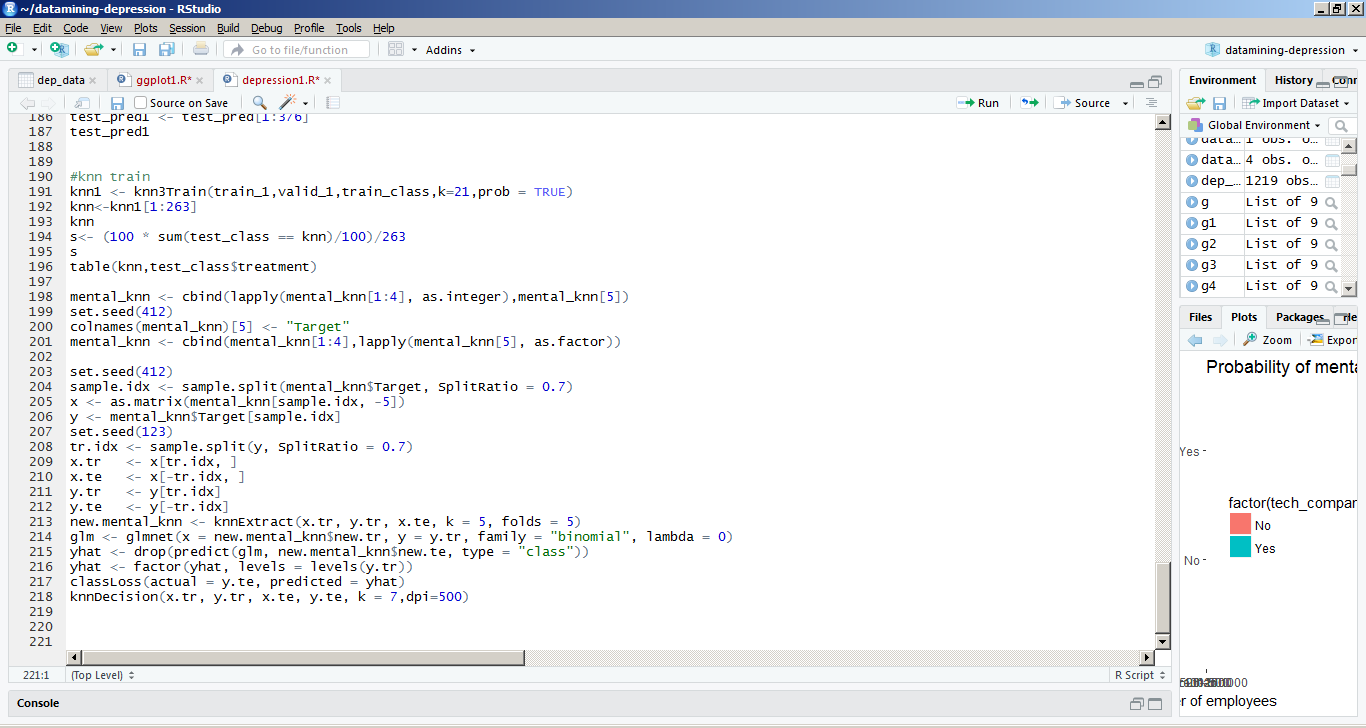


In this example K=4, nearest four samples are used to classify the unknown tuple.

Firstly,normalise the attributes before finding distance to prevent attributes with large ranges from outweighing attributes with smaller ranges. Min-max normalisation can be used. When k=1,the unknown incoming tuple will be grouped to the nearest training tuple class.For categorical attributes compare the tuples. If they are identical then the difference between them is zero.If they are different, then difference is 1. Maximum possible difference is taken for missing attributes.The process is repeated for different values of k and the accuracy is measured. The value of k for which the error rate is minimum is considered.

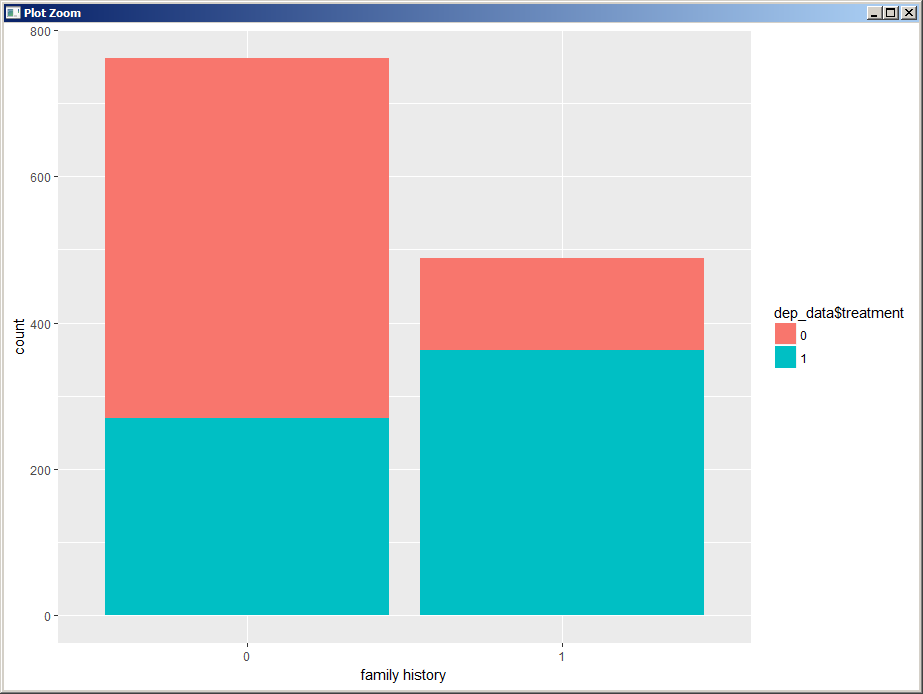
Snapshot of the code





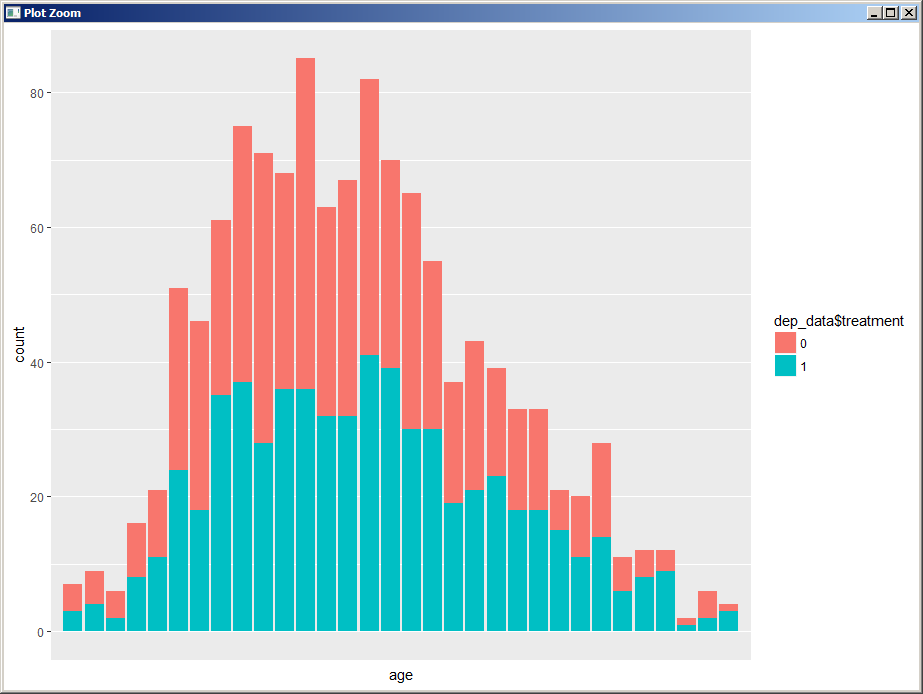
Result Snapshot for each attribute

Family History



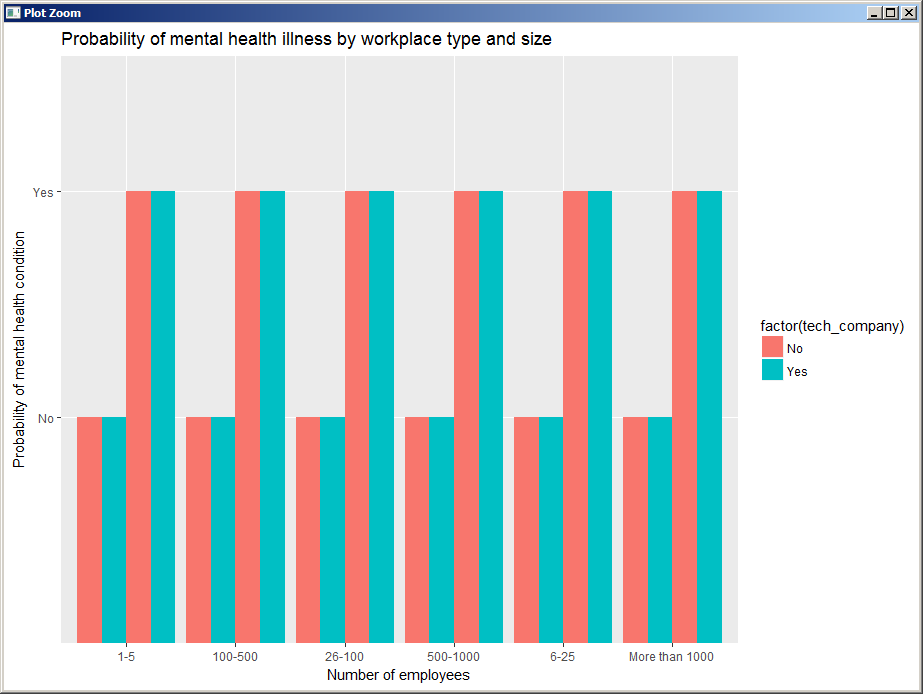
Here in the above graph, 0 represents no treatment is taken and 1 represents treatment is taken.

From the graph,we infer that people with no family history are less likely to take treatment compared to people with family history. Therefore,we can conclude that people with family history,keep track of their mental health and take treatment compared to people with no family history.

Age

This graph represents mental treatment against age groups.

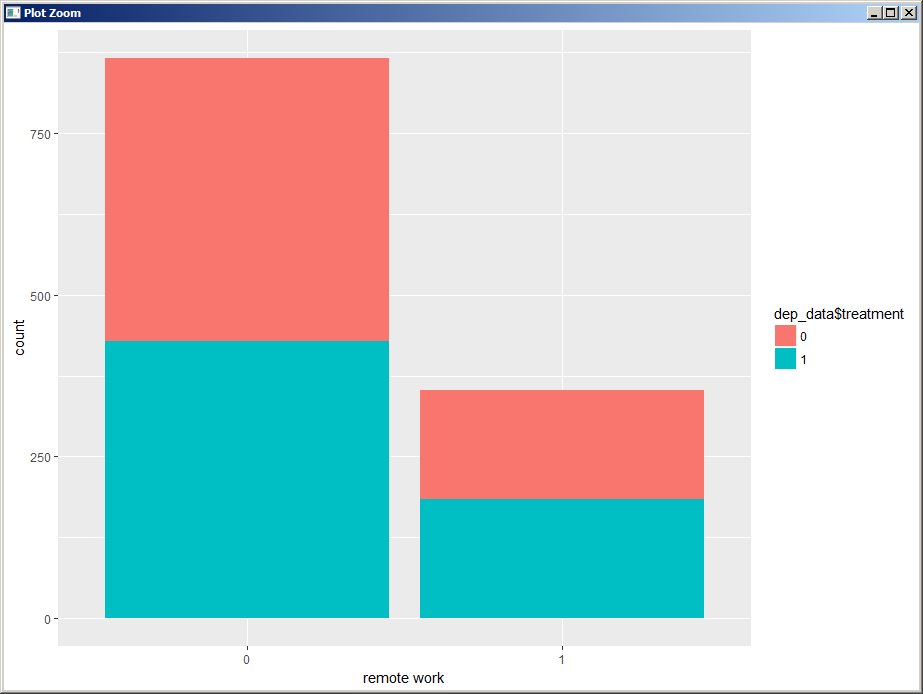
Tech company



Treatment -tech\_company

Form this graph we can infer that both tech company and non tech company have the same probability of taking and not taking treatment.

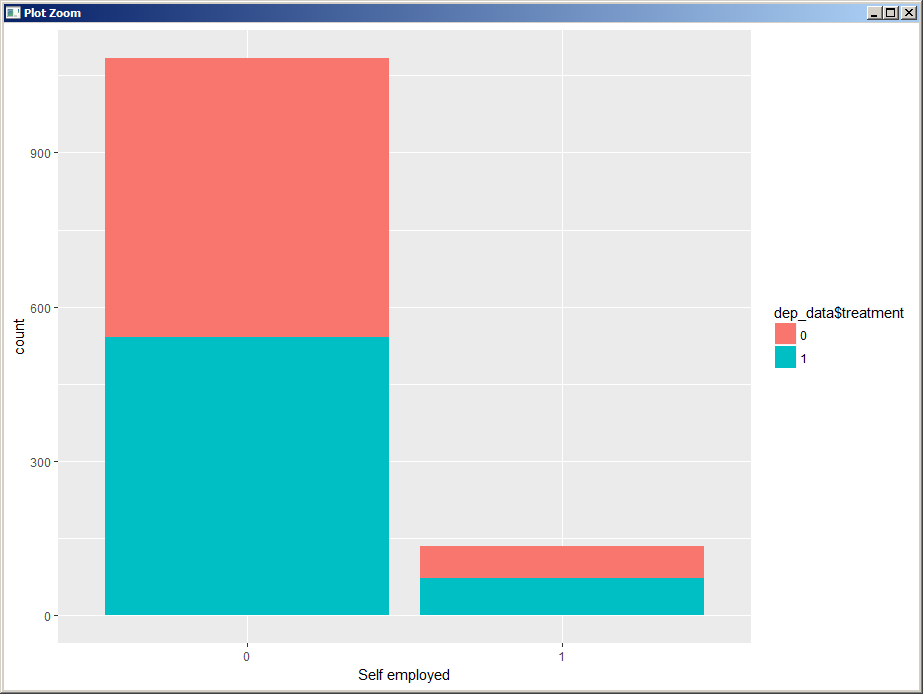
Remote work



Here in the above graph, 0 represents no treatment is taken and 1 represents treatment is taken.

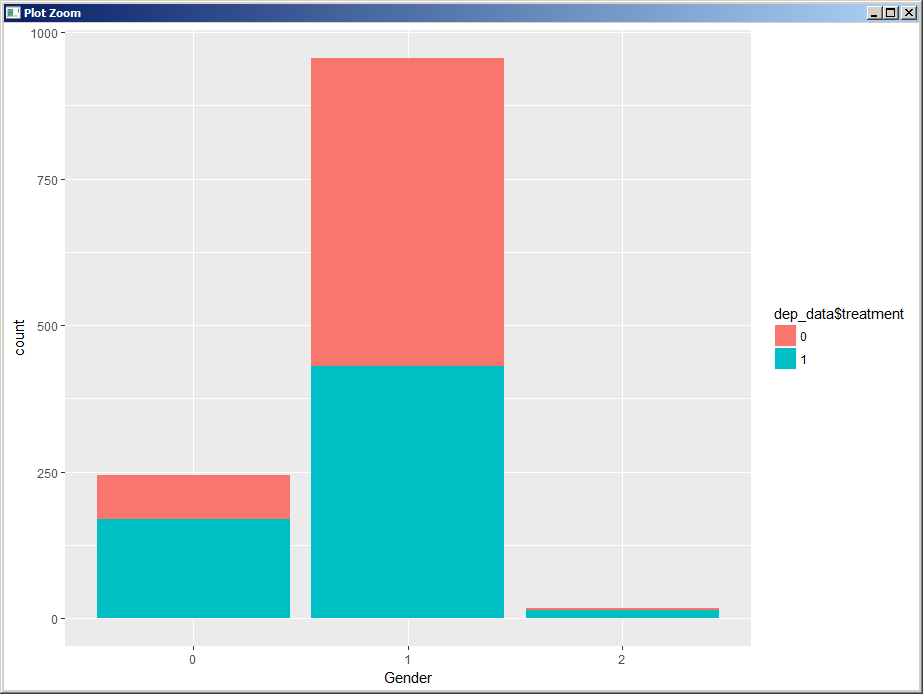
From this graph we can infer that both people doing and not doing remote work have the same probability of taking and not taking treatment.

Self employed



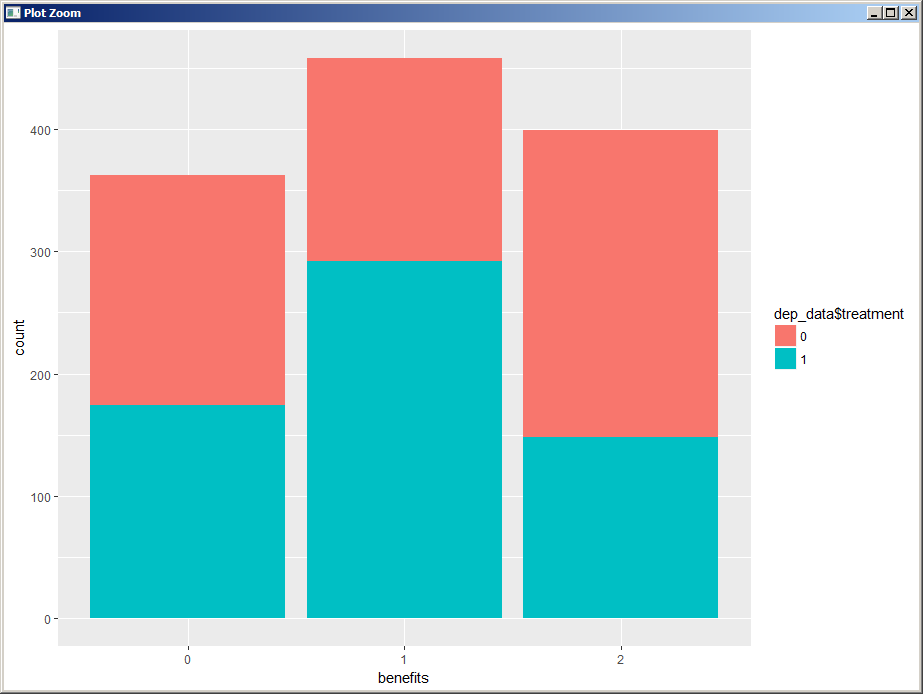
From this graph we can infer that both self employed and not self employed people have the same probability of taking and not taking treatment.

Gender



Most of the transgender people take treatment and they are people who are afraid to talk about these issues even though it affects them more due to social stigma too. The ratio of women who have taken treatment is less . Out of the ppl taken the survey most of them are men and treatment taken ratio is also less.

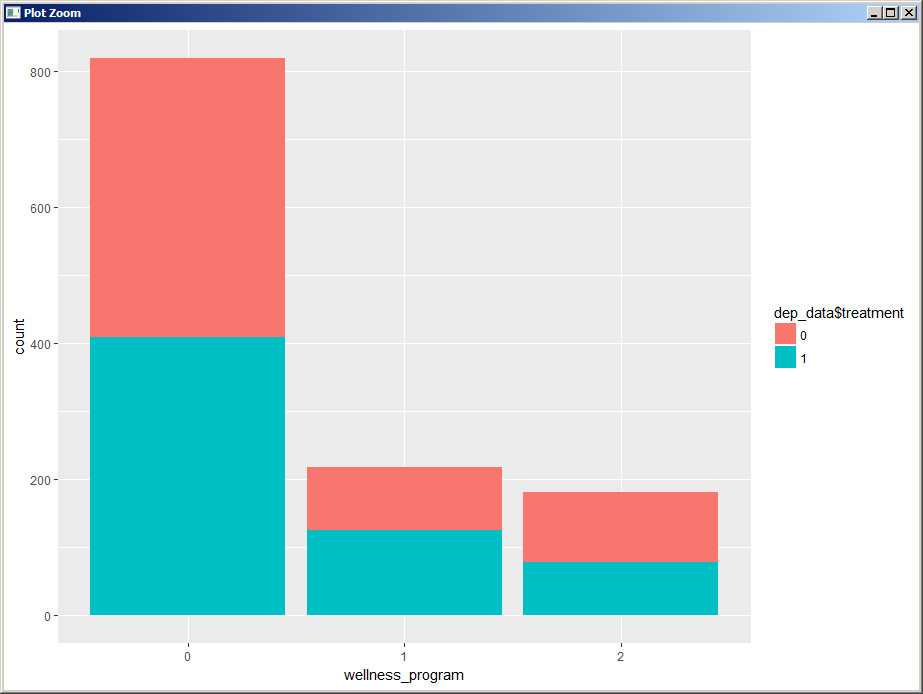
Benefits

****

In this plot the x axis is benefits and y axis is count

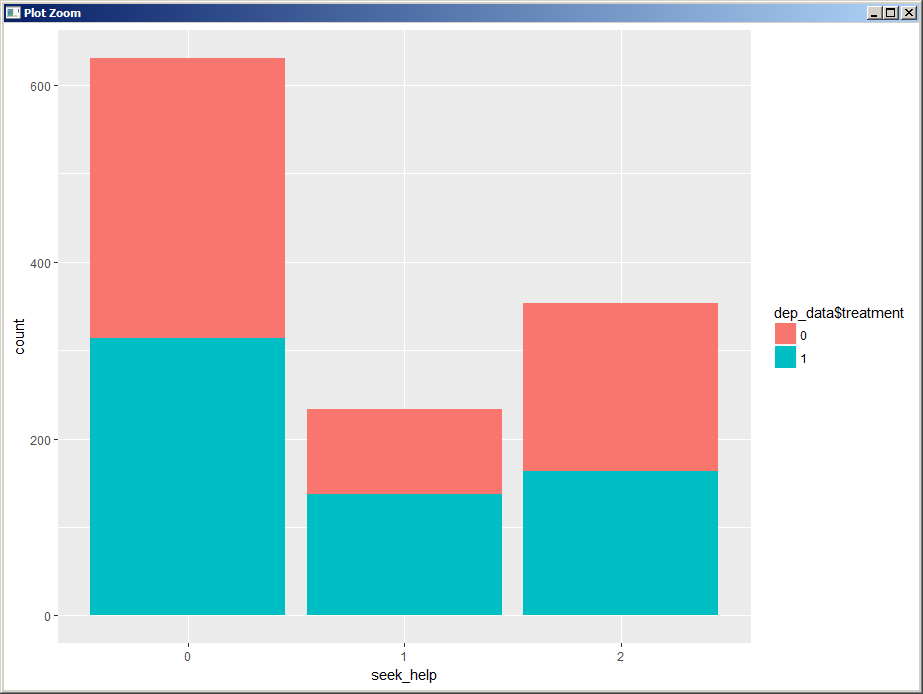
Where benefits is the benefits provided by the employer for mental health.In this plot most of the employers don't provide special benefits for mental health for their employees and most of the people here take treatment for a mental health condition,While many people don't know whether they have such benefits or not hence very less no of people take treatment for their mental health condition.

Wellness program

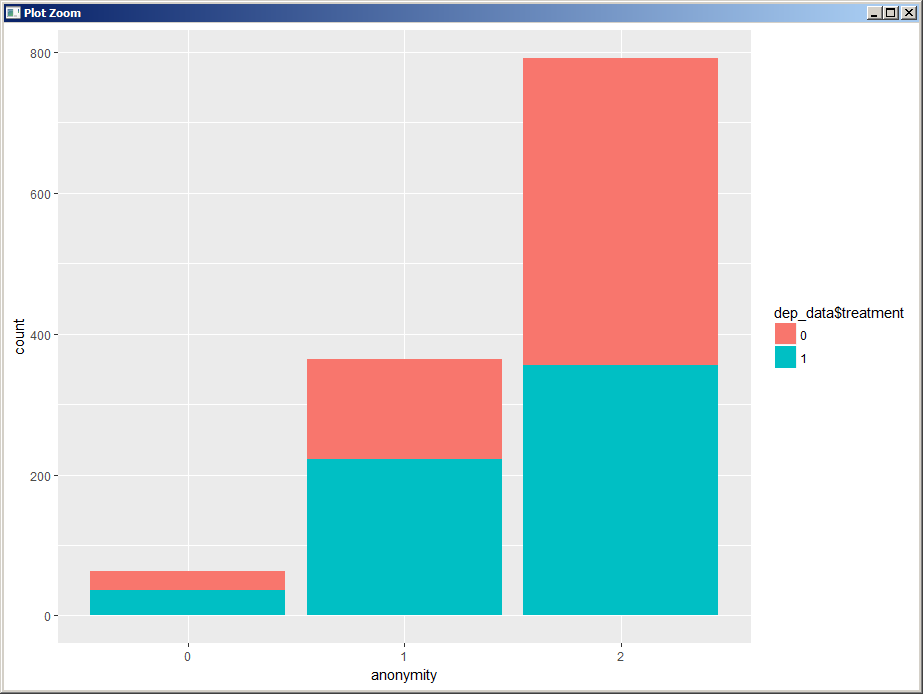


Graph represents treatment against wellness program.Wellness program is a program which helps the employees to take smart and healthy choices.Here red represents no treatment taken and blue represents treatment taken. Employees who don’t choose to include mental health in wellness program are more in number.Among them half take treatment and the other half don’t. Employees who  choose to include mental health in wellness program ,have more than half who takes the treatment and people who don’t know to include in the wellness program ,have less than half who take treatment.

Seek\_help



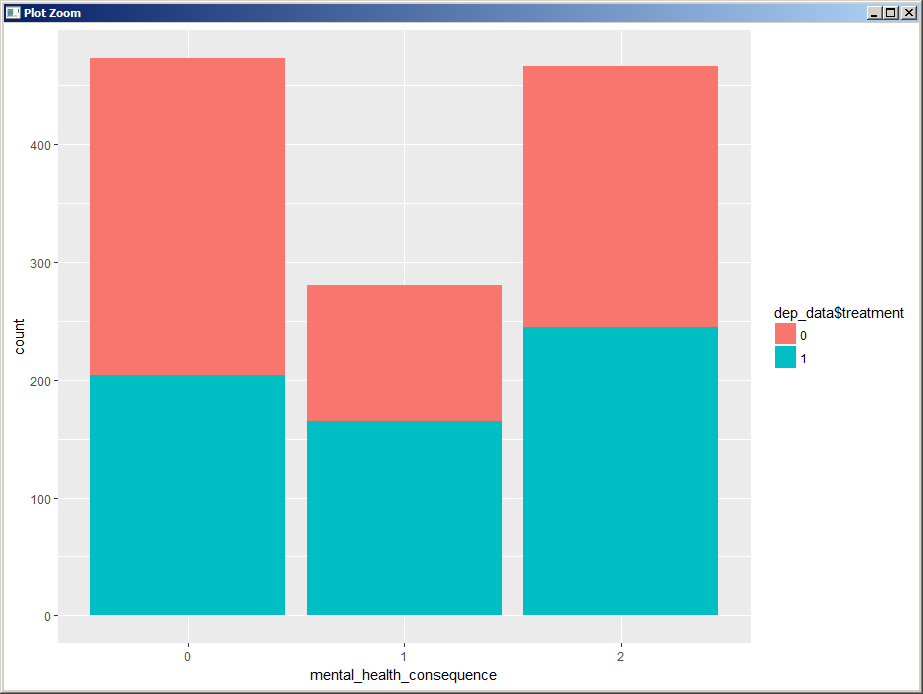
Graph represents treatment against seek help.Seek help means whether the employer takes help in the form of resources to learn more about mental health.Here red represents no treatment taken and blue represents treatment taken. People who don’t seek help is more compared to people who seek help and people who don’t know. Treatment taken by people who don’t seek help is less compared to people who don’t take treatment and who don’t seek help. Treatment taken by people who seek help is more compared to people who don’t seek help and who don’t take treatment.And the people who don’t know to seek help,treatments taken is less compared to treatment not taken.

Anonymity

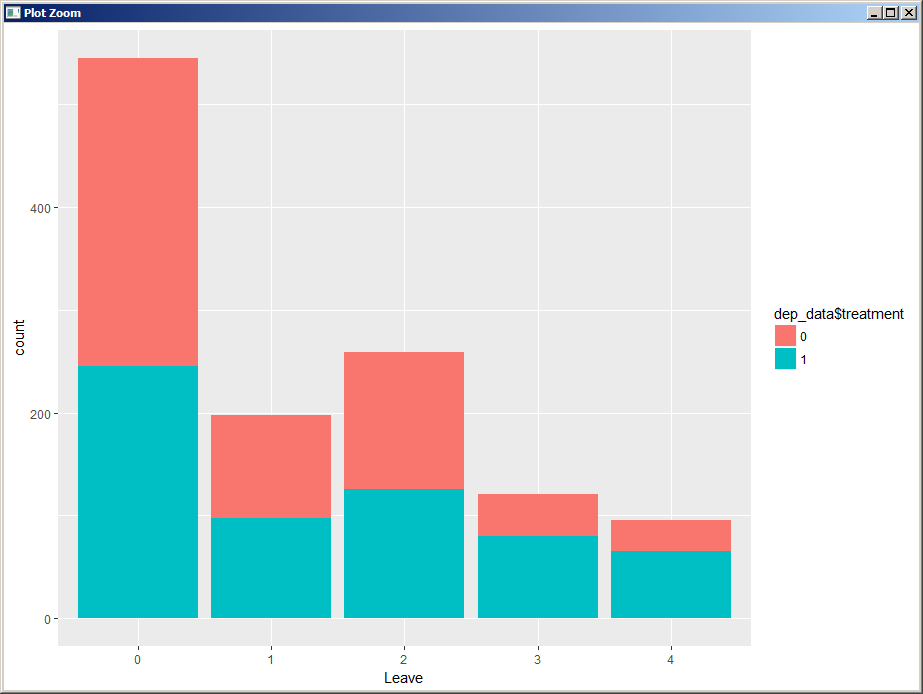
Graph represents treatment against anonymity.Anonymity means  whether being anonymous is protected .Here red represents no treatment taken and blue represents treatment taken.

People who don’t know their anonymity are more compared to people who are sure and and people who deny their anonymity. In People who don’t know their anonymity, More than half don’t take treatment. People who are sure of anonymity ,more than half take treatment. People who deny their anonymity is the least and treatment taken is greater than treatment not taken in that section.

Mental\_health\_consequence



Graph represents treatment against mental health consequence.  Mental health consequence means  whether discussing a mental health issue with your employer would have negative consequences .Here red represents no treatment taken and blue represents treatment taken. People saying yes and don’t know are almost same. In People who say no, taking no treatment is more compared to taking treatment. In People who say don’t know, taking treatment is more compared to taking no treatment. In People who agree, taking treatment is more compared to taking no treatment.

Leave

In this the 0,1,2,3,4 for leave changes

In this plot we plot leave vs count where leave is the ease to take medical leave for a mental health condition here in leave are the responses of people

0-Don't know

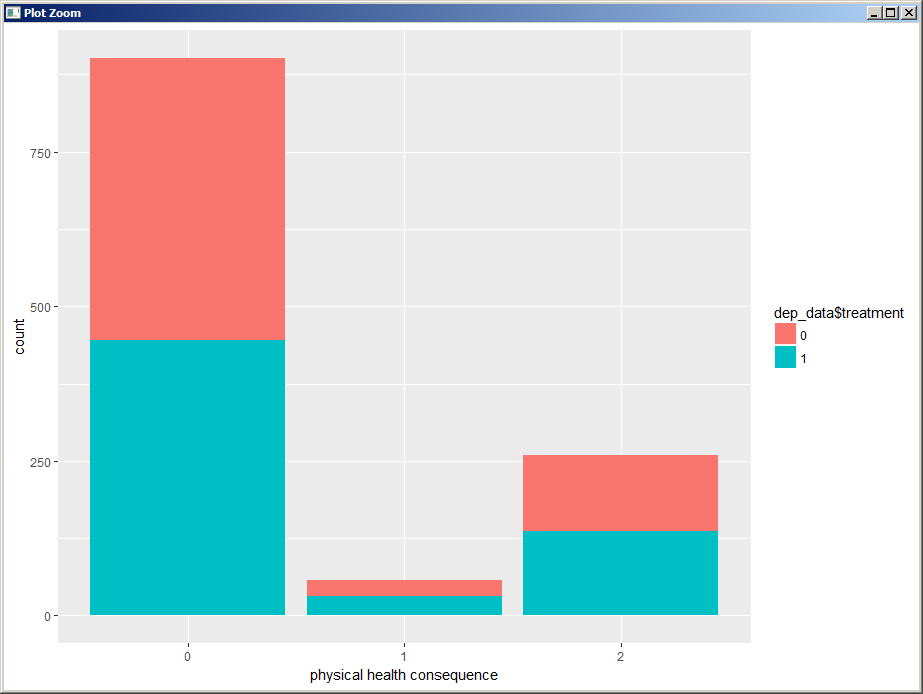
1-Very easy

2-Somewhat easy

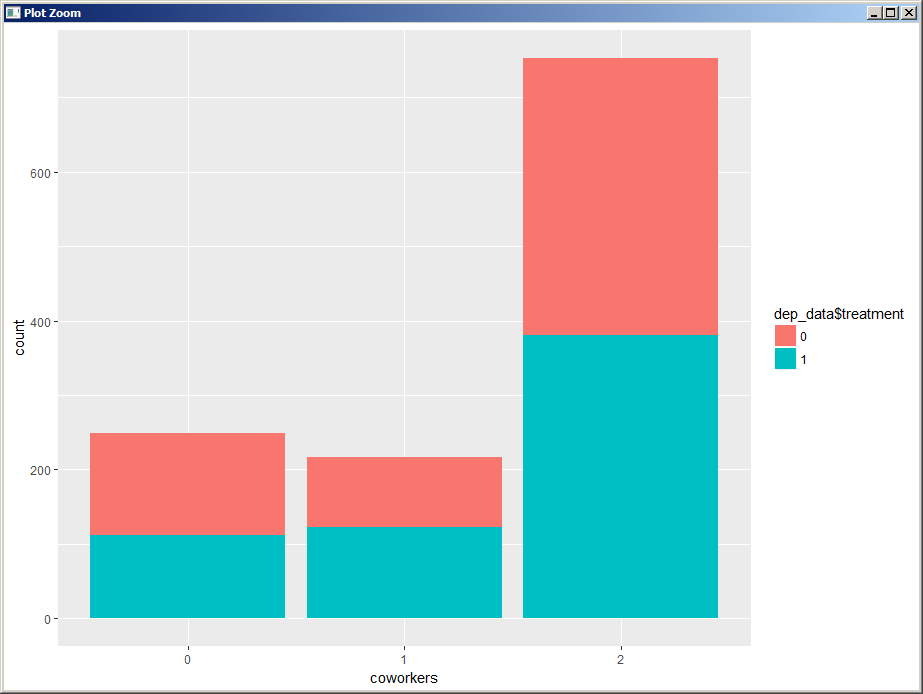
3-Somewhat difficult.

4-Very Difficult

According to the graph most of the people are not aware whether they can avail leave or not.While quite a few people can apply leave somewhat easily.

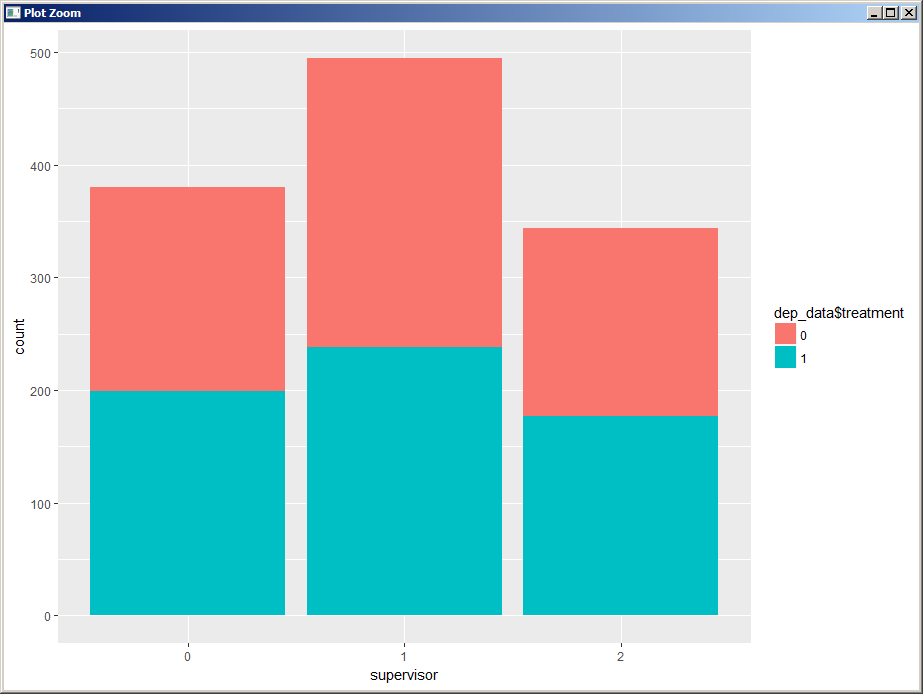
Physical\_health\_consequence

Graph represents treatment against physical health consequence.  Physical  health consequence means  whether discussing a physical  health issue with your employer would have negative consequences .Here red represents no treatment taken and blue represents treatment taken. People saying no are more compared to people agreeing and people who don’t know. In People who say no, taking no treatment is slightly more compared to taking treatment. In People who say don’t know, taking treatment is almost equal to taking no treatment. In People who agree, taking treatment is slightly greater compared to taking no treatment.

Co workers

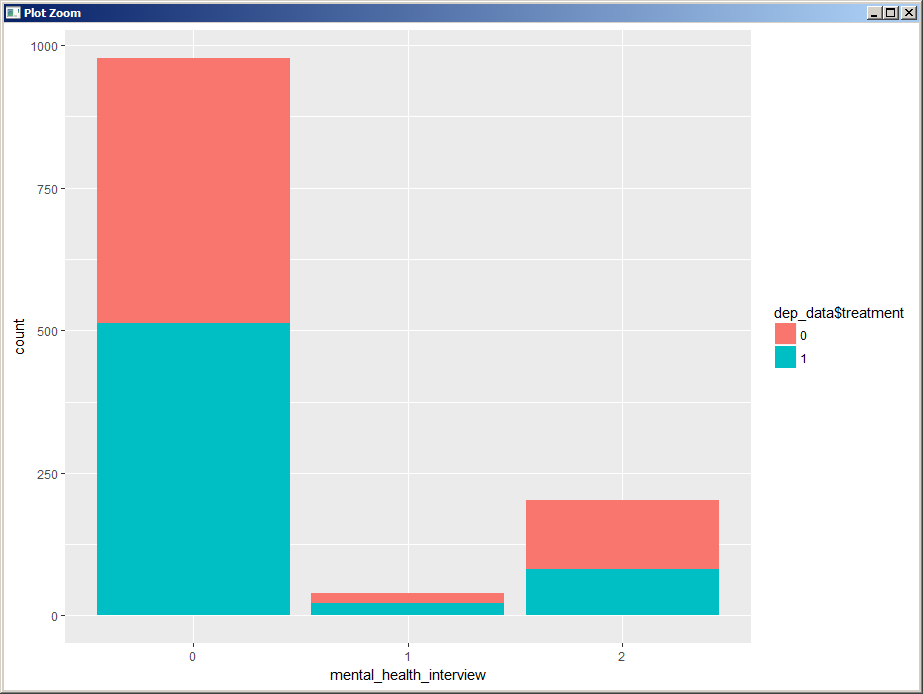
Graph represents treatment against coworkers.  Coworkers in the graph  means  whether you’d willing to discuss a mental health issue with your coworkers. Here red represents no treatment taken and blue represents treatment taken. People saying don’t  know are more compared to people agreeing and people who are disagreeing. In People who say don’t know, taking  treatment is almost equal to taking treatment. In People who say no, taking treatment is slightly less compared to taking no treatment. In People who agree, taking treatment is slightly greater compared to taking no treatment.

Supervisor



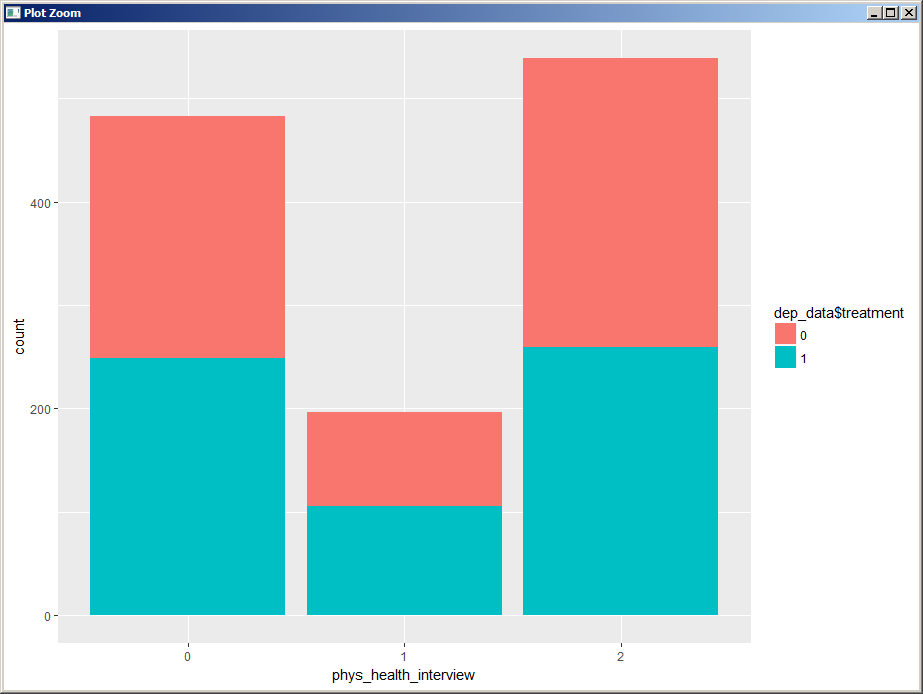
Graph represents treatment against supervisors.  Supervisors in the graph  means  whether you’d willing to discuss a mental health issue directly  with your supervisor .Here red represents no treatment taken and blue represents treatment taken. People saying yes are more compared to people who don’t know and people who are disagreeing. In People who say don’t know, taking  treatment is slightly greater than taking no treatment. In People who say no, taking treatment is slightly greater compared to taking no treatment. In People who agree, taking treatment is slightly greater compared to taking no treatment.

Mental\_health\_interview



Graph represents treatment against mental health interview. Mental health interview in the graph  means  whether you’d willing to bring up a mental health issue with a potential employer in an interview. Here red represents no treatment taken and blue represents treatment taken. People saying yes is the least. In People who say don’t know, taking  treatment is less than taking no treatment. In People who say no, taking treatment is slightly more than taking no treatment. In People who agree, taking treatment is slightly greater compared to taking no treatment.

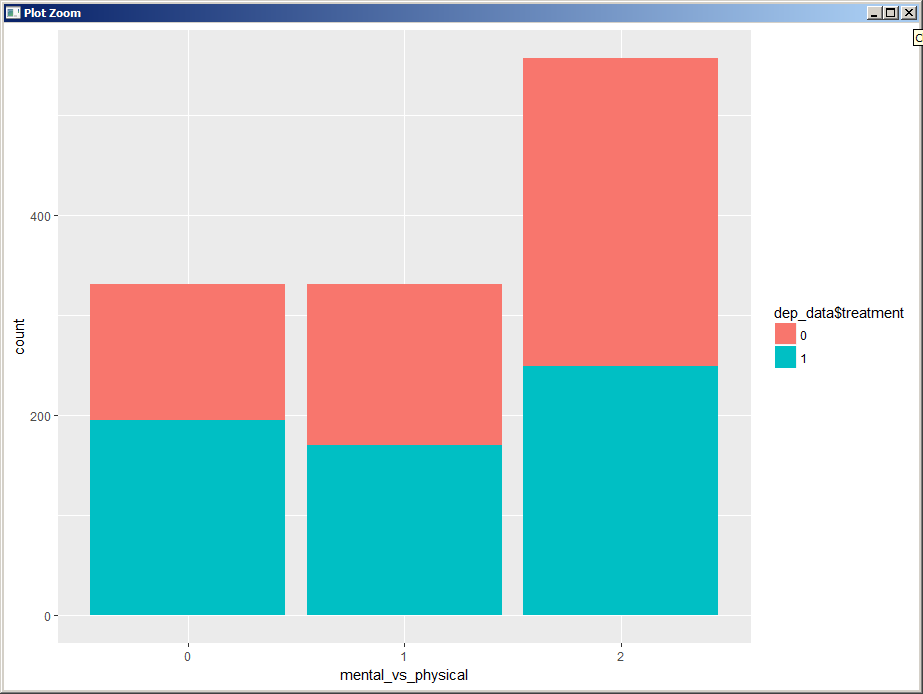
Physical\_health\_interview



This is the plot of phys\_health\_interview vs count where phs\_health\_interview is the response of the people when asked “Would you bring up a physical health issue with a potential employer in an interview?”

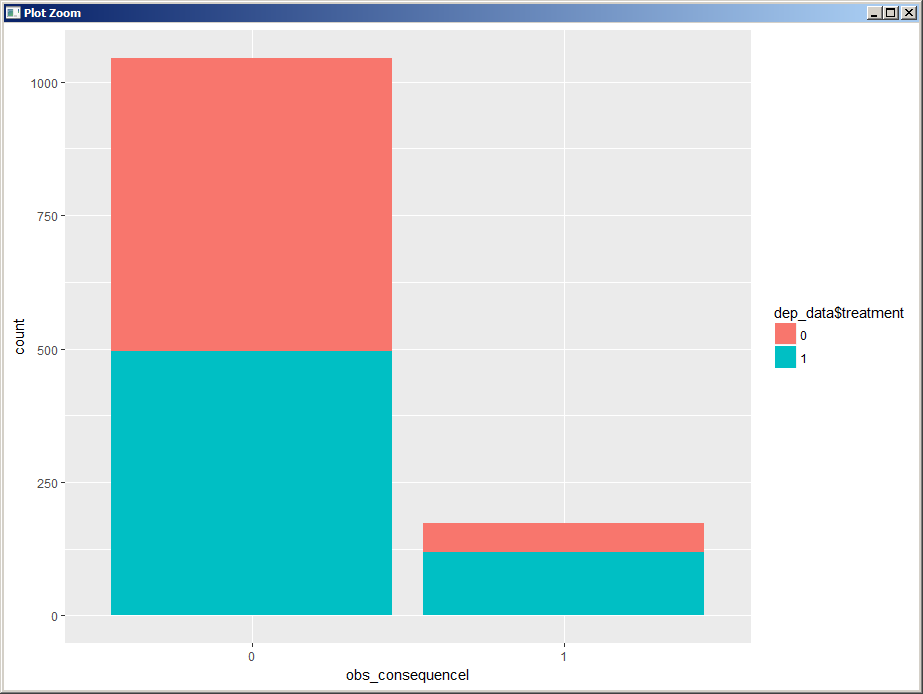
We can infer that most of the people are uncomfortable mentioning their physical illness to a potential employer, whereas very few people are comforatable talking about their physical illness

Mental vs Physical



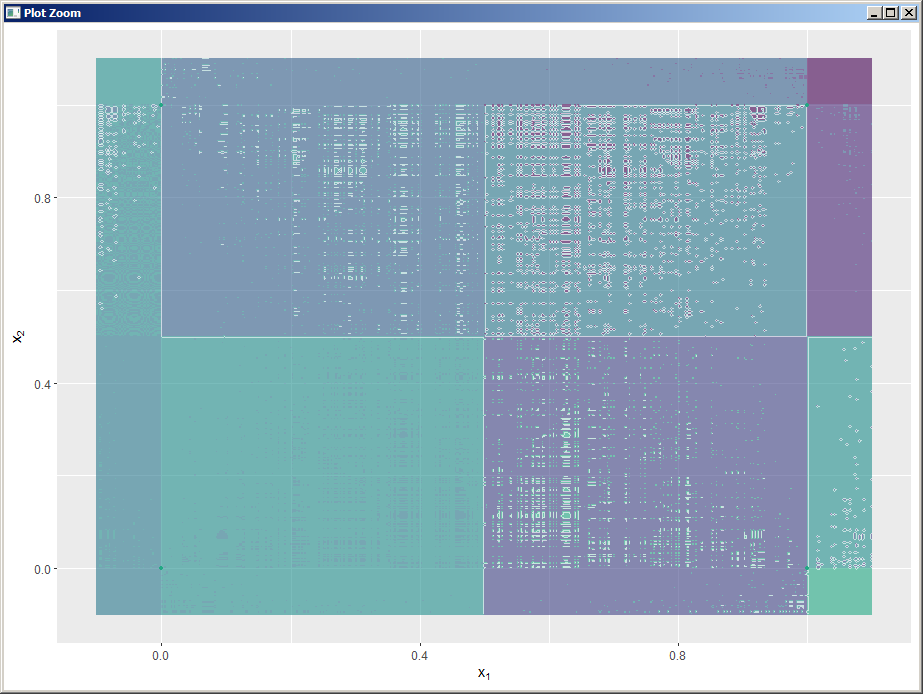
This plot plots mental\_vs\_physical along with count where mental\_vs\_physical is the response given by the people when asked “Do you feel that your employer takes mental health as seriously as physical health?” .Most of the people are unsure about whether their employer considers mental health as important as physical health while equal no of people feel that their employer gives or doesn't give equal importance to mental health and physical health.

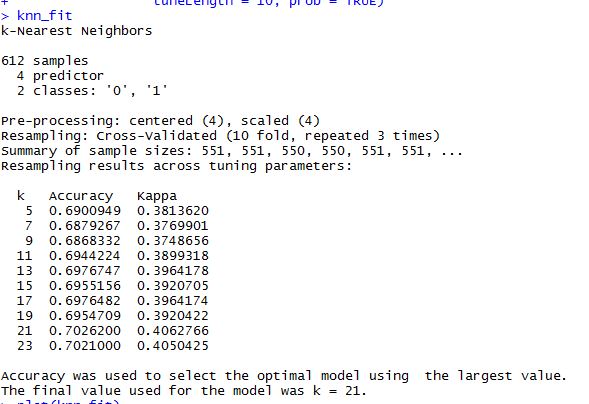
Obs\_consequence



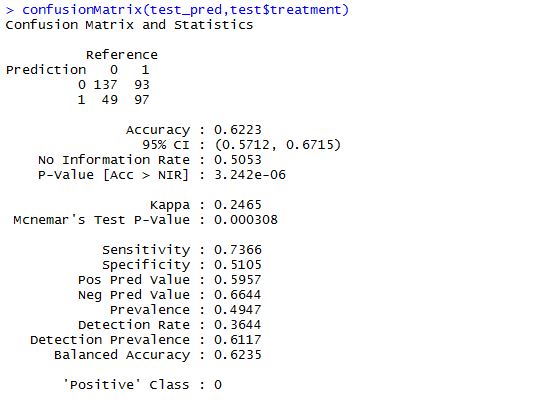
This graph plots obs\_consequence and count according to the treatment where obs\_consequence means observed negative consequences for co workers with mental health conditions in the workplace.From this graph we can infer that most of the people havent seen negative consequences for co workers with mental health condition and the people taking and not taking the treatment are the same. From the set of people who have seen a negative effect most of the people are taking treatment for mental health.

Plot On KNN Decision



Knn Fit

We obtain the highest accuracy with k= 21. Hence for this model k=21 is used.

ConfusionMatrix

We see that  234 tuples have been correctly classified. But the class predicted by our model and the actual class of 142 tuples differ. Hence the accuracy is 62.23 %.

Social Impact

From the survey taken ,we come to know that lots of people undergo mental health problems. We found out that many people are not even comfortable talking about these problems. Males suffer more mental health problems compared to women and transgender. People from age between 20 to 45 undergo various mental health problems cause of stress, anxiety, addictions and worries. Many People are widely affected cause of mental stress which leads them to depression and eventually causing them to hurt themselves physically like suicides. Taking care of mental health has not been given great importance in our society. This survey shows the importance of mental health and puts forth that mental health of a person must be watched and given treatment at the right time as fast as possible to prevent mishaps. This survey also enlightens us the importance that  research work ,treatments & medicines could be improvised to treat mental illness.