Final Project Summary Paper

Cohort B Team 6

Alvaro Chinchayan, Leighton Li, Andrey Lifar, Yoki Liu, Yue Ping, Sherry Zuo
https://github.com/Sherry-Zuo/Capstone-Project
May 1, 2020

I. Problem Statement

In this project, we explored the idea of mental health, with a particular focus on its relation to the workplace environment from a country-level perspective.

We utilize survey data and perform various machine learning models in order to: firstly, predict whether an employee has sought treatment for his/her mental health conditions based on other factors/variables obtained and, secondly, explore whether there is a correlation between mental health and whether the employers provide access to mental health services to their employees.

We truly believe that human resource departments should benefit from our findings in a way that they can start taking into consideration the potential benefits and harms of the existence (or lack of) mental health services.

The importance of research regarding mental health in the workplace environment is extremely relevant to our current situation amidst the COVID-19 pandemic. As employees started working under isolated conditions and constant restructuring, employers must begin to address their employees' mental response to such a rare occurrence in history. Our project would provide a general view of mental health conditions at the workplace to provide better insight and promote better conditions in the workplace environment.

In this project, we will try to answer the following questions: Is it possible to predict whether an employee has sought treatment for his/her mental health conditions based on other factors/variables obtained? Is there a correlation between mental health and whether employers let their employees take a break, or allow them to work from home (remotely)? Is there a relationship between mental health treatment and happiness scores and suicide rates? Could we give advice based on the obtained results?

II. Datasets

In order to conduct and address the objective at hand, we utilized the following datasets:

1. OSMI (Open Sourcing Mental Illness) Mental Health in Tech Survey

https://osmihelp.org/research

Survey from 2014 - Main dataset used for machine learning.

Survey from 2016 to 2019 - Supplemental data used to compare yearly trends.

2. WHO Suicide Statistics

https://www.kaggle.com/szamil/who-suicide-statistics

Happiness scores computed using several explanatory factors such as GDP per capita, degree of freedom, and life expectancy

3. World Happiness Report

https://www.kaggle.com/unsdsn/world-happiness

2015-2019 data

Includes country-level suicide data as well as statistics of gender, population and GDP

III. Exploratory Data Analysis (EDA)

After thorough cleaning of the 2014 survey response data, we obtained 896 rows and 24 columns. As our focus is on a country-level data, it is important to note that we removed countries that did not have sufficient representation (less than 10 survey responses) which shifted our focus to the following nine countries: Australia, Canada, France, Germany, India, Ireland, Netherlands, United Kingdom and the United States of America. In figure 1 we can see a visual representation of this feature. There are additional countries highlighted, but the aforementioned nations were the focus of our research.

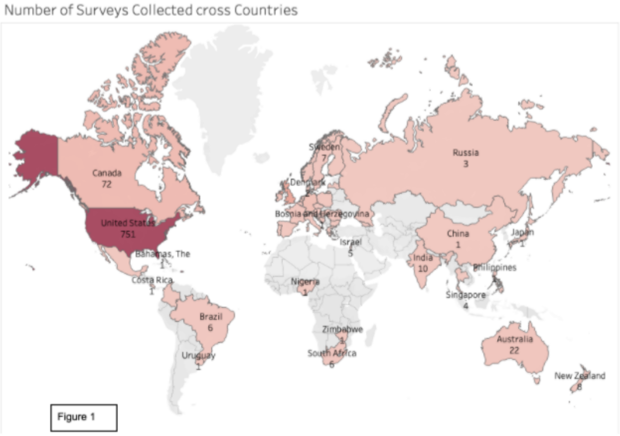


Figure 1

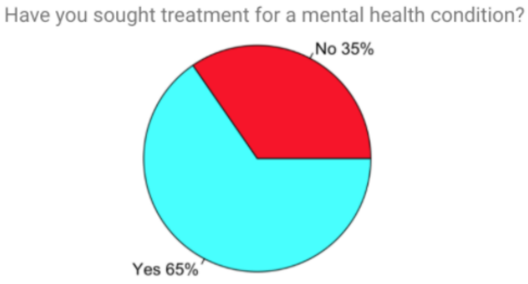
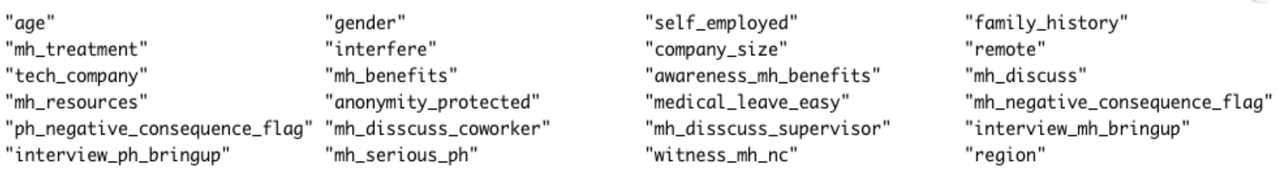


Figure 2

One of our goals is to predict whether an employee would seek mental health treatment. Figure 2 shows the distribution of survey responses to the question “have you sought treatment for a mental health condition”.

The remaining columns are survey responses to individual questions pertaining to mental health (mh abbreviate question including mental health) such as whether the individual have discussions about mental health in the workplace (mh_discuss_coworker), whether they mention mental health issues in an interview (interview_mh_bringup) and others.



```
"age"                          "gender"                   "self_employed"              "family_history"
"mh_treatment"                 "interfere"                "company_size"               "remote"
"tech_company"                 "mh_benefits"              "awareness_mh_benefits"      "mh_discuss"
"mh_resources"                 "anonymity_protected"      "medical_leave_easy"         "mh_negative_consequence_flag"
"ph_negative_consequence_flag" "mh_disscuss_coworker"     "mh_disscuss_supervisor"     "interview_mh_bringup"
"interview_ph_bringup"         "mh_serious_ph"            "witness_mh_nc"              "region"
```

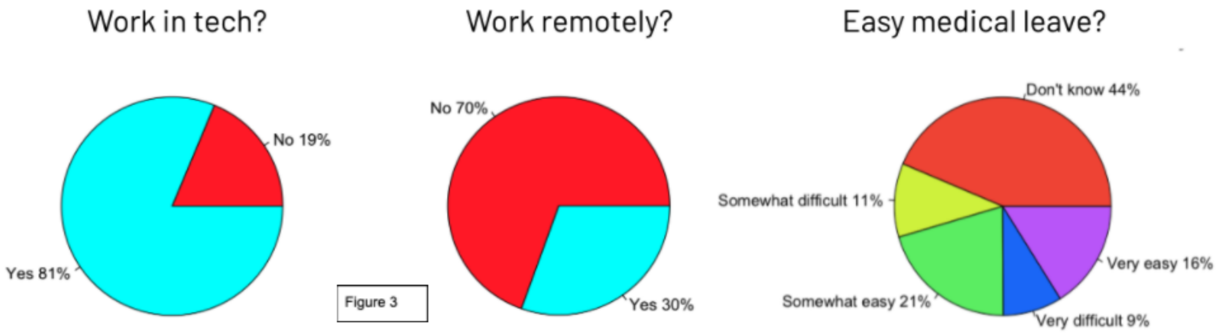


Figure 3

Some features of particular importance in the workplace environment according to the data are illustrated in figure 3. We found that most individuals surveyed work in tech companies, don't work remotely and are uncertain of how feasible it is to request medical leave.

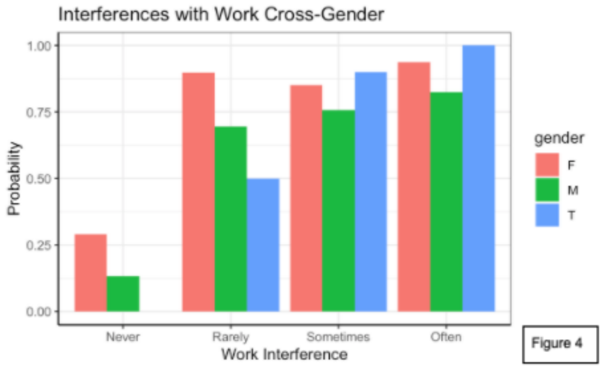


Figure 4

Further exploration of our data led us to understand what the probability was of medical treatment dependent on certain variables. Figure 4 illustrates the rate of seeking treatment for respondents who either feel mental health often/sometimes/rarely/never interferes with work, across different genders.

Figure 5 illustrates the portion of respondents who sought treatment and worked for either tech or non-tech companies across different company sizes.

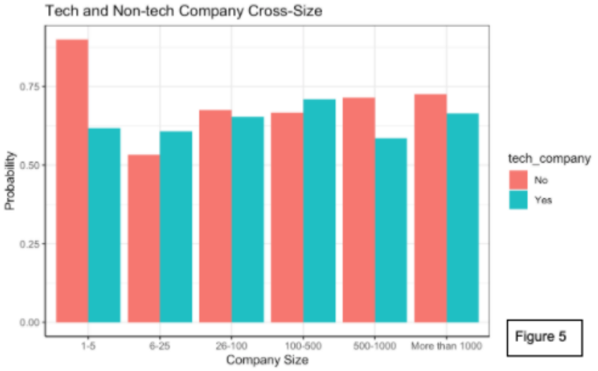


Figure 5

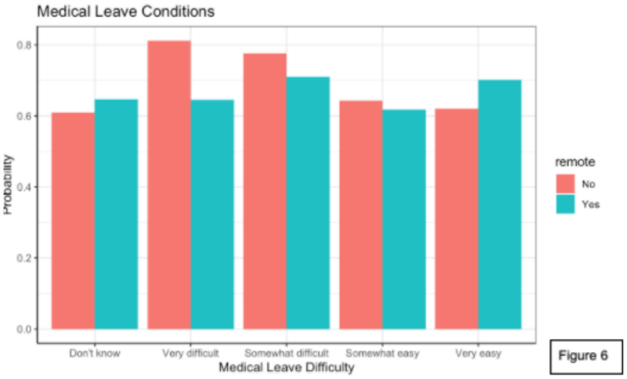


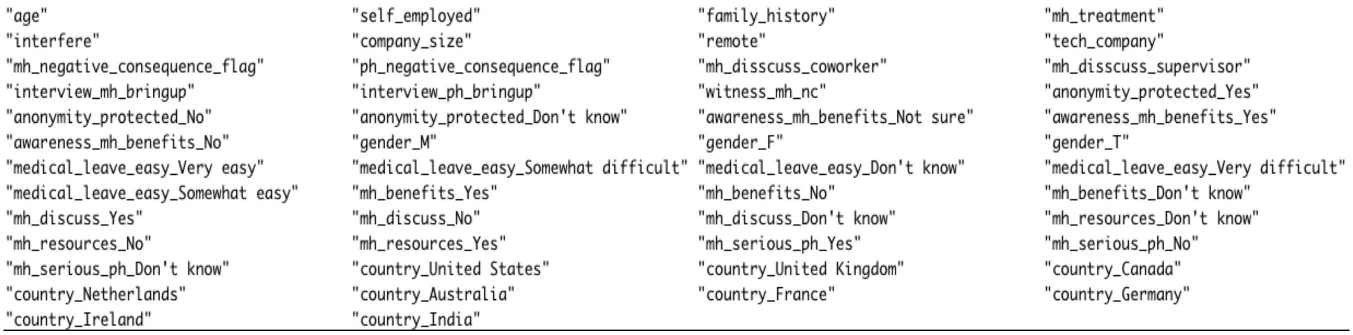
Figure 6

Figure 6 illustrates medical leave conditions. Based on the plot, despite medical leave difficulty, non-remote workers still have a high chance of having sought mental health treatment.

IV. Models

- Predict Mh_treatment with supervised ML

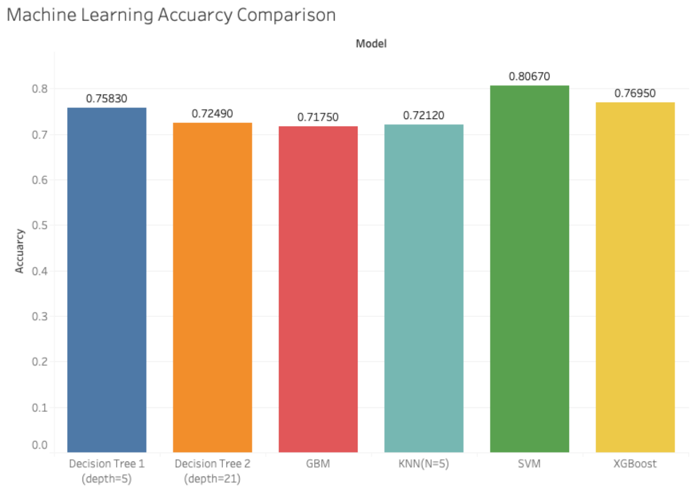
In this part, we tried to answer the following question: is it possible to predict whether an employee has sought mental health treatment? We changed all our dummy variables and created a new dataset for supervised ML models to solve this problem. As a result, we got our target variable (mh_treatment), which is whether the person should be treated. Our independent variables are the remaining 50 variables: company size, awareness, etc..



```
"age"                                   "self_employed"                         "family_history"                        "mh_treatment"
"interfere"                             "company_size"                          "remote"                                "tech_company"
"mh_negative_consequence_flag"          "ph_negative_consequence_flag"          "mh_disscuss_coworker"                  "mh_disscuss_supervisor"
"interview_mh_bringup"                  "interview_ph_bringup"                  "witness_mh_nc"                         "anonymity_protected_Yes"
"anonymity_protected_No"                "anonymity_protected_Don't know"        "awareness_mh_benefits_Not sure"        "awareness_mh_benefits_Yes"
"awareness_mh_benefits_No"              "gender_M"                              "gender_F"                              "gender_T"
"medical_leave_easy_Very easy"          "medical_leave_easy_Somewhat difficult" "medical_leave_easy_Don't know"         "medical_leave_easy_Very difficult"
"medical_leave_easy_Somewhat easy"      "mh_benefits_Yes"                       "mh_benefits_No"                        "mh_benefits_Don't know"
"mh_discuss_Yes"                        "mh_discuss_No"                         "mh_discuss_Don't know"                 "mh_resources_Don't know"
"mh_resources_No"                       "mh_resources_Yes"                      "mh_serious_ph_Yes"                     "mh_serious_ph_No"
"mh_serious_ph_Don't know"              "country_United States"                 "country_United Kingdom"                "country_Canada"
"country_Netherlands"                   "country_Australia"                     "country_France"                        "country_Germany"
"country_Ireland"                       "country_India"
```

We chose 70% of the data to train our model and 30% to test it. We use 5 kinds of supervised ML models to see their accuracy scores on both training and test datasets. These are decision trees, SVM, KNN, GBM and XGboost. Those comparisons are in Table 1 below. We use two kinds of decision trees here: the first is the one with depth limitation of 5, the second one is with no depth limitations and it fits the training data better. We found that its depth is 21 and the accuracy score is close to 1 on the training data, but it also raised the problem of overfitting. As a result, this model doesn't look good on the test dataset. So, it is not a good model for our project.

Model	Accuracy_score - train	Accuracy_score - test
Decision Tree 1 (depth=5) **Decision Tree 2 (depth=21)**	0.8134 Close to 1	0.7583 0.7249
SVM	0.8182	**0.8067**
KNN(N=5)	0.8198	0.7212
GBM	**0.9904**	0.7175
XGBoost	0.8389	0.7695 Table 1



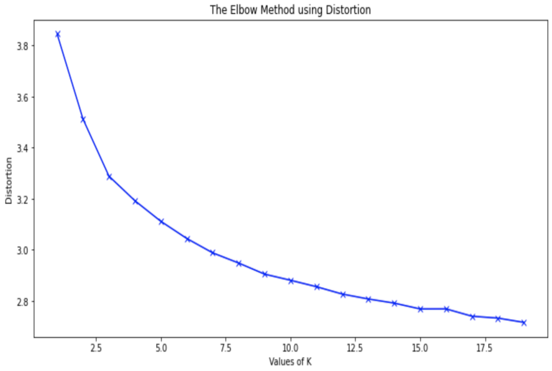
Overall, **SVM** is the most suitable model for our dataset, which performs best on the test dataset with accuracy score of 0.81

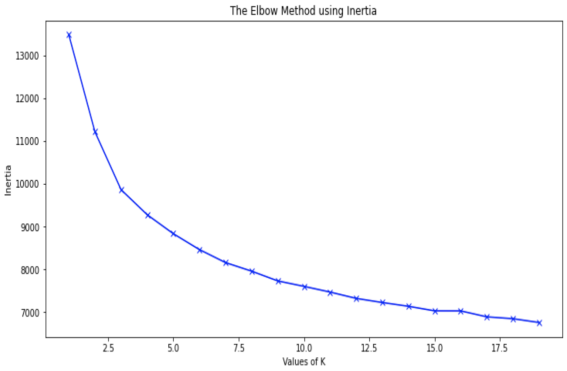
- **Group people with unsupervised ML**

In this part, we used unsupervised machine learning models for grouping people in order to find some common rules between them.

- **K-means**

Firstly, we used K-means for clustering directly with our original dataset. According to the elbow method, we choose K = 6 (see the graph below).

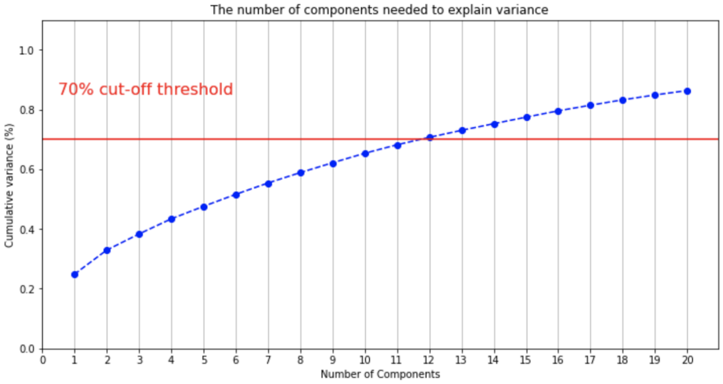




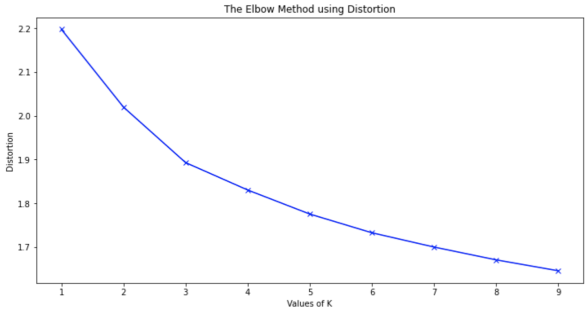
However, since we had too many explanatory variables in our dataset, the results from the direct K-means model do not look good. It is difficult to see that there were any significant group distinctions in 1-3 dimensions, so we did not achieve a suitable visual grouping result.

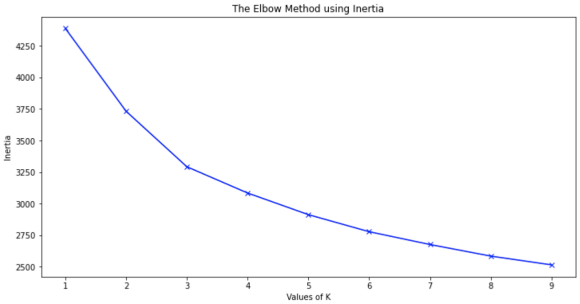
- **PCA + K-means**

In order to reduce the interference of the excess of explanatory variables on the grouping results, we tried to analyze the principal components before using K-means clustering to group people. In the PCA, we scaled the data and decided to choose the first **12 components** in order to get the cumulative variances of 70%.

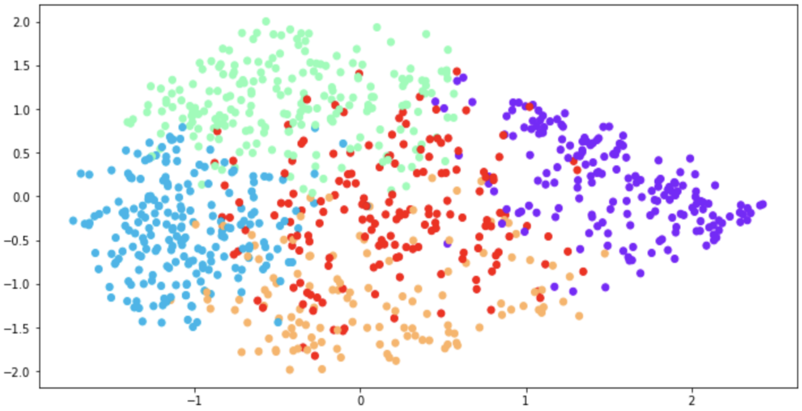


As for the k-means clustering, we choose **K = 5**, as suggested by the elbow method.

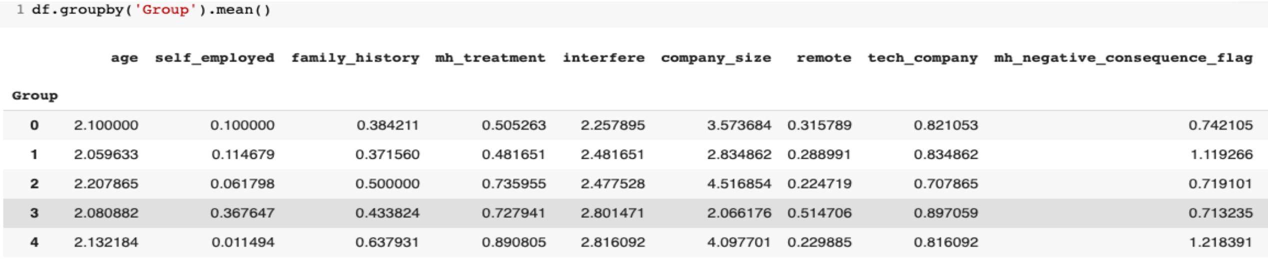




Below, you can see the newly obtained grouping results that look much better, as we can observe exact distinctions between groups. Thus, we want to analyze characteristics for some of those clusters.

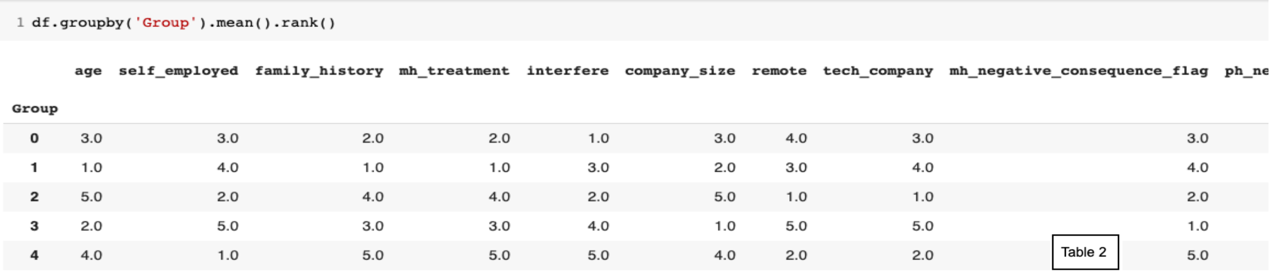


- **Analyze Cluster Characteristics**



```
1 df.groupby('Group').mean()

        age  self_employed  family_history  mh_treatment  interfere  company_size    remote  tech_company  mh_negative_consequence_flag
Group
    0  2.100000      0.100000        0.384211      0.505263   2.257895      3.573684  0.315789      0.821053                      0.742105
    1  2.059633      0.114679        0.371560      0.481651   2.481651      2.834862  0.288991      0.834862                      1.119266
    2  2.207865      0.061798        0.500000      0.735955   2.477528      4.516854  0.224719      0.707865                      0.719101
    3  2.080882      0.367647        0.433824      0.727941   2.801471      2.066176  0.514706      0.897059                      0.713235
    4  2.132184      0.011494        0.637931      0.890805   2.816092      4.097701  0.229885      0.816092                      1.218391
```



```
1 df.groupby('Group').mean().rank()

        age  self_employed  family_history  mh_treatment  interfere  company_size  remote  tech_company  mh_negative_consequence_flag  ph_ne
Group
    0   3.0            3.0             2.0           2.0        1.0           3.0     4.0           3.0                           3.0
    1   1.0            4.0             1.0           1.0        3.0           2.0     3.0           4.0                           4.0
    2   5.0            2.0             4.0           4.0        2.0           5.0     1.0           1.0                           2.0
    3   2.0            5.0             3.0           3.0        4.0           1.0     5.0           5.0                           1.0
    4   4.0            1.0             5.0           5.0        5.0           4.0     2.0           2.0                           5.0
```

Table 2

Results from K-means with PCA:
Table 2 shows the ranking of each cluster for each variable, to summarize some characteristics of them:

1. We found that group G4**,** which has **the highest 'mh_treatment' mean,** consists of individuals who are most likely to seek mental health treatment. Based on our clustering results, the following characteristics are applicable to this group of individuals:

- Less likely to be self-employed
- More likely to have family history of mental health issues
- More likely to feel mental health conditions interfere with work
- More likely to work for larger companies (2nd largest mean)
- Consider discussing mental/physical health issue with an employer to have negative consequences
- Less willing to discuss a mental health issue with coworkers or supervisors
- Less willing to bring up mental/physical issues during an interview
- Have observed or experienced an unsupportive or badly handled response to a mental health issue in workplace (2nd)
- Aware of mental health benefits provided by employers (2nd)
- More likely to have received mental health benefits from previous employers

2. We found that group G1**,** which has **the lowest 'mh_treatment' mean,** consists of individuals who are least likely to seek mental health treatment. Based on our clustering results, the following characteristics are applicable to this group of individuals:

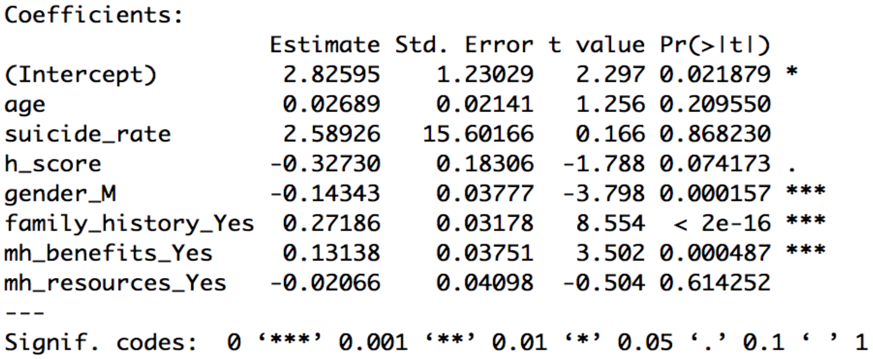
- Younger
- More likely to be self-employed
- More likely to be male and less likely to be female
- Less likely to have family history of mental health issues
- Believe that discussing a mental/physical health issue with the employer would have negative consequences
- Less willing to discuss a mental health issue with coworkers or supervisors
- Less willing to bring up the mental/physical issues in an interview
- Unaware of mental health benefits provided by employers
- Are not easy to take medical leave for a mental health condition
- Less likely to receive mental health benefits from previous employers
- Think that discussing a mental/physical health issue with the employer would have negative consequences

3. We found that group G3 consists of individuals with the highest likelihood of working at a technology company. Based on our clustering results, the following characteristics are applicable to this group of individuals:

- Believe that the mental health condition interferes with work
- Able to work remotely at least 50% of the time
- Do not think that discussing a mental/physical health issue with the employer would have negative consequences
- Willing to discuss a mental health issue with coworkers or supervisors
- Willing to bring up the mental/physical issues in an interview

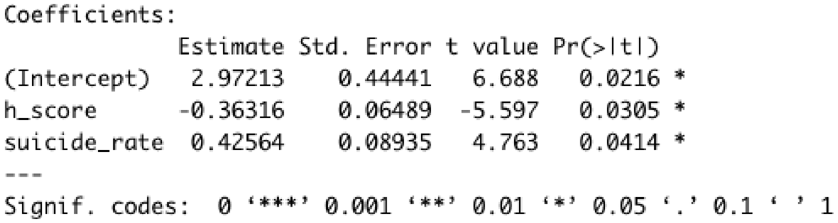
V. Linear Regression, an extension

In order to incorporate world happiness score and suicide statistics into our main question: *whether an employee sought treatment for mental health conditions or not*, we decided to run simple linear regressions on our target variable with other factors. We argue that the two variables mentioned above are helpful to explain our question due to the following: a.) The world happiness score is calculated based on several socio-economic indicators. By using this score, we are potentially looking at several different factors for each country; b.) the suicide rate is one of the most important mental health condition indicators and it gives an overall view of one country's mental health environment. Due to the limited number of observations, it is important to note that our regression results are *not* meant to be prescriptive but to get some general insights regarding elements that could affect employees' decisions on whether seeking mental health treatment.



```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.82595    1.23029   2.297 0.021879 *
age                  0.02689    0.02141   1.256 0.209550
suicide_rate         2.58926   15.60166   0.166 0.868230
h_score             -0.32730    0.18306  -1.788 0.074173 .
gender_M            -0.14343    0.03777  -3.798 0.000157 ***
family_history_Yes   0.27186    0.03178   8.554  < 2e-16 ***
mh_benefits_Yes      0.13138    0.03751   3.502 0.000487 ***
mh_resources_Yes    -0.02066    0.04098  -0.504 0.614252
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first linear regression model is built on an **individual-level** perspective. We find that family history, gender, access to mental health benefits, and happiness score are correlated with whether sought mental health treatment. Notwithstanding, the R-square value is not high enough to conclude that this result is reliable.



```
Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)   2.97213    0.44441   6.688   0.0216 *
h_score      -0.36316    0.06489  -5.597   0.0305 *
suicide_rate  0.42564    0.08935   4.763   0.0414 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The second linear regression is built on a **country-level** perspective, with only the happiness score and suicide rate as explanatory variables. We find the happiness score is negatively correlated with the chance of seeking mental health treatment, whereas the suicide rate is positively related to the chance of seeking mental health treatment. In conclusion, we recognize there exists a variety of factors that either directly or indirectly affect decisions on whether an individual seeks for mental health treatment or not, and these factors can come from both personal perspectives and socio-economic status.

VI. Limitations of the project (Further Research Suggestion)

There are a number of limitations in our project. Firstly, there is a low response rate in the survey dataset, so we had to decrease the number of observations. As a result, the predictive power of all the tests became smaller. Secondly, only the countries with 10 or more survey responses were included in our analysis due to the lack of data in some of them. Because of that, we had to narrow our dataset and the results might be biased towards the developed countries. All the predictions and machine learning models are based on the country-level data, so it cannot always be applicable to individuals. We advise that further research focuses not only on obtaining more individual-level data to get more personalized recommendations on mental health but also encouraging more companies to ask their employees about their mental health.

VII. Takeaways

Companies should raise more awareness for providing access to mental health support to their employees and make them understand that both mental health and physical health are equally important. It is more likely that an employee stays healthy and becomes more productive if the company he or she works for cares about their mental wellbeing. It is very important to encourage employees to utilize and take advantage of existing mental health support programs.