Mental Health in Workplace

Alvaro Chinchayan, Leighton Li, Andrey Lifar, Yoki Liu, Yue Ping, Sherry Zuo

Team 6

Project Objectives

- Explore on people's mental health, specifically in **a corporate environment.**

- Utilize survey data from a non-profit organization and build machine learning models to:
  - Predict if an employee has sought treatment for his/her mental health condition.
  - Explore correlations between mental health conditions and the accessibility of mental health supports in workplaces.

- Provide insights and suggestions to human resources departments regarding current workplace mental health situations and potential improvements in employee health support programs.

Datasets

- Source: OSMI (Open Sourcing Mental Illness) Mental Health in Tech Survey

(https://osmihelp.org/research)

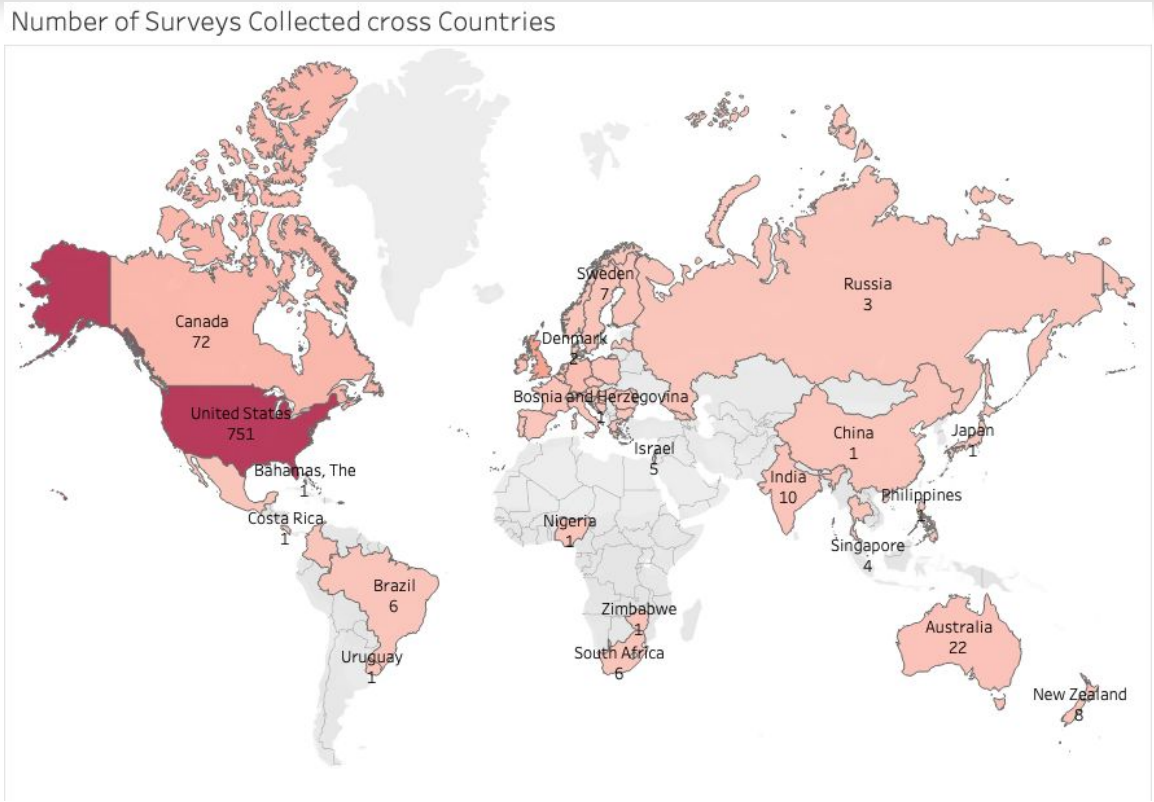
- Survey from 2014:

Main dataset used for machine learning.

- Survey from 2016 to 2019:

Supplemental data used to compare yearly trends.

Data Cleaning



We removed countries that did not have sufficient representation (fewer than **10** survey responses)

Most cases in our dataset are from **United States**

EDA for Survey in 2014

The data has 896 observations and 24 columns (survey questions) after data cleaning.

Our target column is “whether sought treatment”.

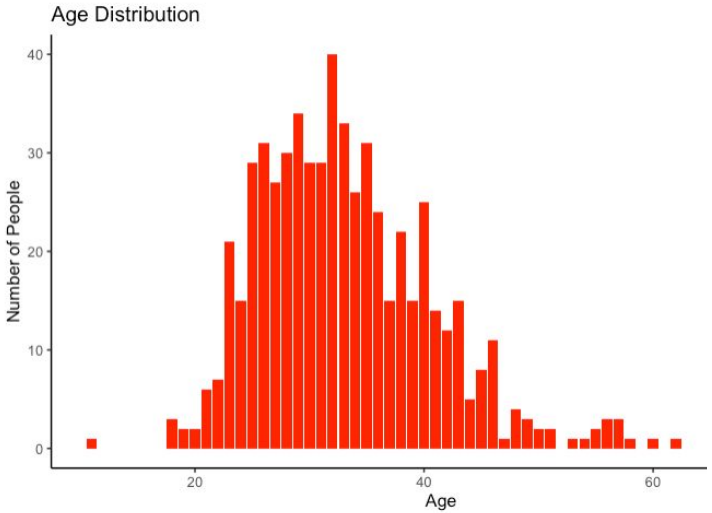
target variable

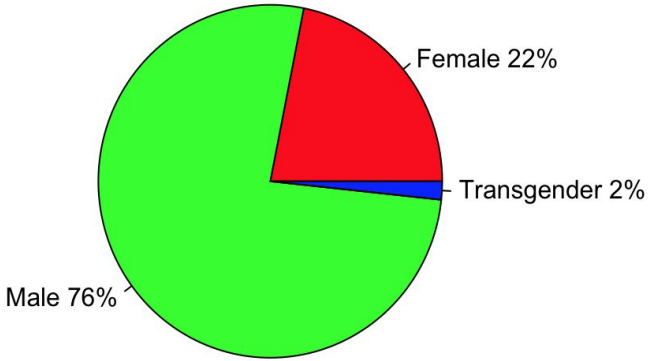
```
 [1] "age"                    "gender"                    "country"
 [4] "self_employed"          "family_history"            "mh_treatment"
 [7] "interfere"              "company_size"              "remote"
[10] "tech_company"           "mh_benefits"               "awareness_mh_benefits"
[13] "mh_discuss"             "mh_resources"              "anonymity_protected"
[16] "medical_leave_easy"     "mh_negative_consequence_flag" "ph_negative_consequence_flag"
[19] "mh_disscuss_coworker"   "mh_disscuss_supervisor"    "interview_mh_bringup"
[22] "interview_ph_bringup"   "mh_serious_ph"             "witness_mh_nc"
```

(Note: A glance of all column names, “mh” stands for “mental health”.)

Age and Gender Distribution

Most respondents are between 20- and 50-year-old and are predominantly males.

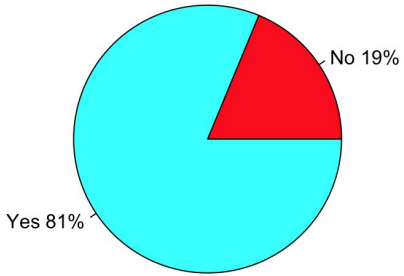




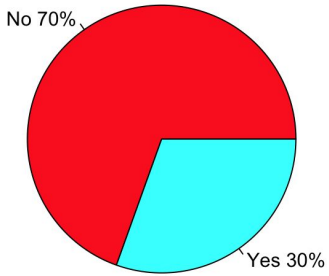
Work Characteristics

Most respondents work for tech companies, are not remote workers, and are either uncertain or pessimistic about medical leave.

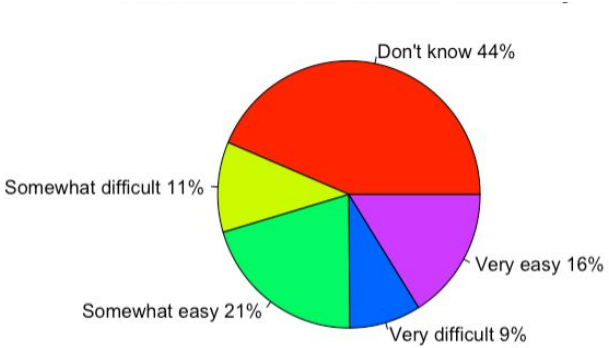
Work in tech?



Work remotely?

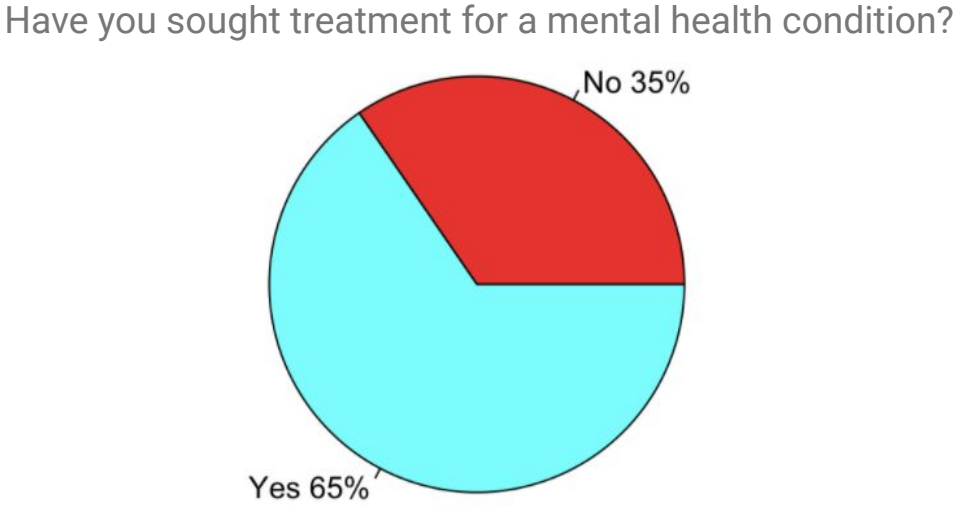


Easy medical leave?

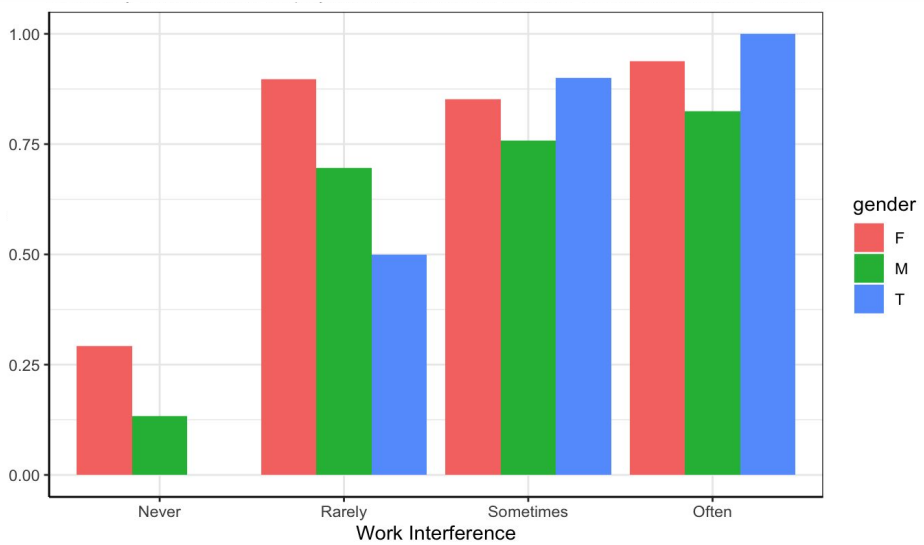


People are Seeking for Mental Health Treatment

More than half of the respondents (65%) sought **mental health treatment** — our target variable

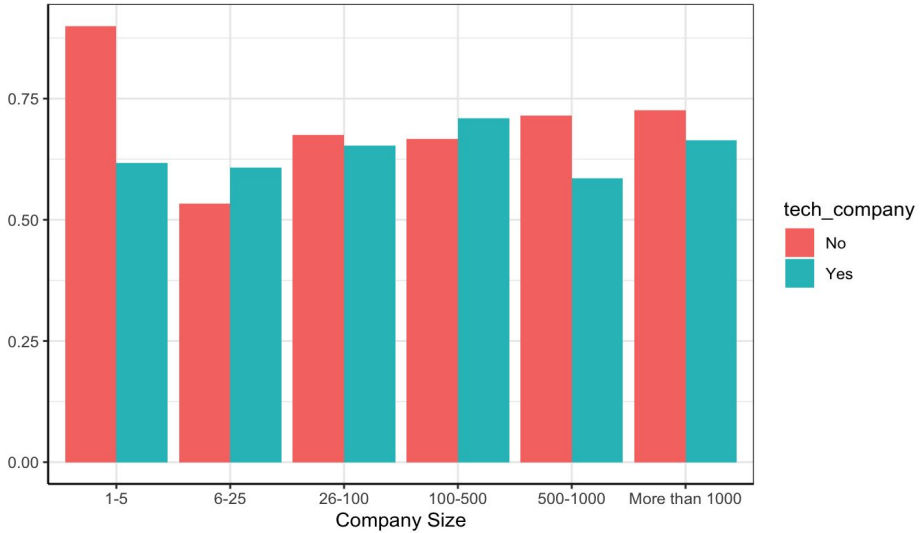


Interferences with Work Cross-Gender



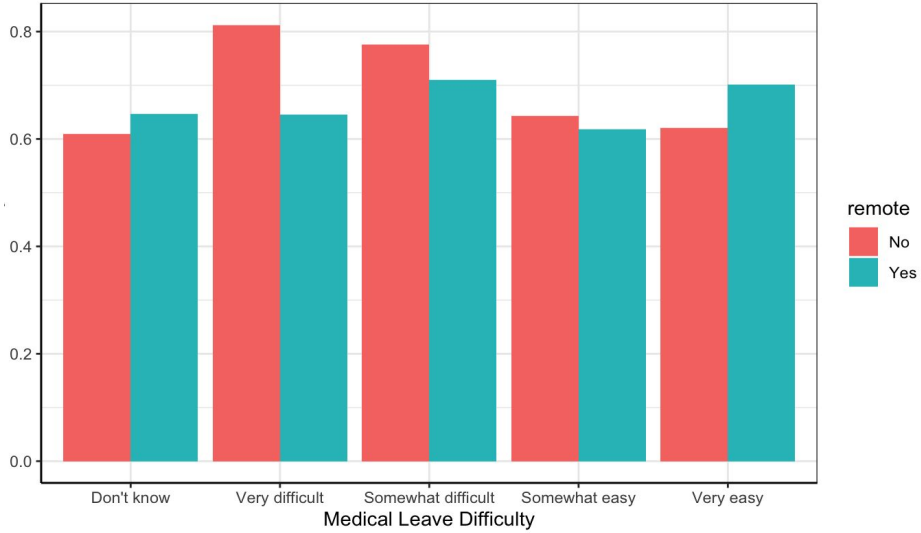
The plot shows the portion of respondents who sought treatment and either felt mental health often/sometimes/rarely/never interferes with work, across different genders.

Tech and Non-tech Company Cross-Size



This plot shows the portion of respondents who sought treatment and worked for either tech or non-tech companies across different company sizes.

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.

Predictive Model

Can we predict whether someone sought mental health treatment given other conditions?

"Mh_treatment"

~

50 other variables

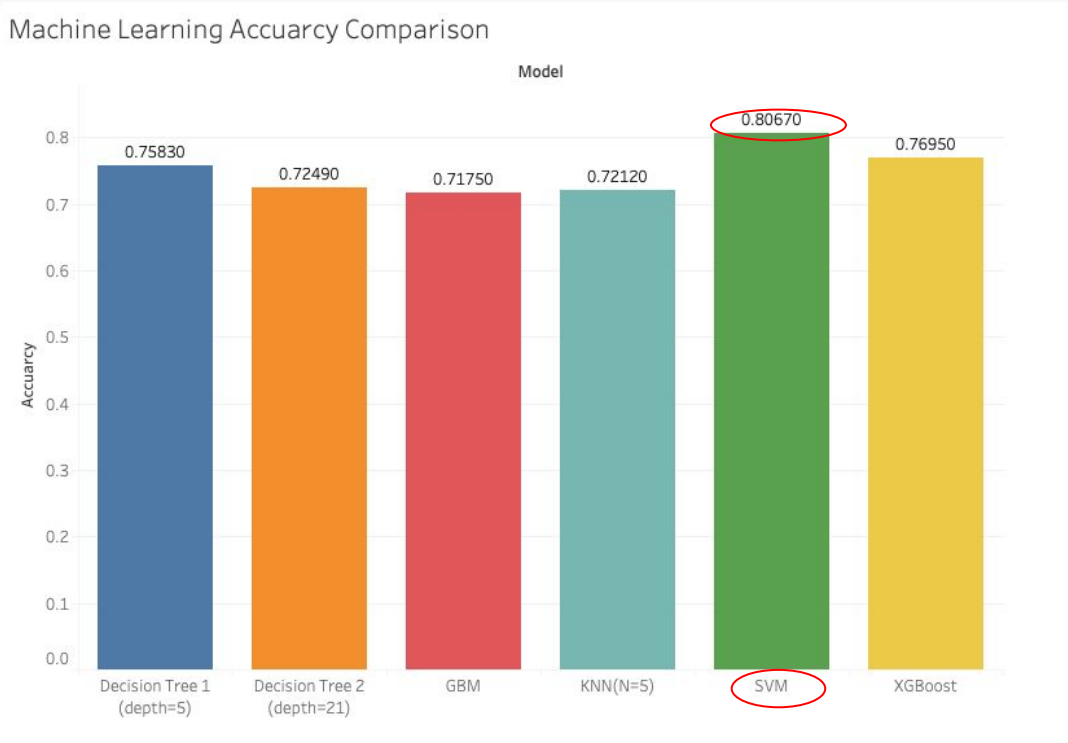
```
['age', 'self_employed', 'family_history', 'interfere', 'company_size',
 'remote', 'tech_company', 'mh_negative_consequence_flag',
 'ph_negative_consequence_flag', 'mh_disscuss_coworker',
 'mh_disscuss_supervisor', 'interview_mh_bringup',
 'interview_ph_bringup', 'witness_mh_nc', 'anonymity_protected_Yes',
 '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'],
```

Supervised Machine Learning Model Accuracy

- Data split into train (70%) and test (30%) sets

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

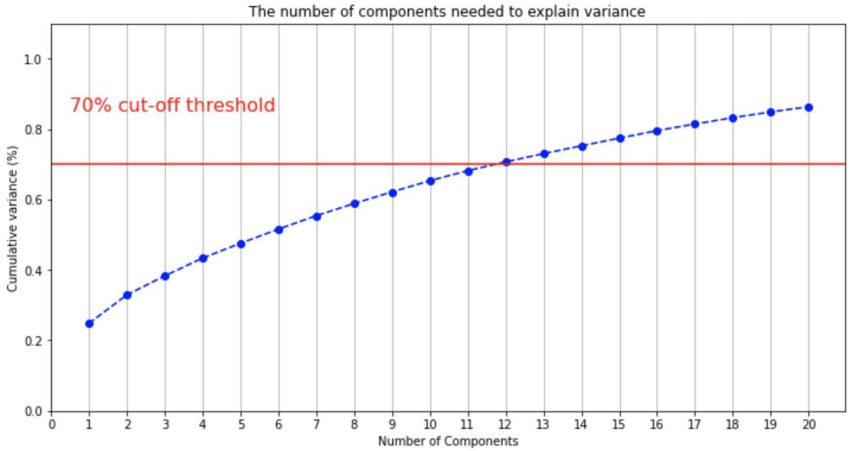
Model Accuracy Comparison



SVM performs best on the test set.

Unsupervised Machine Learning

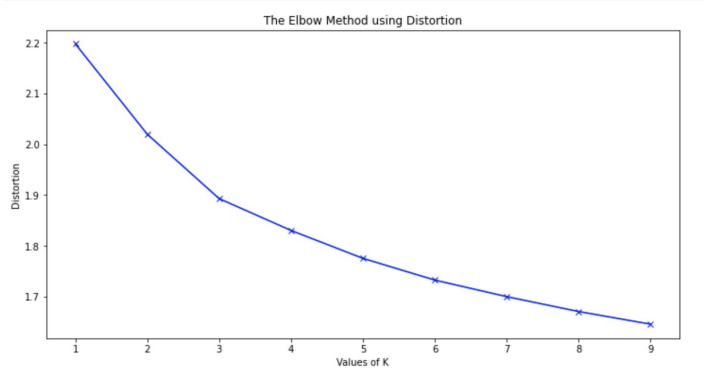
- PCA + k-means
  - PCA - Scale data
  - PCA - Find the optimal number of components

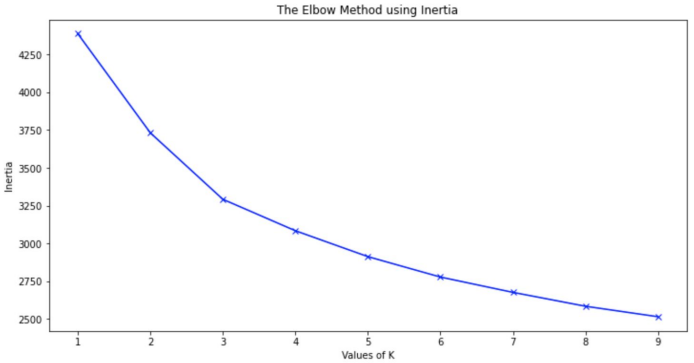


In order to get 70% CV, we decided to choose the first 12 components.

Unsupervised Machine Learning

K-means: k = 5

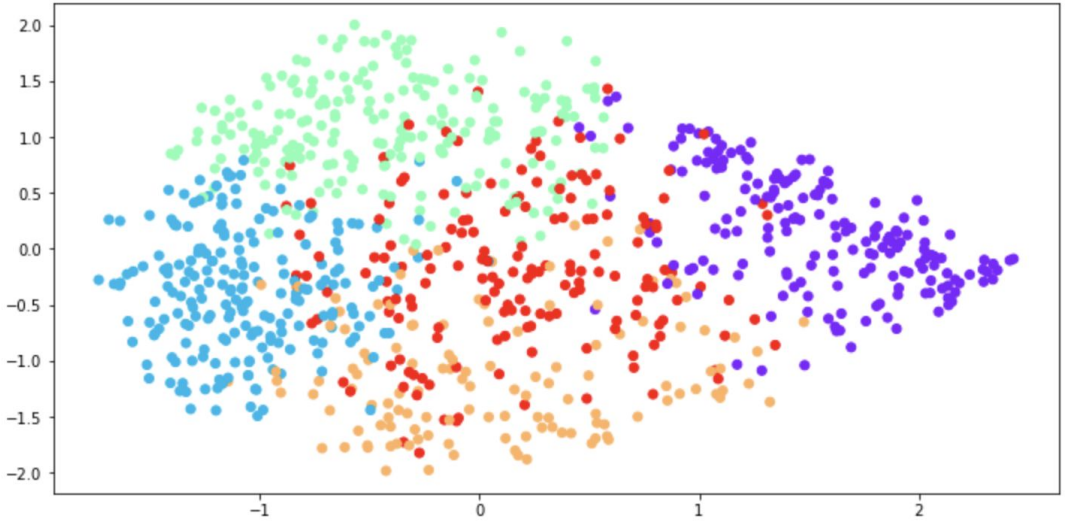




Based on the elbow methods using distortion and inertia, we choose K = 5.

Unsupervised Machine Learning

Five clusters:



The plot above shows five distinct clusters with little overlapping.

Analyze Cluster Characteristics

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

Group	age	self_employed	family_history	mh_treatment	interfere	company_size	remote	tech_company	mh_negative_consequence_flag
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

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

Group	age	self_employed	family_history	mh_treatment	interfere	company_size	remote	tech_company	mh_negative_consequence_flag	ph_ne
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	

Analyze Cluster Characteristics（cont.）

We found that the group has **the highest rate of seeking mental health treatment** on average (G4) and the followings are applicable to this group based on clustering results:

- Are less likely to be self-employed
- Are more likely to have family history of mental health issues
- Are more likely to feel mental health conditions interfere with work
- Are more likely to work for larger companies (2nd largest mean)
- Consider discussing a mental/physical health issue with an employer would have negative consequences
- Are less willing to discuss a mental health issue with coworkers or supervisors
- Are 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)
- Are aware of mental health benefits provided by employers (2nd)
- Are more likely to have received mental health benefits from previous employers

Happiness Score and Suicide Statistics

- World Happiness Score:
  - https://www.kaggle.com/unsdsn/world-happiness
  - Happiness scores computed using several explanatory factors such as GDP per capita, degree of freedom, and life expectancy.
- Suicide Statistics:
  - https://www.kaggle.com/szamil/who-suicide-statistics
  - Includes country-level suicide data as well as statistics of gender, population and GDP.

Linear Regression on Individual-Level

Variables were chosen based on the result from stepwise subset selection.

```
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
```

Happiness Score **−**

Identified as male **−**

Have family history **+**

Have mental health benefits **+**

Linear Regression on Country-Level

Can we find a correlation between happiness scores as well as suicide rates and the chance of seeking mental health treatment?

```
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
```

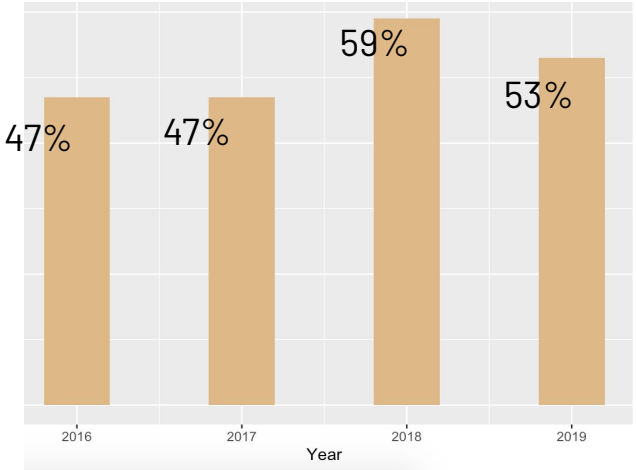
High R-square

Small sample size

Non-prescriptive

Survey Results from 2016 - 2019

Employers provide mental health benefits as part of healthcare coverage.



People feel more comfortable talking about their physical health than they do about their mental health.

Only about **1%** of responders feel comfortable discussing their mental health conditions, while about **60%** of them feel more comfortable discussing their physical health.

Takeaways

- Corporations should raise more awareness for providing access to mental health support;
- Corporations should help employees understand that both mental health and physical health are equally important;
- Corporations should encourage employees to utilize and take advantage of existing mental health support programs

Limitations

- There is a limited response rate in the survey dataset, so we had to decrease the number of observations. As a result, the predictive power became smaller;

- Not all of the countries are represented in our project because of the low response rate in some of them (we only chose the ones with 10 or more observations);

- Not a lot of individual-level data was added to the regressions and other ML models.

Thank you for your attention!

Please share your questions or suggestions with us.

Appendix

1. Project objectives
2. Data exploration
3. Machine learning models
4. Takeaways
5. Limitations