DAC\_Phase3

[1]:

* 1. I. Importing Libraries and Pulling Dataset into the Notebook

*# this will help in making the Python code more structured automatically (good*␣

𝗌*coding practice)*

*#%load\_ext nb\_black*

*# Library to suppress warnings or deprecation notes*

# import warnings

warnings.filterwarnings("ignore")

*# Libraries to help with reading data and working on it (tabular and numerical)*

# import pandas as pd import numpy as np

*# Libraries to perform statistical analysis*

# import scipy.stats as stats import sklearn

*# Library to split data*

**from sklearn.model\_selection import** train\_test\_split

*# libaries to help with data visualization*

# import matplotlib.pyplot as plt import seaborn as sns

*# Removes the limit for the number of displayed columns*

pd.set\_option("display.max\_columns", **None**)

*# Sets the limit for the number of displayed rows*

pd.set\_option("display.max\_rows", 200)

*# statemodels*

# import statsmodels.api as sm

**from statsmodels.tools.tools import** add\_constant

*# To build logistic regression model*

**from sklearn.linear\_model import** LogisticRegression

*# To get diferent metric scores* **from sklearn import** metrics **from sklearn.metrics import** (

f1\_score, accuracy\_score, recall\_score, precision\_score, confusion\_matrix, roc\_auc\_score, precision\_recall\_curve, roc\_curve,

make\_scorer,

)

*# Libraries to build decision tree classifier* **from sklearn.tree import** DecisionTreeClassifier **from sklearn import** tree

*# Libraries to import different ensemble classifiers*

**from sklearn.ensemble import** BaggingClassifier

**from sklearn.ensemble import** RandomForestClassifier

*# To tune different models*

**from sklearn.model\_selection import** GridSearchCV

**import os**

**for** dirname, \_, filenames **in** os.walk('/kaggle/input'):

**for** filename **in** filenames: print(os.path.join(dirname, filename))

/kaggle/input/mental-health-in-tech-survey/survey.csv

[2]:

*# Loading the survey into out notebook*

mhdata = pd.read\_csv('/kaggle/input/mental-health-in-tech-survey/survey.csv')

[3]:

* 1. **II. EDA and Data Preparation**

[3]: (1259, 27)

*# Checking the dimensions of our data*

mhdata.shape

We have 27 columns and 1,259 entries from the survey

[4]:

*# Taking a peak at the first five entries in our dataset*

mhdata.head(5)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [4]: | Timestamp | | | Age | Gender | Country | | state | self\_employed \ | |
|  | 0 2014-08-27 11:29:31 | | | 37 | Female | United States | | IL | NaN | |
|  | 1 2014-08-27 11:29:37 | | | 44 | M | United States | | IN | NaN | |
|  | 2 2014-08-27 11:29:44 | | | 32 | Male | Canada | | NaN | NaN | |
|  | 3 2014-08-27 11:29:46 | | | 31 | Male | United Kingdom | | NaN | NaN | |
|  | 4 2014-08-27 11:30:22 | | | 31 | Male | United States | | TX | NaN | |
|  | | family\_history | treatment | | work\_interfere | | no\_employees | | remote\_work | \ |
| 0 | | No | Yes | | Often | | 6-25 | | No |  |
| 1 | | No | No | | Rarely | | More than 1000 | | No |  |
| 2 | | No | No | | Rarely | | 6-25 | | No |  |
| 3 | | Yes | Yes | | Often | | 26-100 | | No |  |
| 4 | | No | No | | Never | | 100-500 | | Yes |  |

[5]:

*# Checking for data types of each attribute*

mhdata.info()

tech\_company benefits care\_options wellness\_program seek\_help \

1. Yes Yes Not sure No Yes
2. No Don't know No Don't know Don't know
3. Yes No No No No
4. Yes No Yes No No
5. Yes Yes No Don't know Don't know

anonymity leave mental\_health\_consequence \

1. Yes Somewhat easy No
2. Don't know Don't know Maybe
3. Don't know Somewhat difficult No
4. No Somewhat difficult Yes
5. Don't know Don't know No

phys\_health\_consequence coworkers supervisor mental\_health\_interview \

0 No Some of them Yes No

1 No No No No

2 No Yes Yes Yes

3 Yes Some of them No Maybe

4 No Some of them Yes Yes

phys\_health\_interview mental\_vs\_physical obs\_consequence comments

1. Maybe Yes No NaN
2. No Don't know No NaN
3. Yes No No NaN
4. Maybe No Yes NaN
5. Yes Don't know No NaN

<class 'pandas.core.frame.DataFrame'> RangeIndex: 1259 entries, 0 to 1258

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data # | columns (total 27 Column | columns): | 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)

memory usage: 265.7+ KB

It seems that all attributes are objects based on the nature of the data. This is correct but this will be hard to work a model on. This will be explained later on.

# Attribute Information

**Timestamp** - date and time the survey was taken **Age** - age of the surveyee **Gender** - gender of surveyee **Country** - the country which the surveyee resides and works in **State** - which state in the US (if the answer to the country attribute is currently in the US) **self\_employed** - Self-employed? **family\_history** - family history of mental illness present? **treatment** - sought treatment for a mental health condition? **work\_interfere** - does mental health condition (if any) interfere with work? **no\_employees** - Number of employees at work? **remote\_work** - work remotely more than 50% of workshift? **tech\_company** - Is the company you work in primarily a tech company? **benefits** - Does your company provide mental health benefits? **care\_options** - Are you aware of the options for mental health treatment that your company provides? **wellness\_program** - Is mental health included in the health wellness program of your company? **seek\_help** - Does your

[6]:

company assist you or give options on how to seek medical help? **anonymity** - Is your identity protected should you seek medical treatment regarding mental health? **leave** - Is taking a leave due to mental health easy within your company? **mental\_health\_consequence** - Do you feel or think that your company takes it negatively when you talk to them about your mental health condition? **phys\_health\_consequence** - Do you feel or think that your company takes it negatively when you talk to them about your physical health condition? **coworkers** - Are you open to talk about mental health issues with your peers at work? **supervisor** - Are you open to talk about mental health issues with your supervisor? **mental\_health\_interview** - Would you open up about mental health issues with a prospective company during an interview? **phys\_health\_interview**

- Would you open up about physical health issues with a prospective company during an interview? **mental\_vs\_physical** - Do you think that your company’s concern for mental health and physical health are of the same weight? **obs\_consequence** - Have you heard of or observed negative feedback at your workplace for a fellow employee with a mental health condition? **comments** - Additional comments

# Cleaning the Data

We are aware that surveys may not give out the best results for uniformity, so we will take extra steps to clean the data and make it uniform

*# Checking for missing values*

mhdata.isna().apply(pd.value\_counts).T

|  |  |  |  |
| --- | --- | --- | --- |
| [6]: |  | False | True |
|  | Timestamp | 1259.0 | NaN |
|  | Age | 1259.0 | NaN |
|  | Gender | 1259.0 | NaN |
|  | Country | 1259.0 | NaN |
|  | state | 744.0 | 515.0 |
|  | self\_employed | 1241.0 | 18.0 |
|  | family\_history | 1259.0 | NaN |
|  | treatment | 1259.0 | NaN |
|  | work\_interfere | 995.0 | 264.0 |
|  | no\_employees | 1259.0 | NaN |
|  | remote\_work | 1259.0 | NaN |
|  | tech\_company | 1259.0 | NaN |
|  | benefits | 1259.0 | NaN |
|  | care\_options | 1259.0 | NaN |
|  | wellness\_program | 1259.0 | NaN |
|  | seek\_help | 1259.0 | NaN |
|  | anonymity | 1259.0 | NaN |
|  | leave | 1259.0 | NaN |
|  | mental\_health\_consequence | 1259.0 | NaN |
|  | phys\_health\_consequence | 1259.0 | NaN |
|  | coworkers | 1259.0 | NaN |
|  | supervisor | 1259.0 | NaN |
|  | mental\_health\_interview | 1259.0 | NaN |
|  | phys\_health\_interview | 1259.0 | NaN |

|  |  |  |
| --- | --- | --- |
| mental\_vs\_physical | 1259.0 | NaN |
| obs\_consequence | 1259.0 | NaN |
| comments | 164.0 | 1095.0 |

[7]:

*# Checking for duplicates in the dataset*

mhdata.duplicated().any()

1. : False

No duplicated entries are present in the set

1. :

*# Dropping columns not needed/ not useful for EDA and modeling*

mhdata.drop(columns=['Timestamp', 'Country', 'state', 'comments'], inplace =␣

𝗌**True**)

1. :

Dropping the following columns:

* + **Timestamp** - contains date and time of when the surveyee took the survey; not really useful for analysis
  + **Country** - contains information on where the surveyee lives; this would be a useful for a really large dataset with no 1 country dominating more than half of the answers. In our dataset, roughly 59% of the surveyee reside in the USA.
  + **state** - contains information on which state in particular the surveyee lives if they ever answered USA in the Country field. Since only the rest of the 41% of the dataset live outside the USA, we have 41% of null values and would not be organic to include this one
  + **comments** - an optional field for surveyees to answer: any additional comments they would like to make regarding the survey; this will not be included since we only got 14% surveyees who answered this field. We are also not doing a text analysis.

*# Renaming columns for uniformity and ease of understanding. We like uniformity!*

mhdata.rename({'self\_employed' : 'Self\_Employed', 'family\_history' :␣

𝗌'Family\_History',

'treatment' : 'Treatment', 'work\_interfere' : 'Work\_Interfere', 'no\_employees': 'Employee\_Count\_Company', 'remote\_work':␣

𝗌'Remote\_Work', 'tech\_company': 'Tech\_Company',

'benefits': 'Benefits', 'care\_options': 'Care\_Options',␣

𝗌'wellness\_program': 'Wellness\_Program',

'seek\_help': 'Seek\_Help', 'anonymity': 'Anonymity', 'leave':␣

𝗌'Medical\_Leave',

'mental\_health\_consequence': 'Mental\_Health\_Consequence', 'phys\_health\_consequence': 'Physical\_Health\_Consequence',␣

𝗌'coworkers': 'Coworkers\_Reach',

'supervisor': 'Supervisor\_Reach', 'mental\_health\_interview':␣

𝗌'Mental\_Health\_Interview',

'phys\_health\_interview': 'Physical\_Health\_Interview',␣

𝗌'mental\_vs\_physical': 'Mental\_VS\_Physical',

'obs\_consequence': 'Observed\_Consequence\_Workplace'} , inplace =␣

𝗌**True** , axis = 1)

1. :

*# Checking the entries for age*

mhdata['Age'].unique()

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| [10]: array([ | 37, | 44, | 32, | 31, | 33, |
|  | 35, | 39, | 42, | 23, | 29, |
|  | 36, | 27, | 46, | 41, | 34, |
|  | 30, | 40, | 38, | 50, | 24, |
|  | 18, | 28, | 26, | 22, | 19, |
|  | 25, | 45, | 21, | -29, | 43, |
|  | 56, | 60, | 54, | 329, | 55, |
|  | 99999999999, | 48, | 20, | 57, | 58, |
|  | 47, | 62, | 51, | 65, | 49, |
|  | -1726, | 5, | 53, | 61, | 8, |
|  | 11, | -1, | 72]) |  |  |

We have ages that are off (negative and immortal numbers). Let’s fix them by imputing.

1. :

*# calculating the median age* median\_age = mhdata['Age'].median() print(median\_age)

31.0

1. :

*# since some of data are impossible to work with, we will replace impossible*␣

𝗌*values with the median age*

mhdata['Age'].replace([mhdata['Age'][mhdata['Age'] < 15]], median\_age, inplace␣

𝗌= **True**)

mhdata['Age'].replace([mhdata['Age'][mhdata['Age'] > 100]], median\_age, inplace␣

𝗌= **True**)

mhdata['Age'].unique()

1. : array([37, 44, 32, 31, 33, 35, 39, 42, 23, 29, 36, 27, 46, 41, 34, 30, 40,

38, 50, 24, 18, 28, 26, 22, 19, 25, 45, 21, 43, 56, 60, 54, 55, 48,

20, 57, 58, 47, 62, 51, 65, 49, 53, 61, 72])

Replacing impossible values with the median age was our call to use instead of np.nan since we want to avoid missing values in our dataset. Using the median age preserves the central tendency of the data and minimizes the impact of outliers. Here, we set a limit to <15 (minimum working age according to International Labor Organization [ILO] standards) and >100 to avoid outliers

1. :

*# Checking the entries for gender*

mhdata['Gender'].unique()

1. : array(['Female', 'M', 'Male', 'male', 'female', 'm', 'Male-ish', 'maile', 'Trans-female', 'Cis Female', 'F', 'something kinda male?',

'Cis Male', 'Woman', 'f', 'Mal', 'Male (CIS)', 'queer/she/they',

'non-binary', 'Femake', 'woman', 'Make', 'Nah', 'All', 'Enby',

'fluid', 'Genderqueer', 'Female ', 'Androgyne', 'Agender',

'cis-female/femme', 'Guy (-ish) ^\_^', 'male leaning androgynous', 'Male ', 'Man', 'Trans woman', 'msle', 'Neuter', 'Female (trans)', 'queer', 'Female (cis)', 'Mail', 'cis male', 'A little about you', 'Malr', 'p', 'femail', 'Cis Man',

'ostensibly male, unsure what that really means'], dtype=object)

1. :

*# We will only limit our analysis to three categories*

mhdata['Gender'].replace(['Male ', 'male', 'M', 'm', 'Male', 'Cis Male',

'Man', 'cis male', 'Mail', 'Male-ish', 'Male (CIS)',

'Cis Man', 'msle', 'Malr', 'Mal', 'maile', 'Make',],␣

𝗌'Male', inplace = **True**)

mhdata['Gender'].replace(['Female ', 'female', 'F', 'f', 'Woman', 'Female', 'femail', 'Cis Female', 'cis-female/femme', 'Femake',␣

𝗌'Female (cis)',

'woman',], 'Female', inplace = **True**)

mhdata["Gender"].replace(['Female (trans)', 'queer/she/they', 'non-binary', 'fluid', 'queer', 'Androgyne', 'Trans-female', 'male␣

𝗌leaning androgynous',

'Agender', 'A little about you', 'Nah', 'All', 'ostensibly male, unsure what that really means', 'Genderqueer', 'Enby', 'p', 'Neuter', 'something kinda␣

𝗌male?',

'Guy (-ish) ^\_^', 'Trans woman',], 'Queer', inplace =␣

𝗌**True**)

1. :

mhdata['Gender'].value\_counts()

1. : Male 991

Female 247

Queer 21

Name: Gender, dtype: int64

With that fixed, we identify 991 entries from male respondents, 247 entries from female respondents, and 21 entries from queer respondents

1. :

*# we want to make sure that the answers are limited to boolean-like values*

columns\_to\_print = ['Self\_Employed', 'Family\_History','Treatment',␣

𝗌'Work\_Interfere', 'Employee\_Count\_Company', 'Remote\_Work',

'Tech\_Company', 'Benefits', 'Care\_Options',␣

𝗌'Wellness\_Program',

'Seek\_Help', 'Anonymity', 'Medical\_Leave',␣

𝗌'Mental\_Health\_Consequence',

'Physical\_Health\_Consequence', 'Coworkers\_Reach',␣

𝗌'Supervisor\_Reach',

'Mental\_Health\_Interview', 'Physical\_Health\_Interview',␣

𝗌'Mental\_VS\_Physical',

'Observed\_Consequence\_Workplace']

**for** column **in** columns\_to\_print: print(f"**{**column**}**:") print(mhdata[column].value\_counts()) print()

Self\_Employed:

No 1095

Yes 146

Name: Self\_Employed, dtype: int64

Family\_History:

No 767

Yes 492

Name: Family\_History, dtype: int64

Treatment:

Yes 637

No 622

Name: Treatment, dtype: int64

Work\_Interfere:

Sometimes 465

Never 213

Rarely 173

Often 144

Name: Work\_Interfere, dtype: int64

Employee\_Count\_Company:

|  |  |
| --- | --- |
| 6-25 | 290 |
| 26-100 | 289 |
| More than 1000 | 282 |
| 100-500 | 176 |
| 1-5 | 162 |
| 500-1000 | 60 |

Name: Employee\_Count\_Company, dtype: int64 Remote\_Work:

No 883

Yes 376

Name: Remote\_Work, dtype: int64

Tech\_Company: Yes 1031

No 228

Name: Tech\_Company, dtype: int64

Benefits:

Yes 477

Don't know 408

No 374

Name: Benefits, dtype: int64

Care\_Options:

No 501

Yes 444

Not sure 314

Name: Care\_Options, dtype: int64

Wellness\_Program:

No 842

Yes 229

Don't know 188

Name: Wellness\_Program, dtype: int64

Seek\_Help:

No 646

Don't know 363

Yes 250

Name: Seek\_Help, dtype: int64

Anonymity:

Don't know 819

Yes 375

No 65

Name: Anonymity, dtype: int64

Medical\_Leave:

Don't know 563

Somewhat easy 266

Very easy 206

Somewhat difficult 126

Very difficult 98

Name: Medical\_Leave, dtype: int64

Mental\_Health\_Consequence:

No 490

Maybe 477

Yes 292

Name: Mental\_Health\_Consequence, dtype: int64

Physical\_Health\_Consequence:

No 925

Maybe 273

Yes 61

Name: Physical\_Health\_Consequence, dtype: int64

Coworkers\_Reach:

Some of them 774

No 260

Yes 225

Name: Coworkers\_Reach, dtype: int64

Supervisor\_Reach:

Yes 516

No 393

Some of them 350

Name: Supervisor\_Reach, dtype: int64

Mental\_Health\_Interview:

No 1008

Maybe 207

Yes 44

Name: Mental\_Health\_Interview, dtype: int64

Physical\_Health\_Interview:

Maybe 557

No 500

Yes 202

Name: Physical\_Health\_Interview, dtype: int64

Mental\_VS\_Physical:

Don't know 576

Yes 343

No 340

Name: Mental\_VS\_Physical, dtype: int64

Observed\_Consequence\_Workplace:

No 1075

Yes 184

Name: Observed\_Consequence\_Workplace, dtype: int64

[17]:

* + 1. **EDA**

**Target Variable - Treatment**

*# Chart for whether respondent sought medical treatment or not*

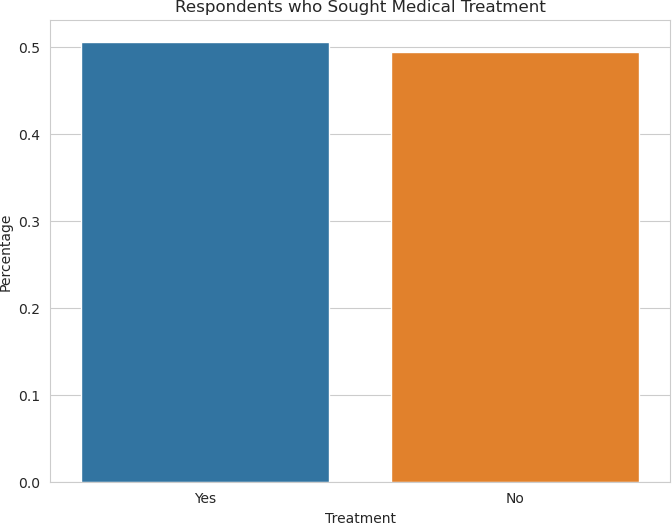
sns.set\_style("whitegrid") plt.figure(figsize = (8,6))

eda\_percentage = mhdata['Treatment'].value\_counts(normalize = **True**).

𝗌rename\_axis('Treatment').reset\_index(name = 'Percentage')

sns.barplot(x = 'Treatment', y = 'Percentage', data = eda\_percentage.head(10)) plt.title('Respondents who Sought Medical Treatment')

plt.show()



**Respondent Demography with Respect to Target**

[18]:

*# Chart for Age* plt.figure(figsize = (18,5)) plt.subplot(1,2,1)

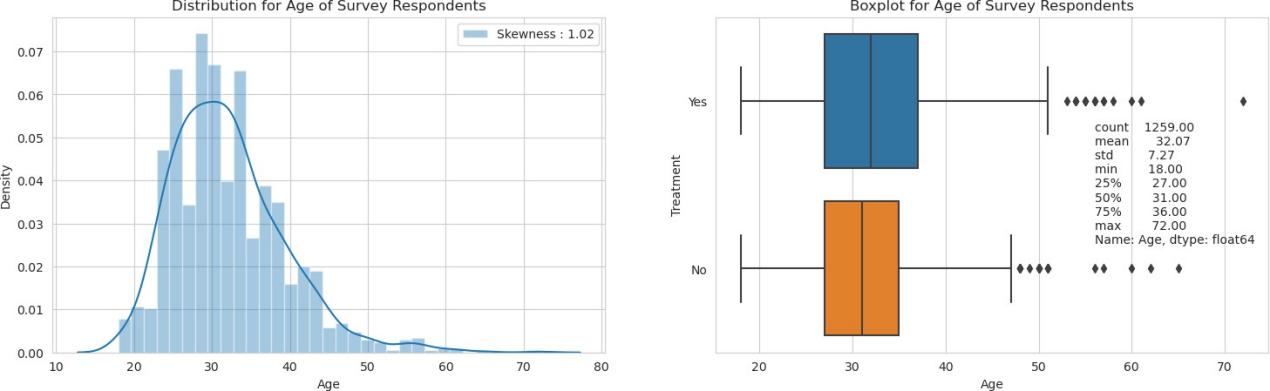
sns.distplot(mhdata['Age'], label = 'Skewness : **%.2f**'%(mhdata['Age'].skew())) plt.legend(loc = 0, fontsize = 10)

plt.title('Distribution for Age of Survey Respondents') plt.subplot(1,2,2)

sns.boxplot(x = "Age", y = "Treatment", data = mhdata) plt.title('Boxplot for Age of Survey Respondents')

age = str(mhdata['Age'].describe().round(2)) plt.text(56, 0.85, age)

plt.show()



[19]:

Distribution and Boxplot has been called to check the profile of the surveyees.

* + - * Results show that age attribute is skewed to the right. We have the highest age at 72 and lowest at 18, the skewness indicates that majority of the tech employees are the young generations ranging from their 20s to late 30s; assuming mid to senior level positions in the tech industry.
      * The boxplot figure presents tech employees that seek medical treatment have much diverse age group, while the group of respondents who did NOT seek medical treatment have a younger range. Both factors are not that significant.

*# Chart for Gender*

plt.figure(figsize = (20, 5))

*# Bar plot for Gender distribution*

plt.subplot(1, 2, 1)

eda\_percentage = mhdata['Gender'].value\_counts(normalize=**True**).

𝗌rename\_axis('Gender').reset\_index(name='Percentage')

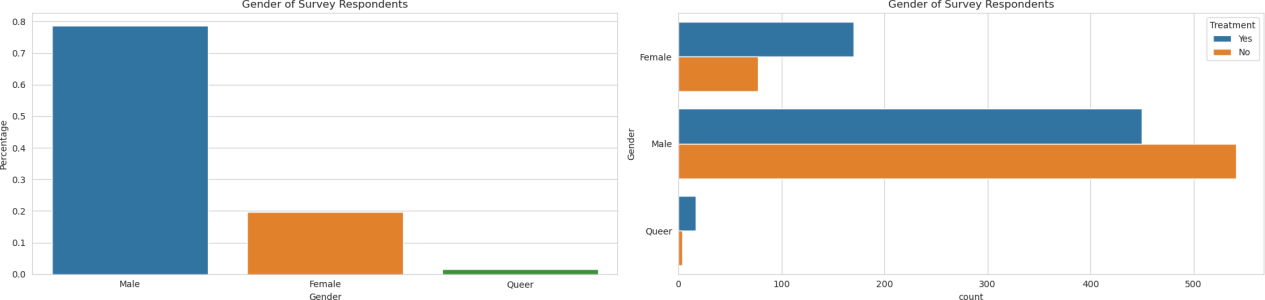
sns.barplot(x = 'Gender', y = 'Percentage', data=eda\_percentage.head(10)) plt.title('Gender of Survey Respondents')

*# Count plot for Gender with hue by Treatment*

plt.subplot(1, 2, 2)

sns.countplot(y = mhdata['Gender'], hue=mhdata['Treatment']) plt.title('Gender of Survey Respondents')

plt.tight\_layout() plt.show()



[20]:

*# Chart for Family History*

plt.figure(figsize = (20,5))

*# Bar plot for Family History distribution*

plt.subplot(1,2,1)

eda\_percentage = mhdata['Family\_History'].value\_counts(normalize = **True**).

𝗌rename\_axis('Family\_History').reset\_index(name = 'Percentage') sns.barplot(x = 'Family\_History', y = 'Percentage', data = eda\_percentage) plt.title('Family History of Survey Respondents')

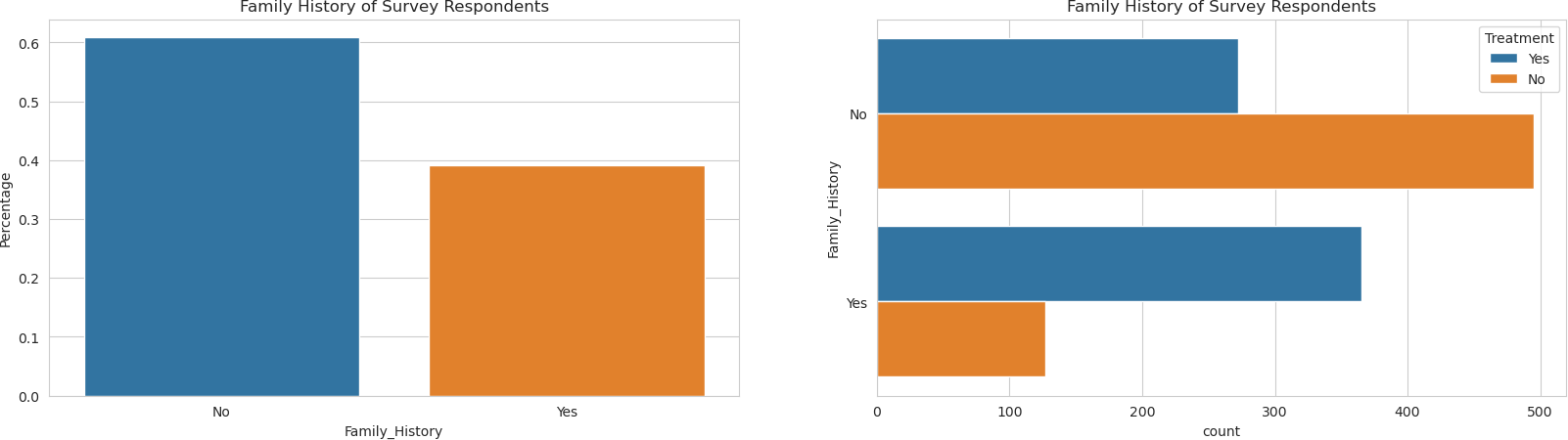
*# Count plot for Family History with hue by Treatment*

plt.subplot(1,2,2)

sns.countplot(y = mhdata['Family\_History'], hue = mhdata['Treatment']) plt.title('Family History of Survey Respondents')

plt.show()

Majority of the surveyees are Male at more than 75%. Male respondents are more likely not to seek medical treatment for mental health compared to other genders.



Majority of the survey respondents recorded that they don’t have a family history of mental health conditions; and respondents who said they have a family history are more likely to seek medical treatment.

# Working Conditions with Respect to Target

[21]:

*# Chart for Work Interference (of mental health)*

plt.figure(figsize = (20,5))

*# Bar plot for Work Interference distribution*

plt.subplot(1,2,1)

eda\_percentage = mhdata['Work\_Interfere'].value\_counts(normalize = **True**).

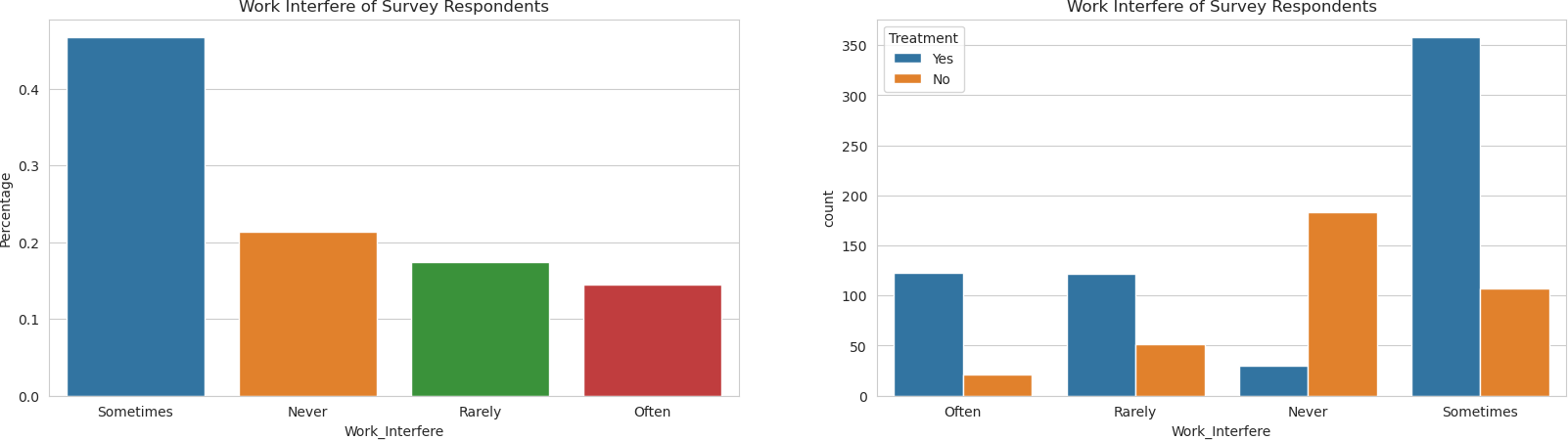
𝗌rename\_axis('Work\_Interfere').reset\_index(name = 'Percentage') sns.barplot(x = 'Work\_Interfere', y = 'Percentage', data = eda\_percentage) plt.title('Work Interfere of Survey Respondents')

*# Count plot for Work Interference with hue by Treatment*

plt.subplot(1,2,2)

sns.countplot(x = mhdata['Work\_Interfere'], hue = mhdata['Treatment']) plt.title('Work Interfere of Survey Respondents')

plt.show()



[22]:

*# Same comment as the previous ones. This will be the last comment for EDA*

plt.figure(figsize = (20,5)) plt.subplot(1,2,1)

eda\_percentage = mhdata['Remote\_Work'].value\_counts(normalize = **True**).

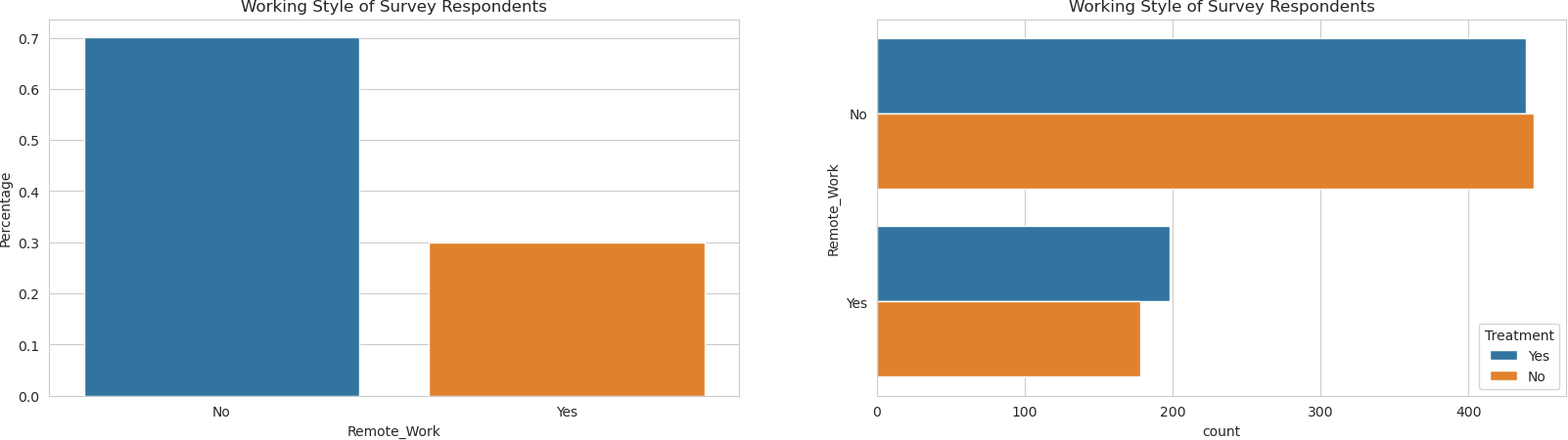
𝗌rename\_axis('Remote\_Work').reset\_index(name = 'Percentage') sns.barplot(x = 'Remote\_Work', y = 'Percentage', data = eda\_percentage) plt.title('Working Style of Survey Respondents')

plt.subplot(1,2,2)

sns.countplot(y = mhdata['Remote\_Work'], hue = mhdata['Treatment']) plt.title('Working Style of Survey Respondents')

plt.show()

Survey reveals that more than 40% of the response say that mental health sometimes interfere with work, with more than 10% of the survey say that it affects them often. Around 20% of the answers reveal that mental health does not affect their work. Regardless of the medical and psychological implication of the responses, it is revealed that surveyees who say that mental health does interfere with work at any degree are more likely to seek medical treatment than those who say that it doesn’t interfere with their work.



[23]:

plt.figure(figsize = (20,5)) plt.subplot(1,2,1)

eda\_percentage = mhdata['Employee\_Count\_Company'].value\_counts(normalize =␣

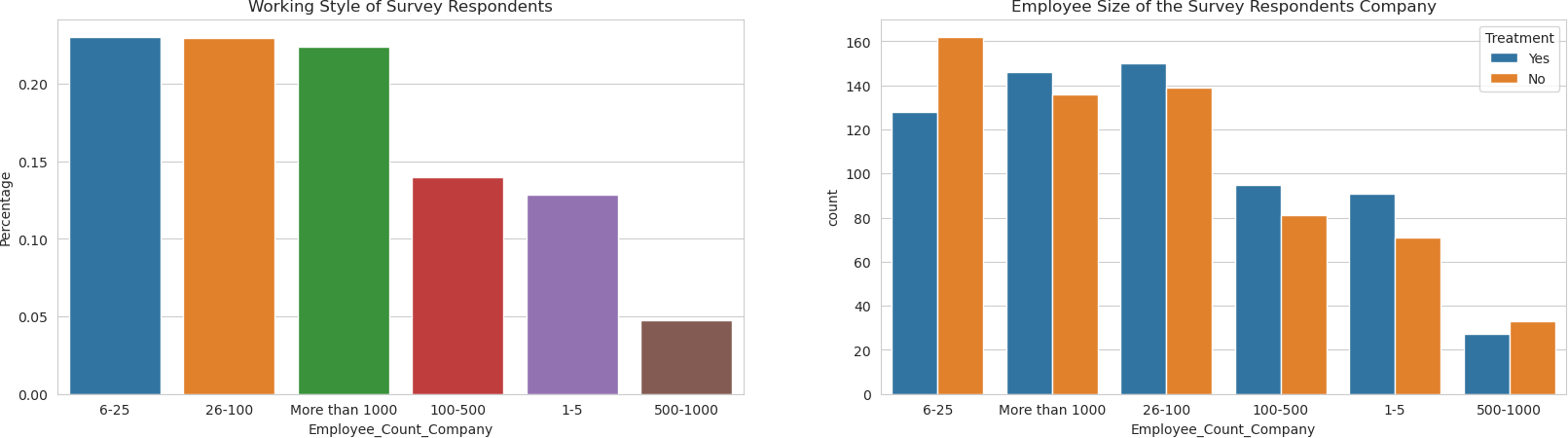
𝗌**True**).rename\_axis('Employee\_Count\_Company').reset\_index(name = 'Percentage') sns.barplot(x = 'Employee\_Count\_Company', y = 'Percentage', data =␣

𝗌eda\_percentage)

plt.title('Working Style of Survey Respondents') plt.subplot(1,2,2)

sns.countplot(x = mhdata['Employee\_Count\_Company'], hue = mhdata['Treatment']) plt.title('Employee Size of the Survey Respondents Company')

plt.show()



[24]:

plt.figure(figsize = (20,5)) plt.subplot(1,2,1)

eda\_percentage = mhdata['Tech\_Company'].value\_counts(normalize = **True**).

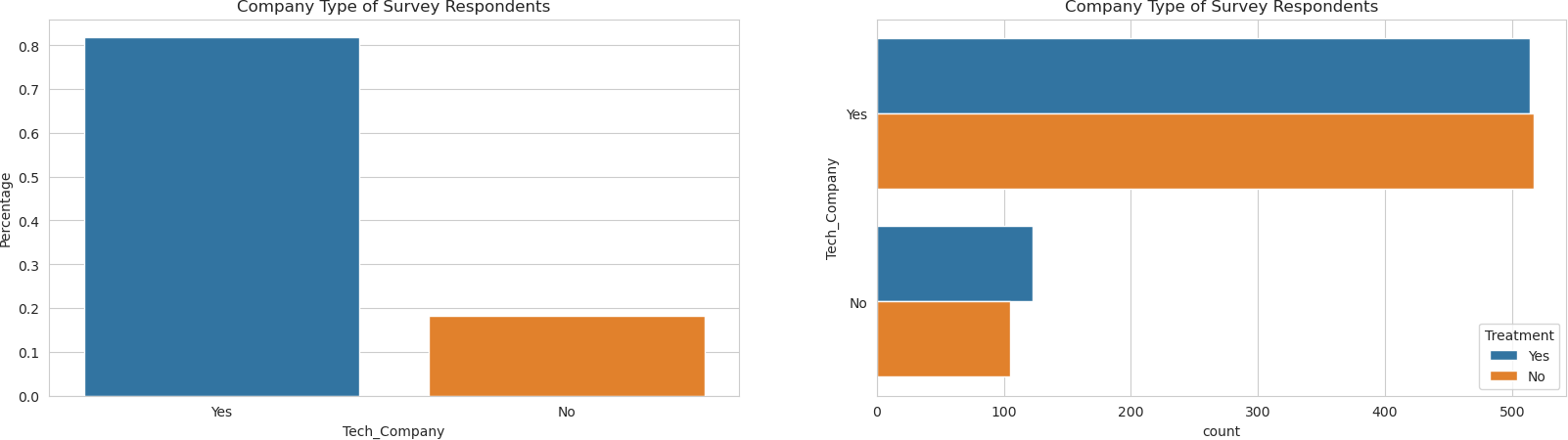
𝗌rename\_axis('Tech\_Company').reset\_index(name = 'Percentage') sns.barplot(x = 'Tech\_Company', y = 'Percentage', data = eda\_percentage) plt.title('Company Type of Survey Respondents')

plt.subplot(1,2,2)

sns.countplot(y = mhdata['Tech\_Company'], hue = mhdata['Treatment'])

There is no significant pattern with regard to employee size of the company where the respondents are working for and getting a treatment.

plt.title('Company Type of Survey Respondents') plt.show()



[25]:

plt.figure(figsize = (20,5)) plt.subplot(1,2,1)

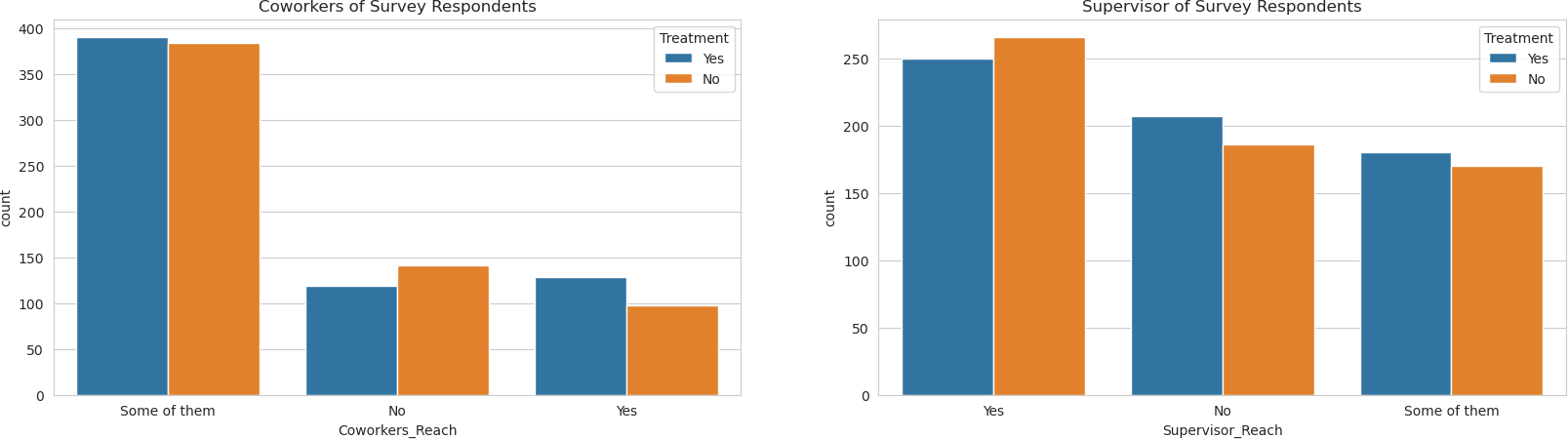
sns.countplot(x = mhdata['Coworkers\_Reach'], hue = mhdata['Treatment']) plt.title('Coworkers of Survey Respondents')

plt.subplot(1,2,2)

sns.countplot(x = mhdata['Supervisor\_Reach'], hue = mhdata['Treatment']) plt.title('Supervisor of Survey Respondents')

plt.show()

Around 18% of the surveyees reveal that they are not with a company that is primarily in the tech field. It also shows that they are more likely to get treatment than those who are working for at a tech company. There are many factors that are at play here. This could also imply that non-tech companies also face a mental health issue among their employees. But it could also suggest that tech companies should promote employee well-being with respect to mental health.



Interestingly, survey shows more positive response with regards to opening up to supervisors than to peers. Respondents who reveal openness to reach out to supervisors are more likely to not get treatment than otherwise. There is no specific pattern with respect to openness to share to coworkers.

[26]:

plt.figure(figsize = (20,5)) plt.subplot(1,2,1)

eda\_percentage = mhdata['Observed\_Consequence\_Workplace'].

𝗌value\_counts(normalize = **True**).rename\_axis('Observed\_Consequence\_Workplace').

𝗌reset\_index(name = 'Percentage')

sns.barplot(x = 'Observed\_Consequence\_Workplace', y = 'Percentage', data =␣

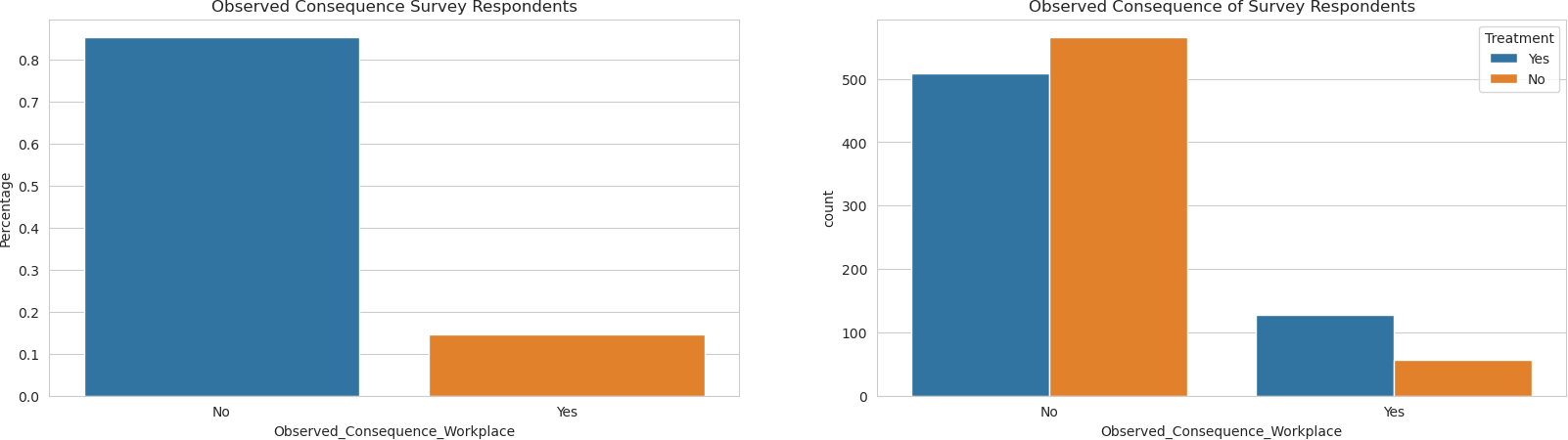
𝗌eda\_percentage)

plt.title('Observed Consequence Survey Respondents') plt.subplot(1,2,2)

sns.countplot(x = mhdata['Observed\_Consequence\_Workplace'], hue =␣

𝗌mhdata['Treatment'])

plt.title('Observed Consequence of Survey Respondents') plt.show()



[27]:

More than 80% respondent that they have not observed negative consequences from coworkers with mental health condition in the workplace. But respondents who say that they do observe negative consequences are more likely to seek medical treatment.

# Workplace benefits, facilities, and confidentiality

plt.figure(figsize = (20,5)) plt.subplot(1,2,1)

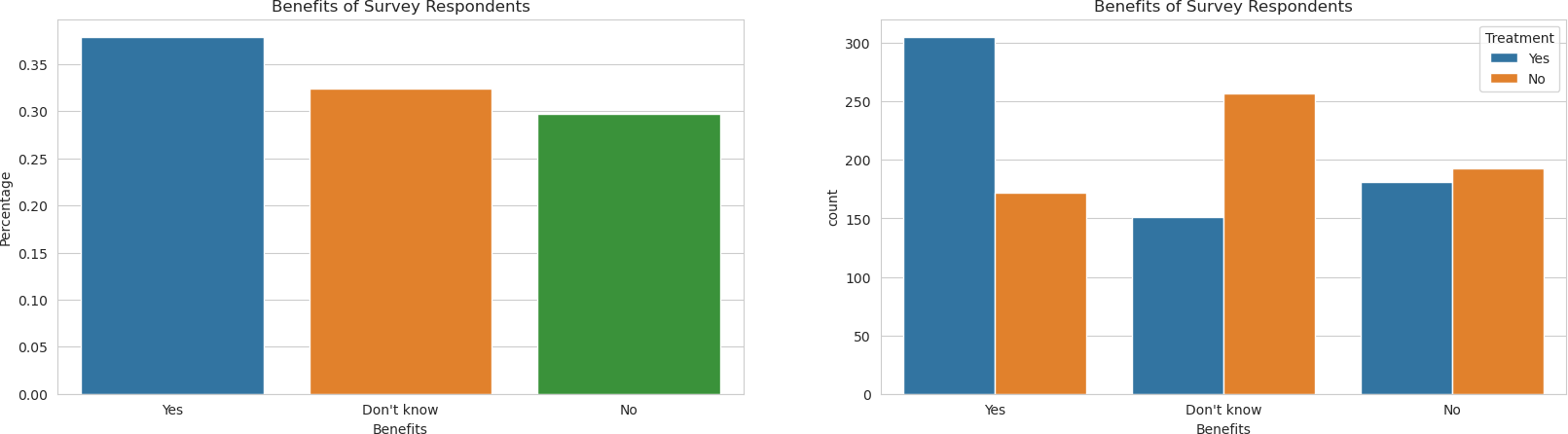
eda\_percentage = mhdata['Benefits'].value\_counts(normalize = **True**).

𝗌rename\_axis('Benefits').reset\_index(name = 'Percentage') sns.barplot(x = 'Benefits', y = 'Percentage', data = eda\_percentage) plt.title('Benefits of Survey Respondents')

plt.subplot(1,2,2)

sns.countplot(x = mhdata['Benefits'], hue = mhdata['Treatment']) plt.title('Benefits of Survey Respondents')

plt.show()



[28]:

plt.figure(figsize = (20,5)) plt.subplot(1,2,1)

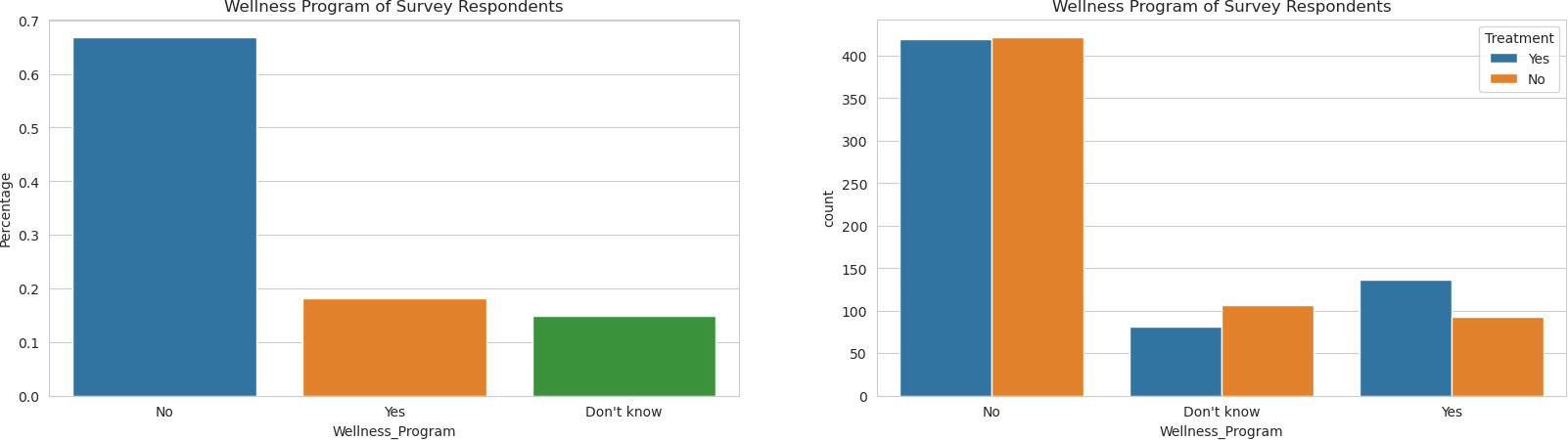
eda\_percentage = mhdata['Wellness\_Program'].value\_counts(normalize = **True**).

𝗌rename\_axis('Wellness\_Program').reset\_index(name = 'Percentage') sns.barplot(x = 'Wellness\_Program', y = 'Percentage', data = eda\_percentage) plt.title('Wellness Program of Survey Respondents')

plt.subplot(1,2,2)

sns.countplot(x = mhdata['Wellness\_Program'], hue = mhdata['Treatment']) plt.title('Wellness Program of Survey Respondents')

plt.show()



[29]:

plt.figure(figsize = (20,5)) plt.subplot(1,2,1)

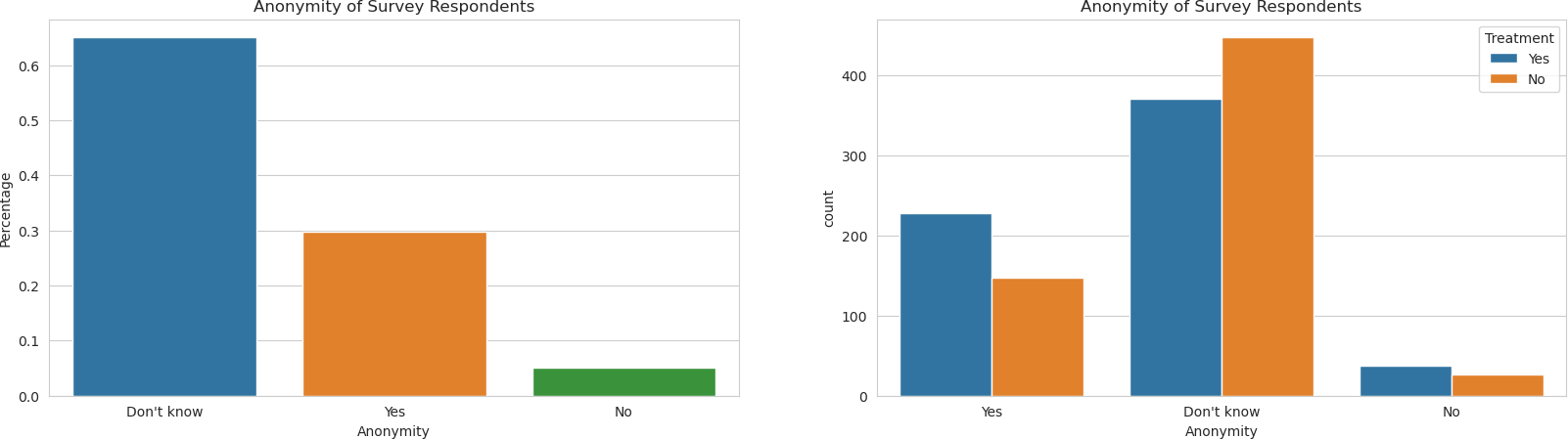
eda\_percentage = mhdata['Anonymity'].value\_counts(normalize = **True**).

𝗌rename\_axis('Anonymity').reset\_index(name = 'Percentage') sns.barplot(x = 'Anonymity', y = 'Percentage', data = eda\_percentage) plt.title('Anonymity of Survey Respondents')

plt.subplot(1,2,2)

sns.countplot(x = mhdata['Anonymity'], hue = mhdata['Treatment']) plt.title('Anonymity of Survey Respondents')

plt.show()



[30]:

[31]:

* It is revealed that only roughly 35% of the respondents are aware of the benefits present in their workplace and those who know are more likely to get medical treatment. The rest of the surveyees either are not aware of the said benefits or there non present. Those who responded otherwise are more likely to not get treatment than those who know the presence of the said benefits.
* Majority of the surveyees reveal the absence of wellness programs in the workplace.
* Majority of the respondents reveal that they are not aware if their identity is protected when consulting or assessing mental health in the workplace. This should not be taken lightly as those who responded that they are aware of the precautionary steps regarding their medical condition are more likely to seek medical condition than those who don’t know.

# Data Pre-processing

mhdata['Treatment'] = np.where(mhdata['Treatment'] == 'Yes', 1, 0)

We now have converted the target variable into boolean values where:

* + - * 1 = Yes, to get medical treatment
      * 2 = No, not getting any medical treatment

mhdata.sample(10)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| [31]: | Age | Gender | Self\_Employed | Family\_History | Treatment | Work\_Interfere | \ |
| 729 | 43 | Male | Yes | No | 1 | Sometimes |  |
| 865 | 34 | Male | No | No | 0 | NaN |  |
| 103 | 33 | Male | No | Yes | 1 | Sometimes |  |
| 1011 | 33 | Female | Yes | No | 0 | Never |  |
| 657 | 32 | Male | Yes | No | 0 | Never |  |
| 456 | 27 | Male | No | No | 0 | Rarely |  |
| 715 | 31 | Male | No | No | 1 | Sometimes |  |
| 604 | 27 | Male | No | No | 0 | NaN |  |
| 459 | 37 | Male | No | No | 0 | NaN |  |
| 38 | 50 | Male | No | No | 0 | NaN |  |

Employee\_Count\_Company Remote\_Work Tech\_Company Benefits Care\_Options \ 729 1-5 Yes Yes No No

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 865 | 6-25 | | | Yes | Yes | No | No | | |
| 103 | 26-100 | | | Yes | Yes | Yes | Yes | | |
| 1011 | 1-5 | | | Yes | Yes | No | Yes | | |
| 657 | 1-5 | | | Yes | Yes | No | No | | |
| 456 | 6-25 | | | No | Yes | No | Yes | | |
| 715 | 26-100 | | | No | No | No | No | | |
| 604 | More than 1000 | | | No | Yes | Yes | Not sure | | |
| 459 | 500-1000 | | | No | Yes | Yes | No | | |
| 38 | 100-500 | | | No | Yes | Yes | Yes | | |
|  | Wellness\_Program | Seek\_Help | | Anonymity | Medical\_Leave | | \ | | |
| 729 | No | No | | Yes | Very difficult | |  | | |
| 865 | No | No | | No | Somewhat difficult | |  | | |
| 103 | Yes | Yes | | Yes | Somewhat easy | |  | | |
| 1011 | Don't know | Don't know | | Yes | Somewhat easy | |  | | |
| 657 | No | No | | Don't know | Don't know | |  | | |
| 456 | No | No | | Don't know | Somewhat easy | |  | | |
| 715 | No | No | | Don't know | Somewhat difficult | |  | | |
| 604 | Don't know | Don't know | | Yes | Don't know | |  | | |
| 459 | No | No | | Don't know | Don't know | |  | | |
| 38 | No | Don't know | | Don't know | Don't know | |  | | |
|  | Mental\_Health\_Consequence | | Physical\_Health\_Consequence | | | Coworkers\_Reach | | | \ |
| 729 | Yes | | No | | | Yes | | |  |
| 865 | Yes | | Maybe | | | No | | |  |
| 103 | No | | No | | | Some of them | | |  |
| 1011 | Maybe | | No | | | No | | |  |
| 657 | No | | No | | | Some of them | | |  |
| 456 | Maybe | | No | | | Some of them | | |  |
| 715 | Yes | | No | | | No | | |  |
| 604 | Maybe | | Maybe | | | No | | |  |
| 459 | Maybe | | Maybe | | | Some of them | | |  |
| 38 | No | | No | | | Some of them | | |  |
|  | Supervisor\_Reach | Mental\_Health\_Interview | | | Physical\_Health\_Interview | | | \ | |
| 729 | No | No | | | Maybe | | |  | |
| 865 | No | No | | | Yes | | |  | |
| 103 | Yes | Maybe | | | Maybe | | |  | |
| 1011 | Some of them | No | | | No | | |  | |
| 657 | Yes | No | | | Maybe | | |  | |
| 456 | Some of them | No | | | Maybe | | |  | |
| 715 | No | No | | | Maybe | | |  | |
| 604 | No | No | | | Maybe | | |  | |
| 459 | Some of them | No | | | No | | |  | |
| 38 | Yes | No | | | Maybe | | |  | |

Mental\_VS\_Physical Observed\_Consequence\_Workplace

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 729 |  | No | Yes |
| 865 |  | No | Yes |
| 103 |  | Yes | No |
| 1011 |  | Yes | No |
| 657 | Don't | know | No |
| 456 |  | No | No |
| 715 | Don't | know | No |
| 604 |  | No | No |
| 459 | Don't | know | No |
| [32]: | 38 | Don't | know | No |

[33]:

*# Let's see how X looks like with the dummy variables*

X.head()

*# Now, we separate our dependent and independent variables.*

X = mhdata.drop(["Treatment"], axis=1) Y = mhdata["Treatment"]

*# The independent variables will be transformed into dummy variables*

X = pd.get\_dummies(X, drop\_first=**True**)

*# adding constant. This is a requirement of Stats Model library. It creates a*␣

𝗌*new column with float value 1*

X = sm.add\_constant(X)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| [33]: | const | Age | Gender\_Male | Gender\_Queer | Self\_Employed\_Yes | \ |
| 0 | 1.0 | 37 | 0 | 0 | 0 |  |
| 1 | 1.0 | 44 | 1 | 0 | 0 |  |
| 2 | 1.0 | 32 | 1 | 0 | 0 |  |
| 3 | 1.0 | 31 | 1 | 0 | 0 |  |
| 4 | 1.0 | 31 | 1 | 0 | 0 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Family\_History\_Yes | Work\_Interfere\_Often | Work\_Interfere\_Rarely | \ |
| 0 0 | 1 | 0 |  |
| 1 0 | 0 | 1 |  |
| 2 0 | 0 | 1 |  |
| 3 1 | 1 | 0 |  |
| 4 0 | 0 | 0 |  |
| Work\_Interfere\_Sometimes Employee\_Count\_Company\_100-500 \ | | | |
| 0 | 0 | 0 | |
| 1 | 0 | 0 | |
| 2 | 0 | 0 | |
| 3 | 0 | 0 | |
| 4 | 0 | 1 | |

Employee\_Count\_Company\_26-100 Employee\_Count\_Company\_500-1000 \

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | | | | 0 | |
| 1 | 0 | | | | 0 | |
| 2 | 0 | | | | 0 | |
| 3 | 1 | | | | 0 | |
| 4 | 0 | | | | 0 | |
|  | Employee\_Count\_Company\_6-25 Employee\_Count\_Company\_More than | | | | 1000 | \ |
| 0 | 1 | | | | 0 |  |
| 1 | 0 | | | | 1 |  |
| 2 | 1 | | | | 0 |  |
| 3 | 0 | | | | 0 |  |
| 4 | 0 | | | | 0 |  |
|  | Remote\_Work\_Yes Tech\_Company\_Yes Benefits\_No Benefits\_Yes | | | | \ |  |
| 0 | 0 | 1 | 0 | 1 | | |
| 1 | 0 | 0 | 0 | 0 | | |
| 2 | 0 | 1 | 1 | 0 | | |
| 3 | 0 | 1 | 1 | 0 | | |
| 4 | 1 | 1 | 0 | 1 | | |
| Care\_Options\_Not sure Care\_Options\_Yes Wellness\_Program\_No \ | | | | | | |

\

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0 | 1 | 0 | | 1 | |
| 1 | 0 | 0 | | 0 | |
| 2 | 0 | 0 | | 1 | |
| 3 | 0 | 1 | | 1 | |
| 4 | 0 | 0 | | 0 | |
| Wellness\_Program\_Yes Seek\_Help\_No Seek\_Help\_Yes Anonymity\_No | | | | | |
| 0 | 0 | 0 | 1 | | 0 |
| 1 | 0 | 0 | 0 | | 0 |
| 2 | 0 | 1 | 0 | | 0 |
| 3 | 0 | 1 | 0 | | 1 |
| 4 | 0 | 0 | 0 | | 0 |
| Anonymity\_Yes Medical\_Leave\_Somewhat difficult \ | | | | | |

\

|  |  |  |  |
| --- | --- | --- | --- |
| 0 | 1 | 0 | |
| 1 | 0 | 0 | |
| 2 | 0 | 1 | |
| 3 | 0 | 1 | |
| 4 | 0 | 0 | |
| Medical\_Leave\_Somewhat easy Medical\_Leave\_Very difficult | | | |
| 0 | 1 | | 0 |
| 1 | 0 | | 0 |
| 2 | 0 | | 0 |
| 3 | 0 | | 0 |
| 4 | 0 | | 0 |

Medical\_Leave\_Very easy Mental\_Health\_Consequence\_No \

|  |  |  |
| --- | --- | --- |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 2 | 0 | 1 |
| 3 | 0 | 0 |
| 4 | 0 | 1 |

Mental\_Health\_Consequence\_Yes Physical\_Health\_Consequence\_No \

0 0 1

1 0 1

2 0 1

3 1 0

4 0 1

Physical\_Health\_Consequence\_Yes Coworkers\_Reach\_Some of them \

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 0 | | 1 | |
| 1 | 0 | | 0 | |
| 2 | 0 | | 0 | |
| 3 | 1 | | 1 | |
| 4 | 0 | | 1 | |
| Coworkers\_Reach\_Yes Supervisor\_Reach\_Some of them Supervisor\_Reach\_Yes | | | | |
| 0 | 0 | 0 | | 1 |
| 1 | 0 | 0 | | 0 |
| 2 | 1 | 0 | | 1 |
| 3 | 0 | 0 | | 0 |
| 4 | 0 | 0 | | 1 |
| Mental\_Health\_Interview\_No Mental\_Health\_Interview\_Yes \ | | | | |

\

\

|  |  |  |  |
| --- | --- | --- | --- |
| 0 | 1 | 0 | |
| 1 | 1 | 0 | |
| 2 | 0 | 1 | |
| 3 | 0 | 0 | |
| 4 | 0 | 1 | |
| Physical\_Health\_Interview\_No Physical\_Health\_Interview\_Yes | | | |
| 0 | 0 | | 0 |
| 1 | 1 | | 0 |
| 2 | 0 | | 1 |
| 3 | 0 | | 0 |
| 4 | 0 | | 1 |
| Mental\_VS\_Physical\_No Mental\_VS\_Physical\_Yes \ | | | |

|  |  |  |
| --- | --- | --- |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 2 | 1 | 0 |

|  |  |  |
| --- | --- | --- |
| 3 | 1 | 0 |
| 4 | 0 | 0 |

Observed\_Consequence\_Workplace\_Yes

0 0

1 0

2 0

3 1

4 0

[34]:

*# checking the shape of our predictor for treatment*

X.shape

[34]: (1259, 45)

From 26 independent variables, replacing them with dummy variables expanded the columns to 45

[35]:

*# Splitting data in train and test sets*

X\_train, X\_test, y\_train, y\_test = train\_test\_split( X, Y, test\_size=0.30, random\_state=1

)

[36]:

*# defining a function to compute different metrics to check performance of a*␣

𝗌*classification model built using statsmodels*

**def** model\_performance\_classification(model, predictors, target, threshold=0.5):

*"""*

*Function to compute different metrics to check classification model*␣

𝗌*performance*

*model: classifier*

*predictors: independent variables target: dependent variable*

*threshold: threshold for classifying the observation as class 1 """*

*# checking which probabilities are greater than threshold*

pred\_temp = model.predict(predictors) > threshold *# rounding off the above values to get classes* pred = np.round(pred\_temp)

acc = accuracy\_score(target, pred) *# to compute Accuracy*

recall = recall\_score(target, pred) *# to compute Recall* precision = precision\_score(target, pred) *# to compute Precision* f1 = f1\_score(target, pred) *# to compute F1-score*

*# creating a dataframe of metrics*

df\_perf = pd.DataFrame(

{

"Accuracy": acc, "Recall": recall, "Precision": precision, "F1": f1,

},

index=[0],

)

**return** df\_perf