



AN ANALYSIS OF

# Mental Health in the Tech Industry

A 2014 dataset dive by Team 1 - Taylor  
Bauer

```
data = pd.read_csv("mental-health.csv")
data.columns
```

```
Index(['Timestamp', 'Age', 'Gender', 'Country', 'state', 'self_employed',
       'family_history', 'treatment', 'work_interfere', 'no_employees',
       'remote_work', 'tech_company', 'benefits', 'care_options',
       'wellness_program', 'seek_help', 'anonymity', 'leave',
       'mental_health_consequence', 'phys_health_consequence', 'coworkers',
       'supervisor', 'mental_health_interview', 'phys_health_interview',
       'mental_vs_physical', 'obs_consequence', 'comments'],
```

	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees	...	leave	mental_health_consequence	
0	2014-08-27 11:29:31	37	Female	United States	IL	NaN	No	Yes	Often	6-25	...	Somewhat easy	No	
1	2014-08-27 11:29:37	44	M	United States	IN	NaN	No	No	Rarely	More than 1000	...	Don't know	Maybe	
2	2014-08-27 11:29:44	32	Male	Canada	NaN	NaN	No	No	Rarely	6-25	...	Somewhat difficult	No	
3	2014-08-27 11:29:46	31	Male	United Kingdom	NaN	NaN	Yes	Yes	Often	26-100	...	Somewhat difficult	Yes	
4	2014-08-27 11:30:22	31	Male	United States	TX	NaN	No	No	Never	100-500	...	Don't know	No	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1254	2015-09-12 11:17:21	26	male	United Kingdom	NaN	NaN	No	No	Yes	NaN	26-100	...	Somewhat easy	No
1255	2015-09-26 01:07:35	32	Male	United States	IL	No	Yes	Yes	Often	26-100	...	Somewhat difficult	No	
1256	2015-11-07 12:36:58	34	male	United States	CA	No	Yes	Yes	Sometimes	More than 1000	...	Somewhat difficult	Yes	
1257	2015-11-30 21:25:06	46	f	United States	NC	No	No	No	NaN	100-500	...	Don't know	Yes	
1258	2016-02-01 23:04:31	25	Male	United States	IL	No	Yes	Yes	Sometimes	26-100	...	Don't know	Maybe	

1259 rows × 27 columns

phys_health_consequence	coworkers	supervisor	mental_health_interview	phys_health_interview	mental_vs_physical	obs_consequence	comments
No	Some of them	Yes		No	Maybe	Yes	No
No	No	No		No	No	Don't know	No
No	Yes	Yes		Yes	Yes	No	No
Yes	Some of them	No	Maybe	Maybe	No	Yes	NaN
No	Some of them	Yes		Yes	Yes	Don't know	No
...	...	...		...	...	...	...
No	Some of them	Some of them		No	No	Don't know	No
No	Some of them	Yes		No	No	Yes	No
Yes	No	No		No	No	No	No
No	No	No		No	No	No	No
No	Some of them	No		No	No	Don't know	No

# The Data: Mental Health in Tech Survey - 2014

- **Origin:**

  - Kaggle

- **Demographic Information:**

  - Age, Gender, State, Country, Employment Status

- **Mental Health Information:**

  - Treatment sought for mental health, family history, work interference

- **Workplace Details:**

  - Tech company affiliation, benefits provided by employer, awareness of mental health care options, and more.

- **Attitudes and Perceptions:**

  - Willingness to discuss mental health with coworkers, supervisors, job interviews, and comparison to attitudes about physical health

- **Several other data features**

# Data Queries

WHAT VARIABLES  
CONTRIBUTE TO SOUGHT  
MENTAL HEALTH  
TREATMENT?



## State-Wise

**Question:** Are there specific states where technology employees more likely to seek mental health treatment?

## Impact of Remote Work

**Question:** Does remote work influence tech employees seeking mental health treatment?

## Awareness and Comfortability

**Question:** Are there differences in sought mental health treatment based on awareness of resources and comfort level of discussing mental health issues with coworkers or supervisors?

## Gender

**Question:** Does gender have an impact on seeking mental health treatment?



## Benefits!

- Inclusive, assistive, and increased workplace for employees.
- Improving remote work conditions and practices.
- Supporting employee well-being to improve productivity.
- Identifying regional disparities.
- **Mental Health Awareness in the workplace.**

# Data Preparation

## CLEANING & ORGANIZING

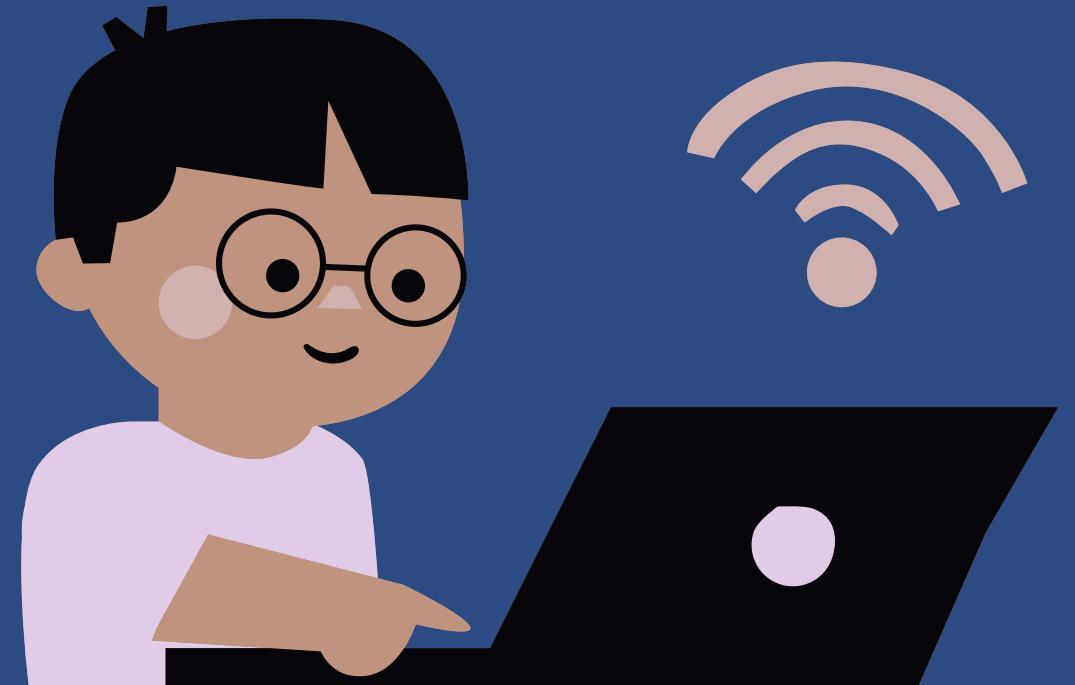
Several of columns were dropped due to the analysis focusing on the following variables:

- treatment (target)
- age
- gender
- country (US)
- state
- remote\_work
- seek\_help
- coworkers
- supervisors
- tech\_company (Yes)

```
# dropping all null values
survey = survey.dropna().reset_index(drop=True)
```

```
survey.head()
```

	Age	Gender	Country	state	self_employed	treatment	remote_work	tech_company	seek_help	coworkers	supervisor
0	46	male	United States	MD		Yes	No	Yes	Yes	Don't know	Yes
1	29	Male	United States	NY		No	Yes	No	Yes	No	Some of them
2	31	male	United States	NC		Yes	No	Yes	Yes	No	Some of them
3	46	Male	United States	MA		No	Yes	Yes	Yes	No	Some of them
4	41	Male	United States	IA		No	Yes	No	No	Don't know	No





# Data Preparation

## CLEANING & ORGANIZING

```
# dropping all null values
survey = survey.dropna().reset_index(drop=True)
survey.head()
```



Dropped all entries with null values - to increase the quality of analysis

```
# only usa entries
df1 = survey[survey["Country"] == "United States"]
df1 = df1.reset_index(drop=True)
```



Focused in on analysis solely for entries in the United States of America

```
# only those who are employed by company
df2 = df1[df1["self-employed"] == "No"]
df2 = df2.reset_index(drop=True)
```



Concentration of company employees - some picked features relate to company environment

```
# work in a primarily tech company
df3 = df2[df2["tech_company"] == "Yes"]
df3 = df3.reset_index(drop=True)
```



Wanted analysis to deal with companies who were primarily regarded as a “Tech Company”

# Data Preparation

## CLEANING & ORGANIZING

```
# checking gender values
gender = df3["Gender"]
gender.unique()

array(['Male', 'male', 'Female', 'female', 'M', 'Male-ish', 'maile',
       'Trans-female', 'F', 'Cis Male', 'm', 'Male (CIS)', 'f',
       'queer/she/they', 'non-binary', 'Femake', 'Make', 'Genderqueer',
       'Female ', 'Male ', 'Man', 'msle', 'Female (trans)',
       'Female (cis)', 'Mail', 'cis male', 'Woman'], dtype=object)
```

```
# replacing female identifying entries with "Female"
f_replace = ["Female", "female", "Trans-female", "F", "f", "Femake", "Female ",
             "Female (trans)", "Female (cis)", "Woman"]
df3["Gender"] = df3["Gender"].replace(f_replace, "Female")

# replacing male identifying entries with "Male"
m_replace = ["Male", "male", "M", "maile", "Cis Male", "m", "Male (CIS)",
             "Make", "Male ", "Man", "msle", "Mail", "cis male"]
df3["Gender"] = df3["Gender"].replace(m_replace, "Male")

# replacing non-conforming identifying entries with "Non-conforming"
nc_replace = ["Male-ish", "queer/she/they", "non-binary", "Genderqueer"]
df3["Gender"] = df3["Gender"].replace(nc_replace, "Non-Conforming")
```

- Due to the gender survey question being free response, there were numerous different gender entries
  - Ultimately, I wanted to change the values to make it easier for future use
  - I struggled with choosing the values to categorize, because I wanted to respect participants gender identities
- 
- Decided these could possibly best fit into three categories, which would help with my future data exploration
  - “Female” maps to those who identified as cis-female or trans-woman
  - “Male” maps to entries who identify with male, as there were no trans-men entries
  - “Non-Conforming” maps to those who were uncertain or non-conforming to gender roles



# Data Preparation

## CLEANING & ORGANIZING

```
# checking age col  
df3["Age"].unique()
```

```
array([ 29,  46,  33,  35,  34,  42,  40,  27,  50,  30,  38,  22,  32,  
       24,  36,  23,  25,  31,  44,  28,  45,  18,  39,  26,  43,  37,  
       41,  60,  54,  329,  21,  55,  57,  58,  48,  47,  62,  56,  49,  
       5,  20,  51,  53,  19], dtype=int64)
```

- In addition to gender, age was also a free response survey question
- This can lead to typos and incorrect values in the data set
- Wanted to ensure all ages were realistic so any descriptive statistics on age were not skewed

```
# dropping entry  
survey_clean = df3[(df3["Age"] != 329) & (df3["Age"] != 5)].reset_index(drop=True)
```

- Found two outliers for the context - entries were dropped

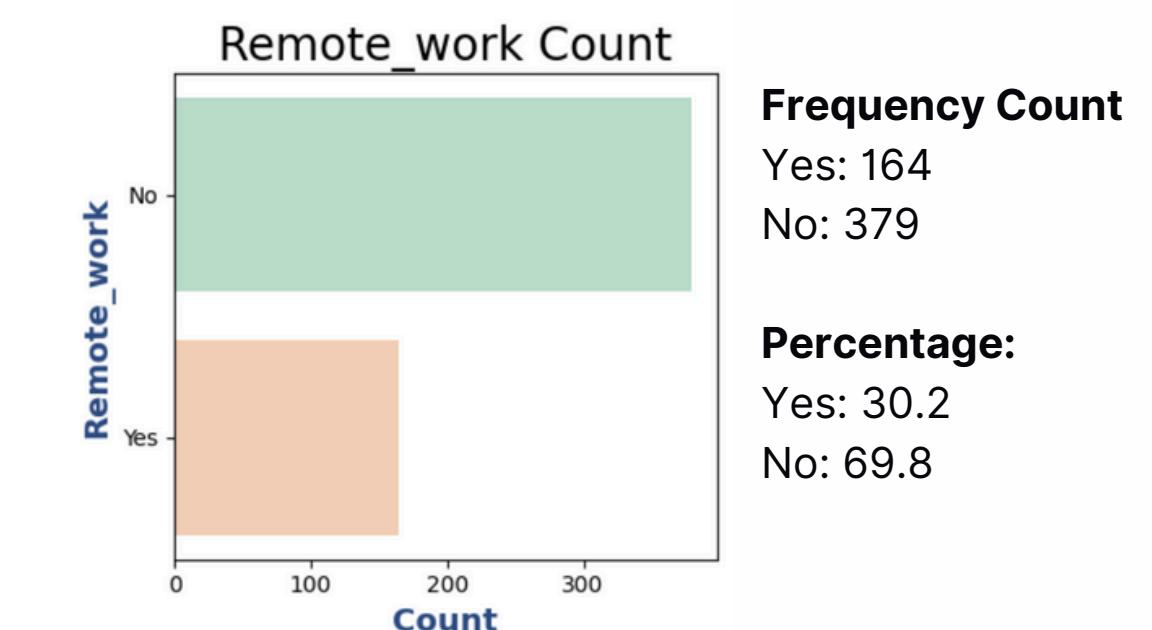
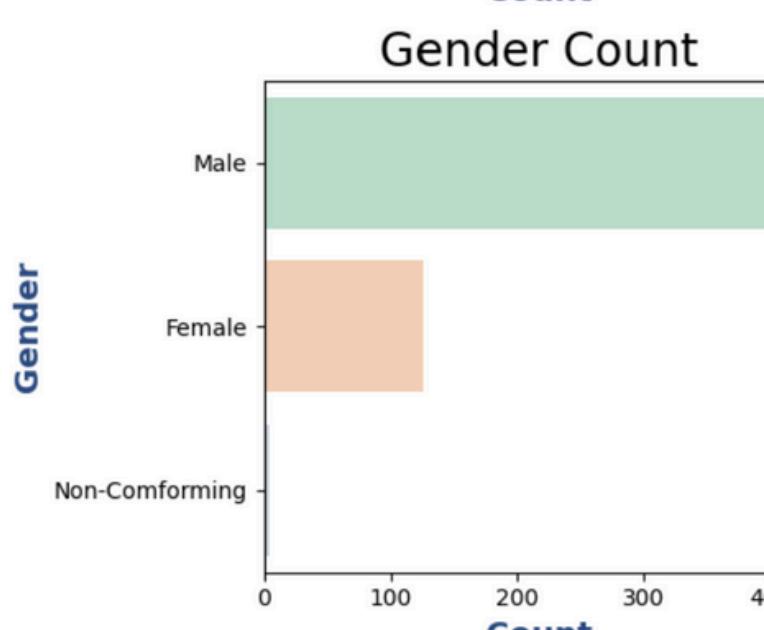
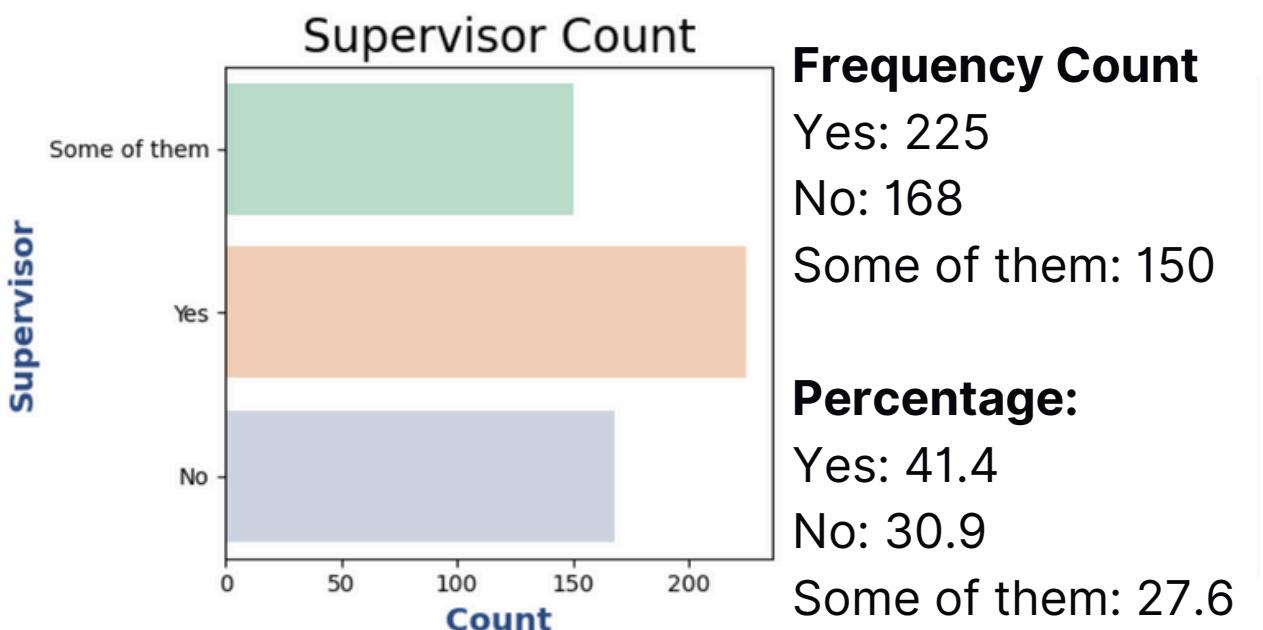
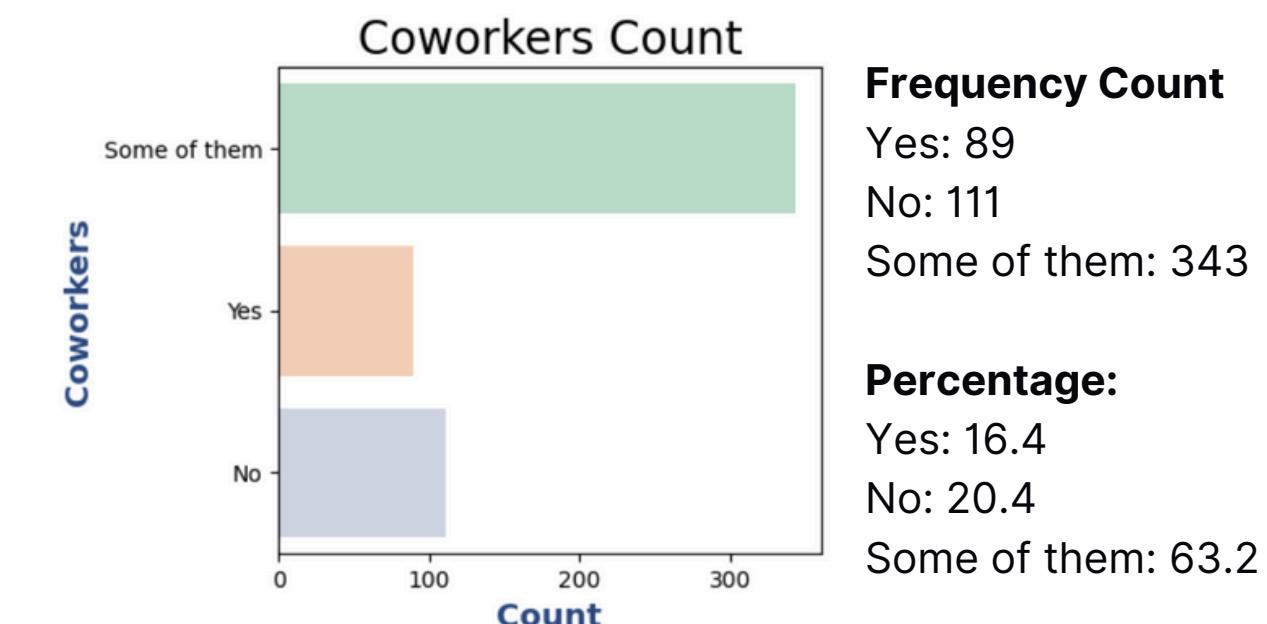
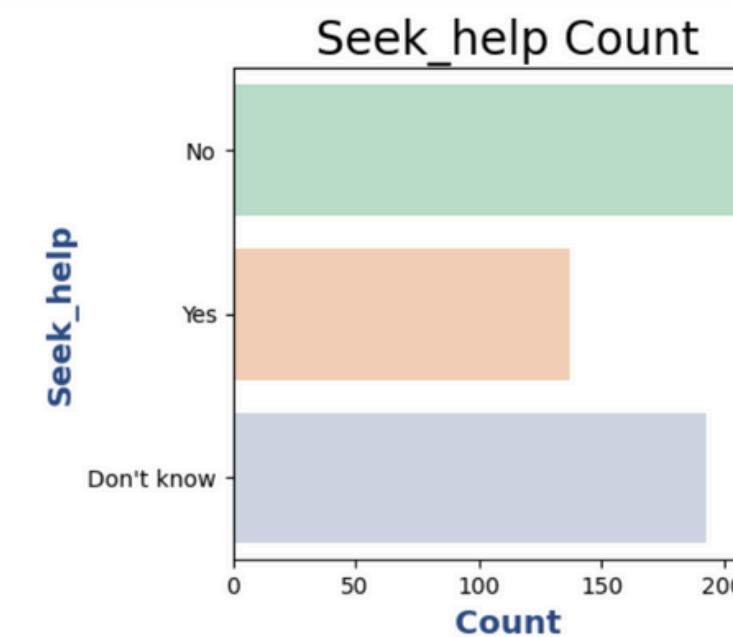
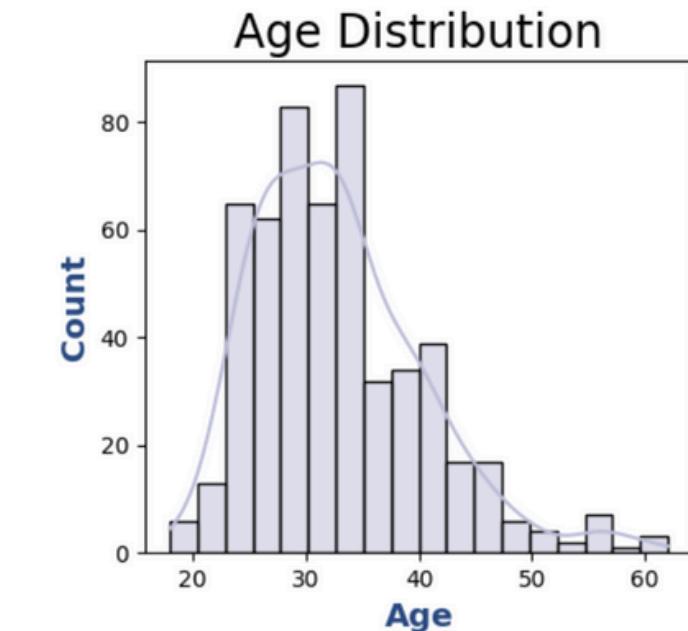
	Age	Gender	Country	state	self_employed	treatment	remote_work	tech_company	seek_help	coworkers	supervisor
0	29	Male	United States	NY	No	Yes	No	Yes	No	Some of them	Some of them
1	46	Male	United States	MA	No	Yes	Yes	Yes	No	Some of them	Yes
2	33	Male	United States	CA	No	Yes	No	Yes	Yes	Yes	Yes
3	33	Male	United States	TN	No	No	No	Yes	Don't know	Some of them	No
4	35	Female	United States	CA	No	Yes	Yes	Yes	Don't know	Yes	Yes

# Hypothesis

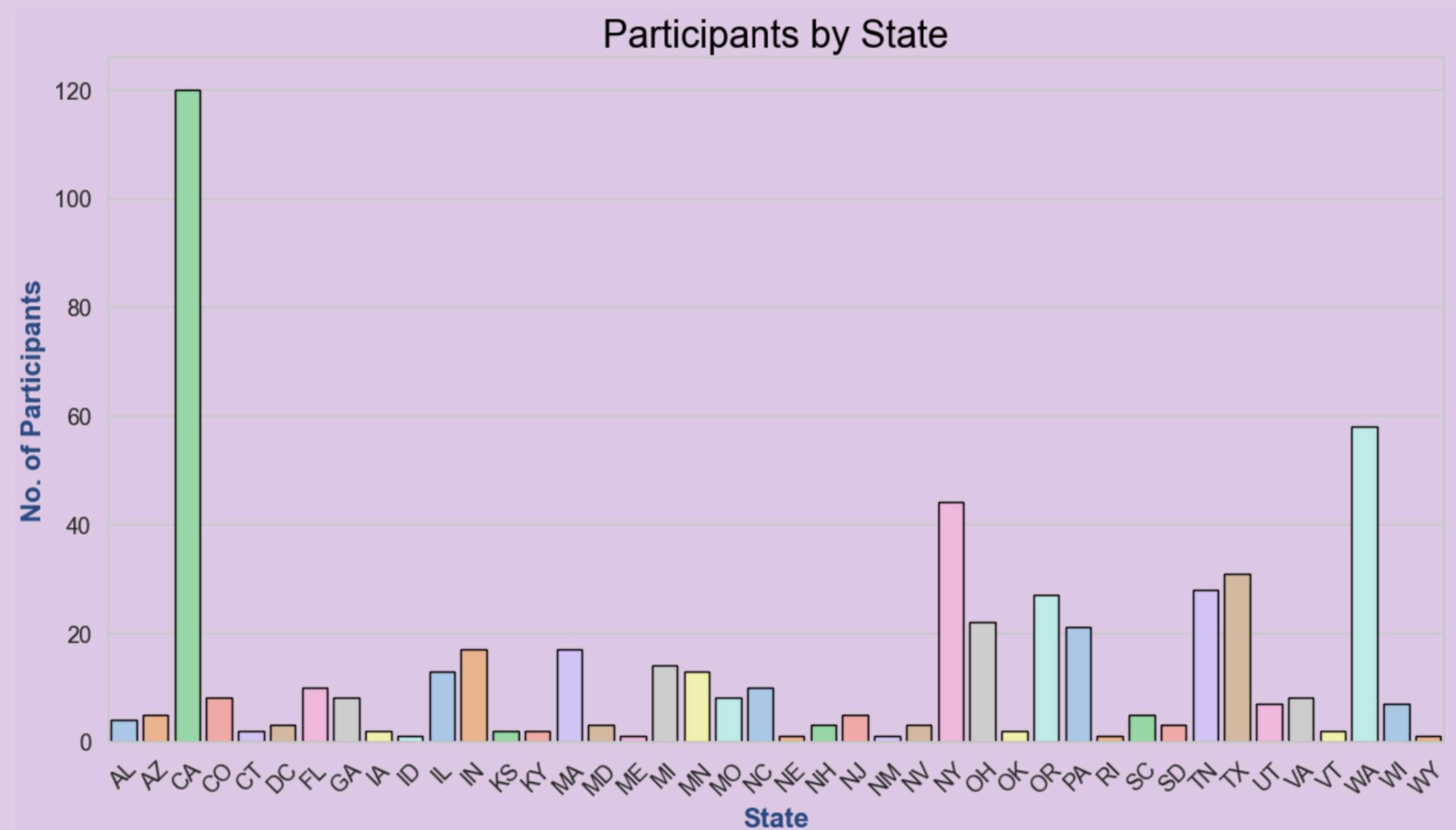
- Remote work will have a fairly substantial impact on whether treatment is sought
- Women will have sought more treatment
- There will be less elder people and more young people
- More supportive and comfortable atmospheres will have less sought treatment



# Descriptive Stats - Variable Distributions



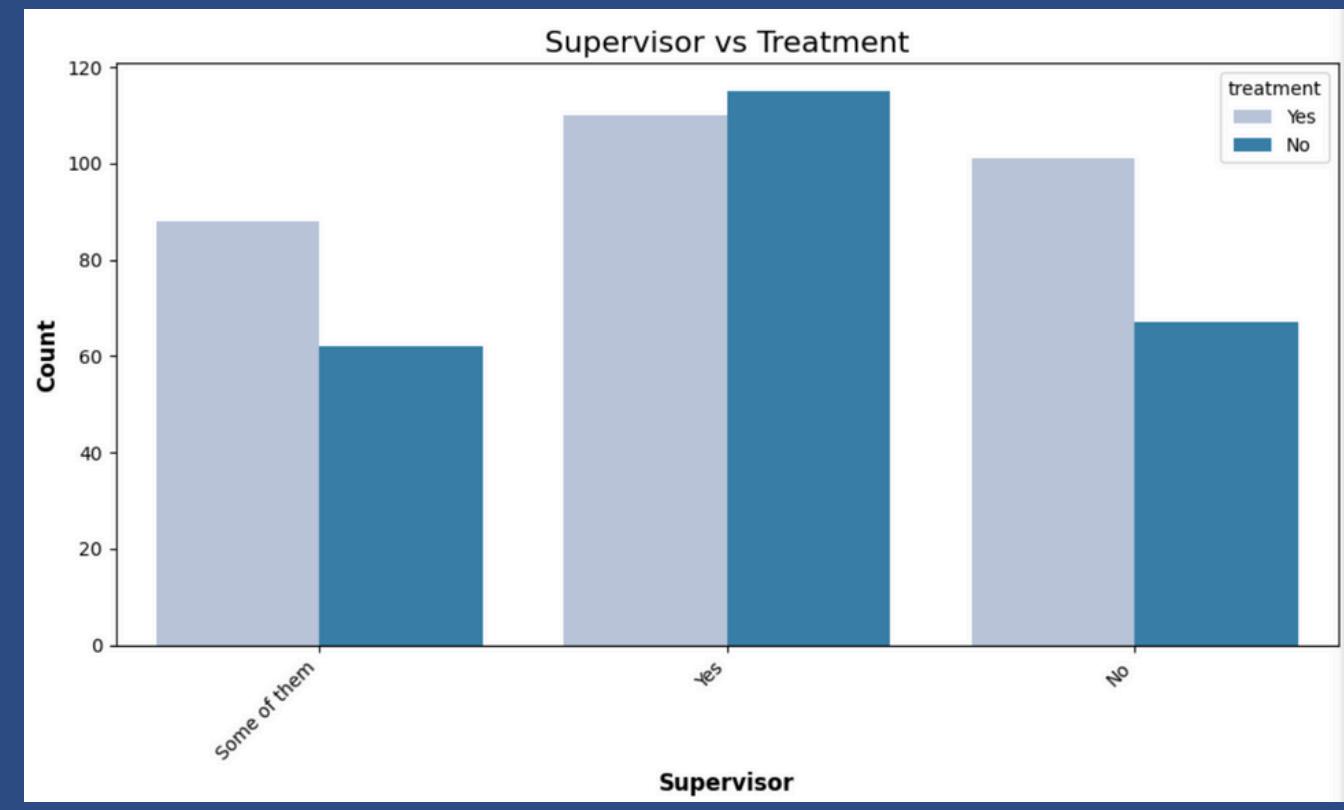
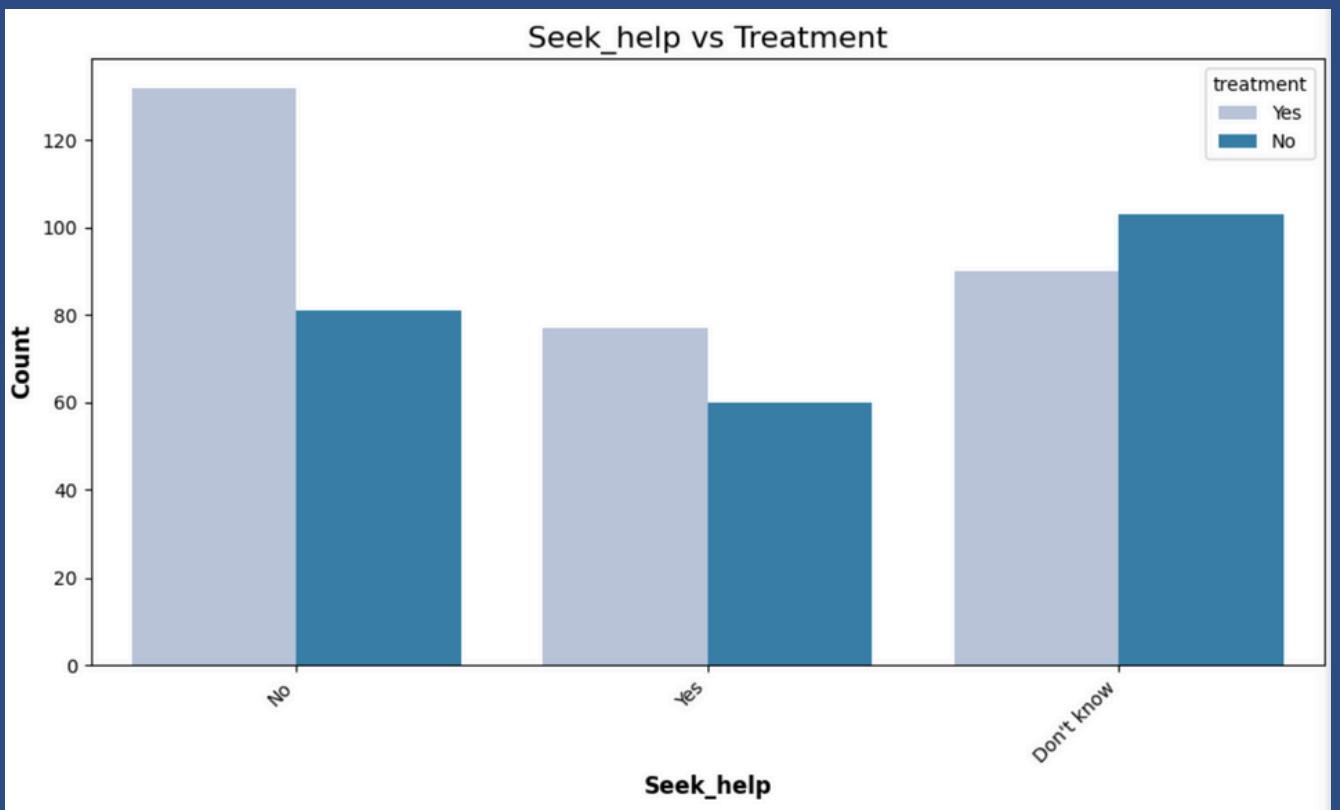
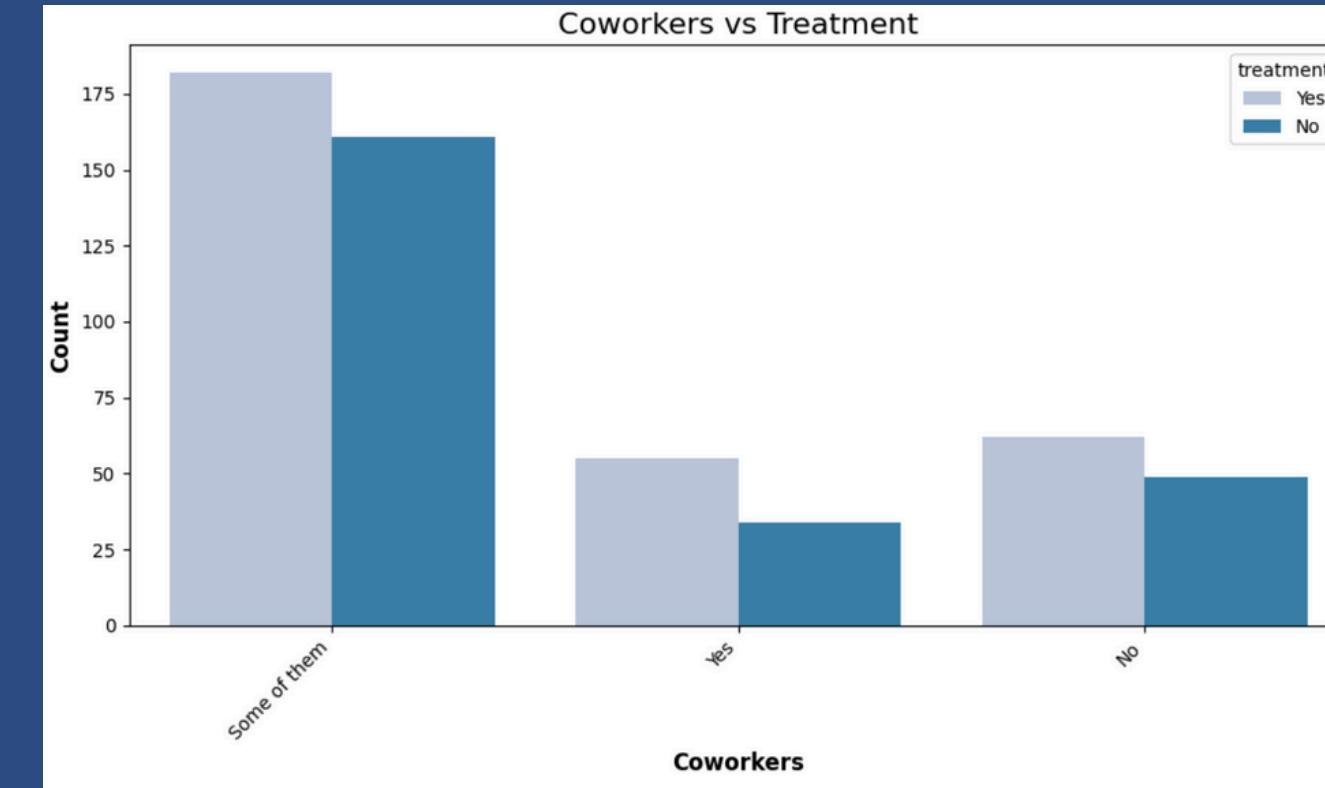
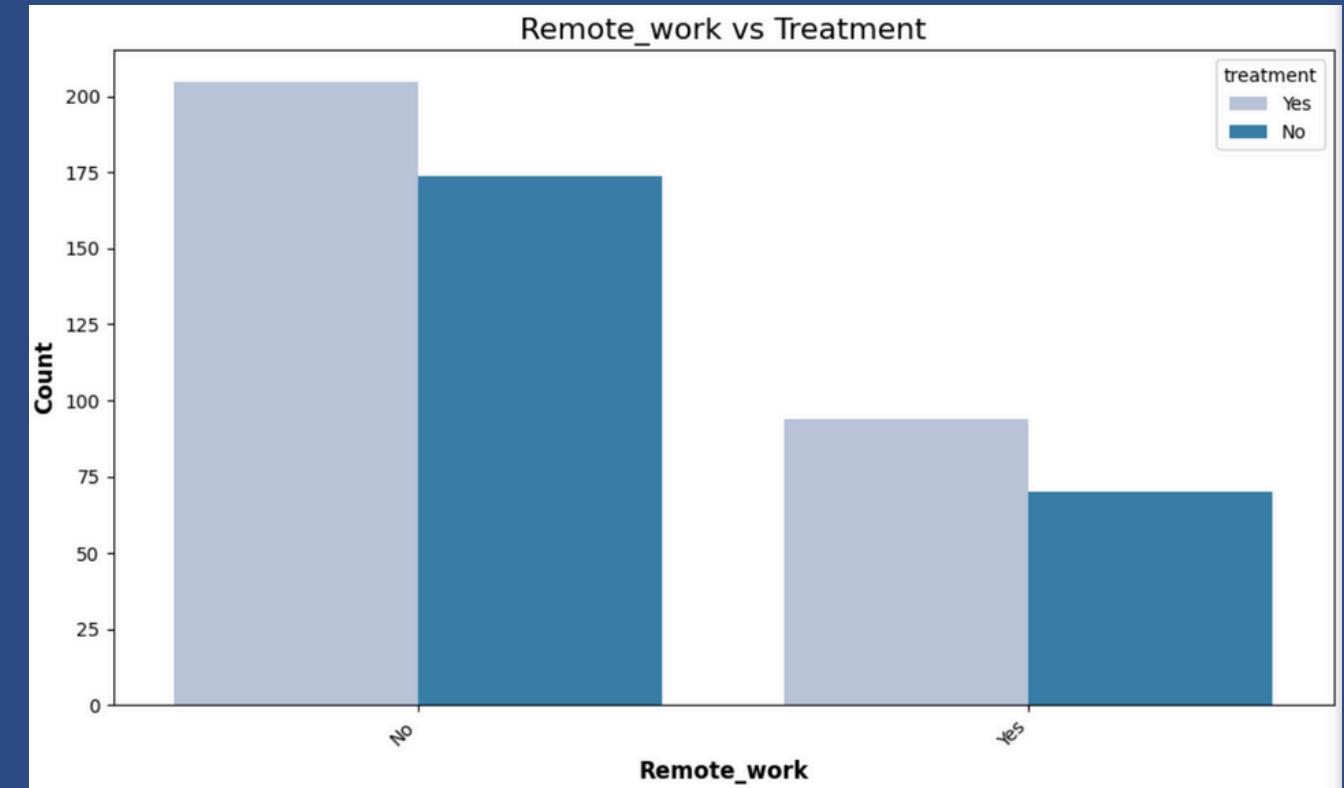
# Descriptive Stats - Variable Distributions



Frequency Counts for States

AL	4	NC	10
AZ	5	NE	1
CA	120	NH	3
CO	8	NJ	5
CT	2	NM	1
DC	3	NV	3
FL	10	NY	44
GA	8	OH	22
IA	2	OK	2
ID	1	OR	27
IL	13	PA	21
IN	17	RI	1
KS	2	SC	5
KY	2	SD	3
MA	17	TN	28
MD	3	TX	31
ME	1	UT	7
MI	14	VA	8
MN	13	VT	2
MO	8	WI	7
WA	58	WY	1

# Variable Interactions



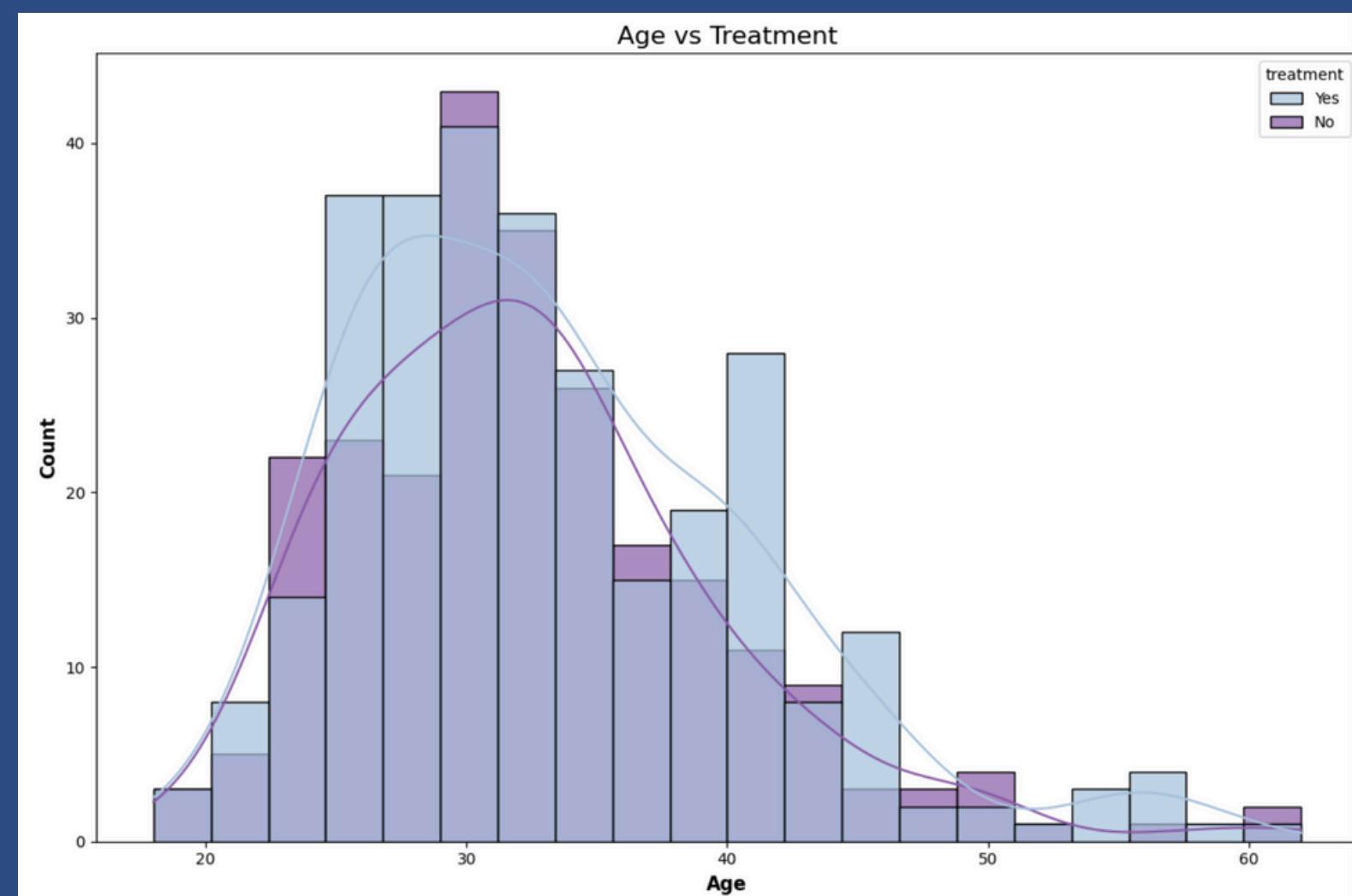
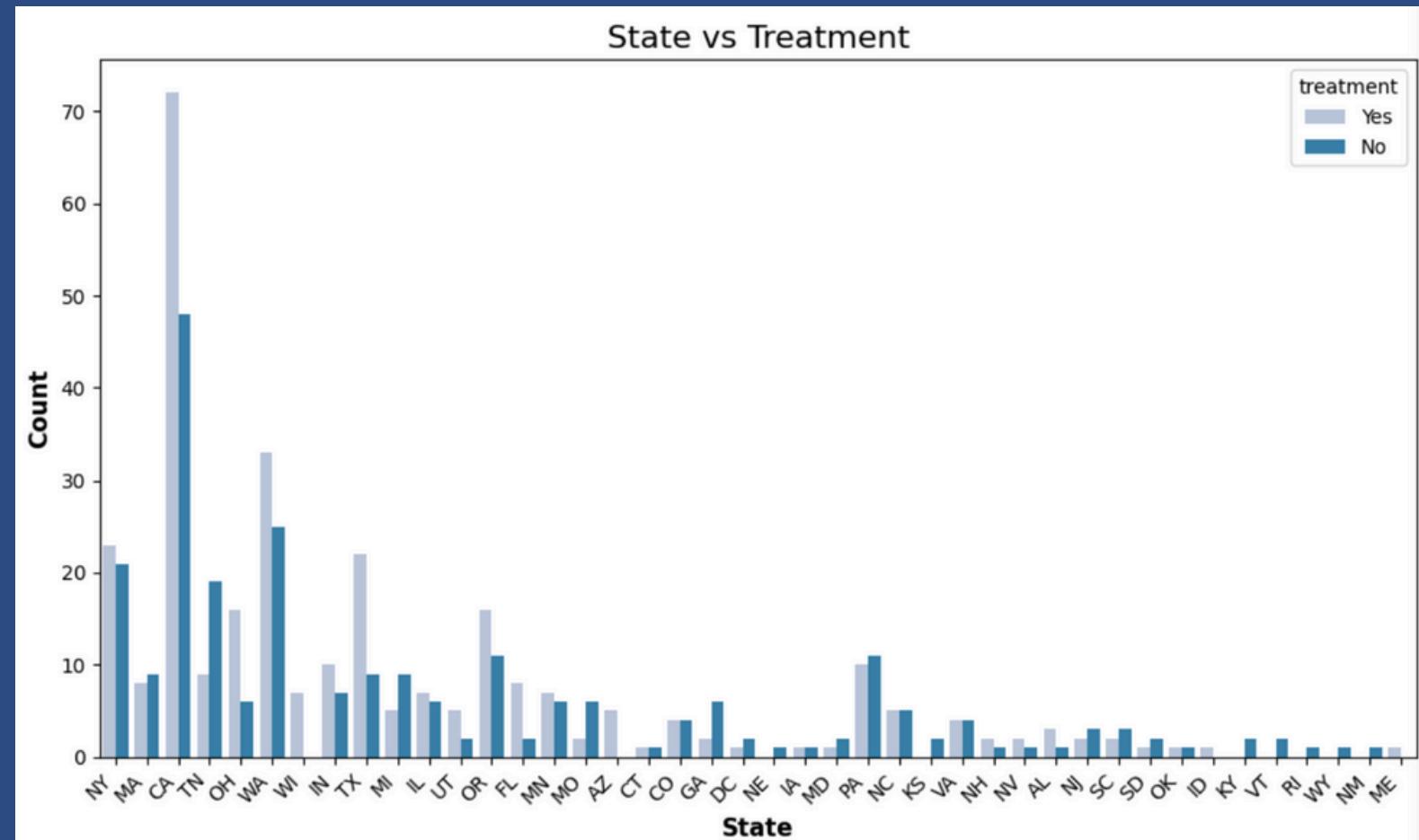
# Variable Interactions

## California:

- Seems to have the highest number of sought treatment.
  - However, also has the largest number of tech workers

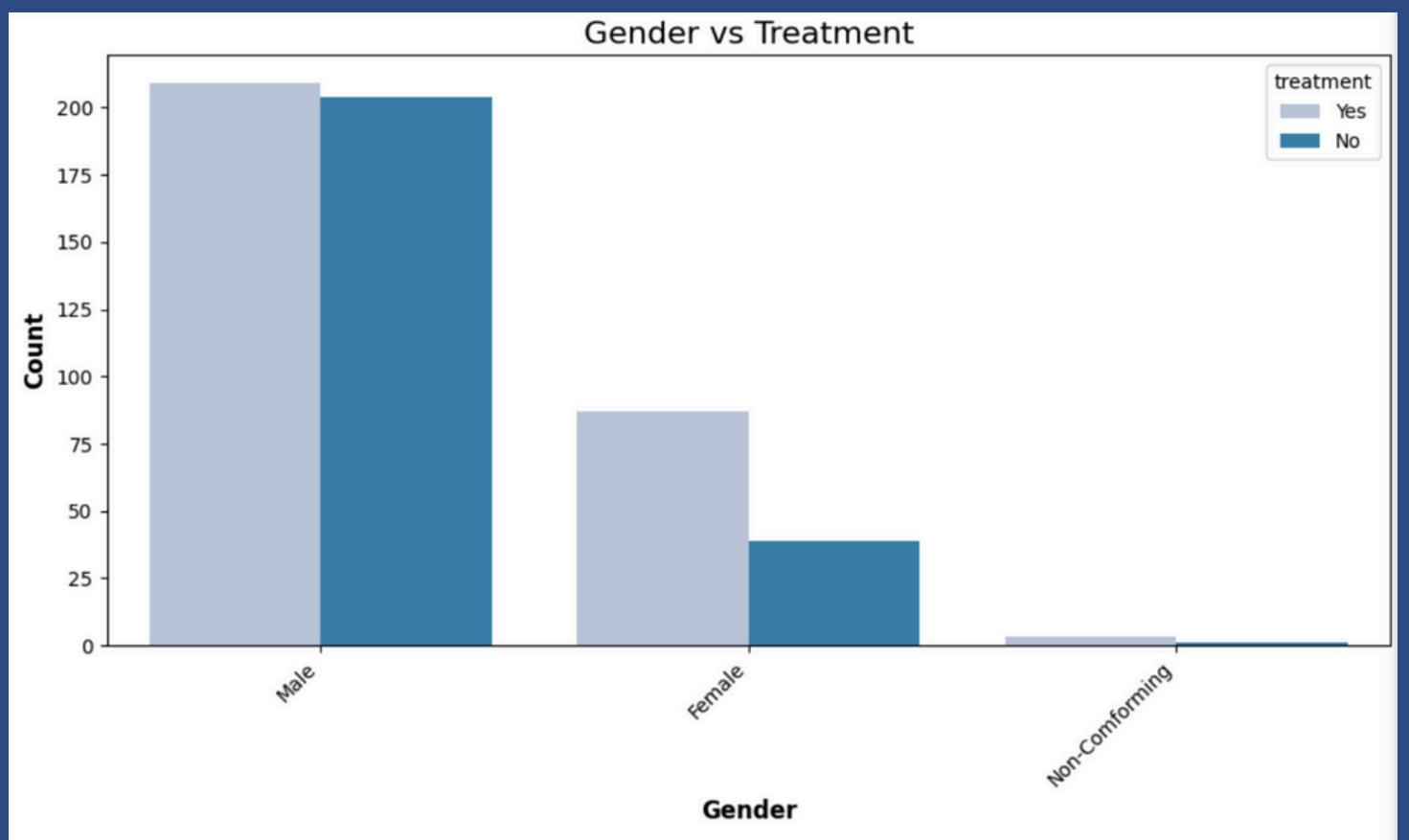
## Mean age for treatment:

- **Yes** (sought treatment):
  - 33.19
- **No** (did not seek treatment)
  - 32.25
- Not a large difference in age
  - Similar to mean age of entire dataset
  - Could be due to numerous tech workers are typically this age in industry



# Variable Interactions

Gender	Yes	No
Female	69.0%	20.0%
Male	50.6%	44.0%
Non-Conforming	75.0%	1.3%

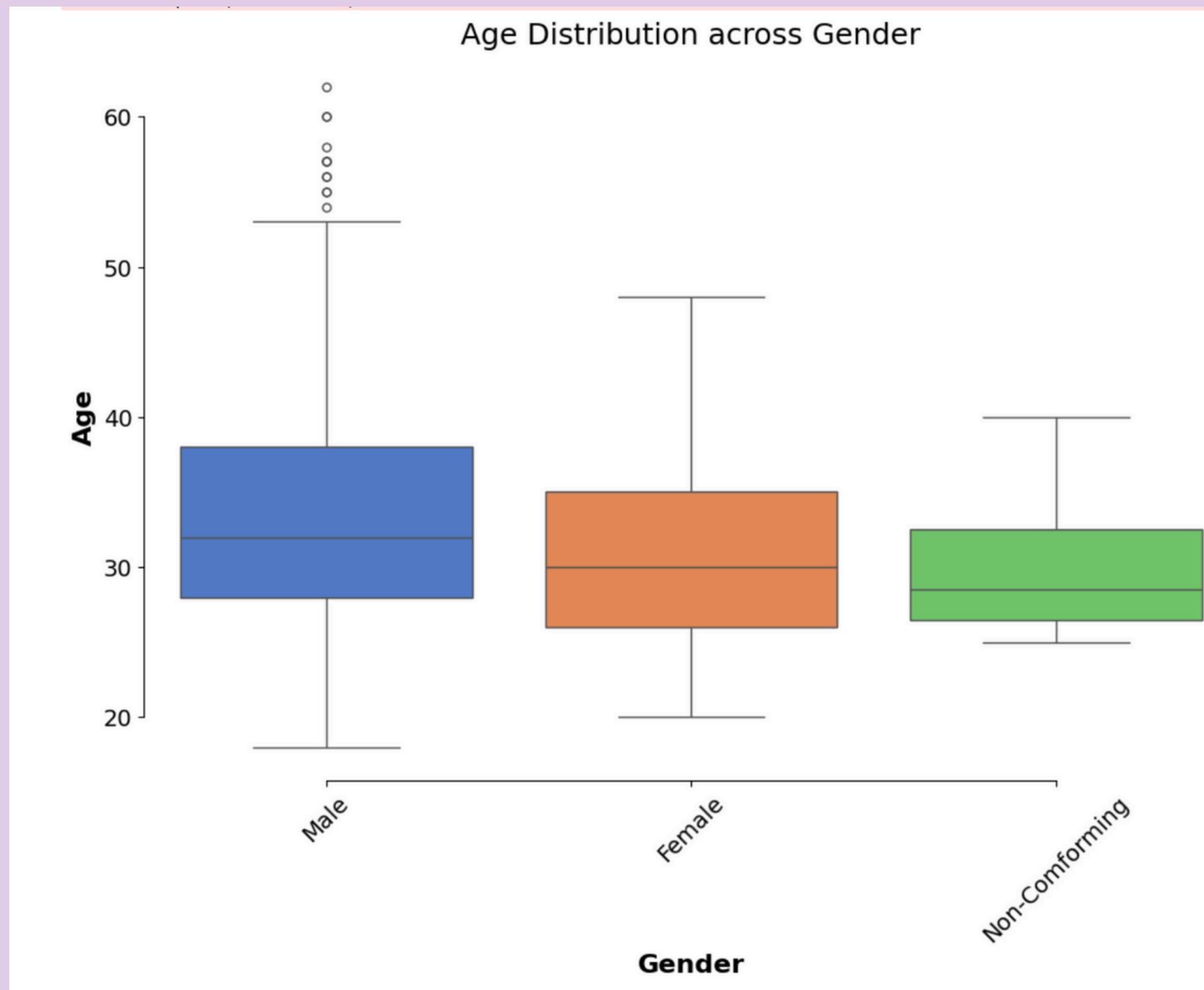


- Percentages are low due to low representation of other genders
- Women have seem to have sought treatment more
- Men seem to border on not seeking vs seeking
- Difficult to draw conclusions on non-conforming genders with such low entries



# Data Anomalies

- High outliers in age category - mainly men
  - Older men typically work more than older women
  - Older prominent men figures in industry
- Majority of variables are categorical Chi-Squared Test
- P-Values for chosen variables vs. Treatment:
  - Age: 0.7508
  - Gender: 0.4275
  - State: 0.5354
  - Remote Work: 0.2010
  - Coworkers: 0.8696
  - Supervisor: 0.9360
  - Seek Help: 0.0676



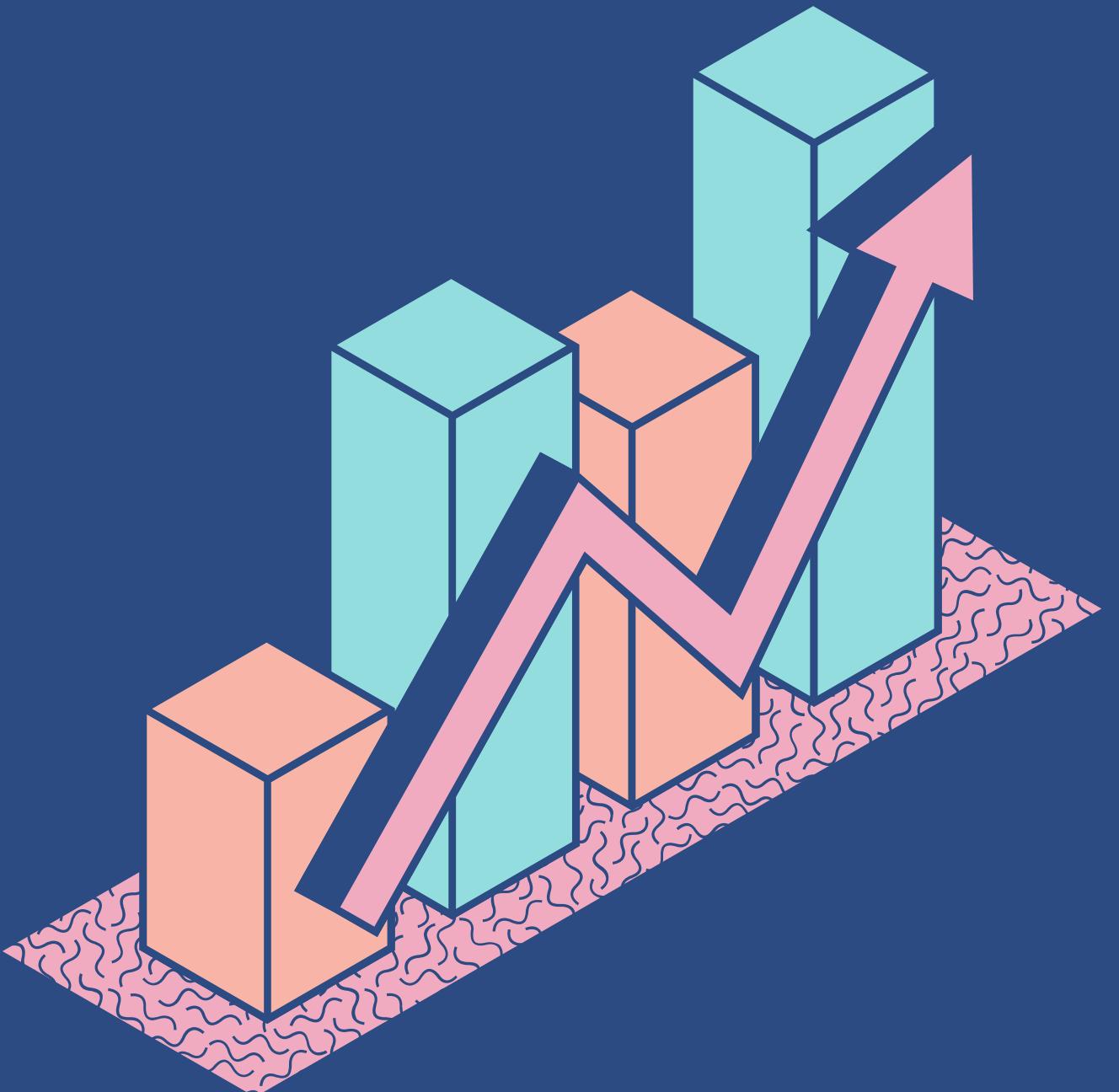
# Cramer's V

## Cramer's V for Treatment versus:

- Age: 0.2681
- Gender: 0.1601
- State: 0.3134
- Remote Work: 0.0258
- Seek Help: 0.1338
- Coworkers: 0.0639
- Supervisor: 0.1050

## What this tells us:

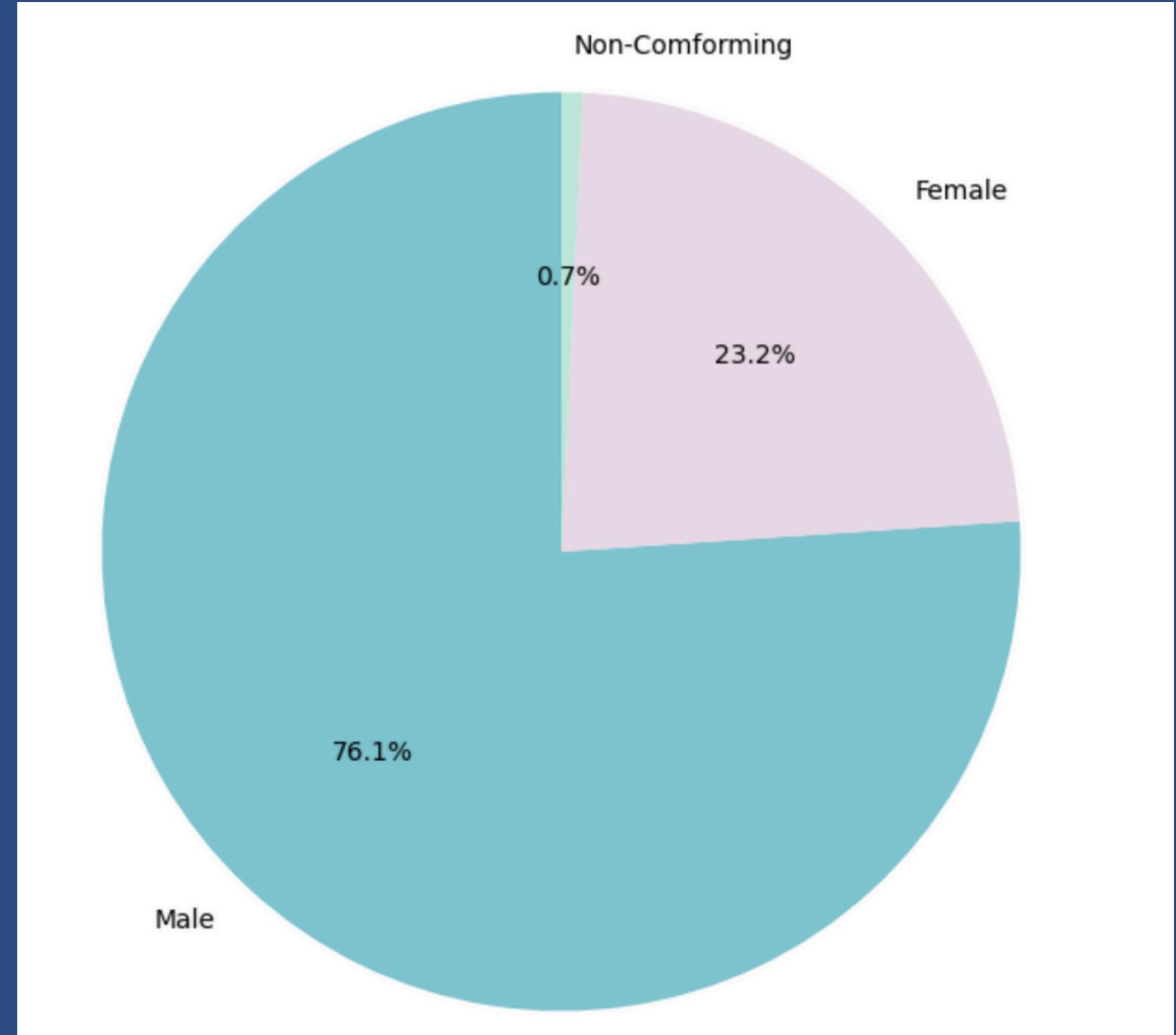
- Strength and relevance of association between variables
- Closer to zero - weaker association
- None of the variables are statistically significant



# Data Anomalies

## GENDER

- **Less significance than expected - P: 0.4275**
  - Could be due to oversaturation of men in dataset
  - There were four non-conforming entries, cannot accurately represent with low numbers
- **Potential fixes:**
  - Introduce more women and non-conforming genders into dataset to assist with imbalance



# Modelling the Data One Hot & Transforming

```
# cols needed
feats = ["Gender", "state", "remote_work", "seek_help", "coworkers", "supervisor"]

# mapping
y = survey_clean["treatment"].map({"Yes": 1, "No": 0})

# one hot encoding
one_hot = OneHotEncoder()

# fit & transform
encoded = one_hot.fit_transform(survey_clean[feats]).toarray()

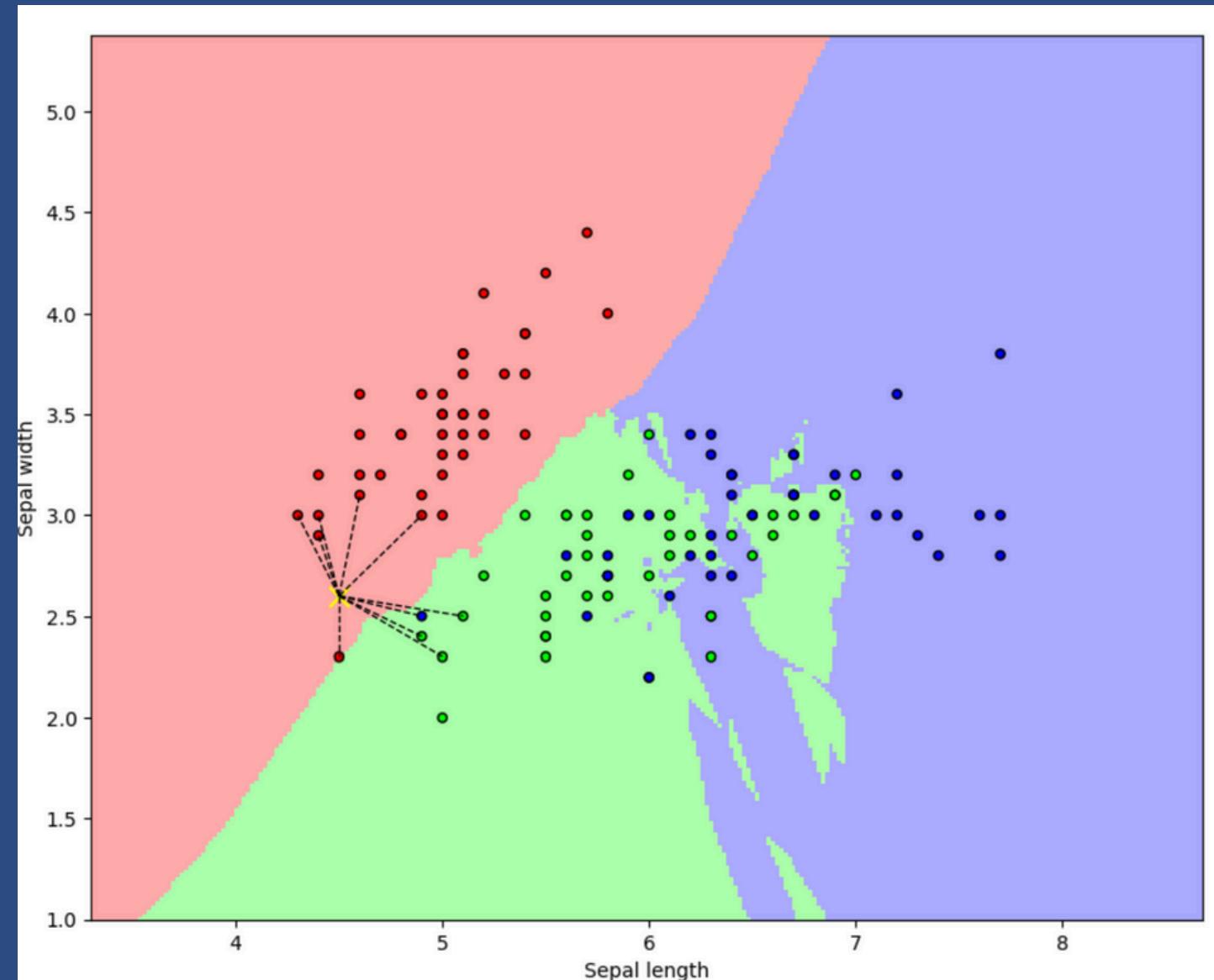
# encoded into df
en_survey = pd.DataFrame(encoded, columns=one_hot.get_feature_names_out(feats))

# reset index
survey_clean.reset_index(drop=True, inplace=True)
df.reset_index(drop=True, inplace=True)

# need to add age since it is already numeric
X = pd.concat([survey_clean[["Age"]], en_survey], axis=1)
```



```
# knn classifier  
knn = KNeighborsClassifier(n_neighbors=5)  
knn.fit(X_train, y_train)
```



Example of KNN from arise.com

# K-Nearest Neighbors

Why not other model?

- KNN is easier to understand - could not find very good resources in regards to survey data and machine learning
- Used for classification and a categorical target variable (which I have)
- KNN does not make assumptions about linearity

# Performance



## Model Accuracy: 0.55

Explanation:

- Correctly predicts 55.05% of time
- Not effective for predictions

## Confusion Matrix

		False positive
True negatives	21	23
False negatives	26	39
		True positive

Classification Report:					
	precision	recall	f1-score	support	
No	0.45	0.48	0.46	44	
Yes	0.63	0.60	0.61	65	
accuracy				0.55	109
macro avg	0.54	0.54	0.54	109	
weighted avg	0.56	0.55	0.55	109	

- Tends to predict yes more than no
- Fixes:
  - Review features used (feature engineering)
  - Changing number of n-neighbors
  - Fine tuning other parameters
- Performance is *moderate* - could try different models

# Thank you!

