DURHAM
COLLEGE
SUCCESS MATTERS

DATA ANALYSIS TOOLS ANALYTICS

(DATA 1202)

DATA REPRODUCIBILITY

ASSIGNMENT - 4

Submitted by: Kartik Sojitra

(100723768)

To Professor: Anthony Ridding

Objective:

we are going to analyze to find how does the frequency of mental health illness and attitudes towards mental health vary by geographic location? We have dataset from a 2014 to 2016 survey that contains opinions towards mental health and frequency of mental health disorders in the tech workplace.

Using the python develop script:

1) Load libraries

```
#Load libraries
import pandas as pd
import matplotlib
```

2) Import dataset

```
#Import data
raw_df = pd.read_csv("survey (1).csv")
```

3) Print dataset

```
#Print data
raw_df.head()
```

4) Build a function for group by

```
#Build a function for group by
group_by_country = raw_df.groupby(["Country", "treatment", "care_options"])["Timestamp"].count().reset_index()
print(group_by_country)
```

5) Show rows specifically for United States

```
#Show rows specifically for United states
country_df = raw_df[raw_df["Country"] == "United States"]
print(country_df)
```

6) Build function for states of United States

```
#Build function for states of United States
states_df = country_df.groupby(["state", "treatment", "care_options"])["Timestamp"].count().reset_index()
print(states_df)
```

7) Sort the states data frame by timestamp

```
#Sort the states dataframe by timestamp
timestamp_frequency_df = states_df.sort_values(["Timestamp"], ascending=False)
print(timestamp_frequency_df)
```

8) Histogram by number of timestamps

```
#Histogram by numbers of timestamp
timestamp_frequency_df.hist()
```

9) Options for the mental health care provided by an employer in a different state

```
#Options for the mental health care provided by an employer in a different state
grp0 = states_df.groupby("state")["care_options"].sum().reset_index()
grp0.sort_values("care_options", ascending=False)
print(grp0)
```

10) Retrieve all columns from the top 10 care options

```
#Retrieve all columns from the top 10 care options
grp1 = grp0.iloc[:10]
print(grp1)
```

Description of Analysis Conducted:

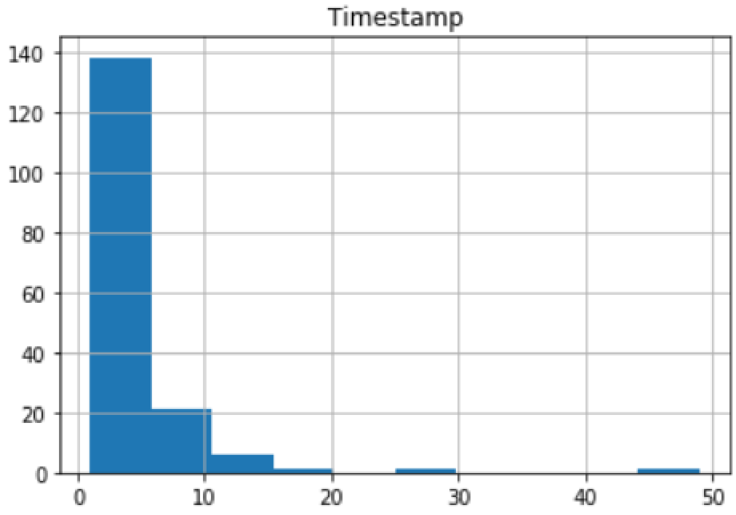
To achieve our objective, we imported libraries pandas and matplotlib in jupyter. We used the dataset **survey (1)** which contains a timestamp, age gender and different attitudes of employers towards mental health disorder.

We build a function for a group by as we need a few columns such as country, treatment, care_options and timestamp for our analysis. After that, we took specific rows of states of America and build group by function for it and sort data frame by timestamps to conduct our analysis.

We plotted histogram by time stamps means the time the survey was submitted of every state of the United States. We also run the script for the mental health care provided by an employer in a different state and retrieve all columns from the top 10 care options in the United States.

Key Findings:

- ➢ The histogram reveals the information of time stamps means the time the survey was submitted of every state of the United States.
- ➢ The highest survey was conducted in California, Washington, and New York from 2014 to 2016.
- ➢ It also describes the highest mental care options provided by employees in California, Washington, and New York.



GitHub link:

Repository name: data1202_final_assignment

Branch Name: assignment004

Commits and Description: 9

Contributor: 1

Merge: As I created only one branch, I didn’t merge any branch via pull request.

https://github.com/Kartik5696/data1202_final_assignment.git

Python Script: Please find the attachment of python script (final_assignment004) with this docx.

References:

(Anthony Ridding) data1202_data analytics tools_week 12 & 13_ tutorials