Week 4 Exercise (group)

- Elysa Alamillo
-

1. Read the data (from week3 exercise)

The dataset, same as week3, is attached in this repository.

In [118…
```
# write your code here, you can add more cells if needed

import pandas as pd
csv_file='pre-course_survey.csv'
df_csv=pd.read_csv(csv_file)
df_csv
```

Out[118]:

	Timestamp	1. On a scale of 1 to 5, how would you rate your current knowledge of Python programming?	2. On a scale of 1 to 5, how would you rate your current knowledge of data science concepts?	3. Have you ever used version control systems such as Git and GitHub?	4. Are you familiar with Jupyter Notebooks or JupyterHub?	5. On a scale of 1 to 5, how would you rate your understanding of reproducible research principles?	6. Ha you e conduc t analysis natu langua process (NL projec
0	4/1/2023 17:50:33	1	1	No	No	1.0	
1	4/1/2023 18:11:14	1	2	No	No	5.0	
2	4/1/2023 18:50:40	1	1	No	No	4.0	
3	4/1/2023 18:53:09	1	1	No	No	1.0	

4	4/1/2023 19:25:36	2	2	No	No	1.0
5	4/1/2023 19:43:48	1	3	No	No	1.0
6	4/1/2023 20:03:11	1	1	No	No	1.0
7	4/1/2023 20:42:55	1	2	No	No	1.0
8	4/1/2023 20:58:16	1	1	No	No	2.0
9	4/1/2023 21:13:10	1	2	No	No	1.0
10	4/1/2023 21:25:53	1	1	No	No	1.0
11	4/2/2023 8:58:39	1	1	No	No	1.0
12	4/2/2023 13:42:42	1	2	No	No	1.0
13	4/2/2023 14:50:47	1	3	No	No	1.0
14	4/2/2023 16:24:49	3	3	Yes	Yes	4.0
15	4/2/2023 18:17:48	1	1	No	No	1.0

16	4/3/2023 12:42:23	1	1	No	No	1.0
17	4/3/2023 12:44:21	1	1	No	No	1.0
18	4/3/2023 15:14:12	1	3	No	No	2.0
19	4/3/2023 18:04:03	1	3	No	Yes	1.0
20	4/3/2023 19:01:13	1	1	No	No	1.0
21	4/3/2023 21:09:48	1	1	No	No	1.0
22	4/3/2023 23:20:47	2	1	No	No	1.0
23	4/4/2023 10:02:25	1	1	No	No	2.0
24	4/4/2023 10:30:57	1	2	No	No	2.0
25	4/4/2023 11:58:00	1	1	No	No	3.0
26	4/4/2023 12:37:37	2	2	No	Yes	3.0

27	4/4/2023 12:46:06	1	1	No	Yes	1.0
28	4/4/2023 12:46:11	1	1	No	No	1.0
29	4/4/2023 12:47:24	3	1	No	Yes	1.0
30	4/4/2023 12:51:54	1	1	No	No	1.0
31	4/4/2023 12:54:07	2	2	No	Yes	2.0
32	4/5/2023 21:02:41	1	1	No	No	1.0
33	4/6/2023 12:40:37	1	1	No	No	1.0
34	4/9/2023 19:47:51	1	1	No	No	1.0
35	4/3/2023 19:01:13	1	1	No	No	1.0
36	4/3/2023 21:09:48	1	1	No	No	1.0
37	4/3/2023 23:20:47	2	1	No	No	NaN
38	4/4/2023 10:02:25	1	1	No	No	2.0

39	4/4/2023 10:30:57	1	2	No	No	2.0
40	4/4/2023 11:58:00	1	1	NaN	No	3.0
41	4/4/2023 12:37:37	2	2	No	Yes	3.0
42	4/4/2023 12:46:06	1	1	No	Yes	1.0
43	4/4/2023 12:46:11	1	1	No	No	1.0
44	4/4/2023 12:47:24	3	1	No	Yes	1.0
45	4/4/2023 12:51:54	1	1	No	No	1.0
46	4/4/2023 12:54:07	2	2	No	Yes	2.0
47	4/5/2023 21:02:41	1	1	NaN	No	1.0
48	4/6/2023 12:40:37	1	1	No	No	1.0
49	4/9/2023 19:47:51	1	1	No	No	1.0

2. Preprocess the data (from week3 exercise)

1. Rename all columns
2. Drop nan
3. Drop duplicates

In [119…
```
def process_data(filename):

    csv_file="pre-course_survey.csv"
    df_csv=pd.read_csv(csv_file)

    old_cols = df_csv.columns
    new_cols = ["time", "python", "concepts", "git", "jupyter", "repro", "nl
    cols_dict = dict(zip(old_cols, new_cols))
    df_csv.rename(columns=cols_dict, inplace=True)

    df_csv.dropna(inplace=True)
    df_csv[df_csv.duplicated()]
    df_csv.drop_duplicates()

    df_csv1 = df_csv.replace({"Yes":1,"No":0})

    return df_csv1
```

In [120…
```
process_data(csv_file)
```

Out[120]:

	time	python	concepts	git	jupyter	repro	nlp	social_media	goals	
0	4/1/2023 17:50:33	1	1	0	0	1.0	0	0	Python and the data science methods described ...	k
1	4/1/2023 18:11:14	1	2	0	0	5.0	0	1	To learn more about data science processes.	Coc
2	4/1/2023 18:50:40	1	1	0	0	4.0	0	0	I'd like to gain more experience in python and...	SQ
3	4/1/2023 18:53:09	1	1	0	0	1.0	0	0	My primary learning goals are to be skillful i...	pro
									Learning how	

4	4/1/2023 19:25:36	2	2	0	0	1.0	0	0	we can use data science in commun...	T
5	4/1/2023 19:43:48	1	3	0	0	1.0	0	1	Learning basic English skills	Not do
6	4/1/2023 20:03:11	1	1	0	0	1.0	0	1	I want to understand data science principles b...	Le inc
7	4/1/2023 20:42:55	1	2	0	0	1.0	1	1	As a senior, I will go to a Marketing Master p...	I hop som
8	4/1/2023 20:58:16	1	1	0	0	2.0	0	0	Understanding what data science for social stu...	cod
9	4/1/2023 21:13:10	1	2	0	0	1.0	0	0	Learning the basics to apply to real world pro...	I hop con
10	4/1/2023 21:25:53	1	1	0	0	1.0	0	0	I don’t even know. To come out of this knowing...	
11	4/2/2023 8:58:39	1	1	0	0	1.0	0	0	Learning and retaining as much information as ...	prog effe
12	4/2/2023 13:42:42	1	2	0	0	1.0	0	1	to know more	how d
13	4/2/2023 14:50:47	1	3	0	0	1.0	0	1	expand my knowledge	
14	4/2/2023 16:24:49	3	3	1	1	4.0	0	0	Gain a better understanding of python	pre ba
15	4/2/2023 18:17:48	1	1	0	0	1.0	0	0	get to know about data science	

16	4/3/2023 12:42:23	1	1	0	0	1.0	0	0	I'd like to further improve my analytical thin...	I'd l
17	4/3/2023 12:44:21	1	1	0	0	1.0	0	0	To learn basic methods of python programming w...	
18	4/3/2023 15:14:12	1	3	0	0	2.0	0	0	I hope to learn how to apply data science in a...	le
19	4/3/2023 18:04:03	1	3	0	1	1.0	0	1	To learn a little more about programming langu...	Prog qu
20	4/3/2023 19:01:13	1	1	0	0	1.0	0	1	I thought learning to produce and analyze data...	som
21	4/3/2023 21:09:48	1	1	0	0	1.0	0	0	To learn more about the subject of data scienc...	I h com
22	4/3/2023 23:20:47	2	1	0	0	1.0	0	0	Learn some basic concepts about data science	
23	4/4/2023 10:02:25	1	1	0	0	2.0	0	0	To learn new skills!	W abl
24	4/4/2023 10:30:57	1	2	0	0	2.0	0	1	Gain applicable skills that I have not been ex...	prac
25	4/4/2023 11:58:00	1	1	0	0	3.0	0	1	Im excited to gain more knowledge on data scie...	I am more
26	4/4/2023 12:37:37	2	2	0	1	3.0	0	0	understanding data science in relation to comm...	tex

27	4/4/2023 12:46:06	1	1	0	1	1.0	0	0	What data science id related to	How
28	4/4/2023 12:46:11	1	1	0	0	1.0	0	0	For future graduate programs preparation in th...	
29	4/4/2023 12:47:24	3	1	0	1	1.0	1	0	I want to acquire basic knowledge of data scie...	mo
30	4/4/2023 12:51:54	1	1	0	0	1.0	0	0	I wish to learn and get a grasp on Python's co...	I hop
32	4/5/2023 21:02:41	1	1	0	0	1.0	0	0	To learn more about data science!	
33	4/6/2023 12:40:37	1	1	0	0	1.0	0	0	Learn how to read big datam	
34	4/9/2023 19:47:51	1	1	0	0	1.0	0	0	coding, data science concepts and experience	e
35	4/3/2023 19:01:13	1	1	0	0	1.0	0	1	I thought learning to produce and analyze data...	som
36	4/3/2023 21:09:48	1	1	0	0	1.0	0	0	To learn more about the subject of data scienc...	I h com
38	4/4/2023 10:02:25	1	1	0	0	2.0	0	0	To learn new skills!	W abl
39	4/4/2023 10:30:57	1	2	0	0	2.0	0	1	Gain applicable skills that I have not been ex...	prac
									understanding	

41	4/4/2023 12:37:37	2	2	0	1	3.0	0	0	data science in relation to comm...	tex
42	4/4/2023 12:46:06	1	1	0	1	1.0	0	0	What data science id related to	How
43	4/4/2023 12:46:11	1	1	0	0	1.0	0	0	For future graduate programs preparation in th...	
45	4/4/2023 12:51:54	1	1	0	0	1.0	0	0	I wish to learn and get a grasp on Python's co...	I hop
48	4/6/2023 12:40:37	1	1	0	0	1.0	0	0	Learn how to read big datam	
49	4/9/2023 19:47:51	1	1	0	0	1.0	0	0	coding, data science concepts and experience	e

In [93]:
```
df_csv.drop_duplicates()
```

Out[93]:

	Timestamp	1. On a scale of 1 to 5, how would you rate your current knowledge of Python programming?	2. On a scale of 1 to 5, how would you rate your current knowledge of data science concepts?	3. Have you ever used version control systems such as Git and GitHub?	4. Are you familiar with Jupyter Notebooks or JupyterHub?	5. On a scale of 1 to 5, how would you rate your understanding of reproducible research principles?	6. Ha you ev conducte te analysis natur langua processi (NL project
0	4/1/2023 17:50:33	1	1	No	No	1.0	N
1	4/1/2023 18:11:14	1	2	No	No	5.0	N
2	4/1/2023 18:50:40	1	1	No	No	4.0	N

3	4/1/2023 18:53:09	1	1	No	No	1.0	N
4	4/1/2023 19:25:36	2	2	No	No	1.0	N
5	4/1/2023 19:43:48	1	3	No	No	1.0	N
6	4/1/2023 20:03:11	1	1	No	No	1.0	N
7	4/1/2023 20:42:55	1	2	No	No	1.0	Y
8	4/1/2023 20:58:16	1	1	No	No	2.0	N
9	4/1/2023 21:13:10	1	2	No	No	1.0	N
10	4/1/2023 21:25:53	1	1	No	No	1.0	N
11	4/2/2023 8:58:39	1	1	No	No	1.0	N
12	4/2/2023 13:42:42	1	2	No	No	1.0	N
13	4/2/2023 14:50:47	1	3	No	No	1.0	N
14	4/2/2023 16:24:49	3	3	Yes	Yes	4.0	N

15	4/2/2023 18:17:48	1	1	No	No	1.0	N
16	4/3/2023 12:42:23	1	1	No	No	1.0	N
17	4/3/2023 12:44:21	1	1	No	No	1.0	N
18	4/3/2023 15:14:12	1	3	No	No	2.0	N
19	4/3/2023 18:04:03	1	3	No	Yes	1.0	N
20	4/3/2023 19:01:13	1	1	No	No	1.0	N
21	4/3/2023 21:09:48	1	1	No	No	1.0	N
22	4/3/2023 23:20:47	2	1	No	No	1.0	N
23	4/4/2023 10:02:25	1	1	No	No	2.0	N
24	4/4/2023 10:30:57	1	2	No	No	2.0	N
25	4/4/2023 11:58:00	1	1	No	No	3.0	N

26	4/4/2023 12:37:37	2	2	No	Yes	3.0	N
27	4/4/2023 12:46:06	1	1	No	Yes	1.0	N
28	4/4/2023 12:46:11	1	1	No	No	1.0	N
29	4/4/2023 12:47:24	3	1	No	Yes	1.0	Y
30	4/4/2023 12:51:54	1	1	No	No	1.0	N
31	4/4/2023 12:54:07	2	2	No	Yes	2.0	N
32	4/5/2023 21:02:41	1	1	No	No	1.0	N
33	4/6/2023 12:40:37	1	1	No	No	1.0	N
34	4/9/2023 19:47:51	1	1	No	No	1.0	N
37	4/3/2023 23:20:47	2	1	No	No	NaN	N
40	4/4/2023 11:58:00	1	1	NaN	No	3.0	N
	4/4/2023						

44	12:47:24	3	1	No	Yes	1.0	Y
47	4/5/2023 21:02:41	1	1	NaN	No	1.0	N

In [121…
```
df_csv.dropna()
```

Out[121]:

	Timestamp	1. On a scale of 1 to 5, how would you rate your current knowledge of Python programming?	2. On a scale of 1 to 5, how would you rate your current knowledge of data science concepts?	3. Have you ever used version control systems such as Git and GitHub?	4. Are you familiar with Jupyter Notebooks or JupyterHub?	5. On a scale of 1 to 5, how would you rate your understanding of reproducible research principles?	6. Ha you e conduc t analysis natu langua process (NL projec
0	4/1/2023 17:50:33	1	1	No	No	1.0	
1	4/1/2023 18:11:14	1	2	No	No	5.0	
2	4/1/2023 18:50:40	1	1	No	No	4.0	
3	4/1/2023 18:53:09	1	1	No	No	1.0	
4	4/1/2023 19:25:36	2	2	No	No	1.0	
5	4/1/2023 19:43:48	1	3	No	No	1.0	
6	4/1/2023 20:03:11	1	1	No	No	1.0	

7	4/1/2023 20:42:55	1	2	No	No	1.0
8	4/1/2023 20:58:16	1	1	No	No	2.0
9	4/1/2023 21:13:10	1	2	No	No	1.0
10	4/1/2023 21:25:53	1	1	No	No	1.0
11	4/2/2023 8:58:39	1	1	No	No	1.0
12	4/2/2023 13:42:42	1	2	No	No	1.0
13	4/2/2023 14:50:47	1	3	No	No	1.0
14	4/2/2023 16:24:49	3	3	Yes	Yes	4.0
15	4/2/2023 18:17:48	1	1	No	No	1.0
16	4/3/2023 12:42:23	1	1	No	No	1.0
17	4/3/2023 12:44:21	1	1	No	No	1.0
18	4/3/2023 15:14:12	1	3	No	No	2.0

19	4/3/2023 18:04:03	1	3	No	Yes	1.0
20	4/3/2023 19:01:13	1	1	No	No	1.0
21	4/3/2023 21:09:48	1	1	No	No	1.0
22	4/3/2023 23:20:47	2	1	No	No	1.0
23	4/4/2023 10:02:25	1	1	No	No	2.0
24	4/4/2023 10:30:57	1	2	No	No	2.0
25	4/4/2023 11:58:00	1	1	No	No	3.0
26	4/4/2023 12:37:37	2	2	No	Yes	3.0
27	4/4/2023 12:46:06	1	1	No	Yes	1.0
28	4/4/2023 12:46:11	1	1	No	No	1.0
29	4/4/2023 12:47:24	3	1	No	Yes	1.0

30	4/4/2023 12:51:54	1	1	No	No	1.0
32	4/5/2023 21:02:41	1	1	No	No	1.0
33	4/6/2023 12:40:37	1	1	No	No	1.0
34	4/9/2023 19:47:51	1	1	No	No	1.0
35	4/3/2023 19:01:13	1	1	No	No	1.0
36	4/3/2023 21:09:48	1	1	No	No	1.0
38	4/4/2023 10:02:25	1	1	No	No	2.0
39	4/4/2023 10:30:57	1	2	No	No	2.0
41	4/4/2023 12:37:37	2	2	No	Yes	3.0
42	4/4/2023 12:46:06	1	1	No	Yes	1.0
43	4/4/2023 12:46:11	1	1	No	No	1.0

45	4/4/2023 12:51:54	1	1	No	No	1.0
48	4/6/2023 12:40:37	1	1	No	No	1.0
49	4/9/2023 19:47:51	1	1	No	No	1.0

3. Data Visualization

After data cleaning, use different (renamed) columns to generate:

1. Line plot
2. Scatter plot
3. Any other types of plot (e.g., you can make a subset of the dataframe by selecting some columns and create a heatmap)
4. Add titles, labels, and/or legends to your plots if needed.
5. Make your plots pretty by customize the color and style.

Use Tuesday's and Thursday's lectures to help choose, create, and customize plots

Line Plot

In [122…

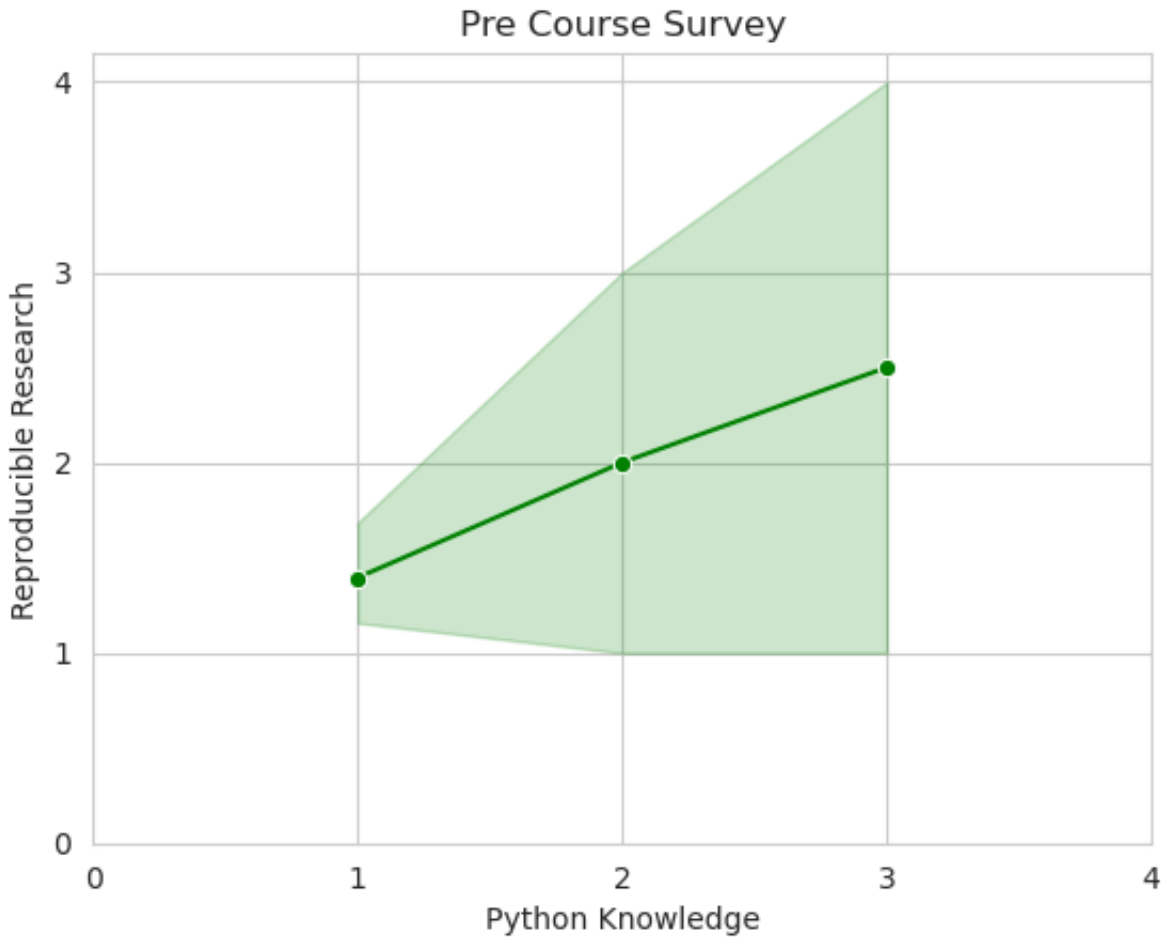
```
# write your code here, you can add more cells if needed


import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

data=process_data(csv_file)
```

In [101…

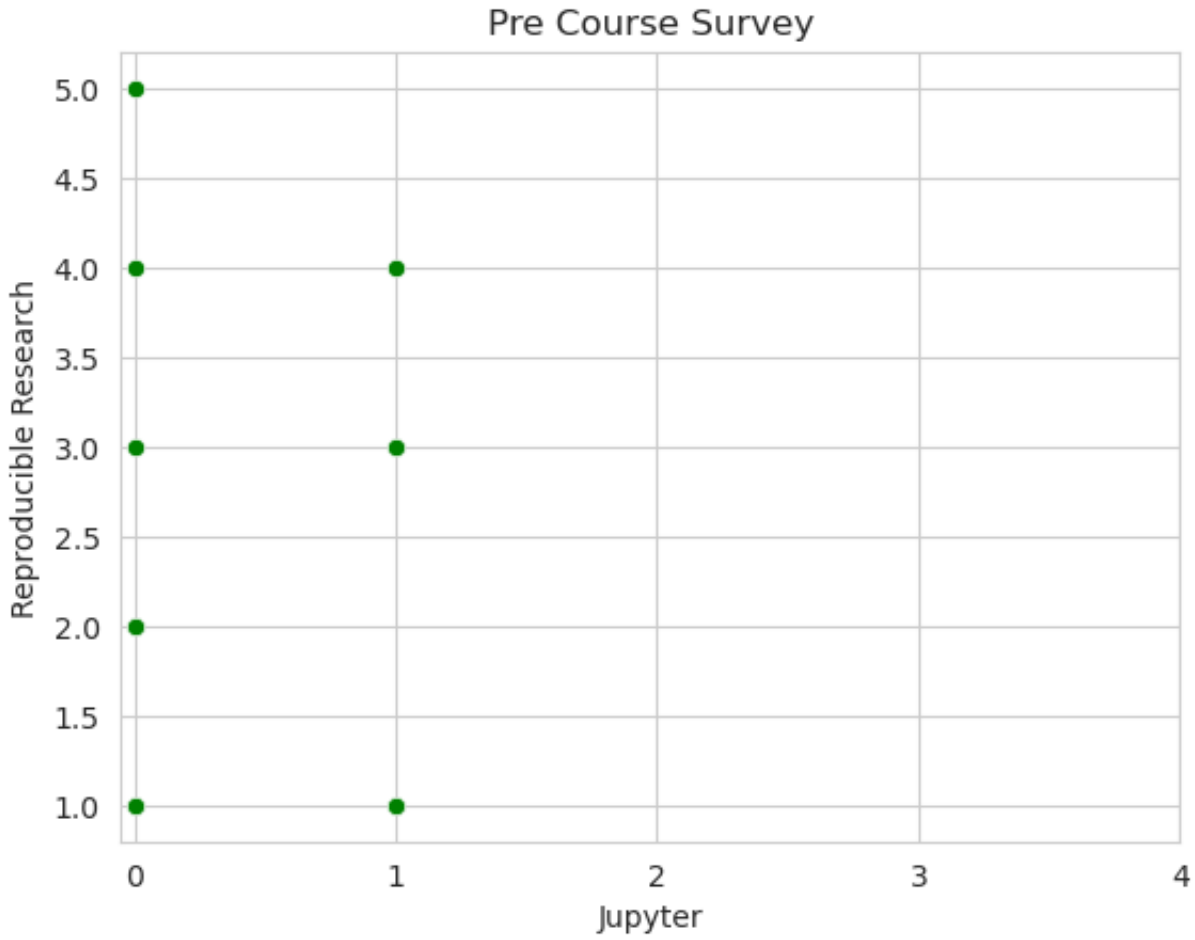
```
sns.set_style("whitegrid")
sns.set_palette("pastel")
sns.lineplot(x="python", y="repro", color='green',marker='o', data=data)
plt.xticks([0, 1, 2, 3, 4,])
plt.yticks([0, 1, 2, 3, 4,])
plt.title('Pre Course Survey')
plt.xlabel('Python Knowledge')
plt.ylabel('Reproducible Research')
plt.show()
```



Scatter Plot

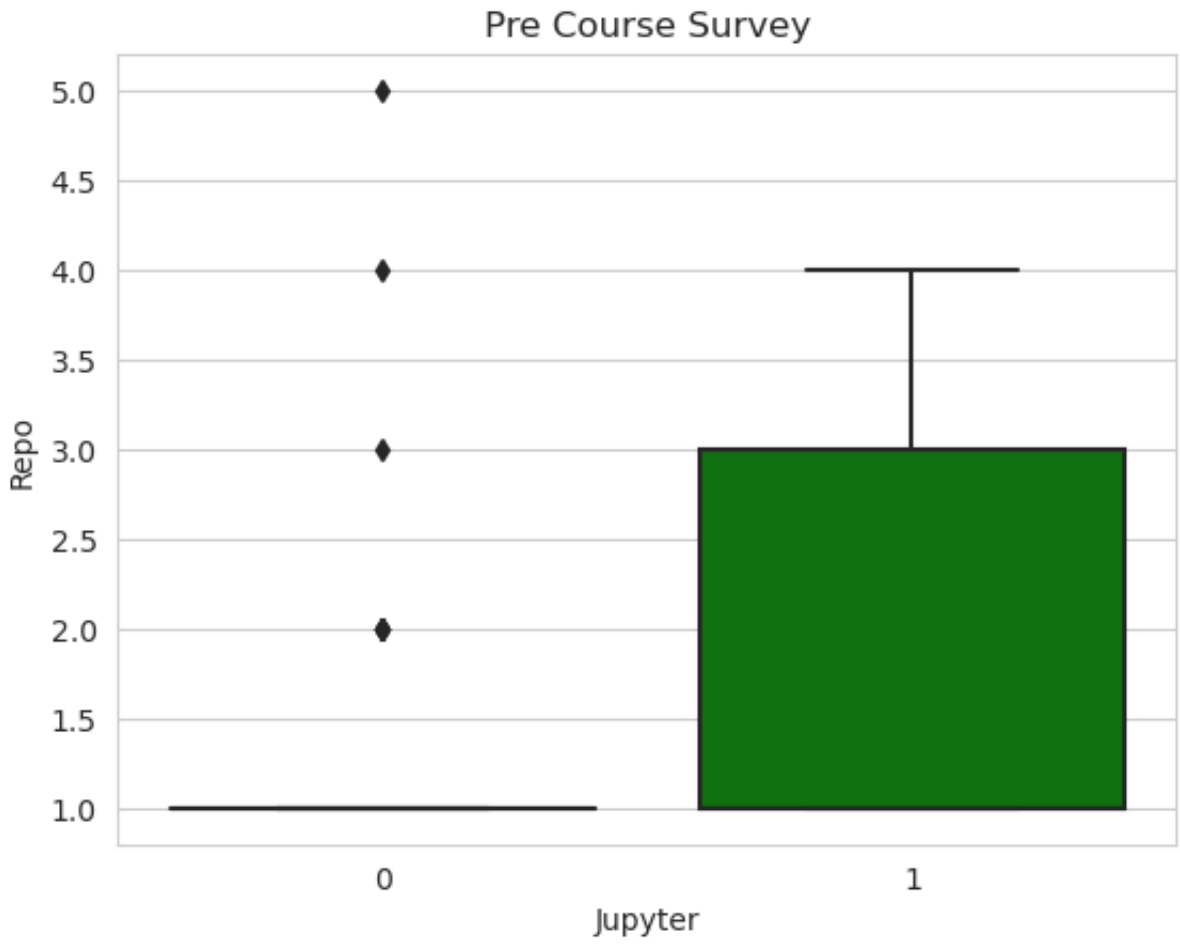
In [127…
```
# write your code here, you can add more cells if needed

data=process_data(csv_file)
sns.set_style("whitegrid")
sns.scatterplot(x="jupyter", y="repro", color= 'green', data=data)
plt.xticks([0, 1,2,3,4])
plt.title('Pre Course Survey')
plt.xlabel('Jupyter')
plt.ylabel('Reproducible Research')
plt.show()
```



Histogram

In [132…
```
data=process_data(csv_file)
sns.set_style("whitegrid")
sns.boxplot(x="jupyter", y="repro", color= 'green', data=data)
plt.title('Pre Course Survey')
plt.xlabel('Jupyter')
plt.ylabel('Repo')
plt.show()
```



4. Data Interpretation

For each plot, write 1-2 sentences to describe key findings.

Write in a Markdown cell

fun experience

5. Push Your Results to GitHub

As you did in previous weeks:

1. `git status`
2. `git add`
3. `git commit –m "type your message here"`
4. `git push`

In [138…
```
git push
```

```
  Cell In[138], line 1
    git push
        ^
SyntaxError: invalid syntax
```

In [137…
```
git status
```

```
  Cell In[137], line 1
    git status
        ^
SyntaxError: invalid syntax
```

In [ ]:
```
## 6. Peer Review (individual)

1. Go to https://github.com/orgs/ReproTeach/repositories, find other teams'
2. Open an issue on their repository (see this [link](https://docs.github.cc
    1. What you like the most about their work (can be anything such as thei
    2. Which part(s) you think they can improve (e.g., bugs in their code)

Every student should go to all other teams' repos and open an issue.

Finish your group project by the deadline (Thursday 04/27 midnight) and ther
```

Have fun coding!