# Lab5-6

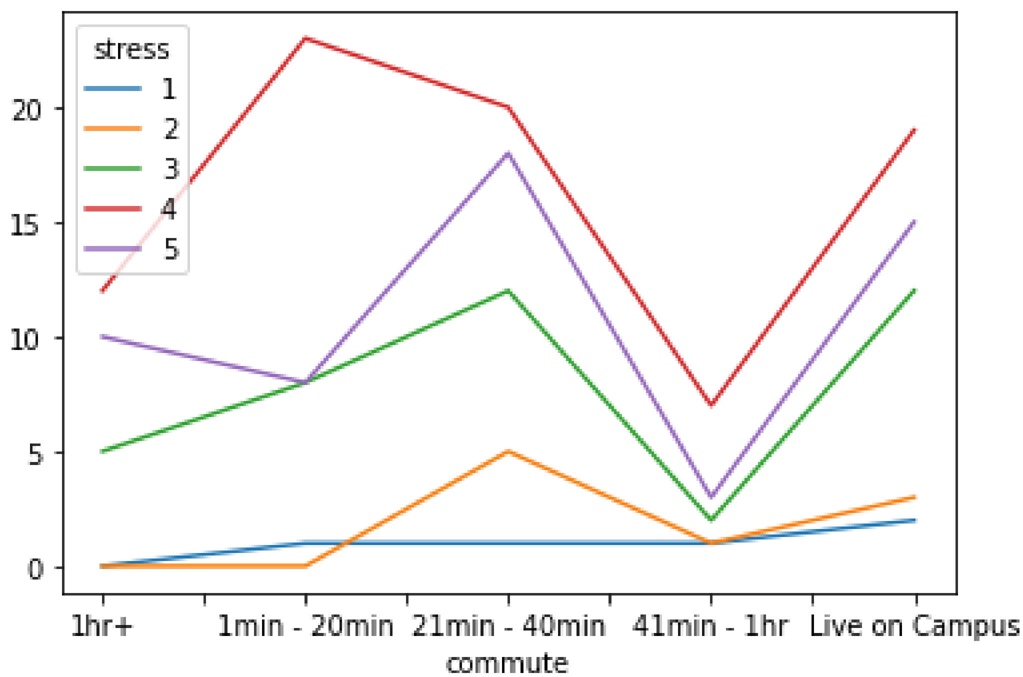
November 9, 2021

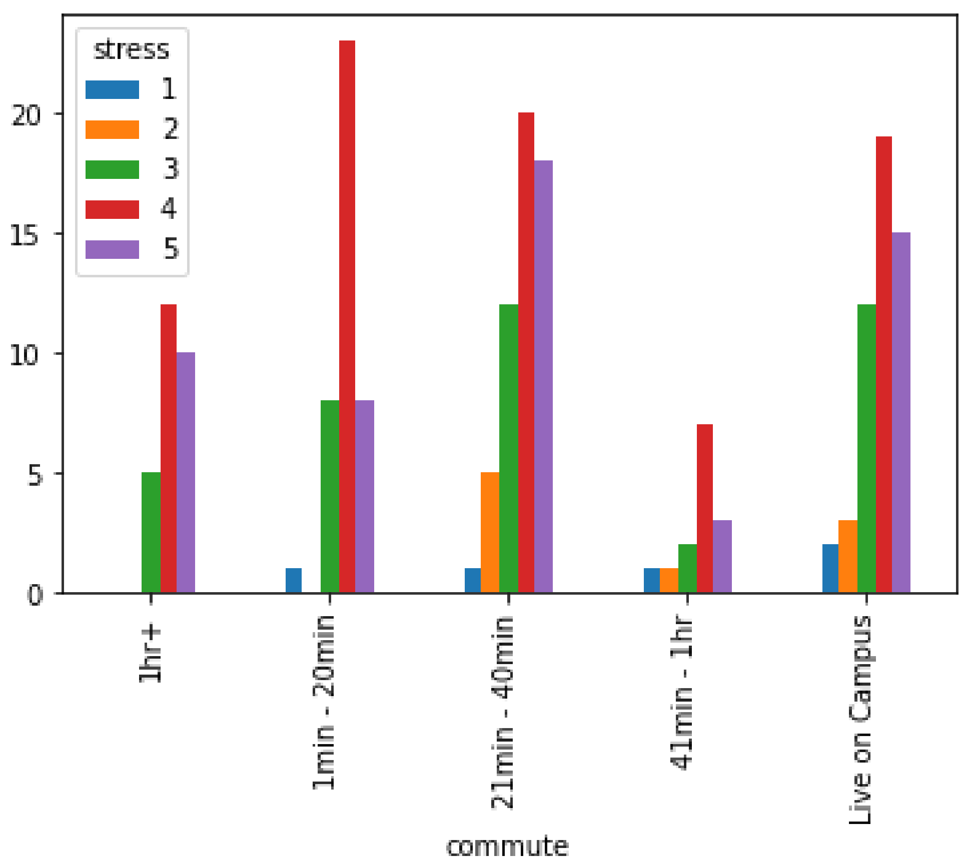
[1]:
```
#1. The data file contains information from UCR students in CS105 and CS111.
#The data used in this lab will be from the average commute time of students␣
↪and their stress levels.
#
#2. From the data, we would like to know whether there is a correlation with␣
↪the
#stress levels of students and their average roundtrip commute time. If so,
#what is that correlation? Additionally, we would also like to see if the␣
↪stress levels and commute times are independent of each other,
#does the amount of time it takes a student to commute to school affect their␣
↪stress levels?
```

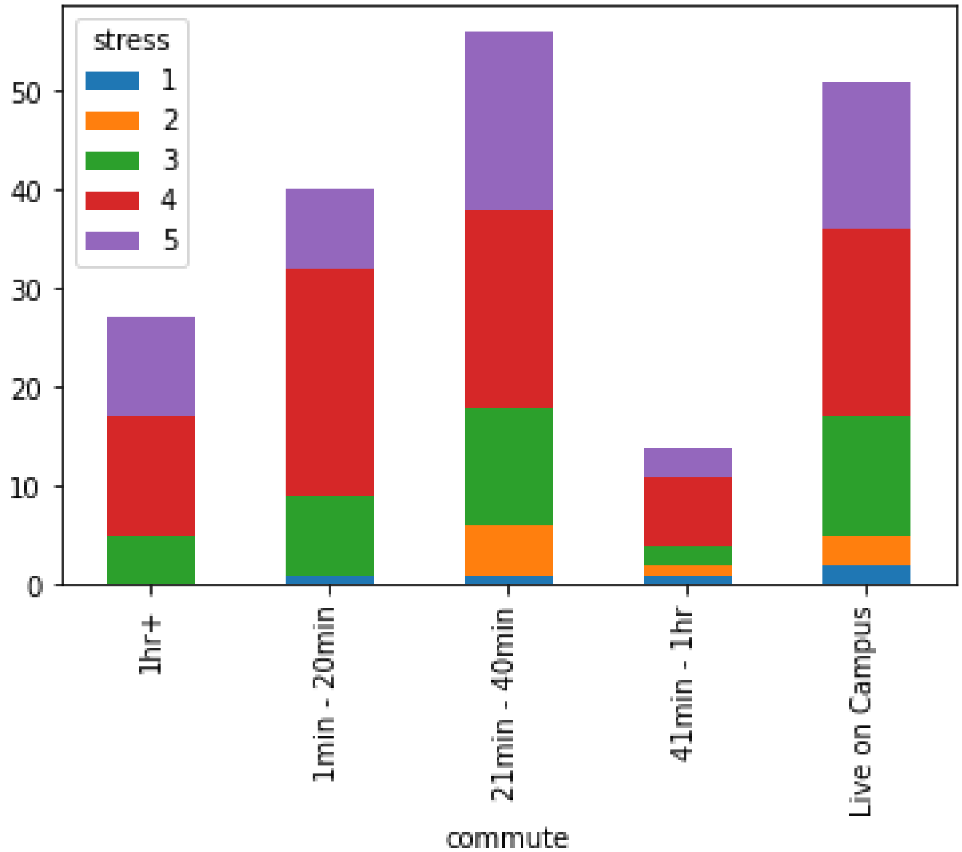
[1]:
```
#3 Compute differing data distributions into visuals
import pandas as pd
import numpy as np

df = pd.read_csv("survey.csv")
```

[3]:
```
df
df['stress']= df.filter(items = ['Rate your stress levels this quarter. 1 being␣
↪not stressed, 5 being the most stressed.'])
df['commute']= df.filter(items = ['How many minutes on average is your␣
↪roundtrip commute to school daily?.1'])
df = df.filter(items = ['commute', 'stress'])
commVstress = pd.crosstab(df.commute, df.stress)
commVstress.plot();
commVstress.plot.bar();
commVstress.plot.bar(stacked = True)
```

[3]:
```
<AxesSubplot:xlabel='commute'>
```

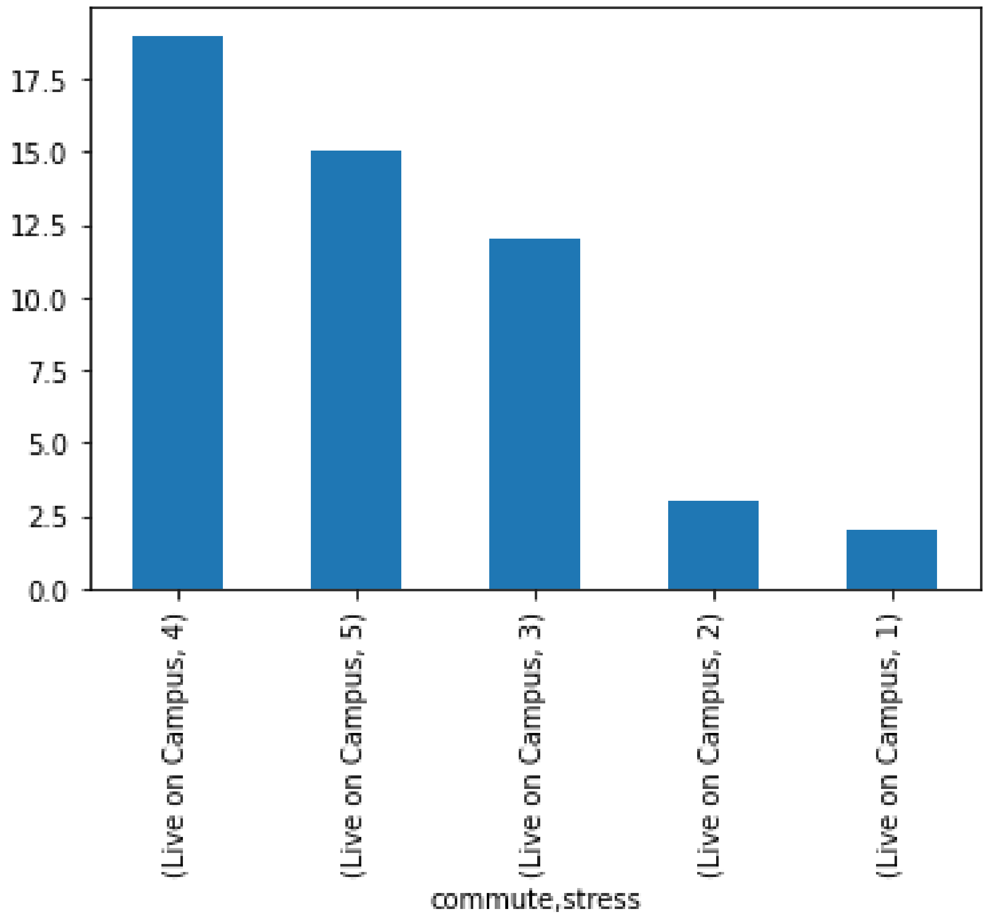

stress
1
2
3
4
5
20
15
10
5
0
1hr+
1min - 20min
21min - 40min
41min - 1hr
Live on Campus
commute


stress
1
2
3
4
5
20
15
10
5
0
1hr+
1min - 20min
21min - 40min
41min - 1hr
Live on Campus
commute



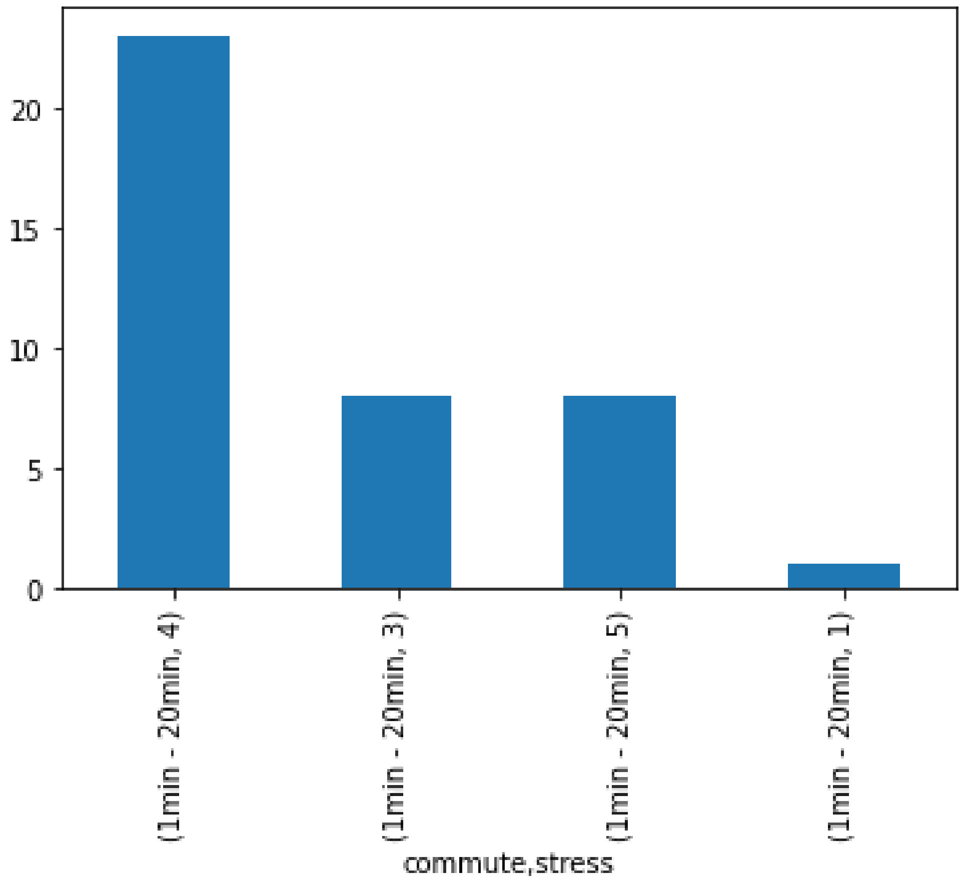

[4]:
```
df0 = df.loc[df['commute'] == 'Live on Campus']
df0.value_counts().plot(kind = 'bar')
```

[4]: <AxesSubplot:xlabel='commute,stress'>



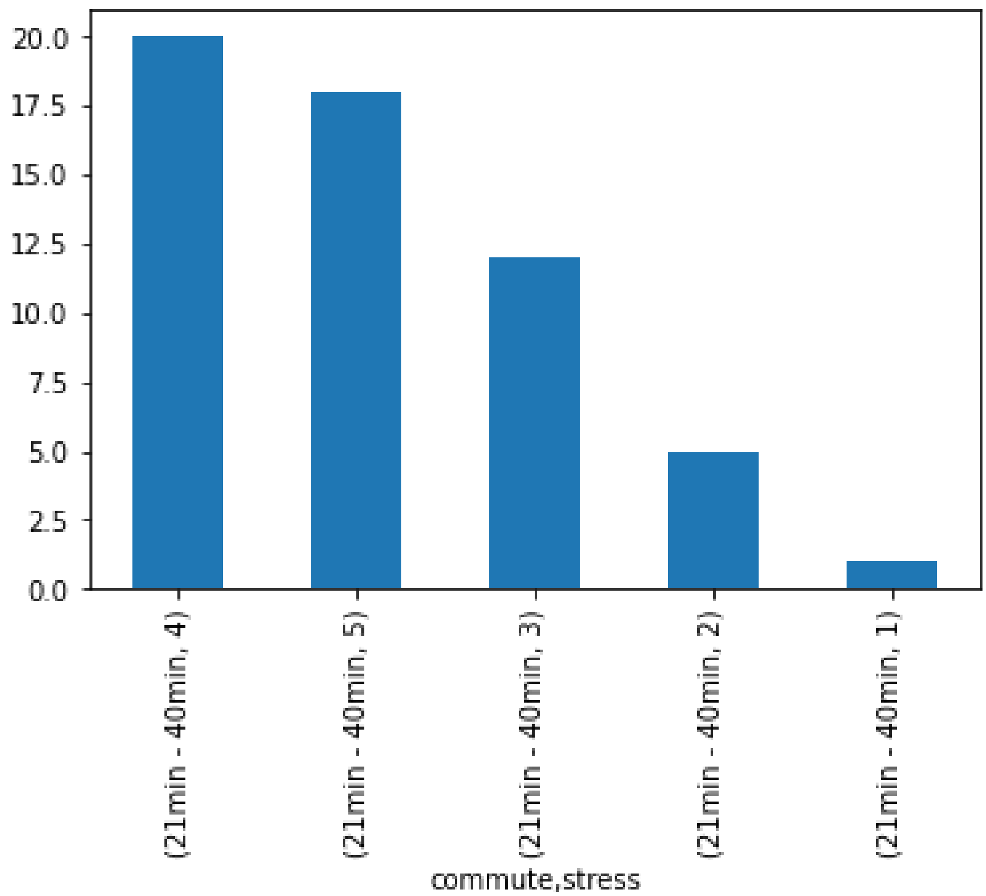

[5]:
```
df1 = df.loc[df['commute'] == '1min - 20min']
df1.value_counts().plot(kind = 'bar')
```

[5]: <AxesSubplot:xlabel='commute,stress'>



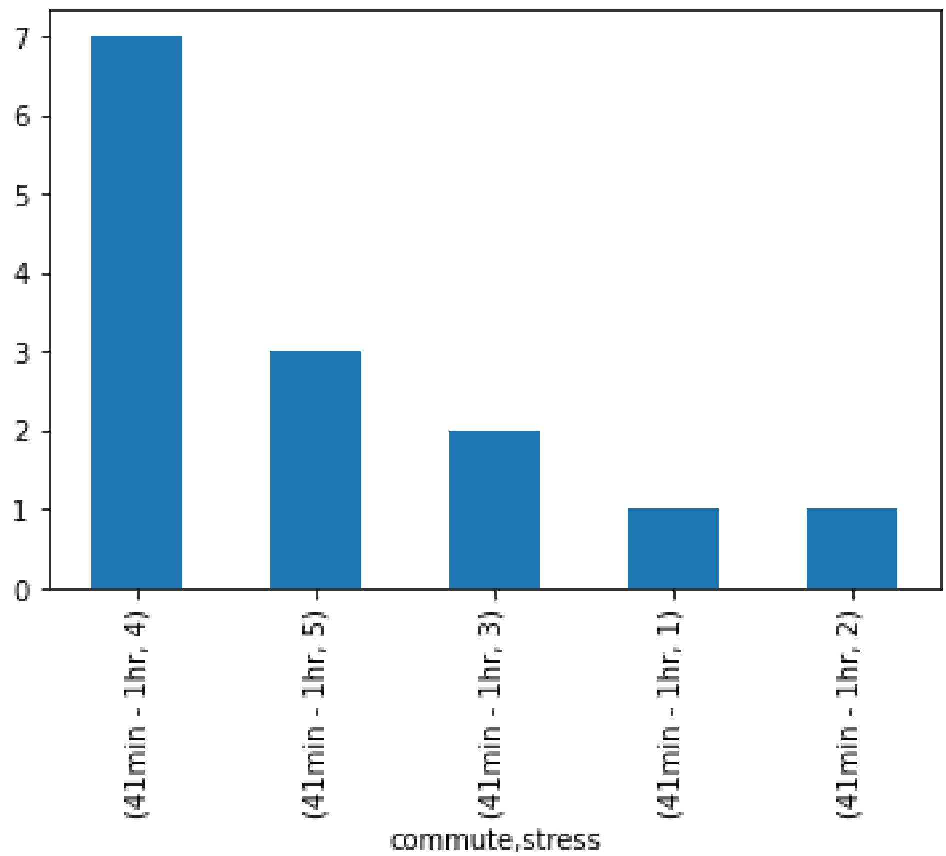

[6]:
```
df2 = df.loc[df['commute'] == '21min - 40min']
df2.value_counts().plot(kind = 'bar')
```

[6]: <AxesSubplot:xlabel='commute,stress'>



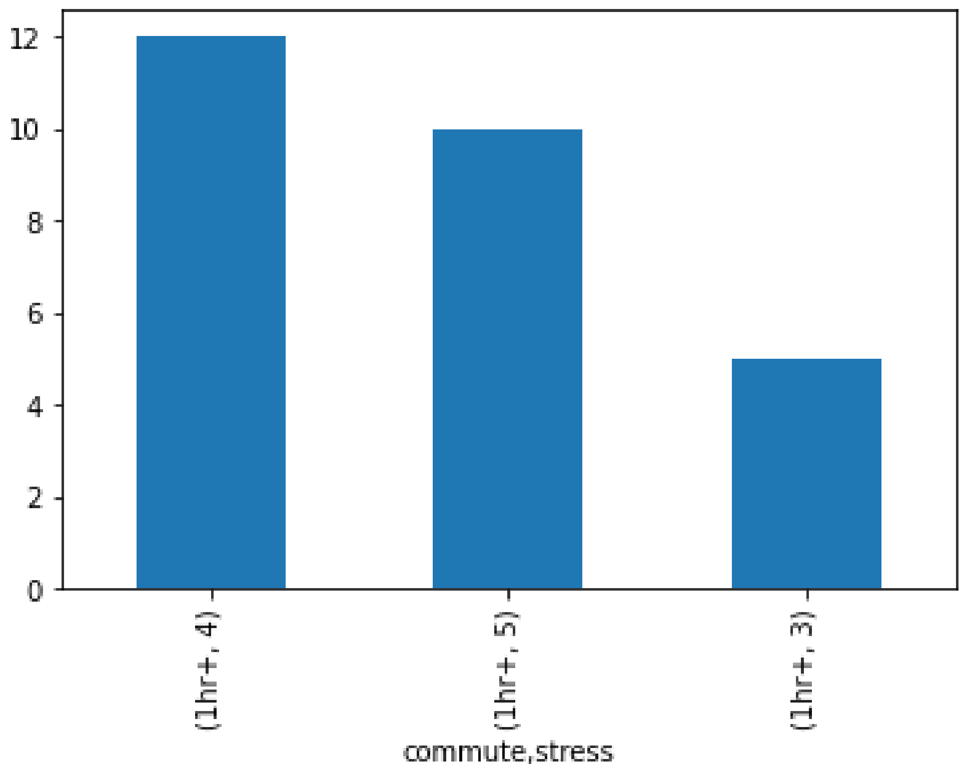

[7]:
```
df3 = df.loc[df['commute'] == '41min - 1hr']
df3.value_counts().plot(kind = 'bar')
```

[7]: <AxesSubplot:xlabel='commute,stress'>




[8]:
```
df4 = df.loc[df['commute'] == '1hr+']
df4.value_counts().plot(kind = 'bar')
```

[8]: <AxesSubplot:xlabel='commute,stress'>




[9]:
```
# we have to filter out the two rows: commute and stress levels and assign␣
↪numerical values to the commute time instead of ranges
# Chi Squared Test: We want to know the correlation between commute time and␣
↪stress level. We will be using 0.05 as our p value

#Our null hypothesis is that commute times and stress levels are dependent␣
↪(commute times influence a student's stress levels).
#Our alternative hypotheses is that commute times and stress levels are␣
↪indepedent to each other (commute times do not influence a student's stress␣
↪levels).

# sample_data = df.loc[df['commute']]
# sample_data

import numpy as np
from scipy.stats import chi2_contingency # for chi-squared test

df['commute'].replace("1hr+", 60, inplace = True)
df['commute'].replace("1min - 20min", 10, inplace = True)
df['commute'].replace("21min - 40min", 30, inplace = True)
df['commute'].replace("41min - 1hr", 50, inplace = True)
df['commute'].replace("Live on Campus", 0, inplace = True)
```

```
sample_data = df[['commute' , 'stress']]
frequency_table = pd.crosstab(sample_data['stress'], sample_data['commute'])
c, p, dof, arr = chi2_contingency(frequency_table)

print("Chi-Squared value is", c)
print("Using Chi-squared table, our critical value for dof: ", dof, " and alpha:
 ↪ 0.05 is 26.296")
print("Thus, the null hypothesis is accepted since our value is less than our␣
 ↪critical value. This means that a students stress levels and commute times␣
 ↪are independent. ")
```

```
Chi-Squared value is 13.719639250445352
Using Chi-squared table, our critical value for dof:  16  and alpha: 0.05 is
26.296
Thus, the null hypothesis is accepted since our value is less than our critical
value. This means that a students stress levels and commute times are
independent.
```

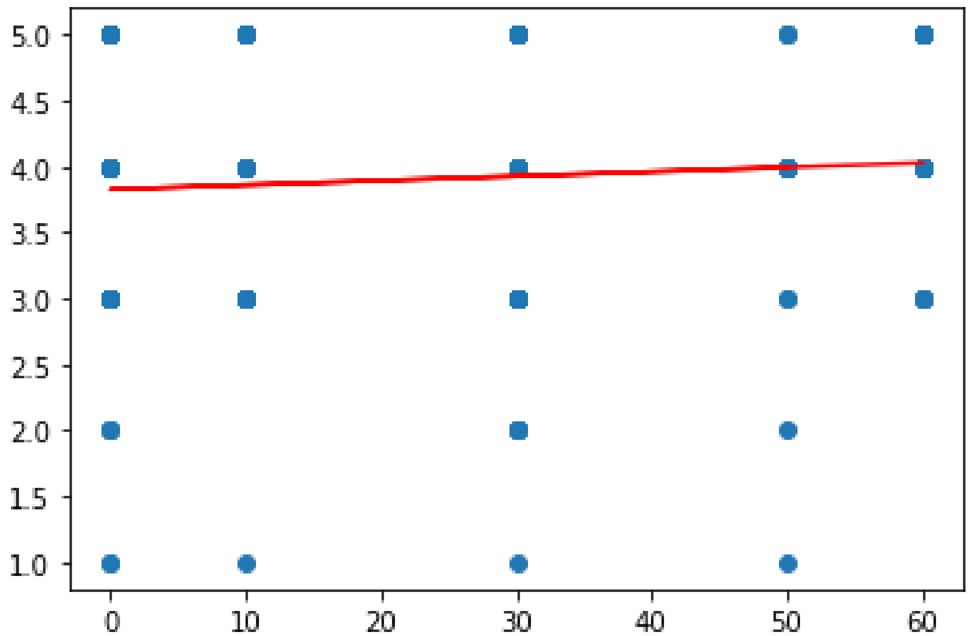
[10]:
```
#Our hypothesis is that there is a strong correlation for commute times and␣
 ↪stress levels, there is a relationship that is displayed by a students'␣
 ↪commute times and stress levels.
#Test with Pearson's correlation test

from scipy.stats import pearsonr
corr, _ = pearsonr(df['commute'], df['stress'])
df['commute']
df['stress']

print('Pearsons correlation: %.3f' % corr)
print("Our correlation is a very weak positive correlation, its low value␣
 ↪indicates that a student's values of stress have very little correlation␣
 ↪with the values of commute times.")
```

```
Pearsons correlation: 0.074
Our correlation is a very weak positive correlation, its low value indicates
that a student's values of stress have very little correlation with the values
of commute times.
```

[11]:
```
#Our hypothesis is that a student's stress levels will be positively correlated␣
 ↪with a student's commute time, as longer commute times will yield higher␣
 ↪stress levels.
#Test with linear regression line.
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression

X = df.iloc[:,0].values.reshape(-1,1)
```

```
Y = df.iloc[:,1].values.reshape(-1,1)
linear_regressor = LinearRegression()
linear_regressor.fit(X,Y)
Y_pred = linear_regressor.predict(X)
plt.scatter(X,Y)
plt.plot(X,Y_pred, color = 'red')
plt.show
print("There is a very little positive relation between commute times and␣
↪stress levels.")
```

```
There is a very little positive relation between commute times and stress
levels.
```




```
[12]: #We can see from the data that we manipulated that our survey questions could
#produce better results if we received open-ended responses for commute times,␣
↪instead of through
#an interval. For the distribution of stress levels, there also may be a skew␣
↪as students in CS111 were also going to receive a quiz after the
# announcement of needing to complete the survey.
#The location and timing of when the data was surveyed could produce biases␣
↪that could also skew our results, which can be why the data shows such low␣
↪values of correlation.
```