Team #4:

Angelica Simityan, Jinseok Lee, Ryan Giron, Steven Nguyen

In [1]:
```
%matplotlib inline
import pandas as pd
import numpy as np
```

In [2]:
```
df = pd.read_csv("Data Needed.csv")
df.head()
```

Out[2]:

	To which gender identity do you most identify?	What year are you?	Are you a transfer student?	What is your current GPA?	What was your GPA in Fall 2021?	What was your GPA in Spring 2021?	What is your current housing situation?	What was your housing situation in Fall 2021?	What is your learning style?	On average, how many hours are you on campus per day?	Vis mater (grap tab pictu e h with learni
0	Man	Fourth Year	No	3.50 to 4.00	3.50 to 4.00	3.50 to 4.00	On-campus housing (Dorms, Glen Mor, etc.)	On-campus housing (Dorms, Glen Mor, etc.)	Kinesthetic (hands-on)	24.0	
1	Man	Third Year	Yes	3.50 to 4.00	3.50 to 4.00	3.50 to 4.00	On-campus housing (Dorms, Glen Mor, etc.)	On-campus housing (Dorms, Glen Mor, etc.)	Visual	24.0	
2	Man	Fourth Year	No	3.50 to 4.00	3.00 to 3.49	3.00 to 3.49	On-campus housing (Dorms, Glen Mor, etc.)	On-campus housing (Dorms, Glen Mor, etc.)	NaN	4.0	
3	Woman	Third Year	No	3.50 to 4.00	3.50 to 4.00	3.50 to 4.00	Living at home	Living at home	Kinesthetic (hands-on)	4.0	
4	Man	Fourth Year	Yes	3.50 to 4.00	3.00 to 3.49	3.00 to 3.49	Living at home	Off-campus housing	Kinesthetic (hands-on)	0.0	

In [3]:
```
df["Overall, I prefer in-person vs. online."].replace(3,0,inplace=True)
#df["Visual materials (graphs, tables, pictures, etc.) help with my learning."].replace
#df["Hands-on activities (worksheets, in-person labs, class assignments) help with my l
df["Social interactions are better in-person vs. online."].replace([1,2,3],[-5,-4,0],in
df["Even though a hybrid option is available, I try to attend in-person classes wheneve
df["I feel more productive when attending class in-person vs. online."].replace([1,2,3]
df["I grasp course material better when attending class in-person vs. online."].replace
df
```

Out[3]:

	To which gender identity do you most identify?	What year are you?	Are you a transfer student?	What is your current GPA?	What was your GPA in Fall 2021?	What was your GPA in Spring 2021?	What is your current housing situation?	What was your housing situation in Fall 2021?	What is your learning style?	On average, how many hours are you on campus per day?	V mate (gra ta pict wit lear
0	Man	Fourth Year	No	3.50 to 4.00	3.50 to 4.00	3.50 to 4.00	On-campus housing (Dorms, Glen Mor, etc.)	On-campus housing (Dorms, Glen Mor, etc.)	Kinesthetic (hands-on)	24.0	
1	Man	Third Year	Yes	3.50 to 4.00	3.50 to 4.00	3.50 to 4.00	On-campus housing (Dorms, Glen Mor, etc.)	On-campus housing (Dorms, Glen Mor, etc.)	Visual	24.0	
2	Man	Fourth Year	No	3.50 to 4.00	3.00 to 3.49	3.00 to 3.49	On-campus housing (Dorms, Glen Mor, etc.)	On-campus housing (Dorms, Glen Mor, etc.)	NaN	4.0	
3	Woman	Third Year	No	3.50 to 4.00	3.50 to 4.00	3.50 to 4.00	Living at home	Living at home	Kinesthetic (hands-on)	4.0	
4	Man	Fourth Year	Yes	3.50 to 4.00	3.00 to 3.49	3.00 to 3.49	Living at home	Off-campus housing	Kinesthetic (hands-on)	0.0	
...	...	...	...	...	...	...	...	...	...	...	...
72	Man	Fourth Year	Yes	3.50 to 4.00	3.50 to 4.00	3.50 to 4.00	Off-campus housing	Off-campus housing	Auditory	1.0	
73	Man	Fourth Year	Yes	3.50 to 4.00	3.50 to 4.00	3.50 to 4.00	Off-campus housing	Off-campus housing	Other	2.0	

	To which gender identity do you most identify?	What year are you?	Are you a transfer student?	What is your current GPA?	What was your GPA in Fall 2021?	What was your GPA in Spring 2021?	What is your current housing situation?	What was your housing situation in Fall 2021?	What is your learning style?	On average, how many hours are you on campus per day?	V mate (gra ta pict wit lear
74	Man	Fourth Year	No	3.50 to 4.00	3.50 to 4.00	3.50 to 4.00	On-campus housing (Dorms, Glen Mor, etc.)	On-campus housing (Dorms, Glen Mor, etc.)	Visual	4.0	
75	Man	Fifth Year or higher	Yes	3.00 to 3.49	3.00 to 3.49	3.00 to 3.49	Living at home	Living at home	Kinesthetic (hands-on)	6.0	
76	Man	Fifth Year or higher	Yes	3.00 to 3.49	3.00 to 3.49	3.00 to 3.49	Living at home	Living at home	Kinesthetic (hands-on)	0.0	

77 rows × 18 columns

In [4]:
```
df.drop([57], inplace=True)
df.drop([65], inplace =True)
```

3.1 Comparing GPA of Fall 2021/Spring 2021 with Overall School Preferrence

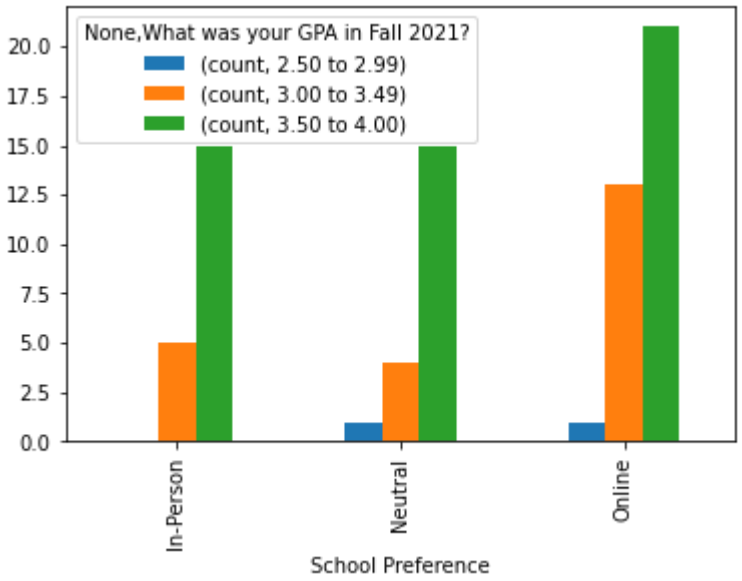
Hypothesis #3.1:

We hypothesize that students prefer online school more than in-person school because they perform better GPAs wise when learning online.

In [5]:
```
def school_preference (row):
   if row['Overall, I prefer in-person vs. online.'] == 0 :
      return 'Neutral'
   if row['Overall, I prefer in-person vs. online.'] == 5 :
      return 'Online'
   if row['Overall, I prefer in-person vs. online.'] == 4 :
      return 'Online'
   if row['Overall, I prefer in-person vs. online.'] == 1 :
      return 'In-Person'
   if row['Overall, I prefer in-person vs. online.'] == 2 :
```

```
        return 'In-Person'

df['School Preference'] = df.apply (lambda row: school_preference(row), axis=1)
df['count'] = 1
df

df_fall_2021_gpa = df.pivot_table(values=["count"], index=["School Preference"], column
df_fall_2021_gpa.fillna(1).plot.bar()
```

Out[5]: <AxesSubplot:xlabel='School Preference'>



In [6]:
```
df_fall_2021_gpa
```

Out[6]:

	count		
What was your GPA in Fall 2021?	**2.50 to 2.99**	**3.00 to 3.49**	**3.50 to 4.00**
School Preference			
In-Person	0	5	15
Neutral	1	4	15
Online	1	13	21

Result:

according to this chart and bar graph, students who prefered online classes performed better than the rest gpa wise
even though fall 2021 was mostly in-person

Test #3.1 (Pearson and Chi-Squared):

In [7]:
```
from scipy.stats import chi2_contingency
chi, p, dfree, expected = chi2_contingency(df_fall_2021_gpa)
print("chi square value: " ,chi)
```

```
print("p value: ", p)
print("degree of freedom: ", dfree)
print("when Fall 2021 GPA and overall school preference are independent: ", expected)
from scipy.stats import chi2
print("for 4 degrees of freedom, the chi-square valued needed to reject the hypothesis
chi2.ppf(0.999, 4)
```

```
chi square value:  3.013750954927425
p value:  0.5555268537649303
degree of freedom:  4
when Fall 2021 GPA and overall school preference are independent:  [[ 0.53333333  5.8666
6667 13.6       ]
 [ 0.53333333  5.86666667 13.6       ]
 [ 0.93333333 10.26666667 23.8       ]]
for 4 degrees of freedom, the chi-square valued needed to reject the hypothesis at the
0.001 significance level:
18.46682695290317
```

Out[7]:

null hypothesis: (3 attributes) 2.5 to 2.99, 3.0 to 3.49, and 3.5 to 4.0 are indepedent alternative hypothesis: (3 attributes) 2.5 to 2.99, 3.0 to 3.49, and 3.5 to 4.0 are correlated we fail to reject our null hypothesis because the chi-square value is 3.013 which is smaller than the value we need of 18.466

In [8]:

```
r1 = df_fall_2021_gpa.corr(method='pearson')
r1
```

Out[8]:

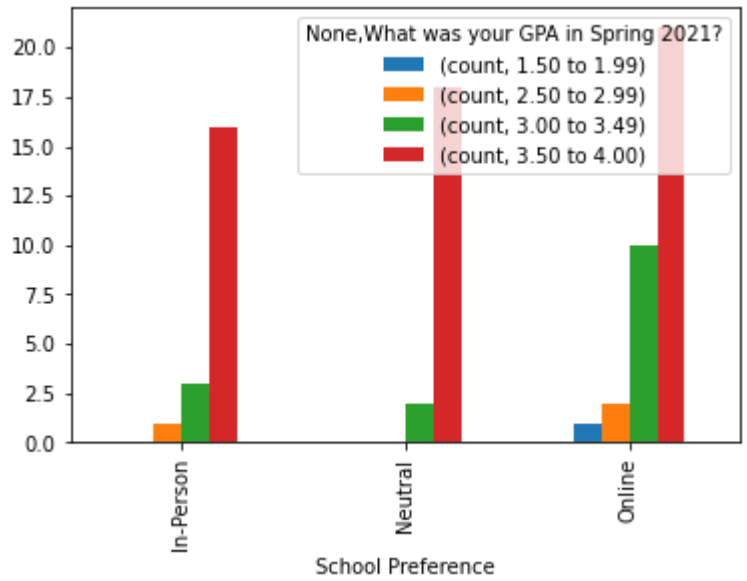
		count		
	What was your GPA in Fall 2021?	**2.50 to 2.99**	**3.00 to 3.49**	**3.50 to 4.00**
	What was your GPA in Fall 2021?			
count	**2.50 to 2.99**	1.000000	0.409644	0.50000
	3.00 to 3.49	0.409644	1.000000	0.99485
	3.50 to 4.00	0.500000	0.994850	1.00000

In [9]:

```
df_spring_2021_gpa = df.pivot_table(values=["count"], index=["School Preference"], colu
df_spring_2021_gpa.fillna(1).plot.bar()
```

Out[9]:

```
<AxesSubplot:xlabel='School Preference'>
```



In [10]:
```
df_spring_2021_gpa
```

Out[10]:

	count			
What was your GPA in Spring 2021?	**1.50 to 1.99**	**2.50 to 2.99**	**3.00 to 3.49**	**3.50 to 4.00**
School Preference				
In-Person	0	1	3	16
Neutral	0	0	2	18
Online	1	2	10	21

Result:

according to this chart and bar graph, students who prefered online classes performed better than the rest gpa wise
spring 2021 was fully online which matches the info that we found
taking both quarters into consideration, it seems that students who prefer online classes do better regardless of school being taught online or in-person
this supports our hypothesis that students prefer online school due to better academic performance

Test #3.1 (Pearson and Chi-Squared):

In [11]:
```
from scipy.stats import chi2_contingency
chi, p, dfree, expected = chi2_contingency(df_spring_2021_gpa)
print("chi square value: " ,chi)
print("p value: ", p)
print("degree of freedom: ", dfree)
print("when Spring 2021 GPA and overall school preference are independent: ", expected)
from scipy.stats import chi2
print("for 6 degrees of freedom, the chi-square valued needed to reject the hypothesis
chi2.ppf(0.999, 6)
```

```
chi square value:  6.497754010695187
p value:  0.3697966807300705
degree of freedom:  6
when Spring 2021 GPA and overall school preference are independent:  [[ 0.27027027  0.81
081081  4.05405405 14.86486486]
 [ 0.27027027  0.81081081  4.05405405 14.86486486]
 [ 0.45945946  1.37837838  6.89189189 25.27027027]]
for 6 degrees of freedom, the chi-square valued needed to reject the hypothesis at the
0.001 significance level:
22.457744484825323
```

Out[11]:

null hypothesis: (4 attributes) 1.5 to 1.99, 2.5 to 2.99, 3.0 to 3.49, and 3.5 to 4.0 are indepedent alternative hypothesis: (3 attributes) 1.5 to 1.99, 2.5 to 2.99, 3.0 to 3.49, and 3.5 to 4.0 are correlated we fail to reject our null hypothesis because the chi-square value is 3.013 which is smaller than the value we need of 22.457

In [12]:
```
r1 = df_spring_2021_gpa.corr(method='pearson')
r1
```

Out[12]:

		count			
	What was your GPA in Spring 2021?	**1.50 to 1.99**	**2.50 to 2.99**	**3.00 to 3.49**	**3.50 to 4.00**
	What was your GPA in Spring 2021?				
count	**1.50 to 1.99**	1.000000	0.866025	0.993399	0.917663
	2.50 to 2.99	0.866025	1.000000	0.917663	0.596040
	3.00 to 3.49	0.993399	0.917663	1.000000	0.866025
	3.50 to 4.00	0.917663	0.596040	0.866025	1.000000

3.2 Comparing Learning Style Preferrence and Overall School Preferrence

Hypothesis #3.2:

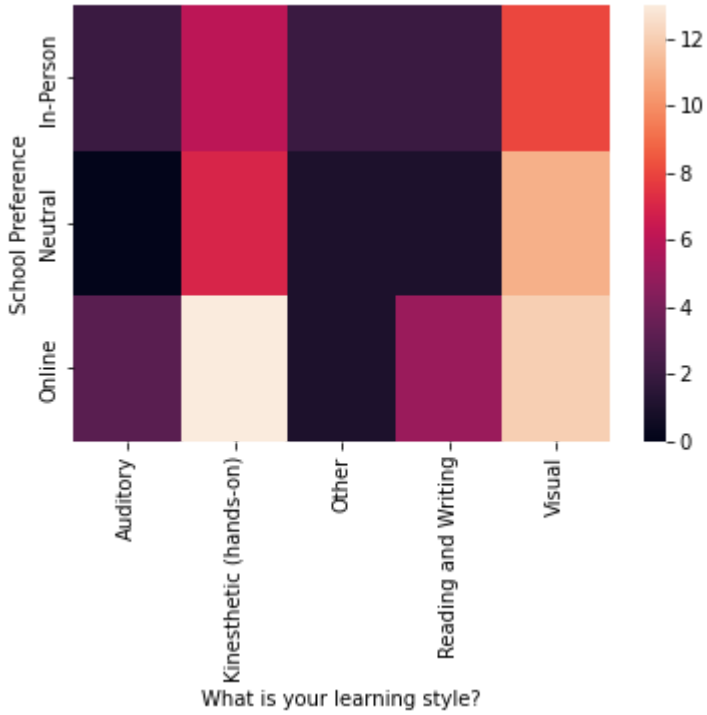
We hypothesize that kinesthetic learners will prefer in-person school due to more hands-on activities and learners who prefer reading/visual would like online school more due to more visuals and the ability to google search and read easily.

In [13]:
```
import seaborn as sns
schoolPref_vs_learningStyle = pd.crosstab(df["School Preference"], df["What is your lea
sns.heatmap(schoolPref_vs_learningStyle)
schoolPref_vs_learningStyle

# according to this heat map and table, it is suprising to see that most of the kinesth
# we can also see that many visual learners prefer online classes which makes sense bec
# Reading and writing is also interesting because most of those learners prefer online
# This visual and information backs up our original hypothesis of kids liking/performin
```

Out[13]:

What is your learning style?	**Auditory**	**Kinesthetic (hands-on)**	**Other**	**Reading and Writing**	**Visual**
School Preference					

What is your learning style?	Auditory	Kinesthetic (hands-on)	Other	Reading and Writing	Visual
School Preference					
In-Person	2	6	2	2	8
Neutral	0	7	1	1	11
Online	3	13	1	5	12



Result:

According to this heat map and table, it is suprising to see that most of the kinesthetic learners prefer online classes instead of in-person classes
We can also see that many visual learners prefer online classes which makes sense because teachers often use slides and other online demonstrations when teaching through Zoom
Reading and writing is also interesting because most of those learners prefer online classes possibly due to the ability to quickly Google search questions and read about them
This visual and information backs up our original hypothesis of kids liking/performing better in school because reading and visual learning is better online

Test #3.2 (Pearson and Chi-Squared):

In [14]:
```
from scipy.stats import chi2_contingency
chi, p, dfree, expected = chi2_contingency(schoolPref_vs_learningStyle)
print("chi square value: " ,chi)
print("p value: ", p)
print("degree of freedom: ", dfree)
print("when learning style and overall school preference are independent: ", expected)
from scipy.stats import chi2
```

```
print("for 8 degrees of freedom, the chi-square valued needed to reject the hypothesis
chi2.ppf(0.999, 8)
```

```
chi square value:  5.594650051087432
p value:  0.6925325138128452
degree of freedom:  8
when learning style and overall school preference are independent:  [[ 1.35135135  7.027
02703  1.08108108  2.16216216  8.37837838]
 [ 1.35135135  7.02702703  1.08108108  2.16216216  8.37837838]
 [ 2.2972973  11.94594595  1.83783784  3.67567568 14.24324324]]
for 8 degrees of freedom, the chi-square valued needed to reject the hypothesis at the
0.001 significance level:
26.12448155837614
```

Out[14]:

null hypothesis: (5 attributes) auditory, kinesthetic, other, reading and writing, and visual are indepedent alternative hypothesis: (5 attributes) auditory, kinesthetic, other, reading and writing, and visual are correlated we fail to reject our null hypothesis because the chi-square value is 5.594 which is smaller than the value we need of 26.124

In [15]:

```
r1 = schoolPref_vs_learningStyle.corr(method='pearson')
r1
```

Out[15]:

What is your learning style?	Auditory	Kinesthetic (hands-on)	Other	Reading and Writing	Visual
What is your learning style?					
Auditory	1.000000	0.662849	0.188982	0.891042	0.052414
Kinesthetic (hands-on)	0.662849	1.000000	-0.609994	0.930501	0.782467
Other	0.188982	-0.609994	1.000000	-0.277350	-0.970725
Reading and Writing	0.891042	0.930501	-0.277350	1.000000	0.500000
Visual	0.052414	0.782467	-0.970725	0.500000	1.000000

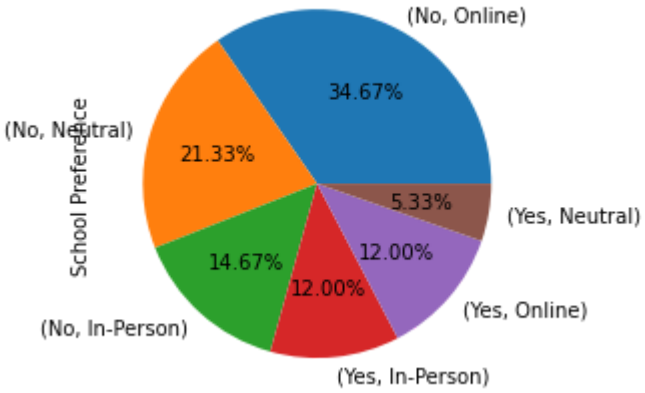
3.3 Relationship between transfer/non transfer student and preference of school method

Hypothesis #3.3:

Students' preferences of in_person and online methods have a correlation with whether students are transfer students or not. We assumed that transfer sutdent would like to prefer in_person and Non-transfer student would like to prepfer online method

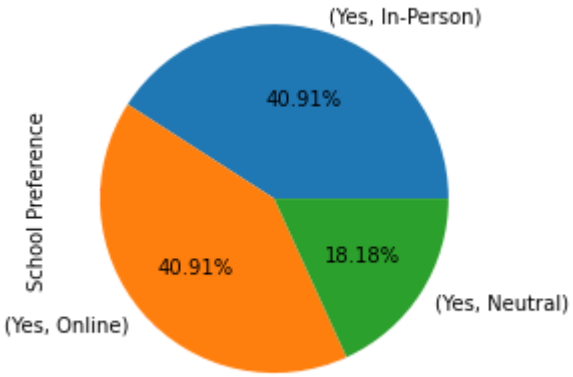
In [16]:

```
# Relationship between transfer/non transfer student and preference of school method.
#caste two columns into string
df["Are you a transfer student?"] = df["Are you a transfer student?"].astype(str)
df["School Preference"] = df["School Preference"].astype(str)
df.dtypes

# Using pie chart and group by function to create visualization between two columns.
Total = df.groupby("Are you a transfer student?")["School Preference"].value_counts().p
```



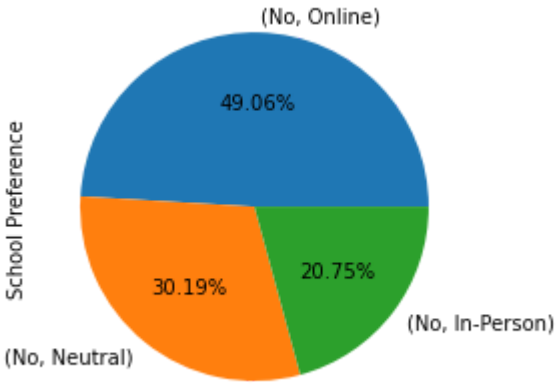
In [17]:

```
#Preference of transfer student.
#Create new data frame for transfer student only
df2 = df[(df["Are you a transfer student?"]=='Yes')]
#Using pie chart and group by function to create visualization of transfer student's pr
Transfer = df2.groupby("Are you a transfer student?")["School Preference"].value_counts
```



In [18]:

```
#Preference of Non transfer student.
#Create new data frame for Non_transfer student only
df3 = df[(df["Are you a transfer student?"]=='No')]
##Using pie chart and group by function to create visualization of Non_transfer student
Nontransfer = df3.groupby("Are you a transfer student?")["School Preference"].value_cou
```



Test: Correlation analysis by pearson

Result:

The pearson correlation hypothesis test result shows that there is no correlation between Overall school preference and whether students are transfer student or not. Most of students prefer to have online class rather than in-person class. Unlike our prediction, the transfer student have the same preference on online and in-person method.

Test #3.3 (Pearson):

In [19]:
```
#Testing Hypotheses
df["Are you a transfer student?"].replace(['Yes','No'],[0,1],inplace=True)

# Create a dataframe for transfer and overall preference only
data = df[["Are you a transfer student?","Overall, I prefer in-person vs. online."]].co
print(data)

#Using pearson correlation test to find out the correlation.
r = data.corr(method='pearson')
r
```

```
    Are you a transfer student?  Overall, I prefer in-person vs. online.
0                             1                                     4.0
1                             0                                     1.0
2                             1                                     5.0
3                             1                                     0.0
4                             0                                     1.0
..                          ...                                     ...
72                            0                                     2.0
73                            0                                     0.0
74                            1                                     4.0
75                            0                                     4.0
76                            0                                     0.0

[75 rows x 2 columns]
```

Out[19]:

	Are you a transfer student?	Overall, I prefer in-person vs. online.
Are you a transfer student?	1.000000	0.033153
Overall, I prefer in-person vs. online.	0.033153	1.000000

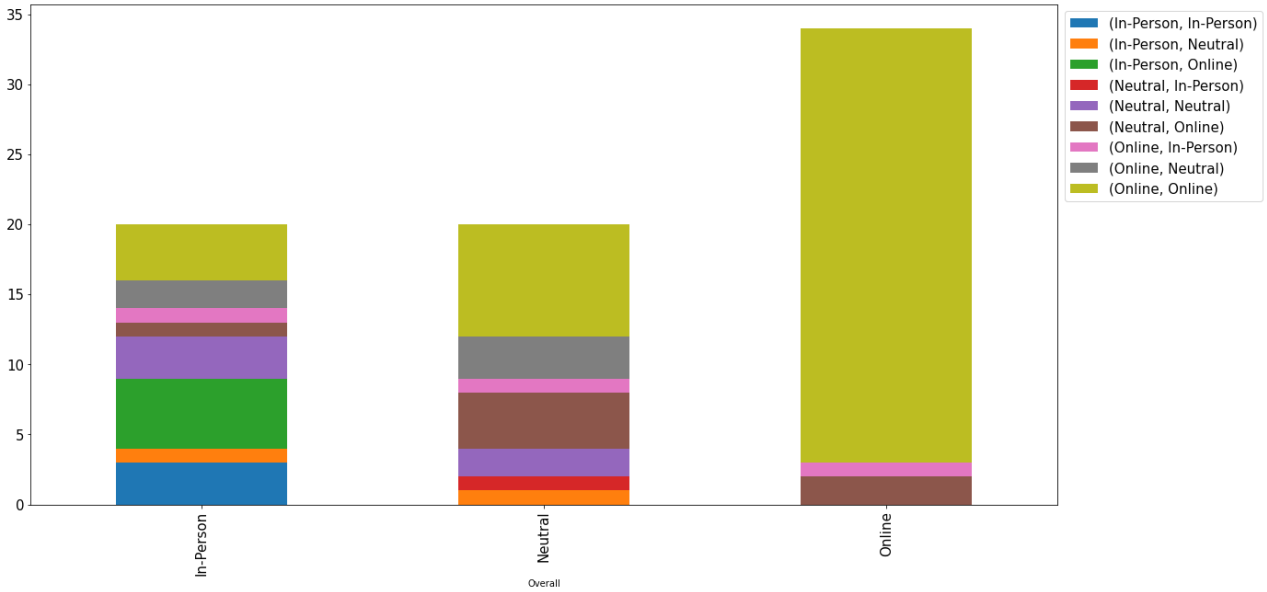
3.4 Comparsion between Productive, Social interaction, and Overall preference

Hypothesis #3.4:

If students prefer online classes, then they will also prefer online method of social interaction and will be more productive on online classes. If student prefer in-person classes, then they will also prefer in-person method of social interaction and will be more productive on in-person classes.

In [20]:
```
# Comparsion between Productive, Social interaction, and Overall preference
#Replacing scales into string
import matplotlib.pyplot as plt

df["Productive"] = df["I feel more productive when attending class in-person vs. online
df["Overall"] = df["Overall, I prefer in-person vs. online."].replace([1,2,0,4,5],['In-
df["Social"] = df["Social interactions are better in-person vs. online."].replace([-5,-

#Shorten the name of columns
Overall= df["Overall"]
Productive = df["Productive"]
Social = df["Social"]

#Create crosstab for three columns
OSP=pd.crosstab(Overall,[Productive,Social])

#Using plot bar to create visualization of three colums relationship
OSP.plot.bar(stacked=True,figsize=(20,10))
plt.legend(bbox_to_anchor=(1.0, 1.0),prop={'size': 15})
plt.xticks(fontsize=15)
plt.yticks(fontsize=15)
```

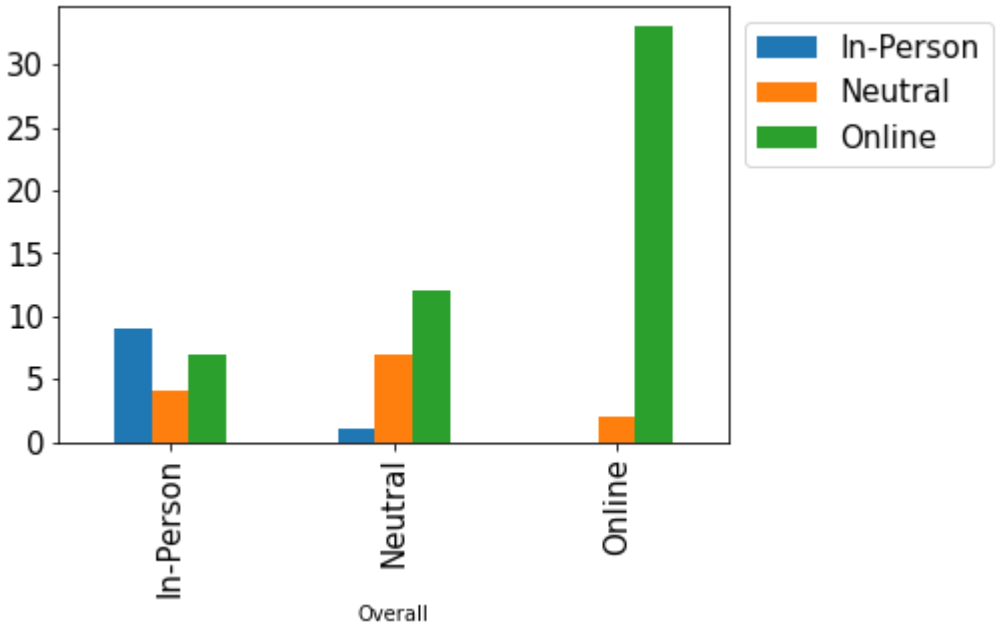
Out[20]:
```
(array([ 0.,  5., 10., 15., 20., 25., 30., 35., 40.]),
 [Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, '')])
```



In [21]:
```
# Relationship between Overall preference and Productive

#Create crosstab for Overall and Prductive columns
OP=pd.crosstab(Overall,Productive)

#Using plot bar to create visualization of two colums' relationship
OP.plot.bar()
plt.legend(bbox_to_anchor=(1.0, 1.0),prop={'size': 15})
```

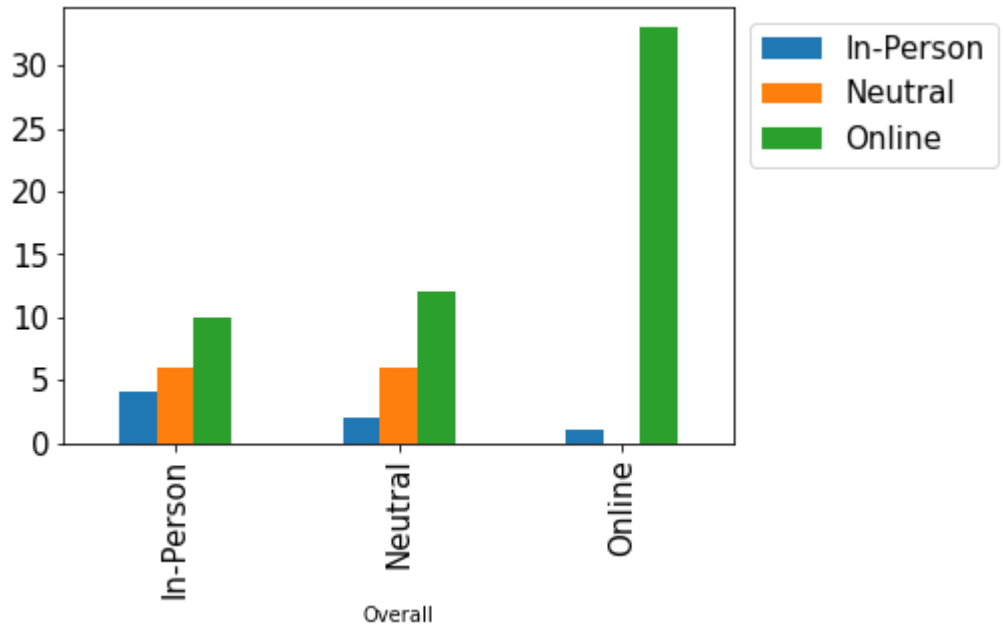
```
plt.xticks(fontsize=15)
plt.yticks(fontsize=15)
```

Out[21]:
```
(array([ 0.,  5., 10., 15., 20., 25., 30., 35.]),
 [Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, '')])
```



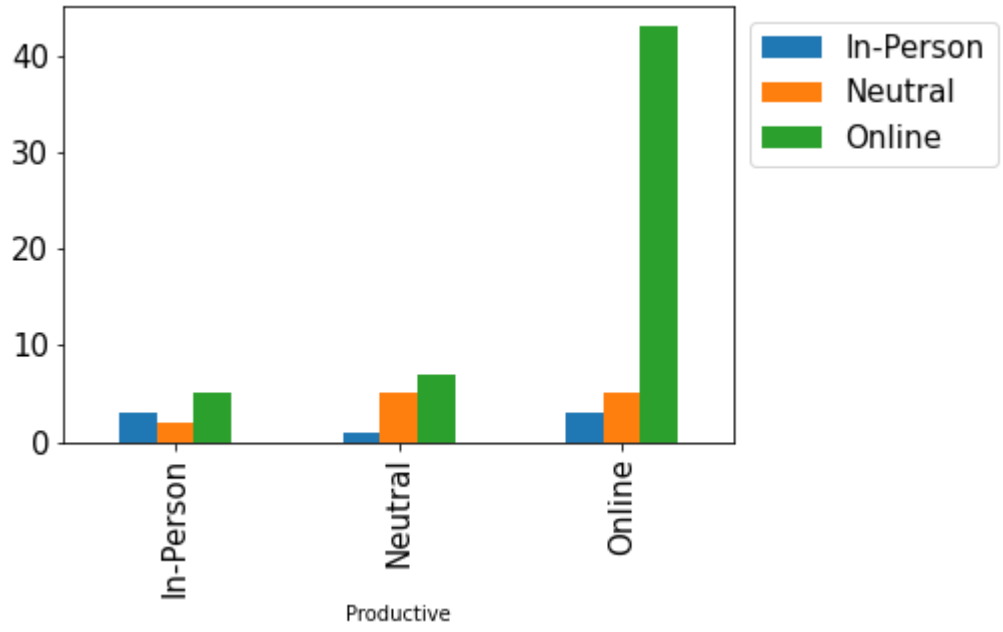
In [22]:
```
#Relationship between Overall preference and Social interaction

#Create cross tab for Overall and Social
OS=pd.crosstab(Overall,Social)

#Using plot bar to create visualization of two colums' relationship
OS.plot.bar()
plt.legend(bbox_to_anchor=(1.0, 1.0),prop={'size': 15})
plt.xticks(fontsize=15)
plt.yticks(fontsize=15)
```

Out[22]:
```
(array([ 0.,  5., 10., 15., 20., 25., 30., 35.]),
 [Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, '')])
```



In [23]:
```
#Relationship between Productive and Social interaction

#Create cross tab for Productive and Social
PS=pd.crosstab(Productive,Social)

#Using plot bar to create visualization of two colums' relationship
PS.plot.bar()
plt.legend(bbox_to_anchor=(1.0, 1.0),prop={'size': 15})
plt.xticks(fontsize=15)
plt.yticks(fontsize=15)
```

Out[23]:
```
(array([ 0., 10., 20., 30., 40., 50.]),
 [Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, '')])
```



Test: Correlation analysis by pearson

Result:

According to the plot bar graphs we find out the students who prefer online class have a tendency to do better on social interaction with online method and to be more productive for online classes. However, students who prefer in-person classes do not have the same tendency for social interaction. Instead they also prefer online method for social interaction. As the result of the pearson test, we find out that there are a weak positive correlations between Overall School preference, Social interaction preference, and be more productive.

Like our prediction we are able to find out positive correlation between More Productive and social interaction questions. There is a weak positive correlations between more productive and Social interactions. It is because the test result is 0.37

Lastly, there is also a weak positive correlation betwwen Overall and Social interactions because the test result came out 0.47.

Test #3.4 (Pearson):

In [24]:
```
# Create new dataframe that contains productive, overall preference, and social interac
data1 = df[["I feel more productive when attending class in-person vs. online.","Overal
data1

# Using pearson method to test Correlation
r1 = data1.corr(method='pearson')
r1
```

Out[24]:

	I feel more productive when attending class in-person vs. online.	Overall, I prefer in-person vs. online.	Social interactions are better in-person vs. online.
I feel more productive when attending class in-person vs. online.	1.000000	0.424188	0.373891
Overall, I prefer in-person vs. online.	0.424188	1.000000	0.444088
Social interactions are better in-person vs. online.	0.373891	0.444088	1.000000

3.5 Relationship Between Gender and Overall School Preferrence

Hypothesis #3.5:

We hypothesize that more students regardless of their gender will prefer online learning as opposed to in-person learning because they save more time, money, and they perform better in school.

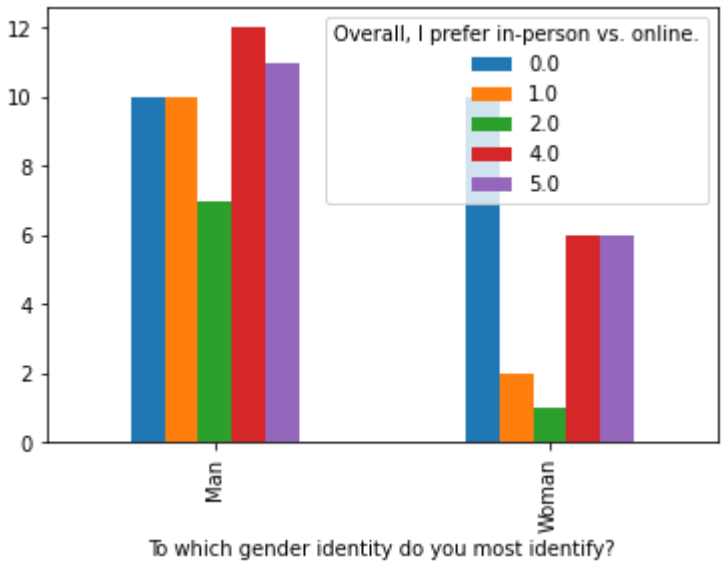
In [25]:
```
Gender_Corr = pd.crosstab(df["To which gender identity do you most identify?"], df["Ove
Gender_Corr
```

Out[25]:

Overall, I prefer in-person vs. online.	0.0	1.0	2.0	4.0	5.0
To which gender identity do you most identify?					
Man	10	10	7	12	11
Woman	10	2	1	6	6

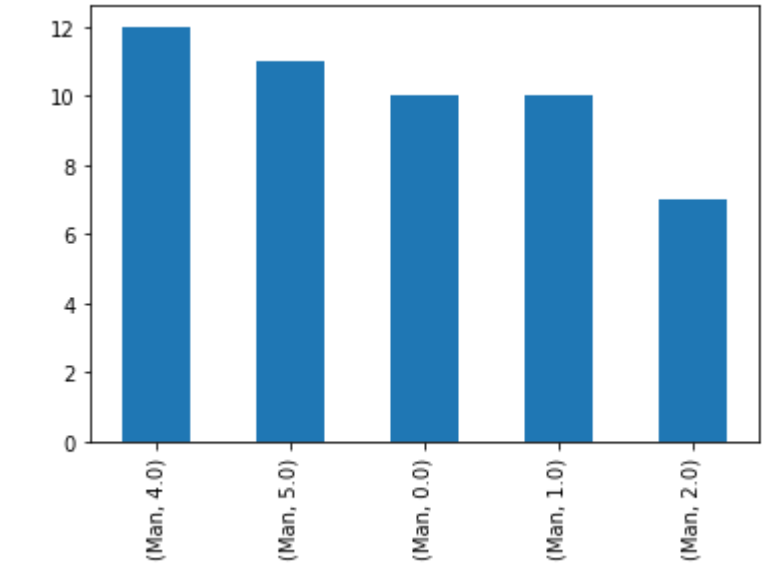
In [26]:
```
Gender_Corr.plot.bar()
```

Out[26]:
```
<AxesSubplot:xlabel='To which gender identity do you most identify?'>
```



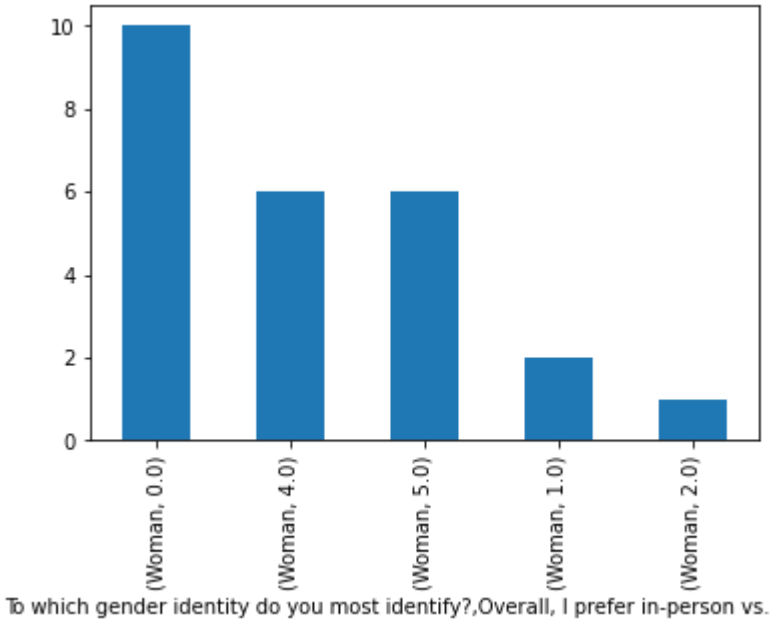
In [27]:
```
data_df2 = df[(df["To which gender identity do you most identify?"]=='Man')]
ManGender = data_df2.groupby("To which gender identity do you most identify?")["Overall
ManGender
```

Out[27]:
```
<AxesSubplot:xlabel='To which gender identity do you most identify?,Overall, I prefer in
-person vs. online.'>
```



In [28]:
```
data_df3 = df[(df["To which gender identity do you most identify?"]=='Woman')]
WomanGender = data_df3.groupby("To which gender identity do you most identify?")["Overa
WomanGender
```

Out[28]:
```
<AxesSubplot:xlabel='To which gender identity do you most identify?,Overall, I prefer in
-person vs. online.'>
```



Here we can see the the data and the histogram visualization/graph, and it shows that out of all the men, the men prefer mostly to be online and out of all the women, the women prefer to be more neutral but also lean towards being online as well. This shows that both genders still do lean more towards learning online, but more women are neutral. So we beleive that gender does not affect the person's decision if they want to be online or in-person so we would conclude that there is no correlation.

In [29]:
```
gendercross = pd.crosstab(df["To which gender identity do you most identify?"], df["Ove
            normalize=True, margins=True)
gendercross
```

Out[29]:

Overall, I prefer in-person vs. online.	0.0	1.0	2.0	4.0	5.0	All
To which gender identity do you most identify?						
Man	0.133333	0.133333	0.093333	0.16	0.146667	0.666667
Woman	0.133333	0.026667	0.013333	0.08	0.080000	0.333333
All	0.266667	0.160000	0.106667	0.24	0.226667	1.000000

Result:

Looking at the the proportions of the graphs, 46% of all men prefer or lean towards online whereas 48% of all women prefer or lean towards online. This data is obtained by combining 4.0 and 5.0. 44% of all men prefer or lean towards in-person whereas 12% of all women prefer or lean towards online. This data is obtained by combining 1.0 and 2.0. 20% of men felt neutral and 40% of women felt neutral. This data is obtained by looking at 0.0. By looking at all of this, both men and women prefer online over in-person. However, the majority of women chose to be neutral (0.0) which was 40% of women and the majority of men chose to be online (5.0) which was 22% of men. Where the men chose in-person the women chose neutral.

Test #3.5 (Pearson and Chi-Squared):

In [30]:

```
df["To which gender identity do you most identify?"].replace(['Man','Woman'],[0,1],inpl

# Create a dataframe for gender and overall preference only
GenderData = df[["To which gender identity do you most identify?","Overall, I prefer in
GenderData

#Using pearson correlation test to find out the correlation.
r = GenderData.corr(method='pearson')
r

print("Correlation between gender and preferring online or inperson is: ", GenderData["
```

```
Correlation between gender and preferring online or inperson is:  -0.051796977028281184
```

Based off the Pearson correlation calculation there is no correlation between gender and the person's preference to attend school online or in-person since the value of the correlation -0.051796977028281184 is really close to 0.

In [31]:

```
from scipy.stats import chi2_contingency
chi, p, dfree, expected = chi2_contingency(Gender_Corr)
print("chi square value: " ,chi)
print("p value: ", p)
print("degree of freedom: ", dfree)
print("expected frequencies when story type and gender is independent: ", expected)
from scipy.stats import chi2
print("for 1 degree of freedom, the chi-square valued needed to reject the hypothesis a
chi2.ppf(0.999, 4)
```

```
chi square value:  5.591911764705882
```

```
p value:  0.23176771813219454
degree of freedom:  4
expected frequencies when story type and gender is independent:  [[13.33333333  8.
5.33333333 12.         11.33333333]
 [ 6.66666667  4.          2.66666667  6.          5.66666667]]
for 1 degree of freedom, the chi-square valued needed to reject the hypothesis at the 0.
001 significance level is:
18.46682695290317
```

Out[31]:

Null Hypothesis: The person's gender and their preference of whether to learn in-person or online are not correlated.

Alternate Hypothesis: The person's gender and their preference of whether to learn in-person or online are correlated.

Since at 0.001 significance level, the p-value of 0.23176771813219454 is greater than the significance level we fail to reject the null hypothesis. Therefore, there is sufficient evidence that the person's gender and their preference of whether to learn in-person or online are not correlated.

3.6 Relation Between Visual/Hands-On Learning and Overal School Preferrence

Hypothisis #3.6:

We hypothesize that students who have a higher prefrence for hands on learning will have a negative correlation to online vs inperson prefrence, meaning students who are strong hands on learners will perefer to have in-person classes. We also think that visual learners will prefer online classes. We also assume that prefering to have in-person classes is dependent on having a strong visual and hands on learning prefrence.

In [32]:
```
VisualLearningCrosstab = pd.crosstab(df['Overall, I prefer in-person vs. online.'], df[
VisualLearningCrosstab
```

Out[32]:

Visual materials (graphs, tables, pictures, etc.) help with my learning.	3.0	4.0	5.0
Overall, I prefer in-person vs. online.			
0.0	0	8	12
1.0	0	3	9
2.0	0	2	6
4.0	0	5	13
5.0	2	6	9

In [33]:
```
HandsOnCrosstab = pd.crosstab(df['Overall, I prefer in-person vs. online.'], df['Hands-
HandsOnCrosstab
```

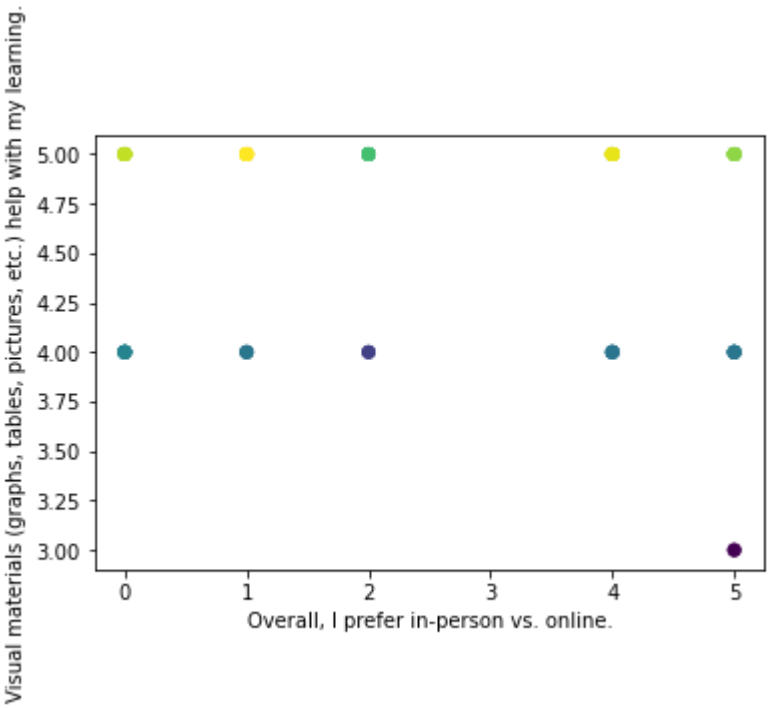
Out[33]:

Hands-on activities (worksheets, in-person labs, class assignments) help with my learning.	1.0	2.0	3.0	4.0	5.0
Overall, I prefer in-person vs. online.					
0.0	0	0	1	5	14
1.0	0	1	0	2	9
2.0	1	0	1	2	4
4.0	0	0	0	1	17
5.0	0	0	1	5	11

For the Scatter plot:

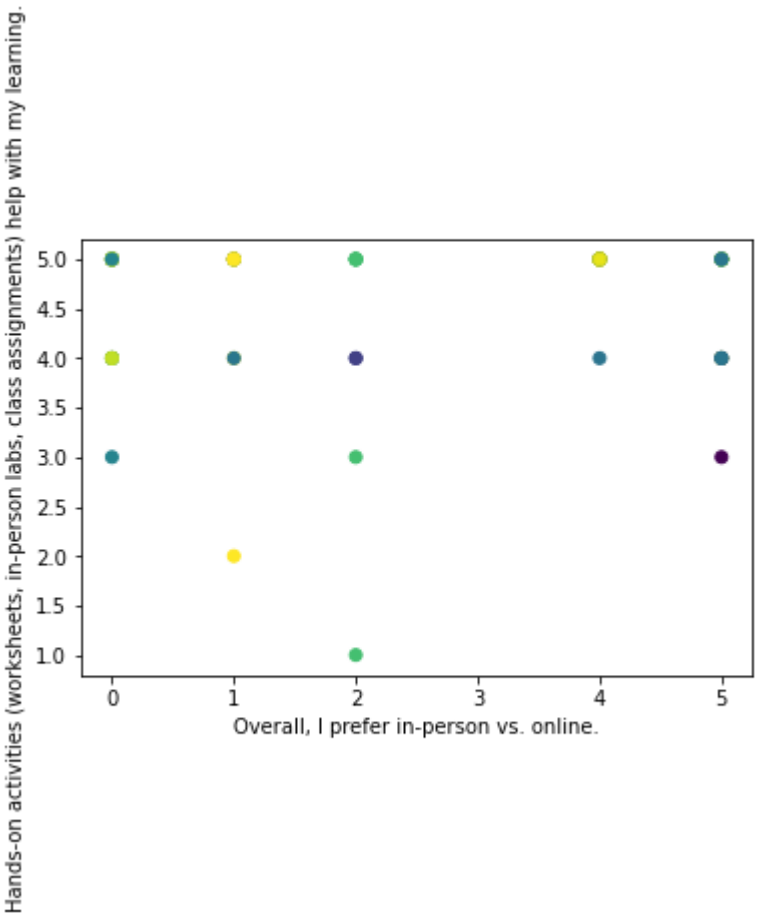
The darker the color the less points there are in one location. The lighter the color the more data points there are in one location

In [34]:
```
import matplotlib.pyplot as plt
from scipy.stats import gaussian_kde
xy = np.vstack([df["Overall, I prefer in-person vs. online."],df["Visual materials (gra
z = gaussian_kde(xy)(xy)
plt.scatter(x=df["Overall, I prefer in-person vs. online."],y=df["Visual materials (gra
plt.ylabel("Visual materials (graphs, tables, pictures, etc.) help with my learning.")
plt.xlabel("Overall, I prefer in-person vs. online.")
plt.show()
```



In [35]:
```
xy = np.vstack([df["Overall, I prefer in-person vs. online."],df["Visual materials (gra
z = gaussian_kde(xy)(xy)
plt.scatter(df["Overall, I prefer in-person vs. online."],df["Hands-on activities (work
plt.ylabel("Hands-on activities (worksheets, in-person labs, class assignments) help wi
```

```
plt.xlabel("Overall, I prefer in-person vs. online.")
plt.show()
```



Test #3.6 (Pearson and Chi-Squared):

In [36]:
```
print("Correlation between Visual learning and Prefering online or inperson is: ",df["O
print("Correlation between Hands-On Learning and Prefering online or inperson is: ",df[
```

```
Correlation between Visual learning and Prefering online or inperson is:  -0.09856506539
828293
Correlation between Hands-On Learning and Prefering online or inperson is:  0.0811216760
6592037
```

In [37]:
```
from scipy.stats import chi2_contingency
data = VisualLearningCrosstab#unChangedDf["Overall, I prefer in-person vs. online.","Ha
chi, p, dfree, expected = chi2_contingency(data)
print("chi square value: " ,chi)
print("p value: ", p)
print("degree of freedom: ", dfree)
print("expected frequencies when visual learning and school preferrence is independent:
from scipy.stats import chi2
print("for 8 degree of freedom, the chi-square valued needed to reject the hypothesis a
chi2.ppf(0.999, 8)
```

```
chi square value:  8.591165632919832
p value:  0.3779486640782953
degree of freedom:  8
```

```
expected frequencies when visual learning and school preferrence is independent:  [[ 0.5
3333333  6.4        13.06666667]
 [ 0.32        3.84        7.84      ]
 [ 0.21333333  2.56        5.22666667]
 [ 0.48        5.76       11.76      ]
 [ 0.45333333  5.44       11.10666667]]
for 8 degree of freedom, the chi-square valued needed to reject the hypothesis at the 0.
001 significance level:
26.12448155837614
```

Out[37]:

In [38]:

```
data = HandsOnCrosstab#unChangedDf["Overall, I prefer in-person vs. online.","Hands-on
chi, p, dfree, expected = chi2_contingency(data)
print("chi square value: " ,chi)
print("p value: ", p)
print("degree of freedom: ", dfree)
print("expected frequencies when hands on learning and school preferrence is independen
from scipy.stats import chi2
print("for 16 degree of freedom, the chi-square valued needed to reject the hypothesis
chi2.ppf(0.999, 16)
```

```
chi square value:  21.413250148544265
p value:  0.16316707645782957
degree of freedom:  16
expected frequencies when hands on learning and school preferrence is independent:  [[
0.26666667  0.26666667  0.8         4.         14.66666667]
 [ 0.16        0.16        0.48        2.4         8.8       ]
 [ 0.10666667  0.10666667  0.32        1.6         5.86666667]
 [ 0.24        0.24        0.72        3.6        13.2       ]
 [ 0.22666667  0.22666667  0.68        3.4        12.46666667]]
for 16 degree of freedom, the chi-square valued needed to reject the hypothesis at the
0.001 significance level:
39.252354790768464
```

Out[38]:

Results

Our initial prediction for the relationship between learning styles and if people prefer online or in-person was that we assumed People who have strong Hands on learning will prefer in-person class, so a positive relation and people who have strong visual learning will prefer online.

However after preforming some analysis on the data using a crosstabulation and scatter plot, we saw that our prediction was wrong. Almost everyone was a visual learner regardless of in-person or online preference, so there is no correlation there.

For hands on learning we again see that we were not correct, a high amount of people that have a hands on learning preference still prefer online.

When we preformed a correlation test, it confirms our analysis being that tests are close to 0 meaning there is almost no correlation between the two sets of columns. We also preformed the chi squared test on the data and saw that the learning styles are independent of preferring online or in-person

3.7 Relationship Between Hours On-Campus and Overall School Preferrence

Hypothisis #3.7:

We hypothesize that people who spend less hours on campus will most likely prefer to have school online, as they probably prefer to do homework and study at home. We also assumed that their prefrence for online vs in-person will be dependent on how many hours they spend on campus.

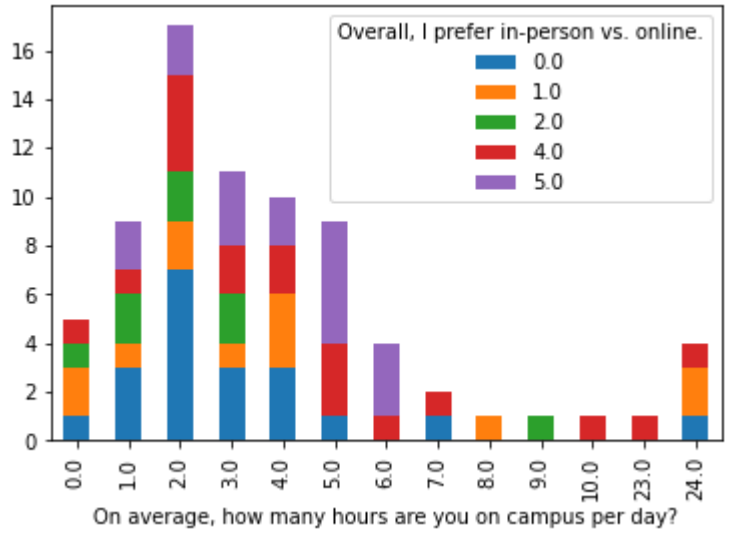
In [39]:
```
HoursOnCampus = pd.crosstab(df["On average, how many hours are you on campus per day?"]
HoursOnCampus
```

Out[39]:

Overall, I prefer in-person vs. online.	0.0	1.0	2.0	4.0	5.0
On average, how many hours are you on campus per day?					
0.0	1	2	1	1	0
1.0	3	1	2	1	2
2.0	7	2	2	4	2
3.0	3	1	2	2	3
4.0	3	3	0	2	2
5.0	1	0	0	3	5
6.0	0	0	0	1	3
7.0	1	0	0	1	0
8.0	0	1	0	0	0
9.0	0	0	1	0	0
10.0	0	0	0	1	0
23.0	0	0	0	1	0
24.0	1	2	0	1	0

In [40]:
```
HoursOnCampus.plot.bar(stacked=True)
```

Out[40]:
```
<AxesSubplot:xlabel='On average, how many hours are you on campus per day?'>
```



Test #3.7 (Pearson and Chi-Squared):

In [41]:
```
print("Correlation between time on campus and Prefering online or inperson is: ",df["Ov
```

```
Correlation between time on campus and Prefering online or inperson is:  0.0349187468643
44395
```

In [42]:
```
data = HoursOnCampus
chi, p, dfree, expected = chi2_contingency(data)
print("chi square value: " ,chi)
print("p value: ", p)
print("degree of freedom: ", dfree)
print("expected frequencies when time spent on campus and school preferrence is indepen
from scipy.stats import chi2
print("for 48 degree of freedom, the chi-square valued needed to reject the hypothesis
chi2.ppf(0.999, 48)
```

```
chi square value:  52.49192691039577
p value:  0.3041567222922587
degree of freedom:  48
expected frequencies when time spent on campus and school preferrence is independent:
[[1.33333333 0.8        0.53333333 1.2        1.13333333]
 [2.4        1.44       0.96       2.16       2.04      ]
 [4.53333333 2.72       1.81333333 4.08       3.85333333]
 [2.93333333 1.76       1.17333333 2.64       2.49333333]
 [2.66666667 1.6        1.06666667 2.4        2.26666667]
 [2.4        1.44       0.96       2.16       2.04      ]
 [1.06666667 0.64       0.42666667 0.96       0.90666667]
 [0.53333333 0.32       0.21333333 0.48       0.45333333]
 [0.26666667 0.16       0.10666667 0.24       0.22666667]
 [0.26666667 0.16       0.10666667 0.24       0.22666667]
 [0.26666667 0.16       0.10666667 0.24       0.22666667]
 [0.26666667 0.16       0.10666667 0.24       0.22666667]
 [1.06666667 0.64       0.42666667 0.96       0.90666667]]
for 48 degree of freedom, the chi-square valued needed to reject the hypothesis at the
0.001 significance level:
```

Out[42]:
```
84.03713371722348
```

Results

Our prediction for a correlation between time spent on campus and in-person v online was that people who spend less time on campus would likely prefer online class and people who spend more time on campus would prefer in-person class. When making a visualization of this relationship using a bar graph, it showed that a majority of people who spend less time on campus wanted to have online classes. However there were also quite a few people who wanted in person classe, but did not spend a lot of time on campus. As for people that spent greater than 6 hours on campus it seemed that a majorty of them were prefering in person or netrual. This somewhat matches our hypothisis, however, we did not account for as many people that wanted in-person to also spend less time on campus.

We can see that doing a correlation test we see the columns are not corrilated. This is likely due to the high number of people that spend 6 hours or less on campus

We also see that in the Chi Squared test that once again the amount of hours students spend on campus is independant of their online vs in-person prefrence

Mini-Project Answers

Question 1:

Our data contains survey responses from students of CS 105 and CS 111.

Question 2:

If there are correlations between overall preferences of in-person and online and other factors. (Such as gender, transfer student or non transfer student, learning style, average hours student stays on the campus, etc...)

Question 3:

1. We found the corresponding values by counting the frequency of GPAs based on a certain school preferrence.
2. We Found the corresponding values by counting the frequency of different learning styles based on a certain school preferrence.
3. We use value_counting functions to find out frequencies from the data. By using Pie chart visualisation, we find out the distribution of student's preferences.
4. We use value_counting functions to find out frequencies from the data. By using bar graphs to compare how frequencies are differently distributed.
5. We created a cross tabulation of the student's gender and compared to school preference. We then made a histogram to visualize if there is a correlation.
6. We created a cross tabulation of visual learning and hands on learning compared to school preference. We then made a scatter plot to visualize if there is a correlation.

7. We created a cross tabulation of hours spent on campus and overall school preference. We then made a bar graph to see the distribution of students and how long they spend on campus.

Question 4:

1. We hypothesize that students prefer online school more than in-person school because they perform better GPAs wise when learning online.
2. We hypothesize that kinesthetic learners will prefer in-person school due to more hands-on activities and learners who prefer reading/visual would like online school more due to more visuals and the ability to google search and read easily.
3. Students' preferences of in_person and online methods have a correlation with whether students are transfer students or not. We assumed that transfer sutdent would like to prefer in_person and Non-transfer student would like to prepfer online method.
4. If students prefer online classes, then they will also prefer online method of social interaction and will be more productive on online classes. If student prefer in-person classes, then they will also prefer in-person method of social interaction and will be more productive on in-person classes
5. We hypothesize that more students regardless of their gender will prefer online learning as opposed to in-person learning because they save more time, money, and they perform better in school.
6. We hypothesize that students who have a higher prefrence for hands on learning will have a negative correlation to online vs inperson prefrence, meaning students who are strong hands on learners will perefer to have in-person classes. We also think that visual learners will prefer online classes. We also assume that prefering to have in-person classes is dependent on having a strong visual and hands on learning prefrence.
7. We hypothesize that people who spend less hours on campus will most likely prefer to have school online, as they probably prefer to do homework and study at home. We also assumed that their prefrence for online vs in-person will be dependent on how many hours they spend on campus.

To verify our hypothesis we used the Chi-Squared Test and the Pearson Correlation Test

Question 5:

Reference:

1. Test #3.1 (Pearson and Chi-Squared)
2. Test #3.2 (Pearson and Chi-Squared)
3. Test #3.3 (Pearson)
4. Test #3.4 (Pearson)
5. Test #3.5 (Pearson and Chi-Squared)
6. Test #3.6 (Pearson and Chi-Squared)
7. Test #3.7 (Pearson and Chi-Squared)

In [15]:
```
import pandas as pd
import numpy as np
from scipy.stats import chi2_contingency
```

In [16]:
```
df = pd.read_csv("Toy data.csv")
df
```

Out[16]:

	Hands-on activities (worksheets, in-person labs, class assignments) help with my learning.	Overall, I prefer in-person vs. online.
0	5	1
1	4	5
2	5	4

In [17]:
```
print("Correlation between Visual and Prefering online or inperson is: ",df["Overall, I
```

```
Correlation between Visual and Prefering online or inperson is:  -0.6933752452815363
```

In [19]:
```
data = df#unChangedDf["Overall, I prefer in-person vs. online.","Hands-on activities (w
chi, p, dfree, expected = chi2_contingency(data)
print("chi square value: " ,chi)
print("p value: ", p)
print("degree of freedom: ", dfree)
print("expected frequencies when hands on learning and school preferrence is independen
from scipy.stats import chi2
print("for 2 degree of freedom, the chi-square valued needed to reject the hypothesis a
chi2.ppf(0.999, 2)
```

```
chi square value:  2.2857142857142856
p value:  0.3189065573239704
degree of freedom:  2
expected frequencies when hands on learning and school preferrence is independent:  [[3.
5  2.5 ]
 [5.25 3.75]
 [5.25 3.75]]
for 2 degree of freedom, the chi-square valued needed to reject the hypothesis at the 0.
001 significance level:
```

Out[19]:
```
13.815510557964274
```

In [ ]:

Pearson test

Wednesday, February 16, 2022 1:36 AM

X = Hands on

6
4
5

Y = Online v in-person

1
5
4

$\bar{X} = 4.6$ $\bar{Y} = 3.3$

a	b	axb	a^2	b^2
.4	-2.3	-.92	.16	5.29
-.6	1.7	-1.02	.36	2.89
.4	.7	.28	.16	.49
Sums		-1.66	.68	8.67

$$r_{xy} = \frac{\sum axb}{\sqrt{\sum a^2 \times \sum b^2}} = \frac{-1.66}{\sqrt{.68 \times 8.67}} = \boxed{-.683}$$

	Hands-on Activities...		Overall I prefer in-...		Total
	5	1,1	1	1,2	6
	4	2,1	5	2,2	9
	5	3,1	4	3,2	9
Total	14		10		24

$$E_i = \frac{\text{Row total} \cdot \text{column total}}{\text{Grand total}}$$

↳ expected data

5 $E_{1,1} = \frac{6 \cdot 14}{24} = 3.5$

4 $E_{2,1} = \frac{9 \cdot 14}{24} = 5.25$

5 $E_{3,1} = \frac{9 \cdot 14}{24} = 5.25$

1 $E_{1,2} = \frac{6 \cdot 10}{24} = 2.5$

5 $E_{2,2} = \frac{9 \cdot 10}{24} = 3.75$

4 $E_{3,2} = \frac{9 \cdot 10}{24} = 3.75$

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$X^2 = \frac{(5-3.5)^2}{3.5} + \frac{(4-5.25)^2}{5.25} + \frac{(5-5.25)^2}{5.25} + \frac{(1-2.5)^2}{2.5} + \frac{(5-3.75)^2}{3.75} + \frac{(4-3.75)^2}{3.75}$$

$$= 2.285714286$$

degrees of freedom = (rows - 1) · (columns - 1)

$$df = (3-1) \cdot (2-1)$$

$$df = 2$$

p-value = $P(X^2 > 2.285714286)$

$= 0.3189065573$ ← rounded

chosen significance level = 0.001

Null hypothesis:
There is no relationship between hands-on activities and overall school preferrence.

Alternate hypothesis:
There is a relationship between hands on activities and overall school preferrence.

Since at 0.001 significance level the p-value of 0.318906557 3 is greater than the significance level, then we fail to reject the null hypothesis.

Therefore, there is sufficient evidence that there is no relationship between hands on activities and overall school preference.