CS105 Mini-Project

Does who a student is living with effect if and how they work jobs?

By:

- Anshul Gupta agupt109@ucr.edu
- Ali Naqvi anaqv007@ucr.edu
- Alex Zhang azhan061@ucr.edu
- Nathan Lee nlee097@ucr.edu

Pre-Data Questions & Analysis

What data do we have?

We have data regarding the living situations of students across multiple CS classes. We asked them who they live with, how many people they live with, where they live, whether they work, how much they work, etc. There are a mix of categorical and quantitative datapoints that we will analyze to answer what we want to know:

What do we want to know?

We want to know how various aspects of a student's home environment go on to affect their employment and school performance. Specifically who they live with and the affects of that.

Hypothesis & Predictions

- Hypothesis 1: There will be a correlation between whether people live with family, friends, or neither and whether or not they work.
- Hypothesis 2: Students who live on-campus are more likely to have roommates of the same major.
- Hypothesis 3: People who live with more people will have a higher GPA on average.

Data Loading & Preprocessing

In [1]:
```
%matplotlib inline
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt

# Load dataframe from data.csv
df = pd.read_csv("data.csv")

# Select relevant columns
df = df.iloc[:, [0, 2, 3, 7, 8, 9, 34, 55, 58, 59, 60, 61, 62, 26, 66]]
df
```

Out[1]:

	Timestamp	What is your current class standing?	What is your age?	Who do you live with?	Do you currently live in a house, apartment, or dorm?	How many people live in your household?	What was your GPA your very first quarter at UCR?	Wha your ca plans gradua
0	2/9/2024 20:12:14	Senior	23+	Neither	House	6	2.73	Get int Job Ind
1	2/9/2024 20:16:34	Junior	20	Both	Apartment	4	3.7	Get int Job Ind
2	2/9/2024 20:18:55	Junior	23+	Friends	House	4	3.75	If no jo to grac sc
3	2/9/2024 20:24:00	Senior	23+	Neither	Apartment	1	3.81	Not Sur
4	2/9/2024 20:26:16	Graduate	22	Neither	Apartment	1	3.23	Get int Job Ind
...	...	...	...	...	...	...	...	
255	2/14/2024 19:46:28	Junior	21	Friends	House	5	4	Get int Job Ind
256	2/15/2024 0:28:38	NaN	21	Family	Apartment	North District 4 bed 2 bath	3.5	Get int Job Ind
257	2/15/2024 8:33:45	Senior	21	Family	House	9	3.7	Attend Sc
258	2/15/2024 16:10:40	Sophomore	21	Family	Apartment	4	3	Get int Job Ind

	Timestamp	What is your current class standing?	What is your age?	Who do you live with?	Do you currently live in a house, apartment, or dorm?	How many people live in your household?	What was your GPA your very first quarter at UCR?	Wha your ca plans gradua
259	2/15/2024 16:14:11	Sophomore	18	Friends	Dorm	3 (room), 8 (hall), ~70 (building)	4	Attend Sc

260 rows × 15 columns

Preprocessing

In [2]:

```
# Fixes empty values
df['Do you currently work?'] = df['Do you currently work?'].fillna('No')

# Replaces custom text answers with appropriate values
df['How many people live in your household?'] = (df['How many people live in
                                                 .fillna(0)
                                                 .replace('4 in total', '4')
                                                 .replace('4 (Including me)'
                                                 .replace('at school 4 inclu
                                                 .replace('3 excluding me',
                                                 .replace('5 including me',
                                                 .replace('North District 4
                                                 .replace('3 (room), 8 (hall
                                                 .astype(int))
df['Who do you live with?'] = df['Who do you live with?'].replace('Family, F
    'Family, Friends, Both', 'Both')
df['Do you currently live in a house, apartment, or dorm?'] = (
    df['Do you currently live in a house, apartment, or dorm?']
    .replace('house (renting)', 'House'))

df.loc[df['What was your GPA your very first quarter at UCR?'].str.contains(
    "I am not sure|idk|I don't know|This is my first quarter|i don't rem|not
df['What was your GPA your very first quarter at UCR?'] = (
    df['What was your GPA your very first quarter at UCR?']
    .replace('Idk, I think 3.2 or something along those lines', '3.2')
    .replace('2.8?', '2.8')
    .replace('3 point something', '3.0')
    .replace('3.67 I think', '3.67')
    .replace('3.0?', '3.0')
    .replace('about 3.0', '3.0')
    .astype(np.float64))
```

```
df.loc[df['How many internship/job applications have you sent out so far?'].
    "A lot|idk|I don't know|More than enough|Not enough|Many|not sure|I dont
df['How many internship/job applications have you sent out so far?'] = (
    df['How many internship/job applications have you sent out so far?']
    .fillna(0)
    .replace('100s', '100')
    .replace(
        'I haven’t sent any internships I think I need to take more courses
        '0')
    .replace('none :(', '0')
    .replace('15+', '15')
    .replace('none for now', '0')
    .replace('20+', '20')
    .replace('Above 50 for this summer but overall over the last 4 years ove
    .replace('5-10', '7')
    .replace('200+', '200')
    .replace('50+', '50')
    .replace('none', '0')
    .replace('25+', '25')
    .replace('~20', '20')
    .replace('100+', '100')
    .replace('50-80 this year', '65')
    .replace('300+', '300')
    .replace('30-40', '35')
    .replace('~60', '60')
    .replace('between 50-100', '75')
    .replace('Less than 10 :( Too many things to do', '10')
    .replace('150-200', '175')
    .replace('>100', '100')
    .replace('~50', '50')
    .replace('Over 20', '20')
    .replace('10-20', '15')
    .astype(int))
# Normalizes non-applicable answers
df.loc[df['Do you currently work?'] == 'No', 'How many hours do you work per
df.loc[df['Do you currently work?'] == 'No', 'Do you work in a department re

df
```

Out[2]:

	Timestamp	What is your current class standing?	What is your age?	Who do you live with?	Do you currently live in a house, apartment, or dorm?	How many people live in your household?	What was your GPA your very first quarter at UCR?	Wha your ca plans gradua
0	2/9/2024 20:12:14	Senior	23+	Neither	House	6	2.73	Get int Job Ind
1	2/9/2024 20:16:34	Junior	20	Both	Apartment	4	3.70	Get int Job Ind
2	2/9/2024 20:18:55	Junior	23+	Friends	House	4	3.75	If no jo to grad s
3	2/9/2024 20:24:00	Senior	23+	Neither	Apartment	1	3.81	Not Sur
4	2/9/2024 20:26:16	Graduate	22	Neither	Apartment	1	3.23	Get int Job Ind
...	...	...	...	...	...	...	...	
255	2/14/2024 19:46:28	Junior	21	Friends	House	5	4.00	Get int Job Ind
256	2/15/2024 0:28:38	NaN	21	Family	Apartment	4	3.50	Get int Job Ind
257	2/15/2024 8:33:45	Senior	21	Family	House	9	3.70	Attend S
258	2/15/2024 16:10:40	Sophomore	21	Family	Apartment	4	3.00	Get int Job Ind

	Timestamp	What is your current class standing?	What is your age?	Who do you live with?	Do you currently live in a house, apartment, or dorm?	How many people live in your household?	What was your GPA your very first quarter at UCR?	Wha your ca plans gradua
259	2/15/2024 16:14:11	Sophomore	18	Friends	Dorm	3	4.00	Attend Sc

260 rows × 15 columns

In [3]:

```
# Working DataFrame
w_df = df[df['Do you currently work?'] == 'Yes']
# Not working DataFrame
nw_df = df[df['Do you currently work?'] == 'No']
w_df
```

Out[3]:

	Timestamp	What is your current class standing?	What is your age?	Who do you live with?	Do you currently live in a house, apartment, or dorm?	How many people live in your household?	What was your GPA your very first quarter at UCR?	Wha your ca plans gradua
0	2/9/2024 20:12:14	Senior	23+	Neither	House	6	2.73	Get int Job Ind
4	2/9/2024 20:26:16	Graduate	22	Neither	Apartment	1	3.23	Get int Job Ind
8	2/9/2024 22:02:49	Junior	20	Friends	House	6	3.40	Get int Job Ind
9	2/9/2024 22:08:43	Senior	22	Family	House	5	NaN	Not Sur
13	2/9/2024 22:15:13	Junior	21	Family	Apartment	4	3.50	Attend S
...	...	...	...	...	...	...	...	
246	2/13/2024 19:37:02	Graduate	23+	Family	House	2	4.00	Get int Job Ind
247	2/13/2024 21:39:14	Senior	21	Friends	Apartment	3	3.60	Get int Job Ind
252	2/14/2024 9:48:12	Junior	20	Family	House	5	3.50	Get int Job Ind
255	2/14/2024 19:46:28	Junior	21	Friends	House	5	4.00	Get int Job Ind

	Timestamp	What is your current class standing?	What is your age?	Who do you live with?	Do you currently live in a house, apartment, or dorm?	How many people live in your household?	What was your GPA your very first quarter at UCR?	Wha your ca plans gradua
258	2/15/2024 16:10:40	Sophomore	21	Family	Apartment	4	3.00	Get int Job Ind

77 rows × 15 columns

```
In [4]: nw_df
```

Out[4]:

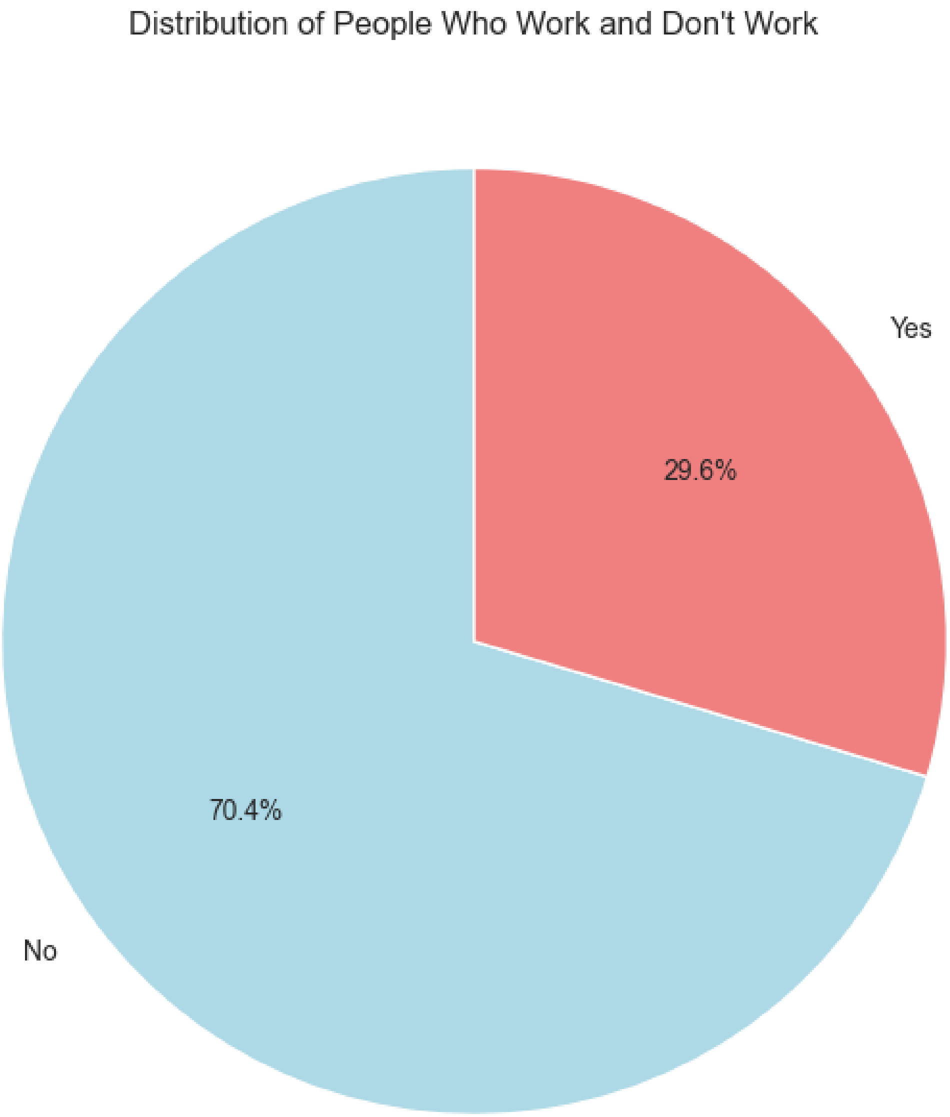
	Timestamp	What is your current class standing?	What is your age?	Who do you live with?	Do you currently live in a house, apartment, or dorm?	How many people live in your household?	What was your GPA your very first quarter at UCR?	Wha your ca plans gradua
1	2/9/2024 20:16:34	Junior	20	Both	Apartment	4	3.70	Get int Job Ind
2	2/9/2024 20:18:55	Junior	23+	Friends	House	4	3.75	If no jc to grac sc
3	2/9/2024 20:24:00	Senior	23+	Neither	Apartment	1	3.81	Not Sur
5	2/9/2024 20:45:09	Junior	21	Both	Apartment	4	4.00	Get int Job Ind
6	2/9/2024 21:55:59	Sophomore	19	Friends	Apartment	4	4.00	Get int Job Ind
...	...	...	...	...	...	...	...	
253	2/14/2024 13:45:45	Senior	21	Family	House	6	4.00	Get int Job Ind
254	2/14/2024 16:26:06	Junior	19	Family	House	5	3.80	Get int Job Ind
256	2/15/2024 0:28:38	NaN	21	Family	Apartment	4	3.50	Get int Job Ind
257	2/15/2024 8:33:45	Senior	21	Family	House	9	3.70	Attend Sc
259	2/15/2024 16:14:11	Sophomore	18	Friends	Dorm	3	4.00	Attend Sc

183 rows × 15 columns

Analysis

In [5]:

```
# Count the number of people who work and don't work
work_counts = df['Do you currently work?'].value_counts()

# Plotting a pie chart
plt.figure(figsize=(8, 8))
plt.pie(work_counts, labels=work_counts.index, autopct='%1.1f%%', startangle
plt.title('Distribution of People Who Work and Don\'t Work')
plt.show()
```



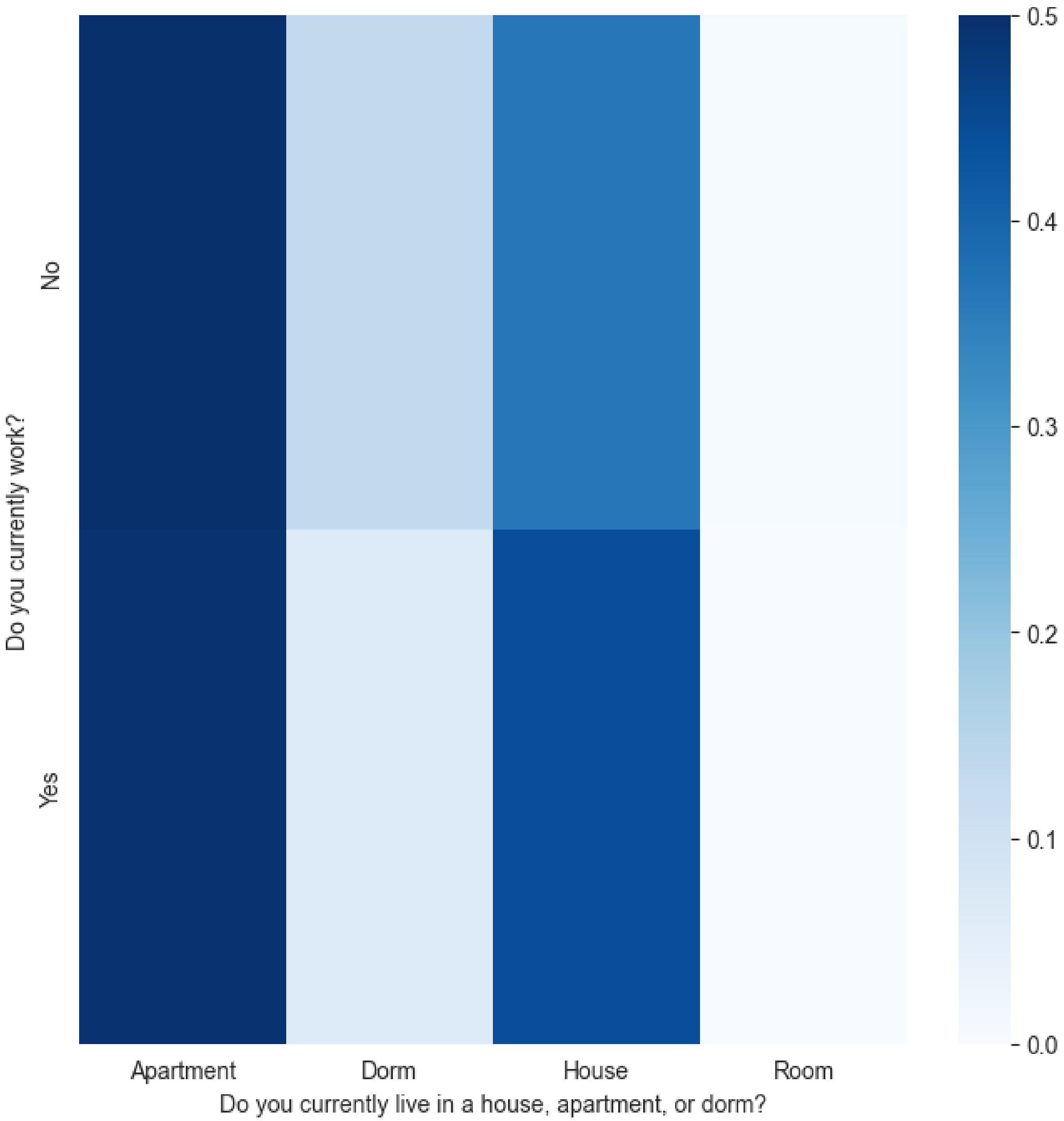
The majority of student respondents (70.4%) do **not** work while attending school.

In [6]:

```
df_2dhist = pd.crosstab(df.loc[:, 'Do you currently work?'],
                        df.loc[:, 'Do you currently live in a house, apartme
                        normalize='index')

# Plot heatmap
plt.subplots(figsize=(8, 8))
sns.heatmap(df_2dhist, cmap="Blues")
plt.xlabel('Do you currently live in a house, apartment, or dorm?')
_ = plt.ylabel('Do you currently work?')
df_2dhist
```

Out[6]:

Do you currently live in a house, apartment, or dorm?	**Apartment**	**Dorm**	**House**	**Room**
Do you currently work?				
No	0.500000	0.131868	0.362637	0.005495
Yes	0.493506	0.064935	0.441558	0.000000

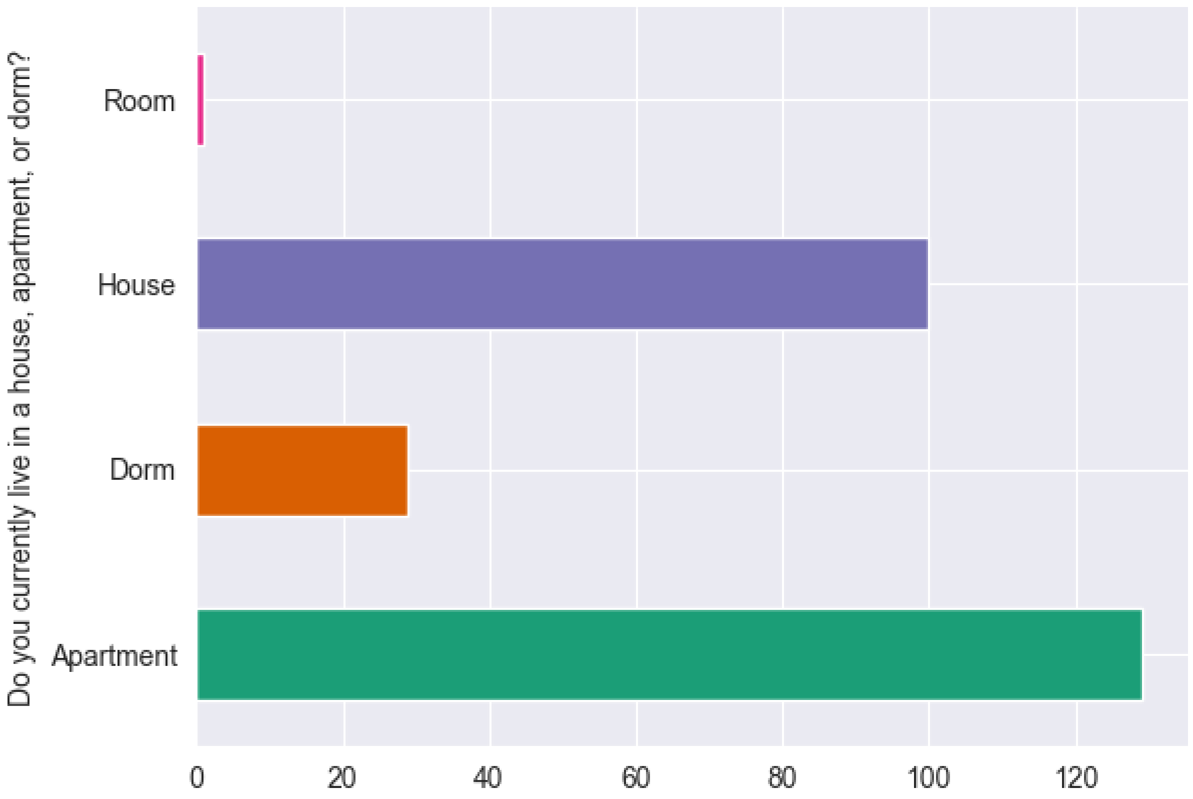


For both working & non-working participants, the proportion who live in an apartment are equivalent (50%).

However, 13% of non-working participants live in a dorm while only 6% of working participants live in a dorm. This 7% drop is matched in participants who live a house, with 44% of working participants living in a house compared to 36% of non-working participants.

This indicates that working participants tend to live off-campus and in living situations that have a higher cost of living.

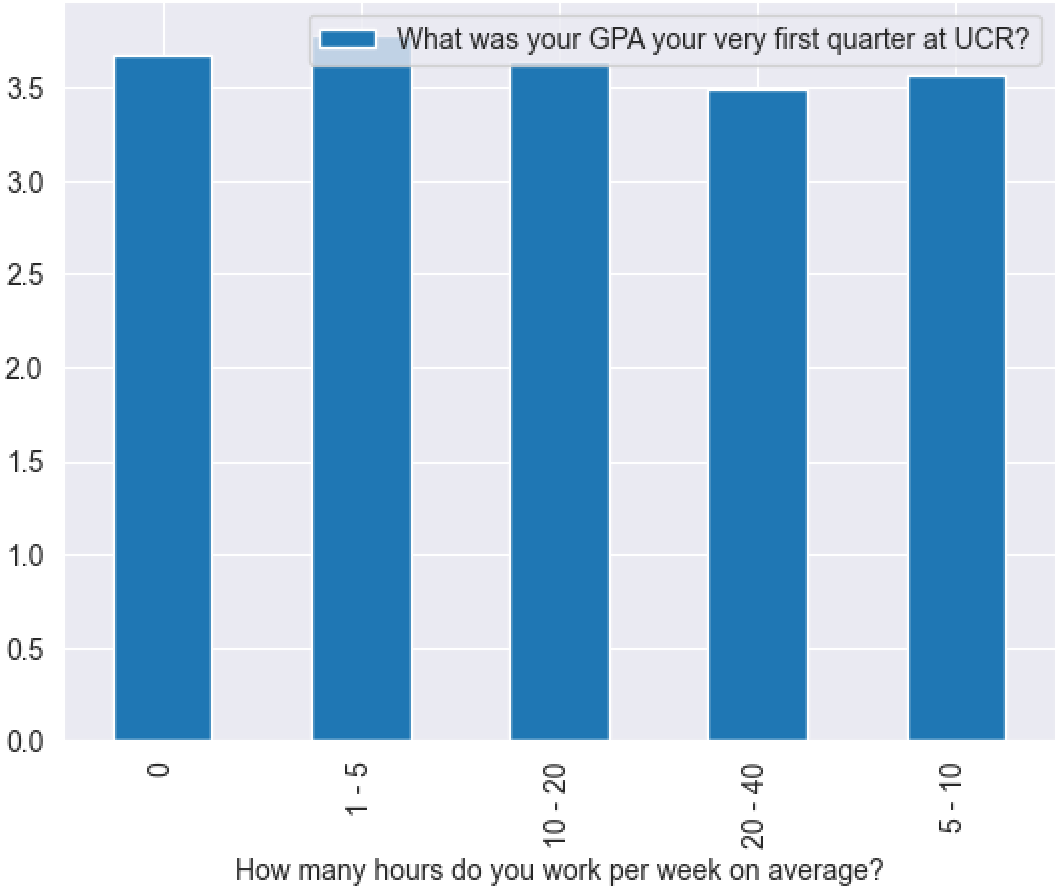
```
In [7]: df.groupby('Do you currently live in a house, apartment, or dorm?').size().p


        plt.gca().spines[['top', 'right', ]].set_visible(False)
```



Most participants live in either an Apartment or a House. This would indicate that most students either live off-campus or on-campus apartments.

```
In [8]: dataTable1 = pd.pivot_table(data=df, values='What was your GPA your very fir
                                    index='How many hours do you work per week on av
        _ = dataTable1.plot(kind='bar')
        print("Total Average GPA: ", df['What was your GPA your very first quarter a
```

```
Total Average GPA:  3.6520247933884296
```



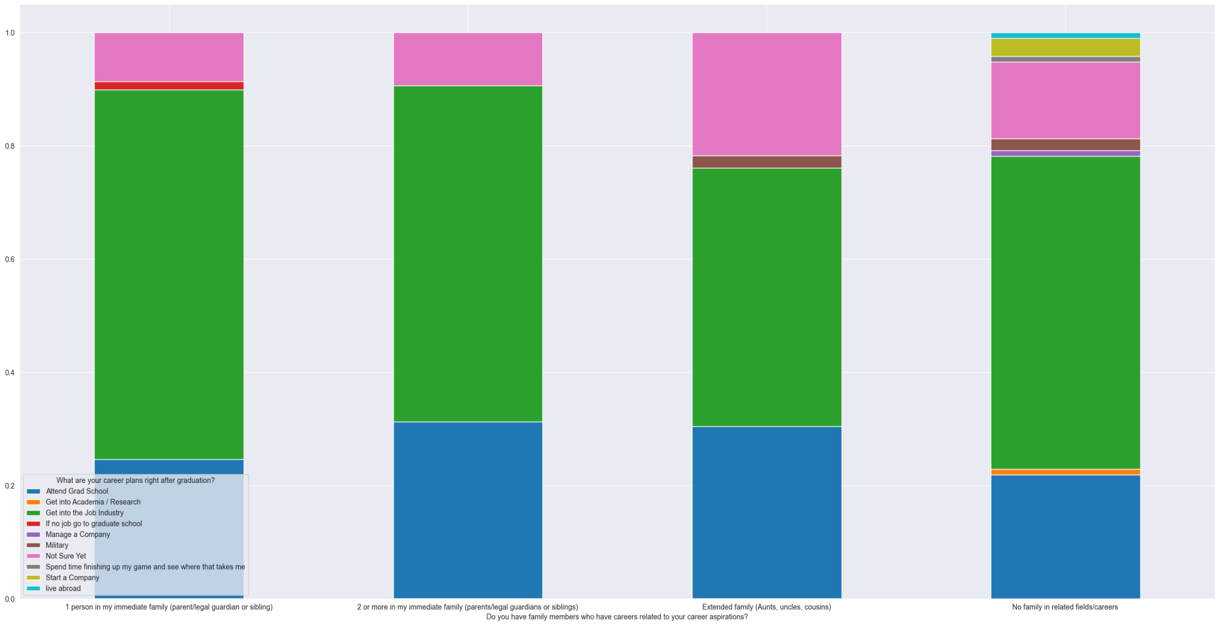
The average GPA seems to be independent in respect to working hours per week. Most students who work less than 20 hours have an equivalent average GPA to the total average GPA of all participants (3.65).

There is a small drop in GPA associated with students who work more than 20 hours (3.5 GPA), which may mean some of those students may struggle maintaining balance between work and school.

This would indicate that most students seem to be able to balance work with school. However, it would also indicate that students who work full-time jobs may struggle slightly in school.

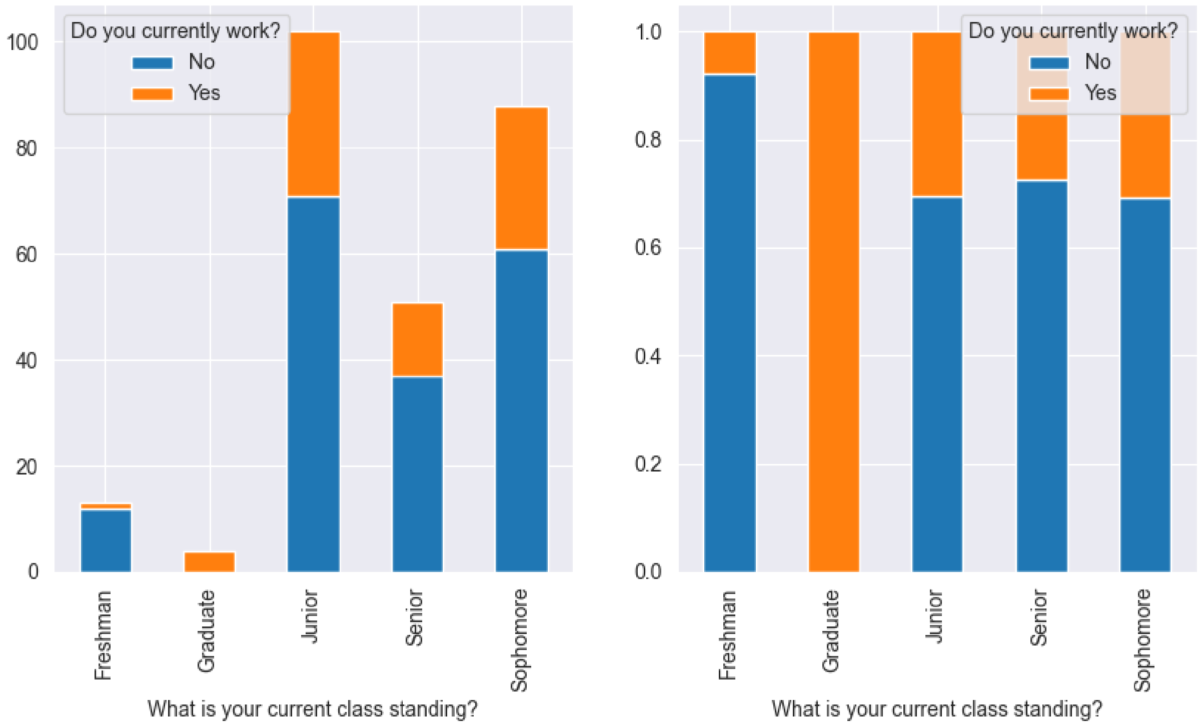
In [9]:

```
stkbar_df = df.dropna()
_ = pd.crosstab(
    df['Do you have family members who have careers related to your career a
    df['What are your career plans right after graduation?'],
    normalize='index'
).plot(kind="bar", stacked=True, rot=0, figsize=(30, 15))
```



Most students across all groups are looking to "get into the job industry". Across all groups, the proportions of "Attending Grad School" and "Get into the job industry" are similar, except for students who have extended family in their career. They are more unsure about their future compared to other groups. Also, students with no family in their field are more diversified in their career plans.

```
In [10]: fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(10, 5))

_ = pd.crosstab(
    df['What is your current class standing?'],
    df['Do you currently work?'],
).plot(kind='bar', stacked=True, ax=axes[0])
_ = pd.crosstab(
    df['What is your current class standing?'],
    df['Do you currently work?'],
    normalize='index'
).plot(kind='bar', stacked=True, ax=axes[1])
```

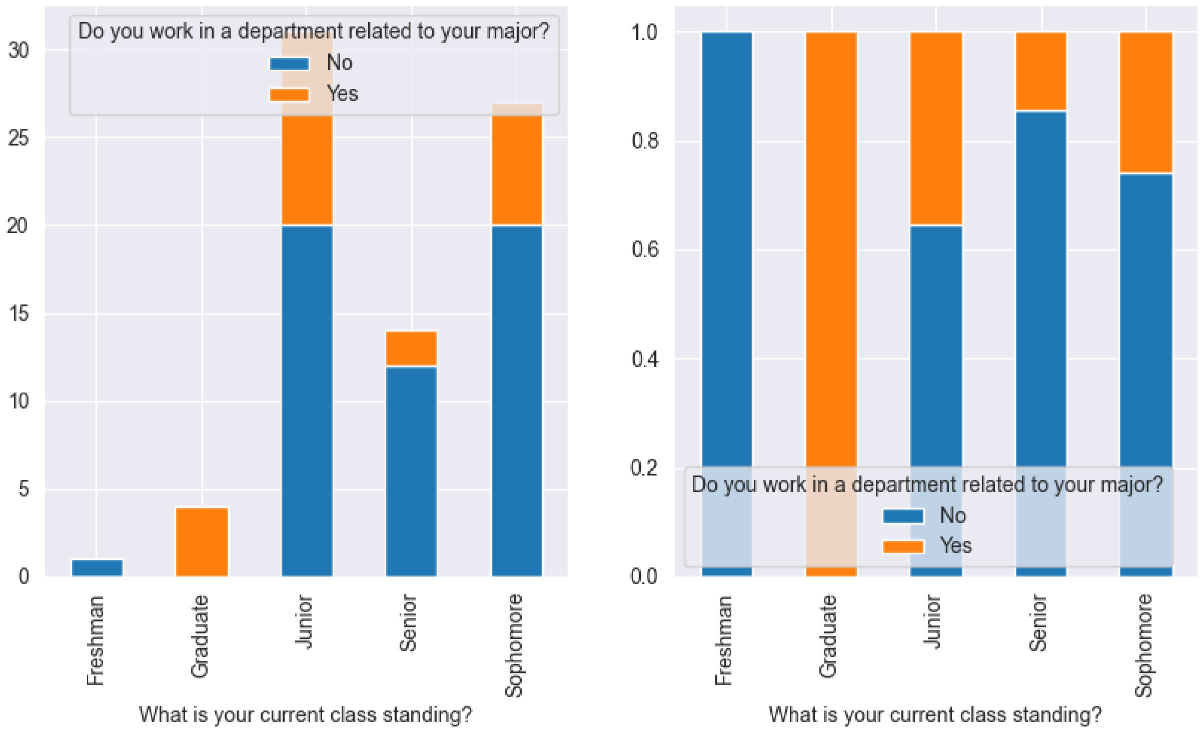


The class standing most likely to work are graduate students, where 100% of participants work. The freshman class is the least likely to work, where 92% of participants do not work.

For Sophomore, Junior, and Senior participants, all 3 groups have similar proportions working with 30% of participants working.

In [11]:

```
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(10, 5))

_ = pd.crosstab(
    w_df['What is your current class standing?'],
    w_df['Do you work in a department related to your major?'],
).plot(kind='bar', stacked=True, ax=axes[0])

_ = pd.crosstab(
    w_df['What is your current class standing?'],
    w_df['Do you work in a department related to your major?'],
    normalize='index',
).plot(kind='bar', stacked=True, ax=axes[1])
```



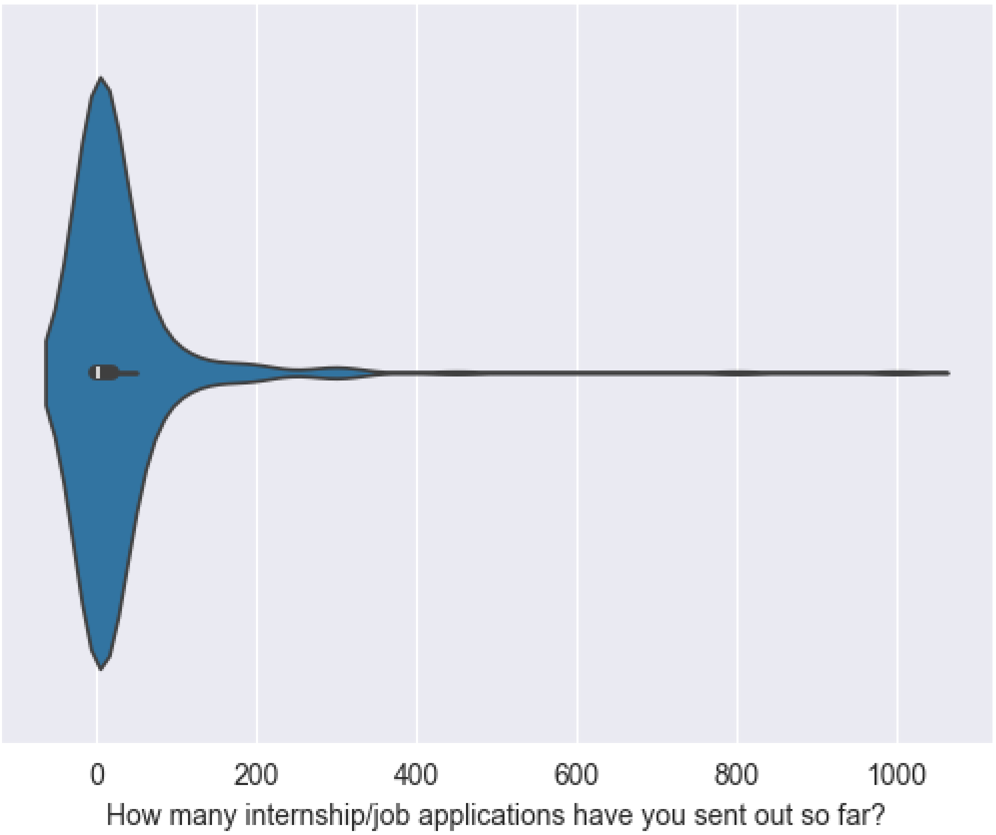
Of the students who responded "yes" to currently working, the above graphs show the proportions of participants who work in a department related to their major. Most students do not work in a department related to their major, indicating that they are working for money rather than job experience. his holds true for all groups except for Graduate students, who all work in a department of their major.

Interestingly, Juniors have a higher rate of working in their major, perhaps indicating internships or students seeking to gain major-related work experience.

In [12]:

```
print("mean internships: ", df["How many internship/job applications have yo
print("median internships: ", df["How many internship/job applications have 
_ = sns.violinplot(x=df["How many internship/job applications have you sent
```

```
mean internships:  31.161538461538463
median internships:  2.0
```



From the violin plot, we can see that few students send out many job applications. Some students have sent out many (up to 1000) applications while most send about 2. The distribution is skewed right with many outliers who have sent out significantly more applications than most students.

Hypotheses

Hypothesis 1: There will be a correlation between whether people live with family, friends, or neither and whether or not they work

Null Hypothesis: There is no relationship between people who live with family, friends, or neither and whether or not they work.

Significance value: 0.1 Degrees of freedom: 3

In [13]:

```
hyp3_major_table = pd.crosstab(df.iloc[:, 3], df.iloc[:, 8], margins=True, m
hyp3_major_table
```

Out[13]:

Do you currently work?	No	Yes	Total
Who do you live with?			
Both	22	5	27
Family	61	33	94
Friends	57	27	84
Neither	42	12	54
Total	182	77	259

In [14]:

```
num_rows, num_cols = hyp3_major_table.shape
# Initialize expected frequencies
expected_frequencies = []
chi_squared = 0
for i in range(num_rows - 1):
    row_totals = hyp3_major_table.iloc[i, -1]
    for j in range(num_cols - 1):
        col_totals = hyp3_major_table.iloc[-1, j]
        expected_frequency = (row_totals * col_totals) / hyp3_major_table.il
        expected_frequencies.append(expected_frequency)
        chi_squared += ((hyp3_major_table.iloc[i, j] - expected_frequency) *

print("Chi-squared value:", chi_squared)
```

```
Chi-squared value: 4.616203438011947
```

With a significance value of 0.1 and 3 degrees of freedom, chi-squared must be greater than 6.25. Since chi-squared of `4.61 < 6.25`, we accept the null hypothesis:

There is no relationship between people who live with family, friends, or neither and whether or not they work.

Hypothesis 2: Students who live on-campus are more likely to have roommates of the same major.

Null Hypothesis: There is no relationship between students who live on-campus and students who have roommates of the same major.

Significance value: 0.1 Degrees of Freedom: 2

In [15]:

```
roommates_major_table = pd.crosstab(df.iloc[:, 4], df.iloc[:, 11], margins=T
roommates_major_table
```

Out[15]:

Do you work in a department related to your major?	No	Yes	Total
Do you currently live in a house, apartment, or dorm?			
Apartment	22	16	38
Dorm	4	1	5
House	27	7	34
Total	53	24	77

In [16]:

```
num_rows, num_cols = roommates_major_table.shape
# Initialize expected frequencies
expected_frequencies = []
chi_squared = 0
for i in range(num_rows - 1):
    row_totals = roommates_major_table.iloc[i, -1]
    for j in range(num_cols - 1):
        col_totals = roommates_major_table.iloc[-1, j]
        expected_frequency = (row_totals * col_totals) / roommates_major_tab
        expected_frequencies.append(expected_frequency)
        chi_squared += ((roommates_major_table.iloc[i, j] - expected_frequer

print("Chi-squared value:", chi_squared)
```

```
Chi-squared value: 4.183390044200403
```

With a significance value of 0.1 and 2 degrees of freedom, chi-squared must be greater than 4.61. Since chi-squared of `4.18 < 4.61` , we accept the null hypothesis:

There is no relationship between students who live on-campus and students who have roommates of the same major.

Hypothesis 3: People who live with more people will have a higher GPA on average.

In [17]:

```
hyp3_major_table = pd.crosstab(df.iloc[:, 5], df.iloc[:, 6], margins=True, m
average_household_size = df.iloc[:, 5].mean(skipna=True)
average_gpa = df.iloc[:, 6].mean(skipna=True)

print("Average Household Size:", average_household_size)
print("Average GPA:", average_gpa)
numerator = 0
denom_x = 0
denom_y = 0
for i in range(260):
    x_i = df.iloc[i, 5]
    y_i = df.iloc[i, 6]
    if not pd.isna(x_i) and not pd.isna(y_i):  # Check for NaN values
        numerator += (x_i - average_household_size) * (y_i - average_gpa)
        denom_x += (x_i - average_household_size) ** 2
        denom_y += (y_i - average_gpa) ** 2
```

```
# Calculate Pearson correlation coefficient
pearson_coefficient = (numerator / ((denom_x * denom_y) ** 0.5))
print("Pearson Correlation Coefficient:", pearson_coefficient)
```

```
Average Household Size: 3.826923076923077
Average GPA: 3.6520247933884296
Pearson Correlation Coefficient: -0.2010052294084673
```

With a Pearson Correlation Coefficient of -0.2, there is a slight negative correlation between household size and average GPA. Students who live alone or with fewer people perform slightly better than those with more roommates.

This goes against our hypothesis that people living with more people will have a higher GPA.

Conclusion

In wanting to figure out in general how various aspects of a student's home environment go on to affect their employment and school performance, we performed 3 different tests: Comparing what kind of people participants lived with and whether or not they work, students who work on campus and their roommates majors, and the number of people participants lived with and the participant's GPA. However, using chi-squared tests and pearson correlation, we discovered that none of our hypotheses had any correlation with it. But from here we can delve deeper into our hypothesis, for example why did working not affect student's GPA's in the first quarter? What are other potential factors as to why that was the case? In addition, Freshmen were found to not work at major related jobs, whereas juniors were the most likely to work at a major related job.

Even though none of our hypotheses were proven to be true, we still learned a lot about the data given to us and we can use it to further more questions and assumptions later down the line.