Shreya Lab 5/6

What information do you have?

In our surevy we asked aboout the previous and future living situationns of UCR students before and after the pandemic. We asked for:
~ where they lived BEFORE covid (On campus apartment, Off campus apartment, Greek House, UCR Dorms)
~ where they want to live AFTER covid (On campus apartment, Off campus apartment, Greek House, UCR Dorms)
~ where they would liive if finances were not a burden (on campus/ off campus)
~ whether they prefer in person, remote learning or both

What would you like to know about the class?

We would like to know if there was a shift in preferences on housing before and after the pandemic. Due to the pandemic many people had to leave and we wanted to know if their experiences during that time of leaving and being away has affected their thought on where they want to stay. When campus closed last march many student had to leave and almost all students living on campus recieved a refund for the portion of payment left in the remaining year however off campus students were not so lucky. We also want to know if more students would live on campus/off campus if finances were not an issue and how that aligns with where students are actually living. Furthermore, we want to see if there is a correlation between where students live and their gender such as are female student more likely to live closer to campus? We also want to know if people's preferences on learning will determine where they live.

Explore the data.

In [1]:

```
import pandas as pd
df = pd.read_csv("Responses Clean.csv")
df.head()
df["Housing BC"] = df["Housing BC"].str.replace('with parents', 'Resident')
df["Housing BC"] = df["Housing BC"].str.replace('With parents', 'Resident')
df["Housing BC"] = df["Housing BC"].str.replace('off-campus, Parents house'
df["Housing BC"] = df["Housing BC"].str.replace('at home', 'Resident')
df["Housing BC"] = df["Housing BC"].str.replace('home', 'Resident')
df["Housing BC"] = df["Housing BC"].str.replace('I lived parents', 'Residen
df["Housing BC"] = df["Housing BC"].str.replace('Off-campus parents', 'Resi
df["Housing BC"] = df["Housing BC"].str.replace('I lived resident', 'Reside
df["Housing BC"] = df["Housing BC"].str.replace('I lived Resident', 'Reside
df["Housing BC"] = df["Housing BC"].str.replace('Local Resident prior', 'Re
df["Housing BC"] = df["Housing BC"].str.replace("Crashed at friends' couche
df["Housing BC"] = df["Housing BC"].str.replace("Same Resident", 'Resident'

df["Housing AC"] = df["Housing AC"].str.replace('with parents', 'Resident')
df["Housing AC"] = df["Housing AC"].str.replace('With parents', 'Resident')
df["Housing AC"] = df["Housing AC"].str.replace('off-campus, Parents house'
df["Housing AC"] = df["Housing AC"].str.replace('at home', 'Resident')
df["Housing AC"] = df["Housing AC"].str.replace('home', 'Resident')
df["Housing AC"] = df["Housing AC"].str.replace('I lived parents', 'Residen
df["Housing AC"] = df["Housing AC"].str.replace('Off-campus parents', 'Resi
df["Housing AC"] = df["Housing AC"].str.replace('I lived resident', 'Reside
df["Housing AC"] = df["Housing AC"].str.replace('I lived Resident', 'Reside
df["Housing AC"] = df["Housing AC"].str.replace('Local Resident prior', 'Re
df["Housing AC"] = df["Housing AC"].str.replace("Crashed at friends' couche
df["Housing AC"] = df["Housing AC"].str.replace("Same Resident", 'Resident'

# df["Housing AC"] = df["Housing AC"].str.replace('Off-campus Resident', 'H
# df["Housing AC"] = df["Housing AC"].str.replace('Home', 'House')
# df["Housing BC"] = df["Housing BC"].str.replace('Off-campus Resident', 'H
# df["Housing BC"] = df["Housing BC"].str.replace('Home', 'House')
# df["Housing AC"] = df["Housing AC"].str.replace("Off campus - House(Not g
# df["Housing BC"] = df["Housing BC"].str.replace("Off campus - House(Not g

comparable_students = df[df['UCR BC'] == "Yes"]
comparable_students.head()
```

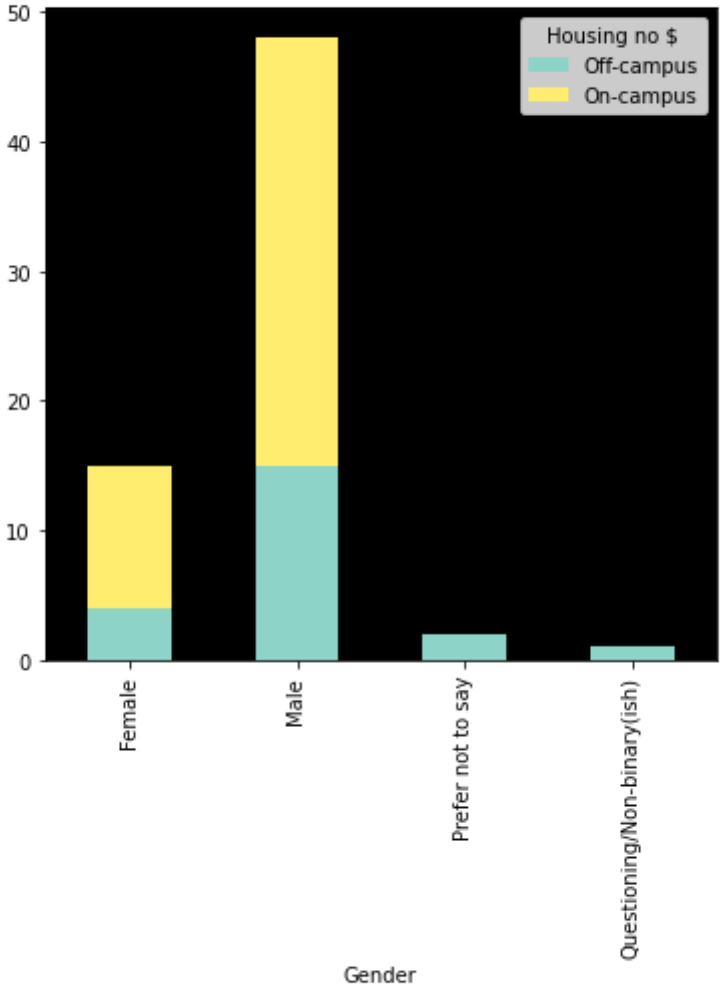
Out[1]:

	Timestamp	Gender	Ethnicity	Academic Year	Major	GPA	UCR BC	Housing BC	Housing no $	Housing AC
0	1/28/2021 10:53:46	Female	White	Sophomore	CS	3.7 - 4.0	Yes	UCR Dorms	On-campus	On Campus Apartment
1	1/30/2021 12:38:43	Male	Asian	Sophomore	Computer Science with Business Applications	3.3 - 3.69	Yes	UCR Dorms	Off-campus	Off - Campus Apartment
2	1/30/2021 12:41:25	Male	Asian	Sophomore	CSBA	3.7 - 4.0	Yes	UCR Dorms	On-campus	On Campus Apartment

	Timestamp	Gender	Ethnicity	Academic Year	Major	GPA	UCR BC	Housing BC	Housing no $	Housing AC
3	1/30/2021 12:46:31	Male	Asian	Sophomore	Computer Science	3.7 - 4.0	Yes	UCR Dorms	On-campus	On Campus Apartment
4	1/30/2021 12:51:38	Female	Asian	Sophomore	Computer Science	3.3 - 3.69	Yes	UCR Dorms	On-campus	On Campus Apartment

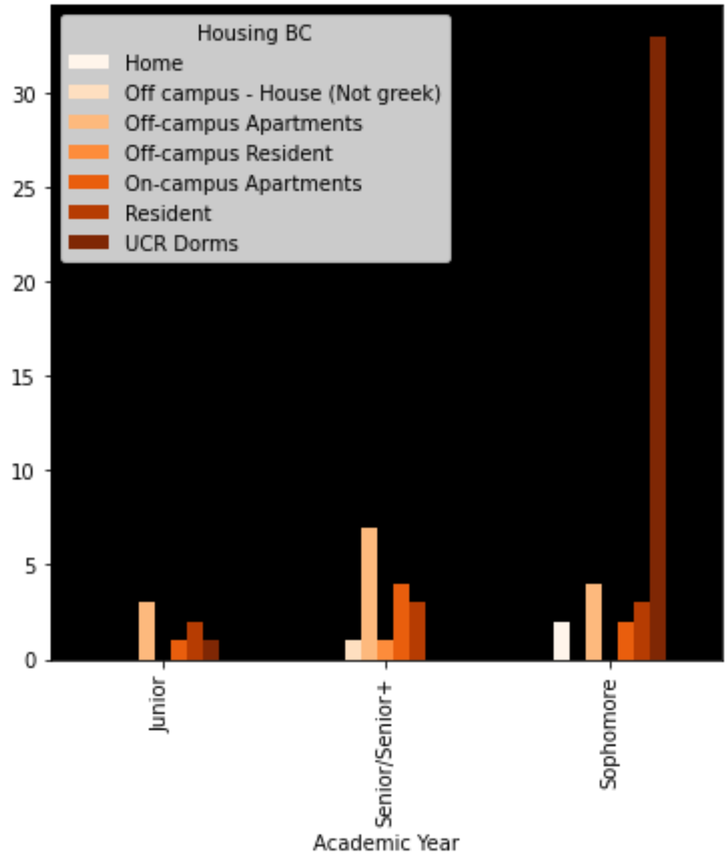
In [2]:

```
#Clean Data
import matplotlib.pyplot as plt
df.head(80)
#Diagram 1 - Gender vs Housing without financial burder
gender_housing = pd.crosstab(comparable_students["Gender"], comparable_stud
gender_housing_plot = gender_housing.plot.bar(stacked=True, colormap='Set3'
gender_housing_plot.set_facecolor('Black')
```

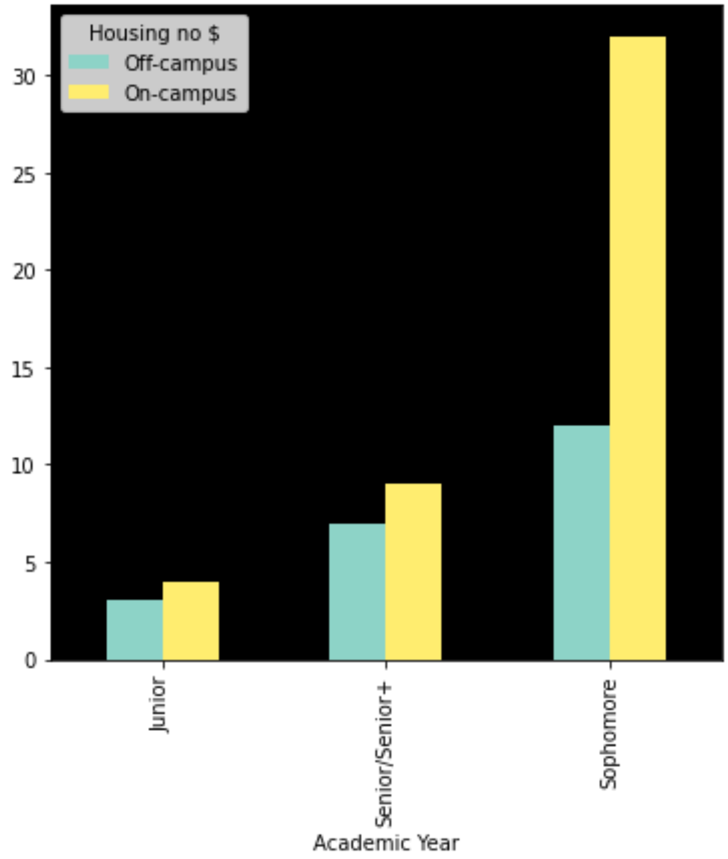


In [3]:

```
#Diagram 2 - academic year vs housing BC
academic_year_housingBC = pd.crosstab(comparable_students["Academic Year"],
academic_year_housingBC_plot = academic_year_housingBC.plot.bar(colormap='O
academic_year_housingBC_plot.set_facecolor('Black')
```

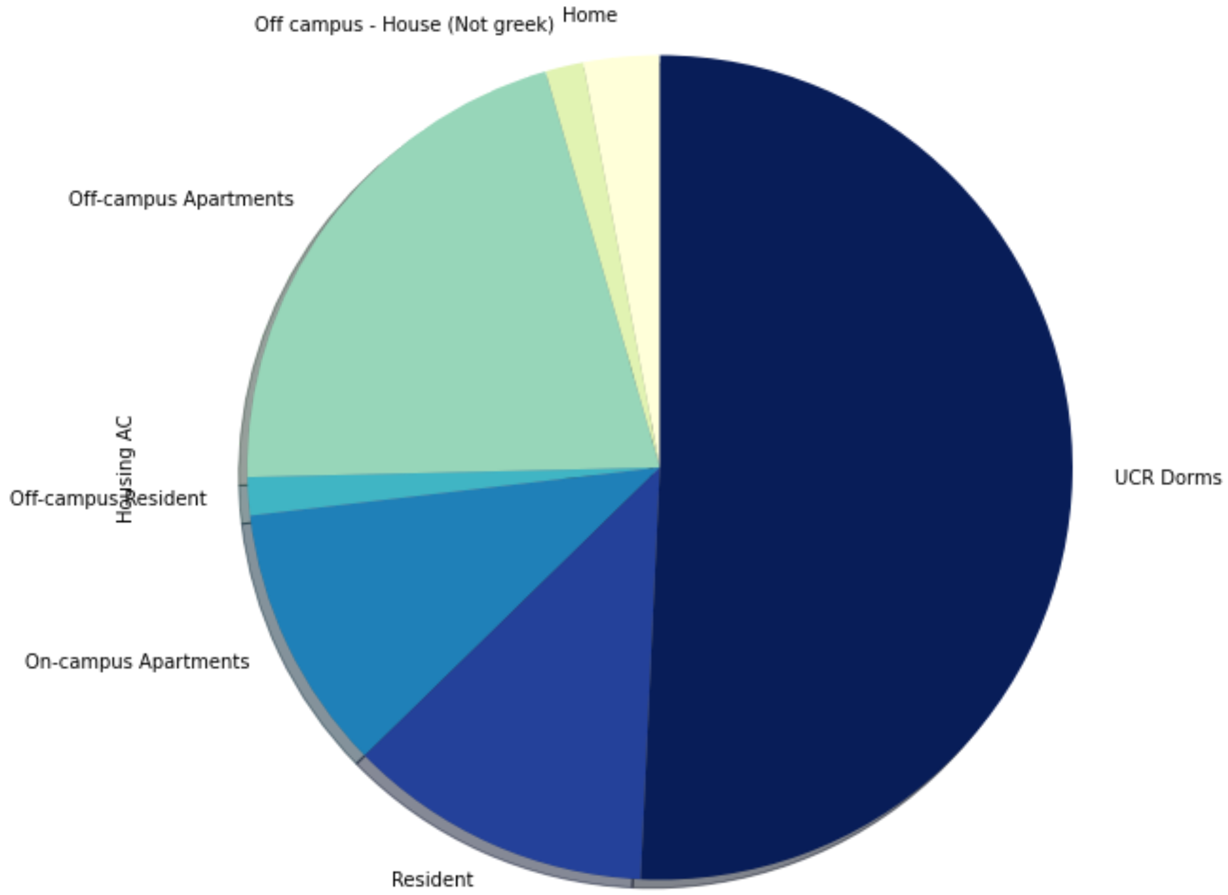


In [4]:
```
#Diagram 3 - academic year vs housing AC
academic_year_housingAC = pd.crosstab(comparable_students["Academic Year"],
academic_year_housingAC_plot = academic_year_housingAC.plot.bar(colormap='S
academic_year_housingAC_plot.set_facecolor('Black')
```

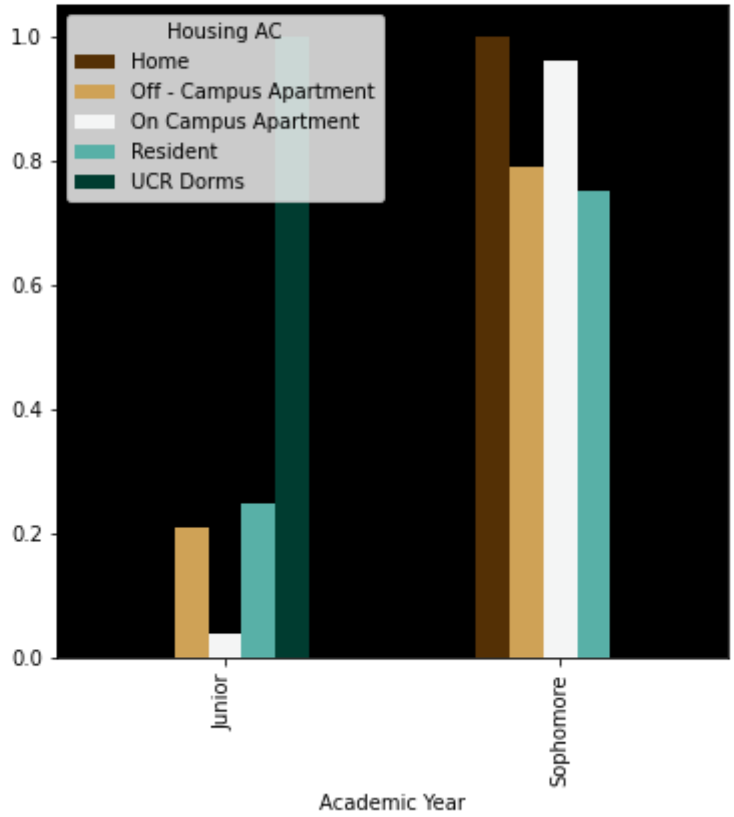


```
In [5]:  #Diagram 4 - total count for housing
         gender_housing = comparable_students.groupby("Housing BC")["Housing AC"].co
         gender_housing.plot(kind='pie', subplots=True, shadow = True,startangle=90,

Out[5]:  array([<AxesSubplot:ylabel='Housing AC'>], dtype=object)
```

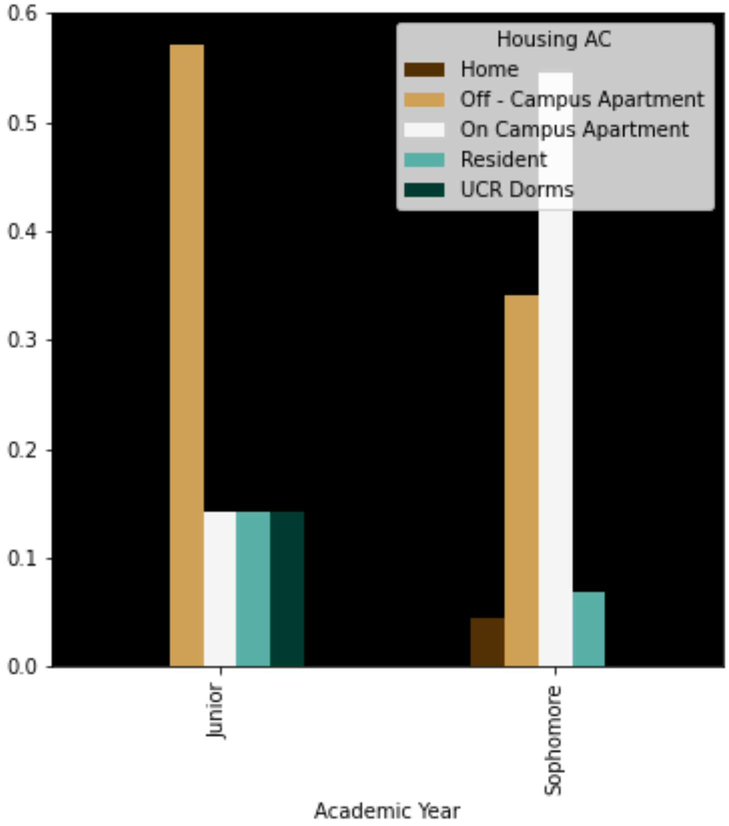


In [6]:
```
sophomore_comparable_students = comparable_students[comparable_students['Ac
sophomore_comparable_students.head()
junior_comparable_students = comparable_students[comparable_students['Acade
junior_comparable_students.head()
sophomore_junior_students = pd.concat([sophomore_comparable_students, junio
#sophomore_junior_students.head(80)
crosstab_AC = pd.crosstab(sophomore_junior_students["Academic Year"], sopho
            normalize=True)
crosstab_AC
#P(Academic year|AC)
AC_counts = crosstab_AC.sum(axis=0)
Academic_year_given_AC = crosstab_AC.divide(AC_counts, axis=1)
Academic_year_given_AC
stacked_Academic_year_given_AC = Academic_year_given_AC.plot.bar(colormap='
stacked_Academic_year_given_AC.set_facecolor('Black')
```



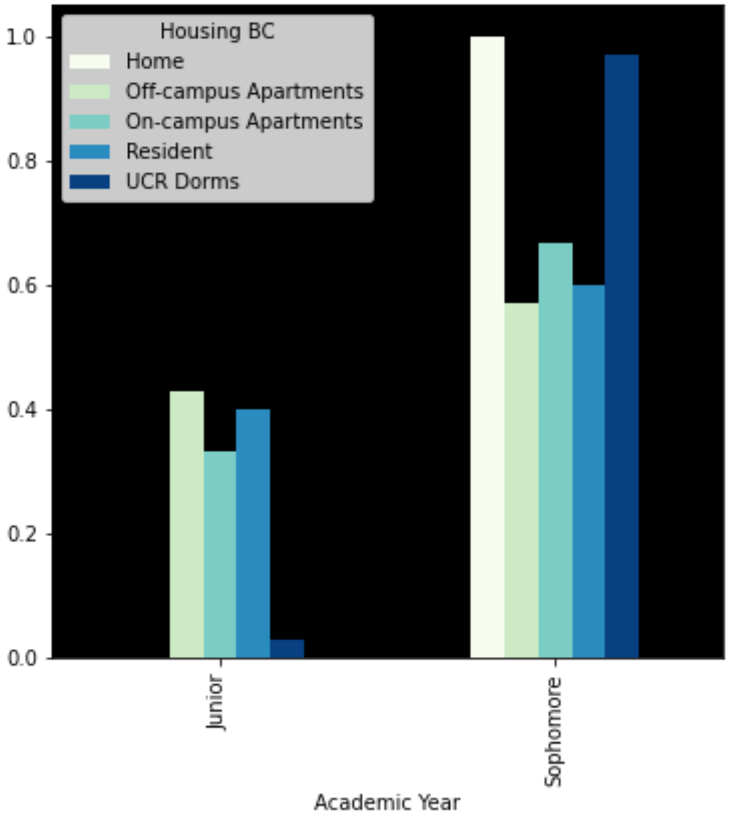
In [7]:

```
#P(AC|Academic year)
AC_counts = crosstab_AC.sum(axis=1)
AC_given_academic_year = crosstab_AC.divide(AC_counts, axis=0)
AC_given_academic_year
stacked_AC_given_academic_year = AC_given_academic_year.plot.bar(colormap='
stacked_AC_given_academic_year.set_facecolor('Black')
```

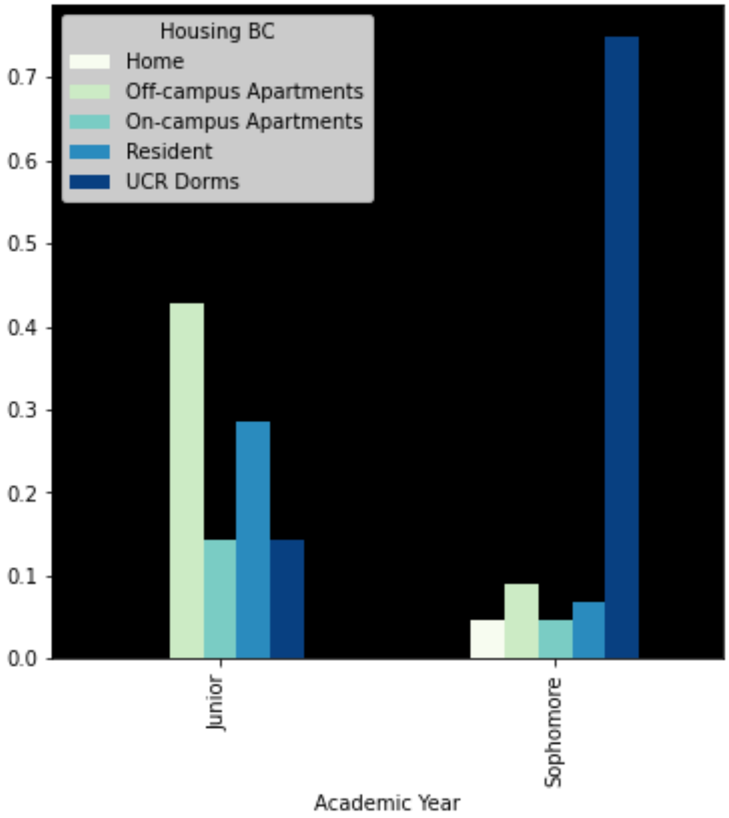


In [8]:

```
crosstab_BC = pd.crosstab(sophomore_junior_students["Academic Year"], sopho
                          normalize=True)
crosstab_BC
#P(Academic year|BC)
BC_counts = crosstab_BC.sum(axis=0)
Academic_year_given_BC = crosstab_BC.divide(BC_counts, axis=1)
Academic_year_given_BC
stacked_Academic_year_given_BC = Academic_year_given_BC.plot.bar(colormap='
stacked_Academic_year_given_BC.set_facecolor('Black')
```



In [9]:
```
#P(BC|Academic year)
BC_counts = crosstab_BC.sum(axis=1)
BC_given_academic_year = crosstab_BC.divide(BC_counts, axis=0)
BC_given_academic_year
stacked_BC_given_academic_year = BC_given_academic_year.plot.bar(colormap='
stacked_BC_given_academic_year.set_facecolor('Black')
```



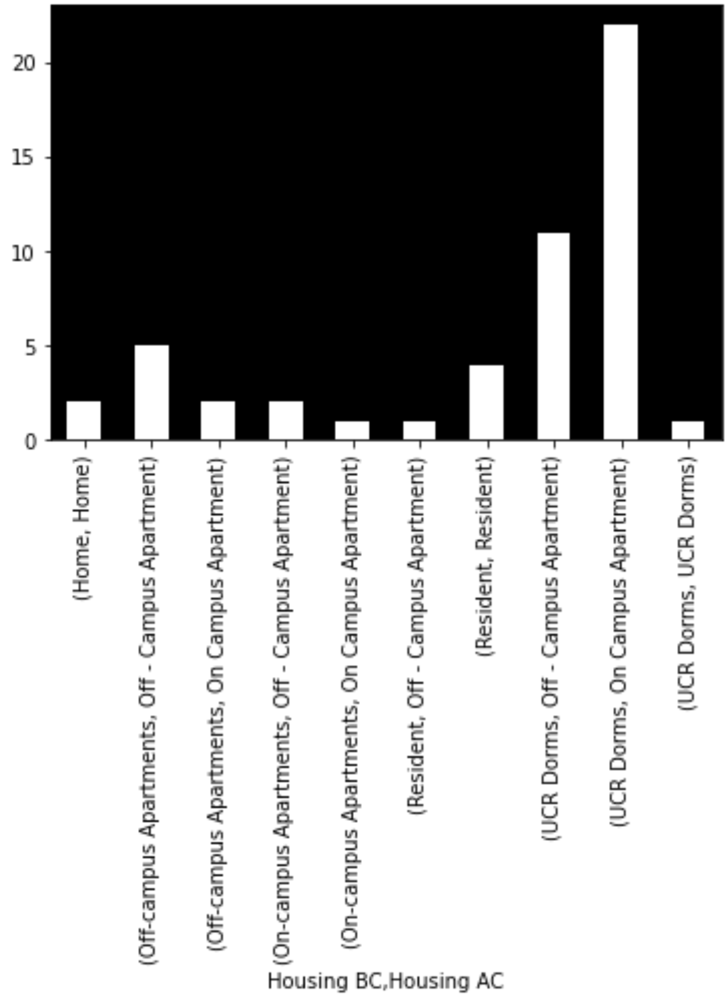
I believe that due to the way covid was handled by school related housing and regular housing people may change where they want to live when they go back. When people had to leave many students were unable to get refunds or their deposit back from off campus residencies. Many places faced legal action and through the consequences of that I think that people may be more favorable towards on-campus housing. To get housing data from pre-covid and after covid I have decided to just look at sophomores and juniors, basically any students that will be there after covid went to Riverside before Covid.

In [10]:
```
sophomore_junior_students[['Housing BC','Housing AC']].describe()
```

Out[10]:

	Housing BC	Housing AC
count	51	51
unique	5	5
top	UCR Dorms	On Campus Apartment
freq	34	25

```
In [11]: housing_change = sophomore_junior_students.groupby(["Housing BC","Housing A
         housing_change_plot = housing_change.plot.bar(stacked=True, color ='White')
         housing_change_plot.set_facecolor('Black')
```



Can you state any hypotheses or make any predictions? Which tests can you apply to verify your hypothesis?

Hypothisis: Gender highly impacts where people live especially among college students. I believe that women are more likely to choose housing that is closer to campus and may choose to live on campus more than off campus. To get housing data from pre-covid and after covid I have decided to just look at sophomores and juniors, basically any students that will be there after covid went to Riverside before Covid.

Null Hypothesis: There is no relationship between gender and on campus vs off campus housing.

I decided to use chi square to determine how much gender really affect housing.

In [12]:

```
from scipy import stats
from scipy.stats import chi2_contingency
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from scipy.stats import pearsonr
from sklearn import preprocessing
import seaborn as sns
```

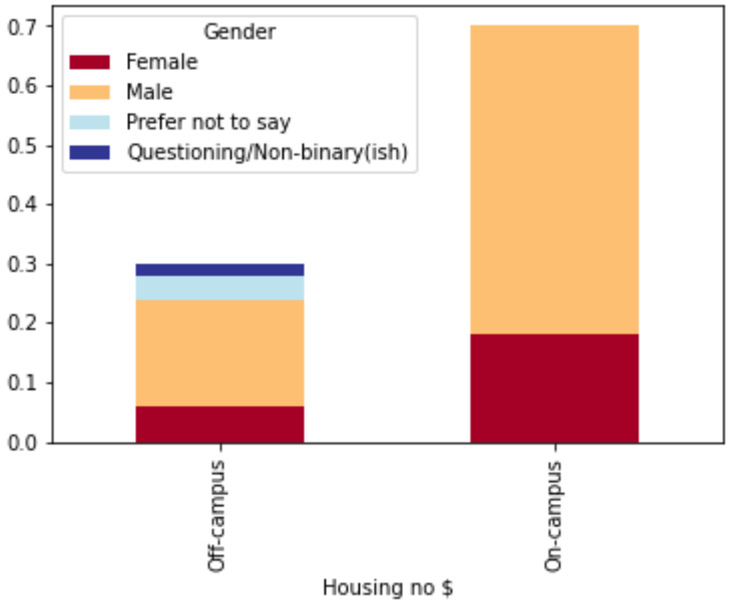
In [13]:

```
gender_housing = sophomore_junior_students.groupby(["Housing no $", "Gender
gender_housing

gender_housing_avg = pd.crosstab(sophomore_junior_students["Housing no $"],
            normalize=True)
gender_housing_avg.plot.bar(stacked=True, colormap='RdYlBu')
gender_housing_avg
# This diagram displays the number of people who would live on-campus/off-
```

Out[13]:

Gender	Female	Male	Prefer not to say	Questioning/Non-binary(ish)
Housing no $				
Off-campus	0.06	0.18	0.04	0.02
On-campus	0.18	0.52	0.00	0.00



In [14]:

```
# Chi-Square Test
chi, p, dof, expected = chi2_contingency(gender_housing_avg)
print(f'Chi-value : {chi}')
print(f'P-Value : {p}')
print(f'Degree of freedom : {dof}')
print('Expectation : ')
#print(expected)
expected = pd.DataFrame(
    expected,
    index=["Off-campus","On-campus"],
    columns=["Female","Male","Prefer not to say","Questtioning/Non-binary(i
expected
```

```
Chi-value : 0.1489795918367347
P-Value : 0.9853721171526216
Degree of freedom : 3
Expectation :
```

Out[14]:

	Female	Male	Prefer not to say	Questtioning/Non-binary(ish)
Off-campus	0.072	0.21	0.012	0.006
On-campus	0.168	0.49	0.028	0.014

□ T-Test

- the difference of two sample means divided by standard of error
- relevant equations: degrees of freedom
- assumptions:
  - random samples
  - independent observation
  - population variances are equal

$$t = \frac{(x_1 - x_2) - (u_1 - u_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Null Hypothesis (H_0): $\mu_f - \mu_m = 0$ → Women aren't more likely to live on campus.

Alt. Hypothesis (H_a) = $\mu_f - \mu_m > 0$ → Women are more likely to live on campus.

□ Chi Square

Null Hypothesis (H_0): $\mu_f - \mu_m = 0$ → Women aren't more likely to live on campus.

Alt. Hypothesis (H_a) = $\mu_f - \mu_m > 0$ → Women are more likely to live on campus.

→ Toy Representation

Housing No $	Female	Male	Prefer Not To Say	Non-binary
Off-campus	0.06	0.18	0.04	0.02
On-campus	0.18	0.52	0	0

→ Let's say Population = 100 & Expected = $\frac{\text{row total} \times \text{column total}}{\text{overall total}} = \frac{x \cdot y}{100}$

Housing No $	Female	Male	Prefer Not To Say	Non-binary	Total
Off-campus	6 (7.2)	18 (21)	4 (1.2)	2 (0.6)	30
On-campus	18 (16.8)	52 (49)	0 (2.8)	0 (1.4)	70
Total	24	70	4	2	100

→ $x^2 = \Sigma_i \frac{(\text{observed val} - \text{expected val})^2}{\text{expected value}}$

$$x^2 = \frac{1.44}{7.2} + \frac{9}{21} + \frac{7.84}{1.2} + \frac{1.96}{0.6} + \frac{1.44}{16.8} + \frac{9}{42} + \frac{7.84}{2.8} + \frac{1.96}{1.4}$$

$= 0.2 + 0.429 + 6.533 + 3.267 + 0.086 + 0.214 + 2.8 + 1.4$

$= 14.929$

→ determine degree of freedom: 3

→ calculate p value to reject / accept hypothesis

□ Pearson

- measures relationship between continuous variables
- correlation as direction of relationship