exploratory_data_analysis

June 22, 2021

[1]:
```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.decomposition import PCA
from sklearn.preprocessing import scale
```

[2]:
```
ski_data = pd.read_csv('ski_data_cleaned.csv')
```

[3]:
```
ski_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 277 entries, 0 to 276
Data columns (total 25 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Name               277 non-null    object 
 1   Region             277 non-null    object 
 2   state              277 non-null    object 
 3   summit_elev        277 non-null    int64  
 4   vertical_drop      277 non-null    int64  
 5   base_elev          277 non-null    int64  
 6   trams              277 non-null    int64  
 7   fastSixes          277 non-null    int64  
 8   fastQuads          277 non-null    int64  
 9   quad               277 non-null    int64  
 10  triple             277 non-null    int64  
 11  double             277 non-null    int64  
 12  surface            277 non-null    int64  
 13  total_chairs       277 non-null    int64  
 14  Runs               274 non-null    float64
 15  TerrainParks       233 non-null    float64
 16  LongestRun_mi      272 non-null    float64
 17  SkiableTerrain_ac  275 non-null    float64
 18  Snow Making_ac     240 non-null    float64
 19  daysOpenLastYear   233 non-null    float64
 20  yearsOpen          277 non-null    float64
```

```
 21  averageSnowfall    268 non-null    float64
 22  AdultWeekend       277 non-null    float64
 23  projectedDaysOpen  236 non-null    float64
 24  NightSkiing_ac     163 non-null    float64
dtypes: float64(11), int64(11), object(3)
memory usage: 54.2+ KB
```

[4]:
```
ski_data.head()
```

[4]:

	Name	Region	state	summit_elev	vertical_drop	\
0	Alyeska Resort	Alaska	Alaska	3939	2500	
1	Eaglecrest Ski Area	Alaska	Alaska	2600	1540	
2	Hilltop Ski Area	Alaska	Alaska	2090	294	
3	Arizona Snowbowl	Arizona	Arizona	11500	2300	
4	Sunrise Park Resort	Arizona	Arizona	11100	1800	

	base_elev	trams	fastSixes	fastQuads	quad	…	TerrainParks	\
0	250	1	0	2	2	…	2.0	
1	1200	0	0	0	0	…	1.0	
2	1796	0	0	0	0	…	1.0	
3	9200	0	1	0	2	…	4.0	
4	9200	0	0	1	2	…	2.0	

	LongestRun_mi	SkiableTerrain_ac	Snow Making_ac	daysOpenLastYear	\
0	1.0	1610.0	113.0	150.0	
1	2.0	640.0	60.0	45.0	
2	1.0	30.0	30.0	150.0	
3	2.0	777.0	104.0	122.0	
4	1.2	800.0	80.0	115.0	

	yearsOpen	averageSnowfall	AdultWeekend	projectedDaysOpen	NightSkiing_ac
0	60.0	669.0	85.0	150.0	550.0
1	44.0	350.0	53.0	90.0	NaN
2	36.0	69.0	34.0	152.0	30.0
3	81.0	260.0	89.0	122.0	NaN
4	49.0	250.0	78.0	104.0	80.0

[5 rows x 25 columns]

[5]:
```
state_summary = pd.read_csv('state_summary.csv')
```

[6]:
```
state_summary.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35 entries, 0 to 34
Data columns (total 8 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
```

```
 0   state                       35 non-null     object
 1   resorts_per_state           35 non-null     int64
 2   state_total_skiable_area_ac 35 non-null     float64
 3   state_total_days_open       35 non-null     float64
 4   state_total_terrain_parks   35 non-null     float64
 5   state_total_nightskiing_ac  35 non-null     float64
 6   state_population            35 non-null     int64
 7   state_area_sq_miles         35 non-null     int64
dtypes: float64(4), int64(3), object(1)
memory usage: 2.3+ KB
```

[7]:
```
state_summary.head()
```

[7]:

	state	resorts_per_state	state_total_skiable_area_ac	state_total_days_open	state_total_terrain_parks	state_total_nightskiing_ac	state_population	state_area_sq_miles
0	Alaska	3	2280.0	345.0	4.0	580.0	731545	665384
1	Arizona	2	1577.0	237.0	6.0	80.0	7278717	113990
2	California	21	25948.0	2738.0	81.0	587.0	39512223	163695
3	Colorado	22	43682.0	3258.0	74.0	428.0	5758736	104094
4	Connecticut	5	358.0	353.0	10.0	256.0	3565278	5543

[8]:
```
state_summary_newind = state_summary.set_index('state')
```

[9]:
```
state_summary_newind.state_area_sq_miles.sort_values(ascending=False).head()
```

[9]:
```
state
Alaska        665384
California    163695
Montana       147040
New Mexico    121590
Arizona       113990
Name: state_area_sq_miles, dtype: int64
```

[10]:
```
state_summary_newind.state_population.sort_values(ascending=False).head()
```

[10]:
```
state
California      39512223
New York        19453561
Pennsylvania    12801989
Illinois        12671821
Ohio            11689100
Name: state_population, dtype: int64
```

[11]:
```
state_summary_newind.resorts_per_state.sort_values(ascending=False).head()
```

[11]:
```
state
New York        33
Michigan        28
Colorado        22
California      21
Pennsylvania    19
Name: resorts_per_state, dtype: int64
```

[12]:
```
state_summary_newind.state_total_skiable_area_ac.sort_values(ascending=False).
↪head()
```

[12]:
```
state
Colorado      43682.0
Utah          30508.0
California    25948.0
Montana       21410.0
Idaho         16396.0
Name: state_total_skiable_area_ac, dtype: float64
```

[13]:
```
state_summary_newind.state_total_nightskiing_ac.sort_values(ascending=False).
↪head()
```

[13]:
```
state
New York        2836.0
Washington      1997.0
Michigan        1946.0
Pennsylvania    1528.0
Oregon          1127.0
Name: state_total_nightskiing_ac, dtype: float64
```

[14]:
```
state_summary_newind.state_total_days_open.sort_values(ascending=False).head()
```

[14]:
```
state
Colorado        3258.0
California      2738.0
Michigan        2389.0
New York        2384.0
```

```
New Hampshire    1847.0
Name: state_total_days_open, dtype: float64
```

[15]:
```
state_summary['resorts_per_100kcapita'] = 100_000 * state_summary.
 ↪resorts_per_state / state_summary.state_population
state_summary['resorts_per_100ksq_mile'] = 100_000 * state_summary.
 ↪resorts_per_state / state_summary.state_area_sq_miles
state_summary.drop(columns=['state_population', 'state_area_sq_miles'],␣
 ↪inplace=True)
state_summary.head()
```

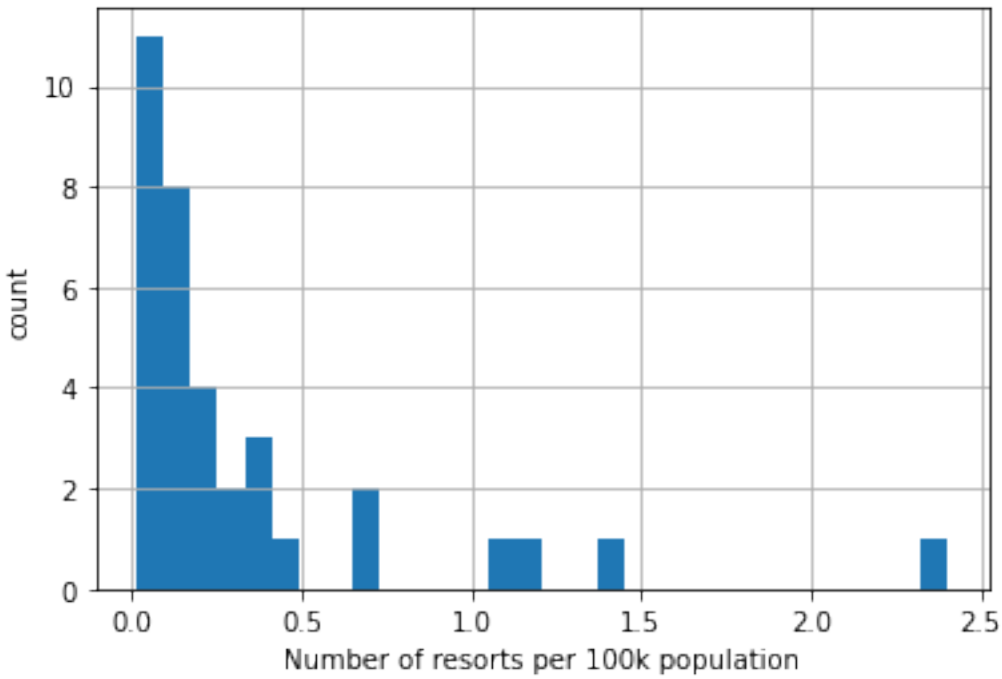
[15]:

	state	resorts_per_state	state_total_skiable_area_ac	\
0	Alaska	3	2280.0	
1	Arizona	2	1577.0	
2	California	21	25948.0	
3	Colorado	22	43682.0	
4	Connecticut	5	358.0	

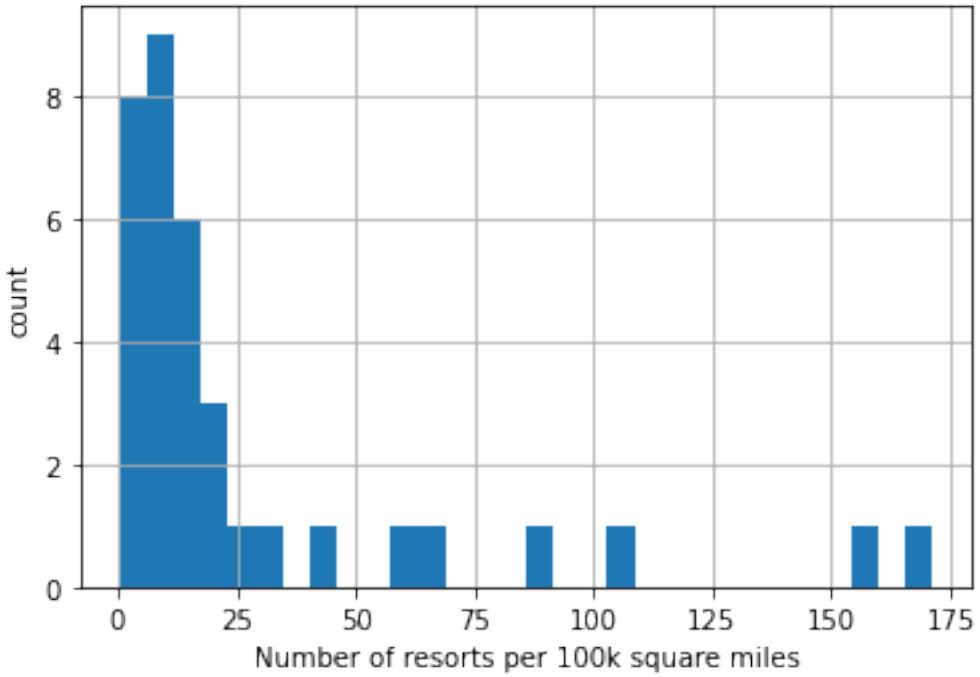
	state_total_days_open	state_total_terrain_parks	\
0	345.0	4.0	
1	237.0	6.0	
2	2738.0	81.0	
3	3258.0	74.0	
4	353.0	10.0	

	state_total_nightskiing_ac	resorts_per_100kcapita	resorts_per_100ksq_mile
0	580.0	0.410091	0.450867
1	80.0	0.027477	1.754540
2	587.0	0.053148	12.828736
3	428.0	0.382028	21.134744
4	256.0	0.140242	90.203861

[16]:
```
state_summary.resorts_per_100kcapita.hist(bins=30)
plt.xlabel('Number of resorts per 100k population')
plt.ylabel('count');
```



[17]:
```
state_summary.resorts_per_100ksq_mile.hist(bins=30)
plt.xlabel('Number of resorts per 100k square miles')
plt.ylabel('count');
```



[18]:
```
state_summary.set_index('state').resorts_per_100kcapita.
↪sort_values(ascending=False).head()
```

[18]:
```
state
Vermont          2.403889
Wyoming          1.382268
New Hampshire    1.176721
Montana          1.122778
Idaho            0.671492
Name: resorts_per_100kcapita, dtype: float64
```

Scale the data

[19]:
```
state_summary_scale = state_summary.set_index('state')
state_summary_index = state_summary_scale.index
state_summary_columns = state_summary_scale.columns
state_summary_scale.head()
```

[19]:
```
             resorts_per_state  state_total_skiable_area_ac  \
state
Alaska                       3                       2280.0
Arizona                      2                       1577.0
California                  21                      25948.0
Colorado                    22                      43682.0
Connecticut                  5                        358.0

             state_total_days_open  state_total_terrain_parks  \
state
Alaska                       345.0                       4.0
Arizona                      237.0                       6.0
California                  2738.0                      81.0
Colorado                    3258.0                      74.0
Connecticut                  353.0                      10.0

             state_total_nightskiing_ac  resorts_per_100kcapita  \
state
Alaska                            580.0                0.410091
Arizona                            80.0                0.027477
California                        587.0                0.053148
Colorado                          428.0                0.382028
Connecticut                       256.0                0.140242

             resorts_per_100ksq_mile
state
Alaska                      0.450867
Arizona                     1.754540
```

```
California                  12.828736
Colorado                    21.134744
Connecticut                 90.203861
```

[20]:
```
state_summary_scale = scale(state_summary_scale)
```

[21]:
```
state_summary_scaled_df = pd.DataFrame(state_summary_scale,␣
↪columns=state_summary_columns)
state_summary_scaled_df.head()
```

[21]:

	resorts_per_state	state_total_skiable_area_ac	state_total_days_open	\
0	-0.806912	-0.392012	-0.689059	
1	-0.933558	-0.462424	-0.819038	
2	1.472706	1.978574	2.190933	
3	1.599351	3.754811	2.816757	
4	-0.553622	-0.584519	-0.679431	

	state_total_terrain_parks	state_total_nightskiing_ac	\
0	-0.816118	0.069410	
1	-0.726994	-0.701326	
2	2.615141	0.080201	
3	2.303209	-0.164893	
4	-0.548747	-0.430027	

	resorts_per_100kcapita	resorts_per_100ksq_mile
0	0.139593	-0.689999
1	-0.644706	-0.658125
2	-0.592085	-0.387368
3	0.082069	-0.184291
4	-0.413557	1.504408

[22]:
```
state_summary_scaled_df.mean()
```

[22]:
```
resorts_per_state           -7.295751e-17
state_total_skiable_area_ac -4.163336e-17
state_total_days_open        7.692260e-17
state_total_terrain_parks    4.599495e-17
state_total_nightskiing_ac   7.612958e-17
resorts_per_100kcapita       5.075305e-17
resorts_per_100ksq_mile      5.075305e-17
dtype: float64
```

[23]:
```
state_summary_scaled_df.std()
```

[23]:
```
resorts_per_state            1.014599
state_total_skiable_area_ac  1.014599
state_total_days_open        1.014599
```

```
state_total_terrain_parks      1.014599
state_total_nightskiing_ac     1.014599
resorts_per_100kcapita         1.014599
resorts_per_100ksq_mile        1.014599
dtype: float64
```

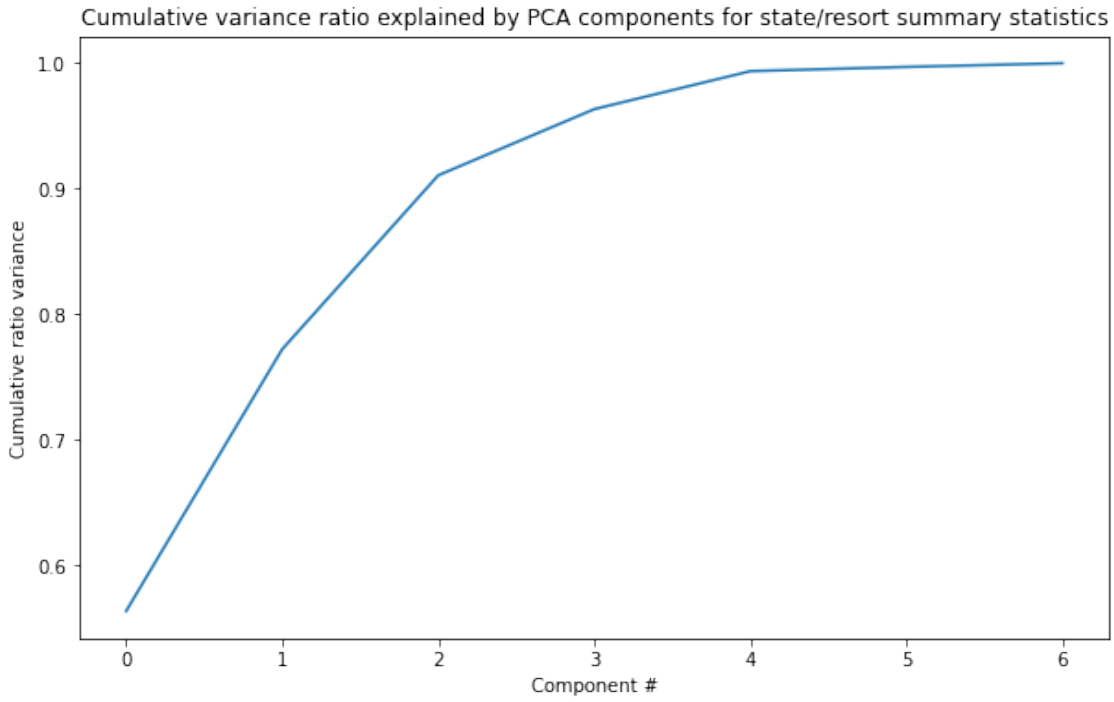
[24]:
```
state_summary_scaled_df.std(ddof=0)
```

[24]:
```
resorts_per_state              1.0
state_total_skiable_area_ac    1.0
state_total_days_open          1.0
state_total_terrain_parks      1.0
state_total_nightskiing_ac     1.0
resorts_per_100kcapita         1.0
resorts_per_100ksq_mile        1.0
dtype: float64
```

Fit the PCA transformation using the scaled data.

[25]:
```
state_pca = PCA().fit(state_summary_scale)
```

[26]:
```
plt.subplots(figsize=(10, 6))
plt.plot(state_pca.explained_variance_ratio_.cumsum())
plt.xlabel('Component #')
plt.ylabel('Cumulative ratio variance')
plt.title('Cumulative variance ratio explained by PCA components for state/
 ↪resort summary statistics');
```

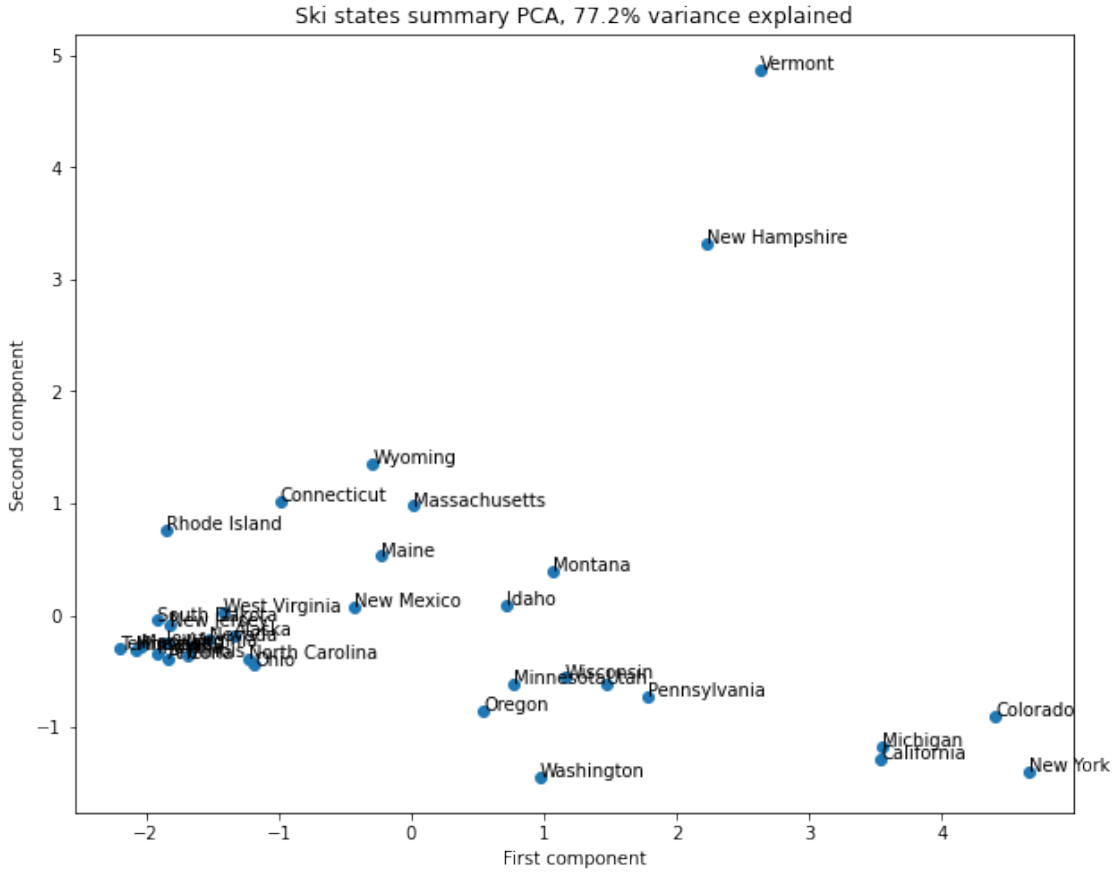


[27]:
```
state_pca_x = state_pca.transform(state_summary_scale)
```

[28]:
```
state_pca_x.shape
```

[28]:
```
(35, 7)
```

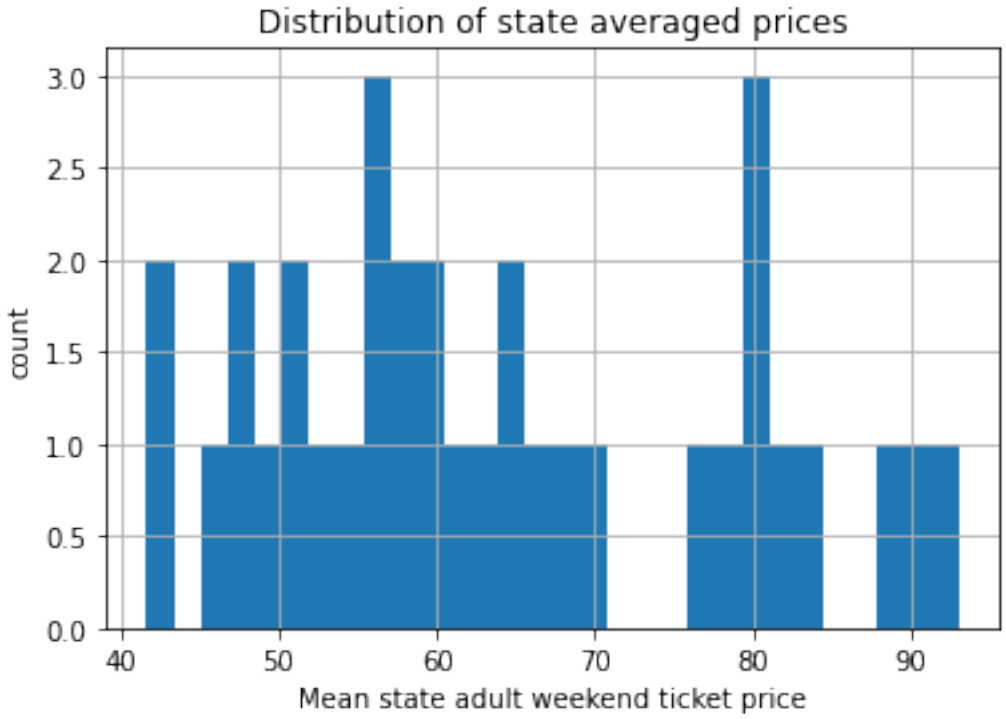
[29]:
```
x = state_pca_x[:, 0]
y = state_pca_x[:, 1]
state = state_summary_index
pc_var = 100 * state_pca.explained_variance_ratio_.cumsum()[1]
plt.subplots(figsize=(10,8))
plt.scatter(x=x, y=y)
plt.xlabel('First component')
plt.ylabel('Second component')
plt.title(f'Ski states summary PCA, {pc_var:.1f}% variance explained')
for s, x, y in zip(state, x, y):
    plt.annotate(s, (x, y))
```



[30]:
```
state_avg_price = ski_data.groupby('state')['AdultWeekend'].mean()
state_avg_price.head()
```

[30]:
```
state
Alaska         57.333333
Arizona        83.500000
California     81.416667
Colorado       90.714286
Connecticut    56.800000
Name: AdultWeekend, dtype: float64
```

[31]:
```
state_avg_price.hist(bins=30)
plt.title('Distribution of state averaged prices')
plt.xlabel('Mean state adult weekend ticket price')
plt.ylabel('count');
```



[32]:
```
pca_df = pd.DataFrame({'PC1':state_pca_x[:, 0], 'PC2': state_pca_x[:, 1]},␣
↪index=state_summary_index)
pca_df.head()
```

[32]:
```
                PC1       PC2
state
Alaska      -1.336533 -0.182208
```

```
Arizona     -1.839049 -0.387959
California   3.537857 -1.282509
Colorado     4.402210 -0.898855
Connecticut -0.988027  1.020218
```

[33]:
```
state_avg_price.head()
```

[33]:
```
state
Alaska         57.333333
Arizona        83.500000
California     81.416667
Colorado       90.714286
Connecticut    56.800000
Name: AdultWeekend, dtype: float64
```

[34]:
```
state_avg_price.to_frame().head()
```

[34]:
```
             AdultWeekend
state
Alaska          57.333333
Arizona         83.500000
California      81.416667
Colorado        90.714286
Connecticut     56.800000
```

[35]:
```
pca_df = pd.concat([pca_df, state_avg_price], axis=1)
pca_df.head()
```

[35]:
```
                  PC1       PC2  AdultWeekend
state
Alaska      -1.336533 -0.182208     57.333333
Arizona     -1.839049 -0.387959     83.500000
California   3.537857 -1.282509     81.416667
Colorado     4.402210 -0.898855     90.714286
Connecticut -0.988027  1.020218     56.800000
```

[36]:
```
pca_df['Quartile'] = pd.qcut(pca_df.AdultWeekend, q=4, precision=1)
pca_df.head()
```

[36]:
```
                  PC1       PC2  AdultWeekend      Quartile
state
Alaska      -1.336533 -0.182208     57.333333  (53.1, 60.4]
Arizona     -1.839049 -0.387959     83.500000  (78.4, 93.0]
California   3.537857 -1.282509     81.416667  (78.4, 93.0]
Colorado     4.402210 -0.898855     90.714286  (78.4, 93.0]
Connecticut -0.988027  1.020218     56.800000  (53.1, 60.4]
```

[37]:
```
pca_df.dtypes
```

[37]:
```
PC1              float64
PC2              float64
AdultWeekend     float64
Quartile        category
dtype: object
```

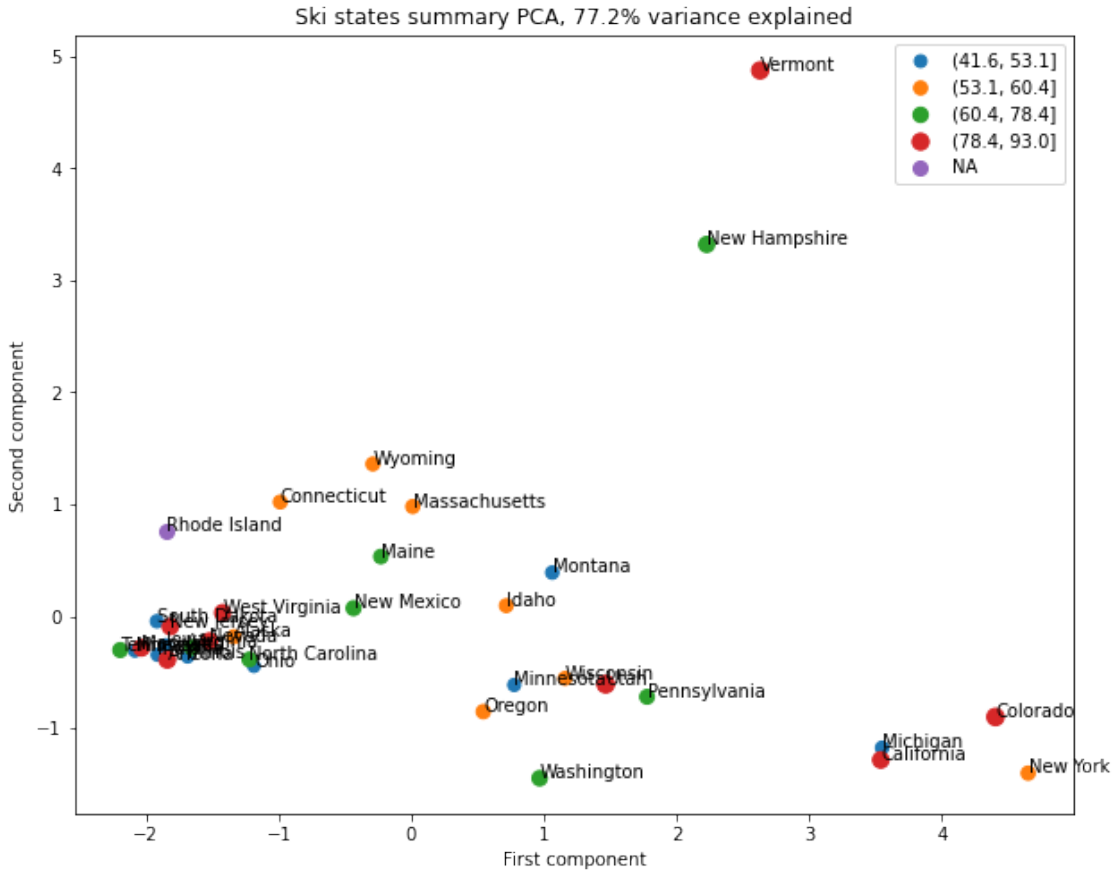
[38]:
```
pca_df[pca_df.isnull().any(axis=1)]
```

[38]:
```
                    PC1       PC2  AdultWeekend Quartile
state
Rhode Island -1.843646  0.761339           NaN      NaN
```

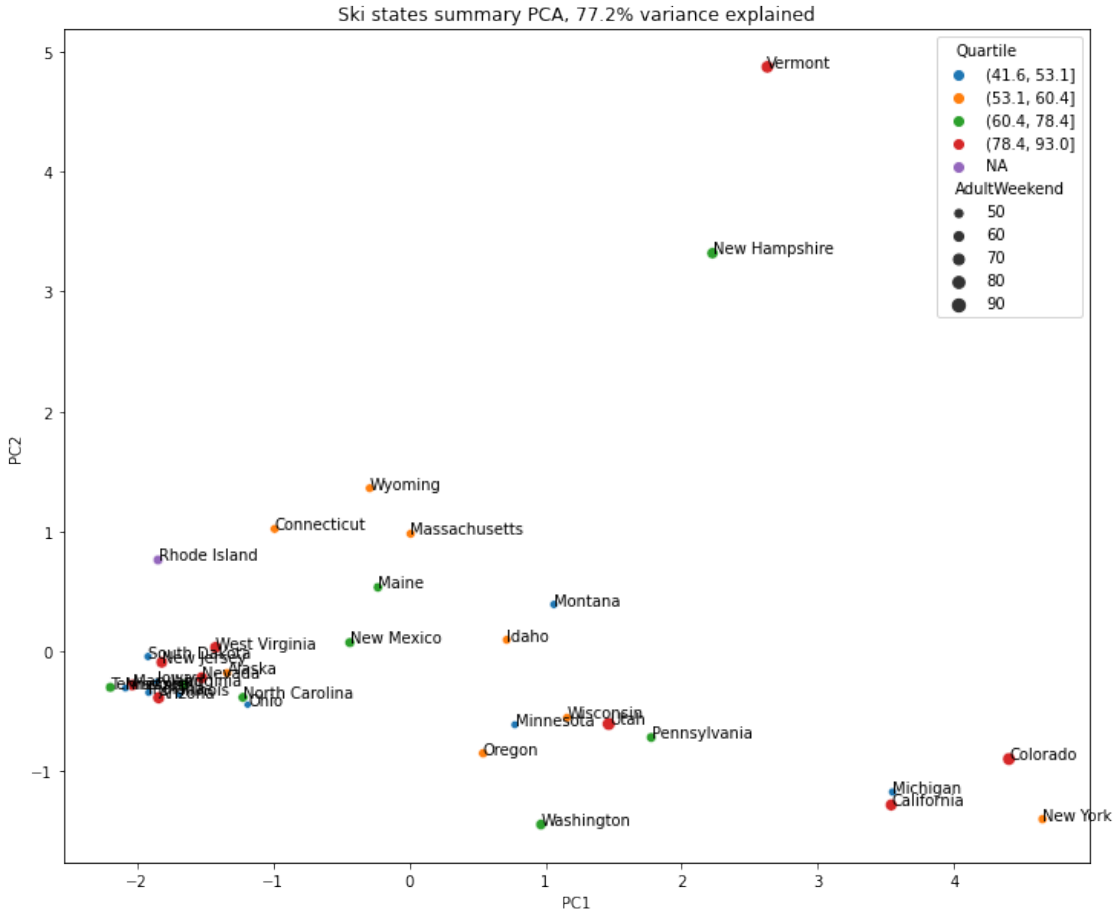
[39]:
```
pca_df['AdultWeekend'].fillna(pca_df.AdultWeekend.mean(), inplace=True)
pca_df['Quartile'] = pca_df['Quartile'].cat.add_categories('NA')
pca_df['Quartile'].fillna('NA', inplace=True)
pca_df.loc['Rhode Island']
```

[39]:
```
PC1             -1.843646
PC2              0.761339
AdultWeekend    64.124388
Quartile               NA
Name: Rhode Island, dtype: object
```

[40]:
```
x = pca_df.PC1
y = pca_df.PC2
price = pca_df.AdultWeekend
quartiles = pca_df.Quartile
state = pca_df.index
pc_var = 100 * state_pca.explained_variance_ratio_.cumsum()[1]
fig, ax = plt.subplots(figsize=(10,8))
for q in quartiles.cat.categories:
    im = quartiles == q
    ax.scatter(x=x[im], y=y[im], s=price[im], label=q)
ax.set_xlabel('First component')
ax.set_ylabel('Second component')
plt.legend()
ax.set_title(f'Ski states summary PCA, {pc_var:.1f}% variance explained')
for s, x, y in zip(state, x, y):
    plt.annotate(s, (x, y))
```



[41]:
```
x = pca_df.PC1
y = pca_df.PC2
state = pca_df.index
plt.subplots(figsize=(12, 10))
sns.scatterplot(x='PC1', y='PC2', size='AdultWeekend', hue='Quartile',
                hue_order=pca_df.Quartile.cat.categories, data=pca_df)
for s, x, y in zip(state, x, y):
    plt.annotate(s, (x, y))
plt.title(f'Ski states summary PCA, {pc_var:.1f}% variance explained');
```



[42]:
```
pd.DataFrame(state_pca.components_, columns=state_summary_columns)
```

[42]:

	resorts_per_state	state_total_skiable_area_ac	state_total_days_open	\
0	0.486079	0.318224	0.489997	
1	-0.085092	-0.142204	-0.045071	
2	-0.177937	0.714835	0.115200	
3	0.056163	-0.118347	-0.162625	
4	-0.209186	0.573462	-0.250521	
5	-0.818390	-0.092319	0.238198	
6	-0.090273	-0.127021	0.773728	

	state_total_terrain_parks	state_total_nightskiing_ac	\
0	0.488420	0.334398	
1	-0.041939	-0.351064	
2	0.005509	-0.511255	
3	-0.177072	0.438912	
4	-0.388608	0.499801	
5	0.448118	0.246196	

```
6                  -0.613576                    0.022185

   resorts_per_100kcapita  resorts_per_100ksq_mile
0                0.187154                 0.192250
1                0.662458                 0.637691
2                0.220359                -0.366207
3                0.685417                -0.512443
4               -0.065077                 0.399461
5                0.058911                -0.009146
6               -0.007887                -0.005631
```

[43]:
```
state_summary[state_summary.state.isin(['New Hampshire', 'Vermont'])].T
```

[43]:
```
                                    17          29
state                       New Hampshire     Vermont
resorts_per_state                      16          15
state_total_skiable_area_ac        3427.0      7239.0
state_total_days_open              1847.0      1777.0
state_total_terrain_parks            43.0        50.0
state_total_nightskiing_ac          376.0        50.0
resorts_per_100kcapita           1.176721    2.403889
resorts_per_100ksq_mile        171.141299  155.990017
```

[44]:
```
state_summary_scaled_df[state_summary.state.isin(['New Hampshire', 'Vermont'])].
↪T
```

[44]:
```
                                 17        29
resorts_per_state          0.839478  0.712833
state_total_skiable_area_ac -0.277128  0.104681
state_total_days_open       1.118608  1.034363
state_total_terrain_parks   0.921793  1.233725
state_total_nightskiing_ac  -0.245050 -0.747570
resorts_per_100kcapita      1.711066  4.226572
resorts_per_100ksq_mile     3.483281  3.112841
```

[45]:
```
ski_data.head().T
```

[45]:
```
                             0                  1                2  \
Name              Alyeska Resort  Eaglecrest Ski Area  Hilltop Ski Area
Region                    Alaska               Alaska            Alaska
state                     Alaska               Alaska            Alaska
summit_elev                 3939                 2600              2090
vertical_drop               2500                 1540               294
base_elev                    250                 1200              1796
trams                          1                    0                 0
fastSixes                      0                    0                 0
fastQuads                      2                    0                 0
```

quad	2	0	0
triple	0	0	1
double	0	4	0
surface	2	0	2
total_chairs	7	4	3
Runs	76.0	36.0	13.0
TerrainParks	2.0	1.0	1.0
LongestRun_mi	1.0	2.0	1.0
SkiableTerrain_ac	1610.0	640.0	30.0
Snow Making_ac	113.0	60.0	30.0
daysOpenLastYear	150.0	45.0	150.0
yearsOpen	60.0	44.0	36.0
averageSnowfall	669.0	350.0	69.0
AdultWeekend	85.0	53.0	34.0
projectedDaysOpen	150.0	90.0	152.0
NightSkiing_ac	550.0	NaN	30.0

	3	4
Name	Arizona Snowbowl	Sunrise Park Resort
Region	Arizona	Arizona
state	Arizona	Arizona
summit_elev	11500	11100
vertical_drop	2300	1800
base_elev	9200	9200
trams	0	0
fastSixes	1	0
fastQuads	0	1
quad	2	2
triple	2	3
double	1	1
surface	2	0
total_chairs	8	7
Runs	55.0	65.0
TerrainParks	4.0	2.0
LongestRun_mi	2.0	1.2
SkiableTerrain_ac	777.0	800.0
Snow Making_ac	104.0	80.0
daysOpenLastYear	122.0	115.0
yearsOpen	81.0	49.0
averageSnowfall	260.0	250.0
AdultWeekend	89.0	78.0
projectedDaysOpen	122.0	104.0
NightSkiing_ac	NaN	80.0

[46]:
```
state_summary.head()
```

[46]:

	state	resorts_per_state	state_total_skiable_area_ac	state_total_days_open	state_total_terrain_parks	state_total_nightskiing_ac	resorts_per_100kcapita	resorts_per_100ksq_mile
0	Alaska	3	2280.0	345.0	4.0	580.0	0.410091	0.450867
1	Arizona	2	1577.0	237.0	6.0	80.0	0.027477	1.754540
2	California	21	25948.0	2738.0	81.0	587.0	0.053148	12.828736
3	Colorado	22	43682.0	3258.0	74.0	428.0	0.382028	21.134744
4	Connecticut	5	358.0	353.0	10.0	256.0	0.140242	90.203861

[47]:
```
# DataFrame's merge method provides SQL-like joins
# here 'state' is a column (not an index)
ski_data = ski_data.merge(state_summary, how='left', on='state')
ski_data.head().T
```

[47]:

	0	1	\
Name	Alyeska Resort	Eaglecrest Ski Area	
Region	Alaska	Alaska	
state	Alaska	Alaska	
summit_elev	3939	2600	
vertical_drop	2500	1540	
base_elev	250	1200	
trams	1	0	
fastSixes	0	0	
fastQuads	2	0	
quad	2	0	
triple	0	0	
double	0	4	
surface	2	0	
total_chairs	7	4	
Runs	76.0	36.0	
TerrainParks	2.0	1.0	
LongestRun_mi	1.0	2.0	
SkiableTerrain_ac	1610.0	640.0	
Snow Making_ac	113.0	60.0	
daysOpenLastYear	150.0	45.0	

yearsOpen	60.0	44.0
averageSnowfall	669.0	350.0
AdultWeekend	85.0	53.0
projectedDaysOpen	150.0	90.0
NightSkiing_ac	550.0	NaN
resorts_per_state	3	3
state_total_skiable_area_ac	2280.0	2280.0
state_total_days_open	345.0	345.0
state_total_terrain_parks	4.0	4.0
state_total_nightskiing_ac	580.0	580.0
resorts_per_100kcapita	0.410091	0.410091
resorts_per_100ksq_mile	0.450867	0.450867

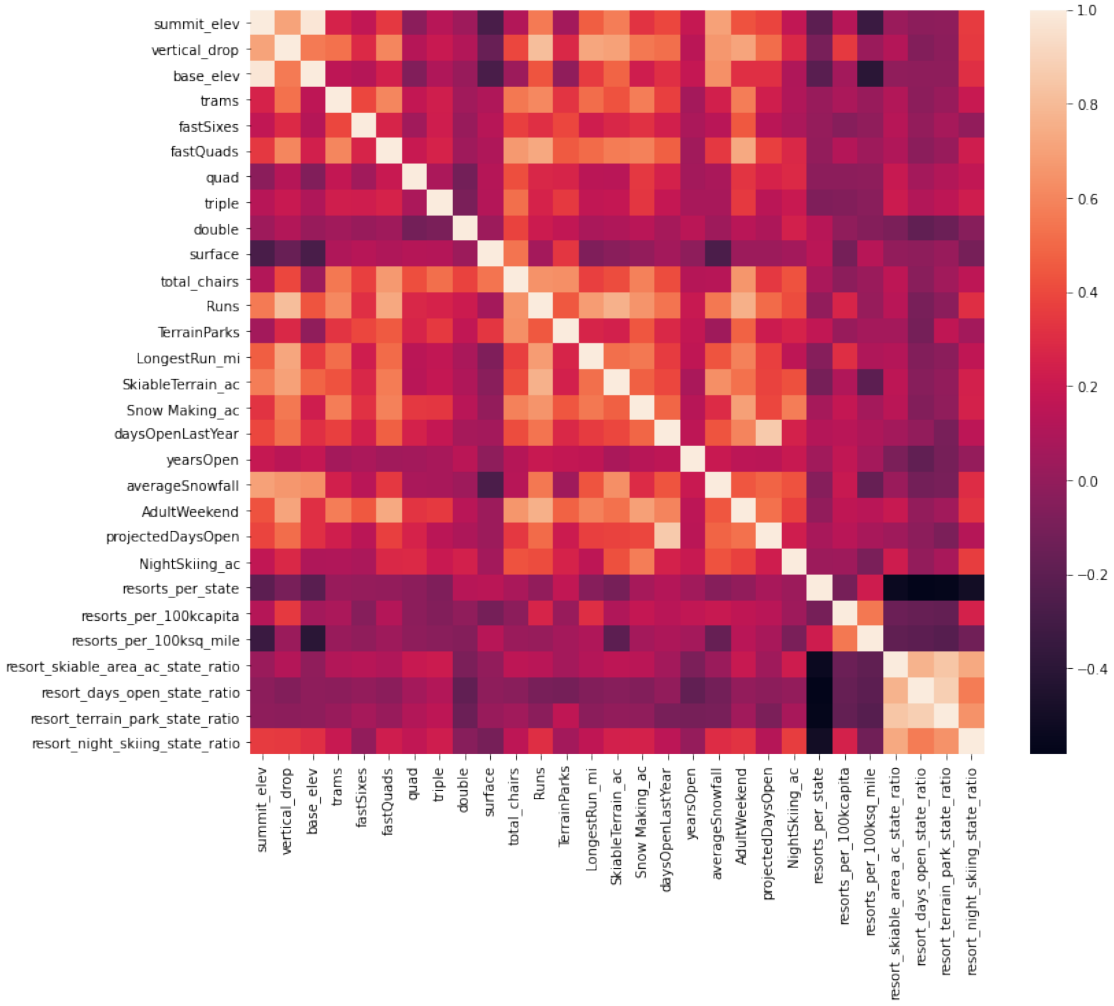
	2	3 \
Name	Hilltop Ski Area	Arizona Snowbowl
Region	Alaska	Arizona
state	Alaska	Arizona
summit_elev	2090	11500
vertical_drop	294	2300
base_elev	1796	9200
trams	0	0
fastSixes	0	1
fastQuads	0	0
quad	0	2
triple	1	2
double	0	1
surface	2	2
total_chairs	3	8
Runs	13.0	55.0
TerrainParks	1.0	4.0
LongestRun_mi	1.0	2.0
SkiableTerrain_ac	30.0	777.0
Snow Making_ac	30.0	104.0
daysOpenLastYear	150.0	122.0
yearsOpen	36.0	81.0
averageSnowfall	69.0	260.0
AdultWeekend	34.0	89.0
projectedDaysOpen	152.0	122.0
NightSkiing_ac	30.0	NaN
resorts_per_state	3	2
state_total_skiable_area_ac	2280.0	1577.0
state_total_days_open	345.0	237.0
state_total_terrain_parks	4.0	6.0
state_total_nightskiing_ac	580.0	80.0
resorts_per_100kcapita	0.410091	0.027477
resorts_per_100ksq_mile	0.450867	1.75454

```
                                        4
Name                      Sunrise Park Resort
Region                                Arizona
state                                 Arizona
summit_elev                             11100
vertical_drop                            1800
base_elev                                9200
trams                                       0
fastSixes                                   0
fastQuads                                   1
quad                                        2
triple                                      3
double                                      1
surface                                     0
total_chairs                                7
Runs                                     65.0
TerrainParks                              2.0
LongestRun_mi                             1.2
SkiableTerrain_ac                       800.0
Snow Making_ac                           80.0
daysOpenLastYear                        115.0
yearsOpen                                49.0
averageSnowfall                         250.0
AdultWeekend                             78.0
projectedDaysOpen                       104.0
NightSkiing_ac                           80.0
resorts_per_state                           2
state_total_skiable_area_ac           1577.0
state_total_days_open                  237.0
state_total_terrain_parks                6.0
state_total_nightskiing_ac              80.0
resorts_per_100kcapita              0.027477
resorts_per_100ksq_mile              1.75454
```

[48]:
```
ski_data['resort_skiable_area_ac_state_ratio'] = ski_data.SkiableTerrain_ac /␣
↪ski_data.state_total_skiable_area_ac
ski_data['resort_days_open_state_ratio'] = ski_data.daysOpenLastYear / ski_data.
↪state_total_days_open
ski_data['resort_terrain_park_state_ratio'] = ski_data.TerrainParks / ski_data.
↪state_total_terrain_parks
ski_data['resort_night_skiing_state_ratio'] = ski_data.NightSkiing_ac /␣
↪ski_data.state_total_nightskiing_ac

ski_data.drop(columns=['state_total_skiable_area_ac', 'state_total_days_open',
                       'state_total_terrain_parks',␣
↪'state_total_nightskiing_ac'], inplace=True)
```

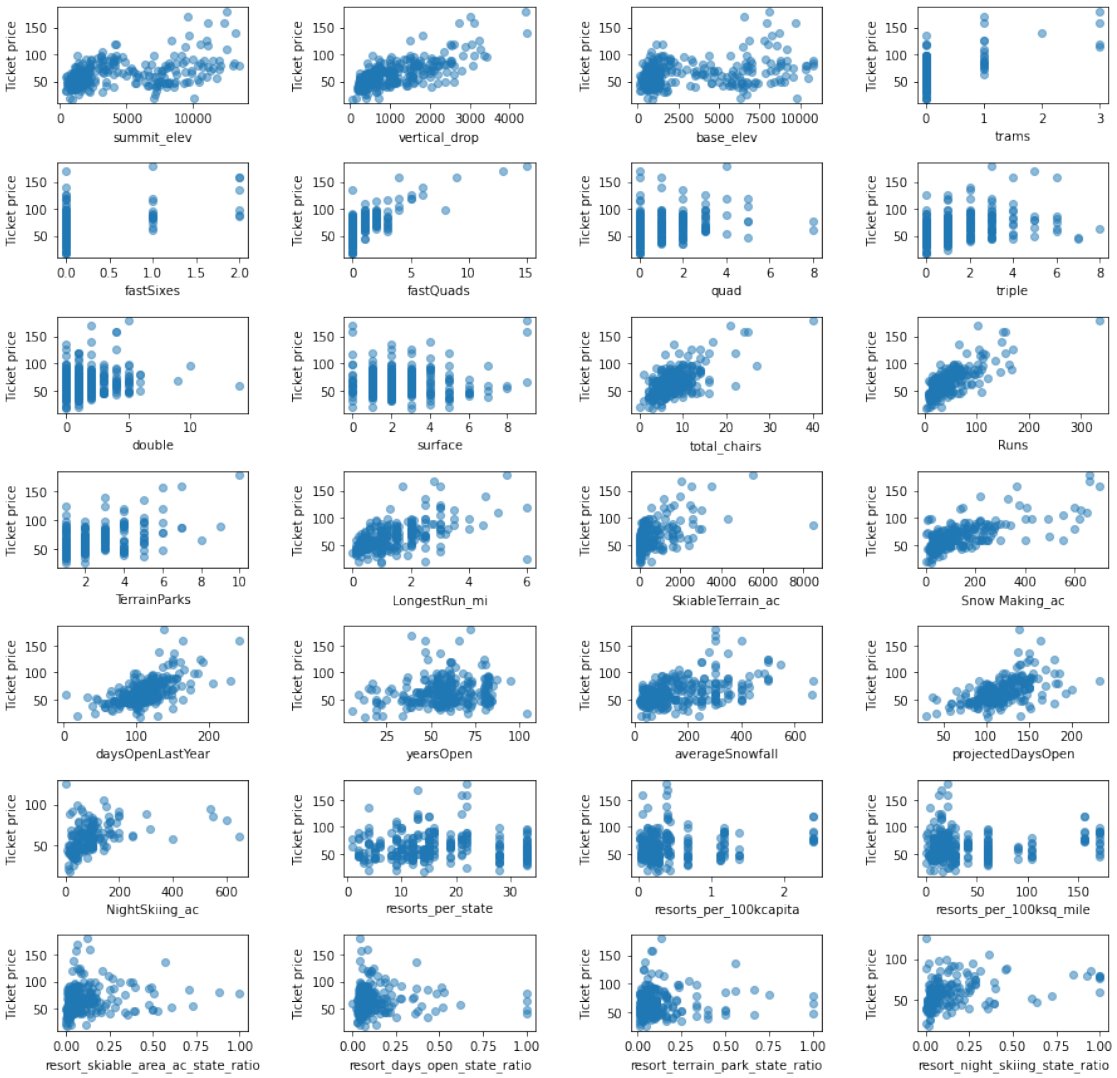
[49]:
```
plt.subplots(figsize=(12,10))
sns.heatmap(ski_data.corr());
```



[50]:
```
def scatterplots(columns, ncol=None, figsize=(15, 8)):
    if ncol is None:
        ncol = len(columns)
    nrow = int(np.ceil(len(columns) / ncol))
    fig, axes = plt.subplots(nrow, ncol, figsize=figsize, squeeze=False)
    fig.subplots_adjust(wspace=0.5, hspace=0.6)
    for i, col in enumerate(columns):
        ax = axes.flatten()[i]
        ax.scatter(x = col, y = 'AdultWeekend', data=ski_data, alpha=0.5)
        ax.set(xlabel=col, ylabel='Ticket price')
    nsubplots = nrow * ncol
    for empty in range(i+1, nsubplots):
        axes.flatten()[empty].set_visible(False)
```

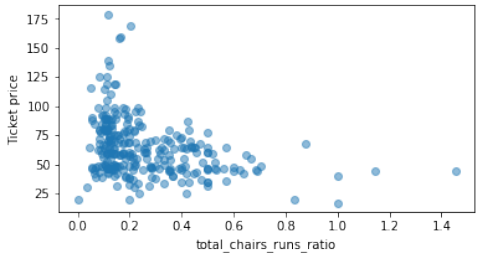
[51]:
```
features = [x for x in ski_data.columns if x not in ['Name', 'Region','state',␣
↪'AdultWeekend']]
```

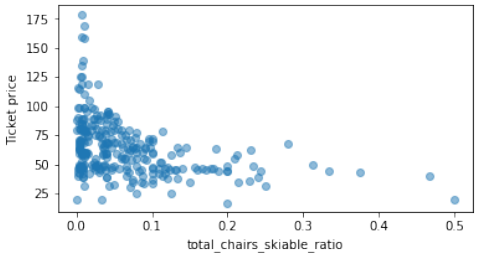
[52]:
```
scatterplots(features, ncol=4, figsize=(15, 15))
```

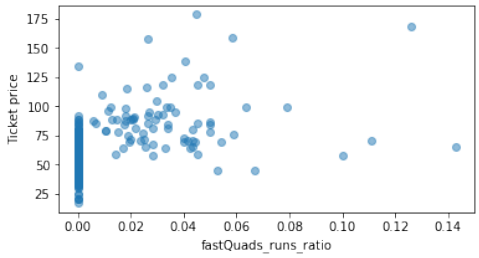


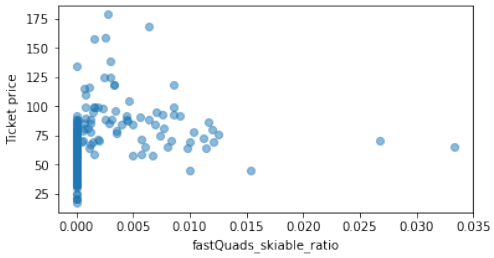
[53]:
```
ski_data['total_chairs_runs_ratio'] = ski_data.total_chairs / ski_data.Runs
ski_data['total_chairs_skiable_ratio'] = ski_data.total_chairs / ski_data.
↪SkiableTerrain_ac
ski_data['fastQuads_runs_ratio'] = ski_data.fastQuads / ski_data.Runs
ski_data['fastQuads_skiable_ratio'] = ski_data.fastQuads / ski_data.
↪SkiableTerrain_ac
```

[54]:
```
scatterplots(['total_chairs_runs_ratio', 'total_chairs_skiable_ratio',
              'fastQuads_runs_ratio', 'fastQuads_skiable_ratio'], ncol=2)
```









In this notebook, we took a deeper look at the data provided to us, most of which are numerical. We started with examining the relationship between state and ticket price by exploring more information on each state. As we can see, New York has the most number of resorts in our data. However, they don't have as many skiing areas as expected, which make sense with New York's high density. We can also see that large states don't necessarily have more resorts. To better understand the data, we calculate the density of resorts in each state: resorts_per_100kcapita and resorts_per_100ksq_mile. After ranking the top states using the ratios, we can see Vermont and New Hampshire are in the top list. The relationship between state and ticket price is not obvious at this point.

To further investigate, we scaled the data on state and performed a PCA transformation. From the PCA, we derived two features and labeled them x and y. We wanted to see if there was a pattern between state and average ticket price. Using matplotlib, we did not discover any obvious pattern, which is why we turned to seaborn, where we found that Vermont and New Hampshire really stand out from the other states. We also discovered that resorts_per_100kcapita and resorts_per_100ksq_mile accounts for quite a lot in why these two states stand out. Overall, we did not find perticular pattern in the state data.

Afterwards, we plotted a heatmap to see the relationship amongst features. We can clearly spot correlations between some features. For example, more night time skiing is provided if the resorts are more densely located with population. We can also find some features, like Runs or Snow, are highly correlated with the ticket price. These features should be used in the subsequence modeling.

[55]:
```
ski_data.head().T
```

[55]:

	0	1
Name	Alyeska Resort	Eaglecrest Ski Area
Region	Alaska	Alaska
state	Alaska	Alaska
summit_elev	3939	2600
vertical_drop	2500	1540
base_elev	250	1200
trams	1	0
fastSixes	0	0
fastQuads	2	0
quad	2	0
triple	0	0
double	0	4
surface	2	0
total_chairs	7	4
Runs	76.0	36.0
TerrainParks	2.0	1.0
LongestRun_mi	1.0	2.0
SkiableTerrain_ac	1610.0	640.0
Snow Making_ac	113.0	60.0
daysOpenLastYear	150.0	45.0
yearsOpen	60.0	44.0
averageSnowfall	669.0	350.0
AdultWeekend	85.0	53.0
projectedDaysOpen	150.0	90.0
NightSkiing_ac	550.0	NaN
resorts_per_state	3	3
resorts_per_100kcapita	0.410091	0.410091
resorts_per_100ksq_mile	0.450867	0.450867
resort_skiable_area_ac_state_ratio	0.70614	0.280702
resort_days_open_state_ratio	0.434783	0.130435
resort_terrain_park_state_ratio	0.5	0.25
resort_night_skiing_state_ratio	0.948276	NaN
total_chairs_runs_ratio	0.092105	0.111111
total_chairs_skiable_ratio	0.004348	0.00625
fastQuads_runs_ratio	0.026316	0.0
fastQuads_skiable_ratio	0.001242	0.0

	2	3
Name	Hilltop Ski Area	Arizona Snowbowl
Region	Alaska	Arizona
state	Alaska	Arizona
summit_elev	2090	11500
vertical_drop	294	2300
base_elev	1796	9200
trams	0	0
fastSixes	0	1

fastQuads	0	0
quad	0	2
triple	1	2
double	0	1
surface	2	2
total_chairs	3	8
Runs	13.0	55.0
TerrainParks	1.0	4.0
LongestRun_mi	1.0	2.0
SkiableTerrain_ac	30.0	777.0
Snow Making_ac	30.0	104.0
daysOpenLastYear	150.0	122.0
yearsOpen	36.0	81.0
averageSnowfall	69.0	260.0
AdultWeekend	34.0	89.0
projectedDaysOpen	152.0	122.0
NightSkiing_ac	30.0	NaN
resorts_per_state	3	2
resorts_per_100kcapita	0.410091	0.027477
resorts_per_100ksq_mile	0.450867	1.75454
resort_skiable_area_ac_state_ratio	0.013158	0.492708
resort_days_open_state_ratio	0.434783	0.514768
resort_terrain_park_state_ratio	0.25	0.666667
resort_night_skiing_state_ratio	0.051724	NaN
total_chairs_runs_ratio	0.230769	0.145455
total_chairs_skiable_ratio	0.1	0.010296
fastQuads_runs_ratio	0.0	0.0
fastQuads_skiable_ratio	0.0	0.0

	4
Name	Sunrise Park Resort
Region	Arizona
state	Arizona
summit_elev	11100
vertical_drop	1800
base_elev	9200
trams	0
fastSixes	0
fastQuads	1
quad	2
triple	3
double	1
surface	0
total_chairs	7
Runs	65.0
TerrainParks	2.0
LongestRun_mi	1.2

```
SkiableTerrain_ac                      800.0
Snow Making_ac                          80.0
daysOpenLastYear                       115.0
yearsOpen                               49.0
averageSnowfall                        250.0
AdultWeekend                            78.0
projectedDaysOpen                      104.0
NightSkiing_ac                          80.0
resorts_per_state                          2
resorts_per_100kcapita              0.027477
resorts_per_100ksq_mile              1.75454
resort_skiable_area_ac_state_ratio  0.507292
resort_days_open_state_ratio        0.485232
resort_terrain_park_state_ratio     0.333333
resort_night_skiing_state_ratio          1.0
total_chairs_runs_ratio             0.107692
total_chairs_skiable_ratio           0.00875
fastQuads_runs_ratio                0.015385
fastQuads_skiable_ratio              0.00125
```

[56]:
```
ski_data
```

[56]:
```
                                 Name   Region    state  summit_elev  \
0                      Alyeska Resort   Alaska   Alaska         3939
1                  Eaglecrest Ski Area   Alaska   Alaska         2600
2                     Hilltop Ski Area   Alaska   Alaska         2090
3                     Arizona Snowbowl  Arizona  Arizona        11500
4                  Sunrise Park Resort  Arizona  Arizona        11100
..                                 …      …      …           …
272                      Hogadon Basin  Wyoming  Wyoming         8000
273          Sleeping Giant Ski Resort  Wyoming  Wyoming         7428
274                   Snow King Resort  Wyoming  Wyoming         7808
275  Snowy Range Ski & Recreation Area  Wyoming  Wyoming         9663
276                White Pine Ski Area  Wyoming  Wyoming         9500

     vertical_drop  base_elev  trams  fastSixes  fastQuads  quad  …  \
0             2500        250      1          0          2     2  …
1             1540       1200      0          0          0     0  …
2              294       1796      0          0          0     0  …
3             2300       9200      0          1          0     2  …
4             1800       9200      0          0          1     2  …
..             …          …    …          …          …   …  …
272            640       7400      0          0          0     0  …
273            810       6619      0          0          0     0  …
274           1571       6237      0          0          0     1  …
275            990       8798      0          0          0     0  …
276           1100       8400      0          0          0     0  …
```

```
     resorts_per_100kcapita  resorts_per_100ksq_mile  \
0                  0.410091                 0.450867
1                  0.410091                 0.450867
2                  0.410091                 0.450867
3                  0.027477                 1.754540
4                  0.027477                 1.754540
..                      …                        …
272                1.382268                 8.178872
273                1.382268                 8.178872
274                1.382268                 8.178872
275                1.382268                 8.178872
276                1.382268                 8.178872

     resort_skiable_area_ac_state_ratio  resort_days_open_state_ratio  \
0                              0.706140                     0.434783
1                              0.280702                     0.130435
2                              0.013158                     0.434783
3                              0.492708                     0.514768
4                              0.507292                     0.485232
..                                  …                            …
272                            0.014104                     0.168994
273                            0.028208                     0.085196
274                            0.061321                     0.168994
275                            0.011498                     0.182961
276                            0.056722                          NaN

     resort_terrain_park_state_ratio  resort_night_skiing_state_ratio  \
0                           0.500000                         0.948276
1                           0.250000                              NaN
2                           0.250000                         0.051724
3                           0.666667                              NaN
4                           0.333333                         1.000000
..                               …                            …
272                         0.071429                              NaN
273                         0.071429                              NaN
274                         0.142857                         1.000000
275                         0.142857                              NaN
276                              NaN                              NaN

     total_chairs_runs_ratio  total_chairs_skiable_ratio  \
0                   0.092105                    0.004348
1                   0.111111                    0.006250
2                   0.230769                    0.100000
3                   0.145455                    0.010296
4                   0.107692                    0.008750
..                       …                           …
```

```
272                 0.071429                    0.021739
273                 0.062500                    0.016304
274                 0.093750                    0.007500
275                 0.151515                    0.066667
276                 0.080000                    0.005405

     fastQuads_runs_ratio  fastQuads_skiable_ratio
0                0.026316                 0.001242
1                0.000000                 0.000000
2                0.000000                 0.000000
3                0.000000                 0.000000
4                0.015385                 0.001250
..                    …                        …
272              0.000000                 0.000000
273              0.000000                 0.000000
274              0.000000                 0.000000
275              0.000000                 0.000000
276              0.000000                 0.000000

[277 rows x 36 columns]
```

[57]:
```
# Save the data
ski_data.to_csv('ski_data_step3_features.csv', index=False)
```

[57]:
```

```

Created in Deepnote