data_wrangling

June 22, 2021

```
[ ]: import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     import os
```

0.0.1 Introducton

In this project, we aim to find a more efficient pricing model for Big Mountain Resort using a dataset with information on 330 different ski resorts across the United States.

Big Mountain Resort, a ski resort located in Montana, has recently installed an additional chair lift, which increases their operating costs by $1,540,000 this season. Big Mountain Resort are looking for guidance on how to select a better value for their ticket price (currently charged based only on market average.)

To find a better pricing strategy for Big Mountain Resort, we will apply the techiques of Linear Regression and Random Forest to make prediction of prices using selected features, and evaluate each model using cross validation.

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

```
[ ]: ski_data.info
```

```
[ ]: <bound method DataFrame.info of                                   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
..                                 …      …      …          …
325               Meadowlark Ski Lodge  Wyoming  Wyoming         9500
326          Sleeping Giant Ski Resort  Wyoming  Wyoming         7428
327                   Snow King Resort  Wyoming  Wyoming         7808
328  Snowy Range Ski & Recreation Area  Wyoming  Wyoming         9663
329                White Pine Ski Area  Wyoming  Wyoming         9500

     vertical_drop  base_elev  trams  fastEight  fastSixes  fastQuads  …  \
0             2500        250      1        0.0          0          2  …
1             1540       1200      0        0.0          0          0  …
2              294       1796      0        0.0          0          0  …
3             2300       9200      0        0.0          1          0  …
4             1800       9200      0        NaN          0          1  …
..             …        …    …        …        …        …  …
325           1000       8500      0        NaN          0          0  …
326            810       6619      0        0.0          0          0  …
327           1571       6237      0        NaN          0          0  …
328            990       8798      0        0.0          0          0  …
329           1100       8400      0        NaN          0          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
..             …                …             …               …
325            1.5              300.0             NaN               NaN
326            1.0              184.0            18.0              61.0
327            1.0              400.0           250.0             121.0
328            0.7               75.0            30.0             131.0
329            0.4              370.0             NaN               NaN

     yearsOpen  averageSnowfall  AdultWeekday  AdultWeekend  \
0         60.0            669.0          65.0          85.0
1         44.0            350.0          47.0          53.0
2         36.0             69.0          30.0          34.0
3         81.0            260.0          89.0          89.0
4         49.0            250.0          74.0          78.0
..         …              …           …           …
325        9.0              NaN           NaN           NaN
326       81.0            310.0          42.0          42.0
327       80.0            300.0          59.0          59.0
328       59.0            250.0          49.0          49.0
329       81.0            150.0           NaN          49.0

     projectedDaysOpen  NightSkiing_ac
0                150.0           550.0
1                 90.0             NaN
2                152.0            30.0
3                122.0             NaN
4                104.0            80.0
..                 …             …
325                NaN             NaN
326               77.0             NaN
```

```
327              123.0           110.0
328                NaN             NaN
329                NaN             NaN

[330 rows x 27 columns]>
```

'AdultWeekday' is the price of an adult weekday ticket. 'AdultWeekend' is the price of an adult weekend ticket. They are the target of our project. The other columns are potential features that could be used to fit our model and predict outcomes.

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

```
[ ]:                   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  fastEight  fastSixes  fastQuads  …  LongestRun_mi  \
     0        250      1        0.0          0          2  …            1.0
     1       1200      0        0.0          0          0  …            2.0
     2       1796      0        0.0          0          0  …            1.0
     3       9200      0        0.0          1          0  …            2.0
     4       9200      0        NaN          0          1  …            1.2

        SkiableTerrain_ac  Snow Making_ac  daysOpenLastYear  yearsOpen  \
     0             1610.0           113.0             150.0       60.0
     1              640.0            60.0              45.0       44.0
     2               30.0            30.0             150.0       36.0
     3              777.0           104.0             122.0       81.0
     4              800.0            80.0             115.0       49.0

        averageSnowfall  AdultWeekday  AdultWeekend  projectedDaysOpen  \
     0            669.0          65.0          85.0              150.0
     1            350.0          47.0          53.0               90.0
     2             69.0          30.0          34.0              152.0
     3            260.0          89.0          89.0              122.0
     4            250.0          74.0          78.0              104.0

        NightSkiing_ac
     0           550.0
     1             NaN
     2            30.0
     3             NaN
     4            80.0
```

```
[5 rows x 27 columns]
```

Information on Big Mountain Resort

```
[ ]: ski_data[ski_data.Name == 'Big Mountain Resort'].T
```

```
[ ]:                                151
     Name               Big Mountain Resort
     Region                         Montana
     state                          Montana
     summit_elev                       6817
     vertical_drop                     2353
     base_elev                         4464
     trams                                0
     fastEight                          0.0
     fastSixes                            0
     fastQuads                            3
     quad                                 2
     triple                               6
     double                               0
     surface                              3
     total_chairs                        14
     Runs                             105.0
     TerrainParks                       4.0
     LongestRun_mi                      3.3
     SkiableTerrain_ac               3000.0
     Snow Making_ac                   600.0
     daysOpenLastYear                 123.0
     yearsOpen                         72.0
     averageSnowfall                  333.0
     AdultWeekday                      81.0
     AdultWeekend                      81.0
     projectedDaysOpen                123.0
     NightSkiing_ac                   600.0
```

```
[ ]: missing = pd.concat([ski_data.isnull().sum(), 100 * ski_data.isnull().mean()],␣
      ↪axis=1)
     missing.columns=['count', '%']
     missing.sort_values(by='count', ascending=False)
```

```
[ ]:                   count          %
     fastEight           166  50.303030
     NightSkiing_ac      143  43.333333
     AdultWeekday         54  16.363636
     AdultWeekend         51  15.454545
     daysOpenLastYear     51  15.454545
     TerrainParks         51  15.454545
```

```
projectedDaysOpen     47  14.242424
Snow Making_ac        46  13.939394
averageSnowfall       14   4.242424
LongestRun_mi          5   1.515152
Runs                   4   1.212121
SkiableTerrain_ac      3   0.909091
yearsOpen              1   0.303030
total_chairs           0   0.000000
Name                   0   0.000000
Region                 0   0.000000
double                 0   0.000000
triple                 0   0.000000
quad                   0   0.000000
fastQuads              0   0.000000
fastSixes              0   0.000000
trams                  0   0.000000
base_elev              0   0.000000
vertical_drop          0   0.000000
summit_elev            0   0.000000
state                  0   0.000000
surface                0   0.000000
```

```
[ ]: ski_data.select_dtypes('object')
```

```
[ ]:                                   Name   Region    state
     0                       Alyeska Resort   Alaska   Alaska
     1                  Eaglecrest Ski Area   Alaska   Alaska
     2                     Hilltop Ski Area   Alaska   Alaska
     3                     Arizona Snowbowl  Arizona  Arizona
     4                  Sunrise Park Resort  Arizona  Arizona
     ..                                 ...      ...      ...
     325               Meadowlark Ski Lodge  Wyoming  Wyoming
     326          Sleeping Giant Ski Resort  Wyoming  Wyoming
     327                   Snow King Resort  Wyoming  Wyoming
     328  Snowy Range Ski & Recreation Area  Wyoming  Wyoming
     329                White Pine Ski Area  Wyoming  Wyoming

     [330 rows x 3 columns]
```

```
[ ]: ski_data['Name'].value_counts().head()
```

```
[ ]: Crystal Mountain                 2
     Mount Bohemia                    1
     Anthony Lakes Mountain Resort    1
     Hunt Hollow Ski Club             1
     Ski Granby Ranch                 1
     Name: Name, dtype: int64
```

There is a duplicated resort name: Crystal Mountain.

```
[ ]: (ski_data['Name'] + ', ' + ski_data['Region']).value_counts().head()
```

```
[ ]: Ski Apache, New Mexico                        1
     Mount Pleasant of Edinboro, Pennsylvania      1
     Bolton Valley, Vermont                        1
     Mulligan's Hollow Ski Bowl, Michigan          1
     Mt. Bachelor, Oregon                          1
     dtype: int64
```

```
[ ]: (ski_data['Name'] + ', ' + ski_data['state']).value_counts().head()
```

```
[ ]: Kirkwood, California                         1
     Mulligan's Hollow Ski Bowl, Michigan         1
     Cranmore Mountain Resort, New Hampshire      1
     Sundown Mountain, Iowa                       1
     Marquette Mountain, Michigan                 1
     dtype: int64
```

```
[ ]: ski_data[ski_data['Name'] == 'Crystal Mountain']
```

```
[ ]:                  Name      Region       state  summit_elev  vertical_drop  \
     104  Crystal Mountain    Michigan    Michigan         1132            375
     295  Crystal Mountain  Washington  Washington         7012           3100

          base_elev  trams  fastEight  fastSixes  fastQuads  …  LongestRun_mi  \
     104        757      0        0.0          0          1  …            0.3
     295       4400      1        NaN          2          2  …            2.5

          SkiableTerrain_ac  Snow Making_ac  daysOpenLastYear  yearsOpen  \
     104              102.0            96.0             120.0       63.0
     295             2600.0            10.0               NaN       57.0

          averageSnowfall  AdultWeekday  AdultWeekend  projectedDaysOpen  \
     104            132.0          54.0          64.0              135.0
     295            486.0          99.0          99.0                NaN

          NightSkiing_ac
     104            56.0
     295             NaN

     [2 rows x 27 columns]
```

There are two different Crystal Mountain Resort

```
[ ]: (ski_data.Region != ski_data.state).count()
[ ]: 330
```

```
[ ]: ski_data['Region'].value_counts()
```

```
[ ]: New York               33
     Michigan               29
     Colorado               22
     Sierra Nevada          22
     Pennsylvania           19
     New Hampshire          16
     Wisconsin              16
     Vermont                15
     Minnesota              14
     Montana                12
     Idaho                  12
     Massachusetts          11
     Washington             10
     Maine                   9
     New Mexico              9
     Wyoming                 8
     Utah                    7
     Salt Lake City          6
     North Carolina          6
     Oregon                  6
     Connecticut             5
     Ohio                    5
     Illinois                4
     Virginia                4
     West Virginia           4
     Mt. Hood                4
     Alaska                  3
     Iowa                    3
     Nevada                  2
     South Dakota            2
     Arizona                 2
     Indiana                 2
     Missouri                2
     New Jersey              2
     Tennessee               1
     Maryland                1
     Northern California     1
     Rhode Island            1
     Name: Region, dtype: int64
```

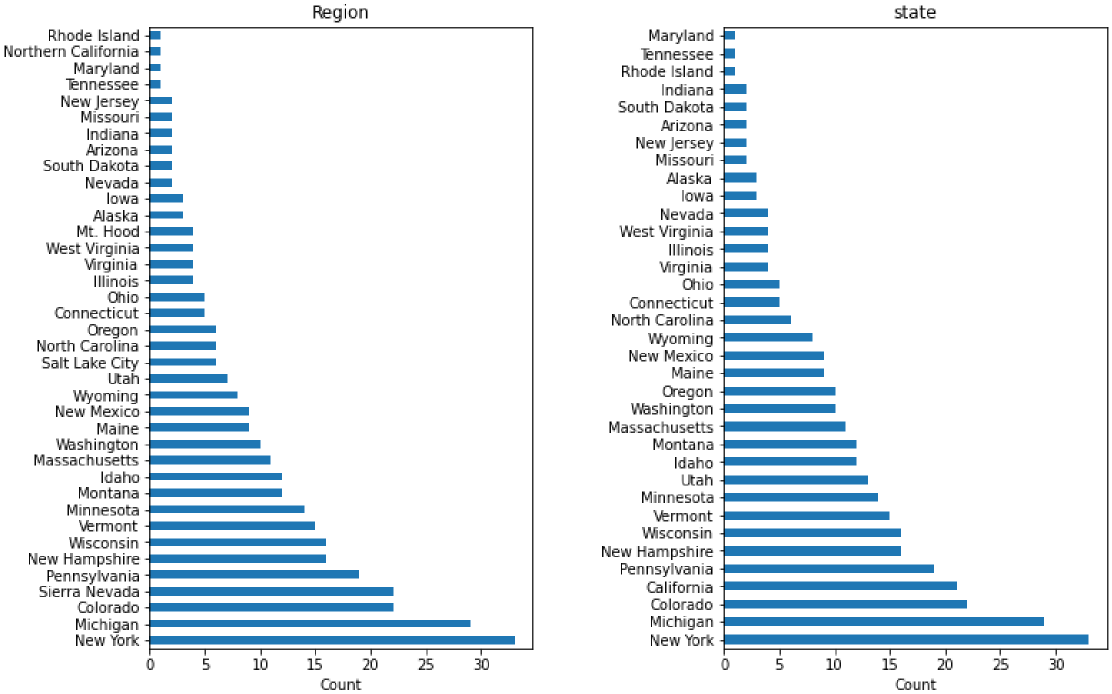
```
[ ]: (ski_data[ski_data.Region != ski_data.state]
      .groupby('state')['Region']
      .value_counts())
```

[ ]:
```
state       Region
California  Sierra Nevada          20
            Northern California     1
Nevada      Sierra Nevada           2
Oregon      Mt. Hood                4
Utah        Salt Lake City          6
Name: Region, dtype: int64
```

[ ]:
```
ski_data[['Region', 'state']].nunique()
```

[ ]:
```
Region    38
state     35
dtype: int64
```

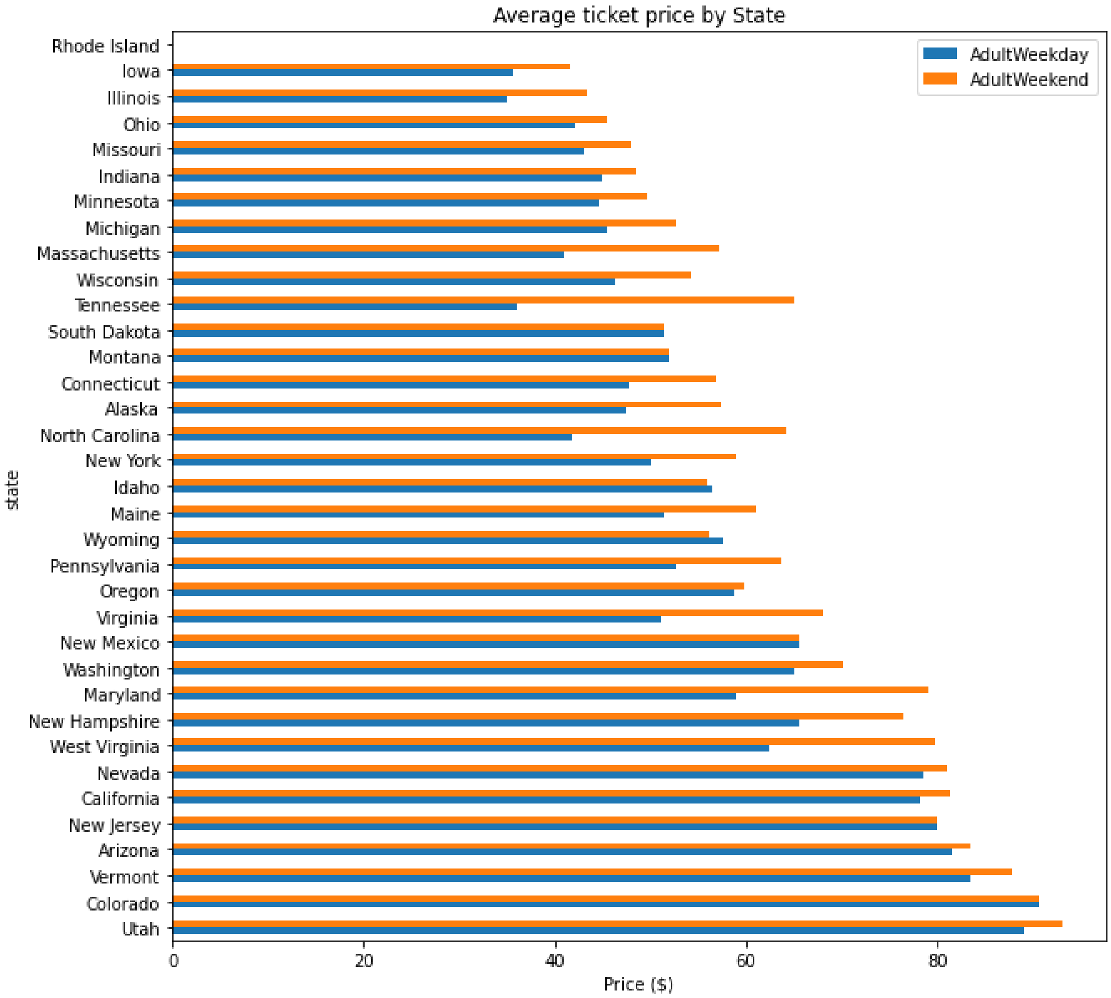
[ ]:
```
fig, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 8))
#Specify a horizontal barplot ('barh') as kind of plot (kind=)
ski_data.Region.value_counts().plot(kind='barh', ax=ax[0])
#Give the plot a helpful title of 'Region'
ax[0].set_title('Region')
#Label the xaxis 'Count'
ax[0].set_xlabel('Count')
#Specify a horizontal barplot ('barh') as kind of plot (kind=)
ski_data.state.value_counts().plot(kind='barh', ax=ax[1])
#Give the plot a helpful title of 'state'
ax[1].set_title('state')
#Label the xaxis 'Count'
ax[1].set_xlabel('Count')
#Give the subplots a little "breathing room" with a wspace of 0.5
plt.subplots_adjust(wspace=0.5);
```



```
[ ]: state_price_means = ski_data.groupby('state')[['AdultWeekday','AdultWeekend']].
      ↪mean()
     state_price_means.head()
```

```
[ ]:              AdultWeekday  AdultWeekend
     state
     Alaska          47.333333     57.333333
     Arizona         81.500000     83.500000
     California      78.214286     81.416667
     Colorado        90.714286     90.714286
     Connecticut     47.800000     56.800000
```

```
[ ]: (state_price_means.reindex(index=state_price_means.mean(axis=1)
         .sort_values(ascending=False)
         .index)
         .plot(kind='barh', figsize=(10, 10), title='Average ticket price by State'))
     plt.xlabel('Price ($)');
```



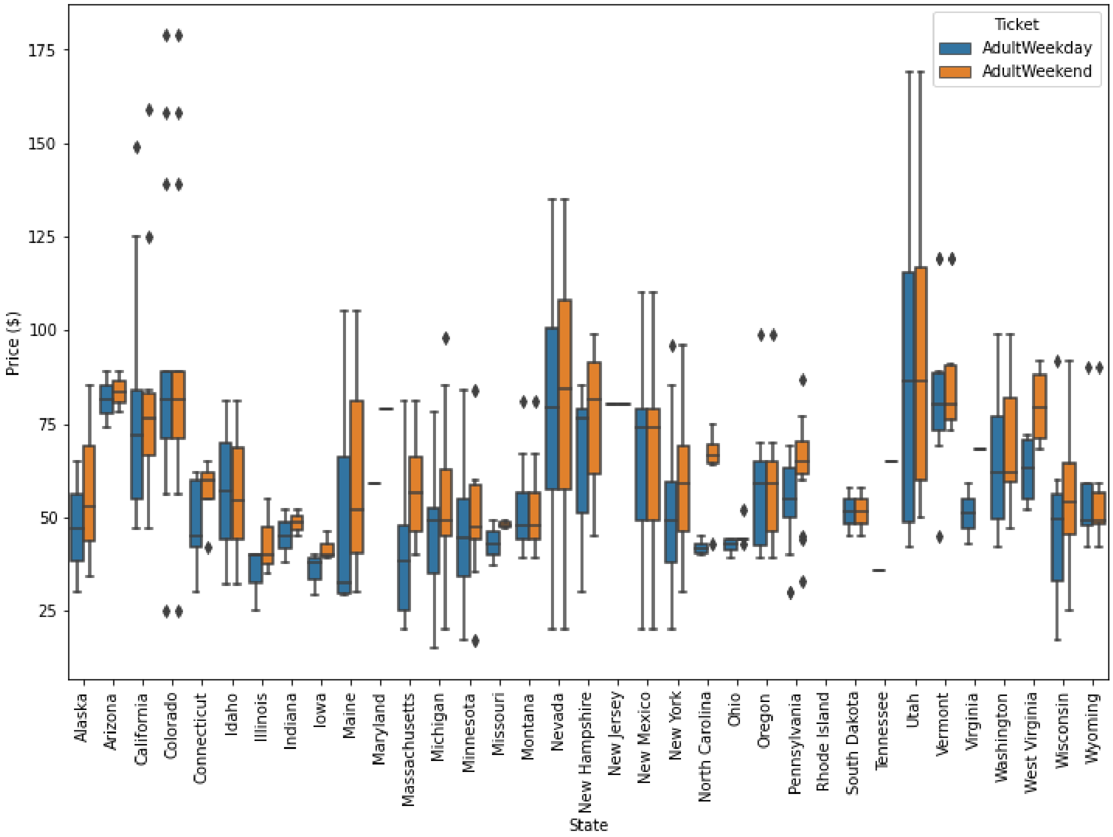
```
[ ]: ticket_prices = pd.melt(ski_data[['state', 'AdultWeekday','AdultWeekend']],
                             id_vars=['state'],
                             var_name= 'Ticket',
                             value_vars=['AdultWeekday','AdultWeekend'],
                             value_name='Price')
```

```
[ ]: ticket_prices.head()
```

```
[ ]:      state        Ticket  Price
     0   Alaska  AdultWeekday   65.0
     1   Alaska  AdultWeekday   47.0
     2   Alaska  AdultWeekday   30.0
     3  Arizona  AdultWeekday   89.0
     4  Arizona  AdultWeekday   74.0
```

[ ]:

```
plt.subplots(figsize=(12, 8))
sns.boxplot(x='state', y='Price', hue='Ticket', data=ticket_prices)
plt.xticks(rotation='vertical')
plt.ylabel('Price ($)')
plt.xlabel('State');
```



[ ]:

```
ski_data.describe().transpose()
```

[ ]:

```
                count         mean          std    min      25%     50%  \
summit_elev     330.0  4591.818182  3735.535934  315.0  1403.75  3127.5
vertical_drop   330.0  1215.427273   947.864557   60.0   461.25   964.5
base_elev       330.0  3374.000000  3117.121621   70.0   869.00  1561.5
trams           330.0     0.172727     0.559946    0.0     0.00     0.0
fastEight       164.0     0.006098     0.078087    0.0     0.00     0.0
fastSixes       330.0     0.184848     0.651685    0.0     0.00     0.0
fastQuads       330.0     1.018182     2.198294    0.0     0.00     0.0
quad            330.0     0.933333     1.312245    0.0     0.00     0.0
triple          330.0     1.500000     1.619130    0.0     0.00     1.0
double          330.0     1.833333     1.815028    0.0     1.00     1.0
surface         330.0     2.621212     2.059636    0.0     1.00     2.0
```

```
total_chairs        330.0    8.266667     5.798683   0.0     5.00    7.0
Runs                326.0   48.214724    46.364077   3.0    19.00   33.0
TerrainParks        279.0    2.820789     2.008113   1.0     1.00    2.0
LongestRun_mi       325.0    1.433231     1.156171   0.0     0.50    1.0
SkiableTerrain_ac   327.0  739.801223  1816.167441   8.0    85.00  200.0
Snow Making_ac      284.0  174.873239   261.336125   2.0    50.00  100.0
daysOpenLastYear    279.0  115.103943    35.063251   3.0    97.00  114.0
yearsOpen           329.0   63.656535   109.429928   6.0    50.00   58.0
averageSnowfall     316.0  185.316456   136.356842  18.0    69.00  150.0
AdultWeekday        276.0   57.916957    26.140126  15.0    40.00   50.0
AdultWeekend        279.0   64.166810    24.554584  17.0    47.00   60.0
projectedDaysOpen   283.0  120.053004    31.045963  30.0   100.00  120.0
NightSkiing_ac      187.0  100.395722   105.169620   2.0    40.00   72.0

                        75%      max
summit_elev         7806.00  13487.0
vertical_drop       1800.00   4425.0
base_elev           6325.25  10800.0
trams                  0.00      4.0
fastEight              0.00      1.0
fastSixes              0.00      6.0
fastQuads              1.00     15.0
quad                   1.00      8.0
triple                 2.00      8.0
double                 3.00     14.0
surface                3.00     15.0
total_chairs          10.00     41.0
Runs                  60.00    341.0
TerrainParks           4.00     14.0
LongestRun_mi          2.00      6.0
SkiableTerrain_ac    690.00  26819.0
Snow Making_ac       200.50   3379.0
daysOpenLastYear     135.00    305.0
yearsOpen             69.00   2019.0
averageSnowfall      300.00    669.0
AdultWeekday          71.00    179.0
AdultWeekend          77.50    179.0
projectedDaysOpen    139.50    305.0
NightSkiing_ac       114.00    650.0
```

[ ]:

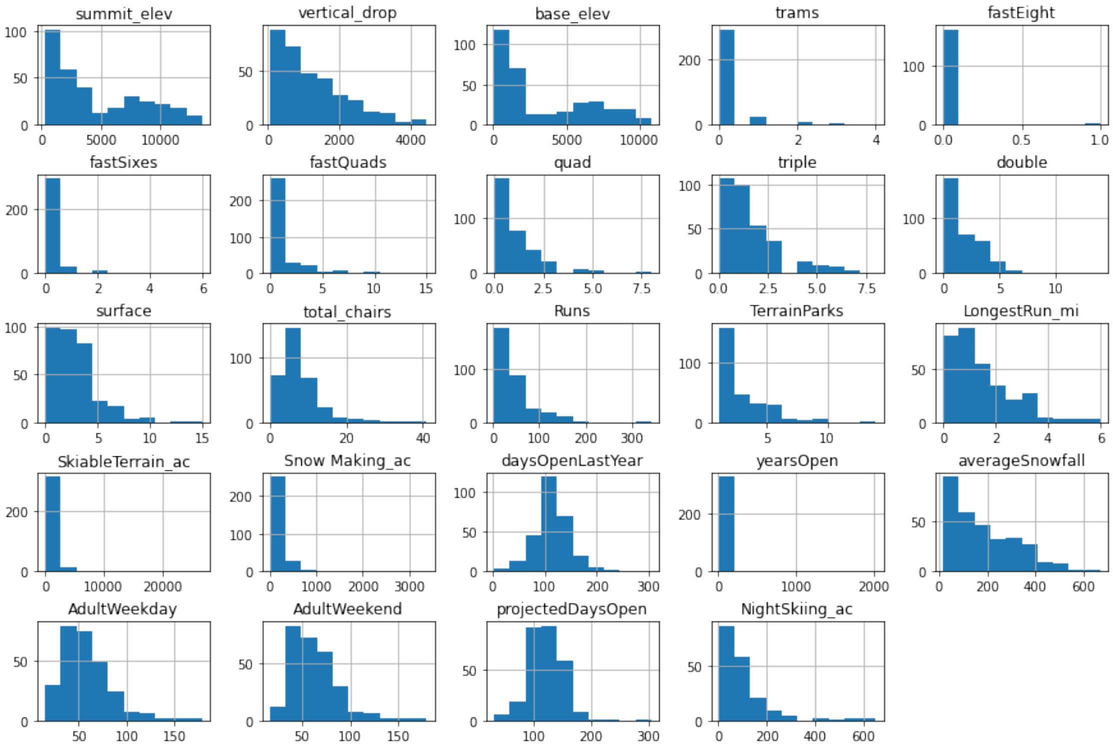
```
missing_price = ski_data[['AdultWeekend', 'AdultWeekday']].isnull().sum(axis=1)
missing_price.value_counts()/len(missing_price) * 100
```

[ ]:

```
0    82.424242
2    14.242424
1     3.333333
dtype: float64
```

[ ]:

```
ski_data.hist(figsize=(15,10))
plt.subplots_adjust(hspace=0.5);
```



[ ]:

```
ski_data.loc[ski_data.SkiableTerrain_ac > 10000]
```

[ ]:

```
                  Name    Region     state  summit_elev  vertical_drop  \
39  Silverton Mountain  Colorado  Colorado        13487           3087

    base_elev  trams  fastEight  fastSixes  fastQuads  …  LongestRun_mi  \
39      10400      0        0.0          0          0  …            1.5

    SkiableTerrain_ac  Snow Making_ac  daysOpenLastYear  yearsOpen  \
39            26819.0             NaN             175.0       17.0

    averageSnowfall  AdultWeekday  AdultWeekend  projectedDaysOpen  \
39            400.0          79.0          79.0              181.0

    NightSkiing_ac
39             NaN

[1 rows x 27 columns]
```

[ ]:

```
ski_data[ski_data. SkiableTerrain_ac > 10000].transpose()
```

```
[ ]:                                39
     Name                 Silverton Mountain
     Region                         Colorado
     state                          Colorado
     summit_elev                       13487
     vertical_drop                      3087
     base_elev                         10400
     trams                                 0
     fastEight                           0.0
     fastSixes                             0
     fastQuads                             0
     quad                                  0
     triple                                0
     double                                1
     surface                               0
     total_chairs                          1
     Runs                                NaN
     TerrainParks                        NaN
     LongestRun_mi                       1.5
     SkiableTerrain_ac               26819.0
     Snow Making_ac                      NaN
     daysOpenLastYear                  175.0
     yearsOpen                          17.0
     averageSnowfall                   400.0
     AdultWeekday                       79.0
     AdultWeekend                       79.0
     projectedDaysOpen                 181.0
     NightSkiing_ac                      NaN
```

```
[ ]: ski_data.loc[39, 'SkiableTerrain_ac']
```

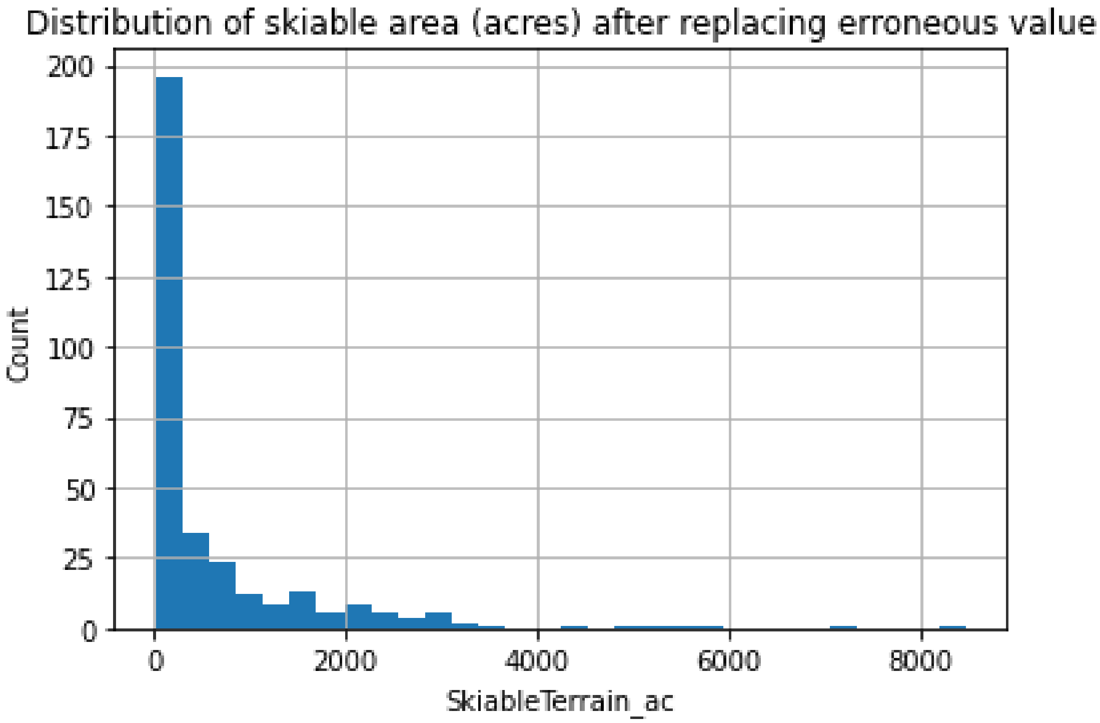
```
[ ]: 26819.0
```

```
[ ]: ski_data.loc[39, 'SkiableTerrain_ac'] = 1819
```

```
[ ]: ski_data.loc[39, 'SkiableTerrain_ac']
```

```
[ ]: 1819.0
```

```
[ ]: ski_data.SkiableTerrain_ac.hist(bins=30)
     plt.xlabel('SkiableTerrain_ac')
     plt.ylabel('Count')
     plt.title('Distribution of skiable area (acres) after replacing erroneous␣
      ↪value');
```



```
[ ]: ski_data['Snow Making_ac'][ski_data['Snow Making_ac'] > 1000]
```

```
[ ]: 11    3379.0
     18    1500.0
     Name: Snow Making_ac, dtype: float64
```

```
[ ]: ski_data[ski_data['Snow Making_ac'] > 3000].T
```

```
[ ]:                                     11
     Name               Heavenly Mountain Resort
     Region                        Sierra Nevada
     state                            California
     summit_elev                           10067
     vertical_drop                          3500
     base_elev                              7170
     trams                                     2
     fastEight                               0.0
     fastSixes                                 2
     fastQuads                                 7
     quad                                      1
     triple                                    5
     double                                    3
     surface                                   8
```

```
total_chairs                          28
Runs                                97.0
TerrainParks                         3.0
LongestRun_mi                        5.5
SkiableTerrain_ac                 4800.0
Snow Making_ac                    3379.0
daysOpenLastYear                   155.0
yearsOpen                           64.0
averageSnowfall                    360.0
AdultWeekday                         NaN
AdultWeekend                         NaN
projectedDaysOpen                  157.0
NightSkiing_ac                       NaN
```

```
[ ]: .6 * 4800
```

```
[ ]: 2880.0
```

```
[ ]: ski_data.fastEight.value_counts()
```

```
[ ]: 0.0    163
     1.0      1
     Name: fastEight, dtype: int64
```

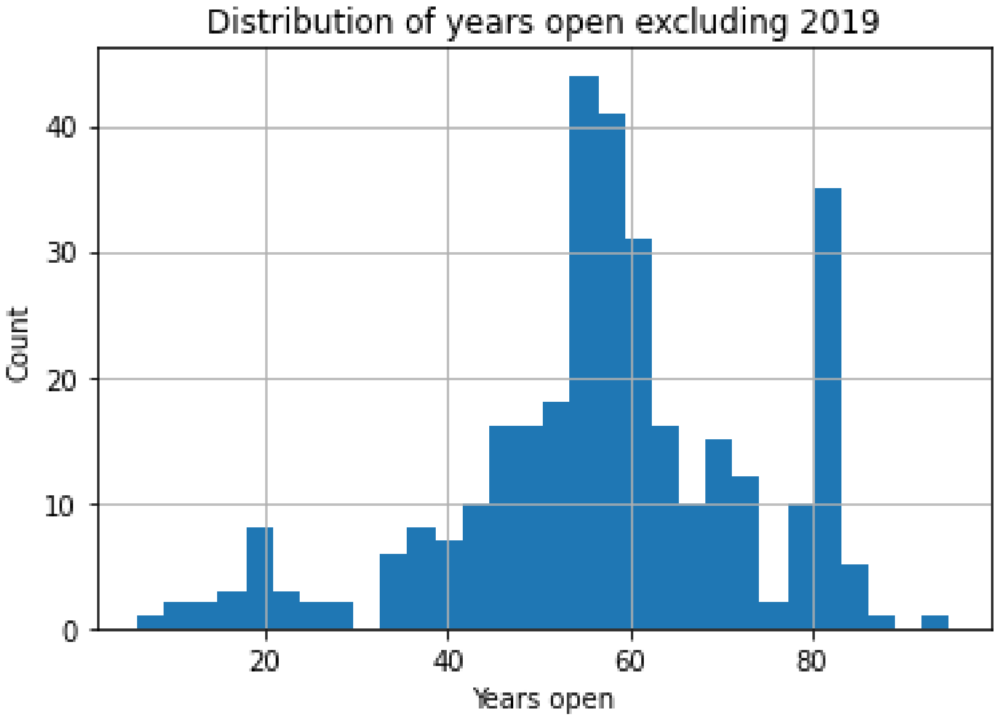
```
[ ]: ski_data.drop(columns='fastEight', inplace=True)
```

```
[ ]: ski_data.loc[ski_data.yearsOpen > 100]
```

```
[ ]:                     Name    Region     state  summit_elev  vertical_drop  \
     34          Howelsen Hill  Colorado  Colorado         7136            440
     115  Pine Knob Ski Resort  Michigan  Michigan         1308            300

          base_elev  trams  fastSixes  fastQuads  quad  …  LongestRun_mi  \
     34        6696      0          0          0     0  …            6.0
     115       1009      0          0          0     0  …            1.0

          SkiableTerrain_ac  Snow Making_ac  daysOpenLastYear  yearsOpen  \
     34                50.0            25.0             100.0      104.0
     115               80.0            80.0               NaN     2019.0

          averageSnowfall  AdultWeekday  AdultWeekend  projectedDaysOpen  \
     34             150.0          25.0          25.0              100.0
     115              NaN          49.0          57.0                NaN

          NightSkiing_ac
     34             10.0
     115             NaN
```

```
[2 rows x 26 columns]
```

[ ]:
```
ski_data['yearsOpen'].loc[ski_data.yearsOpen < 100].hist(bins=30)
plt.xlabel('Years open')
plt.ylabel('Count')
plt.title('Distribution of years open excluding 2019');
```



[ ]:
```
ski_data.yearsOpen[ski_data.yearsOpen < 1000].describe()
```

[ ]:
```
count     328.000000
mean       57.695122
std        16.841182
min         6.000000
25%        50.000000
50%        58.000000
75%        68.250000
max       104.000000
Name: yearsOpen, dtype: float64
```

[ ]:
```
ski_data = ski_data[ski_data.yearsOpen < 1000]
```

[ ]:

```
state_summary = ski_data.groupby('state').agg(
    resorts_per_state=pd.NamedAgg(column='Name', aggfunc='size'), #could pick␣
 ↪any column here
    state_total_skiable_area_ac=pd.NamedAgg(column='SkiableTerrain_ac',␣
 ↪aggfunc='sum'),
    state_total_days_open=pd.NamedAgg(column='daysOpenLastYear', aggfunc='sum'),
    state_total_terrain_parks=pd.NamedAgg(column='TerrainParks', aggfunc='sum'),
    state_total_nightskiing_ac=pd.NamedAgg(column='NightSkiing_ac',␣
 ↪aggfunc='sum')
).reset_index()
state_summary.head()
```

[ ]:

```
         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
0                       580.0
1                        80.0
2                       587.0
3                       428.0
4                       256.0
```

[ ]:

```
missing_price = ski_data[['AdultWeekend', 'AdultWeekday']].isnull().sum(axis=1)
missing_price.value_counts()/len(missing_price) * 100
```

[ ]:

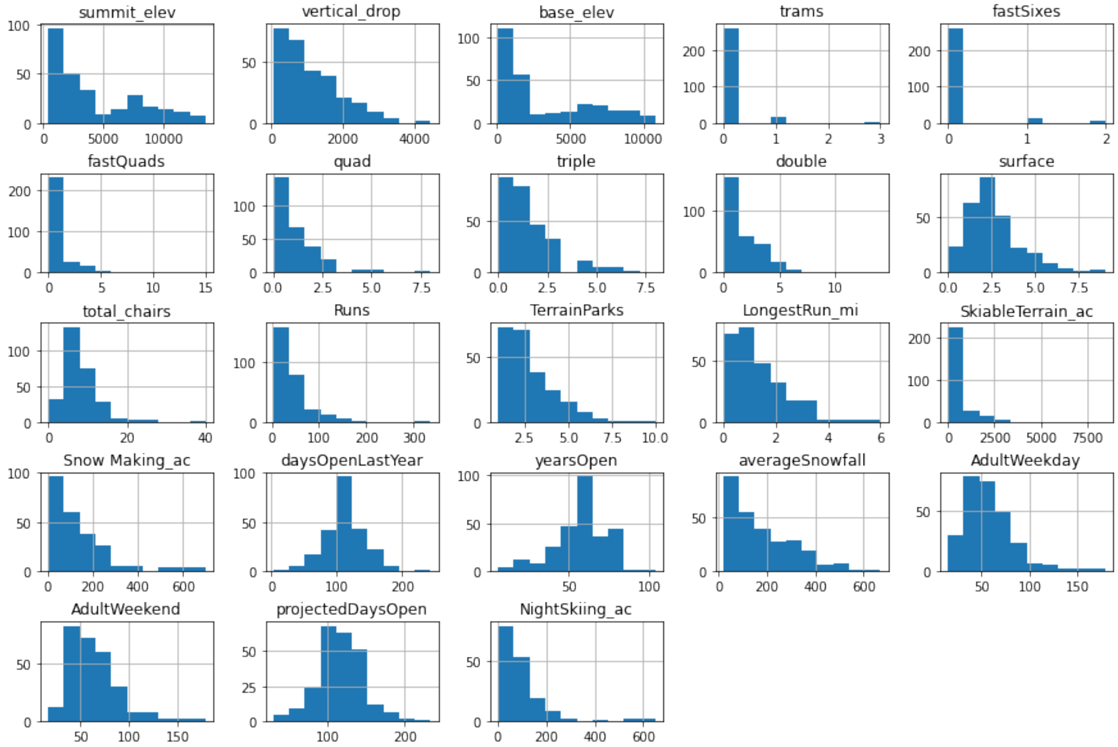
```
0    82.317073
2    14.329268
1     3.353659
dtype: float64
```

[ ]:

```
ski_data = ski_data[missing_price != 2]
```

[ ]:

```
ski_data.hist(figsize=(15, 10))
plt.subplots_adjust(hspace=0.5);
```



```
[ ]: states_url = 'https://simple.wikipedia.org/w/index.php?title=List_of_U.S.
      ↪_states&oldid=7168473'
     usa_states = pd.read_html(states_url)
```

```
[ ]: type(usa_states)
```

```
[ ]: list
```

```
[ ]: len(usa_states)
```

```
[ ]: 1
```

```
[ ]: usa_states = usa_states[0]
     usa_states.head()
```

```
[ ]:   Name &postal abbs. [1]                          Cities              \
       Name &postal abbs. [1] Name &postal abbs. [1].1       Capital   Largest[5]
     0                Alabama                       AL    Montgomery   Birmingham
     1                 Alaska                       AK        Juneau    Anchorage
     2                Arizona                       AZ       Phoenix      Phoenix
     3               Arkansas                       AR   Little Rock  Little Rock
     4             California                       CA    Sacramento  Los Angeles
```

```
  Established[A] Population[B][3] Total area[4]          Land area[4]  \
  Established[A] Population[B][3]           mi2      km2          mi2
0   Dec 14, 1819          4903185         52420   135767        50645
1    Jan 3, 1959           731545        665384  1723337       570641
2   Feb 14, 1912          7278717        113990   295234       113594
3   Jun 15, 1836          3017804         53179   137732        52035
4    Sep 9, 1850         39512223        163695   423967       155779

           Water area[4]         Numberof Reps.
       km2           mi2     km2 Numberof Reps.
0   131171          1775    4597              7
1  1477953         94743  245384              1
2   294207           396    1026              9
3   134771          1143    2961              4
4   403466          7916   20501             53
```

```
[ ]: established = usa_states.iloc[:, 4]
```

```
[ ]: established
```

```
[ ]: 0      Dec 14, 1819
     1       Jan 3, 1959
     2      Feb 14, 1912
     3      Jun 15, 1836
     4       Sep 9, 1850
     5       Aug 1, 1876
     6       Jan 9, 1788
     7       Dec 7, 1787
     8       Mar 3, 1845
     9       Jan 2, 1788
     10     Aug 21, 1959
     11      Jul 3, 1890
     12      Dec 3, 1818
     13     Dec 11, 1816
     14     Dec 28, 1846
     15     Jan 29, 1861
     16      Jun 1, 1792
     17     Apr 30, 1812
     18     Mar 15, 1820
     19     Apr 28, 1788
     20      Feb 6, 1788
     21     Jan 26, 1837
     22     May 11, 1858
     23     Dec 10, 1817
     24     Aug 10, 1821
     25      Nov 8, 1889
     26      Mar 1, 1867
```

```
27    Oct 31, 1864
28    Jun 21, 1788
29    Dec 18, 1787
30     Jan 6, 1912
31    Jul 26, 1788
32    Nov 21, 1789
33     Nov 2, 1889
34     Mar 1, 1803
35    Nov 16, 1907
36    Feb 14, 1859
37    Dec 12, 1787
38    May 29, 1790
39    May 23, 1788
40     Nov 2, 1889
41     Jun 1, 1796
42    Dec 29, 1845
43     Jan 4, 1896
44     Mar 4, 1791
45    Jun 25, 1788
46    Nov 11, 1889
47    Jun 20, 1863
48    May 29, 1848
49    Jul 10, 1890
Name: (Established[A], Established[A]), dtype: object
```

```
[ ]: usa_states_sub = usa_states.iloc[:, [0, 5, 6]].copy()
     usa_states_sub.columns = ['state', 'state_population', 'state_area_sq_miles']
     usa_states_sub.head()
```

```
[ ]:         state  state_population  state_area_sq_miles
     0     Alabama           4903185                52420
     1      Alaska            731545               665384
     2     Arizona           7278717               113990
     3    Arkansas           3017804                53179
     4  California          39512223               163695
```

```
[ ]: missing_states = set(state_summary.state) - set(usa_states_sub.state)
     missing_states
```

```
[ ]: {'Massachusetts', 'Pennsylvania', 'Rhode Island', 'Virginia'}
```

```
[ ]: usa_states_sub.state[usa_states_sub.state.str.
      ↪contains('Massachusetts|Pennsylvania|Rhode Island|Virginia')]
```

```
[ ]: 20    Massachusetts[C]
     37     Pennsylvania[C]
     38     Rhode Island[D]
```

```
45         Virginia[C]
47       West Virginia
Name: state, dtype: object
```

[ ]:
```
usa_states_sub.state.replace(to_replace='\[.*\]', value='', regex=True,␣
 ↪inplace=True)
usa_states_sub.state[usa_states_sub.state.str.
 ↪contains('Massachusetts|Pennsylvania|Rhode Island|Virginia')]
```

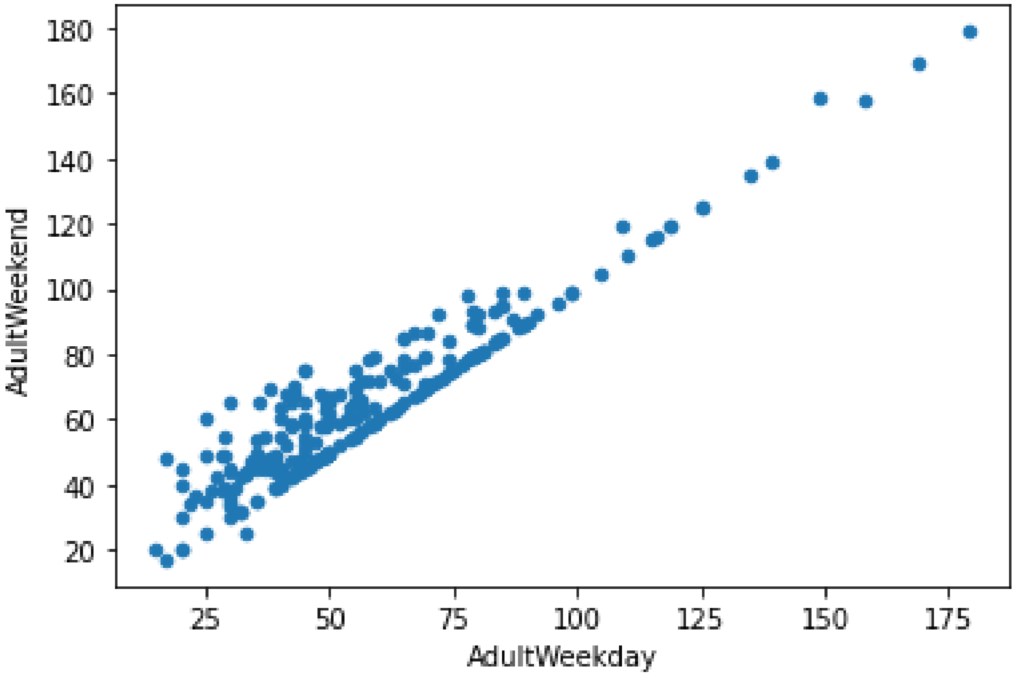
[ ]:
```
20    Massachusetts
37     Pennsylvania
38     Rhode Island
45         Virginia
47    West Virginia
Name: state, dtype: object
```

[ ]:
```
missing_states = set(state_summary.state) - set(usa_states_sub.state)
missing_states
```

[ ]:
```
state_summary = state_summary.merge(usa_states_sub, how='left', on='state')
state_summary.head()
```

[ ]:
```
         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  state_population  state_area_sq_miles
0                       580.0            731545               665384
1                        80.0           7278717               113990
2                       587.0          39512223               163695
3                       428.0           5758736               104094
4                       256.0           3565278                 5543
```

[ ]:
```
ski_data.plot(x='AdultWeekday', y='AdultWeekend', kind='scatter');
```



```
[ ]: ski_data.loc[ski_data.state == 'Montana', ['AdultWeekend', 'AdultWeekday']]
```

```
[ ]:      AdultWeekend  AdultWeekday
     141          42.0          42.0
     142          63.0          63.0
     143          49.0          49.0
     144          48.0          48.0
     145          46.0          46.0
     146          39.0          39.0
     147          50.0          50.0
     148          67.0          67.0
     149          47.0          47.0
     150          39.0          39.0
     151          81.0          81.0
```

```
[ ]: ski_data[['AdultWeekend', 'AdultWeekday']].isnull().sum()
```

```
[ ]: AdultWeekend    4
     AdultWeekday    7
     dtype: int64
```

```
[ ]: ski_data.drop(columns='AdultWeekday', inplace=True)
     ski_data.dropna(subset=['AdultWeekend'], inplace=True)
```

```
[ ]: ski_data.shape
[ ]: (277, 25)
```

```
[ ]: missing = pd.concat([ski_data.isnull().sum(axis=1), 100 * ski_data.isnull().
      ↪mean(axis=1)], axis=1)
     missing.columns=['count', '%']
     missing.sort_values(by='count', ascending=False).head(10)
```

```
[ ]:      count     %
     329      5  20.0
     62       5  20.0
     141      5  20.0
     86       5  20.0
     74       5  20.0
     146      5  20.0
     184      4  16.0
     108      4  16.0
     198      4  16.0
     39       4  16.0
```

```
[ ]: missing['%'].unique()
```

```
[ ]: array([ 0.,  4.,  8., 12., 16., 20.])
```

```
[ ]: missing['%'].value_counts()
```

```
[ ]: 0.0     107
     4.0      94
     8.0      45
     12.0     15
     16.0     10
     20.0      6
     Name: %, dtype: int64
```

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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 277 entries, 0 to 329
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: 56.3+ KB
```

```
[ ]: ski_data.shape
```

```
[ ]: (277, 25)
```

```
[ ]: # save the data to a new csv file
     ski_data.to_csv('ski_data_cleaned.csv',index=False)
```

```
[ ]: # save the state_summary separately.
     state_summary.to_csv('state_summary.csv', index = False)
```

In the original data, there are 329 rows and 27 columns with information on 277 skiing resorts across the nation. After our observation of the histograms of the numeric features of the resorts, it is obvious that some of them are not very plausible.

To begin with, we can see the data for Skiable Terrain_ac are clustering down below 10,000. To investigate further, we can print out the resorts with the value of Skiable Terrain_ac greater than 10,000. It turns out that there is only one resort, Silverton Mountain, has more than 10,000 acres of skiable terrian. The value for Silverton Mountain is 26819, which is suspiciously high compared to other resorts. By searching "silverton mountain skiable area", we can find that the real skiable terrain value for Silverton Mountain is 1819 instead of 26819. We can replace the wrong value with the right one using the .loc acceesor. We can see that the new plot makes more sense with the value change. For the same reason, we delete one row from the data where yearsOpen is 2019.

The fastEight's plot is also strange. Most of the values are 0 and a lot of values are null, which means this feature won't provide us with little information. We can drop the whole column from our data. For the same reason, we drop the rows with no price data.

We also need to set a target feature.Since the goal of the project is to provide a better pricing strategy, we want to use the value of AdultWeekday or AdultWeekend. By plotting a graph, we can see the relationship between AdultWeekday or AdultWeekend is linear (i.e. the higher the value of AdultWeekday, the higher the value of AdultWeekend.) Therefore, we can drop one of the prices. Since AdultWeekend has the least missing value of the two, we choose to drop the AdultWeekday.

After wrangling with the data, we have 277 rows and 25 columns left. Created in Deepnote