lec10

September 20, 2021

[1]:

```
from datascience import *
import numpy as np

%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
```

0.1 Lecture 10

0.2 Apply with Multiple Columns

[2]:

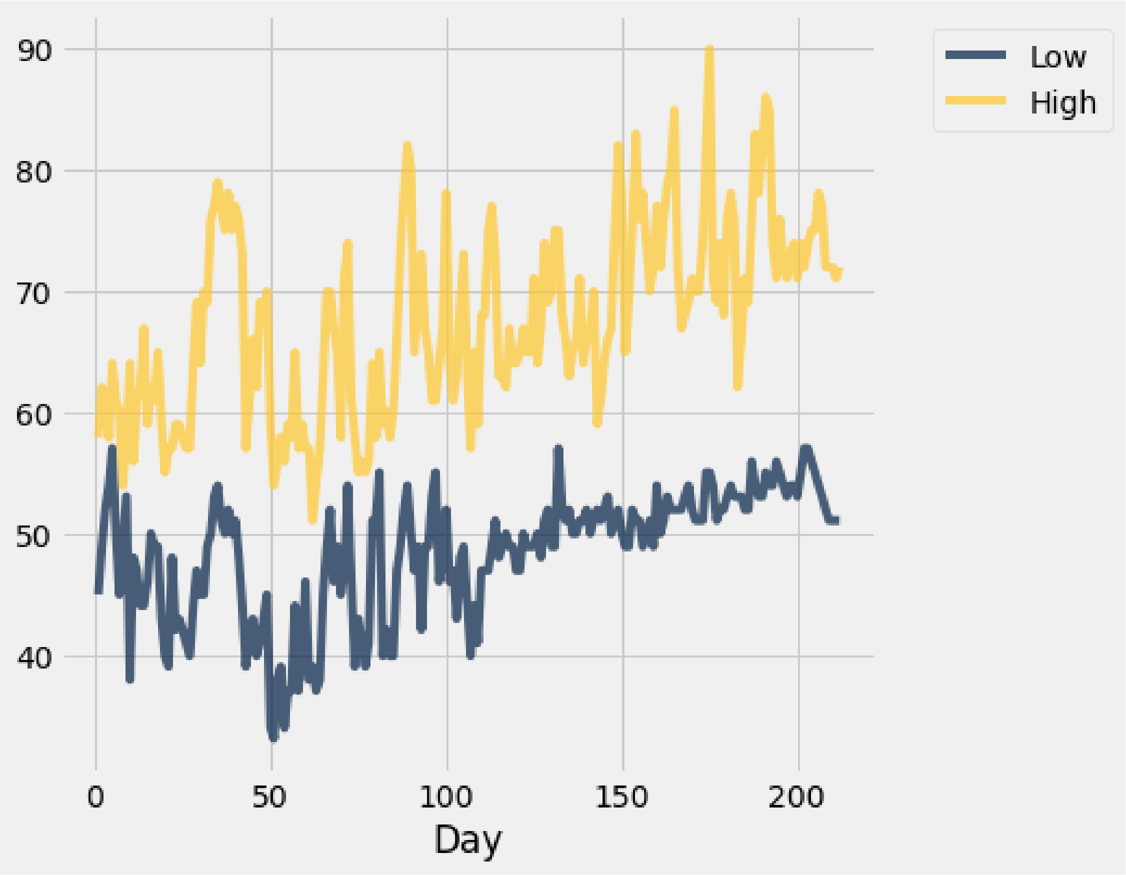
```
temperatures = Table.read_table('temperatures.csv')
temperatures
```

[2]:

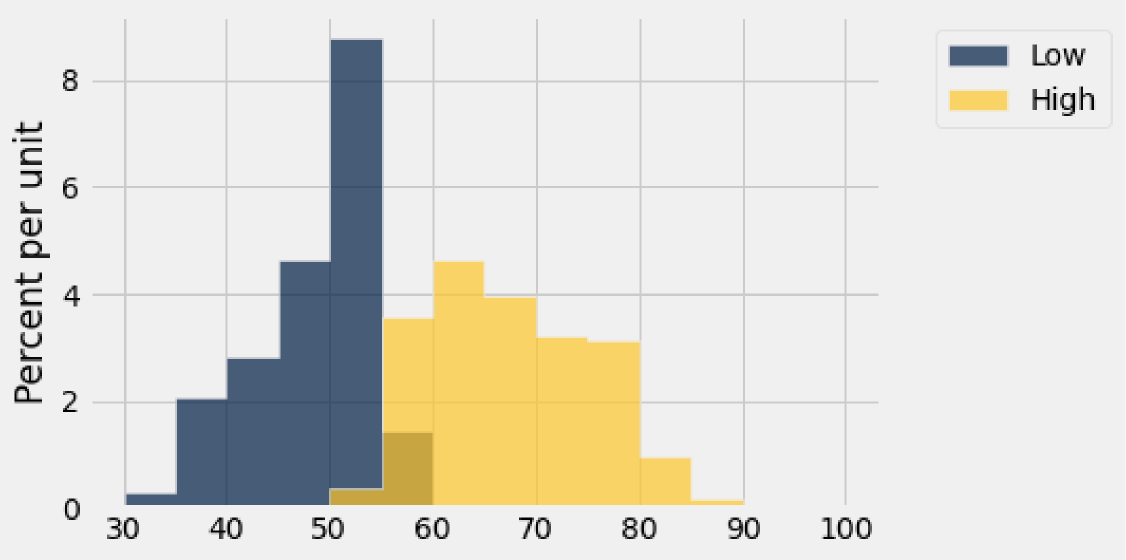
```
Day  | Low   | High
1    | 44.96 | 57.92
2    | 48.92 | 62.06
3    | 51.98 | 60.98
4    | 53.96 | 57.92
5    | 57.02 | 64.04
6    | 50    | 60.98
7    | 44.96 | 60.08
8    | 48.92 | 53.96
9    | 53.06 | 57.92
10   | 37.94 | 64.04
… (202 rows omitted)
```

[3]:

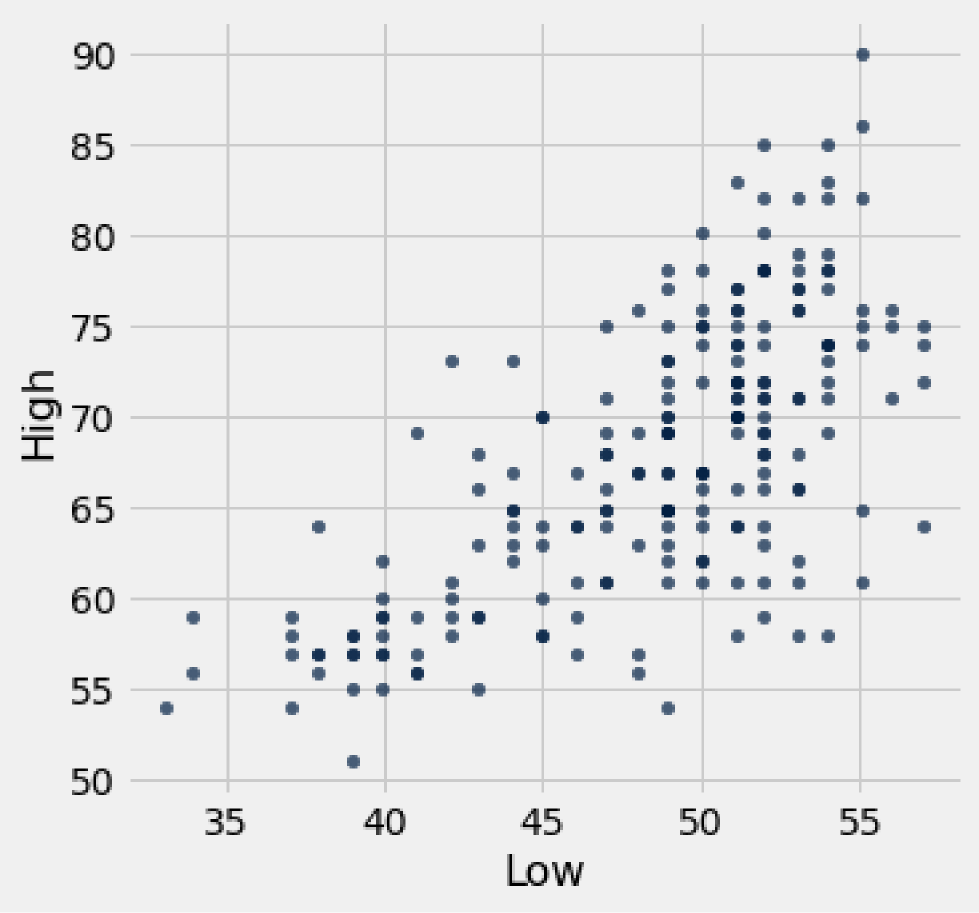
```
temperatures.plot('Day')
```



```
[4]: temperatures.select('Low', 'High').hist(bins=np.arange(30, 105, 5))
```



[5]:
```
temperatures.scatter('Low', 'High')
```



[6]:
```
# Difference between high temp and low temp

def difference(x, y):
    return x-y

difference(65, 54)
```

[6]:
```
11
```

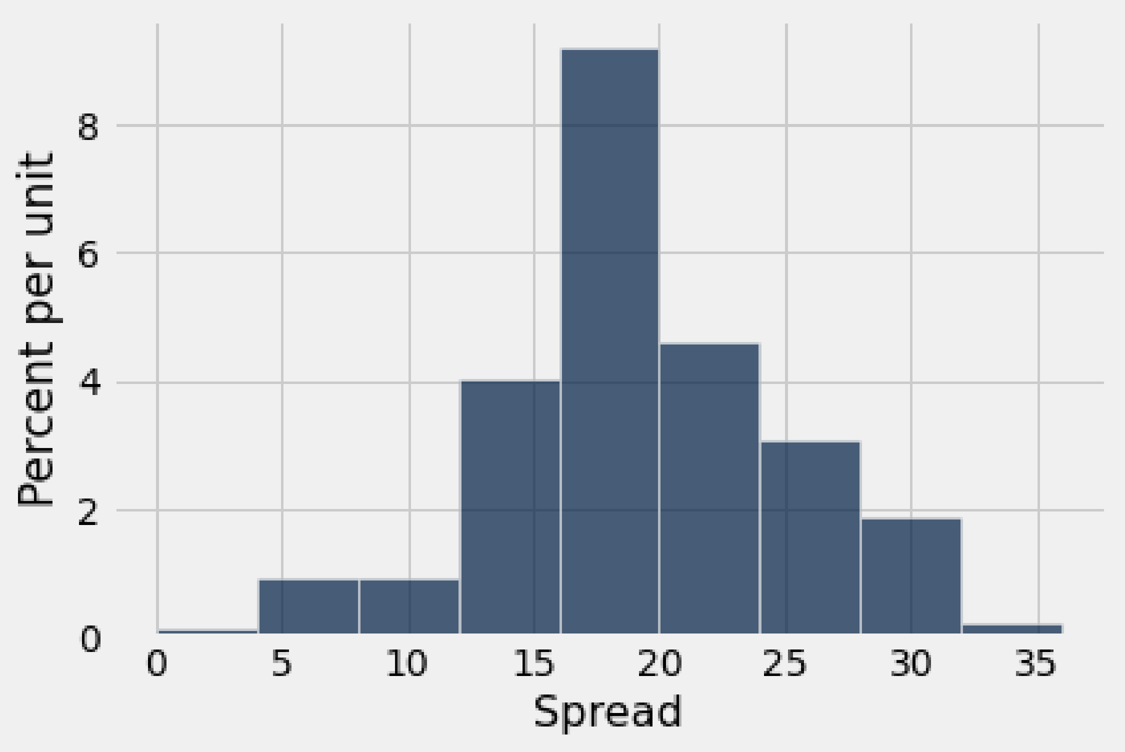
[7]:
```
daily_spread = temperatures.apply(difference, 'High', 'Low')
temperatures = temperatures.with_column('Spread', daily_spread)
temperatures
```

[7]:
```
Day   | Low   | High  | Spread
1     | 44.96 | 57.92 | 12.96
2     | 48.92 | 62.06 | 13.14
3     | 51.98 | 60.98 | 9
```

```
4    | 53.96 | 57.92 | 3.96
5    | 57.02 | 64.04 | 7.02
6    | 50    | 60.98 | 10.98
7    | 44.96 | 60.08 | 15.12
8    | 48.92 | 53.96 | 5.04
9    | 53.06 | 57.92 | 4.86
10   | 37.94 | 64.04 | 26.1
… (202 rows omitted)
```

[8]:

```
temperatures.hist('Spread', bins=np.arange(0, 40, 4))
```



[9]:

```
temperatures.where('Spread', are.above(20)).num_rows / temperatures.num_rows
```

[9]:

```
0.3915094339622642
```

0.3 Function with Optional Arguments

[10]:

```
def percents(s, places):
    return np.round(s/sum(s) * 100, places)
```

[11]:

```
x = make_array(2, 5, 16)
percents(x, 4)
```

```
[11]: array([ 8.6957, 21.7391, 69.5652])
```

```
[12]: def percents(s, places=2):
          return np.round(s/sum(s) * 100, places)
```

```
[13]: percents(x)
```

```
[13]: array([ 8.7 , 21.74, 69.57])
```

0.4 Grouping by Category

```
[14]: all_cones = Table.read_table('cones.csv')
      all_cones
```

```
[14]: Flavor     | Color       | Price
      strawberry | pink        | 3.55
      chocolate  | light brown | 4.75
      chocolate  | dark brown  | 5.25
      strawberry | pink        | 5.25
      chocolate  | dark brown  | 5.25
      bubblegum  | pink        | 4.75
```

```
[15]: cones = all_cones.drop('Color').exclude(5)
      cones
```

```
[15]: Flavor     | Price
      strawberry | 3.55
      chocolate  | 4.75
      chocolate  | 5.25
      strawberry | 5.25
      chocolate  | 5.25
```

```
[16]: cones.group('Flavor')
```

```
[16]: Flavor     | count
      chocolate  | 3
      strawberry | 2
```

```
[17]: cones.group('Flavor', min)
```

```
[17]: Flavor     | Price min
      chocolate  | 4.75
      strawberry | 3.55
```

```
[18]: cones.group('Flavor', list)
```

```
C:\Users\schoend\Anaconda3\lib\site-packages\datascience\tables.py:920:
VisibleDeprecationWarning: Creating an ndarray from ragged nested sequences
(which is a list-or-tuple of lists-or-tuples-or ndarrays with different lengths
or shapes) is deprecated. If you meant to do this, you must specify
'dtype=object' when creating the ndarray
  values = np.array(tuple(values))
```

```
[18]: Flavor     | Price list
      chocolate  | [4.75, 5.25, 5.25]
      strawberry | [3.55, 5.25]
```

```
[19]: cones.group('Flavor', np.average)
```

```
[19]: Flavor     | Price average
      chocolate  | 5.08333
      strawberry | 4.4
```

```
[20]: cones.group('Flavor', min)
```

```
[20]: Flavor     | Price min
      chocolate  | 4.75
      strawberry | 3.55
```

```
[21]: min(cones.where('Flavor', 'chocolate').column('Price'))
```

```
[21]: 4.75
```

```
[22]: def spread(arr):
          return max(arr) - min(arr)

      spread(make_array(7, 10, 2))
```

```
[22]: 8
```

```
[23]: cones.group('Flavor', spread)
```

```
[23]: Flavor     | Price spread
      chocolate  | 0.5
      strawberry | 1.7
```

```
[24]: cones
```

```
[24]: Flavor     | Price
      strawberry | 3.55
      chocolate  | 4.75
      chocolate  | 5.25
      strawberry | 5.25
      chocolate  | 5.25
```

```
[25]: all_cones
```

```
[25]: Flavor     | Color       | Price
      strawberry | pink        | 3.55
      chocolate  | light brown | 4.75
      chocolate  | dark brown  | 5.25
      strawberry | pink        | 5.25
      chocolate  | dark brown  | 5.25
      bubblegum  | pink        | 4.75
```

```
[26]: all_cones.group(['Flavor', 'Color'])
```

```
[26]: Flavor     | Color       | count
      bubblegum  | pink        | 1
      chocolate  | dark brown  | 2
      chocolate  | light brown | 1
      strawberry | pink        | 2
```

```
[27]: all_cones.group(['Flavor', 'Color'], np.average)
```

```
[27]: Flavor     | Color       | Price average
      bubblegum  | pink        | 4.75
      chocolate  | dark brown  | 5.25
      chocolate  | light brown | 4.75
      strawberry | pink        | 4.4
```

0.5 Examples

```
[28]: nba = Table.read_table('nba_salaries.csv').relabeled(3, 'SALARY')
      nba
```

```
[28]: PLAYER            | POSITION | TEAM          | SALARY
      Paul Millsap      | PF       | Atlanta Hawks | 18.6717
      Al Horford        | C        | Atlanta Hawks | 12
      Tiago Splitter    | C        | Atlanta Hawks | 9.75625
      Jeff Teague       | PG       | Atlanta Hawks | 8
      Kyle Korver       | SG       | Atlanta Hawks | 5.74648
      Thabo Sefolosha   | SF       | Atlanta Hawks | 4
      Mike Scott        | PF       | Atlanta Hawks | 3.33333
      Kent Bazemore     | SF       | Atlanta Hawks | 2
      Dennis Schroder   | PG       | Atlanta Hawks | 1.7634
      Tim Hardaway Jr.  | SG       | Atlanta Hawks | 1.30452
      … (407 rows omitted)
```

```
[29]: # total salary paid by each team, highest first
```

```
nba.select('TEAM', 'SALARY').group('TEAM', sum).sort('SALARY sum',␣
 ↪descending=True)
```

[29]:

```
TEAM                  | SALARY sum
Cleveland Cavaliers   | 102.312
Oklahoma City Thunder | 96.8322
Golden State Warriors | 94.0851
Memphis Grizzlies     | 93.7964
Washington Wizards    | 90.0475
Houston Rockets       | 85.2858
San Antonio Spurs     | 84.6521
Charlotte Hornets     | 84.1024
Miami Heat            | 81.5287
New Orleans Pelicans  | 80.5146
… (20 rows omitted)
```

[30]:

```
nba.group('TEAM', sum)
```

[30]:

```
TEAM                  | PLAYER sum | POSITION sum | SALARY sum
Atlanta Hawks         |            |              | 69.5731
Boston Celtics        |            |              | 50.2855
Brooklyn Nets         |            |              | 57.307
Charlotte Hornets     |            |              | 84.1024
Chicago Bulls         |            |              | 78.8209
Cleveland Cavaliers   |            |              | 102.312
Dallas Mavericks      |            |              | 65.7626
Denver Nuggets        |            |              | 62.4294
Detroit Pistons       |            |              | 42.2118
Golden State Warriors |            |              | 94.0851
… (20 rows omitted)
```

[31]:

```
# average salary paid for each position

nba.select('POSITION', 'SALARY').group('POSITION', np.average)
```

[31]:

```
POSITION | SALARY average
C        | 6.08291
PF       | 4.95134
PG       | 5.16549
SF       | 5.53267
SG       | 3.9882
```

[32]:

```
# for each team, average salary paid for each position

nba.drop('PLAYER').group(['TEAM', 'POSITION'], np.average)
```

```
[32]: TEAM           | POSITION | SALARY average
      Atlanta Hawks  | C        | 7.58542
      Atlanta Hawks  | PF       | 11.0025
      Atlanta Hawks  | PG       | 4.8817
      Atlanta Hawks  | SF       | 3
      Atlanta Hawks  | SG       | 1.80969
      Boston Celtics | C        | 2.45046
      Boston Celtics | PF       | 3.08548
      Boston Celtics | PG       | 4.97465
      Boston Celtics | SF       | 4.41716
      Boston Celtics | SG       | 2.00755
      … (137 rows omitted)
```

0.6 Pivot Tables

```
[33]: all_cones
```

```
[33]: Flavor     | Color       | Price
      strawberry | pink        | 3.55
      chocolate  | light brown | 4.75
      chocolate  | dark brown  | 5.25
      strawberry | pink        | 5.25
      chocolate  | dark brown  | 5.25
      bubblegum  | pink        | 4.75
```

```
[34]: all_cones.group(['Flavor', 'Color'])
```

```
[34]: Flavor     | Color       | count
      bubblegum  | pink        | 1
      chocolate  | dark brown  | 2
      chocolate  | light brown | 1
      strawberry | pink        | 2
```

```
[35]: all_cones.pivot('Flavor', 'Color')
```

```
[35]: Color       | bubblegum | chocolate | strawberry
      dark brown  | 0         | 2         | 0
      light brown | 0         | 1         | 0
      pink        | 1         | 0         | 2
```

```
[36]: all_cones.pivot('Flavor', 'Color', values='Price', collect=np.average)
```

```
[36]: Color       | bubblegum | chocolate | strawberry
      dark brown  | 0         | 5.25      | 0
      light brown | 0         | 4.75      | 0
      pink        | 4.75      | 0         | 4.4
```

0.7 Examples

```
[37]: survey = Table.read_table('welcome_survey.csv')
```

```
[38]: survey.show(3)
```

```
<IPython.core.display.HTML object>
```

```
[39]: survey.pivot('Pant leg order', 'Handedness')
```

```
[39]: Handedness    | I don't know | Left leg in first | Right leg in first
      Ambidextrous | 4            | 2                 | 8
      Left-handed  | 15           | 46                | 62
      Right-handed | 181          | 335               | 604
```

```
[40]: nba
```

```
[40]: PLAYER            | POSITION | TEAM          | SALARY
      Paul Millsap      | PF       | Atlanta Hawks | 18.6717
      Al Horford        | C        | Atlanta Hawks | 12
      Tiago Splitter    | C        | Atlanta Hawks | 9.75625
      Jeff Teague       | PG       | Atlanta Hawks | 8
      Kyle Korver       | SG       | Atlanta Hawks | 5.74648
      Thabo Sefolosha   | SF       | Atlanta Hawks | 4
      Mike Scott        | PF       | Atlanta Hawks | 3.33333
      Kent Bazemore     | SF       | Atlanta Hawks | 2
      Dennis Schroder   | PG       | Atlanta Hawks | 1.7634
      Tim Hardaway Jr.  | SG       | Atlanta Hawks | 1.30452
      … (407 rows omitted)
```

```
[41]: # for each team, average salary paid for each position

      nba.pivot('POSITION', 'TEAM', values = 'SALARY', collect = np.average)
```

```
[41]: TEAM                  | C       | PF      | PG      | SF      | SG
      Atlanta Hawks         | 7.58542 | 11.0025 | 4.8817  | 3       | 1.80969
      Boston Celtics        | 2.45046 | 3.08548 | 4.97465 | 4.41716 | 2.00755
      Brooklyn Nets         | 1.3629  | 4.45251 | 3.9     | 13.0403 | 1.74118
      Charlotte Hornets     | 6.77224 | 4.68577 | 4.4853  | 3.76642 | 4.04238
      Chicago Bulls         | 10.4244 | 3.46744 | 11.1715 | 1.95816 | 6.19447
      Cleveland Cavaliers   | 7.75234 | 19.689  | 6.55159 | 22.9705 | 8.98876
      Dallas Mavericks      | 3.23548 | 11.9135 | 4.41818 | 15.3615 | 1.21517
      Denver Nuggets        | 2.6163  | 7.02498 | 3.72362 | 7.19577 | 0.841949
      Detroit Pistons       | 4.0907  | 0       | 13.913  | 1.71622 | 4.58088
      Golden State Warriors | 6.54125 | 7.18637 | 8.45726 | 4.49669 | 9.0005
      … (20 rows omitted)
```

[42]:

```
# CHALLENGE QUESTION: for each team,
# amount paid to "starter" (player earning the most) in each position
```

[ ]:

```

```