1.Exploring your data 1.1 Smokers and drinkers 1.2

## 1. Exploring your data

In [146]:

```
import pandas as pd

import matplotlib.pyplot as plt

import numpy as np


df = pd.read_csv('dob_job_application_filings_subset.csv')

df.tail()
```

```
F:\Program Files\Anaconda\lib\site-packages\IPython\core\interactiveshell.py:3057: DtypeWarning: Columns (16) have mixed types. Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

Out[146]:

| | Job # | Doc # | Borough | House # | Street Name | Block | Lot | Bin # | Job Type | Job Status | ... | Owner's Last Name | Owner's Business Name | O I N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **12841** | 520143988 | 1 | STATEN ISLAND | 8 | NOEL STREET | 5382 | 20 | 5069722 | A2 | D | ... | MALITO | GENO MALITO | 8 |
| **12842** | 121613833 | 1 | MANHATTAN | 724 | 10 AVENUE | 1059 | 4 | 1082503 | A2 | D | ... | CROMAN | 722-724 10TH AVENUE HOLDING LLC | 63 |
| **12843** | 121681260 | 1 | MANHATTAN | 350 | MANHATTAN AVE. | 1848 | 31 | 1055849 | A2 | A | ... | ARYEH | DG UWS LLC | 61 |
| **12844** | 320771704 | 1 | BROOKLYN | 499 | UNION STREET | 431 | 43 | 3007185 | A2 | D | ... | WIGGINS | N/A | 77 |
| **12845** | 520143951 | 1 | STATEN ISLAND | 1755 | RICHMOND ROAD | 887 | 28 | 5022931 | A2 | D | ... | CAMBRIA | RONALD CAMBRIA | 17 |

5 rows × 82 columns

In [147]:

```
df_subset = df.loc[:,['Initial Cost', 'Total Est. Fee']]


df_subset.iloc[:,0].str.replace('$','')


print(df_subset.head())
```

```
  Initial Cost Total Est. Fee
0    $75000.00        $986.00
1        $0.00       $1144.00
2    $30000.00        $522.50
3     $1500.00        $225.00
4    $19500.00        $389.50
```

In [148]:

```
df['Borough'].value_counts(dropna = 'False')
```

Out[148]:

```
MANHATTAN        6310
BROOKLYN         2866
QUEENS           2121
BRONX             974
STATEN ISLAND     575
Name: Borough, dtype: int64
```

In [149]:

```
df.head()
```

Out[149]:

| | Job # | Doc # | Borough | House # | Street Name | Block | Lot | Bin # | Job Type | Job Status | ... | Owner's Last Name | Owner's Business Name | O N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 121577873 | 2 | MANHATTAN | 386 | PARK AVENUE SOUTH | 857 | 38 | 1016890 | A2 | D | ... | MIGLIORE | MACKLOWE MANAGEMENT | 1 |
| **1** | 520129502 | 1 | STATEN ISLAND | 107 | KNOX PLACE | 342 | 1 | 5161350 | A3 | A | ... | BLUMENBERG | NA | 1 |
| **2** | 121601560 | 1 | MANHATTAN | 63 | WEST 131 STREET | 1729 | 9 | 1053831 | A2 | Q | ... | MARKOWITZ | 635 RIVERSIDE DRIVE NY LLC | 6 |
| **3** | 121601203 | 1 | MANHATTAN | 48 | WEST 25TH STREET | 826 | 69 | 1015610 | A2 | D | ... | CASALE | 48 W 25 ST LLC C/O BERNSTEIN | 1 |
| **4** | 121601338 | 1 | MANHATTAN | 45 | WEST 29 STREET | 831 | 7 | 1015754 | A3 | D | ... | LEE | HYUNG-HYANG REALTY CORP | 6 |

5 rows × 82 columns

In [150]:

```
df.columns
```

Out[150]:

```
Index(['Job #', 'Doc #', 'Borough', 'House #', 'Street Name', 'Block', 'Lot',
       'Bin #', 'Job Type', 'Job Status', 'Job Status Descrp',
       'Latest Action Date', 'Building Type', 'Community - Board', 'Cluster',
       'Landmarked', 'Adult Estab', 'Loft Board', 'City Owned', 'Little e',
       'PC Filed', 'eFiling Filed', 'Plumbing', 'Mechanical', 'Boiler',
       'Fuel Burning', 'Fuel Storage', 'Standpipe', 'Sprinkler', 'Fire Alarm',
       'Equipment', 'Fire Suppression', 'Curb Cut', 'Other',
       'Other Description', 'Applicant's First Name', 'Applicant's Last Name',
       'Applicant Professional Title', 'Applicant License #',
       'Professional Cert', 'Pre- Filing Date', 'Paid', 'Fully Paid',
       'Assigned', 'Approved', 'Fully Permitted', 'Initial Cost',
       'Total Est. Fee', 'Fee Status', 'Existing Zoning Sqft',
       'Proposed Zoning Sqft', 'Horizontal Enlrgmt', 'Vertical Enlrgmt',
       'Enlargement SQ Footage', 'Street Frontage', 'ExistingNo. of Stories',
       'Proposed No. of Stories', 'Existing Height', 'Proposed Height',
       'Existing Dwelling Units', 'Proposed Dwelling Units',
       'Existing Occupancy', 'Proposed Occupancy', 'Site Fill', 'Zoning Dist1',
       'Zoning Dist2', 'Zoning Dist3', 'Special District 1',
       'Special District 2', 'Owner Type', 'Non-Profit', 'Owner's First Name',
       'Owner's Last Name', 'Owner's Business Name', 'Owner's House Number',
       'Owner'sHouse Street Name', 'City ', 'State', 'Zip', 'Owner'sPhone #',
       'Job Description', 'DOBRunDate'],
      dtype='object')
```

In [151]:

```
df.shape
```

Out[151]:

```
(12846, 82)
```

In [152]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12846 entries, 0 to 12845
Data columns (total 82 columns):
Job #                           12846 non-null int64
Doc #                           12846 non-null int64
Borough                         12846 non-null object
House #                         12846 non-null object
Street Name                     12846 non-null object
Block                           12846 non-null int64
Lot                             12846 non-null int64
Bin #                           12846 non-null int64
Job Type                        12846 non-null object
Job Status                      12846 non-null object
Job Status Descrp               12846 non-null object
Latest Action Date              12846 non-null object
Building Type                   12846 non-null object
Community - Board               12846 non-null object
Cluster                         0 non-null float64
Landmarked                      2067 non-null object
Adult Estab                     1 non-null object
Loft Board                      65 non-null object
City Owned                      1419 non-null object
Little e                        365 non-null object
PC Filed                        0 non-null float64
eFiling Filed                   12846 non-null object
Plumbing                        12846 non-null object
Mechanical                      12846 non-null object
Boiler                          12846 non-null object
Fuel Burning                    12846 non-null object
Fuel Storage                    12846 non-null object
Standpipe                       12846 non-null object
Sprinkler                       12846 non-null object
Fire Alarm                      12846 non-null object
Equipment                       12846 non-null object
Fire Suppression                12846 non-null object
Curb Cut                        12846 non-null object
Other                           12846 non-null object
Other Description               12846 non-null object
Applicant's First Name          12846 non-null object
Applicant's Last Name           12846 non-null object
Applicant Professional Title    12846 non-null object
Applicant License #             12846 non-null object
Professional Cert               6908 non-null object
Pre- Filing Date                12846 non-null object
Paid                            11961 non-null object
Fully Paid                      11963 non-null object
Assigned                        3817 non-null object
Approved                        4062 non-null object
Fully Permitted                 1495 non-null object
Initial Cost                    12846 non-null object
Total Est. Fee                  12846 non-null object
Fee Status                      12846 non-null object
Existing Zoning Sqft            12846 non-null int64
Proposed Zoning Sqft            12846 non-null int64
Horizontal Enlrgmt              231 non-null object
Vertical Enlrgmt                142 non-null object
Enlargement SQ Footage          12846 non-null int64
Street Frontage                 12846 non-null int64
ExistingNo. of Stories          12846 non-null int64
Proposed No. of Stories         12846 non-null int64
Existing Height                 12846 non-null int64
Proposed Height                 12846 non-null int64
Existing Dwelling Units         12846 non-null object
Proposed Dwelling Units         12846 non-null object
Existing Occupancy              12846 non-null object
Proposed Occupancy              12846 non-null object
Site Fill                       8641 non-null object
Zoning Dist1                    11263 non-null object
Zoning Dist2                    1652 non-null object
Zoning Dist3                    88 non-null object
Special District 1              3062 non-null object
Special District 2              848 non-null object
Owner Type                      0 non-null float64
Non-Profit                      971 non-null object
Owner's First Name              12846 non-null object
Owner's Last Name               12846 non-null object
Owner's Business Name           12846 non-null object
Owner's House Number            12846 non-null object
Owner'sHouse Street Name        12846 non-null object
City                            12846 non-null object
State                           12846 non-null object
Zip                             12846 non-null int64
Owner'sPhone #                  12846 non-null int64
Job Description                 12699 non-null object
DOBRunDate                      12846 non-null object
dtypes: float64(3), int64(15), object(64)
memory usage: 8.0+ MB
```

In [153]:

```
df.describe()
```

Out[153]:

| | Job # | Doc # | Block | Lot | Bin # | Cluster | PC Filed | Existing Zoning Sqft | Proposed Zoning Sqft |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 1.284600e+04 | 12846.000000 | 12846.000000 | 12846.000000 | 1.284600e+04 | 0.0 | 0.0 | 1.284600e+04 | 1.284600e+04 |
| **mean** | 2.426788e+08 | 1.162930 | 2703.834735 | 623.303441 | 2.314997e+06 | NaN | NaN | 1.439973e+03 | 2.007286e+03 |
| **std** | 1.312507e+08 | 0.514937 | 3143.002812 | 2000.934794 | 1.399062e+06 | NaN | NaN | 3.860757e+04 | 4.081570e+04 |
| **min** | 1.036438e+08 | 1.000000 | 1.000000 | 0.000000 | 1.000003e+06 | NaN | NaN | 0.000000e+00 | 0.000000e+00 |
| **25%** | 1.216206e+08 | 1.000000 | 836.000000 | 12.000000 | 1.035728e+06 | NaN | NaN | 0.000000e+00 | 0.000000e+00 |
| **50%** | 2.202645e+08 | 1.000000 | 1411.500000 | 32.000000 | 2.004234e+06 | NaN | NaN | 0.000000e+00 | 0.000000e+00 |
| **75%** | 3.208652e+08 | 1.000000 | 3355.000000 | 59.000000 | 3.343823e+06 | NaN | NaN | 0.000000e+00 | 0.000000e+00 |
| **max** | 5.400246e+08 | 9.000000 | 99999.000000 | 9078.000000 | 5.864852e+06 | NaN | NaN | 2.873107e+06 | 2.873107e+06 |

In [ ]:

In [ ]:

## Uber

In [154]:

```
df_uber = pd.read_csv('nyc_uber_2014.csv')
df_uber.head()
```

Out[154]:

| | Unnamed: 0 | Date/Time | Lat | Lon | Base |
|---|---|---|---|---|---|
| **0** | 0 | 4/1/2014 0:11:00 | 40.7690 | -73.9549 | B02512 |
| **1** | 1 | 4/1/2014 0:17:00 | 40.7267 | -74.0345 | B02512 |
| **2** | 2 | 4/1/2014 0:21:00 | 40.7316 | -73.9873 | B02512 |
| **3** | 3 | 4/1/2014 0:28:00 | 40.7588 | -73.9776 | B02512 |
| **4** | 4 | 4/1/2014 0:33:00 | 40.7594 | -73.9722 | B02512 |

In [ ]:

## Smokers and drinkers

In [155]:

```
df_tips = pd.read_csv('tips.csv')
df_tips.head()
```

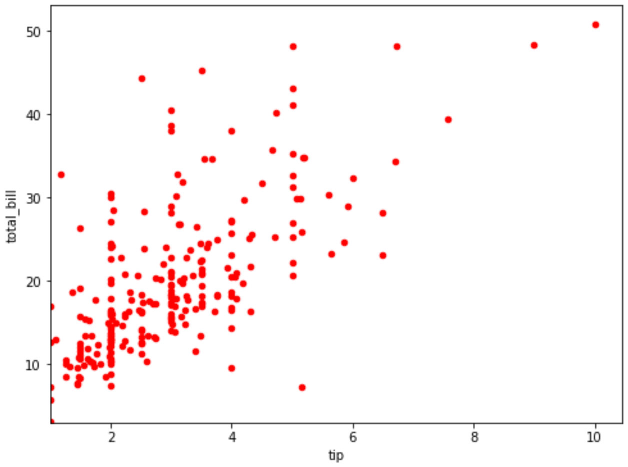
Out[155]:

| | total_bill | tip | sex | smoker | day | time | size |
|---|---|---|---|---|---|---|---|
| **0** | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 |
| **1** | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 |
| **2** | 21.01 | 3.50 | Male | No | Sun | Dinner | 3 |
| **3** | 23.68 | 3.31 | Male | No | Sun | Dinner | 2 |

| | total_bill | tip | sex | smoker | day | time | size |
|---|---|---|---|---|---|---|---|
| **4** | 24.59 | 3.61 | Female | No | Sun | Dinner | 4 |

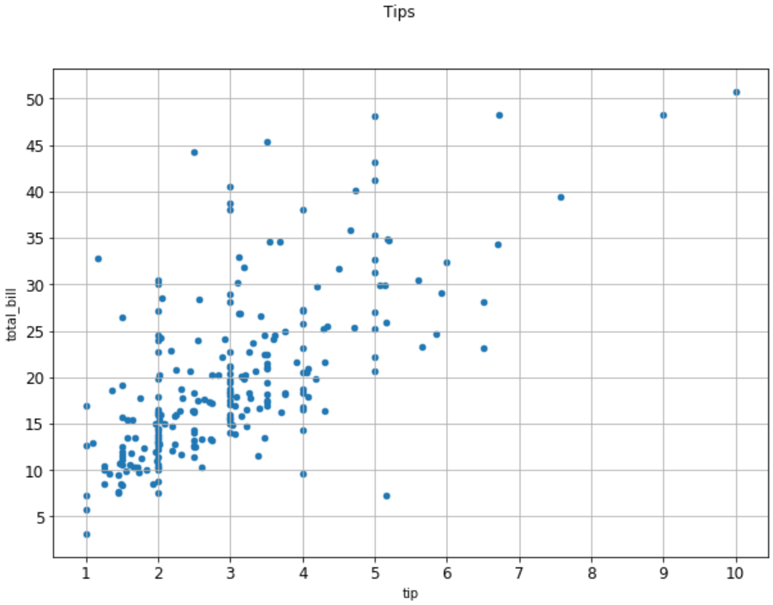
In [156]:

```
df_tips.plot(kind = 'scatter', x = 'tip', y = 'total_bill', figsize = [8,6], color = 'red', ylim = 3, xlim = 1)

plt.show()
```



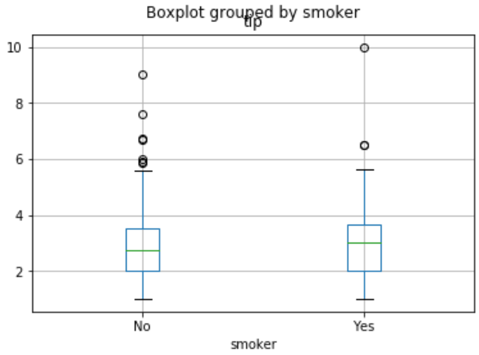
In [157]:

```
df_tips.plot(kind = 'scatter', x = 'tip', y = 'total_bill', title = 'Tips', subplots = True, grid = True, legend = True, xticks = [1,2,3,4,5,6,7,8,9,10], yticks = [5,10, 15, 20, 25,30,35, 40,45,50], fontsize = 12, figsize = [10,7], sharey = False)

plt.show()
```



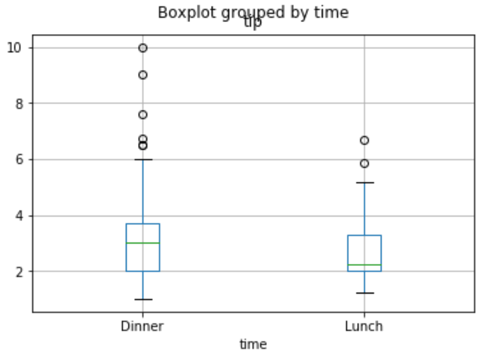
In [158]:

```
df_tips.boxplot(column = 'tip', by = 'smoker')

plt.show()
```



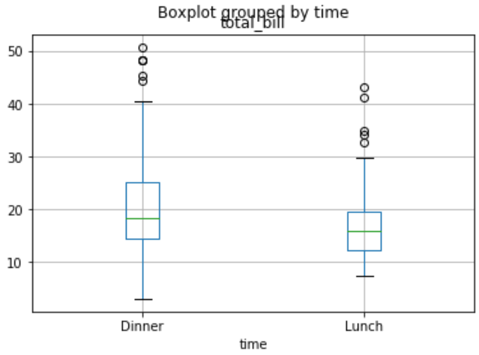
In [159]:

```
df_tips.boxplot(column = 'tip', by = 'time')

plt.show()
```



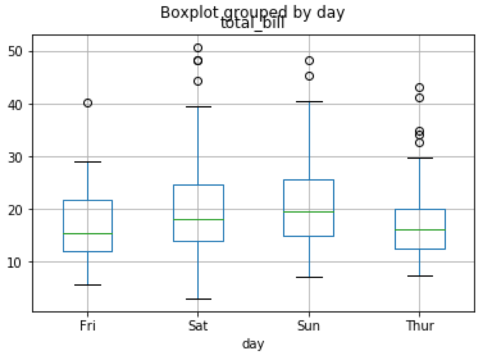
In [160]:

```
df_tips.boxplot(column = 'total_bill', by = 'time')

plt.show()
```



In [161]:

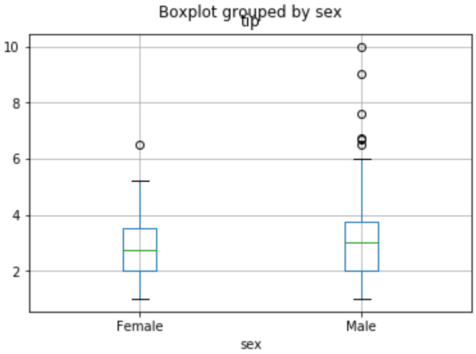
```
df_tips.boxplot(column = 'total_bill', by = 'day')

plt.show()
```



In [162]:

```
df_tips.boxplot(column = 'tip', by = 'sex')
```

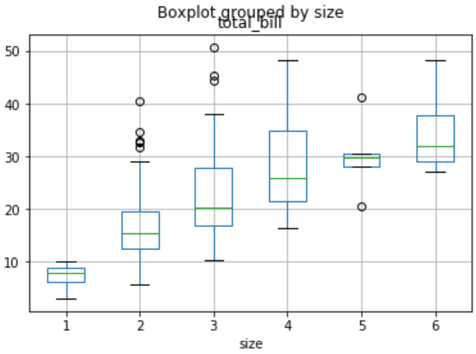
plt.show()



In [163]:

df_tips.boxplot(column = 'total_bill', by = 'size')

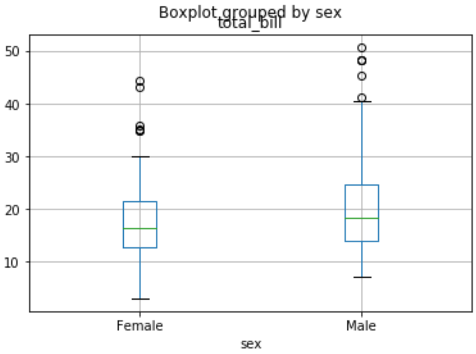
plt.show()



In [164]:

df_tips.boxplot(column = 'total_bill', by = 'sex')

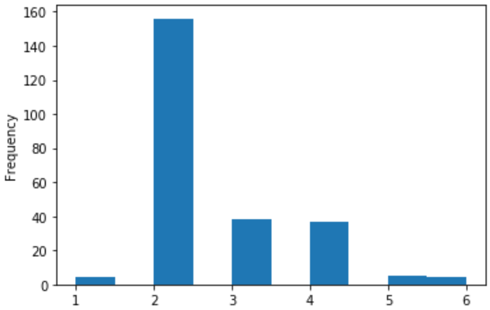
plt.show()



In [165]:

df_tips['size'].plot(kind = 'hist')

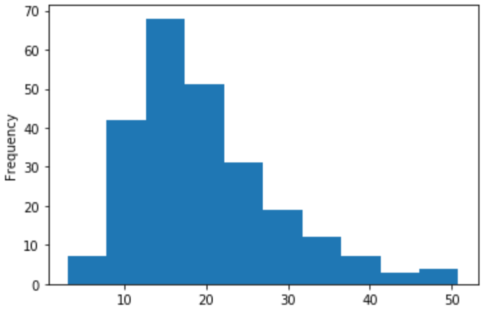
plt.show()



In [166]:

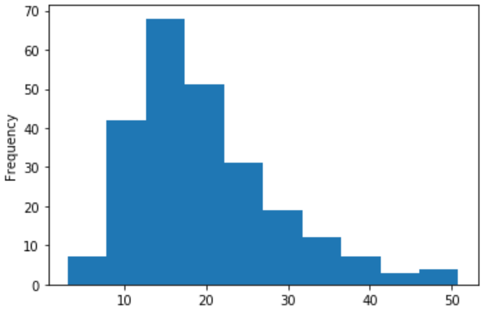
df_tips['total_bill'].plot(kind = 'hist')

```
plt.show()
```



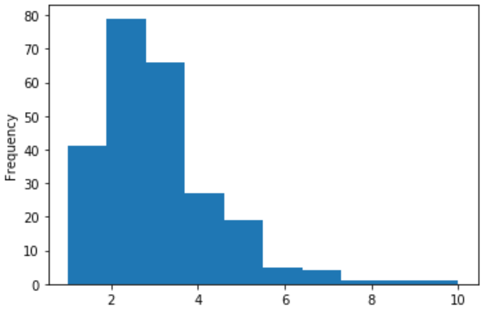
In [167]:

```
df_tips.total_bill.plot(kind = 'hist')

plt.show()
```



In [168]:

```
df_tips.tip.plot(kind = 'hist')

plt.show()
```



In [169]:

```
df_tips.sex.value_counts(dropna = 'False')
```

Out[169]:

```
Male      157
Female     87
Name: sex, dtype: int64
```

In [170]:

```
df_tips.day.value_counts(dropna = 'False')
```

Out[170]:

```
Sat     87
Sun     76
Thur    62
Fri     19
Name: day, dtype: int64
```

In [171]:

```
df_tips.smoker.value_counts()
```

Out[171]:

```
No     151
Yes     93
Name: smoker, dtype: int64
```

In [172]:

```
df_tips.describe()
```

Out[172]:

|  | total_bill | tip | size |
|---|---|---|---|
| **count** | 244.000000 | 244.000000 | 244.000000 |
| **mean** | 19.785943 | 2.998279 | 2.569672 |
| **std** | 8.902412 | 1.383638 | 0.951100 |
| **min** | 3.070000 | 1.000000 | 1.000000 |
| **25%** | 13.347500 | 2.000000 | 2.000000 |
| **50%** | 17.795000 | 2.900000 | 2.000000 |
| **75%** | 24.127500 | 3.562500 | 3.000000 |
| **max** | 50.810000 | 10.000000 | 6.000000 |

In [173]:

```
df_tips.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
total_bill    244 non-null float64
tip           244 non-null float64
sex           244 non-null object
smoker        244 non-null object
day           244 non-null object
time          244 non-null object
size          244 non-null int64
dtypes: float64(2), int64(1), object(4)
memory usage: 13.4+ KB
```

In [ ]:

In [ ]:

## Life expentancy

In [174]:

```
df_gap = pd.read_csv('gapminder.csv')

df_gap.tail()
```

Out[174]:

|  | Unnamed: 0 | 1800 | 1801 | 1802 | 1803 | 1804 | 1805 | 1806 | 1807 | 1808 | ... | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **775** | 255 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **776** | 256 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | 51.1 | 52.3 | 53.1 | 53.7 | 54.7 | 55.6 | 56.3 |
| **777** | 257 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | 47.3 | 48.0 | 49.1 | 51.6 | 54.2 | 55.7 | 57.0 |
| **778** | 258 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **779** | 259 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | 55.6 | 55.8 | 56.0 | 55.9 | 56.0 | 56.0 | 56.1 |

5 rows × 219 columns

In [175]:

```
df_gap.head()
```

Out[175]:

| | Unnamed: 0 | 1800 | 1801 | 1802 | 1803 | 1804 | 1805 | 1806 | 1807 | 1808 | ... | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 201 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **1** | 1 | 28.21 | 28.20 | 28.19 | 28.18 | 28.17 | 28.16 | 28.15 | 28.14 | 28.13 | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **2** | 2 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **3** | 3 | 35.40 | 35.40 | 35.40 | 35.40 | 35.40 | 35.40 | 35.40 | 35.40 | 35.40 | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **4** | 4 | 28.82 | 28.82 | 28.82 | 28.82 | 28.82 | 28.82 | 28.82 | 28.82 | 28.82 | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

5 rows × 219 columns

In [ ]:

In [ ]:

## Air quality

In [176]:

```
df_air = pd.read_csv('airquality.csv')

print(df_air.head())
```

```
   Ozone  Solar.R  Wind  Temp  Month  Day
0   41.0    190.0   7.4    67      5    1
1   36.0    118.0   8.0    72      5    2
2   12.0    149.0  12.6    74      5    3
3   18.0    313.0  11.5    62      5    4
4    NaN      NaN  14.3    56      5    5
```

### *1 Method of melting*

In [177]:

```
df_air_melt = pd.melt(df_air, value_vars = ['Ozone', 'Solar.R', 'Wind', 'Temp'], var_name = 'Criteria', id_vars = ['Month', 'Day'])



print('df_air      - ' + str(df_air_melt.shape))

print('df_air_melt - ' + str(df_air.shape))

display(df_air_melt.tail())

display(df_air_melt.head())
```

```
df_air      - (612, 4)
df_air_melt - (153, 6)
```

| | Month | Day | Criteria | value |
|---|---|---|---|---|
| **607** | 9 | 26 | Temp | 70.0 |
| **608** | 9 | 27 | Temp | 77.0 |
| **609** | 9 | 28 | Temp | 75.0 |
| **610** | 9 | 29 | Temp | 76.0 |
| **611** | 9 | 30 | Temp | 68.0 |

| | Month | Day | Criteria | value |
|---|---|---|---|---|
| **0** | 5 | 1 | Ozone | 41.0 |
| **1** | 5 | 2 | Ozone | 36.0 |
| **2** | 5 | 3 | Ozone | 12.0 |

|   | Month | Day | Criteria | value |
|---|---|---|---|---|
| **3** | 5 | 4 | Ozone | 18.0 |
| **4** | 5 | 5 | Ozone | NaN |

## *2 Method of melting*

In [178]:

```
airquality_melt = pd.melt(df_air, id_vars = ['Month', 'Day'], var_name = 'Measurement', value_name = 'Reading')

airquality_melt.head()
```

Out[178]:

|   | Month | Day | Measurement | Reading |
|---|---|---|---|---|
| **0** | 5 | 1 | Ozone | 41.0 |
| **1** | 5 | 2 | Ozone | 36.0 |
| **2** | 5 | 3 | Ozone | 12.0 |
| **3** | 5 | 4 | Ozone | 18.0 |
| **4** | 5 | 5 | Ozone | NaN |

In [179]:

```
df_air_melt_back = pd.pivot(airquality_melt, columns = 'Measurement', values = 'Reading')

df_air_melt_back
```

Out[179]:

| Measurement | Ozone | Solar.R | Temp | Wind |
|---|---|---|---|---|
| **0** | 41.0 | NaN | NaN | NaN |
| **1** | 36.0 | NaN | NaN | NaN |
| **2** | 12.0 | NaN | NaN | NaN |
| **3** | 18.0 | NaN | NaN | NaN |
| **4** | NaN | NaN | NaN | NaN |
| **5** | 28.0 | NaN | NaN | NaN |
| **6** | 23.0 | NaN | NaN | NaN |
| **7** | 19.0 | NaN | NaN | NaN |
| **8** | 8.0 | NaN | NaN | NaN |
| **9** | NaN | NaN | NaN | NaN |
| **10** | 7.0 | NaN | NaN | NaN |
| **11** | 16.0 | NaN | NaN | NaN |
| **12** | 11.0 | NaN | NaN | NaN |
| **13** | 14.0 | NaN | NaN | NaN |
| **14** | 18.0 | NaN | NaN | NaN |
| **15** | 14.0 | NaN | NaN | NaN |
| **16** | 34.0 | NaN | NaN | NaN |
| **17** | 6.0 | NaN | NaN | NaN |
| **18** | 30.0 | NaN | NaN | NaN |
| **19** | 11.0 | NaN | NaN | NaN |

| Measurement | Ozone | Solar.R | Temp | Wind |
|---|---|---|---|---|
| **20** | 1.0 | NaN | NaN | NaN |
| **21** | 11.0 | NaN | NaN | NaN |
| **22** | 4.0 | NaN | NaN | NaN |
| **23** | 32.0 | NaN | NaN | NaN |
| **24** | NaN | NaN | NaN | NaN |
| **25** | NaN | NaN | NaN | NaN |
| **26** | NaN | NaN | NaN | NaN |
| **27** | 23.0 | NaN | NaN | NaN |
| **28** | 45.0 | NaN | NaN | NaN |
| **29** | 115.0 | NaN | NaN | NaN |
| **...** | ... | ... | ... | ... |
| **582** | NaN | NaN | 91.0 | NaN |
| **583** | NaN | NaN | 92.0 | NaN |
| **584** | NaN | NaN | 93.0 | NaN |
| **585** | NaN | NaN | 93.0 | NaN |
| **586** | NaN | NaN | 87.0 | NaN |
| **587** | NaN | NaN | 84.0 | NaN |
| **588** | NaN | NaN | 80.0 | NaN |
| **589** | NaN | NaN | 78.0 | NaN |
| **590** | NaN | NaN | 75.0 | NaN |
| **591** | NaN | NaN | 73.0 | NaN |
| **592** | NaN | NaN | 81.0 | NaN |
| **593** | NaN | NaN | 76.0 | NaN |
| **594** | NaN | NaN | 77.0 | NaN |
| **595** | NaN | NaN | 71.0 | NaN |
| **596** | NaN | NaN | 71.0 | NaN |
| **597** | NaN | NaN | 78.0 | NaN |
| **598** | NaN | NaN | 67.0 | NaN |
| **599** | NaN | NaN | 76.0 | NaN |
| **600** | NaN | NaN | 68.0 | NaN |
| **601** | NaN | NaN | 82.0 | NaN |
| **602** | NaN | NaN | 64.0 | NaN |
| **603** | NaN | NaN | 71.0 | NaN |
| **604** | NaN | NaN | 81.0 | NaN |
| **605** | NaN | NaN | 69.0 | NaN |
| **606** | NaN | NaN | 63.0 | NaN |
| **607** | NaN | NaN | 70.0 | NaN |
| **608** | NaN | NaN | 77.0 | NaN |
| **609** | NaN | NaN | 75.0 | NaN |

| Measurement | Ozone | Solar.R | Temp | Wind |
|---|---|---|---|---|
| **610** | NaN | NaN | 76.0 | NaN |
| **611** | NaN | NaN | 68.0 | NaN |

612 rows × 4 columns

## Pivotting datas

In [180]:

```
airquality_melt.head()

airquality_pivot = pd.pivot_table(airquality_melt, index = ['Month', 'Day'], columns = 'Measurement', values = 'Reading')

airquality_pivot_reset = airquality_pivot.reset_index()

airquality_pivot_reset.head()
```

Out[180]:

| Measurement | Month | Day | Ozone | Solar.R | Temp | Wind |
|---|---|---|---|---|---|---|
| **0** | 5 | 1 | 41.0 | 190.0 | 67.0 | 7.4 |
| **1** | 5 | 2 | 36.0 | 118.0 | 72.0 | 8.0 |
| **2** | 5 | 3 | 12.0 | 149.0 | 74.0 | 12.6 |
| **3** | 5 | 4 | 18.0 | 313.0 | 62.0 | 11.5 |
| **4** | 5 | 5 | NaN | NaN | 56.0 | 14.3 |

In [181]:

```
df_air.columns
```

Out[181]:

```
Index(['Ozone', 'Solar.R', 'Wind', 'Temp', 'Month', 'Day'], dtype='object')
```

In [182]:

```
df_air.describe()
```

Out[182]:

| | Ozone | Solar.R | Wind | Temp | Month | Day |
|---|---|---|---|---|---|---|
| **count** | 116.000000 | 146.000000 | 153.000000 | 153.000000 | 153.000000 | 153.000000 |
| **mean** | 42.129310 | 185.931507 | 9.957516 | 77.882353 | 6.993464 | 15.803922 |
| **std** | 32.987885 | 90.058422 | 3.523001 | 9.465270 | 1.416522 | 8.864520 |
| **min** | 1.000000 | 7.000000 | 1.700000 | 56.000000 | 5.000000 | 1.000000 |
| **25%** | 18.000000 | 115.750000 | 7.400000 | 72.000000 | 6.000000 | 8.000000 |
| **50%** | 31.500000 | 205.000000 | 9.700000 | 79.000000 | 7.000000 | 16.000000 |
| **75%** | 63.250000 | 258.750000 | 11.500000 | 85.000000 | 8.000000 | 23.000000 |
| **max** | 168.000000 | 334.000000 | 20.700000 | 97.000000 | 9.000000 | 31.000000 |

In [183]:

```
df_air.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 153 entries, 0 to 152
Data columns (total 6 columns):
Ozone      116 non-null float64
Solar.R    146 non-null float64
Wind       153 non-null float64
Temp       153 non-null int64
Month      153 non-null int64
Day        153 non-null int64
dtypes: float64(3), int64(3)
memory usage: 7.2 KB
```

In [184]:

```
df_air.Ozone.value_counts(dropna = 'False').head()
```

Out[184]:

```
23.0    6
16.0    4
13.0    4
14.0    4
18.0    4
Name: Ozone, dtype: int64
```

In [185]:

```
df_air.Month.value_counts(dropna = 'False')
```

Out[185]:

```
8    31
7    31
5    31
9    30
6    30
Name: Month, dtype: int64
```

## 2. Tidying data for analysis

**Pivot table**

In [186]:

```
example = pd.DataFrame({"A": ["foo", "foo", "foo", "foo", "foo","bar", "bar", "bar", "bar"],

                    "B": ["one", "one", "one", "two", "two", "one", "one", "two", "two"],

                    "C": ["small", "large", "large", "small","small", "large", "small", "small", "large"],

                    "D": [1, 2, 2, 3, 3, 4, 5, 6, 7],

                    "E": [2, 4, 5, 5, 6, 6, 8, 9, 9]})

example
```

Out[186]:

|   | A | B | C | D | E |
|---|---|---|---|---|---|
| **0** | foo | one | small | 1 | 2 |
| **1** | foo | one | large | 2 | 4 |
| **2** | foo | one | large | 2 | 5 |
| **3** | foo | two | small | 3 | 5 |
| **4** | foo | two | small | 3 | 6 |
| **5** | bar | one | large | 4 | 6 |
| **6** | bar | one | small | 5 | 8 |
| **7** | bar | two | small | 6 | 9 |
| **8** | bar | two | large | 7 | 9 |

In [187]:

```
table = pd.pivot_table(example, index = ['A', 'B'], values = 'D', columns = 'C', aggfunc = np.sum, fill_value = 0)

table
```

Out[187]:

| | C | large | small |
|---|---|---|---|
| A | B | | |
| bar | one | 4 | 5 |
| | two | 7 | 6 |
| foo | one | 4 | 1 |
| | two | 0 | 6 |

In [188]:

```
table_mean = pd.pivot_table(example, index = ['A','C'], values = ['D','E'], aggfunc = {'D':np.mean, 'E':np.mean})

table_mean
```

Out[188]:

| | | D | E |
|---|---|---|---|
| A | C | | |
| bar | large | 5.500000 | 7.500000 |
| | small | 5.500000 | 8.500000 |
| foo | large | 2.000000 | 4.500000 |
| | small | 2.333333 | 4.333333 |

In [189]:

```
table_math = pd.pivot_table(example, index = ['A','C'], values = ['D', 'E'], aggfunc = {'D':np.mean, 'E':[max, min, np.mean]})

table_math
```

Out[189]:

| | | D | E | | |
|---|---|---|---|---|---|
| | | mean | max | mean | min |
| A | C | | | | |
| bar | large | 5.500000 | 9.0 | 7.500000 | 6.0 |
| | small | 5.500000 | 9.0 | 8.500000 | 8.0 |
| foo | large | 2.000000 | 5.0 | 4.500000 | 4.0 |
| | small | 2.333333 | 6.0 | 4.333333 | 2.0 |

Ebola

In [190]:

```
ebola = pd.read_csv('ebola.csv')

ebola.head()
```

Out[190]:

| | Date | Day | Cases_Guinea | Cases_Liberia | Cases_SierraLeone | Cases_Nigeria | Cases_Senegal | Cases_UnitedStates | C |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1/5/2015 | 289 | 2776.0 | NaN | 10030.0 | NaN | NaN | NaN | N |
| 1 | 1/4/2015 | 288 | 2775.0 | NaN | 9780.0 | NaN | NaN | NaN | N |
| 2 | 1/3/2015 | 287 | 2769.0 | 8166.0 | 9722.0 | NaN | NaN | NaN | N |
| 3 | 1/2/2015 | 286 | NaN | 8157.0 | NaN | NaN | NaN | NaN | N |
| 4 | 12/31/2014 | 284 | 2730.0 | 8115.0 | 9633.0 | NaN | NaN | NaN | N |

In [191]:

```
ebola.describe()
```

Out[191]:

| | Day | Cases_Guinea | Cases_Liberia | Cases_SierraLeone | Cases_Nigeria | Cases_Senegal | Cases_UnitedStates | Ca |
|---|---|---|---|---|---|---|---|---|
| **count** | 122.000000 | 93.000000 | 83.000000 | 87.000000 | 38.000000 | 25.00 | 18.000000 | 16. |
| **mean** | 144.778689 | 911.064516 | 2335.337349 | 2427.367816 | 16.736842 | 1.08 | 3.277778 | 1.0 |
| **std** | 89.316460 | 849.108801 | 2987.966721 | 3184.803996 | 5.998577 | 0.40 | 1.178511 | 0.0 |
| **min** | 0.000000 | 49.000000 | 3.000000 | 0.000000 | 0.000000 | 1.00 | 1.000000 | 1.0 |
| **25%** | 66.250000 | 236.000000 | 25.500000 | 64.500000 | 15.000000 | 1.00 | 3.000000 | 1.0 |
| **50%** | 150.000000 | 495.000000 | 516.000000 | 783.000000 | 20.000000 | 1.00 | 4.000000 | 1.0 |
| **75%** | 219.500000 | 1519.000000 | 4162.500000 | 3801.000000 | 20.000000 | 1.00 | 4.000000 | 1.0 |
| **max** | 289.000000 | 2776.000000 | 8166.000000 | 10030.000000 | 22.000000 | 3.00 | 4.000000 | 1.0 |

In [192]:

```
ebola.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 122 entries, 0 to 121
Data columns (total 18 columns):
Date                   122 non-null object
Day                    122 non-null int64
Cases_Guinea           93 non-null float64
Cases_Liberia          83 non-null float64
Cases_SierraLeone      87 non-null float64
Cases_Nigeria          38 non-null float64
Cases_Senegal          25 non-null float64
Cases_UnitedStates     18 non-null float64
Cases_Spain            16 non-null float64
Cases_Mali             12 non-null float64
Deaths_Guinea          92 non-null float64
Deaths_Liberia         81 non-null float64
Deaths_SierraLeone     87 non-null float64
Deaths_Nigeria         38 non-null float64
Deaths_Senegal         22 non-null float64
Deaths_UnitedStates    18 non-null float64
Deaths_Spain           16 non-null float64
Deaths_Mali            12 non-null float64
dtypes: float64(16), int64(1), object(1)
memory usage: 17.2+ KB
```

In [193]:

```
ebola.columns
```

Out[193]:

```
Index(['Date', 'Day', 'Cases_Guinea', 'Cases_Liberia', 'Cases_SierraLeone',
       'Cases_Nigeria', 'Cases_Senegal', 'Cases_UnitedStates', 'Cases_Spain',
       'Cases_Mali', 'Deaths_Guinea', 'Deaths_Liberia', 'Deaths_SierraLeone',
       'Deaths_Nigeria', 'Deaths_Senegal', 'Deaths_UnitedStates',
       'Deaths_Spain', 'Deaths_Mali'],
      dtype='object')
```

In [194]:

```
ebola_melt = pd.melt(ebola, id_vars = ['Date', 'Day'], var_name = 'type_country', value_name = 'counts')

ebola_melt.head()
```

Out[194]:

| | Date | Day | type_country | counts |
|---|---|---|---|---|
| **0** | 1/5/2015 | 289 | Cases_Guinea | 2776.0 |
| **1** | 1/4/2015 | 288 | Cases_Guinea | 2775.0 |
| **2** | 1/3/2015 | 287 | Cases_Guinea | 2769.0 |
| **3** | 1/2/2015 | 286 | Cases_Guinea | NaN |
| **4** | 12/31/2014 | 284 | Cases_Guinea | 2730.0 |

In [195]:

```
ebola_melt.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1952 entries, 0 to 1951
Data columns (total 4 columns):
Date            1952 non-null object
Day             1952 non-null int64
type_country    1952 non-null object
counts          738 non-null float64
dtypes: float64(1), int64(1), object(2)
memory usage: 61.1+ KB
```

In [196]:

```
ebola_melt['str_split'] = ebola_melt['type_country'].str.split('_')

ebola_melt.head()
```

Out[196]:

| | Date | Day | type_country | counts | str_split |
|---|---|---|---|---|---|
| **0** | 1/5/2015 | 289 | Cases_Guinea | 2776.0 | [Cases, Guinea] |
| **1** | 1/4/2015 | 288 | Cases_Guinea | 2775.0 | [Cases, Guinea] |
| **2** | 1/3/2015 | 287 | Cases_Guinea | 2769.0 | [Cases, Guinea] |
| **3** | 1/2/2015 | 286 | Cases_Guinea | NaN | [Cases, Guinea] |
| **4** | 12/31/2014 | 284 | Cases_Guinea | 2730.0 | [Cases, Guinea] |

In [197]:

```
ebola_melt['type'] = ebola_melt['str_split'].str.get(0)

ebola_melt['country'] = ebola_melt['str_split'].str.get(1)

ebola_melt.head()
```

Out[197]:

| | Date | Day | type_country | counts | str_split | type | country |
|---|---|---|---|---|---|---|---|
| **0** | 1/5/2015 | 289 | Cases_Guinea | 2776.0 | [Cases, Guinea] | Cases | Guinea |
| **1** | 1/4/2015 | 288 | Cases_Guinea | 2775.0 | [Cases, Guinea] | Cases | Guinea |
| **2** | 1/3/2015 | 287 | Cases_Guinea | 2769.0 | [Cases, Guinea] | Cases | Guinea |
| **3** | 1/2/2015 | 286 | Cases_Guinea | NaN | [Cases, Guinea] | Cases | Guinea |
| **4** | 12/31/2014 | 284 | Cases_Guinea | 2730.0 | [Cases, Guinea] | Cases | Guinea |

In [198]:

```
ebola_melt_final = ebola_melt.drop(columns = 'str_split')

ebola_melt_final.head()
```

Out[198]:

| | Date | Day | type_country | counts | type | country |
|---|---|---|---|---|---|---|
| **0** | 1/5/2015 | 289 | Cases_Guinea | 2776.0 | Cases | Guinea |
| **1** | 1/4/2015 | 288 | Cases_Guinea | 2775.0 | Cases | Guinea |

|  | Date | Day | type_country | counts | type | country |
|---|---|---|---|---|---|---|
| **2** | 1/3/2015 | 287 | Cases_Guinea | 2769.0 | Cases | Guinea |
| **3** | 1/2/2015 | 286 | Cases_Guinea | NaN | Cases | Guinea |
| **4** | 12/31/2014 | 284 | Cases_Guinea | 2730.0 | Cases | Guinea |

## 3. Combining data for analysis

In [199]:

```
x = pd.DataFrame({"A": ["foo", "foo", "foo", "foo", "foo","bar", "bar", "bar", "bar"],
                  "B": ["one", "one", "one", "two", "two", "one", "one", "two", "two"],
                  "C": ["small", "large", "large", "small","small", "large", "small", "small", "large"],
                  "D": [1, 2, 2, 3, 3, 4, 5, 6, 7],
                  "E": [2, 4, 5, 5, 6, 6, 8, 9, 9]})
y = pd.DataFrame({"A": ["foo", "bar", "foo", "bar", "foo","bar", "foo", "foo", "bar"],
                  "B": ["two", "one", "one", "two", "one", "one", "two", "two", "two"],
                  "C": ["small", "large", "small", "small","large", "large", "small", "small", "large"],
                  "D": [11, 3, 8, 7, 3, 4, 5, 6, 7],
                  "E": [22, 10, 4, 3, 4, 67, 10, 7, 1]})
display(x)
display(y)
```

|  | A | B | C | D | E |
|---|---|---|---|---|---|
| **0** | foo | one | small | 1 | 2 |
| **1** | foo | one | large | 2 | 4 |
| **2** | foo | one | large | 2 | 5 |
| **3** | foo | two | small | 3 | 5 |
| **4** | foo | two | small | 3 | 6 |
| **5** | bar | one | large | 4 | 6 |
| **6** | bar | one | small | 5 | 8 |
| **7** | bar | two | small | 6 | 9 |
| **8** | bar | two | large | 7 | 9 |

|  | A | B | C | D | E |
|---|---|---|---|---|---|
| **0** | foo | two | small | 11 | 22 |
| **1** | bar | one | large | 3 | 10 |
| **2** | foo | one | small | 8 | 4 |
| **3** | bar | two | small | 7 | 3 |
| **4** | foo | one | large | 3 | 4 |
| **5** | bar | one | large | 4 | 67 |
| **6** | foo | two | small | 5 | 10 |
| **7** | foo | two | small | 6 | 7 |
| **8** | bar | two | large | 7 | 1 |

In [200]:

```
x_concat = pd.concat([x, y], ignore_index = True, axis = 1)
```

```
x_concat
```

Out[200]:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | foo | one | small | 1 | 2 | foo | two | small | 11 | 22 |
| **1** | foo | one | large | 2 | 4 | bar | one | large | 3 | 10 |
| **2** | foo | one | large | 2 | 5 | foo | one | small | 8 | 4 |
| **3** | foo | two | small | 3 | 5 | bar | two | small | 7 | 3 |
| **4** | foo | two | small | 3 | 6 | foo | one | large | 3 | 4 |
| **5** | bar | one | large | 4 | 6 | bar | one | large | 4 | 67 |
| **6** | bar | one | small | 5 | 8 | foo | two | small | 5 | 10 |
| **7** | bar | two | small | 6 | 9 | foo | two | small | 6 | 7 |
| **8** | bar | two | large | 7 | 9 | bar | two | large | 7 | 1 |

In [201]:

```
z = x.merge(y, left_on = 'A', right_on = 'A', how = 'inner')

z.head()
```

Out[201]:

| | A | B_x | C_x | D_x | E_x | B_y | C_y | D_y | E_y |
|---|---|---|---|---|---|---|---|---|---|
| **0** | foo | one | small | 1 | 2 | two | small | 11 | 22 |
| **1** | foo | one | small | 1 | 2 | one | small | 8 | 4 |
| **2** | foo | one | small | 1 | 2 | one | large | 3 | 4 |
| **3** | foo | one | small | 1 | 2 | two | small | 5 | 10 |
| **4** | foo | one | small | 1 | 2 | two | small | 6 | 7 |

## ~ Glob Python

In [202]:

```
import pandas as pd

import glob



pattern = '*.csv'

csv_files = glob.glob(pattern)



display(type(csv_files))

display(csv_files)
```

```
list

['airquality.csv',
 'dob_job_application_filings_subset.csv',
 'ebola.csv',
 'gapminder.csv',
 'nyc_uber_2014.csv',
 'tb.csv',
 'tips.csv',
 'uber1.csv',
 'uber2.csv',
 'uber3.csv']
```

In [203]:

```
pd_tips_file = pd.read_csv(csv_files[-1])

pd_tips_file.head()
```

Out[203]:

| | ,Date/Time,Lat,Lon,Base |
|---|---|
| **0** | 0,6/1/2014 0:00:00,40.7293,-73.992,B02512 |
| **1** | 1,6/1/2014 0:01:00,40.7131,-74.0097,B02512 |
| **2** | 2,6/1/2014 0:04:00,40.3461,-74.661,B02512 |
| **3** | 3,6/1/2014 0:04:00,40.7555,-73.9833,B02512 |
| **4** | 4,6/1/2014 0:07:00,40.688,-74.1831,B02512 |

## ~ Iterating and concatenating all matches

In [204]:

```
import pandas as pd
import glob

list = []
index = 'uber*.csv'
csv_files = glob.glob(index)

for x in csv_files:
    df = pd.read_csv(x, delimiter = ',')
    list.append(df)

uber = pd.concat(list)

display(uber.shape)
display(uber.head())
```

```
(297, 1)
```

| | ,Date/Time,Lat,Lon,Base |
|---|---|
| **0** | 0,4/1/2014 0:11:00,40.769,-73.9549,B02512 |
| **1** | 1,4/1/2014 0:17:00,40.7267,-74.0345,B02512 |
| **2** | 2,4/1/2014 0:21:00,40.7316,-73.9873,B02512 |
| **3** | 3,4/1/2014 0:28:00,40.7588,-73.9776,B02512 |
| **4** | 4,4/1/2014 0:33:00,40.7594,-73.9722,B02512 |

## ~ Merge

In [205]:

```
site = pd.DataFrame({"name": ["DR-1", "DR-3", "MSK-4"],
                     "lat": [-49.85, -47.15, -48.87],
                      "long": [-128.57, -126.72, -123.40]})
visited = pd.DataFrame({"ident": [619, 734, 837],
                   "site": ["DR-1", "DR-3", "MSK-4"],
                   "dated": ["1927-02-08", "1939-01-07", "1932-01-14"]})

display(site)
display(visited)
```

|   | name | lat | long |
|---|---|---|---|
| **0** | DR-1 | -49.85 | -128.57 |
| **1** | DR-3 | -47.15 | -126.72 |
| **2** | MSK-4 | -48.87 | -123.40 |

|   | ident | site | dated |
|---|---|---|---|
| **0** | 619 | DR-1 | 1927-02-08 |
| **1** | 734 | DR-3 | 1939-01-07 |
| **2** | 837 | MSK-4 | 1932-01-14 |

In [206]:

```
o2o = pd.merge(left = site, right = visited, left_on = 'name', right_on = 'site', how = 'inner', validate = 'many_to_many')

o2o
```

Out[206]:

|   | name | lat | long | ident | site | dated |
|---|---|---|---|---|---|---|
| **0** | DR-1 | -49.85 | -128.57 | 619 | DR-1 | 1927-02-08 |
| **1** | DR-3 | -47.15 | -126.72 | 734 | DR-3 | 1939-01-07 |
| **2** | MSK-4 | -48.87 | -123.40 | 837 | MSK-4 | 1932-01-14 |

## 4. Cleaning data for analysis

- *astype() and dtype() works only with DataFrame not Series*
- _category type decrease memory usage

In [216]:

```
type(df_tips)
```

Out[216]:

```
pandas.core.frame.DataFrame
```

In [219]:

```
df_tips2 = df_tips

df_tips2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
total_bill    244 non-null float64
tip           244 non-null float64
sex           244 non-null object
smoker        244 non-null object
day           244 non-null object
time          244 non-null object
size          244 non-null int64
dtypes: float64(2), int64(1), object(4)
memory usage: 13.4+ KB
```

In [220]:

```
type(df_tips2)
```

Out[220]:

```
pandas.core.frame.DataFrame
```

In [208]:

```
df_tips2.head()
```

Out[208]:

| | total_bill | tip | sex | smoker | day | time | size |
|---|---|---|---|---|---|---|---|
| **0** | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 |
| **1** | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 |
| **2** | 21.01 | 3.50 | Male | No | Sun | Dinner | 3 |
| **3** | 23.68 | 3.31 | Male | No | Sun | Dinner | 2 |
| **4** | 24.59 | 3.61 | Female | No | Sun | Dinner | 4 |

In [223]:

```
df_tips2.sex = df_tips2.sex.astype('category')

df_tips2['smoker'] = df_tips2['smoker'].astype('category')
```

In [224]:

```
df_tips2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
total_bill    244 non-null float64
tip           244 non-null float64
sex           244 non-null category
smoker        244 non-null category
day           244 non-null object
time          244 non-null object
size          244 non-null int64
dtypes: category(2), float64(2), int64(1), object(2)
memory usage: 10.3+ KB
```

In [225]:

```
df_tips2.total_bill = pd.to_numeric(df_tips2['total_bill'], errors = 'coerce')
```

In [226]:

```
s = pd.Series(['1.0', '2', -3])

s
```

Out[226]:

```
0    1.0
1      2
2     -3
dtype: object
```

In [227]:

```
pd.to_numeric(s)
```

Out[227]:

```
0    1.0
1    2.0
2   -3.0
dtype: float64
```

In [229]:

```
pd.to_numeric(s, downcast = 'signed')
```

Out[229]:

```
0    1
1    2
2   -3
dtype: int8
```

In [230]:

```
sa = pd.Series(['apple', '1.0', '2', -3])

sa
```

Out[230]:

```
0    apple
1      1.0
2        2
3       -3
dtype: object
```

In [232]:

```
pd.to_numeric(sa,  errors = 'coerce')
```

Out[232]:

```
0    NaN
1    1.0
2    2.0
3   -3.0
dtype: float64
```

In [233]:

```
pd.to_numeric(sa, errors = 'ignore')
```

Out[233]:

```
0    apple
1      1.0
2        2
3       -3
dtype: object
```

In [237]:

```
import re

pattern = re.compile('\d{3}-\d{3}-\d{4}')

match = pattern.match('123-456-7890')

match_2 = pattern.match('1245-953-0123')

print(bool(match))

print(bool(match_2))
```

```
True
False
```

In [239]:

```
import re

matches = re.findall('\d+', 'the recipe calls for 10 strawberries and 1 banana')

print(matches)
```

```
['10', '1']
```

In [240]:

```
pattern1 = bool(re.match(pattern = '\d{3}-\d{3}-\d{4}', string = '123-642-2356'))

print(pattern1)
```

```
True
```

In [241]:

```
pattern2 = bool(re.match(pattern = '\$\d*.\d{2}', string = '$123.45'))

print(pattern2)
```

```
True
```

In [248]:

```
pattern3 = bool(re.match(pattern='\w', string='Australia'))

print(pattern3)
```

```
True
```

In [249]:

```
df_tips.head()
```

Out[249]:

| | total_bill | tip | sex | smoker | day | time | size |
|---|---|---|---|---|---|---|---|
| **0** | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 |
| **1** | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 |
| **2** | 21.01 | 3.50 | Male | No | Sun | Dinner | 3 |
| **3** | 23.68 | 3.31 | Male | No | Sun | Dinner | 2 |
| **4** | 24.59 | 3.61 | Female | No | Sun | Dinner | 4 |

In [275]:

```
def recode_gender(gender):
    import re
    import numpy as np
    if gender == 'Female':
        return 0
    elif gender == 'Male':
        return 1
    else:
        return np.nan
```

In [279]:

```
df_tips2['recode'] = df_tips2.sex.apply(recode_gender)
df_tips2.head()
```

Out[279]:

| | total_bill | tip | sex | smoker | day | time | size | repost | recode |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 | 0 | 0 |
| **1** | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 | 1 | 1 |
| **2** | 21.01 | 3.50 | Male | No | Sun | Dinner | 3 | 1 | 1 |
| **3** | 23.68 | 3.31 | Male | No | Sun | Dinner | 2 | 1 | 1 |
| **4** | 24.59 | 3.61 | Female | No | Sun | Dinner | 4 | 0 | 0 |

## 5. Case study

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: