Στατιστική των Επιχειρήσεων Ι

Διάλεξη 1η

Γιώργος Τσιρογιάννης

Τμήμα Διοίκησης Επιχειρήσεων Αγροτικών Προϊόντων και Τροφίμων,

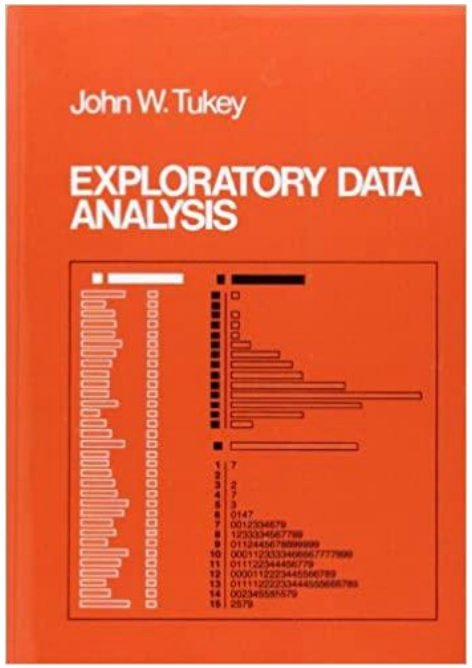
Πανεπιστήμιο Πατρών

Ακαδημαϊκό έτος: 2020-2021

Exploratory Data Analysis

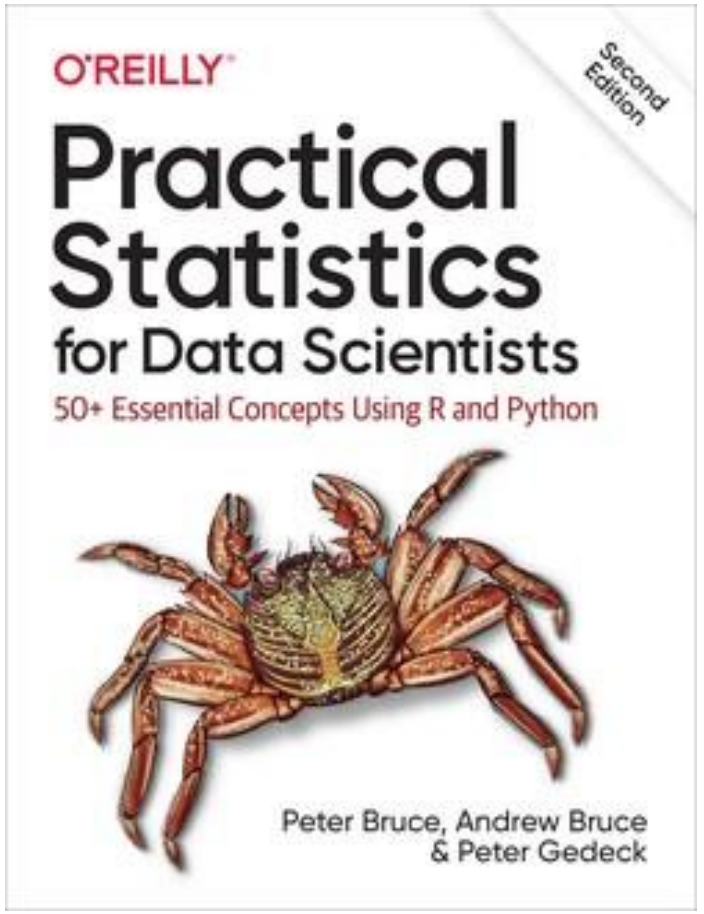
Εξερεύνηση των δεδομένων

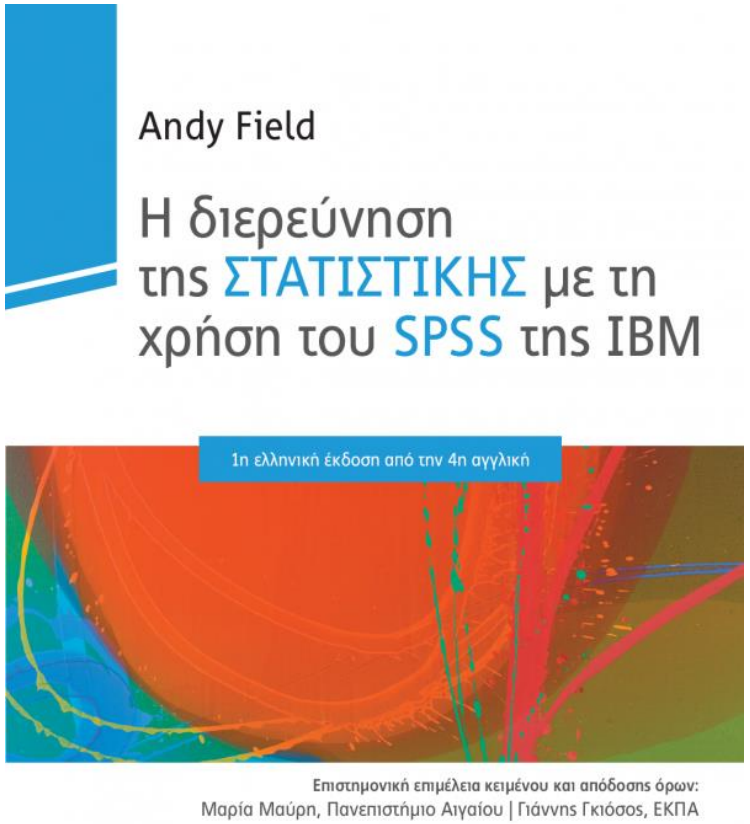
John Wilder Tukey
(1915 – 2000)



1977

Βιβλιογραφία





Πηγές δεδομένων

Κυρίο ενδιαφέρον: Rectangular Data

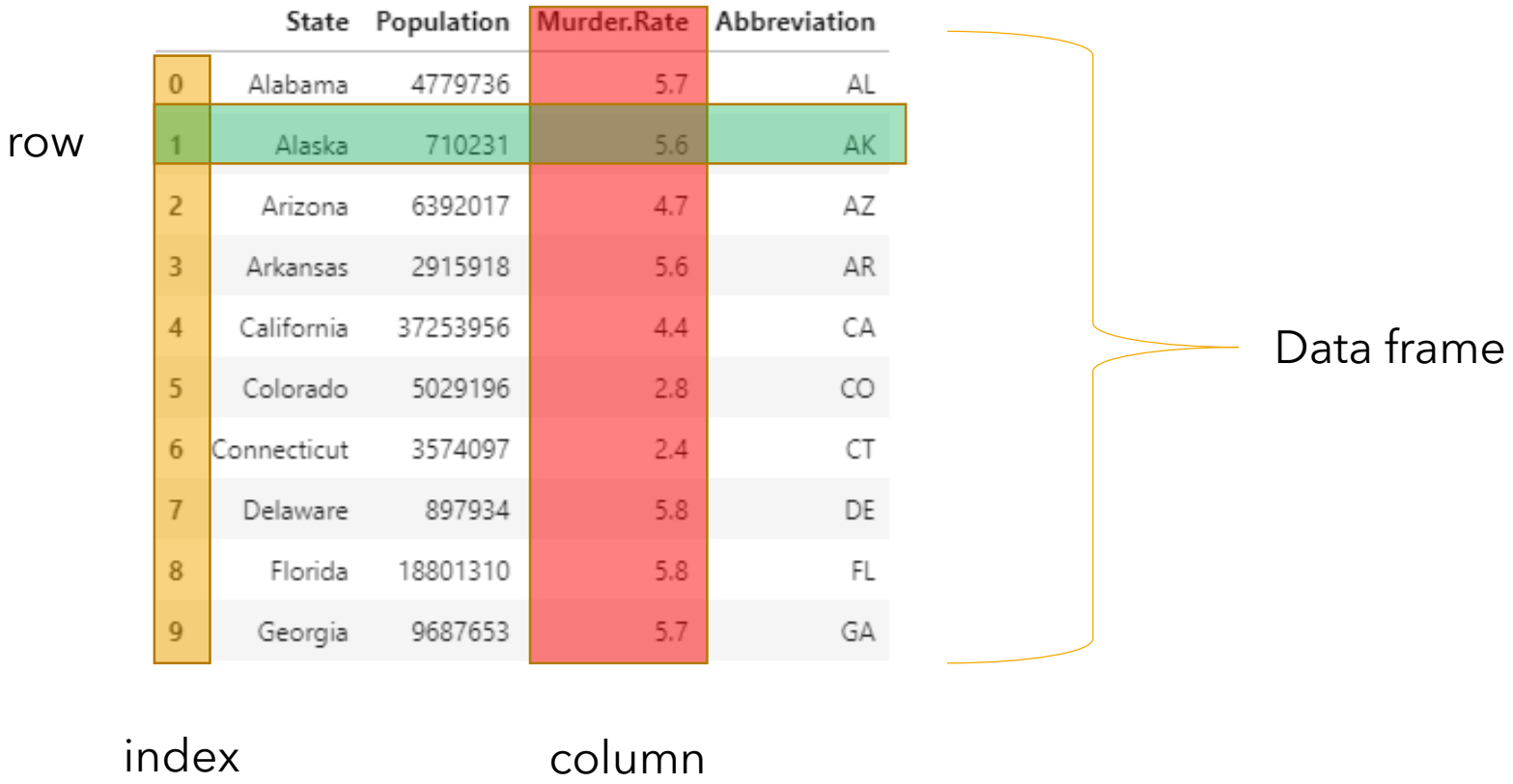
Μετρήσιμα δεδομένα (measured data)

	State	Population	Murder.Rate	Abbreviation
0	Alabama	4779736	5.7	AL
1	Alaska	710231	5.6	AK
2	Arizona	6392017	4.7	AZ
3	Arkansas	2915918	5.6	AR
4	California	37253956	4.4	CA
5	Colorado	5029196	2.8	CO
6	Connecticut	3574097	2.4	CT
7	Delaware	897934	5.8	DE
8	Florida	18801310	5.8	FL
9	Georgia	9687653	5.7	GA

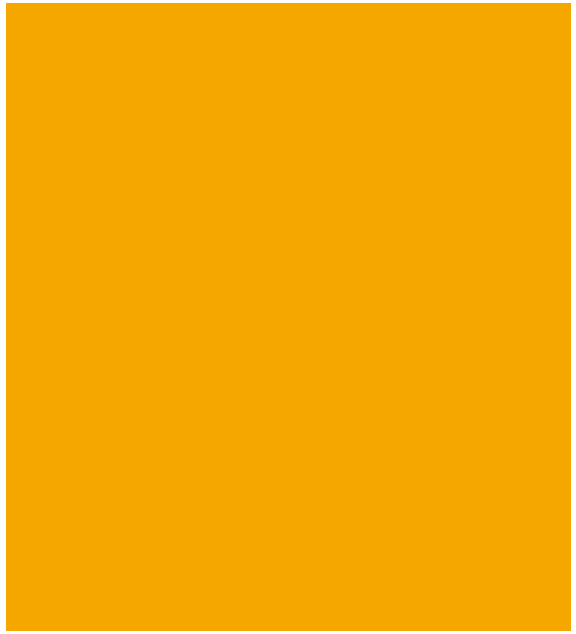
Category	currency	sellerRating	Duration	endDay	ClosePrice	OpenPrice	Competitive?
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	1
Automotive	US	3115	7	Tue	0.01	0.01	1

Κατηγοριοποιημένα δεδομένα (categorical data)

Data frames και indices



Πρακτικό μέρος



Μελέτη της εγκληματικότητάς στις ΗΠΑ

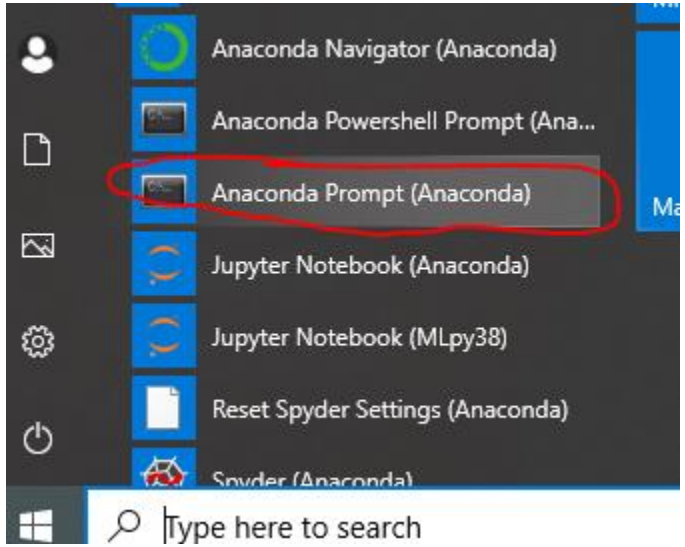
- «Φόρτωση» των βιβλιοθηκών
- «Φόρτωση» των δεδομένων
- Υπολογισμός παραμέτρων θέσης

Μελέτη της εγκληματικότητάς στις ΗΠΑ

- Δημιουργείστε φάκελο MBAStatLab και μετακινηθείτε σε αυτόν
- Κατεβάστε τα αρχεία από το φάκελο data στο https://eclass.upatras.gr/ και αντιγράψετε τα στον δικό σας υπολογιστή στον φάκελο data που θα δημιουργήσετε

Μελέτη της εγκληματικότητάς στις ΗΠΑ

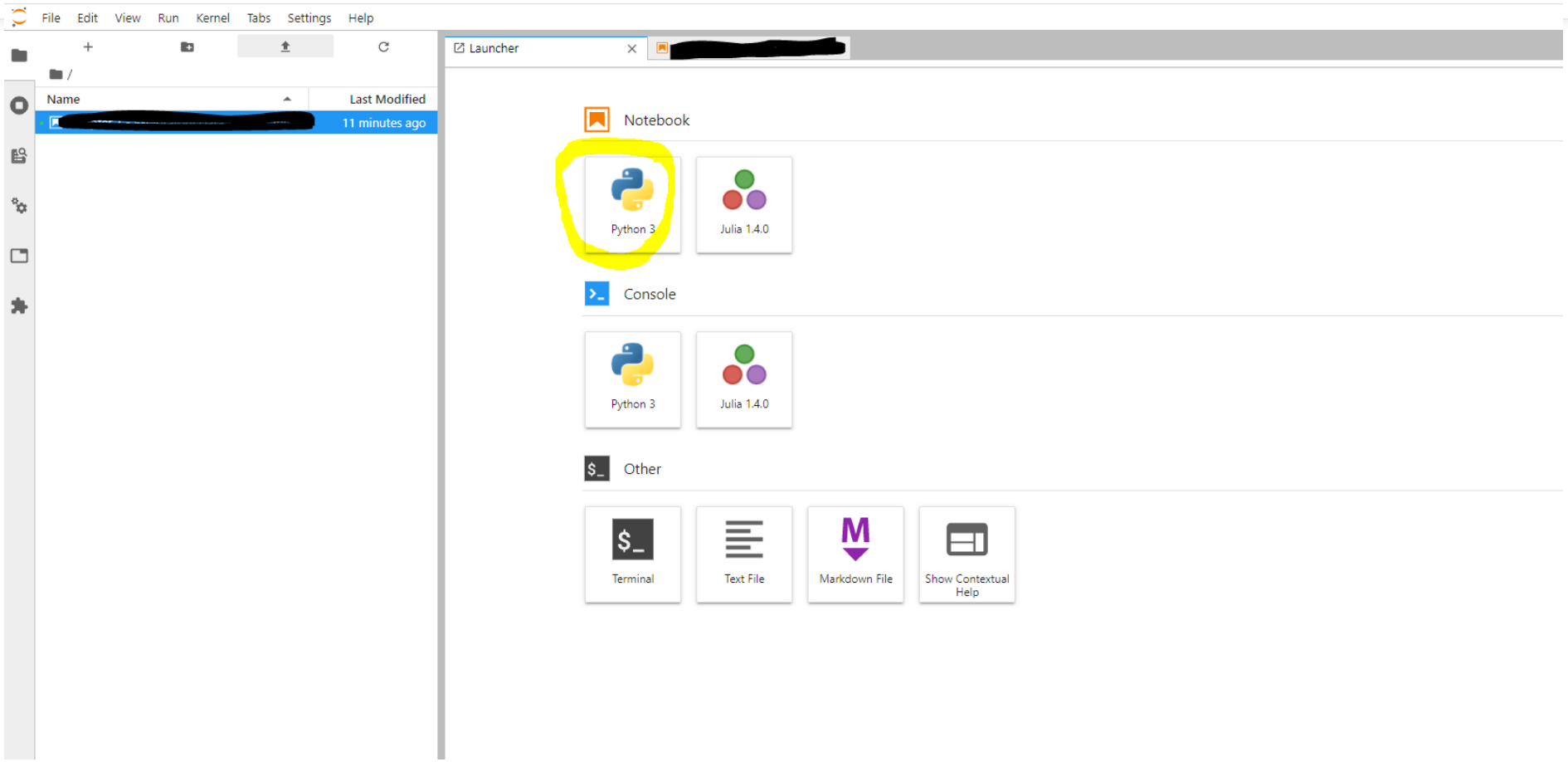
- Ενεργοποιήστε το περιβάλλον στην anaconda



Μελέτη της εγκληματικότητάς στις ΗΠΑ

- Μετακινηθείτε στον κατάλογο που έχετε αποθηκεύσει τα δεδομένα. Πχ:
  cd c:\diafora\folders\data
- Εκτελέστε jupyter lab

Μελέτη της εγκληματικότητάς στις ΗΠΑ



Μελέτη της εγκληματικότητάς στις ΗΠΑ

Φορτωση των βιβλιοθηκών

[1]:
```
from pathlib import Path

import pandas as pd
import numpy as np
from scipy.stats import trim_mean
from statsmodels import robust
#import wquantiles

import seaborn as sns
import matplotlib.pylab as plt
```

Shift Enter

Μελέτη της εγκληματικότητάς στις ΗΠΑ

```
AIRLINE_STATS_CSV = './airline_stats.csv'
KC_TAX_CSV = './kc_tax.csv'
LC_LOANS_CSV = './lc_loans.csv'
AIRPORT_DELAYS_CSV = './dfw_airline.csv'
SP500_DATA_CSV = './sp500_data.csv'
SP500_SECTORS_CSV = './sp500_sectors.csv'
STATE_CSV = './state.csv'
```

Μελέτη της εγκληματικότητάς στις ΗΠΑ

Φόρτωση δεδομένων

[4]:
```
state = pd.read_csv(STATE_CSV)
```

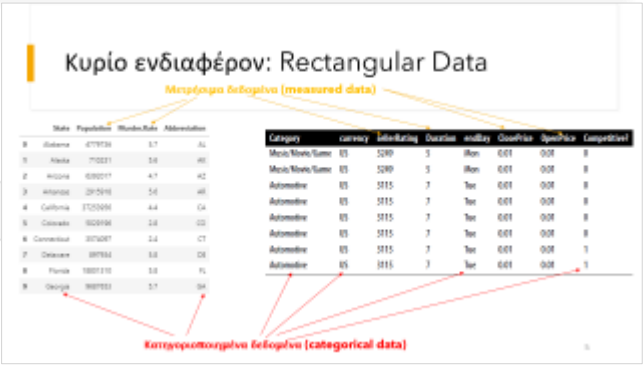
Προεσκόπηση του dataframe

[5]:
```
state.head(10)
```

[5]:

	State	Population	Murder.Rate	Abbreviation
0	Alabama	4779736	5.7	AL
1	Alaska	710231	5.6	AK
2	Arizona	6392017	4.7	AZ
3	Arkansas	2915918	5.6	AR
4	California	37253956	4.4	CA
5	Colorado	5029196	2.8	CO
6	Connecticut	3574097	2.4	CT
7	Delaware	897934	5.8	DE
8	Florida	18801310	5.8	FL
9	Georgia	9687653	5.7	GA

Μελέτη της εγκληματικότητάς στις ΗΠΑ



Μέγεθος τους dataframe

```
[6]: print(f'Αριθμός γραμμών: {state.shape[0]} και στηλών: {state.shape[1]}')
```

```
Αριθμός γραμμών: 50 και στηλών: 4
```

Index και ονοματα στηλών

```
[7]: print(f'Index του dataframe:{state.index.to_list()}')
```

```
Index του dataframe:[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 3
2, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49]
```

```
[8]: print(f'Ονοματα στηλών: {state.columns.to_list()}')
```

```
Ονοματα στηλών: ['State', 'Population', 'Murder.Rate', 'Abbreviation']
```

Εκτιμητές θέσης (Estimates of Location)

Μέσος (mean)

$$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

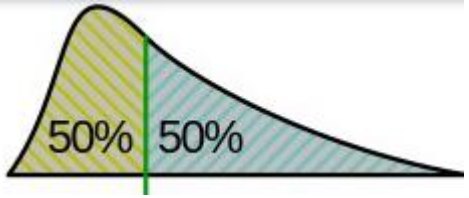
Σταθμισμένος μέσος (weighted mean)

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Trimmed mean

$$\text{Trimmed mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}$$

Διάμεσος (median)



Μελέτη της εγκληματικότητάς στις ΗΠΑ

Παράμετροι θέσης

Mean

[9]:
```
mean_v = state['Population'].mean()
print(f'Population Mean: {mean_v}')
```

```
Population Mean: 6162876.3
```

Trimmed mean (10%)

[10]:
```
trim_mean_v = trim_mean(state['Population'],0.1)
print(f'Population Trimmed Mean: {trim_mean_v}')
```

```
Population Trimmed Mean: 4783697.125
```

Median

[11]:
```
median_v = state['Population'].median()
print(f'Population Median: {median_v}')
```

```
Population Median: 4436369.5
```

Μελέτη της εγκληματικότητάς στις ΗΠΑ

Weighted mean

[12]:
```
print(f"Murder rate Mean: {state['Murder.Rate'].mean()}")
weighted_mean_v=np.average(state['Murder.Rate'], weights=state['Population'])
print(f"Murder weighted mean: {weighted_mean_v}")
```

```
Murder rate Mean: 4.066
Murder weighted mean: 4.445833981123393
```

Εκτιμητές διακύμανσης (Estimates of Variability)

Mean absolute deviation

$$\frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Standard deviation

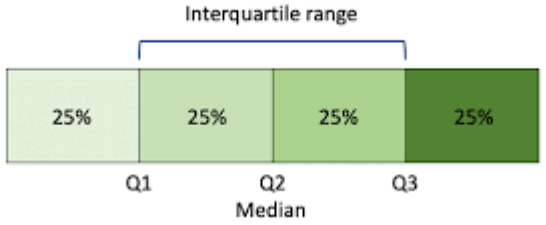
$$s = \sqrt{\text{Variance}}$$

Median absolute deviation (MAD)

$$\text{Median}(|x_1 - m|, |x_2 - m|, ..., |x_N - m|)$$

m =median

Interquartile range (IQR)



Μελέτη της εγκληματικότητάς στις ΗΠΑ

Εκτιμητές διακύμανσης

Standard deviation

[14]:
```
std_v = state['Population'].std()
print(f'Standard deviation: {std_v}')
```

```
Standard deviation: 6848235.347401142
```

Interquartile range (IQR) 25% - 75%

[16]:
```
IQR_v = state['Population'].quantile(0.75) - state['Population'].quantile(0.25)
print(f'IQR 25% - 75% : {IQR_v}')
```

```
IQR 25% - 75% : 4847308.0
```

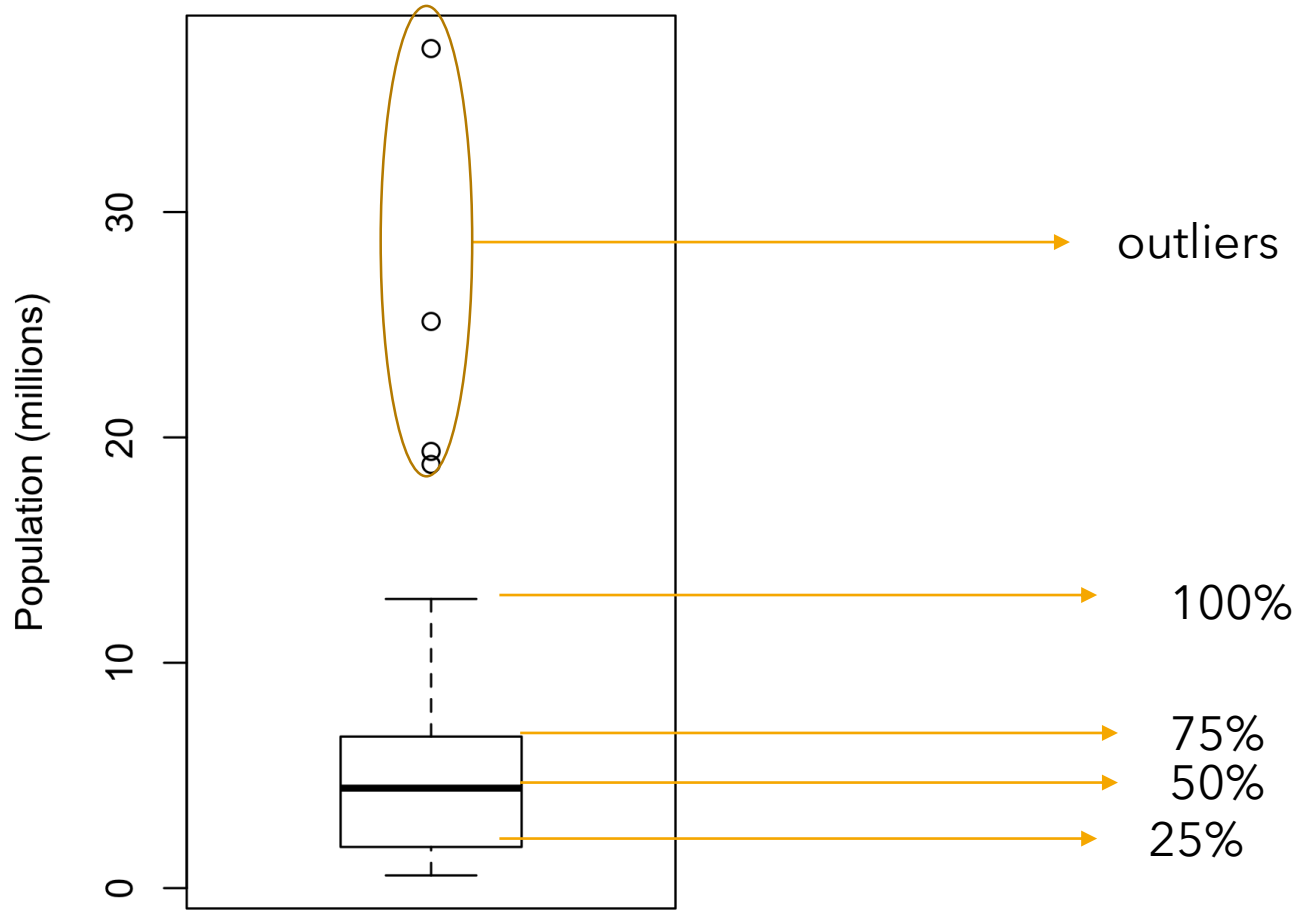
Median absolute deviation (MAD)

[18]:
```
MAD_v = robust.scale.mad(state['Population'])
print(f'MAD : {MAD_v}')
```

```
MAD : 3849876.1459979336
```

Εξερευνώντας όλη της κατανομή

Boxplot



Εξερευνώντας όλη της κατανομή

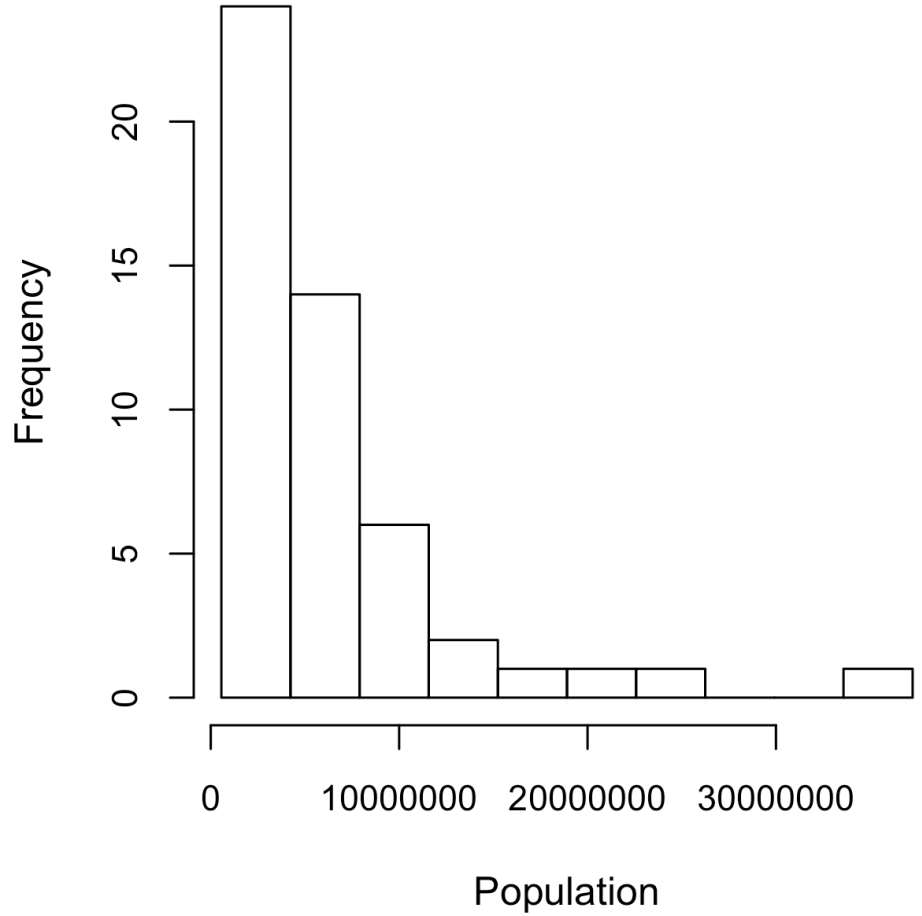
Πίνακας συχνοτήτων

BinNumber	BinRange	Count	States
1	563,626–4,232,658	24	WY,VT,ND,AK,SD,DE,MT,RI,NH,ME,HI,ID,NE,WV,NM,NV,UT,KS,AR,MS,IA,CT,OK,OR
2	4,232,659–7,901,691	14	KY,LA,SC,AL,CO,MN,WI,MD,MO,TN,AZ,IN,MA,WA
3	7,901,692–11,570,724	6	VA,NJ,NC,GA,MI,OH
4	11,570,725–15,239,757	2	PA,IL
5	15,239,758–18,908,790	1	FL
6	18,908,791–22,577,823	1	NY

Διαμερίσει σε διαστήματα και κατηγοριοποίηση

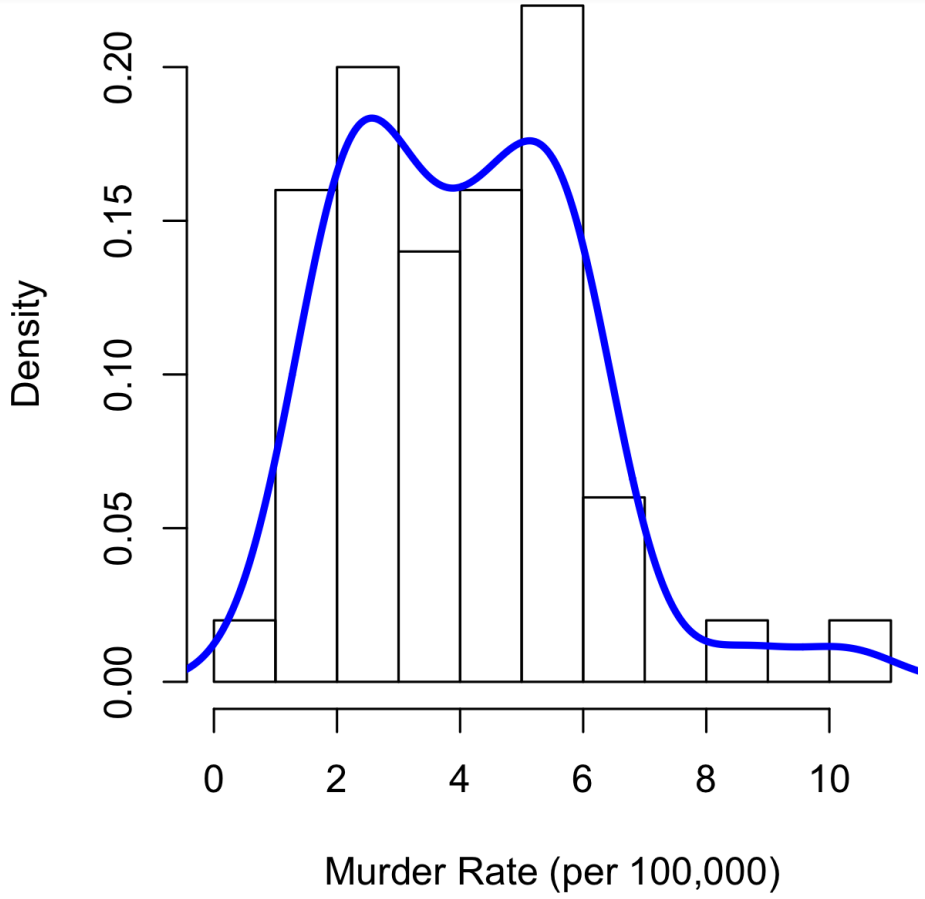
Εξερευνώντας όλη της κατανομή

Histogram



Εξερευνώντας όλη της κατανομή

Density plot



Μελέτη της εγκληματικότητάς στις ΗΠΑ

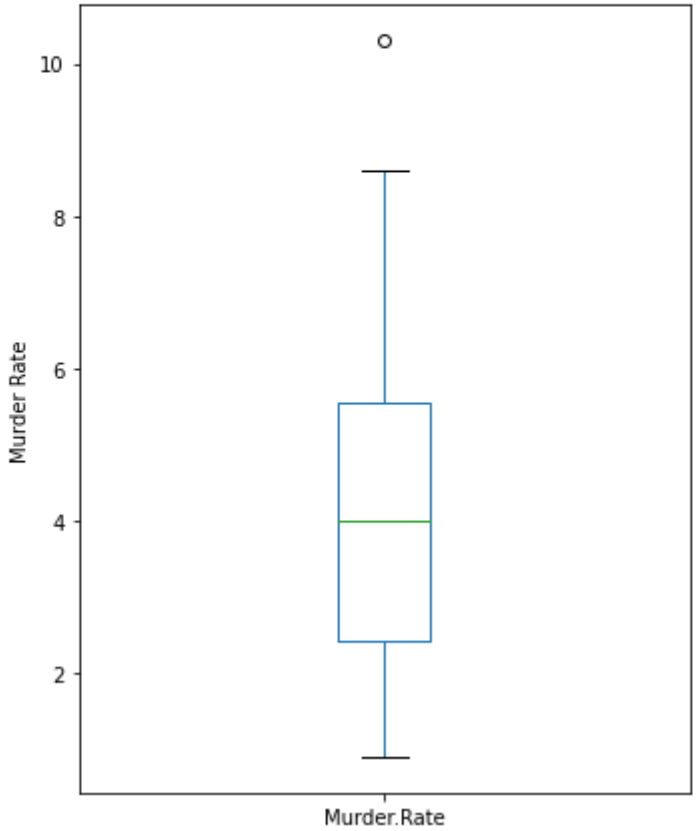
Percentiles και Boxplots

[22]:
```
quantiles_v = state['Murder.Rate'].quantile([0.05, 0.25, 0.5, 0.75, 0.95])
print(f'Quantiles:\n{quantiles_v}')
```

```
Quantiles:
0.05    1.600
0.25    2.425
0.50    4.000
0.75    5.550
0.95    6.510
Name: Murder.Rate, dtype: float64
```

Μελέτη της εγκληματικότητάς στις ΗΠΑ

[37]:
```
ax = (state['Murder.Rate']).plot.box(figsize=(5, 6))
ax.set_ylabel('Murder Rate')
plt.tight_layout()
plt.show()
```



Μελέτη της εγκληματικότητάς στις ΗΠΑ

Πίνακες συχνοτήτων και Histograms

Πίνακες συχνοτήτων

```
[44]: binnedPopulation = pd.cut(state['Population'], 8)
      print(f'Διστήματα πληθυσμού:\n{binnedPopulation.value_counts()}')
```

```
Διστήματα πληθυσμού:
(526935.67, 5149917.25]      29
(5149917.25, 9736208.5]      13
(9736208.5, 14322499.75]      4
(32667664.75, 37253956.0]     1
(23495082.25, 28081373.5]     1
(18908791.0, 23495082.25]     1
(14322499.75, 18908791.0]     1
(28081373.5, 32667664.75]     0
Name: Population, dtype: int64
```

Μελέτη της εγκληματικότητάς στις ΗΠΑ

[47]:
```
binnedPopulation.name = 'binnedPopulation'
df = pd.concat([state, binnedPopulation], axis=1)
df = df.sort_values(by='Population')
groups = []
for group, subset in df.groupby(by='binnedPopulation'):
    groups.append({
        'BinRange': group,
        'Count': len(subset),
        'States': ','.join(subset.Abbreviation)
    })
groups_df = pd.DataFrame(groups)
groups_df.head(8)
```

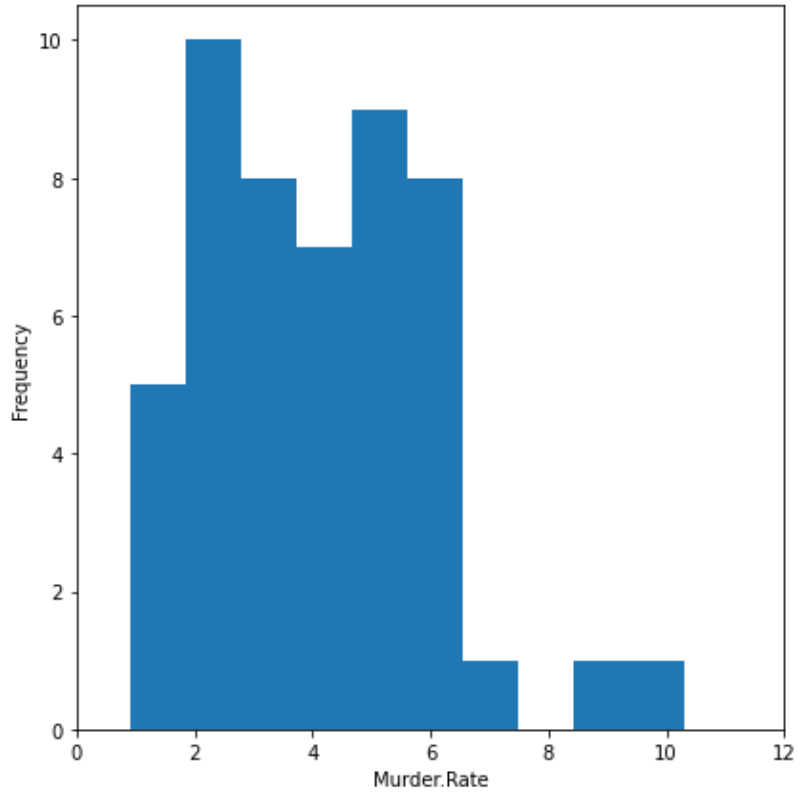
[47]:

	BinRange	Count	States
0	(526935.67, 5149917.25]	29	WY,VT,ND,AK,SD,DE,MT,RI,NH,ME,HI,ID,NE,WV,NM,N...
1	(5149917.25, 9736208.5]	13	MN,WI,MD,MO,TN,AZ,IN,MA,WA,VA,NJ,NC,GA
2	(9736208.5, 14322499.75]	4	MI,OH,PA,IL
3	(14322499.75, 18908791.0]	1	FL
4	(18908791.0, 23495082.25]	1	NY
5	(23495082.25, 28081373.5]	1	TX
6	(28081373.5, 32667664.75]	0	
7	(32667664.75, 37253956.0]	1	CA

Μελέτη της εγκληματικότητάς στις ΗΠΑ

Histograms

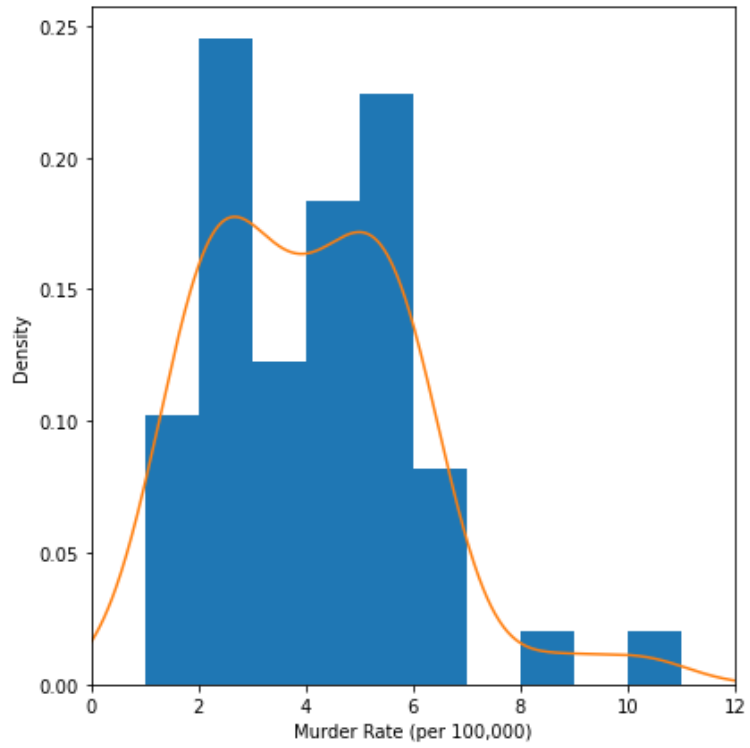
[36]:
```
ax = (state['Murder.Rate']).plot.hist(figsize=(6, 6), xlim=[0, 12])
ax.set_xlabel('Murder.Rate')
plt.tight_layout()
plt.show()
```



Μελέτη της εγκληματικότητάς στις ΗΠΑ

Density plots

```
[35]: ax = state['Murder.Rate'].plot.hist(density=True, xlim=[0, 12],
                                          bins=range(1,12), figsize=(6, 6))
state['Murder.Rate'].plot.density(ax=ax)
ax.set_xlabel('Murder Rate (per 100,000)')
plt.tight_layout()
plt.show()
```



Καθυστερήσεις πτήσεων

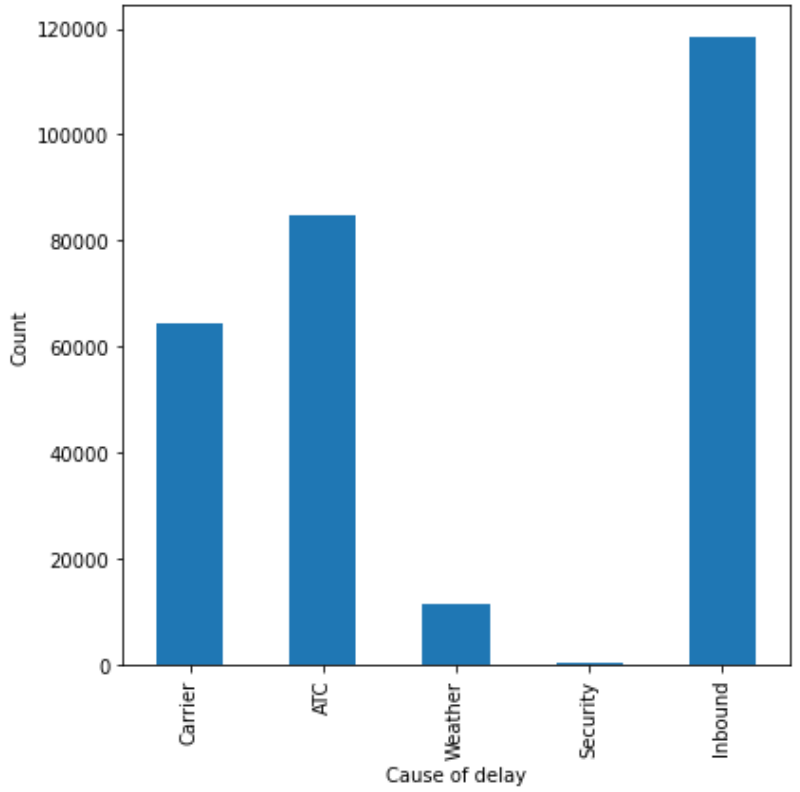
Bar Plots

[50]:
```
delays = pd.read_csv(AIRPORT_DELAYS_CSV)
delays.head()
```

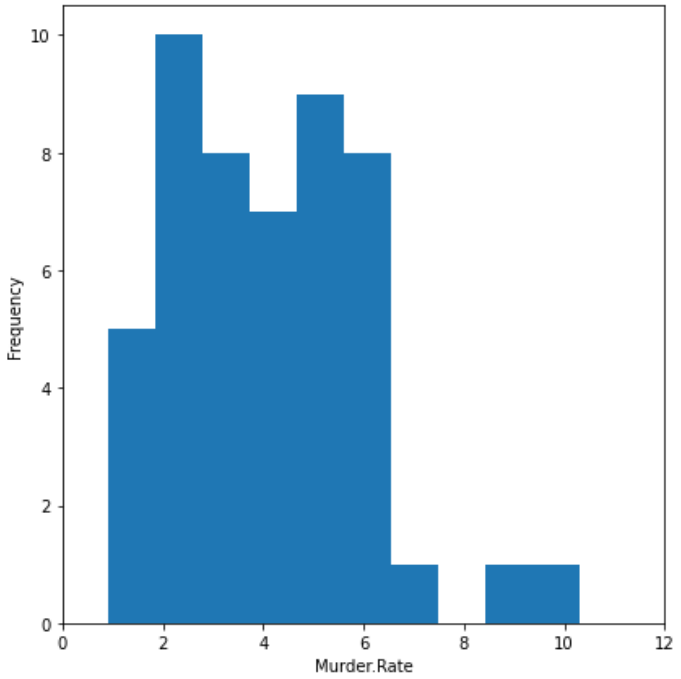
[50]:

	Carrier	ATC	Weather	Security	Inbound
0	64263.16	84856.5	11235.42	343.15	118427.82

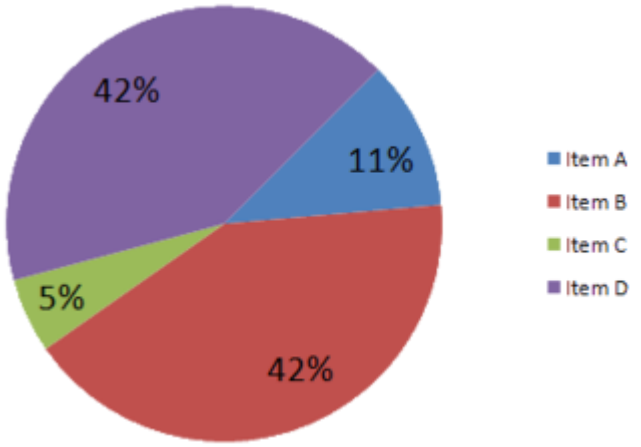
[51]:
```
ax = delays.transpose().plot.bar(figsize=(6, 6), legend=False)
ax.set_xlabel('Cause of delay')
ax.set_ylabel('Count')
plt.tight_layout()
plt.show()
```



Δελεαστικό αλλά απαράδεκτο

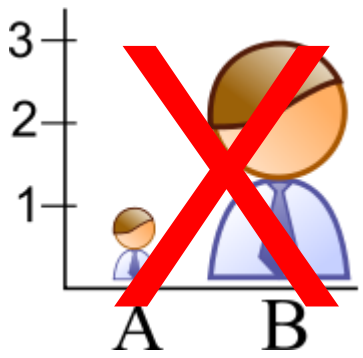


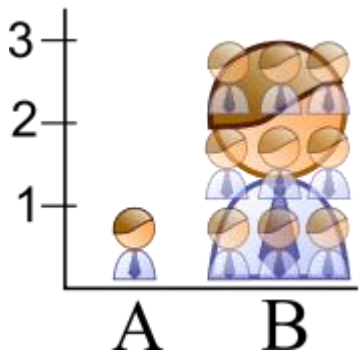
Δελεαστικό αλλά απαράδεκτο

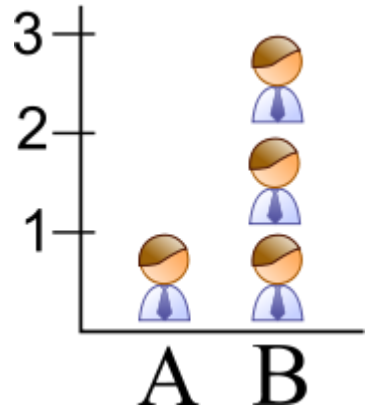


https://en.wikipedia.org/wiki/Misleading_graph

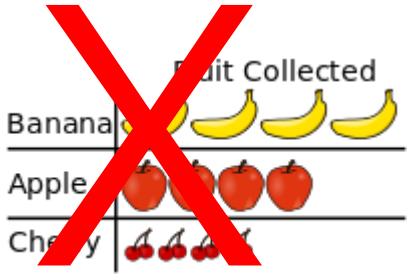
Δελεαστικό αλλά απαράδεκτο

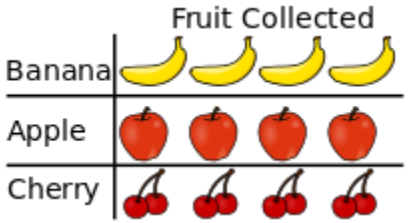




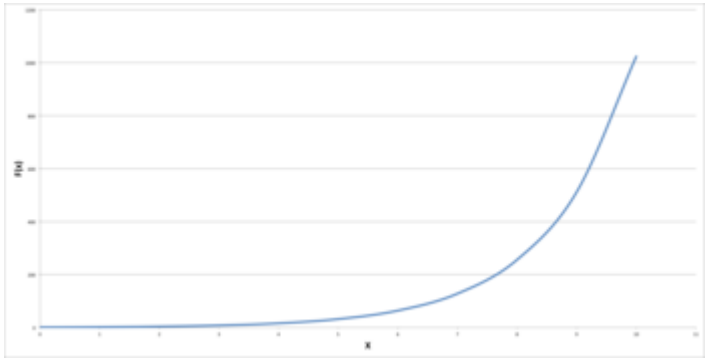


Δελεαστικό αλλά απαράδεκτο

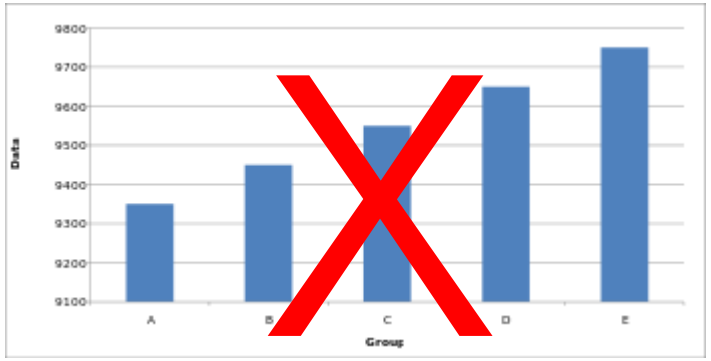


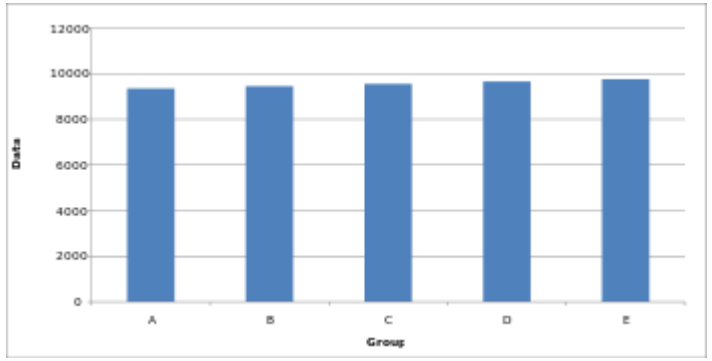


Δελεαστικό αλλά απαράδεκτο

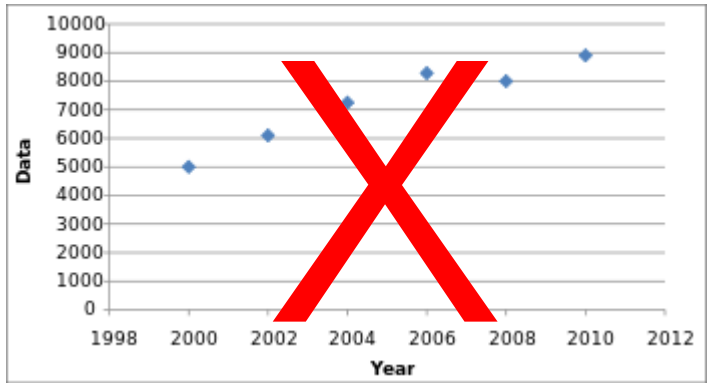


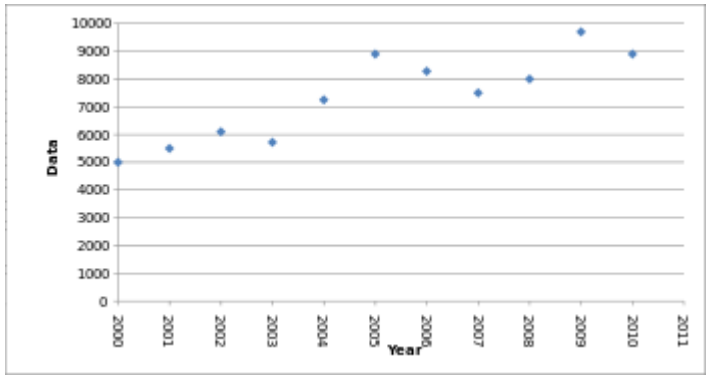
Δελεαστικό αλλά απαράδεκτο





Δελεαστικό αλλά απαράδεκτο





Συσχέτιση

- Ένα ζεύγος μεταβλητών X,Y:
  - Είναι θετικά σχετιζόμενες όταν το X αυξάνει το Y αυξάνει με γραμμικό τρόπο
  - Είναι αρνητικά σχετιζόμενες όταν το X αυξάνει το Y μειώνεται με γραμμικό τρόπο
- Ποσοτικά πχ Pearson's correlation coefficient:

$$r = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Δεδομένα του χρηματιστηρίου

Μέτρα συσχέτισης

Φόρτωση δεδομένων

[26]:
```
sp500_sym = pd.read_csv(SP500_SECTORS_CSV)
sp500_px = pd.read_csv(SP500_DATA_CSV, index_col=0)
```

Μορφή δεδομένων

Επισκόπηση των δεδομένων

[28]:
```
sp500_sym.head(5)
```

[28]:

	sector	sector_label	sub_sector	symbol
0	information_technology	Technology	data_processing_&_outsourced_services	ADS
1	information_technology	Technology	systems_software	CA
2	information_technology	Technology	systems_software	MSFT
3	information_technology	Technology	systems_software	RHT
4	information_technology	Technology	it_consulting_&_services	CTSH

[29]:
```
sp500_px.head(5)
```

[29]:

	ADS	CA	MSFT	RHT	CTSH	CSC	EMC	IBM	XRX	ALTR	...	WAT	ALXN	AMGN	BXLT	BIIB	CELG	GILD	REGN	VRTX	HSIC
1993-01-29	0.0	0.060124	-0.022100	0.0	0.0	0.018897	0.007368	0.092165	0.259140	-0.007105	...	0.0	0.0	0.34716	0.0	0.04167	0.00000	0.015564	1.75	0.1250	0.0
1993-02-01	0.0	-0.180389	0.027621	0.0	0.0	0.018889	0.018425	0.115207	-0.100775	0.063893	...	0.0	0.0	-0.23144	0.0	0.00000	-0.01041	0.007782	1.25	0.1250	0.0
1993-02-02	0.0	-0.120257	0.035900	0.0	0.0	-0.075573	0.029482	-0.023041	0.028796	-0.014192	...	0.0	0.0	-0.11572	0.0	0.00000	0.00000	-0.007792	-0.25	0.0000	0.0
1993-02-03	0.0	0.060124	-0.024857	0.0	0.0	-0.151128	0.003689	-0.253454	-0.043190	-0.007105	...	0.0	0.0	-0.08679	0.0	0.04167	-0.04167	-0.038919	-0.50	0.0625	0.0
1993-02-04	0.0	-0.360770	-0.060757	0.0	0.0	0.113350	-0.022114	0.069862	0.000000	-0.007096	...	0.0	0.0	0.14465	0.0	-0.04166	-0.03126	-0.046711	0.00	0.0625	0.0

5 rows × 517 columns

Επιλογή υποσυνόλου ενδιαφέροντος

Επιλογή δεδομένων

[30]:
```
# τηλεπικοινωνίες
telecomSymbols = sp500_sym[sp500_sym['sector'] == 'telecommunications_services']['symbol']
# μετά το H2 2017
telecom = sp500_px.loc[sp500_px.index >= '2012-07-01', telecomSymbols]
```

[37]:
```
print(f'Τηλεπικοινωνιακές εταιρείες:\n{list(telecomSymbols)}')
```

```
Τηλεπικοινωνιακές εταιρείες:
['T', 'CTL', 'FTR', 'VZ', 'LVLT']
```

[38]:
```
telecom.head(5)
```

[38]:

	T	CTL	FTR	VZ	LVLT
2012-07-02	0.422496	0.140847	0.070879	0.554180	-0.519998
2012-07-03	-0.177448	0.066280	0.070879	-0.025976	-0.049999
2012-07-05	-0.160548	-0.132563	0.055128	-0.051956	-0.180000
2012-07-06	0.342205	0.132563	0.007875	0.140106	-0.359999
2012-07-09	0.136883	0.124279	-0.023626	0.253943	0.180000

Πίνακας συσχετίσεων

Υπολογισμός πίνακα συσχετίσεων

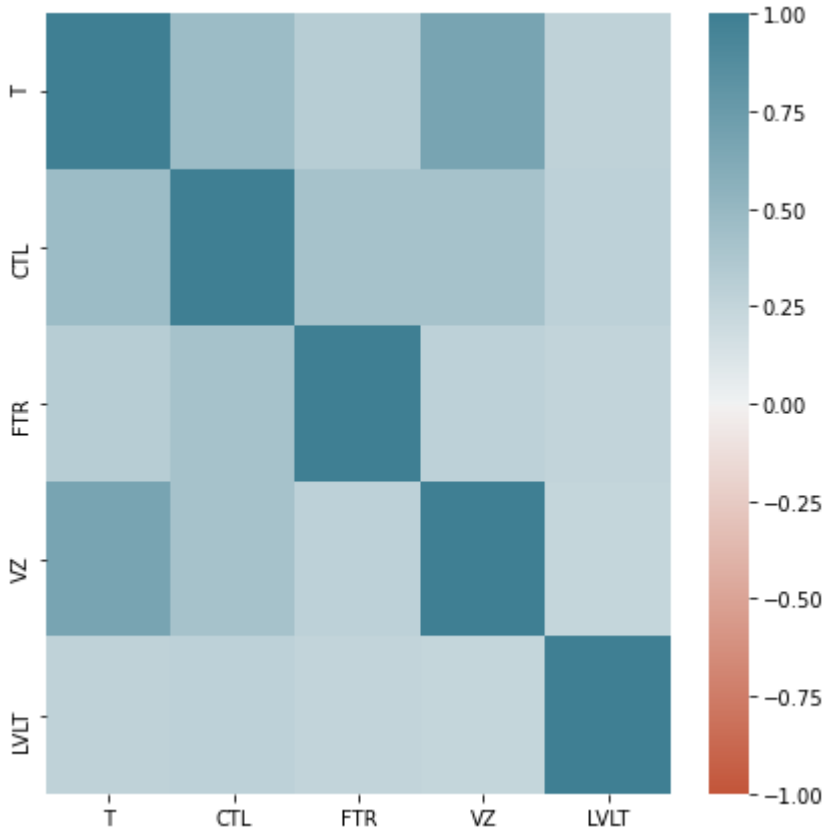
[40]:
```
telecom_corr = telecom.corr()
print(telecom_corr)
```

```
              T       CTL       FTR        VZ      LVLT
T      1.000000  0.474683  0.327767  0.677612  0.278626
CTL    0.474683  1.000000  0.419757  0.416604  0.286665
FTR    0.327767  0.419757  1.000000  0.287386  0.260068
VZ     0.677612  0.416604  0.287386  1.000000  0.242199
LVLT   0.278626  0.286665  0.260068  0.242199  1.000000
```

Γραφική αναπαράσταση πίνακα συσχετίσεων

Γραφική αναπαράσταση πίνακα συσχετίσεων

[43]:
```
fig, ax = plt.subplots(figsize=(6, 6))
ax = sns.heatmap(telecom_corr, vmin=-1, vmax=1,
                 cmap=sns.diverging_palette(20, 220, as_cmap=True),
                 ax=ax)

plt.tight_layout()
plt.show()
```



Άσκηση

- Για τα δεδομένα που αντιστοιχούν στα sp500_sym και sp500_px
- Υπολογίστε των πίνακα συσχετίσεων και την γραφική τους απεικόνιση για τις μετοχές του κλάδου των exchange traded fund (κωδικός etf) για την περίοδο της α)τελευταίας δεκαετίας (1/12/2010) καθώς και την β)περίοδο 1/1/2010-31/12/2012.
- Να ποσοτικοποιήσετε τις διαφοροποιήσεις των συσχετίσεων στις περιόδους. Τι παρατηρείτε;

Άσκηση: περίοδος α

Επιλογή δεομένων

```
[50]: # etf
etfSymbols = sp500_sym[sp500_sym['sector'] == 'etf']['symbol']
# μετά το 1/1/2010
etf_period1 = sp500_px.loc[sp500_px.index >= '2010-01-01', etfSymbols]
```

Υπολογισμός συσχετίσεων

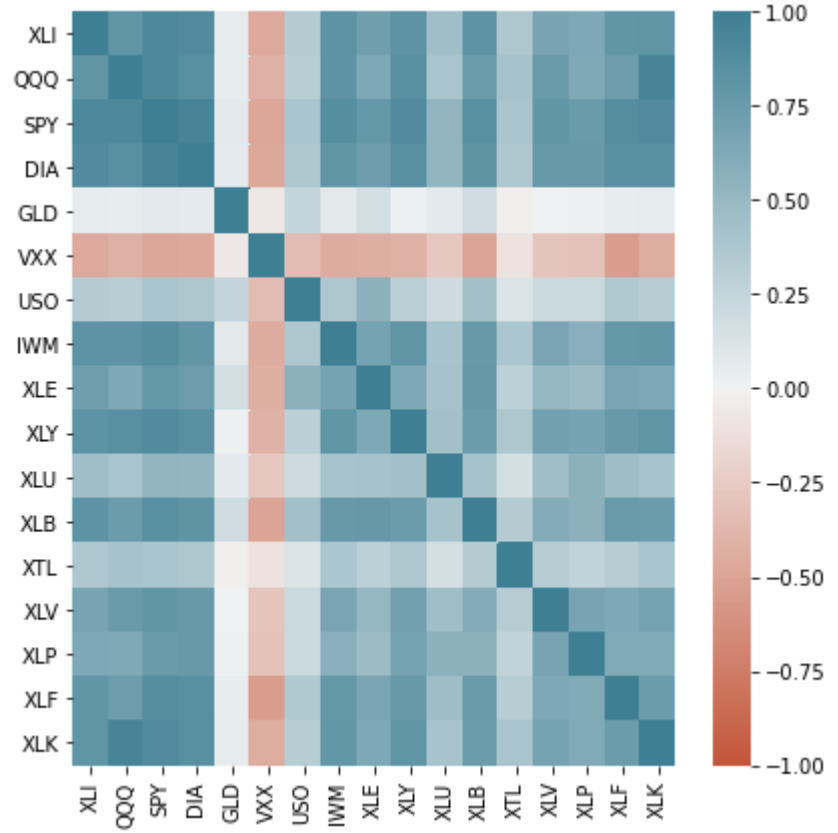
```
[51]: etf_period1_corr = etf_period1.corr()
print(etf_period1_corr)
```

```
          XLI       QQQ       SPY       DIA       GLD       VXX       USO  \
XLI  1.000000  0.810654  0.908960  0.894576  0.058281 -0.463916  0.339243
QQQ  0.810654  1.000000  0.910479  0.844504  0.052627 -0.413374  0.312938
SPY  0.908960  0.910479  1.000000  0.960449  0.075316 -0.491029  0.390772
DIA  0.894576  0.844504  0.960449  1.000000  0.065750 -0.471257  0.367450
GLD  0.058281  0.052627  0.075316  0.065750  1.000000 -0.070221  0.254995
VXX -0.463916 -0.413374 -0.491029 -0.471257 -0.070221  1.000000 -0.342259
USO  0.339243  0.312938  0.390772  0.367450  0.254995 -0.342259  1.000000
IWM  0.826674  0.825026  0.868769  0.802741  0.078353 -0.449434  0.374186
XLE  0.719189  0.644080  0.783781  0.739586  0.177902 -0.431351  0.556834
XLY  0.827128  0.858607  0.896541  0.857940  0.016209 -0.418040  0.299458
XLU  0.447937  0.401213  0.538585  0.530041  0.074534 -0.269019  0.200635
XLB  0.820344  0.746891  0.851822  0.815363  0.192808 -0.496403  0.434781
XTL  0.373309  0.415997  0.398989  0.373221 -0.022173 -0.093965  0.119464
XLV  0.679303  0.761216  0.799276  0.770879  0.009840 -0.289544  0.218657
XLP  0.634650  0.626777  0.751165  0.766398  0.023367 -0.310964  0.212372
XLF  0.799502  0.737757  0.881729  0.850399  0.047296 -0.533126  0.360146
XLK  0.806268  0.954522  0.899784  0.857896  0.060232 -0.439559  0.326921
```

Άσκηση: περίοδος α

Γραφική αναπαράσταση

[62]:
```
fig, ax = plt.subplots(figsize=(6, 6))
ax = sns.heatmap(etf_period1_corr, vmin=-1, vmax=1,
                 cmap=sns.diverging_palette(20, 220, as_cmap=True),
                 ax=ax)
plt.tight_layout()
plt.show()
```



Άσκηση: περίοδος β

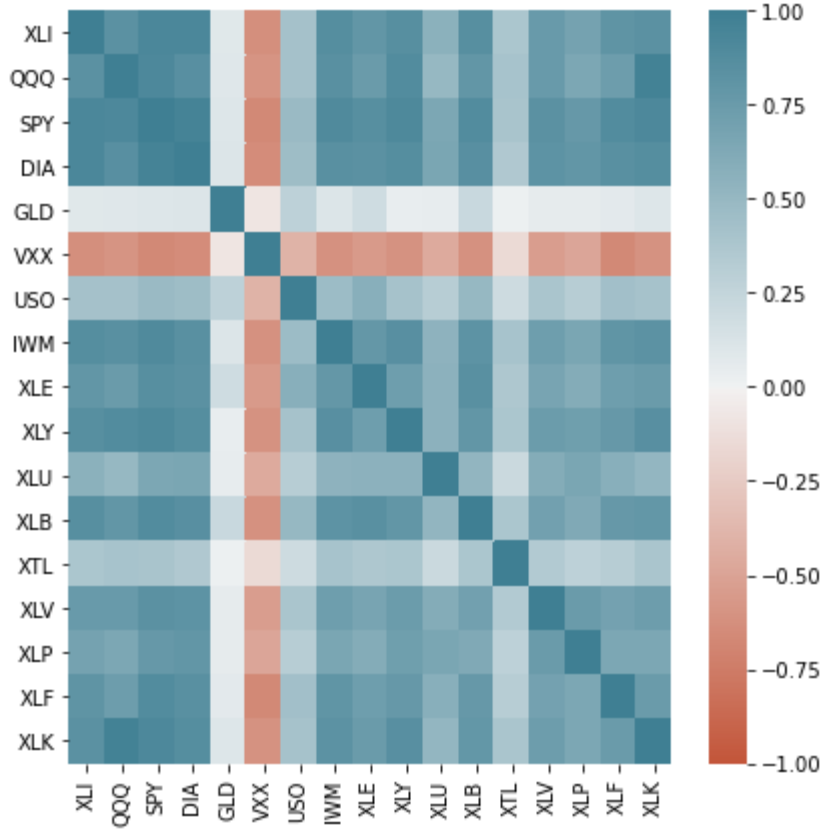
Περίοδος β

[63]:
```
# etf
etfSymbols = sp500_sym[sp500_sym['sector'] == 'etf']['symbol']
# περιοδος 1/1/2010-31/12/2012
etf_period2 = sp500_px.loc[(sp500_px.index >= '2010-01-01') & (sp500_px.index <= '2012-12-31'), etfSymbols]
```

[64]:
```
etf_period2_corr = etf_period2.corr()
print(etf_period2_corr)
```

Άσκηση: περίοδος β

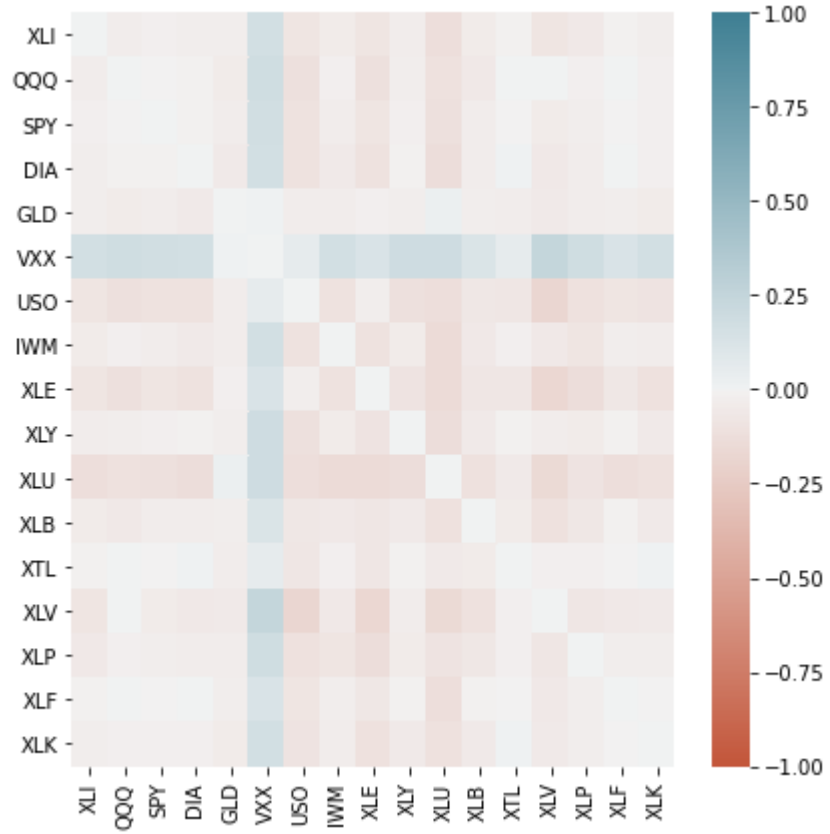
[65]:
```
fig, ax = plt.subplots(figsize=(6, 6))
ax = sns.heatmap(etf_period2_corr, vmin=-1, vmax=1,
                 cmap=sns.diverging_palette(20, 220, as_cmap=True),
                 ax=ax)
plt.tight_layout()
plt.show()
```



Άσκηση

- Να ποσοτικοποιήσετε τις διαφοροποιήσεις των συσχετίσεων στις περιόδους

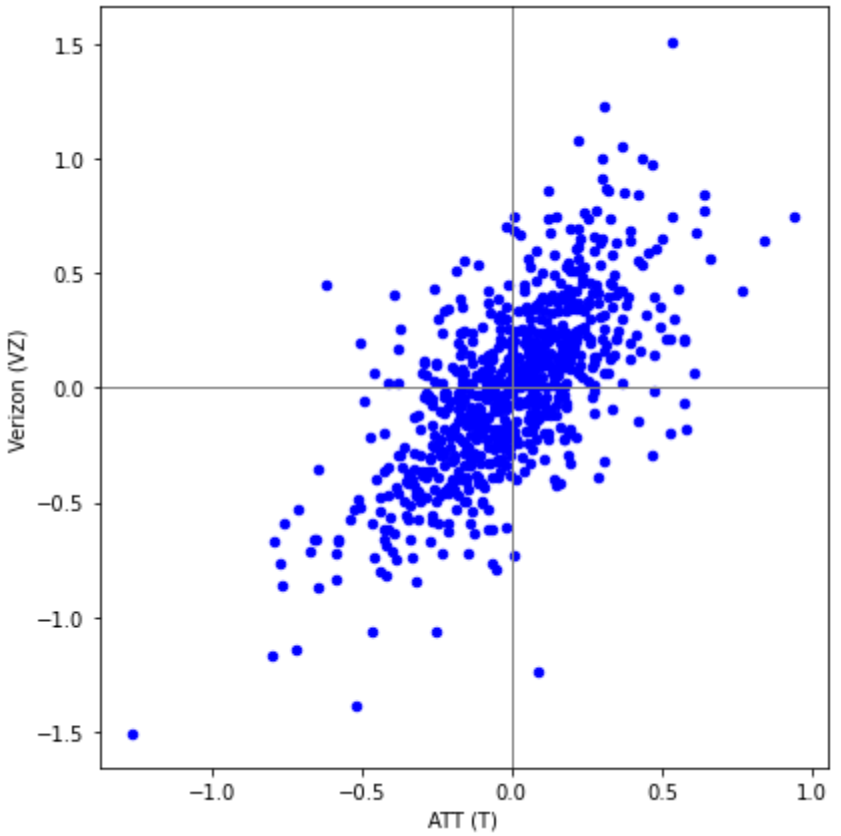
```
[67]: fig, ax = plt.subplots(figsize=(6, 6))
ax = sns.heatmap(etf_period1_corr - etf_period2_corr, vmin=-1, vmax=1,
                 cmap=sns.diverging_palette(20, 220, as_cmap=True),
                 ax=ax)
plt.tight_layout()
plt.show()
```



Προβολή αλληλεπιδράσεων μεταβλητών

Scatterplot

```
[76]: ax = telecom.plot.scatter(x='T', y='VZ', figsize=(6, 6), c='Blue')
ax.set_xlabel('ATT (T)')
ax.set_ylabel('Verizon (VZ)')
ax.axhline(0, color='grey', lw=1)
ax.axvline(0, color='grey', lw=1)
plt.tight_layout()
plt.show()
```



Προβολή αλληλεπιδράσεων μεταβλητών (μεγάλο πλήθος δεδομένων)

Φόρτωση δεδομένων (τιμές ακινήτων)

[80]:
```
kc_tax = pd.read_csv(KC_TAX_CSV)
kc_tax0 = kc_tax.loc[(kc_tax.TaxAssessedValue < 750000) &
                     (kc_tax.SqFtTotLiving > 100) &
                     (kc_tax.SqFtTotLiving < 3500), :]
print(f'Μέγεθος των δεδομένων: {kc_tax0.shape}')
```

```
Μέγεθος των δεδομένων: (432693, 3)
```

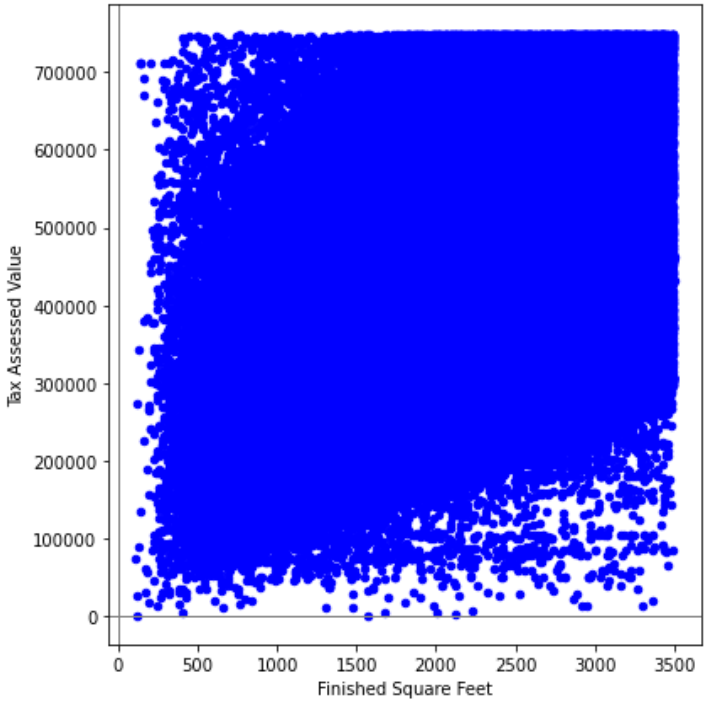
[82]:
```
kc_tax0.head(5)
```

[82]:

	TaxAssessedValue	SqFtTotLiving	ZipCode
1	206000.0	1870	98002.0
2	303000.0	1530	98166.0
3	361000.0	2000	98108.0
4	459000.0	3150	98108.0
5	223000.0	1570	98032.0

Προβολή αλληλεπιδράσεων μεταβλητών (μεγάλο πλήθος δεδομένων)

[84]:
```
ax = kc_tax0.plot.scatter(x='SqFtTotLiving', y='TaxAssessedValue', figsize=(6, 6), c='Blue')
ax.set_xlabel('Finished Square Feet')
ax.set_ylabel('Tax Assessed Value')
ax.axhline(0, color='grey', lw=1)
ax.axvline(0, color='grey', lw=1)
plt.tight_layout()
plt.show()
```

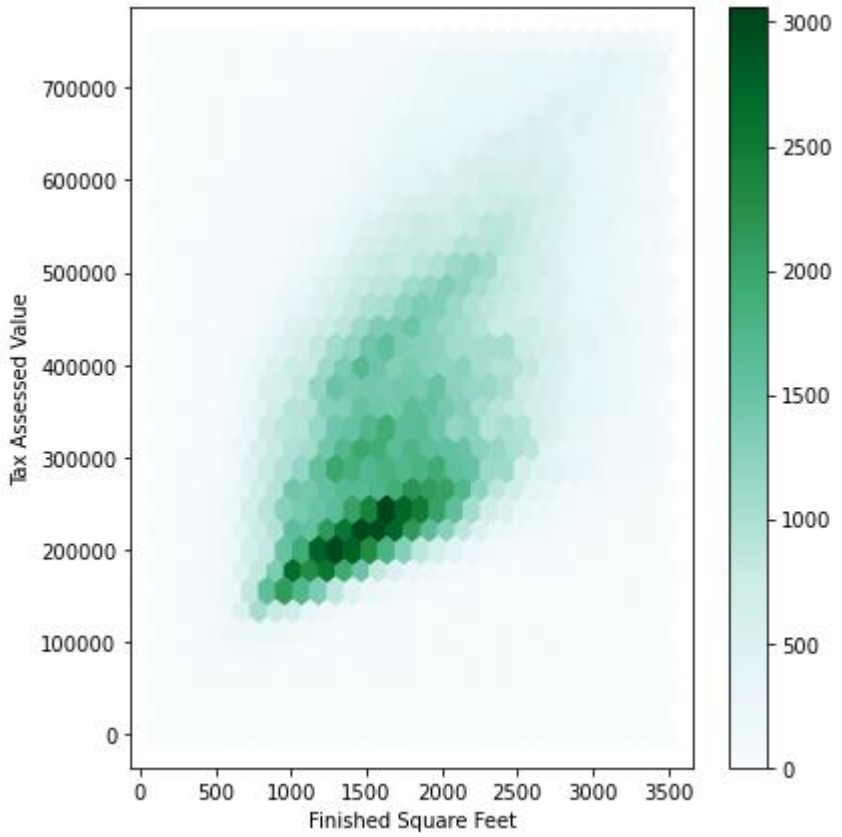


Περιορισμένη κατανόηση

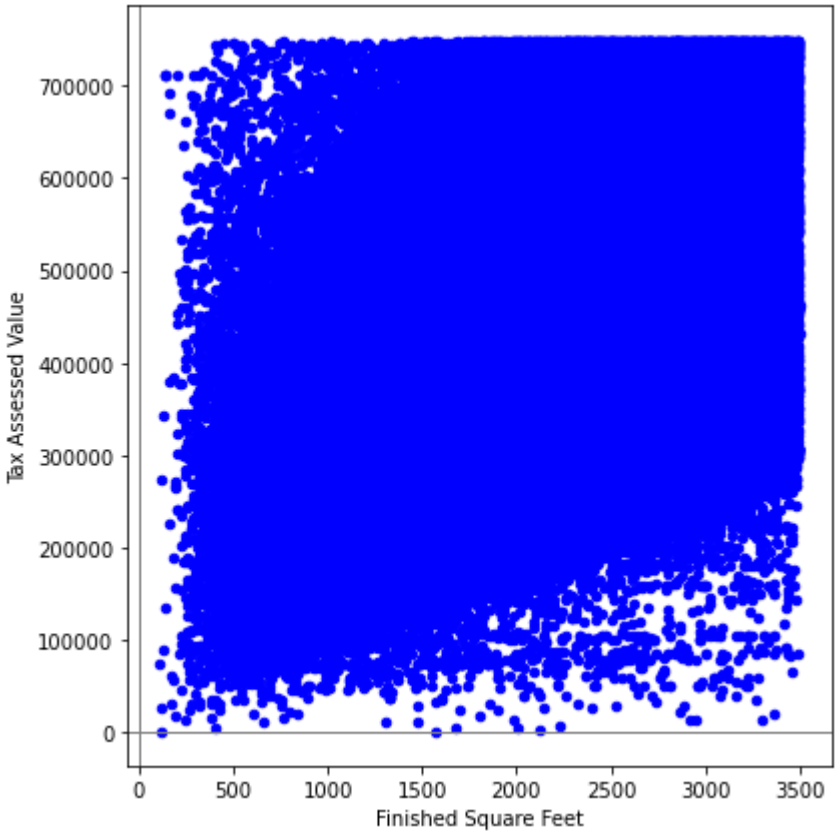
Hexagonal binning

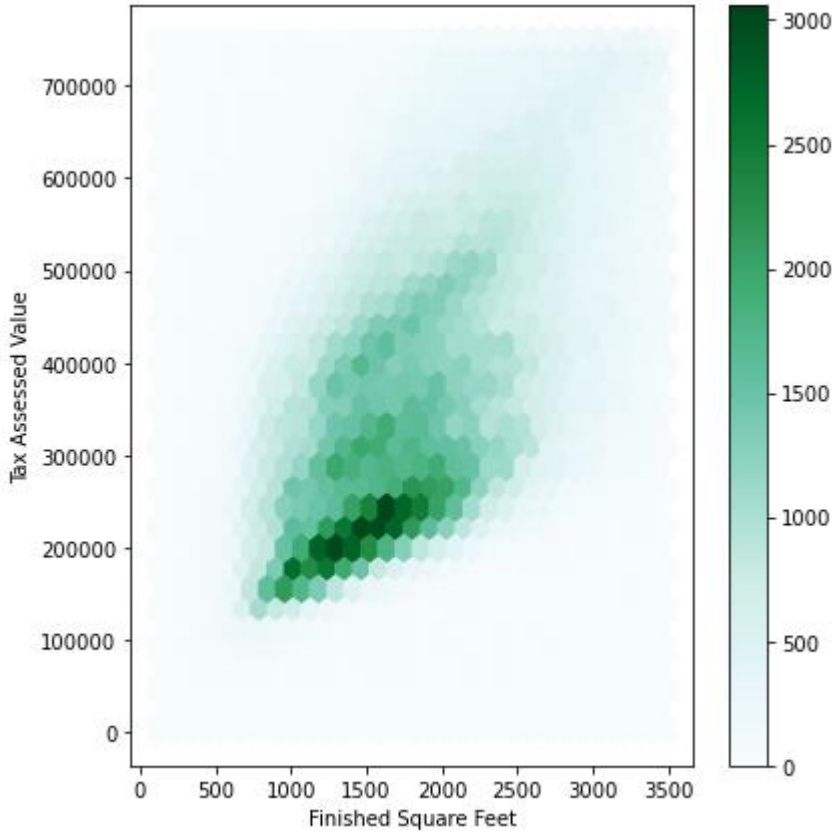
Hexagonal binning

[88]:
```
ax = kc_tax0.plot.hexbin(x='SqFtTotLiving', y='TaxAssessedValue',
                         gridsize=30, sharex=False, figsize=(6, 6))
ax.set_xlabel('Finished Square Feet')
ax.set_ylabel('Tax Assessed Value')
plt.tight_layout()
plt.show()
```



Προβολή αλληλεπιδράσεων μεταβλητών (μεγάλο πλήθος δεδομένων)





Κατηγοριοποιημένα δεδομένα

Φόρτωση δεδομένων

[102…]
```
lc_loans = pd.read_csv(LC_LOANS_CSV)
lc_loans.head()
```

[102…]

	status	grade
0	Fully Paid	B
1	Charged Off	C
2	Fully Paid	C
3	Fully Paid	C
4	Current	B

[105…]
```
print(f'Μέγεθος αρχείου δανείων:{lc_loans.shape}')
```

```
Μέγεθος αρχείου δανείων:(450961, 2)
```

Οδήγηση - Ομαδοποίηση (pivoting)

Οδήγηση - Ομαδοποίηση (pivoting)

[104…]
```
crosstab = lc_loans.pivot_table(index='grade', columns='status',
                                aggfunc=lambda x: len(x), margins=True)
print(crosstab)
```

```
status  Charged Off  Current  Fully Paid  Late     All
grade
A              1562    50051       20408   469   72490
B              5302    93852       31160  2056  132370
C              6023    88928       23147  2777  120875
D              5007    53281       13681  2308   74277
E              2842    24639        5949  1374   34804
F              1526     8444        2328   606   12904
G               409     1990         643   199    3241
All           22671   321185       97316  9789  450961
```

Οδήγηση - Ομαδοποίηση (pivoting)
Ποσοστιαία μέρη

[140…]
```
crosstab_perc = crosstab[['Charged Off','Current', 'Fully Paid', 'Late']].div(crosstab['All'], axis=0)*100
crosstab_perc.columns = crosstab_perc.columns + ' %'
```

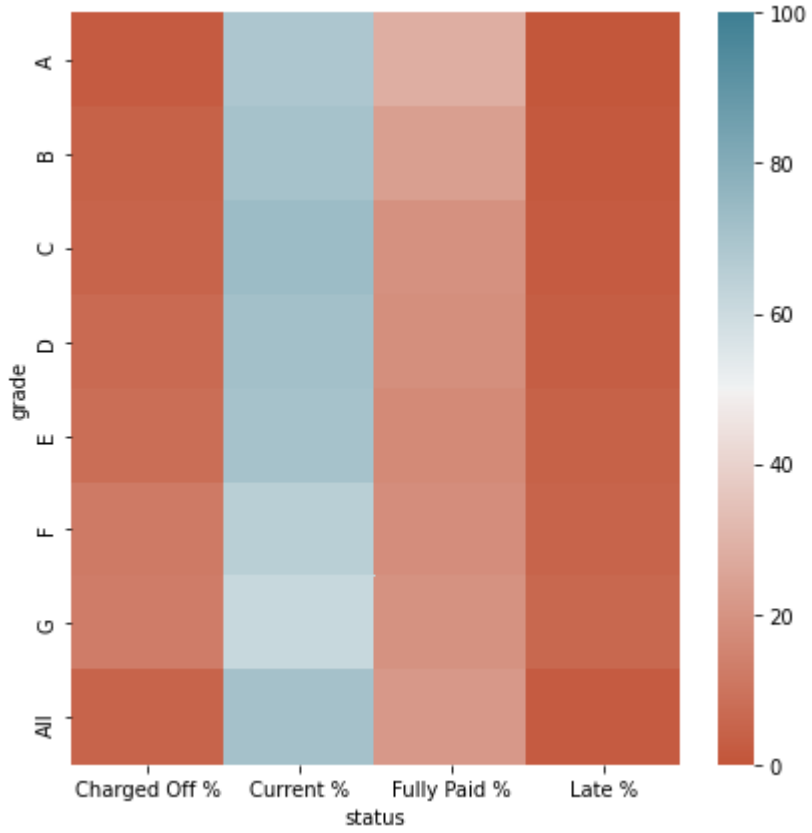
[141…]
```
crosstab_perc
```

[141…]

status	Charged Off %	Current %	Fully Paid %	Late %
grade				
A	2.154780	69.045386	28.152849	0.646986
B	4.005439	70.901262	23.540077	1.553222
C	4.982834	73.570217	19.149535	2.297415
D	6.740983	71.732838	18.418891	3.107288
E	8.165728	70.793587	17.092863	3.947822
F	11.825790	65.437074	18.040918	4.696218
G	12.619562	61.400802	19.839556	6.140080
All	5.027264	71.222345	21.579693	2.170698

Οδήγηση – Ομαδοποίηση
Γραφική αναπαράσταση

[142…
```
fig, ax = plt.subplots(figsize=(6, 6))
ax = sns.heatmap(crosstab_perc, vmin=0, vmax=100,
                 cmap=sns.diverging_palette(20, 220, as_cmap=True),
                 ax=ax)

plt.tight_layout()
plt.show()
```



Μίξη κατηγοριοποιημένων και αριθμητικών δεδομένων

Φόρτωση δεδομένων

[146...
```
airline_stats = pd.read_csv(AIRLINE_STATS_CSV)
airline_stats.head()
```

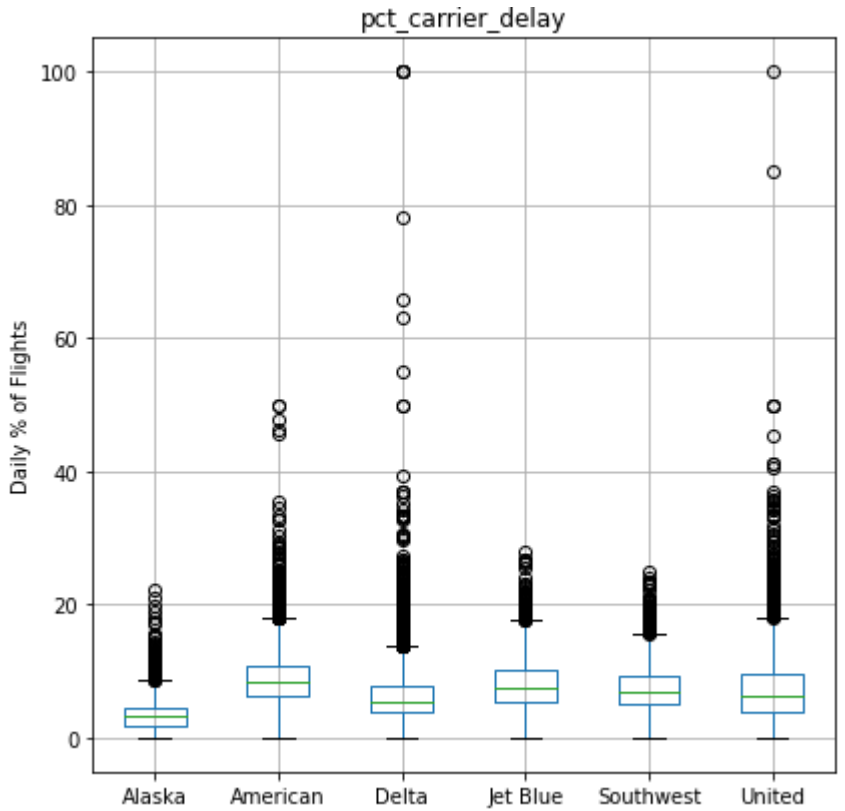
[146...

	pct_carrier_delay	pct_atc_delay	pct_weather_delay	airline
0	8.153226	1.971774	0.762097	American
1	5.959924	3.706107	1.585878	American
2	7.157270	2.706231	2.026706	American
3	12.100000	11.033333	0.000000	American
4	7.333333	3.365591	1.774194	American

Μίξη κατηγοριοποιημένων και αριθμητικών δεδομένων

Boxplots

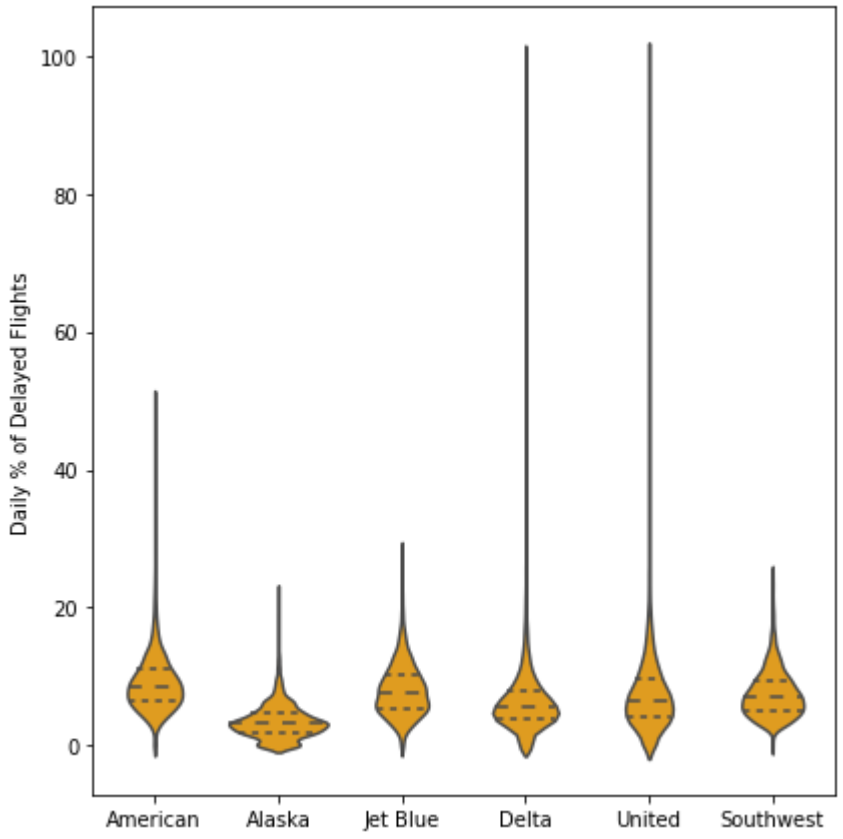
```
[149…  ax = airline_stats.boxplot(by='airline', column='pct_carrier_delay',
                                  figsize=(6, 6))
       ax.set_xlabel('')
       ax.set_ylabel('Daily % of Flights')
       plt.suptitle('')

       plt.tight_layout()
       plt.show()
```



Μίξη κατηγοριοποιημένων και αριθμητικών δεδομένων

Violinplot

```
[155…  fig, ax = plt.subplots(figsize=(6, 6))
sns.violinplot(x=airline_stats.airline, y=airline_stats.pct_carrier_delay,
               ax=ax, inner='quartile', color='orange')
ax.set_xlabel('')
ax.set_ylabel('Daily % of Delayed Flights')
plt.tight_layout()
plt.show()
```



Αλληλεπίδραση για περισσότερες από δύο μεταβλητές

Επιλογή υποσυνόλου δεδομένων

[157…]
```
zip_codes = [98188, 98105, 98108, 98126]
kc_tax_zip = kc_tax0.loc[kc_tax0.ZipCode.isin(zip_codes),:]
kc_tax_zip.head()
```

[157…]

	TaxAssessedValue	SqFtTotLiving	ZipCode
3	361000.0	2000	98108.0
4	459000.0	3150	98108.0
10	202000.0	830	98108.0
11	210000.0	1130	98108.0
12	193000.0	1560	98108.0

Αλληλεπίδραση για περισσότερες από δύο μεταβλητές

Γραφική παράσταση πολλών μεταβλητών

[166…

```
def hexbin(x, y, color, **kwargs):
    cmap = sns.light_palette(color, as_cmap=True)
    plt.hexbin(x, y, gridsize=25, cmap=cmap, **kwargs)

g = sns.FacetGrid(kc_tax_zip, col='ZipCode', col_wrap=2)
g.map(hexbin, 'SqFtTotLiving', 'TaxAssessedValue',
      extent=[0, 3500, 0, 700000])
g.set_axis_labels('Finished Square Feet', 'Tax Assessed Value')
g.set_titles('Zip code {col_name:.0f}')

plt.tight_layout()
plt.show()
```