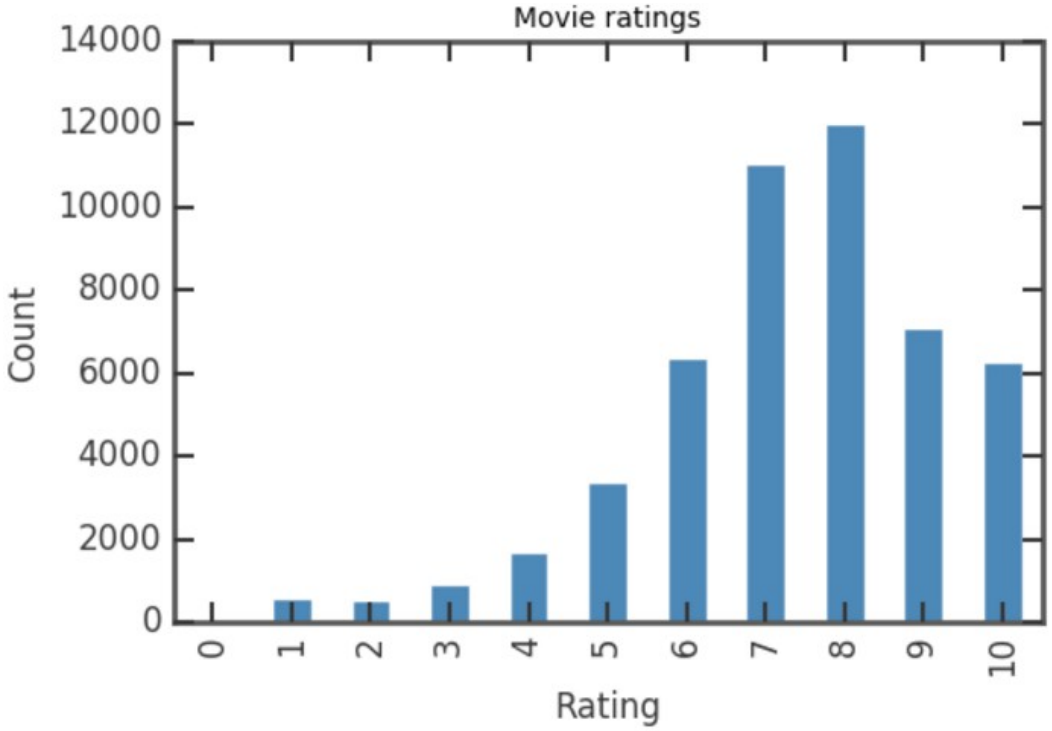
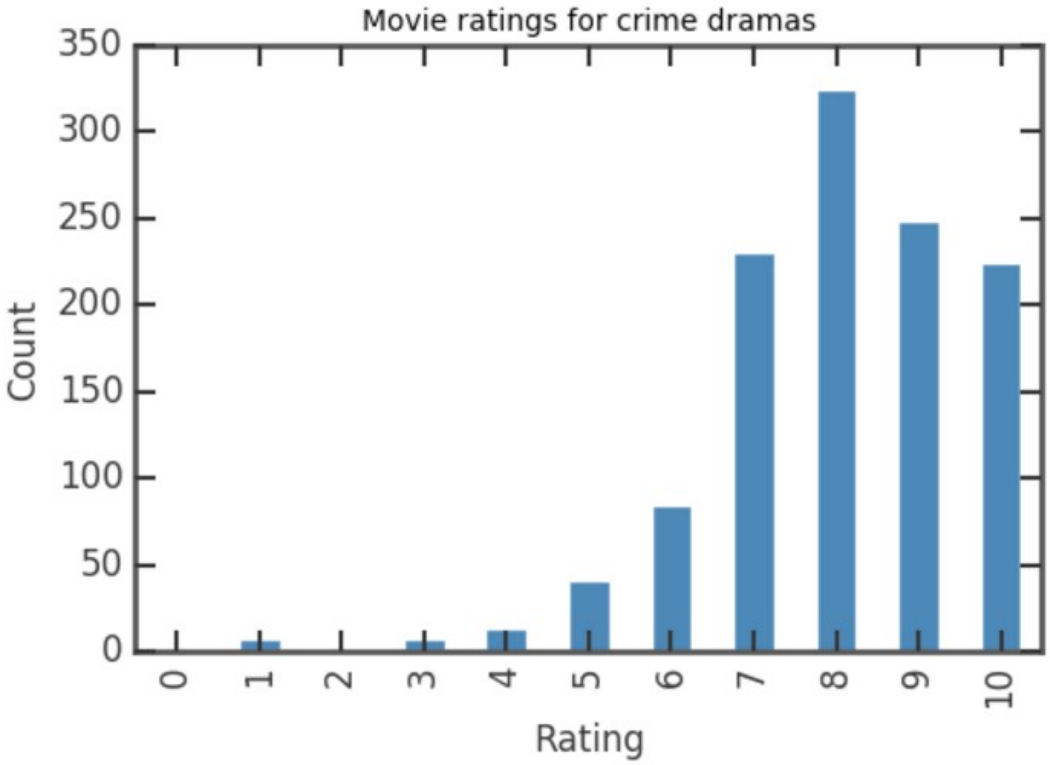
## Chapter 1: Tools of the Trade

| | user id | item id | rating | timestamp |
|---|---|---|---|---|
| **0** | 1 | 1074638 | 7 | 1365029107 |
| **1** | 1 | 1853728 | 8 | 1366576639 |
| **2** | 2 | 104257 | 8 | 1364690142 |
| **3** | 2 | 1259521 | 8 | 1364118447 |
| **4** | 2 | 1991245 | 7 | 1364117717 |

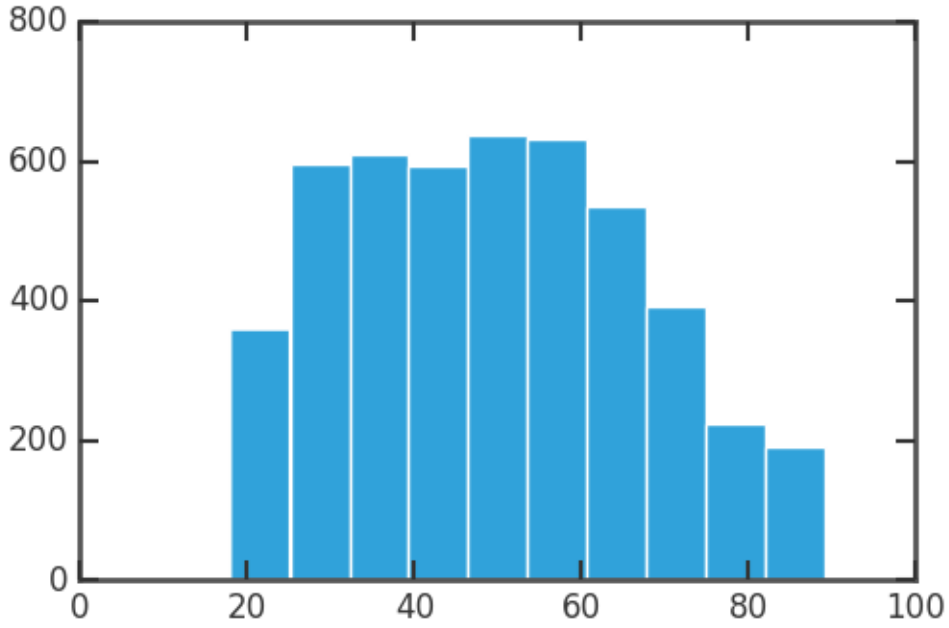
```
8     12012
7     11063
9      7119
6      6373
10     6281
5      3399
4      1696
3       924
1       595
2       533
0         5
Name: rating, dtype: int64
```

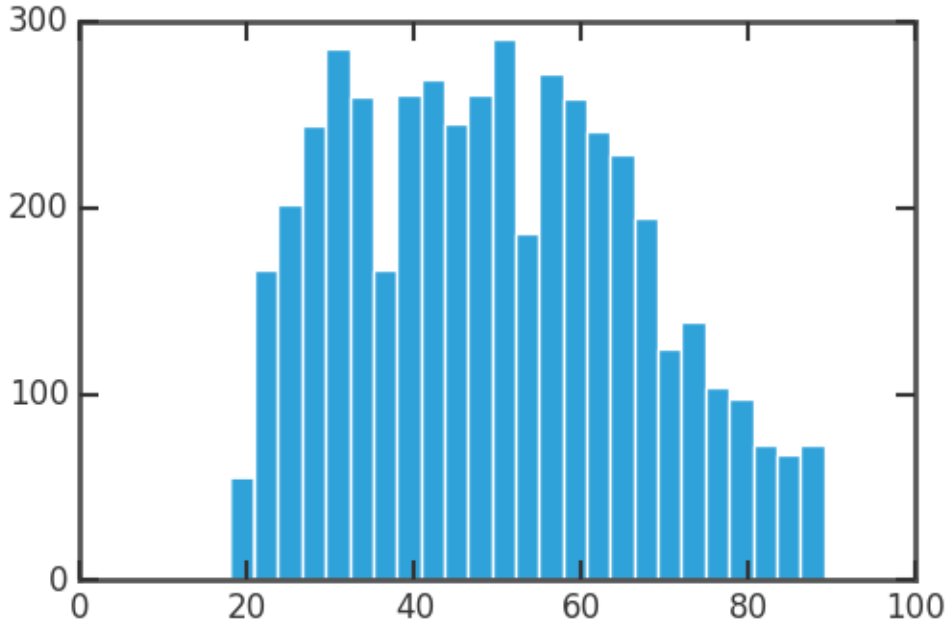


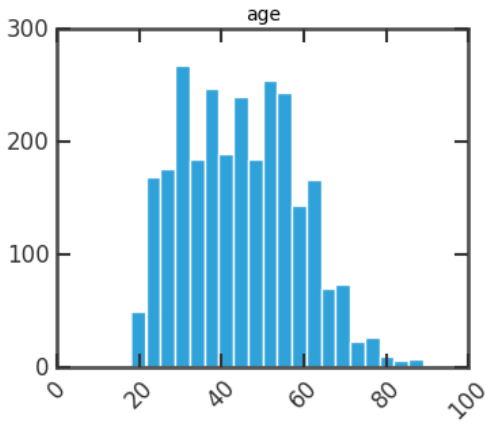

```
0     True
1    False
2    False
3    False
4    False
Name: genre, dtype: bool
```

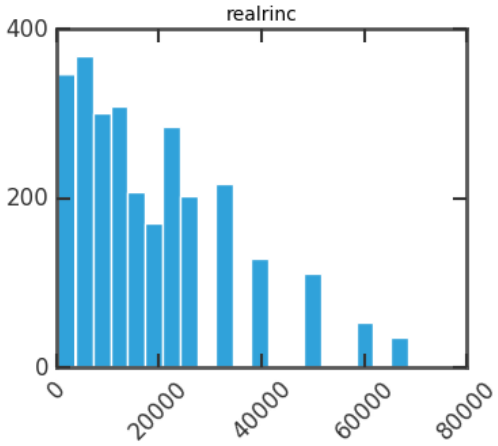

Movie ratings for crime dramas
Count
Rating
0
50
100
150
200
250
300
350
0
1
2
3
4
5
6
7
8
9
10

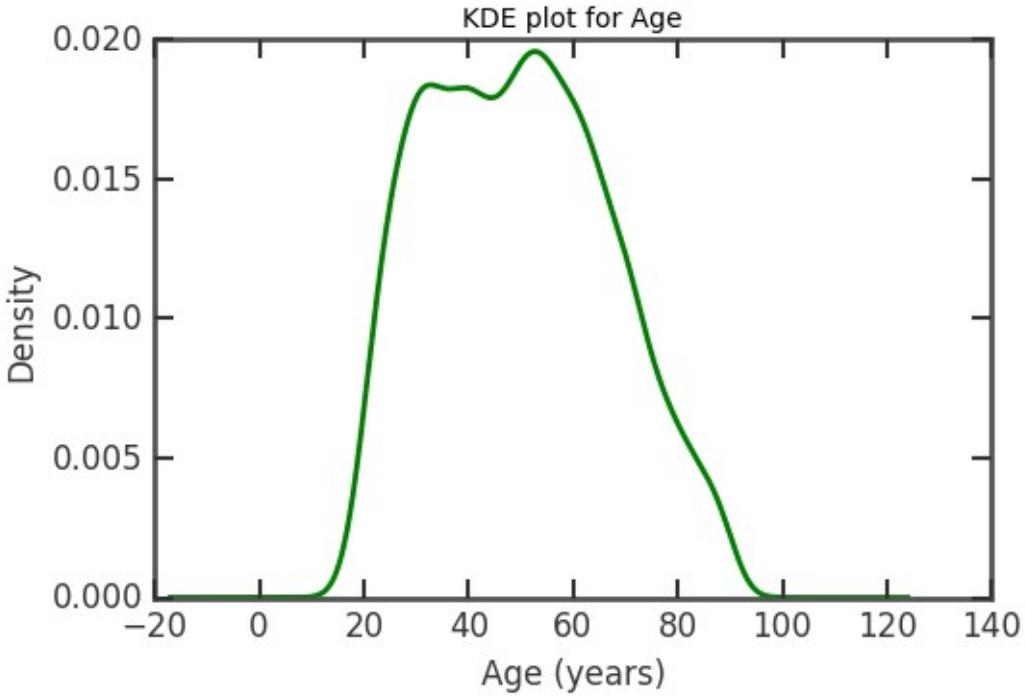
## Chapter 2: Exploring Data

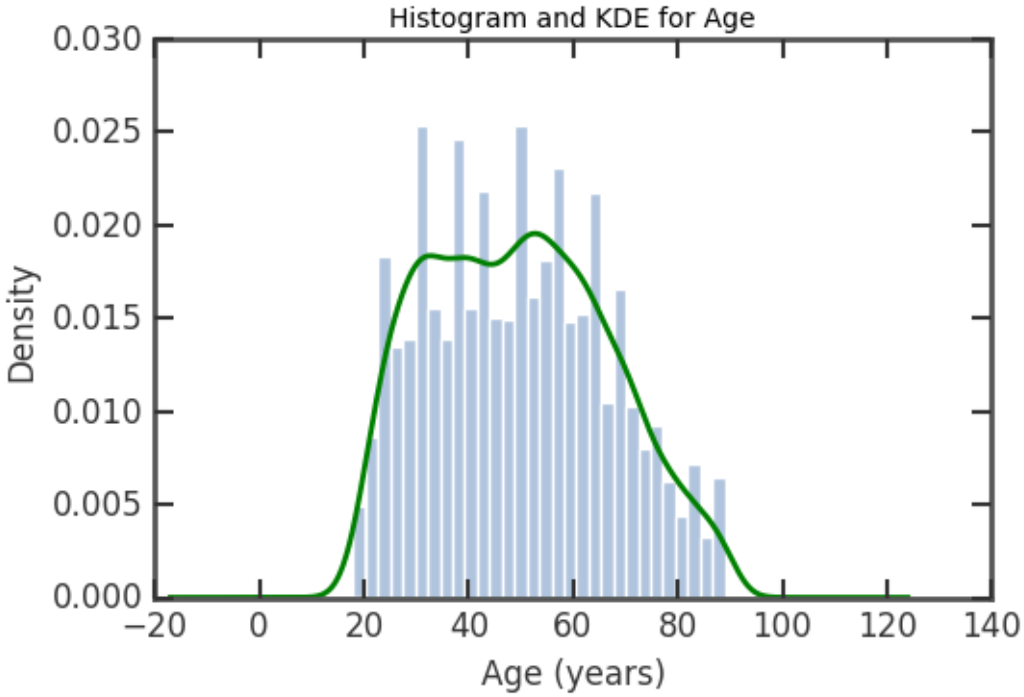


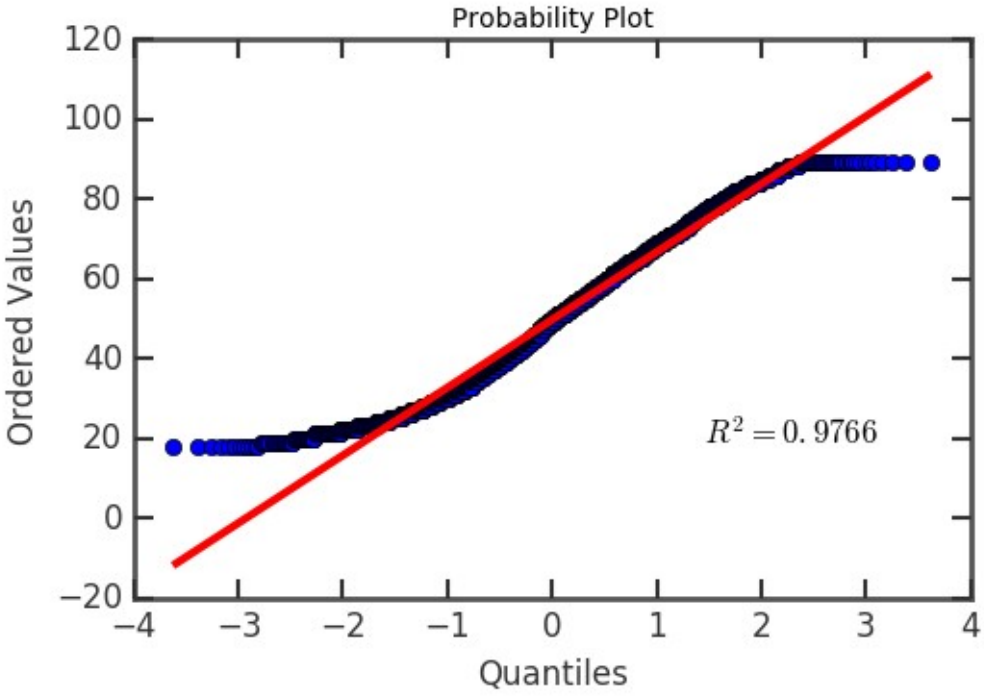

300
200
100
0
0
20
40
60
80
100

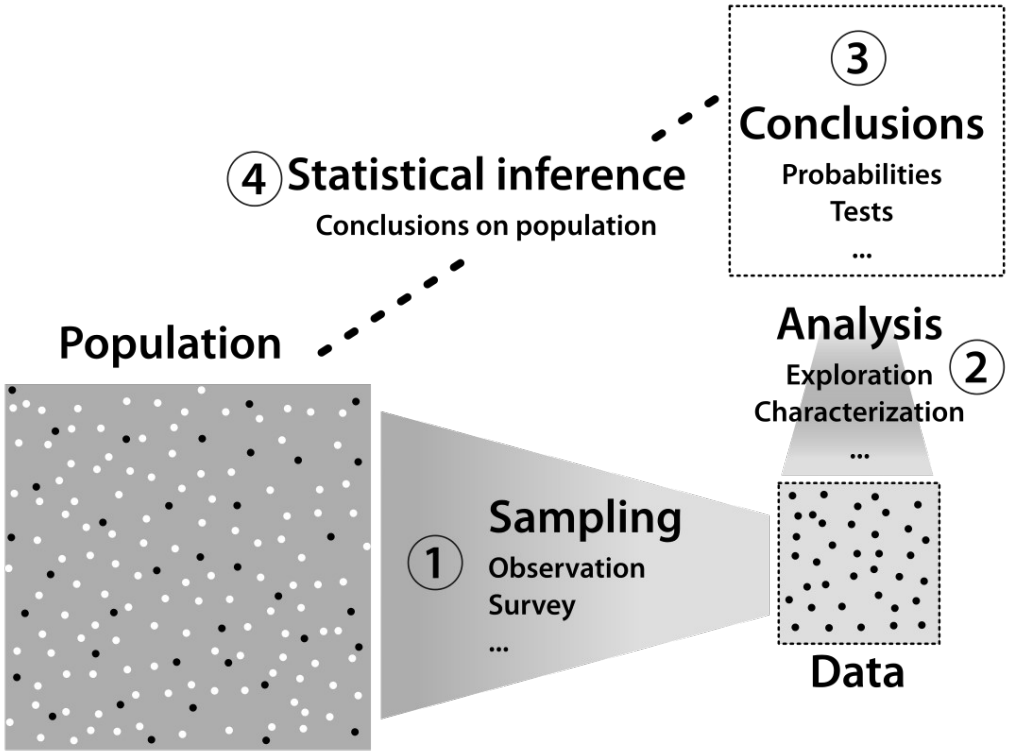

age
300
200
100
0
0
20
40
60
80
100

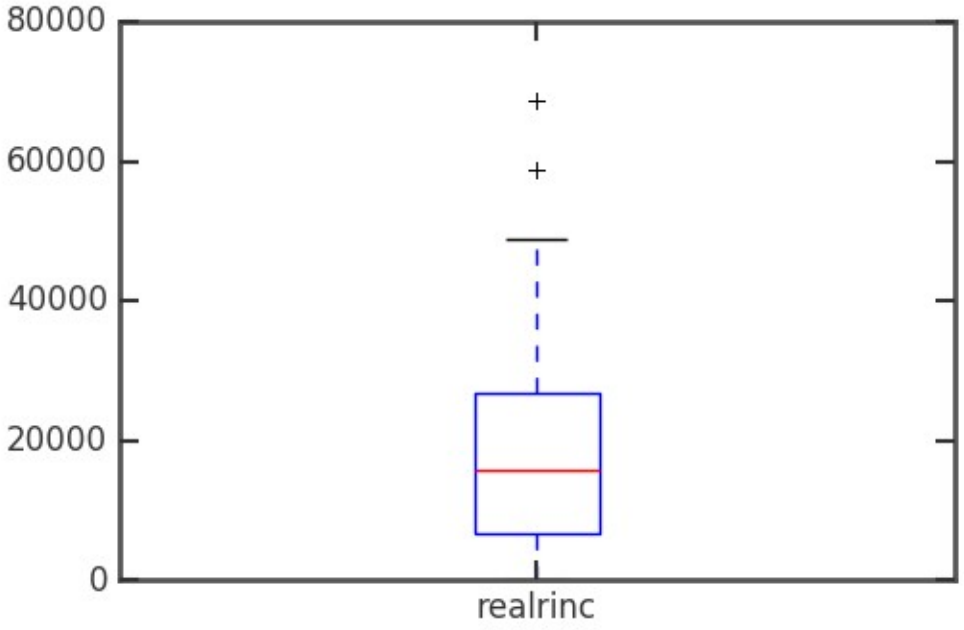

realrinc
400
200
0
0
20000
40000
60000
80000

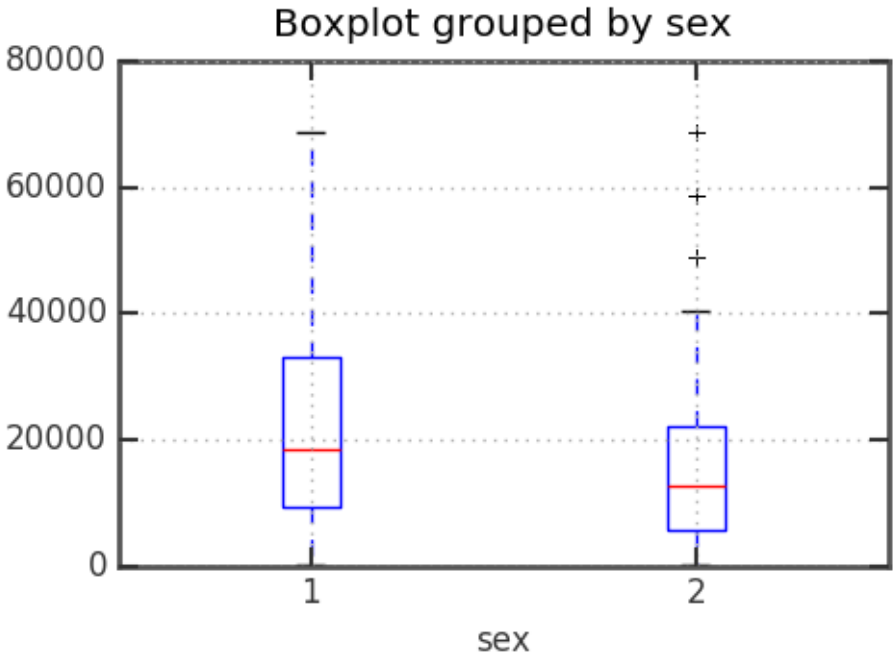

KDE plot for Age
0.020
0.015
0.010
0.005
0.000
Density
−20
0
20
40
60
80
100
120
140
Age (years)

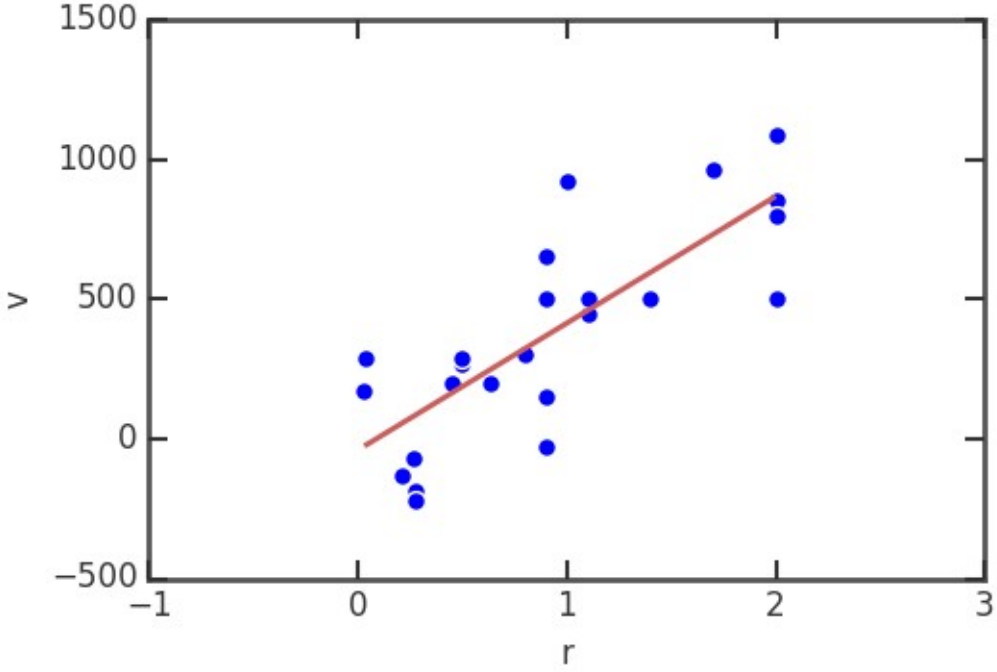

Histogram and KDE for Age
Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
−20
0
20
40
60
80
100
120
140
Age (years)


Probability Plot
120
100
80
60
40
20
0
−20
−4
−3
−2
−1
0
1
2
3
4
Ordered Values
Quantiles
$R^2 = 0.9766$

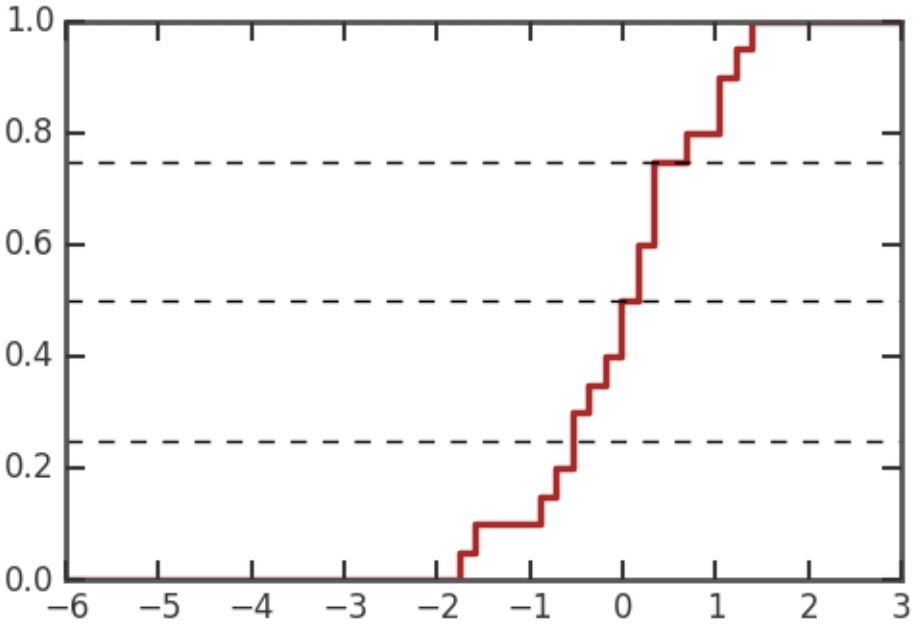

3
Conclusions
Probabilities
Tests
...
4 Statistical inference
Conclusions on population
Population
Analysis
Exploration
Characterization
...
2
1
Sampling
Observation
Survey
...
Data

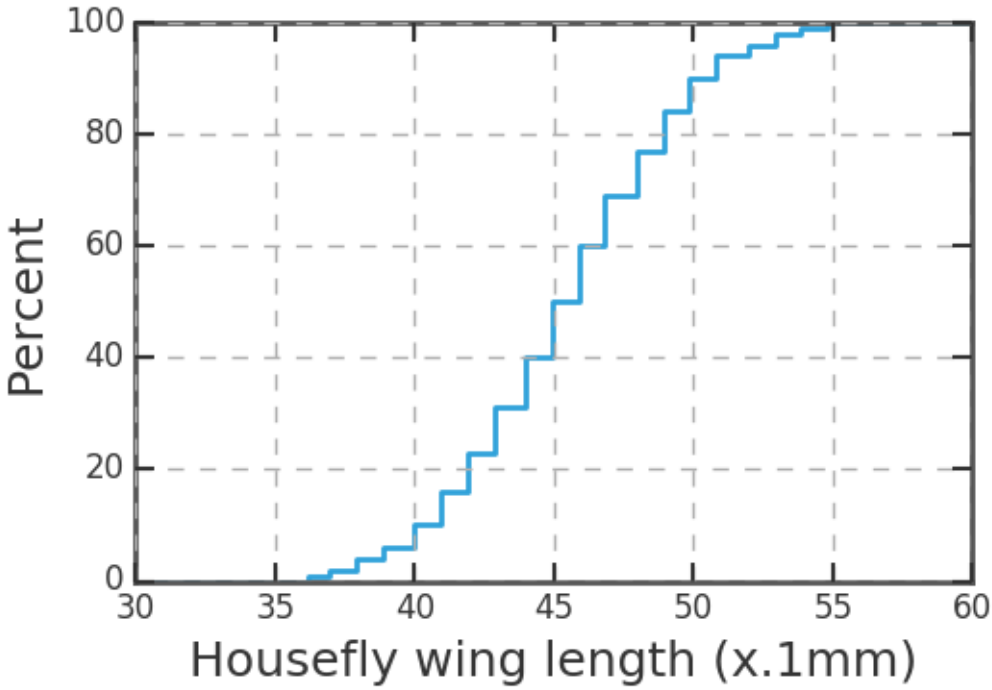

80000
60000
40000
20000
0
realrinc

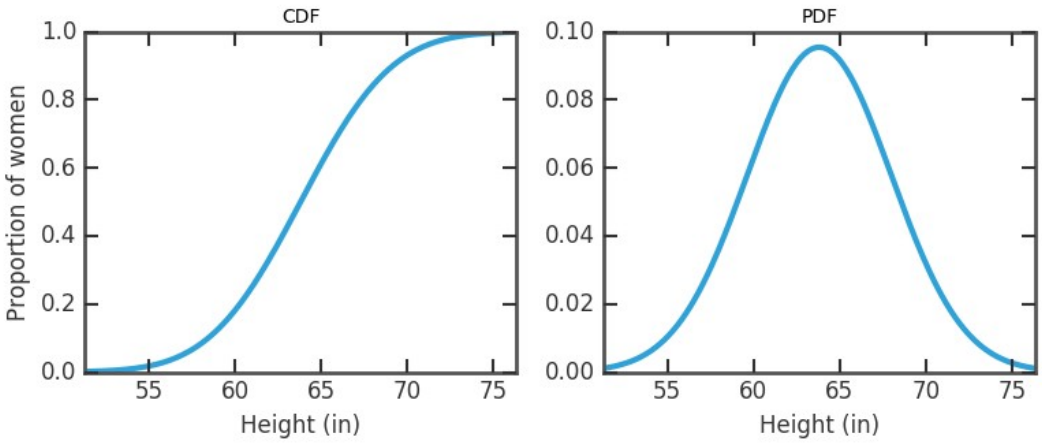

Boxplot grouped by sex
80000
60000
40000
20000
0
1
2
sex

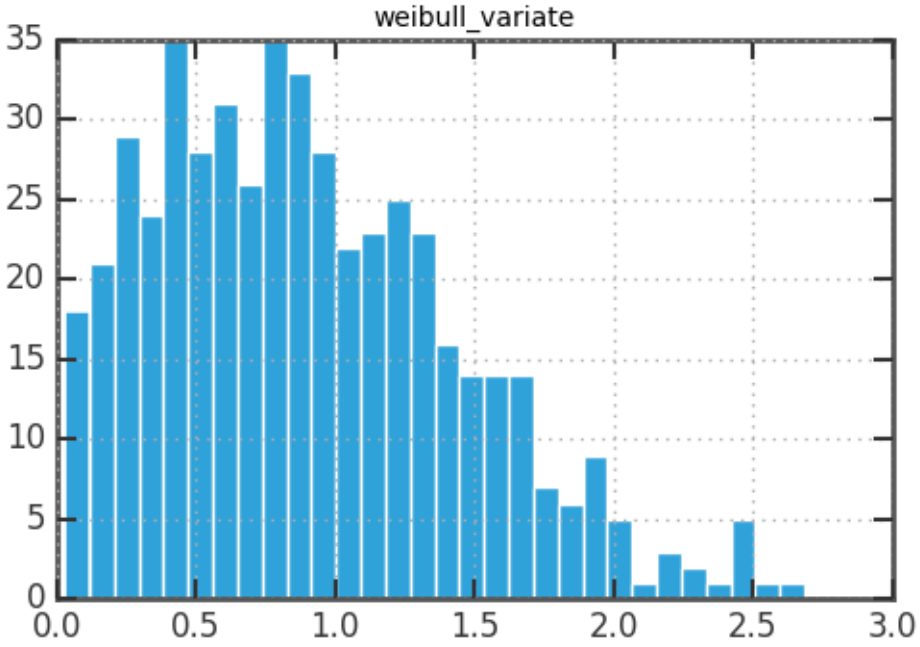

1500
1000
500
0
−500
v
−1
0
1
2
3
r

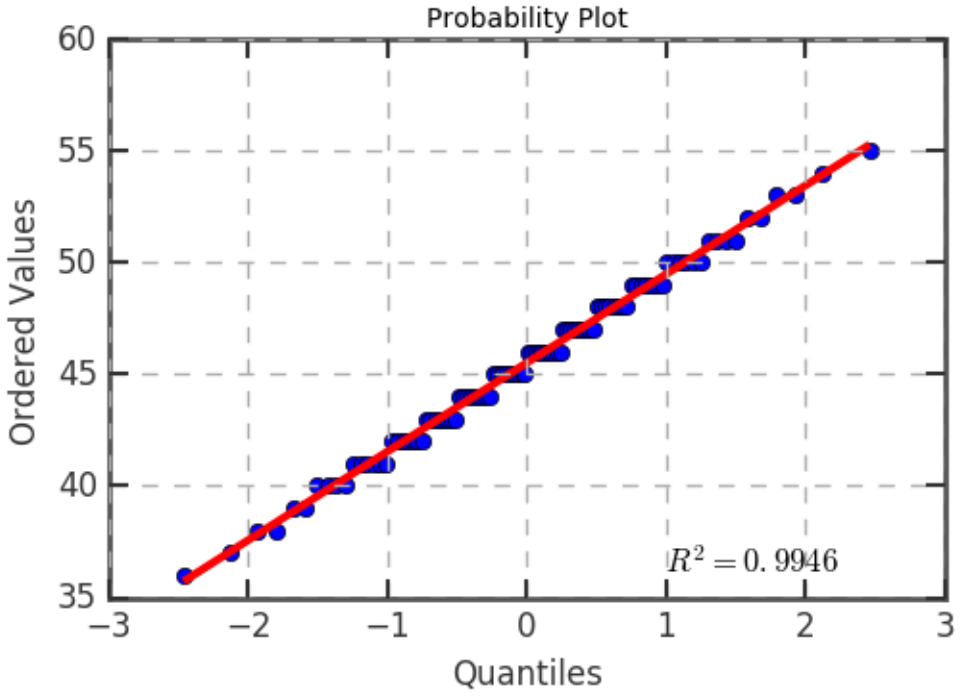
## Chapter 3: Learning About Models

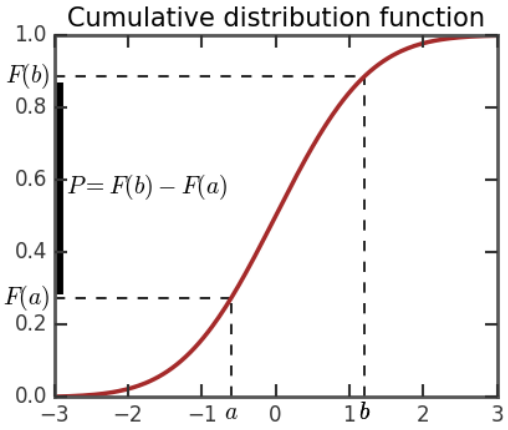

1.0
0.8
0.6
0.4
0.2
0.0
−6
−5
−4
−3
−2
−1
0
1
2
3

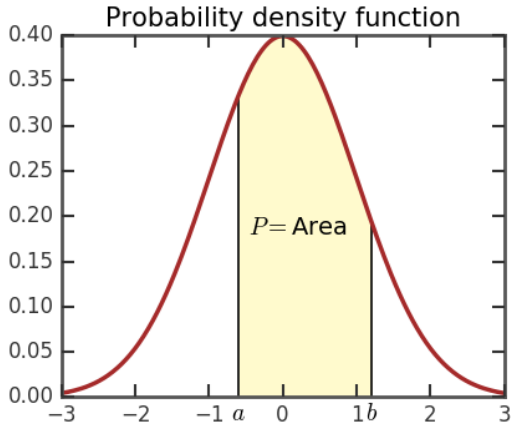

100
80
60
40
20
0
Percent
30
35
40
45
50
55
60
Housefly wing length (x.1mm)

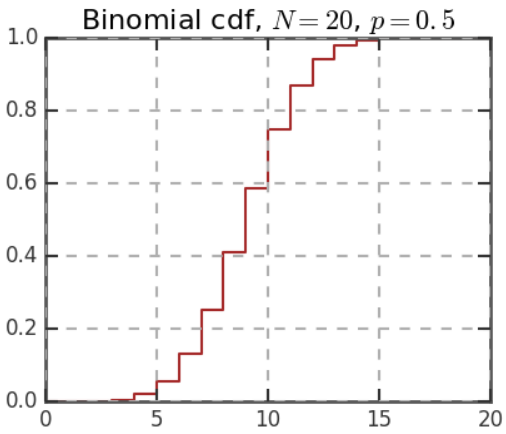

CDF
PDF
Proportion of women
1.0
0.8
0.6
0.4
0.2
0.0
0.10
0.08
0.06
0.04
0.02
0.00
55
60
65
70
75
Height (in)
Height (in)

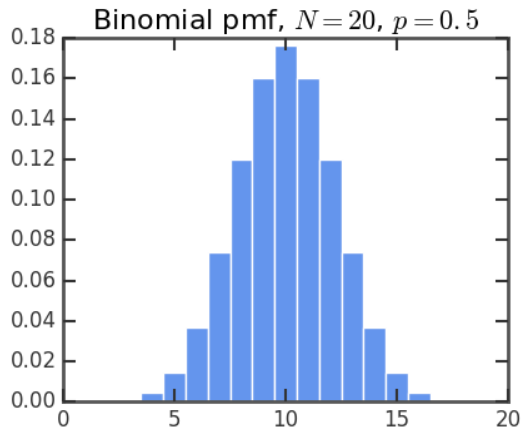

weibull_variate
35
30
25
20
15
10
5
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

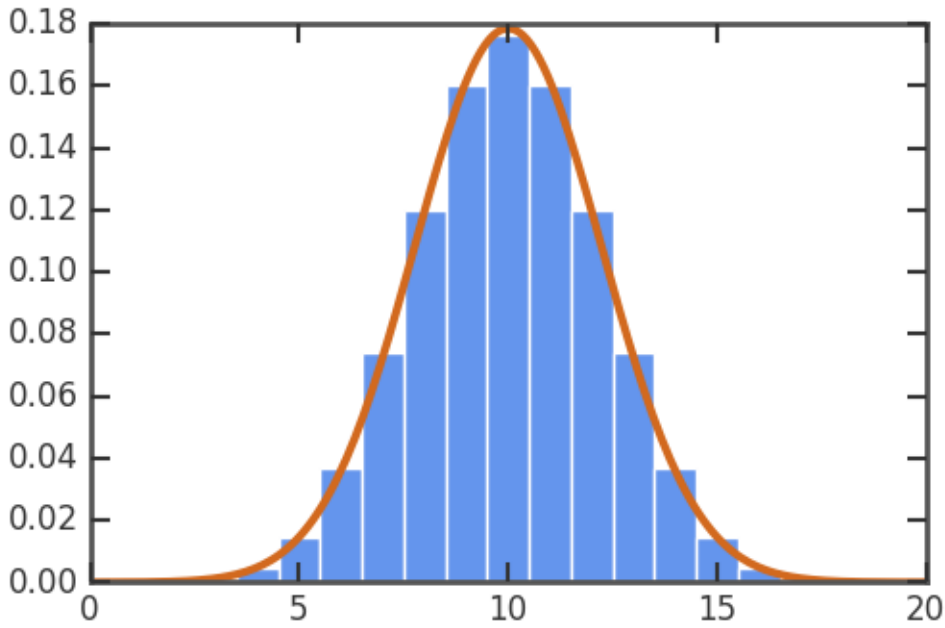

Probability Plot
Ordered Values
Quantiles
$R^2 = 0.9946$

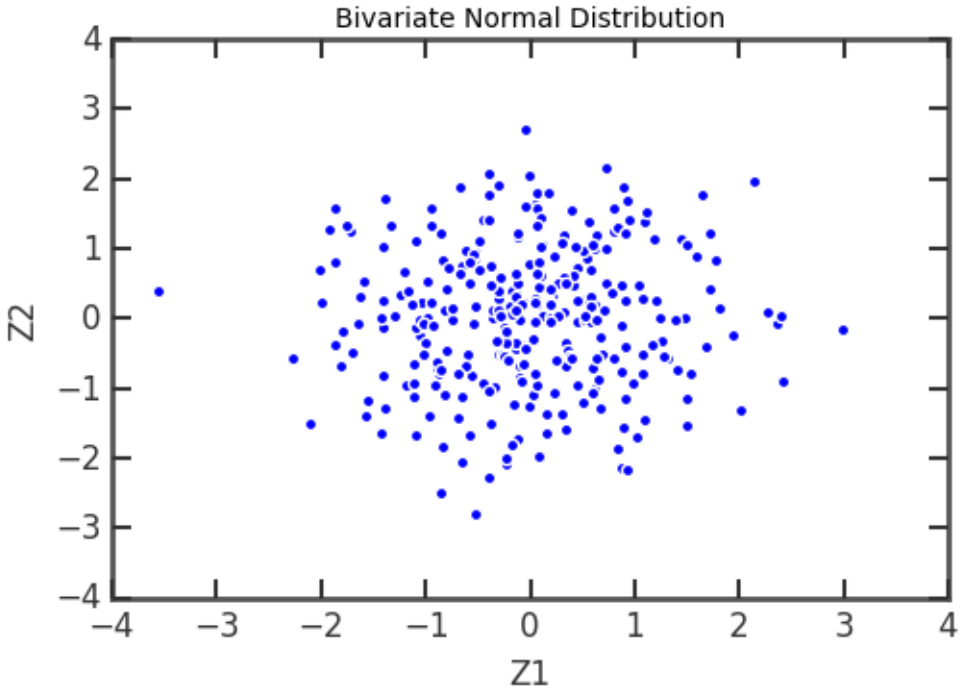

Cumulative distribution function
$F(b)$
$F(a)$
$P = F(b) - F(a)$
$a$
$b$


Probability density function
$P =$ Area
$a$
$b$


Binomial cdf, $N = 20$, $p = 0.5$
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20

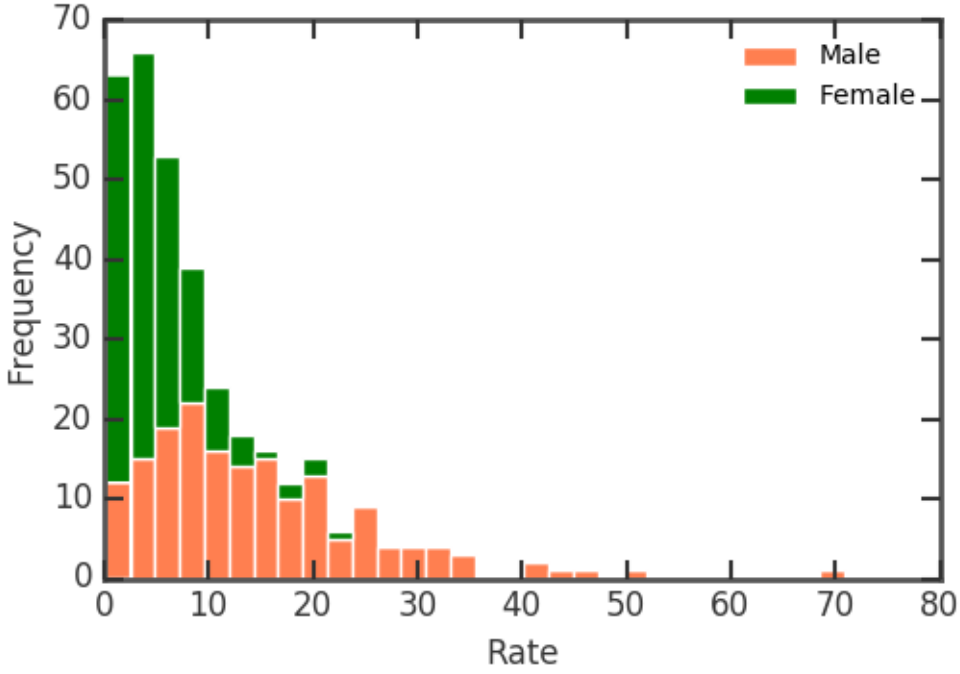

Binomial pmf, $N = 20$, $p = 0.5$
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20

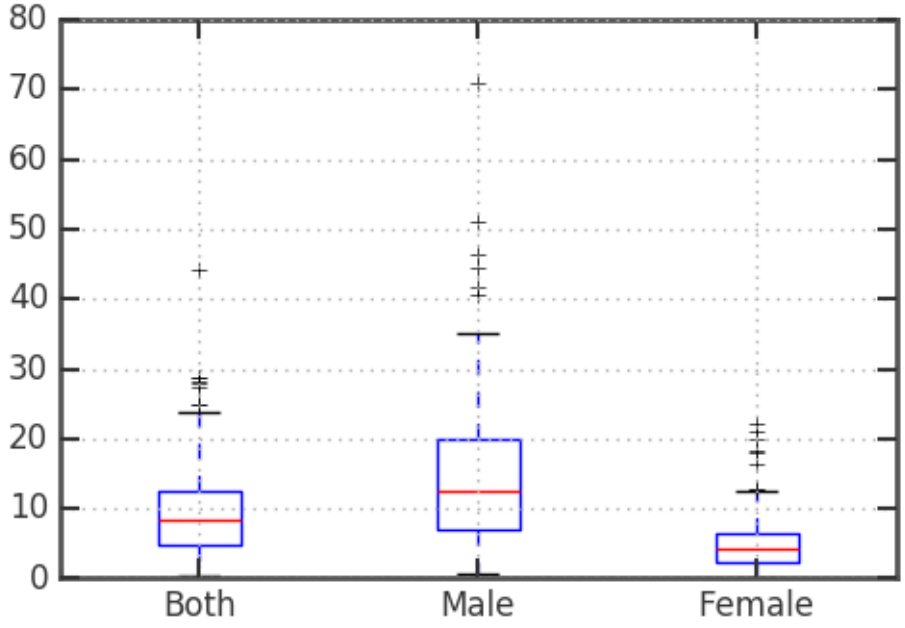

0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20

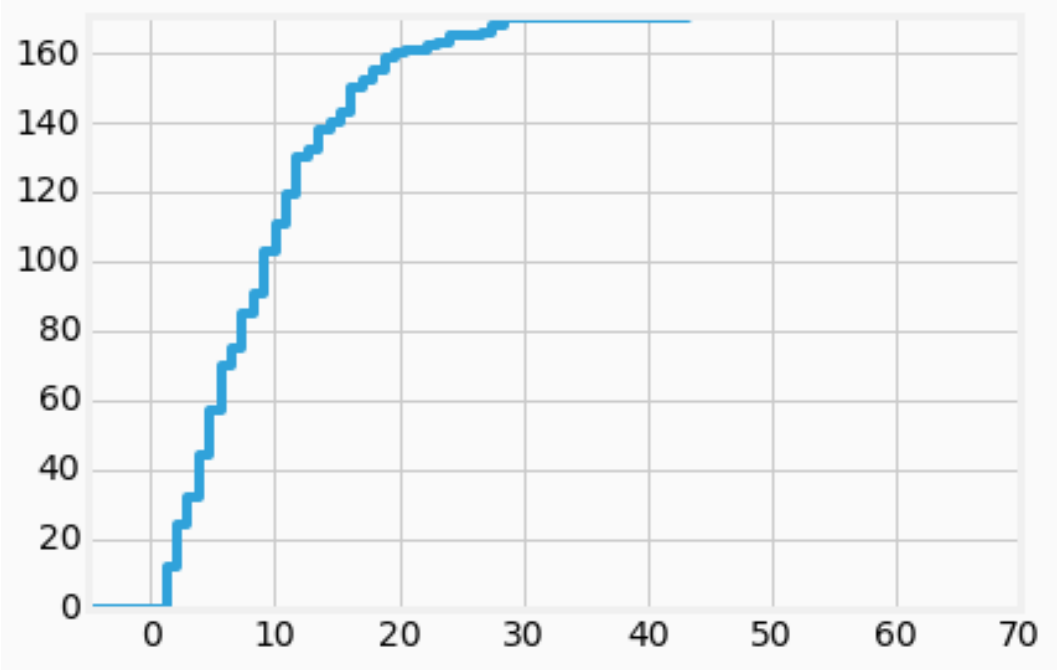

Bivariate Normal Distribution
Z2
Z1
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4

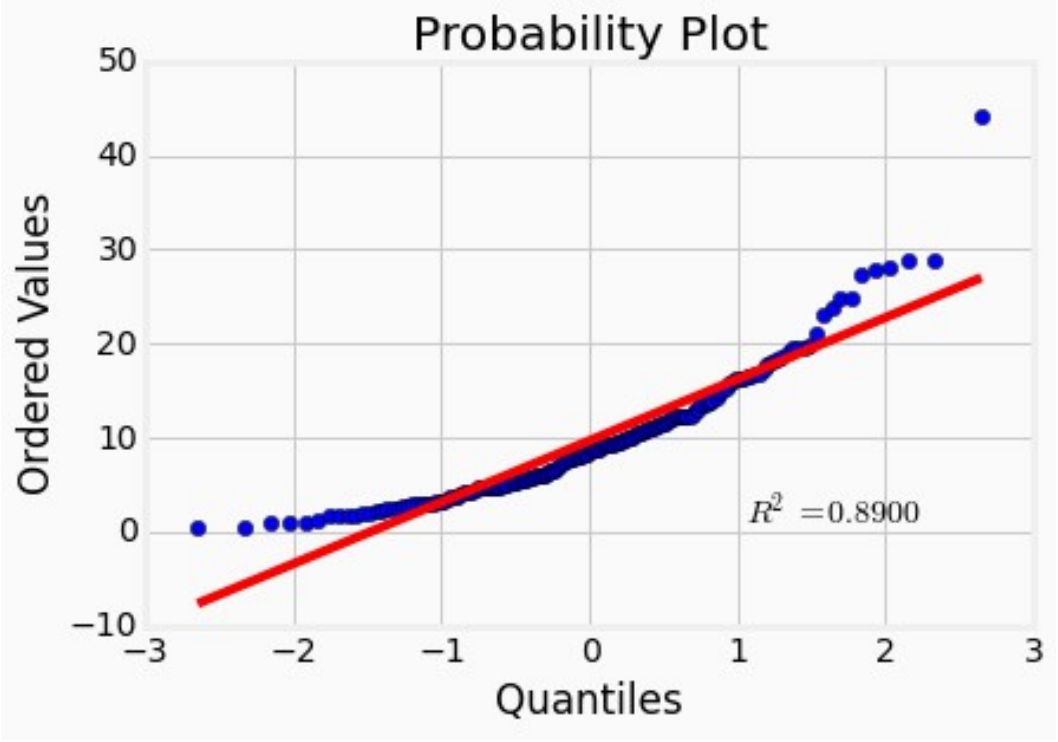
# Chapter 4: Regression

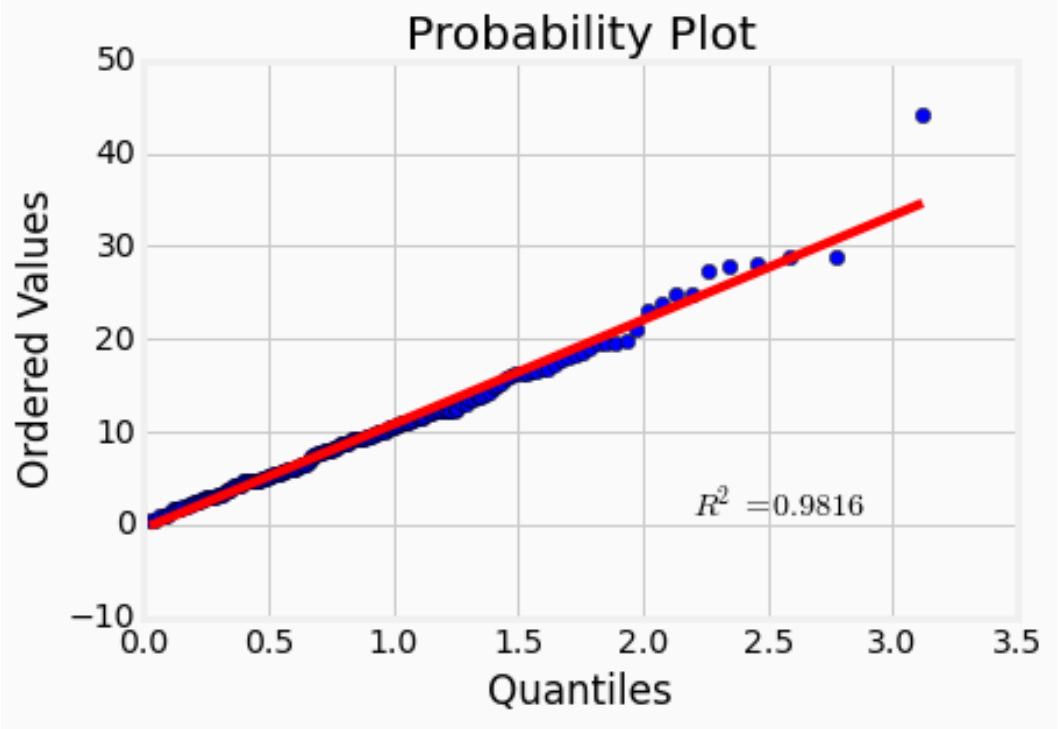
| | Country | Both | Female | Male |
|---|---|---|---|---|
| **0** | Afghanistan | 5.7 | 5.3 | 6.2 |
| **1** | Albania | 5.9 | 5.2 | 6.6 |
| **2** | Algeria | 1.9 | 1.5 | 2.3 |
| **3** | Angola | 13.8 | 7.3 | 20.7 |
| **4** | Argentina | 10.3 | 4.1 | 17.2 |
| **5** | Armenia | 2.9 | 0.9 | 5.0 |
| **6** | Australia | 10.6 | 5.2 | 16.1 |
| **7** | Austria | 11.5 | 5.4 | 18.2 |
| **8** | Azerbaijan | 1.7 | 1.0 | 2.4 |
| **9** | Bahamas | 2.3 | 1.3 | 3.6 |

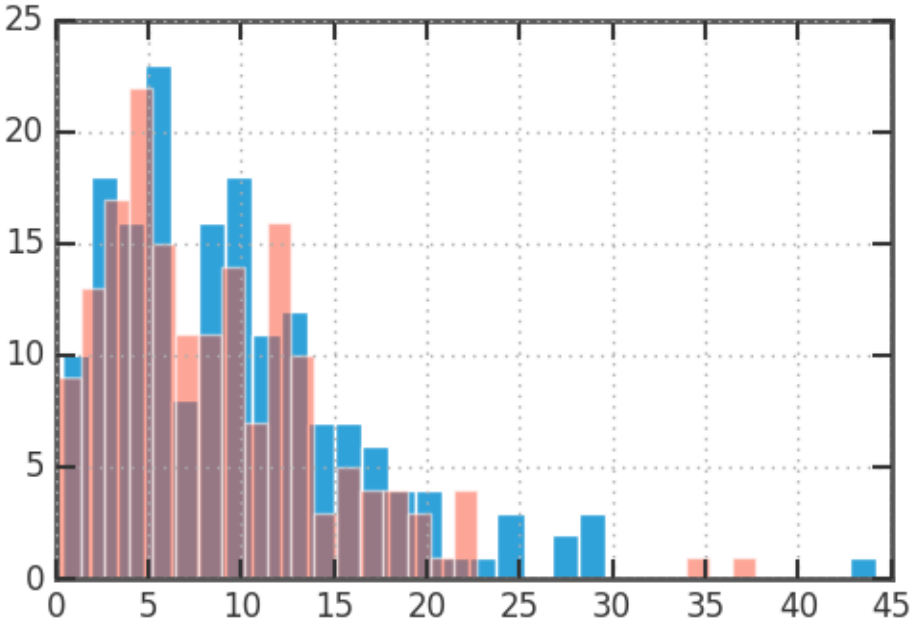

Male
Female
Frequency
Rate
0
10
20
30
40
50
60
70
80

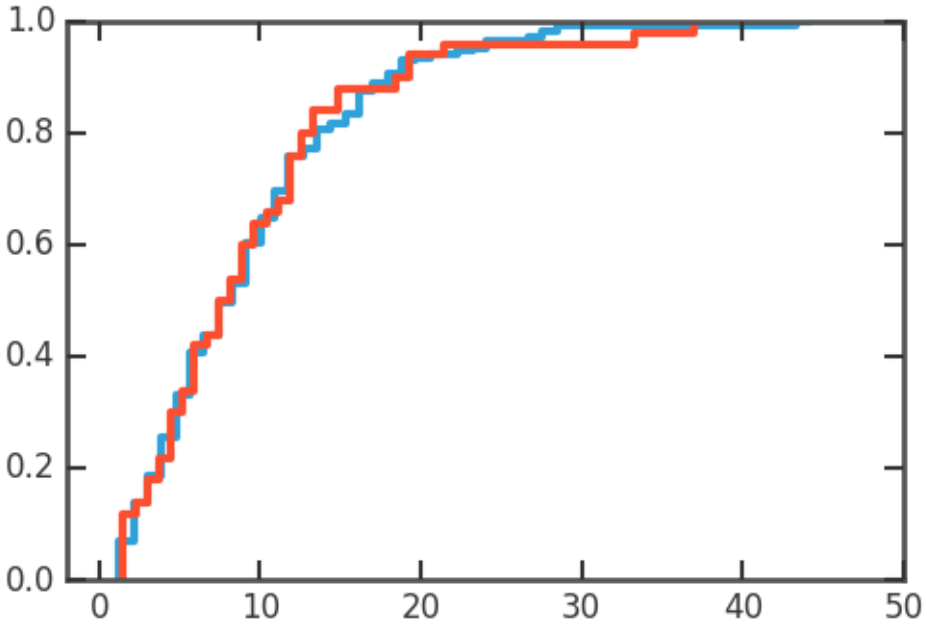

Both
Male
Female
0
10
20
30
40
50
60
70
80


160
140
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70


Probability Plot
Ordered Values
Quantiles
$R^2$ =0.8900
50
40
30
20
10
0
−10
−3
−2
−1
0
1
2
3


Probability Plot
Ordered Values
Quantiles
$R^2$ =0.9816
50
40
30
20
10
0
−10
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

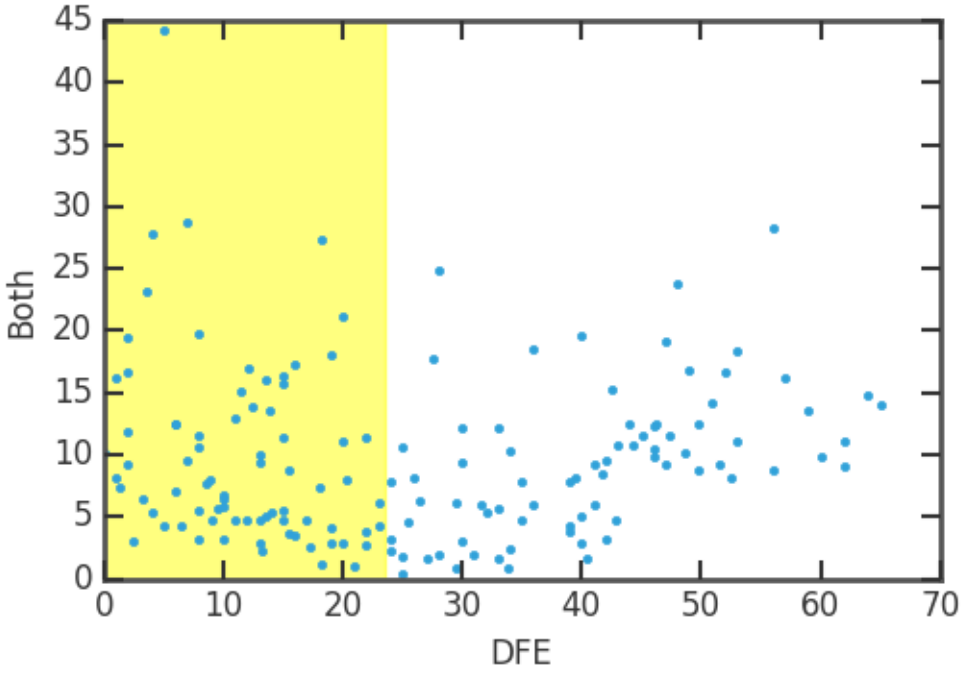

25
20
15
10
5
0
0
5
10
15
20
25
30
35
40
45

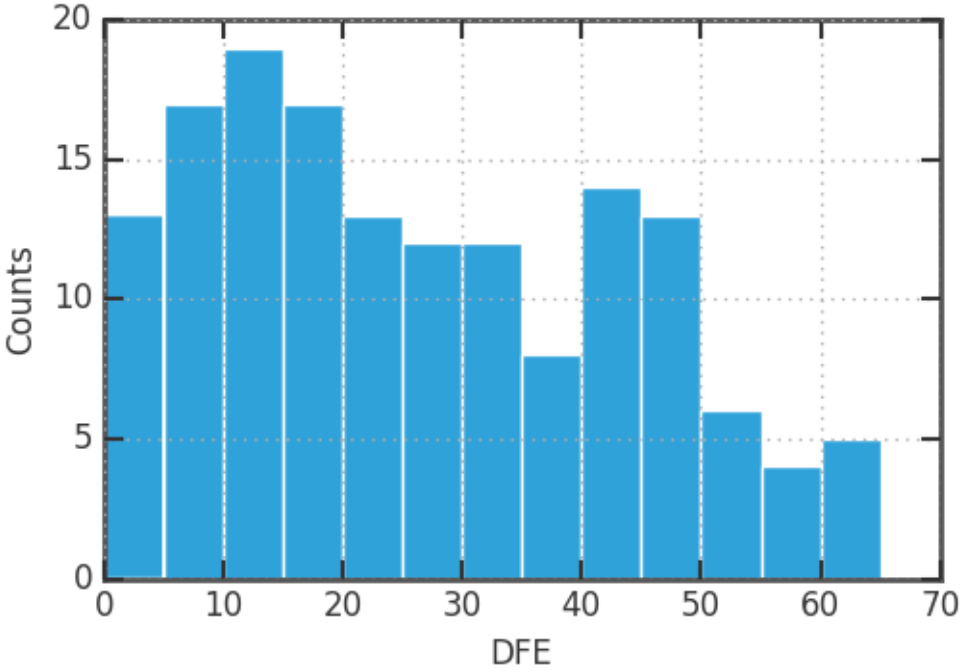


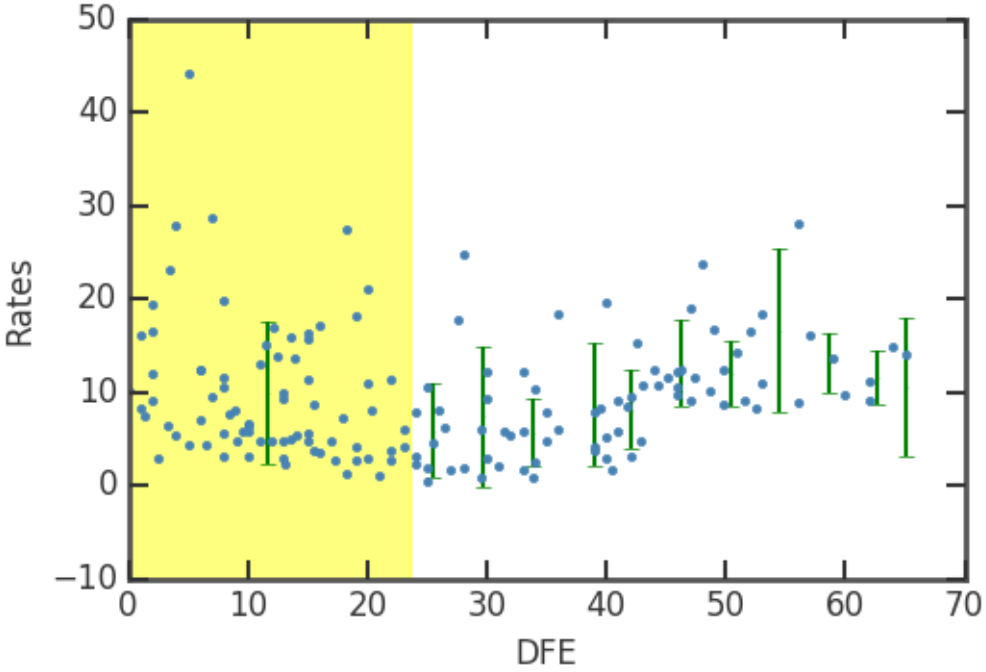

```
Index(['LAT', 'LONG', 'DMS_LAT', 'DMS_LONG', 'MGRS', 'JOG', 'DSG', 'AFFIL',
       'FIPS10', 'SHORT_NAME', 'FULL_NAME', 'MOD_DATE', 'ISO3136'],
      dtype='object')
```

| | LAT | LONG | DMS_LAT | DMS_LONG | MGRS | JOG | DSG | AFFIL | FIPS10 | SHORT_NAME | FULL_NAME | MOD_DATE | ISO3136 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 33.000000 | 66.0 | 330000 | 660000 | 42STB1970055286 | NI42-09 | PCLI | NaN | AF | Afghanistan | Islamic Republic of Afghanistan | 2009-04-10 | AF |
| **1** | 41.000000 | 20.0 | 410000 | 200000 | 34TDL1589839239 | NK34-08 | PCLI | NaN | AL | Albania | Republic of Albania | 2007-02-28 | AL |
| **2** | 28.000000 | 3.0 | 280000 | 30000 | 31REL0000097202 | NH31-15 | PCLI | NaN | AG | Algeria | People's Democratic Republic of Algeria | 2011-03-03 | DZ |
| **3** | -14.333333 | -170.0 | -142000 | -1700000 | 1802701 | NaN | PCLD | US | AS | American Samoa | Territory of American Samoa | 1998-10-06 | AS |
| **4** | 42.500000 | 1.5 | 423000 | 13000 | 31TCH7675006383 | NK31-04 | PCLI | NaN | AN | Andorra | Principality of Andorra | 2007-02-28 | AD |

| | Country | Both | Female | Male | Lat | Lon |
|---|---|---|---|---|---|---|
| **0** | Afghanistan | 5.7 | 5.3 | 6.2 | 33 | 66 |
| **1** | Albania | 5.9 | 5.2 | 6.6 | 41 | 20 |
| **2** | Algeria | 1.9 | 1.5 | 2.3 | 28 | 3 |
| **3** | Angola | 13.8 | 7.3 | 20.7 | -12.5 | 18.5 |
| **4** | Argentina | 10.3 | 4.1 | 17.2 | -34 | -64 |




20
15
10
5
0
Counts
0
10
20
30
40
50
60
70
DFE




```
slope:0.3204
intercept:-4.2373
rvalue:0.5102
pvalue:0.0000
stderr:0.0715
```

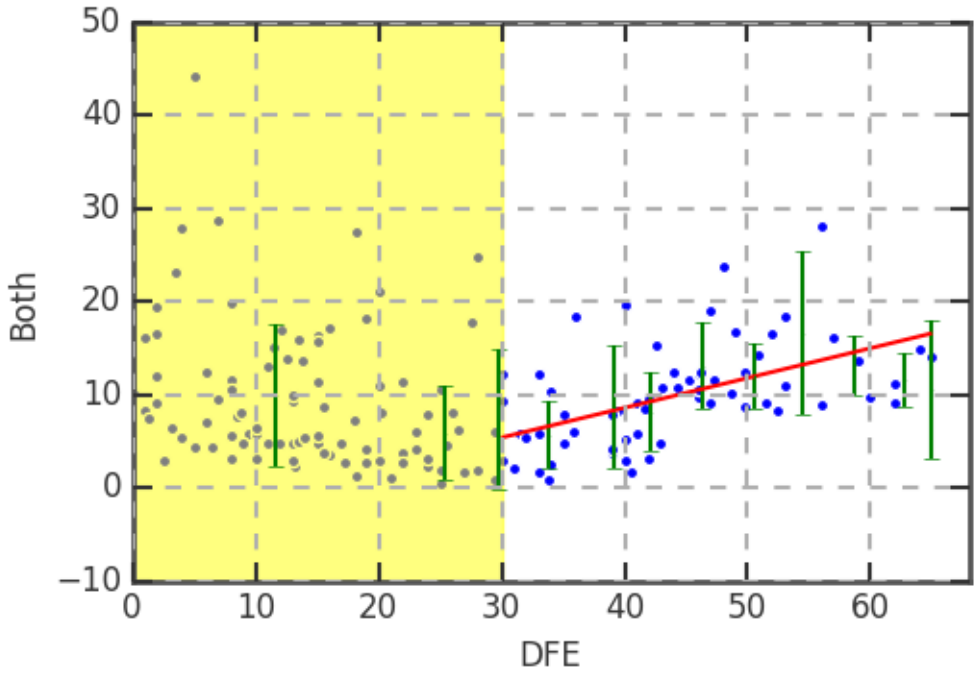
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Both | R-squared: | 0.260 |
| Model: | OLS | Adj. R-squared: | 0.247 |
| Method: | Least Squares | F-statistic: | 20.06 |
| Date: | Mon, 21 Dec 2015 | Prob (F-statistic): | 3.65e-05 |
| Time: | 03:40:19 | Log-Likelihood: | -175.72 |
| No. Observations: | 59 | AIC: | 355.4 |
| Df Residuals: | 57 | BIC: | 359.6 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [95.0% Conf. | Int.] |
|---|---|---|---|---|---|---|
| Intercept | -4.2373 | 3.272 | -1.295 | 0.201 | -10.789 | 2.315 |
| DFE | 0.3204 | 0.072 | 4.479 | 0.000 | 0.177 | 0.464 |

| | | | |
|---|---|---|---|
| Omnibus: | 13.615 | Durbin-Watson: | 2.424 |
| Prob(Omnibus): | 0.001 | Jarque-Bera (JB): | 14.566 |
| Skew: | 1.099 | Prob(JB): | 0.000687 |
| Kurtosis: | 4.047 | Cond. No. | 238. |

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.



| | id | name |
|---|---|---|
| **680** | 6.0.GDPpc_constant | GDP per capita, PPP (constant 2011 internation... |
| **4611** | GDPPCKD | GDP per Capita, constant US$, millions |
| **4612** | GDPPCKN | Real GDP per Capita (real local currency units... |
| **6390** | NE.GDI.FTOT.CR | GDP expenditure on gross fixed capital formati... |
| **6478** | NV.AGR.PCAP.KD.ZG | Real agricultural GDP per capita growth rate (%) |
| **6600** | NY.GDP.PCAP.CD | GDP per capita (current US$) |
| **6601** | NY.GDP.PCAP.CN | GDP per capita (current LCU) |
| **6602** | NY.GDP.PCAP.KD | GDP per capita (constant 2005 US$) |
| **6603** | NY.GDP.PCAP.KD.ZG | GDP per capita growth (annual %) |
| **6604** | NY.GDP.PCAP.KN | GDP per capita (constant LCU) |

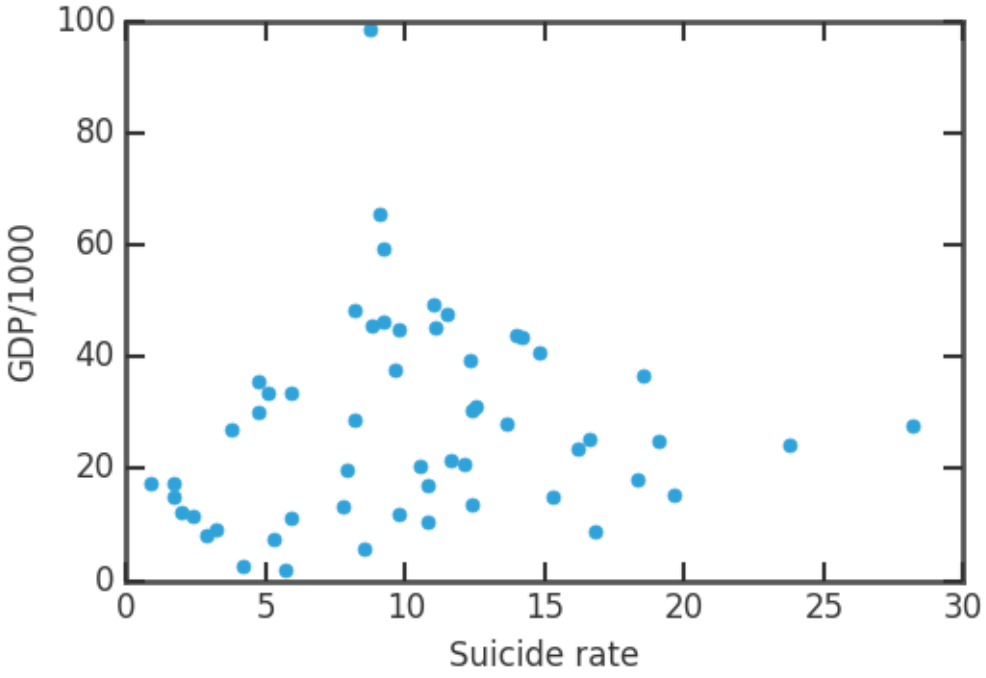
| | | NY.GDP.PCAP.PP.CD |
|---|---|---|
| **country** | **year** | |
| **Arab World** | **2014** | 15975.039211 |
| **Caribbean small states** | **2014** | 15231.111124 |
| **Central Europe and the Baltics** | **2014** | 23884.797208 |
| **East Asia & Pacific (all income levels)** | **2014** | 14853.204148 |
| **East Asia & Pacific (developing only)** | **2014** | 11922.720831 |

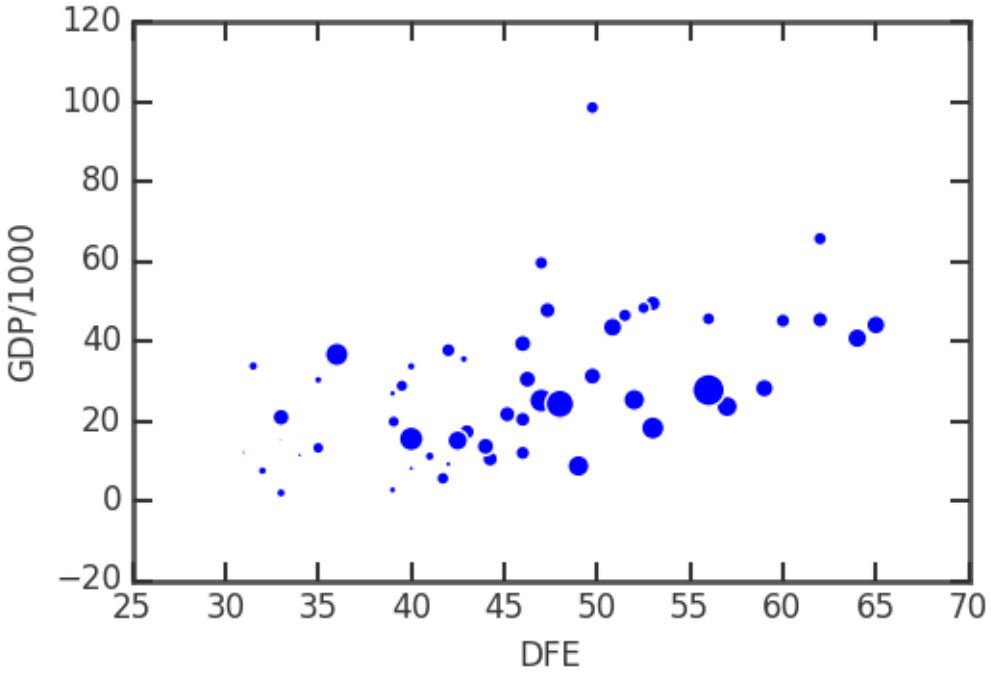
```
country                                 year
Arab World                              2014    15975.039211
Caribbean small states                  2014    15231.111124
Central Europe and the Baltics          2014    23884.797208
East Asia & Pacific (all income levels) 2014    14853.204148
East Asia & Pacific (developing only)   2014    11922.720831
Name: NY.GDP.PCAP.PP.CD, dtype: float64
```

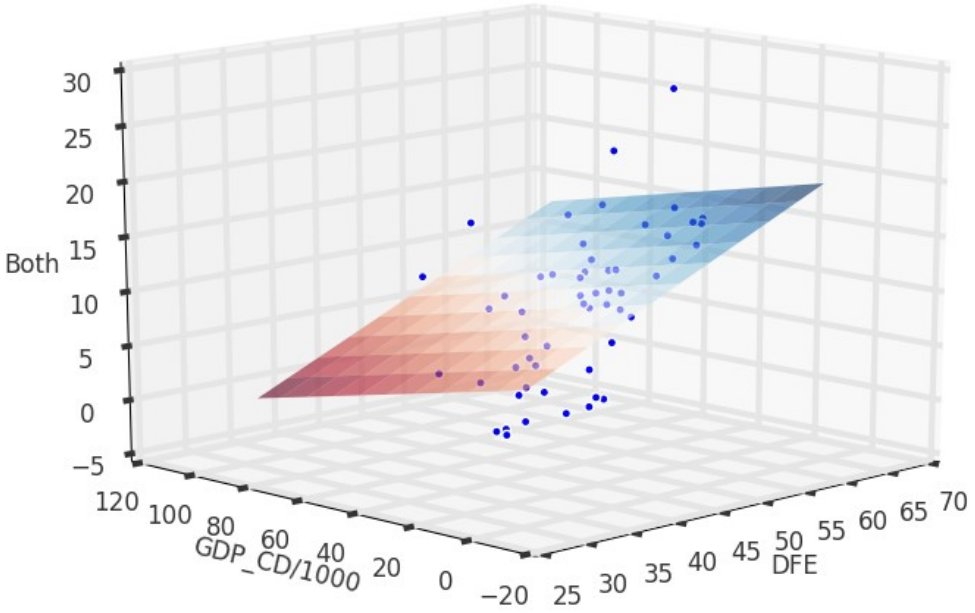
| | country | gdp |
|---|---|---|
| 0 | Arab World | 15975.039210528601 |
| 1 | Caribbean small states | 15231.111124481498 |
| 2 | Central Europe and the Baltics | 23884.797208032996 |
| 3 | East Asia & Pacific (all income levels) | 14853.204148331899 |
| 4 | East Asia & Pacific (developing only) | 11922.7208309251 |

```
    Country  Both  Female  Male  Lat  Lon  DFE GDP_CD
146  Sweden  11.1     6.1  16.2   62   15   62  45183
    country                 gdp
218  Sweden  45183.0196872713
[ 45183.01968727]
```

| | Country | Both | Female | Male | Lat | Lon | DFE | GDP_CD |
|---|---|---|---|---|---|---|---|---|
| **146** | Sweden | 11.1 | 6.1 | 16.2 | 62 | 15 | 62 | 45183 |




120
100
80
60
40
20
0
−20
GDP/1000
25
30
35
40
45
50
55
60
65
70
DFE


Both
30
25
20
15
10
5
0
−5
120
100
80
60
40
20
0
−20
GDP_CD/1000
25
30
35
40
45
50
55
60
65
70
DFE

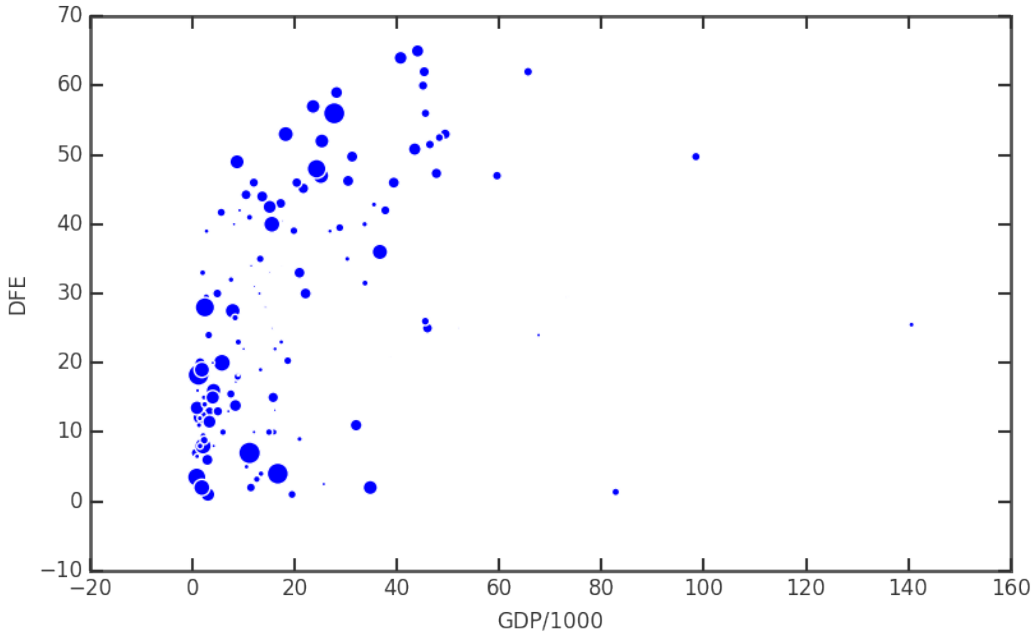
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Both | R-squared: | 0.288 |
| Model: | OLS | Adj. R-squared: | 0.260 |
| Method: | Least Squares | F-statistic: | 10.33 |
| Date: | Mon, 21 Dec 2015 | Prob (F-statistic): | 0.000171 |
| Time: | 03:40:42 | Log-Likelihood: | -161.90 |
| No. Observations: | 54 | AIC: | 329.8 |
| Df Residuals: | 51 | BIC: | 335.8 |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

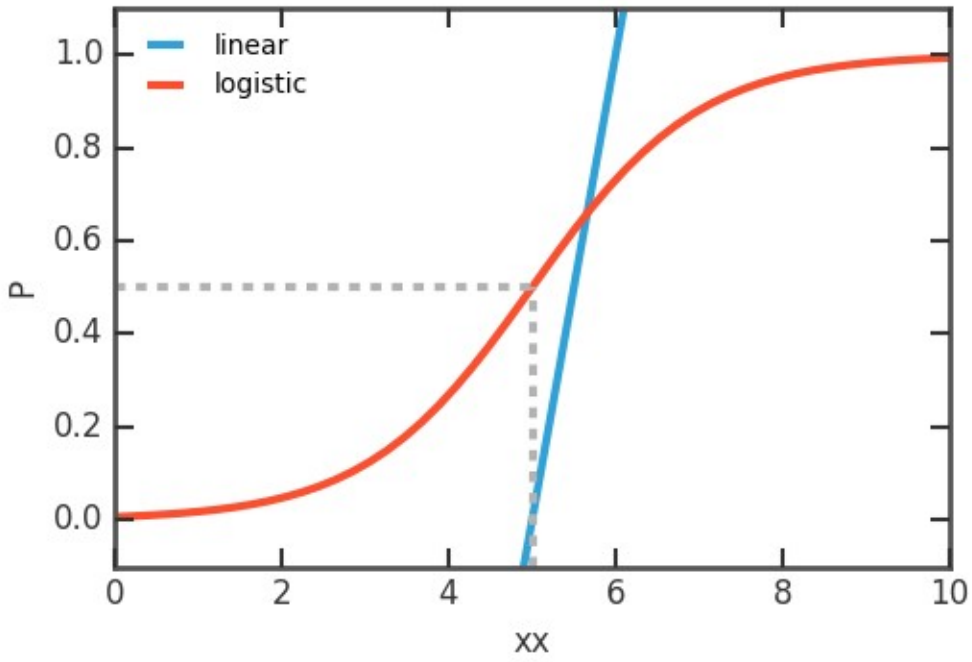
| | coef | std err | t | P>\|t\| | [95.0% Conf. | Int.] |
|---|---|---|---|---|---|---|
| const | -5.9298 | 3.626 | -1.635 | 0.108 | -13.210 | 1.350 |
| DFE | 0.3942 | 0.090 | 4.397 | 0.000 | 0.214 | 0.574 |
| GDP_CD | -6.238e-05 | 4.55e-05 | -1.371 | 0.176 | -0.000 | 2.9e-05 |

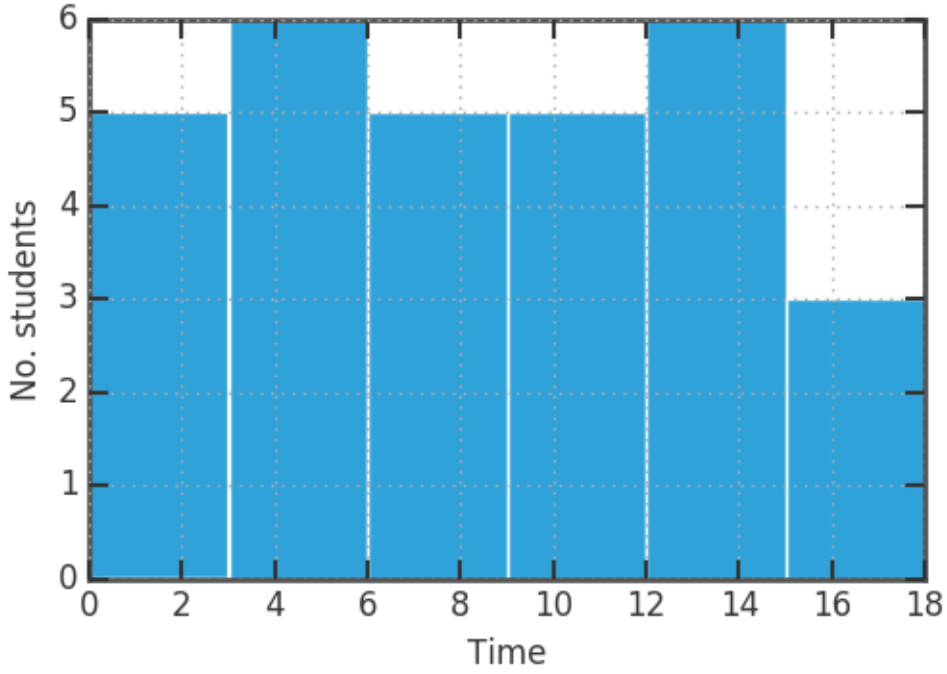
| | | | |
|---|---|---|---|
| Omnibus: | 11.836 | Durbin-Watson: | 2.272 |
| Prob(Omnibus): | 0.003 | Jarque-Bera (JB): | 11.958 |
| Skew: | 1.062 | Prob(JB): | 0.00253 |
| Kurtosis: | 3.899 | Cond. No. | 1.72e+05 |

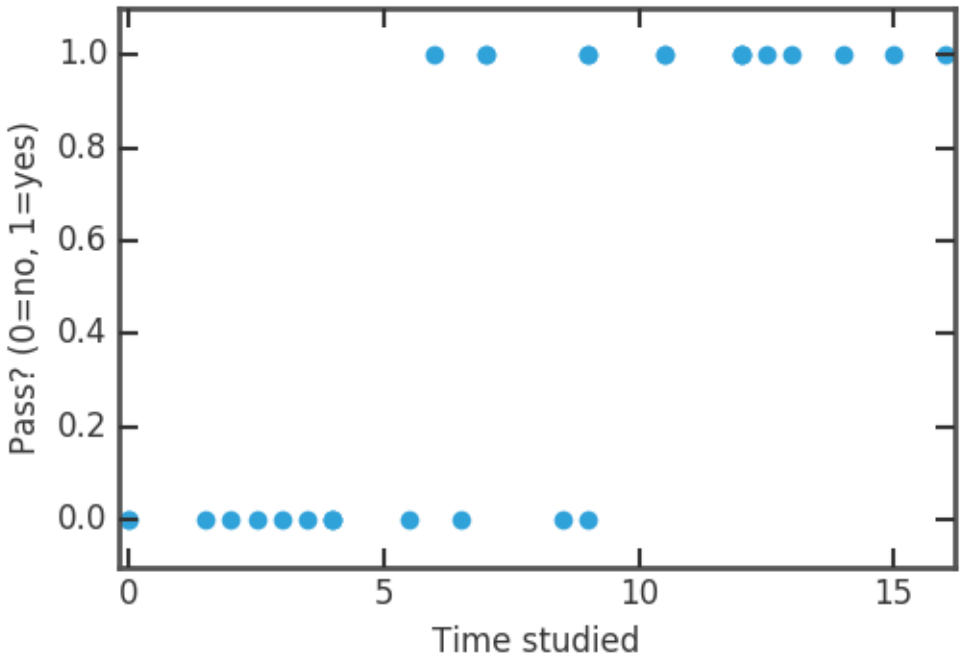
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.72e+05. This might indicate that there are
strong multicollinearity or other numerical problems.



| | Country | Both | Male | Female | GDP_CD | DFE |
|---|---|---|---|---|---|---|
| **0** | Afghanistan | 5.7 | 6.2 | 5.3 | 1932.89 | 33.0 |
| **1** | Albania | 5.9 | 6.6 | 5.2 | 10304.7 | 41.0 |
| **2** | Algeria | 1.9 | 2.3 | 1.5 | 14193.4 | 28.0 |
| **3** | Angola | 13.8 | 20.7 | 7.3 | NaN | 12.5 |
| **4** | Argentina | 10.3 | 17.2 | 4.1 | NaN | 34.0 |




No. students
Time
0
1
2
3
4
5
6
0
2
4
6
8
10
12
14
16
18


Pass? (0=no, 1=yes)
Time studied
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15

```
Optimization terminated successfully.
         Current function value: 0.251107
         Iterations 8
```

Logit Regression Results

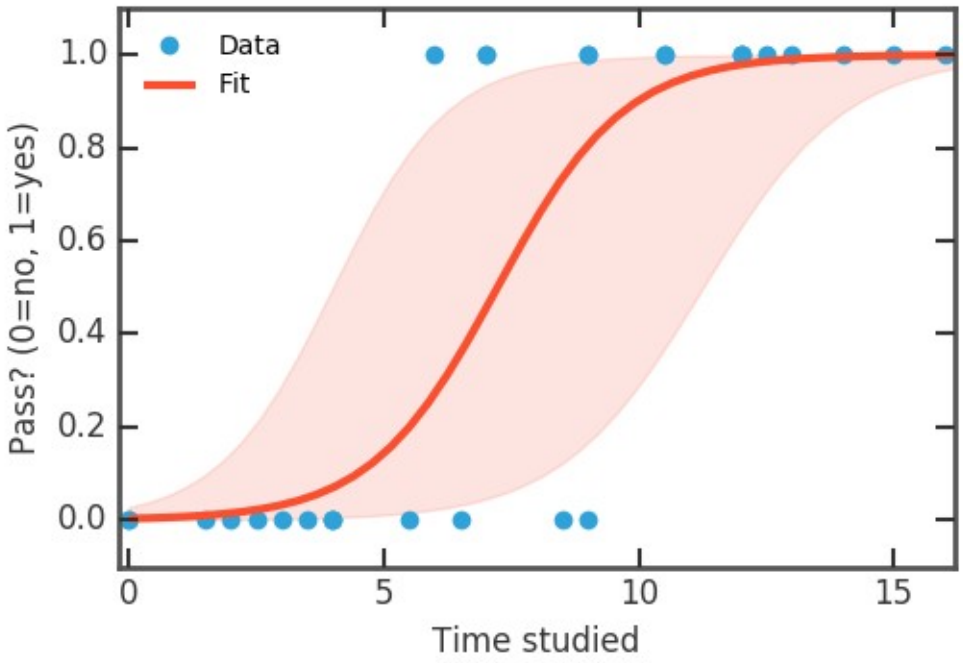
| | | | |
|---|---|---|---|
| Dep. Variable: | Pass | No. Observations: | 30 |
| Model: | Logit | Df Residuals: | 28 |
| Method: | MLE | Df Model: | 1 |
| Date: | Mon, 21 Dec 2015 | Pseudo R-squ.: | 0.6366 |
| Time: | 03:40:43 | Log-Likelihood: | -7.5332 |
| converged: | True | LL-Null: | -20.728 |
| | | LLR p-value: | 2.791e-07 |

| | coef | std err | z | P>\|z\| | [95.0% Conf. | Int.] |
|---|---|---|---|---|---|---|
| const | -5.7980 | 2.240 | -2.588 | 0.010 | -10.188 | -1.408 |
| Time | 0.8020 | 0.297 | 2.703 | 0.007 | 0.220 | 1.384 |

| | const | Time |
|---|---|---|
| **const** | 5.017663 | -0.635081 |
| **Time** | -0.635081 | 0.088035 |

```
-5.79798670884 0.801979232718
```

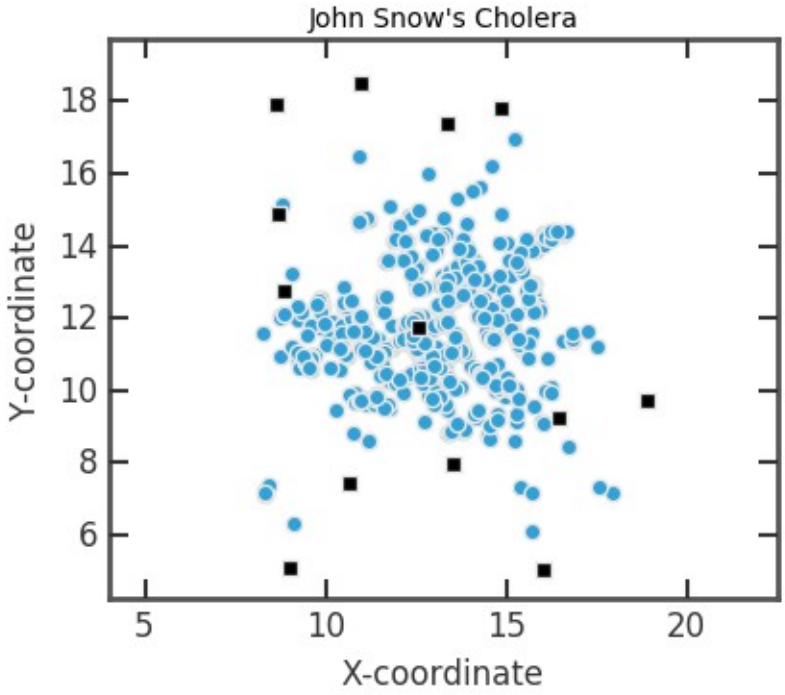
| | 0 | 1 |
|---|---|---|
| **const** | -10.188333 | -1.407640 |
| **Time** | 0.220444 | 1.383514 |

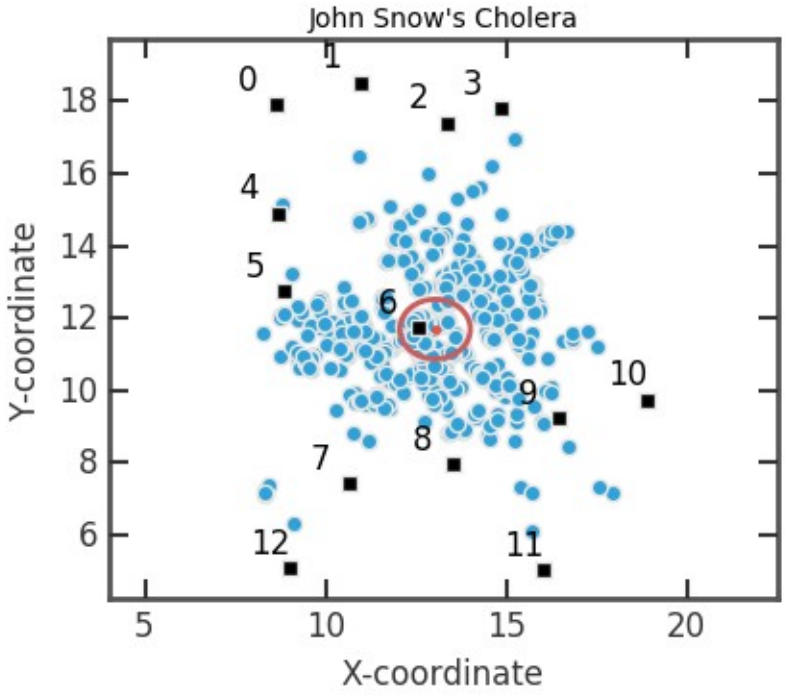

Data
Fit
Pass? (0=no, 1=yes)
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
Time studied

## Chapter 5: Clustering

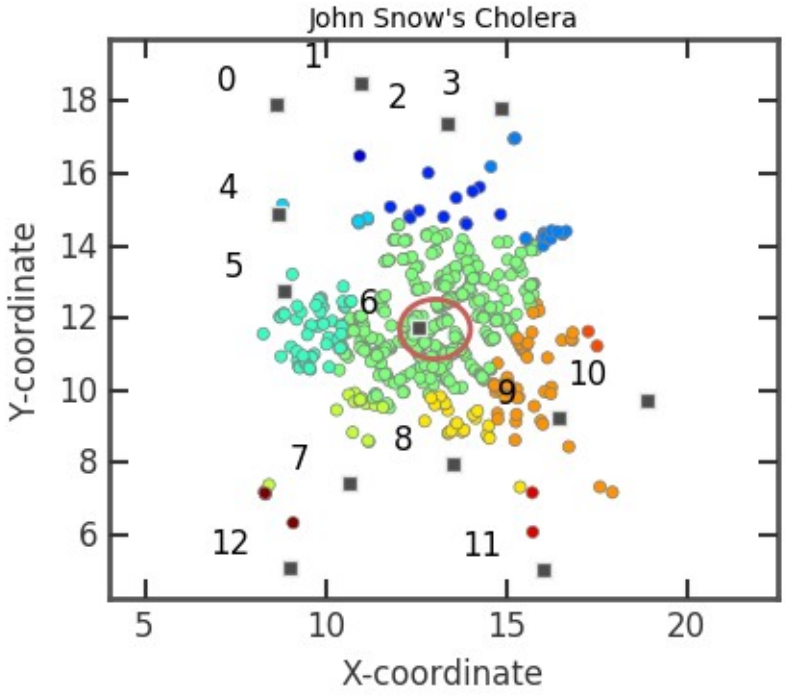
| | X | Y |
|---|---|---|
| **0** | 13.588010 | 11.095600 |
| **1** | 9.878124 | 12.559180 |
| **2** | 14.653980 | 10.180440 |
| **3** | 15.220570 | 9.993003 |
| **4** | 13.162650 | 12.963190 |

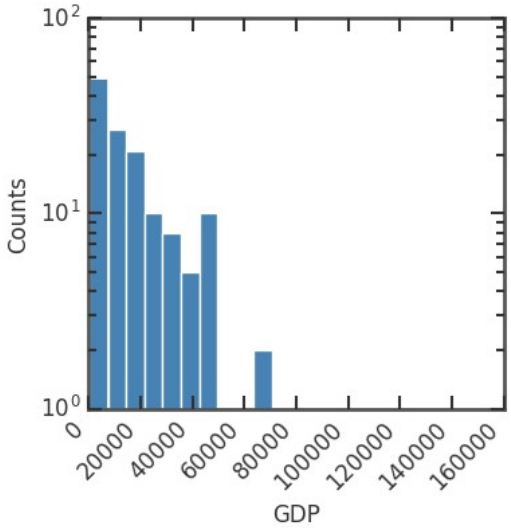
| | X | Y |
|---|---|---|
| **0** | 8.651201 | 17.891600 |
| **1** | 10.984780 | 18.517851 |
| **2** | 13.378190 | 17.394541 |
| **3** | 14.879830 | 17.809919 |
| **4** | 8.694768 | 14.905470 |

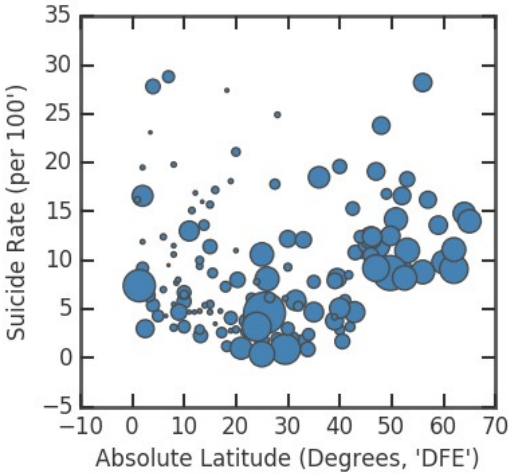

John Snow's Cholera
Y-coordinate
X-coordinate
18
16
14
12
10
8
6
5
10
15
20

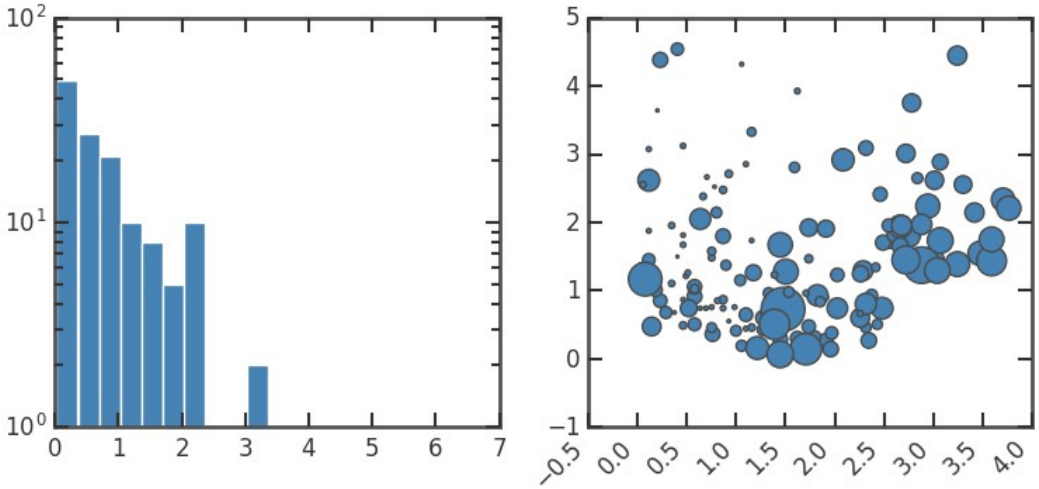


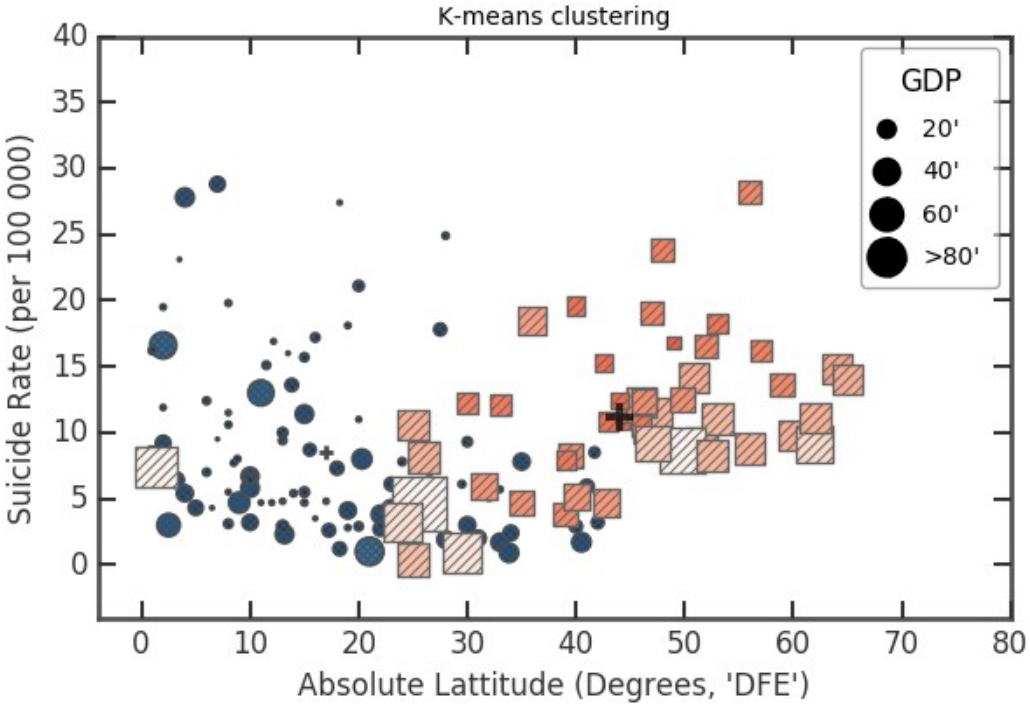

| | X | Y | C |
|---|---|---|---|
| **0** | 13.588010 | 11.095600 | 6 |
| **1** | 9.878124 | 12.559180 | 5 |
| **2** | 14.653980 | 10.180440 | 9 |
| **3** | 15.220570 | 9.993003 | 9 |
| **4** | 13.162650 | 12.963190 | 6 |


John Snow's Cholera
Y-coordinate
X-coordinate
0
1
2
3
4
5
6
7
8
9
10
11
12


Counts
GDP

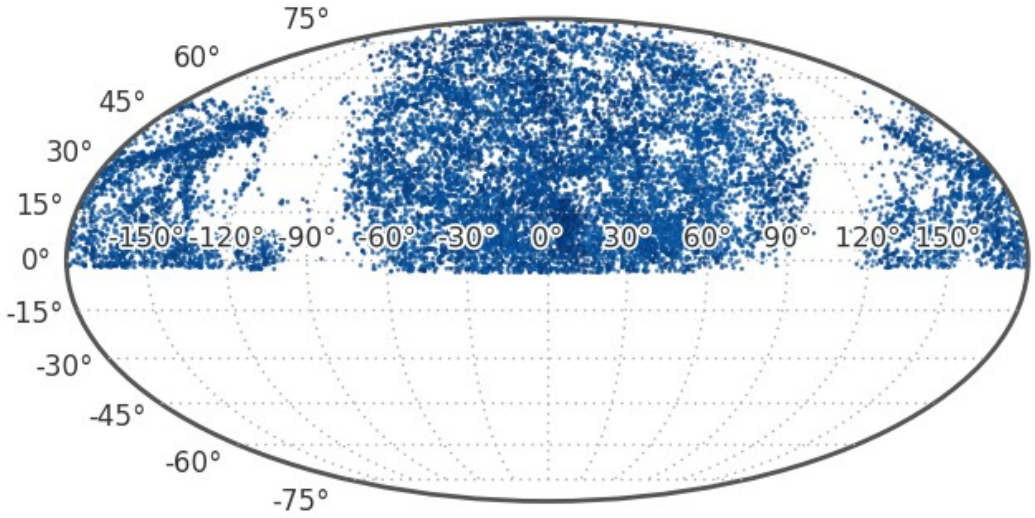

Suicide Rate (per 100')
Absolute Latitude (Degrees, 'DFE')

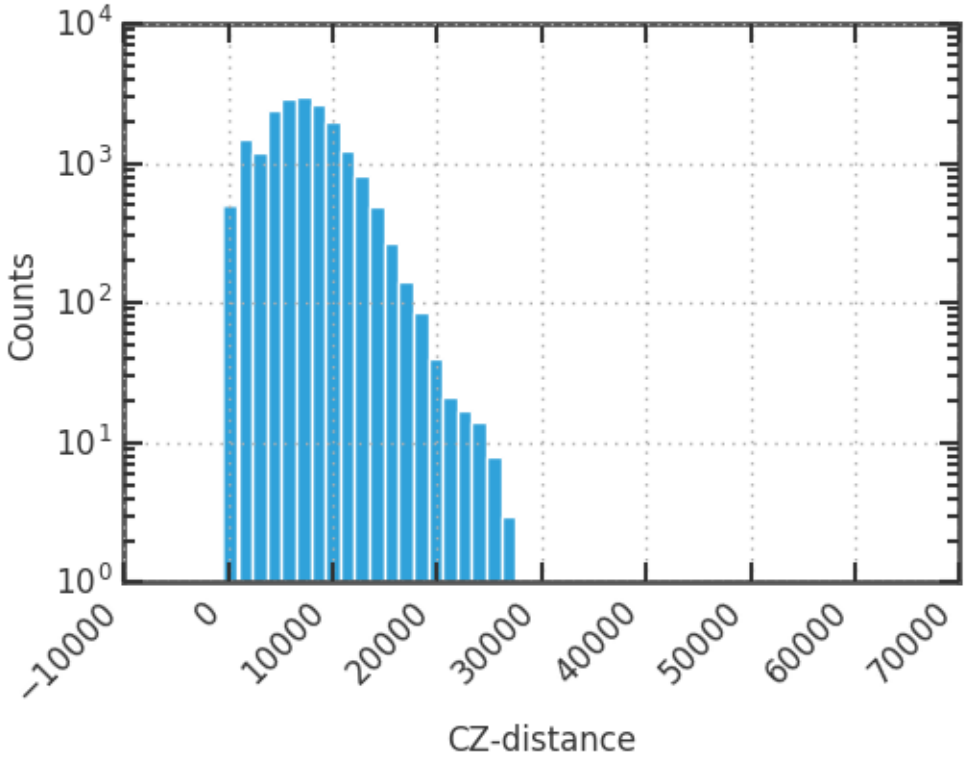

$10^2$
$10^1$
$10^0$
0 1 2 3 4 5 6 7
5 4 3 2 1 0 −1
−0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

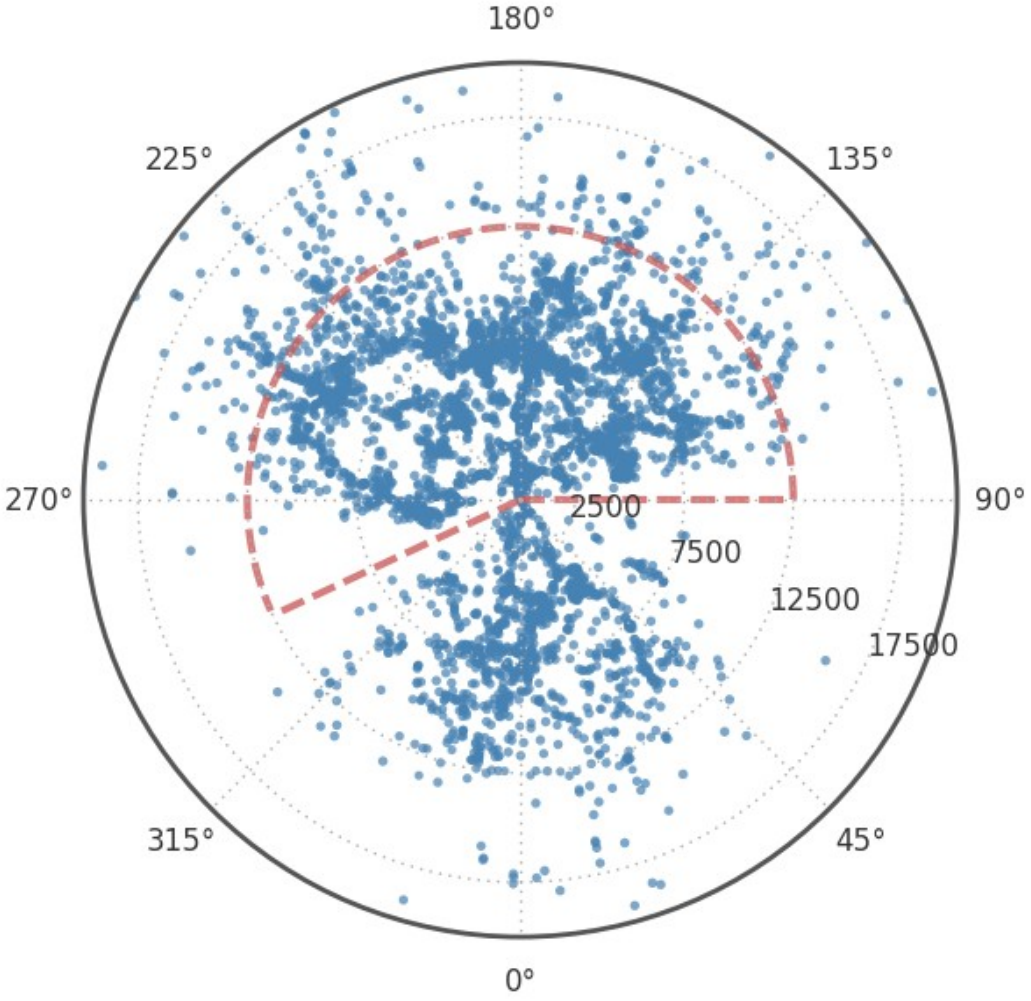

K-means clustering
GDP
20'
40'
60'
>80'
Suicide Rate (per 100 000)
0 5 10 15 20 25 30 35 40
0 10 20 30 40 50 60 70 80
Absolute Lattitude (Degrees, 'DFE')

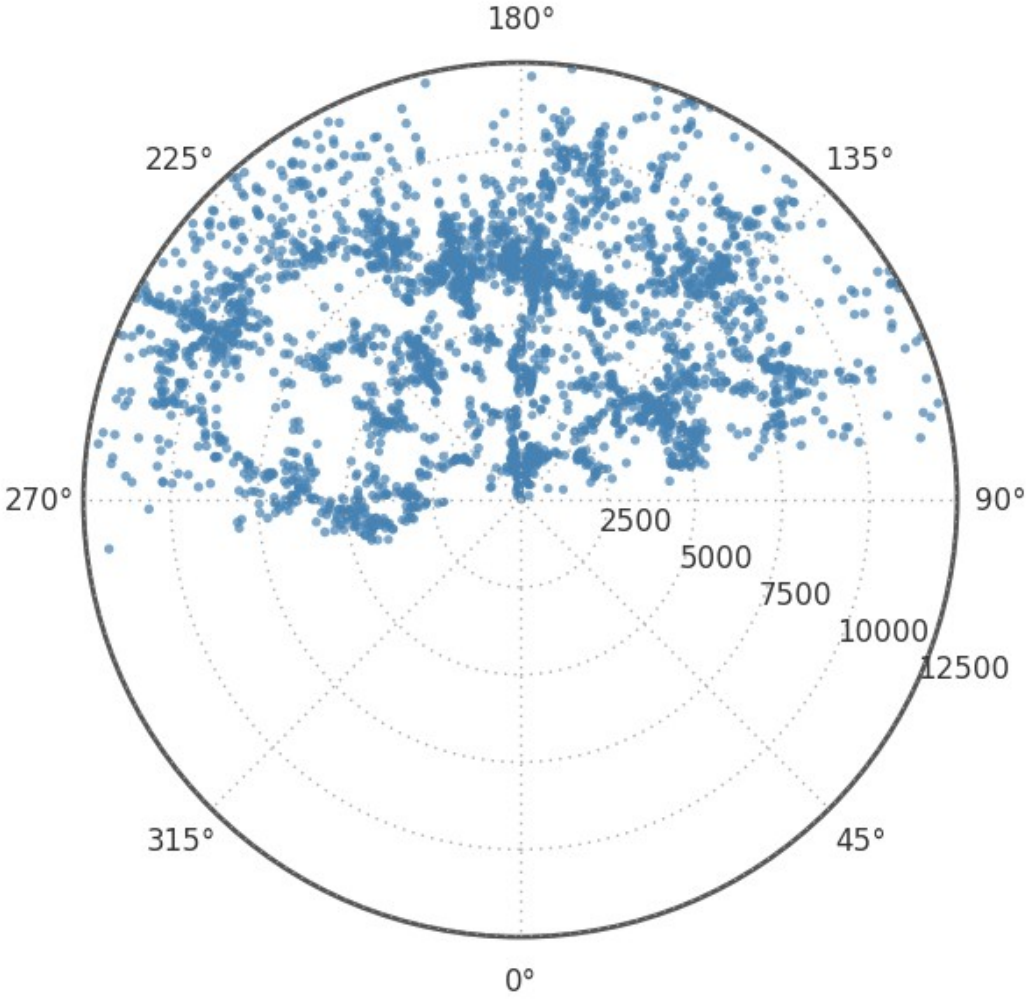
| | ra | dec | Zmag | cz | cze | T | U | Ne | Zname | C | Ref | Oname | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 000237.9 | +163838 | 14.9 | 6350 | 19 | A | 1 | 0 | 000000+16220 | F | | I5378S | | |
| **1** | 000246.3 | +185310 | 14.8 | 7864 | 47 | A | 0 | 0 | 000012+18370 | Z | 0650 | 00002+1837 | | |
| **2** | 000257.0 | +041231 | 15.5 | 8695 | 40 | E | 0 | 0 | 000030+03560 | Z | 2700 | 00005+0356 | | |
| **3** | 000302.9 | +185221 | 15.5 | 8007 | 39 | E | 0 | 0 | 000030+18360 | Z | 0650 | 00005+1836 | | |
| **4** | 000305.6 | -015450 | 14.3 | 7298 | 42 | B | 0 | 0 | 000036-02110 | Z | 2218 | 00006-0211 | | |

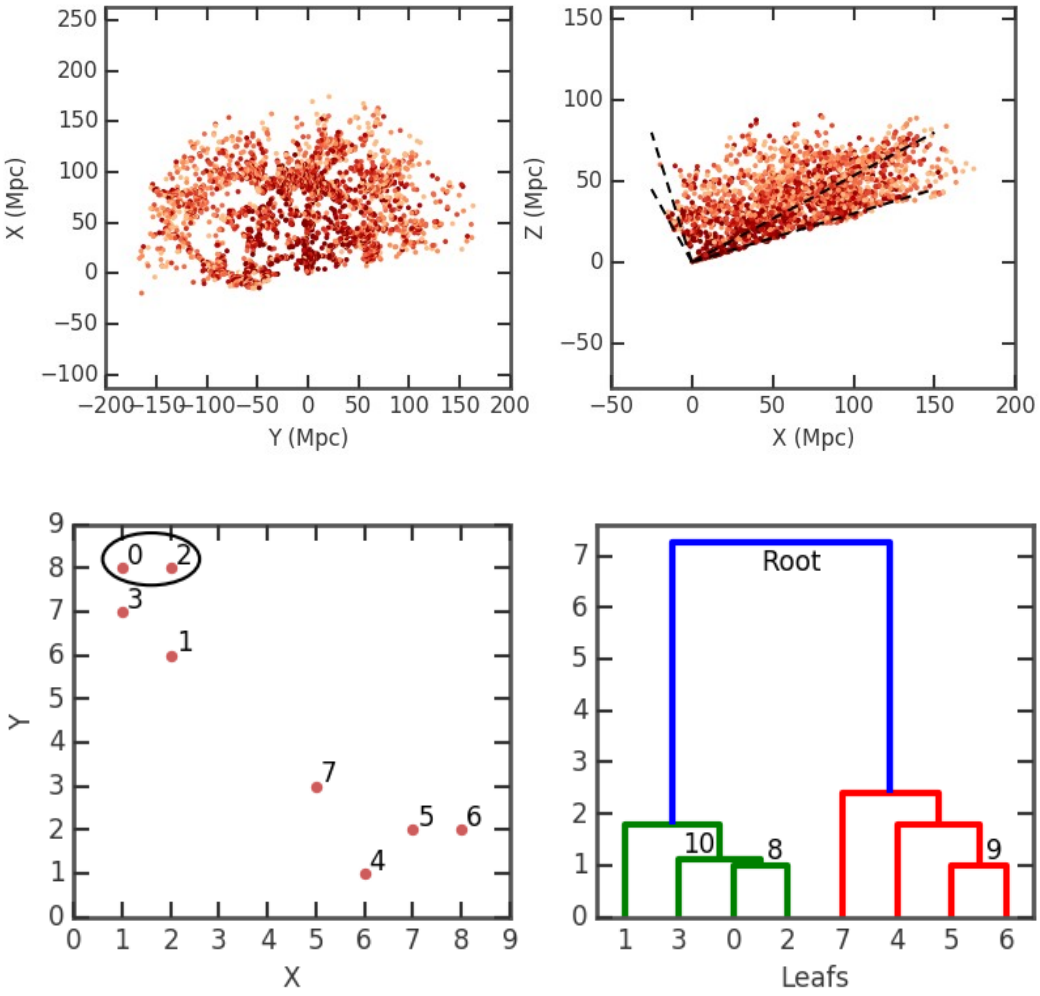
| | ra | dec | Zmag | cz | cze | T | U | Ne | Zname | C | Ref | Oname | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 00h02m37.9s | +16d38m38s | 14.9 | 6350 | 19 | A | 1 | 0 | 000000+16220 | F | | I5378S | | |
| **1** | 00h02m46.3s | +18d53m10s | 14.8 | 7864 | 47 | A | 0 | 0 | 000012+18370 | Z | 0650 | 00002+1837 | | |
| **2** | 00h02m57.0s | +04d12m31s | 15.5 | 8695 | 40 | E | 0 | 0 | 000030+03560 | Z | 2700 | 00005+0356 | | |
| **3** | 00h03m02.9s | +18d52m21s | 15.5 | 8007 | 39 | E | 0 | 0 | 000030+18360 | Z | 0650 | 00005+1836 | | |
| **4** | 00h03m05.6s | -01d54m50s | 14.3 | 7298 | 42 | B | 0 | 0 | 000036-02110 | Z | 2218 | 00006-0211 | | |

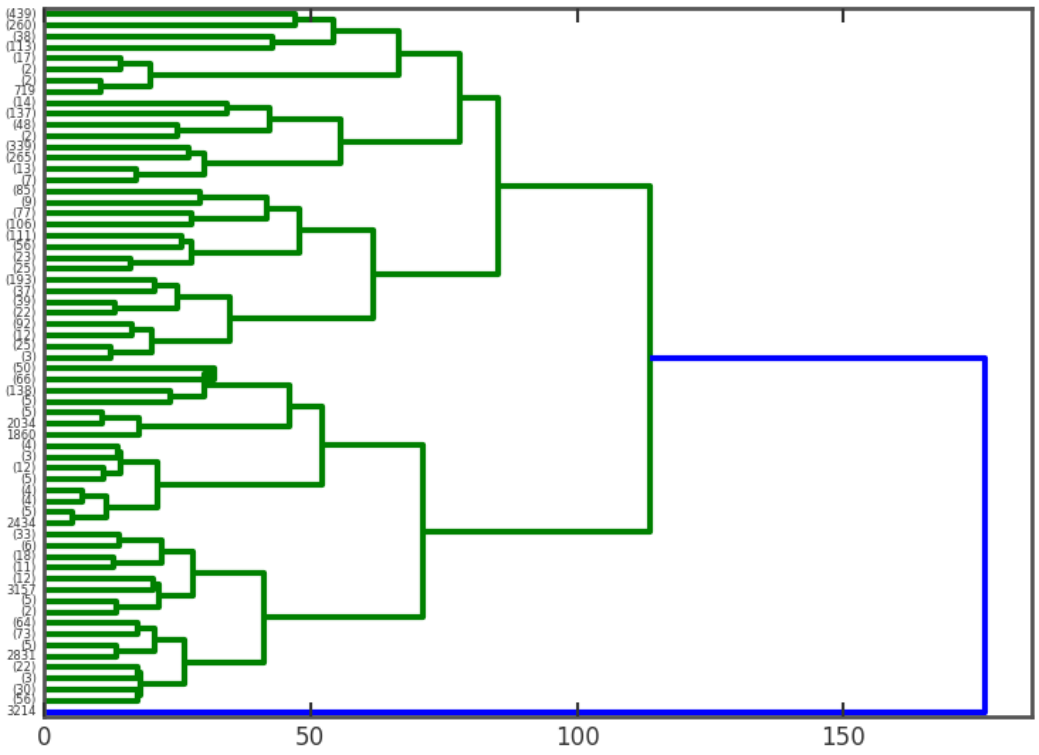


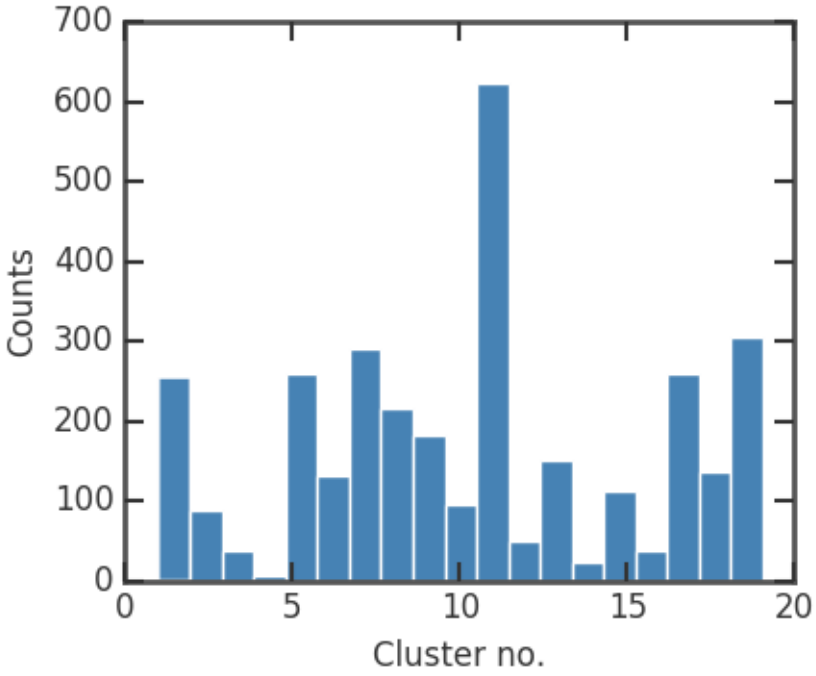

$10^4$
$10^3$
$10^2$
$10^1$
$10^0$
Counts
−10000
0
10000
20000
30000
40000
50000
60000
70000
CZ-distance

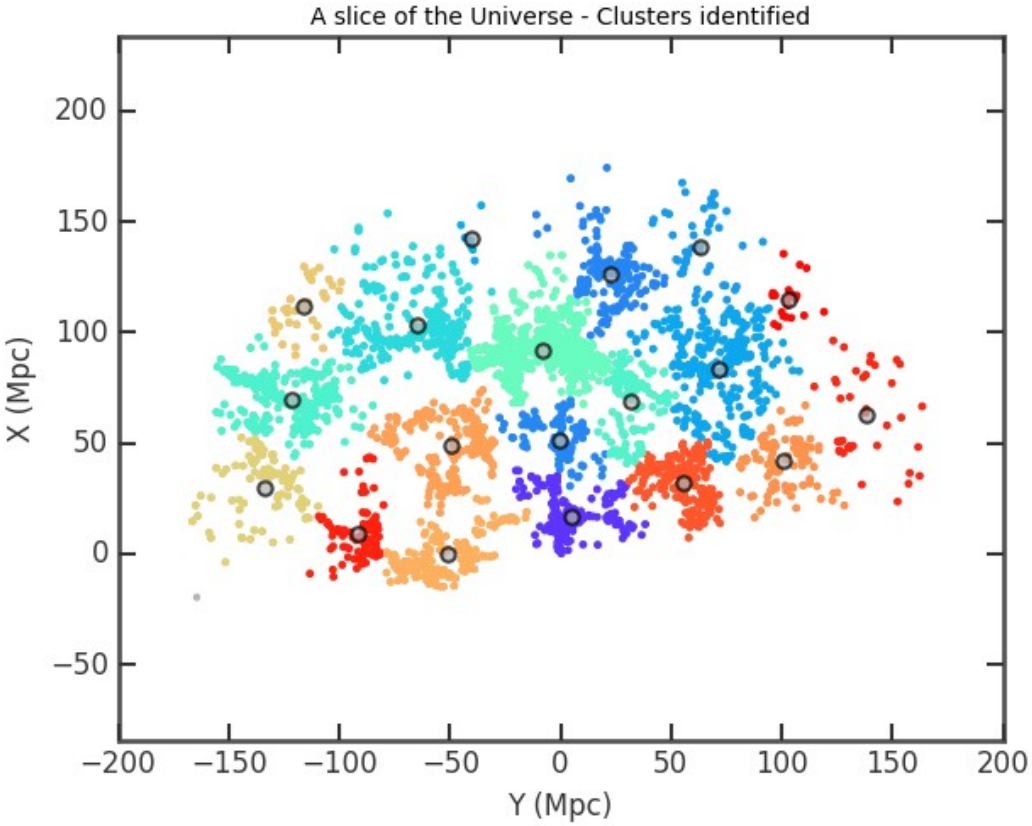

180°
225°
135°
270°
90°
315°
45°
0°
2500
7500
12500
17500


180°
225°
135°
270°
90°
2500
5000
7500
10000
12500
315°
45°
0°


X (Mpc)
Y (Mpc)
Z (Mpc)
X (Mpc)
Y
X
0
2
3
1
7
5
6
4
Root
10
8
9
Leafs

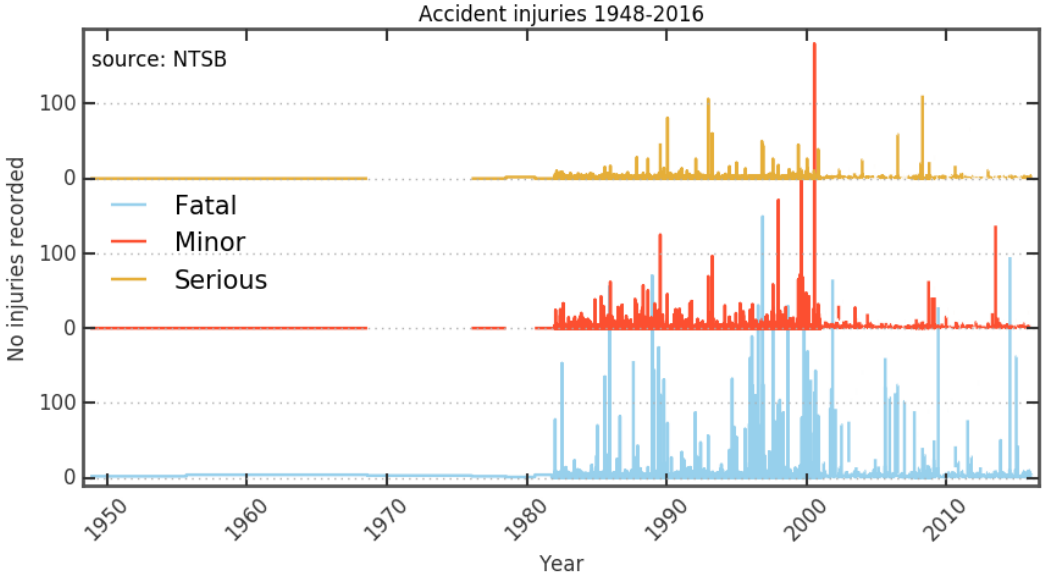

(439)
(260)
(38)
(113)
(17)
(2)
(2)
779
(14)
347
(137)
(48)
(2)
(339)
(265)
(13)
(7)
(85)
(9)
(77)
(106)
(111)
(56)
(23)
(25)
(193)
(37)
(39)
(22)
(92)
(15)
(25)
(3)
(50)
(66)
(66)
(138)
(5)
(5)
2034
1860
(4)
(3)
(12)
(5)
(4)
(4)
(5)
2434
(33)
(6)
(18)
(11)
(12)
315
(5)
(2)
(64)
(73)
(5)
2831
(22)
(3)
(30)
(58)
3214
0
50
100
150

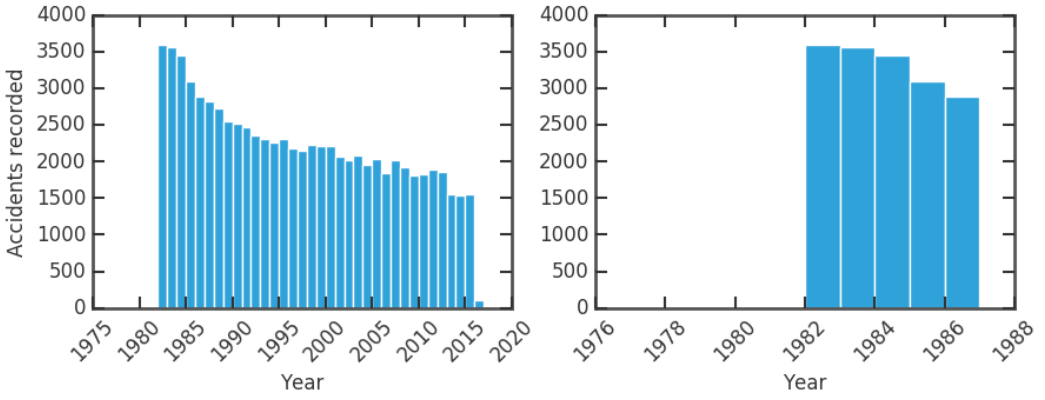

700
600
500
400
300
200
100
0
Counts
0
5
10
15
20
Cluster no.

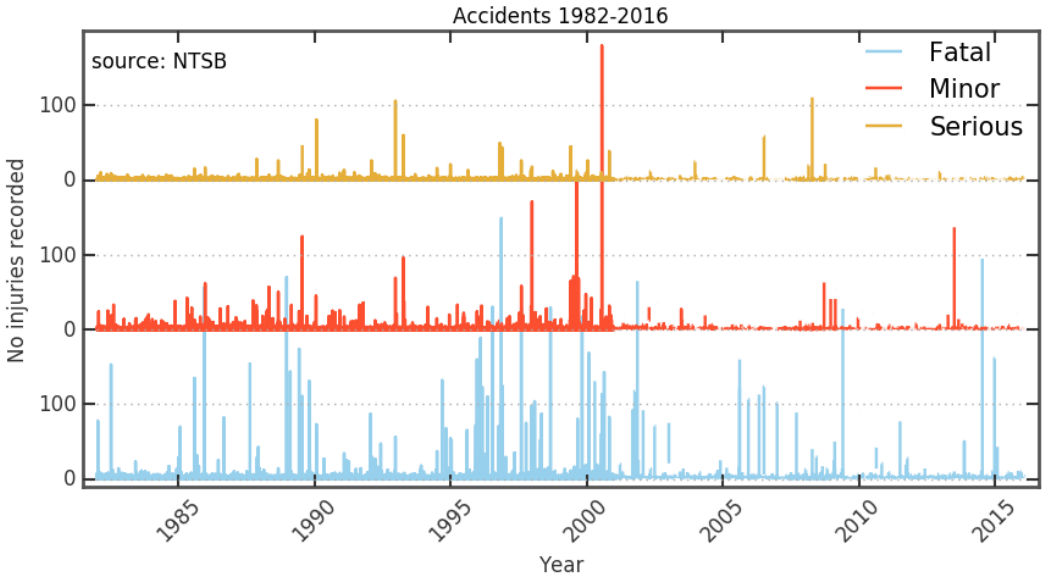

A slice of the Universe - Clusters identified
X (Mpc)
Y (Mpc)
200
150
100
50
0
−50
−200
−150
−100
−50
0
50
100
150
200

# Chapter 6: Bayesian Methods

```
Index(['id', 'type', 'number', 'date', 'location', 'country', 'lat', 'long',
       'airport_code', 'airport_name', 'injury_severity', 'aircraft_damage',
       'aircraft_cat', 'reg_no', 'make', 'model', 'amateur_built',
       'no_engines', 'engine_type', 'FAR_desc', 'schedule', 'purpose',
       'air_carrier', 'fatal', 'serious', 'minor', 'uninjured', 'weather',
       'broad_phase', 'report_status', 'pub_date', 'none'],
      dtype='object')
```

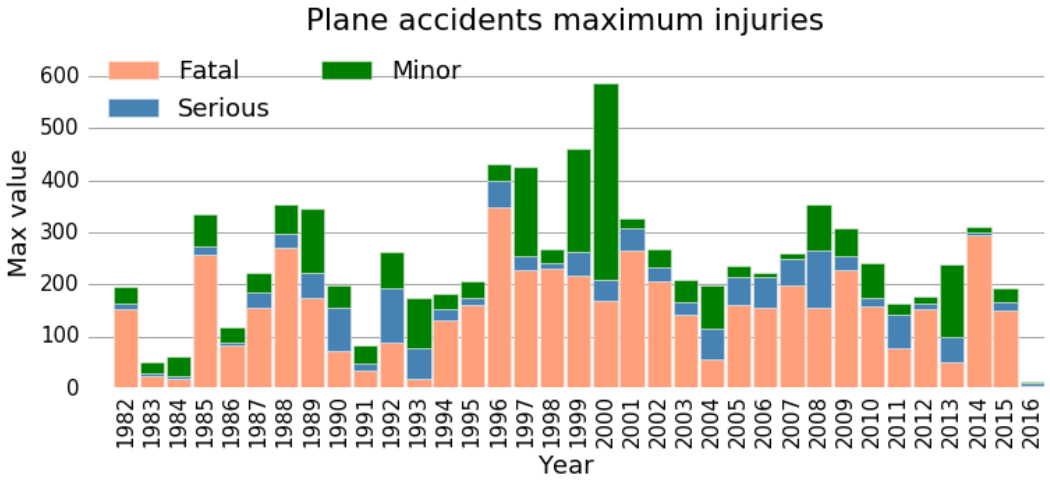




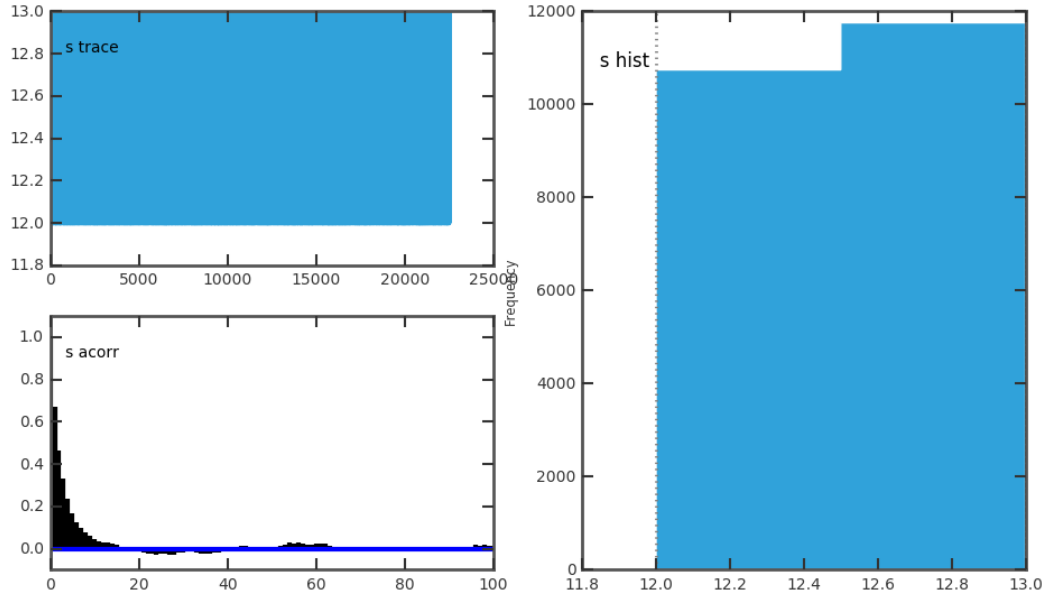

| | lat | long | fatal | serious | minor | uninjured | month | year | decyear |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 30.757778 | -88.355555 | 0.443978 | 0.203699 | 0.279474 | 2.317168 | 6.488450 | 1982 | 1982.495213 |
| **2** | 47.080556 | -117.368611 | 0.358996 | 0.190059 | 0.296213 | 4.258810 | 6.652137 | 1983 | 1983.508864 |
| **3** | NaN | NaN | 0.356749 | 0.202322 | 0.303919 | 3.621739 | 6.553659 | 1984 | 1984.502076 |
| **4** | NaN | NaN | 0.534198 | 0.198379 | 0.359507 | 3.663855 | 6.477390 | 1985 | 1985.494267 |
| **5** | NaN | NaN | 0.410435 | 0.215454 | 0.338097 | 4.138531 | 6.480556 | 1986 | 1986.495181 |

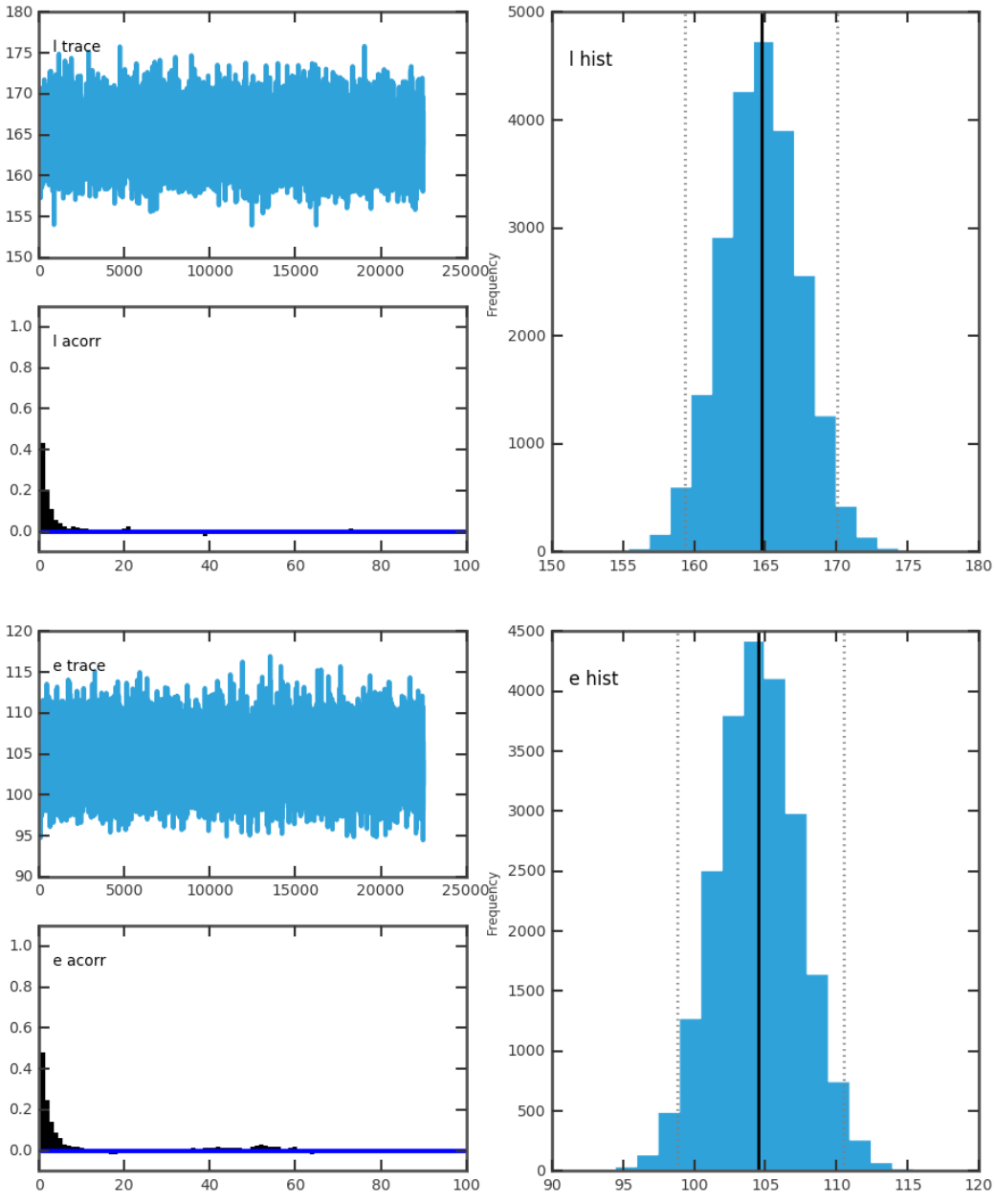
```
array([ 1982.,  1983.,  1984.,  1985.,  1986.,  1987.,  1988.,  1989.,
        1990.,  1991.,  1992.,  1993.,  1994.,  1995.,  1996.,  1997.,
        1998.,  1999.,  2000.,  2001.,  2002.,  2003.,  2004.,  2005.,
        2006.,  2007.,  2008.,  2009.,  2010.,  2011.,  2012.,  2013.,
        2014.,  2015.,  2016.])
```

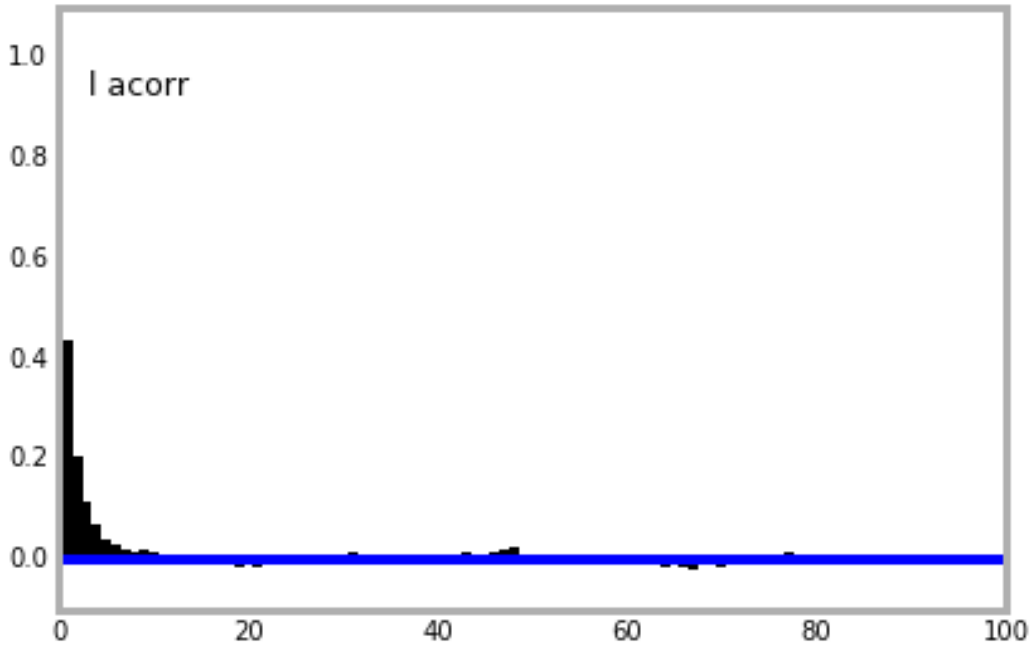


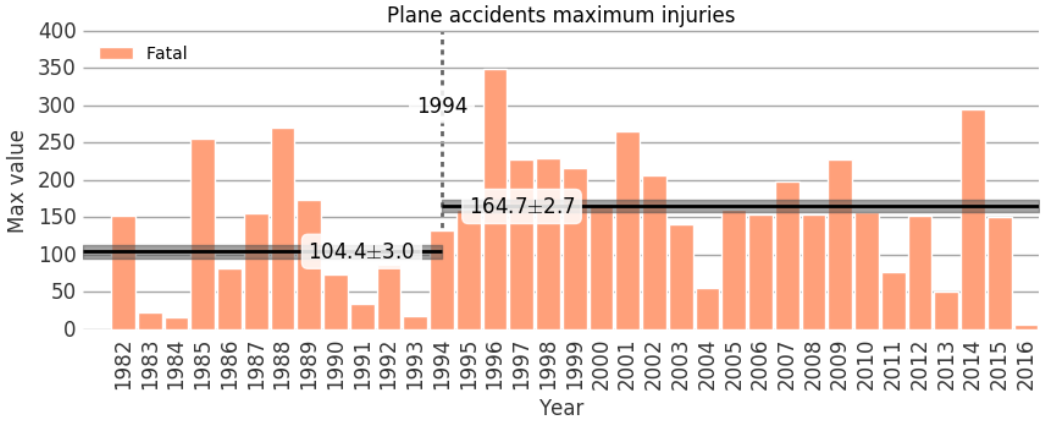

```
[-----------------100%-----------------] 50000 of 50000 complete in 8.6 sec
```

```
{<pymc.distributions.new_dist_class.<locals>.new_class 'l' at 0x7f841e56b470>: [<pymc.StepMethods.Metropolis at 0x7f
841e56b390>],
 <pymc.distributions.new_dist_class.<locals>.new_class 'e' at 0x7f841e56b358>: [<pymc.StepMethods.Metropolis at 0x7f
841e56b208>],
 <pymc.distributions.new_dist_class.<locals>.new_class 's' at 0x7f841e56b3c8>: [<pymc.StepMethods.DiscreteMetropolis
at 0x7f841e56b2e8>]}
```

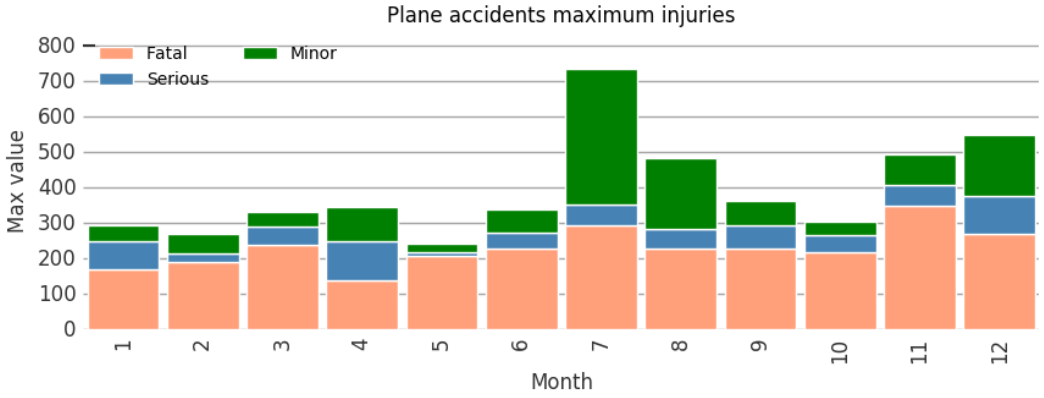


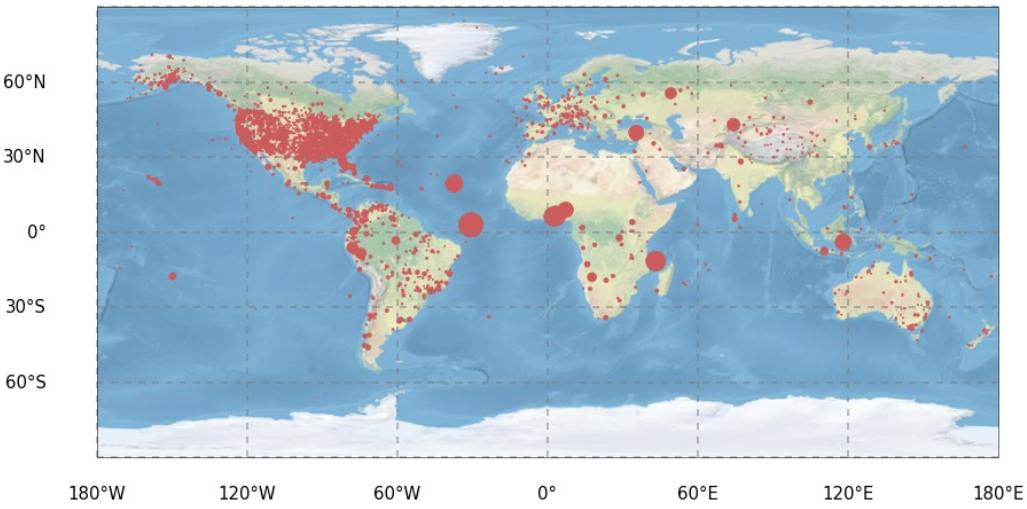

l trace
l acorr
l hist
Frequency
e trace
e acorr
e hist
Frequency

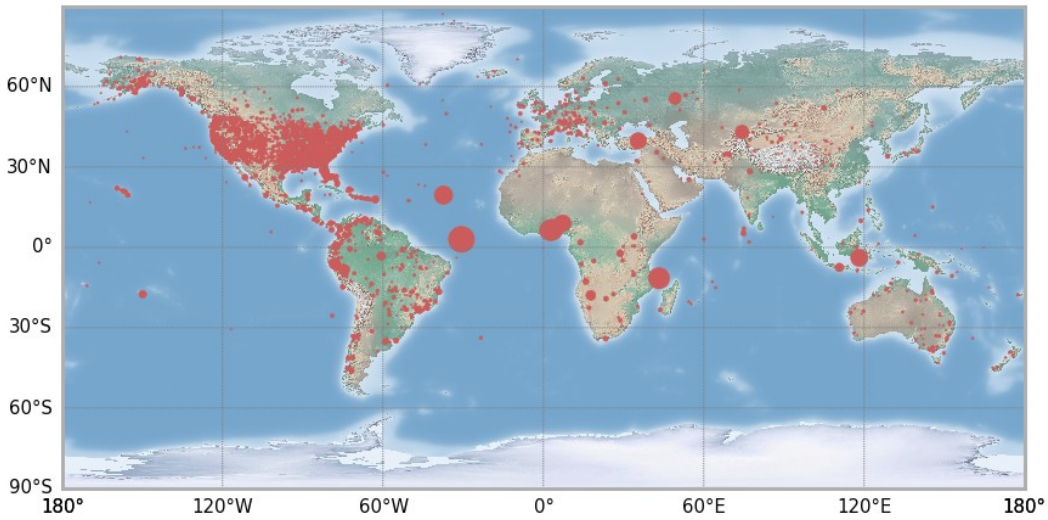

l acorr
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100


Plane accidents maximum injuries
Fatal
1994
164.7±2.7
104.4±3.0
Max value
Year
400
350
300
250
200
150
100
50
0
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016

| | lat | long | fatal | serious | minor | uninjured | month | year | decyear |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 35.355070 | -91.799283 | 1.024984 | 0.351452 | 0.482369 | 7.505215 | 1 | 1996.595841 | 1996.637186 |
| **2** | 35.211711 | -92.921433 | 0.890277 | 0.295262 | 0.486586 | 6.861646 | 2 | 1996.190641 | 1996.313377 |
| **3** | 36.439443 | -93.719725 | 0.673851 | 0.281289 | 0.428362 | 6.491056 | 3 | 1996.299847 | 1996.503531 |
| **4** | 37.187038 | -93.330245 | 0.651183 | 0.304870 | 0.489868 | 5.455664 | 4 | 1996.292600 | 1996.579883 |
| **5** | 37.816921 | -94.954229 | 0.671199 | 0.272957 | 0.445699 | 5.194848 | 5 | 1996.382683 | 1996.754983 |



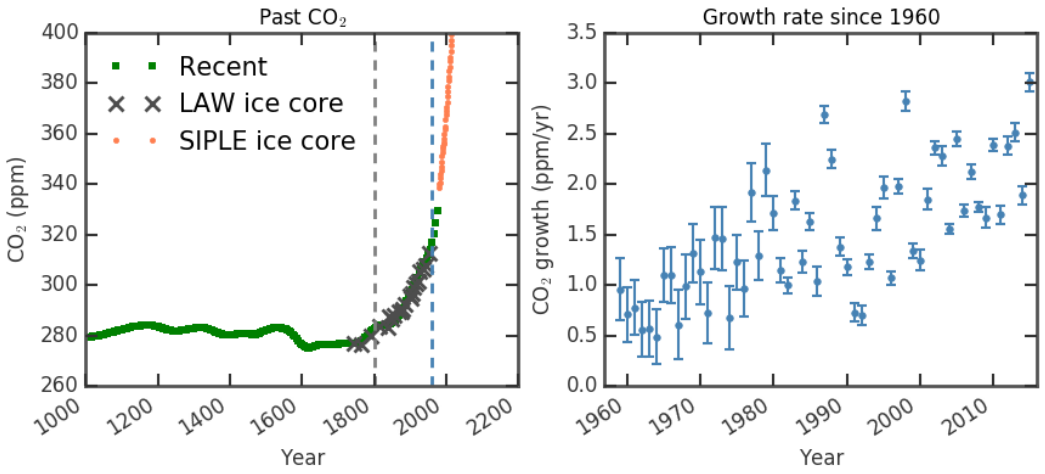



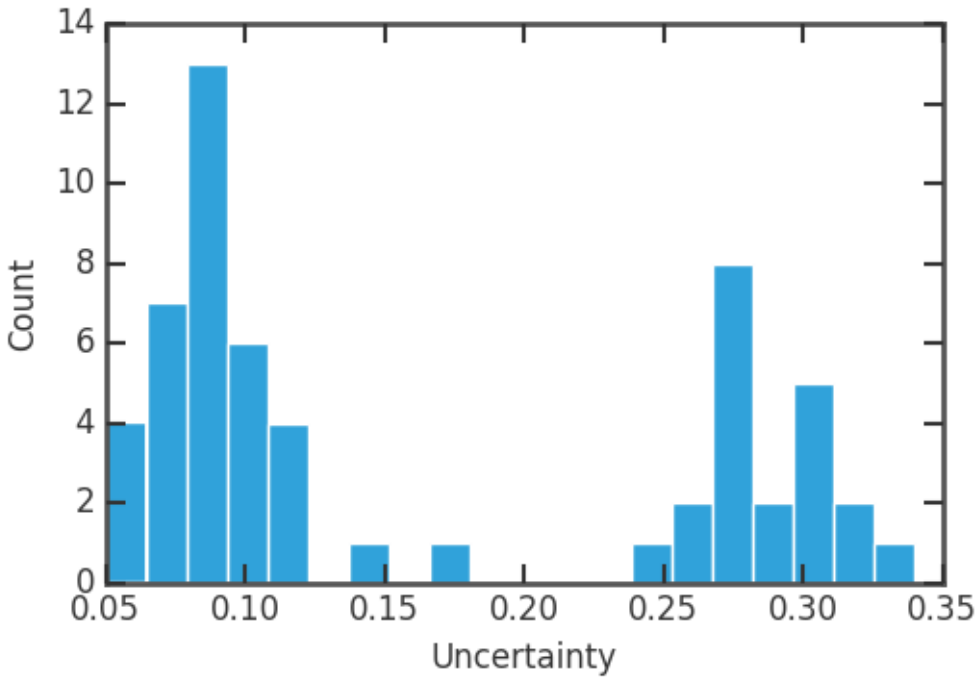


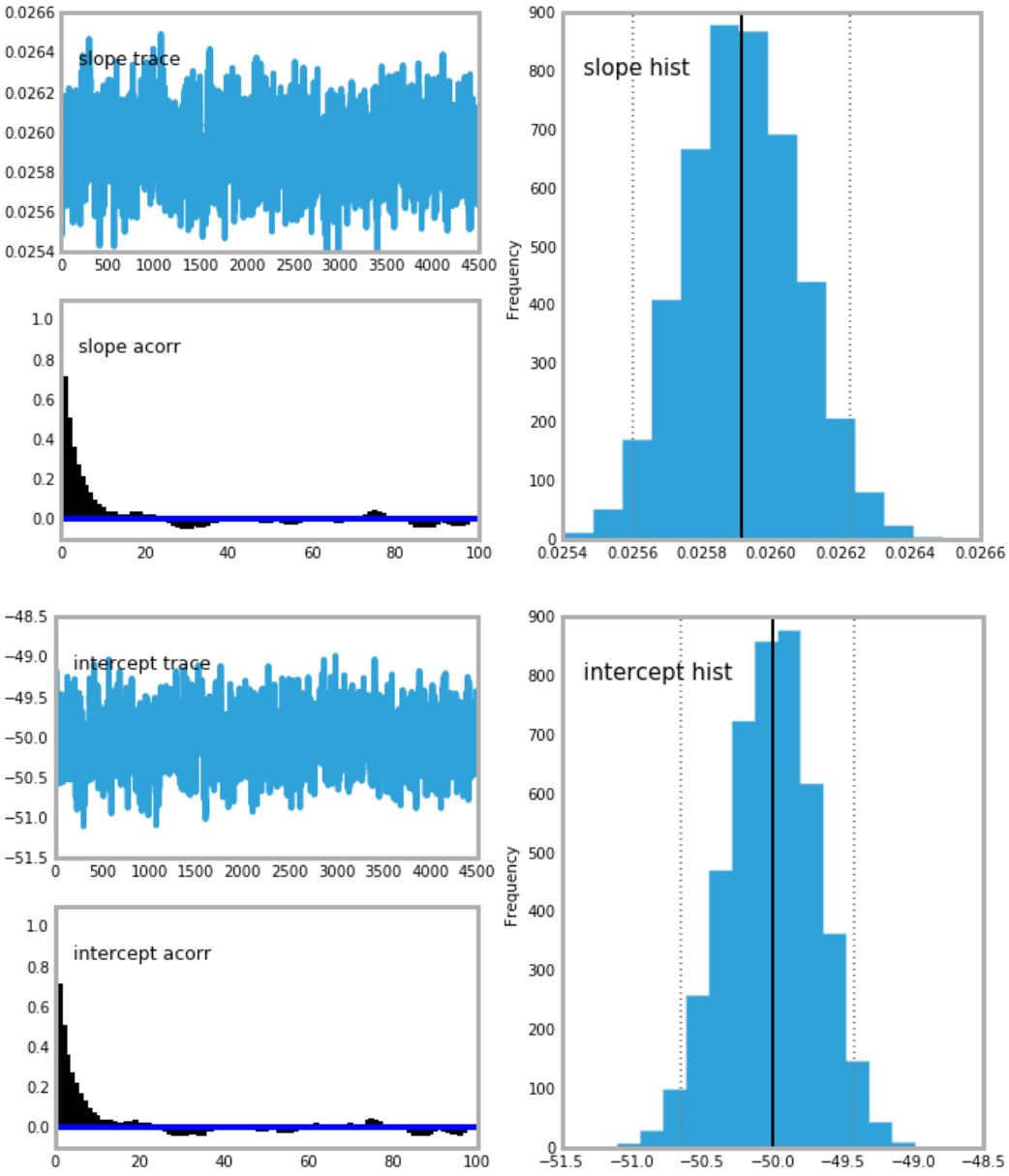

| | depth | year | co2 |
|---|---|---|---|
| **20** | 86.80 | 1943 | 307.9 |
| **21** | 81.22 | 1953 | 312.7 |
| **22** | Data | in | the |
| **23** | table | were | published |
| **24** | CO2 | concentrations | are |

```
object float64 float64 float64
```

```
(dtype('float64'), dtype('int64'))
```


Past $CO_2$
Recent
LAW ice core
SIPLE ice core
$CO_2$ (ppm)
Year
Growth rate since 1960
$CO_2$ growth (ppm/yr)
Year

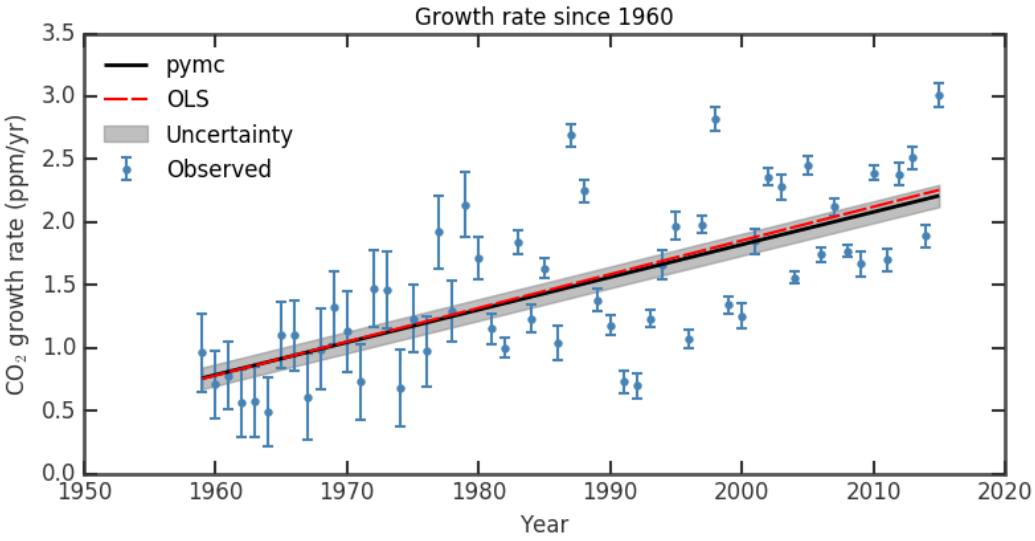

Count
Uncertainty



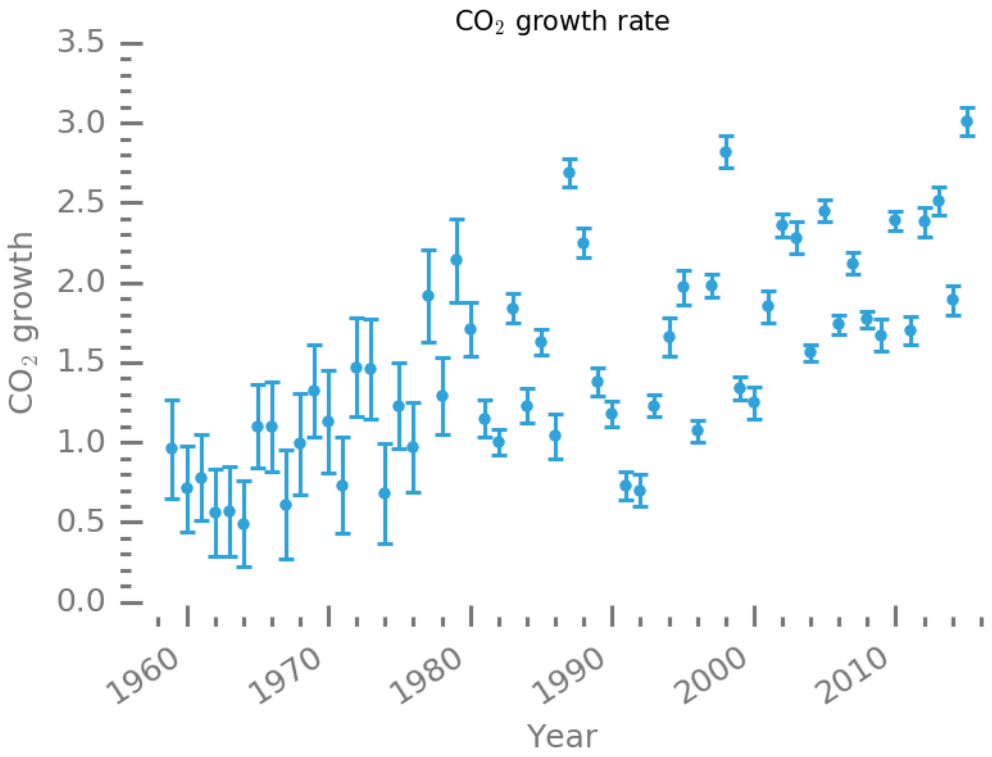

OLS: slope:0.027, intercept:-51.81
Bay: slope:0.026, intercept:-50.01

| | 0 | 1 |
|---|---|---|
| **Intercept** | -66.531103 | -37.092365 |
| **year** | 0.019425 | 0.034240 |

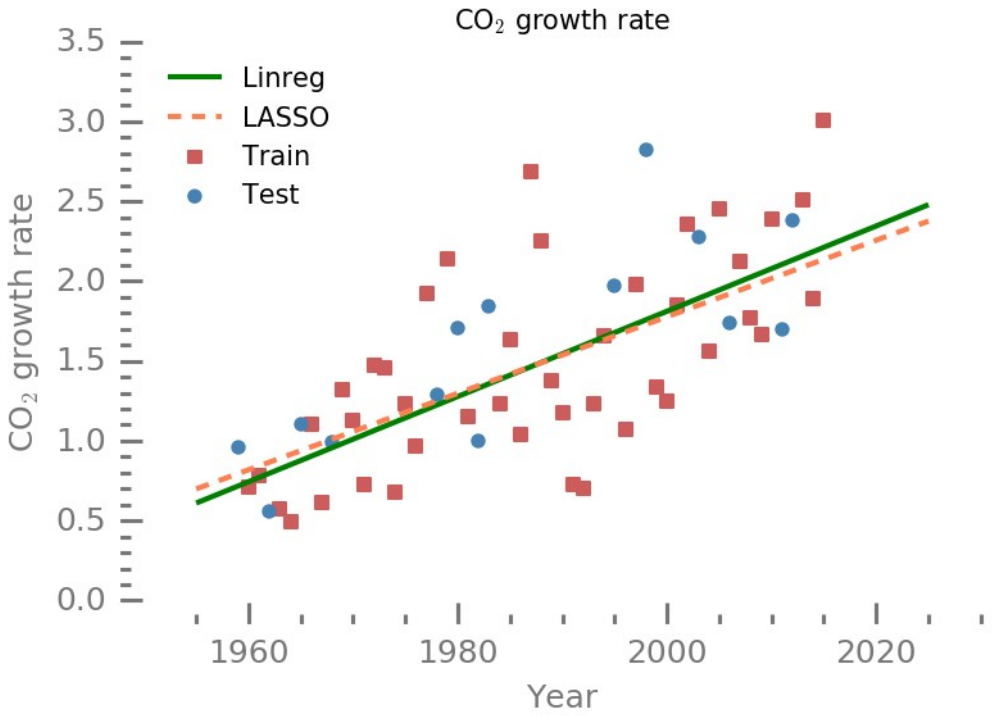

```
{'intercept': {'95% HPD interval': array([-50.64990877, -49.42168758]),
  'mc error': 0.011931753405578262,
  'mean': -50.008066085637815,
  'n': 4500,
  'quantiles': {2.5: -50.646760270081515,
   25: -50.228993605017173,
   50: -50.004119841092979,
   75: -49.789140434449351,
   97.5: -49.412790669964195},
  'standard deviation': 0.31970980437231761},
 'slope': {'95% HPD interval': array([ 0.02560229,  0.02622579]),
  'mc error': 6.0141748688389682e-06,
  'mean': 0.0259157602742307,
  'n': 4500,
  'quantiles': {2.5: 0.025609913276748567,
   25: 0.025803373826353462,
   50: 0.02591258446734945,
   75: 0.026027359901668552,
   97.5: 0.026237852465415799},
  'standard deviation': 0.00016186303140381465}}
```

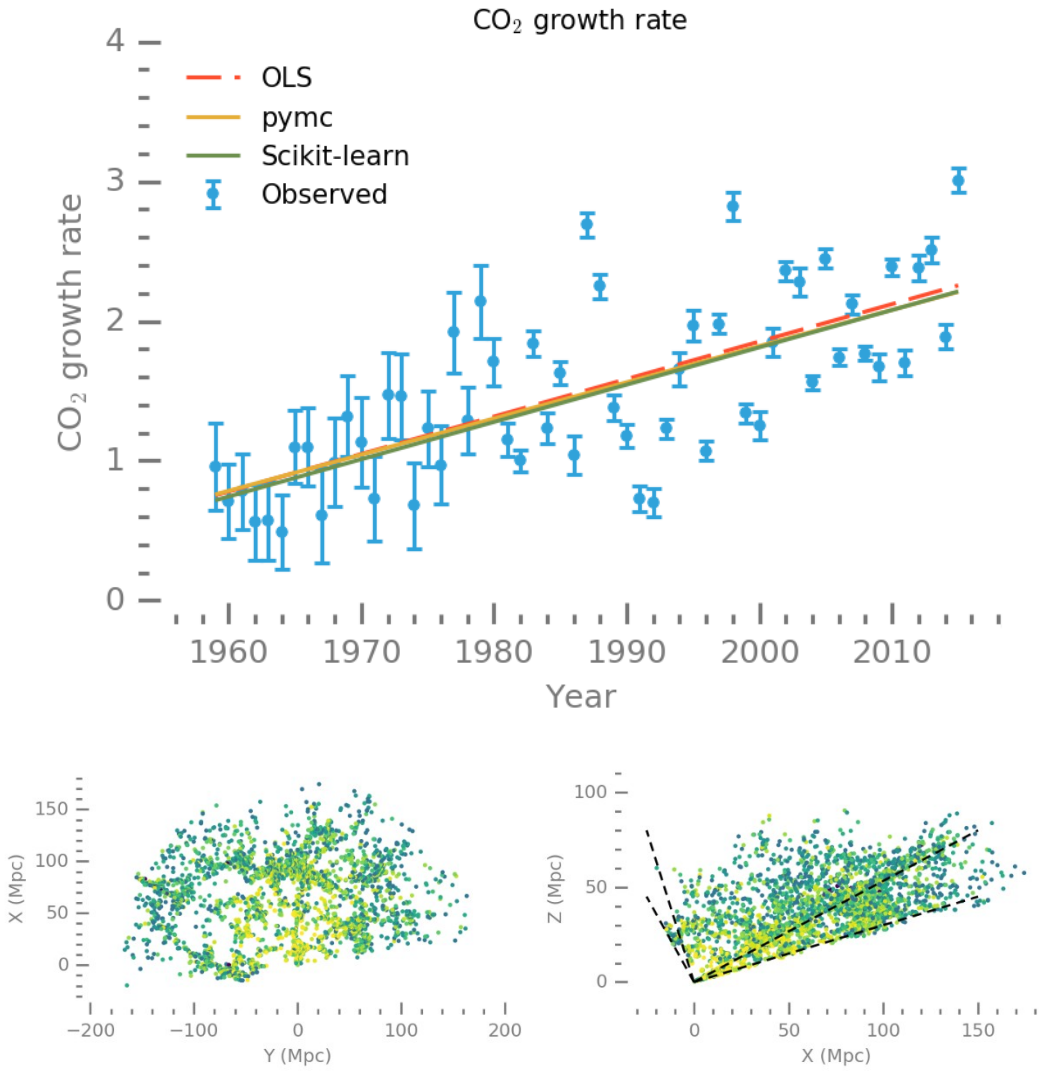
# Chapter 7: Supervised and Unsupervised Learning



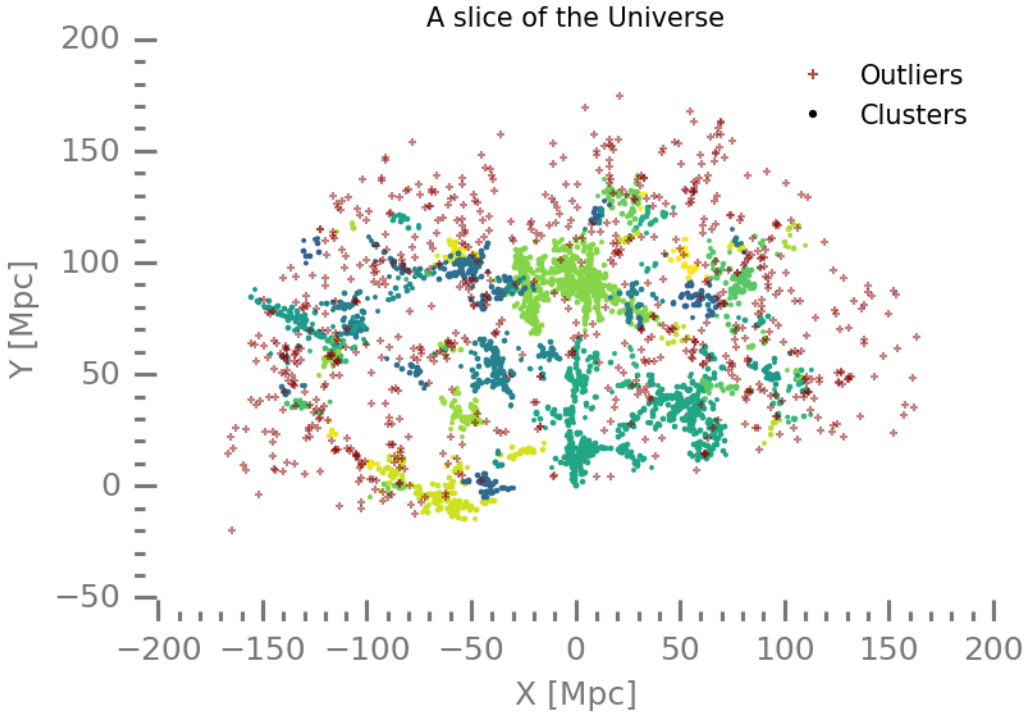

```
++++++ LinearRegression ++++++
Slope: 0.027 Intercept:-51.60
Mean squared residuals: 0.17
Variance score: 0.56
++++++ Lasso ++++++
Slope: 0.024 Intercept:-46.16
Mean squared residuals: 0.17
Variance score: 0.56
```




```
LinearSVC: 0.49
LASSO:     0.48

      Slope  Intercept
ML : 0.027 -51.597
OLS: 0.027 -51.812
Bay: 0.026 -49.998
```

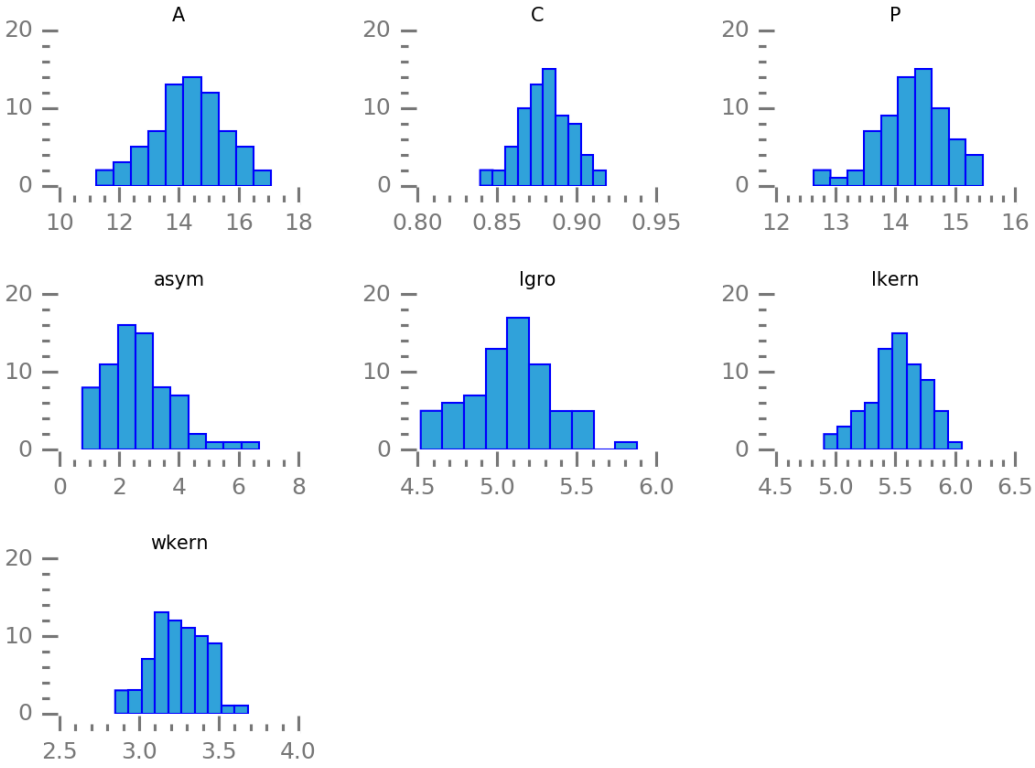


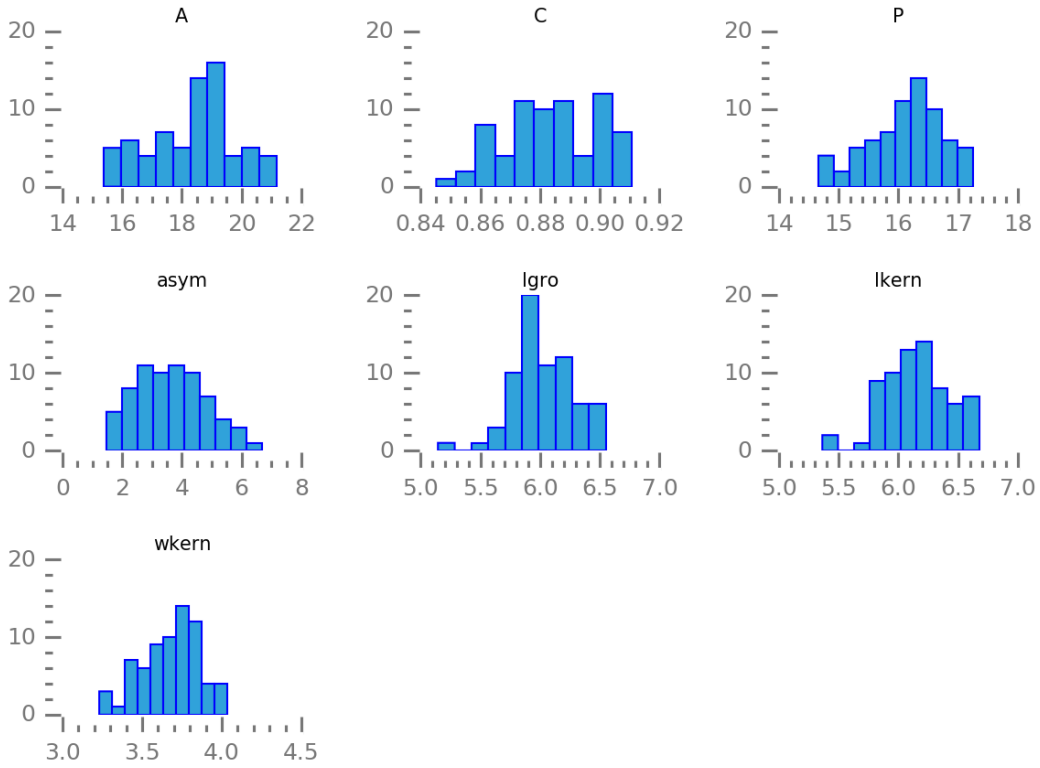

```
Estimated number of clusters: 8
Silhouette Coefficient: -0.143
```

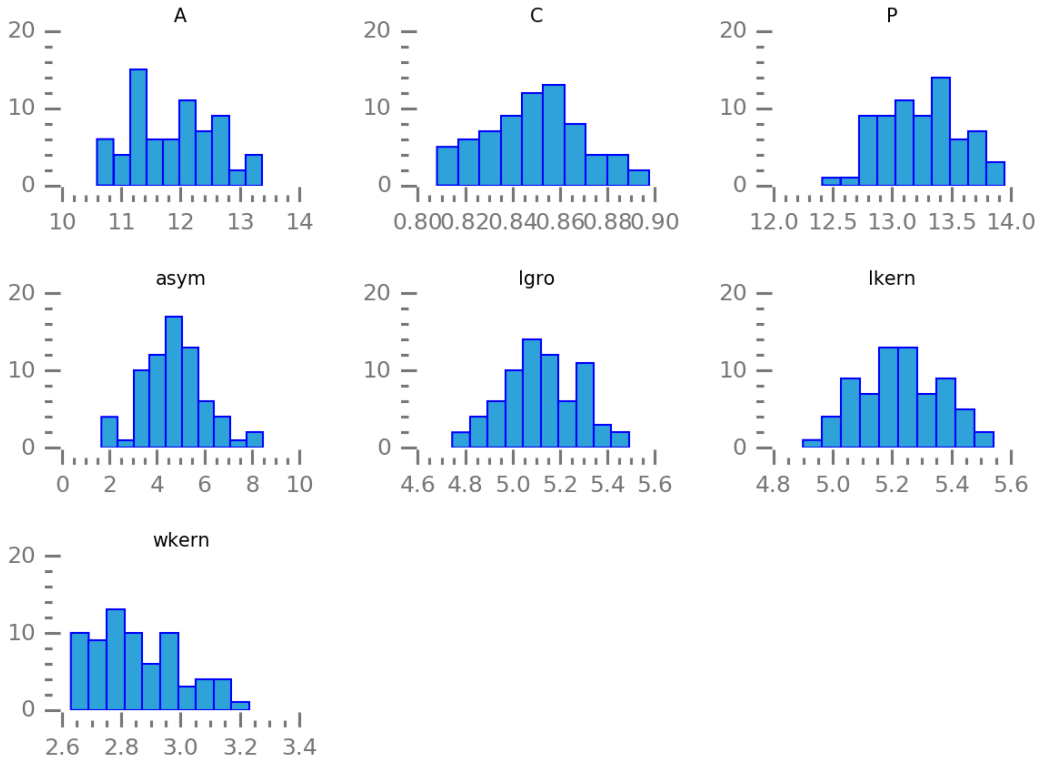


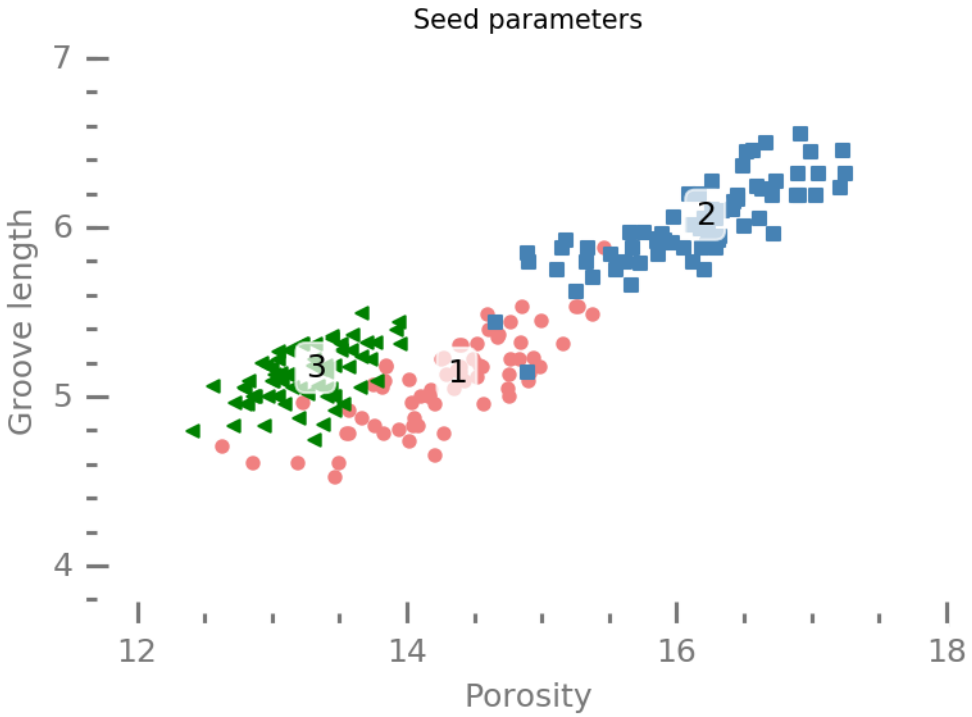

1. area A,
2. perimeter P,
3. compactness C = 4*pi*A/P^2,
4. length of kernel,
5. width of kernel,
6. asymmetry coefficient
7. length of kernel groove.
8. group

| | A | P | C | lkern | wkern | asym | lgro | gr |
|---|---|---|---|---|---|---|---|---|
| **0** | 15.26 | 14.84 | 0.8710 | 5.763 | 3.312 | 2.221 | 5.220 | 1 |
| **1** | 14.88 | 14.57 | 0.8811 | 5.554 | 3.333 | 1.018 | 4.956 | 1 |
| **2** | 14.29 | 14.09 | 0.9050 | 5.291 | 3.337 | 2.699 | 4.825 | 1 |
| **3** | 13.84 | 13.94 | 0.8955 | 5.324 | 3.379 | 2.259 | 4.805 | 1 |
| **4** | 16.14 | 14.99 | 0.9034 | 5.658 | 3.562 | 1.355 | 5.175 | 1 |



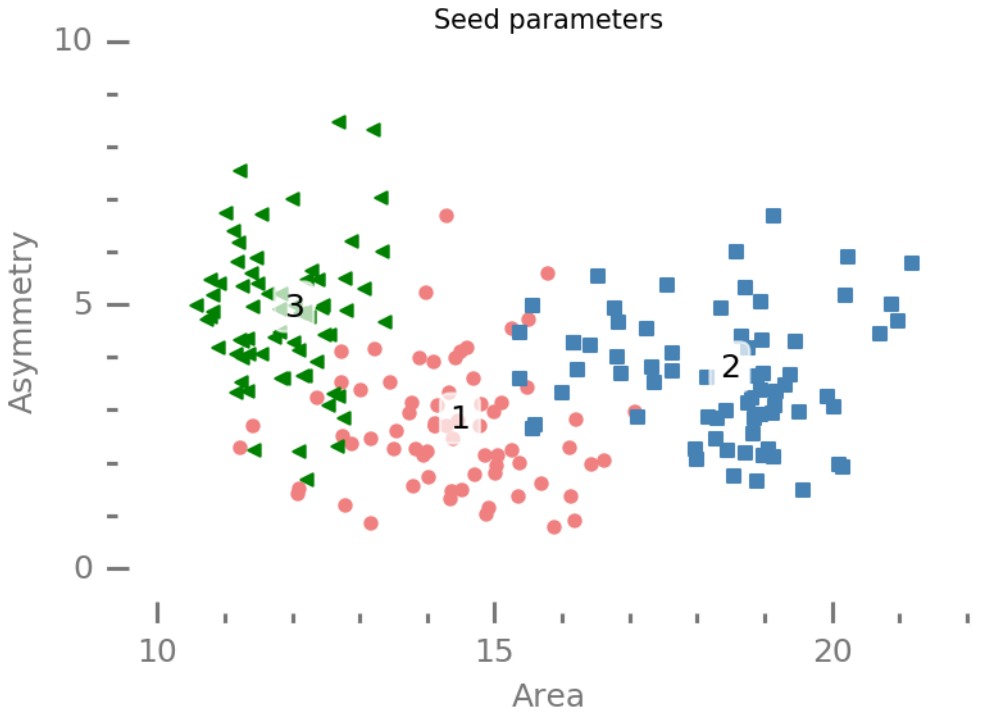

A
20
10
0
14
16
18
20
22
C
20
10
0
0.84
0.86
0.88
0.90
0.92
P
20
10
0
14
15
16
17
18
asym
20
10
0
0
2
4
6
8
lgro
20
10
0
5.0
5.5
6.0
6.5
7.0
lkern
20
10
0
5.0
5.5
6.0
6.5
7.0
wkern
20
10
0
3.0
3.5
4.0
4.5

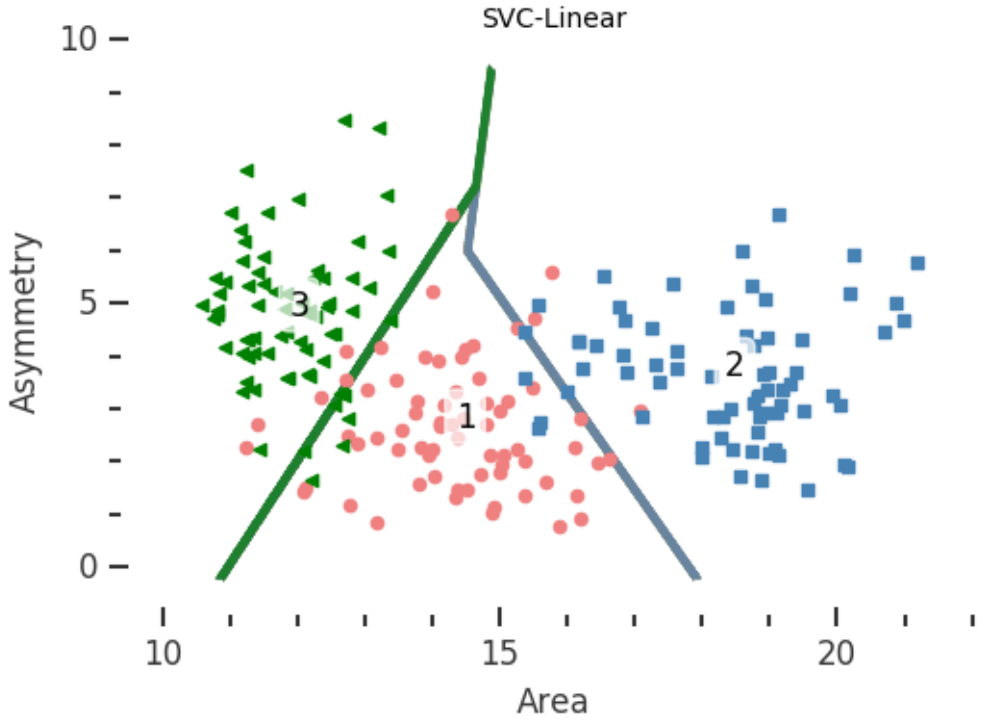

A
20
10
0
10 11 12 13 14
C
20
10
0
0.80 0.82 0.84 0.86 0.88 0.90
P
20
10
0
12.0 12.5 13.0 13.5 14.0
asym
20
10
0
0 2 4 6 8 10
lgro
20
10
0
4.6 4.8 5.0 5.2 5.4 5.6
lkern
20
10
0
4.8 5.0 5.2 5.4 5.6
wkern
20
10
0
2.6 2.8 3.0 3.2 3.4

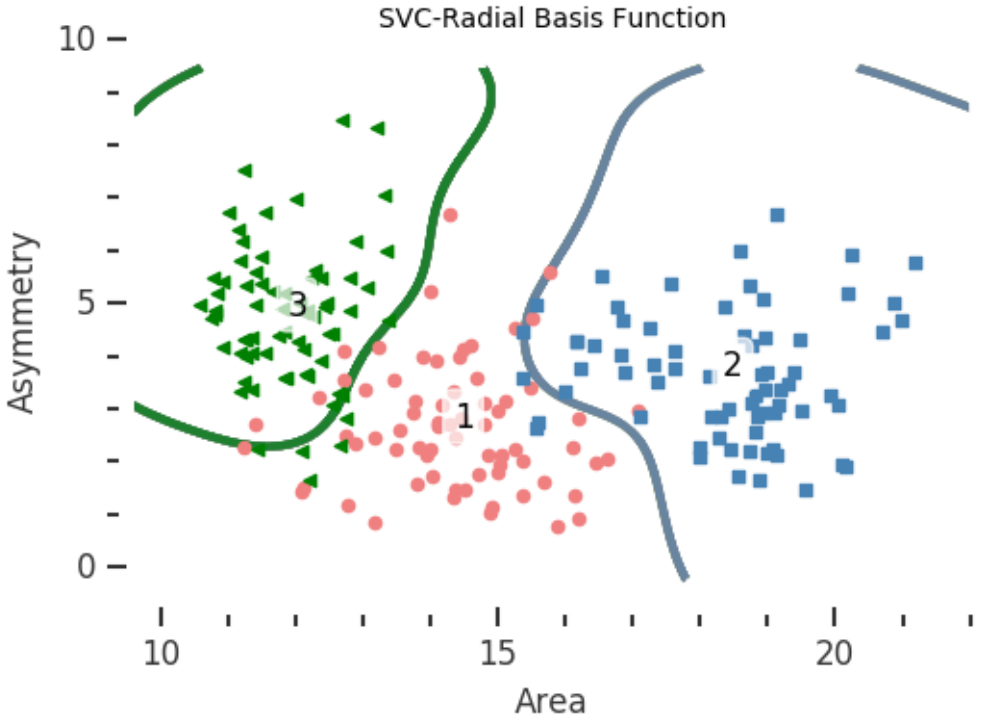


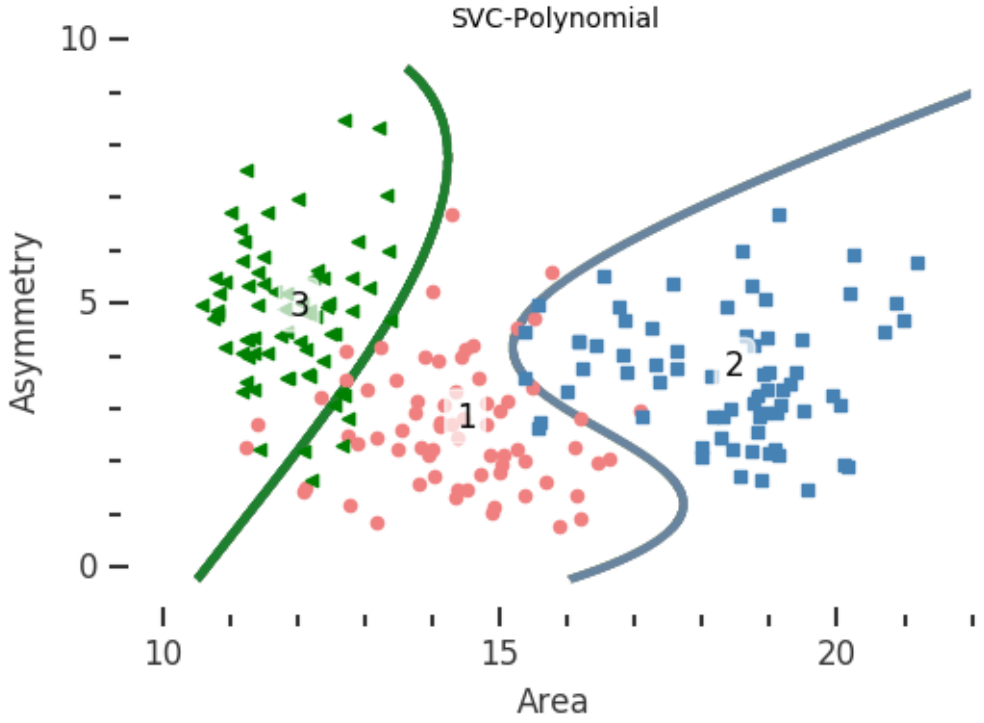

```
array([[ 15.26 ,   2.221],
       [ 14.88 ,   1.018],
       [ 14.29 ,   2.699],
       [ 13.84 ,   2.259],
       [ 16.14 ,   1.355]])
```

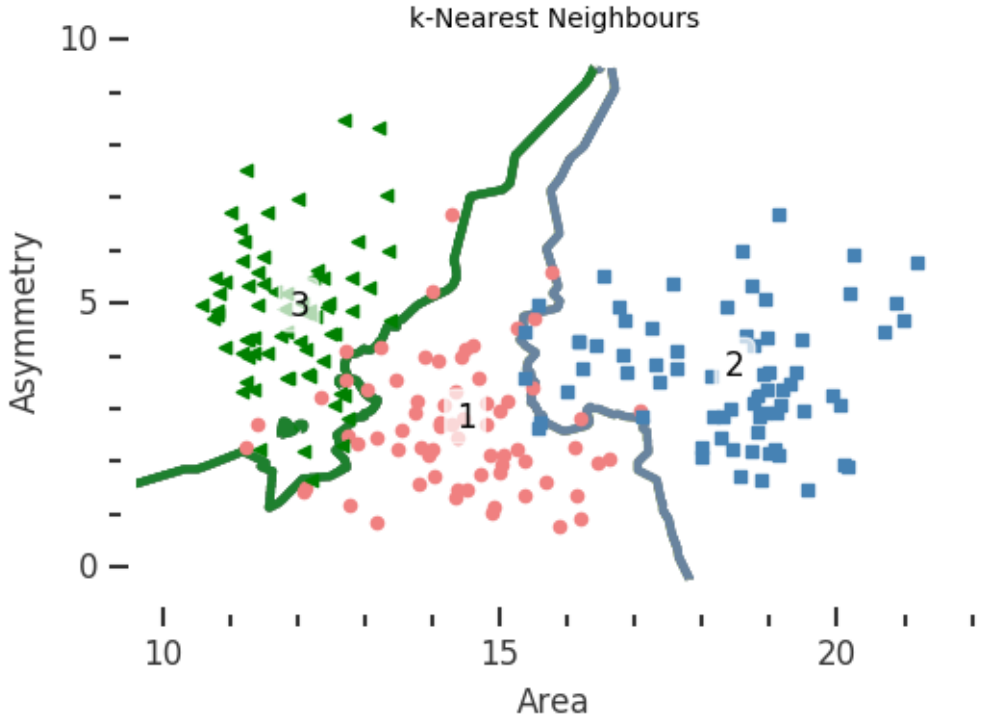
| | A | P | C | lkern | wkern | asym | lgro | gr |
|---|---|---|---|---|---|---|---|---|
| **0** | 15.26 | 14.84 | 0.8710 | 5.763 | 3.312 | 2.221 | 5.220 | 1 |
| **1** | 14.88 | 14.57 | 0.8811 | 5.554 | 3.333 | 1.018 | 4.956 | 1 |
| **2** | 14.29 | 14.09 | 0.9050 | 5.291 | 3.337 | 2.699 | 4.825 | 1 |
| **3** | 13.84 | 13.94 | 0.8955 | 5.324 | 3.379 | 2.259 | 4.805 | 1 |
| **4** | 16.14 | 14.99 | 0.9034 | 5.658 | 3.562 | 1.355 | 5.175 | 1 |

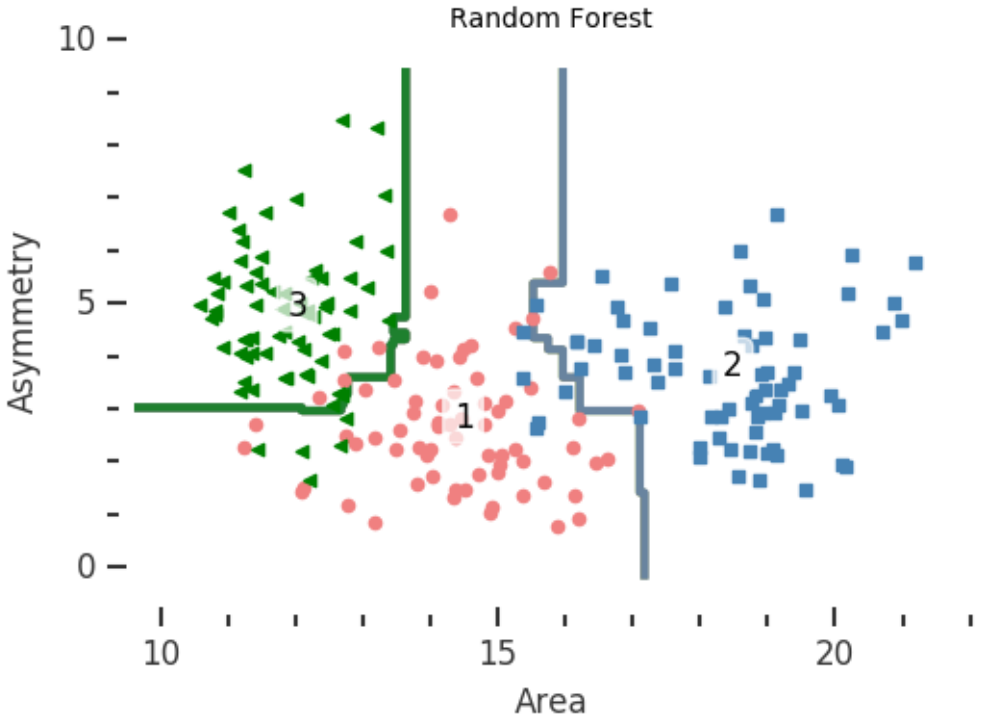

Seed parameters
10
5
0
Asymmetry
3
1
2
10
15
20
Area


SVC-Linear
10
5
0
Asymmetry
3
1
2
10
15
20
Area


SVC-Radial Basis Function
10
5
0
Asymmetry
3
1
2
10
15
20
Area


SVC-Polynomial
10
5
0
Asymmetry
3
1
2
10
15
20
Area

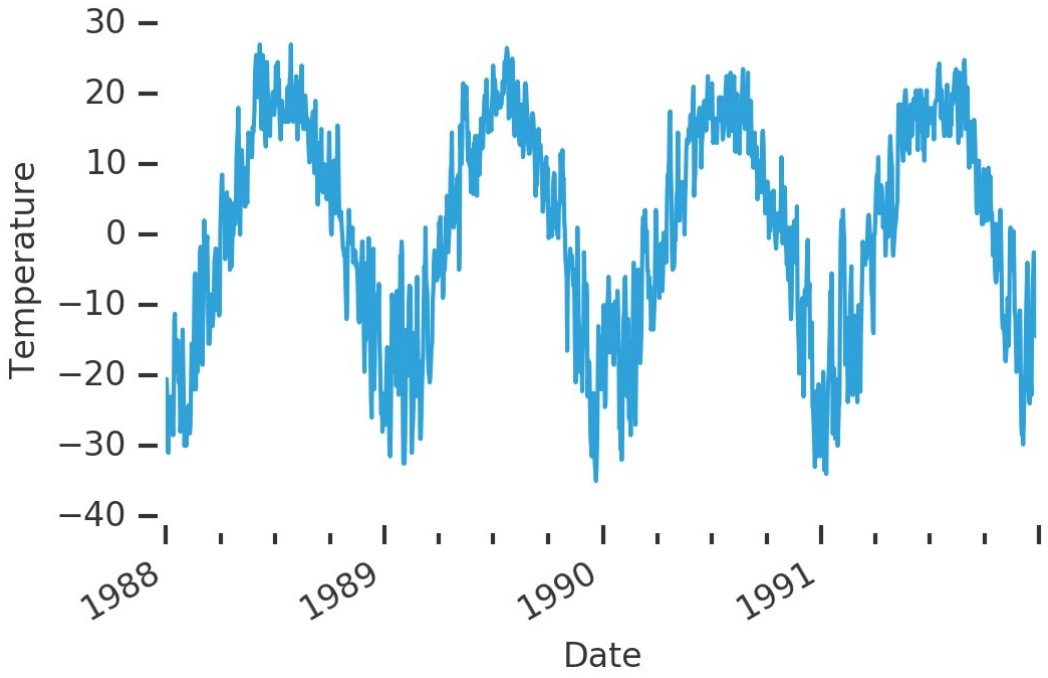

k-Nearest Neighbours
10
5
0
Asymmetry
3
1
2
10
15
20
Area


Random Forest
10
5
0
Asymmetry
3
1
2
10
15
20
Area

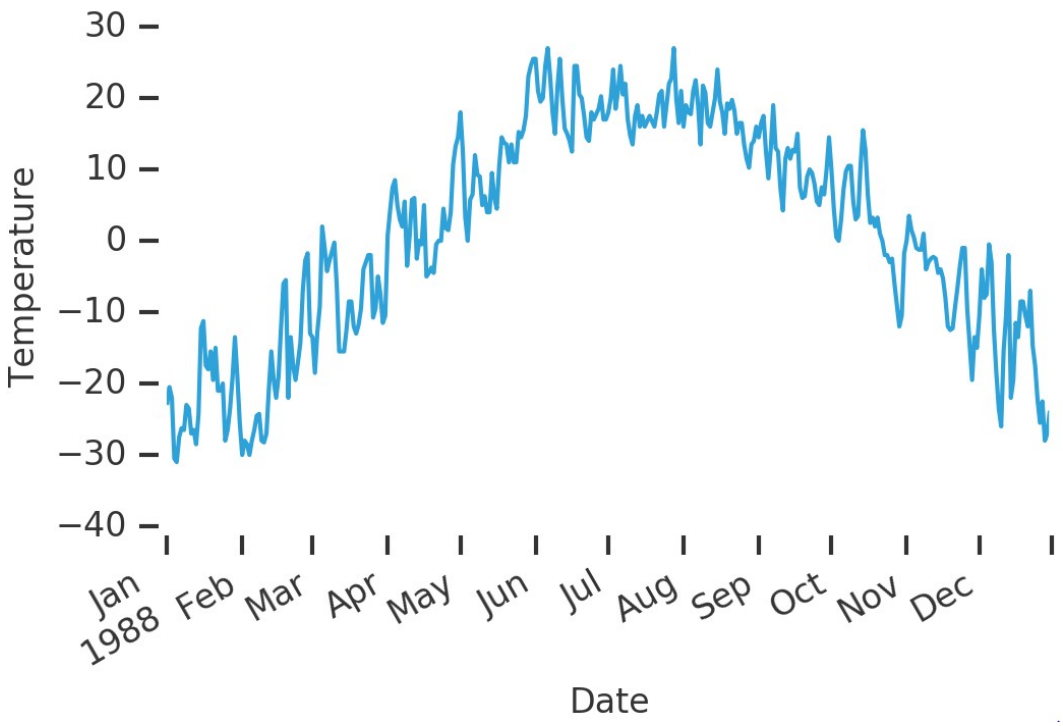
# Chapter 8: Time Series Analysis

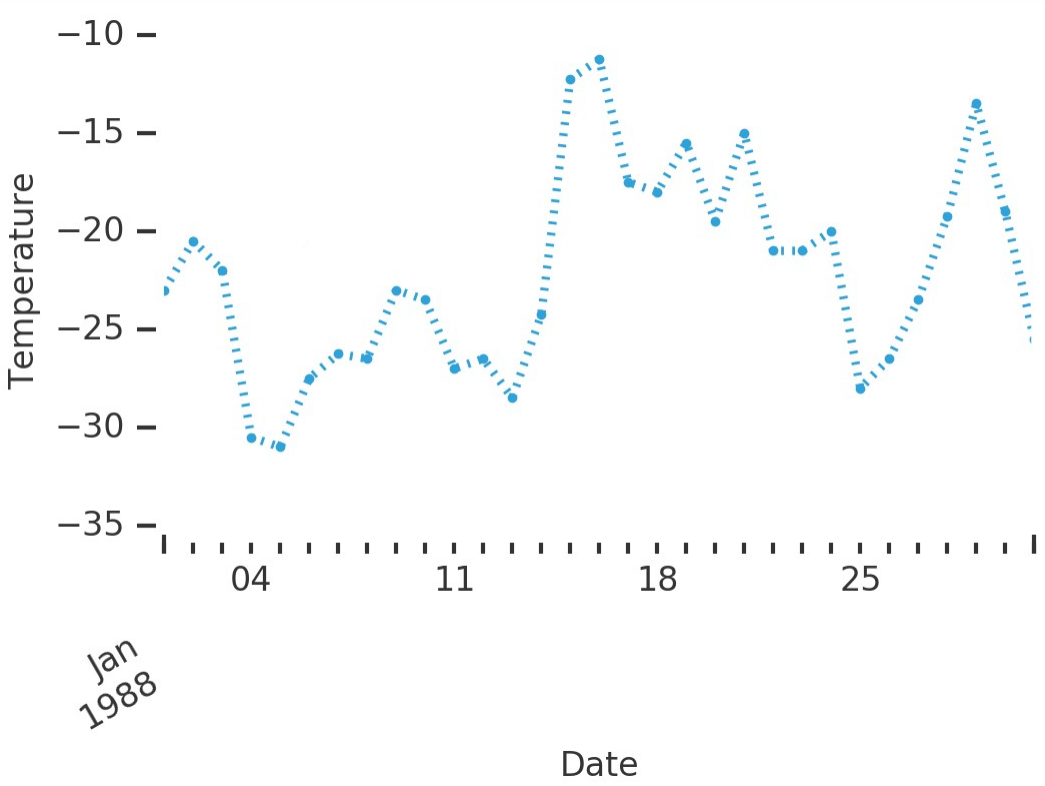
| | Temp |
|---|---|
| **Date** | |
| **1988-01-01** | -23.0 |
| **1988-01-02** | -20.5 |
| **1988-01-03** | -22.0 |
| **1988-01-04** | -30.5 |
| **1988-01-05** | -31.0 |

```
DatetimeIndex(['1988-01-01', '1988-01-02', '1988-01-03', '1988-01-04',
               '1988-01-05', '1988-01-06', '1988-01-07', '1988-01-08',
               '1988-01-09', '1988-01-10',
               ...
               '1991-12-22', '1991-12-23', '1991-12-24', '1991-12-25',
               '1991-12-26', '1991-12-27', '1991-12-28', '1991-12-29',
               '1991-12-30', '1991-12-31'],
              dtype='datetime64[ns]', name='Date', length=1461, freq=None)
```

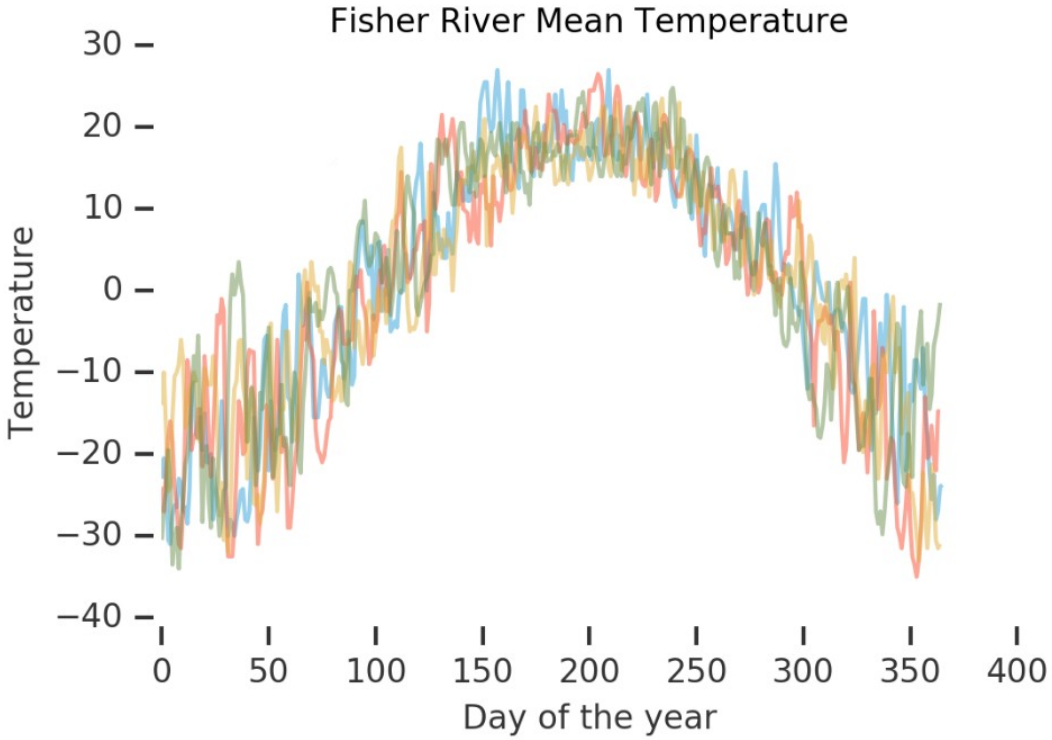

30
20
10
0
−10
−20
−30
−40
Temperature
1988
1989
1990
1991
Date

```
count    1461.000000
mean        0.803320
std        15.154634
min       -35.000000
25%       -11.250000
50%         2.000000
75%        14.500000
max        27.000000
Name: Temp, dtype: float64
```

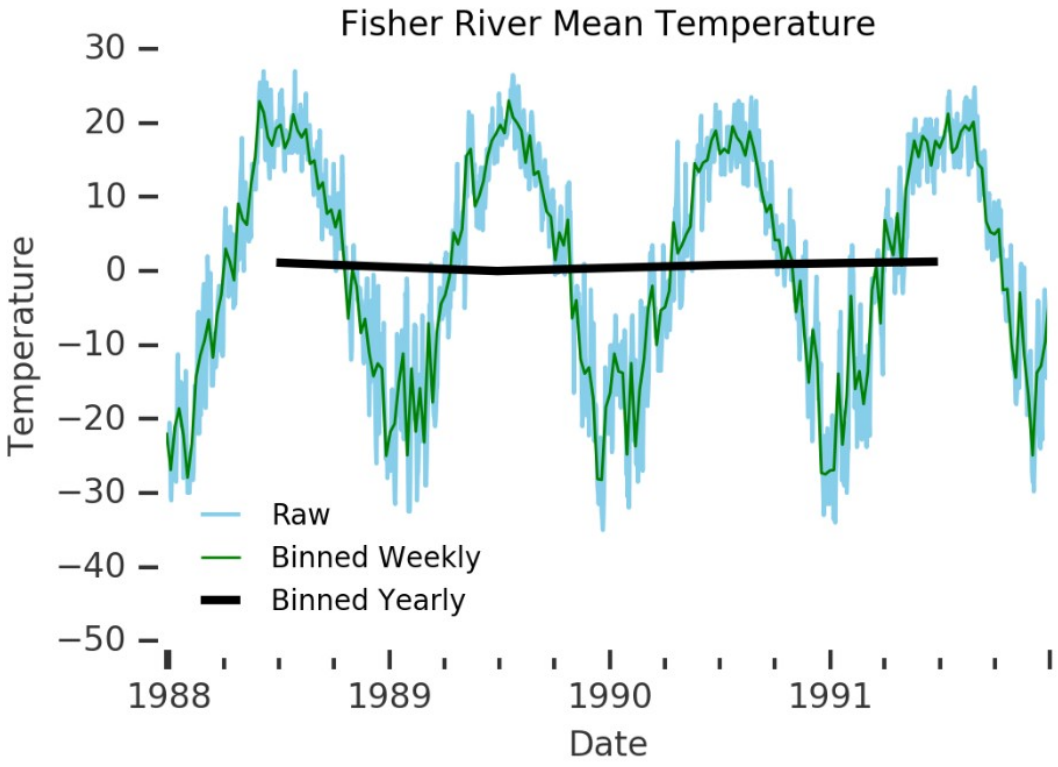

30
20
10
0
−10
−20
−30
−40
Temperature
Jan
1988
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Date

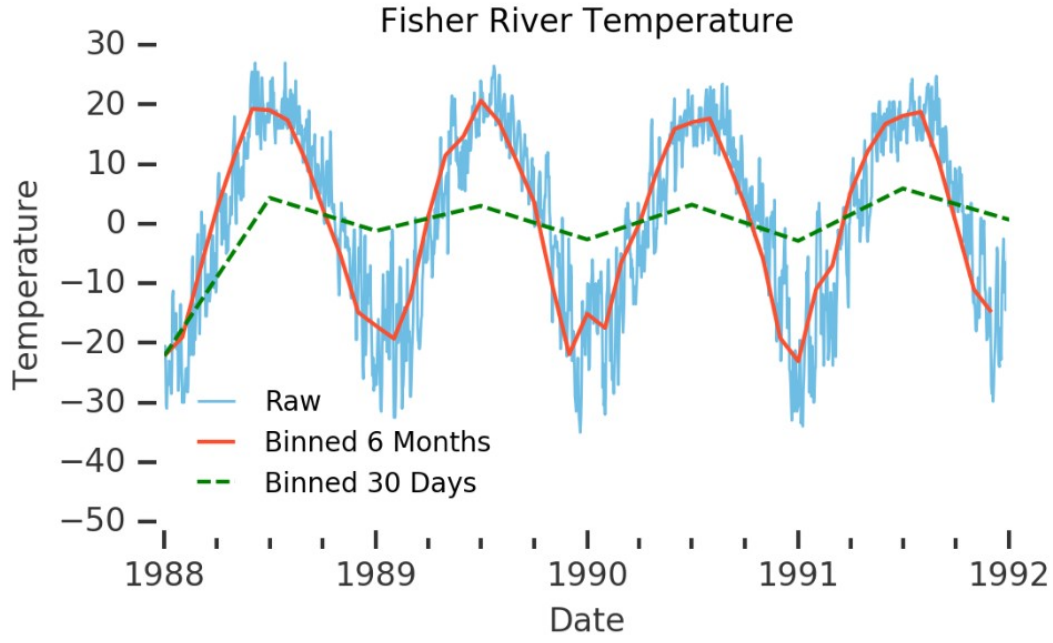


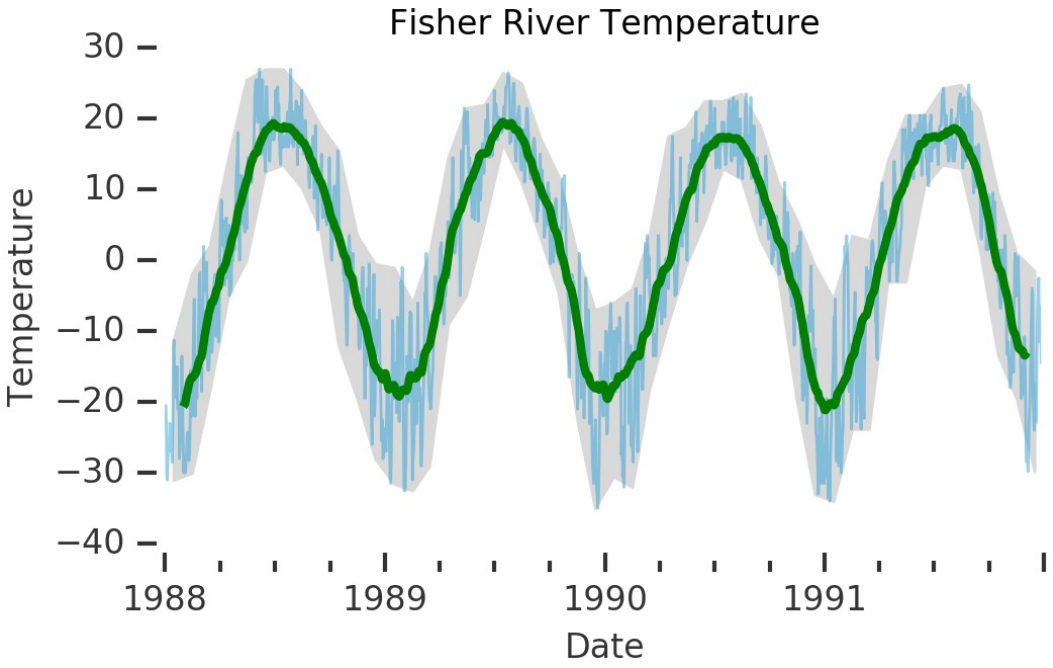

```
Date
1988-01-04   -30.50
1988-01-05   -31.00
1988-01-06   -27.50
1988-01-07   -26.25
1988-01-08   -26.50
Name: Temp, dtype: float64
```

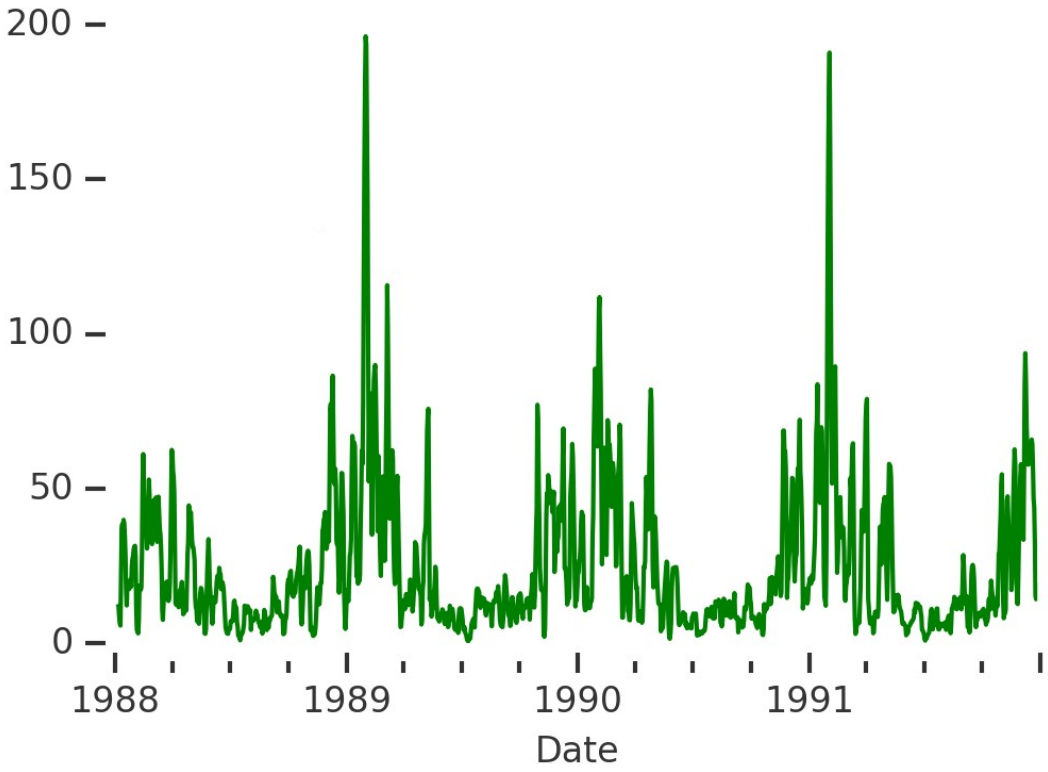


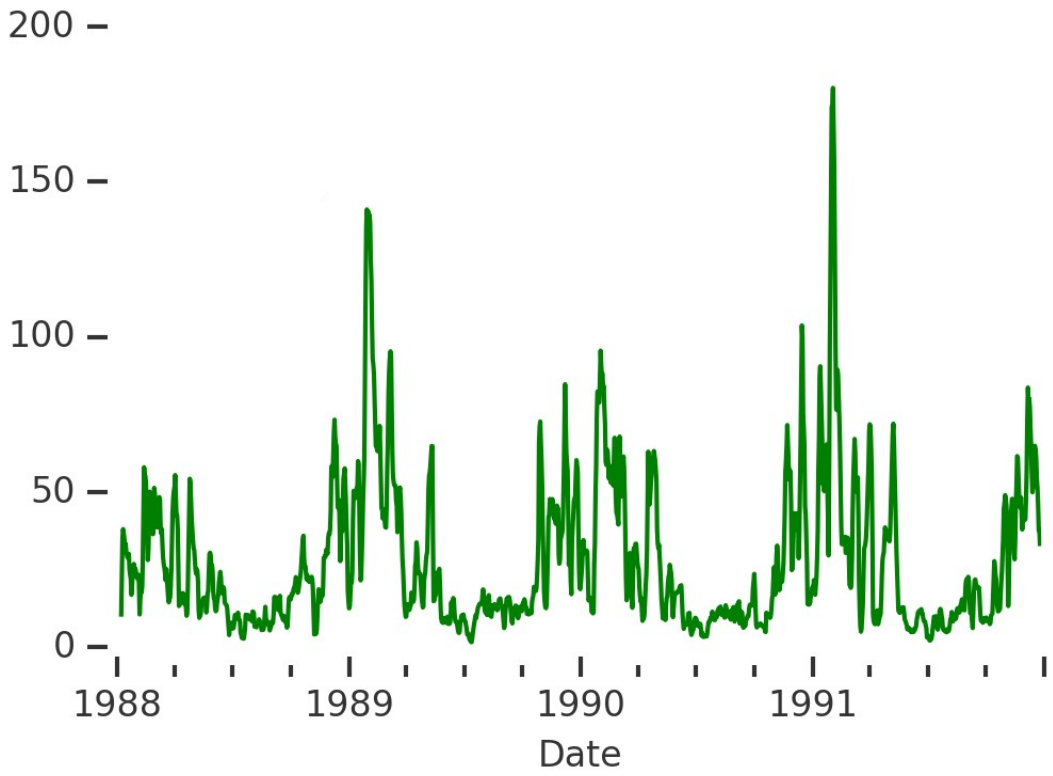

```
Date
1988-12-31    1.138661
1989-12-31   -0.006164
1990-12-31    0.815753
1991-12-31    1.264110
Freq: A-DEC, Name: Temp, dtype: float64
```

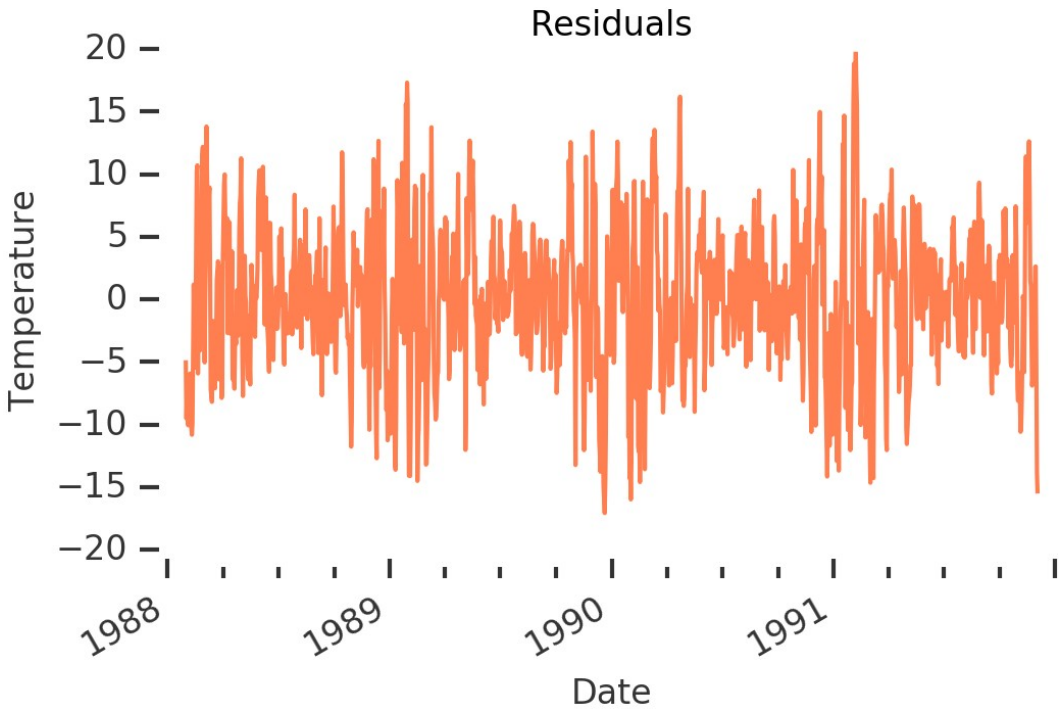

Fisher River Mean Temperature
30
20
10
0
−10
−20
−30
−40
−50
Temperature
Raw
Binned Weekly
Binned Yearly
1988
1989
1990
1991
Date

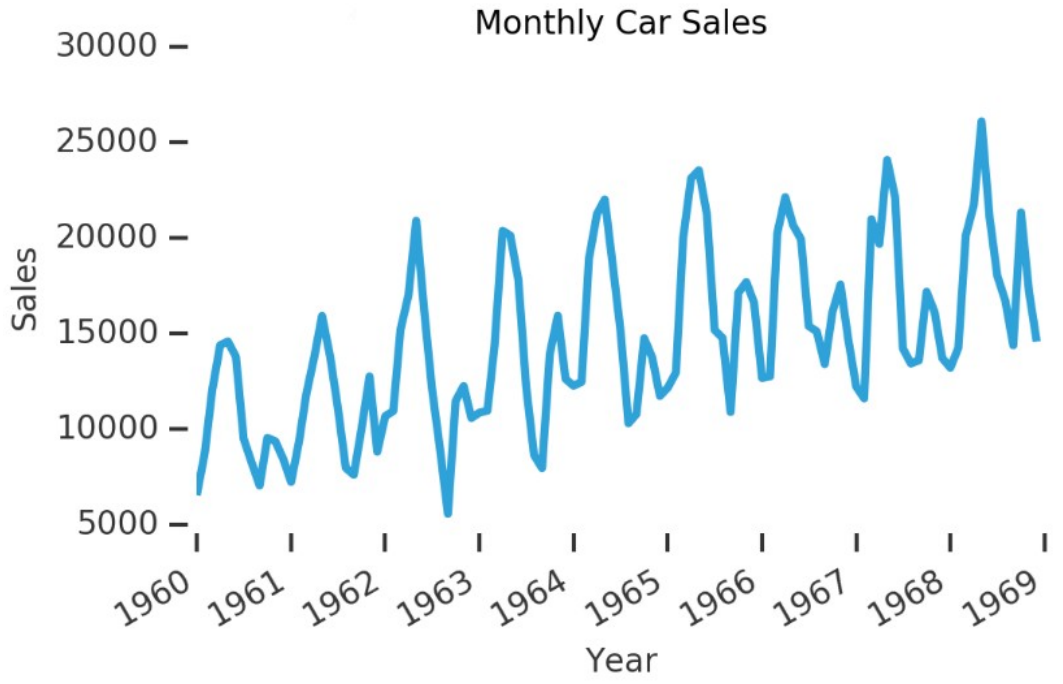

Fisher River Temperature
30
20
10
0
−10
−20
−30
−40
−50
Temperature
Raw
Binned 6 Months
Binned 30 Days
1988
1989
1990
1991
1992
Date


Fisher River Temperature
30
20
10
0
−10
−20
−30
−40
Temperature
1988
1989
1990
1991
Date

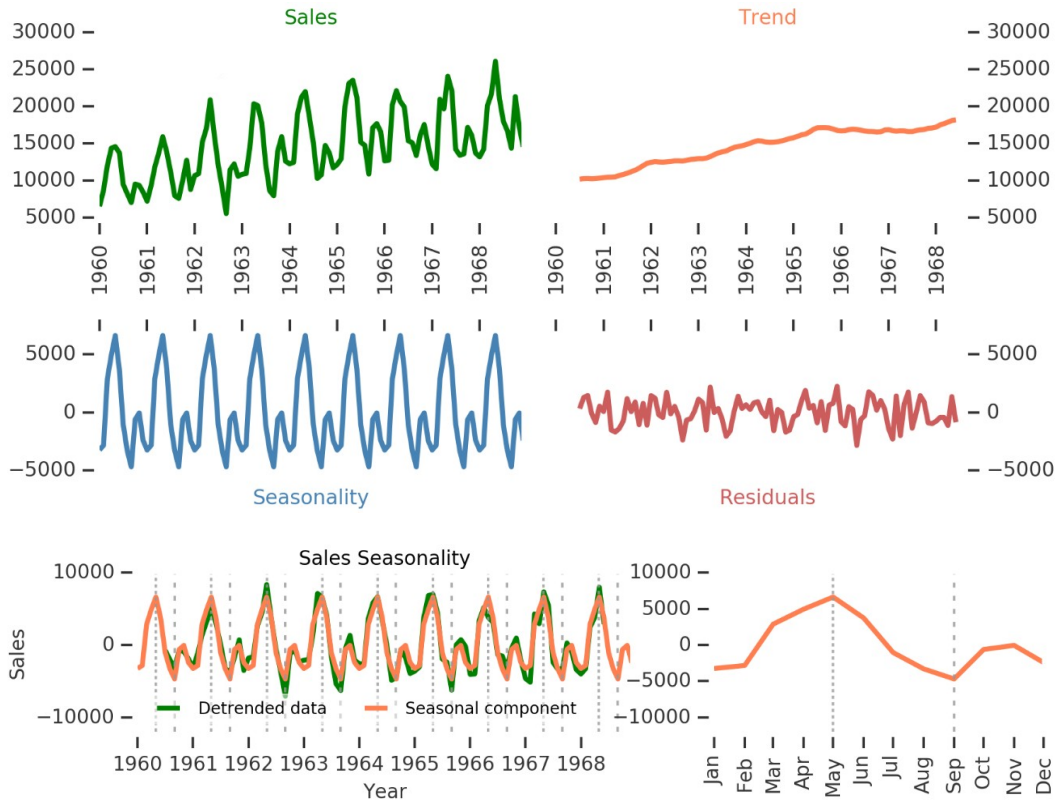

200
150
100
50
0
1988
1989
1990
1991
Date


200
150
100
50
0
1988
1989
1990
1991
Date

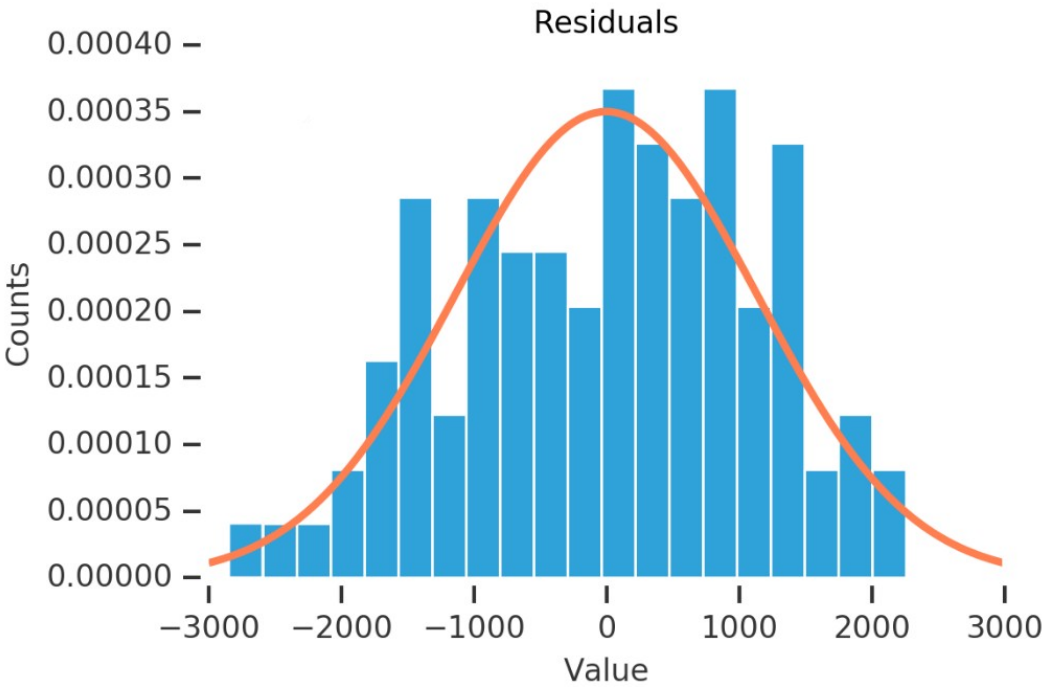

Residuals
Temperature
20
15
10
5
0
−5
−10
−15
−20
1988
1989
1990
1991
Date

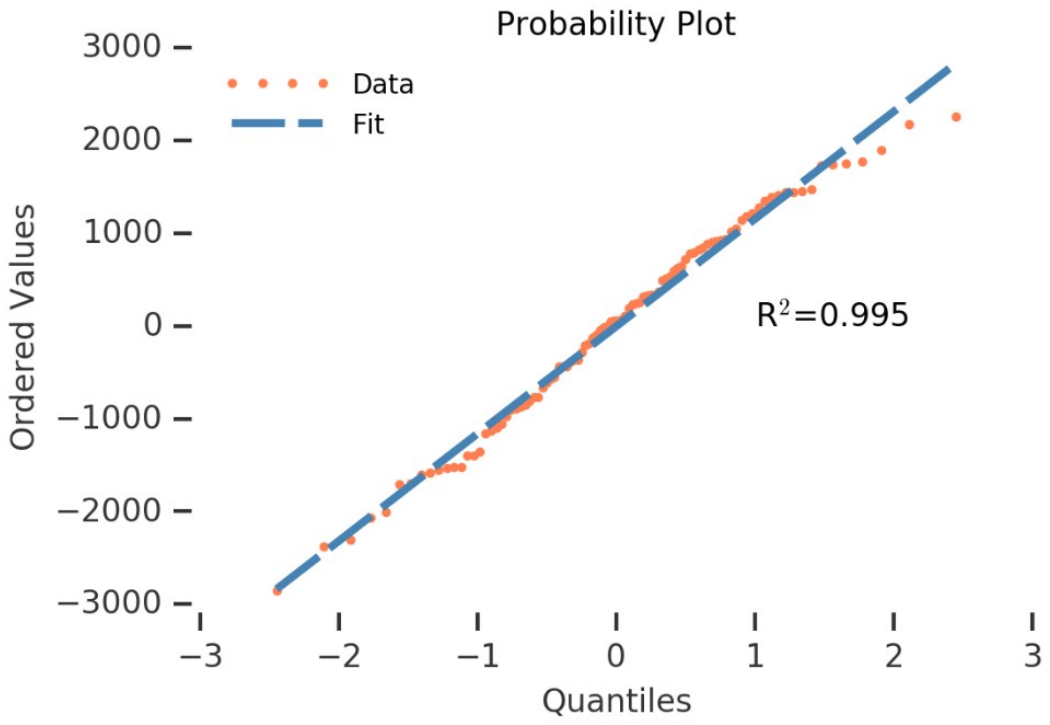

Monthly Car Sales
Sales
30000
25000
20000
15000
10000
5000
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
Year

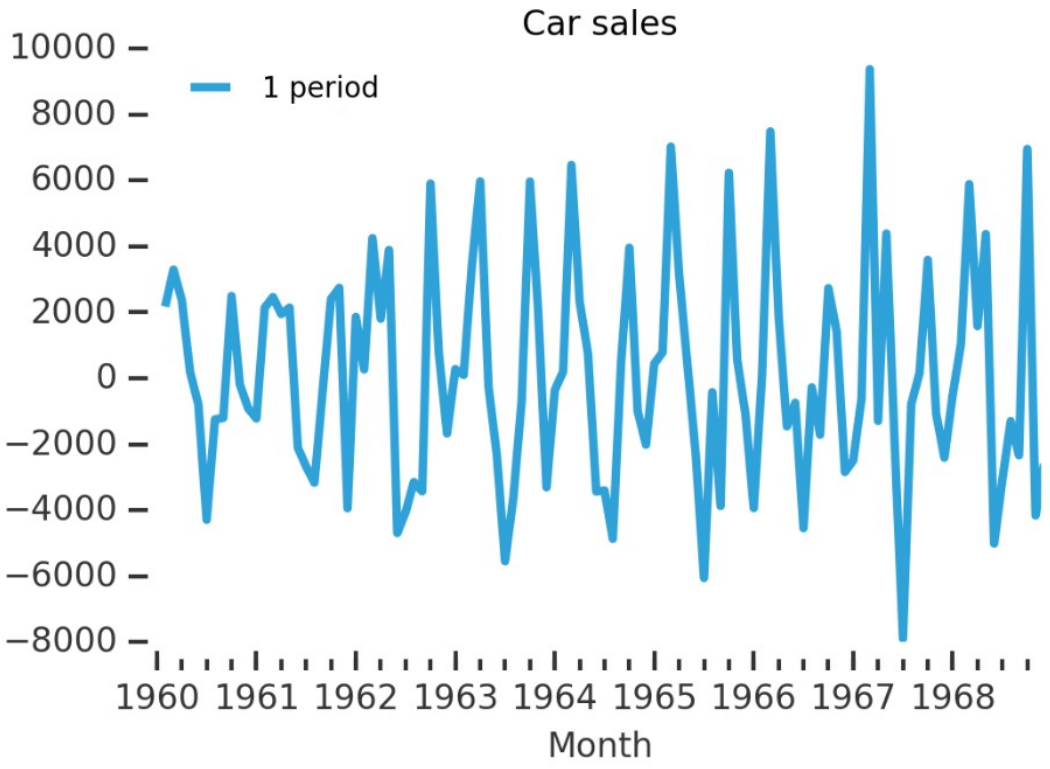
```
adf               -1.673
p                 0.763
crit. val.        1%: -4.060, 5%: -3.459, 10%: -3.155
stationary?       false
```



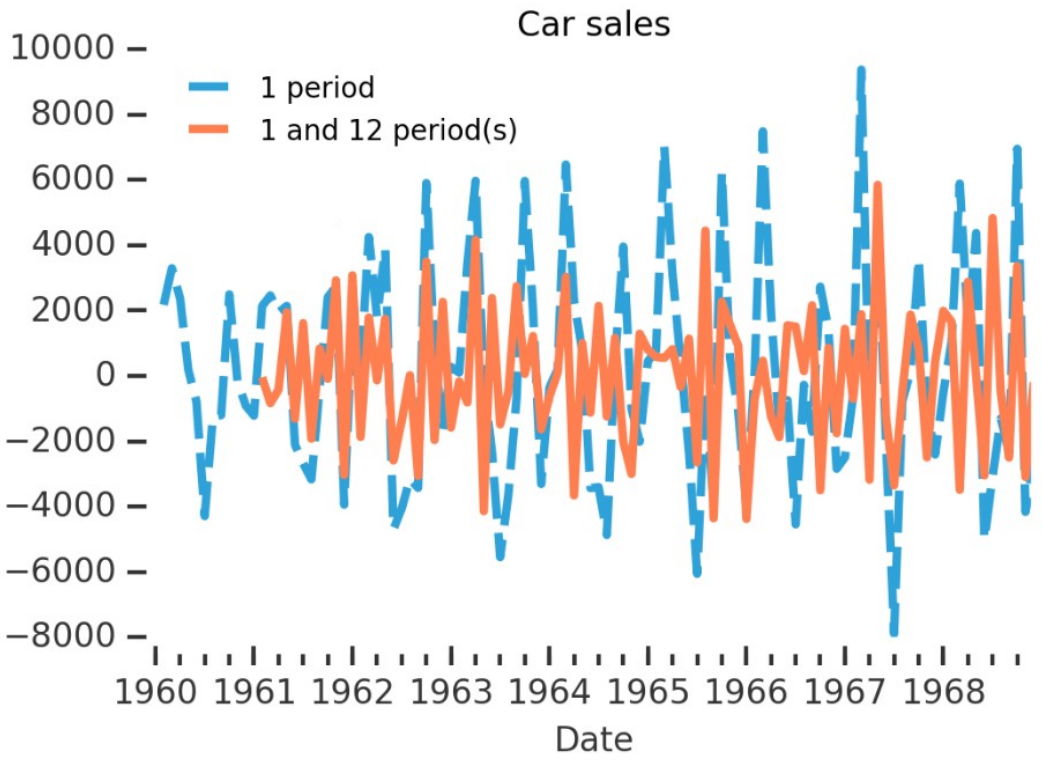

```
adf               -4.501
p                 0.0015
crit. val.        1%: -4.072, 5%: -3.465, 10%: -3.159
stationary?       true
```


Residuals
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
Counts
−3000
−2000
−1000
0
1000
2000
3000
Value

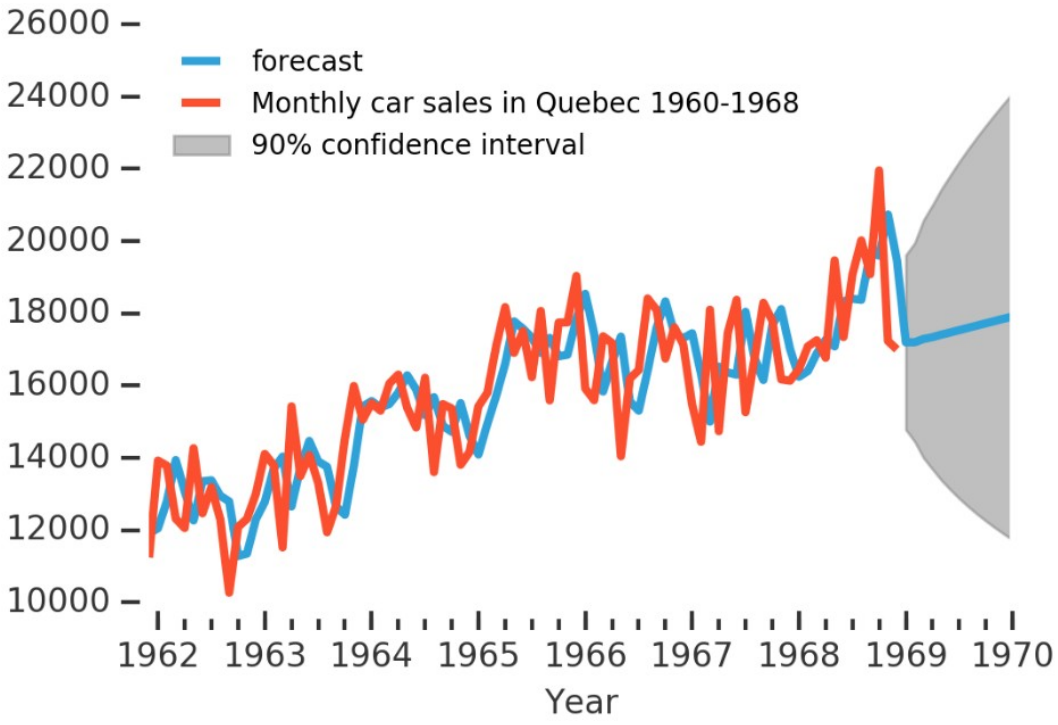

Probability Plot
Data
Fit
$R^2$=0.995
Ordered Values
Quantiles
3000
2000
1000
0
−1000
−2000
−3000
−3
−2
−1
0
1
2
3

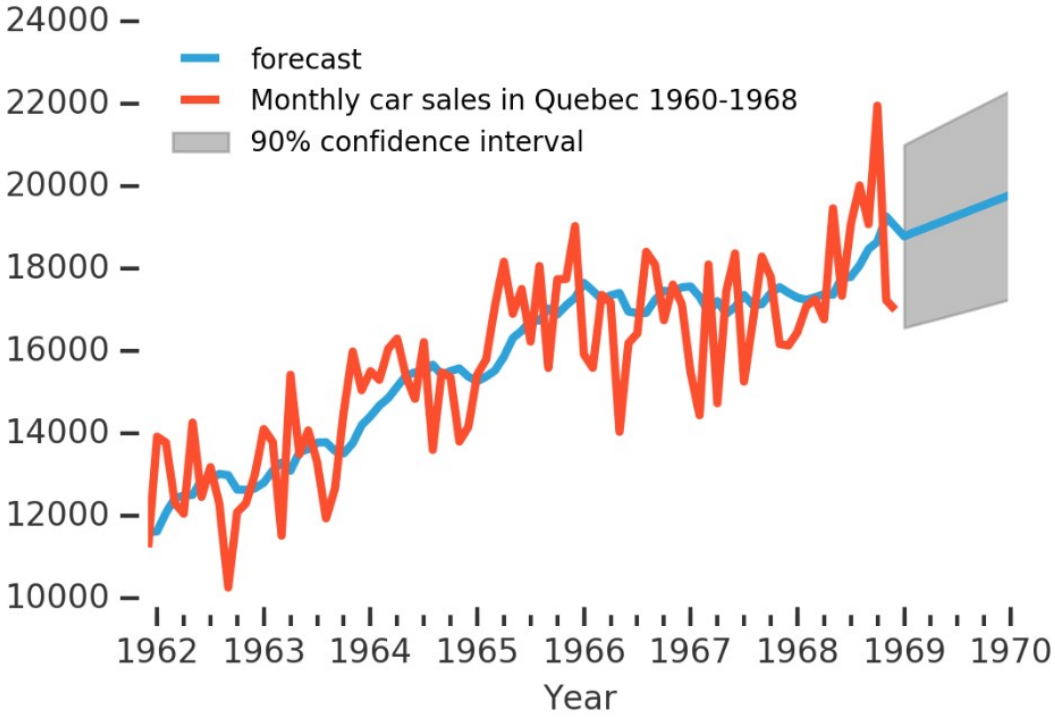



| | |
|---|---|
| adf | -3.124 |
| p | 0.101 |
| crit. val. | 1%: -4.061, 5%: -3.459, 10%: -3.156 |
| stationary? | false |

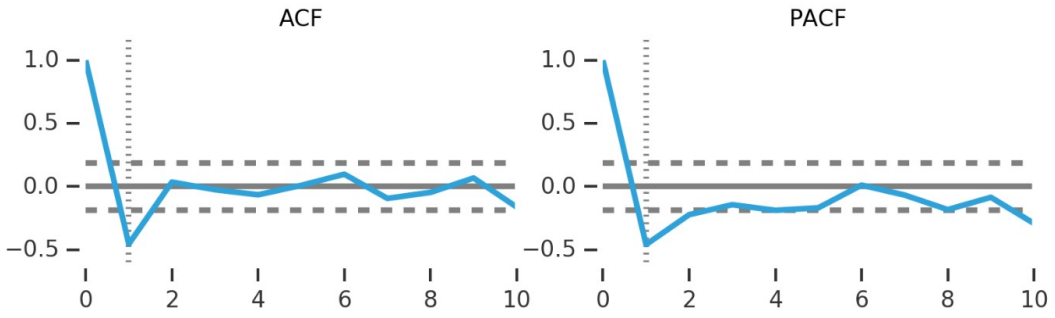


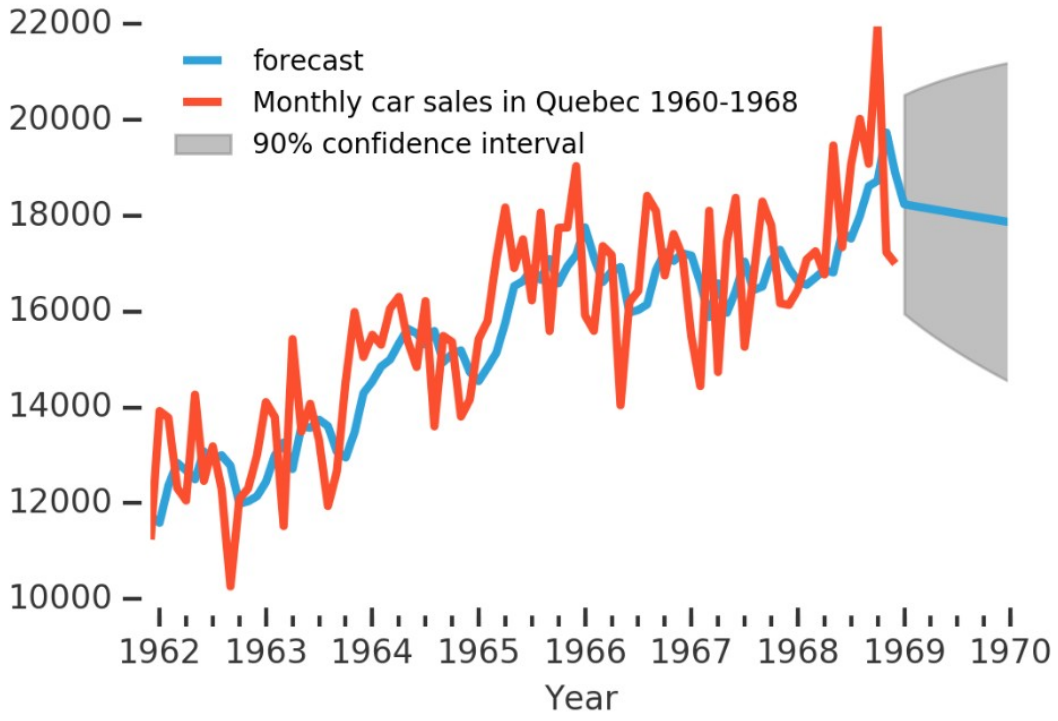

```
adf              -3.875
p                0.0131
crit. val.       1%: -4.077, 5%: -3.467, 10%: -3.160
stationary?      true

adf              -3.611
p                0.0289
crit. val.       1%: -4.061, 5%: -3.459, 10%: -3.156
stationary?      true
```


26000
24000
22000
20000
18000
16000
14000
12000
10000
forecast
Monthly car sales in Quebec 1960-1968
90% confidence interval
1962
1963
1964
1965
1966
1967
1968
1969
1970
Year

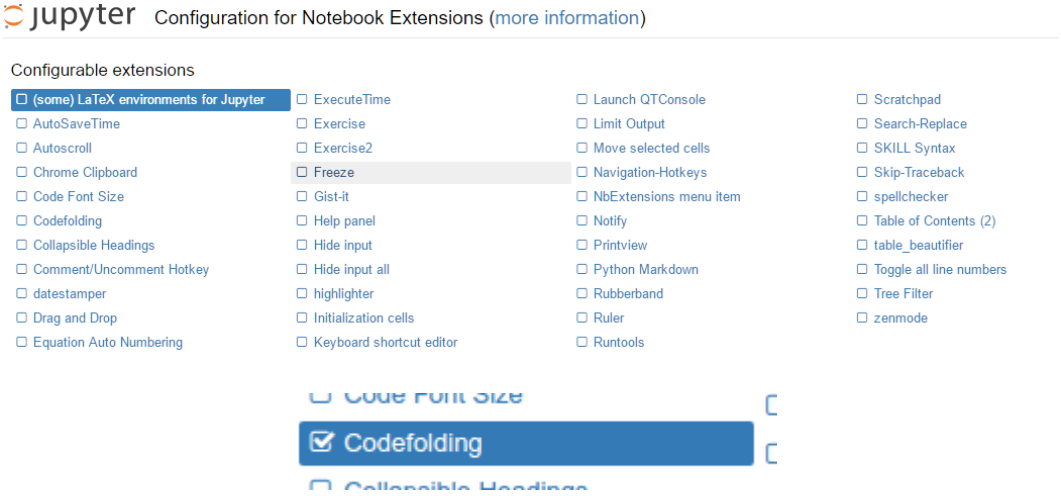



```
{'aic':              0            1            2
 0  1893.258581  1853.075312  1853.070371
 1  1870.333181  1852.567740  1853.880900
 2  1866.617420  1853.644132          NaN, 'aic_min_order': (1, 1)}
```



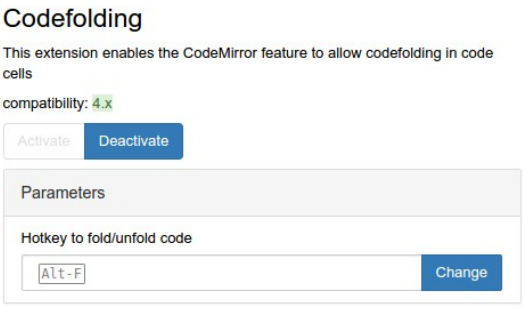

22000
20000
18000
16000
14000
12000
10000
forecast
Monthly car sales in Quebec 1960-1968
90% confidence interval
1962
1963
1964
1965
1966
1967
1968
1969
1970
Year

# Appendix: More on Jupyter Notebook and matplotlib Styles











/nbextensions/usability/codefolding/readme.md?v=20160501205159

This extension adds codefolding functionality from CodeMirror to a codecell.

After clicking on the gutter (left margin of codecell) or typing `Alt+F`, the code gets folded. See the examples below.
The folding status is saved in the cell metadata of the notebook, so reloading of a notebook will restore the folding view.

## Supported modes

Three different folding modes are supported:

```
def function_one(p1,p2):
    sel = [p1, p2][np.random.randint(2)]
    return sel
```

```
def function_one(p1,p2):↔
```

```
# it is possible to use comments
# to hide code as well
function_one(10,30)
```

```
# it is possible to use comments ↔
```

## Collapsible Headings

Allows notebook to have collapsible sections, separated by headings

compatibility: 4.x

Activate | Deactivate

Parameters

- [ ] Add a toolbar button to collapse the closest header cell

- [x] Add command-mode keyboard shortcuts to collapse/uncollapse the selected heading cell

Command-mode shortcut to collapse the selected heading cell

Left | Change

Command-mode shortcut to uncollapse (expand) the selected heading cell

Right | Change

**heading h2**

**collapsed heading h3**

# 1 A report

This document gives short examples on how to work with the Jupyter Notebook and some of the extensions that I hope that you have now activated. It also shows that you can actually write nice summaries of the analysis, perhaps not for the final version of a report, but for early drafts it should be ideal.

## 1.1 Some example text and code

In a Markdown cell you can for example

- Create headings
- Type normal text, just like any text editor and

# 1 A report

[...]

### Help panel

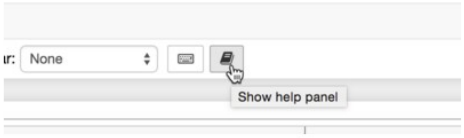
Add a toolbar button to display a help panel showing shortcuts to the right side of the notebook

compatibility: 3.x, 4.x

Activate Deactivate

Parameters

add a toolbar button to open the shortcuts dialog/panel

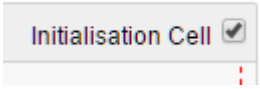



### Initialization cells

Add a cell toolbar selector to mark cells as 'initialization' cells. Such initialization cells are run on notebook load, or on clicking the provided button in the main toolbar

compatibility: 3.x, 4.x

Activate Deactivate

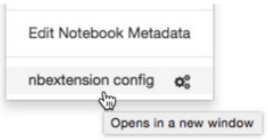



## NbExtensions menu item

Add an edit-menu item to open the NbExtensions config page

compatibility: 4.x

Activate | Deactivate

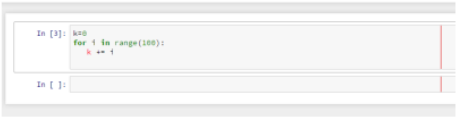



## Ruler

This extension enables the Ruler CodeMirror feature

compatibility: 4.x

Activate | Deactivate



Parameters

Column where ruler is displayed

78

Ruler color

Ruler style, e.g. solid, dashed

dashed

```
%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
In [33]: values = (1+np.array([0, 1.e-15]))*1.e27
         plt.plot(values)
```

```
C:\Users\[redacted]\Anaconda3\lib\site-packages\matplotlib\ticker.py in _set_form
at(self, vmin, vmax)
    608         thresh = 1e-3 * 10 ** loc_range_oom
    609         while sigfigs >= 0:
--> 610             if np.abs(locs - np.round(locs, decimals=sigfigs)).max()
< thresh:
    611                 sigfigs -= 1
    612             else:

C:\Users\[redacted]\Anaconda3\lib\site-packages\numpy\core\fromnumeric.py in roun
d_(a, decimals, out)
   2791     except AttributeError:
   2792         return _wrapit(a, 'round', decimals, out)
-> 2793     return round(decimals, out)
   2794
   2795

AttributeError: 'float' object has no attribute 'rint'
```

```
In [34]: values = (1+np.array([0, 1.e-15]))*1.e27
         plt.plot(values)
```

```
Out[34]: [<matplotlib.lines.Line2D at 0x2910e536128>]

Error in callback <function install_repl_displayhook.<locals>.post_execute at 0
x00000291040B30D0> (for post_execute):

AttributeError: 'float' object has no attribute 'rint'

AttributeError: 'float' object has no attribute 'rint'

<matplotlib.figure.Figure at 0x2910e405828>
```

## Table of Contents (2)

The ToC2 extension displays a floating (draggable) table of contents of the notebooks headers. Optionally, it also allows to automatically number all notebook's sections, and to add a table of Contents cell at the top of the notebook.

compatibility: 4.x

Activate | Deactivate

Parameters

- [x] Automatically number notebook's sections

Maximum level of nested sections to display on the tables of contents

4

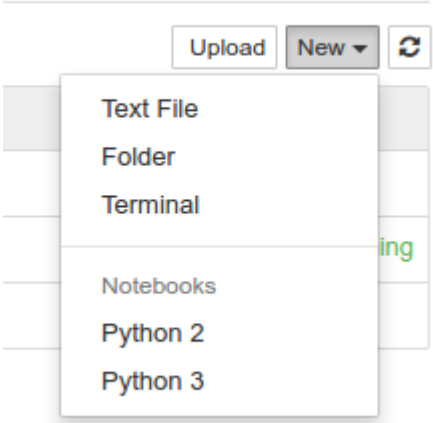
- [ ] Add a Table of Contents at the top of the notebook
- [x] Display toc window at startup

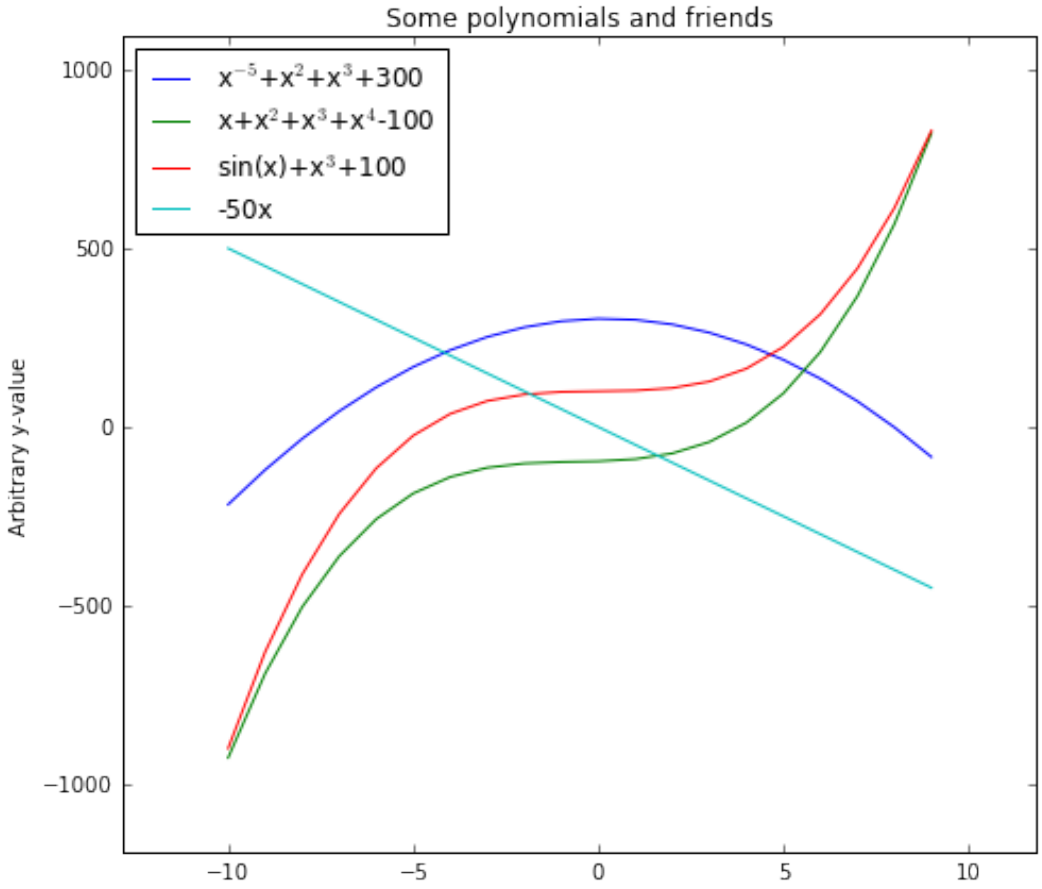
Contents [-] ↻ n t

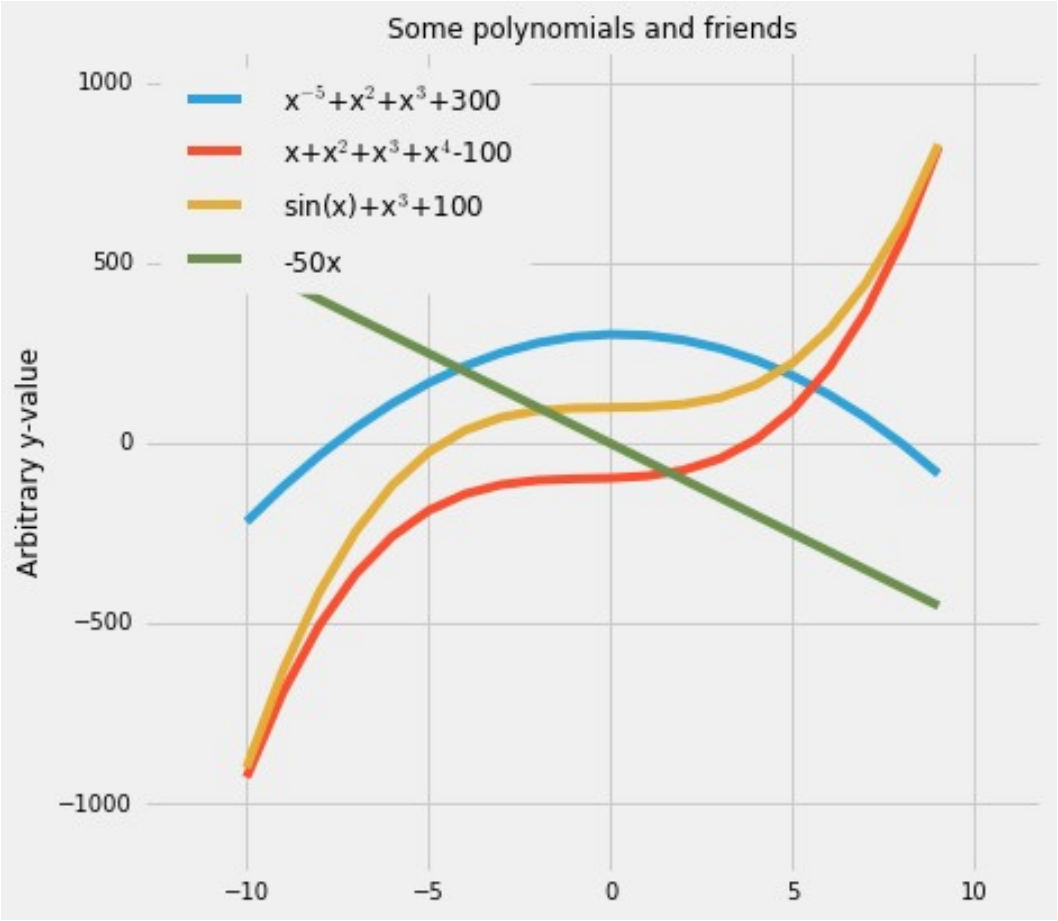
# Table of Contents

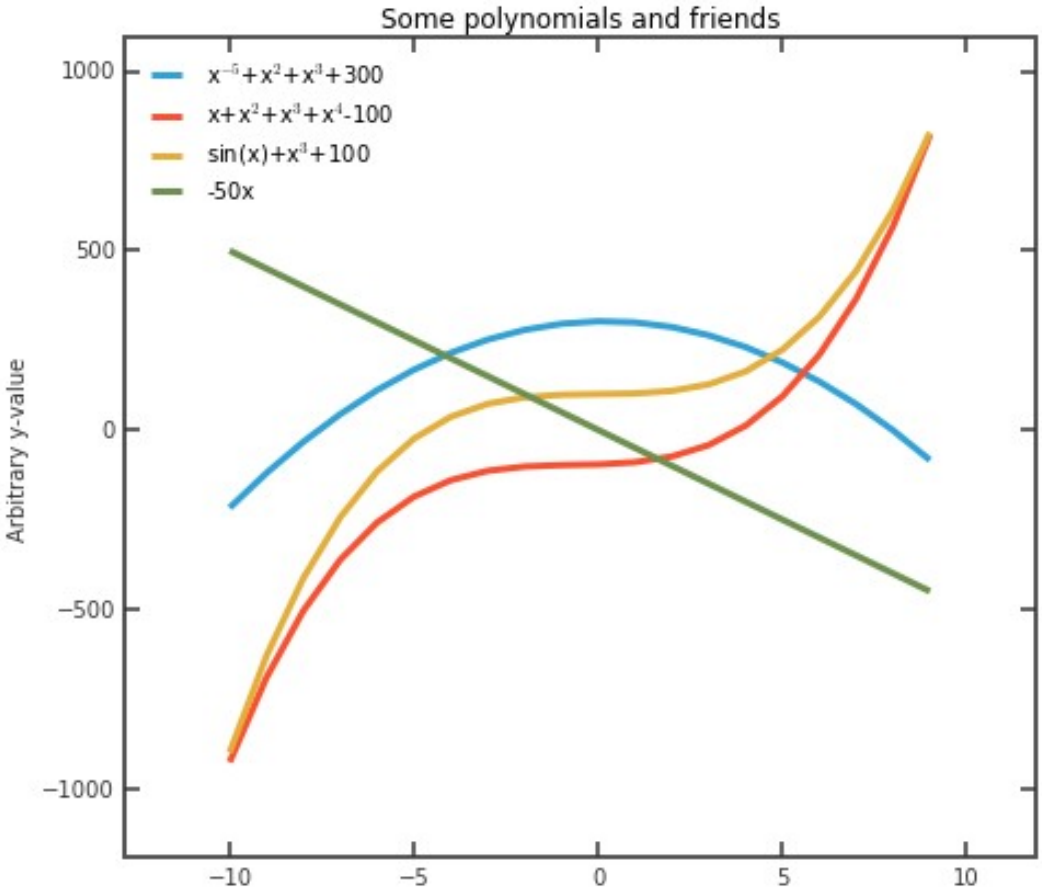


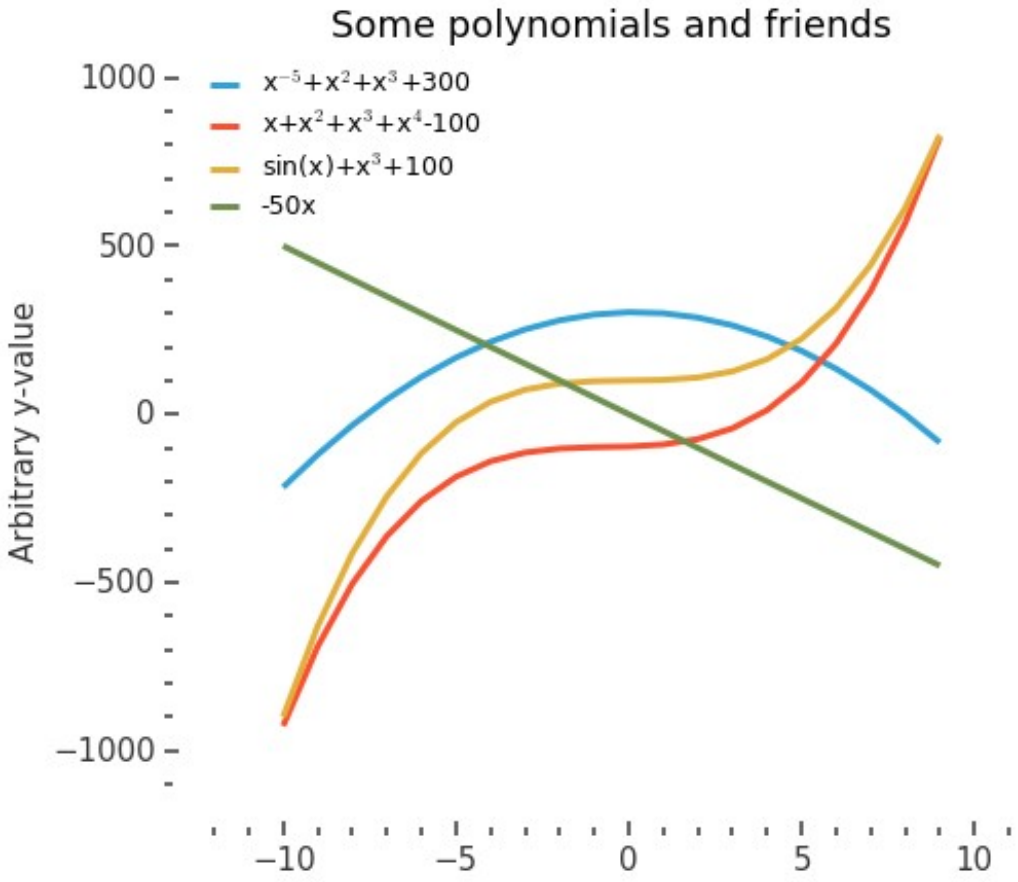
In [1]:
```
import warnings
```




Some polynomials and friends
x^-5+x^2+x^3+300
x+x^2+x^3+x^4-100
sin(x)+x^3+100
-50x
Arbitrary y-value
1000
500
0
−500
−1000
−10
−5
0
5
10


Some polynomials and friends
x⁻⁵+x²+x³+300
x+x²+x³+x⁴-100
sin(x)+x³+100
-50x
Arbitrary y-value
1000
500
0
−500
−1000
−10
−5
0
5
10


Some polynomials and friends
$x^{-5}+x^2+x^3+300$
$x+x^2+x^3+x^4-100$
$\sin(x)+x^3+100$
-50x
Arbitrary y-value
1000
500
0
−500
−1000
−10
−5
0
5
10


Some polynomials and friends
x⁻⁵+x²+x³+300
x+x²+x³+x⁴-100
sin(x)+x³+100
-50x
Arbitrary y-value
1000
500
0
−500
−1000
−10
−5
0
5
10