Data Mining

Lecture 2: data and data preprocessing

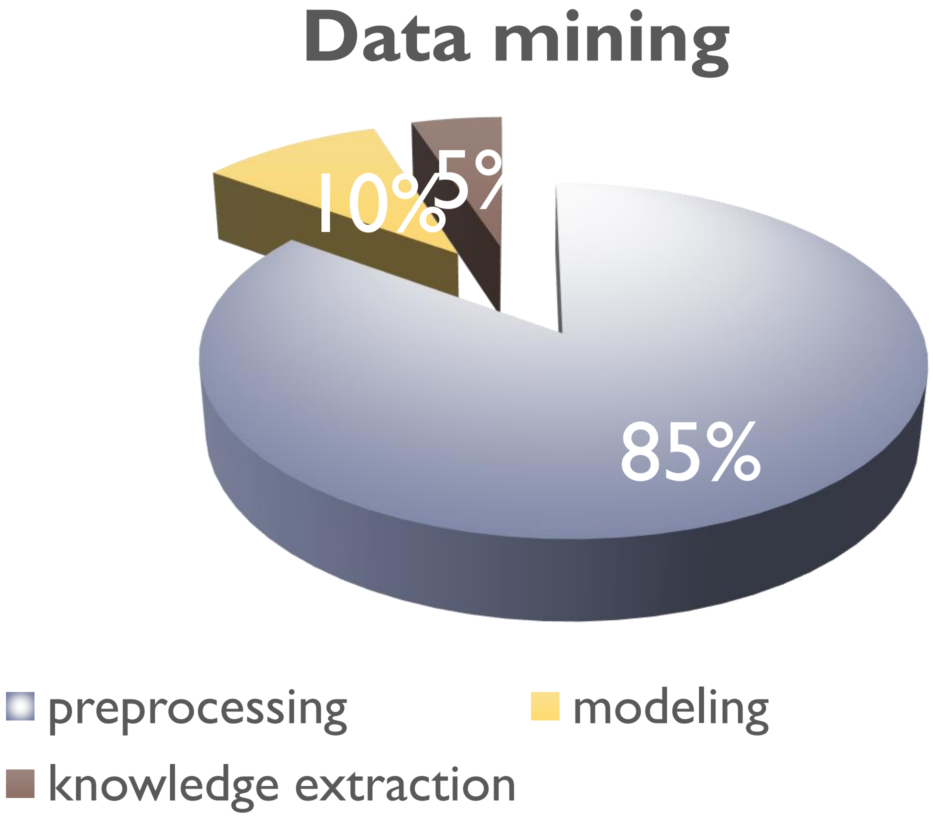
Data mining analysis steps CRISP-dm

1. Determining the purpose of the analysis,
2. Preliminary selection of variables, data cleaning (preprocessing),
3. Analytical form selection, on the basis of descriptive data analysis, correlation matrix, and knowledge,
4. Final variable and model selection,
5. Model parameters estimation,
6. Model verification (statistical, substantive),
7. Practical model uses (dependence analysis, prediction).

GIGO

Garbage in- Garbage out

Time-consuming data mining phases



Data preprocessing

- Preparation of data variables from the preliminary, raw data and the final data set that will be used in all subsequent phases.
- Cases selection and variables that will be analysed and which are suitable for analysis.
- Variable transformations, if necessary.
- Clear raw data so that it is ready to be used by modelling tools.

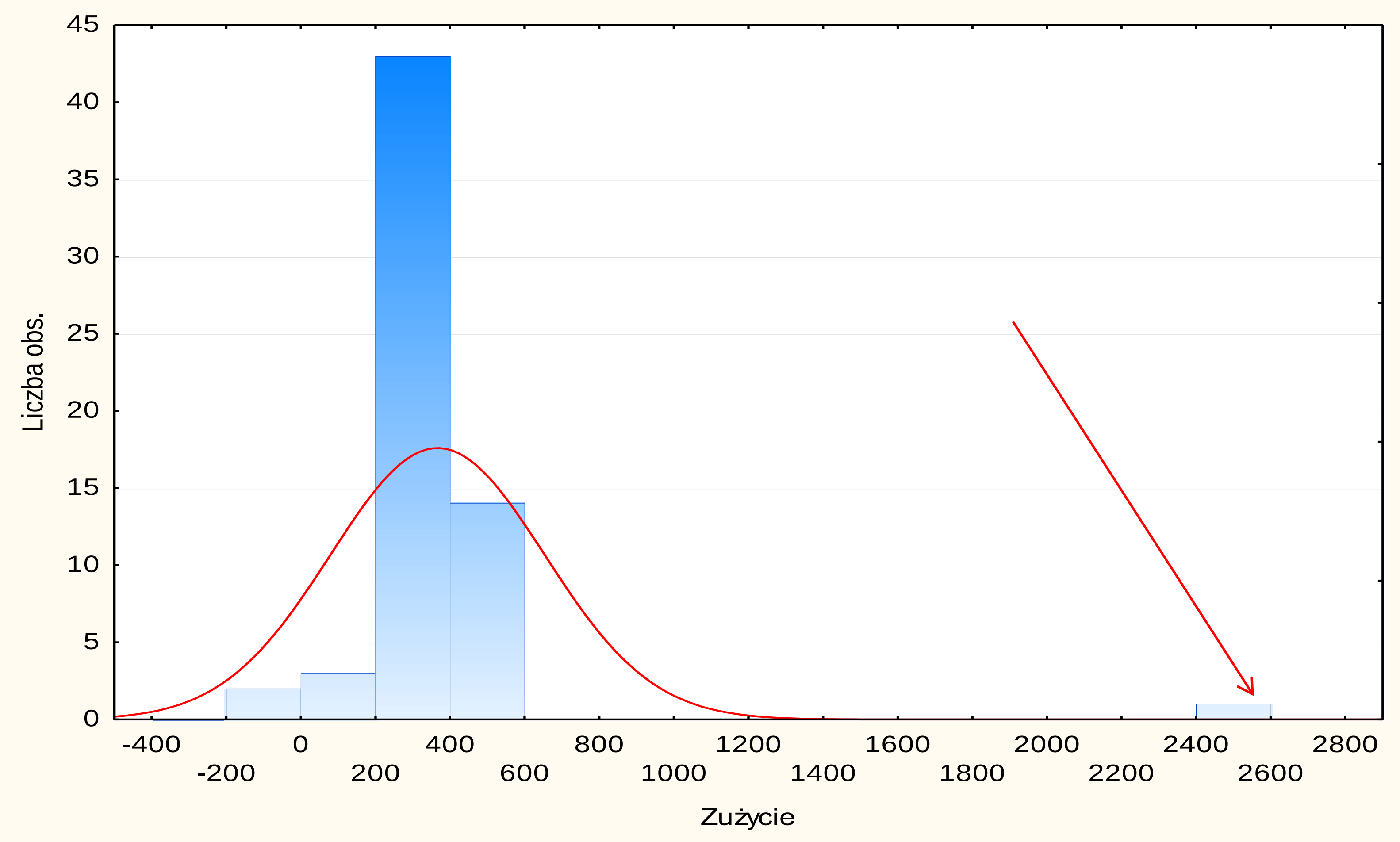
Motivation of data preprocessing

- Preparation of data variables from the preliminary, raw data and the final data set that will be used in all subsequent phases.
- Cases selection and variables that will be analysed and which are suitable for analysis.
- Variable transformations, if necessary.
- Clear the raw data so that it is ready to be used by the modelling tools.

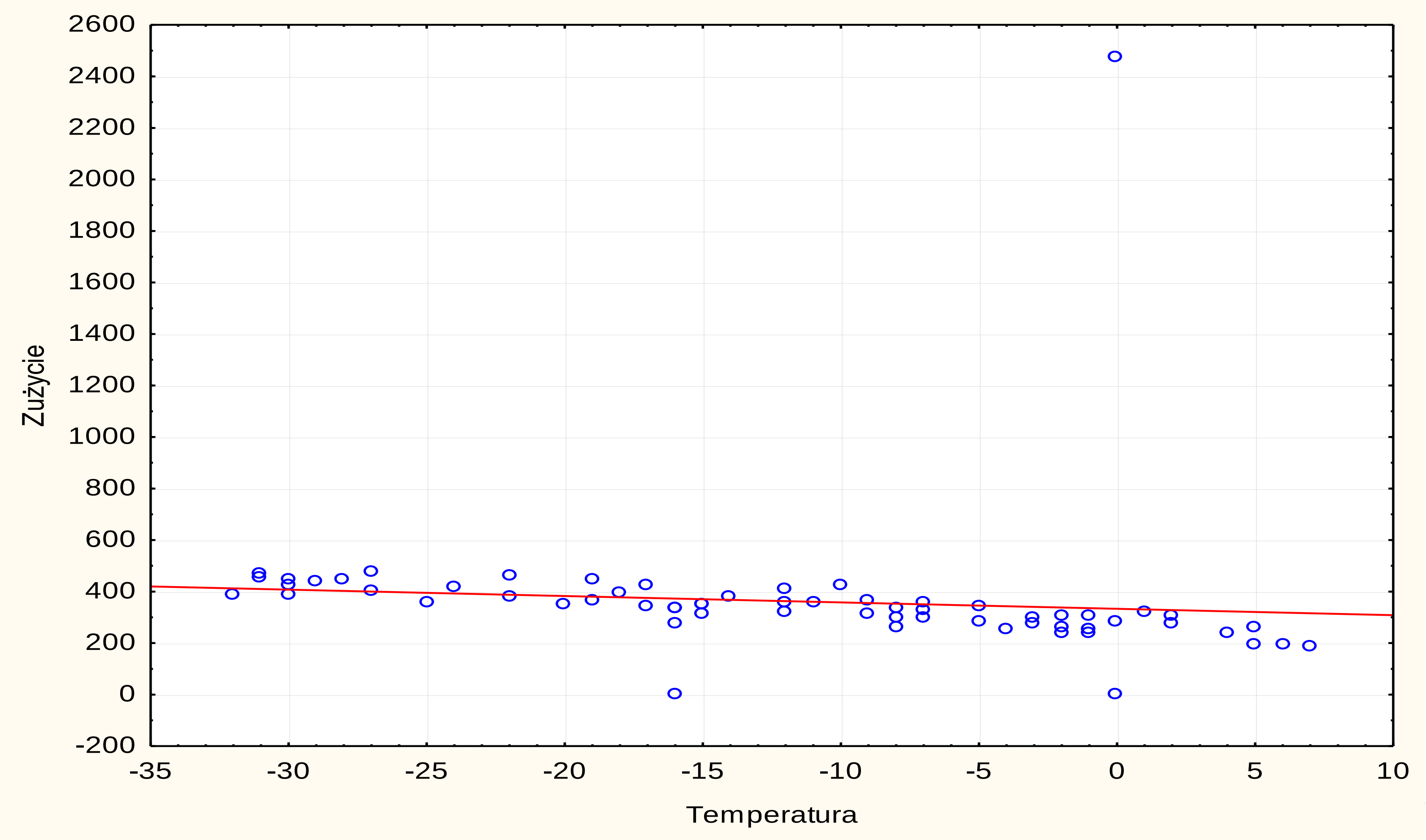
Data cleaning

- Analytical and graphical detection of erroneous and unusual observations,
- Handling and replacing missing data,
- Identification and removal of duplicate records.

Analytical and graphical detection of erroneous and unusual observations



Analytical and graphical detection of erroneous and unusual observations

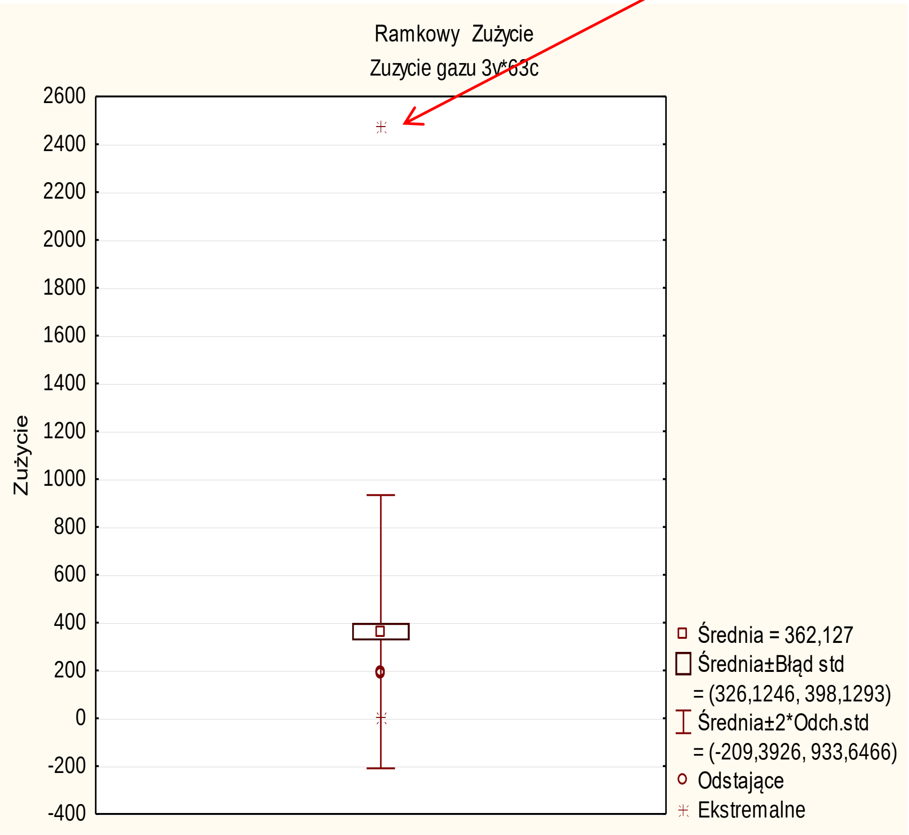


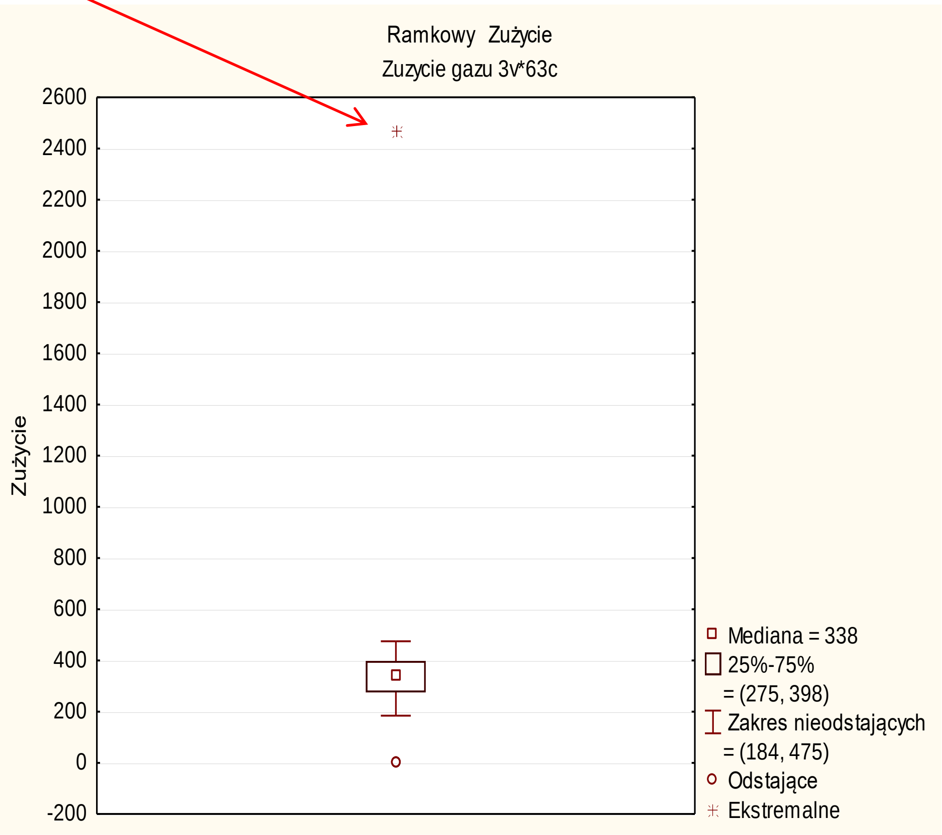
Analytical and graphical detection of erroneous and unusual observations

Descriptive statistics						
	N	Mean	Median	Min.	Max.	Std. Dev.
Dataset	63	362,1270	338,0000	0,00	2471,000	285,7598

Analytical and graphical detection of erroneous and unusual observations

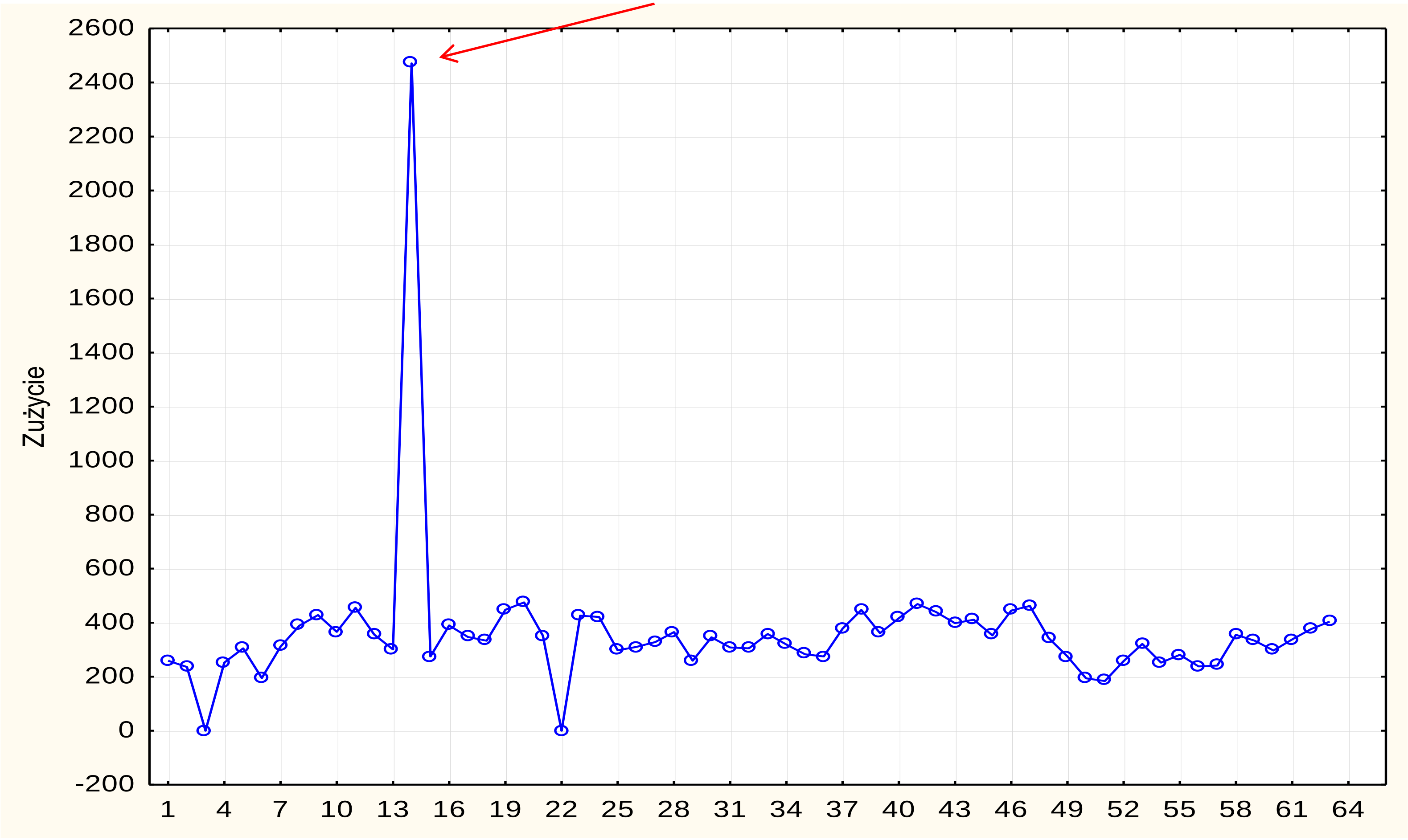
Extreme values





Analytical and graphical detection of erroneous and unusual observations

Extreme values



Missing data

Methods for missing data:

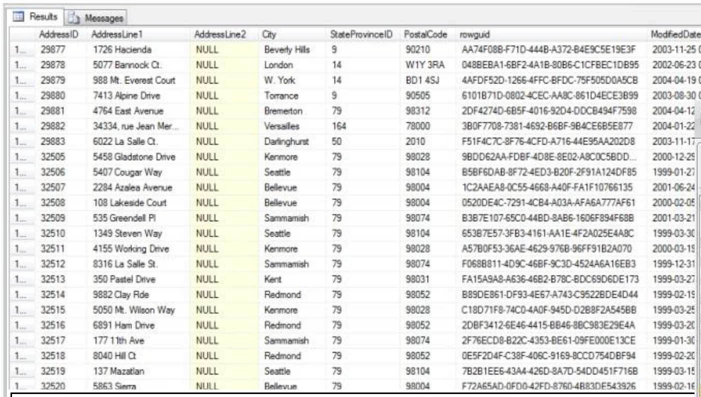
- Omit rows with empty values,
- Replacement of the missing value determined by the analyst,
- Replace with mean or median - for numbers,
- Modal value substitution for qualitative variables,
- Generation of a random value from the observed distribution of the variable,
- Link the blank data to the rest of the object data and give the most probable value.

Missing data

Omit rows with empty values- only for „table data”,

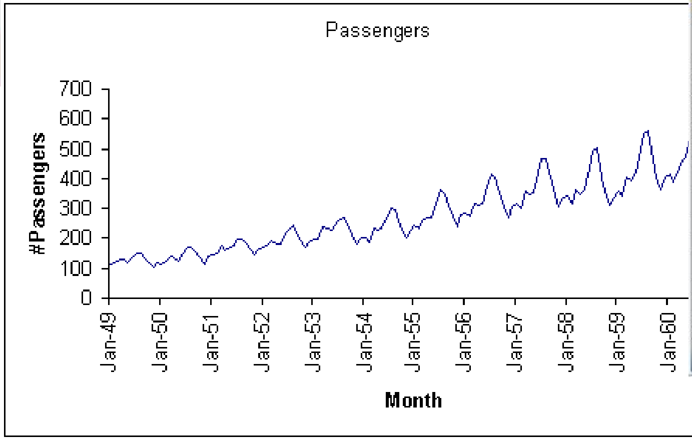
Omit columns with empty values- user criterion,

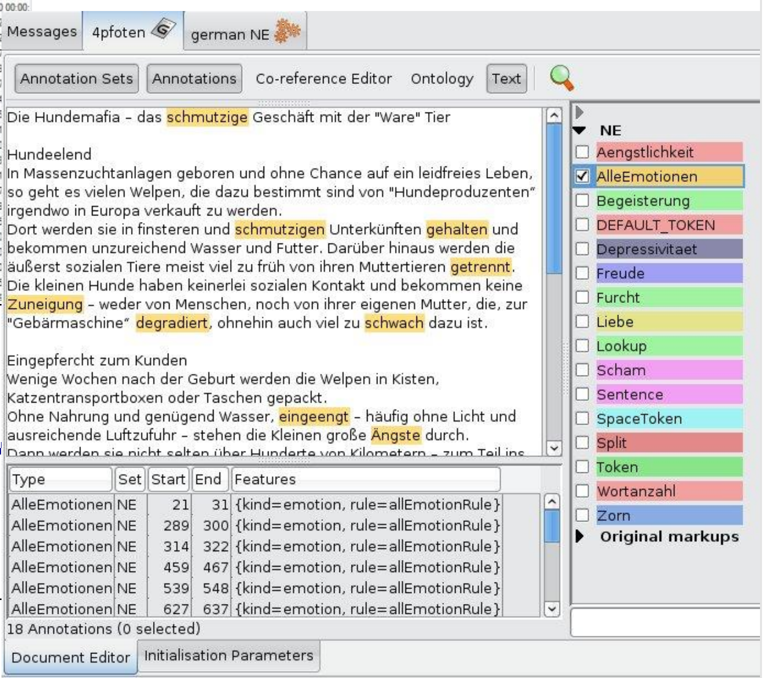
UDL, LDL (ADL, UDL), right censored, left censored.



Results | Messages

	AddressID	AddressLine1	AddressLine2	City	StateProvinceID	PostalCode	rowguid	ModifiedDate
1...	29877	1726 Hacienda	NULL	Beverly Hills	9	90210	AA74F08B-F71D-444B-A372-B4E9C5E19E3F	2003-11-25 00:00:
1...	29878	5077 Bannock Ct.	NULL	London	14	W1Y 3RA	048BEBA1-6BF2-4A1B-80B6-C1CFBEC1DB95	2002-06-23 00:00:
1...	29879	988 Mt. Everest Court	NULL	W. York	14	BD1 4SJ	4AFDF52D-1266-4FFC-BFDC-75F505D0A5CB	2004-04-19 00:00:
1...	29880	7413 Alpine Drive	NULL	Torrance	9	90505	6101B71D-0802-4CEC-AA8C-861D4ECE3B99	2003-08-30 00:00:
1...	29881	4764 East Avenue	NULL	Bremerton	79	98312	2DF4274D-6B5F-4016-92D4-DDCB494F7598	2004-04-12
1...	29882	34334, rue Jean Mer...	NULL	Versailles	164	78000	3B0F7708-7381-4692-B6BF-9B4CE6B5E877	2004-01-22
1...	29883	6022 La Salle Ct.	NULL	Darlinghurst	50	2010	F51F4C7C-8F76-4CFD-A716-44E95AA202D8	2003-11-17
1...	32505	5458 Gladstone Drive	NULL	Kenmore	79	98028	9BDD62AA-FDBF-4D8E-8E02-A8C0C5BDD...	2000-12-29
1...	32506	5407 Cougar Way	NULL	Seattle	79	98104	B5BF6DAB-8F72-4ED3-B20F-2F91A124DF85	1999-01-27
1...	32507	2284 Azalea Avenue	NULL	Bellevue	79	98004	1C2AAEA8-0C55-4668-A40F-FA1F10766135	2001-06-24
1...	32508	108 Lakeside Court	NULL	Bellevue	79	98004	0520DE4C-7291-4CB4-A03A-AFA6A777AF61	2000-02-05
1...	32509	535 Greendell Pl	NULL	Sammamish	79	98074	B3B7E107-65C0-44BD-8AB6-1606F894F68B	2001-03-21
1...	32510	1349 Steven Way	NULL	Seattle	79	98104	653B7E57-3FB3-4161-AA1E-4F2A025E4A8C	1999-03-30
1...	32511	4155 Working Drive	NULL	Kenmore	79	98028	A57B0F53-36AE-4629-976B-96FF91B2A070	2000-03-19
1...	32512	8316 La Salle St.	NULL	Sammamish	79	98074	F068B811-4D9C-46BF-9C3D-4524A6A16EB3	1999-12-31
1...	32513	350 Pastel Drive	NULL	Kent	79	98031	FA15A9A8-A636-46B2-B78C-BDC69D6DE173	1999-03-27
1...	32514	9882 Clay Rde	NULL	Redmond	79	98052	B89DE861-DF93-4E67-A743-C9522BDE4D44	1999-02-19
1...	32515	5050 Mt. Wilson Way	NULL	Kenmore	79	98028	C18D71F8-74C0-4A0F-945D-D2B8F2A545BB	1999-03-25
1...	32516	6891 Ham Drive	NULL	Redmond	79	98052	2DBF3412-6E46-4415-BB46-8BC983E29E4A	1999-03-20
1...	32517	177 11th Ave	NULL	Sammamish	79	98074	2F76ECD8-B22C-4353-BE61-09FE000E13CE	1999-01-30
1...	32518	8040 Hill Ct	NULL	Redmond	79	98052	0E5F2D4F-C38F-406C-9169-8CCD754DBF94	1999-02-20
1...	32519	137 Mazatlan	NULL	Seattle	79	98104	7B2B1EE6-43A4-426D-8A7D-54DD451F716B	1999-03-15
1...	32520	5863 Sierra	NULL	Bellevue	79	98004	F72A65AD-0FD0-42FD-8760-4B83DE543926	1999-02-16





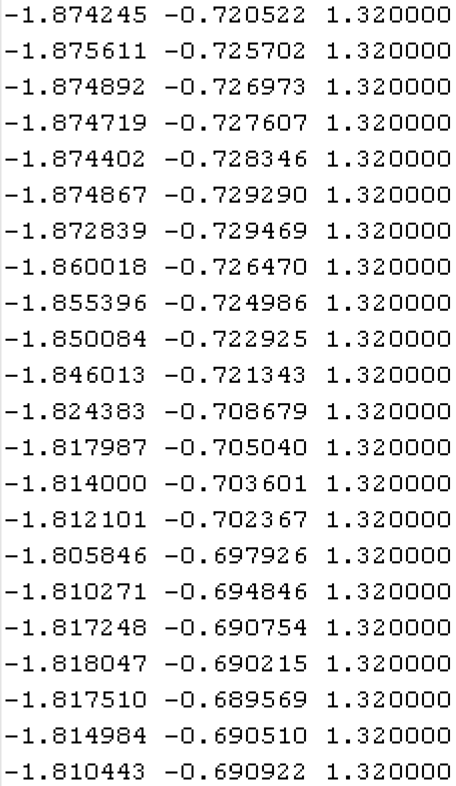
Die Hundemafia – das schmutzige Geschäft mit der "Ware" Tier

Hundeelend
In Massenzuchtanlagen geboren und ohne Chance auf ein leidfreies Leben, so geht es vielen Welpen, die dazu bestimmt sind von "Hundeproduzenten" irgendwo in Europa verkauft zu werden.
Dort werden sie in finsteren und schmutzigen Unterkünften gehalten und bekommen unzureichend Wasser und Futter. Darüber hinaus werden die äußerst sozialen Tiere meist viel zu früh von ihren Muttertieren getrennt.
Die kleinen Hunde haben keinerlei sozialen Kontakt und bekommen keine Zuneigung – weder von Menschen, noch von ihrer eigenen Mutter, die, zur "Gebärmaschine" degradiert, ohnehin auch viel zu schwach dazu ist.

Eingepfercht zum Kunden
Wenige Wochen nach der Geburt werden die Welpen in Kisten, Katzentransportboxen oder Taschen gepackt.
Ohne Nahrung und genügend Wasser, eingeengt – häufig ohne Licht und ausreichende Luftzufuhr – stehen die Kleinen große Ängste durch.
Dann werden sie nicht selten über Hunderte von Kilometern – zum Teil ins

Type	Set	Start	End	Features
AlleEmotionen	NE	21	31	{kind=emotion, rule=allEmotionRule}
AlleEmotionen	NE	289	300	{kind=emotion, rule=allEmotionRule}
AlleEmotionen	NE	314	322	{kind=emotion, rule=allEmotionRule}
AlleEmotionen	NE	459	467	{kind=emotion, rule=allEmotionRule}
AlleEmotionen	NE	539	548	{kind=emotion, rule=allEmotionRule}
AlleEmotionen	NE	627	637	{kind=emotion, rule=allEmotionRule}

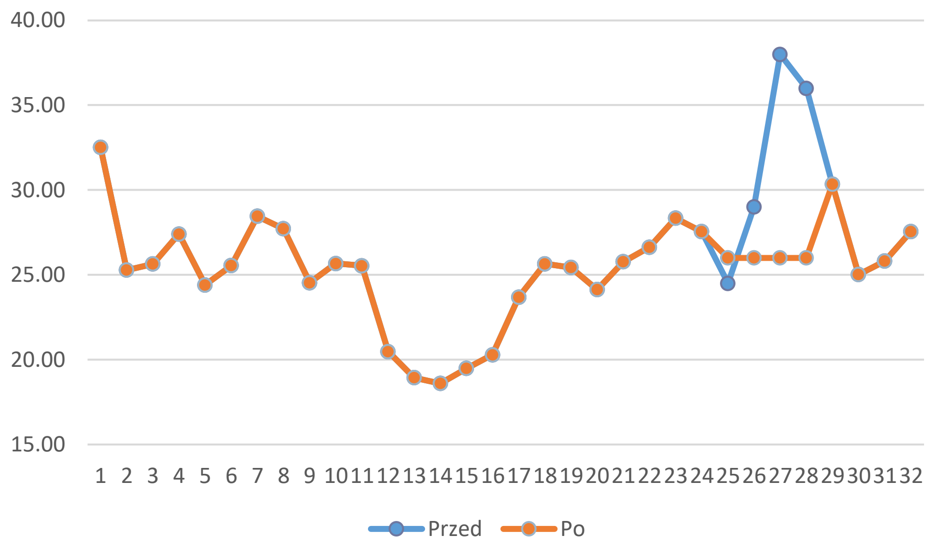
18 Annotations (0 selected)



```
-1.874245 -0.720522 1.320000
-1.875611 -0.725702 1.320000
-1.874892 -0.726973 1.320000
-1.874719 -0.727607 1.320000
-1.874402 -0.728346 1.320000
-1.874867 -0.729290 1.320000
-1.872839 -0.729469 1.320000
-1.860018 -0.726470 1.320000
-1.855396 -0.724986 1.320000
-1.850084 -0.722925 1.320000
-1.846013 -0.721343 1.320000
-1.824383 -0.708679 1.320000
-1.817987 -0.705040 1.320000
-1.814000 -0.703601 1.320000
-1.812101 -0.702367 1.320000
-1.805846 -0.697926 1.320000
-1.810271 -0.694846 1.320000
-1.817248 -0.690754 1.320000
-1.818047 -0.690215 1.320000
-1.817510 -0.689569 1.320000
-1.814984 -0.690510 1.320000
-1.810443 -0.690922 1.320000
```

Missing data

Replacement of the missing value determined by the analyst



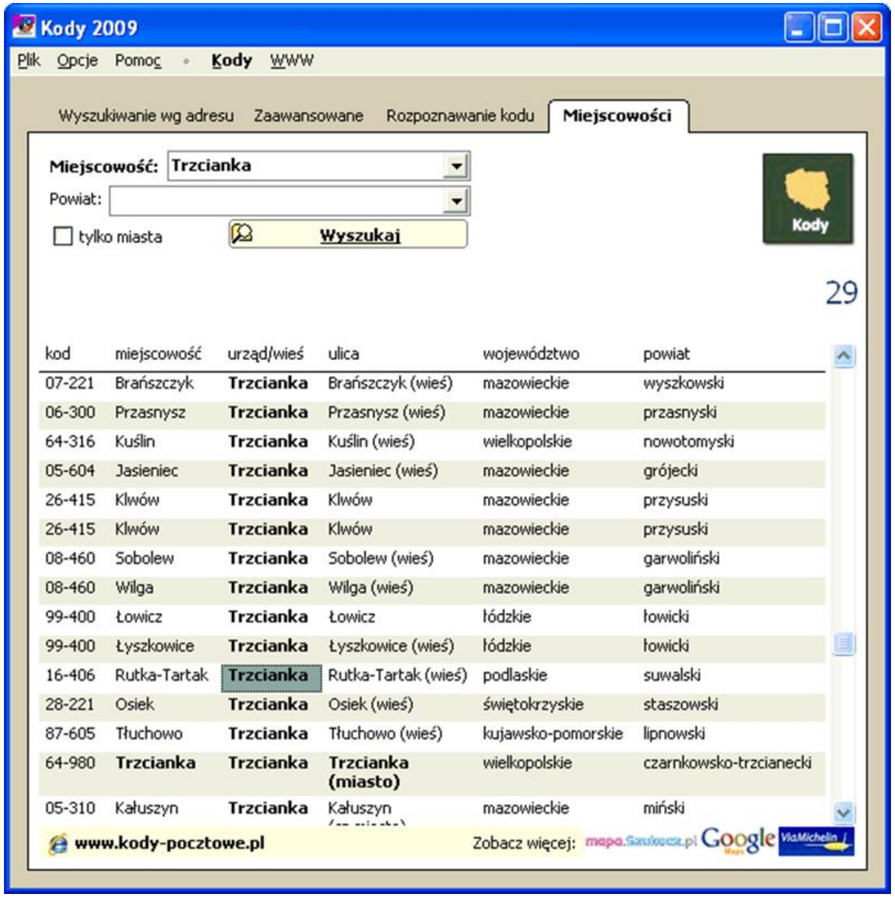
Data transformations

Calculation of derived variables:

- Differentiation of time series,
- Logarithms,
- Roots,
- Calculation of indicators.

Data transformations

Change assignment category (transcoding)



Miejscowość: Trzcianka

Powiat:

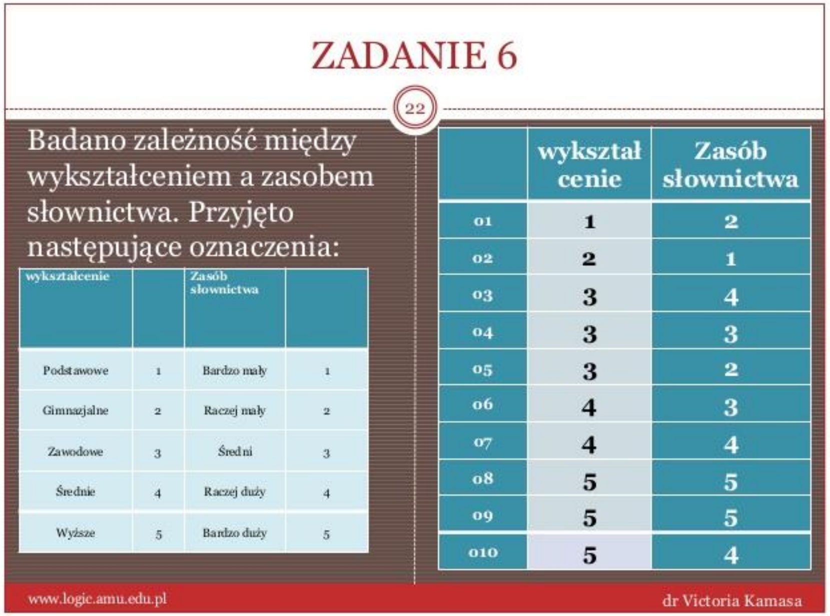
□ tylko miasta

29

kod	miejscowość	urząd/wieś	ulica	województwo	powiat
07-221	Brańszczyk	**Trzcianka**	Brańszczyk (wieś)	mazowieckie	wyszkowski
06-300	Przasnysz	**Trzcianka**	Przasnysz (wieś)	mazowieckie	przasnyski
64-316	Kuślin	**Trzcianka**	Kuślin (wieś)	wielkopolskie	nowotomyski
05-604	Jasieniec	**Trzcianka**	Jasieniec (wieś)	mazowieckie	grójecki
26-415	Klwów	**Trzcianka**	Klwów	mazowieckie	przysuski
26-415	Klwów	**Trzcianka**	Klwów	mazowieckie	przysuski
08-460	Sobolew	**Trzcianka**	Sobolew (wieś)	mazowieckie	garwoliński
08-460	Wilga	**Trzcianka**	Wilga (wieś)	mazowieckie	garwoliński
99-400	Łowicz	**Trzcianka**	Łowicz	łódzkie	łowicki
99-400	Łyszkowice	**Trzcianka**	Łyszkowice (wieś)	łódzkie	łowicki
16-406	Rutka-Tartak	**Trzcianka**	Rutka-Tartak (wieś)	podlaskie	suwalski
28-221	Osiek	**Trzcianka**	Osiek (wieś)	świętokrzyskie	staszowski
87-605	Tłuchowo	**Trzcianka**	Tłuchowo (wieś)	kujawsko-pomorskie	lipnowski
64-980	**Trzcianka**	**Trzcianka**	**Trzcianka (miasto)**	wielkopolskie	czarnkowsko-trzcianecki
05-310	Kałuszyn	**Trzcianka**	Kałuszyn	mazowieckie	miński

Data transformations

Rank assignment



ZADANIE 6

22

Badano zależność między wykształceniem a zasobem słownictwa. Przyjęto następujące oznaczenia:

wykształcenie		Zasób słownictwa	
Podstawowe	1	Bardzo mały	1
Gimnazjalne	2	Raczej mały	2
Zawodowe	3	Średni	3
Średnie	4	Raczej duży	4
Wyższe	5	Bardzo duży	5

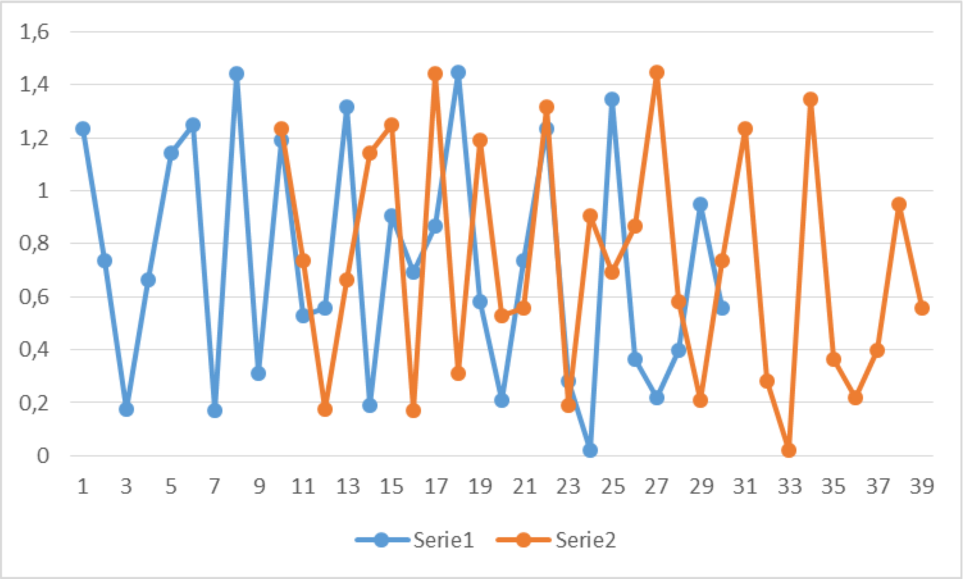
	wykształcenie	Zasób słownictwa
o1	1	2
o2	2	1
o3	3	4
o4	3	3
o5	3	2
o6	4	3
o7	4	4
o8	5	5
o9	5	5
o10	5	4

www.logic.amu.edu.pl

dr Victoria Kamasa

Data transformations

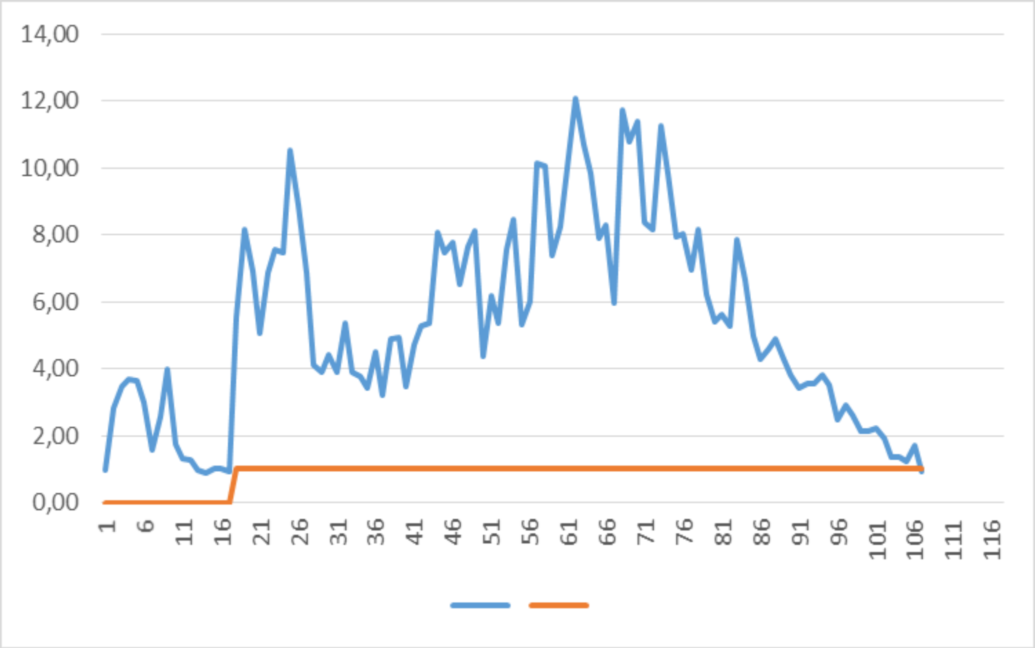
Latent variables



Data transformations

Text variables operations:

- Counting,
- Conversion to numbers (recoding),
- Complete the information,
- Analysis of descriptions.



Data analysis

It is a knowledge discovery about:

- pure data, data type,
- data collection,
- data preparation,
- factors that affect the development of the phenomenon,
- seasonality, trend, events and incidents.

Data in Matrix Form

This data type having one (dependent) variable described by one or many (independent /predictors) variables.

Example: Runners finish time on 100 meters (dependent variable) described by: height, weight, number of workouts, leg length (independent variables).

Regression problems can also be written as binary variables (they replace all variables that are not quantitative).

Time-Series Data

Time-series database

- Consists of sequences of values or events changing with time,
- Data is recorded at regular intervals,
- Characteristic time-series components,

Trend, cycle, seasonal, irregular.

Applications

- Financial: stock price, inflation,
- Industry: power consumption,
- Scientific: experiment results,
- Meteorological: precipitation.

Categories of Time-Series Movements

Categories of Time-Series Movements

- Long-term or trend movements (trend curve): general direction in which a time series is moving over a long interval of time,
- Cyclic movements or cycle variations: long term oscillations about a trend line or curve,

  e.g., business cycles, may or may not be periodic.
- Seasonal movements or seasonal variations,

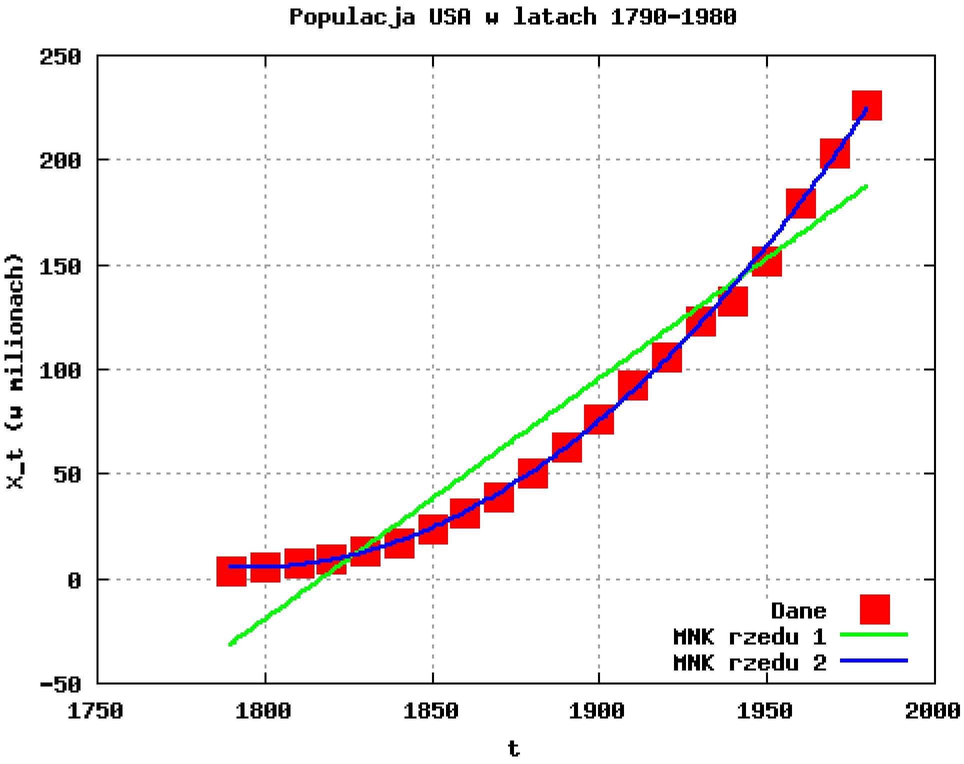
  almost identical patterns that a time series appears to follow during corresponding months of successive years.
- Irregular or random movements,

Time series analysis: decomposition of a time series into these four basic movements:

- Additive Modal: TS = T + C + S + I
- Multiplicative Modal: TS = T × C × S × I

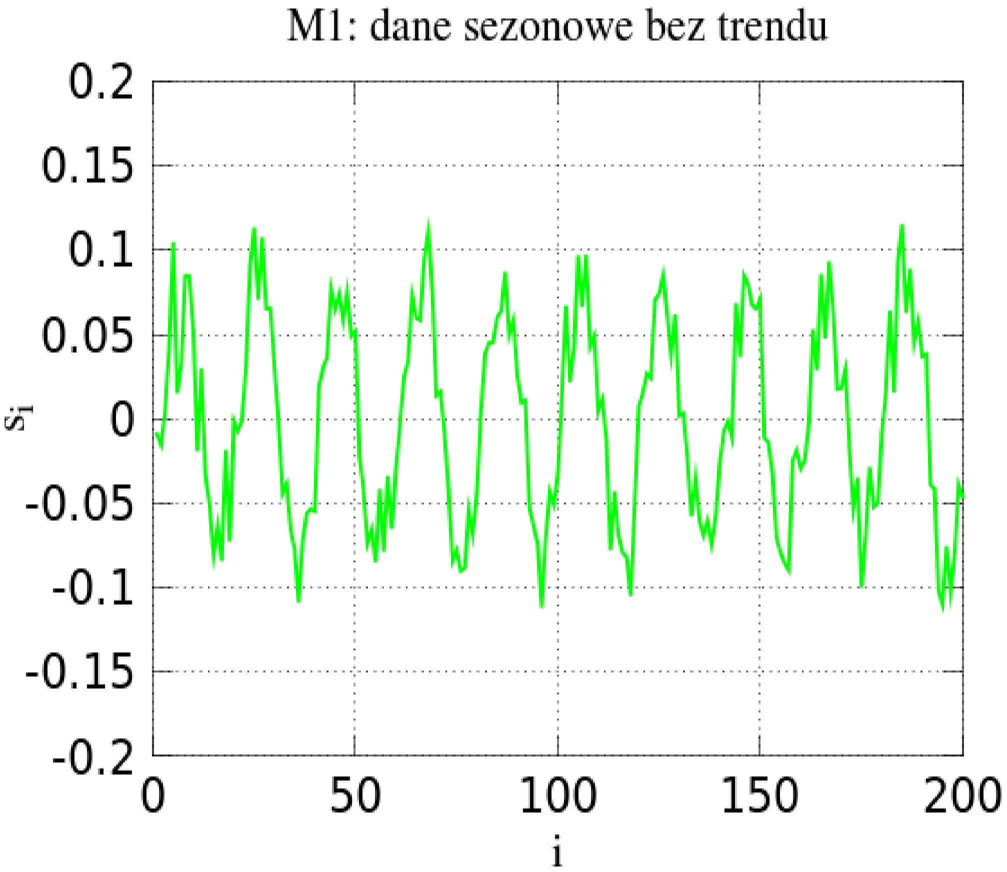
Categories of Time-Series Movements

- Long-term or trend movements,
- Average level,



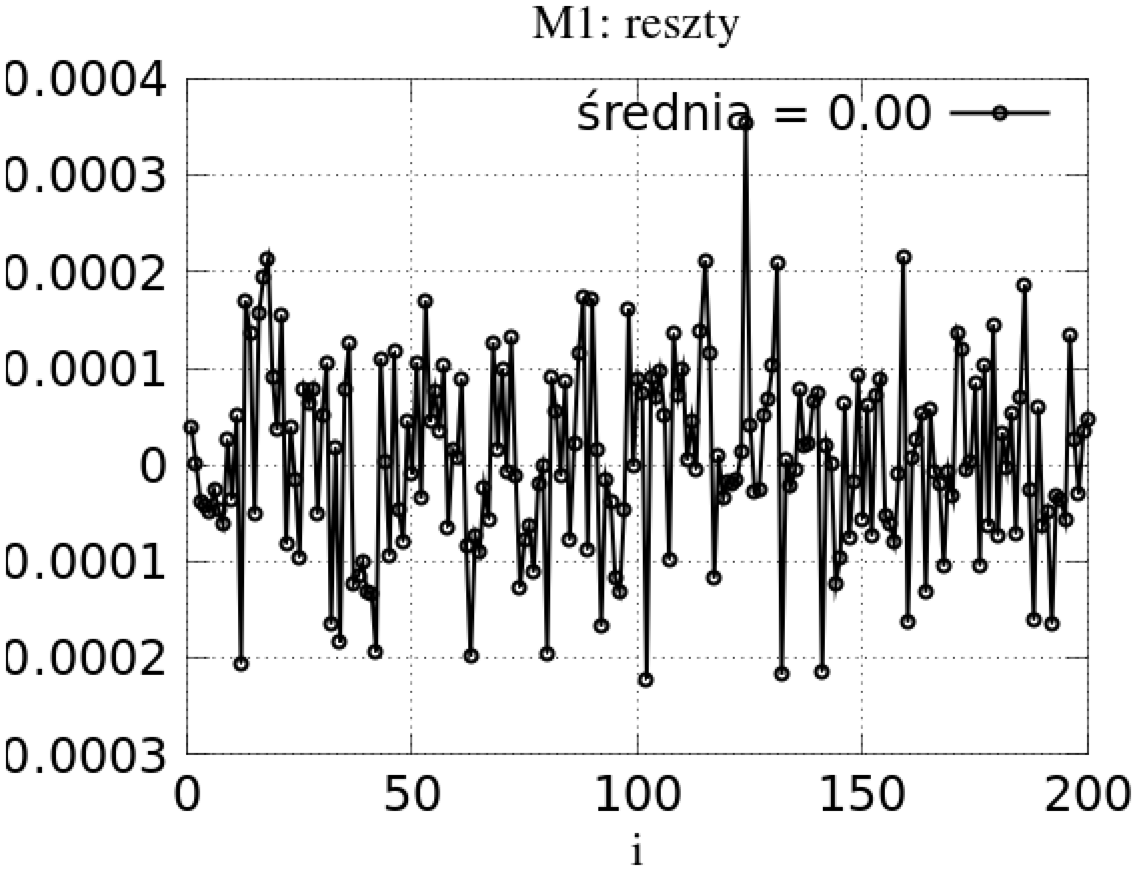
Categories of Time-Series Movements

- Cyclic movements or cycle variations



Categories of Time-Series Movements

- Irregular or random movements

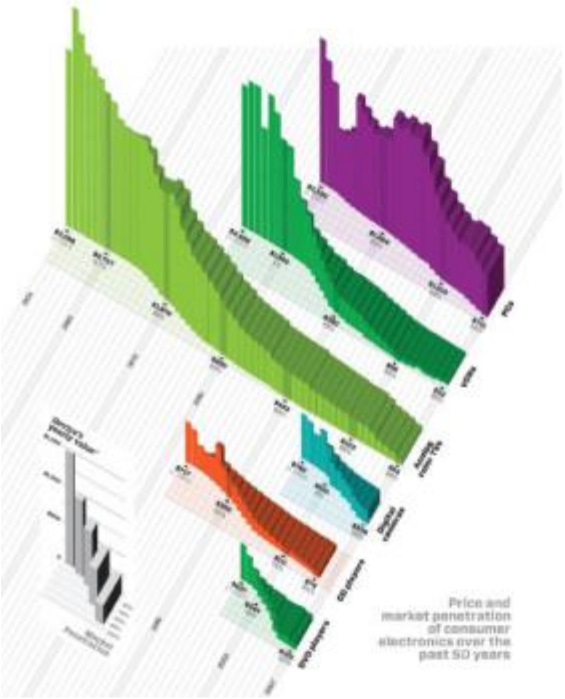


Aim of the Time series analysis

- Modelling of a certain phenomenon / process based on observed changes in some measurable quantities describing this process,
- Isolation and measurement of time series components (decomposition of time series),
- Prediction of future values using the obtained model.

How to get information about data?

- Descriptive analysis,
- Data visualizations: graphs, plots, histograms.



Data analysis, what to do?

1. Plots: linear, spot, column, box and whiskers,

2. Descriptive statistics (mean, median, max., min., std. dev.,…), tables, crosstabulation tables,

3. Histograms, normality graph,

4. ACF and PACF graphs, Fourier analysis – for time series,

5. Correlation matrices,

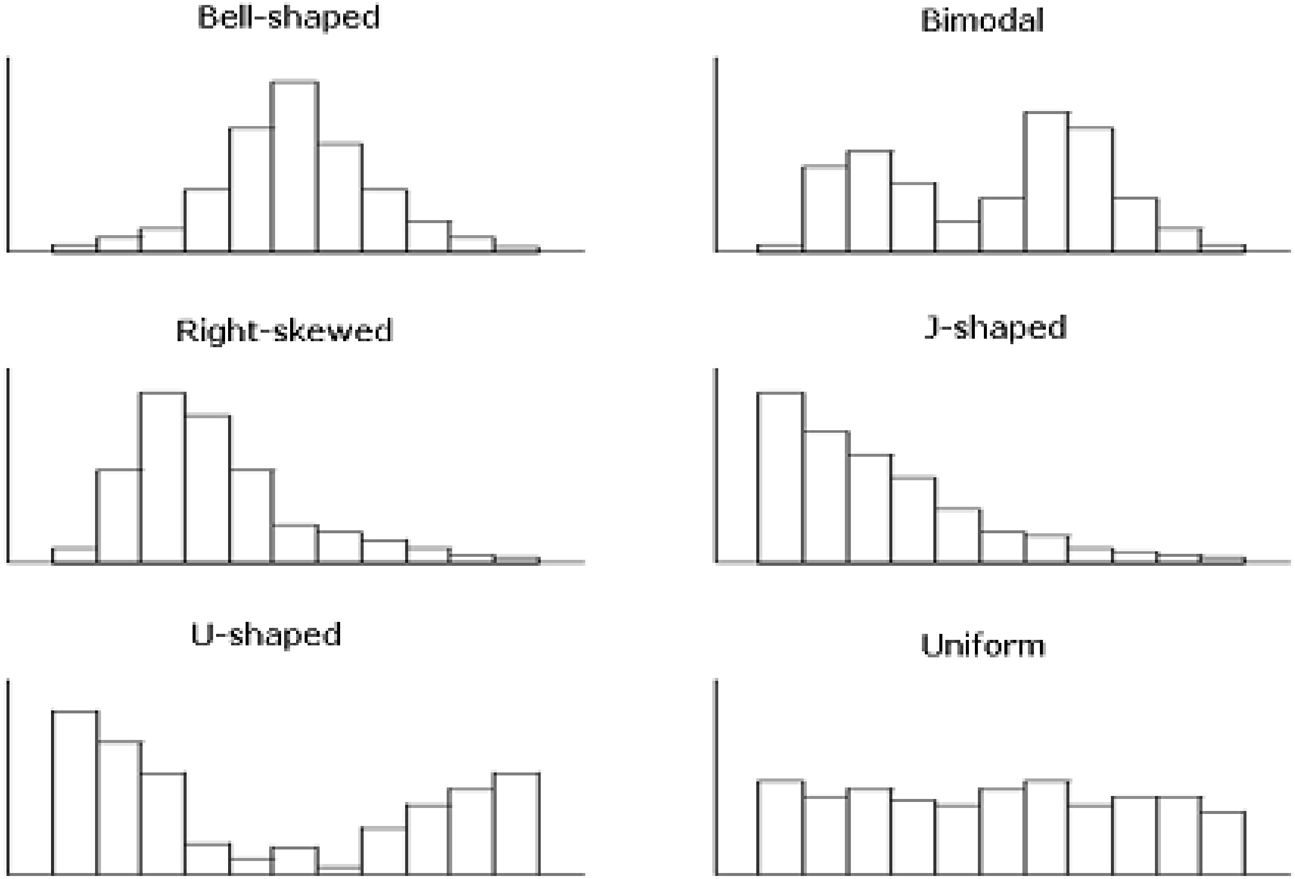
6. Cluster analysis.

Descriptive statistics

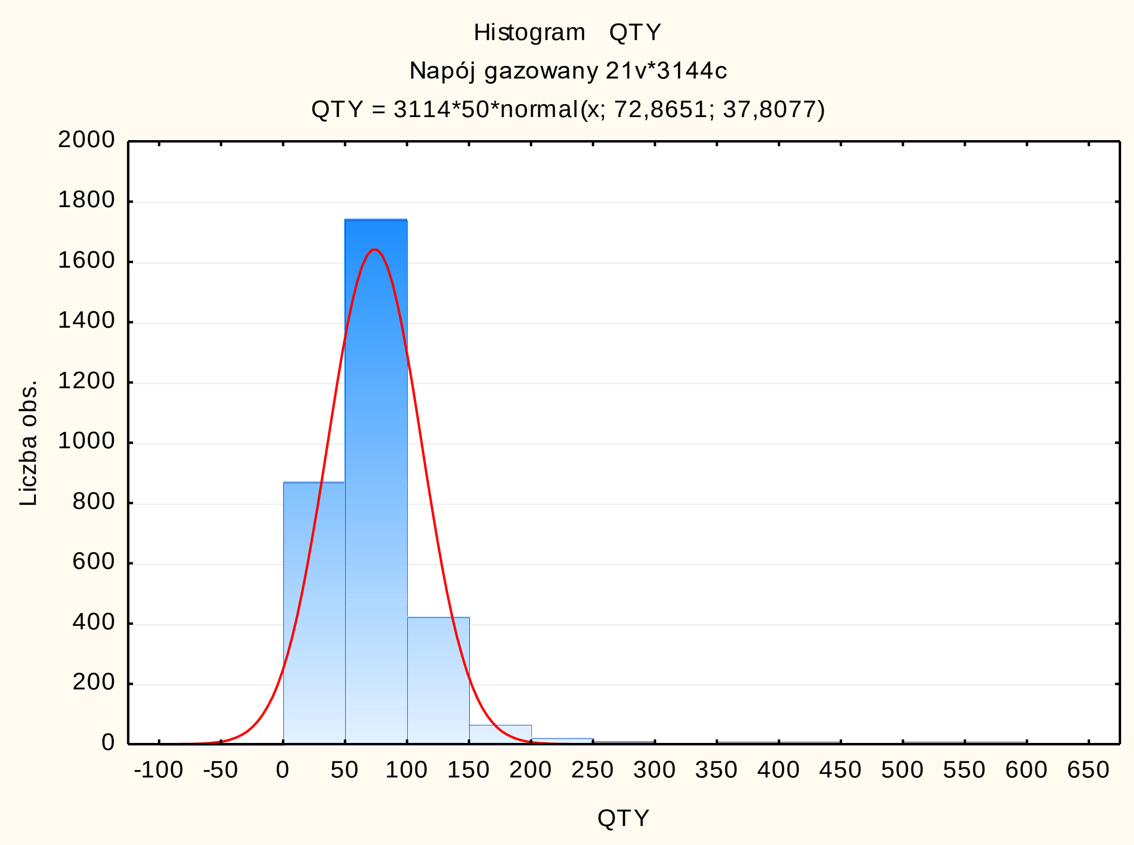
Descriptive statistics (chronologie jesiony)

	N	Mean	Median	Mode	Min.	Max.	Variance	Std. dev.	Variation coefficient	Skewness
ry1fr_n	116	0,977	0,961	Wielokr.	0,499	1,650	0,033	0,181	18,486	0,249
ry1fr_r	116	1,002	1,003	Wielokr.	0,547	1,632	0,029	0,172	17,156	0,365
ry1fr_c	116	0,963	0,954	Wielokr.	0,236	1,754	0,106	0,326	33,862	-0,193

Histograms -shapes

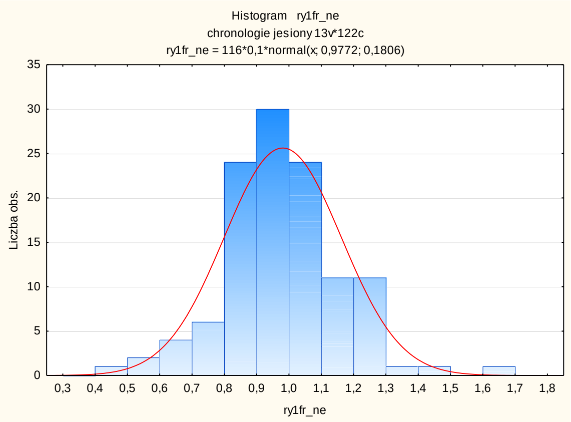


Data analysis

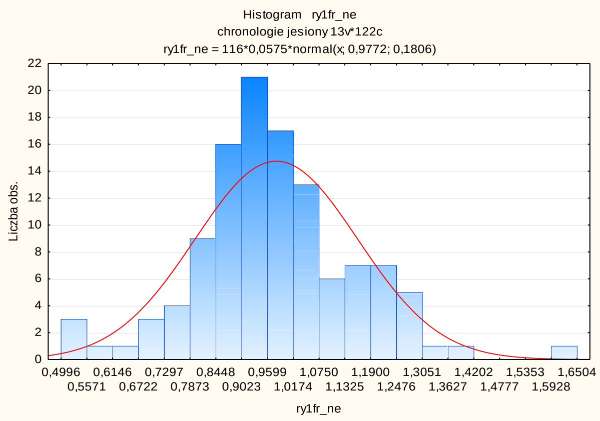


Data analysis

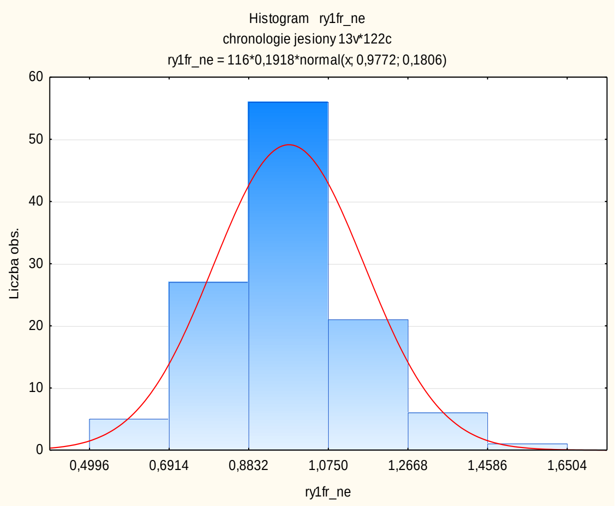
Histogram - numbers of bars



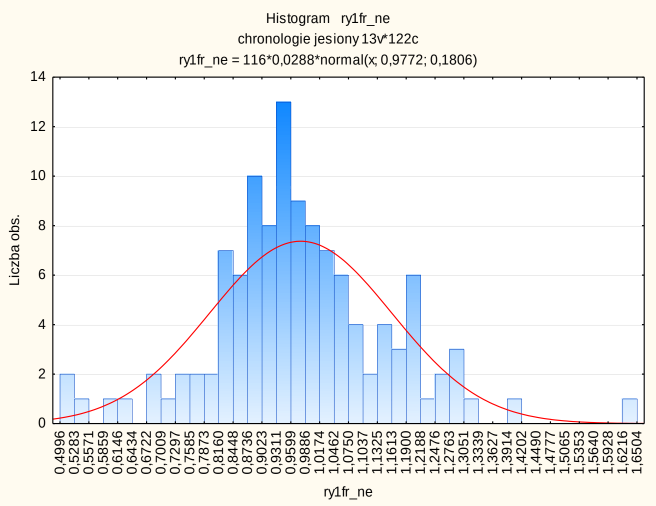
Histogram ry1fr_ne
chronologie jesiony 13v*122c
ry1fr_ne = 116*0,1*normal(x; 0,9772; 0,1806)



Histogram ry1fr_ne
chronologie jesiony 13v*122c
ry1fr_ne = 116*0,0575*normal(x; 0,9772; 0,1806)



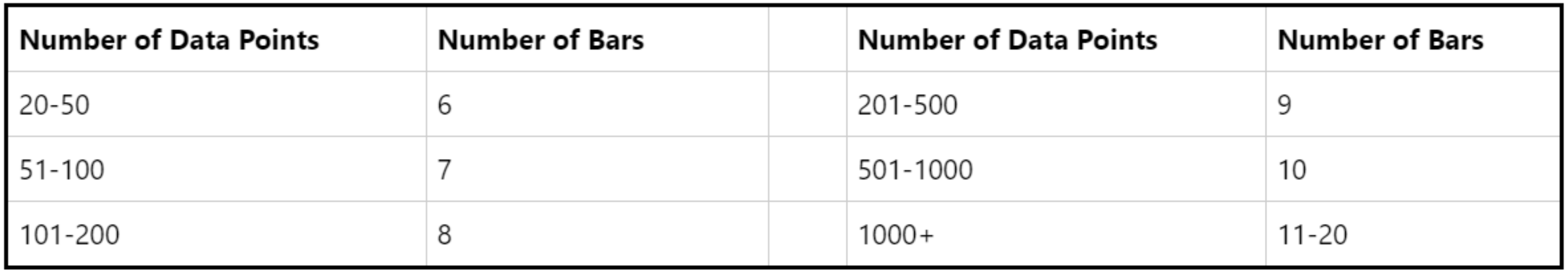
Histogram ry1fr_ne
chronologie jesiony 13v*122c
ry1fr_ne = 116*0,1918*normal(x; 0,9772; 0,1806)



Histogram ry1fr_ne
chronologie jesiony 13v*122c
ry1fr_ne = 116*0,0288*normal(x; 0,9772; 0,1806)

Histograms- numer of bars

Juran's Quality Control Handbook provides these guidelines for the number of bars and states that they are not "rigid" and should be adjusted when necessary.

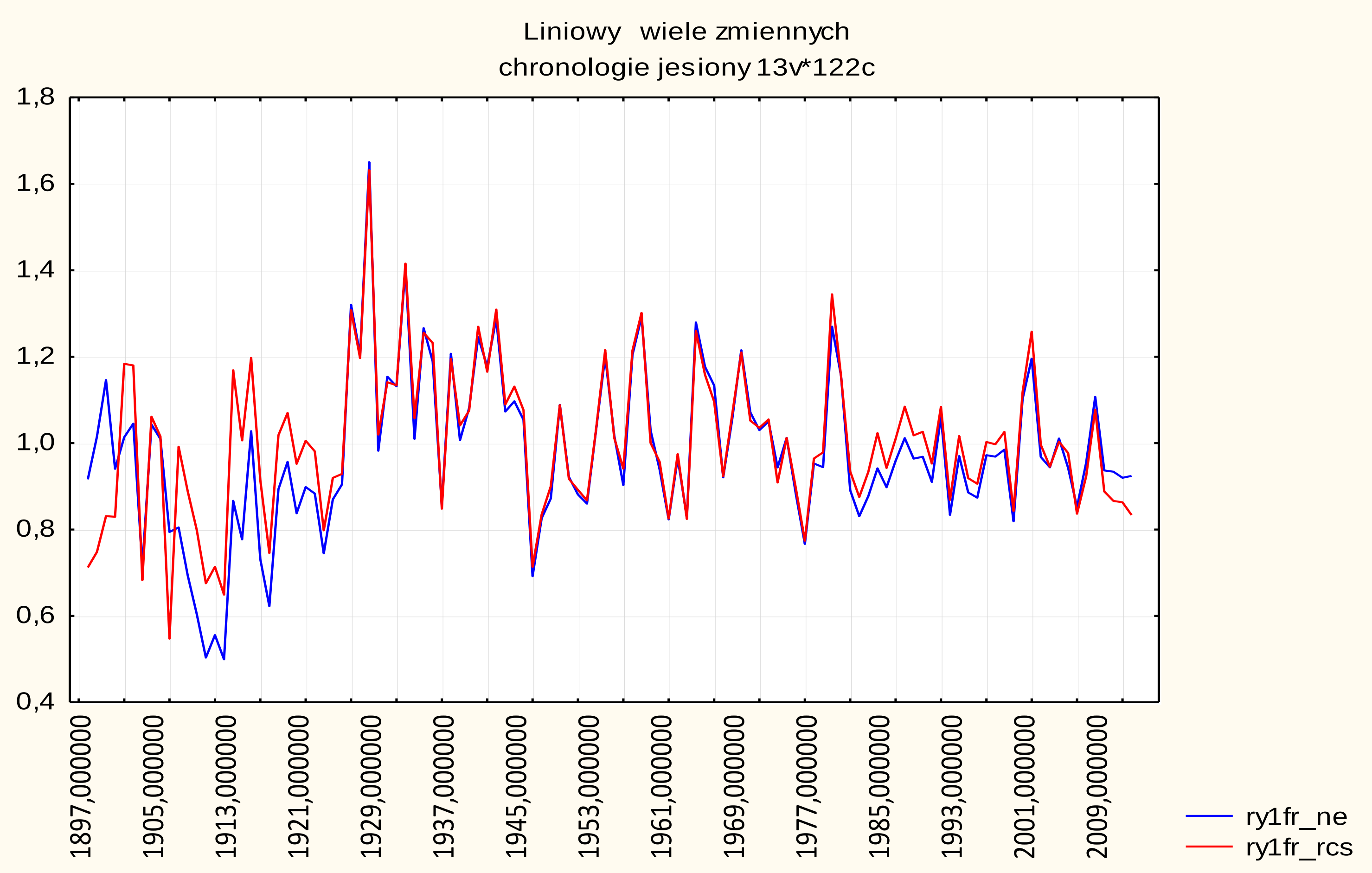


Number of Data Points	Number of Bars		Number of Data Points	Number of Bars
20-50	6		201-500	9
51-100	7		501-1000	10
101-200	8		1000+	11-20

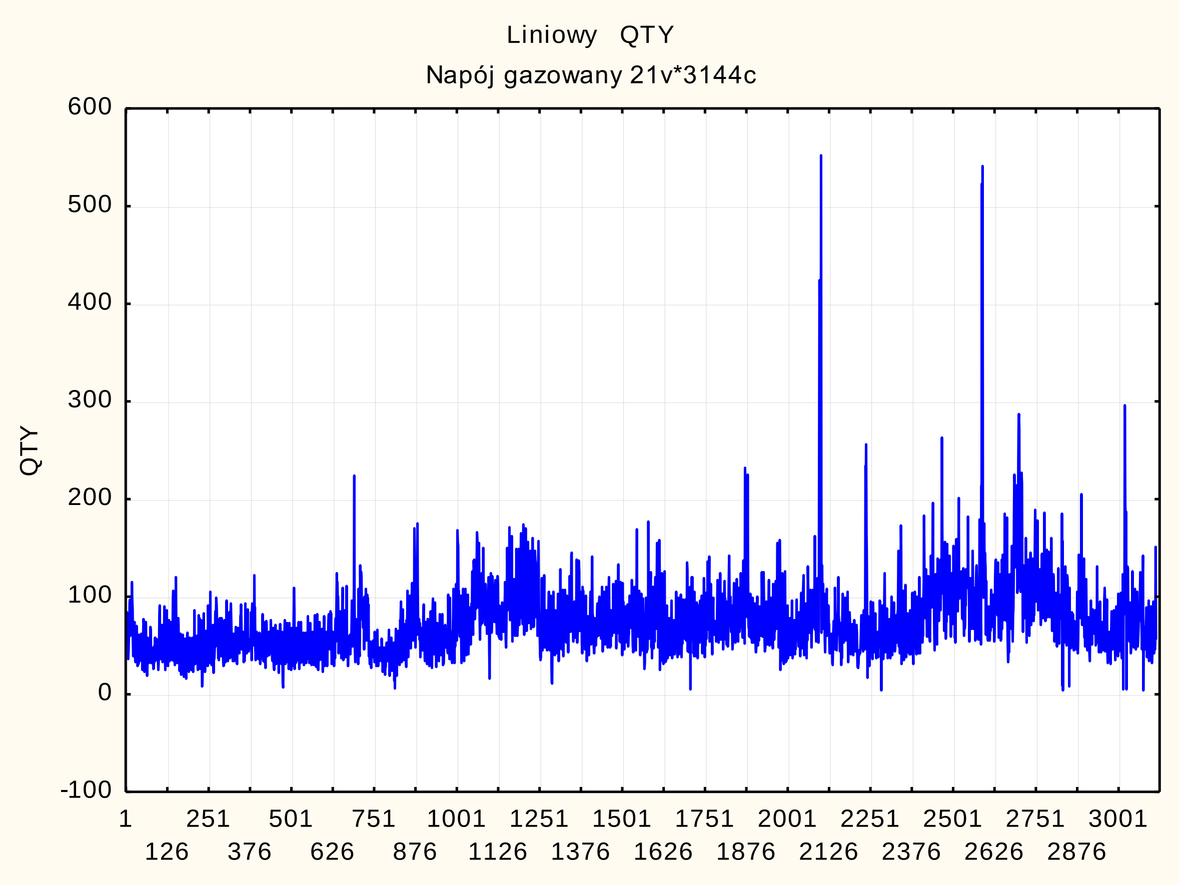
$$k < 5\log n \qquad k \approx \sqrt{n} \qquad k \approx 1 + 3{,}3\log n \qquad k \le \frac{n}{10}$$

Szerokość klasy: $\approx(\mathbf{x}_{\max}-\mathbf{x}_{\min})/k$

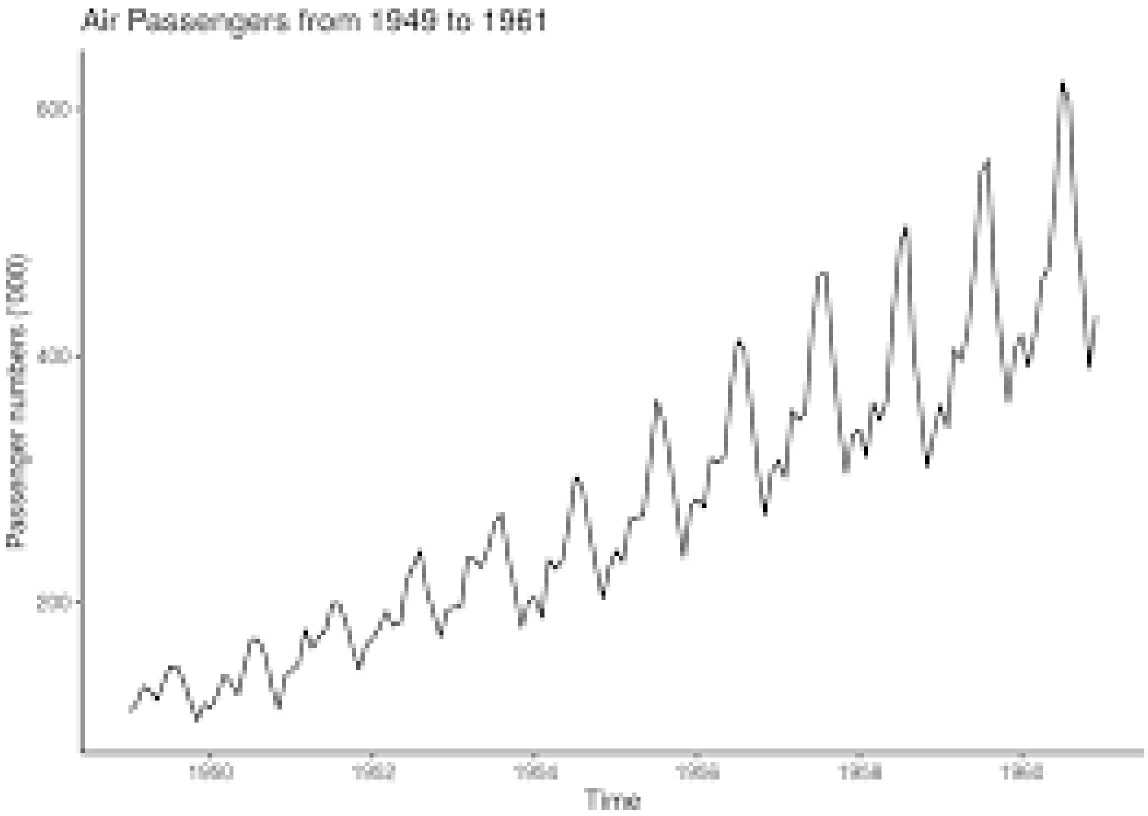
Data analysis



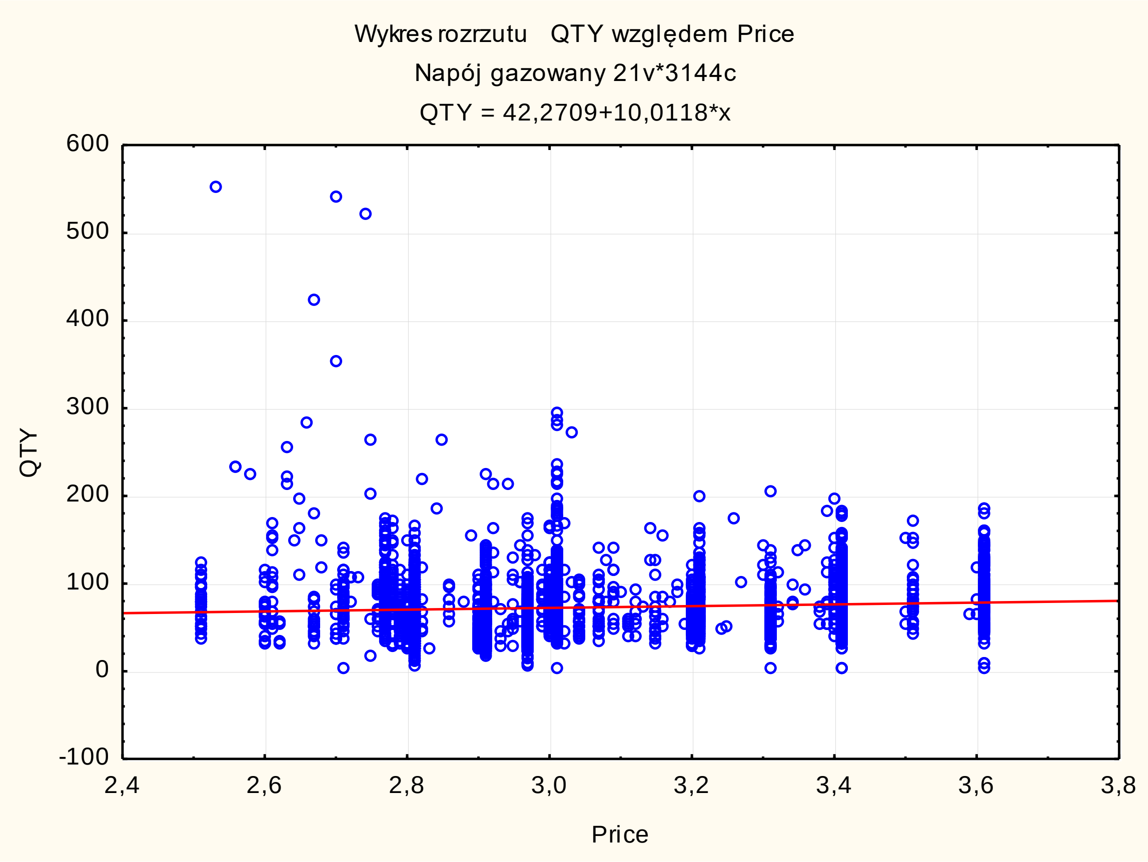
Data analysis



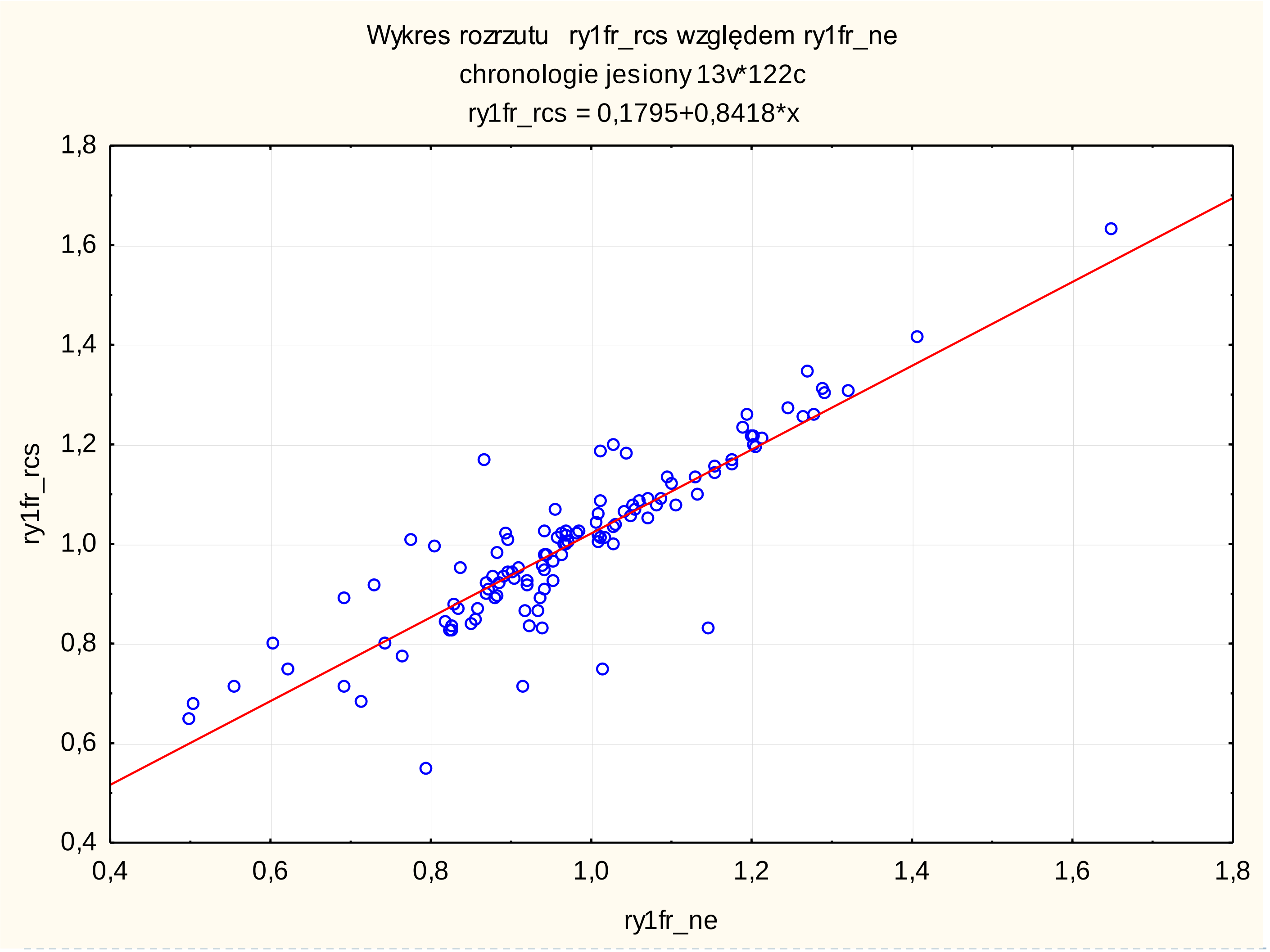
Data analysis



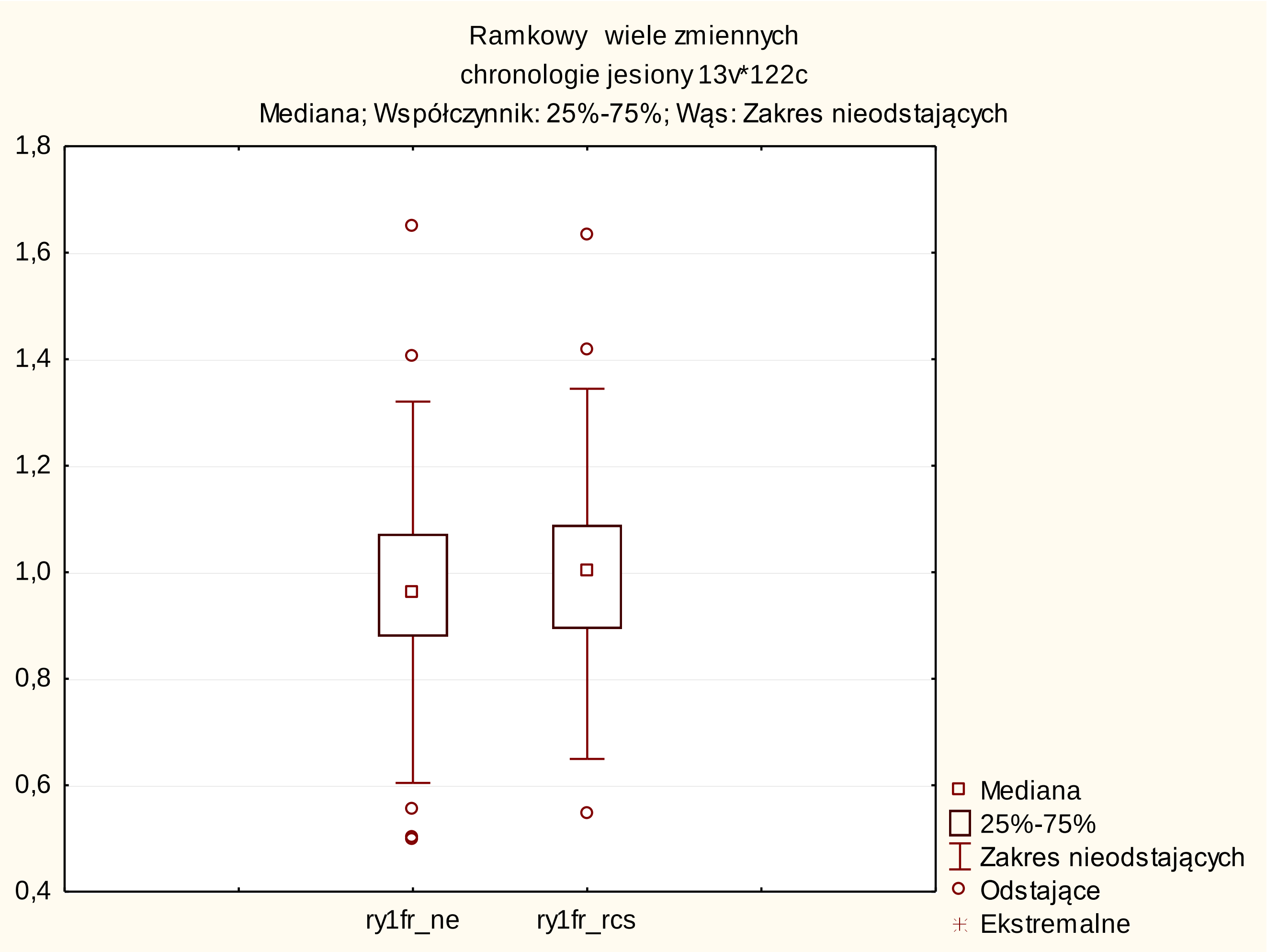
Data analysis



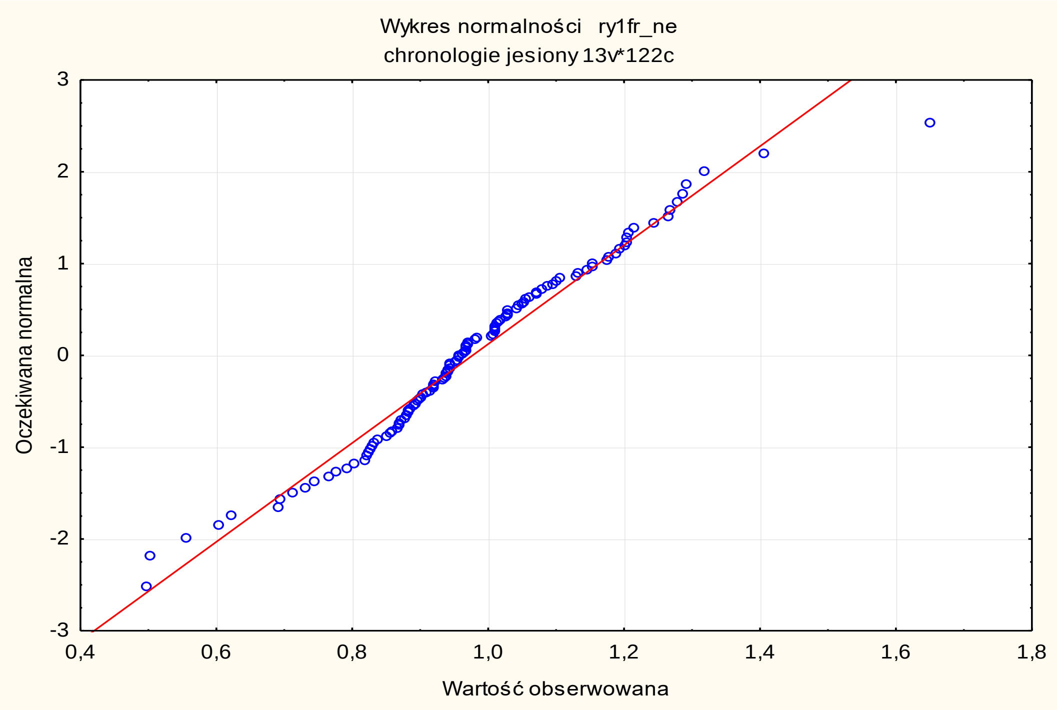
Data analysis



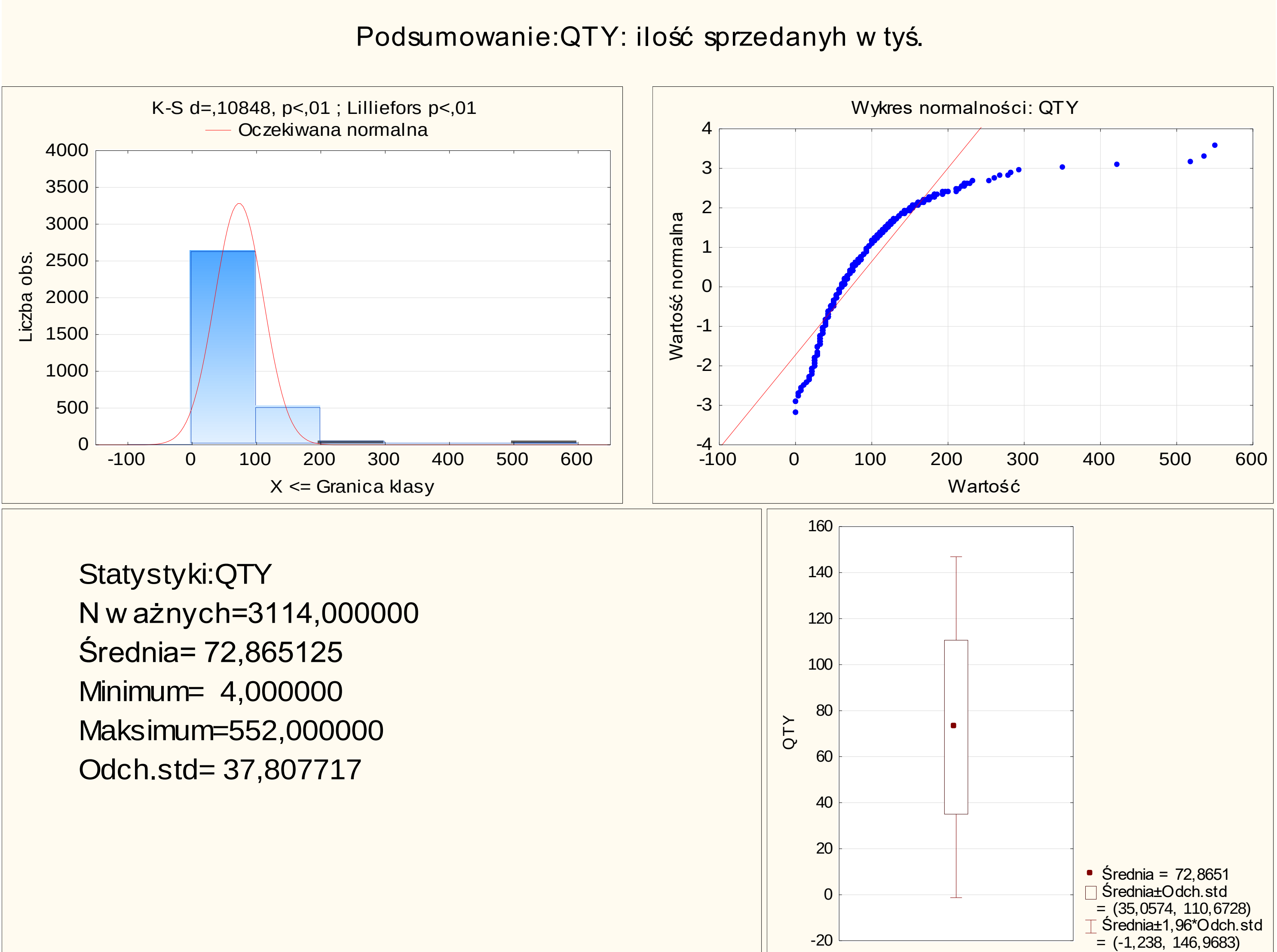
Data analysis



Data analysis



Data analysis



Podsumowanie:QTY: ilość sprzedanyh w tyś.

Statystyki:QTY
N ważnych=3114,000000
Średnia= 72,865125
Minimum= 4,000000
Maksimum=552,000000
Odch.std= 37,807717

Autocorrelation

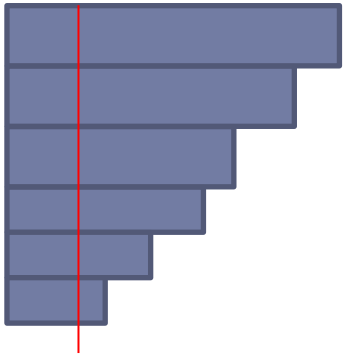
Autocorrelation correlogram. Seasonal patterns of time series can be examined via correlograms. The correlogram (autocorrelogram) displays graphically and numerically the autocorrelation function (*ACF*), that is, serial correlation coefficients (and their standard errors) for consecutive lags in a specified range of lags (e.g., 1 through 30). Ranges of two standard errors for each lag are usually marked in correlograms but typically the size of auto correlation is of more interest than its reliability because we are usually interested only in very strong (and thus highly significant) autocorrelations.

Partial autocorrelations

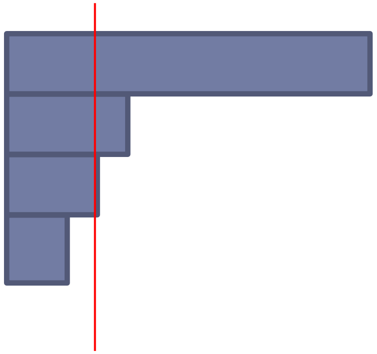
Partial autocorrelations. Another useful method to examine serial dependencies is to examine the partial autocorrelation function (*PACF*) - an extension of autocorrelation, where the dependence on the intermediate elements (those *within* the lag) is removed. In other words, the partial autocorrelation is similar to autocorrelation, except that when calculating it, the (auto) correlations with all the elements within the lag are partially out.

If a lag of 1 is specified (i.e., there are no intermediate elements within the lag), then partial autocorrelation is equivalent to autocorrelation. In a sense, the partial autocorrelation provides a "cleaner" picture of serial dependencies for individual lags (not confounded by other serial dependencies).

Autocorelation



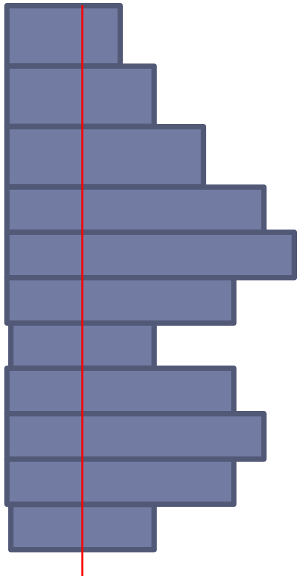
ACF



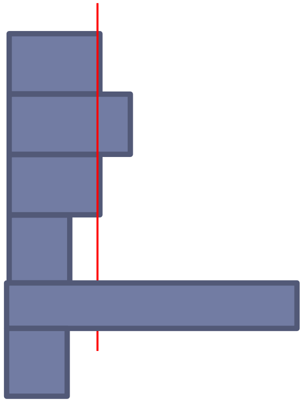
PACF

TREND

Autocorelation



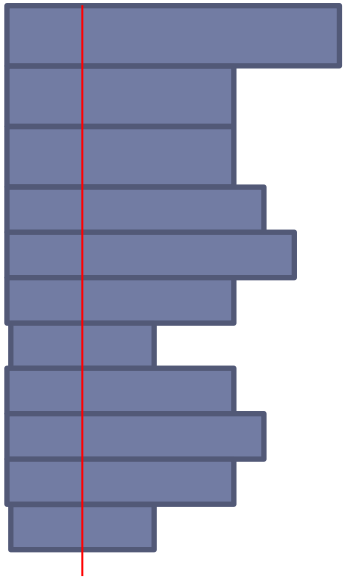
ACF



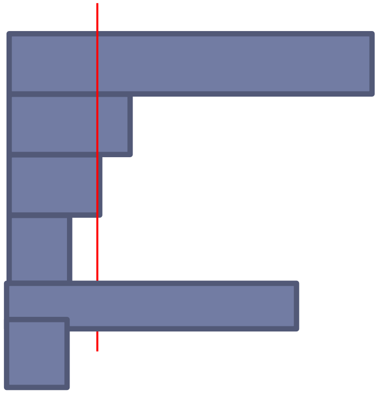
PACF

seasonality

Autocorelation



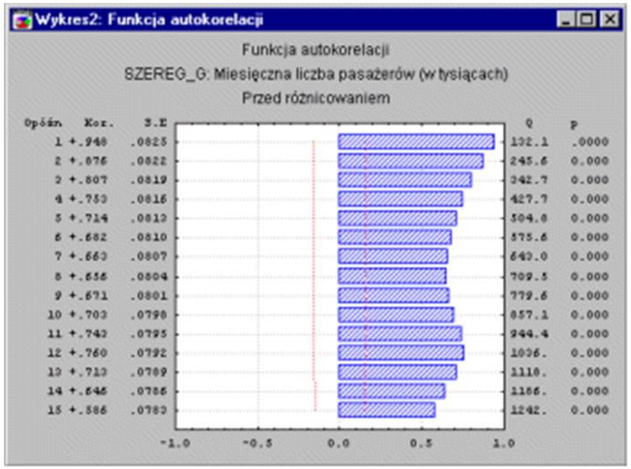
ACF



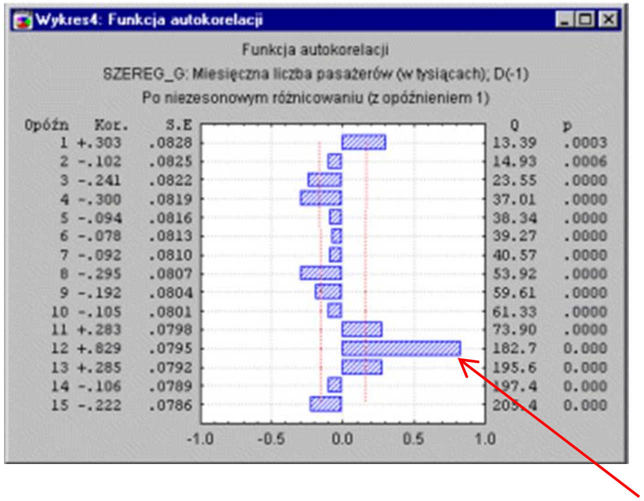
PACF

TREND + SEASONALITY

Autocorrelations/ Partial autocorrelations

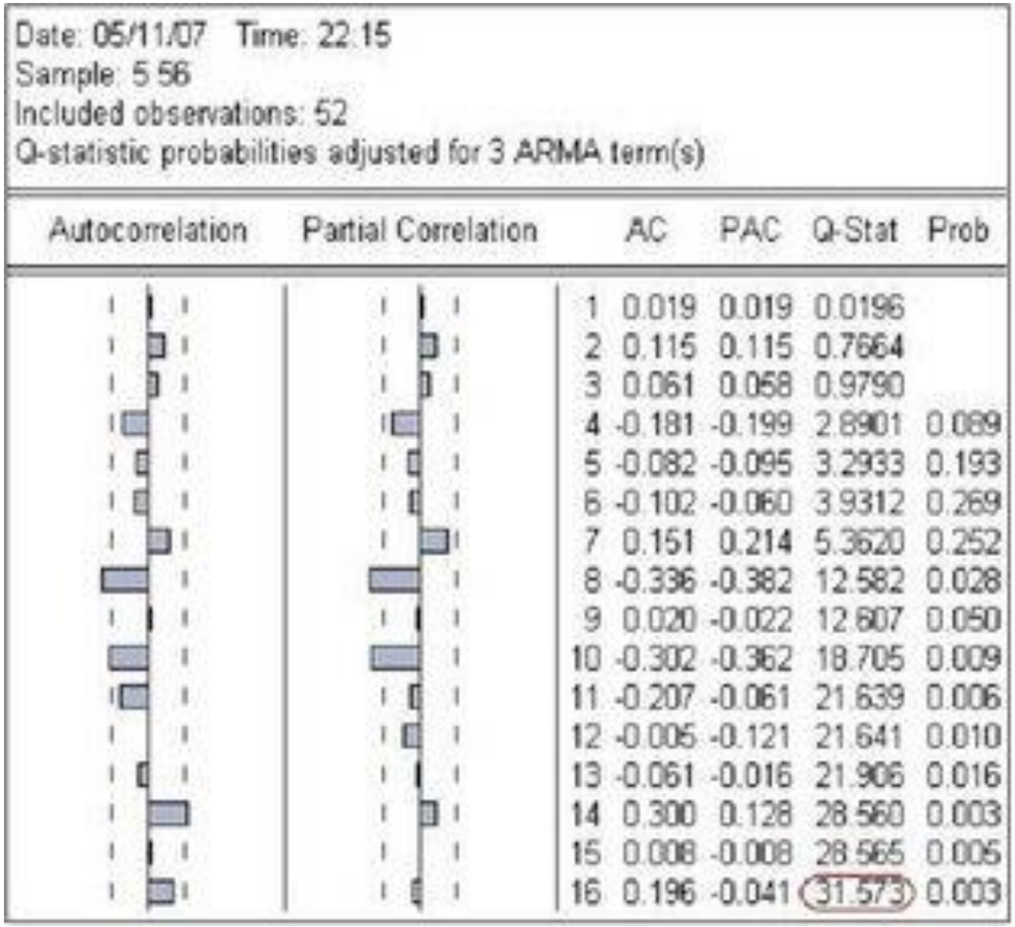

Wykres2: Funkcja autokorelacji
Funkcja autokorelacji
SZEREG_G: Miesięczna liczba pasażerów (w tysiącach)
Przed różnicowaniem
-1.0
-0.5
0.0
0.5
1.0

Opóźn	Kor.	S.E	Q	p
1	+.948	.0825	132.1	.0000
2	+.876	.0822	245.6	0.000
3	+.807	.0819	342.7	0.000
4	+.753	.0816	427.7	0.000
5	+.714	.0813	504.8	0.000
6	+.682	.0810	575.6	0.000
7	+.663	.0807	643.0	0.000
8	+.656	.0804	709.5	0.000
9	+.671	.0801	779.6	0.000
10	+.703	.0798	857.1	0.000
11	+.743	.0795	944.4	0.000
12	+.760	.0792	1036.	0.000
13	+.713	.0789	1118.	0.000
14	+.646	.0786	1186.	0.000
15	+.586	.0783	1242.	0.000


Wykres4: Funkcja autokorelacji
Funkcja autokorelacji
SZEREG_G: Miesięczna liczba pasażerów (w tysiącach); D(-1)
Po niezesonowym różnicowaniu (z opóźnieniem 1)
-1.0
-0.5
0.0
0.5
1.0

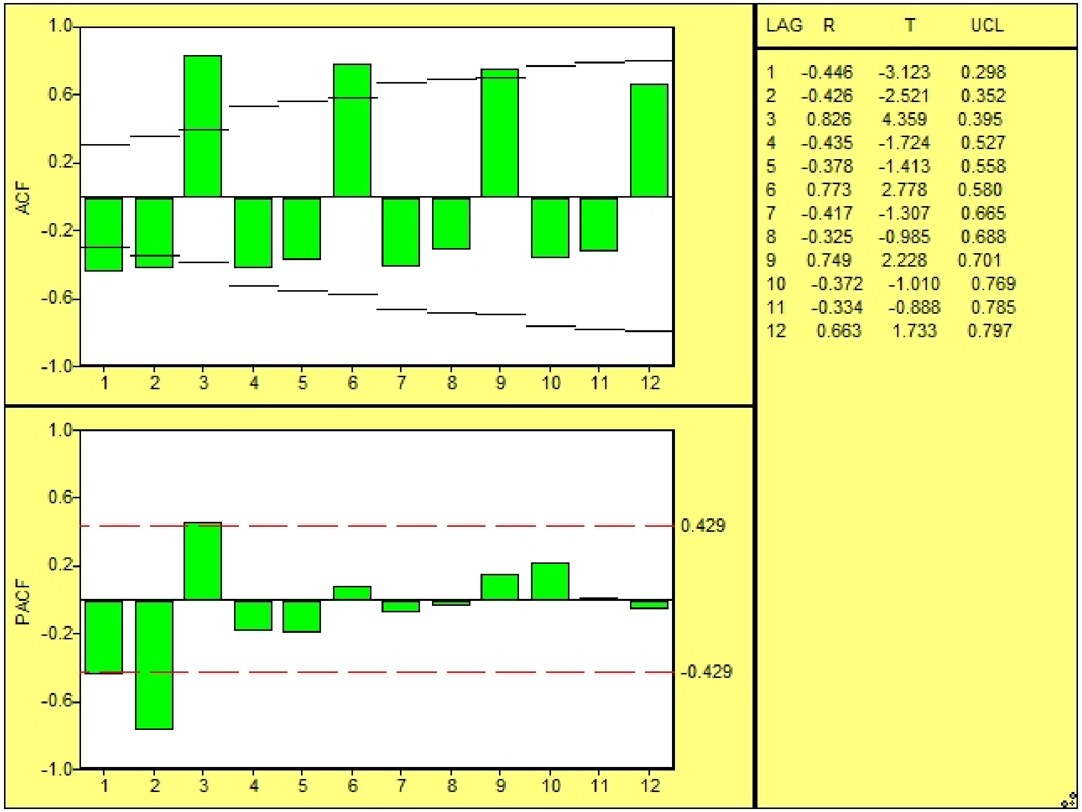
Opóźn	Kor.	S.E	Q	p
1	+.303	.0828	13.39	.0003
2	-.102	.0825	14.93	.0006
3	-.241	.0822	23.55	.0000
4	-.300	.0819	37.01	.0000
5	-.094	.0816	38.34	.0000
6	-.078	.0813	39.27	.0000
7	-.092	.0810	40.57	.0000
8	-.295	.0807	53.92	.0000
9	-.192	.0804	59.61	.0000
10	-.105	.0801	61.33	.0000
11	+.283	.0798	73.90	.0000
12	+.829	.0795	182.7	0.000
13	+.285	.0792	195.6	0.000
14	-.106	.0789	197.4	0.000
15	-.222	.0786	205.4	0.000

Brak trendu i sezonowości


Date: 05/11/07 Time: 22:15
Sample: 5 56
Included observations: 52
Q-statistic probabilities adjusted for 3 ARMA term(s)

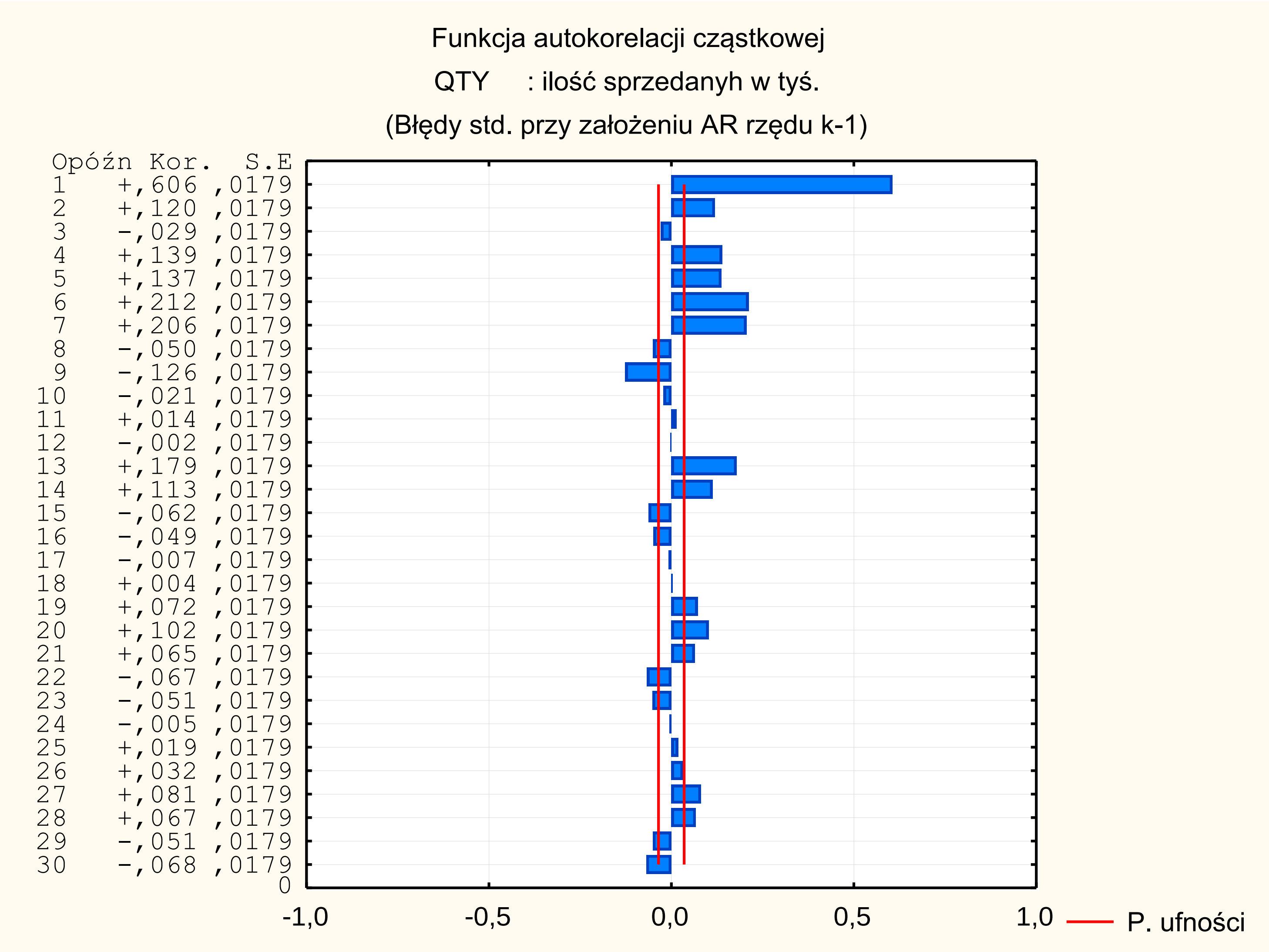
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.019	0.019	0.0196	
		2	0.115	0.115	0.7664	
		3	0.061	0.058	0.9790	
		4	-0.181	-0.199	2.8901	0.089
		5	-0.082	-0.095	3.2933	0.193
		6	-0.102	-0.060	3.9312	0.269
		7	0.151	0.214	5.3620	0.252
		8	-0.336	-0.382	12.582	0.028
		9	0.020	-0.022	12.607	0.050
		10	-0.302	-0.362	18.705	0.009
		11	-0.207	-0.061	21.639	0.006
		12	-0.005	-0.121	21.641	0.010
		13	-0.061	-0.016	21.906	0.016
		14	0.300	0.128	28.560	0.003
		15	0.008	-0.008	28.565	0.005
		16	0.196	-0.041	31.573	0.003

Autocorrelations/ Partial autocorrelations



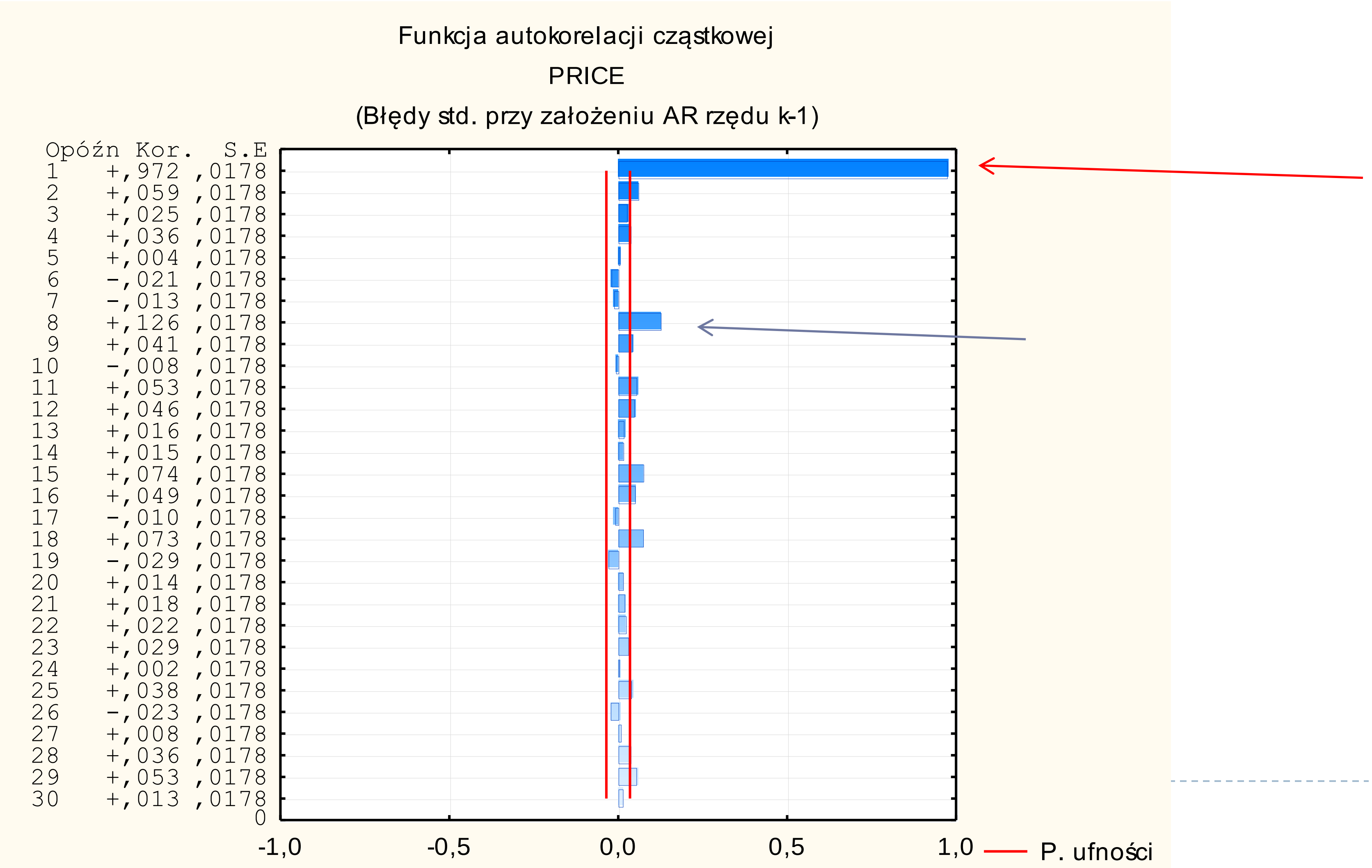
LAG	R	T	UCL
1	-0.446	-3.123	0.298
2	-0.426	-2.521	0.352
3	0.826	4.359	0.395
4	-0.435	-1.724	0.527
5	-0.378	-1.413	0.558
6	0.773	2.778	0.580
7	-0.417	-1.307	0.665
8	-0.325	-0.985	0.688
9	0.749	2.228	0.701
10	-0.372	-1.010	0.769
11	-0.334	-0.888	0.785
12	0.663	1.733	0.797

Partial autocorrelations



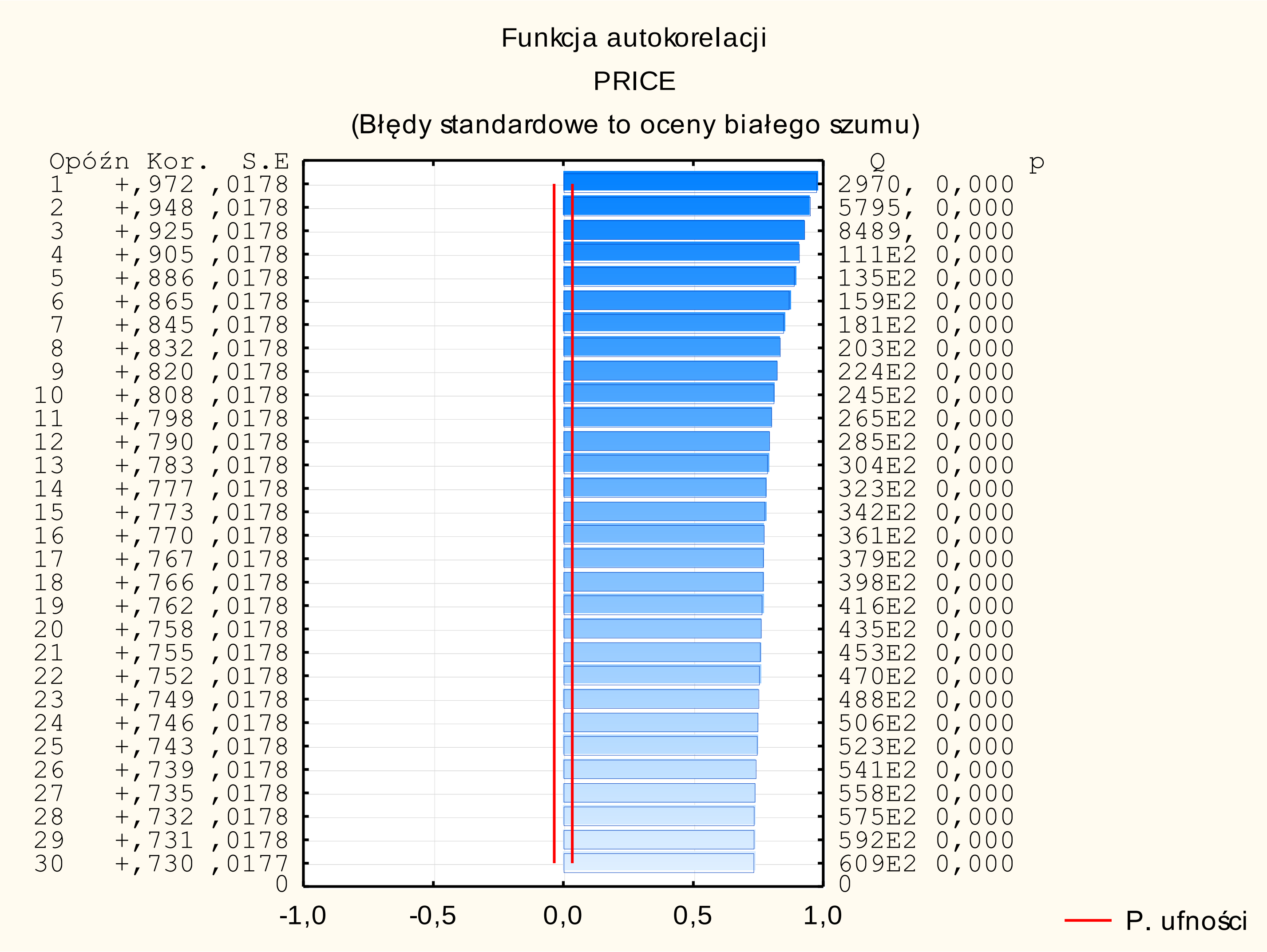
```
Opóźn Kor.  S.E
 1   +,606 ,0179
 2   +,120 ,0179
 3   -,029 ,0179
 4   +,139 ,0179
 5   +,137 ,0179
 6   +,212 ,0179
 7   +,206 ,0179
 8   -,050 ,0179
 9   -,126 ,0179
10   -,021 ,0179
11   +,014 ,0179
12   -,002 ,0179
13   +,179 ,0179
14   +,113 ,0179
15   -,062 ,0179
16   -,049 ,0179
17   -,007 ,0179
18   +,004 ,0179
19   +,072 ,0179
20   +,102 ,0179
21   +,065 ,0179
22   -,067 ,0179
23   -,051 ,0179
24   -,005 ,0179
25   +,019 ,0179
26   +,032 ,0179
27   +,081 ,0179
28   +,067 ,0179
29   -,051 ,0179
30   -,068 ,0179
```

Partial autocorrelations



Opóźn	Kor.	S.E
1	+,972	,0178
2	+,059	,0178
3	+,025	,0178
4	+,036	,0178
5	+,004	,0178
6	-,021	,0178
7	-,013	,0178
8	+,126	,0178
9	+,041	,0178
10	-,008	,0178
11	+,053	,0178
12	+,046	,0178
13	+,016	,0178
14	+,015	,0178
15	+,074	,0178
16	+,049	,0178
17	-,010	,0178
18	+,073	,0178
19	-,029	,0178
20	+,014	,0178
21	+,018	,0178
22	+,022	,0178
23	+,029	,0178
24	+,002	,0178
25	+,038	,0178
26	-,023	,0178
27	+,008	,0178
28	+,036	,0178
29	+,053	,0178
30	+,013	,0178

Autocorrelations



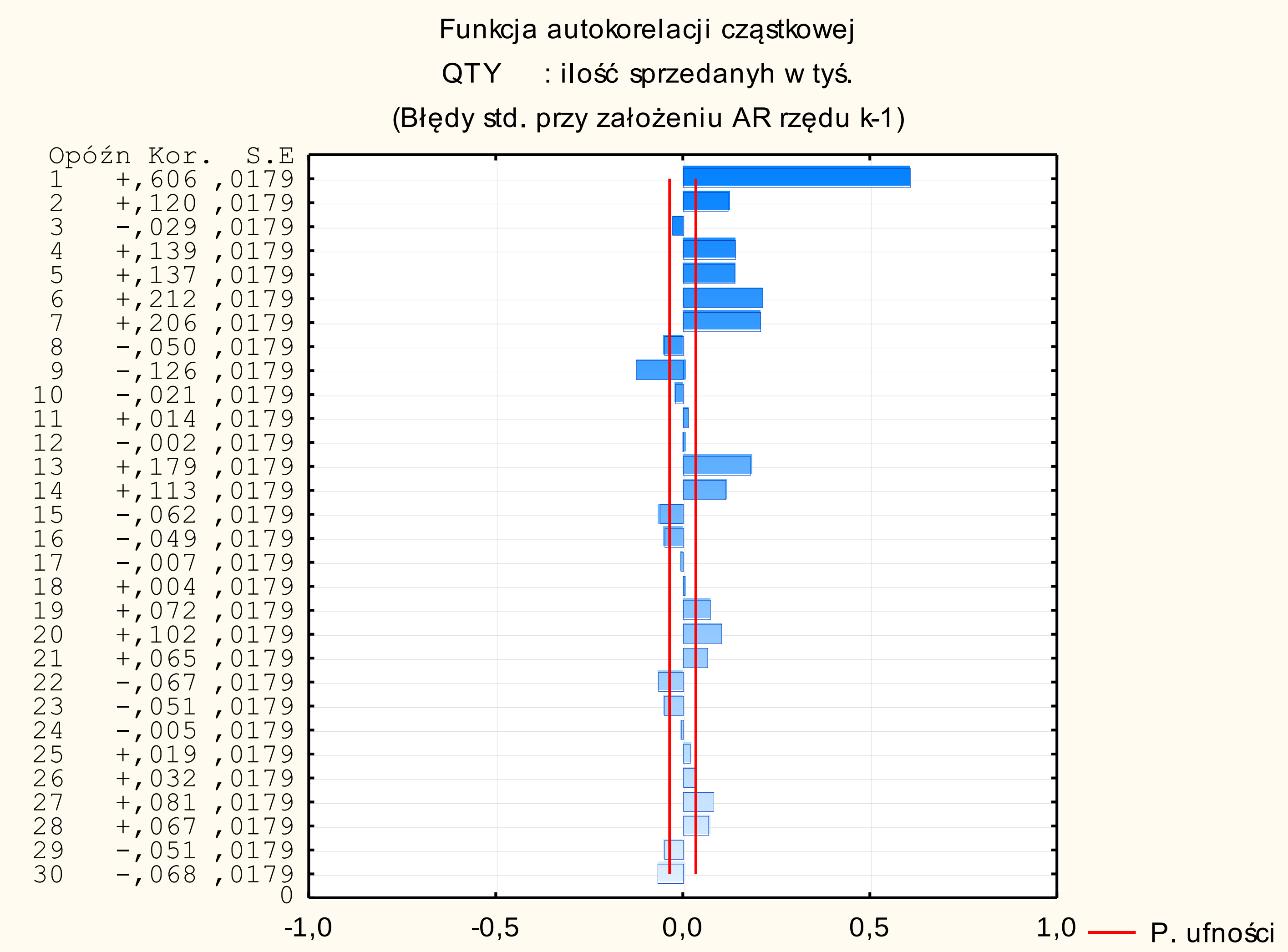
Funkcja autokorelacji

PRICE

(Błędy standardowe to oceny białego szumu)

Opóźn	Kor.	S.E	Q	p
1	+,972	,0178	2970,	0,000
2	+,948	,0178	5795,	0,000
3	+,925	,0178	8489,	0,000
4	+,905	,0178	111E2	0,000
5	+,886	,0178	135E2	0,000
6	+,865	,0178	159E2	0,000
7	+,845	,0178	181E2	0,000
8	+,832	,0178	203E2	0,000
9	+,820	,0178	224E2	0,000
10	+,808	,0178	245E2	0,000
11	+,798	,0178	265E2	0,000
12	+,790	,0178	285E2	0,000
13	+,783	,0178	304E2	0,000
14	+,777	,0178	323E2	0,000
15	+,773	,0178	342E2	0,000
16	+,770	,0178	361E2	0,000
17	+,767	,0178	379E2	0,000
18	+,766	,0178	398E2	0,000
19	+,762	,0178	416E2	0,000
20	+,758	,0178	435E2	0,000
21	+,755	,0178	453E2	0,000
22	+,752	,0178	470E2	0,000
23	+,749	,0178	488E2	0,000
24	+,746	,0178	506E2	0,000
25	+,743	,0178	523E2	0,000
26	+,739	,0178	541E2	0,000
27	+,735	,0178	558E2	0,000
28	+,732	,0178	575E2	0,000
29	+,731	,0178	592E2	0,000
30	+,730	,0177	609E2	0,000
		0	0	

Partial autocorrelations



Funkcja autokorelacji cząstkowej

QTY : ilość sprzedanyh w tyś.

(Błędy std. przy założeniu AR rzędu k-1)

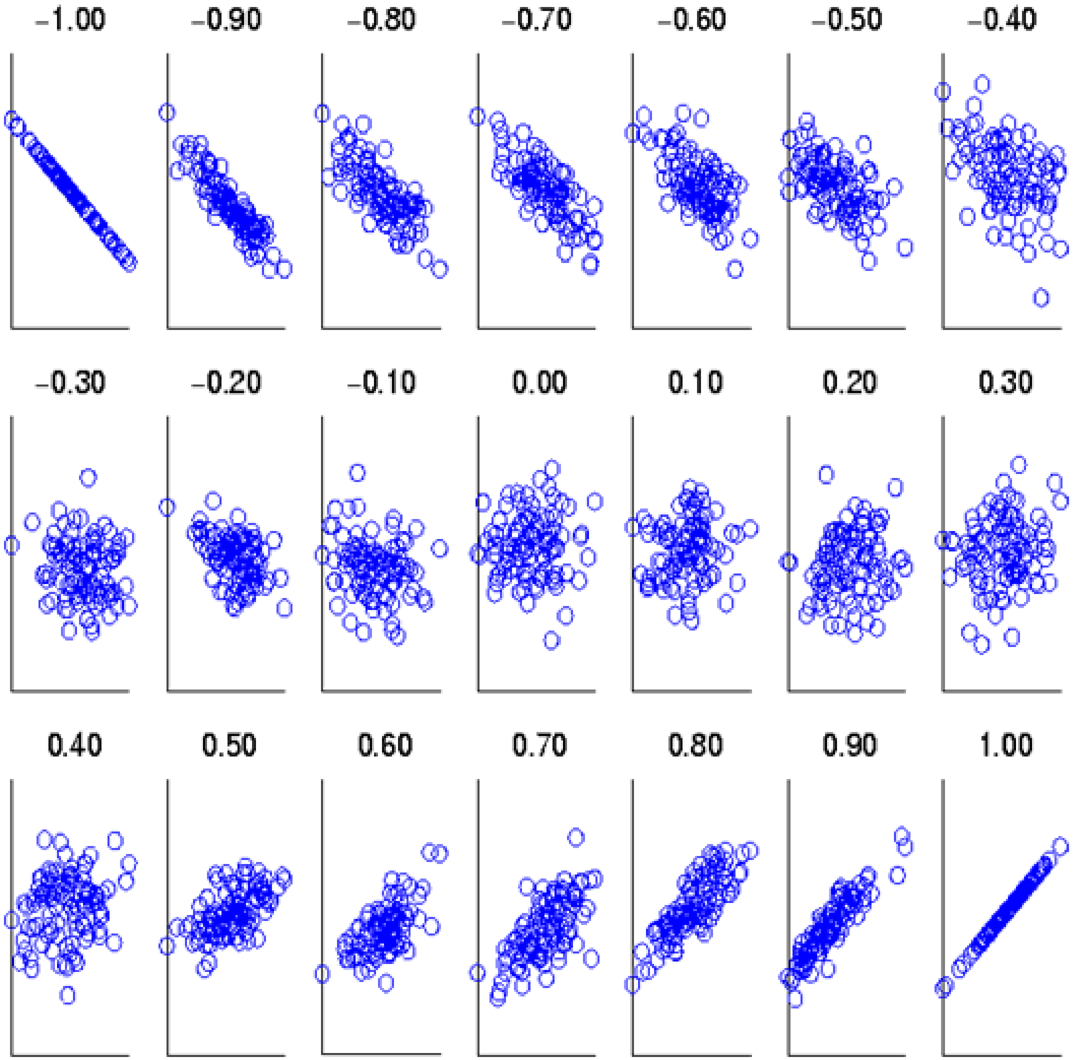
Opóźn	Kor.	S.E
1	+,606	,0179
2	+,120	,0179
3	-,029	,0179
4	+,139	,0179
5	+,137	,0179
6	+,212	,0179
7	+,206	,0179
8	-,050	,0179
9	-,126	,0179
10	-,021	,0179
11	+,014	,0179
12	-,002	,0179
13	+,179	,0179
14	+,113	,0179
15	-,062	,0179
16	-,049	,0179
17	-,007	,0179
18	+,004	,0179
19	+,072	,0179
20	+,102	,0179
21	+,065	,0179
22	-,067	,0179
23	-,051	,0179
24	-,005	,0179
25	+,019	,0179
26	+,032	,0179
27	+,081	,0179
28	+,067	,0179
29	-,051	,0179
30	-,068	,0179

Correlation

Correlation is a measure of the relation between two or more variables. The measurement scales used should be at least interval scales, but other correlation coefficients are available to handle other types of data. Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect *negative* correlation while a value of +1.00 represents a perfect *positive* correlation. A value of 0.00 represents a lack of correlation.

Both datasets should have the same numer of observations.

Correlation

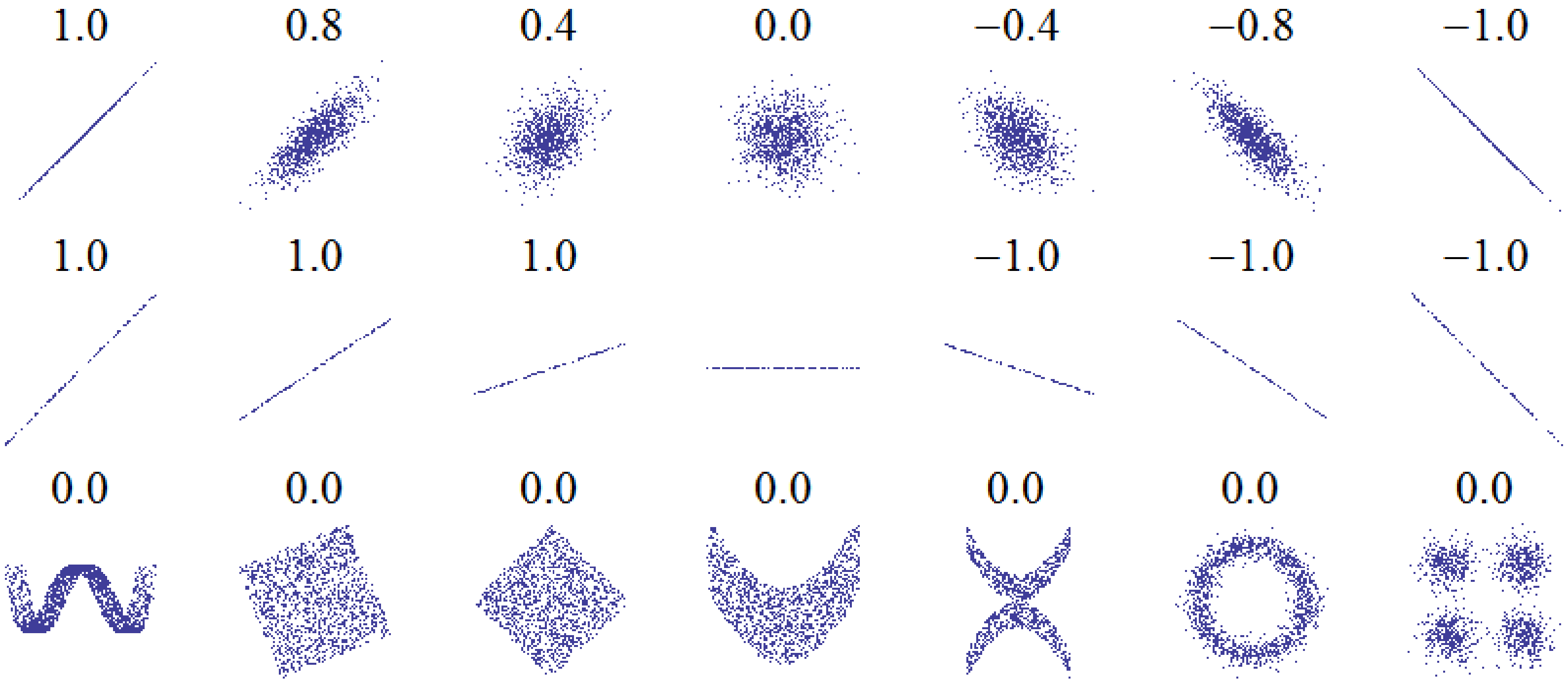


Pearson linear correlation coefficient

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}},$$

Pearson linear correlation coefficient



Nonparametric correlations

- Could be used to find nonlinear dependences in two datasets for many data types.
- Correlation coefficients can range from -1.00 to +1.00.
- The following are three types of commonly used nonparametric correlation coefficients (Spearman R, Kendall Tau, and Gamma coefficients).

Spearman R.

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

$$d_i = \mathrm{R}\,x_i - \mathrm{R}\,y_i$$

di is a difference between ranks variables *x* and *y* for *i* observation

Spearman R.

IQ, X_i	Hours of TV per week, Y_i	Rank x_i	Rank y_i	d_i	d_i^2
86	0	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Spearman R.

$$\rho = 1 - \frac{6 \times 194}{10(10^2 - 1)}$$

ρ = -29/165 = −0.175757575... With p = 0.627188 (using t distribution)

tau Kendal correlation

To calculate Kendal correlations, data should be compiled into all possible pairs and then divide these pairs into three possible categories:

- **Concordant pairs-** ordered in the same way (P),
- **Discordant pairs** – ordered differently (Q),
- **Bonded pairs** – the same values in pair for both pairs (T).

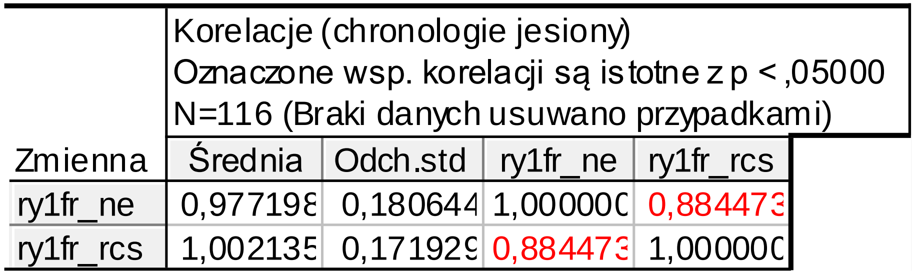
tau Kendal correlation

$$\tau = \frac{P - Q}{P + Q + T}$$

$$P + Q + T = \binom{N}{2} = \frac{N(N-1)}{2}$$

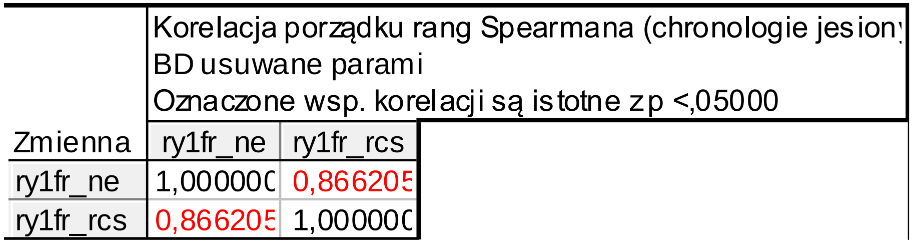
$$\tau = 2\frac{P - Q}{N(N-1)}$$

Correlations



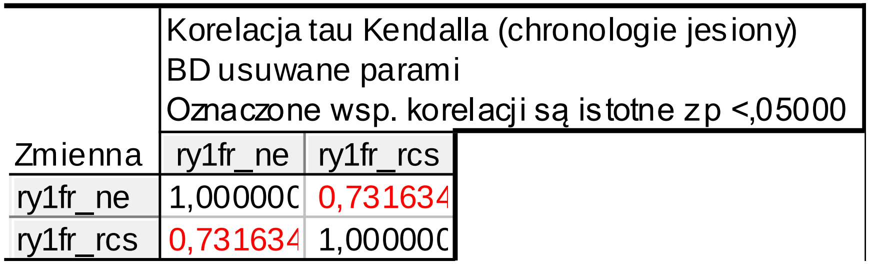
Korelacje (chronologie jesiony)
Oznaczone wsp. korelacji są istotne z p < ,05000
N=116 (Braki danych usuwano przypadkami)

Zmienna	Średnia	Odch.std	ry1fr_ne	ry1fr_rcs
ry1fr_ne	0,977198	0,180644	1,000000	0,884473
ry1fr_rcs	1,002135	0,171929	0,884473	1,000000



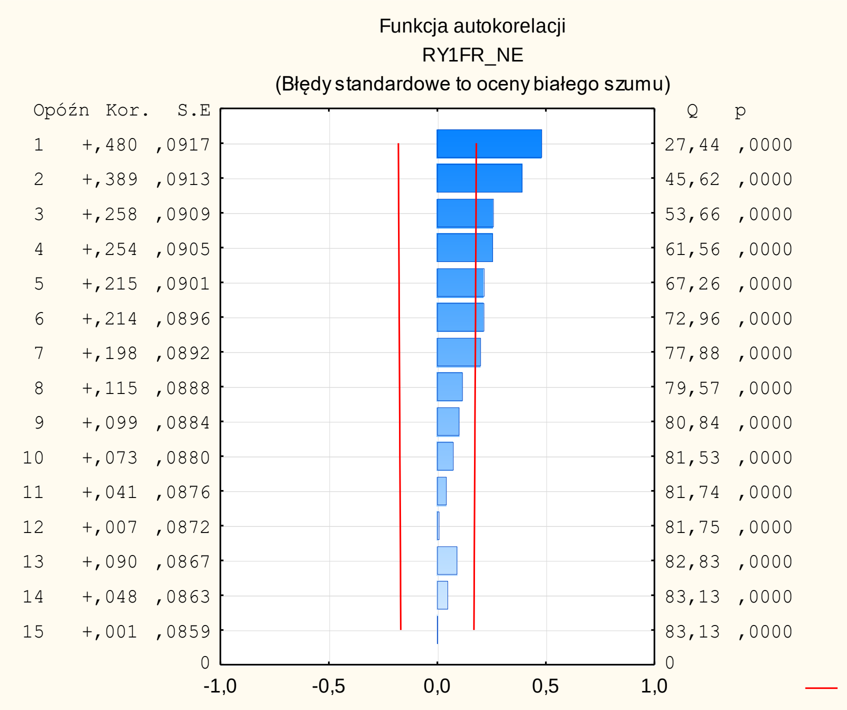
Korelacja porządku rang Spearmana (chronologie jesion
BD usuwane parami
Oznaczone wsp. korelacji są istotne z p <,05000

Zmienna	ry1fr_ne	ry1fr_rcs
ry1fr_ne	1,000000	0,866205
ry1fr_rcs	0,866205	1,000000



Korelacja tau Kendalla (chronologie jesiony)
BD usuwane parami
Oznaczone wsp. korelacji są istotne z p <,05000

Zmienna	ry1fr_ne	ry1fr_rcs
ry1fr_ne	1,000000	0,731634
ry1fr_rcs	0,731634	1,000000



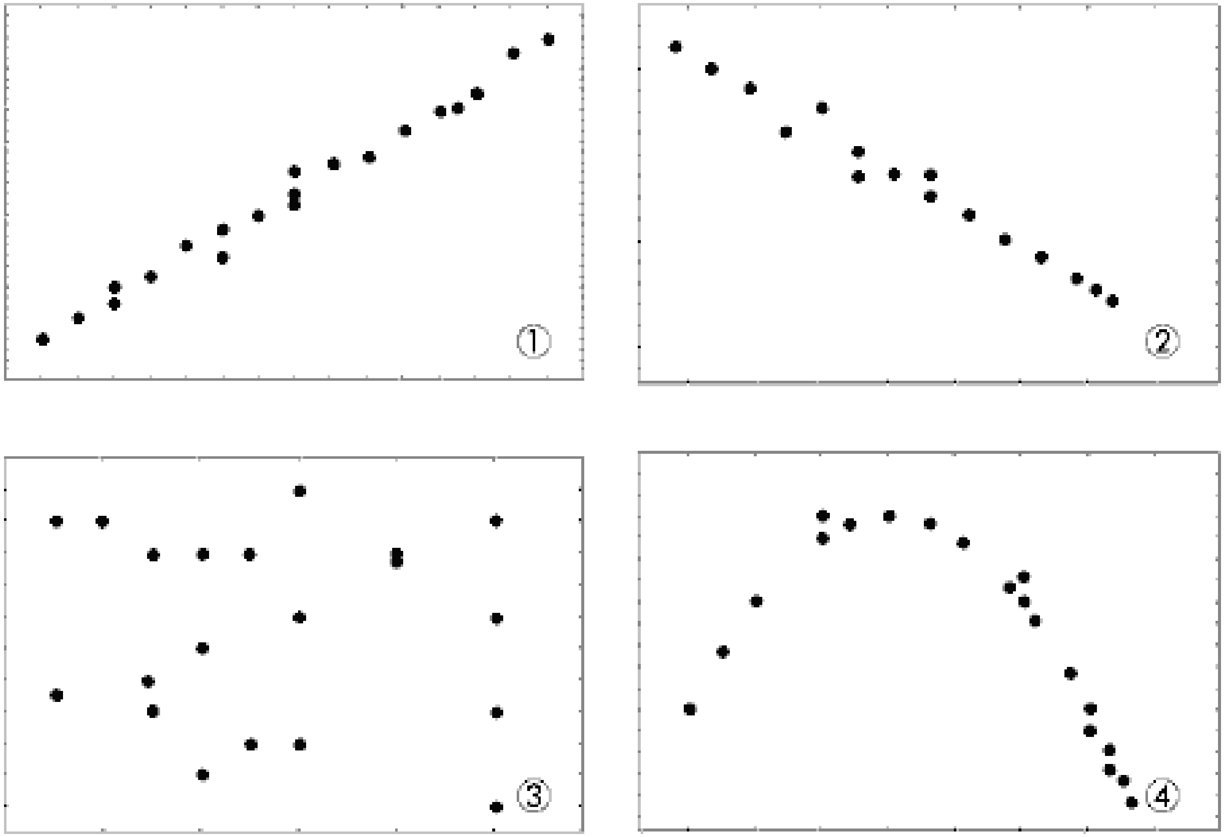
Funkcja autokorelacji
RY1FR_NE
(Błędy standardowe to oceny białego szumu)

Opóźn	Kor.	S.E	Q	p
1	+,480	,0917	27,44	,0000
2	+,389	,0913	45,62	,0000
3	+,258	,0909	53,66	,0000
4	+,254	,0905	61,56	,0000
5	+,215	,0901	67,26	,0000
6	+,214	,0896	72,96	,0000
7	+,198	,0892	77,88	,0000
8	+,115	,0888	79,57	,0000
9	+,099	,0884	80,84	,0000
10	+,073	,0880	81,53	,0000
11	+,041	,0876	81,74	,0000
12	+,007	,0872	81,75	,0000
13	+,090	,0867	82,83	,0000
14	+,048	,0863	83,13	,0000
15	+,001	,0859	83,13	,0000

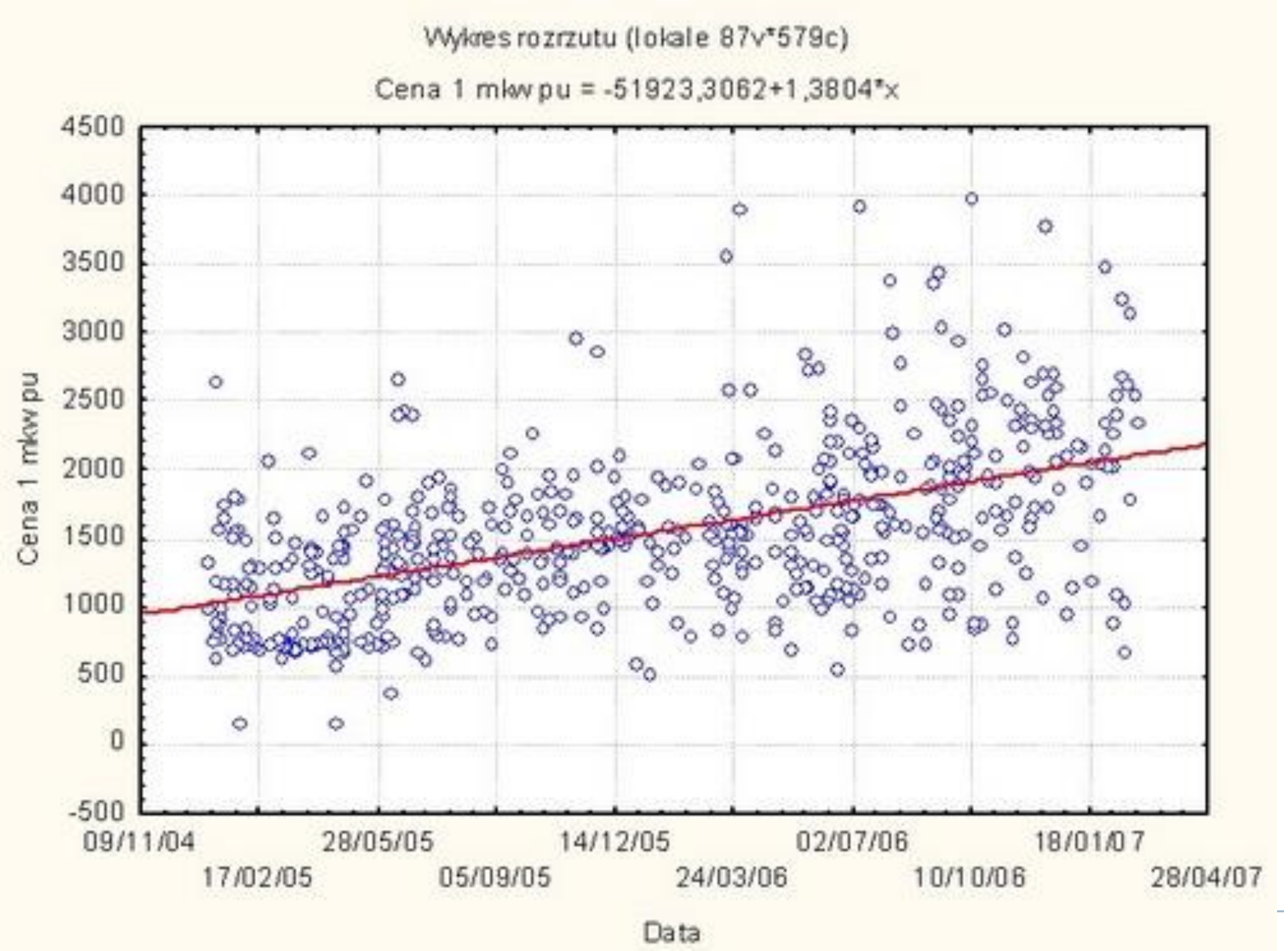
Measures similar to correlation:

- Fi coefficient,
- V Cramer’s coefficient.

Scatter plot- correlation

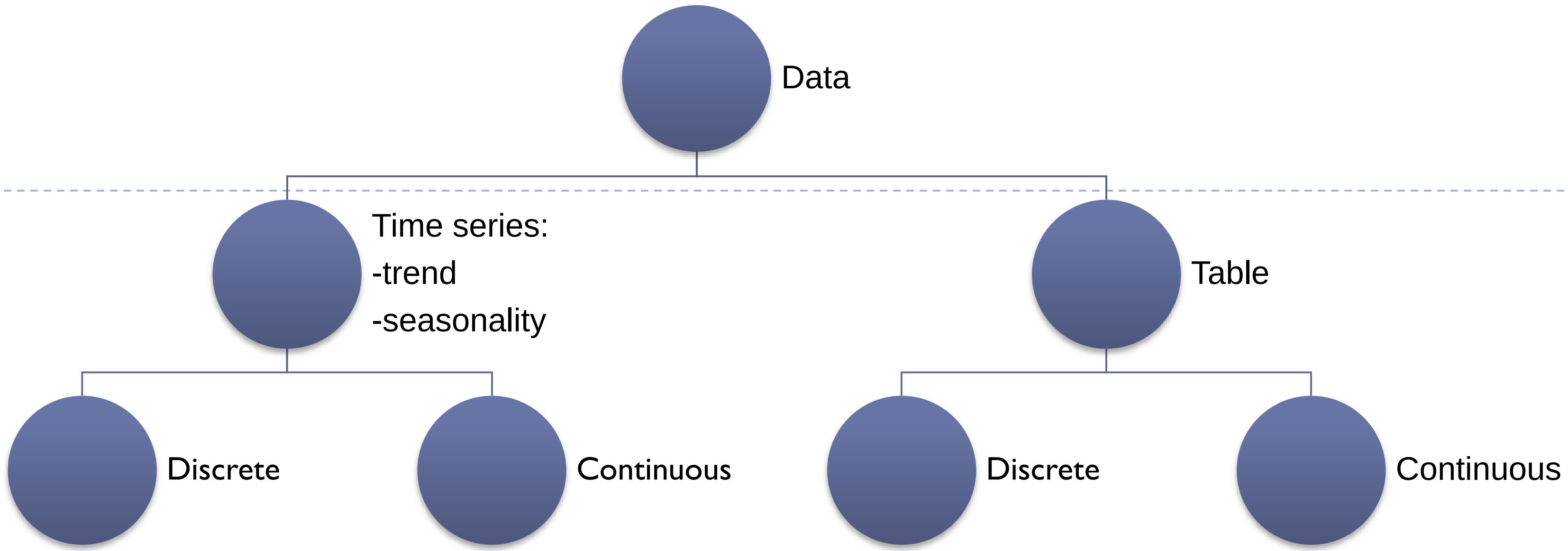


Scatter plot- correlation



More correlation coefficients

The Point-Biserial Correlation Coefficient is a correlation measure of the strength of association between a continuous-level variable (ratio or interval data) and a binary variable,
The phi coefficient (or mean square contingency coefficient) is a measure of association for two binary variables. It is known as the Matthews correlation coefficient (MCC),
Tetrachoric Correlation: Used to calculate the correlation between binary categorical variables.
Polychoric Correlation: Used to calculate the correlation between ordinal categorical variables.
Cramer’s V: Used to calculate the correlation between nominal categorical variables.



Data

Time series:
-trend
-seasonality

Table

Discrete

Continuous

Discrete

Continuous

NO TREND, SEAS.
-mode
-random v.
-choosen v.

TREND, SEAS.
-interpolation from Neighborhood v./seas.

NO TREND, SEAS.
-mean, median…
-random v.
-choosen v.

TREND, SEAS.
-interpolation from Neighborhood v./seas.

-mode
-random v.
-choosen v.
-row deletion

-mean, median
-random v.
-choosen v.
-row deletion

Pytania?

Do zastanowienia się:
-Jaki jest wpływ wykonania preprocessingu na dalsze modelowanie
-Jaką różnicę w wynikach możemy uzyskać pracując z próbą zamiast populacji i czym jest ona spowodowana

Na ćwiczenia wiadomości:
-Preprocessing w zależności od typu danych i ich formy
-W jaki sposób analizujemy dane (na co zwracamy uwagę, jakie analizy wykonujemy)

Lecture 3:
Regression model