

Chapter 1: Foreseeing Variable Problems When Building ML Models

← → ↻ https://archive.ics.uci.edu/ml/machine-learning-databases/kddcup98-mld/epsilon_mirror/

Index of /ml/machine-learning-databases/kddcup98-mld/epsilon_mirror

- [Parent Directory](#)
- [cup98dic.txt](#)
- [cup98doc.txt](#)
- [cup98lrm.txt.Z](#)
- [cup98lrm.zip](#) 
- [cup98que.txt](#)
- [cup98val.txt.Z](#)
- [cup98val.zip](#)
- [instruct.txt](#)
- [readme](#)
- [testing.txt](#)
- [valtarget.readme](#)
- [valtarget.txt](#)

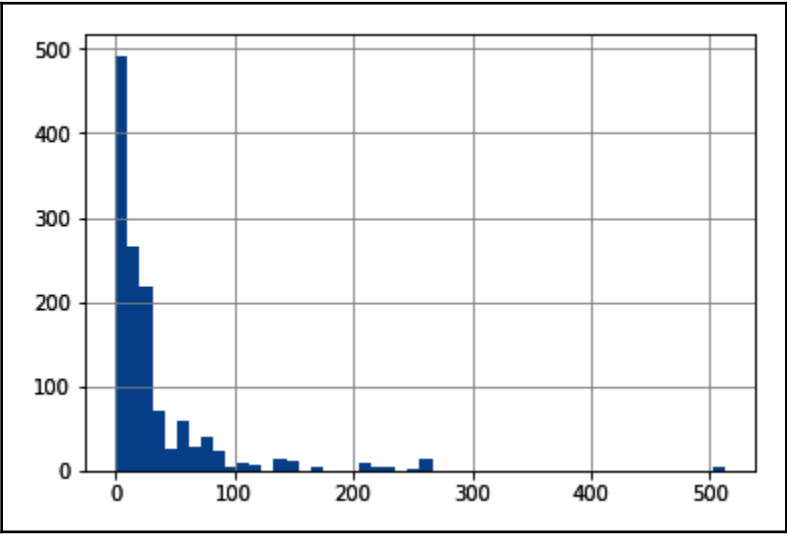
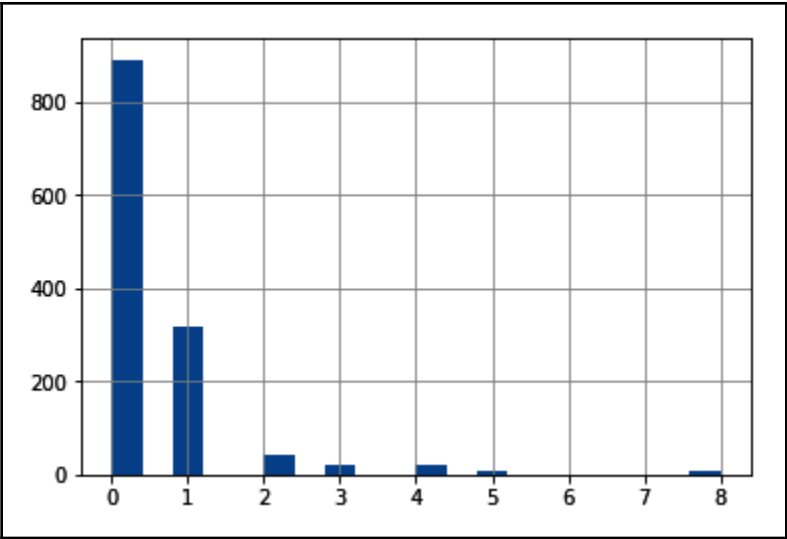
Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips SVN/1.7.14 Phusion_Passenger/4.0.53 mod_perl/2.0.10 Perl/v5.16.3 Server at archive.ics.uci.edu Port 443

← → ↻ <https://archive.ics.uci.edu/ml/machine-learning-databases/car/>

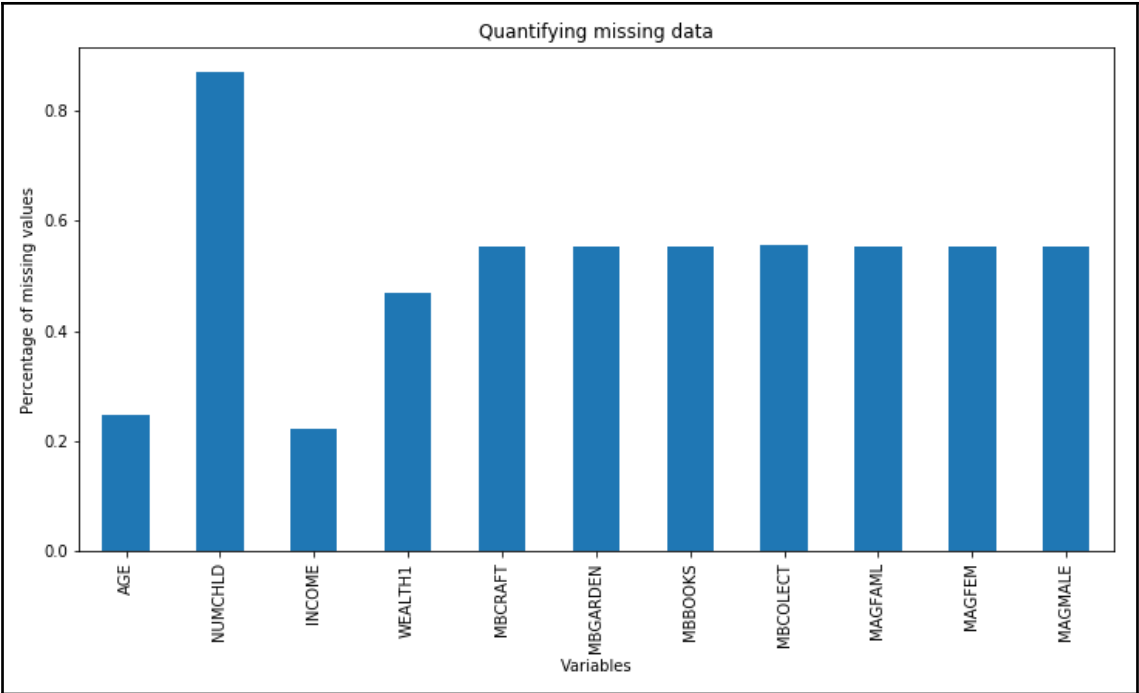
Index of /ml/machine-learning-databases/car

- [Parent Directory](#)
- [car.c45-names](#)
- [car.data](#) 
- [car.names](#)

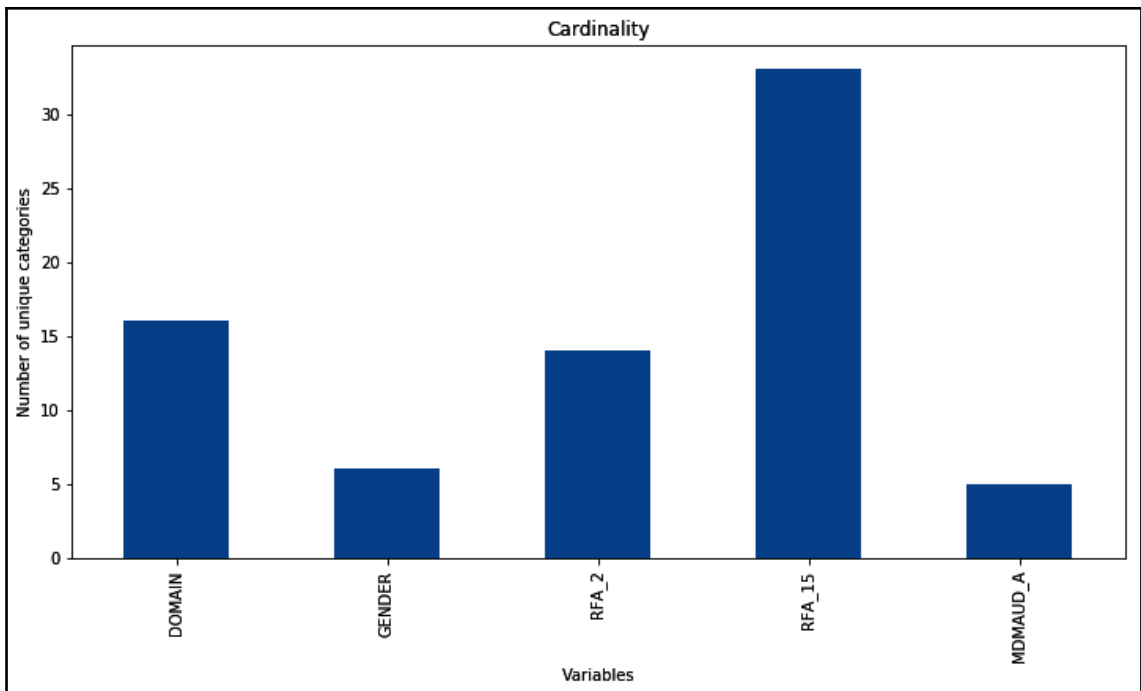
Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips SVN/1.7.14 Phusion_Passenger/4.0.53 mod_perl/2.0.10 Perl/v5.16.3 Server at archive.ics.uci.edu Port 443



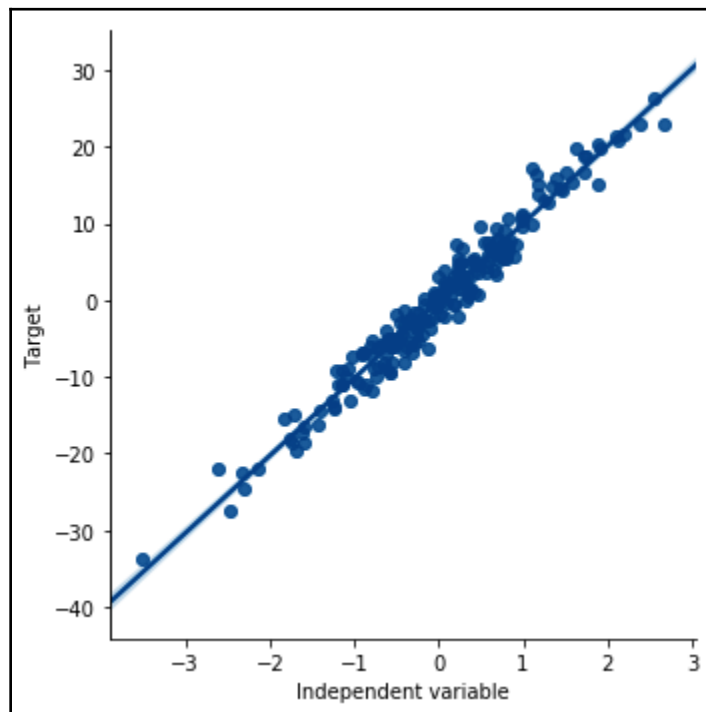
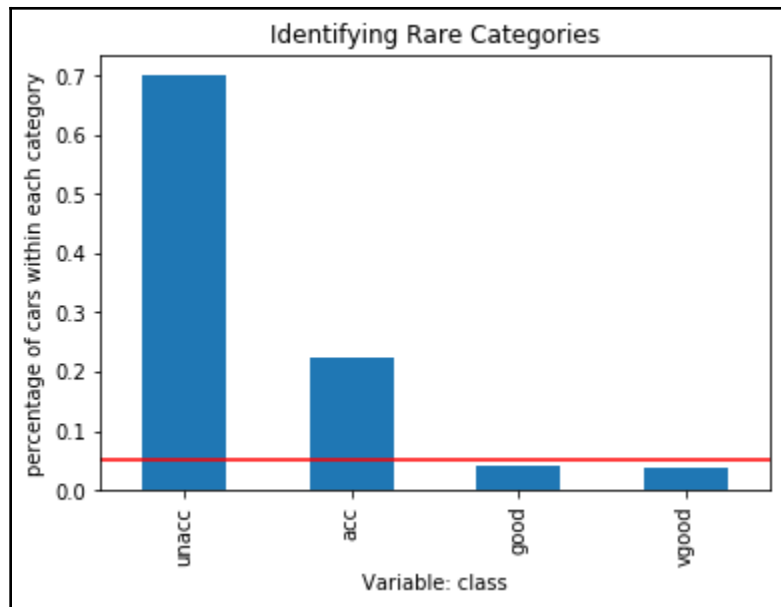
	AGE	NUMCHLD	INCOME	WEALTH1	MBCRAFT	MBGARDEN	MBBOOKS	MBCOLECT	MAGFAML	MAGFEM	MAGMALE
0	60.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	46.0	1.0	6.0	9.0	0.0	0.0	3.0	1.0	1.0	1.0	0.0
2	NaN	NaN	3.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
3	70.0	NaN	1.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	78.0	1.0	3.0	2.0	1.0	0.0	9.0	0.0	4.0	1.0	0.0

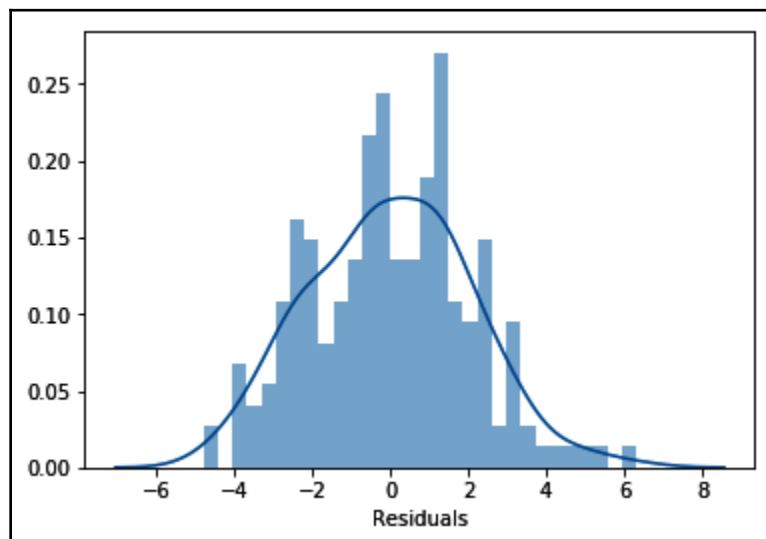
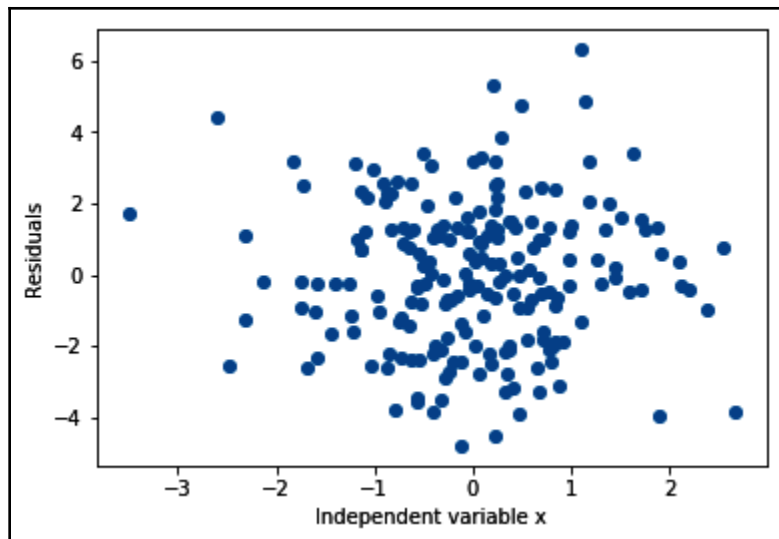


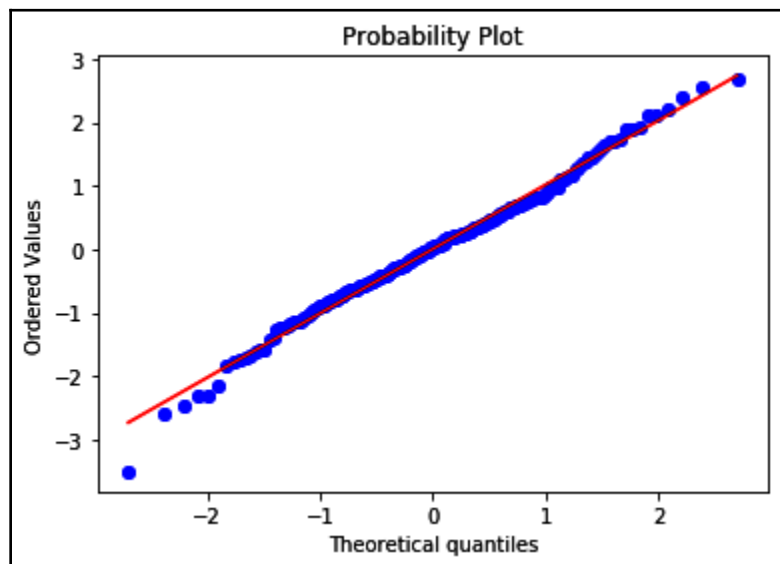
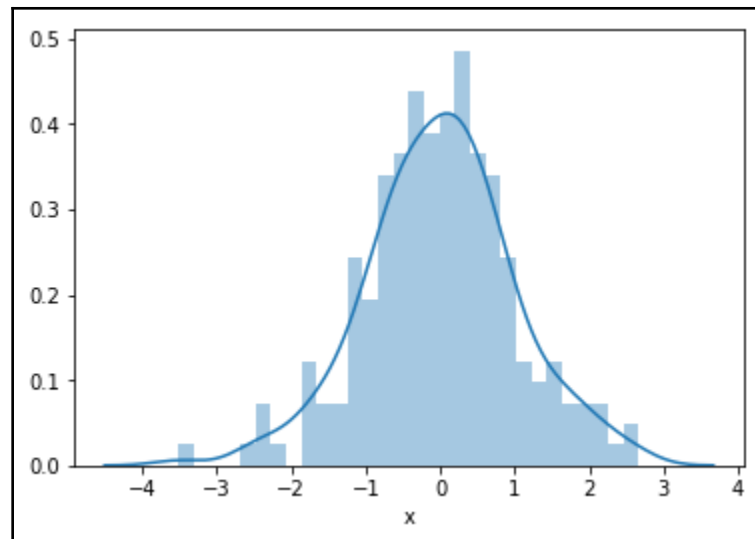
	DOMAIN	GENDER	RFA_2	RFA_15	MDMAUD_A
0	T2	F	L4E	S4E	X
1	S1	M	L2G	NaN	X
2	R2	M	L4E	S4F	X
3	R2	F	L4E	S4E	X
4	S2	F	L2F	NaN	X

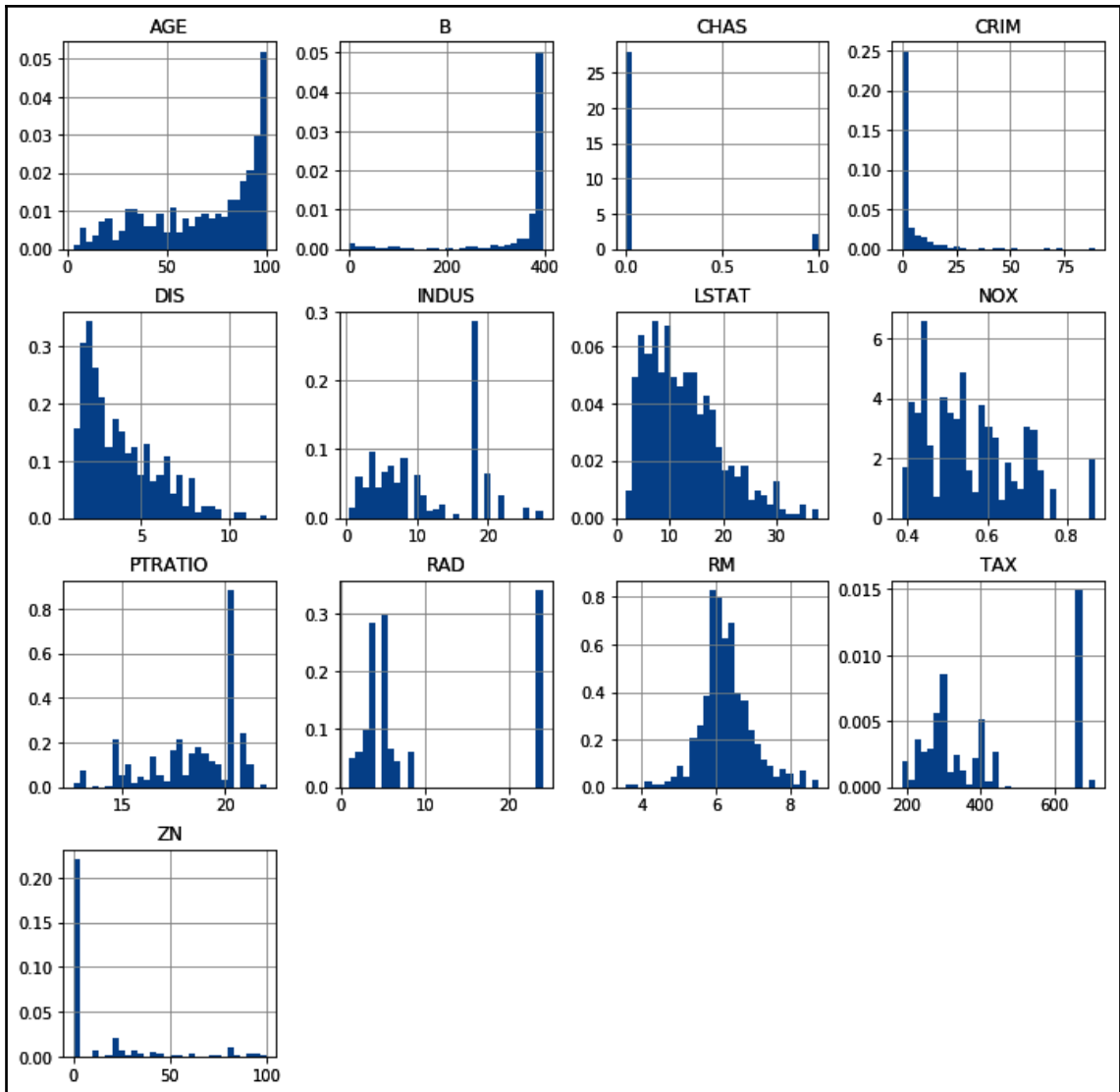


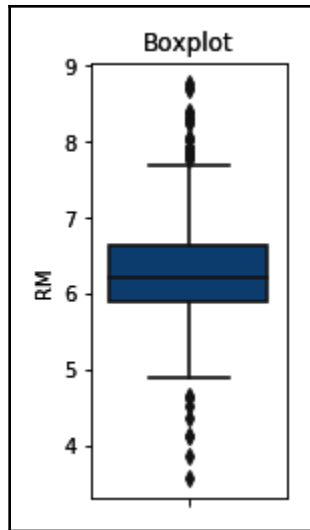
	buying	maint	doors	persons	lug_boot	safety	class
0	vhigh	vhigh	2	2	small	low	unacc
1	vhigh	vhigh	2	2	small	med	unacc
2	vhigh	vhigh	2	2	small	high	unacc
3	vhigh	vhigh	2	2	med	low	unacc
4	vhigh	vhigh	2	2	med	med	unacc











	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000

Chapter 2: Imputing Missing Data

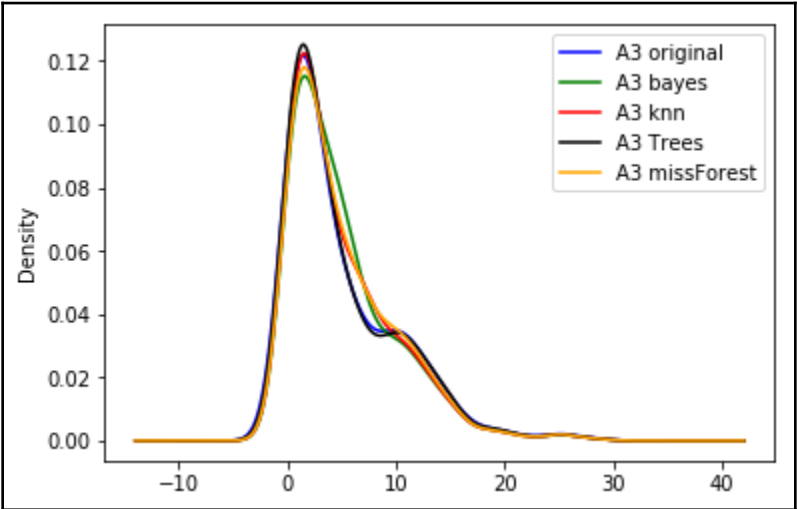
Not secure | archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/

Index of /ml/machine-learning-databases/credit-screening

- [Parent Directory](#)
- [Index](#)
- [credit.lisp](#)
- [credit.names](#)
- [crx.data](#)
- [crx.names](#)

Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips SVN/1.7.14 Phusion_Passenger/4.0.53 mod_perl/2.0.10 Perl/v5.16.3 Server at archive.ics.uci.edu Port 80

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	...	A13	A14	A15	A1_NA	A3_NA	A4_NA	A5_NA	A6_NA	A7_NA	A8_NA
596	a	46.08	3.000	u	g	c	v	2.375	t	t	...	g	396.0	4159	0	0	0	0	0	0	0
303	a	15.92	2.875	u	g	q	v	0.085	f	f	...	g	120.0	0	0	0	0	0	0	0	0
204	b	36.33	2.125	y	p	w	v	0.085	t	t	...	g	50.0	1187	0	0	0	0	0	0	0
351	b	22.17	0.585	y	p	ff	ff	0.000	f	f	...	g	100.0	0	0	0	0	0	0	0	0
118	b	57.83	7.040	u	g	m	v	14.000	t	t	...	g	360.0	1332	0	0	0	0	0	0	0



Chapter 3: Encoding Categorical Variables

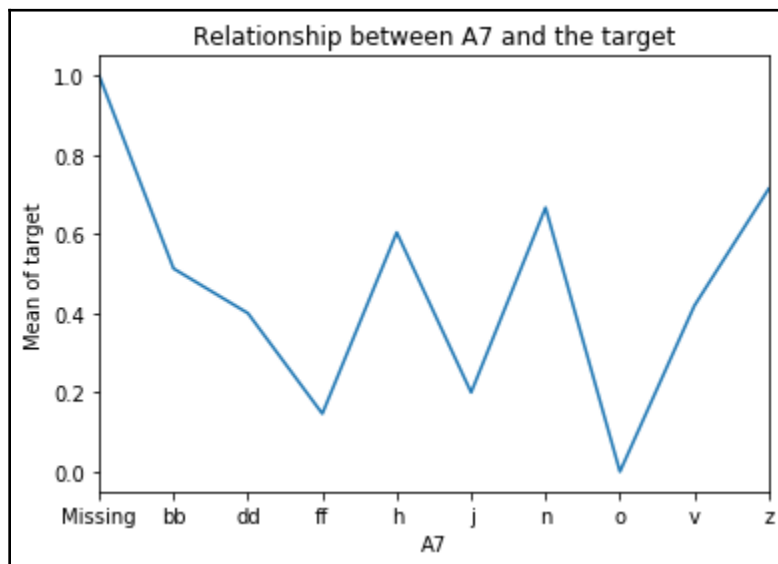
[←](#) [→](#) [↻](#) [🔒](#) Not secure | [archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/](#)

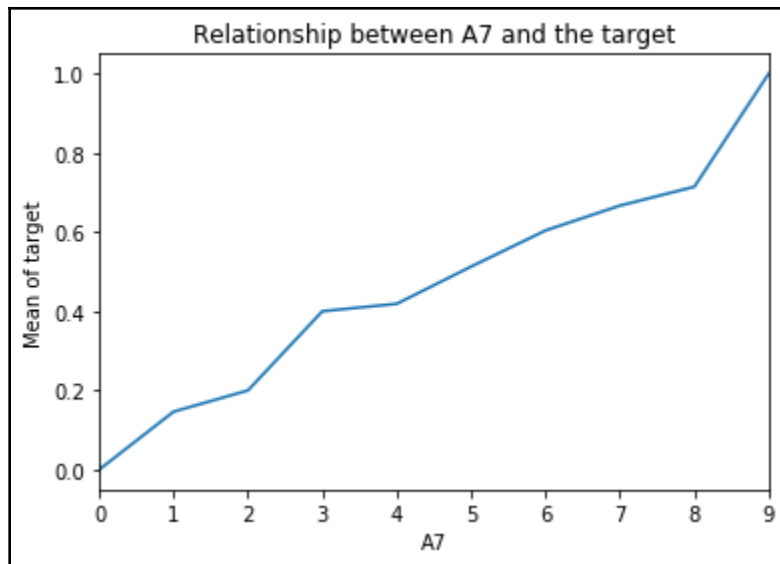
Index of /ml/machine-learning-databases/credit-screening

- [Parent Directory](#)
- [Index](#)
- [credit.lisp](#)
- [credit.names](#)
- [crx.data](#) 
- [crx.names](#)

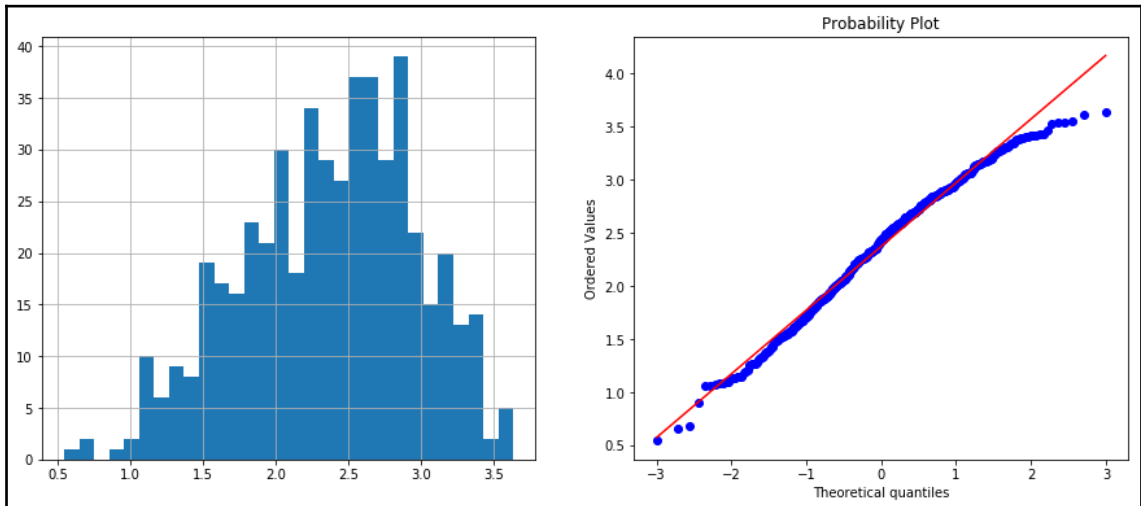
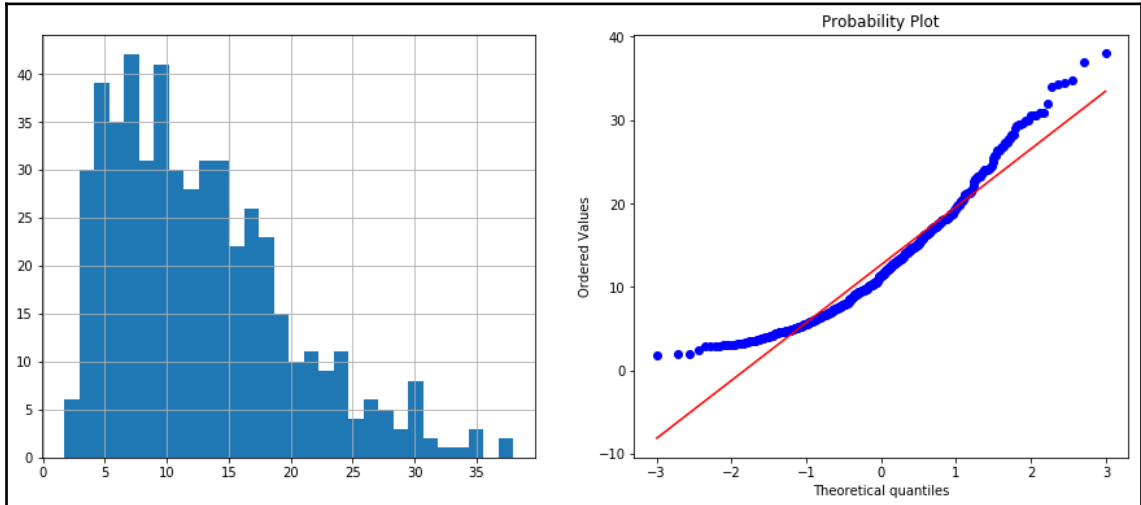
Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips SVN/1.7.14 Phusion_Passenger/4.0.53 mod_perl/2.0.10 Perl/v5.16.3 Server at archive.ics.uci.edu Port 80

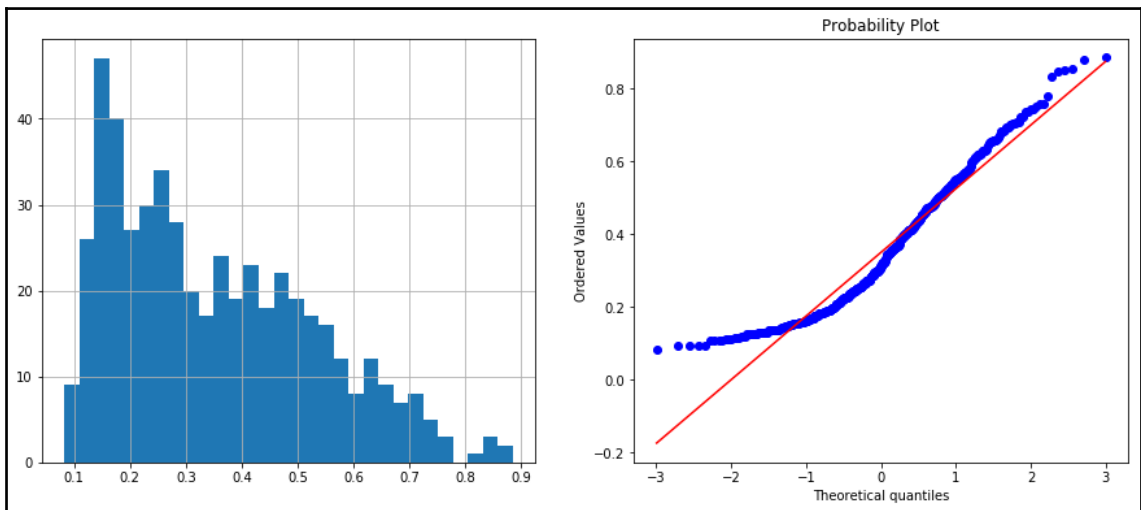
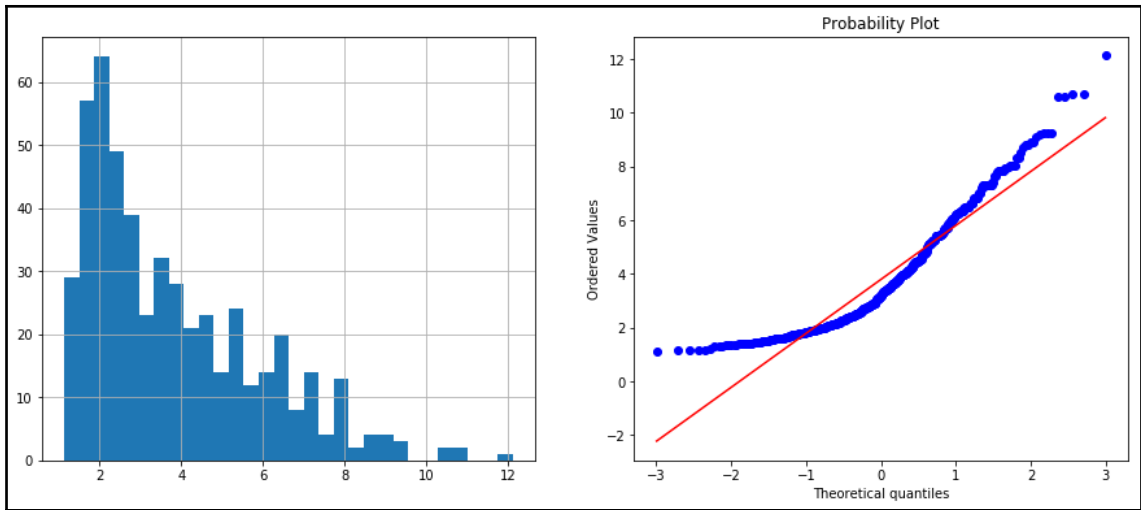
	A2	A3	A8	A11	A14	A15	A1_a	A1_b	A4_u	A4_y	...	A7_z	A7_bb	A7_j	A7_Missing	A7_n	A9_t	A10_t	A12_t	A13_g	A13_s
596	46.08	3.000	2.375	8	396.0	4159	1	0	1	0	...	0	0	0	0	0	1	1	1	1	0
303	15.92	2.875	0.085	0	120.0	0	1	0	1	0	...	0	0	0	0	0	0	0	0	1	0
204	36.33	2.125	0.085	1	50.0	1187	0	1	0	1	...	0	0	0	0	0	1	1	0	1	0
351	22.17	0.585	0.000	0	100.0	0	0	1	0	1	...	0	0	0	0	0	0	0	0	1	0
118	57.83	7.040	14.000	6	360.0	1332	0	1	1	0	...	0	0	0	0	0	1	1	1	1	0



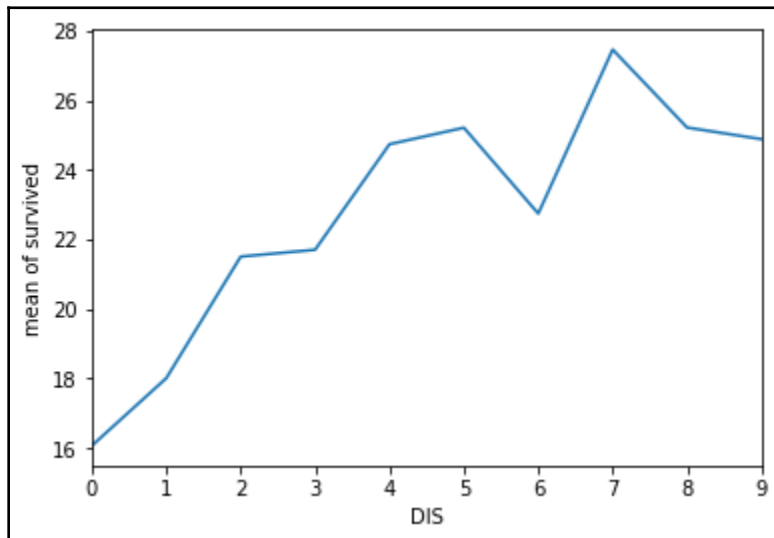
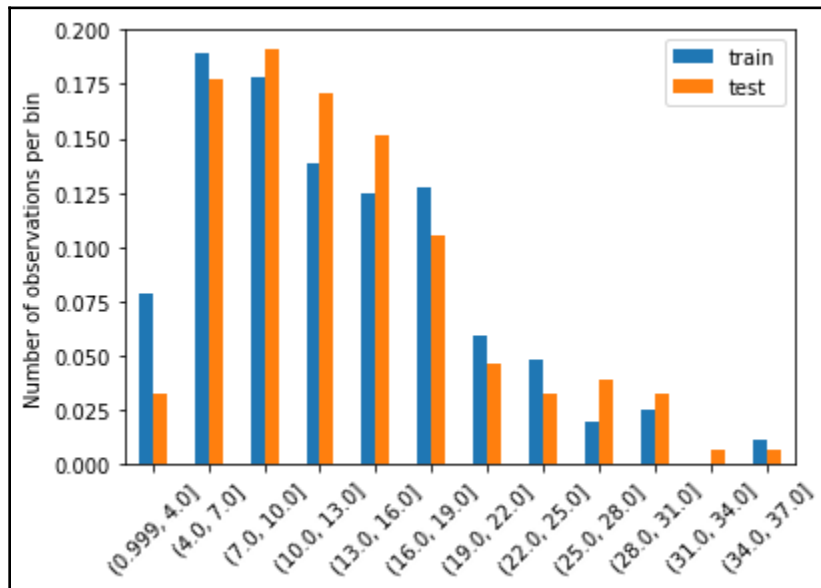


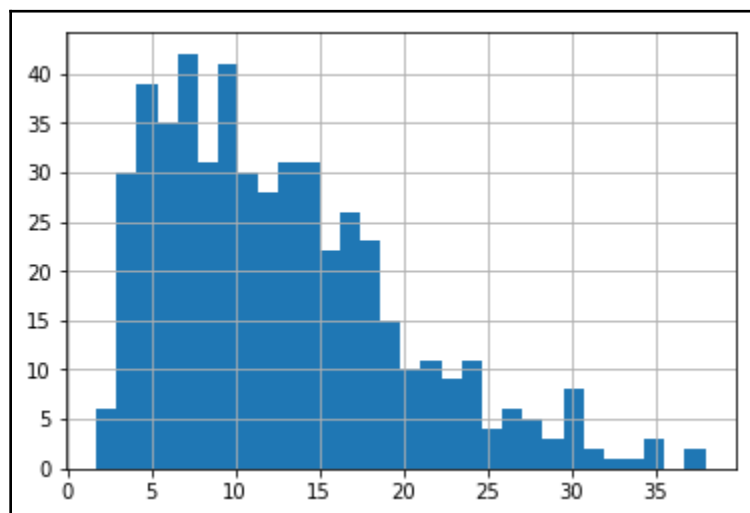
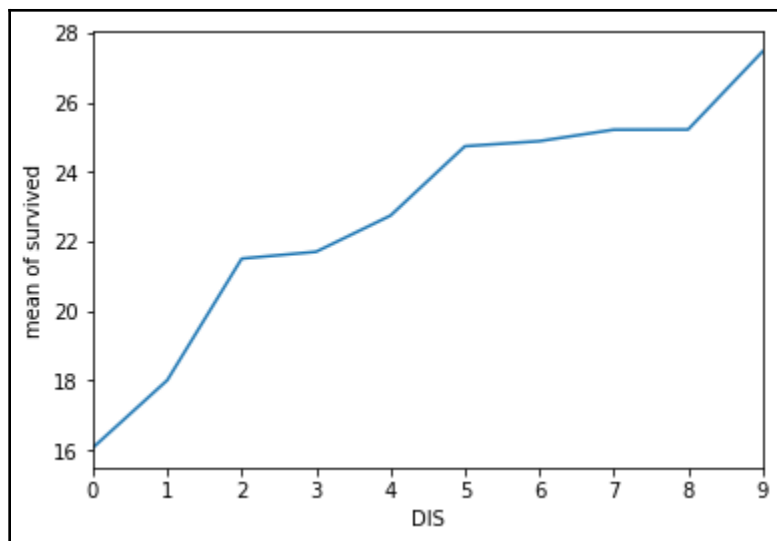
Chapter 4: Transforming Numerical Variables

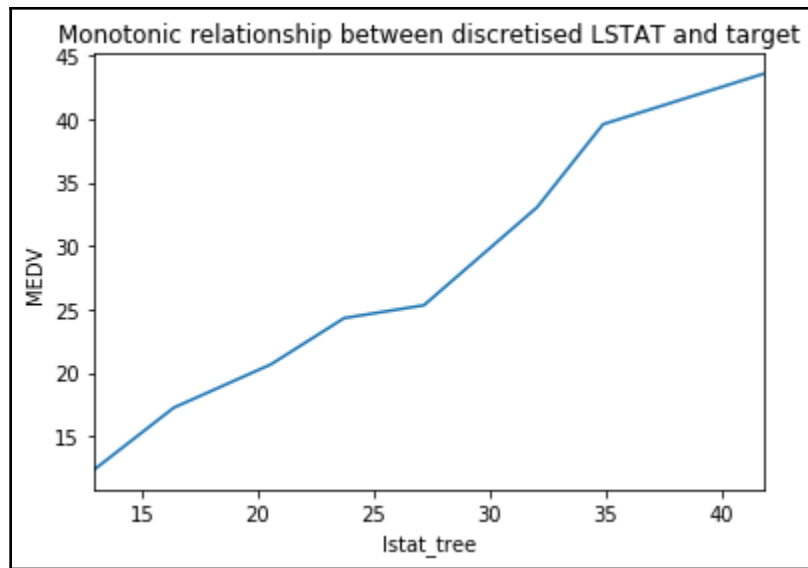




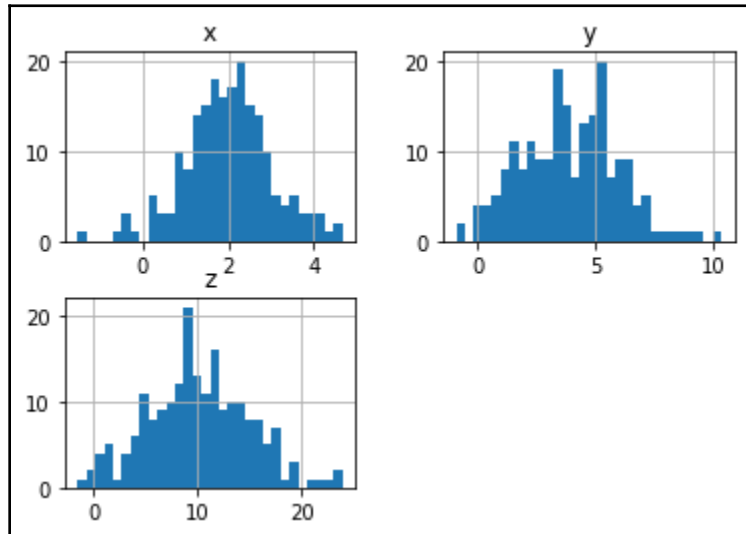
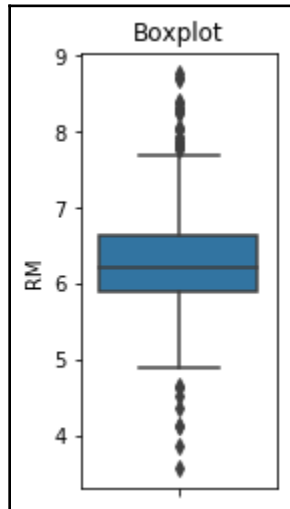
Chapter 5: Performing Variable Discretization







Chapter 6: Working with Outliers



Chapter 7: Deriving Features from Dates and Time Variables

	date
0	2019-03-05 00:00:00
1	2019-03-05 00:01:00
2	2019-03-05 00:02:00
3	2019-03-05 00:03:00
4	2019-03-05 00:04:00

	date_var	datetime_var	date	time
0	Jan-2015	2015-01-01	2015-01-01	00:00:00
1	Apr-2013	2013-04-01	2013-04-01	00:00:00
2	Jun-2014	2014-06-01	2014-06-01	00:00:00
3	Jan-2015	2015-01-01	2015-01-01	00:00:00

date	
0	2019-03-31
1	2019-04-30
2	2019-05-31
3	2019-06-30
4	2019-07-31

	date	year
0	2019-03-31	2019
1	2019-04-30	2019
2	2019-05-31	2019
3	2019-06-30	2019
4	2019-07-31	2019

	date	year	month
0	2019-03-31	2019	3
1	2019-04-30	2019	4
2	2019-05-31	2019	5
3	2019-06-30	2019	6
4	2019-07-31	2019	7

	date	year	month	quarter
0	2019-03-31	2019	3	1
1	2019-04-30	2019	4	2
2	2019-05-31	2019	5	2
3	2019-06-30	2019	6	2
4	2019-07-31	2019	7	3

	date	year	month	quarter	semester
0	2019-03-31	2019	3	1	1
1	2019-04-30	2019	4	2	1
2	2019-05-31	2019	5	2	1
3	2019-06-30	2019	6	2	1
4	2019-07-31	2019	7	3	2

date	
0	2019-03-05
1	2019-03-06
2	2019-03-07
3	2019-03-08
4	2019-03-09

date day_mo		
0	2019-03-05	5
1	2019-03-06	6
2	2019-03-07	7
3	2019-03-08	8
4	2019-03-09	9

	date	day_mo	day_week
0	2019-03-05	5	1
1	2019-03-06	6	2
2	2019-03-07	7	3
3	2019-03-08	8	4
4	2019-03-09	9	5

	date	day_mo	day_week	day_week_name
0	2019-03-05	5	1	Tuesday
1	2019-03-06	6	2	Wednesday
2	2019-03-07	7	3	Thursday
3	2019-03-08	8	4	Friday
4	2019-03-09	9	5	Saturday

	date	day_mo	day_week	day_week_name	is_weekend
0	2019-03-05	5	1	Tuesday	0
1	2019-03-06	6	2	Wednesday	0
2	2019-03-07	7	3	Thursday	0
3	2019-03-08	8	4	Friday	0
4	2019-03-09	9	5	Saturday	1

	date	day_mo	day_week	day_week_name	is_weekend	week
0	2019-03-05	5	1	Tuesday	0	10
1	2019-03-06	6	2	Wednesday	0	10
2	2019-03-07	7	3	Thursday	0	10
3	2019-03-08	8	4	Friday	0	10
4	2019-03-09	9	5	Saturday	1	10

	date
0	2019-03-05 00:00:00
1	2019-03-05 01:15:10
2	2019-03-05 02:30:20
3	2019-03-05 03:45:30
4	2019-03-05 05:00:40

	date	hour	min	sec
0	2019-03-05 00:00:00	0	0	0
1	2019-03-05 01:15:10	1	15	10
2	2019-03-05 02:30:20	2	30	20
3	2019-03-05 03:45:30	3	45	30
4	2019-03-05 05:00:40	5	0	40

	date	hour	min	sec	h	m	s
0	2019-03-05 00:00:00	0	0	0	0	0	0
1	2019-03-05 01:15:10	1	15	10	1	15	10
2	2019-03-05 02:30:20	2	30	20	2	30	20
3	2019-03-05 03:45:30	3	45	30	3	45	30
4	2019-03-05 05:00:40	5	0	40	5	0	40

	date	hour	min	sec	h	m	s	is_morning
0	2019-03-05 00:00:00	0	0	0	0	0	0	0
1	2019-03-05 01:15:10	1	15	10	1	15	10	0
2	2019-03-05 02:30:20	2	30	20	2	30	20	0
3	2019-03-05 03:45:30	3	45	30	3	45	30	0
4	2019-03-05 05:00:40	5	0	40	5	0	40	0

	date1	date2
0	2019-03-05 00:00:00	2019-03-31
1	2019-03-05 01:00:00	2019-04-30
2	2019-03-05 02:00:00	2019-05-31
3	2019-03-05 03:00:00	2019-06-30
4	2019-03-05 04:00:00	2019-07-31

	date1	date2	elapsed_days
0	2019-03-05 00:00:00	2019-03-31	26
1	2019-03-05 01:00:00	2019-04-30	55
2	2019-03-05 02:00:00	2019-05-31	86
3	2019-03-05 03:00:00	2019-06-30	116
4	2019-03-05 04:00:00	2019-07-31	147

	date1	date2	elapsed_days	months_passed
0	2019-03-05 00:00:00	2019-03-31	26	1.0
1	2019-03-05 01:00:00	2019-04-30	55	2.0
2	2019-03-05 02:00:00	2019-05-31	86	3.0
3	2019-03-05 03:00:00	2019-06-30	116	4.0
4	2019-03-05 04:00:00	2019-07-31	147	5.0

	date1	date2	elapsed_days	months_passed	diff_seconds	diff_minutes
0	2019-03-05 00:00:00	2019-03-31	26	1.0	2246400.0	37440.0
1	2019-03-05 01:00:00	2019-04-30	55	2.0	4834800.0	80580.0
2	2019-03-05 02:00:00	2019-05-31	86	3.0	7509600.0	125160.0
3	2019-03-05 03:00:00	2019-06-30	116	4.0	10098000.0	168300.0
4	2019-03-05 04:00:00	2019-07-31	147	5.0	12772800.0	212880.0

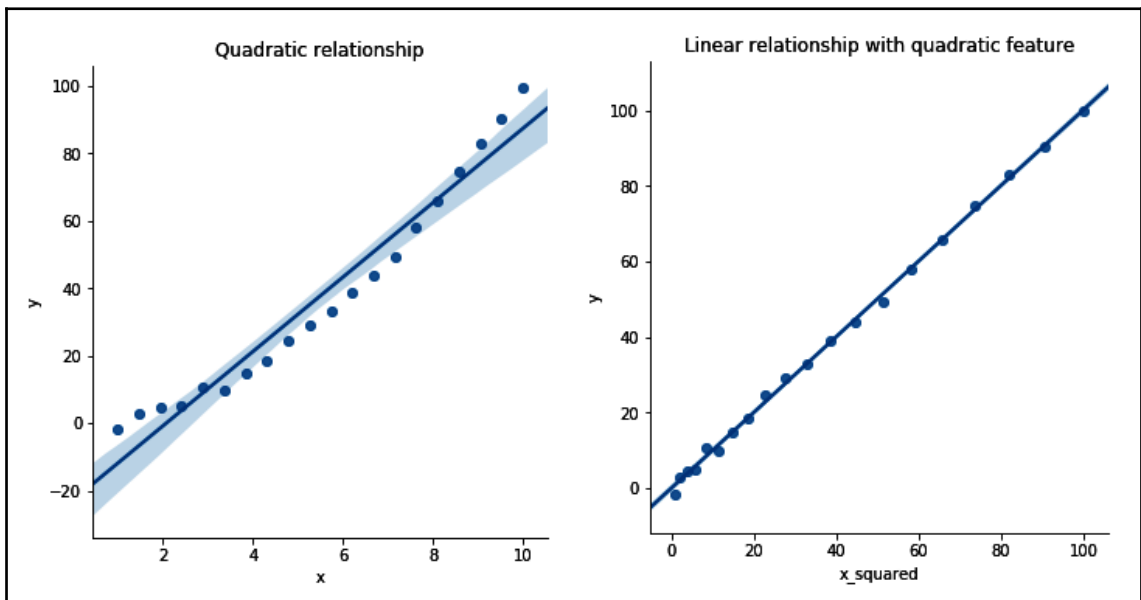
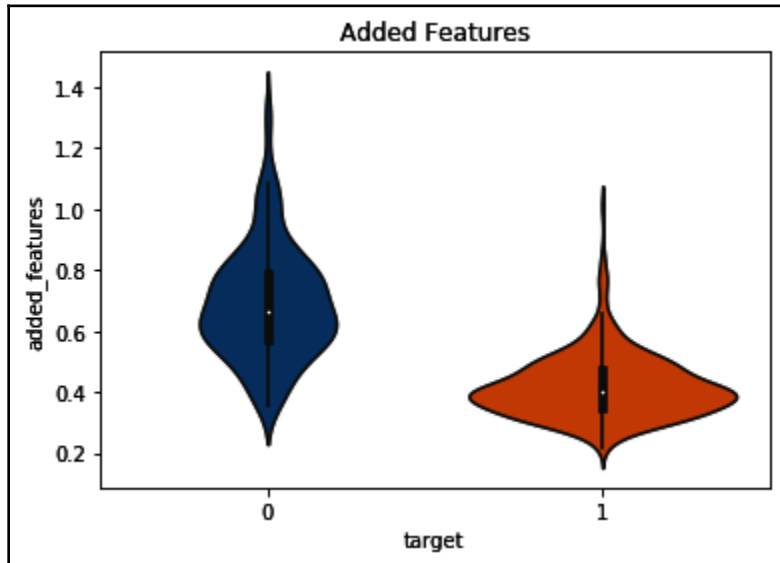
	date1	date2	elapsed_days	months_passed	diff_seconds	diff_minutes	to_today
0	2019-03-05 00:00:00	2019-03-31	26	1.0	2246400.0	37440.0	234 days 16:46:28.095694
1	2019-03-05 01:00:00	2019-04-30	55	2.0	4834800.0	80580.0	234 days 15:46:28.095694
2	2019-03-05 02:00:00	2019-05-31	86	3.0	7509600.0	125160.0	234 days 14:46:28.095694
3	2019-03-05 03:00:00	2019-06-30	116	4.0	10098000.0	168300.0	234 days 13:46:28.095694
4	2019-03-05 04:00:00	2019-07-31	147	5.0	12772800.0	212880.0	234 days 12:46:28.095694

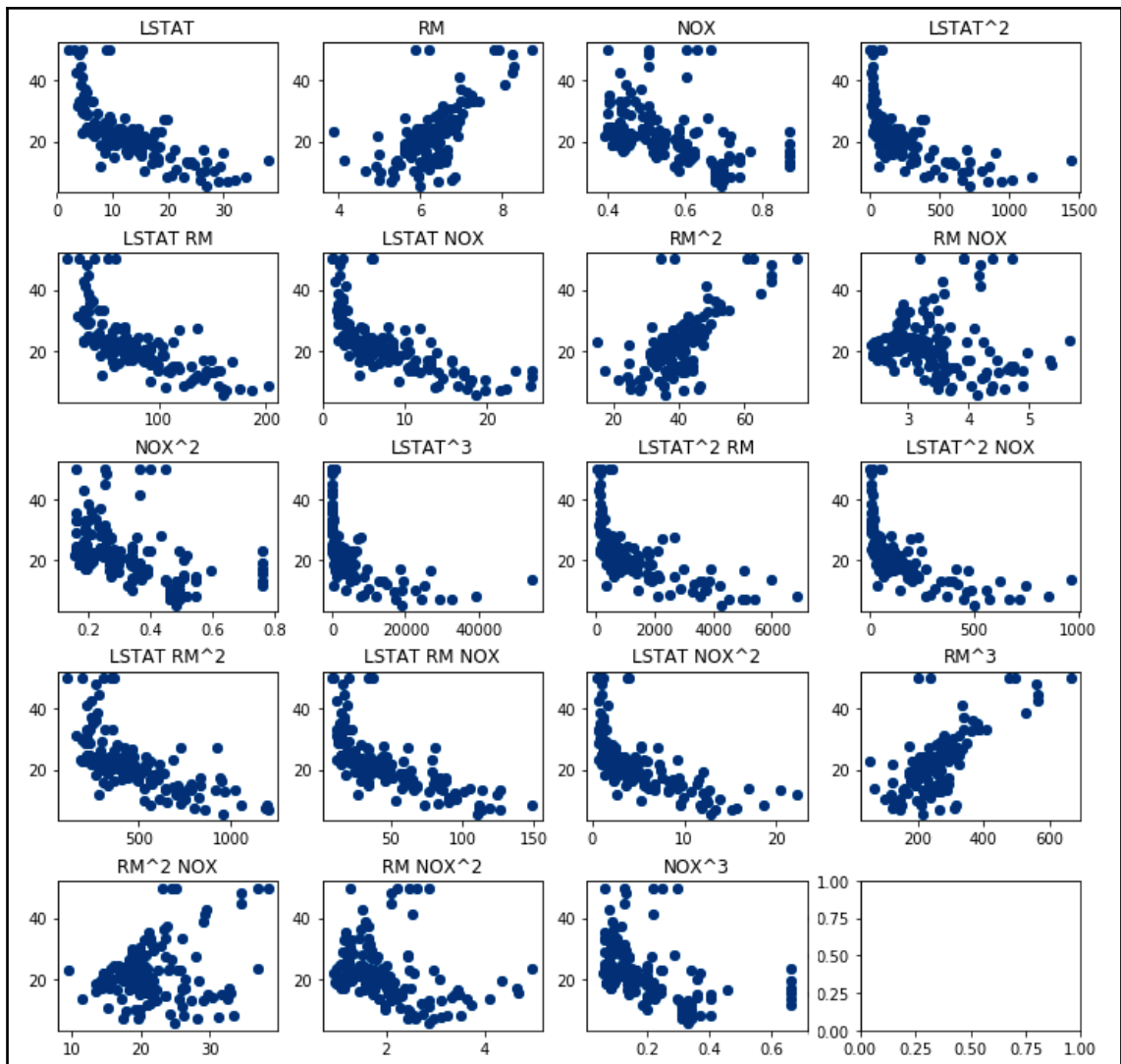
	time1	time2
0	2015-06-10 09:00:00+02:00	2015-07-01 09:00:00+02:00
1	2015-06-10 10:00:00+02:00	2015-07-01 10:00:00+02:00
2	2015-06-10 11:00:00+02:00	2015-07-01 11:00:00+02:00
0	2015-09-10 09:00:00-05:00	2015-08-01 09:00:00-05:00
1	2015-09-10 10:00:00-05:00	2015-08-01 10:00:00-05:00
2	2015-09-10 11:00:00-05:00	2015-08-01 11:00:00-05:00

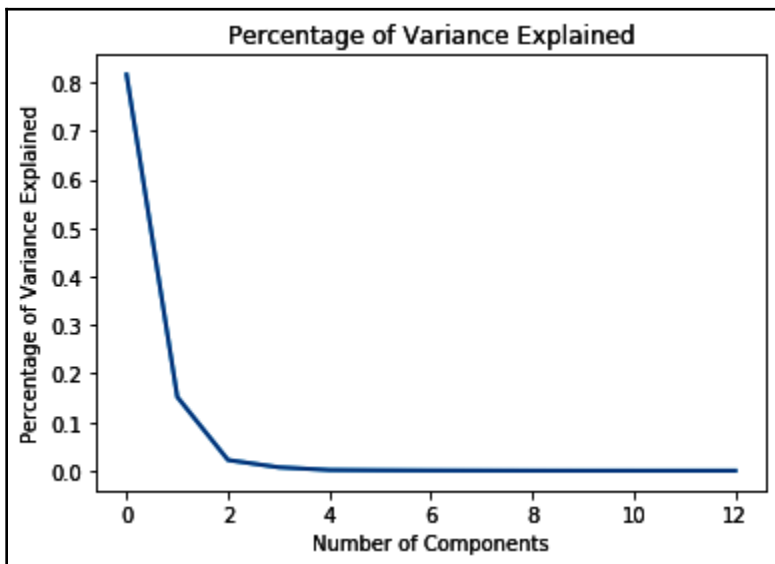
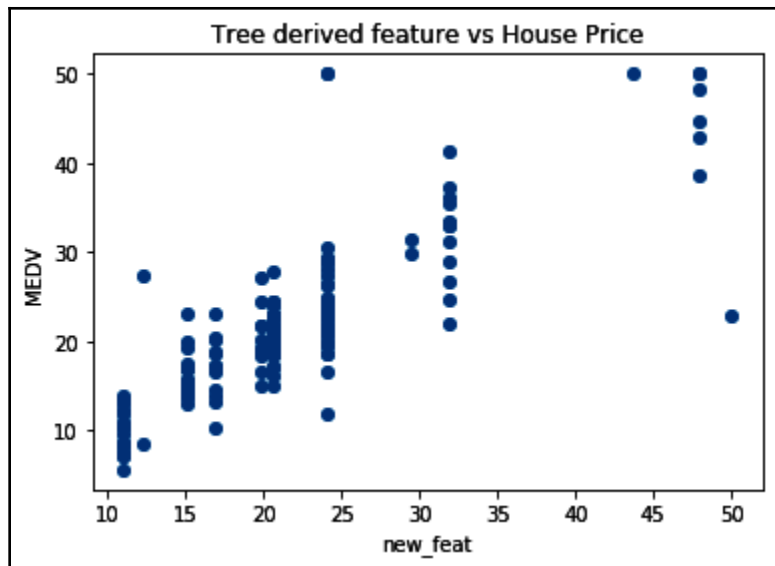
	time1	time2	time1_utc	time2_utc
0	2015-06-10 09:00:00+02:00	2015-07-01 09:00:00+02:00	2015-06-10 07:00:00+00:00	2015-07-01 07:00:00+00:00
1	2015-06-10 10:00:00+02:00	2015-07-01 10:00:00+02:00	2015-06-10 08:00:00+00:00	2015-07-01 08:00:00+00:00
2	2015-06-10 11:00:00+02:00	2015-07-01 11:00:00+02:00	2015-06-10 09:00:00+00:00	2015-07-01 09:00:00+00:00
0	2015-09-10 09:00:00-05:00	2015-08-01 09:00:00-05:00	2015-09-10 14:00:00+00:00	2015-08-01 14:00:00+00:00
1	2015-09-10 10:00:00-05:00	2015-08-01 10:00:00-05:00	2015-09-10 15:00:00+00:00	2015-08-01 15:00:00+00:00
2	2015-09-10 11:00:00-05:00	2015-08-01 11:00:00-05:00	2015-09-10 16:00:00+00:00	2015-08-01 16:00:00+00:00

	time1_london	time2_berlin
0	2015-06-10 08:00:00+01:00	2015-06-10 09:00:00+02:00
1	2015-06-10 09:00:00+01:00	2015-06-10 10:00:00+02:00
2	2015-06-10 10:00:00+01:00	2015-06-10 11:00:00+02:00
0	2015-09-10 15:00:00+01:00	2015-09-10 16:00:00+02:00
1	2015-09-10 16:00:00+01:00	2015-09-10 17:00:00+02:00
2	2015-09-10 17:00:00+01:00	2015-09-10 18:00:00+02:00

Chapter 9: Applying Mathematical Computations to Features







Chapter 10: Creating Features with Transactional and Time Series Data

←

→

↺

🏠

🔒  archive.ics.uci.edu/ml/machine-learning-databases/00374/

Index of /ml/machine-learning-databases/00374

• [Parent Directory](#)

• [energydata_complete.csv](#) 

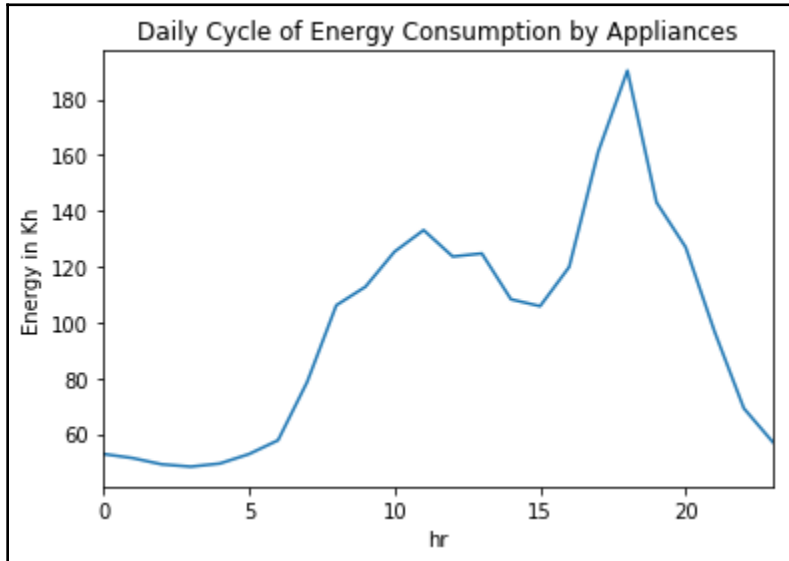
Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips SVN/1.7.14 Phusion_Passenger/4.0.53 mod_perl/2.0.10 Perl/v5.16.3 Server at archive.ics.uci.edu Port 80

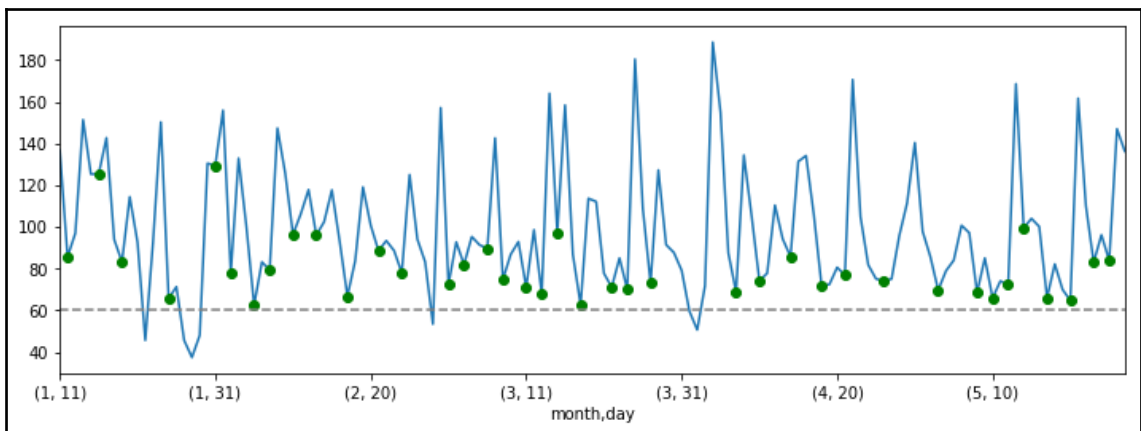
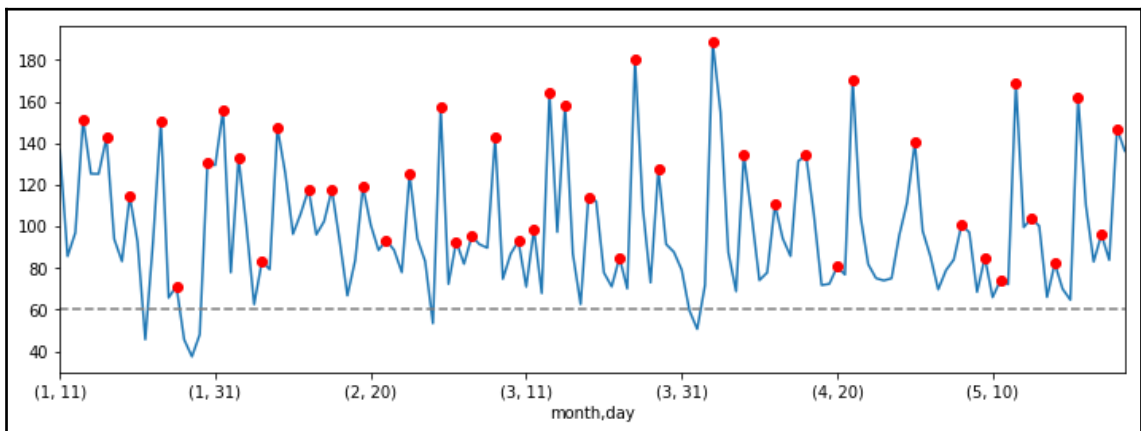
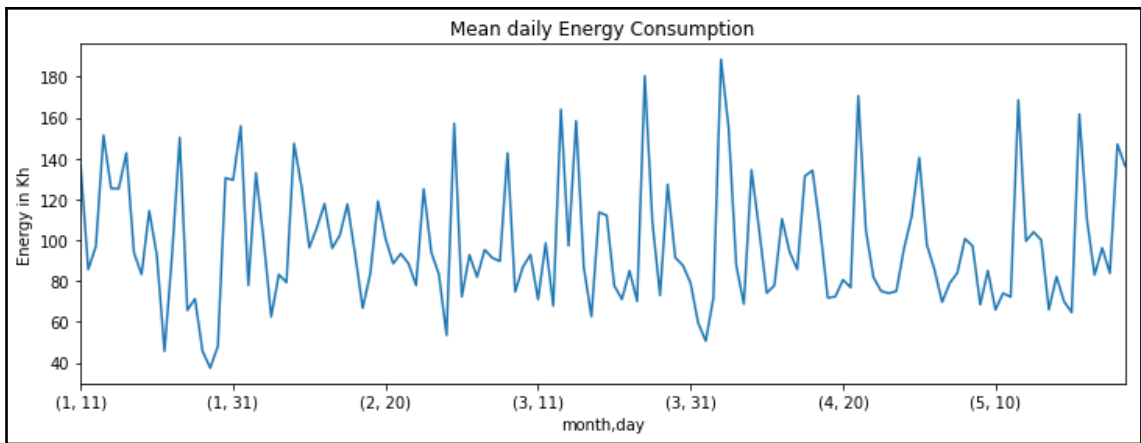
	total_amount	max_amount	min_amount	mean_amount	median_amount	std_amount	number of transactions
customer_id							
1	9025.62	139.43	5.81	71.631905	69.715	40.442059	126
2	7200.28	146.81	8.73	77.422366	75.960	37.705178	93
3	6236.62	149.15	5.89	67.060430	58.930	43.683296	93
4	8727.68	149.95	5.73	80.070459	81.410	45.068765	109
5	6349.66	149.02	7.55	80.375443	78.870	44.095630	79

	SUM(transactions.amount)	MAX(transactions.amount)	MIN(transactions.amount)	MEAN(transactions.amount)	MEDIAN(transactions.amount)	STD
customer_id						
2	7200.28	146.81	8.73	77.422366	75.960	
5	6349.66	149.02	7.55	80.375443	78.870	
4	8727.68	149.95	5.73	80.070459	81.410	
1	9025.62	139.43	5.81	71.631905	69.715	
3	6236.62	149.15	5.89	67.060430	58.930	

	date	Appliances	lights
0	2016-01-11 17:00:00	NaN	NaN
1	2016-01-11 17:10:00	NaN	NaN
2	2016-01-11 17:20:00	NaN	NaN
3	2016-01-11 17:30:00	NaN	NaN
4	2016-01-11 17:40:00	NaN	NaN
5	2016-01-11 17:50:00	55.000000	35.000000
6	2016-01-11 18:00:00	55.000000	38.333333
7	2016-01-11 18:10:00	55.000000	41.666667
8	2016-01-11 18:20:00	56.666667	43.333333
9	2016-01-11 18:30:00	60.000000	43.333333

date	Appliances						lights					
	sum	max	min	mean	median	std	sum	max	min	mean	median	std
2016-01-11 17:00:00	60.0	60.0	60.0	60.000000	60.0	NaN	30.0	30.0	30.0	30.000000	30.0	NaN
2016-01-11 17:10:00	120.0	60.0	60.0	60.000000	60.0	0.000000	60.0	30.0	30.0	30.000000	30.0	0.000000
2016-01-11 17:20:00	170.0	60.0	50.0	56.666667	60.0	5.773503	90.0	30.0	30.0	30.000000	30.0	0.000000
2016-01-11 17:30:00	220.0	60.0	50.0	55.000000	55.0	5.773503	130.0	40.0	30.0	32.500000	30.0	5.000000
2016-01-11 17:40:00	280.0	60.0	50.0	56.000000	60.0	5.477226	170.0	40.0	30.0	34.000000	30.0	5.477226
2016-01-11 17:50:00	330.0	60.0	50.0	55.000000	55.0	5.477226	210.0	40.0	30.0	35.000000	35.0	5.477226
2016-01-11 18:00:00	330.0	60.0	50.0	55.000000	55.0	5.477226	230.0	50.0	30.0	38.333333	40.0	7.527727
2016-01-11 18:10:00	330.0	60.0	50.0	55.000000	55.0	5.477226	250.0	50.0	30.0	41.666667	40.0	7.527727
2016-01-11 18:20:00	340.0	60.0	50.0	56.666667	60.0	5.163978	260.0	50.0	40.0	43.333333	40.0	5.163978
2016-01-11 18:30:00	360.0	70.0	50.0	60.000000	60.0	6.324555	260.0	50.0	40.0	43.333333	40.0	5.163978





	find_no_peaks	find_no_valleys
customer_id		
1	2.0	2.0
2	1.0	0.0
3	1.0	2.0
4	1.0	2.0
5	1.0	1.0

	month	day	Appliances	year
0	1	14	151.388889	2016
1	1	17	142.708333	2016
2	1	20	114.444444	2016
3	1	24	150.277778	2016
4	1	26	71.319444	2016

	month	day	Appliances	year	date
0	1	14	151.388889	2016	2016-01-14
1	1	17	142.708333	2016	2016-01-17
2	1	20	114.444444	2016	2016-01-20
3	1	24	150.277778	2016	2016-01-24
4	1	26	71.319444	2016	2016-01-26

	month	day	Appliances	year	date	peak_distance
0	1	14	151.388889	2016	2016-01-14	NaN
1	1	17	142.708333	2016	2016-01-17	3.0
2	1	20	114.444444	2016	2016-01-20	3.0
3	1	24	150.277778	2016	2016-01-24	4.0
4	1	26	71.319444	2016	2016-01-26	2.0

Chapter 11: Extracting Features from Text Variables

