

Learn Distribution & Missing Value Synthesis

Computational ClassNotes

dara@lossogenerality.com

In[1]:= data = ResourceData["Sample Data: Boston Homes"]

Out[]:=

CRIM	ZN	INDUS	CHAS
0.00632	18	2.31	tract does not bound Charles river
0.02731	0	7.07	tract does not bound Charles river
0.02729	0	7.07	tract does not bound Charles river
0.03237	0	2.18	tract does not bound Charles river
0.06905	0	2.18	tract does not bound Charles river
0.02985	0	2.18	tract does not bound Charles river
0.08829	12.5	7.87	tract does not bound Charles river
0.14455	12.5	7.87	tract does not bound Charles river
0.21124	12.5	7.87	tract does not bound Charles river
0.17004	12.5	7.87	tract does not bound Charles river
0.22489	12.5	7.87	tract does not bound Charles river
0.11747	12.5	7.87	tract does not bound Charles river
0.09378	12.5	7.87	tract does not bound Charles river
0.62976	0	8.14	tract does not bound Charles river
0.63796	0	8.14	tract does not bound Charles river
0.62739	0	8.14	tract does not bound Charles river
1.05393	0	8.14	tract does not bound Charles river
0.7842	0	8.14	tract does not bound Charles river
0.80271	0	8.14	tract does not bound Charles river
0.7258	0	8.14	tract does not bound Charles river

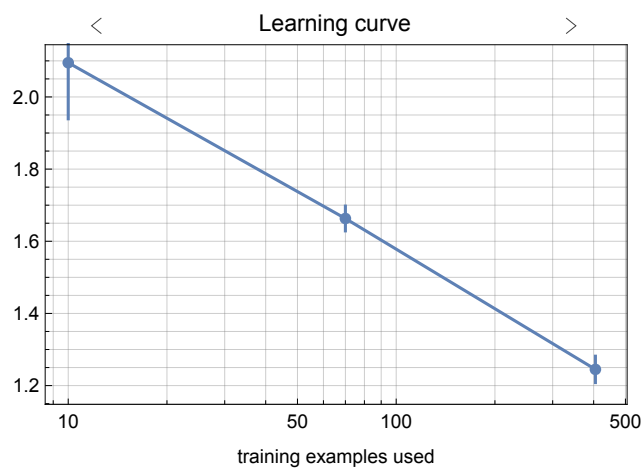
K < showing 1-20 of 506 > X

Learn Distribution

Distribution information

Data type	Mixed (number: 14)
Dimension	14
Method	KernelDensityEstimation
Loss	1.24 ± 0.037
Entropy	15.2 ± 0.30
PDF query speed	497. examples/s
Sampling speed	12.4 examples/ms
Model memory	241. kB
Training examples used	506 examples
Training time	3.19 s

Out[8]=



```
In[ ]:= RandomVariate[ld]
```

```
Out[ ]:=
```

CRIM	5.75561
ZN	-3.60632
INDUS	10.7493
CHAS	tract bounds Charles river
NOX	0.574599 ppm
RM	5.12254
AGE	40.9354
DIS	2.37326
RAD	2.05688
TAX	297.306
PTRATIO	18.6227
BLACK	367.094
LSTAT	11.9863%
MEDV	13.39

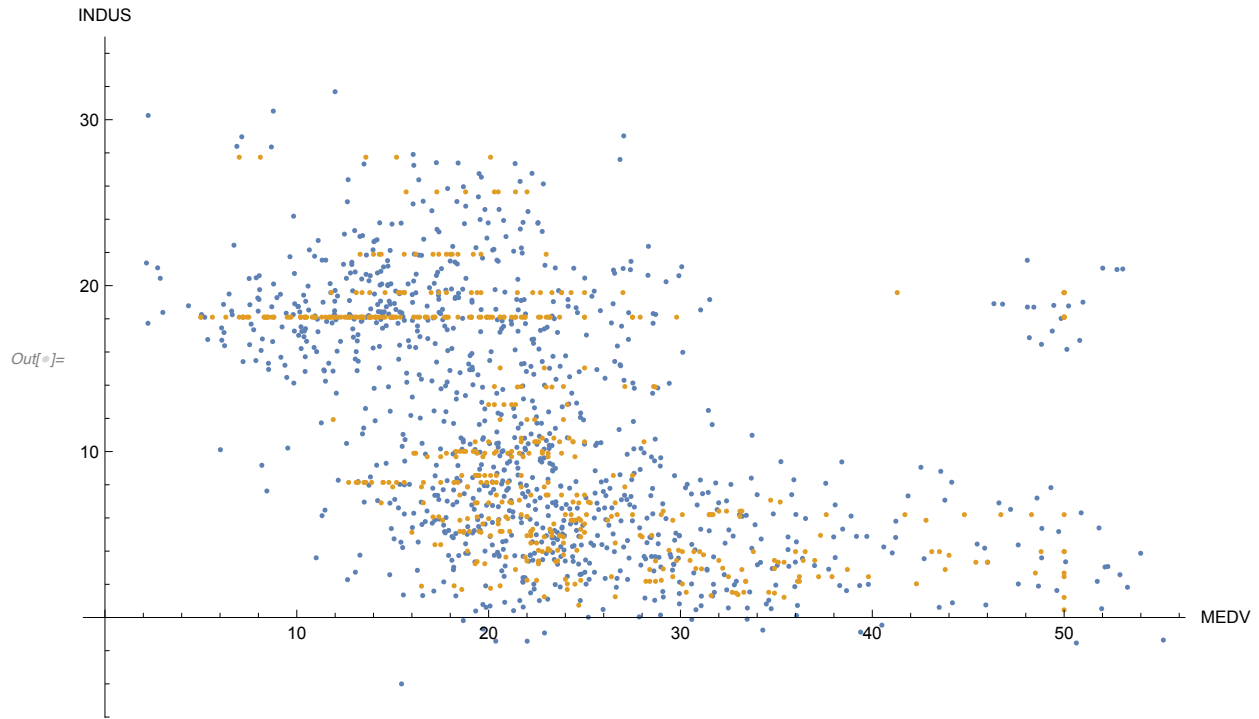
Random Variate New Variables

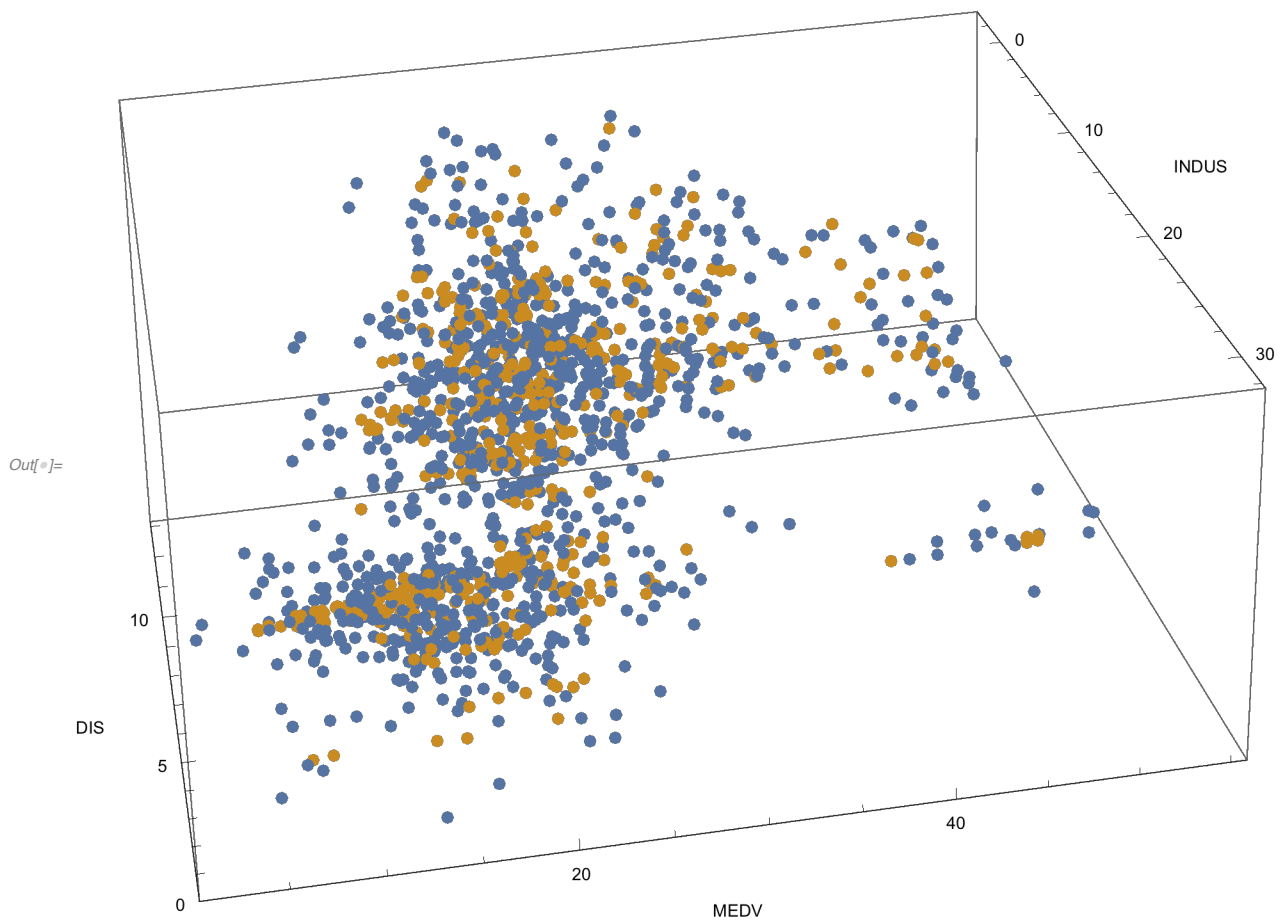
```
In[ ]:= RandomVariate[ld, 4]
```

```
Out[ ]:=
```

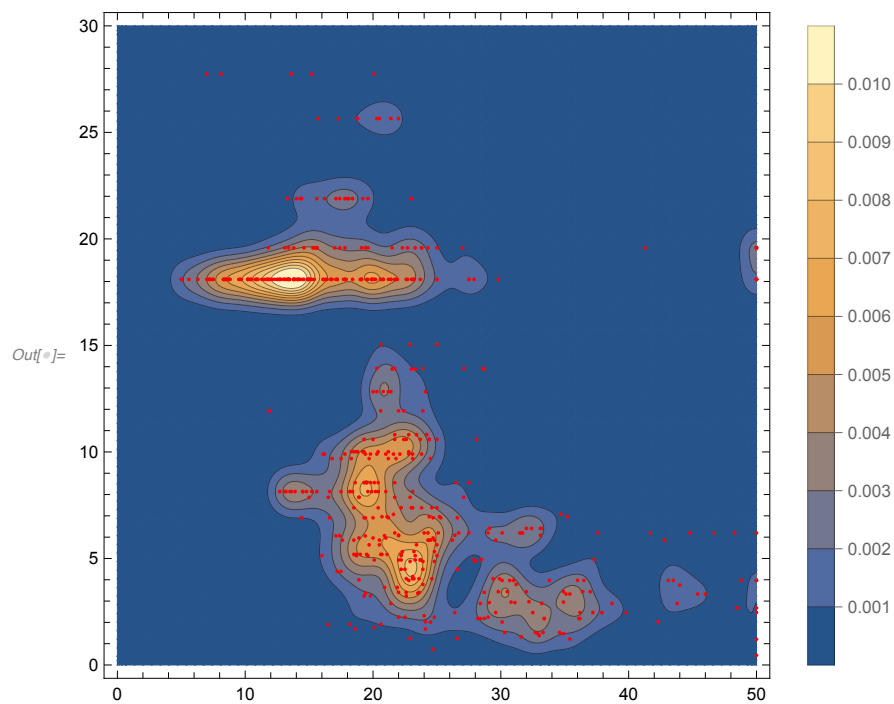
CRIM	ZN	INDUS	CHAS
0.0411203	8.83712	4.77143	tract does not bound Charles river
0.0455871	19.347	3.39643	tract does not bound Charles river
9.27904	2.97157	17.2305	tract bounds Charles river
22.9233	-1.54632	15.8158	tract does not bound Charles river

Blue dots are new random **variated** and golden dots are the original data. As you can see there is nice dense distribution to mimic the original data's distribution.





MEDV vs. INDUS probability distribution function



Missing Value Synthesis

Out[]:=

CRIM
0.00632
0.02731
0.02729
0.03237
0.06905
0.02985
0.08829
0.14455
0.21124
0.17004
0.22489
0.11747
0.09378
0.62976
0.63796
0.62739
1.05393
0.7842
0.80271
0.7258

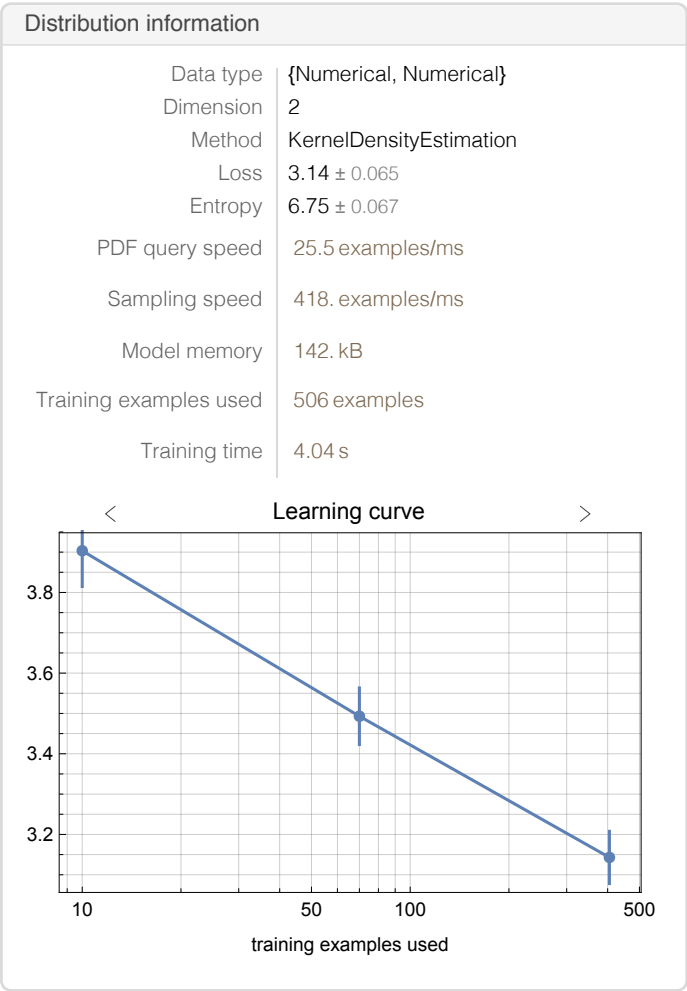
showing 1-20 of 506

Remove from data from CRIM i.e. missing data

Out[]=

CRIM	ZN	INDUS	CHAS
0.00632	18	2.31	tract does not bound Charles river
0.02731	0	7.07	tract does not bound Charles river
0.02729	0	7.07	tract does not bound Charles river
0.03237	0	2.18	tract does not bound Charles river
—	0	2.18	tract does not bound Charles river
0.02985	0	2.18	tract does not bound Charles river
—	12.5	7.87	tract does not bound Charles river
0.14455	12.5	7.87	tract does not bound Charles river
0.21124	12.5	7.87	tract does not bound Charles river
0.17004	12.5	7.87	tract does not bound Charles river
0.22489	12.5	7.87	tract does not bound Charles river
0.11747	12.5	7.87	tract does not bound Charles river
0.09378	12.5	7.87	tract does not bound Charles river
0.62976	0	8.14	tract does not bound Charles river
0.63796	0	8.14	tract does not bound Charles river
0.62739	0	8.14	tract does not bound Charles river
1.05393	0	8.14	tract does not bound Charles river
0.7842	0	8.14	tract does not bound Charles river
0.80271	0	8.14	tract does not bound Charles river
—	0	8.14	tract does not bound Charles river
K < showing 1–20 of 506 > X			

Learn the distribution for CRIM, assume order matters.



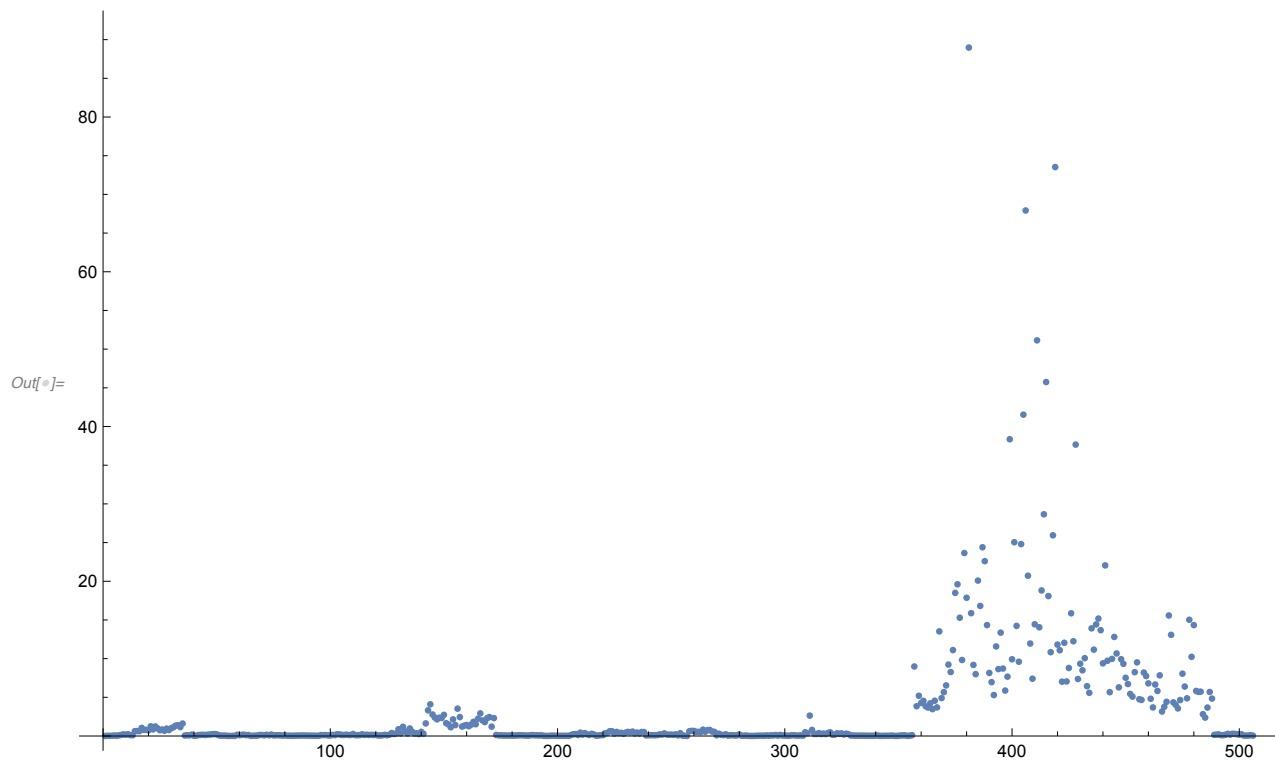
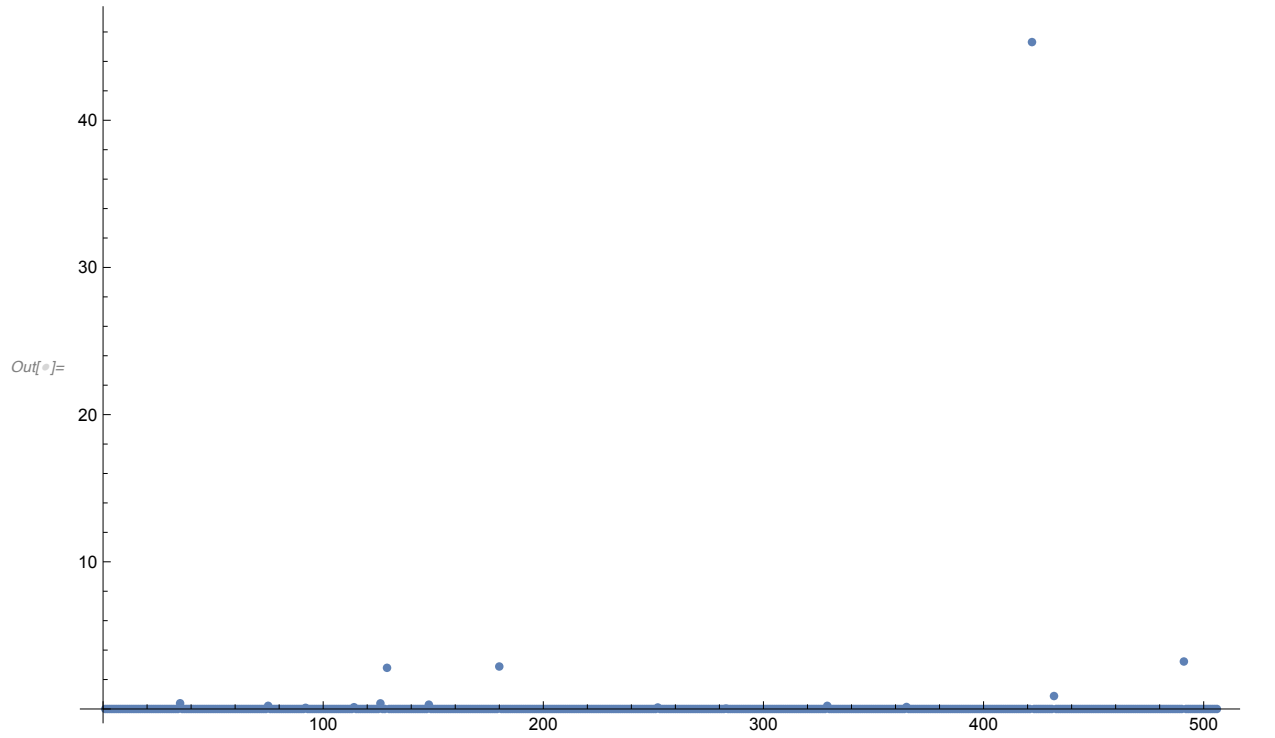
Synthesize the missing data by the learned distribution

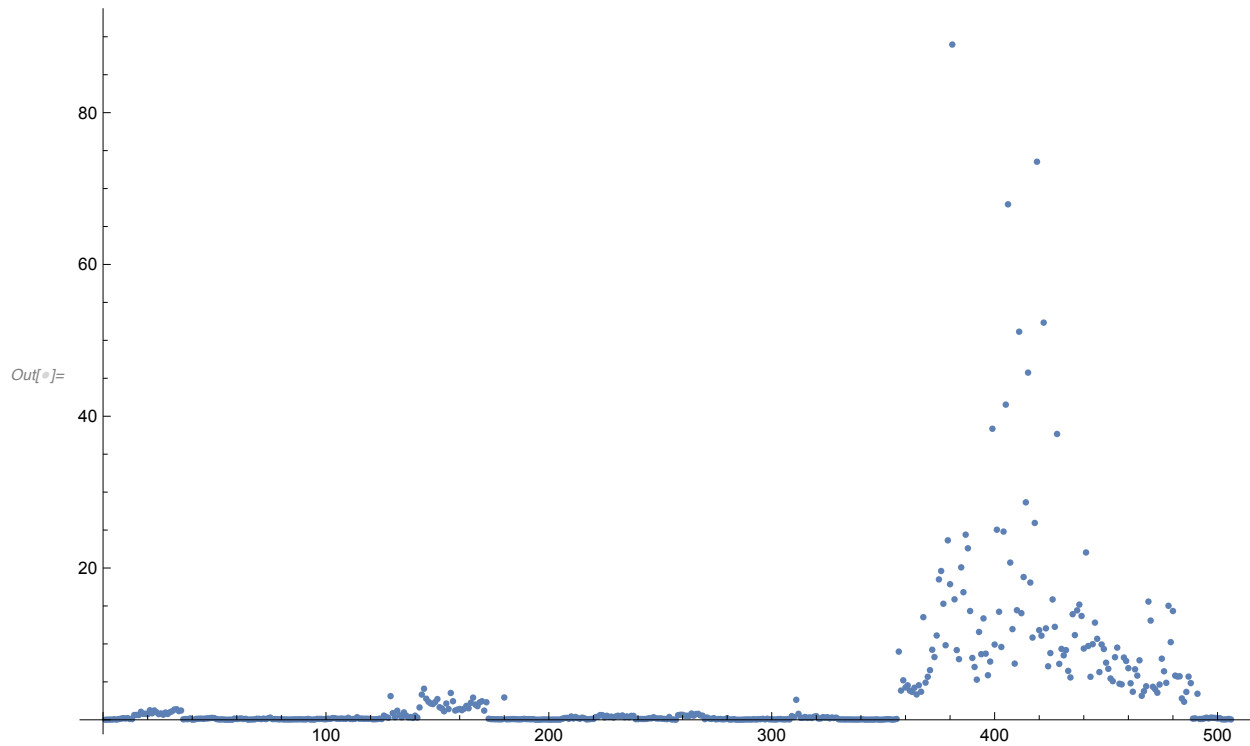
Out[]:=

CRIM
0.00632
0.02731
0.02729
0.03237
0.06905
0.02985
0.08829
0.14455
0.21124
0.17004
0.22489
0.11747
0.09378
0.62976
0.63796
0.62739
1.05393
0.7842
0.80271
0.7258

K < showing 1-20 of 506 > X

Compare the error plot vs. the original CRIM vs. the CRIM with missing values replaced.





Learn Date Distribution

Learn a distribution for some dates around Aug-Sept 2016 and 2017.

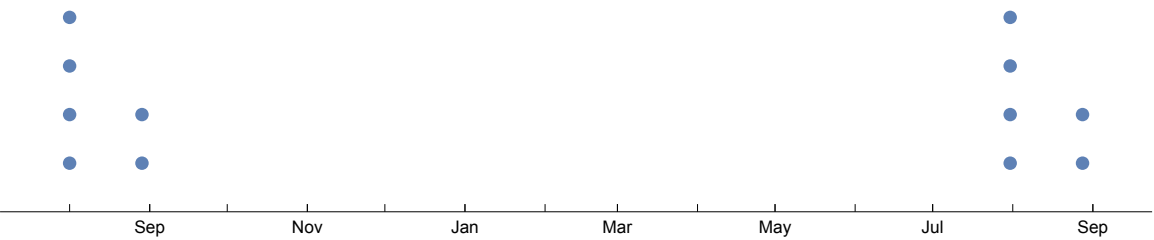
```
times = {DateObject[{2016, 8, 1}, "Week"], DateObject[{2016, 8, 2}, "Week"],
  DateObject[{2016, 8, 3}, "Week"], DateObject[{2016, 8, 4}, "Week"],
  DateObject[{2016, 9, 1}, "Week"], DateObject[{2016, 9, 2}, "Week"],

  DateObject[{2017, 8, 1}, "Week"], DateObject[{2017, 8, 2}, "Week"],
  DateObject[{2017, 8, 3}, "Week"], DateObject[{2017, 8, 4}, "Week"],
  DateObject[{2017, 9, 1}, "Week"], DateObject[{2017, 9, 2}, "Week"]

}
```

Out[24]=

Distribution information		
	Data type	Date
	Dimension	6
	Method	KernelDensityEstimation
	Loss	0.435 ± 0.052
	Entropy	2.41 × 10 ⁴ ± 1.2 × 10 ³
	PDF query speed	425. examples/s
	Sampling speed	113. examples/ms
	Model memory	100. kB
	Training examples used	12 examples
	Training time	1.08 s

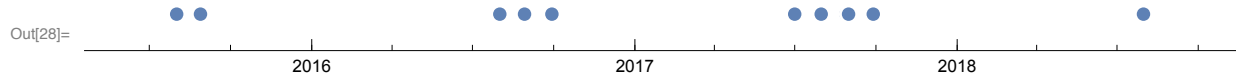


Simulate some dates with the same distribution

Out[25]= {  Tue 31 Jul 2018 00:00:00 GMT+1. ,  Mon 31 Jul 2017 00:00:59 GMT+1. ,
 Mon 29 Aug 2016 00:00:59 GMT+1. ,  Mon 1 Aug 2016 00:00:59 GMT+1. ,
 Sat 30 Jul 2016 23:00:00 GMT+1. ,  Wed 27 Sep 2017 23:00:00 GMT+1. ,
 Sat 1 Jul 2017 00:00:00 GMT+1. ,  Sat 1 Aug 2015 00:59:59 GMT+1. ,
 Wed 30 Aug 2017 23:59:59 GMT+1. ,  Fri 31 Jul 2015 23:59:00 GMT+1. ,
 Wed 28 Sep 2016 23:59:59 GMT+1. ,  Mon 1 Aug 2016 23:59:59 GMT+1. ,
 Mon 31 Jul 2017 00:00:00 GMT+1. ,  Mon 1 Aug 2016 00:00:00 GMT+1. ,
 Tue 31 Jul 2018 01:00:00 GMT+1. ,  Thu 27 Aug 2015 23:59:59 GMT+1. ,
 Mon 31 Jul 2017 00:00:00 GMT+1. ,  Sun 27 Aug 2017 23:59:59 GMT+1. ,
 Sat 29 Aug 2015 00:00:00 GMT+1. ,  Mon 31 Jul 2017 23:59:59 GMT+1. }

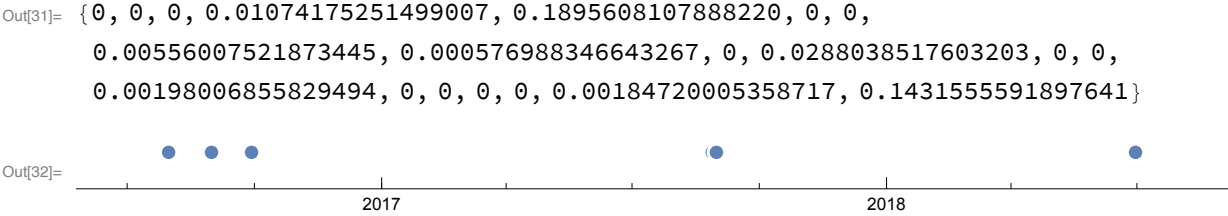
Out[26]= { 0.13224, 0.105608, 0.1199, 0.230701, 0.124495, 0.0958818,
0.110114, 0.115144, 0.178139, 0.226953, 0.0601836, 0.0807734 }

Out[27]= { 0.00484489608046761, 0, 0, 0, 0, 0, 1.56314152900341 $\times 10^{-6}$,
0, 0, 0, 0, 0, 0.1539065590067127, 0.1065410517403622,
0.000185246049743286, 0, 0.0497305385568624, 0, 0.0817872758693852, 0 }

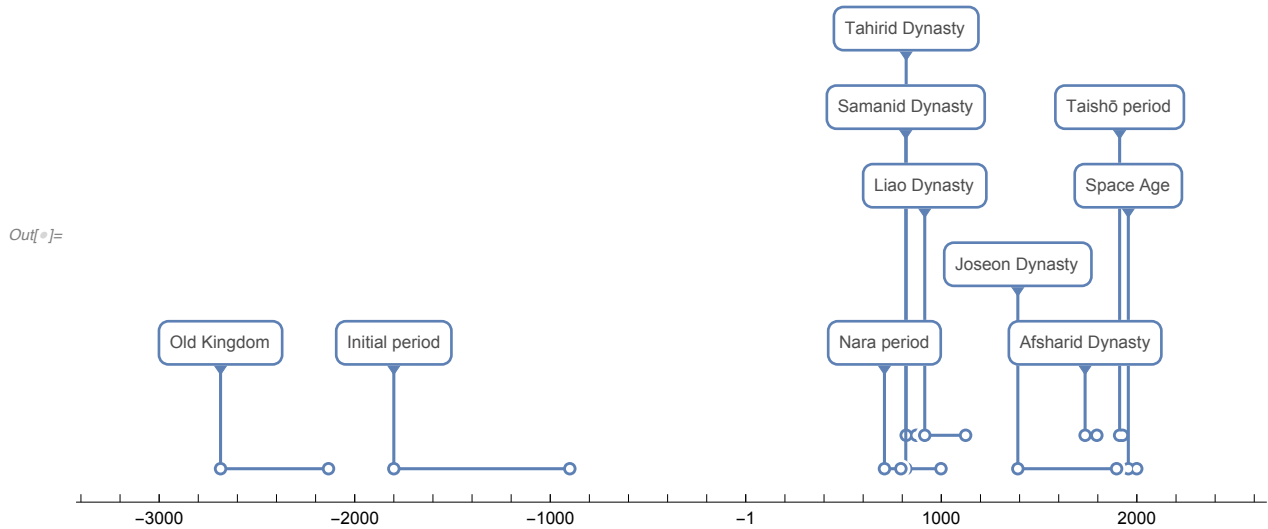


Out[29]= {  Sat 30 Jul 2016 23:59:59 GMT+1. ,  Sat 30 Jul 2016 23:59:59 GMT+1. ,
 Mon 1 Aug 2016 23:59:59 GMT+1. ,  Wed 31 Aug 2016 00:00:00 GMT+1. ,
 Mon 1 Aug 2016 00:00:00 GMT+1. ,  Sat 30 Jul 2016 23:59:59 GMT+1. ,
 Sun 28 Aug 2016 23:59:59 GMT+1. ,  Thu 31 Aug 2017 00:01:00 GMT+1. ,
 Sat 30 Jun 2018 00:00:00 GMT+1. ,  Sun 31 Jul 2016 22:59:59 GMT+1. ,
 Thu 31 Aug 2017 00:00:00 GMT+1. ,  Wed 28 Sep 2016 23:00:00 GMT+1. ,
 Sun 27 Aug 2017 23:58:59 GMT+1. ,  Thu 1 Sep 2016 00:00:00 GMT+1. ,
 Mon 29 Aug 2016 23:00:00 GMT+1. ,  Fri 29 Jun 2018 23:59:59 GMT+1. ,
 Sun 27 Aug 2017 23:00:00 GMT+1. ,  Sun 31 Jul 2016 00:00:59 GMT+1. ,
 Tue 29 Aug 2017 00:00:00 GMT+1. ,  Mon 1 Aug 2016 00:00:00 GMT+1. }

Out[30]= { 0.169529, 0.167998, 0.0907247, 0.124404, 0.0812242, 0.0708844,
0.207371, 0.116551, 0.185578, 0.127964, 0.0232922, 0.0391345 }



```
In[ ]:= TimelinePlot[{Afsharid Dynasty HISTORICAL PERIOD , Initial period HISTORICAL PERIOD ,  
Joseon Dynasty HISTORICAL PERIOD , Liao Dynasty HISTORICAL PERIOD ,  
Nara period HISTORICAL PERIOD , Old Kingdom HISTORICAL PERIOD ,  
Samanid Dynasty HISTORICAL PERIOD , Space Age HISTORICAL PERIOD ,  
Tahrid Dynasty HISTORICAL PERIOD , Taishō period HISTORICAL PERIOD }]]
```



```

In[ ]:= TimelinePlot[<|"1.0" →  Day: Thu 23 Jun 1988 , "1.1" →  Day: Mon 31 Oct 1988 ,
"1.2" →  Day: Tue 1 Aug 1989 , "2.0" →  Day: Tue 15 Jan 1991 ,
"2.1" →  Day: Mon 15 Jun 1992 , "2.2" →  Day: Tue 1 Jun 1993 ,
"3.0" →  Day: Tue 3 Sep 1996 , "4.0" →  Day: Wed 19 May 1999 ,
"4.1" →  Day: Thu 2 Nov 2000 , "4.2" →  Day: Fri 1 Nov 2002 ,
"5.0" →  Day: Thu 12 Jun 2003 , "5.1" →  Day: Mon 25 Oct 2004 ,
"5.2" →  Day: Mon 20 Jun 2005 , "6.0" →  Day: Tue 1 May 2007 ,
"7.0" →  Day: Tue 18 Nov 2008 , "8.0" →  Day: Mon 15 Nov 2010 ,
"9.0" →  Day: Wed 28 Nov 2012 , "10.0" →  Day: Wed 9 Jul 2014 ,
"10.1" →  Day: Mon 30 Mar 2015 , "10.2" →  Day: Tue 14 Jul 2015 ,
"10.3" →  Day: Thu 15 Oct 2015 , "10.4" →  Day: Wed 2 Mar 2016 ,
"11.0" →  Day: Mon 8 Aug 2016 , "11.1" →  Day: Thu 16 Mar 2017 ,
"11.2" →  Day: Thu 14 Sep 2017 , "11.3" →  Day: Thu 8 Mar 2018 |>]

```

Out[]:=

