Welkom in deze basis ML cursus

Contents

Data Cleaning

In deze cursus zullen we de basis principes van Machine Learning belichten en bespreken. We zullen verschillende topics licht aanraken en hopen een volledig holistisch beeld te geven over de basis principes. Deze cursus werkt met life code examples.

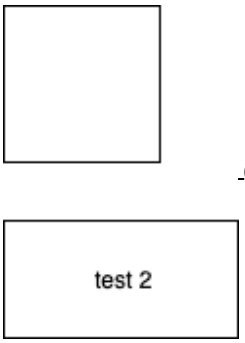
Dit zijn de topics die we gaan bekijken.

```
print('test')
```

```
test
```

Data Cleaning

You can also create content with Jupyter Notebooks. This means that you can include code blocks and their outputs in your book.



(_images/test.drawio.png)

Fig. 1 Some architecture

```
import numpy as np
import pandas as pd
from pandas_profiling import ProfileReport

df = pd.DataFrame(np.random.rand(100, 5), columns=["a", "b", "c", "d", "e"])
```

```
profile = ProfileReport(df, title="Pandas Profiling Report")

profile.config.html.navbar_show = False
```

```
from IPython.core.display import display, HTML
display(HTML(profile.to_html()))
```

Overview

Dataset statistics

Number of variables	5
Number of observations	100
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	4.0 KiB
Average record size in memory	41.3 B

Warnings

e is highly correlated with b	**High correlation**
b is highly correlated with e	**High correlation**
a has unique values	**Unique**
b has unique values	**Unique**
c has unique values	**Unique**
d has unique values	**Unique**
e has unique values	**Unique**

Reproduction

Analysis started	
Analysis finished	
Duration	
Software version	
Download configuration	8,%7B%22title%22%3A%20%22Pandas%20Profiling%20Report%22%2C%20%22dataset%22%3A%20%7E

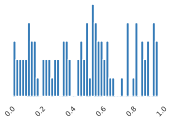
Variables

a

Real number ($\mathbb{R}_{\geq 0}$)

`UNIQUE`

Distinct	1	**Minimum**	0.018637
Distinct (%)	100.0	**Maximum**	0.976000
Missing		**Zeros**	
Missing (%)	0.0	**Zeros (%)**	
Infinite		**Negative**	
Infinite (%)	0.0	**Negative (%)**	
Mean	0.48892145	**Memory size**	92

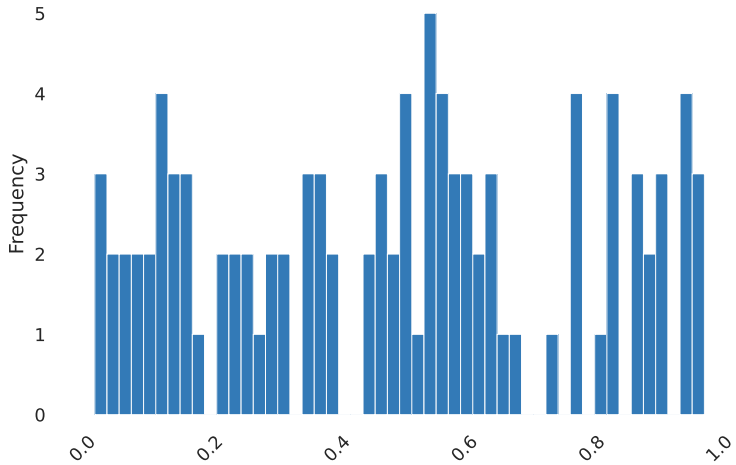


Quantile statistics

Minimum	0.01863781815
5-th percentile	0.0600468312
Q1	0.2364742718
median	0.5075457962
Q3	0.698722507
95-th percentile	0.9430209829
Maximum	0.9760005096
Range	0.9573626915
Interquartile range (IQR)	0.4622482352

Descriptive statistics

Standard deviation	0.2850460218
Coefficient of variation (CV)	0.5830098481
Kurtosis	-1.128119121
Mean	0.4889214526
Median Absolute Deviation (MAD)	0.2582513964
Skewness	0.06043034897
Sum	48.89214526
Variance	0.08125123455
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
0.8789580073	1	1.0%
0.3852844112	1	1.0%
0.4743185557	1	1.0%
0.230612358	1	1.0%
0.04450222217	1	1.0%
0.2518389107	1	1.0%
0.7784752198	1	1.0%
0.2384282431	1	1.0%
0.9143658594	1	1.0%
0.4645200299	1	1.0%
Other values (90)	90	90.0%

Value	Count	Frequency (%)
0.01863781815	1	1.0%
0.0193058458	1	1.0%
0.03699403575	1	1.0%
0.04450222217	1	1.0%
0.05664978566	1	1.0%
0.06022562307	1	1.0%
0.06667730756	1	1.0%
0.09381564513	1	1.0%
0.0940088209	1	1.0%
0.1033454994	1	1.0%

Value	Count	Frequency (%)
0.9760005096	1	1.0%
0.9759091585	1	1.0%
0.9572411122	1	1.0%
0.9548071685	1	1.0%
0.9454224594	1	1.0%
0.9428945894	1	1.0%
0.9405438851	1	1.0%
0.916814053	1	1.0%
0.9143658594	1	1.0%
0.914042992	1	1.0%

b

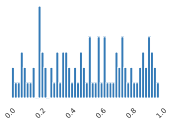
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION (This variable has a high correlation with 1 fields: e)

UNIQUE

Distinct	1
Distinct (%)	100.0
Missing	
Missing (%)	0.0
Infinite	
Infinite (%)	0.0
Mean	0.51197787

Minimum	0.006509
Maximum	0.98516
Zeros	
Zeros (%)	
Negative	
Negative (%)	
Memory size	9

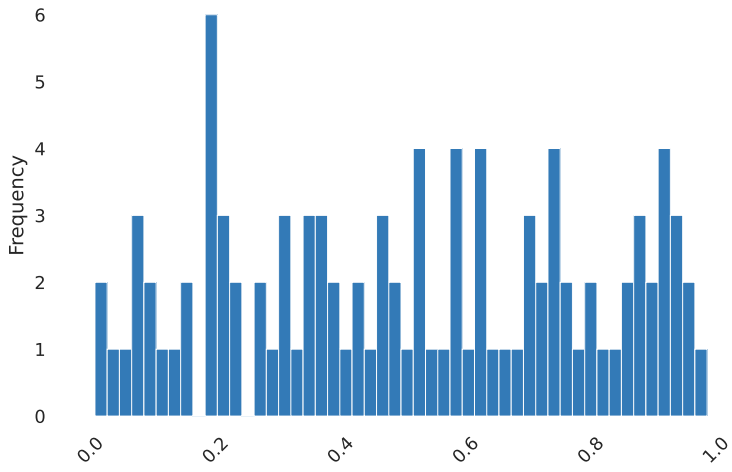


Quantile statistics

Descriptive statistics

Minimum	0.006509098205
5-th percentile	0.07920263672
Q1	0.2692457856
median	0.5198696259
Q3	0.7451787293
95-th percentile	0.9292412896
Maximum	0.9851641865
Range	0.9786550883
Interquartile range (IQR)	0.4759329437

Standard deviation	0.2818424917
Coefficient of variation (CV)	0.5504974011
Kurtosis	-1.200767154
Mean	0.5119778789
Median Absolute Deviation (MAD)	0.233582135
Skewness	-0.0398632784
Sum	51.19778789
Variance	0.07943519015
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
0.1884279066	1	1.0%
0.7168257846	1	1.0%
0.006509098205	1	1.0%
0.7468078584	1	1.0%
0.1923460766	1	1.0%
0.2922368685	1	1.0%
0.4630055204	1	1.0%
0.5242201023	1	1.0%
0.7446356863	1	1.0%
0.5869803099	1	1.0%
Other values (90)	90	90.0%

Value	Count	Frequency (%)
0.006509098205	1	1.0%
0.02252102267	1	1.0%
0.02618680483	1	1.0%
0.05887603179	1	1.0%
0.06907094958	1	1.0%
0.07973588341	1	1.0%
0.0840777849	1	1.0%
0.09035251008	1	1.0%
0.09685886099	1	1.0%
0.1065508176	1	1.0%

Value	Count	Frequency (%)
0.9851641865	1	1.0%
0.9510888561	1	1.0%
0.9509403283	1	1.0%
0.9389056165	1	1.0%
0.9384230195	1	1.0%
0.9287580407	1	1.0%
0.9262456971	1	1.0%
0.9235217976	1	1.0%
0.9100670787	1	1.0%
0.9082698258	1	1.0%

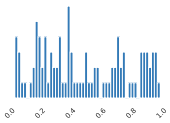
C

Real number ($\mathbb{R}_{\geq 0}$)

`UNIQUE`

Distinct	1
Distinct (%)	100.0
Missing	
Missing (%)	0.0
Infinite	
Infinite (%)	0.0
Mean	0.48603349

Minimum	0.0319041
Maximum	0.99074
Zeros	
Zeros (%)	
Negative	
Negative (%)	
Memory size	92

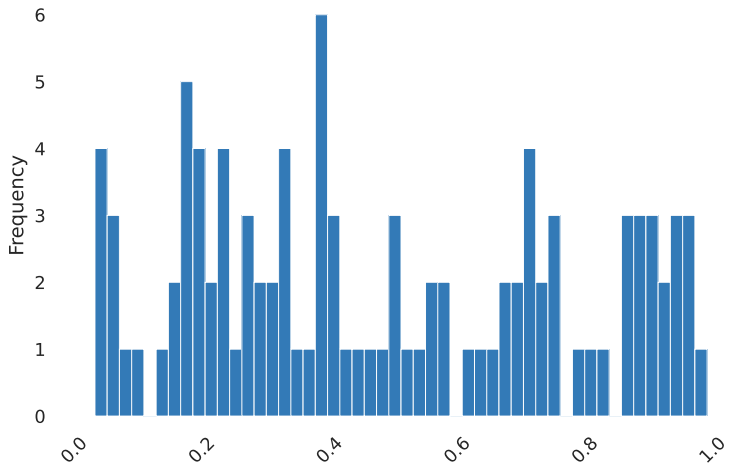


Quantile statistics

Descriptive statistics

Minimum	0.03190411161
5-th percentile	0.0568078603
Q1	0.2375779607
median	0.4183085392
Q3	0.7244614964
95-th percentile	0.9483966008
Maximum	0.990740819
Range	0.9588367074
Interquartile range (IQR)	0.4868835358

Standard deviation	0.2886738619
Coefficient of variation (CV)	0.5939382102
Kurtosis	-1.251207609
Mean	0.4860334912
Median Absolute Deviation (MAD)	0.2392638431
Skewness	0.2062087227
Sum	48.60334912
Variance	0.08333259853
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
0.4013505218	1	1.0%
0.402687468	1	1.0%
0.6909327429	1	1.0%
0.8769519661	1	1.0%
0.8920660923	1	1.0%
0.234349196	1	1.0%
0.7875627508	1	1.0%
0.7195961524	1	1.0%
0.9236370292	1	1.0%
0.3155761892	1	1.0%
Other values (90)	90	90.0%

Value	Count	Frequency (%)
0.03190411161	1	1.0%
0.0427773222	1	1.0%
0.04691651125	1	1.0%
0.04993939951	1	1.0%
0.05517628529	1	1.0%
0.05689373267	1	1.0%
0.05868180749	1	1.0%
0.08258646091	1	1.0%
0.09389298493	1	1.0%
0.1321242481	1	1.0%

Value	Count	Frequency (%)
0.990740819	1	1.0%
0.9686204898	1	1.0%
0.9528968307	1	1.0%
0.9526239527	1	1.0%
0.9522475253	1	1.0%
0.9481939206	1	1.0%
0.9472647842	1	1.0%
0.9278662787	1	1.0%
0.9236370292	1	1.0%
0.8999740386	1	1.0%

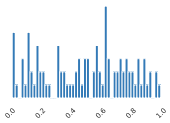
d

Real number ($\mathbb{R}_{\geq 0}$)

UNIQUE

Distinct	1
Distinct (%)	100.0
Missing	
Missing (%)	0.0
Infinite	
Infinite (%)	0.0
Mean	0.49406827

Minimum	0.0122359
Maximum	0.995262
Zeros	
Zeros (%)	
Negative	
Negative (%)	
Memory size	92

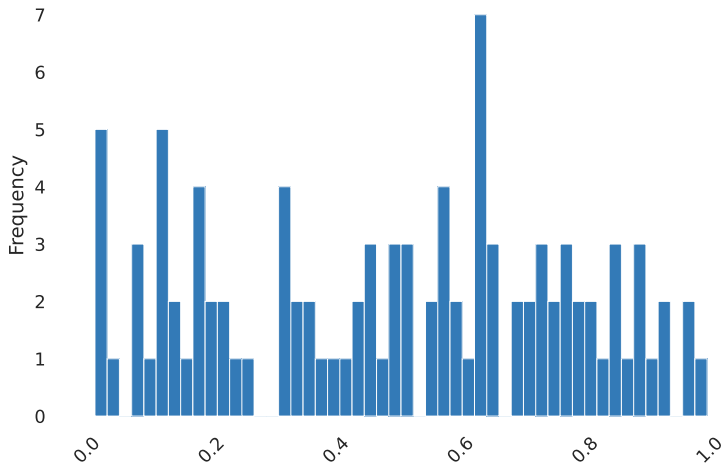


Quantile statistics

Descriptive statistics

Minimum	0.01223595601
5-th percentile	0.04932439312
Q1	0.2159359376
median	0.5110839082
Q3	0.7258316909
95-th percentile	0.9159420373
Maximum	0.9952625833
Range	0.9830266272
Interquartile range (IQR)	0.5098957532

Standard deviation	0.2808410044
Coefficient of variation (CV)	0.5684255001
Kurtosis	-1.168965338
Mean	0.4940682716
Median Absolute Deviation (MAD)	0.2446427503
Skewness	-0.1137620653
Sum	49.40682716
Variance	0.07887166973
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
0.7656867404	1	1.0%
0.6552999385	1	1.0%
0.02103871913	1	1.0%
0.3397323975	1	1.0%
0.5000892399	1	1.0%
0.8870149788	1	1.0%
0.1762646647	1	1.0%
0.6286415596	1	1.0%
0.07680352272	1	1.0%
0.6338592278	1	1.0%
Other values (90)	90	90.0%

Value	Count	Frequency (%)
0.01223595601	1	1.0%
0.02054682623	1	1.0%
0.02103871913	1	1.0%
0.02580243255	1	1.0%
0.02777470476	1	1.0%
0.05045858724	1	1.0%
0.07420877827	1	1.0%
0.07680352272	1	1.0%
0.08084453089	1	1.0%
0.1046241613	1	1.0%

Value	Count	Frequency (%)
0.9952625833	1	1.0%
0.9735132842	1	1.0%
0.9622377113	1	1.0%
0.9305529964	1	1.0%
0.9291570377	1	1.0%
0.9152465109	1	1.0%
0.8925624861	1	1.0%
0.8873660378	1	1.0%
0.8870149788	1	1.0%
0.8688061473	1	1.0%

e

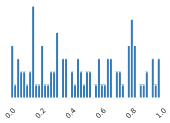
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION (This variable has a high correlation with 1 fields: b)

UNIQUE

Distinct	1
Distinct (%)	100.0
Missing	
Missing (%)	0.0
Infinite	
Infinite (%)	0.0
Mean	0.47049194

Minimum	0.0003219
Maximum	0.9925
Zeros	
Zeros (%)	
Negative	
Negative (%)	
Memory size	

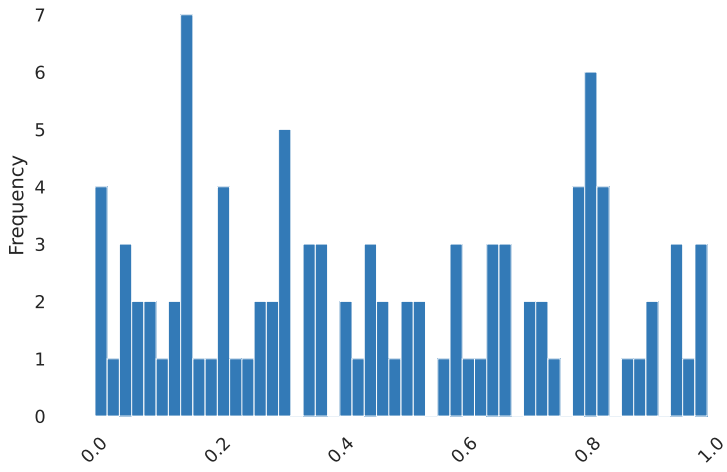


Quantile statistics

Descriptive statistics

Minimum	0.0003219040324
5-th percentile	0.04420765636
Q1	0.2028685927
median	0.4486478797
Q3	0.752542013
95-th percentile	0.946571387
Maximum	0.9925560152
Range	0.9922341112
Interquartile range (IQR)	0.5496734203

Standard deviation	0.2981249551
Coefficient of variation (CV)	0.6336451875
Kurtosis	-1.286223447
Mean	0.4704919424
Median Absolute Deviation (MAD)	0.2724700296
Skewness	0.09649435472
Sum	47.04919424
Variance	0.08887848883
Monotonicity	Not monotonic



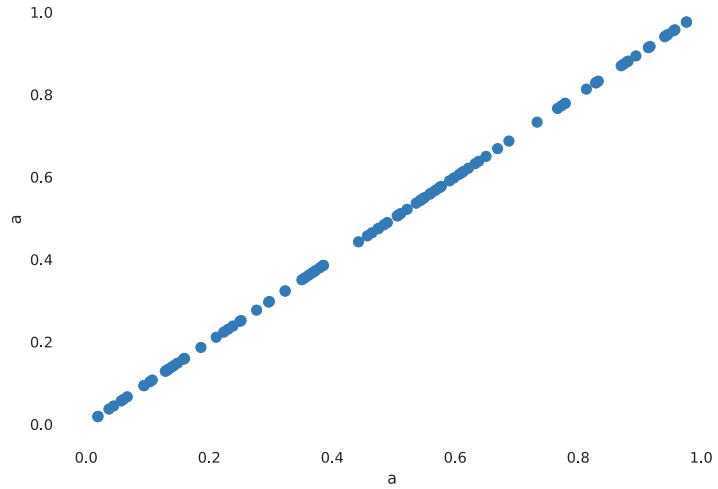
Histogram with fixed size bins (bins=50)

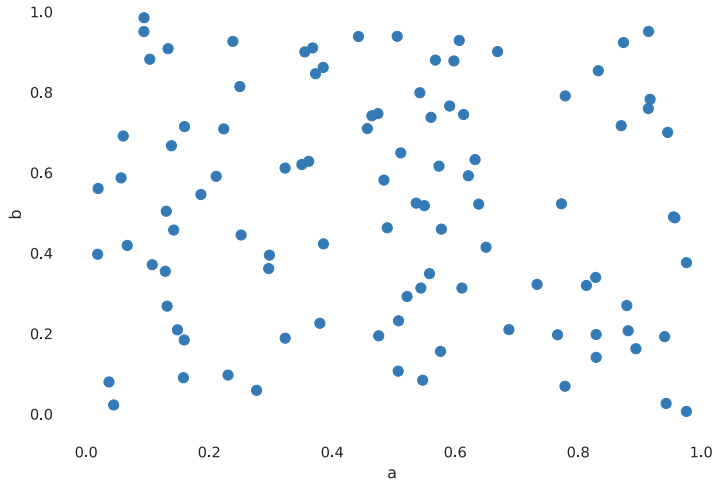
Value	Count	Frequency (%)
0.6690943647	1	1.0%
0.2136346713	1	1.0%
0.5792908849	1	1.0%
0.2763320622	1	1.0%
0.03452908108	1	1.0%
0.8019927495	1	1.0%
0.2913747076	1	1.0%
0.9543189282	1	1.0%
0.133494912	1	1.0%
0.0003219040324	1	1.0%
Other values (90)	90	90.0%

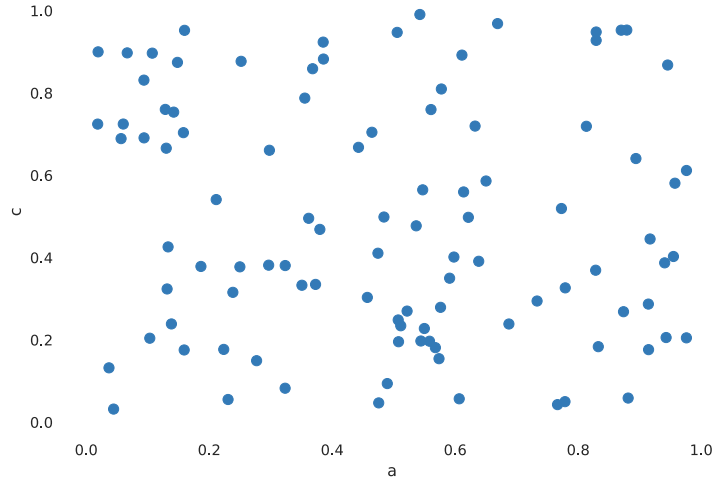
Value	Count	Frequency (%)
0.0003219040324	1	1.0%
0.00213601801	1	1.0%
0.005256986642	1	1.0%
0.007236904362	1	1.0%
0.03452908108	1	1.0%
0.04471705506	1	1.0%
0.05148909728	1	1.0%
0.05441638016	1	1.0%
0.06137133543	1	1.0%
0.06693420819	1	1.0%

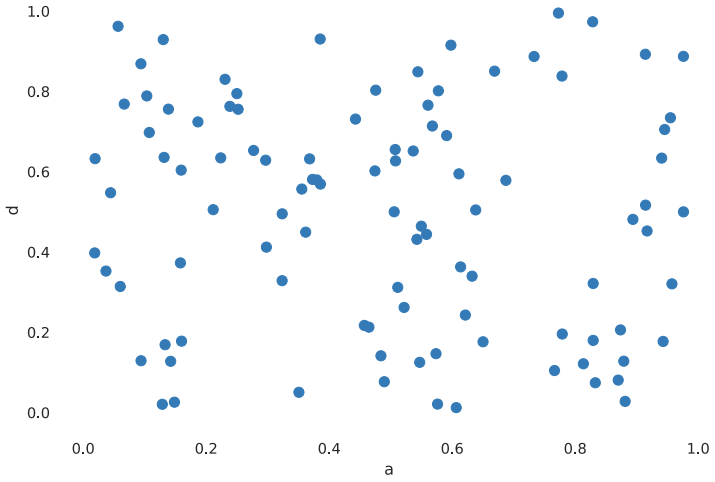
Value	Count	Frequency (%)
0.9925560152	1	1.0%
0.9782755368	1	1.0%
0.972921715	1	1.0%
0.9543189282	1	1.0%
0.9470676883	1	1.0%
0.9465452658	1	1.0%
0.9450627265	1	1.0%
0.9108799965	1	1.0%
0.9092464346	1	1.0%
0.8874639287	1	1.0%

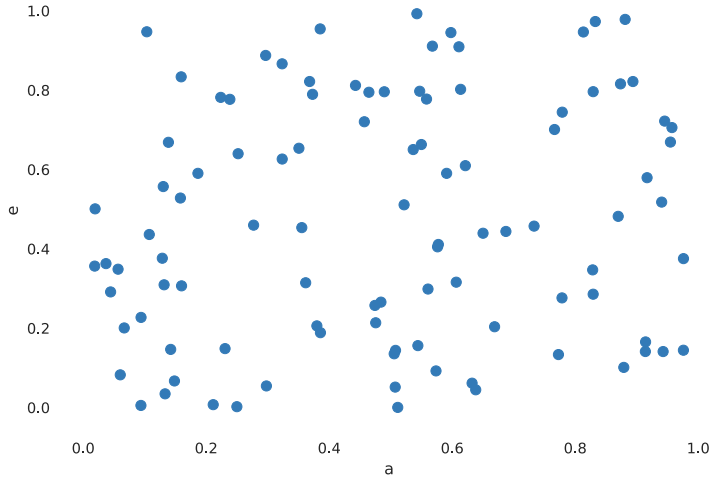
Interactions

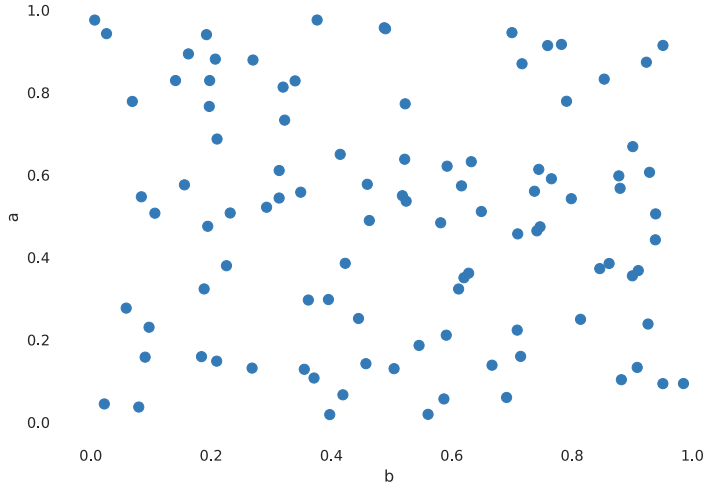

1.0
0.8
0.6
0.4
0.2
0.0
a
0.0
0.2
0.4
0.6
0.8
1.0
a

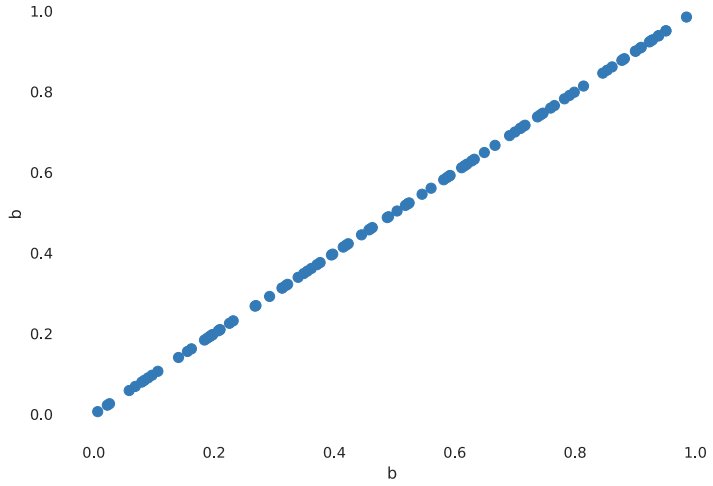

1.0
0.8
0.6
0.4
0.2
0.0
b
0.0
0.2
0.4
0.6
0.8
1.0
a

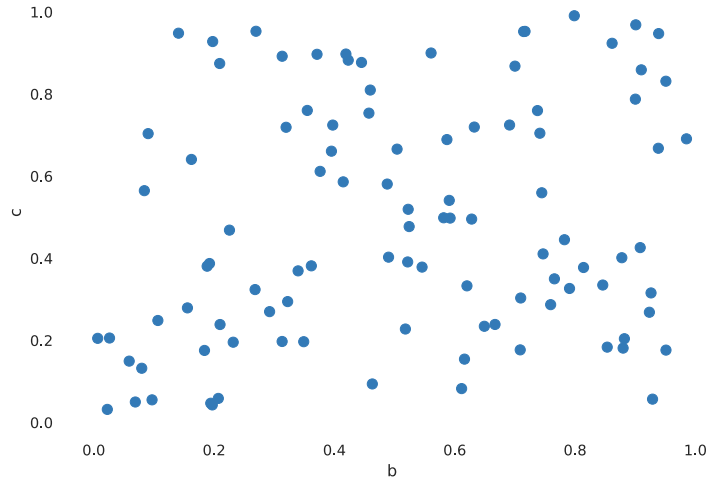

1.0
0.8
0.6
0.4
0.2
0.0
c
0.0
0.2
0.4
0.6
0.8
1.0
a

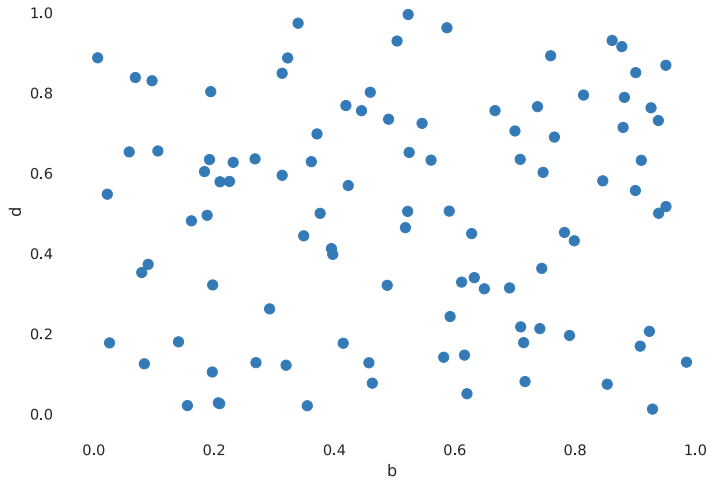

1.0
0.8
0.6
0.4
0.2
0.0
d
0.0
0.2
0.4
0.6
0.8
1.0
a

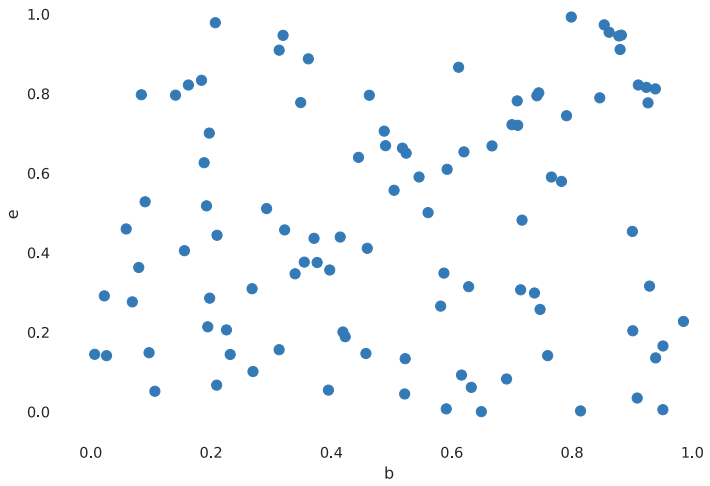

1.0
0.8
0.6
0.4
0.2
0.0
e
0.0
0.2
0.4
0.6
0.8
1.0
a

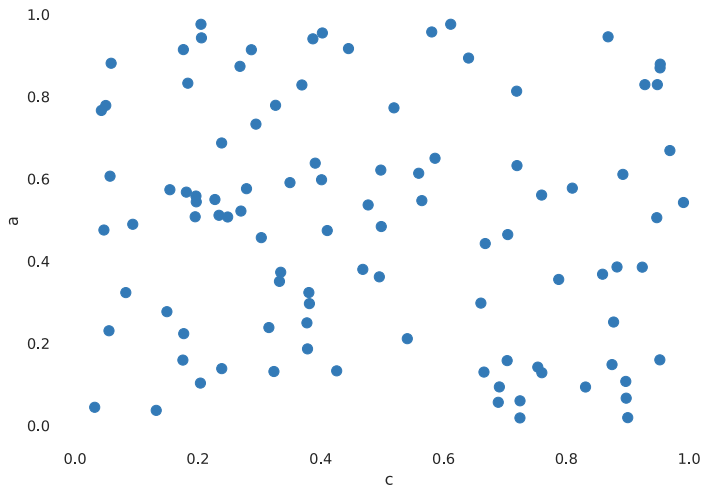

1.0
0.8
0.6
0.4
0.2
0.0
a
0.0
0.2
0.4
0.6
0.8
1.0
b

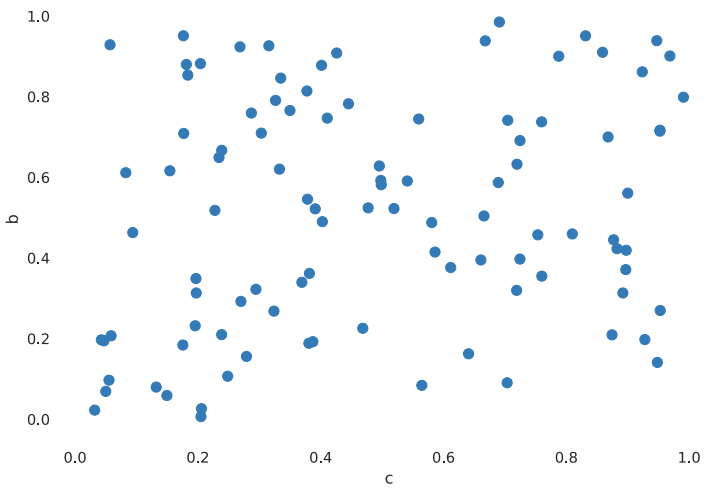

1.0
0.8
0.6
0.4
0.2
0.0
b
0.0
0.2
0.4
0.6
0.8
1.0
b

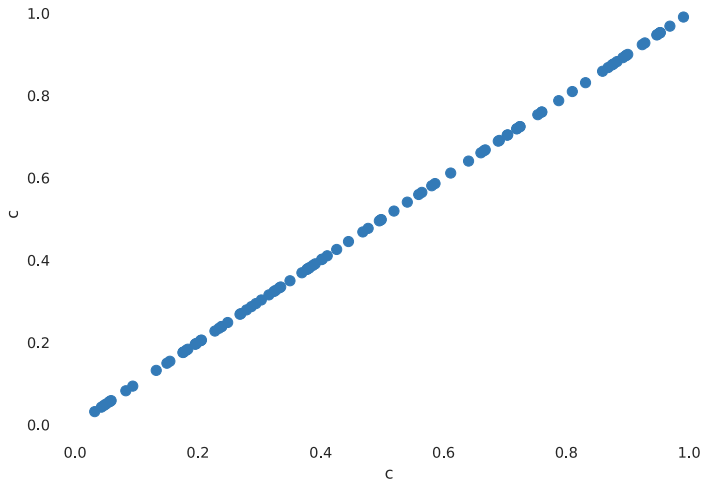

1.0
0.8
0.6
0.4
0.2
0.0
c
0.0
0.2
0.4
0.6
0.8
1.0
b

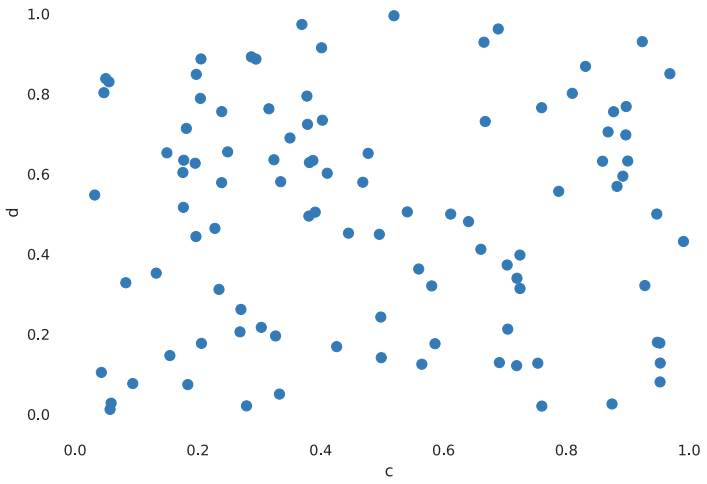

1.0
0.8
0.6
0.4
0.2
0.0
d
0.0
0.2
0.4
0.6
0.8
1.0
b

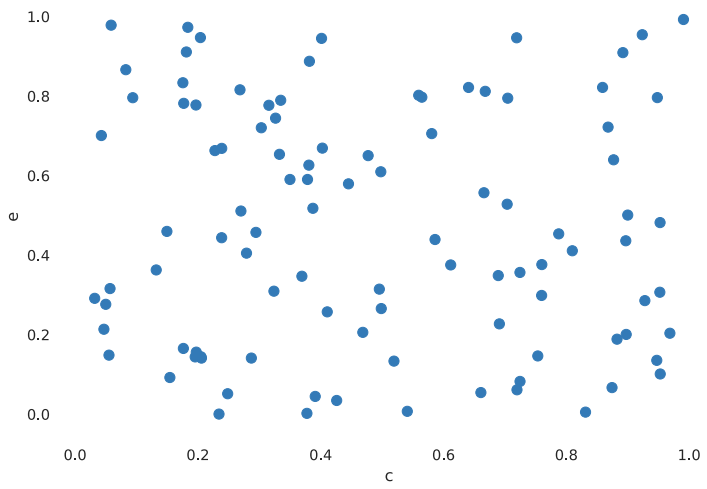

1.0
0.8
0.6
0.4
0.2
0.0
e
0.0
0.2
0.4
0.6
0.8
1.0
b

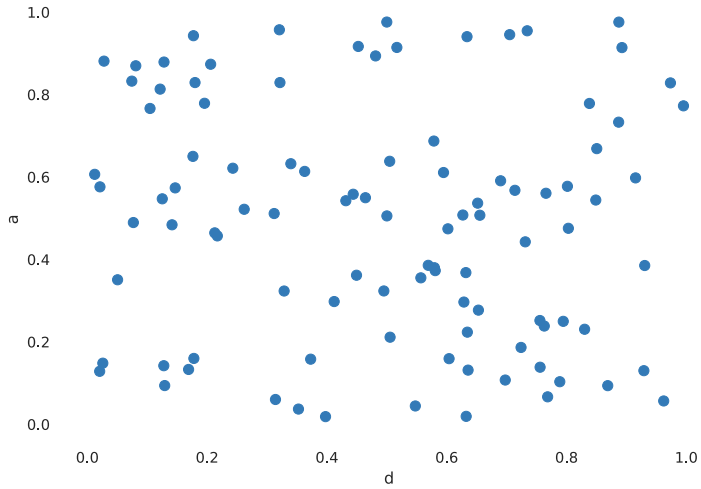

1.0
0.8
0.6
0.4
0.2
0.0
a
0.0
0.2
0.4
0.6
0.8
1.0
c

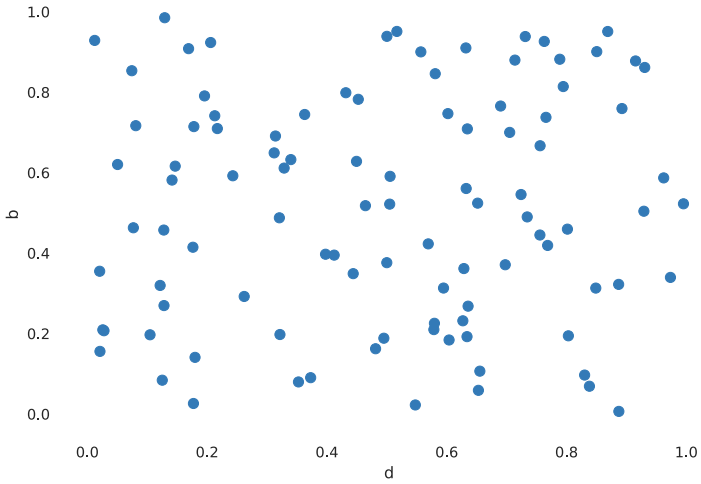

1.0
0.8
0.6
0.4
0.2
0.0
b
0.0
0.2
0.4
0.6
0.8
1.0
c

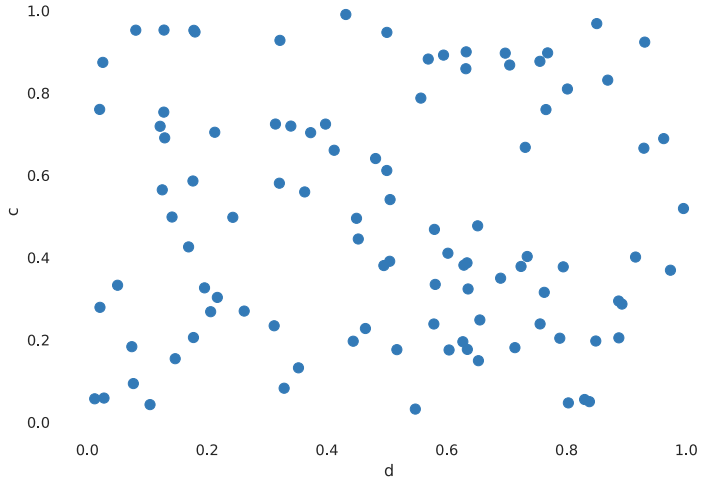

1.0
0.8
0.6
0.4
0.2
0.0
c
0.0
0.2
0.4
0.6
0.8
1.0
c

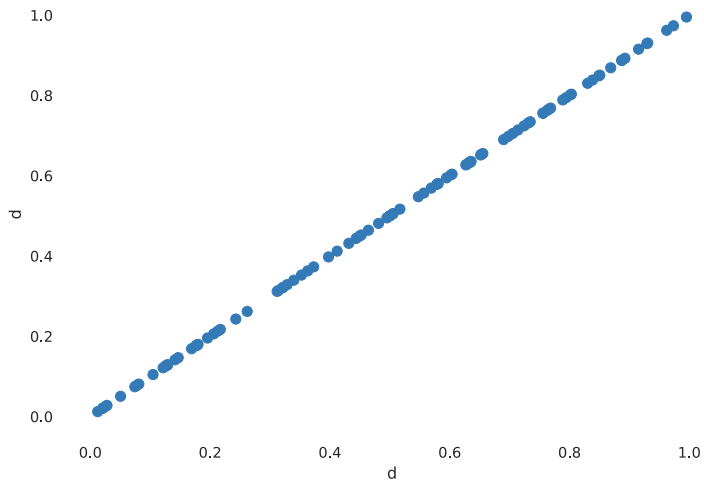

1.0
0.8
0.6
0.4
0.2
0.0
d
0.0
0.2
0.4
0.6
0.8
1.0
c

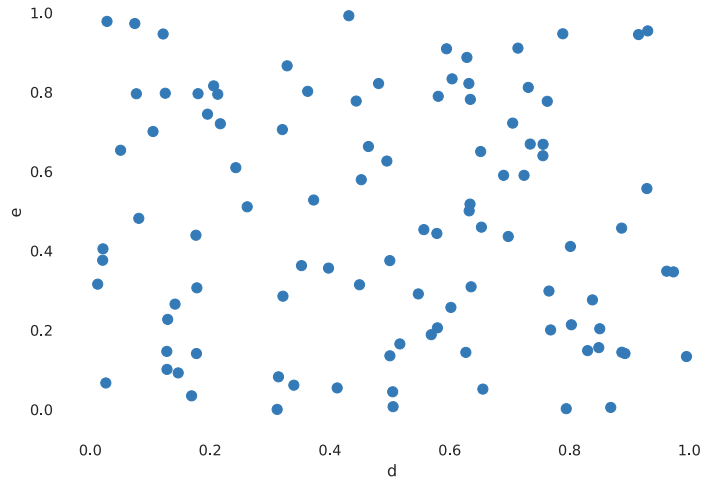

1.0
0.8
0.6
0.4
0.2
0.0
e
0.0
0.2
0.4
0.6
0.8
1.0
c

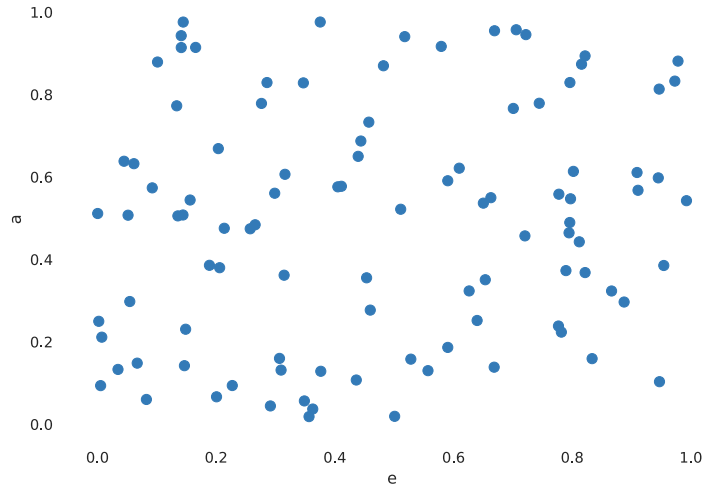

1.0
0.8
0.6
0.4
0.2
0.0
a
0.0
0.2
0.4
0.6
0.8
1.0
d

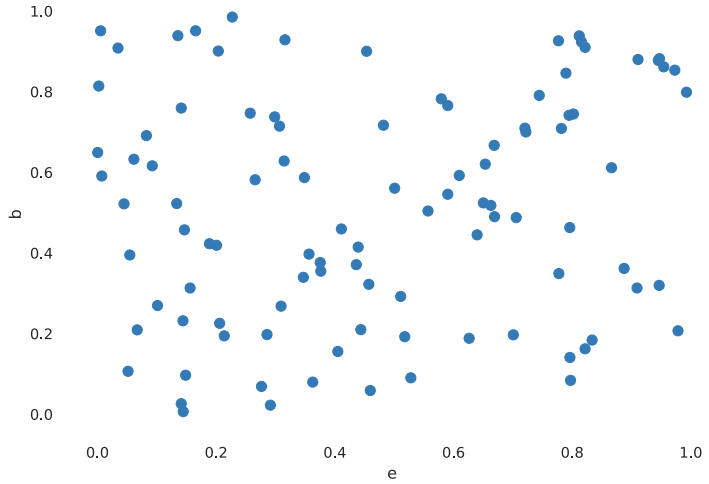

1.0
0.8
0.6
0.4
0.2
0.0
b
0.0
0.2
0.4
0.6
0.8
1.0
d

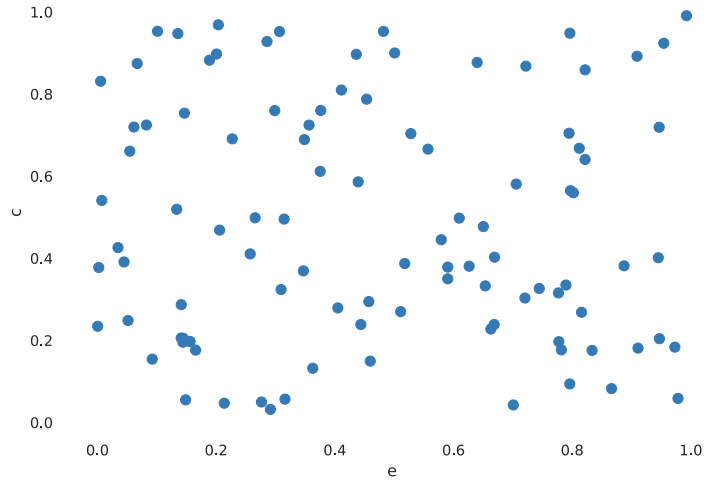

1.0
0.8
0.6
0.4
0.2
0.0
c
0.0
0.2
0.4
0.6
0.8
1.0
d

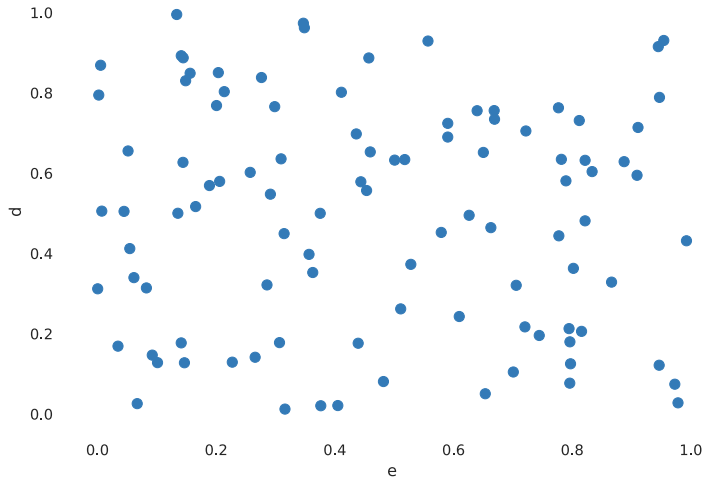

1.0
0.8
0.6
0.4
0.2
0.0
d
0.0
0.2
0.4
0.6
0.8
1.0
d

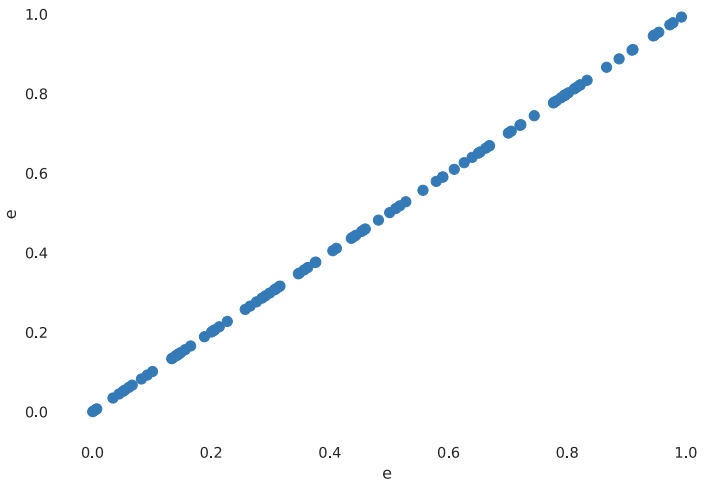

1.0
0.8
0.6
0.4
0.2
0.0
e
0.0
0.2
0.4
0.6
0.8
1.0
d


1.0
0.8
0.6
0.4
0.2
0.0
a
0.0
0.2
0.4
0.6
0.8
1.0
e

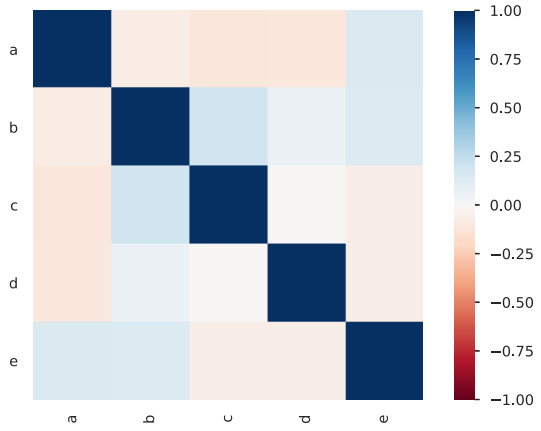

1.0
0.8
0.6
0.4
0.2
0.0
b
0.0
0.2
0.4
0.6
0.8
1.0
e

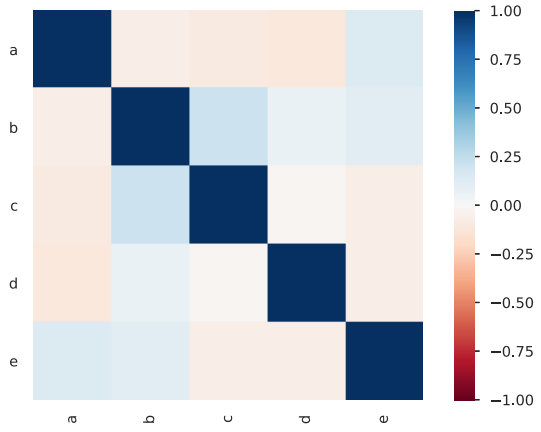

1.0
0.8
0.6
0.4
0.2
0.0
c
0.0
0.2
0.4
0.6
0.8
1.0
e

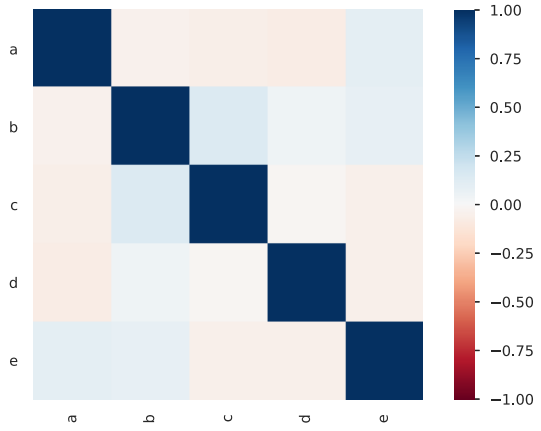

1.0
0.8
0.6
0.4
0.2
0.0
d
0.0
0.2
0.4
0.6
0.8
1.0
e

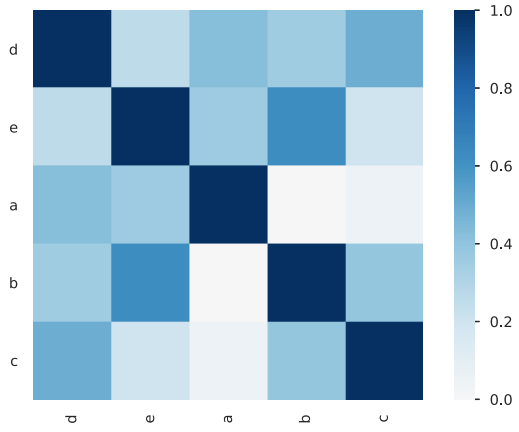


Correlations

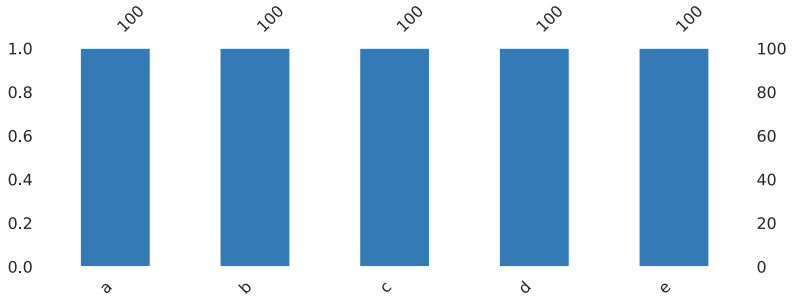




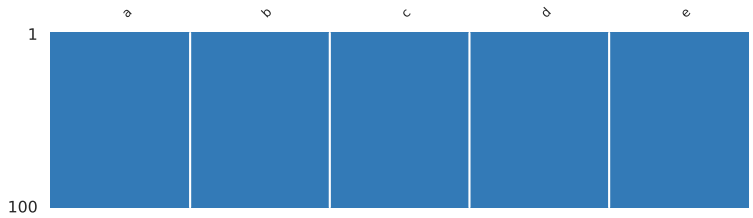




Missing values



A simple visualization of nullity by column.



Nullity matrix is a data-dense display which lets you quickly visually pick out patterns in data completion.

Sample

First rows

	a	b	c	d	e
0	0.474319	0.746808	0.410638	0.601951	0.257351
1	0.036994	0.079736	0.132124	0.352477	0.362746
2	0.573613	0.616188	0.154347	0.146668	0.092282
3	0.590882	0.765660	0.349927	0.689958	0.590261
4	0.650082	0.414614	0.586145	0.176265	0.439305
5	0.940544	0.192346	0.387238	0.633859	0.517776
6	0.323544	0.188428	0.380691	0.495017	0.626307
7	0.361696	0.628085	0.495461	0.449375	0.314486
8	0.914366	0.951089	0.176332	0.516746	0.165373
9	0.355386	0.900249	0.787563	0.556787	0.453485

Last rows

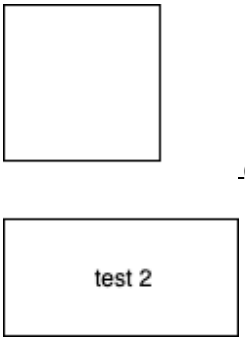
	a	b	c	d	e
90	0.094009	0.985164	0.690933	0.129119	0.227148
91	0.107413	0.371081	0.896800	0.697707	0.436130
92	0.954807	0.489968	0.402687	0.734371	0.669094
93	0.577448	0.459604	0.809721	0.801519	0.411013
94	0.916814	0.782351	0.445211	0.452333	0.579291
95	0.484077	0.581595	0.498598	0.141363	0.265662
96	0.475572	0.194423	0.046917	0.803091	0.213635
97	0.128581	0.354901	0.760073	0.020547	0.376285
98	0.542505	0.798708	0.990741	0.431543	0.992556
99	0.957241	0.487789	0.580769	0.320590	0.705632

Report generated with pandas-profiling (https://github.com/pandas-profiling/pandas-profiling).

There is a lot more that you can do with outputs (such as including interactive outputs) with your book. For more information about this, see the Jupyter Book documentation (https://jupyterbook.org)

Data Cleaning

You can also create content with Jupyter Notebooks. This means that you can include code blocks and their outputs in your book.



(_images/test.drawio.png)

Fig. 2 Some architecture

```
import $ivy.`org.apache.spark::spark-sql:2.4.0` // Or use any other 2.x version here
import $ivy.`sh.almond::almond-spark:0.10.9` // Not required since almond 0.7.0 (will be
 automatically added when importing spark)
```

```
import $ivy.$                                    // Or use any other 2.x version here

import $ivy.$                                  // Not required since almond 0.7.0 (will b
e automatically added when importing spark)
```

```
import org.apache.log4j.{Level, Logger}
Logger.getLogger("org").setLevel(Level.OFF)
```

```
import org.apache.log4j.{Level, Logger}
```

```
import org.apache.spark.sql._

val spark = {
  NotebookSparkSession.builder()
    .master("local[*]")
    .getOrCreate()
}
```

```
Loading spark-stubs
```

```
Getting spark JARs
```

```
java.lang.Exception: Error starting class server at http://x86_64-conda-linux-gnu:38551
  org.apache.spark.sql.ammonitesparkinternals.AmmoniteClassServer.<init>(AmmoniteClassS
erver.scala:60)
  org.apache.spark.sql.ammonitesparkinternals.AmmoniteSparkSessionBuilder.getOrCreate(A
mmoniteSparkSessionBuilder.scala:268)
  org.apache.spark.sql.almondinternals.NotebookSparkSessionBuilder.getOrCreate(Notebook
SparkSessionBuilder.scala:62)
  ammonite.$sess.cmd2$Helper.<init>(cmd2.sc:5)
  ammonite.$sess.cmd2$.<init>(cmd2.sc:7)
  ammonite.$sess.cmd2$.<clinit>(cmd2.sc:-1)
java.net.SocketException: Unresolved address
  sun.nio.ch.Net.translateToSocketException(Net.java:131)
  sun.nio.ch.Net.translateException(Net.java:157)
  sun.nio.ch.Net.translateException(Net.java:163)
  sun.nio.ch.ServerSocketAdaptor.bind(ServerSocketAdaptor.java:76)
  org.eclipse.jetty.server.ServerConnector.openAcceptChannel(ServerConnector.java:345)
  org.eclipse.jetty.server.ServerConnector.open(ServerConnector.java:310)
  org.eclipse.jetty.server.AbstractNetworkConnector.doStart(AbstractNetworkConnector.ja
va:80)
  org.eclipse.jetty.server.ServerConnector.doStart(ServerConnector.java:234)
  org.eclipse.jetty.util.component.AbstractLifeCycle.start(AbstractLifeCycle.java:72)
  org.eclipse.jetty.server.Server.doStart(Server.java:386)
  org.eclipse.jetty.util.component.AbstractLifeCycle.start(AbstractLifeCycle.java:72)
  org.apache.spark.sql.ammonitesparkinternals.AmmoniteClassServer.<init>(AmmoniteClassS
erver.scala:57)
  org.apache.spark.sql.ammonitesparkinternals.AmmoniteSparkSessionBuilder.getOrCreate(A
mmoniteSparkSessionBuilder.scala:268)
  org.apache.spark.sql.almondinternals.NotebookSparkSessionBuilder.getOrCreate(Notebook
SparkSessionBuilder.scala:62)
  ammonite.$sess.cmd2$Helper.<init>(cmd2.sc:5)
  ammonite.$sess.cmd2$.<init>(cmd2.sc:7)
  ammonite.$sess.cmd2$.<clinit>(cmd2.sc:-1)
java.nio.channels.UnresolvedAddressException
  sun.nio.ch.Net.checkAddress(Net.java:101)
  sun.nio.ch.ServerSocketChannelImpl.bind(ServerSocketChannelImpl.java:218)
  sun.nio.ch.ServerSocketAdaptor.bind(ServerSocketAdaptor.java:74)
  org.eclipse.jetty.server.ServerConnector.openAcceptChannel(ServerConnector.java:345)
  org.eclipse.jetty.server.ServerConnector.open(ServerConnector.java:310)
  org.eclipse.jetty.server.AbstractNetworkConnector.doStart(AbstractNetworkConnector.ja
va:80)
  org.eclipse.jetty.server.ServerConnector.doStart(ServerConnector.java:234)
  org.eclipse.jetty.util.component.AbstractLifeCycle.start(AbstractLifeCycle.java:72)
  org.eclipse.jetty.server.Server.doStart(Server.java:386)
  org.eclipse.jetty.util.component.AbstractLifeCycle.start(AbstractLifeCycle.java:72)
  org.apache.spark.sql.ammonitesparkinternals.AmmoniteClassServer.<init>(AmmoniteClassS
erver.scala:57)
  org.apache.spark.sql.ammonitesparkinternals.AmmoniteSparkSessionBuilder.getOrCreate(A
mmoniteSparkSessionBuilder.scala:268)
  org.apache.spark.sql.almondinternals.NotebookSparkSessionBuilder.getOrCreate(Notebook
SparkSessionBuilder.scala:62)
  ammonite.$sess.cmd2$Helper.<init>(cmd2.sc:5)
  ammonite.$sess.cmd2$.<init>(cmd2.sc:7)
  ammonite.$sess.cmd2$.<clinit>(cmd2.sc:-1)
```

```
val test = "test"
```

```
test: String = "test"
```

```
println(test)
```

```
test
```

There is a lot more that you can do with outputs (such as including interactive outputs) with your book. For more information about this, see the Jupyter Book documentation (https://jupyterbook.org)

By Tom Michiels