



# 帅器学习

TechYoung Machine Learning Class

## 数据集

# 本工作的意义

收集一些常见的数据集，尝试解释这些数据集的意义，拓展视野，理解数据分析科学在试图解决社会，经济，政治，和文化问题时的思路和思考方式。

<http://archive.ics.uci.edu/ml/index.php>

# UC Irvine Machine Learning Repository

← → ↺ 不安全 | archive.ics.uci.edu/ml/index.php

UCI Machine Learning Repository  
Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact



Search

Repository Web

**View ALL Data Sets**

Welcome to the UC Irvine Machine Learning Repository!


We currently maintain 481 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About](#) page. For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).

Supported By:  In Collaboration With: 








Latest News:

- 09-24-2018: Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidoul
- 04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
- 03-01-2010: [Note](#) from donor regarding Netflix data
- 10-16-2009: Two new data sets have been added.
- 09-14-2009: Several data sets have been added.
- 03-24-2008: New data sets have been added!
- 06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope








Featured Data Set: [Statlog \(Landsat Satellite\)](#)

 Task: Classification  
Data Type: Multivariate  
# Attributes: 36  
# Instances: 6435

Newest Data Sets:

- 07-30-2019:  [PPG-DaLiA](#)
- 07-24-2019:  [Divorce Predictors data set](#)
- 07-22-2019:  [Alcohol QCM Sensor Dataset](#)
- 07-14-2019:  [Incident management process enriched event log](#)
- 06-30-2019:  [Wave Energy Converters](#)
- 06-22-2019:  [Query Analytics Workloads Dataset](#)
- 06-17-2019:  [Opinion Corpus for Lebanese Arabic Reviews \(OCLAR\)](#)

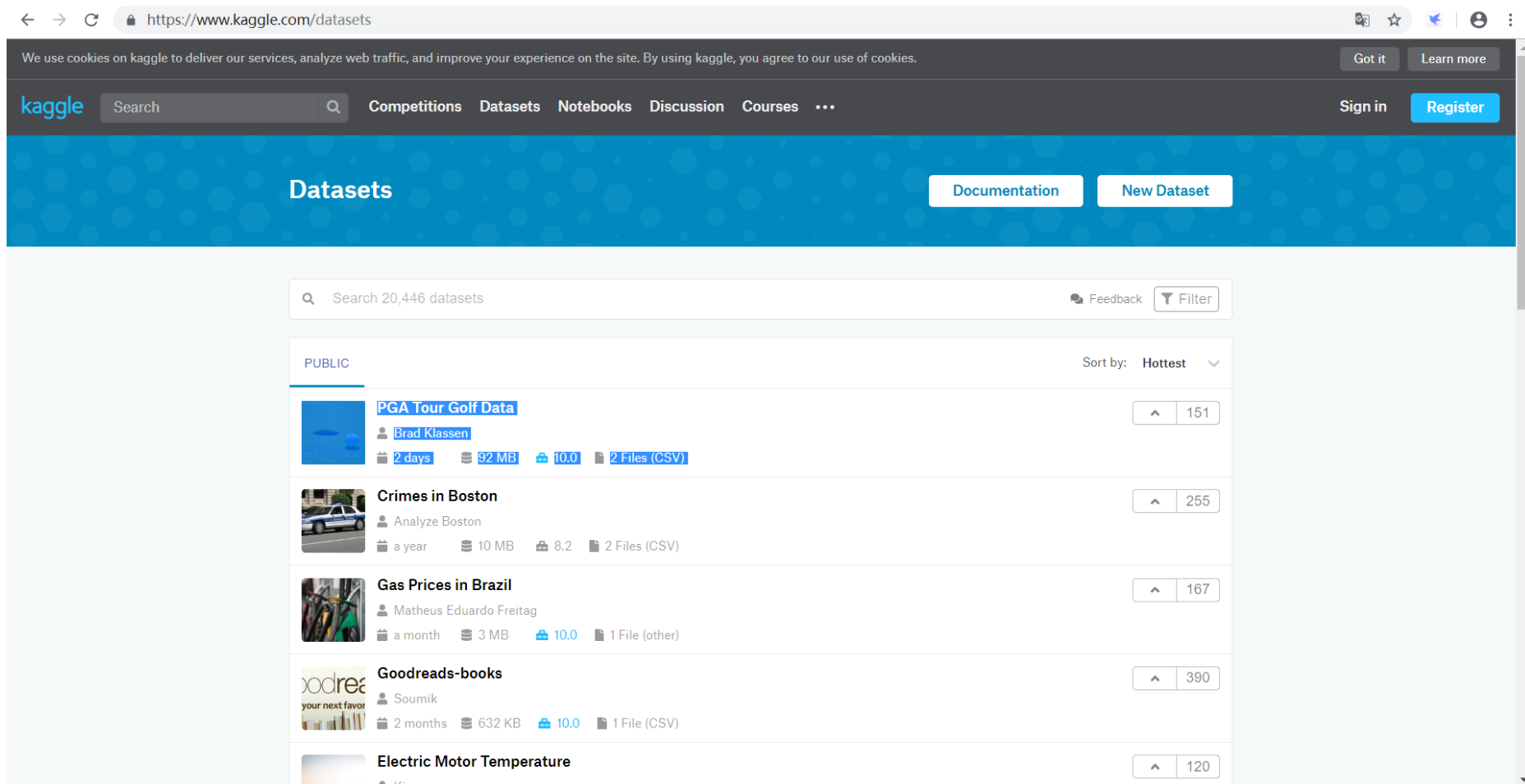
Most Popular Data Sets (hits since 2007):

- 2781342:  [Iris](#)
- 1558791:  [Adult](#)
- 1208821:  [Wine](#)
- 1022797:  [Car Evaluation](#)
- 999942:  [Wine Quality](#)
- 988352:  [Heart Disease](#)
- 978837:  [Breast Cancer Wisconsin \(Diagnostic\)](#)

点击这里查看所有数据集

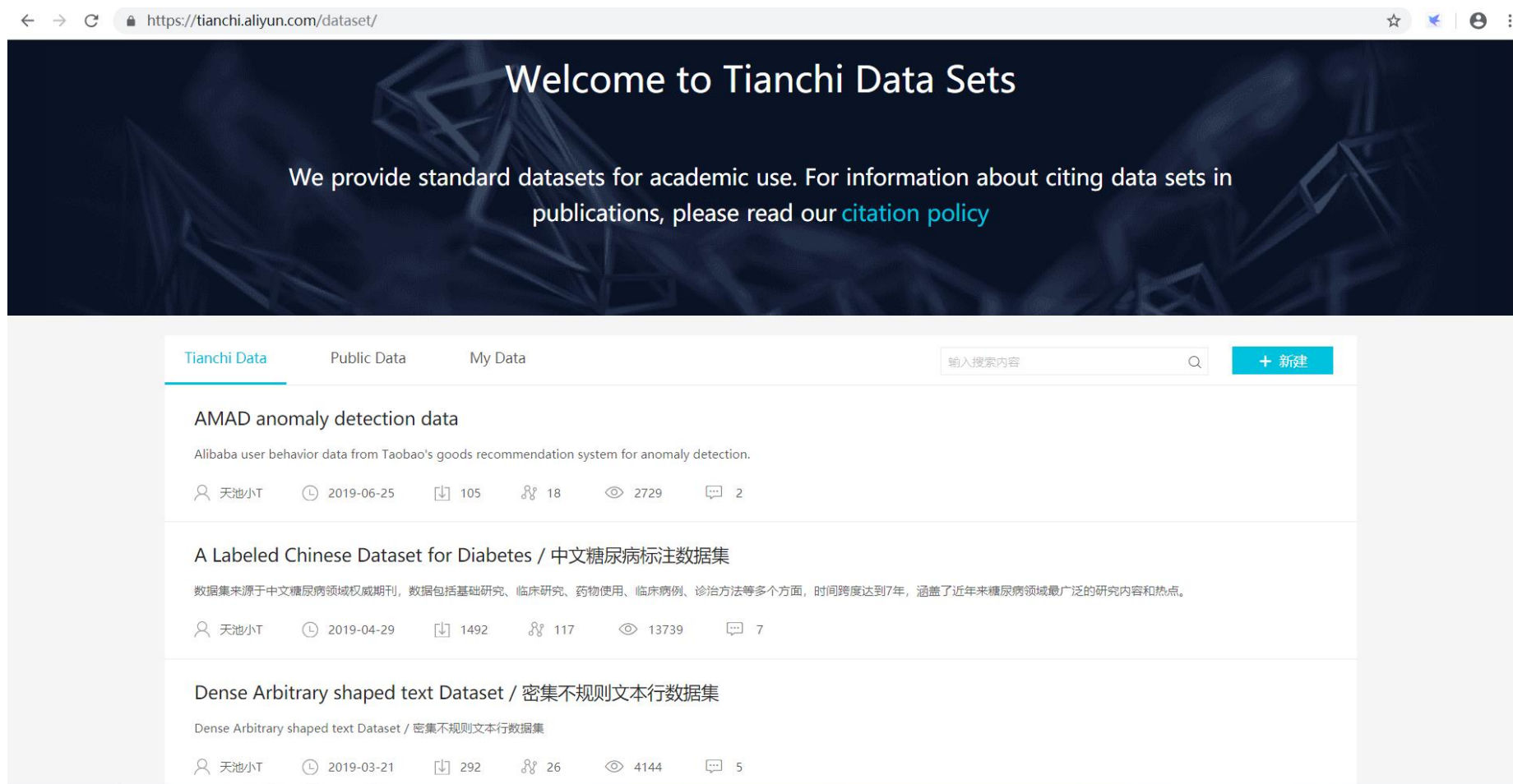
<http://archive.ics.uci.edu/ml/index.php>

# Kaggle Datasets



<https://www.kaggle.com/datasets>

# Tianchi Data Sets



<https://tianchi.aliyun.com/dataset/>

# 样例-共享单车数据

特征	解释
dteday	单车用户使用日期
season	单车用户使用季节 (春夏秋冬: 1-4)
temp	用户使用时温度 (t-t_min)/(t_max-t_min)
atemp	用户使用时体感温度 (t-t_min)/(t_max-t_min)
hum	当日能见度: 能见度/100
windspeed	风速: 风速/67
casual	该时段未注册用户使用人数
registered	该时段注册用户使用人数
cnt	该时段总用户数量
大量投放	是(>1500):1 否(<1500):0

dteday	season	temp	atemp	hum	windspeed	casual	registered	cnt	大量投放
2011/1/1	1	0.344167	0.363625	0.805833	0.160446	331	654	985	0
2011/1/2	1	0.363478	0.353739	0.696087	0.248539	131	670	801	0
2011/1/3	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349	0
2011/4/15	2	0.446667	0.441913	0.67125	0.226375	642	2484	3126	1
2011/4/16	2	0.430833	0.425492	0.888333	0.340808	121	674	795	0
2011/4/17	2	0.456667	0.445696	0.479583	0.303496	1558	2186	3744	1
2011/7/11	3	0.7625	0.729804	0.635833	0.282337	723	3363	4086	1
2011/7/12	3	0.794167	0.739275	0.559167	0.200254	662	3596	4258	1
2011/7/13	3	0.746667	0.689404	0.631667	0.146133	748	3594	4342	1
2011/9/27	4	0.636667	0.574525	0.885417	0.118171	477	3643	4120	1
2011/9/28	4	0.635	0.575158	0.84875	0.148629	480	3427	3907	1
2011/9/29	4	0.616667	0.574512	0.699167	0.172883	653	4186	4839	1

说明	无
意义	利用气象和其他数据可以预测在该时段和季节是否需要大量投放单车, 且可以通过使用用户在线总人数的数据来监测城市的流动性, 使得单车合理的投放
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset">http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset</a>

# 数据集-共享单车数据

特征	解释
dteday	单车用户使用日期
season	单车用户使用季节 (春夏秋冬: 1-4)
temp	用户使用时温度 $(t - t_{\min}) / (t_{\max} - t_{\min})$
atemp	用户使用时体感温度 $(t - t_{\min}) / (t_{\max} - t_{\min})$
hum	当日能见度: 能见度/100
windspeed	风速: 风速/67
casual	该时段未注册用户使用人数
registered	该时段注册用户使用人数
cnt	该时段总用户数量
大量投放	是(>1500):1 否(<1500):0

dteday	season	temp	atemp	hum	windspeed	casual	registered	cnt	大量投放
2011/1/1	1	0.344167	0.363625	0.805833	0.160446	331	654	985	0
2011/1/2	1	0.363478	0.353739	0.696087	0.248539	131	670	801	0
2011/1/3	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349	0
2011/4/15	2	0.446667	0.441913	0.67125	0.226375	642	2484	3126	1
2011/4/16	2	0.434783	0.425492	0.888333	0.340808	121	674	795	0
2011/4/17	2	0.456667	0.445696	0.479583	0.303496	1558	2186	3744	1
2011/7/11	3	0.7625	0.729804	0.635833	0.282337	723	3363	4086	1
2011/7/12	3	0.794167	0.739275	0.559167	0.200254	662	3596	4258	1
2011/7/13	3	0.746667	0.689404	0.631667	0.146133	748	3594	4342	1
2011/9/27	4	0.636667	0.574525	0.885417	0.118171	477	3643	4120	1
2011/9/28	4	0.635	0.575158	0.84875	0.148629	480	3427	3907	1
2011/9/29	4	0.616667	0.574512	0.699167	0.172883	653	4186	4839	1

说明	无
意义	利用气象和其他数据可以预测在该时段和季节是否需要大量投放单车, 且可以通过使用用户在线总人数的数据来监测城市的流动性, 使得单车合理的投放
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset">http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset</a>



# 数据集-欧洲信用卡交易记录（诈骗交易分析）

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11~V25	V26	V27	V28	Amount	Class
- 1.35981	- 0.07278	2.53634 7	1.37815 5	- 0.33832	0.46238 8	0.23959 9	0.09869 8	0.36378 7	0.09079 4	.....	- 0.18911	0.13355 8	- 0.02105	149.62	0
1.19185 7	0.26615 1	0.16648	0.44815 4	0.06001 8	- 0.08236	-0.0788	0.08510 2	- 0.25543	-0.16697	.....	0.12589 5	- 0.00898	0.0147 24	2.69	0
- 1.35835	- 1.34016	1.77320 9	0.37978	-0.5032	1.80049 9	0.79146 1	0.24767 6	- 1.51465	0.20764 3	.....	-0.1391	- 0.05535	- 0.05975	378.66	0
- 0.96627	- 0.18523	1.79299 3	- 0.86329	- 0.01031	1.24720 3	0.23760 9	0.37743 6	- 1.38702	-0.05495	.....	- 0.22193	0.06272 3	0.0614 58	123.5	0
- 1.15823	0.87773 7	1.54871 8	0.40303 4	- 0.40719	0.09592 1	0.59294 1	- 0.27053	0.81773 9	0.75307 4	.....	0.50229 2	0.21942 2	0.2151 53	69.99	1

特征	解释	说明	数据集包含2013年9月由欧洲持卡人通过信用卡进行的交易。在284,807笔交易中有492个欺诈。它仅包含数字输入变量，这是PCA变换的结果。遗憾的是，由于机密性问题，我们无法获取有关数据的原始功能和更多背景信息。唯一未经PCA转换的特征是“数量”（即交易金额）。数据集高度不平衡，采样工作尤为重要。
V1~V28	V1, V2, ... V28 是使用PCA获得的主要feature		
Amount	交易金额		
Class	1表示欺诈性交易，否则为0	数据来源	<a href="https://www.kaggle.com/mlgulg/creditcardfraud">https://www.kaggle.com/mlgulg/creditcardfraud</a>



# 数据集-泰国的Facebook Live Sellers

特征	解释	Type	Published	Reactions	Comments	Shares	Like	Love	Wow	Haha	Sad	Angry
Type	发布广告类型											
Published	发布时间											
Reactions	观看者对于广告的反应数量	video	4/22/2018 6:00	529	512	262	432	92	3	1	1	0
Comments	观看者对于广告的评价数量	photo	4/21/2018 22:45	150	0	0	150	0	0	0	0	0
		video	4/21/2018 6:17	227	236	57	204	21	1	1	0	0
Shares	观看者转发数	photo	4/21/2018 2:29	111	0	0	111	0	0	0	0	0
Like	Reactions中Like的数量	photo	4/18/2018 3:22	213	0	0	204	9	0	0	0	0
		photo	4/18/2018 2:14	217	6	0	211	5	1	0	0	0
Love	Reactions中Love的数量	video	4/18/2018 0:24	503	614	72	418	70	10	2	0	3
Wow	Reactions中Wow的数量	video	4/17/2018 7:42	295	453	53	260	32	1	1	0	1
		photo	4/17/2018 3:33	203	1	0	198	5	0	0	0	0
Haha	Reactions中Haha的数量	说明	Reactions=Like+Love+Wow+Haha+Sad+Angry，标签为发布类型Type。									
Sad	Reactions中Sad的数量	意义	意义：分析了消费者参与的可变性，强调了Facebook Live的使用带来的变化。通过研究不同时间段不同敬业度指标的平均值，分析了季节因素。最后，确定统计上的异常点，并对其销售方法和活动进行进一步的定性分析。									
Angry	Reactions中Angry的数量	URL	<a href="https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand">https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand</a>									

# 数据集-台湾信用卡客户违约与支付预测

序号	给定信用额	性别	教育	婚姻状况	年龄	9月还款状况	8月还款状况	7月还款状况	9月账单金额	8月账单金额	7月账单金额	9月还款金额	8月还款金额	7月还款金额	Y
1	20000	2	2	1	24	2	2	-1	3913	3102	689	0	689	0	1
2	120000	2	2	2	26	-1	2	0	2682	1725	2682	0	1000	1000	1
3	90000	2	2	2	34	0	0	0	29239	14027	13559	1518	1500	1000	0
4	50000	2	2	1	37	0	0	0	46990	48233	49291	2000	2019	1200	0
5	50000	1	2	1	57	-1	0	-1	8617	5670	35835	2000	36681	10000	0

特征		解释	说明	本数据集研究台湾客户违约给付的数据及用户信息。
给定信用额		用户信用卡可使用额度		
性别		1-男 2-女		
教育		1-研究生 2-大学 3-高中 4-其他		
婚姻状况		1-已婚 2-单身 3-其他		
X月还款状况		-2-完全还款 -1-部分还款 1-延迟一月 2-延迟两月 9-延迟九个月及以上		
标签Y (label)		下月该用户是否违约 1=yes 2=no	URL	<a href="http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients">http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients</a>

# 数据集-匹兹堡桥梁

特征		解释	河流	地点 (1-52)	建造年代	目的用途	长度	.....	材质
所在河流		A M O	M	3	B	公路	缺乏数据		木头
建造地点		1—52	A	25	B	公路	1037		木头
建造年代		A:早于1800    B:1801~1900 C:晚于1900	A	39	B	沟渠	缺乏数据		木头
			A	29	B	公路	1000		木头
目的用途		公路  渠路  人行路  RR	M	23	B	公路	缺乏数据		木头
建造材质		木头 铁 钢	A	27	B	公路	990		木头
说明	该数据集包括匹兹堡所建造桥梁特征数据。		A	28	B	沟渠	1000		铁
			M	3	B	公路	1500		铁
			A	39	B	沟渠	缺乏数据		木头
			A	29	B	公路	1000		木头
			M	17	C	公路	2250		钢
			A	33	C	公路	980		钢
意义	给出某座桥梁的一些特征，可以计算出桥梁在某个时间段建造的概率，这对实地考察一些年代信息缺失的桥梁有很大参考价值，有助于考古或者当地文化局开展相关工作。								
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Pittsburg_h+Bridges">http://archive.ics.uci.edu/ml/datasets/Pittsburg_h+Bridges</a>								

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46

# 数据集-钻石数据-part2/2

特征	解释
carat	钻石克拉
cut	钻石切割质量
color	钻石成色
clarity	钻石净度
depth	钻石全深比
table	钻石台宽比
price	钻石价格（美元）
X	长度（mm）
Y	宽度（mm）
Z	深度（mm）

说明	台宽比：圆形钻石台面的最大宽度与钻石腰部平均直径之比，花式钻石则以台面宽度和钻石宽度之比。 全深比：圆钻的全深百分比是将深度除以最小与最大直径的平均数，花式钻石将深度除以宽度即得，而长度不在计算之内。
意义	可以建立一个钻石交易平台，卖方投入钻石，经平台专业技术鉴定确认克拉、切割质量、成色、净度等数据后，依靠此数据确切地报出钻石价格。一方面提高买方信任度，增加交易成功机会，扩大钻石交易市场；另一方面通过这个公开透明的数据平台催化钻石成为极浓缩、会增值、生利息的无国界自由货币。
URL	<a href="https://www.kaggle.com/shivam2503/diamonds">https://www.kaggle.com/shivam2503/diamonds</a>

帅器学习  
TechYoung Machine Learning Class

说明	“普查收入”数据集。根据人口普查数据预测年收入是否超过5万美元/年。
意义	意义：此数据集包含各国成年人的自身特征与年收入分类，揭示了其相关性。通过年龄、种族等特征判断年收入是否到达50k。
URL	<a href="https://archive.ics.uci.edu/ml/datasets/Adult">https://archive.ics.uci.edu/ml/datasets/Adult</a>

# 数据集-小麦籽粒形状特征与种类数据

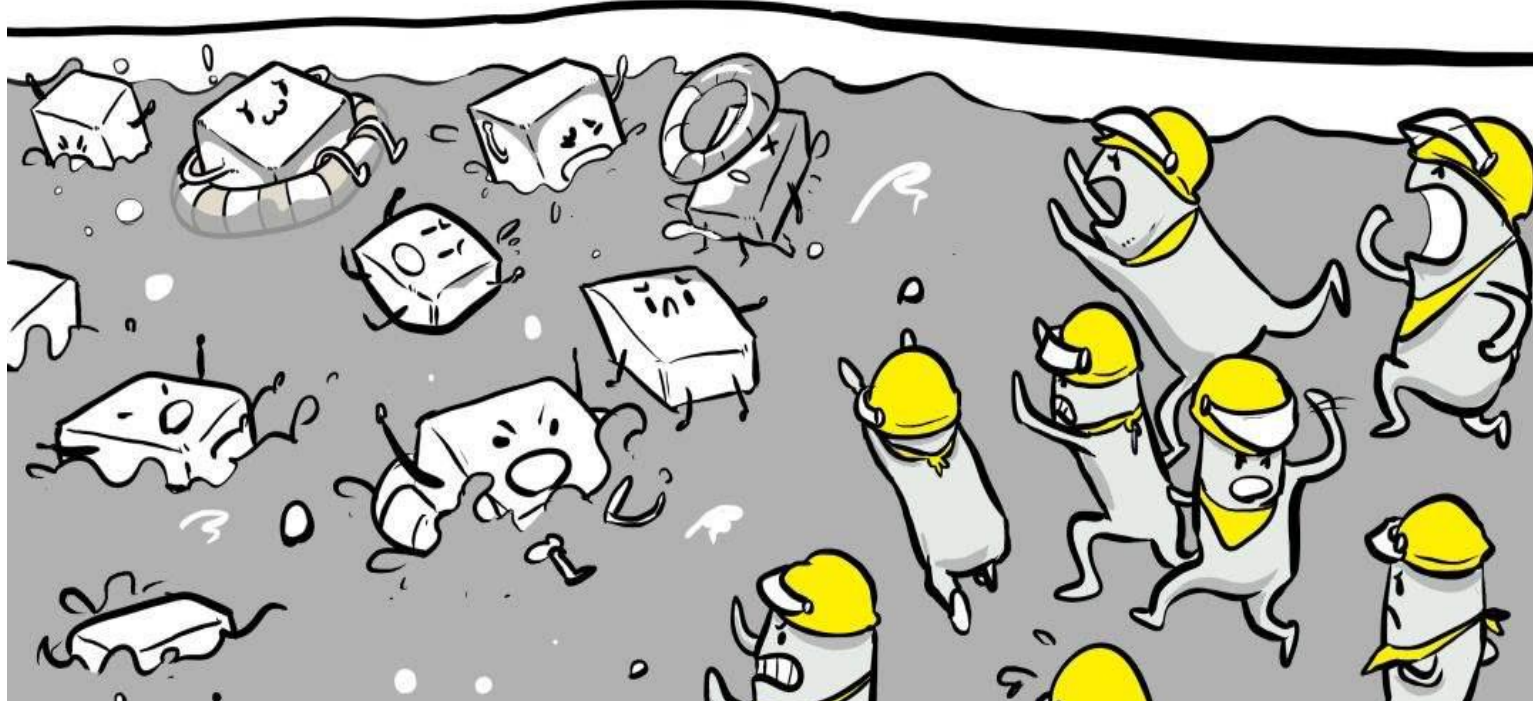
特征		解释	A	P	C	LK	WK	AC	LKG	K
A		面积	13.45	14.02	0.8604	5.516	3.065	3.531	5.097	1
P		周长	14.09	14.41	0.8529	5.717	3.186	3.92	5.299	1
C		紧凑度	15.05	14.68	0.8779	5.712	3.328	2.129	5.36	1
LK		籽粒长度	15.38	14.77	0.8857	5.662	3.419	1.999	5.222	1
WK		籽粒宽度	12.11	13.47	0.8392	5.159	3.032	1.502	4.519	1
AC		不对称系数	12.73	13.75	0.8458	5.412	2.882	3.533	5.067	1
LKG		核槽的长度	17.63	15.98	0.8673	6.191	3.561	4.076	6.06	2
K (种类)		1-Kama 2-Rosa 3-Canadian	16.84	15.67	0.8623	5.998	3.484	4.675	5.877	2
说明	此数据集包含三种不同小麦籽粒样本，并提取了小麦籽粒的特征数据，根据特征数据对样本进行分类。		17.26	15.73	0.8763	5.978	3.594	4.539	5.791	2
			19.11	16.26	0.9081	6.154	3.93	2.936	6.079	2
意义	此数据集揭示了小麦籽粒种类与籽粒各种形状特征的关联性，因此通过以上形状特征即可对小麦籽粒种类进行分类。该关联性可通过计算机技术加以运用，实现小麦籽粒自动分类，能够有效改善人工观察分类所产生的效率低、客观性差和准确率低等问题。		21.18	17.21	0.8989	6.573	4.033	5.78	6.231	2
			20.88	17.05	0.9031	6.45	4.032	5.016	6.321	2
URL	<a href="http://archive.ics.uci.edu/ml/datasets/seeds">http://archive.ics.uci.edu/ml/datasets/seeds</a>		18.36	16.52	0.8452	6.666	3.485	4.933	6.448	2
			13.07	13.92	0.848	5.472	2.994	5.304	5.395	3
			13.32	13.94	0.8613	5.541	3.073	7.035	5.44	3
			13.34	13.95	0.862	5.389	3.074	5.995	5.307	3
			12.22	13.32	0.8652	5.224	2.967	5.469	5.221	3
			11.21	13.13	0.8167	5.279	2.687	6.169	5.275	3
			11.43	13.13	0.8335	5.176	2.719	2.221	5.132	3
			12.49	13.46	0.8658	5.267	2.967	4.421	5.002	3



# 数据集-Pima印第安人糖尿病数据-part1/2

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0

# 数据集-Pima印第安人糖尿病数据-part2/2

特征	解释	说明	无
Pregnancies	怀孕次数	意义	数据集的目的是根据数据集中包含的某些诊断测量值，即怀孕次数、舒张压、血糖浓度、年龄等，诊断性地预测患者是否患有糖尿病。
Glucose	口服葡萄糖耐量试验中2小时的血糖浓度		
BloodPressure	舒张压（mm Hg）	URL	<a href="https://www.kaggle.com/uciml/pima-indians-diabetes-database">https://www.kaggle.com/uciml/pima-indians-diabetes-database</a>
SkinThickness	三头肌皮褶厚度（mm）		
Insulin	2-Hour serum insulin (mu U/ml)		
BMI	体重指数（重量单位：kg/（高度单位：m）^2）		
DiabetesPedigreeFunction	糖尿病谱系功能		
Age	年龄（岁）		
Outcome	768的类变量（0或1） 268为1，其他为0		

帅器学习  
TechYoung Machine Learning Class

说明	无
意义	通过对class, age, menopause, inv-nodes, node-caps和 irradiat的分析, 可以得到乳腺癌肿瘤大小的情况, 进而有助于即使治疗。
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Breast+Cancer">http://archive.ics.uci.edu/ml/datasets/Breast+Cancer</a>

# 数据集-学生自身特征与数学考试表现数据

特征	解释	gender	race	parental level of education	lunch	test preparation course	passing
Gender	性别						
Race	种族 A B C D E	female	group B	bachelor's degree	standard	none	64
parental level of education	父母教育水平	female	group C	some college	standard	completed	38
lunch	午饭standard free/reduced	female	group B	master's degree	standard	none	58
test preparation course	none/completed	male	group A	associate's degree	free	none	40
		male	group C	some college	standard	none	78
math score	数学成绩分数	female	group B	associate's degree	standard	none	62

说明	该数据集包含学生自身特征与入学数学考试是否及格，成绩由自美国的高中生
意义	通过此数据集可以了解家长背景、考试准备等因素对学生成绩的影响，并且可以用来对某个学生数学成绩是否及格进行判断和预测，从而采取相应措施侧重提高不能及格者的成绩。
URL	<a href="https://www.kaggle.com/spscientist/students-performance-in-exams">https://www.kaggle.com/spscientist/students-performance-in-exams</a>

# 数据集-波士顿犯罪数据

特征	解释	所属警区	案发地点	星期	小时	类别
所属警区	警区号码	D14	808	Sunday	13	Larceny
		C11	347	Tuesday	0	Vandalism
		D4	151	Monday	19	Towed
案发地点	数字表示案发准确地点	D4	272	Monday	21	Investigate Property
星期	Monday - Sunday	B3	421	Monday	21	Investigate Property
小时	0 - 24	C11	398	Monday	21	Motor Vehicle Accident Response
		B2	330	Monday	21	Auto Theft
犯罪类别	无	B2	584	Monday	20	Verbal Disputes
		C6	177	Monday	20	Robbery

说明

此数据集统计了波士60000+犯人的信息及其特征数据。

意义

给出波士顿某次犯罪的一些特征，可以推测出大概率的罪犯作案时间，有助于案件侦破和波士顿安防的调整。

URL

<https://www.kaggle.com/AnalyzeBoston/crimes-in-boston/downloads/crimes-in-boston.zip/2>

# 数据集-蘑菇形态及毒性特征

特征	解释
蘑菇盖形状	凸面型/凸面型/钟型
表面质地	光滑/鳞状
蘑菇盖颜色	棕色/黄色/白色
气味	辛辣/杏仁味/茴香味
族群数量	零散/许多
生长地区	市区/牧场/草
是否有毒	有/无

说明	此数据集进行的蘑菇记录来自奥杜邦协会北美蘑菇野外指南(1981)。包括蘑菇的物理特性及分类。
意义	通过此数据集可以了解北美蘑菇的毒性与以上形态特征之间的相关性，可以用于通过观察它的以上特征数据判断一个陌生蘑菇是否有毒。
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Mushroom">http://archive.ics.uci.edu/ml/datasets/Mushroom</a>

蘑菇盖形状	蘑菇盖表面	蘑菇盖颜色	气味	族群数量	生长地区	.....	有毒/无毒
凸面型	光滑	棕色	辛辣	零散	市区		有
凸面型	光滑	黄色	杏仁味	许多	草		无
钟型	光滑	白色	茴香味	许多	牧场		无
凸面型	鳞状	白色	辛辣	零散	市区		有
凸面型	光滑	灰色	无味	许多	草		无
凸面型	鳞状	黄色	杏仁味	许多	草		无
钟型	光滑	白色	杏仁味	许多	牧场		无
钟型	鳞状	白色	茴香味	零散	牧场		无
凸面型	鳞状	白色	辛辣	零散	草		有
钟型	光滑	黄色	杏仁味	零散	牧场		无
凸面型	鳞状	黄色	茴香味	许多	草		无
凸面型	鳞状	黄色	杏仁味	零散	牧场		无
钟型	光滑	黄色	杏仁味	零散	草		无
凸面型	鳞状	白色	辛辣	零散	市区		有
凸面型	纤维状	棕色	无味	许多	草		无
凹陷型	纤维状	灰色	无味	单个	市区		无



# 数据集-丹佛市犯罪数据

特征	解释
X/Y坐标	该地区在一定区域地图上的横纵坐标。
经度纬度	该地区在地球上的位置。
区号	犯罪行为所在区号码
辖区号	犯罪行为所在辖区号码
是否犯罪	1-犯罪 0-无犯罪

说明	该数据集包括丹佛市过去5年的犯罪行为，这些数据基于国家基于事件的报告系统(NIBRS)。
意义	可通过此数据集得到丹佛市地理位置与恶性事件的相关性，可以进行对某地区的犯罪概率的判断从而调整丹佛市安防分配以预防犯罪，降低犯罪率。
URL	<a href="https://www.kaggle.com/paultimothymooney/denver-crime-data">https://www.kaggle.com/paultimothymooney/denver-crime-data</a>

X坐标	Y坐标	经度	维度	区号	辖区号	是否犯罪
3193983	1707251	-104.81	39.77319	5	521	1
3201943	1711852	-104.781	39.78565	5	522	1
3152762	1667011	-104.957	39.66349	3	314	1
3157162	1681320	-104.941	39.7027	3	312	1
3153211	1686545	-104.955	39.71711	3	311	1
3151310	1696020	-104.962	39.74315	6	622	1
3133441	1692147	-105.026	39.73279	1	122	1
3145202	1688799	-104.984	39.72342	3	311	1
3142965	1693682	-104.992	39.73686	6	611	0
3136231	1701209	-105.015	39.75763	1	113	1
3161788	1672521	-104.925	39.67846	3	323	1
3141343	1678270	-104.998	39.69458	4	422	1
3156992	1676337	-104.942	39.68902	3	312	1
3134022	1700050	-105.023	39.75448	1	121	1
3168039	1680120	-104.903	39.69921	3	321	1
3136548	1703738	-105.014	39.76456	1	113	1



# 数据集-输血服务中心献血数据

特征	解释	Recency	Frequency	Monetary	Time	Label
Recency	距上次献血间隔月份	2	50	12500	98	1
Frequency	总共献血次数	0	13	3250	28	1
Time	据第一次献血间隔月份	1	24	6000	77	0
		4	11	2750	28	0
Monetary	总共的献血量 (CC)	4	9	2250	26	1
		9	9	2260	16	0
Label	2007年5月份是否献血 1—献血 0—未献血	0	2	500	4	0
		14	7	1750	26	0

说明	此数据集来源于台湾新竹的服务中心献血者的献血信息记录。
意义	通过此数据可以根据当下献血情况预测未来某一时间点的血量贮存状况，从而预先采取若干措施，避免出现血量供应不足等情况。还可以根据当前献血人信息预测一年后该人是否献血。
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center">http://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center</a>

# 数据集-动物分类 (Part1)

animal name:	hair	feather	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	cat size	type
aardvark	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	1
antelope	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	1
bass	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	4
bear	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	1
boar	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	1
buffalo	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	1
calf	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1	1
carp	0	0	1	0	0	1	0	1	1	0	0	1	0	1	1	0	4
catfish	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	4
cavy	1	0	0	1	0	0	0	1	1	1	0	0	4	0	1	0	1
cheetah	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	1
chicken	0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0	2

说明	此数据集包含多种动物的17个形态特征属性。他们分属于7个不同的class。
意义	根据动物的特征对其进行分类。
URL	<a href="https://archive.ics.uci.edu/ml/datasets/Zoo">https://archive.ics.uci.edu/ml/datasets/Zoo</a>

属性特征:	解释:	属性特征:	解释:
hair	是否有毛发	feather	是否有羽毛
predator	是否为捕食者	eggs	有否卵生
toothed	是否有牙齿	milk	是否产奶
backbone	是否有脊椎	airborne	是否飞行
breathes	是否在陆地呼吸	aquatic	是否水栖
venomous	是否分泌毒液	fins	是否有鳍
legs	腿的数量 (2, 4, 6, 8)	tail	是否有尾
domestic	是否驯化	catsize	是否猫科
type	种类 1-7 1 -- (41) aardvark, antelope, bear, boar, buffalo, calf, cavy, cheetah, deer, dolphin, elephant, fruitbat, giraffe, girl, goat, gorilla, hamster, hare, leopard, lion, lynx, mink, mole, mongoose, opossum, oryx, platypus, polecat, pony, porpoise, puma, pussycat, raccoon, reindeer, seal, sealion, squirrel, vampire, vole, wallaby,wolf 2 -- (20) chicken, crow, dove, duck, flamingo, gull, hawk, kiwi, lark, ostrich, parakeet, penguin, pheasant, rhea, skimmer, skua, sparrow, swan, vulture, wren 3 -- (5) pitviper, seasnake, slowworm, tortoise, tuatara 4 -- (13) bass, carp, catfish, chub, dogfish, haddock, herring, pike, piranha, seahorse, sole, stingray, tuna 5 -- (4) frog, frog, newt, toad 6 -- (8) flea, gnat, honeybee, housefly, ladybird, moth, termite, wasp 7 -- (10) clam, crab, crayfish, lobster, octopus, scorpion, seawasp, slug, starfish, worm		

帅器学习  
TechYoung Machine Learning Class

特征	解释	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphatase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Data set
Age	年龄											
Gender	性别											
Total_Bilirubin	总胆红素	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
		62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
Direct_Bilirubin	直接胆红素	62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
Alkaline_Phosphatase	碱性磷酸酶	58	Male	1	0.4	182	14	20	6.8	3.4	1	1
		72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
Alamine_Aminotransferase	阿拉明转氨酶	17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	2
		55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
Aspartate_Aminotransferase	天冬氨酸盐转氨酶	57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8	1
Total_Protiens	总蛋白质	说明	该数据集包含来自印度安德拉邦东北地区的416份肝脏病人记录和167份非肝脏病人记录。“Dataset” 列是一个类标签，用于将组划分为肝病患者(肝病)或非肝病患者(无疾病)。该数据集包含441例男性患者记录和142例女性患者记录。shi'ya（年龄超过89岁的病人被列为90岁）									
Albumin	白蛋白											
Albumin_and_Globulin_Ratio	白蛋白和球蛋白的比值	意义	由于过量饮酒、吸入有害气体、食用受污染的食品、腌菜和药物，肝病患者不断增加。该数据集用于评估预测算，食药改善身体状况避免患病，以减轻医生的负担。									
Dataset	病人是否换肺病	URL	<a href="https://www.kaggle.com/uciml/indian-liver-patient-records/downloads/indian-liver-patient-records.zip/1">https://www.kaggle.com/uciml/indian-liver-patient-records/downloads/indian-liver-patient-records.zip/1</a>									

帅器学习  
TechYoung Machine Learning Class

特征	解释	ambient	coolant	u_d	u_q	motor_speed	torque	i_d	i_q	pm
ambient	定子热传感器测定的周围温度	-0.752	-1.118	0.328	-1.298	-1.222	-0.250	1.030	-0.246	-2.522
		-0.771	-1.117	0.330	-1.298	-1.222	-0.249	1.030	-0.246	-2.522
coolant	水冷电机出水口温度	-0.783	-1.117	0.333	-1.302	-1.222	-0.249	1.029	-0.246	-2.523
		-0.781	-1.117	0.334	-1.302	-1.222	-0.249	1.033	-0.247	-2.522
		-0.774	-1.117	0.335	-1.303	-1.222	-0.249	1.032	-0.247	-2.522
u_d	D组件电压	-0.763	-1.117	0.335	-1.303	-1.222	-0.248	1.031	-0.246	-2.522
		-0.749	-1.116	0.335	-1.302	-1.222	-0.248	1.030	-0.246	-2.523
u_q	Q组件电压	-0.738	-1.114	0.336	-1.305	-1.222	-0.248	1.030	-0.246	-2.523
		-0.731	-1.112	0.335	-1.304	-1.222	-0.248	1.030	-0.246	-2.523
		-0.727	-1.109	0.336	-1.306	-1.222	-0.248	1.030	-0.246	-2.523
motor_speed	电机速度	-0.752	-1.118	0.328	-1.298	-1.222	-0.250	1.030	-0.246	-2.522
		-0.771	-1.117	0.330	-1.298	-1.222	-0.249	1.030	-0.246	-2.522
i_d	D组件电流	说明	A comprehensive csv files containing all measurement sessions and features. Each row represents one snapshot of sensor data at a certain time step.							
i_q	Q组件电流									
torque	电流引发的扭矩	意义	利用温度传感器检测电动机各部分的温度、电压电流等数据，监控电动机核心温度高低，以保证运行安全。 This dataset is mildly anonymized.							
pm	红外热成像单元测定的永磁体表面温度（视为转子温度）									
		URL	https://www.kaggle.com/wkirgsn/electric-motor-temperature							

Age	workclass	fnlwgt	education	marital	occupation	relationship	race	sex	Capital_gain	Capital_loss	hr_per_week	country	income
39	State-gov	77516	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

特征	解释
Age	年龄
workclass	工作类别
fnlwgt	人口普查ID
education	受教育程度
marital	婚姻情况
occupation	职业
relationship	家庭身份
race	种族
sex	性别
capital_gain	资本收益
capital_loss	资本损失
hr_per_week	每周工作时长
country	国籍（初始）
income	收入
标签：收入大于50K或小于50K	

说明	此数据集包含32561不同身份的人的样本，提取人的特征数据包括：
意义	此数据集揭示了一个人的年收入与其各种社会和个人特征的关联性，因此通过以上特征可以对一个人的年收入是否超过50K美金进行预测。
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Adult">http://archive.ics.uci.edu/ml/datasets/Adult</a>





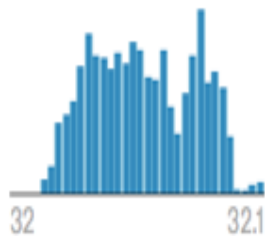
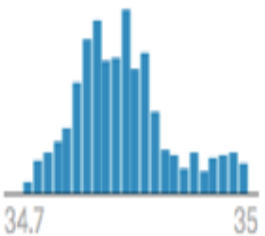
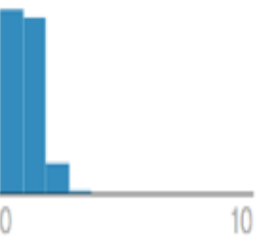
**说明** 此数据详细显示了共享车辆的位置，其中样本特征包含：

- 1.停放的时间
- 2.所在的纬度
- 3.所在的经度
- 4.当地停放的车辆数目

样本的标签： 车辆列表

**意义** 通过分析样本数据可以得知详细的可用共享车辆的位置，给人们出行带来了很大的方便，目的在于有助于减少私家车的购买以及人们对私家车的依赖。

**URL** <http://www.kaggle.com/doit-intl/autotel-shared-car-locations/downloads/autotel-shared-car-locations.zip/1>

	timestamp	latitude	longitude	# total_cars	carsList
	18998 unique values				<div><div></div>46%</div> <div>[120]0%</div> <div>Other (8848)53%</div>
1	2019-01-10 11:45:55.070781 UTC	32.09995	34.78794	1	[182]
2	2019-01-10 11:45:55.070781 UTC	32.06567	34.79612	1	[268]
3	2019-01-10 11:45:55.070781 UTC	32.06465	34.80322	1	[106]
4	2019-01-10 11:45:55.070781 UTC	32.05978	34.81034	1	[180]
5	2019-01-10 11:45:55.070781 UTC	32.05133	34.75089	1	[16]
6	2019-01-10 11:45:55.070781 UTC	32.04223	34.7742	1	[72]
7	2019-01-10 11:45:55.070781 UTC	32.04156	34.77128	1	[160]
8	2019-01-10 11:45:55.070781 UTC	32.12373	34.81346	1	[210]

# 数据集-LOL地图资源占有情况对比赛结果影响

特征	解释	Game Duration	seasonId	First RiftHerald	firstBlood	firstTower	firstInhibitor	firstBaron	firstDragon	winner
Game Duration	游戏进行时长	1949	9	2	2	1	1	1	1	1
		1851	9	1	1	1	1	0	1	1
seasonId	游戏赛季	1493	9	0	2	1	1	1	2	1
		1758	9	0	1	1	1	1	1	1
winner	本局比赛中游戏胜利方	2094	9	0	2	1	1	1	1	1
		2059	9	0	2	2	1	1	2	1
firstBlood	拿下本局比赛中第一滴血的队伍	1993	9	1	1	2	1	1	1	1
		1334	9	1	1	1	0	0	2	1
firstTower	摧毁本局比赛中第一座防御塔的队伍	1387	9	2	2	2	2	0	2	2
		2681	9	0	2	2	2	2	2	2
firstInhibitor	本局比赛中第一次使用抑制剂的队伍	1391	9	2	2	2	2	2	2	2
firstBaron	本局比赛中完成对男爵第一次击杀的队伍	说明 此数据中紫色方=1，蓝色方=2，未有队伍进行该项目=0								
firstDragon	本局比赛中完成对远古龙第一次击杀的队伍	意义 通过此数据集的数据，可以得出S9赛季在排位赛中开局时地图资源（如：第一滴血，第一座防御塔，第一次击杀远古龙等）被双方队伍占有的情况对比赛结果的影响，可用来在比赛进行中对比赛结果实时预测，以及完成战队的战术运用规划。								
firstRiftHerald	本局比赛中完成对峡谷先锋第一次击杀的队伍	URL <a href="https://www.kaggle.com/datasnaek/league-of-legends">https://www.kaggle.com/datasnaek/league-of-legends</a>								

特征	解释	Isoperimetric Factor	Smoothness	Eccentricity	Aspect ratio	Elongation	种类
		0.7893	0.0050	0.7839	1.6719	0.4057	Quercus suber
		0.6575	0.0080	0.8924	2.3298	0.5697	Salix atrocinera
Isoperimetric Factor	等周因子	0.6484	0.0222	0.6311	1.1271	0.3481	Populus nigra
		0.7335	0.0092	0.5870	1.3131	0.2822	Alnus sp.
Smoothness	顺滑度	0.3456	0.0185	0.8847	1.9502	0.6420	Quercus robur
		0.2629	0.0054	0.5402	1.1417	0.6351	Crataegus monogyna
Eccentricity	偏心率	0.5251	0.0109	0.8470	1.7847	0.5235	Ilex aquifolium
Aspect ratio	纵横比	说明	无				
		意义	此数据集揭示了不同叶子种类的各类形状和性状的数据，因此可通过以上形状和性状的特征值对叶子种类进行分类。				
Elongation	伸长率	URL	<a href="http://archive.ics.uci.edu/ml/datasets/Leaf">http://archive.ics.uci.edu/ml/datasets/Leaf</a>				

# 数据集-网络文章受欢迎程度（所占份额）

特征	解释	文章字数	特殊符号	KW	发布日	主观	客观	份额
文章字数	文章的内容总字数	219	0.663	5	Mon	0.521	0.092	593
		255	0.604	4	Mon	0.341	0.148	711
特殊符号	文章中特殊符号的字数	211	0.575	6	Mon	0.702	0.323	1500
		531	0.503	7	Mon	0.429	0.100	1200
KW	metadata中关键词数量	288	0.589	6	Tue	0.449	0.155	755
		414	0.606	6	Tue	0.343	0.010	468
发布日	文章发布的星期日	134	0.702	10	Tue	0.406	0.135	4600
主观	文章的主观程度	281	0.610	4	Tue	0.434	0.899	1300
客观	文章的客观程度							
份额	文章在网络市场中的受欢迎程度	说明	无					
		意义	该数据集总结了Mashable在两年内发表的关于文章的异构特征集。目标是预测社交网络中文章的受欢迎程度					
		URL	<a href="http://archive.ics.uci.edu/ml/datasets/Online+N">http://archive.ics.uci.edu/ml/datasets/Online+N</a>					

特征	解释
social smoker	是否抽烟 (1有 0无)
social drinker	是否喝酒 (1有 0无)
Distance	通勤距离 (公里)
Age	年龄
Son	子女数量
Reason for absence drinker	有理由请假数量
Expense	通勤开销
education	教育程度 (1高中 2大学 3研究生 4博士生)
pet	宠物数量
time	旷工时长 (小时)

social drinker	social smoker	Distance	Age	Son	pet	Reason for absence	Expense	education	time
1	0	36	33	2	1	26	289	1	4
1	0	13	50	1	0	0	118	1	0
1	0	51	38	0	0	23	179	1	2
1	1	5	39	2	0	7	279	1	4
1	0	36	33	2	1	23	289	1	2
1	0	51	38	0	0	23	179	1	2
1	0	52	28	1	4	22	361	1	8
1	0	50	36	4	0	23	260	1	4
1	0	12	34	2	0	19	155	1	40
0	0	11	37	1	1	22	235	3	8
1	0	50	36	4	0	1	260	1	8
1	0	50	36	4	0	1	260	1	8
说明	此数据集包含工人旷工的时长以及其个人相关特征数据。								
意义	意义：通过对工人日常生活与工作相关数据的分析，来预测工人的旷工时长。通过相应措施来减少工人旷工时间。								
URL	<a href="http://archive.ics.uci.edu/ml/machine-learning-databases/00445/">http://archive.ics.uci.edu/ml/machine-learning-databases/00445/</a>								

# 数据集-根据体征判断是否患有心律不齐

年龄	以年为单位
性别	0=男性; 1=女性
身高	以厘米为单位
体重	以千克为单位
QRS时限	心电图心室除极的时间, 平均QRS时限
标签	是否患有心律不齐 0=无; 1=有 以上为被测人的心律不齐患病相关特征参数
意义	区分有无心律失常, 利用年龄、性别、QRS时限等数据可以推测该对象是否患有心率异常这一种病。通过以上一些便于测得的数据判断可减少评估成本, 加快评估效率。
UCL	<a href="http://archive.ics.uci.edu/ml/datasets/Arrhythmia">http://archive.ics.uci.edu/ml/datasets/Arrhythmia</a>

年龄	性别	身高	体重	QRS 时限	是否患有 心律不齐
54	1	165	64	81	1
56	0	172	95	138	0
54	0	175	94	100	1
55	0	190	80	88	0
75	0	169	51	100	0
13	1	160	52	77	1
40	1	162	54	78	1
49	0	168	56	84	0
44	1	167	67	89	1
50	0	170	72	102	0
62	1	165	86	77	1
45	1	172	58	78	0

# 数据集-某便利店分店在不同地区销售商品种类数据

特征	解释
Fresh	生鲜
Milk	乳制品
Beverage	饮料
Frozen	冻鲜
Washing Materials	洗化用品
Delicatessen	熟食
Region	地区

Fresh	Milk	Beverage	Frozen	Washing Materials	Delicatessen	Region
12669	9656	7561	214	2674	1338	北京
7057	9810	9568	1762	3293	1776	北京
22615	5410	7198	3915	1777	5185	北京
5963	3648	6192	425	1716	750	北京
30624	7209	4897	18711	763	2876	济南
1537	3748	5838	1859	3381	806	济南
5550	12729	16767	864	12420	797	济南
29635	2335	8280	3046	371	117	上海
2137	3737	19172	1274	17120	142	上海
9784	925	2405	4447	183	297	上海
10617	1795	7647	1483	857	1233	上海
1479	14982	11924	662	3891	3508	上海
说明	数据集中的数据代表顾客在不同商品类别上以货币单位表示的年度支出。					
意义	此数据集分析记录某便利店在不同城市主要销售商品的种类，经营者可分析不同地区商品需求的偏好，有针对性地进行供货。					
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Wholesale+customers">http://archive.ics.uci.edu/ml/datasets/Wholesale+customers</a>					



# 数据集——巴西导游公司员工旷工程度

特征	解释	Transportation expense	Distance from Residence to Work	Service time	Age	Education	Social drinker	Social smoker	Degree
Transportation expense	员工上班所需的交通费用								
Distance from Residence to Work	公司与员工住所的距离	118	13	18	50	1	1	0	1
		260	50	11	36	1	1	0	1
		225	26	9	28	1	0	0	1
Service time	员工的工作时长	289	36	13	33	1	1	0	1
Age	员工的年龄	118	13	18	50	1	1	0	1
		369	17	12	31	1	1	0	1
Education	员工受教育的程度 (高中-1, 本科-2, 研究生-3, 博士-4)	118	13	18	50	1	1	0	1
		118	10	10	37	1	0	0	1
		235	29	12	48	1	0	1	1
Social drinker	员工是否喝酒 (喝酒-1, 不喝酒-2)	279	5	14	39	1	1	1	1
		330	16	4	28	2	0	0	1
Social smoker	员工是否吸烟 (吸烟-1, 不吸烟-2)	378	49	11	36	1	0	1	1
Degree	员工旷工等级 (旷工0h-1, 旷工1~5h-2, 旷工6~10h-3, 旷工11~100h-4, 旷工>100h-5)								
		说明	此数据集包含巴西导游公司员工旷工时长等级及相关特征数据。						
		意义	利用员工的年龄、工作时长、是否抽烟、是否喝酒、上班距离、所耗交通费用和受教育程度可以预测员工的旷工程度，这样可以在企业进行招聘时对员工的留用做出更加准确的判断。						
		URL	<a href="http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work">http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work</a>						

# 数据集-美国驻北京大使馆附近PM2.5监测

特征	解释
No	数据编号
year	检测年份
month	检测月份
day	检测日期
hour	监测发生的整点时间
pm2.5	PM2.5浓度值 (ug/m <sup>3</sup> )
DEWP	固定气压之下, 空气中所含的气态水达到饱和而凝结成液态水需要降至的温度( $\hat{a}, f$ )
TEMP	检测时周围气温( $\hat{a}, f$ )
PRES	监测时空气压力(hPa)
cbwd	监测时组合风向方位
lws	风速 (m/s)
ls	降雪累计小时数
lr	降雨累计小时数
rank	污染等级 (0: 无, 1: 轻度, 2: 中度, 3: 重度)

No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	lws	ls	lr	rank
10	2010	1	1	9	0	-20	-8	1017	NW	27.28	0	0	0
20	2010	1	1	19	0	-17	-4	1017	NW	1.79	0	0	0
30	2010	1	2	5	109	-7	-6	1022	SE	7.14	3	0	1
40	2010	1	2	15	154	-9	-5	1025	SE	35.75	0	0	1
50	2010	1	3	1	63	-8	-6	1026	SE	61.69	5	0	1
60	2010	1	3	11	98	-11	-9	1022	SE	102.8	16	0	1
70	2010	1	3	21	72	-11	-10	1023	NW	7.15	27	0	1
80	2010	1	4	7	29	-21	-13	1027	NW	73.75	0	0	1
90	2010	1	4	17	30	-20	-11	1032	NW	159.56	0	0	1
100	2010	1	5	3	25	-27	-18	1035	NW	213.65	0	0	1

说明	此数据集包含美国驻北京大使馆附近的空气检测数据以及污染等级评测。
意义	利用气象数据可以预测美国驻北京大使馆周边PM2.5数据, 且可以通过以上特征指标评估空气污染程度。由此可以有针对性和市民生活出行进行有效预警, 对相关环境部门有效应对有参考作用。
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data">http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data</a>

	name	release_date	english	developer	publisher	platforms	require_d_age	categories	genres	steamspy_tags	achievements	positive_ratings	negative_ratings	average_playtime	median_playtime	owners	price
10	Counter-Strike	2000/11/1	1	Valve	Valve	windows;mac;linux	0	Multi-player;Online Multi-Player;Local Multi-Player;Valve Anti-Cheat enabled	Action	Action;FPS;Multiplayer	0	124534	3339	17612	317	10000000-20000000	7.19
20	Team Fortress Classic	1999/4/1	1	Valve	Valve	windows;mac;linux	0	Multi-player;Online Multi-Player;Local Multi-Player;Valve Anti-Cheat enabled	Action	Action;FPS;Multiplayer	0	3318	633	277	62	5000000-10000000	3.99
30	Day of Defeat	2003/5/1	1	Valve	Valve	windows;mac;linux	0	Multi-player;Valve Anti-Cheat enabled	Action	FPS;World War II;Multiplayer	0	3416	398	187	34	5000000-10000000	3.99
40	Deathmatch Classic	2001/6/1	1	Valve	Valve	windows;mac;linux	0	Multi-player;Online Multi-Player;Local Multi-Player;Valve Anti-Cheat enabled	Action	Action;FPS;Multiplayer	0	1273	267	258	184	5000000-10000000	3.99
50	Half-Life: Opposing Force	1999/11/1	1	Gearbox Software	Valve	windows;mac;linux	0	Single-player;Multi-player;Valve Anti-Cheat enabled	Action	FPS;Action;Sci-fi	0	5250	288	624	415	5000000-10000000	3.99

特征	解释
appid	对应id
Name	游戏名字
Release date	发行日期
English	是否支持英语
Developer	开发商
publisher	发行商
Platforms	支持平台
Required age	年龄限制
Categories	游戏种类；例如单人或者联机
Genres	用分号分隔的游戏类型，例如运动或者冒险
Steamspy tags	由steamspy推荐的标签
achievements	游戏内部成就数量
Positive ratings	积极评价数量
Negative ratings	负面评价数量
Average playtime	平均游玩时间
Median playtime	游玩时间中位数
Owners	估计的游戏拥有者
price	游戏价格

说明	此数据集包含STEAM多种游戏的特征属性。
意义	应用该数据集可通过对STEAM的发行游戏进行市场研究，例如 游戏的流行类型进行研究，或者可以发现异常情况以便深入市场研究，比如对负面评价较多但是估计拥有者依旧呈现上涨趋势的游戏进行筛选，而后再另外调查是否有其他因素（如广告或者游戏改良等原因）导致异常发生。
URL	<a href="https://www.kaggle.com/nikdavis/steam-store-games">https://www.kaggle.com/nikdavis/steam-store-games</a>



# 数据集-共享单车数据

特征	解释
dteday	单车用户使用日期
season	单车用户使用季节 (春夏秋冬: 1-4)
temp	用户使用时温度 (t-t_min)/(t_max-t_min)
atemp	用户使用时体感温度 (t-t_min)/(t_max-t_min)
hum	当日能见度: 能见度/100
windspeed	风速: 风速/67
casual	该时段未注册用户使用人数
registered	该时段注册用户使用人数
cnt	该时段总用户数量
大量投放	是(>1500):1 否(<1500):0

dteday	season	temp	atemp	hum	windspeed	casual	registered	cnt	大量投放
2011/1/1	1	0.344167	0.363625	0.805833	0.160446	331	654	985	0
2011/1/2	1	0.363478	0.353739	0.696087	0.248539	131	670	801	0
2011/1/3	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349	0
2011/4/15	2	0.446667	0.441913	0.67125	0.226375	642	2484	3126	1
2011/4/16	2	0.430833	0.425492	0.888333	0.340808	121	674	795	0
2011/4/17	2	0.456667	0.445696	0.479583	0.303496	1558	2186	3744	1
2011/7/11	3	0.7625	0.729804	0.635833	0.282337	723	3363	4086	1
2011/7/12	3	0.794167	0.739275	0.559167	0.200254	662	3596	4258	1
2011/7/13	3	0.746667	0.689404	0.631667	0.146133	748	3594	4342	1
2011/9/27	4	0.636667	0.574525	0.885417	0.118171	477	3643	4120	1
2011/9/28	4	0.635	0.575158	0.84875	0.148629	480	3427	3907	1
2011/9/29	4	0.616667	0.574512	0.699167	0.172883	653	4186	4839	1
说明	此数据集为共享单车的流动性数据特征。								
意义	利用气象和其他数据可以预测在该时段和季节是否需要大量投放单车, 且可以通过使用用户在线总人数的数据来监测城市的流动性, 使得单车合理的投放								
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset">http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset</a>								



# 数据集—IBM员工绩效评价数据

特征	解释
性别	F-女 M-男
出差频率	1-Frequently 2-Rarely
教育程度	1-Below College 2-College 3-Above College
职位水平	1'Very Low'' 2 'Low' 3'Medium' 4 'High' 5'Very High'
同事关系满意度	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
标签	解释
工作表现	'Low'、'Good'、 'Excellent'、 'Outstanding'

性别	出差频率	所在部门	教育程度	职位水平	同事关系满意度	工作表现
F	1	Research & Development	1	2	4	Outstanding
F	1	sale	2	3	4	Excellent
F	2	Research & Development	1	1	3	Excellent
F	2	sale	3	2	3	Excellent
M	1	Research & Development	3	2	4	Outstanding
M	1	sale	2	3	3	Excellent
M	2	Research & Development	2	4	4	Excellent
M	2	sale	3	5	4	Excellent

说明	此数据揭示IBM员工表现情况与员工自身背景的关系
意义	可以作为公司对员工未来表现预测的参考数据，通过对员工的个人背景及工作状况的调查，推断出该员工在公司的总体表现情况
URL	<a href="https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset">https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset</a>

# 数据集-共享汽车使用量

特征		解释	时间	纬度	经度	共享汽车被使用的数量
时间		共享汽车被使用的开始时间，为UTC时间，范围：2019年1月10日00：00至23：59	11:45:55.070781 UTC	32.110	34.841	1
			11:45:55.070781 UTC	32.056	34.784	0
			10:12:12.721107 UTC	32.037	34.763	4
			11:50:04.533343 UTC	32.093	34.785	0
纬度		范围：32-32.1	11:50:04.533343 UTC	32.103	34.806	4
经度		范围：34.7-35	11:47:59.684216 UTC	32.091	34.790	2
			09:38:52.434456 UTC	32.084	34.775	0
说明		本数据集中的特征数据为不同时间段不同地点的共享汽车使用的数量。	12:15:03.425963 UTC	32.085	34.797	3
			12:42:10.963106 UTC	32.120	34.830	1
			12:27:33.105647 UTC	32.043	34.781	1
意义		该数据可用于帮助共享汽车公司合理分配汽车的投放位置并根据时间和地点做出具体规划。	11:58:24.326574 UTC	32.116	34.788	0
			09:45:07.265499 UTC	32.121	34.833	1
			10:30:56.837264 UTC	32.063	34.776	1
URL		<a href="https://www.kaggle.com/doi-intl/autotel-shared-car-locations">https://www.kaggle.com/doi-intl/autotel-shared-car-locations</a>	09:55:31.692807 UTC	32.067	34.782	2
			12:50:30.672668 UTC	32.087	34.780	0



Temp	Rain_1h	Clouds_all	Weather_main	Date_time	Time_index	Traffic_volume	Volume_index
287.25	0	1	5	2013/5/13 19:00	19	3205	4
284.73	0	1	5	2013/5/13 21:00	21	2568	3
281.11	0	1	5	2013/5/13 23:00	23	958	1
292.67	0.51	90	6	2018/8/24 5:00	6	2950	3
291.89	0.98	90	3	2018/8/24 6:00	7	5259	6
292.04	0	90	4	2013/5/18 1:00	1	727	1
291.39	0	90	3	2013/5/18 2:00	2	501	1
286.78	0	90	6	2013/5/19 0:00	0	1475	2
290.05	3.65	100	6	2013/6/13 12:00	12	5126	6
291.52	0	90	8	2013/5/20 9:00	9	5114	6
294.64	0.51	36	3	2013/6/12 12:00	12	4989	5
290.41	8.64	92	3	2013/6/13 9:00	9	5144	6

数据来源：

# 数据集-Metro Interstate Traffic Volume数据集 (Part2)

特征	解释	Weather_main	代码
Temp	平均温度 - 以开尔文为单位	Clouds	1
Rain_1h	降雨量_h - 以小时为单位的降雨量	Thunderstorm	2
Clouds_all	云量百分比	Rain	3
Weather_main	当前天气简短描述	Drizzle	4
Date_time	收集的数据的时间	Clear	5
Time_index	Date_time对应的指数 e.g.2012-10-11 24: 00 == 24	Mist	6
Traffic_volume	每小时西行交通量	Fog	7
Volume_index	每小时西行交通量千位数代表的指数	Haze	8
		Snow	9
		Smoke	10

说明	本数据集中的特征数据为不同时间段不同地点的共享汽车使用的数量。标签： <b>Traffic_volume(回归) / Volume_index (聚类)</b>
意义	数据集中的数据为大致位于明尼阿波利斯和明尼苏达州圣保罗之间的94号州际公路不同天气特征下的每小时西行交通量，包括下述7项天气特征、1项日期特征和每小时西行交通量，目的为在天气特征、日期特征与每小时西行交通量建立联系，预测每小时天气功能和假期包括对交通量的影响。
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume">http://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume</a>

# 数据集-日本旅社数据库

特征	解释
Hotel.name	旅馆名称
City	宿舍所在城市名称
Price.from	一晚住宿的最低价格
Distance	距市中心距离
Summary.score	摘要评分
Atmosphere	空气评分
Cleanliness	清洁度评分
Facilities	设备评分
Location.y	位置评分
Security	安全评级得分
Staff	员工评分
Valueformoney	物有所值评分
Rating	总体评价 1-Very Good 2-Superb 3-Fabulous

City	Price .from	Distance	Sum mary. score	Atm osp here	Clea nline ss	Fac iliti es	Loca tion. y	Sec uri ty	Staff	Valu efor mon ey	R at ing
Osaka	3300	2.9km from city centre	9.2	8.9	9.4	9.3	8.9	9	9.4	9.4	2
Tokyo	3600	7.8km from city centre	8.7	8	7	9	8	10	10	9	3
Tokyo	2600	8.7km from city centre	7.4	8	7.5	7.5	7.5	7	8	6.5	1
Tokyo	1500	10.5km from city centre	9.4	9.5	9.5	9	9	9.5	10	9.5	2
Tokyo	2200	3km from city centre	7.7	6.7	7.2	6.8	8.5	7.8	8.5	8.1	1
Osaka	1600	9.7km from city centre	9.2	9.5	9.1	8.7	8.8	8.9	9.8	9.5	2
Osaka	2700	1.9km from city centre	9	8.7	9.2	8.9	8.8	9.1	9.2	9	2
Osaka	2600	4km from city centre	8.6	7.5	9	8	8.5	10	8	9	3
Hiroshima	2600	1.9km from city centre	9.4	9	9.5	9.2	9.4	9.5	9.7	9.3	2
Fukuoka-City	4300	0.3km from city centre	8.6	8	10	6	10	8	10	8	3

说明	次数据集包含日本旅社的特征及评分以及总体评价。
意义	通过应用该数据集可以预测顾客对旅馆整体满意度在各个特征下所占比重并对其按照重要性排序以达到吸引更多顾客的目的来创造更多经济效益.
URL	<a href="https://www.kaggle.com/koki25ando/hostel-world-dataset">https://www.kaggle.com/koki25ando/hostel-world-dataset</a>

# 数据集-用户知识水平模型数据库

特征	解释
STG	目标对象学习时间的长短
SCG	目标对象重复次数的程度
STR	用户对与目标对象相关的事物的学习时间
LPR	用户对与目标对象相关的事物的考试成绩
PEG	用户的目标对象的考试成绩
标签	解释
UNS	用户知识水平

说明	以上是用户知识水平和目标对象学习及相关对象时间、重复次数、考试成绩的数据库
意义	通过收集特征值的数据、建立适当模型，可以分析预测用户的知识水平，以此提供个性化的推送、产品、服务等
URL	<a href="https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling">https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling</a>

STG	SCG	STR	LPR	PEG	UNS
0	0	0	0	0	very_low
0.08	0.08	0.1	0.24	0.9	High
0.06	0.06	0.05	0.25	0.33	Low
0.1	0.1	0.15	0.65	0.3	Middle
0.08	0.08	0.08	0.98	0.24	Low
0.09	0.15	0.4	0.1	0.66	Middle
0.1	0.1	0.43	0.29	0.56	Middle
0.15	0.02	0.34	0.4	0.01	very_low
0.2	0.14	0.35	0.72	0.25	Low
0	0	0.5	0.2	0.85	High
0.18	0.18	0.55	0.3	0.81	High
0.06	0.06	0.51	0.41	0.3	Low
0.1	0.1	0.52	0.78	0.34	Middle
0.1	0.1	0.7	0.15	0.9	High
0.2	0.2	0.7	0.3	0.6	Middle

# 数据集-阿尔巴尼亚自杀率

特征	解释	年份	性别	年龄	总GDP	人均DPG	自杀率
年份	无	1987	0	3	21.56	796	5.19
		1987	1	4	21.56	796	0
性别	男性=0 女性=1	1988	1	2	21.26	769	1.91
		1989	0	1	23.35	833	3.71
		1989	0	4	23.35	833	4.22
年龄	5-14岁=0 15-24岁=1 25-34岁=2 35-54岁=3 55-74岁=4	1993	0	1	12.28	437	7.4
		1994	0	3	19.85	697	2.48
		1996	1	2	33.14	1127	3.73
		1997	0	3	23.59	793	7.88
总GDP	单位为亿美元	1998	1	0	27.07	899	0.27
		2000	1	3	36.32	1299	1.51
人均GDP	单位为美元	说明 本数据集反映了阿尔巴尼亚1987年-2010年自杀率与经济的关系。					
标签	解释	意义 分析阿尔巴尼亚经济环境与自杀率的关系，寻找不同性别、不同年龄段自杀率上升的信号，有计划性地进行预防与干预。					
自杀率	每10万人口自杀率	URL <a href="https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016">https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016</a>					

# 数据集-膀胱炎症数据

特征	解释
Temp.	Temperature of patient 病人体温
Nau.	Occurrence of nausea 是否有恶心症状 (是为1, 否为0)
Lum.p	Lumbar pain 腰椎是否疼痛 (是为1, 否为0)
Uri.Pu	Urine pushing(continuous need for urination) 是否有持续性尿意 (是为1, 否为0)
Mic.p	Micturition pains 是否有尿痛症状 (是为1, 否为0)
Ure.o	Burning of urethra, itch, swelling of urethra outlet 尿道是否有灼烧感, 是否发痒, 尿道口是否肿胀 (是为1, 否为0)
标签	解释
Infl.	Inflammation of urinary baldder 是否有膀胱炎 (是为1, 否为0)

Temp.	Nau.	Lum.p	Uri.Pu	Mic.p	Ure.o	Infl.
35.5	0	1	0	0	0	0
35.9	0	0	1	1	1	1
35.9	0	1	0	0	0	0
36	0	0	1	1	1	1
36	0	1	0	0	0	0
36	0	1	0	0	0	0
36.2	0	0	1	1	1	1
36.2	0	1	0	0	0	0
36.3	0	0	1	1	1	1
36.6	0	0	1	1	1	1
36.6	0	0	1	1	1	1
36.6	0	1	0	0	0	0

说明	该数据集展示了膀胱炎患者的各个身体特征数据。
意义	该数据集的主要意义在于测试专家系统, 该系统可通过患者症状对“患者是否患有膀胱炎”做出推定诊断。
URL	<a href="https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations">https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations</a>

# 数据集-玻璃类型数据

特征	解释
RI	玻璃折光率
Na	玻璃中钠元素的含量
Mg	玻璃中镁元素的含量
Al	玻璃中铝元素的含量
Si	玻璃中硅元素的含量
K	玻璃中钾元素的含量
Ca	玻璃中钙元素的含量
Ba	玻璃中钡元素的含量
Fe	玻璃中铁元素的含量
type	1 建筑窗口浮动处理 2 建筑窗口不浮动处理 3 车辆窗口浮动处理 4 车辆窗口不浮动处理 5 容器 6 餐具 7 前照灯

RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	type
1.523	13.31	3.58	0.82	71.99	0.12	10.17	0	0.03	1
1.51574	14.86	3.67	1.74	71.87	0.16	7.36	0	0.12	2
1.51934	13.64	3.54	0.75	72.65	0.16	8.89	0.15	0.24	3
1.52211	14.19	3.78	0.91	71.36	0.23	9.14	0	0.37	3
1.51514	14.01	2.68	3.5	69.89	1.68	5.87	2.2	0	5
1.51915	12.73	1.85	1.86	72.69	0.6	10.09	0	0	5
1.52171	11.56	1.88	1.56	72.86	0.47	11.41	0	0	5
1.52151	11.03	1.71	1.56	73.44	0.58	11.62	0	0	5
1.51969	12.64	0	1.65	73.75	0.38	11.53	0	0	5
1.51905	14	2.39	1.56	72.37	0	9.57	0	0	6
1.51937	13.79	2.41	1.19	72.76	0	9.77	0	0	6
1.51623	14.14	0	2.88	72.61	0.08	9.18	1.06	0	7
说明	此数据集展示了各种不同的玻璃的特征及所含元素占比量。								
意义	利用折光率、玻璃中各种元素的含量等信息可以预测玻璃的类型，基于类型的判断，可以得知该玻璃的用途。								
URL	<a href="https://archive.ics.uci.edu/ml/datasets/Glass+Identification">https://archive.ics.uci.edu/ml/datasets/Glass+Identification</a>								



# 数据集-具有声学特征的帕金森数据

特征	解释	试验者	性别	RAP	Shim	HNR	MFC	RPDE	DFA	PPE	GNE	是否患病
RAP	相对平均扰动	CONT-01	1	0.0015	0.0303	59.4380	1.3195	0.2504	0.5993	0.0054	0.9004	0
Shim	幅度扰动	CONT-02	0	0.0011	0.0247	62.1796	1.4952	0.2614	0.6252	0.0305	0.8760	0
HNR	谐波噪声比	CONT-03	1	0.0014	0.0172	76.2689	1.5019	0.2621	0.5727	0.0059	0.9821	0
MFC	谱测量	CONT-04	1	0.0082	0.0307	68.6250	1.5771	0.2954	0.7428	0.7824	0.9083	0
RPDE	复发期密度熵	CONT-05	0	0.0042	0.0392	78.9730	1.3735	0.3193	0.7669	0.2566	0.9128	0
DFA	趋势波动	PARK-01	0	0.0032	0.0404	52.7494	1.3174	0.2451	0.6597	0.5605	0.8817	1
PPE	间距周期熵	PARK-02	0	0.0014	0.0260	84.1942	1.3882	0.2465	0.5161	0.0053	0.9140	1
GNE	噪声激发比	PARK-03	0	0.0039	0.1114	60.0669	1.5176	0.3743	0.5528	0.3940	0.9169	1
标签	解释	PARK-04	0	0.0012	0.0151	71.6610	1.8056	0.2194	0.5476	0.0287	0.9711	1
是否患病	0-否      1-是	PARK-05	0	0.0011	0.0143	57.8317	1.2720	0.2286	0.6273	0.2336	0.8911	1
说明		本数据集提取10个人（含5名帕金森患者）的声学特征信息										
意义		揭示了声学特征与是否患有帕金森疾病之间的关系，通过提取一个人的相关信息，可以判断他是否患有帕金森										
URL		<a href="https://archive.ics.uci.edu/ml/datasets/Parkinson/">https://archive.ics.uci.edu/ml/datasets/Parkinson/</a>										

## 实时卖家帖子参与度

特征		解释		实时卖家帖子参与度								
status_type	发布类型	status_type	status_published	num_reactions	num_comments	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys
status_published	发布时间	video	4/22/2018 6:00	529	512	262	432	92	3	1	1	0
num_reactions	反应人数	photo	4/21/2018 22:45	150	0	0	150	0	0	0	0	0
num_comments	评论次数	video	4/21/2018 6:17	227	236	57	204	21	1	1	0	0
num_shares	分享次数	photo	4/21/2018 2:29	111	0	0	111	0	0	0	0	0
num_likes	点赞次数	photo	4/18/2018 3:22	213	0	0	204	9	0	0	0	0
num_loves/wows/hahas/sads/angrys	发布后各参与状态人数	photo	4/18/2018 2:14	217	6	0	211	5	1	0	0	0
		video	4/18/2018 0:24	503	614	72	418	70	10	2	0	3
		video	4/17/2018 7:42	295	453	53	260	32	1	1	0	1
		photo	4/17/2018 3:33	203	1	0	198	5	0	0	0	0
		photo	4/11/2018 4:53	170	9	1	167	3	0	0	0	0
说明	本数据集分析了Facebook页面上的10家泰国时尚化妆品零售商家中不同性质的帖子(视频、照片、状态和链接)。这些帖子的参与度量由评论、共享和反应组成。											
意义	分析消费者参与的可变性，强调了Facebook 的使用带来的变化。确定统计上的异常点，并对其销售方法和活动进行进一步的定性分析。											
URL	<a href="https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand">https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand</a>											

# 数据集-航天飞机降落参数及推荐方式

特征	解释	稳定度	信号	风力干扰	震感	可见度	推荐降落方式
稳定度	稳定/不稳定	稳定	pp	头部	弱	可见	自动
信号	pp/nn	稳定	pp	头部	中	可见	自动
风力干扰	头部/尾部	稳定	pp	尾部	弱	可见	自动
震感	强/中/弱	稳定	pp	尾部	中	可见	自动
可见度	可见/不可见	稳定	pp	头部	强	可见	手动
推荐降落方式	自动/手动	稳定	pp	尾部	强	可见	自动
		不稳定	nn	头部	强	不可见	自动
		不稳定	nn	尾部	弱	不可见	自动
		不稳定	nn	头部	中	不可见	自动
		稳定	nn	尾部	中	可见	手动
		稳定	pp	尾部	强	可见	手动

说明	此数据集收集了航天飞机降落时的一些参数作为特征数据。
意义	通过此数据集可以找到航天飞机推荐降落方式与各个降落参数之间的相关性。用于推测某次降落是否需要宇航员来进行手动操作。
URL	<a href="https://archive.ics.uci.edu/ml/datasets/Shuttle+Landing+Control">https://archive.ics.uci.edu/ml/datasets/Shuttle+Landing+Control</a>

# 数据集-收入普查数据

特征	解释
age	公民年龄
workclass	公民工作类型
occupation	公民职业
sex	公民性别
worktime	公民工作时长
标签	解释
conclusion	公民收入（0表示年薪<50k, 1表示年薪>=50k）

age	workclass	occupation	sex	worktime	conclusion
39	3	6	1	40	0
50	7	10	1	13	0
38	8	8	1	40	0
53	8	8	1	40	0
28	8	9	0	40	0
37	8	10	0	40	0
49	8	12	0	16	0
52	7	10	1	45	1
31	8	9	0	50	1
42	8	10	1	40	1
37	8	10	1	80	1
30	3	9	1	40	1
意义	探究工作类型、年龄、工作时长等对公民收入的影响，可为准备选取专业和求职的人群提供参考，也可根据目前各职业年龄层分布、性别分布、工作时长等探究民生、就业问题				
URL	<a href="https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations">https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations</a>				

# 数据集-ILPD（印度肝病患者数据集）

特征	解释
Age	年龄
Gender	性别
TB	总红胆素
AAP	碱性磷酸酶
Sgpt	丙氨酸氨基转移酶
Sgot	天冬氨酸氨基转移酶
TP	总蛋白
ALB	白蛋白
A/G	比率白蛋白和球蛋白比率
标签	解释
是否患病	0-否      1-是

Age	Gender	TB	DB	AAP	Sgpt	Sgot	TP	ALB	A/G	是否患病
17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	0
55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8	1
72	Male	2.7	1.3	260	31	56	7.4	3	0.6	1
64	Male	0.9	0.3	310	61	58	7	3.4	0.9	0
74	Female	1.1	0.4	214	22	30	8.1	4.1	1	1
61	Male	0.7	0.2	145	53	41	5.8	2.7	0.87	1
25	Male	0.6	0.1	183	91	53	5.5	2.3	0.7	0
38	Male	1.8	0.8	342	168	441	7.6	4.4	1.3	1
33	Male	1.6	0.5	165	15	23	7.3	3.5	0.92	0
说明	此数据集包含10个实验者（6个患有肝脏疾病）的生物特征数据，目的是用于肝脏疾病的诊断。									
意义	该数据集表现了生物特征与是否患有肝脏之间的关系，通过提取一个人的特征，可以判断其是否患病。									
URL	<a href="https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29">https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29</a>									

# 数据集-气球与拿气球人的特征数据

特征	解释
颜色	黄色，紫色
大小	小，大
行为	拉紧，浸水
年龄	成年人，小孩
是否充气	是，否

说明	这是一组曾经用于认知心理学实验的数据，此数据集包含拿气球的行为特征数据。
意义	通过此数据集可以实现根据气球的与拿气球人的特征来判断气球是否为充气气球。
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Balloons">http://archive.ics.uci.edu/ml/datasets/Balloons</a>

颜色	大小	行为	年龄	是否充气
黄色	小	拉紧	成年人	是
黄色	小	拉紧	小孩	否
黄色	小	浸水	成年人	否
黄色	小	浸水	小孩	否
黄色	大	拉紧	成年人	是
黄色	大	拉紧	小孩	否
黄色	大	浸水	成年人	否
黄色	大	浸水	小孩	否
紫色	小	拉紧	成年人	是
紫色	小	拉紧	小孩	否
紫色	小	浸水	成年人	否
紫色	小	浸水	小孩	否
紫色	大	拉紧	成年人	是
紫色	大	拉紧	小孩	否
紫色	大	浸水	成年人	否
紫色	大	浸水	小孩	否

特征	解释	Match_id	assists	deaths	kills	gpm	level	win
Match_id	比赛ID	3870602242	7	7	15	469	18	0
		3870453145	14	5	8	503	21	1
Assists	助攻数	3868388262	24	5	10	603	25	1
Deaths	死亡数	3868240894	6	5	16	620	25	1
		3866278803	7	5	6	572	21	1
Kills	击杀数	3866190562	2	3	4	495	20	0
Gpm	每分钟金钱	3866060233	18	6	12	624	25	1
		3862555760	9	2	9	598	24	1
Level	英雄等级	3862410871	24	5	13	511	22	1
Win	0-失败, 1-胜利	3860298305	11	5	15	623	23	1
		3860181472	12	1	13	597	24	1

说明	此数据集分析记录DOTA2职业赛中单选手的赛场游戏参数。
意义	通过对中单选手每场比赛相关数据的分析, 来预测比赛的胜负。同时可以对对其训练做出指导, 通过训练提高某一方面数据来提高比赛获胜几率。
URL	<a href="https://www.kaggle.com/siegmundwang/dota2">https://www.kaggle.com/siegmundwang/dota2</a>



# 数据集 - 鲍鱼体态及性别特征数据

特征	解释	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings	Sex
Length	体长，鲍鱼带壳长度	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	F
Diameter	直径，与体长垂直的长度	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	M
Height	鲍鱼的总体高度	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	I
		0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	I
Whole weight	鲍鱼总体重量	0.525	0.38	0.14	0.6065	0.194	0.1475	0.21	14	F
		0.43	0.35	0.11	0.406	0.1675	0.081	0.135	10	F
Shucked weight	去壳重，鲍鱼肉的重量	0.49	0.38	0.135	0.5415	0.2175	0.095	0.19	11	M
Viscera weight	鲍鱼脏器重量	0.535	0.405	0.145	0.6845	0.2725	0.171	0.205	10	F
		0.47	0.355	0.1	0.4755	0.1675	0.0805	0.185	10	F
Shell weight	鲍鱼壳重量	0.5	0.4	0.13	0.6645	0.258	0.133	0.24	12	M
		0.355	0.28	0.085	0.2905	0.095	0.0395	0.115	7	M
Rings	鲍鱼生长环数	说明	无							
Sex	鲍鱼性别 I-婴儿 F-雌性 M- 雄性 婴儿时期的鲍鱼没有性别的区分	意义	意义：可以通过对此数据集的数据分析，对样本以外的鲍鱼进行性别预测，判断鲍鱼的性别是M,F还是I。达到以非专业手段简单判别鲍鱼性别的目的							
		URL	<a href="http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/">http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/</a>							

# 数据集-鸢尾属植物特征及所属类别数据

特征	解释	萼片长度	萼片宽度	花瓣长度	花瓣宽度	所属类别
萼片长度	植物属性，单位CM	5.1	3.5	1.4	0.2	setosa
		6.3	3.3	4.7	1.6	versicolor
萼片宽度	植物属性，单位CM	7.6	3.0	6.6	2.1	virginica
		5.0	3.6	1.4	0.2	setosa
花瓣长度	植物属性，单位CM	5.9	3.0	4.2	1.5	versicolor
		6.5	3.2	5.1	2.0	virginica
花瓣宽度	植物属性，单位CM	5.4	3.7	1.5	0.2	setosa
		6.2	2.2	4.5	1.5	versicolor
所属类别	鸢尾属植物类别，包括三类：setosa，versicolor与virginica。	7.7	3.8	6.7	2.2	virginica
		5.4	3.9	1.3	0.4	setosa
		6.8	2.8	4.8	1.4	versicolor



说明	此数据集由R.A. Fisher在1988年创建，包含鸢尾属植物的三个类别各50例，提取样本的四个属性信息：萼片长度，萼片宽度，花瓣长度，花瓣宽度。标签为所属鸢尾属植物类别，包括三类：setosa，versicolor与virginica。
意义	意义：通过此数据集可以观察了解鸢尾属植物的以上四个特征信息与类别之间的相关性。可以用于对某一鸢尾属植物通过属性信息预测其所属类别。
URL	<a href="https://archive.ics.uci.edu/ml/datasets/Iris">https://archive.ics.uci.edu/ml/datasets/Iris</a>

Date	Time	CO(G T)	PT08. S1(C O)	NM HC(G T)	C6H 6(GT )	PT08.S 2(NM HC)	NOx(GT )	PT08. S3(N Ox)	NO2( GT)	PT08.S 4(NO2 )	PT08.S 5(O3)	T	RH	AH
2004/3/10	18:00:00	2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9	0.7578
2004/3/10	19:00:00	2	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7	0.7255
2004/3/10	20:00:00	2.2	1402	88	9.0	939	131	1140	114	1555	1074	11.9	54.0	0.7502
2004/3/10	21:00:00	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11.0	60.0	0.7867
2004/3/10	22:00:00	1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6	0.7888
2004/3/11	1:00:00	1	1136	31	3.3	672	62	1453	76	1333	730	10.7	60.0	0.7702
2004/3/11	2:00:00	0.9	1094	24	2.3	609	45	1579	60	1276	620	10.7	59.7	0.7648
2004/3/11	3:00:00	0.6	1010	19	1.7	561	-200	1705	-200	1235	501	10.3	60.2	0.7517
2004/3/11	4:00:00	-200	1011	14	1.3	527	21	1818	34	1197	445	10.1	60.5	0.7465
2004/3/11	5:00:00	0.7	1066	8	1.1	512	16	1918	28	1182	422	11.0	56.2	0.7366

# 数据集-空气各成分浓度与污染程度（Part2）

特征	解释	特征	解释
Date	DD/MM/YYYY	NOx(GT)	ppb中NOx真实小时平均浓度
Time	HH.MM.SS	PT08.S3(NOx)	PT08.S3(氧化钨)每小时平均传感器响应(NOx靶向)
CO(GT)	mg/m^3中CO的真实平均浓度	NO2(GT)	mg/m^3中NO2真实小时平均浓度
PT08.S1(CO)	PT08.S1(氧化锡)每小时平均传感器响应(CO靶向)	PT08.S4(NO2)	PT08.S4(氧化钨)每小时平均传感器响应(NO2靶向)
NMHC(GT)	mg/m^3中总非亚甲基烃类浓度的真实小时浓度	PT08.S5(O3)	PT08.S5(氧化钨)每小时平均传感器响应(名义上O3目标)
C6H6(GT)	mg/m^3中真实每小时平均苯浓度(参考分析仪)	T	°C的温度
PT08.S2(NMHC)	PT08.S2(二氧化钛)每小时平均传感器响应(NMHC靶向)	RH	相对湿度(%)
		AH	绝对湿度

说明	本数据集的数据来源于部署在意大利一个城市现场的气体多传感器设备的响应，真实记录地面实际每小时CO，非亚甲基碳氢化合物，苯，总氮氧化物(NOx)和二氧化氮(NO2)的平均浓度。
意义	可以通过调查城市空气的以上参数浓度，判断该城市空气受污染的严重程度（轻度/重度）。并库判断空气污染程度与哪一种物质的相关性更大，从而更好得治理污染。
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Air+Quality">http://archive.ics.uci.edu/ml/datasets/Air+Quality</a>

帅器学习  
TechYoung Machine Learning Class

特征	解释	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	PH	sulphates	alcohol	quality
fixed acidity	非挥发性酸	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
volatile acidity	可挥发性酸	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
citric acid	柠檬酸	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
residual sugar	红酒甜度	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
chlorides	氯化物	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
free sulfur dioxide	游离二氧化硫	7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4
total sulfur dioxide	二氧化硫总量	6.9	0.4	0.14	2.4	0.085	21	40	0.9968	3.43	0.63	9.7	6
density	葡萄酒密度	6.3	0.39	0.16	1.4	0.08	11	23	0.9955	3.34	0.56	9.3	5
PH	液体酸碱度	7.6	0.41	0.24	1.8	0.08	4	11	0.9962	3.28	0.59	9.5	5
quality	红酒质量 (0-10)	说明 本数据集中的特征数据为不同种类的红酒所具有的各种成分含量。数据来自对红酒的物理化学分析。											
		意义 该数据集可用于建构红酒质量评级模型，根据某种红酒中各种成分的含量，推测这种红酒的质量评分。											
		URL <a href="http://archive.ics.uci.edu/ml/datasets/Wine+Quality">http://archive.ics.uci.edu/ml/datasets/Wine+Quality</a>											

# 数据集-心脏病患者特征数据

特征	解释
年龄	患者年龄
性别	1-男 0-女
休息时的血压	血压测量值
胆固醇	血液中胆固醇含量 (mg/dl)
空腹血糖	是否大于120mg/dl(是=1, 否=0)
最高心率	游离二氧化硫
胸痛类型	1- typical angina 2- atypical angina 3- non-anginal pain 4- asymptomatic
说明	本数据集包括匈牙利心脏病患者的特征数据。
意义	通过分析样本数据可以得出患者所患的胸痛类型与以上数据特征之间的相关性，用于通过特征参数判断新患者具体是何种胸痛类型。
URL	<a href="http://archive.ics.uci.edu/ml/datasets/Heart+Disease">http://archive.ics.uci.edu/ml/datasets/Heart+Disease</a>

年龄	性别	休息 血压	胆固 醇	空腹 血糖	最高心 率	.....	胸痛 类型
63	1	145	233	1	150		3
37	1	130	250	0	187		2
41	0	130	204	0	172		1
56	1	120	236	0	178		1
57	0	120	354	0	163		0
57	1	140	192	0	148		0
56	0	140	294	0	153		1
44	1	120	263	0	173		1
52	1	172	199	1	162		2
57	1	150	168	0	174		2
66	0	150	226	0	114		3
43	1	150	247	0	171		0



# 数据集-男性两种急性炎症与症状数据

特征	解释
A1	病人体温
A2	是否有恶心症状
A3	是否有腰椎疼痛
A4	是否持续需要排尿
A5	是否排尿疼痛
A6	是否尿道烧灼、发痒、尿道出口肿胀
D1	D1-患有膀胱炎
D2	患有肾盂肾炎
说明	该数据集由医学专家创建，每个实例代表一个患者。
意义	作为测试专家系统的数据集，该系统可通过患者症状对两种泌尿系统疾病进行推定诊断。
UCL	<a href="http://archive.ics.uci.edu/ml/datasets/Acute+Inflammations">http://archive.ics.uci.edu/ml/datasets/Acute+Inflammations</a>

A1	A2	A3	A4	A5	A6	D1	D2
35,5	no	yes	no	no	no	no	no
35,9	no	no	yes	yes	yes	yes	no
36,7	no	no	yes	yes	yes	yes	no
36,7	no	yes	no	no	no	no	no
36,7	no	yes	no	no	no	no	no
36,9	no	yes	no	no	no	no	no
37,1	no	no	yes	no	no	yes	no
37,2	no	no	yes	yes	no	yes	no
37,9	no	no	yes	yes	no	yes	no
37,9	no	yes	no	no	no	no	no
37,9	no	no	yes	yes	yes	yes	no
37,9	no	no	yes	no	no	yes	no
38,0	no	yes	yes	no	yes	no	yes
38,0	no	yes	yes	no	yes	no	yes
38,1	no	yes	yes	no	yes	no	yes
38,3	no	yes	yes	no	yes	no	yes
38,5	no	yes	yes	no	yes	no	yes
38,7	no	yes	yes	no	yes	no	yes
40,0	yes	yes	yes	yes	yes	yes	yes
40,0	yes	yes	yes	yes	no	yes	yes



# 数据集-纽约市官方招聘网站招聘信息及其标签数据

特征	解释	Agency	# Of Positions	Job Category	Salary Range From	Salary Range To	Do I like it
Agency	空缺职位所在市代理机构名称	NYC HOUSING AUTHORITY	52	Maintenance & Operations	51907.68	54580.32	0
# Of Positions	岗位空缺数目						
Job Category	工作类别	NYC HOUSING AUTHORITY	52	Maintenance & Operations	51907.68	54580.32	0
SalaryRangeFrom	薪资底限						
Salary Range To	薪资上限	DEPT OF HEALTH/MENTAL HYGIENE	1	Finance, Accounting, & Procurement	50598	85053	0
Do I like it	是否喜欢这份工作 1—喜欢    0—不喜欢						
<b>说明</b>	<b>该数据集包含纽约市官方招聘网站上的当前职位招聘信息。城市雇员可获得的内部职位和公众可获得的外部职位也包括在内</b>	DEPT OF ENVIRONMENT PROTECTION	1	Health Public Safety, Inspections, & Enforcement	50623	75083	0
		NYC HOUSING AUTHORITY	1	Engineering, Architecture, & Planning	90000	110000	1
意义	想要找工作时，招聘网站可根据找工作者平时着重浏览的工作或其喜欢的工作特征对其进行相关推荐。						
UCL	<a href="https://www.kaggle.com/new-york-city/new-york-city-current-job-postings">https://www.kaggle.com/new-york-city/new-york-city-current-job-postings</a>						

# 数据集-泰坦尼克号生存情况

特征	解释
Pclass	乘客级别 1-3
Sex	性别 F- 女性 M-男性
age	乘客年龄
Sibsp	兄弟姐妹加上配偶的总人数
Parch	父母加上孩子的总人数
Ticket	船票编号
Cabin	船舱编号
Embarked	上船港口代号C = Cherbourg, Q = Queenstown, S = Southampton
survived	乘客是否生还 1—乘客生还; 0—乘客未生还

Name	survived	Pclass	Sex	Age	SibSp
Braund, Mr. Owen Harris	0	3	M	22	1
Cumings, Mrs. John Bradley	1	1	F	38	1
Heikkinen, Miss. Laina	1	3	F	26	0
Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	1	F	35	1
Allen, Mr. William Henry	0	3	M	35	0

Parch	Ticket	Fare	Cabin	Embarked
1	A/5 21171	7.25		S
0	PC 17599	71.2833	C85	C
0	STON/O2. 3101282	7.925		S
0	113803	53.1	C123	S
0	373450	8.05		S

说明	此数据集为泰坦尼克号事故中生存情况的特征数据调查。
意义	可以根据数据集中各参数与是否生还的相关性，根据泰坦尼克号乘客的以上特征参数判别其是否生还。
UCL	<a href="https://www.kaggle.com/c/titanic/data">https://www.kaggle.com/c/titanic/data</a>

# 数据集-蒙泰西尼奥公园森林火灾数据（Part1）

X轴	Y轴	月份	日	FFMC	DMC	DC	ISI	温度	湿度	风速	雨量	燃烧面积
4	3	9月	周二	91.9	111.7	770.3	6.5	15.9	53	2.2	0	2.93
7	5	8月	周二	96.1	181.1	671.2	14.3	27.3	63	4.9	6.4	10.82
6	5	9月	周二	91.9	111.7	770.3	6.5	19.6	45	3.1	0	20.03
4	3	5月	周五	89.6	25.4	73.7	5.7	18	40	4	0	38.48
8	3	1月	周一	88.2	96.2	229	4.7	14.3	79	4	0	1.94
6	5	9月	周六	91.2	94.3	744.4	8.4	16.8	47	4.9	0	12.64
6	5	8月	周五	91	166.9	752.6	7.1	18.2	62	5.4	0	0.43
8	6	8月	周日	81.6	56.7	665.6	1.9	27.8	35	2.7	0	0
7	4	10月	周六	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
8	6	2月	周五	91.7	33.3	77.5	9	8.3	97	4	0.2	0
6	5	9月	周六	92.5	121.1	674.4	8.6	25.1	27	4	0	1090.84
7	5	4月	周日	81.9	3	7.9	3.5	13.4	75	1.8	0	0

# 数据集-蒙泰西尼奥公园森林火灾数据 (Part2)

X轴	Montesinho公园地图中的 X轴空间坐标: 1到9	DMC	FWI系统的DMC (粗腐殖质湿度码) 指数 1.1到291.3
Y轴	Montesinho公园地图中的Y轴空间坐标: 2到9	DC	FWI系统的DC (干旱码) 指数7.9到860.6
年月	1月-12月	ISI	来自FWI系统的ISI指数 (火灾初始蔓延指数) :0.0到56.10
日	周一-周日	温度	2.2至33.30度
FFMC	FWI系统 (加拿大森林火险气候指数系统) 的FFMC (细小可燃物湿度码) 指数18.7到96.20	相对湿度	相对湿度(%):15.0 ~ 100
风速	风速(公里/小时):0.40至9.40	雨量	室外雨量(毫米/平方米):0.0至6.4

说明	以为蒙泰西尼奥公园森林的火灾相关特征参数 标签为火灾的最终燃烧面积
意义	利用气象和其他数据可以预测葡萄牙东北地区森林火灾的烧毁面积。且可以通过森林的以上特征指标来评估森林发生火灾的可能性位置和火灾程度, 这样就可以有针对性地对特定地点的森林进行监控, 使得有限的监控资源得到充分有效的利用。
URL	<a href="https://archive.ics.uci.edu/ml/datasets/Forest+Fires">https://archive.ics.uci.edu/ml/datasets/Forest+Fires</a>

数据集-Goodreads网站书籍数据-part1/2									<div> <div>帅器学习</div> <div>TechYoung Machine Learning Class</div> </div>
bookID	title	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count
1	Harry Potter and the Half-Blood Prince (Harry Potter #6)	J.K. Rowling-Mary GrandPré	4.56	0439785960	9780439785969	eng	652	1944099	26249
2	Harry Potter and the Order of the Phoenix (Harry Potter #5)	J.K. Rowling-Mary GrandPré	4.49	0439358078	9780439358071	eng	870	1996446	27613
3	Harry Potter and the Sorcerer's Stone (Harry Potter #1)	J.K. Rowling-Mary GrandPré	4.47	0439554934	9780439554930	eng	320	5629932	70390
4	Harry Potter and the Chamber of Secrets (Harry Potter #2)	J.K. Rowling-Mary GrandPré	4.41	0439554896	9780439554893	eng	352	6267	272
5	Harry Potter and the Prisoner of Azkaban (Harry Potter #3)	J.K. Rowling-Mary GrandPré	4.55	043965548X	9780439655484	eng	435	2149872	33964

特征	解释
bookID	该网站对书籍的唯一标识
title	书籍的标题
authors	书籍的作者
average_rating	该网站上书籍的平均得分
isbn	国际标准书号
isbn13	新的国际标准书号
language_code	书籍的语言
num_pages	书籍的页数
ratings_count	总评分数
text_reviews_count	总评论数



说明	此数据集包含了从Goodreads网站获取的13.7k条书籍数据，主要包括书籍的基本信息以及网站的评论数、评分数以及总得分。
意义	Goodreads是全球最大的在线读书交流社区，拥有极为丰富的图书信息，为人们推荐优秀的书籍，研究相关评论和评分可以帮助我们进行图书推荐，还可以研究图书得分与其他属性的相关性。
URL	<a href="https://www.kaggle.com/jealousleopard/goodreadsbooks">https://www.kaggle.com/jealousleopard/goodreadsbooks</a>

# 数据集-全国各省疫苗流向数据

name	src	create_company	report_company	prov	year	price
23价肺炎球菌多糖疫苗	进口	Merck Sharp & Dohme Corp.	Merck Sharp & Dohme Corp.	北京市	2018	0
重组乙型肝炎疫苗	国产	华北制药金坦生物技术股份有限公司	华北制药金坦生物技术股份有限公司	上海市	2018	93.5
支水痘减毒活疫苗	国产	上海生物制品研究所有限责任公司	上海生物制品研究所有限责任公司	陕西省	2016	136
流感病毒裂解疫苗	国产	长春生物制品研究所有限责任公司	长春生物制品研究所有限责任公司	海南省	2017	29
肠道病毒71型灭活疫苗(人二倍体细胞)	国产	中国医学科学院医学生物学研究所	中国医学科学院医学生物学研究所	广东省	2018	188

特征	解释	说明
name	疫苗名称	此数据集包含了2016-2018年全国各个省份的第二类疫苗采购公示数据，第二类疫苗指由公民自费并且自愿受种的其他疫苗，而第一类疫苗是政府免费提供按照规定受种的疫苗。
src	疫苗来源（国产/进口）	
create_company	生产企业	意义 长春长生这家药企被发现所造疫苗连续造假，人们对于疫苗质量有很大的担忧，通过对各省采购疫苗的数据可以追踪疫苗的来源，分析其分布特征，也可以观察到我国疫苗的供应情况。
report_company	申报企业	
prov	省份	
year	中标年份	URL <a href="https://tianchi.aliyun.com/dataset/dataDetail?dataId=1427">https://tianchi.aliyun.com/dataset/dataDetail?dataId=1427</a>
price	疫苗价格	



# 数据集-助教的教学绩效数据

Whether of not the TA is a native English speaker	Course instructor	course	Summer /regular semester	Class size	Class attribute
1	23	3	1	19	3
2	15	3	1	17	3
1	23	3	2	49	3
1	5	2	2	33	3
2	7	11	2	55	3
2	23	3	1	20	3
2	9	5	2	19	3
2	6	17	2	42	2
1	18	7	2	48	1
2	20	15	2	18	2
2	7	11	1	20	3
2	9	2	2	39	3

特征	解释
Whether of not the TA is a native English speaker	助教是否是英语母语者 1-英语, 2-非英语
Course instructor	教员编号
course	课程编号
Summer/regular semester	小学期还是正常学期 1-summer, 2-regular
Class size	课程人数
Class attribute	课程评价 1-low, 2-medium, 3-high

说明	此数据集包含了the University of Wisconsin-Madison统计系的151位助教的教学相关数据。
意义	通过对教学情况数据的分析了解影响助教教学绩效的影响因素，可以帮助预测未来某个助教的教学结果，帮助进行选择助教的决策，也给选择做助教的同学提供参考意义。
URL	<a href="https://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation">https://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation</a>

# 数据集-新闻主题分类数据

category	headline	author	link	Short_desc ription	date	特征	解释
						category	新闻的主体分类
						headline	文章题目
						author	课程编号
						link	文章发布的链接
						Short_description	文章简介
						date	发表时间
CRIME	There Were 2 Mass Shootings In Texas Last Week, But Only 1 On TV	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-amanda-painter-mass-shooting_us_5b081ab4e4b0802d69caad89	She left her husband. He killed their children. Just another day in America.	2018-05-26	说明	此数据集包含了2012到2018年200k条从HuffPost网站发布的新闻信息。
ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2...	Andy McDonald	https://www.huffingtonpost.com/entry/will-smit...	Of course it has a song.	2018-05-26	意义	通过对这些新闻相关数据的分析，可以帮助给未标签的新闻或文章进行打标签操作，可以结合语义分析技术帮助识别文章的类别。
ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-gran...	The actor and his longtime girlfriend Anna Ebe...	2018-05-26	URL	<a href="https://www.kaggle.com/rmisra/news-category-dataset">https://www.kaggle.com/rmisra/news-category-dataset</a>

# 数据集-银行市场营销活动数据

age	job	marital	education	default	housing	loan	contact	...	y
30	blue-collar	married	basic.9y	no	yes	no	cellular		no
39	sevices	single	high.school	no	no	no	telephone		no
25	sevices	married	high.school	no	yes	no	telephone		no
38	sevices	married	basic.9y	no	unknown	unknown	telephone		no
47	admin.	married	university.degree	no	yes	no	cellular		no
32	sevices	single	university.degree	no	no	no	cellular		no

说明	此数据集包含了某葡萄牙银行机构的直接市场营销活动的相关数据，目的是预测客户是否会订购定期存款这一项产品。
意义	银行作为经营货币金融产品和服务的特殊企业，也需制定市场营销战略,该数据集收集了客户数据、营销活动数据以及相关宏观经济数据，可以帮助我们分析营销活动的效果，分析影响客户订购的因素，预测客户是否会订购产品。
URL	<a href="https://archive.ics.uci.edu/ml/datasets/Bank+Marketing">https://archive.ics.uci.edu/ml/datasets/Bank+Marketing</a>

特征	解释
age	年龄
job	工作类型
marital	婚姻状况
education	教育背景
default	是否有过信贷违约
housing	是否有住房贷款
loan	是否有个人贷款
contact	联络方式
month	最后一次联络的月份
day_of_week	最后一次联络的星期
duration	最后一次联络的时长
campaign	与该客户在此次营销活动期间联络的次数
pdays	距离上次营销活动最后一次联络的天数
previous	与该客户在上次营销活动期间联络的次数
poutcome	上次营销活动的结果
emp.var.rate	就业变动率（季度指标）
cons.price.idx	CPI居民消费价格指数（月度指标）
cons.conf.idx	消费者信心指数（月度指标）
euribor3m	3个月欧元同业拆借利率（日度指标）
nr.employed	就业人数（季度指标）
y	是否订购定期存款

帅器学习  
TechYoung Machine Learning Class

©2018 TechYoung Machine Learning Class. All rights reserved.