



温州大學
WENZHOU UNIVERSITY

机器学习-机器学习项目流程

黄海广 副教授

2021年06月

本章目录

2

- 01** 机器学习项目流程概述
- 02** 数据清洗
- 03** 特征工程
- 04** 数据建模

1.机器学习项目流程概述

3

01 机器学习项目流程概述

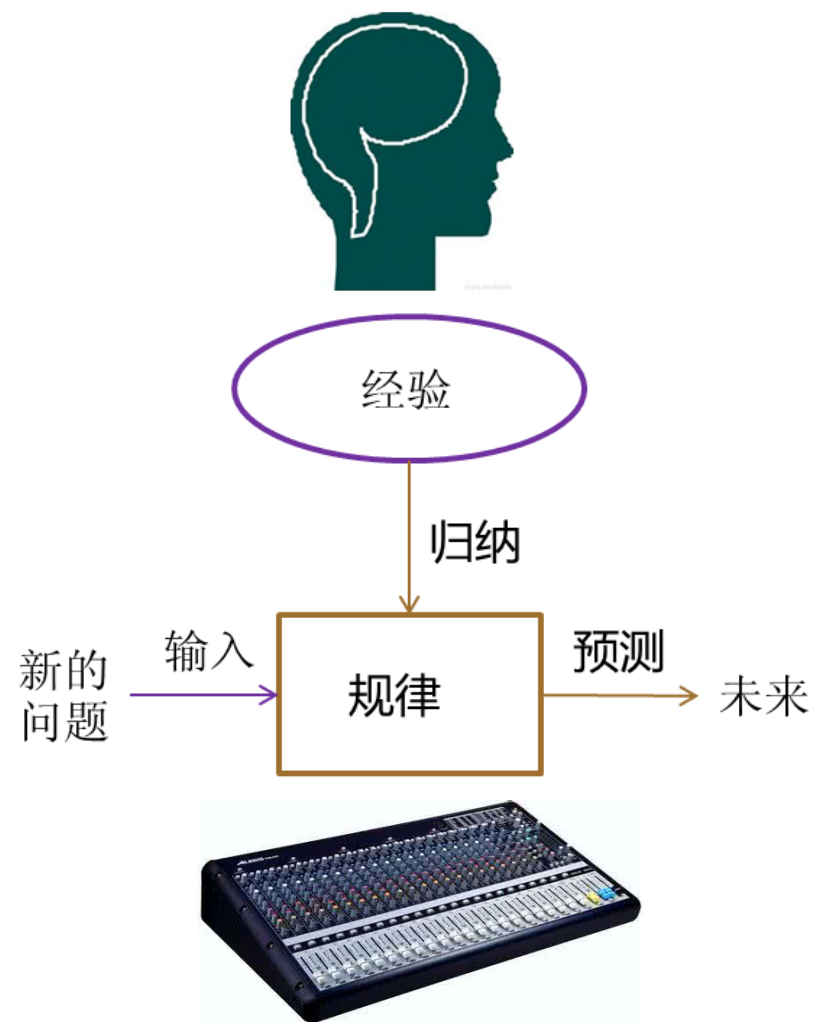
02 数据清洗

03 特征工程

04 数据建模

机器学习的一般步骤

4



机器学习的一般步骤

5

数据搜集



数据清洗



特征工程



数据建模



机器学习的一般步骤

6

数据搜集



数据清洗



特征工程



数据建模

- 网络下载
- 网络爬虫
- 数据库读取
- 开放数据
-

- 数据清理和格式化
- 探索性数据分析(EDA)

- 特征工程
- 特征选择

- 基于性能指标比较几种机器学习模型
- 对最佳模型执行超参数调整
- 在测试集上评估最佳模型
- 解释模型结果
- 得出结论

2.数据清洗

7

01 机器学习项目流程概述

02 数据清洗

03 特征工程

04 数据建模

2.数据清洗

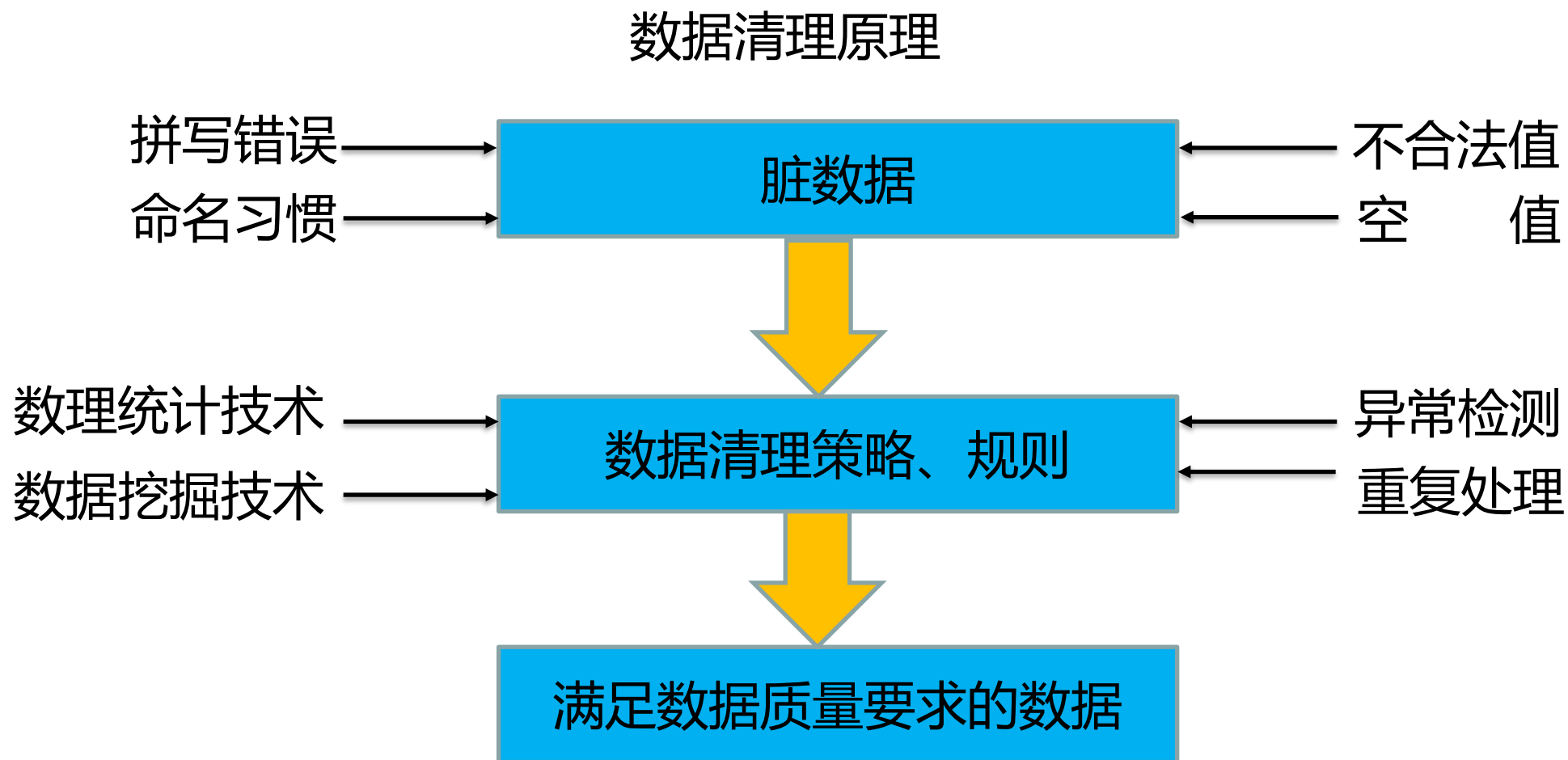
8

什么是数据清洗？

数据清洗是指发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等。与问卷审核不同，录入后的数据清理一般是由计算机而不是人工完成。

2.数据清洗

9



探索性数据分析(EDA)

10

探索性数据分析(EDA)

探索性数据分析 (EDA) 是一个开放式流程，我们制作绘图并计算统计数据，以便探索我们的数据。

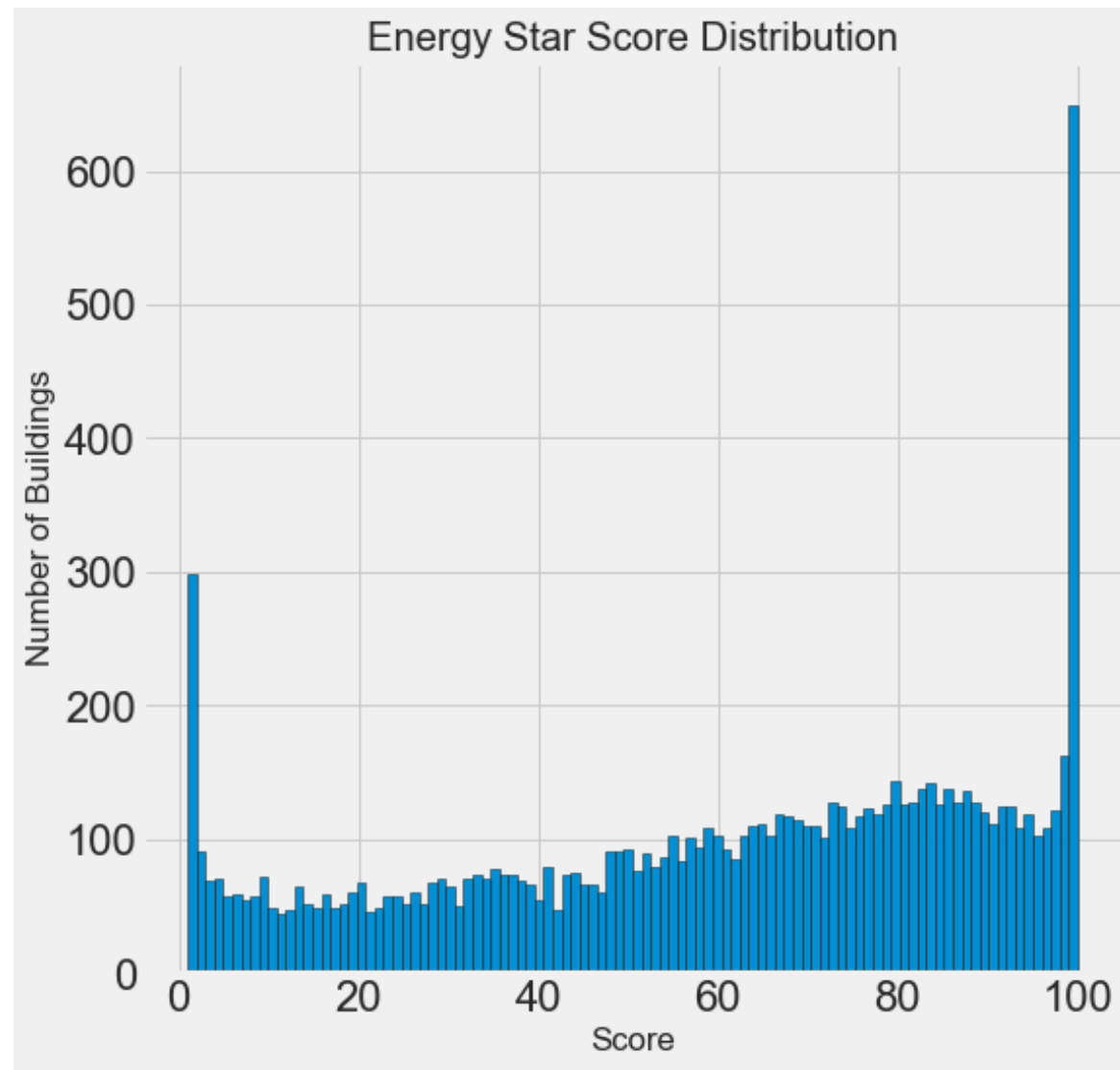
- 目的是找到异常，模式，趋势或关系。这些可能是有趣的（例如，找到两个变量之间的相关性），或者它们可用于建模决策，例如使用哪些特征。
- 简而言之，EDA的目标是确定我们的数据可以告诉我们什么！

探索性数据分析(EDA)

11

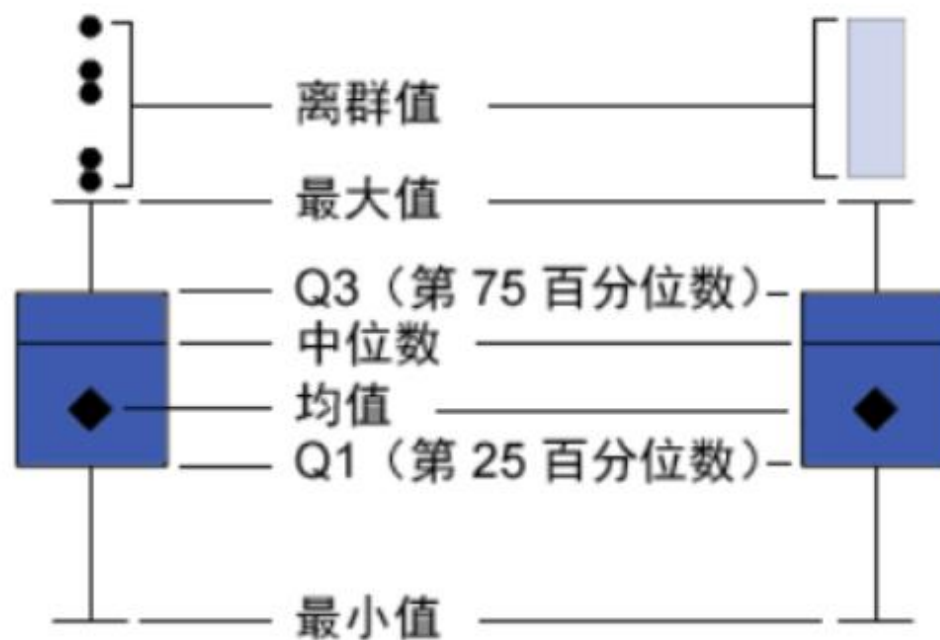
单变量图显示此变量的分布

plt.hist()可以显示单变量图，也叫直方图



探索性数据分析(EDA)

12



boxplot : 箱型图又称为盒须图、盒式图或箱线图，是一种用作显示一组数据分散情况资料的统计图。它能显示出一组数据的**最大值**、**最小值**、**中位数**及**上下四分位数**。

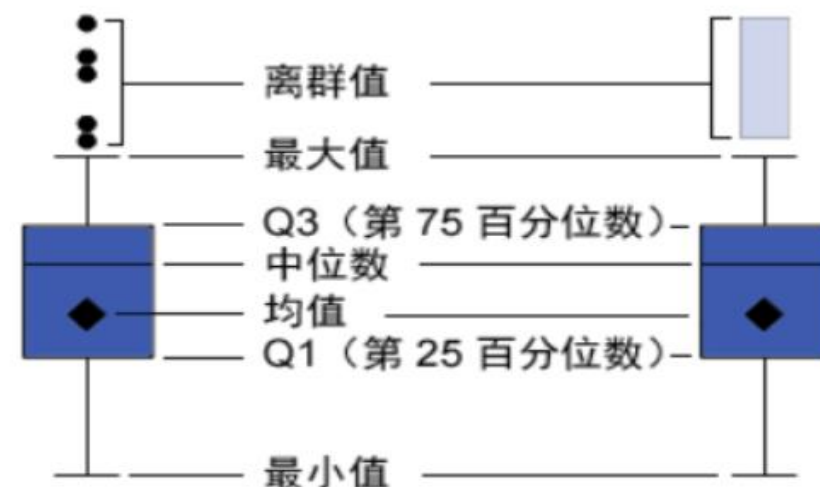
探索性数据分析(EDA)

13

$IQR = Q3 - Q1$ ，即上四分位数与下四分位数之间的差，也就是盒子的长度。

最小观测值为 $\min = Q1 - 1.5 * IQR$ ，如果存在离群点小于最小观测值，则下限为最小观测值，离群点单独以点汇出。

最大观测值为 $\max = Q3 + 1.5 * IQR$ ，如果存在离群点大于最大观测值，则上限为最大观测值，离群点单独以点汇出。如果没有比最大观测值大的数，则上限为最大值。



探索性数据分析(EDA)

14

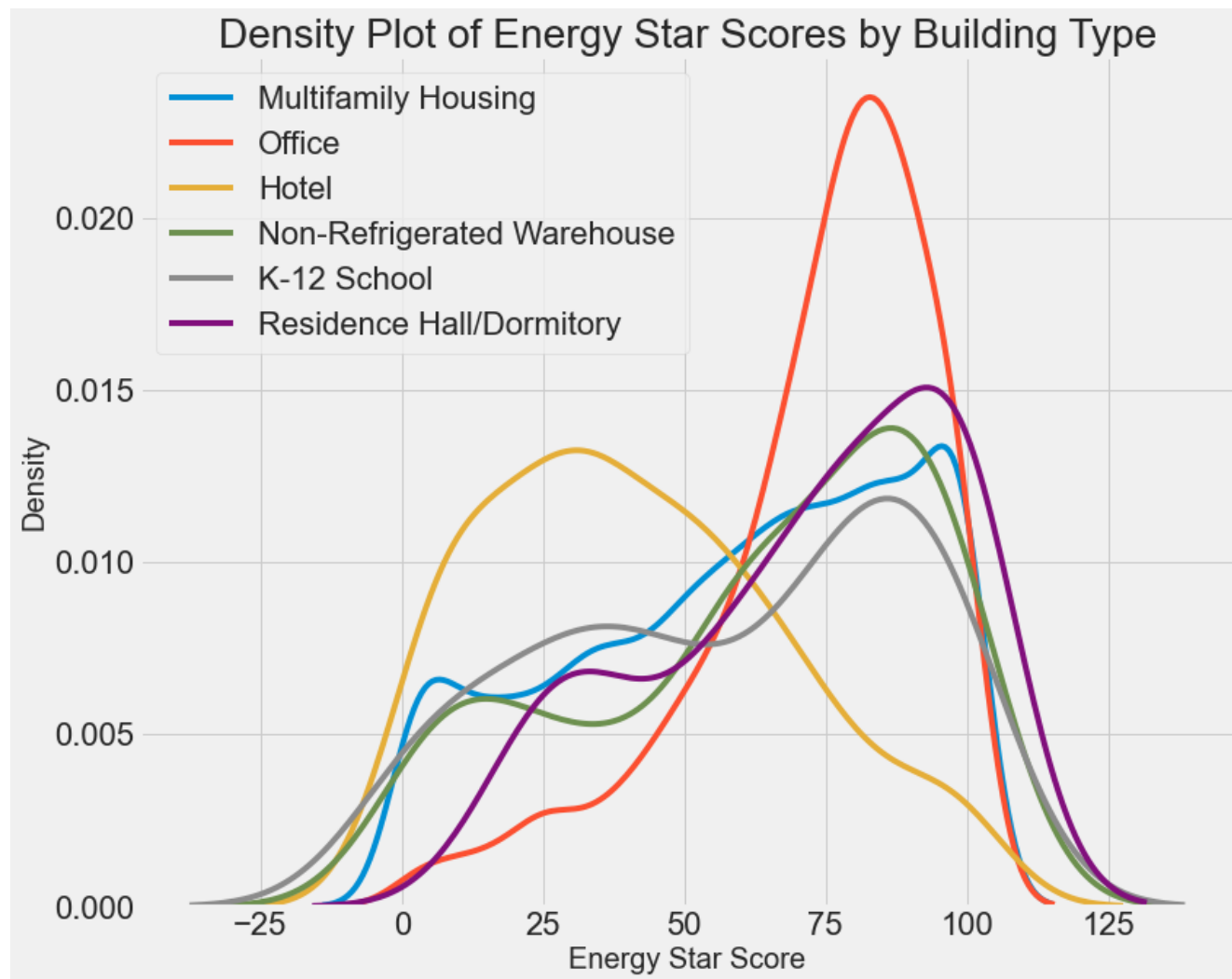
寻找关系

为了查看**分类变量 - categorical variables**对分数的影响，我们可以通过**分类变量**的值来绘制**密度图**。密度图还显示单个变量的分布，可以认为是平滑的直方图。如果我们通过为**分类变量**密度曲线着色，这将向我们展示分布如何基于类别变化的。

探索性数据分析(EDA)

15

这幅图我们可以看到**建筑类型**对 Energy Star Score 有重大影响。办公楼往往有较高的分数，而酒店的分数较低。



探索性数据分析(EDA)

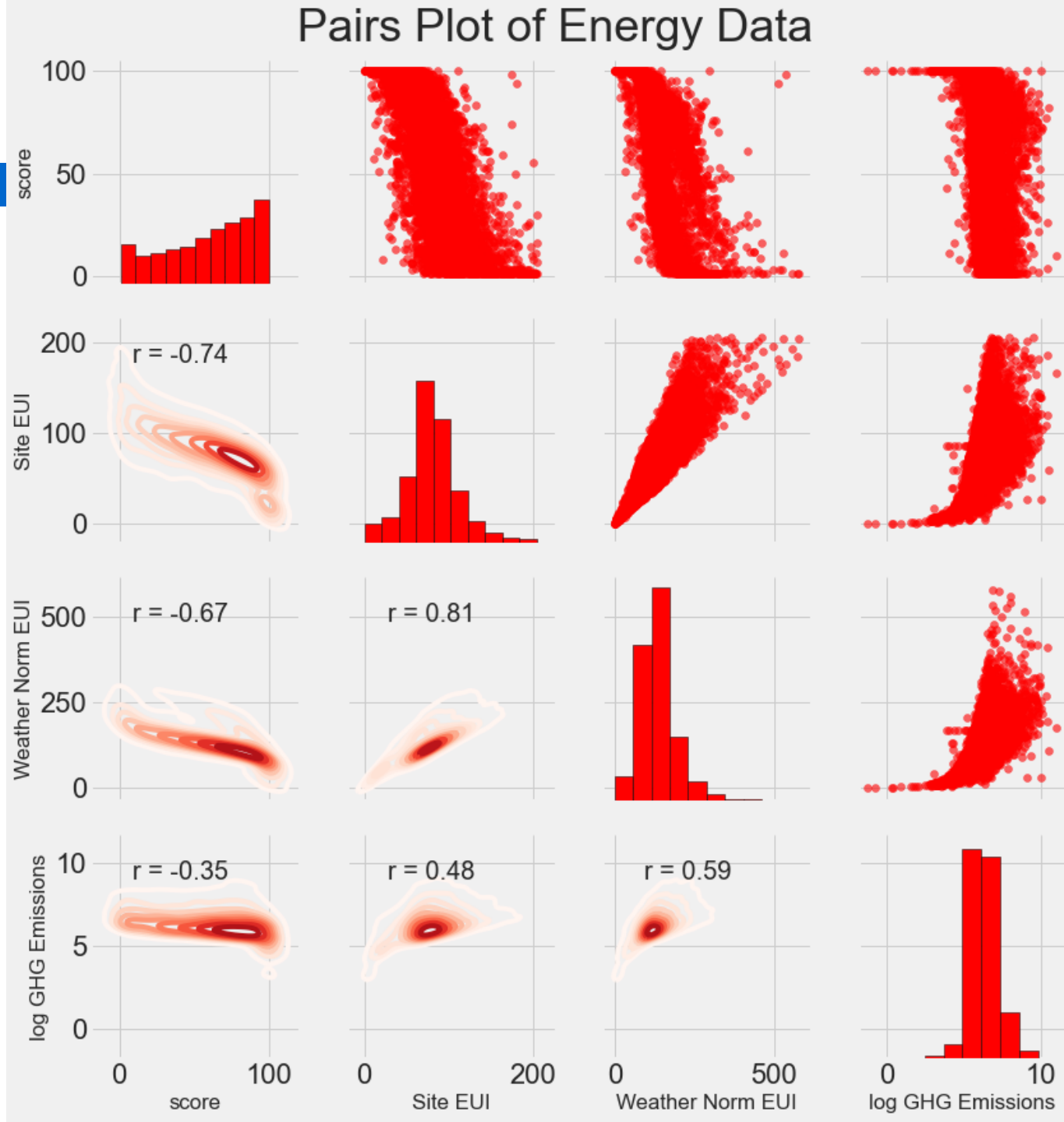
16

现在我们有了正确的列数据类型，我们可以通过**查看每列中缺失值的百分比来开始分析**。当我们进行探索性数据分析时，缺失的值很好，但是必须使用机器学习方法进行填写。

探索性数据分析(EDA)

17

Pairs Plot是一次检查多个变量的好方法，因为它显示了对角线上的变量对和单个变量直方图之间的散点图。



3.特征工程

18

01 机器学习项目流程概述

02 数据清洗

03 特征工程

04 数据建模

3.特征工程

19

特征工程和特征选择

•**特征工程**: 获取原始数据并提取或创建新特征的过程。这可能意味着需要对变量进行变换, 例如自然对数和平方根, 或者对分类变量进行one-hot编码, 以便它们可以在模型中使用。一般来说, 我认为特征工程是从原始数据创建附加特征。

•**特征选择**: 选择数据中最相关的特征的过程。在特征选择中, 我们删除特征以帮助模型更好地总结新数据并创建更具可解释性的模型。一般来说, 特征选择是减去特征, 所以我们只留下那些最重要的特征。

3.特征工程

20

特征工程

特征工程在数据挖掘中有举足轻重的位置数据领域一致认为：
数据和特征决定了机器学习的上限，而模型和算法只能逼近这个上限而已。

特征工程重要性：

特征越好，灵活性越强； 特征越好，模型越简单； 特征越好，性能越出色； 好特征即使使用一般的模型，也能得到很好的效果！

主要方法
离散型变量处理
分箱/分区
交叉特征
特征缩放
特征提取
.....

3.特征工程

21

特征选择

特征选择主要有两个功能：

- 1.减少特征数量、降维，使模型泛化能力更强，减少过拟合
- 2.增强对特征和特征值之间的理解

主要方法

去除变化小的特征

去除共线特征

去除重复特征

主成分分析 ([PCA](#))

.....

3.特征工程

22

数据划分



不考虑时间因素，通常打乱数据



时间序列



4.数据建模

23

01 机器学习项目流程概述

02 数据清洗

03 特征工程

04 数据建模

数据建模

24

- 基于性能指标比较几种机器学习模型
- 对最佳模型执行超参数调整
- 在测试集上评估最佳模型
- 解释模型结果
- 得出结论

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
0	Extreme Gradient Boosting	2666.8675	22721899.5305	4764.4228	0.8410	0.4428	0.3151	0.0429
1	Gradient Boosting Regressor	2671.5927	23019681.2661	4794.6037	0.8393	0.4439	0.3143	0.0683
2	CatBoost Regressor	2814.6048	24757340.4659	4973.7765	0.8265	0.4734	0.3427	1.1286
3	Random Forest	2779.2026	25351757.1506	5032.2587	0.8218	0.4816	0.3432	0.2087
4	Light Gradient Boosting Machine	3018.9895	25515012.3051	5049.8492	0.8192	0.5534	0.3876	0.0815
5	Extra Trees Regressor	2755.9265	28180447.2658	5299.6566	0.8043	0.4875	0.3255	0.1496
6	AdaBoost Regressor	4366.1001	29298215.0087	5411.0606	0.7915	0.6478	0.7662	0.0195
7	Ridge Regression	4339.6093	38542499.6202	6196.4891	0.7343	0.6348	0.4429	0.0036
8	Bayesian Ridge	4343.5006	38542310.2536	6196.4607	0.7343	0.6405	0.4436	0.0058
9	Linear Regression	4332.7658	38549952.0026	6197.0842	0.7343	0.6369	0.4415	0.0043
10	Lasso Regression	4332.6327	38543897.4692	6196.6074	0.7343	0.6404	0.4416	0.0038
11	TheilSen Regressor	4124.3658	38946435.2631	6224.8917	0.7327	0.5337	0.3743	0.7617
12	Least Angle Regression	4323.4578	40017870.2286	6312.0115	0.7250	0.5647	0.4242	0.0062
13	Lasso Least Angle Regression	4322.4466	40023599.4550	6312.5498	0.7249	0.5401	0.4245	0.0060
14	Decision Tree	3184.9728	44561182.4569	6663.2248	0.6826	0.5343	0.3523	0.0050
15	Huber Regressor	3478.8635	49170605.5859	6997.8228	0.6590	0.4873	0.2212	0.0472
16	Random Sample Consensus	3467.4036	52056856.3074	7203.0774	0.6382	0.4970	0.2175	0.0831
17	Orthogonal Matching Pursuit	5760.0475	57656797.2076	7580.0224	0.6026	0.7426	0.8996	0.0034
18	Passive Aggressive Regressor	4817.2726	59211213.8323	7652.9707	0.5928	0.7577	0.4334	0.0068
19	Elastic Net	6399.4702	72811792.6577	8506.2813	0.5021	0.6789	0.8016	0.0030
20	K Neighbors Regressor	6858.1227	105272520.3363	10228.2497	0.2784	0.7524	0.7450	0.0036
21	Support Vector Machine	8401.1273	163965107.0052	12732.6249	-0.1092	0.9303	1.0323	0.0308

1. <https://github.com/WillKoehrsen/machine-learning-project-walkthrough>

谢谢!

