数据集

1. 信用卡欺诈检测

数据集包含2013年9月欧洲持卡人通过信用卡进行的交易

<https://www.kaggle.com/mlg-ulb/creditcardfraud>

字段展示

time 每个事务与数据集中第一个事务之间的秒数

V1、V2、...V28 使用PCA获得的数值

Amount 事物金额

Class 发生欺诈时值为1，否则为0

样本数据

图片包含 应用程序

描述已自动生成

2.用于欺诈检测的合成财务数据集Synthetic Financial Datasets For Fraud Detection

<https://www.kaggle.com/ealaxi/paysim1>

字段展示

Step 映射真实世界中的时间单位。在这种情况下，一步是一小时的时间。总步骤744（30天模拟）。

type 交易类型

amount 以当地货币表示的交易金额

nameOrig 交易的客户

oldbalanceOrg交易前的初始余额

newbalanceOrg交易后的余额

nameDest 交易的收款人ID

oldbalanceDest交易前收款人的初始余额。没有以M（商家）开头的客户信息

NewBalanceTest 交易后的收款人余额。没有以M（商家）开头的客户信息

isFraud 这是模拟中欺诈代理进行的交易。在这个特定的数据集中，代理的欺诈行为旨在通过控制或客户账户获利，并试图通过转移到另一个账户然后从系统中套现来清空资金（识别欺诈交易）

isFlaggedFraud 该业务模式旨在控制从一个帐户到另一个帐户的大规模转账，并标记非法尝试。此数据集中的非法尝试是试图在单个事务中传输超过200.000的数据（识别欺诈交易）

样本数据



# 3.信用卡交易欺诈检测数据集Credit Card Transactions Fraud Detection Dataset

# https://www.kaggle.com/kartik2112/fraud-detection

# 这是一个信用卡交易数据集，包含2019年1月1日至2020年12月31日期间的合法和欺诈交易。它涵盖了1000名客户与800名商家进行交易的信用卡

# 字段展示

# index 索引

trans*date*trans\_time 交易日期时间

cc\_num 客户信用卡号码

merchant 商家名称

category 商家类别

amt 交易金额

first 持卡人名字

last 持卡人姓氏

gender  持卡人性别

street 持卡人所居住的街道

city 持卡人居住的城市

state  持卡人居住的州

zip 邮区号（位置信息）

lat 持卡人维度

long  持卡人经度

city\_pop城市人口

job工作

dob 出生日期

trans\_num交易编号

unix\_time时间戳

merch\_lat商家纬度

merch\_long 商家经度

is\_fraud欺诈标志（目标类）

# 样本数据

# 图形用户界面 描述已自动生成

# 图形用户界面, 应用程序 描述已自动生成

# 4.来自金融支付系统的合成数据Synthetic data from a financial payment system

# <https://www.kaggle.com/ealaxi/banksim1>

# 1）字段展示

# Step步幅

# Customer顾客

# Age年龄

# Gender性别

# zipcodeOri邮政编码

# merchant商家

# zipMerchant商家邮政编码

# category交易类别

# amount金额

# fraud是否欺诈

# 样本数据

# 图片包含 窗户, 窗帘, 游戏机, 大 描述已自动生成

# 2）字段展示

# Source付款人

# Target收款人

# Weight金额

# Typetrans交易类型

# Fraud是否欺诈

# 样本数据

# 表格 描述已自动生成

# 5.伦敦自行车共享系统

# <https://www.kaggle.com/edenau/london-bike-sharing-system-data?select=stations.csv>

# 图形用户界面, 文本, 应用程序, 电子邮件 描述已自动生成

# 字段展示

Journey Duration骑行时间

Journey ID骑行ID

End Date结束日期

End Month结束月份

End Year结束年份

End Hour结束小时

End Minute结束分钟

End Station ID还车车站ID

Start Date开始日期

Start Month开始月份

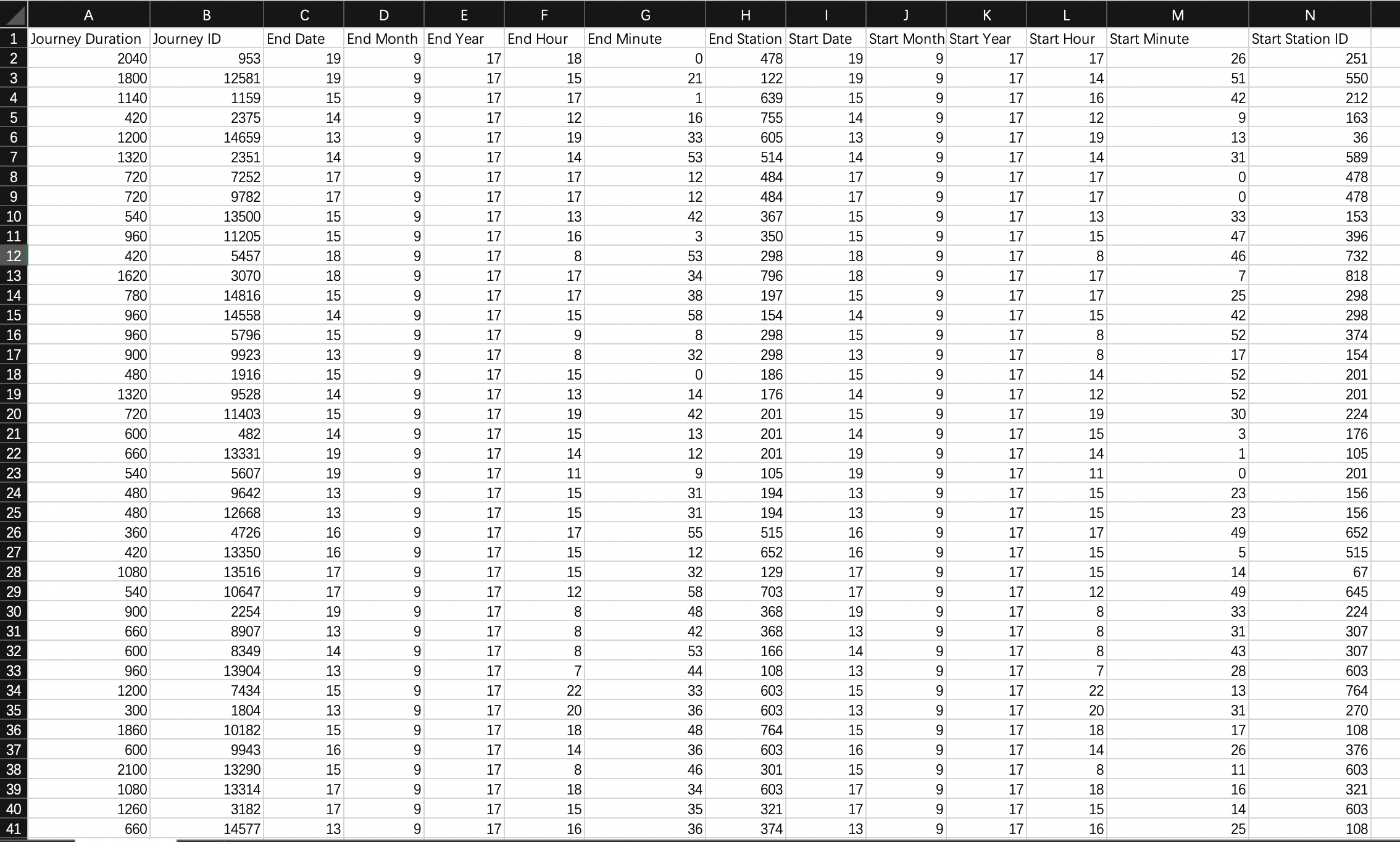
Start Year开始年份

Start Hour开始小时

Start Minute开始分钟

Start Station ID借车车站ID

样本数据



# 车站信息

# 

# 6.美国交通事故数据集US-Accidents: A Countrywide Traffic Accident Dataset

# <https://smoosavi.org/datasets/us_accidents>

提供了2016年2月至2020年12月收集的数据。在这个数据集中，大约有**1亿**起交通事故

字段展示

ID事故记录的唯一标识符

Severity显示事故的严重程度，数字介于1到4之间。

Start\_Time当时时区的事故开始时间

End\_Time当时时区的事故结束时间，指事故对交通流量影响消失的时间

Start\_Lat起点的GPS坐标纬度

Start\_Lng起点的GPS坐标经度

End\_Lat终点的GPS坐标纬度

End\_Lng终点的GPS坐标经度

Distance(mi) 受事故影响的道路长度

Description事故的描述

Number地址的街道编号

Street地址的街道名称

Side地址中街道的相对一侧（左/右）

City城市

County县

State州

Zipcode邮政编码

Country国家

Timezone事故地点（东部、中部等）显示时区

Airport\_Code表示距离事故地点最近的机场气象站

Weather\_Timestamp天气观测记录的时间戳（当地时间）

Temperature(F)温度（华氏度为单位）

Wind\_Chill(F)风寒（华氏度为单位）

Humidity(%)湿度

Pressure(in)气压

Visibility(mi)可见度

Wind\_Direction风向

Wind\_Speed(mph)风速

Precipitation(in)降水

Weather\_Condition天气状况

Amenity附近位置存在便利设施

Bump附近存在减速

Crossing附近存在十字路口

Give\_Way附近存在让路

Junction附近存在交叉路口

No\_Exit附近禁止通行

Railway附近存在铁路

Roundabout附近存在环岛

Station附近存在车站

Stop附近有刹车

Traffic\_Calming附近是否交通平静

Traffic\_Signal附近是否有交通信号

Turning\_Loop附近是否有转弯环

Sunrise\_Sunset根据日出/日落显示白天或夜晚的时段

Civil\_Twilight（时间）

Nautical\_Twilight（时间）

Astronomical\_Twilight（时间）

样本数据

图片包含 窗户, 游戏机, 大

描述已自动生成

表格

描述已自动生成

# 表格 描述已自动生成

# 7.主要城市的日温度Daily Temperature of Major Cities

# https://www.kaggle.com/sudalairajkumar/daily-temperature-of-major-cities

# 字段展示

# Region地区

Country国家

City城市

Month月份

Day日期

Year年份

AvgTemperature平均气温

# 样本数据

# 表格 描述已自动生成

# 8.结束暴力推文ENDviolence Tweets

# https://www.kaggle.com/mathurinache/endviolence-tweets

# 字段展示

user\_name用户名

user\_location用户位置

user\_description用户描述

user\_created用户创建日期

user\_followers用户粉丝

user\_friends用户朋友

user\_favourites用户收藏夹

user\_verified用户验证

date发文日期

text文本

hashtags标签

source客户端

is\_retweet是否转发

# 样本数据

# 表格 描述已自动生成

# 9.南美洲的新冠疫情Covid in South America - Latest Data

# <https://www.kaggle.com/anandhuh/covid-in-south-america-latest-data>

数据集包含截至2021年11月15日的南美洲新冠病毒数据

# 字段展示

**Country** 南美洲国家、领土和岛屿的名称

**Total Cases** 病例总数

**Total Deaths** 死亡人数

**Total Recovered**康复病例总数

**Active Cases** 现有病例总数

**Total Cases/1 mil population每百万人口总病例**

**Deaths/1 mil population** 每百万人口死亡人数

**Total Tests** 完成新冠检测总人数

**Tests/1 mil population**每百万人进行新冠病毒检测人数

**Population** 人口

# 样本数据

# 表格 描述已自动生成

# 10.美国自行车共享数据US Bikeshare Data

# https://www.kaggle.com/walidsayed/us-bikeshare-data?select=chicago.csv

# 字段展示

# Start time 开始时间

# End time结束时间

# Trip duration使用时间

# Start station出发站

# End station结束站

# User type用户类型

# Gender性别

# Birth year出生年份

# 样本数据

# 表格 描述已自动生成

# 11．纽约机动车与人碰撞 NYC Motor Vehicle Collisions to Person

# https://www.kaggle.com/kukuroo3/nyc-motor-vehicle-collisions-to-person

# 字段展示

* CRASH\_DATE 碰撞日期
* CRASH\_TIME 碰撞时间
* PERSON\_INJURY 人受伤情况
* PERSON\_AGE 年龄
* BODILY\_INJURY 受伤的身体区域
* SAFETY\_EQUIPMENT使用的安全设备（如安全带、安全气囊等）
* PERSON\_SEX 性别
* PERSON\_TYPE 骑自行车的人、机动车乘客、行人
* PED\_LOCATION 行人的位置（是否在十字路口）
* CONTRIBUTINGFACTOR2导致车辆碰撞的因素
* EJECTION表示以下情况：未弹出、部分弹出或从车辆中弹出
* COMPLAINT身体受伤类型（如脑震荡、严重烧伤、严重出血等）
* EMOTIONAL\_STATUS明显死亡、无意识、半意识等
* VEHICLE\_ID车辆ID
* PERSON\_ID系统分配的人员识别码
* CONTRIBUTINGFACTOR1导致指定车辆碰撞的因素1
* POSITIONINVEHICLE 座位位置（即司机、前排乘客等）
* PED\_ROLE目击者
* UNIQUE\_ID系统生成的记录码
* PED\_ACTION 车祸时行人在做什么
* COLLISIONID 车祸识别码

# 样本数据

# 图形用户界面, 应用程序, 表格, Excel 描述已自动生成

# 12. 电影数据库

# rotten tomatoes movie database

# <https://www.kaggle.com/ayushkalla1/rotten-tomatoes-movie-database>

# 字段展示

# Cast1-6:演员1-6

# Description：电影描述

# Director1-3:导演1-3

# Genre：电影类型

# Rating：评级

# Release date：发布日期

# Runtime：时长

# Studio：工作室

# Title：电影标题

# Writer1-4:作家1-4

# Year：电影上映年份

# 样本数据

# 

# 13.芝加哥Airbnb开放数据

# https://www.kaggle.com/jinbonnie/chicago-airbnb-open-data?select=listings.csv

# 字段展示

### Columns Information

* id – airbnb的ID
* name – adirbnb的名称
* host\_id – 主机ID
* host\_name – 主机名称
* neighborhood – Airbnb所属地区
* latitude – 位置纬度
* longitude – 位置经度
* room\_type – 房子类型（整套、公寓、私人房间或其他）
* price – 公寓每天价格
* minimum\_nights – 你需要预定的最少夜晚
* numberofreviews – 数据集的评论总数
* last\_review – 上次评论时间
* reviewspermonth – 每月评论
* calculated\_host\_listings\_count – 主机的总列表号
* availability\_365 – 可用天数

# 样本数据

# 图形用户界面, 应用程序, 表格, Excel 描述已自动生成

# 14. Bay Area Bike Sharing Trips

# https://www.kaggle.com/datasets?search=DC+area

# 字段展示

Month骑行月份

trip\_duration\_sec骑行时间（秒）

start\_station\_id出发站ID

start\_station\_name出发站名称

end\_station\_id结束站ID

end\_station\_name结束站名称

bike\_id车辆ID

user\_type用户类别

member\_birth\_year出生年份

member\_gender性别

# 样本数据

# 图形用户界面, 应用程序, 表格, Excel 描述已自动生成

# 15.权利的游戏剧集数据

# https://www.kaggle.com/dasbootstrapping/game-of-thrones-episode-data

# 字段展示

# Season季别

# Episode Number剧集编号

# Number in Season所在季别的剧集编号

# Episode Name剧集名称

# Director导演

# Writer作家

# Original Air Date首播日期

# US viewers (million)美国观众（百万）

# Runtime (mins)时长

IMDB Description描述

IMDB votes投票

Imdb Rating评级

Notable Death Count显著死亡（其中每颗子弹算作1例显著死亡）

# 样本数据

# 图形用户界面, 应用程序, 表格, Excel 描述已自动生成

# 16.心力衰竭预测数据集

# https://www.kaggle.com/fedesoriano/heart-failure-prediction

# 字段展示 Age: 病人年龄

Sex: 患者的性别[M：男性，F：女性]

ChestPainType: 胸痛类型[TA：典型心绞痛，ATA：非典型心绞痛，NAP：非心绞痛，ASY：无症状]

RestingBP: 休息血压[mm汞]

Cholesterol: 血清胆固醇[mm/dl]

FastingBS: 空腹血糖[1: if FastingBS > 120 mg/dl, 0: otherwise]

RestingECG: 静息心电图

MaxHR: 达到最高心率[数值在60到202之间]

ExerciseAngina: 运动诱导心绞痛[Y：是，N：否]

Oldpeak: oldpeak = ST [在低落时测量的数值]

ST\_Slope: 峰值运动ST段的斜率[向上：向上倾斜，平坦：平坦，向下：向下倾斜]

HeartDisease: 输出类[1：心脏病，0：正常]

# 样本数据

# 图形用户界面, 应用程序, 表格, Excel 描述已自动生成

# 17.丙型肝炎预测数据集

# <https://www.kaggle.com/fedesoriano/hepatitis-c-dataset>

# 字段展示

1) X（患者ID/编号）  
2) category诊断类别

3）age（以岁为单位）  
4）sex（f，m）  
属性5至14参考实验室数据：  
5）ALB  
6）ALP  
7）ALT  
8）AST  
9）BIL  
10）CHE  
11）CHOL  
12）CREA  
13）GGT  
14）PROT

# 样本数据

# 

# 18.身体脂肪预测数据集

# https://www.kaggle.com/fedesoriano/body-fat-prediction-dataset

# 字段展示

1. Density determined from underwater weighing水下称重确定的密度
2. Percent body fat from Siri's (1956) equation身体密度百分比
3. Age (years)年龄
4. Weight (lbs)体重（磅）
5. Height (inches)身高（英寸）
6. Neck circumference (cm)颈围
7. Chest circumference (cm)胸围
8. Abdomen 2 circumference (cm)腹部2周长
9. Hip circumference (cm)臀围
10. Thigh circumference (cm)腿围
11. Knee circumference (cm)膝盖周长
12. Ankle circumference (cm)脚踝周长
13. Biceps (extended) circumference (cm)二头肌周长
14. Forearm circumference (cm)前臂周长
15. Wrist circumference (cm)手腕周长

# 样本数据

# 

# 19.1993-2021全球海平面数据

# <https://www.kaggle.com/kkhandekar/global-sea-level-1993-2021>

# 字段展示

Col1 - Year  
Col2 - Total Weighted Observations 加权观测值  
Col3 - GMSL (Global Isostatic Adjustment (GIA) not applied) variation (mm) with respect to 20-year TOPEX/Jason collinear mean reference   
Col4 - Standard Deviation of GMSL (GIA not applied) variation estimate (mm)  
Col5 - Smoothed (60-day Gaussian type filter) GMSL (GIA not applied) variation (mm)    
Col6 - GMSL (Global Isostatic Adjustment (GIA) applied) variation (mm) with respect to 20-year TOPEX/Jason collinear mean reference   
Col7 - Standard deviation of GMSL (GIA applied) variation estimate (mm)  
Col8 - Smoothed (60-day Gaussian type filter) GMSL (GIA applied) variation (mm)  
Col9 - Smoothed (60-day Gaussian type filter) GMSL (GIA applied) variation (mm); annual and semi-annual signal removed

# 样本数据

# 图形用户界面, 应用程序, 表格, Excel 描述已自动生成

# 20.流量预测数据集

# https://www.kaggle.com/fedesoriano/traffic-prediction-dataset

# 字段展示

# DataTime时间

# Junction路口数

# Vehicles一小时的车辆数量

# ID

# 样本数据

# 图形用户界面, 应用程序, 表格, Excel 描述已自动生成