

美国 2018 年商业航班准点表现抽样数据

一、 数据来源

美国交通部运输统计局: https://www.transtats.bts.gov/Fields.asp?Table_ID=236

二、 数据抽样

该数据集统计局统计了自 1987 年来到现在（截止 2019 年 8 月份）的美国航班飞行数据，包括起降时间、出发地情况、到达地情况、飞机情况、准点表现等多个指标。数据集中最近完整年度 2018 年共有近 700 万条数据，涉及 18 家航空公司，5000 多架飞机。为简化研究问题，本次数据抽样共随机抽取 100 架飞机，抽样后共得到 127143 条数据。

三、 数据介绍

每条数据为一趟航班信息记录，共选取 29 个指标，具体指标介绍如下：

因变量：

- ArrDelay: 到达延误（以分钟计），提前到达该指标值为负；

其他指标：

- Year: 年份，数据集选取时间为 2018 年；
- Month: 月份；
- Day_of_Month: 具体为月份中哪一天；
- Day_of_Week: 表示星期几；
- Op_Unique_Carrier: 航空公司代码，具体代码对应航空公司在 `L_UNIQUE_CARRIERS.csv` 文件中，可另行查看；
- Tail_Number: 对应每一架飞机；
- Op_Carrier_Fl_Num: 航班编号；
- Origin: 出发机场代码，具体代码对应机场在 `L_AIRPORT_Origin.csv` 文件中，可自行查看；
- Dest: 到达地机场代码，具体代码对应机场在 `L_AIRPORT_Dest.csv` 文件中，可自行查看；
- CRSDepTime: 计划起飞时间，为当地时间，数据格式为(hhmm)；
- DepTime: 实际起飞时间，为当地时间，数据格式为(hhmm)；
- DepDelay: 起飞延误（以分钟计），提前起飞该指标值为负；
- TaxiOut: 飞机起飞前滑行时间（以分钟计）；
- CRSArrTime: 计划降落时间，为当地时间，数据格式为(hhmm)；
- ArrTime: 实际降落时间，为当地时间，数据格式为(hhmm)；

- TaxiIn: 飞机降落后滑行时间（以分钟计）；
- Cancelled: 数值为 1 表示航班取消，0 则没有取消；
- CancellationCode: 取消原因，共有 A/B/C/D 四种，具体见 L_CANCEL-LATION.csv 文件；
- Diverted: 数值为 1 表示航班绕道，0 则没有；
- CRSElapsedTime: 系统预定飞行时长（以分钟计）；
- ActualElapsedTime: 系统实际飞行时长（以分钟计）；
- AirTime: 在空中时长（以分钟计）；
- Distance: 始末机场距离，单位为英里；
- CarrierDelay: 因航空公司延误（以分钟计）；
- WeatherDelay: 因天气原因延误（以分钟计）；
- NASDelay: 因国家航空系统延误（以分钟计）；
- SecurityDelay: 因安全问题延误（以分钟计）；
- LateAircraftDelay: 因相同飞机在前一机场晚点到达造成的延误（以分钟计）；

四、 分析任务

请以判断飞机是否到达延误为研究目的（延误则取值为 1，否则取为 0，指标计算以 ArrDelay 为标准），对本数据集进行分析。