2020中国数据马拉松大赛

决赛说明及指导

欢迎您参加2020中国数据马拉松大赛！在此文件中，我们会向您详细阐述此次中国数据马拉松大赛的主题, 比赛中将会使用的数据集概述，以及赛事指导和建议。

## 背景

（DELAG）Deutsche Luftschiffahrts -Aktiengesellschaft 成立于 1909 年11月，作为第一家客运航空公司，当时的DELAG在政府的协助下开始运营由齐柏林飞艇公司制造的飞艇。随后，虽然更多的航空公司成立了，但是直到第二次世界大战左右，航空公司才开始大量投资包括客运和货运在内的民用航空运输。

如今，航空业已成为世界经济的重要引擎。尽管经历了动荡和危机，航空公司仍然是经济的重要支柱。航空出行不仅有益于全球经济合作，也使人们得以感受截然不同的山川和文化，带来了终身难忘的回忆和欢乐。

商业旅行的晴雨表通常可以用作消费者可自由支配支出和休闲活动的主要指标。 随着世界之间的联系越来越紧密，我们找到了更好的方式来研究个中关联，并将其可视化——航空业一直都是人们活动和兴趣的热点。

## 任务

**目标：**

根据2017年美国商业航空公司的航班流量数据 （如下所述） ，分析商业航空公司行业发展与消费者大趋势和全球重大事件之间的联系。

我们已经整理了一些补充数据集供参考使用，包括额外的商业航空公司旅行数据、航空公司乘客票价和股票价格的数据、2017年美国重大事件的信息以及来自于美国机场的6小时天气数据。

**你最应该关注的：**研究的创新性，数据分析的质量，你的洞察力及研究的深度。

**你具体的任务：**提出自己的问题（即研究的方向），分析现有的数据集（请参看第3部分“数据集”）来回答你设立的问题。**您不需要太全面，在提出的问题范围内，质量数据分析更为重要**。

你提交的报告可以是预测性的，例如通过机器学习和（或）时间序列分析来预测或建模航空旅行趋势。你提交的报告也可以是启发性的，例如提供可视化数据来展示你的发现，或者用统计检验来验证你的结论。总之，这是一场开放式的竞赛，考验你是否有独立思考和实践自己想法的能力。

可以研究以下示例问题之一 ，或者创建自己的研究题目。 鼓励创造性地提出自己的问题，**然而，创新立题也应注重分析的深度、精确度和严谨性，创新和严谨同等重要。**

**示例问题1**：航空公司的总体飞行模式（例如，流量，目的地选择）与该航空公司的财务/存量表现如何相关？ 是否可以发现特定趋势以区分表现最佳的公司与表现最差的公司？

**示例问题2**：天气的严重程度与对航空公司航班延误的实际影响有何关系？ 是否存在天气严重程度的断点，在该断点处航班受到的影响更大？

**示例问题3**：延迟/取消模式是否会完全影响库存/财务绩效？ 航空公司在天气差于平均水平的季度中的财务表现如何？

**示例问题4**：美国国内重大事件如何影响航空运输和旅客票价模式？

## 数据集

提供的数据集将分布在7个表格中。你应该根据所设立的问题来选择相关的表格进行分析。**请在比赛前了解下面的数据集概述**，这样可以帮助你预判你在比赛中可能会使用哪些表格，可能会需要的分析，以及可能会需要建立的模型等。

注：每张表中我们都会标注出处，你也可以使用原始数据，但我们强烈建议你使用我们提供的数据集，因为我们已经初步整理和清洗过这些原始数据。

**3.1 主要数据**

我们提供了以下数据：

* *航空公司（****airlines****）*

航空公司ID到名称的映射。

61行 & 2列。 大小： 约0.1MB。来源：互联网。

* *机场（****airports****）*

美国各机场的重要详细信息（名称，州，标识符，纬度，经度等）。

322行 & 6列。 大小： 约0.1MB。来源：美国运输部（[US Department of Transportation](https://www.transportation.gov/)）。

* *事件（****events\_US****）*

2017年全年，美国各地的公共事件。

1,151行 & 4列。 大小： 约0.1MB。来源：[Shore Fire Media](https://shorefire.com/)。

* *票价（****fares****）*

2017年每个季度、航线、航空公司组合的票价分布，价格为10美元。

80,823行 & 255列。 大小： 约44MB。来源：美国运输部（[US Department of Transportation](https://www.transportation.gov/)）。

* *飞行交通（****flight\_traffic****）*

有关2017年美国国内航班延误的信息。

约600,000行 & 24列。 大小： 约130MB压缩版，约550MB正常版。来源：美国运输部的交通统计局（[Bureau of Transportation Statistics](https://www.bts.gov/)）。

* *股票价格（****stock\_prices****）*

从2016年末到2018年初，美国多家航空公司的每日收盘价。

380行 & 10列。 大小： 约0.1MB。来源：[Alpha Vantage](https://www.alphavantage.co/)。

* *天气（****weather****）*

2017年，每6小时在美国各机场收集的天气数据（温度，风，降水，云量等）

353,864行 & 12列。 大小： 约48MB。来源：美国国家环境信息中心（[National Centers for Environmental Information](https://www.ncei.noaa.gov/)）。

**3.2 数据表格模式概要**

**请阅读第7部分的“数据表格模式概要”**，每一个文件都会有一个相应的格式说明。

**3.3准备和设置**

* 我们**强烈建议你安装 Jupyter Notebook** （<http://jupyter.org/install.html>），它在数据工程、探索和建模方面非常好用。Jupyter Notebook 提供交互式、实时的开发环境，因此不会像标准版的“terminal + text editor”环境那样产生许多痛点，而且它与 Python和R都兼容。
* **请确保您的电脑有至少1GB的可用内存**。否则，您的电脑将很难运行我们提供的数据集。

**3.4数据陷阱**

尽管我们已经对数据进行了一些整理和清洗，但仍有一些陷阱需要注意：

* 你需要自行决定如何处理缺失值、异常值、错误值和非规范化字段等，或者如何进行原始数据的转换，如何根据研究方向做出一些合理的决定等。
* 有时候你需要通过查找一些相关资料，来了解变量背后的一些背景知识，例如对美国一些特定区域的了解。这可能会影响你随后如何整理和使用数据。
* 提供的数据集中可能没有足够的某种类型的数据点，例如美国某些地区的数据点很少甚至没有。你将不得不根据要回答的问题来决定如何解决，或者处理数据不足的缺陷。

## 提交要求

**4.1 提交内容**

**提交的内容必须包括报告、相关代码、短视频（硬性要求）**。您也可以提交补充材料（附录）。

1. **报告（限于15页）**：报告分两部分，非技术性阐述和技术性阐述。如果摘要统计和数据可视化可以帮助你更好解释你的观点，我们会非常建议你使用。另外，**必须**在报告中回答以下问题：
   1. **非技术性阐释**

* 您研究的问题是什么？
* 您主要发现或结论是什么？
* 为什么这些发现或结论很重要？
  1. **技术性阐释**
* 您使用的方法论/方法是什么？
* 请描述您的数据处理和探索过程。
* 请描述您的分析和建模步骤。

1. **代码**：**请提交用于生成你研究结果的相关代码。**虽然你的代码不会被打分，但是你必须提交代码，否则你的提交不予接受，即自动放弃比赛。
2. **视频（限于5分钟）**：请录制自己的视频，其中包括：

* 个人介绍–请告诉我们您是谁以及为什么您参加中国数据马拉松
* 报告摘要–请简要阐述您的论题，发现和结论。

4）**补充材料（非硬性要求）**

补充材料应该是进一步能帮助展示你的洞察力、深度和创新的，或者更清晰地解释结论，否则补充材料就没有实际意义。比如，分析过程中遇到的障碍、对于报告中某一处进行补充说明、下一步想要研究的领域、或者分析过程中失败、错误的分析方法等。**补充材料请放于报告后的附录中。**

* 1. **格式**

1. **报告**

* 具体形式不限，但必须是通用的形式，比如HTML、PDF、PPT、网页链接等。**请确保不需要通过专用的软件来打开。**
* 你可以使用你偏好的工具来生成报告，比如Python Notebook、Shiny Application、Microsoft Office 等。但是，你的报告必须是上述提到的通用形式。例 如，如果用 Python Notebook，那么导出的报告应该是 HTML 格式。如果用 Shiny Application，那么报告应该是发布在一个可访问的网页链接上。
* **请提交用于生成报告的源文件**。例如，如果你提交的PDF报告包含MathType、方程式或特殊符号，那么你就应该提交一份 LaTeX 源文件。

**2) 代码**

**将代码打包成一个单独的压缩文件包，并和报告分开。**

1. **视频**

**本次比赛由HRTPS提供在线视频录制工具（小程序）。**

**请查收您的邮箱接收视频工具链接。**

**4.3 截止时间**

**必须在7月6日周一早上8:00 之前提交。**

**4.4 提交方法**

**详情请见决赛群内公告。**

## 评估标准

|  |  |  |
| --- | --- | --- |
| 非技术性阐释 | 你的目标是清楚定义你所研究的问题，阐述关键论点及其意义，汇总统计数据和可视化效果。一个好的阐述是能非常精准、准确和清晰地表达你的见解。 | |
| **评分内容** | **评分标准** |
| **见解深刻** | 是否挖掘更深层意义的结论或超出预期的观察 |
| **内容相关** | 得出的结论是否与本次大赛出题相关 |
| **表达精准** | 是否得出非常精准详细而非宽泛空洞的结论 |
| **描述清晰** | 是否有清楚表达所研究问题及结论 |
| 技术性阐释 | 您的目标是解释研究的方法论/方法，描述数据处理和探索的过程，以及展示分析和建模的步骤。在合适的地方强烈建议使用可视化数据。一个好的技术性阐述表现为，不借助任何其他解释，别人就能非常清楚地读懂你的过程和结论。 | |
| **评分内容** | **评分标准** |
| **数据整理和**  **工程处理** | 1. 进行了适当的数据质量控制，例如缺失值、异常值、错误值和非规范化字段的处理。 2. 如何对不同字段进行转化（如有），从而实现数据的结合使用，总之需要证明你使用的非常规数据处理的技巧。 3. 讨论特征工程的性能以及解释你为何要选择这些特征。 |
| **调查深度** | 1. 多步骤的探索性数据分析-每一步是否有数据可视化，解释如何得出结论及为何使用，并如何帮助得出后续的决定。 2. 生成和检验你的假设，并推断结论。 3. 能较好地汇总分析结果，而不仅仅是罗列不同的分析。 |
| **分析与模型**  **的严谨性** | 1. 定量与定性分析的呈现同样重要，比如异常值、残差、中介变量/工具变量的分析等。 2. 做出假设和选择的逻辑合理，特别是特征选择的过程。 3. 能够对模型做出性能与缺点的分析。 4. 解释对所用的可视化与统计数据的原因及所得结论。 |
| **创造力** | 这部分内容是加分项，但是我们也不希望你本末倒置，花太多时间追求创造性，而牺牲了你本次大赛的主要目标-即在短时间内确立一个研究话题并通过数据分析支持你的结论。以下是几个展现创造力的例子：   1. 一些很有意思的外部数据集。 2. 能够与一个完全不同的情况进行非常有逻辑地类比。 3. 除上述所有列举的情况外，展现出的特殊技能。 |

## 6. 指导与建议

**6.1 时间分配**

您将有60个小时来进行决赛。时间线如下：

**比赛开始：2020年7月3日，晚上8:00**：您将收到问题说明和数据集的副本。

**比赛结束：2020年7月6日，早上8:00**：此时您必须以电子方式提交所有可交付成果（书面报告，代码和视频）。

主要的步骤：

* 研究此份比赛说明，清楚比赛规则。决定哪些您感兴趣的研究方向，并且确定这几个研究方向的排序。
* 研究数据集概述并形成您的数据分析策略，计划您数据探索的过程。
* 大部分时间应该用在这几个方面：研究数据，进行定性和定量分析，并撰写你的过程和结果。
* 计划何时开始写您的报告。**友情建议：不要晚于7月5日12点才开始撰写报告。**
  1. **如何确立一个好的研究题目**

不可否认，这部分非常具有挑战性，事实上每位选手都花了许多精力和时间来确立研究题目。但大多数选手都在比赛结束后反馈，花一定时间确立题目是非常有必要的，因为如果题目设计的合理，剩下的部分相对更偏实际操作和解决问题。我们建议你利用比赛的第一晚上的时间来确立一些潜在的题目，然后在开始验证你的假设，进而确定你的最终题目。

我们建议你：

1. 详细研究我们提供的“数据表格模式概要”，来帮助你识别不同数据字段之间的潜在关系，并探索如何将整体数据细分为更具体的层次。
2. 在网上搜索有关美国的一些背景信息，帮你了解事件的关键问题是什么。
3. 对数据执行初始的探索性数据分析（摘要统计、数据可视化和简单回归等），来帮你验证题目至关重要的某些假设是否是正确的。如果是正确的，那么你可以继续进行下一步操作；如果不正确，你应该考虑放弃这个题目并进入下一个题目的验证。

**去年得奖选手的建议：**

1. 不要忘记你独特的背景。
2. 尽可能不要尝试学习新工具，学会克服“临时抱佛脚”的心态；相反，你应该充分利用已掌握的技能，思考如何使用已学的知识尽可能从数据中提取有用的信息和发现。
3. “花一定的时间探索数据来帮助确定合适的题目”是非常有必要的。俗话说的好：“磨刀不误砍柴工”。 通常，大家在决赛开始的第一个小时（甚至更长的时间）来研究如何提出一个比较不错的问题。
   1. **最佳实践与应避免事项**

|  |  |
| --- | --- |
| **最佳实践** | **应避免** |
| **1.** 研究立题时，应该专注于假设检验。 | **1.** 不要试图在这么短的几个小时里，尝试所有你知道的模型，幻想交叉验证得到一个理想的精确度。 |
| **2.**至少花6个小时来写你的报告，确保你有不错的可视化数据和书面形式来展现你的最后结果。 | **2.** 不要违背统计模型的假设。有时，特定的模型需要一定的特征，因此请确保这些前提是满足的。 |
| **3.** 进行适当的因果分析。比如你的模型通过了标准的交叉验证，但这并不说明这里有因果关系。 | **3.** 不要盲目的拿来一个研究陈述，然后一直钻研或者尝试各种方法往这个研究陈述上靠。有时候你会发现，进一步的数据探索能够轻易地帮你验证这个陈述是错的或者不值得研究。 |

**6.4 寻求帮助**

如果你有任何疑惑，可以直接在决赛微信群内进行提问。

## 数据表格模式概要

***航空公司****（****airlines****）*

航空公司ID到名称的映射。

61行 & 2列。 大小： 约0.1MB。来源：互联网。

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 说明 |
| airline\_id | 字符串 | 2个字母的航空公司标识符 |
| airline\_name | 字符串 | 航空公司名称 |

***机场****（****airports****）*

美国各机场的重要详细信息（名称，州，标识符，纬度，经度等）。

322行 & 6列。 大小： 约0.1MB。来源：美国运输部（[US Department of Transportation](https://www.transportation.gov/)）。

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 说明 |
| airport\_id | 字符串 | 3个字母的机场标识符 |
| airport\_name | 字符串 | 机场名称 |
| city | 字符串 | 机场的城市 |
| state | 字符串 | 机场的州 |
| latitude | 浮点数 | 机场的维度 |
| longitude | 浮点数 | 机场的经度 |

***事件****（****events\_US****）*

2017年全年，美国各地的公共事件。

1,151行 & 4列。 大小： 约0.1MB。来源：[Shore Fire Media](https://shorefire.com/)。

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 说明 |
| date | 字符串 | 日期，格式DD/MM/YYYY和DD/MM/YY |
| city | 字符串 | 事件发生在的城市名称 |
| state | 字符串 | 事件发生在的州名称 |
| event\_name | 字符串 | 事件名称 |

***票价****（****fares****）*

2017年每个季度、航线、航空公司组合的票价分布，价格为10美元。

80,823行 & 255列。 大小： 约44MB。来源：美国运输部（[US Department of Transportation](https://www.transportation.gov/)）。

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 说明 |
| quarter | 整数 | 基准财务季度 |
| airline\_id | 字符串 | 航空公司标识符 |
| origin\_airport | 字符串 | 出发机场的3个字母标识符 |
| destination\_airport | 字符串 | 到达机场的3个字母标识符 |
| distance | 整数 | 机场之间的距离，以英里为单位 |
| fare\_bucket\_x  (合计249列) | 整数 | 票价在x到x + $10美元之间的乘客数量。第一列不包括低于$ 10的价值。最后一列包括$2500美元以上的所有价值 |

***飞行交通****（****flight\_traffic****）*

有关2017年美国国内航班延误的信息。

约600,000行 & 24列。 大小： 约130MB压缩版，约550MB正常版。来源：美国运输部的交通统计局（[Bureau of Transportation Statistics](https://www.bts.gov/)）。

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 说明 |
| year | 整数 | 年 |
| month | 整数 | 月 |
| day | 整数 | 日 |
| airline\_id | 字符串 | 航空公司标识符 |
| origin\_airport | 字符串 | 出发机场标识符 |
| destination\_airport | 字符串 | 到达机场标识符 |
| scheduled\_departure | 整数 | 预定的出发时间。 格式HHMM |
| actual\_departure | 整数 | 实际出发时间。 格式HHMM |
| taxi\_out | 整数 | 滑行出发时间，以分钟为单位 |
| wheels\_off | 整数 | 离场时间。 格式HHMM |
| wheels\_on | 整数 | 落地时间。 格式HHMM |
| taxi\_in | 整数 | 滑行到达时间，以分钟为单位 |
| scheduled\_arrival | 整数 | 预定的到达时间。 格式HHMM |
| actual\_arrival | 整数 | 实际到达时间。 格式HHMM |
| cancelled | 二进制 | 如果取消航班则为1，否则为0 |
| diverted | 二进制 | 如果改到飞航则为1，否则为0 |
| scheduled\_elapsed | 整数 | 计划的飞行时间，以分钟为单位 |
| actual\_elapsed | 整数 | 实际的飞行时间，以分钟为单位 |
| distance | 整数 | 机场之间的距离，以英里为单位 |
| airline\_delay | 整数 | 由于航空公司造成的延误，以分钟为单位 |
| weather\_delay | 整数 | 由于天气原因延迟，以分钟为单位 |
| air\_system\_delay | 整数 | 由于国家航空系统而延迟，以分钟为单位 |
| security\_delay | 整数 | 由于安全原因延迟，以分钟为单位 |
| aircraft\_delay | 整数 | 由于飞机晚点造成的延误，以分钟为单位 |

***股票价格****（****stock\_prices****）*

从2016年末到2018年初，美国多家航空公司的每日收盘价。

380行 & 10列。 大小： 约0.1MB。来源：[Alpha Vantage](https://www.alphavantage.co/)。

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 说明 |
| Timestamp | 字符串 | 日期，格式DD/MM/YYYY |
| airline\_stock\_price (合计9列) | 浮点数 | 收盘价 |

***天气****（****weather****）*

2017年，每6小时在美国各机场收集的天气数据（温度，风，降水，云量等）

353,864行 & 12列。 大小： 约48MB。来源：美国国家环境信息中心（[National Centers for Environmental Information](https://www.ncei.noaa.gov/)）。

|  |  |  |
| --- | --- | --- |
| Field | Type | Description |
| airport\_id | 字符串 | 3个字母的机场标识符 |
| airport\_name | 字符串 | 机场名称 |
| latitude | 浮点数 | 机场的维度 |
| longitude | 浮点数 | 机场的经度 |
| elevation | 浮点数 | 机场海拔，以米为单位 |
| datetime | 字符串 | 测量的日期和时间。 格式YYYY-MM-DDTHH：MM：SS |
| temperature | 浮点数 | 机场周围的气温，以摄氏度为单位 |
| visibility | 整数 | 能见度，以米为单位 |
| wind\_direction | 整数 | 水平风向，以北为顺时针方向 |
| wind\_speed | 浮点数 | 水平风速，以米/秒为单位 |
| snow\_depth | 浮点数 | 积雪深度，以厘米为单位 |
| cloud\_status | 浮点数 | 表示云覆盖状态的代码（0 – 9） |

***cloud\_status代码：***

0: 无

1: ACSL (Altocumulus Standing Lenticular) **高积站立式双凸透镜云**

2: ACCAS (Altocumulus Castelanus) **高积云Castelanus**

3: TCU (Towering Cumulus) **塔式积云**

4: MDT CU (Moderate Cumulus) **中等累积云**

5: CB/CB MAM DISTANT (Cumulonimbus or Cumulonimbus Mammatus visible) **可见积雨云或积雨云**

6: CB/CBMAM (Cumulonimbus or Cumulonimbus Mammatus within 20 nautical miles) **20海里以内的积雨云母或积雨云母**

7: Unknown 未知

9: Missing 缺失