

# 过去空气质素健康指数记录（英文版）的资料传送规格

 English Version

数据集地址

## [字段解释]

属性	描述	备注
Date	空气质素健康指数的公布日期	资料规格： 16/10/2018
Hour	根据 Date 里的一天分成小时 「其中一天的最后有一天中最大的 AQHI」	资料规格：1
Central/Western	中西区监测站的空气质素健康指数	
Eastern	东区监测站的空气质素健康指数	
Kwun Tong	观塘监测站的空气质素健康指数	
Sham Shui Po	深水步监测站的空气质素健康指数	
Kwai Chung	葵涌监测站的空气质素健康指数	
Tsuen Wan	荃湾监测站的空气质素健康指数	
Tseung Kwan O	将军澳监测站的空气质素健康指数	
Yuen Long	元朗监测站的空气质素健康指数	
Tuen Mun	屯门监测站的空气质素健康指数	
Tung Chung	东涌监测站的空气质素健康指数	
Tai Po	大埔监测站的空气质素健康指数	
Sha Tin	沙田监测站的空气质素健康指数	
Tap Mun	塔门监测站的空气质素健康指数	

属性	描述	备注
Causeway Bay	铜锣湾监测站的空气质素健康指数	
Central	中环监测站的空气质素健康指数	
Mong Kok	旺角监测站的空气质素健康指数	

# EDA Exploratory Data Analysis 「探索性数据分析」

## 整体思路

- 1. 数据质量评估: 检查缺失值、重复值和异常值。
- 2. 单变量分析: 深入理解每一个值
- 3. 多变量分析与时空分析: 探索属性之间的关系, 以及数据在时间和空间上的模式.

## 1. 数据质量评估

- 分析内容:
  - 缺失值分析
  - 异常值检测 「主要是 AQHI 的异常值」
  - 日期与时间格式分析

## 2. 单变量分析

### 理解每个变量自身的分布和统计特征

属性: Data : 某一年的某一月的某一日

属性: Hour : 某一日的某一小时

- 小时段: 01-24
- Daily Max 「某一日的某一小时的最大 AQHI」

属性: AQHI : 各个检测站的 AQHI(Central/Western, Eastern, Kwun Tong, Sham Shui Po, Kwai Chung, Tsuen Wan, Tseung Kwan O, Yuen Long, Tuen Mun, Tung Chung, Tai Po, Sha Tin, Tap Mun, Causeway Bay, Central, Mong Kok)

### 分析的核心

分析内容:

- 中心趋势与离散度: 计算每个站点点描述性统计: 均值, 中位数, 众数, 标准差, 方差, 极值等
- 分布形态: 识别每个站点 AQHI 的分布特征

- 如何分析: 使用直方图和密度图可视化每个站点的 AQHI 分布, 识别偏态和峰态
- 箱线图: 识别每个站点 AQHI 的异常值和四分位数范围

### 3. 多变量分析与时空分析

探索属性之间的关系, 以及数据在时间和空间上的模式

时间序列分析:

- 分析内容:
  - 长期趋势: 将 Data 进行聚合(例如计算每日平均 AQHI), 绘制时间序列图. 观察整体空气质量的变化趋势, 是否有季节性规律

空间分析(监测站之间的关系):

- 分析内容:
  - 相关性分析: 计算不同监测站 AQHI 之间的相关系数矩阵, 识别站点之间的关系强度, 并用热力图进行可视化「热力图: 使用热力图可视化不同监测站 AQHI 之间的相关性」
    - 目的: 找出哪些区域的空气质量变化模式高度一致(相关性高), 哪些区域相对独立. 例如: 地理上接近的站点相关性高
  - “最差”区域排名:
    - 计算每个监测站的平均 AQHI, 并进行排名, 识别空气质量最差的区域
    - 可视化: 使用条形图展示各监测站的平均 AQHI 排名

## Feature Engineering + Impact Analysis 「特征工程 + 影响分析」

### 1. 回顾 EDA 报告

#### 1. 数据质量评估

- i. 缺失值很少(总共 113 个缺失值, 占总数的 0.17%), 其中 Tap Mun 站点缺失值较多(占该站点数据的 5.6%)
- ii. 没有发现超出理论范围(1-10)的异常值
- iii. 使用 IQR 方法检测到 1401 个异常值, 占总数据的 2.1%

#### 2. 单变量分析

- i. 每个站点的 AQHI 均值大致在 3-4 之间, 中位数( $\approx 4$ ) 略高于均值, 表明分布略偏左
- ii. 各个站点的均值 AQHI 在 3.447(Tsuen Wan) 到 4.126(Kwai Chung) 之间

#### 3. 多变量分析与时空分析

- i. 站点之间的相关性较高(平均 0.914), 尤其是地理上接近的站点, 最高相关性在 0.98 左右
- ii. 最低相关性出现在 Tap Mun

- iii. 聚类分析将站点分组,说明地理位置对空气质量有影响
- iv. 时间分析显示, AQHI 在一天波动中, 峰值时间在 17:00, 谷值时间在 08:00, 周末的 AQHI 低于工作日

## 2. 特征工程

### 1. 处理缺失值

- 缺失值很少, 可以使用插值法填补

### 2. 时间特征工程

- 从 DateTime 中提取一下特征:
  - 小时段
  - 一天中的时间段(早晨, 下午, 晚上, 深夜)
  - 是否是周末 「周末 AQHI 较低」
  - 月份 「季节性影响」
  - 星期几(周一到周日)

### 3. 空间特征工程

- 站点高度相关性, 可以考虑使用主成分分析(PCA)或聚类分析将多个站点的 AQHI 合并为综合指标, 降低维度, 减少冗余信息
- 可以根据聚类分析的结果, 将站点分组, 并计算每个群集的平均 AQHI 作为新的特征

### 4.

## 3. 影响分析

### 1. 特征重要性分析:

- 在训练模型后, 使用特征重要性评分(如基于树模型的特征重要性)来评估新特征对模型性能的贡献

### 2. 时间特征影响:

- 分析小时, 周末等时间特征对预测的影响, 验证是否与 EDA 中的模式一致.(例如, 17:00 的峰值, 周末的较低值)

### 3. 空间特征影响:

- 检测不同站点特征的重要性, 以及区域特征(PCA 或聚类结果)对模型的贡献