

重大地震数据分析实验报告

数据文件: earthquakes-2025-10-29_21-16-05_+0800.tsv; 记录数: 6616

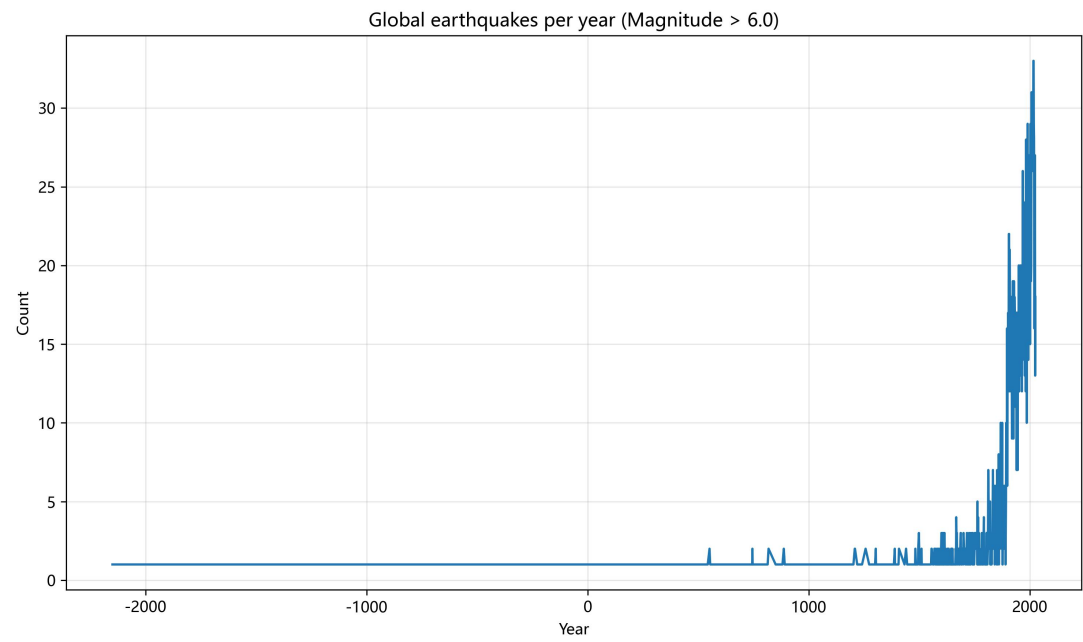
数据来源: Significant Earthquake Database (公元前 2150 年至今)。读取 TSV 后进行统计与绘制。

1.1 各国因地震造成的死亡总人数（前十）

排名	国家	死亡总人数
1	CHINA	2139210
2	TURKEY	1199742
3	IRAN	1014453
4	ITALY	498219
5	SYRIA	419226
6	HAITI	323484
7	AZERBAIJAN	319251
8	JAPAN	242445
9	ARMENIA	191890
10	PAKISTAN	145083

数据文件: ../tables/top10_deaths.csv

1.2 每年全球震级 > 6.0 的地震总数



图像: eq_counts_gt6.png

仅展示最近 10 年

年份	次数 (M>6.0)
2016	33
2017	33
2018	29
2019	28
2020	16
2021	23

年份	次数 (M>6.0)
2022	21
2023	27
2024	13
2025	18

趋势观察：

- 长期平均率较稳定（受板块构造控制），不易出现单调趋势。
- 早期记录不完整，现代仪器时代记录更充分，导致统计随时间改善。
- 年度波动可能由偶发大震年或余震序列造成。

1.3 各国家地震总次数与最大地震日期（降序，前十）

排名	国家	地震总次数	最大地震日期
1	CHINA	623	公元 1668 年-7-25
2	JAPAN	424	公元 2011 年-3-11
3	INDONESIA	421	公元 2004 年-12-26
4	IRAN	388	公元 856 年-12-22
5	TURKEY	358	公元 1939 年-12-26
6	ITALY	333	公元 1915 年-1-13

排名	国家	地震总次数	最大地震日期
7	GREECE	289	公元 365 年-7-21
8	USA	280	公元 1964 年-3-28
9	PHILIPPINES	230	公元 1897 年-9-21
10	MEXICO	214	公元 1787 年-3-28

数据文件: `..\tables\country_stats_top10.csv`

方法与实现

- 读取 TSV 并构建通用列映射（兼容 Country/Deaths/Mag/Year 等同义列）。
 - 死亡总数：按国家聚合空值视为 0。
 - 年度计数：统计 Year、Mag 两列齐备且 Mag > 6.0 的记录。
 - CountEq_LargestEq：返回该国记录总数与最大震级事件的日期（优先使用显式日期列，否则 Year-Mo-Dy 拼接）。
 - 绘图使用英文标题与坐标标签：Global earthquakes per year (Magnitude > 6.0)、Year、Count。

复现实验

```
python scripts/sig_eqs_analysis.py --file  
"data/earthquakes-2025-10-29_21-16-05_+0800.tsv"
```

质量与局限

- 早期历史记录不完整，可能低估某些时期事件/死亡。
- 不同国家的边界与地名可能随历史变动，聚合存在语义差异。
- 仅以震级阈值过滤，未考虑震源深度与台站覆盖的影响。

2010–2020 年深圳风速分析报告

站点：深圳宝安国际机场 (ID: 2281305, CALL_SIGN: ZGSZ)

数据来源与范围

- 数据集：NOAA Integrated Surface Dataset (ISD), 逐小时记录。
- 时间范围：2010-01 至 2020-09。
- 数据文件：./2281305.csv (项目根目录, 使用相对路径)

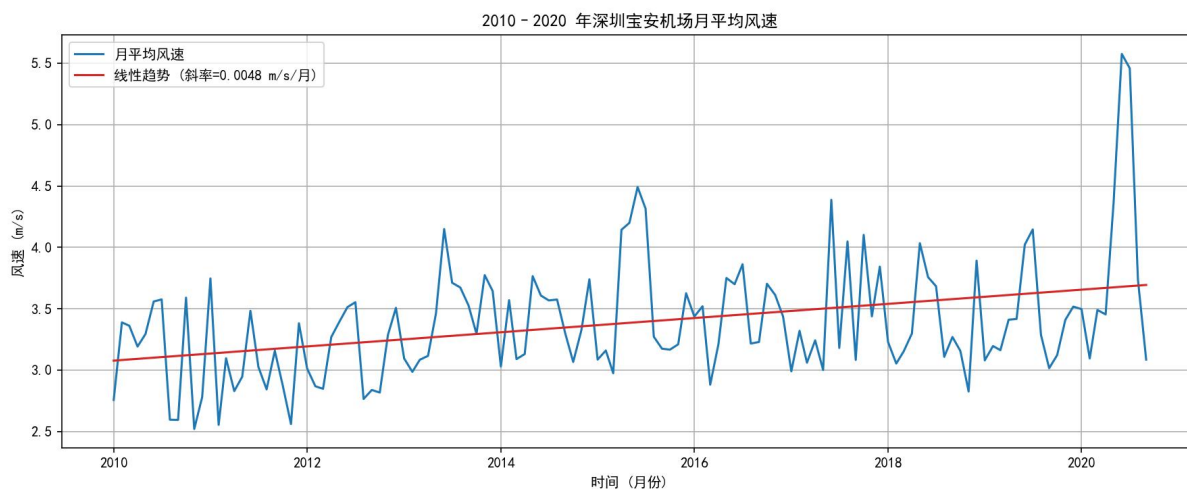
风数据筛选方法

风数据来自字段 WND, 其结构为 方向, 方向质量, 观测类型, 风速 (十分位 m/s), 风速质量。根据用户指南 (参考文献), 风速数值以十分位的米/秒表示, 例如 0060 表示 6.0 m/s。

本次分析采用如下筛选规则：

- 风速缺失值剔除：当风速为 9999/999/99999 时视为缺失。
- 质量控制：仅保留风速质量码为 0 或 1 的记录。
- 观测类型不作限制：包含正常 (N)、可变风向 (V) 与静风 (C), 静风风速记为 0.0 m/s。
- 时间过滤：仅保留 2010-01 至 2020-12 的记录。

结果与趋势



使用每月起始时间点进行重采样 (MS) 计算月平均风速，并对时间序列进行线性拟合。拟合斜率为 **0.0048 m/s/每月**，总体呈 **上升** 趋势。

文件组织与复现

- 脚本: `./脚本/深圳风速分析.py`
- 图表: `./图表/月平均风速_2010-2020.png`
- 报告: `./报告/深圳风速报告.html`
- 数据集: `./2281305.csv` (已包含在项目中)

参考文献

- Visual Crossing. How We Process the NOAA Integrated Surface Database Historical Weather Data.
<https://www.visualcrossing.com/resources/documentation/weather-data/how-we-process-integrated-surface-database-historical-weather-data/>
- NOAA NCEI. Global Hourly - Integrated Surface Database (ISD).
<https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database>

NCEI Climate at a Glance — Global Land+Ocean Monthly Temperature Anomalies

数据源：NOAA NCEI Climate at a Glance Global Time Series — CSV 直链：
https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/global/time-series/globe/tavg/land_ocean/1/0/1850-2025/data.csv

任务概述

依据题目“探索一个数据集”，从允许站点选择含时间信息的 CSV 数据，完成：

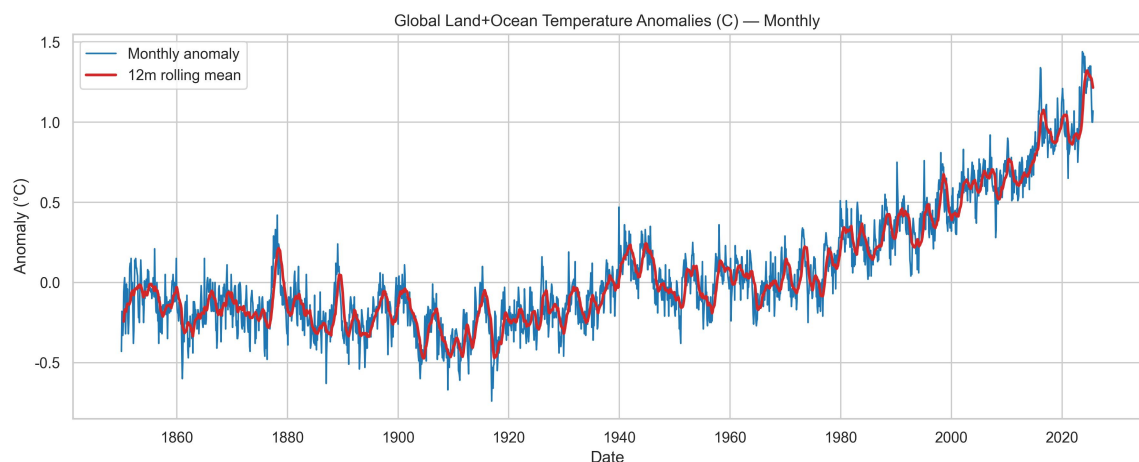
- **3.1** 载入并清理缺失或质量较差数据点
- **3.2** 绘制某变量的时间序列图
- **3.3** 至少 5 项统计检查并报告发现

3.1 数据载入与清理

- 忽略 CSV 注释行（以#开头），解析 Date(YYYYMM) 与 Anomaly(°C)
- 删除缺失与非数值记录；按日期排序并去重
- 依据 z-score 阈值 ($|z| > 4$) 剔除极端异常值

3.2 时间序列图

下图展示月度距平（蓝）与 12 个月滚动均值（红），用于观察长期趋势。



3.3 统计检查 (>=5 项)

样本量	2108
清洗后缺失	0
均值 (°C)	0.066
中位数 (°C)	-0.035
标准差 (°C)	0.382
最小值 (°C)	-0.740
最大值 (°C)	1.440
线性趋势 每十年 (°C)	0.058
拟合优度 R^2	0.595
自相关 滞后 12 个月	0.891
3 σ 异常值个数	23
均值(1901–1930)	-0.260
均值(1991–2020)	0.627
差值(近代-早期) (°C)	0.888

结论要点

- 全球陆海月平均温度距平呈显著上升趋势，线性拟合每十年约 0.058 °C， $R^2=0.595$
- 时间序列在一年期尺度存在明显正自相关（滞后 12 个月 0.891），滚动均值持续抬升
- 清洗后极端异常值很少（3 σ 之外 23 个），数据质量总体良好
- 近代（1991–2020）均值 0.627 °C 显著高于早期（1901–1930）均值 -0.260 °C，差值 0.888 °C

复现说明

安装依赖： `.\.venv\Scripts\pip.exe install pandas matplotlib seaborn`
运行分析： `.\.venv\Scripts\python.exe analyze_ncei.py` 输出目录：

outputs\ (包含 time_series.png, report.txt, report.html)

引用: NOAA National Centers for Environmental Information (NCEI), Climate at a Glance:
Global Time Series,

<https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/global/time-series>