# مستندات پروژه تحلیل مصرف آب

## 1. عنوان و نمای کلی پروژه

* **عنوان پروژه:** داشبورد تحلیل مصرف آب و پیش‌پردازش داده
* **هدف:** این پروژه بر تحلیل جامع داده‌های مصرف آب برای سال‌های ۱۴۰۱، ۱۴۰۲ و ۱۴۰۳ متمرکز است. این پروژه شامل بارگذاری، پیش‌پردازش، پاکسازی داده‌ها، تحلیل داده‌های اکتشافی و توسعه یک داشبورد وب تعاملی برای بصری‌سازی معیارهای کلیدی و شناسایی مشکلات کیفیت داده و الگوهای مصرف می‌باشد.
* **اجزا:**
  + فایل‌های داده خام (Excel/CSV)
  + نوت‌بوک Jupyter (preprocessing and cleaning.ipynb) برای پیش‌پردازش داده و تحلیل اولیه.
  + خلاصه تحلیل آماری (preprocessing\_summary.md) که یافته‌های کلیدی نوت‌بوک را خلاصه می‌کند.
  + کد پایتون (analyzer.py) که توابع قابل استفاده مجدد برای تحلیل و بصری‌سازی داده‌ها را فراهم می‌کند.
  + اپلیکیشن Streamlit (app.py) که یک داشبورد وب تعاملی برای کاوش داده‌ها ایجاد می‌کند.

## 2. توضیحات داده

این پروژه از داده‌های مصرف آب ارائه شده در قالب Excel/CSV استفاده می‌کند که سال‌های ۱۴۰۱، ۱۴۰۲ و ۱۴۰۳ را پوشش می‌دهد. فایل‌های داده خام (1401.xlsx، 1402.xlsx، 1403.xlsx) حاوی اطلاعات مفصلی با ساختار هدر MultiIndex در دو ردیف هستند.

داده‌های هر سال به نظر می‌رسد ترکیبی از موارد زیر است:

1. **اطلاعات ثابت مشتری/کنتور:** در ستون‌های ابتدایی قرار دارد. این شامل جزئیاتی مانند نام مشترک، کد اشتراک، نوع پروانه، موقعیت مکانی (استان، شهرستان، امور آب، محدوده مطالعاتی)، جزئیات کنتور (سریال، سایز، تاریخ نصب)، زمان آخرین اتصال و وضعیت قطع می‌باشد.
2. **داده‌های ماهانه مصرف:** پس از اطلاعات ثابت، ستون‌های ماهانه، داده‌های سری زمانی برای هر مشتری را ارائه می‌دهند. این ستون‌های ماهانه زیر عنوان ماه مربوطه (مثلاً ‘۱۴۰۱/۰۱’، ‘۱۴۰۱/۰۲’) گروه‌بندی شده‌اند و چندین معیار را برای آن دوره زمانی مشخص ارائه می‌دهند.

**ستون‌های کلیدی شناسایی شده (با نام‌های ترجمه شده استفاده شده در نوت‌بوک):**

* **اطلاعات ثابت:**
  + نام مشترک / Customer Name
  + کد اشتراک / Subscription Code
  + نوع پروانه / License Type
  + استان / Province
  + شهرستان / County
  + امور آب / Water Authority
  + محدوده مطالعاتی / Study Area
  + سریال کنتور / Meter Serial
  + سایز کنتور / Meter Size
  + تاریخ نصب / Installation Date
  + زمان آخرین اتصال / Last Connection Time
  + قطع کن / Disconnect Status
  + مصرف بازه (m³) / Consumption in Period (m³) (بر اساس موقعیت در بخش اطلاعات ثابت، این به نظر می‌رسد مصرف کل در یک دوره *بزرگتر* باشد، نه ماهانه).
  + ساعت کارکرد بازه (h) / Operating Hours in Period (h) (مشابهاً، مجموع ساعت کارکرد در یک دوره بزرگتر).
  + میانگین دبی بازه l/s / Average Flow Rate in Period (l/s) (میانگین دبی در یک دوره بزرگتر).
* **مصرف ماهانه (برای هر ماه YYYY/MM):**
  + دبی l/s / Flow Rate (l/s) (میانگین دبی برای ماه مشخص).
  + تعداد دبی منفی / Number of Negative Flows (تعداد نمونه‌های دبی منفی در ماه).
  + درصد دبی منفی / Percentage of Negative Flows (درصد نمونه‌های دبی منفی در ماه).
  + ساعت کارکرد (h) / Operating Hours (h) (مجموع ساعت کارکرد برای ماه).
  + مصرف (m³) (m³) / Consumption (m³) (مصرف کل برای ماه).
  + تعداد اطلاعات موجود / Number of Available Data Points (تعداد نقاط داده معتبر در ماه).
  + تعداد مورد انتظار / Expected Number of Data Points (تعداد مورد انتظار نقاط داده در ماه).
  + درصد اطلاعات موجود / Percentage of Available Data Points (درصد نقاط داده معتبر در ماه).

## 3. اهداف

اهداف این پروژه عبارتند از: \* پاکسازی و پیش‌پردازش داده‌های خام از چند سال و قالب‌های مختلف. \* شناسایی و مدیریت مشکلات کیفیت داده، به ویژه مقادیر اشتباه مصرف منفی و ساعت کارکرد. \* تبدیل داده‌ها به قالبی مناسب برای تحلیل سری‌های زمانی. \* انجام تحلیل داده‌های اکتشافی برای درک الگوهای مصرف، توزیع‌ها و روابط بین متغیرها. \* ارائه یک داشبورد تعاملی برای کاربران جهت کاوش آسان داده‌ها و بصری‌سازی‌ها. \* آماده‌سازی زمینه برای کارهای آتی احتمالی، مانند تشخیص ناهنجاری یا مدل‌سازی پیش‌بینانه.

## 4. فرآیند پیش‌پردازش و پاکسازی (preprocessing and cleaning.ipynb)

نوت‌بوک preprocessing and cleaning.ipynb مراحل انجام شده برای آماده‌سازی داده‌های خام برای تحلیل را شرح می‌دهد.

1. **بارگذاری داده‌های خام:**
   * فایل‌های اکسل (1401.xlsx، 1402.xlsx، 1403.xlsx) با استفاده از pd.read\_excel(..., header=[0, 1]) بارگذاری می‌شوند تا هدر MultiIndex دو ردیفی به درستی مدیریت شود.
   * DataFrames جداگانه (df\_1401, df\_1402, df\_1403) برای داده‌های کامل هر سال ایجاد می‌شود.
   * ستون‌های اطلاعات ثابت و ستون‌های مصرف ماهانه به صورت مفهومی بر اساس محدوده‌های ایندکس ستون جدا می‌شوند.
2. **بررسی و کاوش اولیه داده‌ها:**
   * بررسی‌های اولیه روی DataFrames بارگذاری شده انجام می‌شود، از جمله نمایش چند ردیف اول (.head())، بررسی انواع داده و تعداد مقادیر غیرتهی (.info()) و تولید آمار توصیفی (.describe()) برای ستون‌های عددی.
   * شکل هر DataFrame بررسی می‌شود.
   * مقادیر گمشده با استفاده از .isnull().sum() شناسایی و شمارش می‌شوند (مانند سلول ۳۵ برای df\_info). این بلافاصله ستون‌هایی با مقادیر گمشده را مشخص می‌کند.
   * بصری‌سازی‌ها (هیستوگرام، باکس‌پلات، ماتریس همبستگی) برای درک توزیع ویژگی‌های عددی و شناسایی مشکلات احتمالی مانند ناهنجاری‌ها استفاده می‌شوند. کتابخانه missingno برای بصری‌سازی الگوهای داده گمشده استفاده می‌شود.
3. **بازسازی ساختار داده (فرمت Long):**
   * یک مرحله حیاتی، تبدیل داده‌های ماهانه از قالب Wide به قالب Long است که برای تحلیل سری‌های زمانی و رسم نمودار در طول ماه‌ها مناسب‌تر است.
   * ستون کد اشتراک استخراج می‌شود.
   * ستون‌های ماهانه با استفاده از سطوح MultiIndex شناسایی می‌شوند.
   * یک حلقه از طریق ماه‌ها تکرار می‌شود و معیارهای ماهانه برای هر مشتری stack می‌شود.
   * DataFrame حاصل (df\_long) دارای ردیف‌هایی است که هر ترکیب مشتری-ماه را نشان می‌دهند، با ستون‌هایی برای ‘Subscription Code’، ‘month’ و معیارهای ماهانه (‘Flow Rate (l/s)’, ‘Operating Hours (h)’, ‘Consumption (m³)’, و غیره). این فرآیند برای ۱۴۰۱ (سلول‌های ۲۴، ۳۱)، ۱۴۰۳ (سلول ۲۵، ۲۶) و ۱۴۰۲ (سلول ۲۸) نشان داده شده است.
   * .info() روی df\_long فرمت Long را تأیید می‌کند و حجم داده‌های گمشده در ستون‌های معیارهای ماهانه را نشان می‌دهد (مانند سلول ۳۱ که NaNs زیادی را برای Flow Rate، Operating Hours، Consumption نشان می‌دهد).
4. **تغییر نام ستون‌ها:**
   * یک دیکشنری column\_translations برای نگاشت نام ستون‌های فارسی به نام‌های انگلیسی توصیفی‌تر تعریف می‌شود (سلول ۳۵).
   * df\_info و df\_long با استفاده از این دیکشنری تغییر نام داده می‌شوند (سلول‌های ۳۸، ۳۹). توجه داشته باشید که یک عدم تطابق جزئی وجود دارد که در آن ‘کد اشتراک’ به ‘Subscription Code’ در دیکشنری نگاشت شده است، اما پس از فرآیند stacking و تغییر نام در برخی سلول‌ها (مانند سلول ۲۳) به صورت ‘subscrition code’ (حروف کوچک) در df\_long ظاهر می‌شود. به نظر می‌رسد این یک اثر جانبی reset\_index و مرحله تغییر نام بعدی باشد. برای حفظ یکپارچگی در مستندات، از نام استاندارد ‘Subscription Code’ استفاده می‌شود مگر اینکه به صراحت به نام موقت در خروجی df\_long اشاره شود.
5. **مدیریت داده‌های اشتباه (مقادیر منفی):**
   * بر اساس دانش دامنه، مصرف آب و ساعت کارکرد منفی نامعتبر هستند.
   * ردیف‌های در DataFrames اطلاعاتی برای ۱۴۰۲ و ۱۴۰۳ (df\_info\_1402, df\_info\_1403) که در آن‌ها ‘Consumption in Period (m³)’ منفی است، شناسایی می‌شوند (سلول‌های ۹۲ و ۹۳ خروجی‌هایی برای ردیف‌های مصرف منفی در DataFrames اطلاعاتی ۱۴۰۲ و ۱۴۰۳ نشان می‌دهند).
   * مقادیر ‘Subscription Code’ متناظر با این رکوردهای اشتباه در لیست‌هایی جمع‌آوری می‌شوند (customers\_to\_drop\_1402, customers\_to\_drop\_1403).
   * DataFrames اطلاعاتی پاکسازی شده (df\_info\_clean\_1402, df\_info\_clean\_1403) با فیلتر کردن ردیف‌هایی که ‘Subscription Code’ آن‌ها در این لیست‌ها قرار دارد، ایجاد می‌شوند (سلول ۹۴). این مورد جدی‌ترین مشکل کیفیت داده در اطلاعات خلاصه را برطرف می‌کند.
   * مقادیر منفی در داده‌های ماهانه مصرف (df\_long) برای ‘Consumption (m³)’ و ‘Operating Hours (h)’ شناسایی می‌شوند (سلول‌های ۳۳، ۳۴). اگرچه این موارد به عنوان ناهنجاری لیست شده‌اند، اما نوت‌بوک آن‌ها را *شناسایی* می‌کند ولی در اسنیپت‌های ارائه شده صراحتاً آن‌ها را از df\_long *حذف نمی‌کند* قبل از ذخیره داده‌ها در قالب Long. ایده‌های تحلیل اشاره به بررسی/مدیریت این موارد دارند.
6. **ذخیره داده‌های پاکسازی شده/پردازش شده:**
   * DataFrames اطلاعاتی پاکسازی شده (df\_info\_clean\_1402, df\_info\_clean\_1403) به صورت فایل‌های CSV ذخیره می‌شوند (1402\_clean\_info.csv, 1403\_clean\_info.csv) (سلول ۹۶).
   * DataFrames پردازش شده در قالب Long (df\_long برای ۱۴۰۱، df\_long\_1402, df\_long\_1403) به صورت فایل‌های CSV ذخیره می‌شوند (long\_usage1401.csv, long\_usage1402.csv, long\_usage1403.csv) (سلول‌های ۱۰۰، ۱۰۱، ۱۰۲).

## 5. تحلیل داده‌های اکتشافی (EDA) و یافته‌های کلیدی

نوت‌بوک شامل مراحل مختلفی برای درک ویژگی‌های داده است:

1. **آمار توصیفی کلی:** خلاصه آمار برای ستون‌های عددی در df\_long بینشی در مورد تمایل مرکزی، پراکندگی و دامنه معیارهای ماهانه ارائه می‌دهد (سلول ۳۰). به ویژه، این نشان می‌دهد که میانگین مصرف به دلیل ناهنجاری‌های منفی شدید، منفی است.
2. **آمار در سطح مشتری:** گروه‌بندی df\_long بر اساس ‘Subscription Code’ و محاسبه میانگین، بینشی در مورد رفتار میانگین هر مشتری در طول سال ارائه می‌دهد (سلول ۱۱۷).
3. **معیارهای کیفیت داده:** تحلیل ‘Number of Negative Flows’ و ‘Percentage of Available Data Points’ به اندازه‌گیری کامل بودن داده و نرخ خطا در ترکیب‌های مشتری-ماه کمک می‌کند (سلول‌های ۴۰، ۴۱). درصد قابل توجهی از رکوردها دسترسی به داده پایینی دارند (< 50%).
4. **تحلیل همبستگی:** یک ماتریس همبستگی برای ستون‌های عددی در df\_long روابط را نشان می‌دهد (سلول ۴۲). یک همبستگی منفی قوی بین ‘Operating Hours (h)’ و ‘Consumption (m³)’ مشاهده می‌شود که غیرمعمول است و یک پرچم برای مشکل کیفیت داده است که نیاز به بررسی بیشتر دارد.
5. **شناسایی موارد پرت:** مشتریان برتر بر اساس میانگین مصرف ماهانه و تعداد دبی منفی شناسایی می‌شوند (سلول‌های ۴۳، ۴۵)، که مناطق احتمالی برای بررسی عمیق‌تر را برجسته می‌کنند.
6. **شناسایی ناهنجاری‌ها:**
   * ناهنجاری‌های واضح مانند مصرف/ساعت کارکرد منفی به صراحت در df\_long شناسایی می‌شوند (سلول ۴۸).
   * روش IQR برای تشخیص موارد پرت احتمالی در ‘Consumption (m³)’ (با حذف مقادیر منفی برای استحکام محاسبه) در df\_long اعمال می‌شود (سلول ۴۹). تعداد قابل توجهی موارد پرت یافت می‌شود.
   * از متدهای کلاس InteractiveDataAnalyzer برای تولید بصری‌سازی برای تشخیص ناهنجاری (امتیاز Z)، توزیع (هیستوگرام) و روابط (heatmap همبستگی، parallel coordinates) روی df\_info\_clean استفاده می‌شود (سلول‌های ۹۸، ۹۹، ۱۰۱). این تحلیل بصری، کجی توزیع‌ها و وجود موارد پرت در معیارهایی مانند ‘Consumption in Period (m³)’ و ‘Average Flow Rate in Period (l/s)’ را تأیید می‌کند.

**یافته‌های کلیدی و نکات کیفیت داده:** \* داده‌ها دارای ساختار MultiIndex هستند که نیاز به بارگذاری و پردازش خاصی دارد. \* تبدیل به فرمت Long برای تحلیل سری‌های زمانی ضروری است. \* داده‌های گمشده در ستون‌های معیارهای ماهانه رایج هستند. \* مقادیر منفی در مصرف و ساعت کارکرد وجود دارند و به عنوان داده‌های اشتباه در نظر گرفته می‌شوند، به ویژه در داده‌های اطلاعاتی که رکوردهای مشتری متناظر حذف می‌شوند. \* یک همبستگی منفی قوی و غیرمعمول بین ساعت کارکرد ماهانه و مصرف وجود دارد، که نشان‌دهنده مشکلات احتمالی داده یا الگوهای عملیاتی غیرمعمول است. \* تحلیل بصری، توزیع‌های بسیار کج و موارد پرت در معیارهای عددی کلیدی را تأیید می‌کند.

## 6. داشبورد تعاملی (app.py و analyzer.py)

فایل app.py یک داشبورد تحلیل تعاملی را با استفاده از فریم‌ورک Streamlit پیاده‌سازی می‌کند و از کلاس InteractiveDataAnalyzer تعریف شده در analyzer.py بهره می‌برد.

**analyzer.py:** این فایل پایتون کلاس InteractiveDataAnalyzer را تعریف می‌کند. این کلاس توابع مختلفی را برای تحلیل و بصری‌سازی یک DataFrame پانداس کپسوله می‌کند. متدهای آن شامل موارد زیر هستند: \* فیلتر کردن داده‌ها بر اساس مقدار، محدوده یا بازه تاریخی. \* تولید نمودارهای تعاملی با استفاده از Plotly (هیستوگرام، باکس‌پلات، نمودار پراکندگی، نمودار خطی، تشخیص ناهنجاری، نمودار دایره‌ای، نمودار میله‌ای، heatmap همبستگی و نمودار خاص مصرف بر اساس نوع پروانه). \* انجام خلاصه‌های آماری (معادل .describe() و خلاصه تفصیلی شامل کجی و کشیدگی). \* نمایش تعداد مقادیر گمشده. \* پیاده‌سازی تشخیص ناهنجاری با استفاده از روش Z-score. \* ارائه قابلیت‌های مقیاس‌بندی/نرمال‌سازی داده‌ها (scale\_column, normalize\_all\_numeric) با استفاده از پیش‌پردازشگرهای scikit-learn.

این کلاس منطق اصلی مورد استفاده توسط اپلیکیشن Streamlit را برای انجام تحلیل و تولید بصری‌سازی به صورت پویا بر اساس انتخاب کاربر فراهم می‌کند.

**app.py:** اپلیکیشن Streamlit به عنوان رابط کاربری تعاملی برای پروژه عمل می‌کند. \* به کاربران امکان می‌دهد یک یا چند فایل اکسل یا CSV را بارگذاری کنند. \* پس از بارگذاری فایل‌ها، یک نوار کناری ظاهر می‌شود که به کاربران امکان انتخاب یک مجموعه داده از فایل‌های بارگذاری شده را می‌دهد. \* فایل انتخاب شده در یک DataFrame پانداس بارگذاری می‌شود. \* یک شیء InteractiveDataAnalyzer با DataFrame بارگذاری شده نمونه‌سازی می‌شود. \* نوار کناری گزینه‌های پیکربندی را ارائه می‌دهد: \* انتخاب یک ستون عددی. \* انتخاب یک ستون دسته‌ای. \* انتخاب یک نوع نمودار از لیست از پیش تعریف شده (شامل ‘None’, ‘Histogram’, ‘Boxplot’, ‘Line Chart’, ‘Outlier Detection’, ‘Pie Chart’, ‘Bar Chart’, ‘Correlation Heatmap’). \* بر اساس نوع نمودار و ستون‌های انتخاب شده، اپلیکیشن متد مربوطه را از شیء InteractiveDataAnalyzer فراخوانی کرده و نمودار Plotly را با استفاده از st.plotly\_chart نمایش می‌دهد. \* اپلیکیشن همچنین نمایش می‌دهد: \* خلاصه آماری برای ستون عددی انتخاب شده (analyzer.statistical\_summary). \* جدول مقادیر گمشده در هر ستون (analyzer.show\_missing\_values). \* بخشی برای پیش‌پردازش که به کاربران امکان انتخاب یک روش مقیاس‌بندی (‘standard’, ‘minmax’, ‘robust’) و اعمال آن بر روی یک ستون عددی انتخاب شده (analyzer.scale\_column) را می‌دهد و نتیجه را در یک DataFrame نمایش می‌دهد. \* مدیریت خطا برای بارگذاری فایل گنجانده شده است.

این ساختار به کاربران امکان می‌دهد به صورت پویا جنبه‌های مختلف داده را کاوش کرده، توزیع‌ها و روابط مختلف را بصری‌سازی کرده و به سرعت به آمار کلیدی و اطلاعات کیفیت داده از طریق یک رابط کاربری دوستانه دسترسی پیدا کنند. این برنامه از داده‌های پاکسازی شده تولید شده توسط نوت‌بوک (با فرض اینکه کاربر فایل‌های CSV پاکسازی شده را بارگذاری می‌کند یا خود برنامه داده‌های پاکسازی شده را به صورت داخلی بارگذاری می‌کند) برای پشتیبانی از بصری‌سازی‌ها و ویژگی‌های خود استفاده می‌کند.

## 7. اجرای پروژه

برای اجرای اپلیکیشن Streamlit و تعامل با داده‌ها: 1. اطمینان حاصل کنید که پایتون و کتابخانه‌های لازم نصب شده‌اند (streamlit, pandas, plotly, seaborn, matplotlib, scipy, scikit-learn, missingno در صورت نیاز برای نوت‌بوک). می‌توانید آن‌ها را با استفاده از pip نصب کنید: bash pip install streamlit pandas plotly seaborn matplotlib scipy scikit-learn missingno openpyxl 2. کدهای ارائه شده را در فایل‌هایی با نام‌های app.py و analyzer.py در یک دایرکتوری ذخیره کنید. 3. فایل‌های داده پاکسازی شده (مانند long\_usage1401.csv, 1402\_clean\_info.csv, و غیره - یا فایل‌های .xlsx اصلی، بسته به اینکه قصد دارید برنامه چه فایلی را بارگذاری کند) را در مکانی که توسط اسکریپت قابل دسترسی باشد (مثلاً همان دایرکتوری) قرار دهید. 4. ترمینال یا Command Prompt خود را باز کنید. 5. به دایرکتوری که فایل‌ها را در آن ذخیره کرده‌اید بروید. 6. اپلیکیشن Streamlit را با دستور زیر اجرا کنید: bash streamlit run app.py 7. مرورگر وب شما به طور خودکار باز شده و داشبورد Streamlit را نمایش می‌دهد. از قسمت بارگذاری فایل برای آپلود فایل‌های داده خود و از نوار کناری برای تعامل با گزینه‌های تحلیل استفاده کنید.

## 8. کارهای آتی احتمالی

بر اساس وضعیت فعلی پروژه و یافته‌ها، کارهای آتی می‌تواند شامل موارد زیر باشد:

* **پاکسازی داده قوی‌تر:**
  + پیاده‌سازی مدیریت پیچیده‌تر برای مصرف/ساعت کارکرد منفی در داده‌های ماهانه (df\_long)، فراتر از صرفاً شناسایی. این می‌تواند شامل جایگزینی (imputation)، محدود کردن (capping) یا پرچم‌گذاری برای تحلیل جداگانه بر اساس تخصص دامنه باشد.
  + توسعه استراتژی‌هایی برای مدیریت مقادیر گمشده در معیارهای ماهانه (مثلاً روش‌های جایگزینی خاص سری‌های زمانی).
* **تشخیص ناهنجاری پیشرفته:**
  + پیاده‌سازی روش‌های تشخیص ناهنجاری خاص سری‌های زمانی (مثلاً بر اساس تجزیه فصلی، آمار متحرک یا الگوریتم‌های پیشرفته‌تر) برای شناسایی الگوهای مصرف غیرمعمول در طول زمان برای مشتریان منفرد.
  + کاوش تکنیک‌های تشخیص ناهنجاری چند متغیره.
* **مهندسی ویژگی (Feature Engineering):**
  + استخراج ویژگی‌های جدید از داده‌های سری زمانی (مانند تغییر مصرف ماهانه، شاخص‌های فصلی، معیارهای بهره‌وری عملیاتی).
  + ترکیب موثر ویژگی‌ها از DataFrames اطلاعاتی و Long.
* **تحلیل و مدل‌سازی سری‌های زمانی:**
  + تحلیل روندها و فصلی بودن مصرف در سطح کلی و برای بخش‌های خاص مشتریان.
  + توسعه مدل‌هایی برای پیش‌بینی مصرف آب در آینده.
  + ساخت مدل‌هایی برای خوشه‌بندی مشتریان بر اساس الگوهای مصرف آن‌ها.
* **توسعه داشبورد:**
  + اضافه کردن بصری‌سازی‌های تعاملی و گزینه‌های فیلتر بیشتر.
  + ادغام نتایج مدل‌های خوشه‌بندی یا پیش‌بینی در داشبورد.
  + بهبود مدیریت و نمایش مقادیر منفی یا ناهنجاری‌ها، شاید با اجازه دادن به کاربران برای تغییر وضعیت نمایش آن‌ها یا مشاهده خلاصه‌هایی به طور خاص برای رکوردهای دارای مشکل.
  + اضافه کردن مستندات واضح یا توضیحات (tooltips) در داخل برنامه که معیارها و بصری‌سازی‌ها را توضیح دهند.

## 9. نتیجه‌گیری

این پروژه پایه محکمی برای تحلیل داده‌های مصرف آب فراهم می‌کند. نوت‌بوک با موفقیت به پاکسازی اولیه داده‌ها و بازسازی ساختار پرداخته و مشکلات کلیدی کیفیت داده مانند مصرف منفی و همبستگی‌های غیرمنتظره را شناسایی کرده است. توسعه کلاس InteractiveDataAnalyzer و اپلیکیشن Streamlit ابزار ارزشمندی را برای کاوش و بصری‌سازی تعاملی ایجاد می‌کند و داده‌ها را قابل دسترس‌تر و قابل فهم‌تر می‌سازد. یافته‌های تحلیل EDA مناطق حیاتی (مقادیر منفی، ناهنجاری‌های همبستگی، داده‌های گمشده) را برجسته می‌کنند که نیاز به بررسی بیشتر برای ساخت مدل‌های قابل اعتماد یا استنتاج قطعی در مورد الگوهای مصرف آب دارند.