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، حاوی سال‌های ۱۴۰۱ (1401.xlsx، 1402.xlsx، 1403.xlsx) ۱۴۰۲ و ۱۴۰۳ را پوشش می‌دهد. فایل‌های داده خام 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:**
  - Customer Name / نام مشترک
  - Subscription Code / کد اشتراک
  - License Type / نوع پروانه
  - Province / استان
  - County / شهرستان
  - Water Authority / امور آب
  - Study Area / محدوده مطالعاتی
  - Meter Serial / سریال کنتور
  - Meter Size / سایز کنتور
  - Installation Date / تاریخ نصب
  - Last Connection Time / زمان آخرین اتصال
  - Disconnect Status / قطع کن

- 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 نقاط داده معتبر / Number of Available Data Points (تعداد اطلاعات موجود در ماه).
  - تعداد مورد انتظار نقاط داده / Expected Number of Data Points (تعداد مورد انتظار در ماه).
  - درصد نقاط داده / Percentage of Available Data Points (درصد اطلاعات موجود 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)

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

1. **بارگذاری داده‌های خام:**

- فایل‌های اکسل (1401.xlsx، 1402.xlsx، 1403.xlsx) با استفاده از pd.read_excel(...,) دو ردیفی به درستی مدیریت MultiIndex بارگذاری می‌شوند تا هدر (header=[0, 1]) شود.
- برای داده‌های کامل هر سال (df_1401, df_1402, df_1403) جداگانه DataFrames 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:**

- بارگذاری شده انجام می‌شود، از جمله نمایش DataFrames بررسی‌های اولیه روی و تولید (info().) بررسی انواع داده و تعداد مقادیر غیرتهی ،(head().) چند ردیف اول برای ستون‌های عددی (describe().) آمار توصیفی.
- شکل هر DataFrame بررسی می‌شود.
- مقادیر گمشده با استفاده از isnull().sum() شناسایی و شمارش می‌شوند (مانند این بلافاصله ستون‌هایی با مقادیر گمشده را مشخص .(df_info سلول ۳۵ برای می‌کند.

- بصری‌سازی‌ها (هیستوگرام، باکس‌پلات، ماتریس همبستگی) برای درک توزیع ویژگی‌های عددی و شناسایی مشکلات احتمالی مانند ناهنجاری‌ها استفاده می‌شوند. برای بصری‌سازی الگوهای داده گمشده استفاده می‌شود missingno کتابخانه.

3. **بازسازی ساختار داده (فرمت Long):**

- است که برای Long به قالب Wide یک مرحله حیاتی، تبدیل داده‌های ماهانه از قالب تحلیل سری‌های زمانی و رسم نمودار در طول ماه‌ها مناسب‌تر است.
- ستون کد اشتراک استخراج می‌شود.
- شناسایی می‌شوند MultiIndex ستون‌های ماهانه با استفاده از سطوح.
- stack یک حلقه از طریق ماه‌ها تکرار می‌شود و معیارهای ماهانه برای هر مشتری می‌شود.
- دارای ردیف‌هایی است که هر ترکیب مشتری-ماه را (df_long) حاصل DataFrame و معیارهای ‘month’، ‘Subscription Code’ نشان می‌دهند، با ستون‌هایی برای ماهانه (‘Flow Rate (l/s)’, ‘Operating Hours (h)’, ‘Consumption (m^3)’, و این فرآیند برای ۱۴۰۱ (سلول‌های ۲۴، ۳۱)، ۱۴۰۳ (سلول ۲۵، ۲۶) و ۱۴۰۲ (سلول ۲۸) .(غیره نشان داده شده است.
- را تأیید می‌کند و حجم داده‌های گمشده در Long فرمت df_long روی ()info. زیادی را برای NaNs مانند سلول ۳۱ که) ستون‌های معیارهای ماهانه را نشان می‌دهد (نشان می‌دهد Flow Rate، Operating Hours، Consumption.

4. **تغییر نام ستون‌ها:**

- برای نگاشت نام ستون‌های فارسی به نام‌های column_translations یک دیکشنری انگلیسی توصیفی‌تر تعریف می‌شود (سلول ۳۵).
- با استفاده از این دیکشنری تغییر نام داده می‌شوند (سلول‌های ۳۸، df_long و df_info ۳۹). توجه داشته باشید که یک عدم تطابق جزئی وجود دارد که در آن ‘کد اشتراک’ به stacking در دیکشنری نگاشت شده است، اما پس از فرآیند ‘Subscription Code’

‘subscrition code‘ و تغییر نام در برخی سلول‌ها (مانند سلول ۲۳) به صورت ظاهر می‌شود. به نظر می‌رسد این یک اثر جانبی df_long در (حروف کوچک) و مرحله تغییر نام بعدی باشد. برای حفظ یکپارچگی در مستندات، از نام reset_index استفاده می‌شود مگر اینکه به صراحت به نام ‘Subscription Code‘ استاندارد اشاره شود df_long موقت در خروجی.

5. **مدیریت داده‌های اشتباه (مقادیر منفی):**

- بر اساس دانش دامنه، مصرف آب و ساعت کارکرد منفی نامعتبر هستند.
- ردیف‌های در DataFrames ۱۴۰۲ و ۱۴۰۳ (df_info_1402, df_info_1403) اطلاعاتی برای منفی است، ‘Consumption in Period (m^3)‘ که در آن‌ها سلول‌های ۹۲ و ۹۳ خروجی‌هایی برای ردیف‌های مصرف منفی در) شناسایی می‌شوند (اطلاعاتی ۱۴۰۲ و ۱۴۰۳ نشان می‌دهند DataFrames.
- متناظر با این رکوردهای اشتباه در لیست‌هایی ‘Subscription Code‘ مقادیر (customers_to_drop_1402, customers_to_drop_1403) جمع‌آوری می‌شوند.
- DataFrames (df_info_clean_1402, df_info_clean_1403) اطلاعاتی پاکسازی شده آن‌ها در این لیست‌ها قرار دارد، ‘Subscription Code‘ با فیلتر کردن ردیف‌هایی که ایجاد می‌شوند (سلول ۹۴). این مورد جدی‌ترین مشکل کیفیت داده در اطلاعات خلاصه را برطرف می‌کند.
- و ‘Consumption (m^3)‘ برای (df_long) مقادیر منفی در داده‌های ماهانه مصرف شناسایی می‌شوند (سلول‌های ۳۳، ۳۴). اگرچه این موارد به ‘Operating Hours (h)‘ عنوان ناهنجاری لیست شده‌اند، اما نوت‌بوک آن‌ها را *شناسایی* می‌کند ولی در *حذف نمی‌کند* قبل از ذخیره داده‌ها df_long اسنیپت‌های ارائه شده صراحتاً آن‌ها را از ایده‌های تحلیل اشاره به بررسی/مدیریت این موارد دارند .Long در قالب.

6. **ذخیره داده‌های پاکسازی شده/پردازش شده:**

- اطلاعاتی پاکسازی شده DataFrames (df_info_clean_1402, df_info_clean_1403) ذخیره می‌شوند CSV به صورت فایل‌های (1402_clean_info.csv, 1403_clean_info.csv) (سلول ۹۶).
- DataFrames پردازش شده در قالب Long (df_long ۱۴۰۱، برای df_long_1402, df_long_1403) ذخیره می‌شوند CSV به صورت فایل‌های (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 تمایل بینشی در مورد 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 'Percentage 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 (m^3)' 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 df_long در (مقادیر منفی برای استحکام محاسبه اعمال می‌شود (سلول ۴۹). تعداد موارد پرت یافت می‌شود قابل توجهی.
- برای تولید بصری‌سازی برای تشخیص InteractiveDataAnalyzer از متدهای کلاس parallel ،همبستگی heatmap) توزیع (هیستوگرام) و روابط ،(Z امتیاز) ناهنجاری coordinates) روی df_info_clean استفاده می‌شود (سلول‌های ۹۸، ۹۹، ۱۰۱). این 'Consumption' تحلیل بصری، کجی توزیع‌ها و وجود موارد پرت در معیارهایی مانند 'Average Flow Rate in Period (l/s)' و 'in Period (m³)' 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. * مقادیر منفی در مصرف و ساعت کارکرد ماهانه رایج هستند. * داده‌های گمشده در ستون‌های معیارهای ماهانه 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 (analyzer.py و app.py)

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

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

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

به عنوان رابط کاربری تعاملی برای پروژه عمل می‌کند. * به کاربران Streamlit اپلیکیشن **app.py:** را بارگذاری کنند. * پس از بارگذاری فایل‌ها، یک نوار CSV امکان می‌دهد یک یا چند فایل اکسل یا کناری ظاهر می‌شود که به کاربران امکان انتخاب یک مجموعه داده از فایل‌های بارگذاری شده را پانداس بارگذاری می‌شود. * یک شیء DataFrame می‌دهد. * فایل انتخاب شده در یک بارگذاری شده نمونه‌سازی می‌شود. * نوار کناری DataFrame با InteractiveDataAnalyzer گزینه‌های پیکربندی را ارائه می‌دهد: * انتخاب یک ستون عددی. * انتخاب یک ستون دسته‌ای. * ‘None’, ‘Histogram’, ‘Boxplot’, شامل) انتخاب یک نوع نمودار از لیست از پیش تعریف شده ‘Line Chart’, ‘Outlier Detection’, ‘Pie Chart’, ‘Bar Chart’, ‘Correlation Heatmap’). * بر اساس نوع نمودار و ستون‌های انتخاب شده، اپلیکیشن متد مربوطه را از شیء

نمایش st.plotly_chart را با استفاده از Plotly فراخوانی کرده و نمودار InteractiveDataAnalyzer 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 (analyzer.scale_column) شده نمایش می‌دهد. * DataFrame را می‌دهد و نتیجه را در یک 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 پاکسازی شده را بارگذاری می‌کند یا خود برنامه داده‌های CSV با فرض اینکه کاربر فایل‌های) نوت‌بوک برای پشتیبانی از بصری‌سازی‌ها و ویژگی‌های خود (پاکسازی شده را به صورت داخلی بارگذاری می‌کند استفاده می‌کند.

7. اجرای پروژه

و تعامل با داده‌ها: 1. اطمینان حاصل کنید که پایتون و کتابخانه‌های Streamlit برای اجرای اپلیکیشن (streamlit, pandas, plotly, seaborn, matplotlib, scipy, scikit-learn, missingno) لازم نصب شده‌اند. می‌توانید آن‌ها را با استفاده از pip نصب کنید (در صورت نیاز برای نوت‌بوک): bash

```
pip install streamlit pandas plotly seaborn matplotlib scipy scikit-learn missingno openpyxl
```

2. در یک دایرکتوری analyzer.py و app.py کدهای ارائه شده را در فایل‌هایی با نام‌های مانند long_usage1401.csv, 1402_clean_info.csv و ...) ذخیره کنید. 3. فایل‌های داده پاکسازی شده را در مکانی (اصلی، بسته به اینکه قصد دارید برنامه چه فایلی را بارگذاری کند xlsx. غیره - یا فایل‌های Command که توسط اسکریپت قابل دسترسی باشد (مثلاً همان دایرکتوری) قرار دهید. 4. ترمینال یا 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

را نمایش می‌دهد. از قسمت بارگذاری فایل برای آپلود فایل‌های Streamlit خودکار باز شده و داشبورد .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)، جایگزینی (imputation) فراتر از صرفاً شناسایی. این می‌تواند شامل یا پرچم‌گذاری برای تحلیل جداگانه بر اساس تخصص دامنه (capping) محدود کردن .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)
  - .Long اطلاعاتی و DataFrames ترکیب موثر ویژگی‌ها از
- **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 ابزار ارزشمندی را برای کاوش و بصری‌سازی تعاملی ایجاد می‌کند و داده‌ها را قابل دسترس‌تر و قابل فهم‌تر می‌سازد. یافته‌های تحلیل EDA (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.
