EDA + STATS PROJECT

Task Description

- Analyze the data and extract useful insights.
- Validate the below assumptions using the corresponding statistical tests:
    - Does the region significantly affect energy consumption?
    - Is there a relationship between the time of day and energy consumption?
- Assuming that we can shutdown few cells were the consumption is low, what is the threshold that you would recommend and how much KW and EGP it would save accordingly?

Data Description

1. Energy Consumption.csv:

This dataset contains energy consumption data collected from network sites, to help us monitor and track sites' energy consumption, the dataset has 5 columns:

- **DateTime**
- **Site_id:** unique site identifier contains more than one cell.
- **Cell_id**: unique cell identifier.
- **Region:** geographical area that identifies site location.
- **KWH/hh (per half hour):** Energy consumed in half hour intervals.

2. Power Demand.csv:

This dataset contains energy consumption demand every half hour interval, the dataset has 2 columns:

- **DemandDateTime**
- **Demand:** energy consumption demand for the time interval

Demand prices for calculating cost:

- High (77.21p/kWh)
- Normal (19.46p/kWh)
- Low (6.89p/kWh)

Deliverables Expectations

1. **Power point file:**
    - Your methodology of carrying out the analysis.
    - Challenges and how did you overcome it.
    - Pre-modeling insights.
    - Post-modeling insights.

2. **Share your code in a clean and organized way (Public Google Colab Link or Jupyter Notebook is preferred).**

3. **A Dashboard (using any suitable tool)** that can be used by management to understand energy costs using tariffs table. **(Optional – Bonus)**

Scoring Rubric

Data Exploration and Analysis (EDA) (25 points)

- **Data Cleaning and Preparation (5 points):**
  - Properly handling missing values, outliers, and inconsistencies in the data.
  - Transforming variables as needed (e.g., creating time-based features).
  - **Scoring:**
    - 5 points: Excellent data cleaning and preparation with clear explanations.
    - 3-4 points: Good data cleaning with minor issues or missing explanations.
    - 1-2 points: Basic data cleaning with significant issues or lack of clarity.
    - 0 points: Inadequate or no data cleaning.
- **Exploratory Analysis (10 points):**
  - Creating informative visualizations (histograms, scatter plots, box plots, etc.) to illustrate key findings.
  - Providing clear and concise interpretations of the exploratory analysis results.
  - **Scoring:**
    - 10 points: Excellent EDA with insightful visualizations and interpretations.
    - 7-9 points: Good EDA with relevant visualizations and interpretations.
    - 4-6 points: Basic EDA with some visualizations and interpretations.
    - 1-3 points: Limited EDA with minimal visualizations and interpretations.
    - 0 points: Inadequate or no EDA.
- **Feature Engineering (5 points):**
  - Creating new relevant features from existing ones (e.g., time of day categories, weekend/weekday indicators) to improve analysis and modeling.
  - **Scoring:**
    - 5 points: Excellent feature engineering with creative and relevant new features.
    - 3-4 points: Good feature engineering with some useful new features.
    - 1-2 points: Basic feature engineering with limited new features.
    - 0 points: Inadequate or no feature engineering.
- **Data Insights (5 points):**
  - Extracting meaningful and actionable insights from the data exploration process.
  - Clearly communicating the insights and their potential implications.
  - **Scoring:**
    - 5 points: Excellent insights that are relevant, actionable, and well-explained.
    - 3-4 points: Good insights with some relevance and clarity.
    - 1-2 points: Basic insights with limited relevance or clarity.
    - 0 points: Inadequate or no insights.

Statistical Hypothesis Testing (25 points)

- **Hypothesis Formulation (5 points):**
  - Clearly stating the null and alternative hypotheses for each research question.
  - **Scoring:**
    - 5 points: Clear and accurate hypotheses for both research questions.
    - 3-4 points: Mostly accurate hypotheses with minor issues.

        - 1-2 points: Vague or incorrect hypotheses.
        - 0 points: Missing or completely incorrect hypotheses.
- **Test Selection and Application (10 points):**
    - Selecting the appropriate statistical tests (e.g., t-test, ANOVA, correlation) for each hypothesis.
    - Correctly applying the tests and interpreting the results (p-values, confidence intervals).
    - **Scoring:**
        - 10 points: Correct test selection and application with accurate interpretation of results.
        - 7-9 points: Mostly correct test selection and application with minor errors in interpretation.
        - 4-6 points: Some correct test selection and application with significant errors in interpretation.
        - 1-3 points: Incorrect test selection or application with flawed interpretation.
        - 0 points: No attempt to perform hypothesis testing.
- **Assumptions and Limitations (5 points):**
    - Checking the assumptions of the statistical tests.
    - Discussing any limitations of the analysis or the chosen tests.
    - **Scoring:**
        - 5 points: Thoroughly addresses assumptions and limitations.
        - 3-4 points: Addresses most assumptions and limitations with minor gaps.
        - 1-2 points: Briefly mentions assumptions and limitations with significant gaps.
        - 0 points: No discussion of assumptions or limitations.
- **Clarity and Presentation of Results (5 points):**
    - Presenting the hypothesis testing results in a clear and concise manner.
    - Using tables and visualizations to effectively communicate the findings.
    - **Scoring:**
        - 5 points: Clear and well-organized presentation of results with effective visualizations.
        - 3-4 points: Mostly clear presentation with some room for improvement.
        - 1-2 points: Confusing or poorly organized presentation of results.
        - 0 points: Results are not presented or are incomprehensible.

Threshold Recommendation and Cost Savings (20 points)

- **Threshold Determination (10 points):**
    - Developing a logical and data-driven approach to determine the threshold for shutting down cells.
    - Justifying the chosen threshold based on the analysis and insights.
    - **Scoring:**
        - 10 points: Well-justified and data-driven threshold with a clear explanation of the methodology.
        - 7-9 points: Reasonable threshold with some justification based on the analysis.
        - 4-6 points: Threshold with weak justification or limited connection to the analysis.
        - 1-3 points: Arbitrary threshold with no clear justification.
        - 0 points: No attempt to determine a threshold.
- **Cost Savings Calculation (10 points):**
    - Accurately calculating the potential KW and EGP savings based on the chosen threshold.
    - Considering the different demand prices (high, normal, low) in the calculations.

- **Scoring:**
  - 10 points: Accurate and detailed cost savings calculations considering different demand prices.
  - 7-9 points: Mostly accurate calculations with minor errors or omissions.
  - 4-6 points: Significant errors or omissions in the cost savings calculations.
  - 1-3 points: Flawed or incomplete calculations.
  - 0 points: No attempt to calculate cost savings.

Communication and Presentation (30 points)

- **PowerPoint Presentation (15 points):**
  - Clear and concise presentation of the methodology, challenges, and insights.
  - Effective use of visualizations to communicate the findings.
  - Professional and well-organized presentation.
  - **Scoring:**
    - 15 points: Excellent presentation that is clear, concise, and engaging.
    - 11-14 points: Good presentation with minor areas for improvement.
    - 7-10 points: Adequate presentation with some organizational or clarity issues.
    - 3-6 points: Poorly organized or confusing presentation.
    - 0 points: No presentation or a presentation that is incomprehensible.
- **Code Quality (10 points):**
  - Well-structured and commented code.
  - Use of appropriate coding practices and libraries.
  - Easy to follow and understand.
  - **Scoring:**
    - 10 points: Excellent code quality that is well-structured, commented, and efficient.
    - 7-9 points: Good code quality with minor areas for improvement.
    - 3-6 points: Code is functional but poorly structured or difficult to understand.
    - 0-2 points: Code is not provided or is not functional.
- **Dashboard (5 points - Bonus):**
  - Creates a functional and user-friendly dashboard to visualize energy costs.
  - Effectively uses the provided tariffs table in the dashboard.
  - **Scoring:**
    - 5 points: Excellent dashboard that is insightful, user-friendly, and well-designed.
    - 3-4 points: Good dashboard with some areas for improvement in design or functionality.
    - 1-2 points: Basic dashboard with limited functionality or design flaws.
    - 0 points: No dashboard or a dashboard that is not functional.

Total Points: 100
Grading Scale:

- 90-100: A
- 80-89: B
- 70-79: C
- 60-69: D
- Below 60: F