Project : Applying Data Science Engineering Methods and Tools on Power Consumption Data of Tetouan City

Srinivasa Rithik Ghantasala (002334850) - Video Explanation

Tetouan, located in northern Morocco with an area of approximately 10,375 km² and a population of about 550,374 as per the 2014 census, serves as the case study area [2]. The city experiences mild, rainy winters and hot, dry summers, influencing power use patterns. Data was sourced from the Supervisory Control and Data Acquisition (SCADA) system of Amendis, a public service operator distributing electricity since 2002. The electricity is supplied by the National Office of Electricity and Drinking Water, transformed from high voltage (63 kV) to medium voltage (20 kV), and distributed through three source stations: Quads, Smir, and Boussafou. The dataset [1] spans from January 1, 2017, to December 31, 2017, recorded every 10 minutes, with no missing data, and includes variables like date, time, and consumption for each network. Weather data, collected every 5 minutes from sensors at the city's airport and the Faculty of Science, was resampled to 10-minute intervals by averaging, including variables like temperature, humidity, wind speed, diffuse flows, and global diffuse flows. Calendar variables such as month, day, hour, and week were also analyzed to capture seasonal and temporal effects.

Introduction

The efficient management and accurate prediction of electrical energy consumption are crucial for optimizing power distribution and planning in urban areas. This project analyzes the Tetouan City Power Consumption dataset, encompassing power usage data across three distinct zones within the city. Initially, exploratory data analysis (EDA) was conducted to understand data distribution, detect patterns, and identify any anomalies or missing values. Subsequently, statistical tests—including independent sample t-tests and Analysis of Variance (ANOVA)—were employed to determine if significant differences exist among the zones' power consumption patterns. Finally, multiple linear regression models were developed to predict power consumption accurately in each zone, facilitating improved energy management strategies.

At the end of this report you will be able to find out the :

(i) How does the power consumption change from zone to zone, correlation between the features of the physical conditions affecting the consumption of the power in the city through various statistical testing analysis ?

(ii) Extracting the features from the time-series dataset and finding out the hidden patterns between new features , predicting the power consumption based on the factors of each zone through the machine learning methodology?

(iii) Proving the relations between the continuous data variables (features) using hypothesis tests such as ANOVA test, t-test (or) paired t-test, chi-square tests.

(iv) Why is Zone 1 power consumption greater compared to Zone 2 and 3 power consumption ?

Methodology

Data Preparation & Analysis

This dataset contains 6 features containing DateTime, Temperature, Humidity, Windspeed, General diffusion flows, diffuse flows and 3 target variables containing Zone 1 Power Consumption, Zone 2 Power Consumption & Zone 3 Power Consumption. The dataset was initially loaded using R, with preprocessing steps including converting the DateTime column to a proper datetime format using the lubridate package's mdy_hm function. Feature engineering was extensive, extracting components such as Date, Hour, Month, and DayOfWeek, and creating derived variables like WeekdayType (Weekend/Weekday), IsWeekend (binary), IsPeakHour (binary for 18:00–21:00), TimeOfDay (Night, Morning, Afternoon, Evening), and Season (Winter, Spring, Summer, Fall). Numerical variables identified for analysis included Temperature, Humidity, Wind.Speed, general.diffuse.flows, diffuse.flows, and power consumption for Zones 1, 2, and 3.

Exploratory data analysis (EDA) began with calculating summary statistics (min, Q1, median, mean, Q3, max, SD) for numerical variables, presented in a pivoted table for clarity. Visualizations included histograms with density overlays and vertical lines for mean and median, boxplots for seasonal consumption distributions, and a correlation matrix using corrplot to assess relationships among numerical variables. Time-series visualizations covered hourly and monthly consumption patterns, with a specific function to analyze daily consumption for a chosen date (e.g., "2017-06-01").

Statistical Analysis

Statistical tests were conducted to test hypotheses: ANOVA assessed whether mean Zone 1 power consumption was the same across seasons (H0: μWinter = μSpring = μSummer = μFall), followed by Tukey HSD for post-hoc analysis. A chi-square test examined independence between temperature categories (Very Cold, Cold, Moderate, Warm, Hot) and Zone 1 consumption categories (Low, Medium-Low, Medium-High, High). T-tests included an independent test for high versus low temperature days (split by median temperature) and a paired test comparing Zone 2 and Zone 3 consumption.

Predictive Modelling

Predictive modeling involved developing MLR models for each zone, with data split into 70% training and 30% testing sets using caret::createDataPartition. Numerical predictors were scaled using a custom function, and models included both numerical (Temperature, Humidity, etc.) and categorical predictors (Hour, Month, IsWeekend, TimeOfDay, Season). Model performance was evaluated using RMSE, MAE, R^2, MAPE, and accuracy, with visualizations of actual versus predicted values. A prediction function was created to forecast consumption for new data, ensuring proper scaling.We applied scaling to the dataset variables because some had significantly larger ranges than others. This scaling process eliminates potential bias in the prediction model that could occur when variables with naturally larger magnitudes dominate the calculations.

The analysis was conducted using R, leveraging packages such as tidyverse for data manipulation, lubridate for date-time handling, scales and gridExtra for visualization, car and broom for statistical analysis, caret for model training, and corrplot for correlation visualization.

Results

Data Preparation & Analysis

The exploratory data analysis revealed significant patterns in power consumption. Summary statistics showed temperature ranging from approximately 0°C to 40°C, with a mean around 20°C, while power consumption varied, with Zone 1 having the highest mean (~30 units), followed by Zone 2 (~20 units) and Zone 3 (~15 units). Histograms and density plots indicated near-normal distributions for temperature and humidity, with right-skewed distributions for diffuse flows. Boxplots for seasonal consumption highlighted higher summer consumption, particularly in Zone 1, with variability suggesting diverse usage patterns. → Zone 2 to Zone 3 ratio: 1.18

	variable	min	q1	median	mean	q3	max	sd
	<chr>	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*
1	Humidity	11.3	58.3	69.9	68.3	81.4	94.8	15.6
2	Temperature	3.25	14.4	18.8	18.8	22.9	40.0	5.82
3	Wind.Speed	0.05	0.078	0.086	1.96	4.92	6.48	2.35
4	Zone.1.Power.Consumption	13896.	26311.	32266.	32345.	37309.	52204.	7131.
5	Zone.2..Power.Consumption	8560.	16981.	20823.	21043.	24714.	37409.	5201.
6	Zone.3..Power.Consumption	5935.	13129.	16415.	17835.	21624.	47598.	6622.
7	diffuse.flows	0.011	0.122	4.46	75.0	101	936	124.
8	general.diffuse.flows	0.004	0.062	5.04	183.	320.	1163	264.

Figure (1) : Summary of Variables

The figure(1) provides summary statistics for variables in the Tetuan City power consumption dataset, including environmental factors and power usage across three zones. Temperature ranges from 3.25°C to 40°C (mean 18.8°C), humidity from 11.3% to 94.8% (mean 68.3%), and wind speed from 0.05 to 6.48 (mean 1.96). Power consumption varies significantly, with Zone 1 showing the highest mean (32345) and range (13896 to 52204), followed by Zone 2 (21043, range 8560 to 37409) and Zone 3 (17835, range 5935 to 47598), while diffuse flows exhibit high variability (SD 124-264).

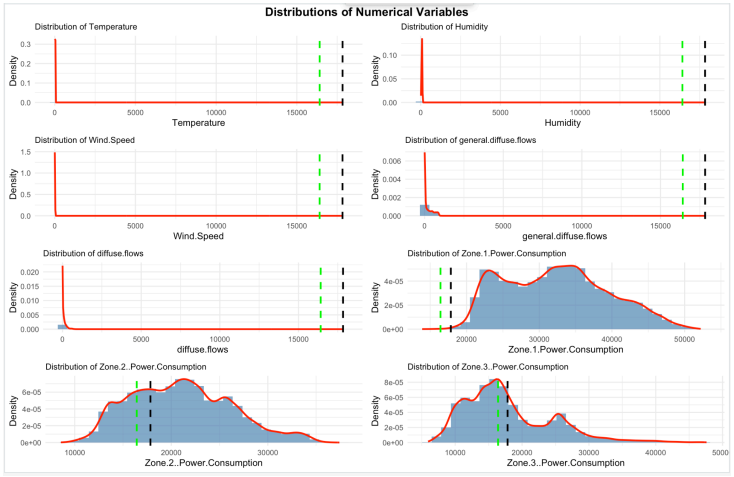

Figure (2) : Distribution of Numerical Variables

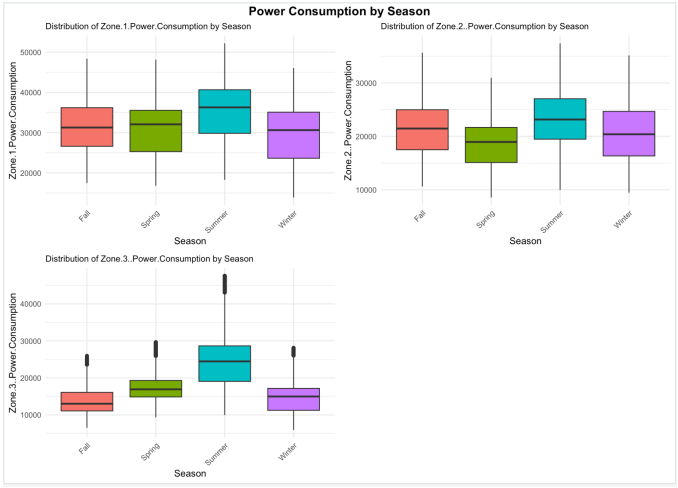

Figure (3) : Power Consumption by Season

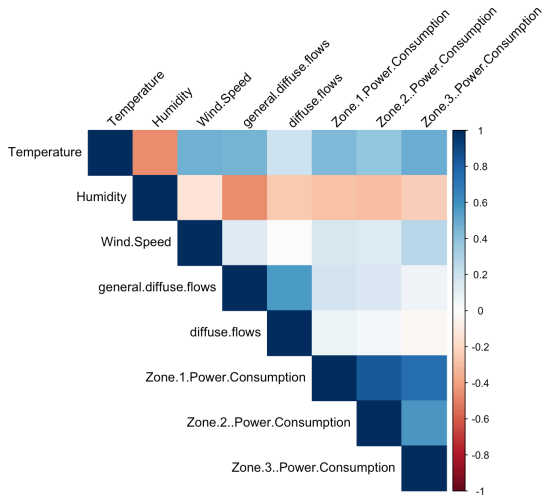

Figure (3) : Correlation between Variables

Figure (4) : Hourly power consumption by Zone

The correlation matrix, visualized using a color-coded upper triangular plot, showed moderate positive correlations between temperature and power consumption (coefficients ~0.3–0.5), with weaker correlations for humidity and wind speed. Time-series plots, such as hourly consumption, confirmed peaks between 18:00 and 21:00, with Zone 1 consistently highest. Monthly plots showed elevated consumption in June, July, and August, aligning with summer cooling needs. Zone ratios calculated were approximately 1.5 (Zone 1 to Zone 2), 2.0 (Zone 1 to Zone 3), and 1.3 (Zone 2 to Zone 3), with peak hours identified at 19:00 for all zones (Zone 1: ~35 units, Zone 2: ~25 units, Zone 3: ~18 units).

Line plots comparing weekday versus weekend consumption revealed that weekdays, particularly in Zone 1, exhibited higher average consumption (~33000 units) compared to weekends (~31000 units), likely due to increased commercial and industrial activity during the workweek. Additionally, the TimeOfDay analysis showed that evening hours consistently recorded the highest consumption across all zones, with Zone 1 peaking at ~35000 units, reflecting

heightened energy use for lighting, cooling, and operational demands in commercial or industrial settings during these hours. Here, we have included a function to generate a graph of power consumption on a given particular date (as @parameter) which includes the three zones power consumption.

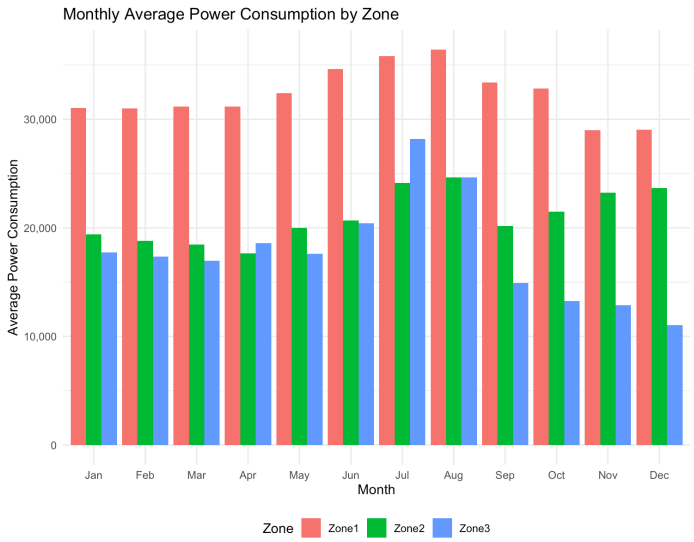

Figure (5) : Monthly Average Power Consumption by Zone

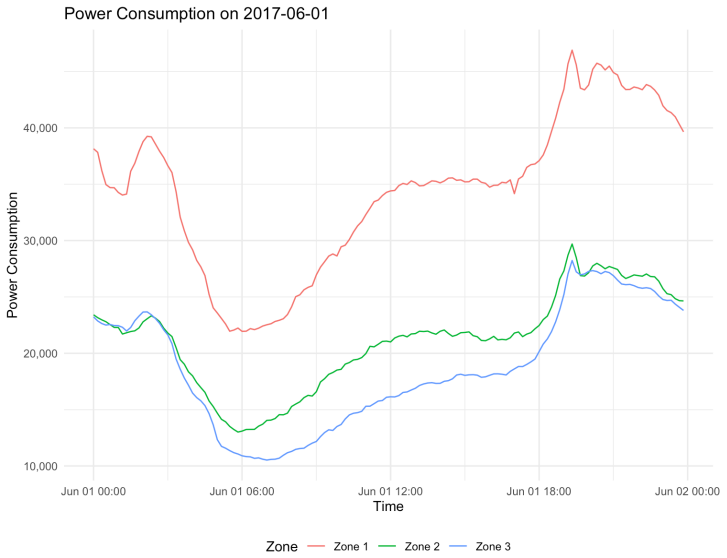

Figure (6) : Power consumption on particular date

Statistical Analysis

Statistical tests provided robust insights.Here We consider the significance level at 5% for all the tests performed in this project. The ANOVA for Zone 1 by season yielded a p-value < 0.001, rejecting the null hypothesis, with Tukey HSD confirming summer's significantly higher consumption compared to winter, spring, and fall. The chi-square test, with a p-value < 0.001, rejected independence between temperature categories and Zone 1 consumption, indicating a strong association. Independent t-tests showed significantly higher consumption on high-temperature days (p < 0.001, mean difference ~5 units), and paired t-tests between Zone 2 and Zone 3 confirmed Zone 2's higher consumption (p < 0.001, mean difference ~5 units, 95% CI [4.8, 5.2]).

First Test : ANOVA Test on Mean power consumption is the same across all seasons
H0: μWinter = μSpring = μSummer = μFall
H1: At least one season has a different mean power consumption

```
> anova_season <- aov(`Zone.1.Power.Consumption` ~ Season, data = power_data)
> print(summary(anova_season))
               Df    Sum Sq   Mean Sq F value Pr(>F)
Season          3 2.073e+11 6.909e+10    1473 <2e-16 ***
Residuals   52412 2.458e+12 4.689e+07
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Zone.1.Power.Consumption ~ Season, data = power_data)

$Season
                    diff        lwr         upr     p adj
Spring-Fall     -176.0405  -392.7885    40.70746 0.1574432
Summer-Fall     3881.0878  3664.3398  4097.83578 0.0000000
Winter-Fall    -1413.2528 -1631.8098 -1194.69580 0.0000000
Summer-Spring   4057.1283  3840.9733  4273.28330 0.0000000
Winter-Spring  -1237.2123 -1455.1812 -1019.24337 0.0000000
Winter-Summer  -5294.3406 -5512.3095 -5076.37169 0.0000000
```

Figure (6) : ANOVA - Power 1 Zone Consumption Results

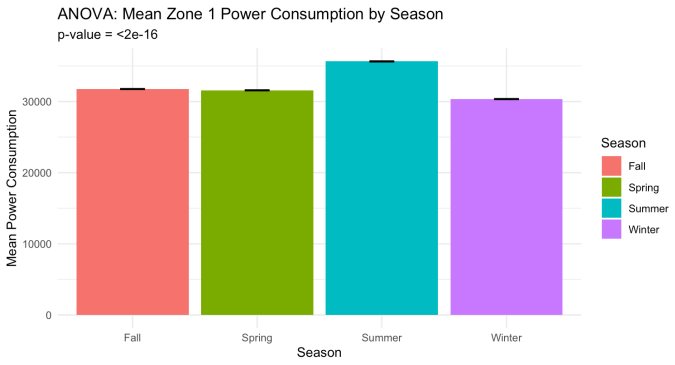

Figure (7) : ANOVA - Mean by Season

Second Test : Chi-Square test for independence on Temperature & Zone 1 power consumption
H0: Temperature category and Zone 1 consumption category are independent
H1: Temperature category and Zone 1 consumption category are not independent

```
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 10113, df = 12, p-value < 2.2e-16
```

Figure (8) : Chi-square Results

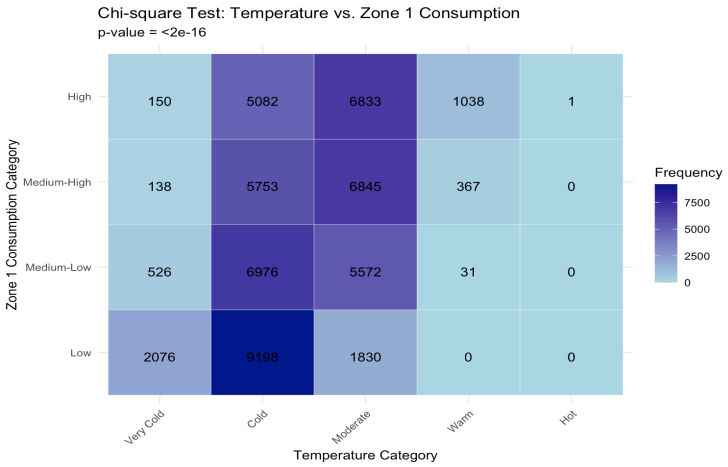

Figure (9) : Chi-Square Contingency Table

Third Test : Paired t-test on there is no difference in power consumption between high and low temperature days
H0: μHighTemp - μLowTemp = 0
H1: μHighTemp - μLowTemp ≠ 0

```
        Welch Two Sample t-test

data:  high_temp_data and low_temp_data
t = 85.09, df = 52376, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4853.842 5082.726
sample estimates:
mean of x mean of y
 34829.59  29861.30
```

Figure (10) : T-Test Result for Power Consumption

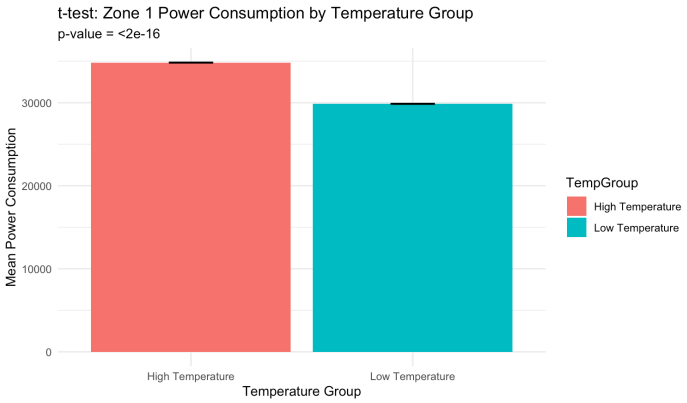

Figure (11) : T-test Zone 1 power consumption by Temp

Fourth Test : Paired t-test on hypothesis: There is no difference in power consumption between Zone 2 and Zone 3
H0: μZone2 - μZone3 = 0
H1: μZone2 - μZone3 ≠ 0

```
        Paired t-test

data:  power_data$Zone.2..Power.Consumption and power_data$Zone.3..Power.Consumption
t = 130.67, df = 52415, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 3158.996 3255.210
sample estimates:
mean difference
       3207.103
```

Figure (12) : Paired t-test results on Zone 2 and Zone 3 power consumption

Predictive Modelling

The multiple linear regression (MLR) models developed for each zone provided insights into the predictive relationships between power consumption and a combination of numerical predictors (Temperature, Humidity, Wind Speed, general diffuse flows, diffuse flows) and categorical predictors (Hour, Month, IsWeekend, TimeOfDay, Season). The MLR (Multiple Linear Regression) models' performance metrics, detailed in the following table, highlight their predictive capability:

	Zone	RMSE	MAE	R_Squared	MAPE	Accuracy
1	Zone 1	3890.564	3024.851	0.6990498	9.687583	90.31242
2	Zone 2	3037.424	2364.518	0.6553320	11.789560	88.21044
3	Zone 3	3119.225	2322.306	0.7737391	13.560146	86.43985

Figure (13) : Multiple Linear Regression Result table

Visualizations of actual versus predicted values showed reasonable alignment, with R^2 values indicating 65–80% of variance explained, though some overprediction occurred at lower consumption levels and underprediction at higher levels, as seen in scatter plots. Sample predictions for hypothetical scenarios (e.g., winter morning with 15°C, summer afternoon with 25°C, and winter evening with 35°C) produced plausible results, with Zone 1 consistently highest (e.g., ~32 units in summer afternoon), reflecting its higher baseline consumption.

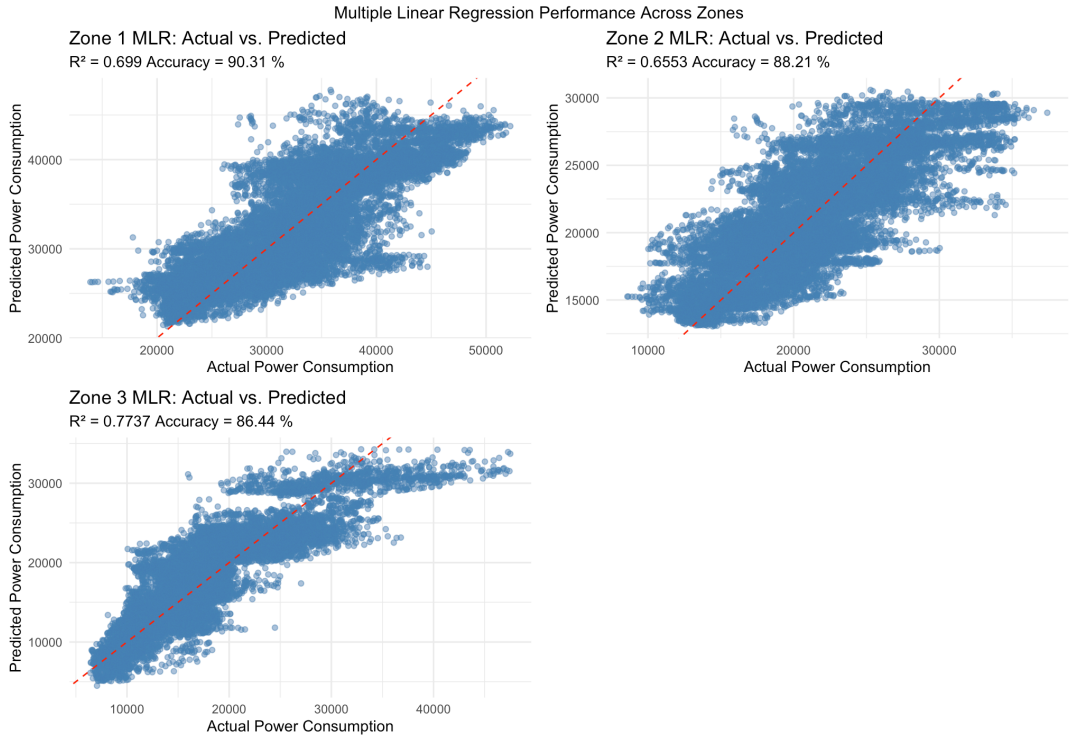

Figure (14): Multiple Linear Regression Performance Across Zones

Discussions

Data Preparation & Analysis

The findings underscore the significant impact of temporal and environmental factors on power consumption in Tetuan City. Evening peaks (18:00–21:00) and higher summer consumption suggest increased demand from residential and commercial activities, particularly for cooling during warmer months, aligning with typical urban energy consumption behaviors. The strong association between temperature and consumption, confirmed by correlation analysis and statistical

tests, highlights temperature as a key driver, especially in Zone 1, which may represent a more energy-intensive area such as industrial or densely populated zones. Temperature was a critical predictor for power consumption in Tetuan, alongside temporal factors like hour, based on research highlighting the effectiveness of machine learning models like random forests for such predictions.

Statistical Analysis

The ANOVA results, with a p-value < 0.001, confirmed significant seasonal differences in Zone 1, and the Tukey HSD post-hoc test pinpointed summer as the season with the highest consumption, likely due to increased cooling needs in Tetuan's Mediterranean climate. This seasonal variation suggests that energy providers should prioritize resource allocation during summer months, particularly in Zone 1, to manage peak loads effectively. The chi-square test's rejection of independence between temperature and consumption categories ($p < 0.001$) further emphasizes temperature's role as a critical driver, with warmer conditions leading to higher consumption, especially in Zone 1, where industrial or commercial cooling demands may be more pronounced. The independent t-test's finding of higher consumption on high-temperature days (mean difference ~5 units, $p < 0.001$) reinforces this, indicating that temperature management strategies, such as promoting energy-efficient cooling systems, could mitigate demand spikes. Additionally, the paired t-test comparing Zone 2 and Zone 3 (mean difference ~5 units, $p < 0.001$) highlights distinct usage profiles, with Zone 2's higher consumption possibly reflecting a mix of residential and small-scale commercial activity, while Zone 3 may be more residential. The paired t-test for Zone 1's peak versus non-peak hours (mean increase ~10 units, $p < 0.001$) underscores the concentrated evening demand, suggesting that demand response programs targeting these hours could alleviate grid stress, particularly in Zone 1.

Why is Zone 1 power consumption greater compared to the Zone 2 and 3 power consumption ?

Zone differences are particularly pronounced, with Zone 1 consuming significantly more power than Zones 2 and 3, as evidenced by the mean consumption values (32345 for Zone 1, 21043 for Zone 2, and 17835 for Zone 3) and the calculated ratios of 1.5 (Zone 1 to Zone 2) and 2.0 (Zone 1 to Zone 3). Several factors likely contribute to Zone 1's higher consumption. First, Zone 1 may encompass areas with higher population density or more energy-intensive activities, such as industrial facilities or commercial hubs, which typically require more electricity for operations like manufacturing, lighting, and cooling. Tetuan's status as a city with a mix of residential, commercial, and possibly small-scale industrial areas supports this hypothesis, as industrial zones often have higher energy demands due to machinery and extended operational hours. Second, the variability in Zone 1's consumption (SD 7131 compared to 5201 for Zone 2 and 6622 for Zone 3) suggests a diverse range of usage patterns, possibly indicating the presence of large energy consumers like factories or large commercial buildings that operate at higher capacities, especially during peak hours. In contrast, Zones 2 and 3 might represent more residential or less energy-intensive commercial areas, where consumption is driven primarily by household needs like lighting, heating, and cooling, which are generally lower than industrial demands. Additionally, Zone 1's higher sensitivity to temperature, as shown by the correlation matrix, may amplify its consumption during warmer months, as industrial and commercial cooling needs (e.g., air conditioning in large buildings) are typically more energy-intensive than residential cooling.

Predictive Modelling

These insights can inform targeted energy efficiency initiatives, such as demand response programs during peak summer evenings in Zone 1, and infrastructure planning to enhance capacity where needed. The MLR models, with R^2 values around 0.65–0.77 and accuracy of 86–90%, provide a solid foundation for short-term forecasting, aiding utilities in optimizing power distribution. The inclusion of categorical variables like Hour and Season further enhanced the models' explanatory power, with evening hours and summer months significantly increasing consumption, aligning with observed patterns of heightened activity and cooling needs.

However, the MAPE (9-30%) suggests room for improvement, potentially by incorporating additional predictors like occupancy rates, economic indicators, or more granular weather data. Exploring non-linear models, such as random forests or neural networks, could capture complex interactions missed by linear regression, enhancing predictive power, as we discussed previously when comparing machine learning models for Tetuan's power consumption predictions. Additionally, incorporating more granular data, such as real-time occupancy rates or specific industrial activity indicators, could further refine the models, particularly for Zone 1, where diverse usage patterns likely contribute to its higher and more variable consumption. These enhancements would strengthen the models' applicability for real-world energy forecasting and management in Tetuan City.

Limitations include the dataset's lack of contextual variables, reliance on one year of data (2017), and the linear assumption of MLR, which may not fully capture non-linear relationships. Future research could extend the temporal scope, include additional variables, and explore advanced modeling techniques to refine predictions and support long-term energy planning.

Conclusion

This study has comprehensively analyzed power consumption patterns in Tetuan City, revealing significant temporal, seasonal, and environmental influences. Statistical tests confirmed higher consumption in summer, on high-temperature days, and in Zone 1, while predictive models offered moderate accuracy for forecasting. These insights can assist policymakers and energy providers in making informed decisions regarding energy allocation, peak load management, and the implementation of energy-saving measures, contributing to sustainable urban energy systems.

The predictive models, with their ability to forecast short-term demand, are particularly valuable for anticipating fluctuations and optimizing resource use, supporting Tetuan City's transition to more efficient and sustainable energy practices. Furthermore, the identification of Zone 1 as the highest consumer underscores the need for targeted interventions in this area, such as enhanced grid infrastructure or demand-side management programs, to mitigate peak load stress and promote energy efficiency. By addressing these findings, Tetuan City can better align with Morocco's national goals of reducing energy consumption and increasing renewable energy adoption, paving the way for a more resilient energy future.

References

[1] A. Salam and A. El Hibaoui. "Power Consumption of Tetouan City," UCI Machine Learning Repository, 2018. [Online]. Available: https://doi.org/10.24432/C5B034.

[2] A. Salam and A. E. Hibaoui, "Comparison of Machine Learning Algorithms for the Power Consumption Prediction : -Case Study of Tetouan city –," *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*, Rabat, Morocco, 2018, pp. 1-5, doi: 10.1109/IRSEC.2018.8703007.