Business Report

Team Name : Chocolate chip cookies

Team Members : Armand Hubler, Aswin Subramanian Maheswaran, Emili Khachatryan, Riyad Mazari and Ronald Beltran.

1. About Data

ACCIONA's Sustainability Datathon challenges us to create a 7-day predictive model for Villarubia's water consumption by sector. The goal is to reduce water loss, prevent environmental damage, and avoid unexpected shortages, aligning with the Sustainable Development Goals for a sustainable and resilient water supply. To tackle ACCIONA's challenge, the biggest step was to understand and interpret the data as we had six datasets corresponding to different aspects of the water cycle in Villarubia.

During data analysis, we focused on “Caudales.csv”, a key dataset detailing measurement levels, consumption times, and city sectors. We connected the “bissioCode” from “gis.csv” to “Sector_Neta” in “Caudales.csv” to map out sectors. As suppliers can be assigned to several zones, we considered the highest hierarchical level as the dominant one. Our exploratory analysis classified sectors using “gis.csv”, which identifies canonical types for each of them by their geographical data. Recognizing “Totalizador” as a quality benchmark, we formulated a 'Zones' class encompassing Villarubia's various sectors.

2. Our Model

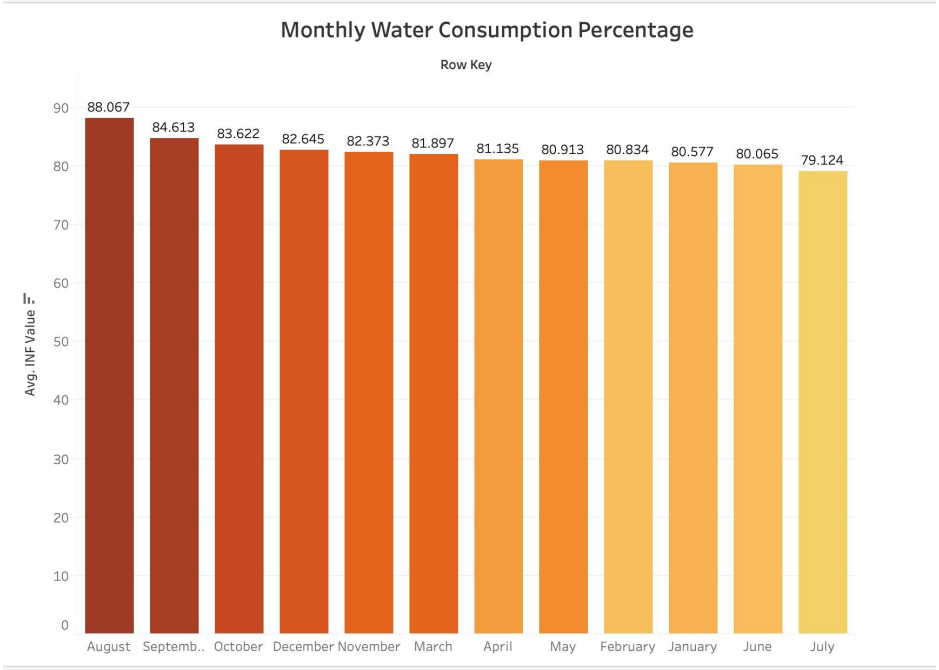
The main goal was to harmonise the datasets using a unified identifier. This common variable, termed the "Canonical", would have measurements from various hierarchical levels, ensuring that data from each of these levels could be integrated seamlessly. The canonical variable chosen was *"PRESION_ENTRADA_SECTOR"*. This decision was informed by two primary factors, one which covers the variable across various sectors, making it an ideal choice to ensure representation from all hierarchical levels, secondly, it displays significant resemblance in terms of the sector-aspect, making it a suitable option for merging datasets. This enhanced DataFrame was then integrated with descriptive attributes sourced from the clima.csv file. In essence, our approach was a systematic data integration strategy, leveraging a key variable to merge and analyse datasets across multiple hierarchical levels.

Hierarchical Level 1	**Hierarchical Level 2**	**Hierarchical Level 3**
LS Urda	LS-V Sureste	LS Centro
		LS Churruca
		LS Planta

In our model, we implemented a stacked generalisation model which unifies multiple individual models into a single meta-model, and this provides a better prediction accuracy than any of the individual base models. We initiated the Random Forest, XGBoost, and CatBoost models into our training dataset to identify the patterns in our data. Then, we combined predictions from different techniques which captures the diversity of our model. This approach tends to have better accuracy in predicting variables and reduces bias, hence decreasing the chances of overfitting in our models.

3. Business Aspects

Seasonality



OLS Regression Results

Dep. Variable:	Demand	R-squared:	0.677
Model:	OLS	Adj. R-squared:	0.675
Method:	Least Squares	F-statistic:	366.4
Date:	Sun, 22 Oct 2023	Prob (F-statistic):	2.68e-128
Time:	19:41:26	Log-Likelihood:	-4611.2
No. Observations:	529	AIC:	9230.
Df Residuals:	525	BIC:	9247.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.937e+05	795.405	369.183	0.000	2.92e+05	2.95e+05
temp	191.1708	14.378	13.296	0.000	162.925	219.417
humidity	13.0999	7.692	1.703	0.089	-2.011	28.211
solarenergy	133.6405	20.860	6.407	0.000	92.661	174.620

Omnibus:	70.323	Durbin-Watson:	0.062
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27.086
Skew:	0.335	Prob(JB):	1.31e-06
Kurtosis:	2.117	Cond. No.	871.

The "Monthly Water Consumption Percentage" graph, combined with the regression model's insights, offers a holistic understanding of water usage dynamics.

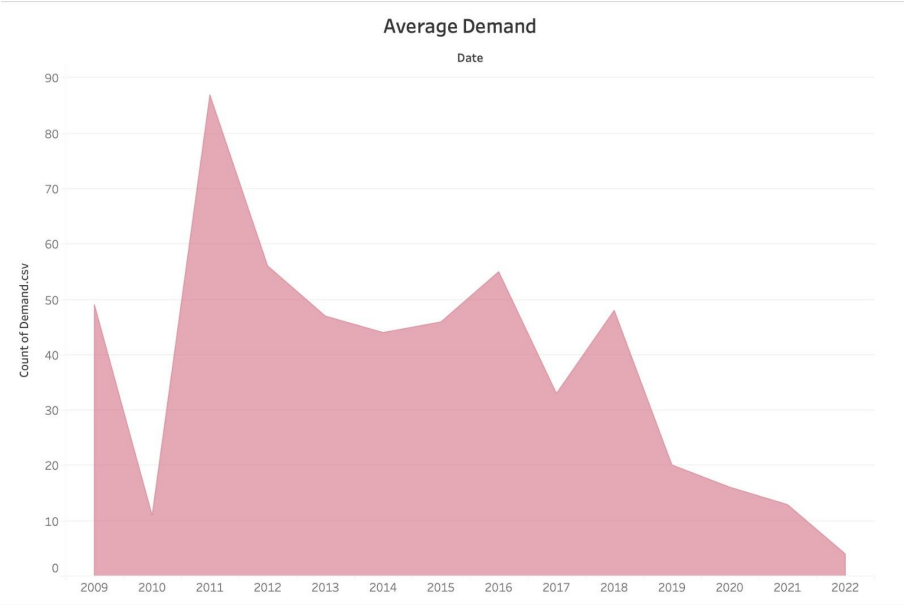
The water consumption is noticeably the highest in August, potentially hinting towards increased water usage during this month, which could be due to a variety of factors like the peak of summer or specific cultural events. As per our predictions, a strong influencer on water demands is Temperature, as the water demand surges by approximately 191.1708 units. This is statistically significant, confirming the assumption that warmer temperatures heighten water usage. Also, the solar energy rises by around 133.6405 units, a relationship underscored by its statistically significant p-value of 0.000, confirming our hypothesis that demand is high during the specified months.

Meanwhile, the consistency from September to May suggests stable and predictable water usage patterns during these periods. Interestingly, despite being summer months, June and July register the lowest consumption percentages. This might be indicative of increased water conservation efforts, fewer outdoor activities, or other factors that need to be explored. Overall, understanding these monthly variations is pivotal for water management strategies to ensure efficient resource allocation and to prepare for peak demand periods. Our regression analysis with its F-statistic value of 366.4 coupled with a minuscule p-value signifies that the model is statistically robust. An R-squared of 0.677 implies that about 67.7% of the variability in water demand is explained by the predictors in the model - producing a commendable good fit.

Demand

In order to bolster Acciona's customer base and ensure long-term loyalty, a meticulous analysis was undertaken on monthly contract trends. The below findings highlighted that new contract sign-ups peaked in February and June. However, as time progressed, terminated contracts surpassed the new enrollments. Armed with this insight into customer acquisition and retention, Acciona can refine its strategies by upgrading their current offerings, promoting their renewables and focusing on customer feedback. For the people of Villarubia, this means enhanced service offerings with more tailored solutions, and a commitment from Acciona to meet their evolving needs and priorities.

The below graph highlights the average demand over a period of time :



- 2009-2011: A clear surge in demand, peaking in 2011, likely indicating a period of rapid growth or increased consumption.
- 2011-2014: Sharp decrease in demand, possibly due to conservation initiatives, economic factors, or shifts in public awareness.
- 2014-2016: Demand stabilises with a minor uptick, suggesting a balance between consumption and conservation efforts.
- 2016-2022: Steady decline, hinting at long-term sustainable consumption strategies or technological efficiencies.

The fluctuating demand patterns emphasise the importance of a regression analysis to pinpoint key influencing factors. Acciona can utilise this analysis to craft strategic responses to these observed trends and anticipate future shifts in demand.

Prevention Plan

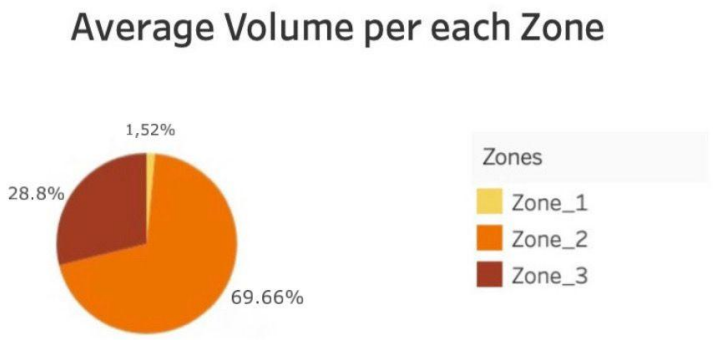
As mentioned earlier, Acciona supplies water for 3 different hierarchical levels with each having varying levels of water consumption. Our goal in this approach is to find the sectors classified under the hierarchical zones which require support and maintenance. This method will allow Acciona to prevent water loss and avoid shortages for families.
The underlying focus is on analysing the demand by zones and calculating related metrics. We filter data for each zone, determining new and terminated services for each date, and calculating the adjusted accumulated demand for the zone from the required years of 2021 and 2022. In order to measure the level of water, we utilise the *Canonical* type, "*Volumen Diario"*(daily volume), the number of measurements, and the average demand for each zone into a comprehensive dataframe. This data frame serves as a summary of key metrics for each zone. Finally, the summary data frame "Zone_consumption", results provided in the image below contains the key metrics for each zone needed for inspection. Given the size of different sizes of pipes (*calibre)*, the 13 one is the average size of a pipe and is connected to each household through the canonical elements which supply the water. Overall, this code is

a systematic approach to analysing and understanding demand patterns across different zones, using various data sources and criteria.

```
    Zones  Average volumen_Diario  Number of volumen_Diario measures  \
0  Zone_1               48.515252                                476
1  Zone_2             2239.161914                                494
2  Zone_3              926.725988                                815

   Average demand
0       61.500000
1     2248.750000
2     3224.857143
```

Our analysis of these results display three basic interpretations :



- Zone 1 : The average volume of water stored and utilised is lowest and the demand units are low, hinting that one sector part of Zone 1 does not require new development of pipes, but ensures constant checks to supply water efficiently.

- Zone 2 : The average volume utilised is the highest with a greater demand suggesting that pipes which hold this water volume require constant maintenance on the pipes as there are chances of water leak. Zone 2 consists of two different sectors.

- Zone 3 : The average volume utilised is the second highest with the greatest demand indicating that the sectors in Zone 3 require more water supply, hence the pressure of water flowing and the daily volume consumption of water is relatively proportional indicating that these areas need new pipes to hold more water.