Análisis del consumo energético de un centro comercial

Programa de experto en Data Science U-TAD



**Autor: Director:**

Raúl Amarelle Valera Carlos Gil Bellosta

03/09/2018

Índice

# Resumen ejecutivo

El presente proyecto trata del análisis de un data set que contiene, entre sus variables, datos del consumo energético de un centro comercial a lo largo de un periodo de dos años.

El objeto de este proyecto es analizar dicha serie temporal y la influencia del resto de variables, para en la medida de lo posible, realizar un modelo de predicción del consumo energético.

El proyecto se estructura en tres bloques principales:

1. Lectura y tratamiento de datos.
2. Visualización de datos y otras relaciones.
3. Casos de análisis de las series temporales y predicciones

La serie y el cálculo de predicciones se han realizado aplicando cuatro métodos:

1. Impacto causal utilizando modelos bayesianos de series de tiempo
2. Utilización del método Holt-Winters
3. Método ARIMA de predicción
4. Árboles de clasificación y regresión, RPART Tree.

El empleo de diferentes técnicas ha permitido comparar diferentes resultados de predicción y cómo hemos tratado de mejorarlo, los cuales se explicarán en el presente documento.

# Problemática y objetivos

Como sabemos, la electricidad es un bien que no puede ser almacenado, por lo que es de mucha utilidad conocer la previsión de consumo, aunque sea a corto plazo (un día) con respecto el consumo real.

El objetivo principal sería desarrollar un modelo predictivo que estime nuevas predicciones, compararla con el modelo predictivo actual y decidir cuál de las dos opciones predice mejor, con mejor error.

Otros objetivos específicos:

* Análisis y tratamiento del dataset
* Generación de otra información que pudiera ser relevante.
* Analizar diferentes modelos estadísticos para la previsión de la demanda así como un análisis comparativo de la calidad predictiva de cada modelo.

# Metodología

Tal y como hemos avanzado, el proyecto se estructura en los siguientes bloques:

1. Lectura, análisis y tratamiento de datos.
2. Visualización de datos y otras relaciones entre variables.
3. Casos de análisis de las series temporales y predicciones.
4. Conclusiones

Para el cálculo de predicciones he utilizado cuatro métodos diferentes:

1. Impacto causal utilizando modelos bayesianos de series de tiempo
2. Utilización del método Holt-Winters
3. Método ARIMA de predicción
4. Árboles de clasificación y regresión, RPART Tree.

# Tecnologías utilizadas

En la parte técnica, se ha utilizado R como lenguaje de programación.

Análisis y tratamiento de datos: uso de algunas librerías representativas como ggplot2, dplyr, data.table, zoo, etc.

Machine Learning: uso de librerías como bsts, Holt-Winters, forecast, tseries, MLmetrics, etc.

# Análisis del dataset inicial

Este dataset es una serie temporal con los siguientes datos relativos a un centro comercial.

Breve explicación de las variables:

1. Fecha: fecha de la medición, sólo hay una medición diaria
2. Estimado: sólo aparece “No”, no nos aporta información
3. Kwh: potencia real consumida
4. LB: Línea base, es decir, la predicción de potencia consumida
5. De 5 a 10. CCDD o CHDD: Estas seis columnas son de temperaturas, 3 de Cooling-Degree Day y 3 de Cooling Heating-Degree Day. Es la diferencia entre el promedio diario de temperatura y una determinada temperatura base de referencia, que suele ser la exterior.

CHDD18 = 8.5, quiere decir que sobre la temperatura de 18º, el centro comercial ha tenido que climatizar 8.5º. Por eso CHDD19=9.5º y CHDD20=10.5º, porque si la temperatura de referencia sube un grado, hay que calentar un grado más, la diferencia aumenta ese grado.

El centro comercial tiene dos modos de climatización: calefacción o refrigeración y en la práctica o funciona en un modo o en otro, por eso de las 6 columnas, son siempre 3 ceros vs 3 números.

1. Afluencia: Número de asistentes al centro comercial.

# Tratamiento de datos

De las 11 variables iniciales, podemos eliminar la 2 (“Estimado”) y 4 columnas de temperaturas. De las 3+3 columnas de temperaturas que tenemos, 2+2 son redundantes. Con quedarnos una columna de cada modo de climatización, es suficiente.

Escogemos CCDD20 y CHDD18, porque son las columnas más ventajosas. Es decir, porque si hay que calentar, calentar a 18º (CHDD18) es la línea de referencia mínima. Y si hay que refrigerar, refrigerar a 20º (CCDD20) es la referencia mínima.

De la primera exploración de las 6 variables resultantes llama la atención un aspecto importante que conviene resaltar. La columna FECHA la reconoce como cadena de caracteres (chr) y conviene tratarla como fecha, para que pase a ser Date. Esto es importante para posteriores representaciones gráficas de la serie, porque normalmente el eje X reflejará las fechas, para ver la evolución de los valores.

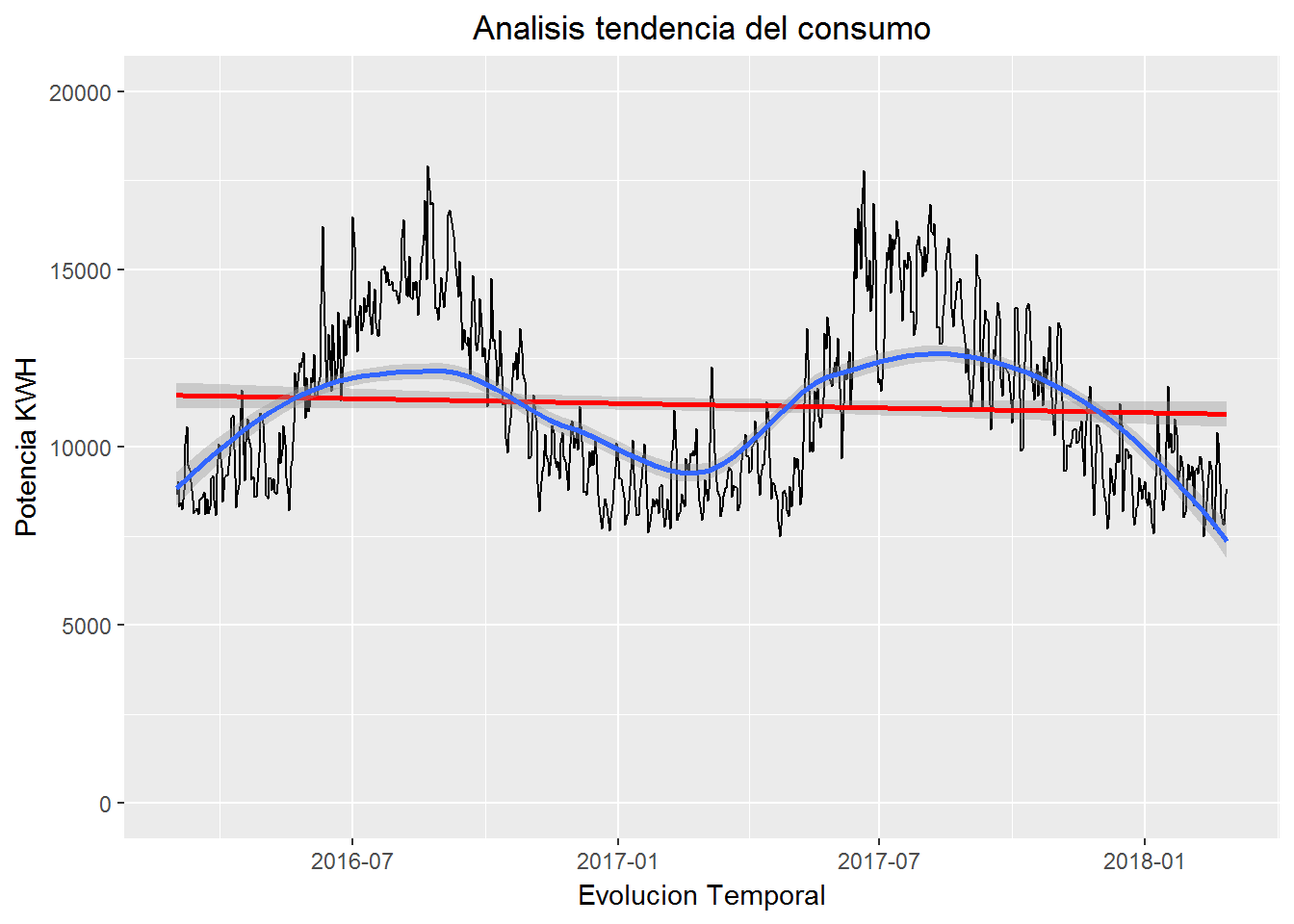
Dentro de este apartado de Tratamiento de Datos, se incluyen una serie de variables creadas ad-hoc y de las que se puede extraer información relevante. Hemos realizado algunos análisis de la información disponible, pero más con la idea de ilustrar las posibilidades que ofrece el dataset. Lógicamente se podría haber profundizado más, pero el objetivo del proyecto es analizar una serie de modelos predictivos y no tanto en detallar la casuística que nos ofrece el dataset.

Caso aparte también merece el Tratamiento de Missing Values. De 729 muestras, sólo hay 4 NA, por lo que son tan pocas, que lo más práctico es eliminarlas. Además, en nuestra serie tampoco tiene sentido realizar alguna simulación para rellenar los huecos con media por columna o similar, porque las mediciones de consumo no guardan relación entre sí.

# Visualización de datos, variables y otras relaciones

Voy a recopilar en este punto algunas gráficas que he generado y considero de interés para explicar información relevante.

## Análisis tendencia del consumo

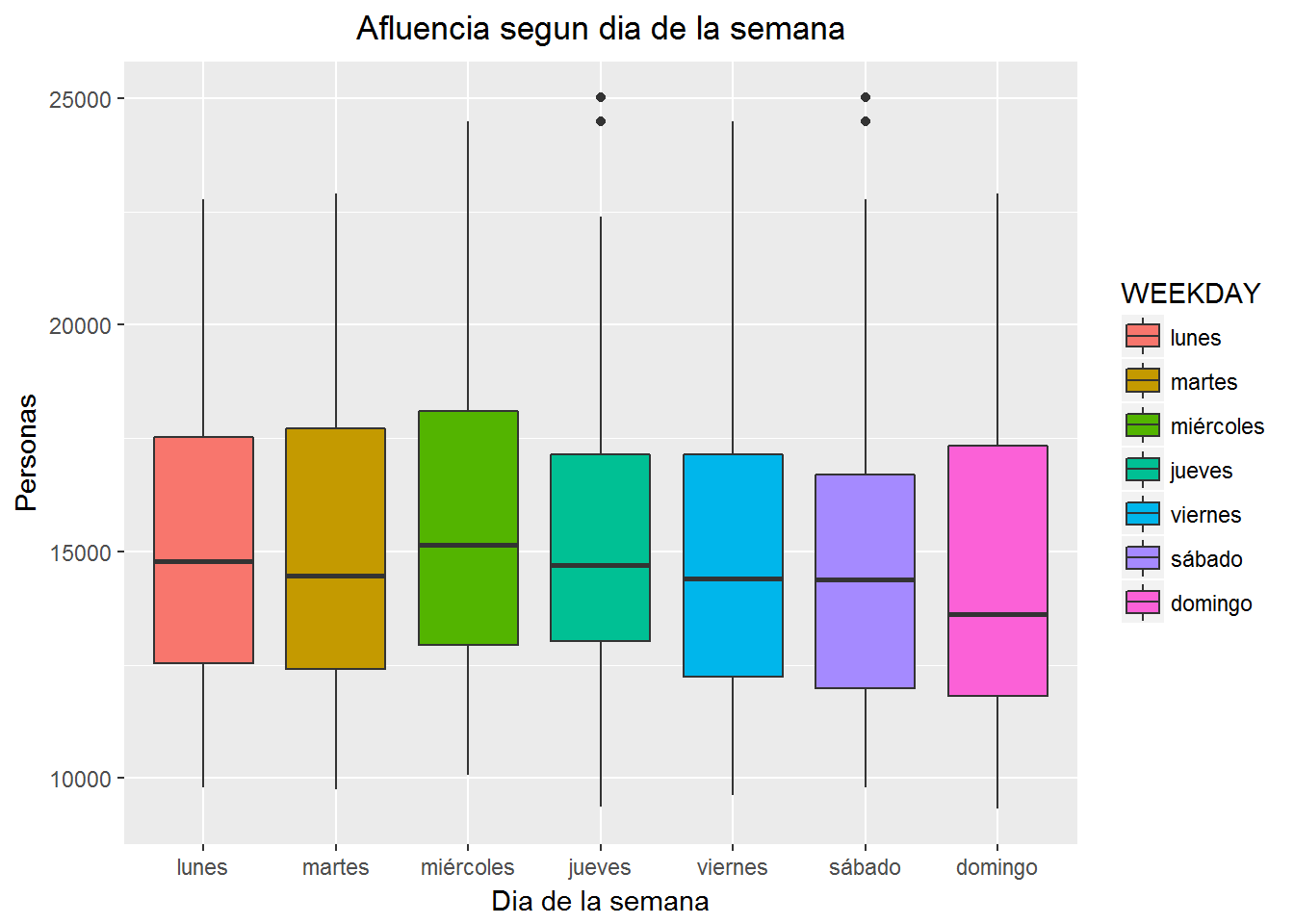


Si nos fijamos en la línea roja, se aprecia que hay una ligera tendencia de disminución del consumo a lo largo de estos dos años, motivada por la bajada de consumo de los últimos 3-4 meses de nuestros datos.

Para extraer alguna conclusión más solida habría que analizar un periodo más amplio, pues al ser la disminución más pronunciada al final, podría ser por un motivo coyuntural en una época del año con extraordinario buen tiempo. Nuestra serie está compuesta justo por dos años completos, que es el requisito mínimo para que se pueda afirmar.

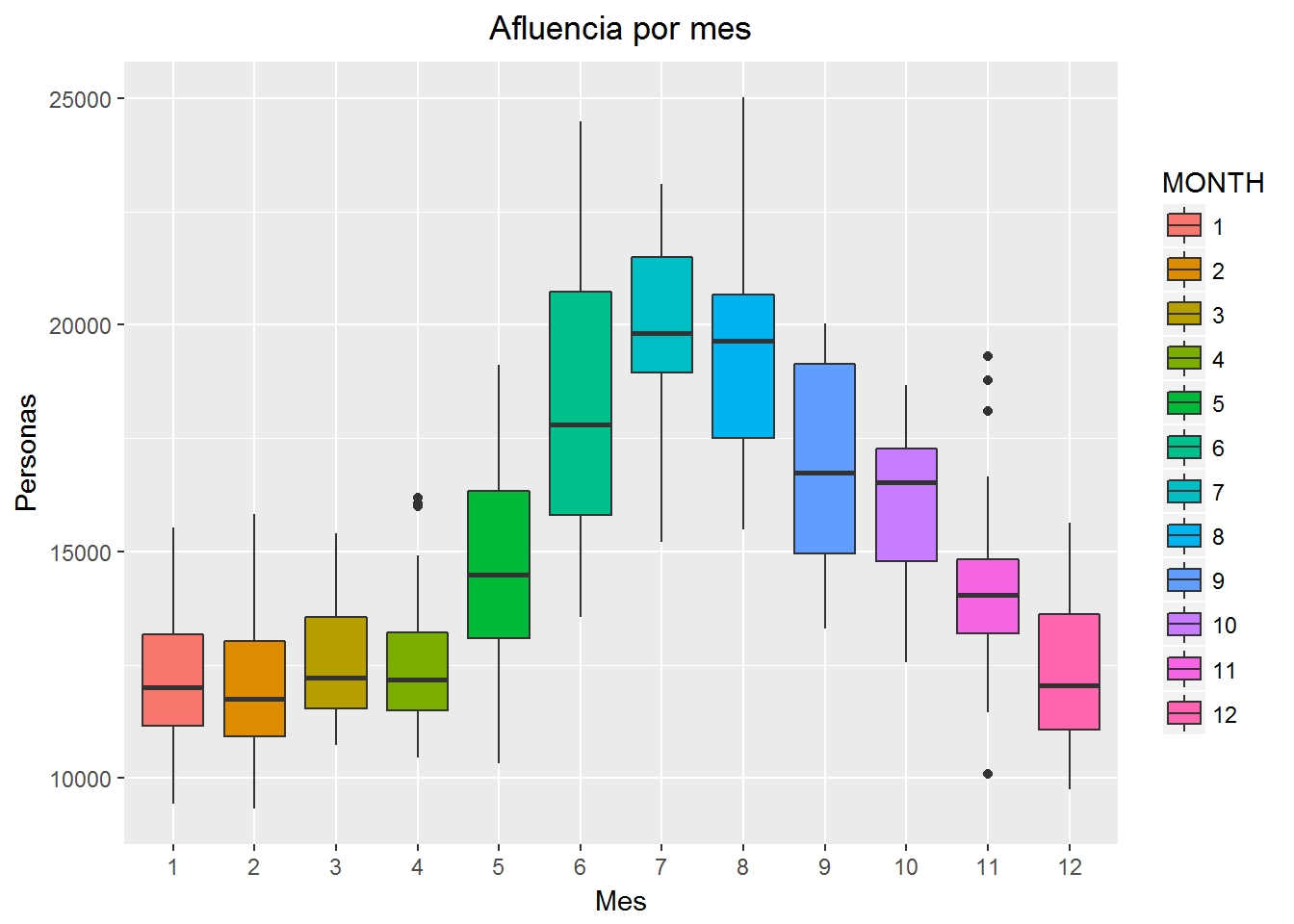
## Análisis de la variable Afluencia

En primer lugar analizamos afluencia según el día de la semana.



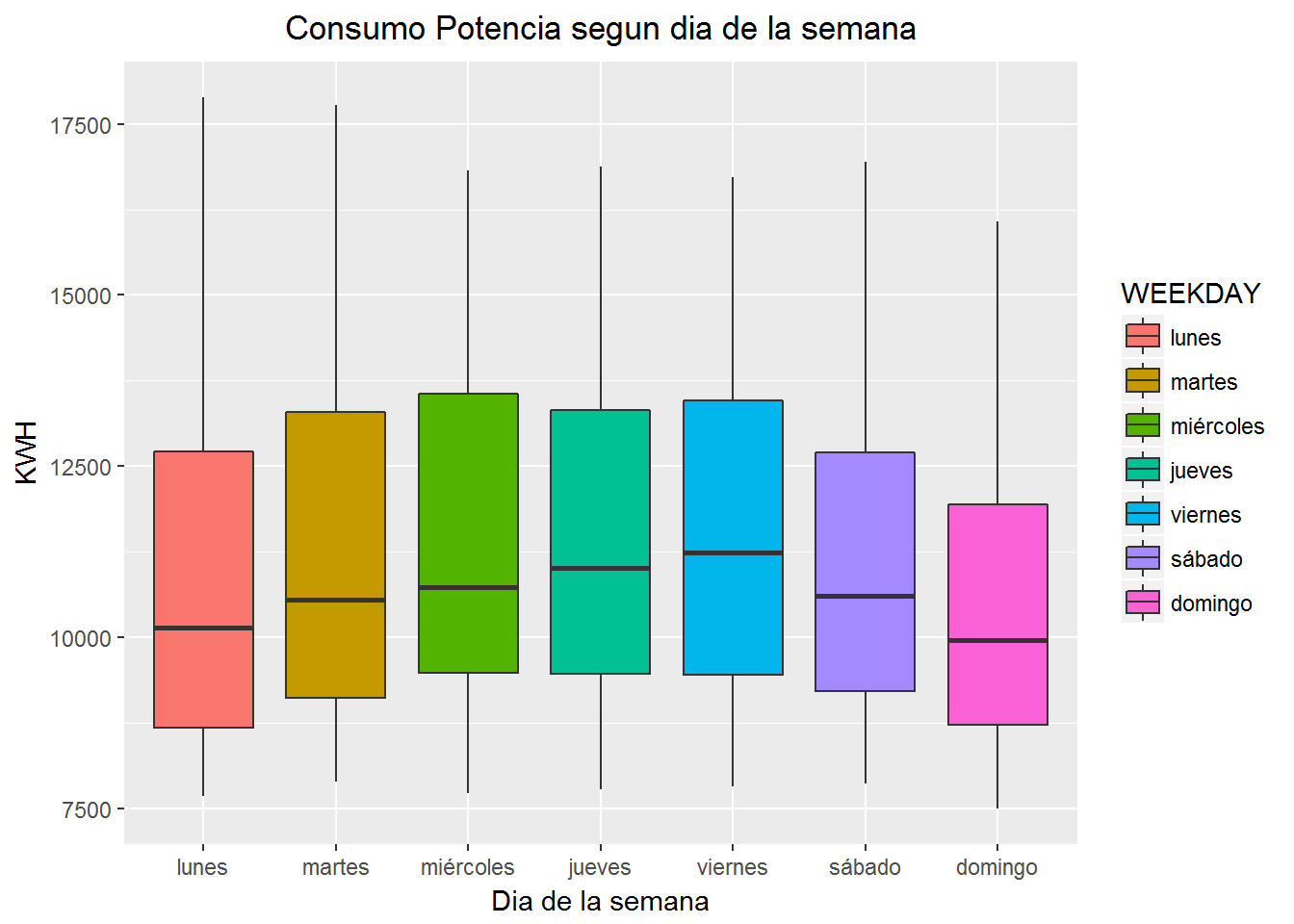
En mi opinión, sorprendentemente, el día de mayor asistencia es el miércoles, cuando inconscientemente esperaba que fuera el sábado.

A continuación, analizamos afluencia según el mes del año. Pero también en mi opinión de manera sorprendente, los meses de mayor afluencia son julio y agosto de manera muy destacada, cuando imaginas que la gente está de vacaciones y sale más con el buen tiempo. Y es que casi dobla, de unas 12.000 personas durante los meses de invierno a casi 20.000 en verano.



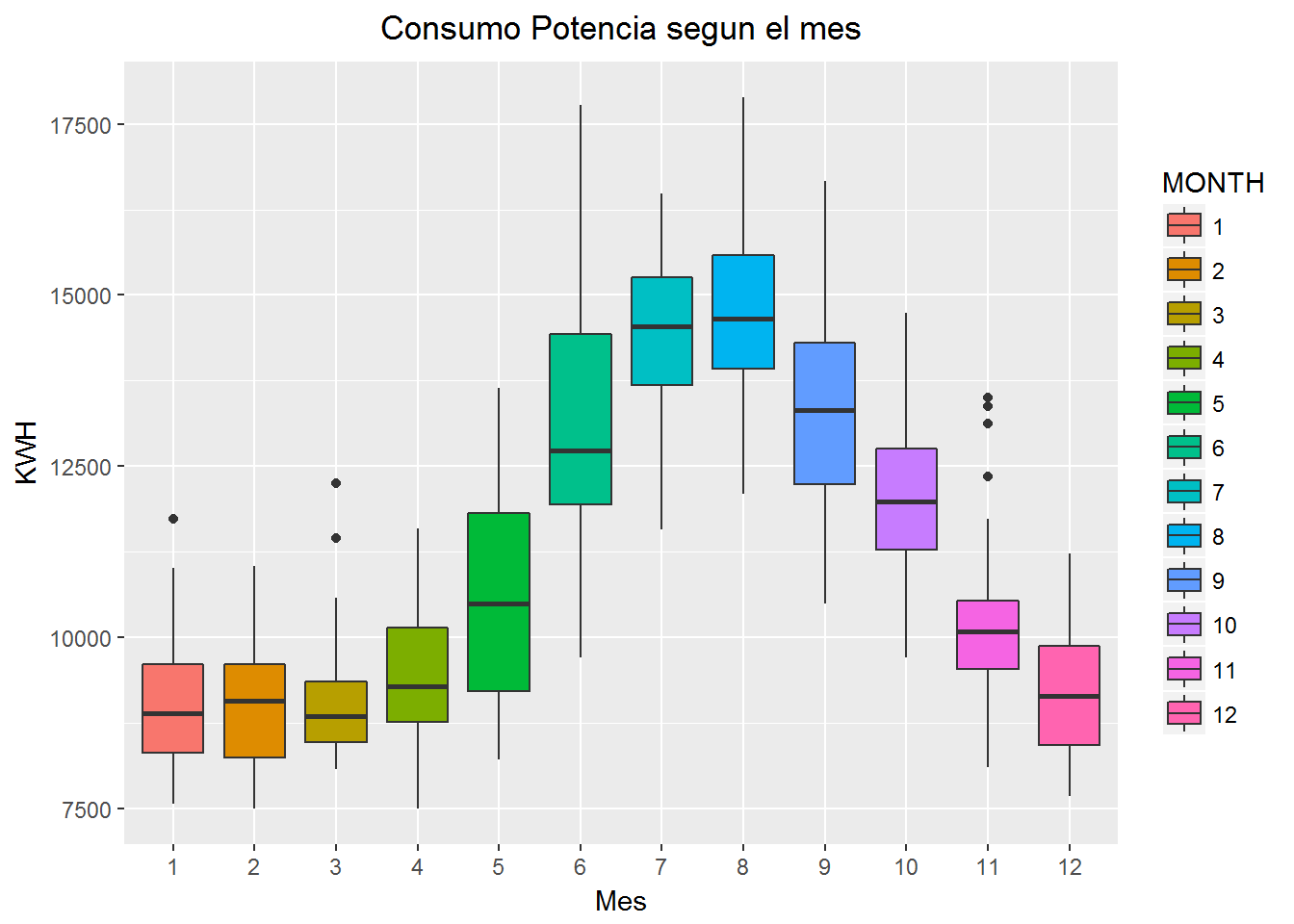
## Análisis de la variable potencia

En primer lugar analizamos el consumo de potencia según el día de la semana.



Hay un contraste entre el día de la semana que más consumo hay (viernes) con el de mayor asistencia (miércoles). También vemos que hay un aumento progresivo de desde el lunes (min) a viernes (max).

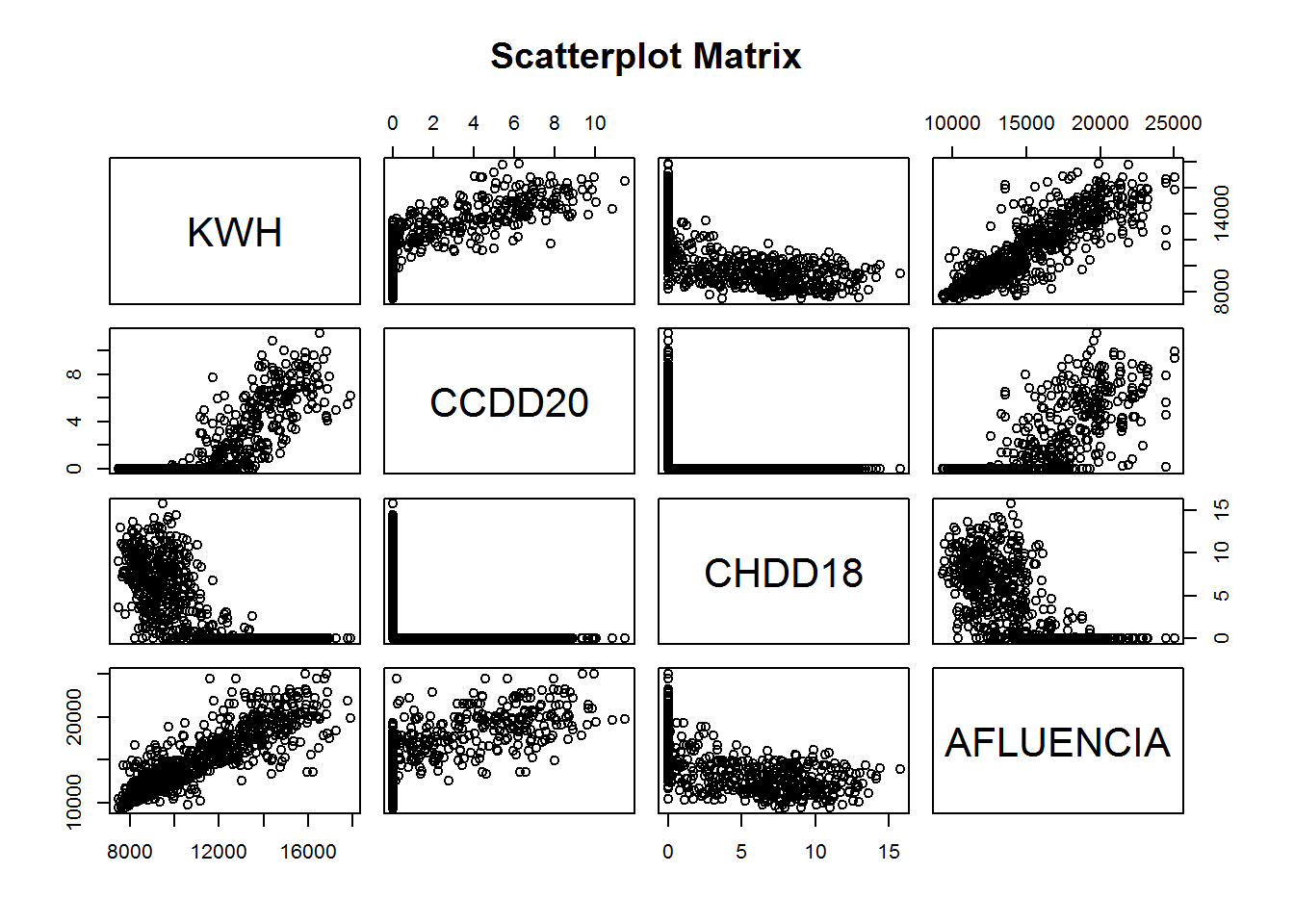
Y ahora el consumo de potencia según el mes del año.



Los meses de mayor consumo son julio y agosto, que coinciden con los meses de mayor afluencia. Hay una relación directa entre la temperatura de confort en su interior y la afluencia al centro comercial.

## Análisis de correlaciones

Vamos a ver la matriz de correlaciones entre algunas de las variables más interesantes: KWH (2), CCDD20(4), CHDD18(5) y AFLUENCIA (6).



Aunque hay bastante dispersión en los datos, en primer lugar destaca una fuerte dependencia entre Consumo vs Afluencia. También llama la atención la relación entre Consumo y CCDD20 (Refrigeración). Esto es coherente, cuanta mayor diferencia de temperatura con el exterior, mayor consumo. Ya vimos anteriormente que en los meses de verano hay mucho más consumo.

En cambio, en modo Calefacción (CHDD18), al ser una línea plana significa que para el mismo consumo hay días de poca diferencia de temperatura y otros días con mayor gradiente térmico.

Con la tabla de correlaciones numéricas para confirmar toda esta información:

## KWH CCDD20 CHDD18 AFLUENCIA

## KWH 1.0000000 0.8072247 -0.7207035 0.8513208

## CCDD20 0.8072247 1.0000000 -0.5228964 0.7284759

## CHDD18 -0.7207035 -0.5228964 1.0000000 -0.6668167

## AFLUENCIA 0.8513208 0.7284759 -0.6668167 1.0000000

# Análisis de la serie temporal según 4 casos diferentes

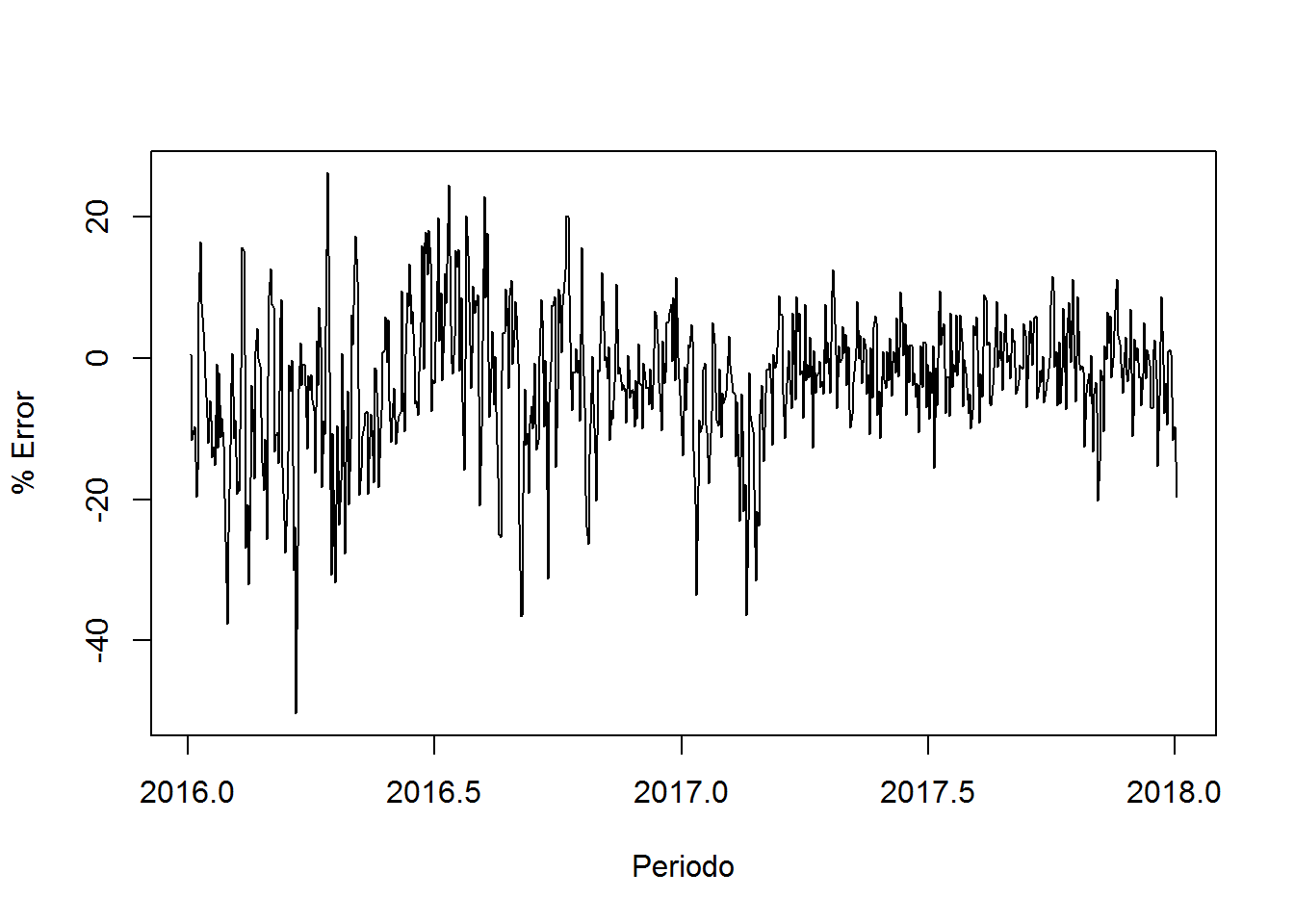
El objetivo principal es desarrollar un modelo predictivo que prediga nuestro Consumo Real (KWH) y mejorar en la medida de lo posible, los resultados de la predicción actual (LB).

Para conseguirlo vamos a definir nuestro plan:

1. Aplicar diferentes métodos de análisis a nuestra serie.
2. Analizar la calidad predictiva de cada modelo.
3. Escoger el modelo que mejor prediga.
4. Comparar el error entre su modelo predictivo y nuestra mejor propuesta.
5. Comprobar cuál de las dos opciones es mejor.

La dificultad de analizar esta serie viene dada porque está compuesta por muestras diarias en un ciclo estacional de dos años. Al tener justo el ciclo de dos años, se observa claramente una estacionalidad anual. Por un lado, tampoco es una serie de amplio rango lo suficientemente larga (ej. diez años). Por otro, la estacionalidad para periodos más cortos no resulta evidente.

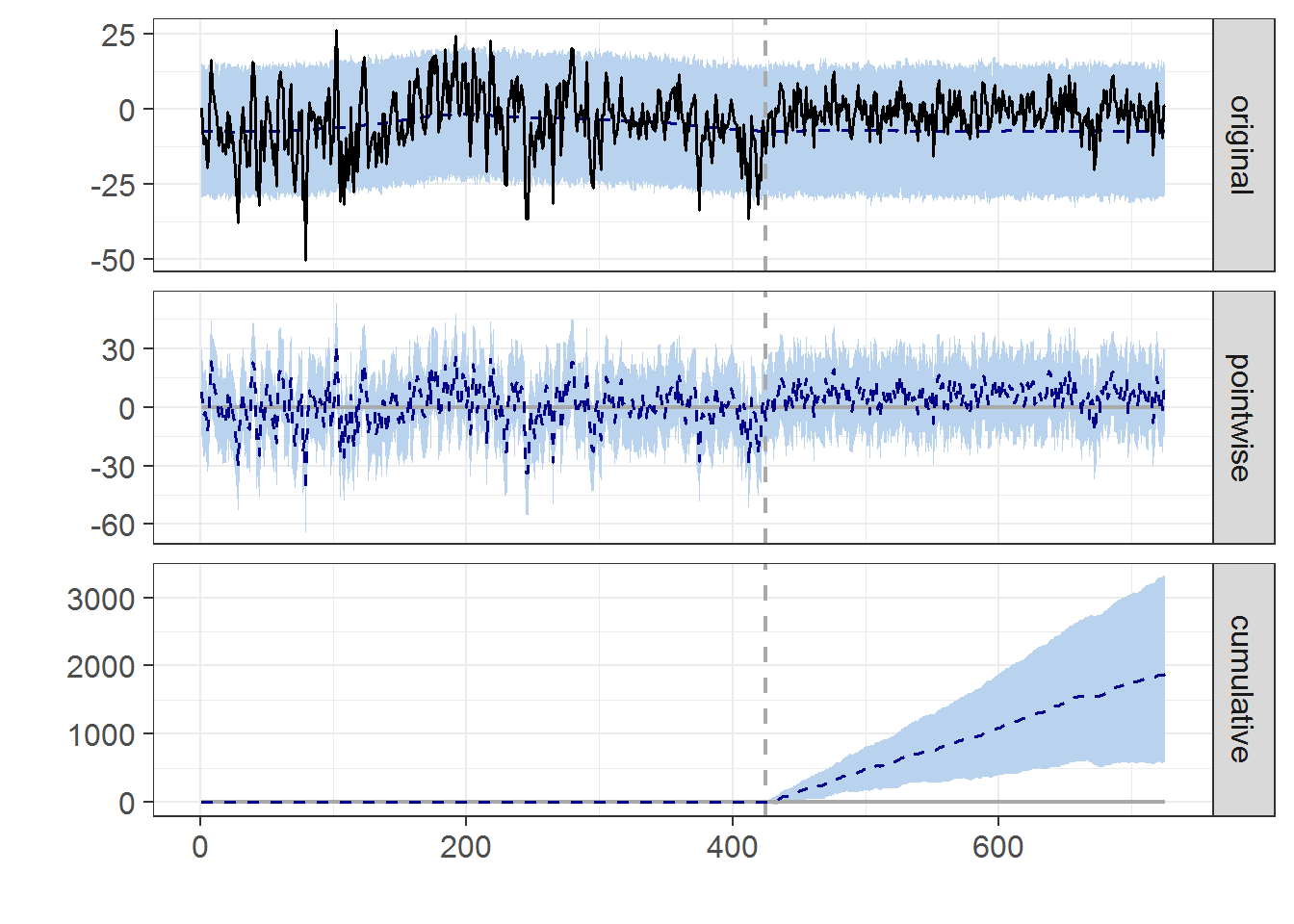
Empezamos analizamos la evolución de la variable Error:



Como primera aproximación, podemos observar que el modelo de predicción ha mejorado a partir de la segunda mitad de 2017, ya que el % Error es mucho menor, tiende a concentrarse en la horquilla de -10% a +10%.

## CASO 1: Impacto causal utilizando modelos bayesianos de series

Vamos a suponer que a partir de la medición 426, coincidiendo con el 01/05/2017, se han introducido mejoras en el algoritmo de predicción y vamos a comparar la mejora que ha supuesto aplicando Causal Impact.



## Posterior inference {CausalImpact}

##

## Average Cumulative

## Actual -0.91 -273.72

## Prediction (s.d.) -7.2 (2.4) -2153.3 (722.7)

## 95% CI [-12, -2.9] [-3609, -866.9]

##

## Absolute effect (s.d.) 6.3 (2.4) 1879.6 (722.7)

## 95% CI [2, 11] [593, 3335]

##

## Relative effect (s.d.) -87% (-34%) -87% (-34%)

## 95% CI [-28%, -155%] [-28%, -155%]

##

## Posterior tail-area probability p: 0.00235

## Posterior prob. of a causal effect: 99.76471%

##

## For more details, type: summary(impact, "report")

El resultado del análisis nos dice que la intervención en el modelo redujo la previsión de ERROR en un 87%.

Ahora vamos a calcular la predicción utilizando un modelo personalizado con bsts.model.

Una serie tiene dos componentes principales, una parte lineal y otra estacional. La parte lineal está clara, pero para la estacional vamos a analizar cuál de los supuestos encaja mejor: semanal o mensual.