Maestría Oficial en Big Data y Data Science

Actividad 2.- Clustering y Serie Temporales

06MBID. - Estadística Avanzada

Alumno: Marrero Carrión, Ricardo Alberto

Fecha de entrega: 29/09/2021

Índice

Índice de figuras

Índice de tablas

1. Clustering

La primera parte de la actividad comprende la **aplicación de dos técnicas de Clustering**, analizando sus resultados en base al dataset escogido.

Para la aplicación del Clustering, seguiremos trabajando con el dataset desarrollado en la Actividad1, donde mediante una serie de variables y datos obtenidos del Complejo Turístico Walt Disney World Orlando (FL), **buscamos la probabilidad de predecir cual es la mejor semana de visitar el parque para el turista, o agrupar estos datos en grupos con bastante similitud, para que una empresa de turismo pueda aplicar campañas de marketing.**

Esto último es lo que trataremos en este capítulo.

1.1. Descripción de los datos

Como se comentó anteriormente, por dar continuidad a lo estudiado en esta asignatura y la de Minería de Datos, seguiremos usando el dataset transformado mediante el proceso KDD y utilizado en la Actividad 1; aunque este dataset desarrollado tiene 14 atributos y 2119 observaciones, **el alcance de esta parte de la Actividad 2 solo abarcará el análisis de dos de sus variables**, creando un el siguiente subset:

COLUMNA	TIPO	DESCRIPCION	FACTOR DE DECISION
SEMANA_ANYO (variable dependiente)	Entero	Número de la semana (0...53)	Fecha
RATIO_CAPACIDAD_FALLOS	Numérico	Grado de pérdida de capacidad del parque por fallos en las atracciones (0...1)	Ocupación a Afluencia
RATIO_HORAS_APERTURA	Numérico	Razón de total horario de apertura de los parques (0...1)	Horario de Apertura

Tabla 1. Atributos del subset Clustering. Elaboración propia.

La razón de usar **estas dos variables en que en su conjunto podrían explicar en qué semana de año los parques tendrían un mayor rango horario y una menor tasa de fallos**, para así predecir la probabilidad de tomar las decisiones explicadas en el apartado anterior.

1.2. Motivación del análisis estadístico

Nos centraremos en este capítulo en **analizar patrones predictivos que nos permita segmentar las variables del dataset, en tres categorías:**

Segmento	Detalle
S1	Temporada en el año óptima
S2	Temporada en el año media
S3	Temporada en el año no óptima

Tabla 2. Segmentos de los resultados esperados. Elaboración propia.

Para este análisis usaremos las dos técnicas estudiadas en la asignatura.

Por un lado, aplicaremos la **técnica del K-means**, con K=3 para ver si este enfoque satisface el objetivo de agrupar en tres segmentos las semanas del año en función a la tasa de fallos y las horas de apertura.

Posteriormente, mediante el **Clustering Jerárquico**, buscaremos mediante un dendograma, confirmar la probabilidad de que el enfoque de segmentar en tres categorías es óptimo para nuestro estudio.

1.3. Estadística Descriptiva

Del análisis estadístico, para las variables escogida en el dataset de esta actividad, encontramos que:

	RATIO_CAPACIDAD_FALLOS	RATIO_HORAS_APERTURA
Valor mínimo	0.76	0.47
1er Cuartil	0.86	0.67
Mediana	0.94	0.71
Media	0.91	0.71
3er Cuartil	0.95	0.76
Máximo	1.00	1.00

Tabla 3. Estadísticos variables capacidad fallos y horas de apertura. Elaboración propia.

Considerando que estos atributos se encuentran normalizados, podemos observar que:

- Por pérdida en la capacidad de atender a los visitantes, por fallos en las atracciones; **el parque tiene de media un ratio capacidad//fallos del 91%**
- Y como mínimo funcionarán a una capacidad del 76%

- En cuanto a las horas de apertura del parque, existe una mayor variación de su ratio; y, por ende, del total de horas abiertas, porque dependerá del día de la semana y la estación del año.
- Pero **en media abren en un 71% del día con más hora de apertura**; y como mínimo un 47%.

Y del diagrama de bigote o boxplot, observamos que esta variabilidad en las horas de apertura hace que se presenten valores Outliers.

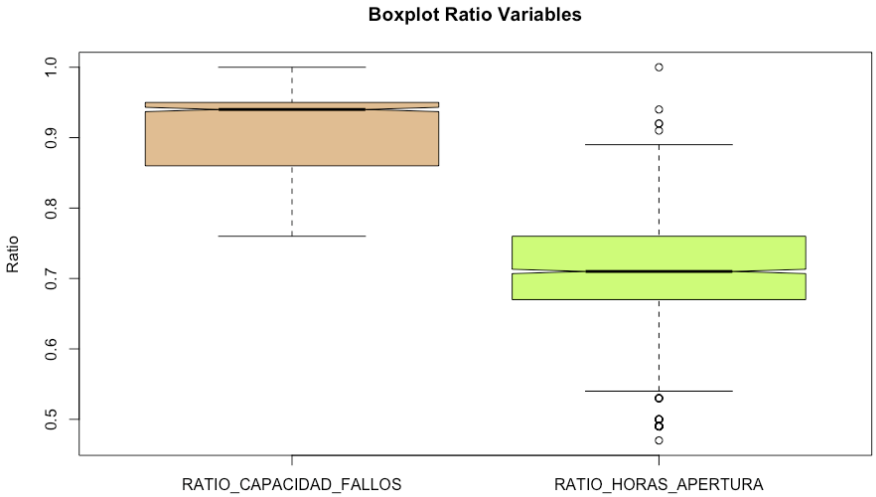

Ilustración 1. Boxplot de las variables analizadas

Este hecho lo podemos comprobar por medio de sendos histogramas sobre las variables a analizar.

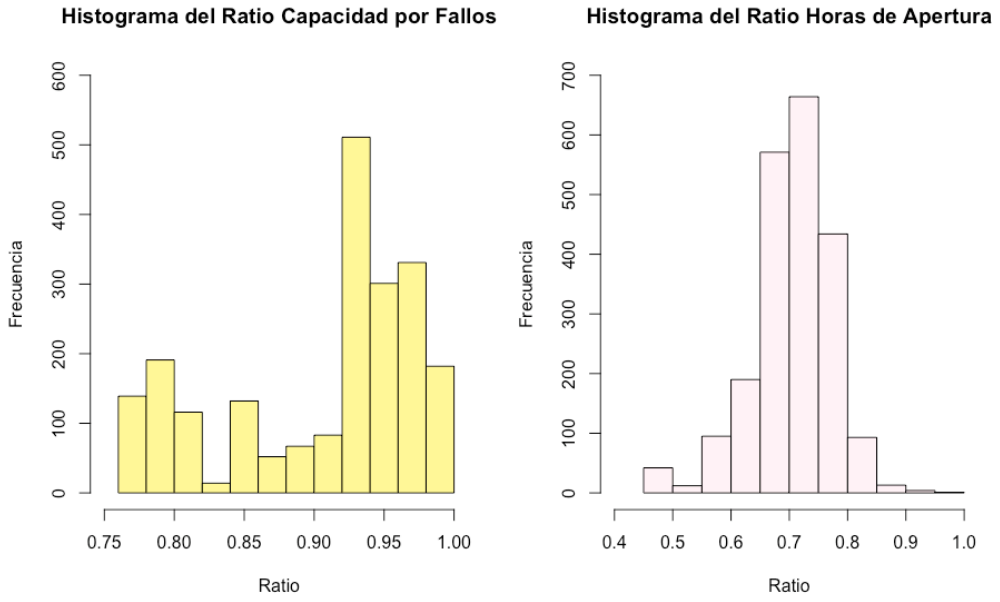

Ilustración 2. Histogramas de frecuencia capacidad fallos y horas de apertura

Estos gráficos nos ayudan a concluir con las observaciones del dataset, **el parque esta preparado para funcionar mayormente entre 90% y 95% de su capacidad**, y que su ratio de apertura contiene observaciones de baja frecuencia que pudiésemos considerar como outliers.

Posteriormente analizaremos si estos Outliers tienen alguna relevancia en los Cluster de nuestro modelo.

1.4. Técnicas de Clustering

Pasamos entonces a aplicar dos técnicas de Clustering para evaluar si nuestro supuesto inicial de agrupar nuestras observaciones en 3 clúster es válida.

1.4.1. K-Means

Evaluamos por medio de la técnica K-Means, con K=3 las variables de RATIO_CAPACIDAD_FALLOS vs RATIO_HORAS_APERTURA, y, por otro lado, vemos como se distribuyen las observaciones de las SEMANA_ANYO en este modelo.

Recordemos que buscamos agrupar nuestras observaciones en tres categorías para determinar cual es la semana óptima del año.

Aplicando las técnicas de K-Means detalladas en el Anexo (Script R) obtenemos:

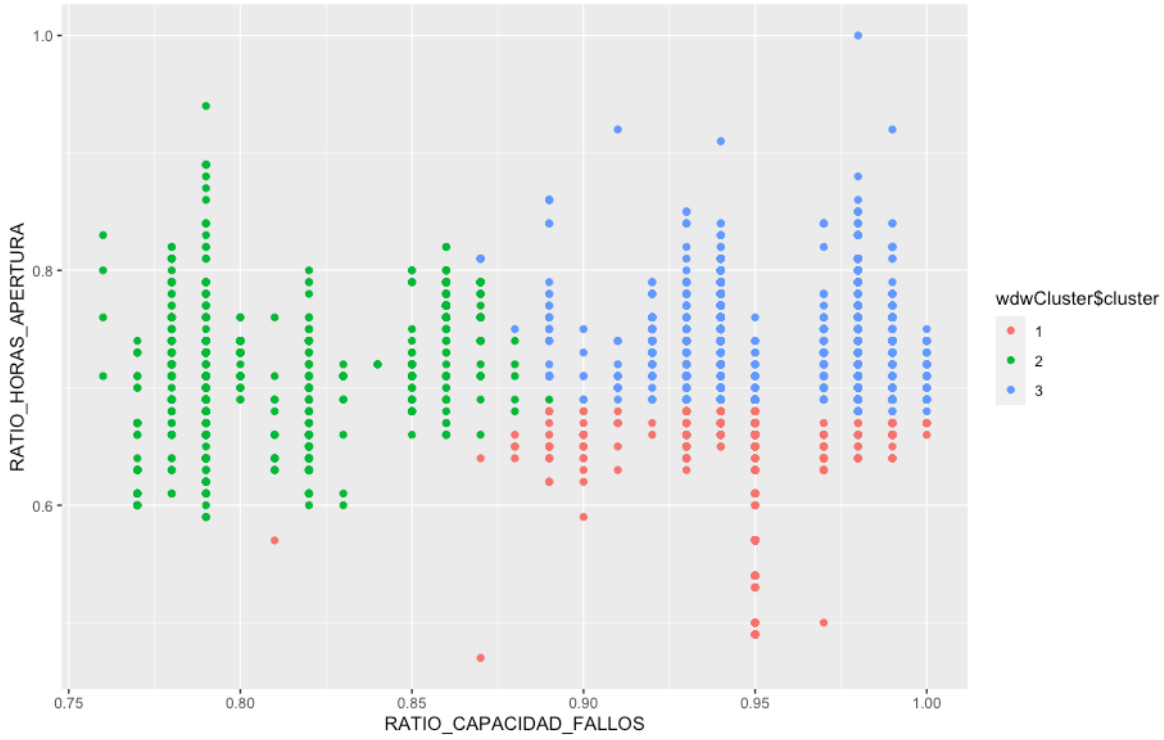

Ilustración 3. Técnica K-Means K = 3

Observamos que el K-Means con 3 clúster, agrupa las observaciones en tres categorías, que pusieren coincidir con nuestro análisis:

- Clúster 1 -> Grupo con observaciones donde el parque abre por debajo del 50% de su máximo, pero tiene una tasa alta de capacidad ante fallos (menos atracciones fuera de servicio) mas del 87%
- Clúster 2 -> Grupo en una zona media de horario de apertura de los parques, y que adicionalmente presentan una tasa baja de capacidad (mayor número de atracciones fuera de servicio). En esta categoría se observa un mayor número de observaciones. **(Grupo Dominante)**
- Clúster 3 -> Grupo en una zona media alta de horario de apertura, pero con una tasa favorable de capacidad ante fallos (menos atracciones fuera de servicio)

Con esta aproximación, entonces vemos en que clúster encajan las semanas del año del dataset observado

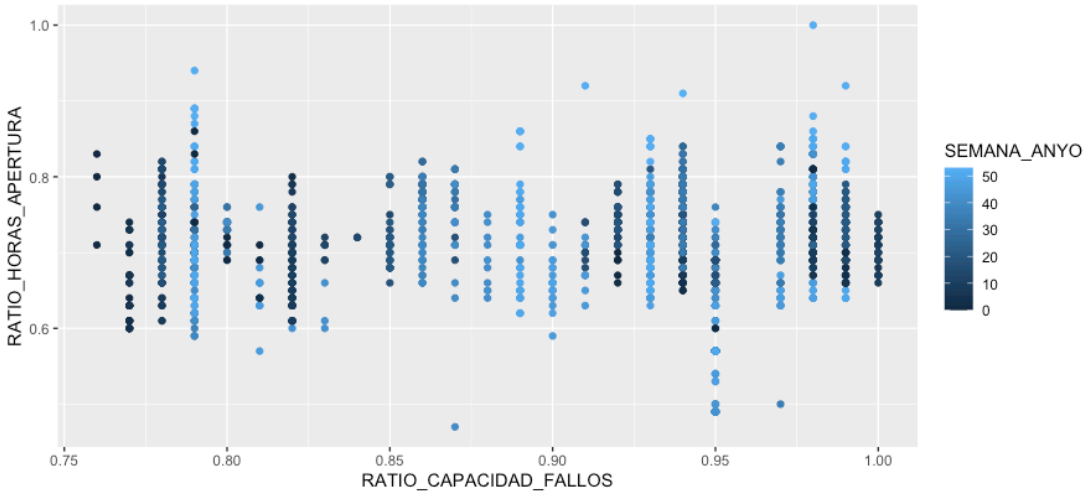

Ilustración 4. Evaluación de la variable dependiente

¿Que podemos observar de la comparación de estos dos diagramas? Podríamos entonces valorar o analizar la probabilidad de cuales semanas del año podrían estar en los Clúster antes descritos. Por ejemplo:

- Las primeras semanas del año, aproximadamente entre W1 y W10, estarían dentro del Clúster 2, donde los parques abren menos horas y hay mas atracciones fuera de servicio; que pudiese coincidir con un período de mantenimiento
- Y las últimas semanas del año en el Clúster 3, atracciones a pleno rendimiento y mayor número de horas de apertura; que bien pudiesen coincidir con la pausa navideña.

Viendo estos dos diagramas, pudiésemos decir, que **los supuestos valores Outliers si son significativos en la muestra,** porque representan una franja de las semanas finales del año.

Ahora bien, ¿Es óptima la selección de K=3?, veamos la comprobación mediante el diagrama del codo.

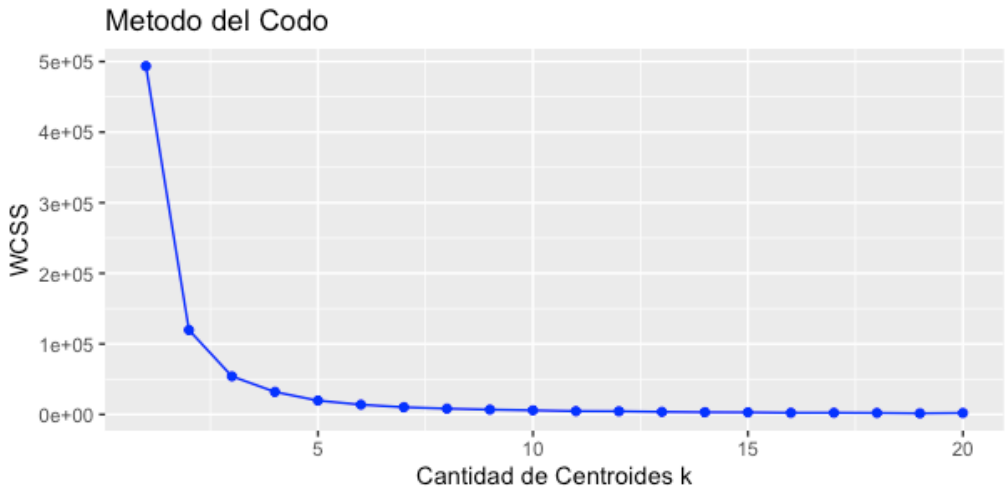

Ilustración 5. Método del codo

Pues efectivamente vemos que **la curva se suaviza justo en la tercera iteración**; por lo que podríamos dar por válido nuestro modelo.

Ya por último, vemos como quedan los clúster con sus respectivos centroides, y distribución final.

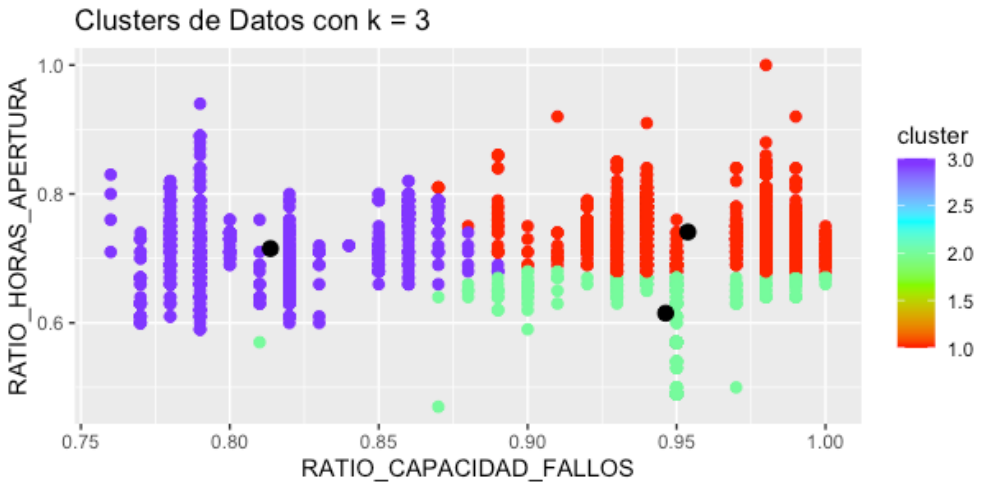

Ilustración 6. Modelo de Clustering K-Means con Centroides

A continuación, parasemos a desarrollar el mismo estudio, pero con una técnica de Clustering Jerárquico.

1.4.2. Clustering Jerárquico

Para ilustrar como se representa el análisis de nuestro modelo a través de un modelo de Clustering Jerárquico, hemos usado la función de R “hclust” que por defecto utiliza la técnica de complete linkage.

Es nuestro caso es adecuado, porque buscamos un método robusto, sin que tenga en cuenta los Outliers, ya que en el punto anterior indicamos que si eran observaciones significativas.

Para realizar el gráfico hemos usado la función “fviz_dend”, del paquete (factoextra); quedando nuestro análisis de la siguiente manera:

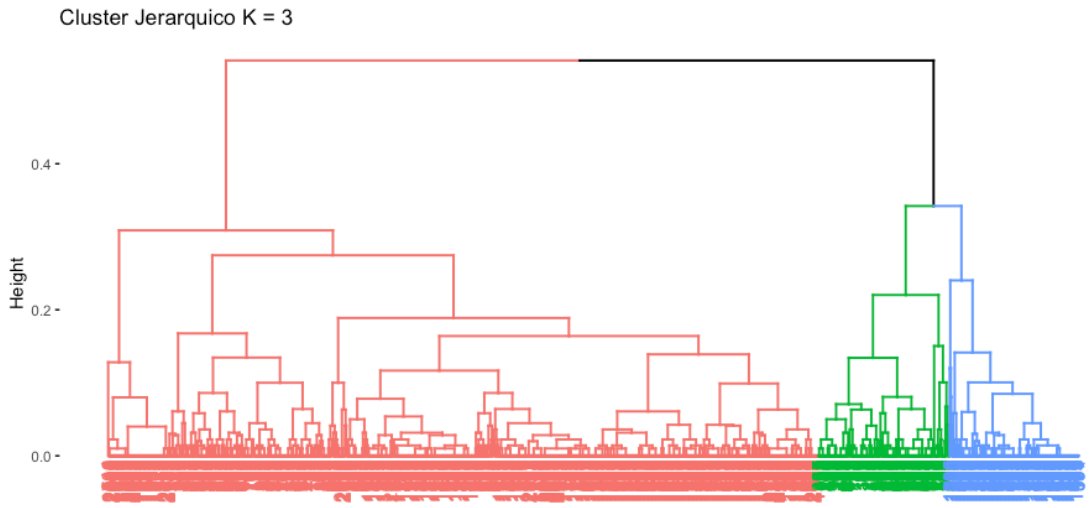

Ilustración 7. Clustering Jerárquico

Donde las observaciones han quedado agrupadas de la siguiente forma:

Segmento	Detalle	# Observaciones
S1	Temporada en el año óptima	286
S2	Temporada en el año media	1551
S3	Temporada en el año no óptima	282

Tabla 4. Relación Segmentos con el Clustering Jerárquico. Elaboración propia.

Que **en una primera valoración se aproxima al análisis resultante del método K-Means**.

1.5. Análisis de Resultados

Mediante el uso de herramientas de Clustering, hemos podido **observar que por medio de a la asociación de variables de un dataset, podemos hacer un análisis estadístico de como agrupar observaciones con el fin de analizar similitudes.**

En este capítulo escogimos solo dos de las variables de nuestro dataset original (trabajado en la Actividad 1) y ver su comportamiento y asociación con una variable dependiente como puede ser la Semana de Año.

Para ello usamos dos técnicas K-Means y Clustering Jerárquico, y que partiendo de nuestra asunción de dividir los segmentos en tres categorías **se logró obtener un modelo lo suficientemente ajustado para lograr este fin.**

Con el resultado del K-Means pudimos obtener dos resultados:

- Poder **valorar cuando es la semana óptima de visita en función a la tasa de atracciones operativos (capacidad/fallos) vs horas de apertura del parque. Ilustración 3 y 4.**
- Y de cara a un operador turístico ver en **que semana de pueden aplicar campañas de marketing tomando en cuenta las variables anteriores**.

Con el resultado del **Clustering Jerárquico, pudimos ilustrar cual es el Cluster o categoría dominante.**

Sobre el resultado de la agrupación de ambas técnicas, el resultado fue bastante similar; y con la finalidad de explicar el objetivo de este trabajo que era buscar una semana óptima para visitar un Parque de Atracciones; considero que **es posible predecir la variable dependiente, tal y como indica la Ilustración 4 y la Tabla 4 de los puntos anteriores.**

Como mejora a este análisis, tenemos:

- *Intentar ver si la normalización de las variables es óptima* para este modelo; recordemos que dichos valores vienen de un proceso KDD desarrollado en la asignatura de Minería de Datos
- *Buscar un algoritmo o modelo que nos permita analizar mas variables independientes*, para hacer mas robusto nuestro análisis.

2. Series Temporales

Siguiendo el hilo conductor del análisis del dataset descrito en el Capítulo 1, y orientándolo al **estudio de la probabilidad y predicción de una variable dependiente en el tiempo**, vamos a hacer a continuación el análisis del atributo TEMPERATURA_MEDIA de dicho conjunto de datos.

2.1. Descripción de los datos

Recordemos, que el dataset que hemos estado usando en esta Asignatura y la de Minería de Datos, trata de un conjunto de variables que nos ofrecían una serie de características diarias sobre los parques Walt Disney World en Orlando (FL) y que todo el estudio hecho hasta ahora se basa en determinar una semana óptima de viajar, para un turista; y la posibilidad de ofertar campañas de marketing para una empresa de turismo.

El dataset original una vez tratado por el proceso KDD consta de 2119 observaciones y 14 variables, y la que usaremos en el estudio de la serie temporal será el siguiente:

COLUMNA	TIPO	DESCRIPCION	FACTOR DE DECISION
TEMPERATURA_MEDIA	Numérico	Media de la temperatura en ºC	Factores Medioambientales

Tabla 5. Atributos del subset Serie Temporal. Elaboración propia.

La información que contiene el dataset original, **abarca desde el 01/01/2015 al 02/06/2021**; pero, **para darle un formato de ciclo al análisis de la serie temporal**; en el subset que desarrollemos en este capitulo, solo **utilizaremos los datos de temperatura entre el 01/01/2015 y el 31/12/2020**; es decir, 2079 observaciones.

Queda entonces supuesta que tendremos **365 observaciones por año en un ciclo anual y que el subset abarcará 6 años de estudio.**

2.2. Motivación del Análisis Descriptivo

Queremos entonces **valorar el comportamiento de este atributo en el tiempo, y si se trata de una serie estacional, si tiene ciclos periódicos, si presenta una tendencia; y, por ende, si esta información nos permite usar a esta variable como un dato significativo a la hora de escoger la mejor semana**. Es decir, decidir según nuestros gustos, cuando podríamos ir.

Esto añadiría mas robustez a nuestra decisión, recordemos que el capítulo anterior ya vimos mediante el Clustering, que semana se agrupa en cada clúster según la tasa de fallos de las atracciones y la cantidad de horas que abren los parques.

2.3. Estadística Descriptiva

Analizando descriptivamente esta variable tenemos:

	TEMPERATURA_MEDIA (ºC)
Valor mínimo	4.31
1er Cuartil	20.48
Mediana	24.66
Media	23.42
3er Cuartil	27.52
Máximo	30.98

Tabla 6. Estadísticos variable temperatura media. Elaboración propia.

¿Que información podemos obtener el Boxplot e Histograma de la variable?

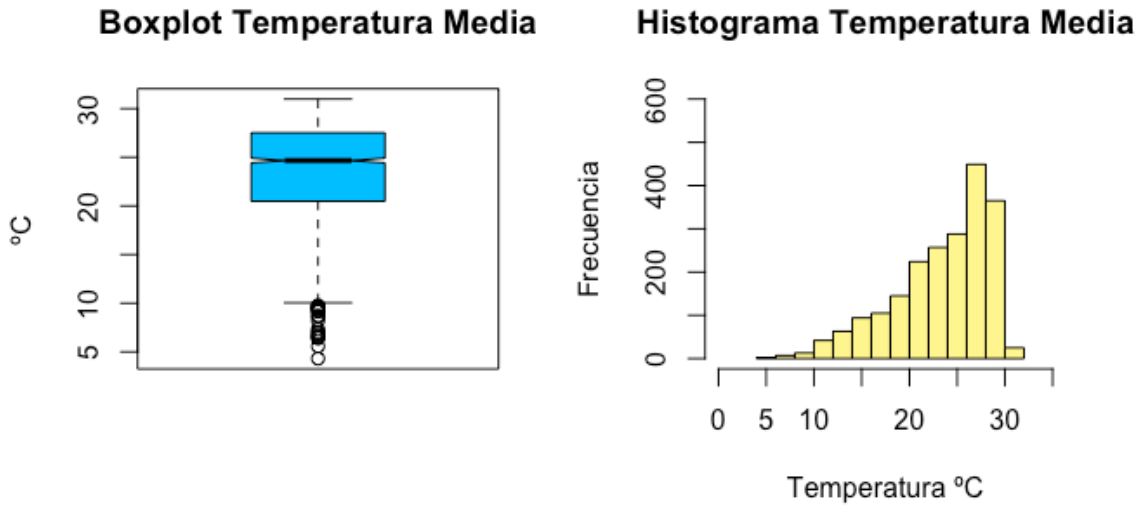

Ilustración 8. Boxplot e Histograma de la variable temperatura media

De ambos gráficos podemos sacar los siguientes supuestos:

- Por un lado, **existe posible Outliers de temperaturas bajas** durante la observación. ¿Alguna ola de frio?
- Tanto la media, como la mediana está muy cercana y que **el rango de variación de temperatura en la ciudad de Orlando es bastante estable**; lo que nos podría hacer intuir que tendremos un comportamiento estacional.
- El histograma confirma que gran parte de **las observaciones se ubicarán en un rango aproximado de 20ºC y 30ºC**

2.4. Moving Average

Vamos a partir analizando la componente de serie temporal de los datos que ofrece nuestra variable.

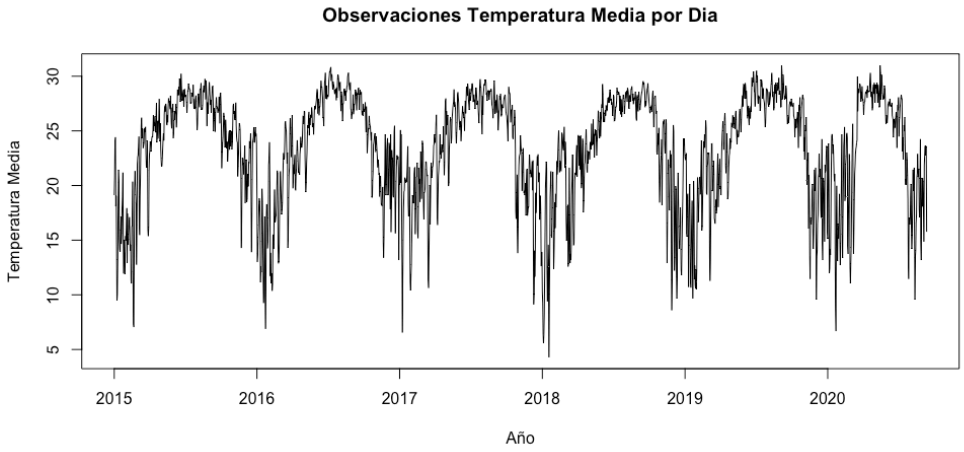

Ilustración 9. Plot Serie Temporal

Observamos entonces que los seis ciclos (2015-2020) que comentamos anteriormente, **muestra un comportamiento estacionario a lo largo de los años**; solo añadiendo valores presuntamente anómalos al inicio de 2018.

El modelo que usaremos en el análisis y descomposición de nuestra serie temporal, **serie del tipo ARIMA**; donde por simplicidad de la actividad, nos hemos basado en la función *"auto.arima"* del paquete **"forecast"** que nos ofrece R.

Así, por medio de un proceso iterativo, la función nos ayudará a escoger cuales son los mejores parámetros del modelo.

Para la elaboración del modelo hemos eliminado la estacionalidad de la serie:

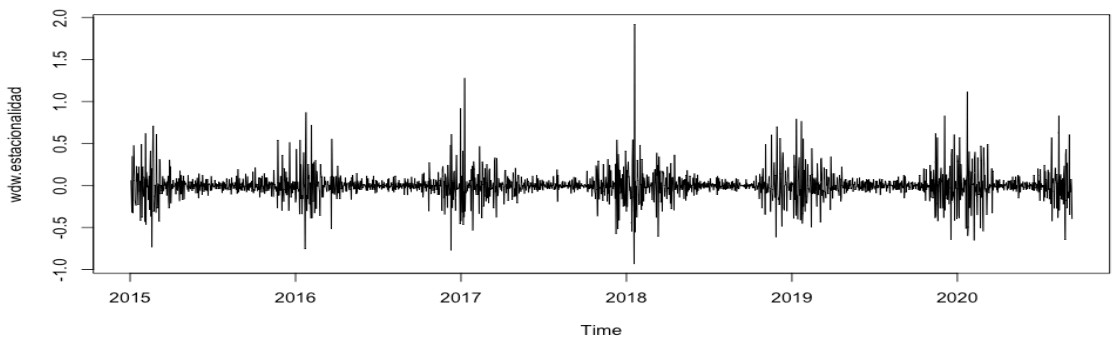

Ilustración 10. Serie temporal eliminado la estacionalidad

Obteniendo los siguientes resultados:

```
> wdw.auto.arima
Series: wdw.estacionalidad
ARIMA(5,0,0) with zero mean

Coefficients:
          ar1      ar2      ar3      ar4      ar5
      -0.6530  -0.7066  -0.6208  -0.4099  -0.2581
s.e.   0.0213   0.0240   0.0252   0.0240   0.0213

sigma^2 estimated as 0.01811:  log likelihood=1220.25
AIC=-2428.5   AICc=-2428.45   BIC=-2394.66
>
```

Ilustración 11. Resultado función auto.arima

Por lo tanto la ejecución de esta función, arroja que **el mejor modelo para el estudio de la predicción de la variable Temperatura Media es ARIMA(5,0,0).**

Sería entonces un modelo de orden 5, sin aplicar diferenciación; es decir, no es necesario eliminar la tendencia.

2.5. Descomposición de la serie temporal.

Ya por último, la descomposición de nuestra serie queda de la siguiente manera:

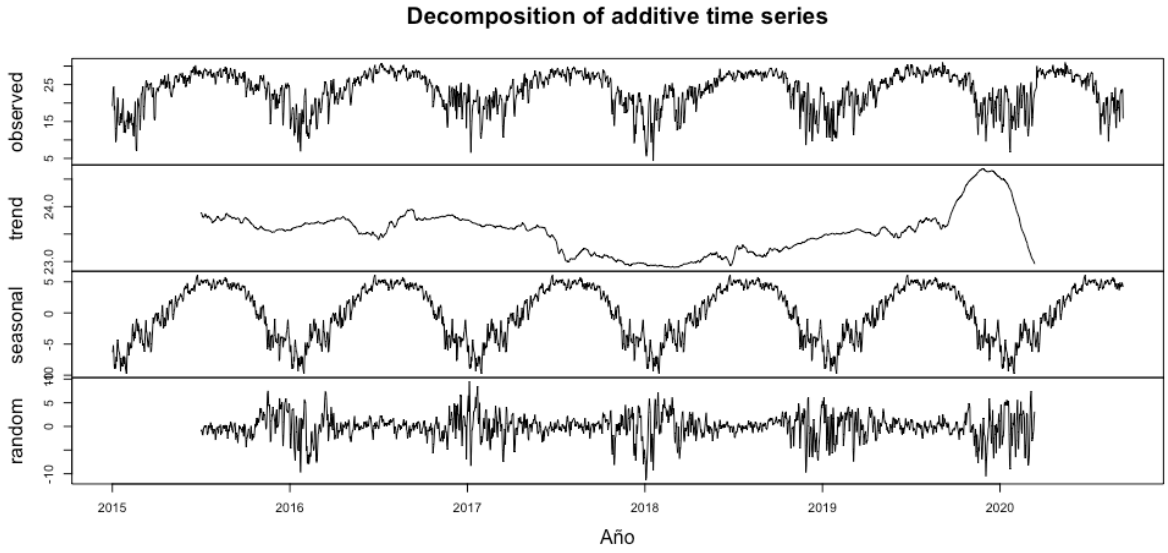

Ilustración 12. Descomposición de la serie temporal

Donde observamos que:

- **Los datos observados mantienen un comportamiento periódico**
- **Existe una tendencia casi nula**, salvo probables bajadas de temperatura en 2018 y un repunte en 2020.
- **Es totalmente estacional**, en ciclos anuales
- Existe un **posible ruido blanco en la transición de los años**, posiblemente por los Outliers de baja temperatura que vimos en el análisis estadístico. Recordemos que el rango intercuartilico está entre 20ºC y 30ºC.

3. Anexo (Script R)

El código en R utilizado en esta actividad se entrega como fichero anexo, al estar construido a través de R Markdown y exportado en .pdf.