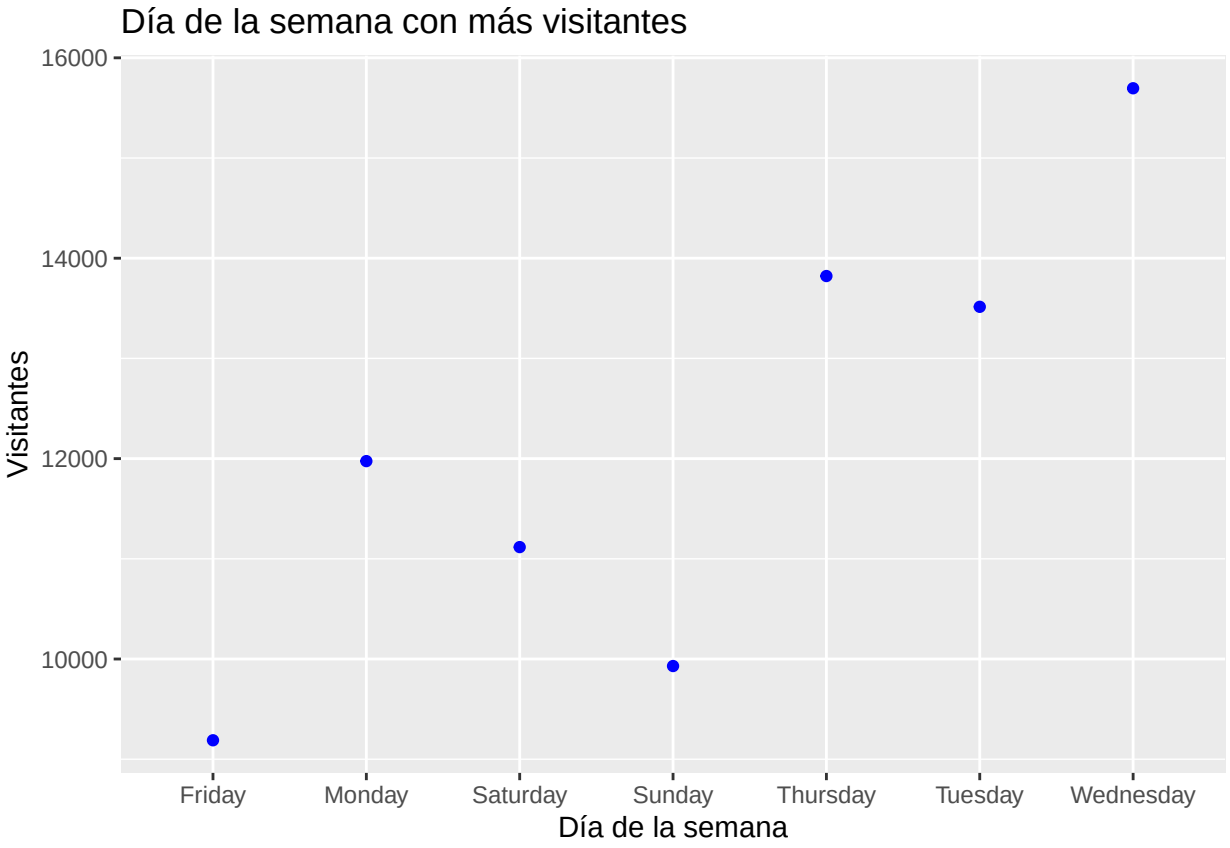
Proyecto Knn - Klustera

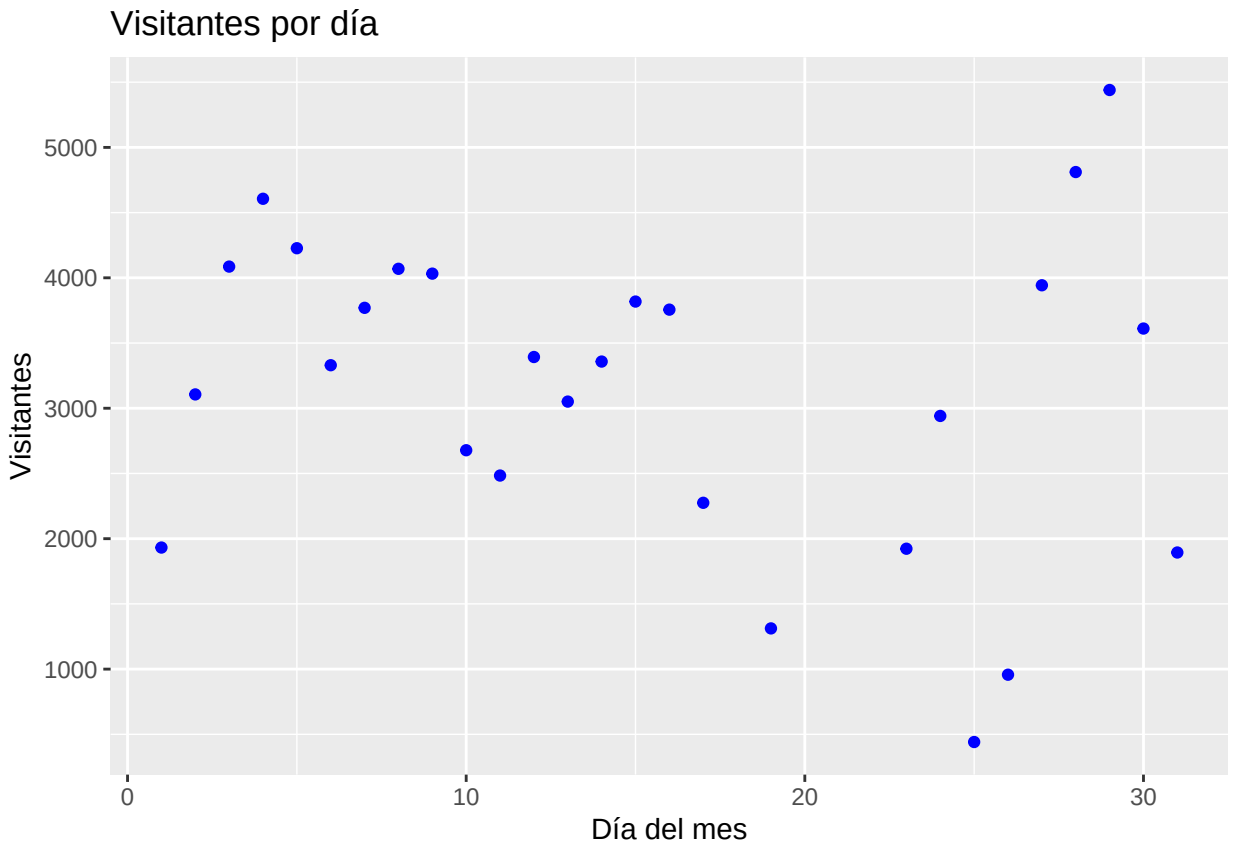
2020-11-13

Planteamiento de las preguntas y sus gráficas

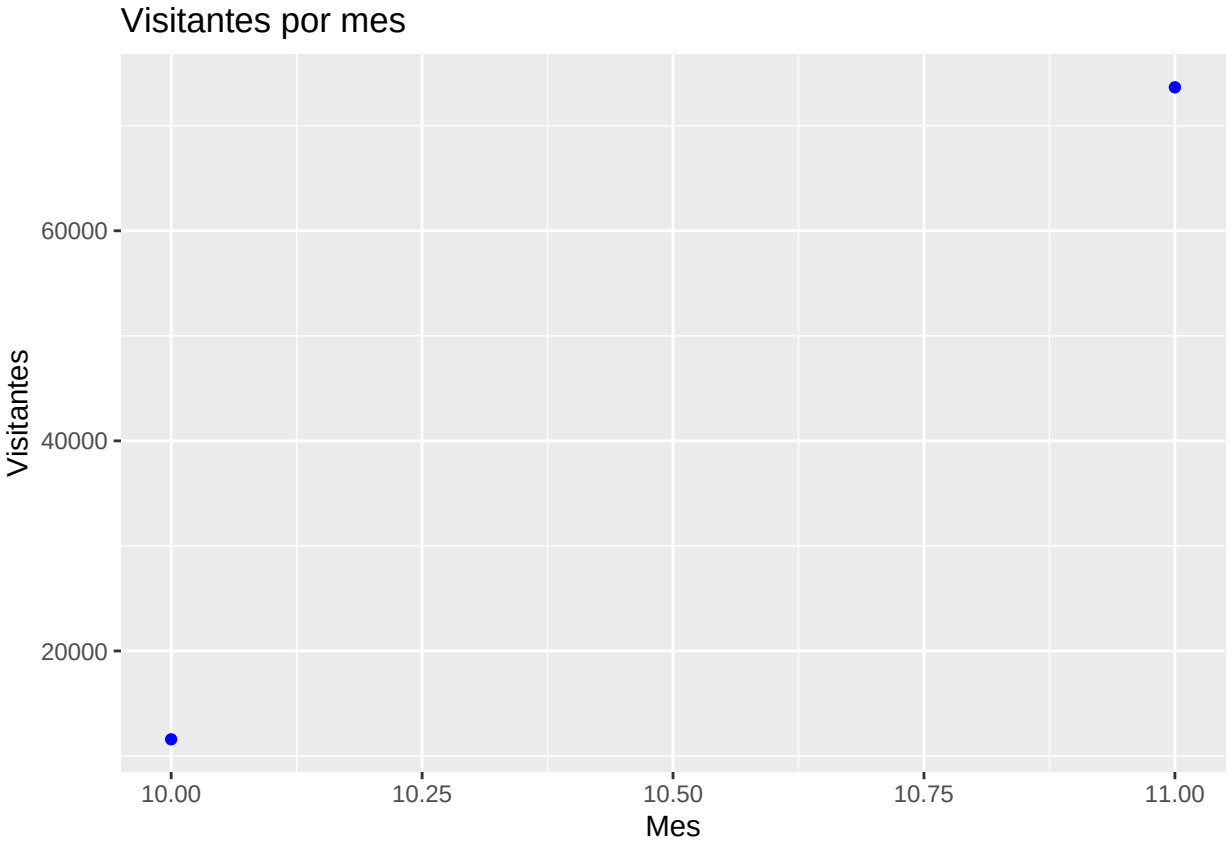
¿Que día de la semana contamos con mayor número de visitantes?



¿Cuantos visitantes hay por día?


Visitantes por día
Visitantes
5000
4000
3000
2000
1000
0
10
20
30
Día del mes

¿Que mes cuenta con mayor número de visitantes?



Entrenamiento con la Clasificación de datos mediante la categoría de los "vecinos" más cercanos (KNN)

En el proyecto de Klustera, utilizamos la Clasificación de datos mediante la categoría de los "vecinos" más cercanos (KNN)

Utilizaremos dos bases de datos de registros de quienes son y no son visitantes, que le llamaremos e y v. La BD e, tiene un columna llamada visitor, el cual ya identifica a partir de las observaciones que tiene, quienes si cumplen con la condición de visitante y quienes no. Esta tabla nos servirá para poder sacar el modelo y usarla con v.

Usando las columnas tiempodeses, day_tz, hour_tz, para usarlo en nuestro modelo como datos base para identificar a los visitantes, les aplicacamos una normalización por la gran diferencia entre mínimos y máximos de cada columna como se puede observar a continuación

Columna tiempodeses

Min. : 0 , 1st Qu.: 0 , Median : 0 , Mean : 2375 , 3rd Qu.: 421 , Max. :68062

Columna day_tz

Min. : 1.0 , 1st Qu.: 7.0 , Median :13.0 , Mean :14.8 , 3rd Qu.:24.0 , Max. :31.0

Columna hour_tz

Min. : 0.00 , 1st Qu.:10.00 , Median :14.00 , Mean :13.64 , 3rd Qu.:18.00 , Max. :23.00

Al normalizarlos, proseguimos en usarlos para obtener las muestras para nuestro modelo el cual tomamos una relación de 80-20 para los datos de entrenamiento y prueba.

Usamos la formula de Knn

```
e_test_pred <- knn(train = e_train, test = e_test, cl = e_train_labels, k = 3)
```

Usamos k=3 porque la raíz cuadra de 249556 es 499 y la consola nos indicaba que eran muchos puntos para analizar. Por lo que empezamos con un número menor. Obtenemos los siguientes datos:

El total de false y true que tenemos en la base de datos e y los porcentajes que representan en visitantes y no visitantes.

```
##
##  false   true
## 164312  85244
```

```
##
##    visitante No Visitante
##         34.2         65.8
```

Aquí observamos la tabla de confusión

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  49912
##
##
##                | e_test_pred
## e_test_labels |    visitante | No Visitante |    Row Total |
## --------------|--------------|--------------|--------------|
##     visitante |        15686 |         1164 |        16850 |
##               |        0.941 |        0.035 |              |
##               |        0.314 |        0.023 |              |
## --------------|--------------|--------------|--------------|
##  No Visitante |          989 |        32073 |        33062 |
##               |        0.059 |        0.965 |              |
##               |        0.020 |        0.643 |              |
## --------------|--------------|--------------|--------------|
##  Column Total |        16675 |        33237 |        49912 |
##               |        0.334 |        0.666 |              |
## --------------|--------------|--------------|--------------|
##
##
```

Viendo los valores de la cantidad de no visitantes en visitantes, quise buscar disminuir la cantidad de no visitantes en visitantes por lo que eleve el dato de k en 100 y da el siguiente resultado

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  49912
##
##
##                | e_test_pred
## e_test_labels |    visitante | No Visitante |    Row Total |
## --------------|--------------|--------------|--------------|
##     visitante |        11542 |         5308 |        16850 |
##               |        0.998 |        0.138 |              |
##               |        0.231 |        0.106 |              |
## --------------|--------------|--------------|--------------|
##  No Visitante |           27 |        33035 |        33062 |
##               |        0.002 |        0.862 |              |
##               |        0.001 |        0.662 |              |
## --------------|--------------|--------------|--------------|
##  Column Total |        11569 |        38343 |        49912 |
##               |        0.232 |        0.768 |              |
## --------------|--------------|--------------|--------------|
##
##
```

por ultimo intentamos con k = 200 y obtenemos un No visitante = 0

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  49912
##
##
##                | e_test_pred
## e_test_labels |    visitante | No Visitante |    Row Total |
## --------------|--------------|--------------|--------------|
##     visitante |         7355 |         9495 |        16850 |
##               |        1.000 |        0.223 |              |
##               |        0.147 |        0.190 |              |
## --------------|--------------|--------------|--------------|
##  No Visitante |            0 |        33062 |        33062 |
##               |        0.000 |        0.777 |              |
##               |        0.000 |        0.662 |              |
## --------------|--------------|--------------|--------------|
##  Column Total |         7355 |        42557 |        49912 |
##               |        0.147 |        0.853 |              |
## --------------|--------------|--------------|--------------|
##
##
```

Hacemos el mismo procedimiento para la base de datos v donde debemos probar nuestro modelo que obtuvimos en e y usamos un valor en k de 80 para obtener 0 de igual forma:

```
##
##  false   true
## 164312  85244
```

```
##
##    visitante No Visitante
##         34.2         65.8
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  9000
##
##
##                | v_test_pred
## e_test_labels |    visitante | No Visitante |    Row Total |
## --------------|--------------|--------------|--------------|
##     visitante |            1 |         3236 |         3237 |
##               |        1.000 |        0.360 |              |
##               |        0.000 |        0.360 |              |
## --------------|--------------|--------------|--------------|
##  No Visitante |            0 |         5763 |         5763 |
##               |        0.000 |        0.640 |              |
##               |        0.000 |        0.640 |              |
## --------------|--------------|--------------|--------------|
##  Column Total |            1 |         8999 |         9000 |
##               |        0.000 |        1.000 |              |
## --------------|--------------|--------------|--------------|
##
##
```

Al final, agregamos a esa base de datos la columna de visitor para agregar los datos de visitantes y no visitantes a la tabla.