Universidad Autónoma de Yucatán

Maestría en Ciencias de la Computación

Métodos Estadísticos de Machine Learning

Proyecto 2

Autor: Mario Herrera Almira

20 de marzo del 2023

En la página web https://datasciencedojo.com/blog/datasets-data-science-skills/?utm_campaign=DSD%20blogs%202022&utm_content=223545409&utm_medium=social&utm_source=twitter&hss_channel=tw-1318985240 se presenta un repositorio de datos, que se pueden utilizar con el objetivo de mejorar las habilidades en el área de Ciencia de Datos.

Para este proyecto van a tomar la base de datos 23 y proponer un modelo de regresión lineal para el problema propuesto.

Se debe escribir un reporte que incluya al menos la siguiente información:

1. Un análisis descriptivo de los datos, que incluya:
    a. Identificación de la variable dependiente y la independiente (5 puntos)
    b. Identificar la relación que existe entre la variable dependiente y cada variable independiente (15 puntos)
2. El modelo de regresión estimado: su ecuación y el método utilizado para llegar al modelo (10 puntos)
3. Determinar cuáles variables son significativas en presencia de las otras en el modelo estimado (10 puntos)
4. Evaluar el modelo: Presenta el valor del R^2 y su interpretación. Presenta el valor del MSE. (10 puntos)

Respuesta

Para resolver este problema solamente se tiene en cuenta la base de datos de “Hours”, la base de datos “Day” no se tiene en cuenta porque contiene la misma información que “Hours” pero con una columna menos que es la de “hr”, esta columna en mi opinión si es importante tenerla en cuenta porque la hora del día si puede ser influyente sobre la cantidad de bicicletas que se pueden rentar, ya que hay horarios del día en los que normalmente las personas no rentan bicicletas.

1-)

a-)

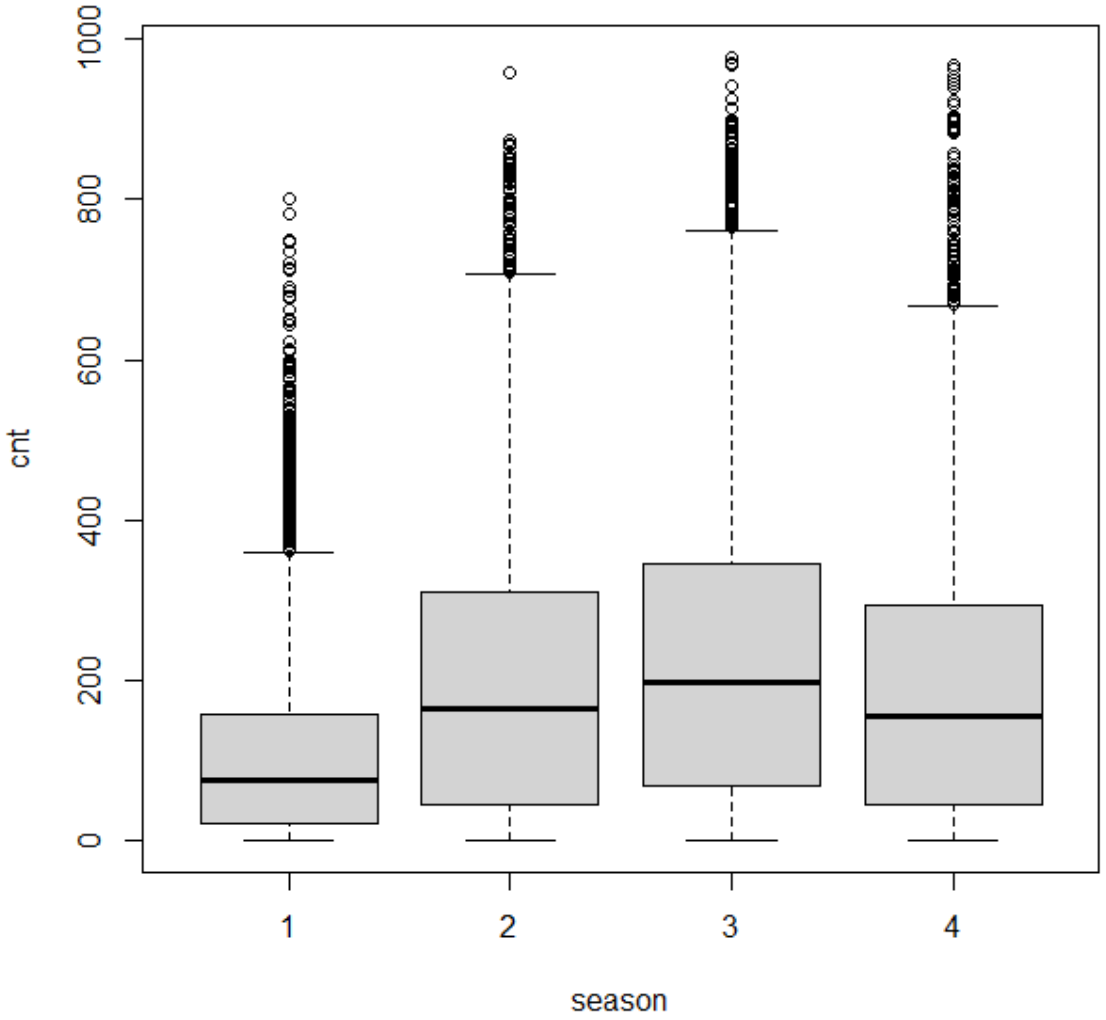
La **variable dependiente** es la cantidad de bicicletas rentadas (cnt) ya que es el valor que se quiere predecir en este ejercicio utilizando regresión lineal y además es la única que puede depender de cómo se comporten el resto de las variables.

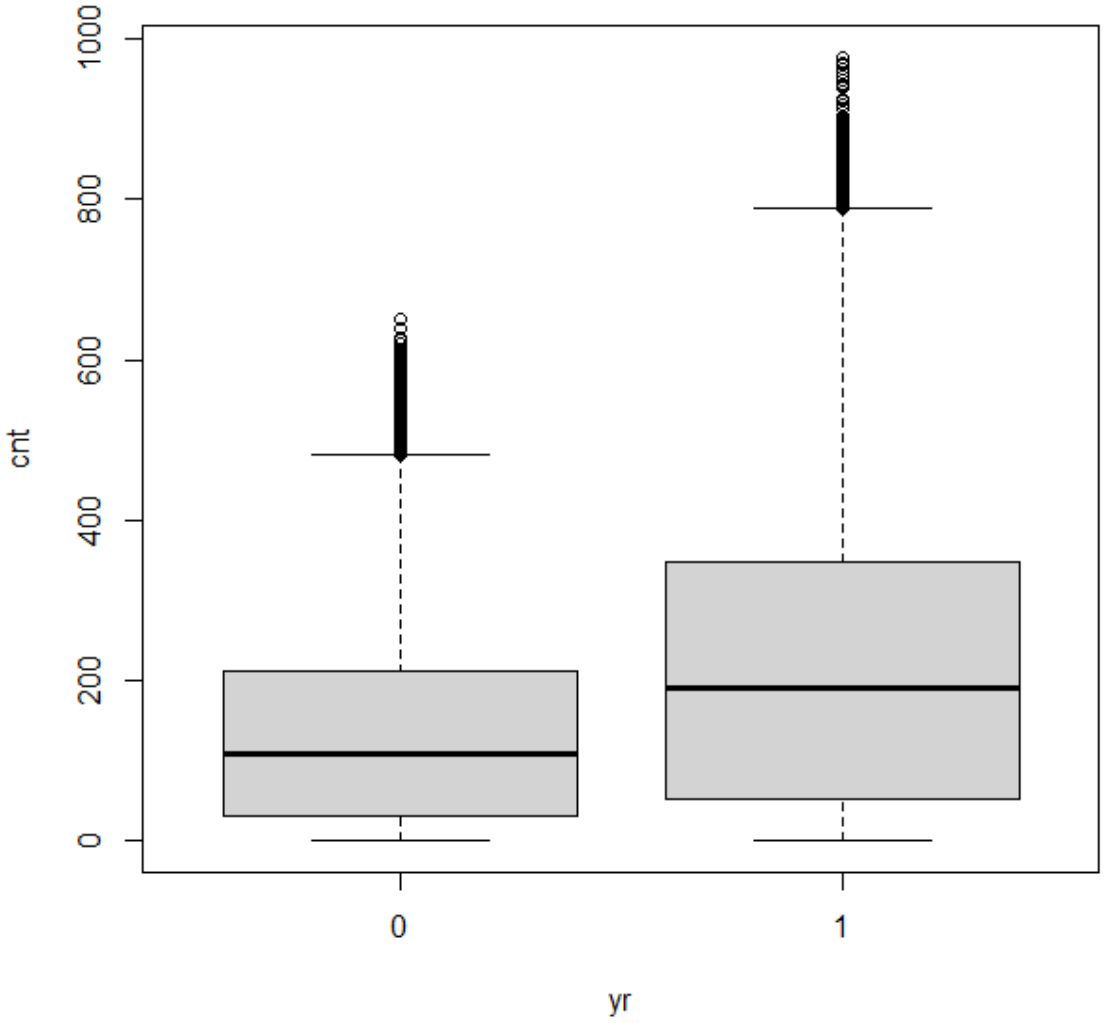
Las **variables independientes** son todas las demás sin contar a instant, dteday, casual ni registered. Instant no se tiene en cuenta porque esta variable es solo un valor de identificación para cada uno de los registros, no tiene ninguna influencia en el comportamiento de los datos. Dteday también es una variable que no aporta información sobre el comportamiento d ellos datos ya que solo registra la fecha en la que se hizo la operación, además que el año y el mes ya se encuentran en columnas independientes. Las variables casual y registered no son factores medio

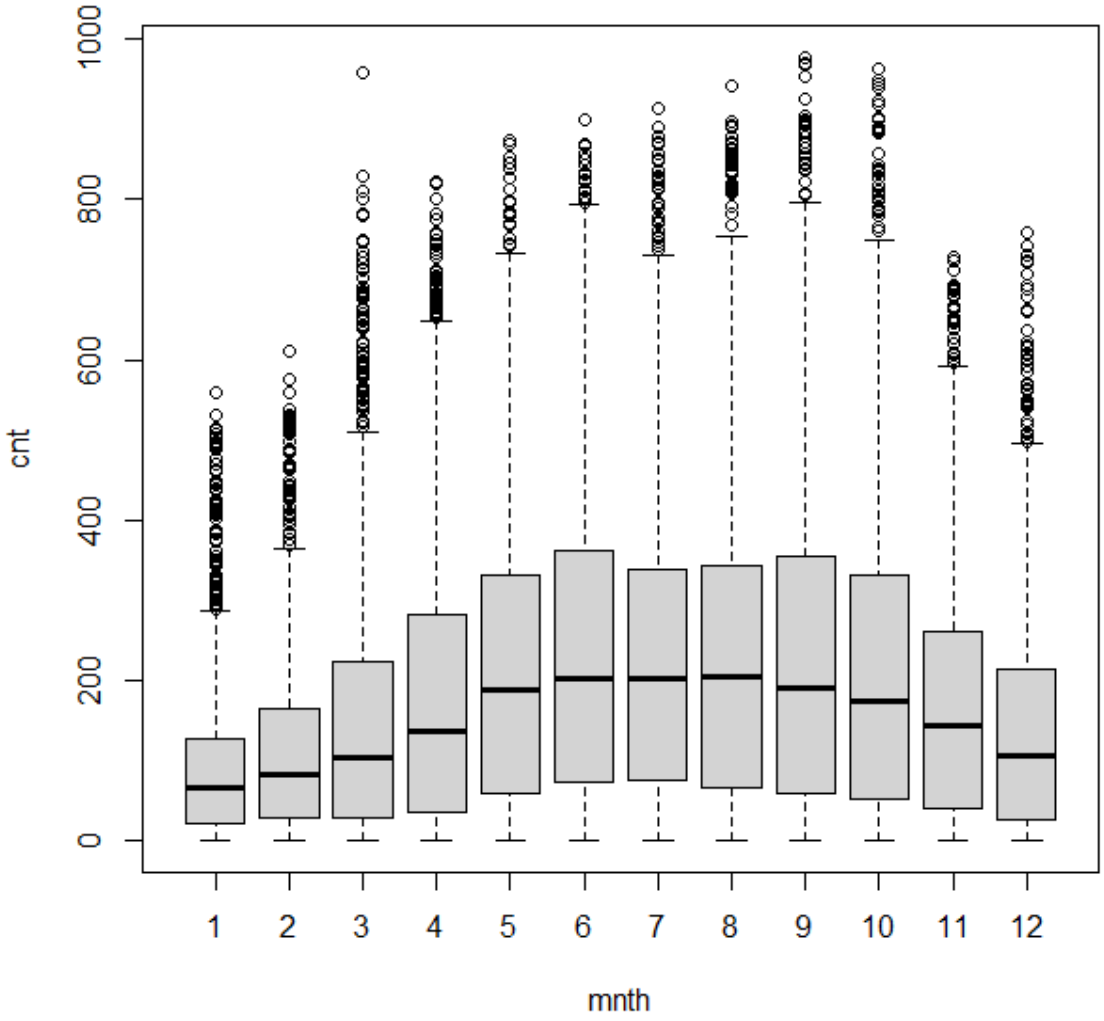
ambientales que influyan en la la cantidad de bicicletas que se rentan por lo que no se tienen en cuenta tampoco para el modelo.

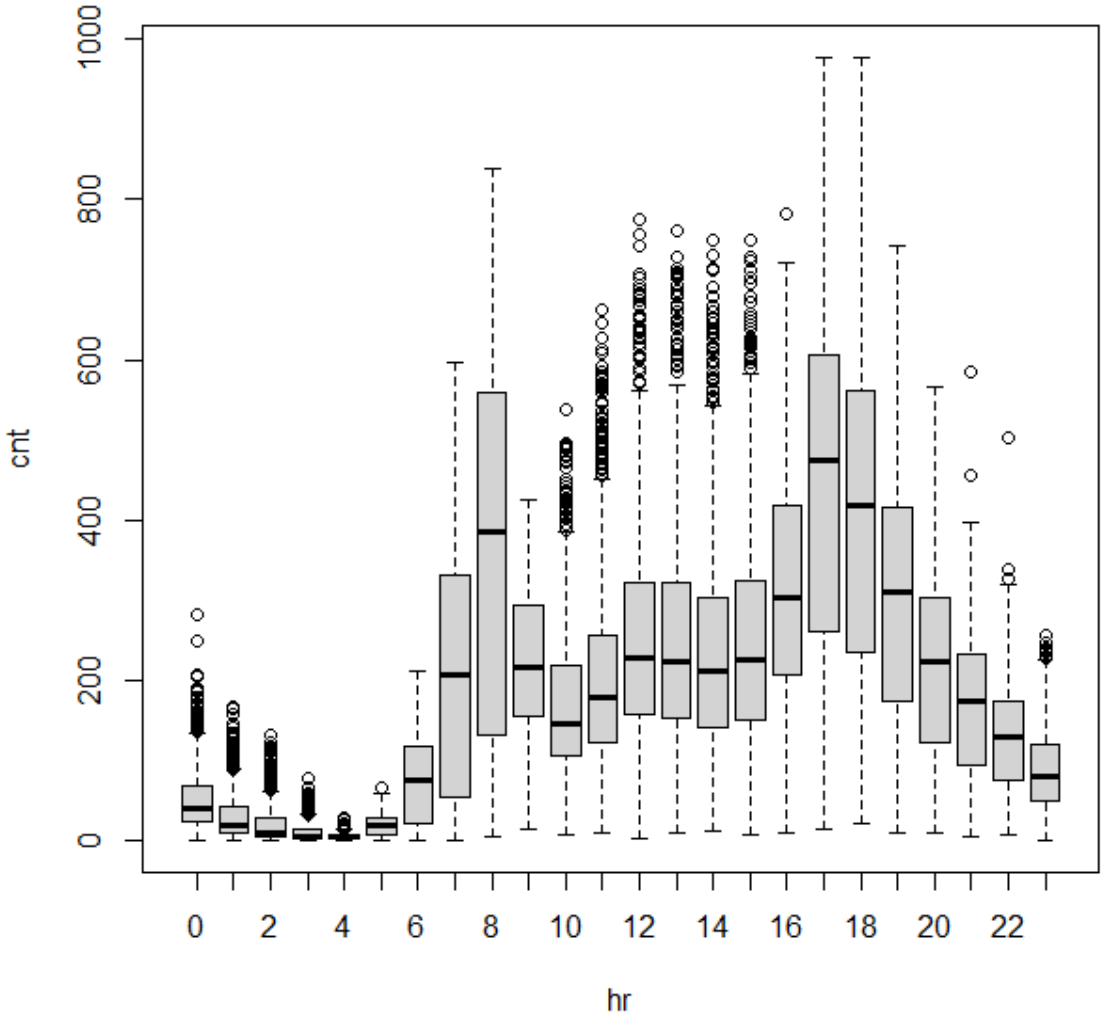
b-)

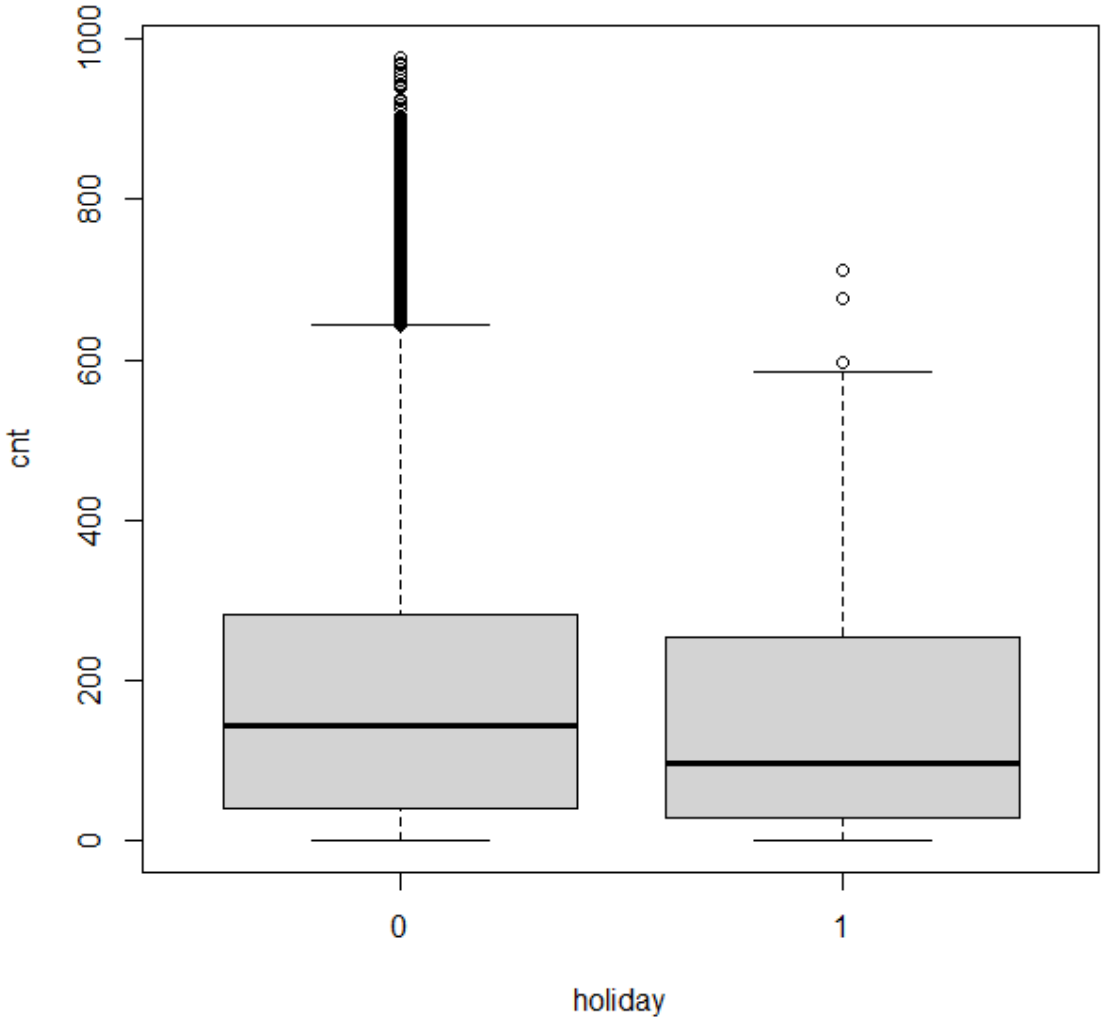
En las gráficas siguientes se puede observar que la variable dependiente “cnt” posee una correlación relativamente considerable con las variables independientes: “season”, “yr”, “mnth”, “hr”, “weekday” y “wheathersit” ya que sus gráficos de cajas y bigotes muestran que la cantidad de bicicletas rentadas varía de forma visible con respecto a cada una de las categorías. Sin embargo, las variables “workingday” y “holiday” muestran muy poca variación en los centros de los gráficos por lo que es posible que estas variables queden fuera del modelo final cuando se analice su correlación en presencia de otras variables.

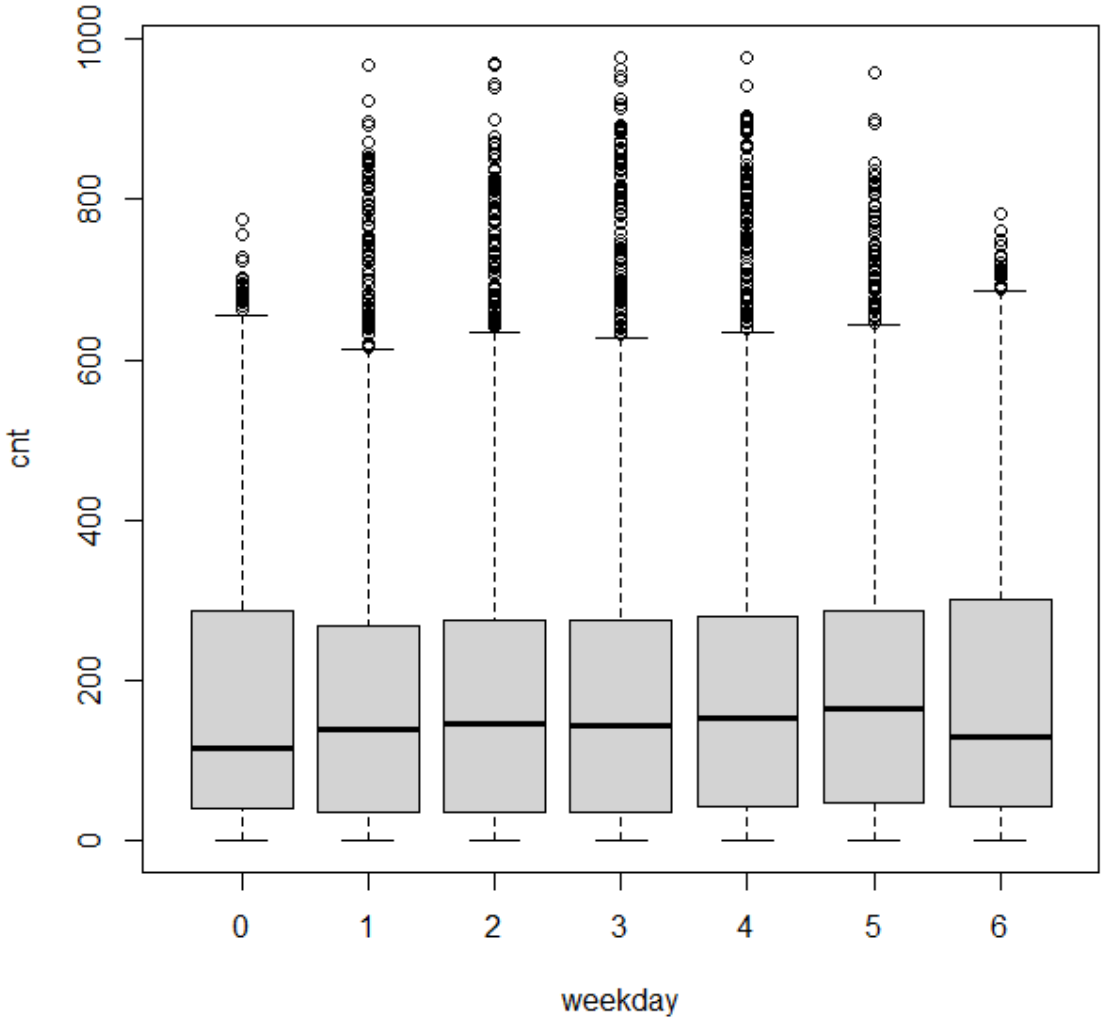


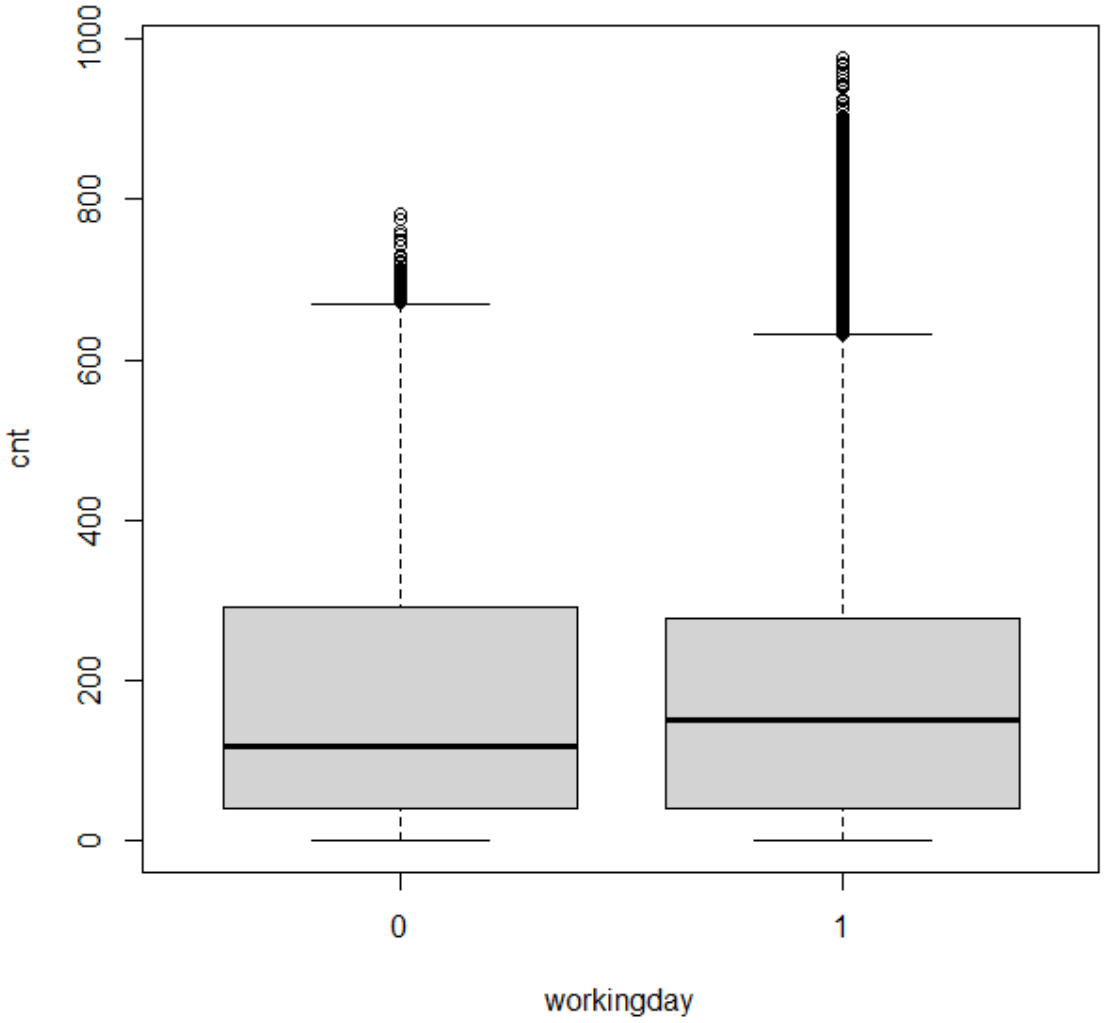

cnt
0
200
400
600
800
1000
0
1
yr

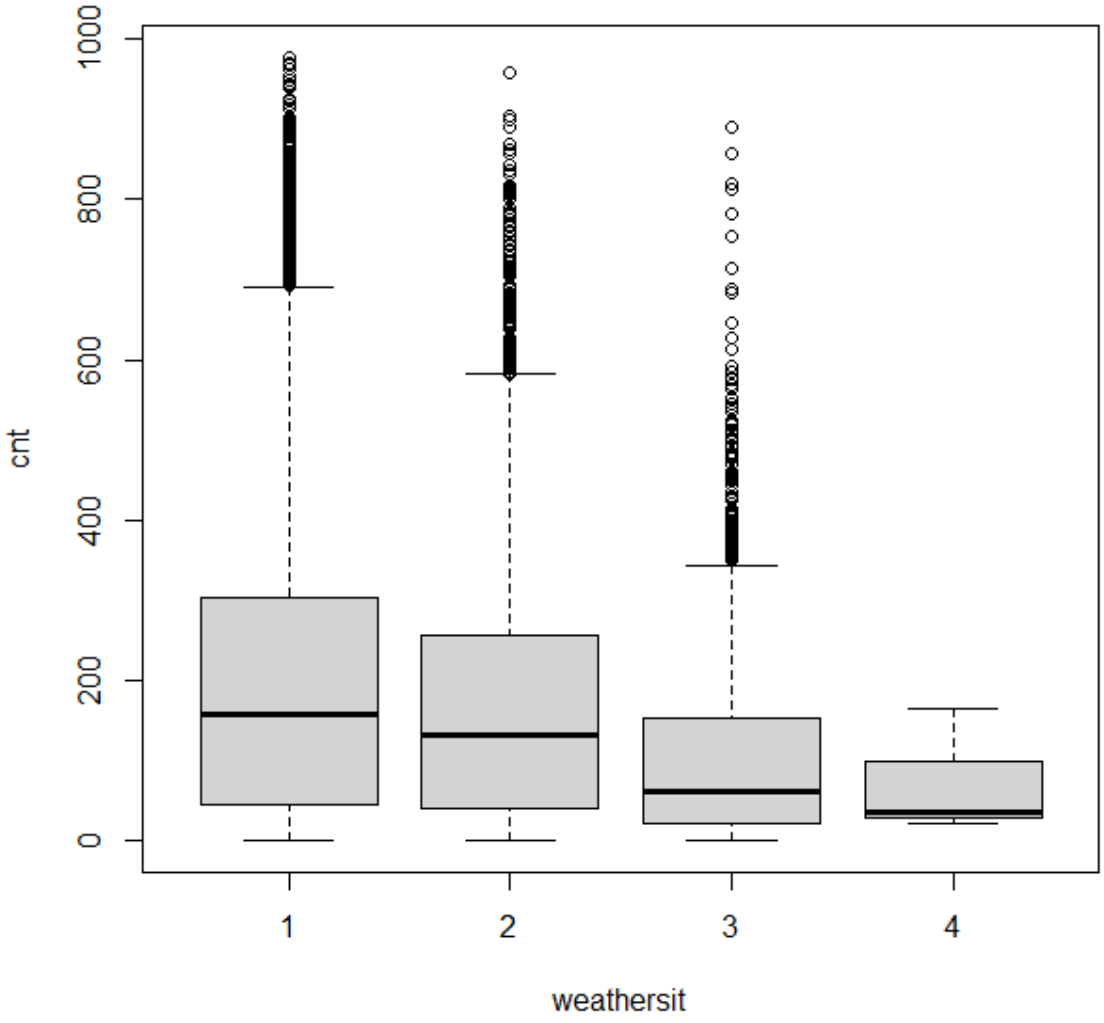

cnt
0
200
400
600
800
1000
1
2
3
4
5
6
7
8
9
10
11
12
mnth


cnt
0
200
400
600
800
1000
0
2
4
6
8
10
12
14
16
18
20
22
hr

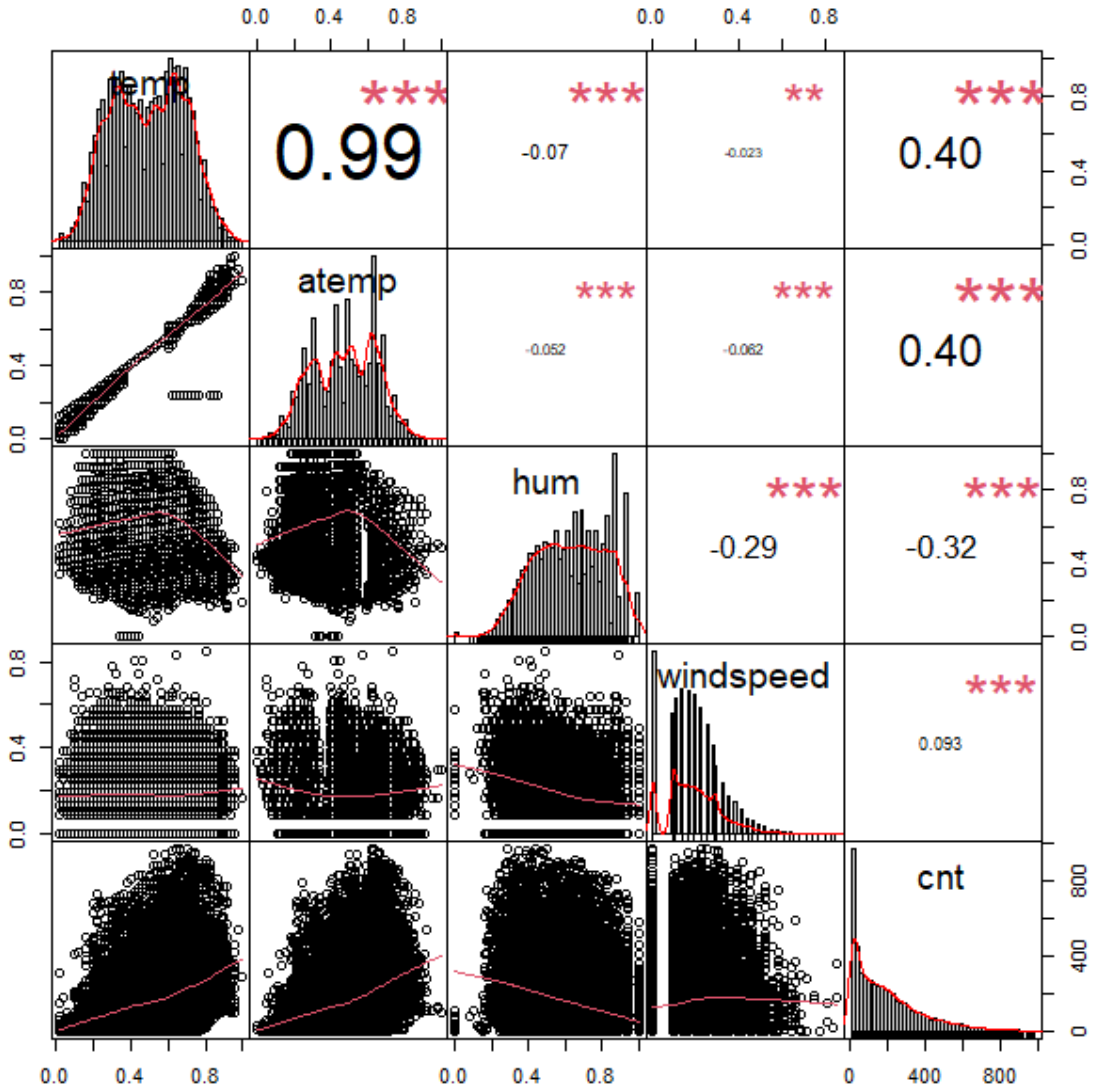

cnt
0
200
400
600
800
1000
0
1
holiday


cnt
0
200
400
600
800
1000
0
1
2
3
4
5
6
weekday


cnt
0
200
400
600
800
1000
0
1
workingday



En la tabla siguiente se puede observar que hay dos variables que poseen una fuerte correlación con “cnt”, estas son “temp” y “atemp”, con correlaciones de 0.40 cada una. El resto de las variables en esta tabla no poseen una fuerte correlación con “cnt”. En las dos gráficas de correlación se puede observar que la relación de estas variables con “cnt” es creciente.



Utilizando esta información se puede concluir que las variables que más relacionadas se encuentran con el comportamiento de “cnt” son: “hr”, “temp” y “atemp”. Por lo que estas variables podrían ser parte del modelo final en dependencia de que tanto estén correlacionadas entre ellas.

2-)

El método utilizado para llegar al modelo que le da solución a este problema fue el de “Selección hacia adelante” utilizando como variables predictoras todas aquellas presentes en el modelo menos las cuatro que fueron descartas al inicio por las razones mencionadas (instant, dteday, casual y registered).

En el punto 3 de este ejercicio se explica cuál fue la variable que se eliminó por no tener una influencia significativa en presencia de las demás. Luego de eliminar dicha variable la ecuación propuesta para resolver el problema es la siguiente:

$$cnt = -83.63 + (-17.29 * hr1) + (-26.37 * hr2) + (-37.11 * hr3) + (-40.26 * hr4) + (-23.50 * hr5) + (35.39 * hr6) + (170.42 * hr7) + (310.80 * hr8) + (163.10 * hr9) + (108.44 * hr10) + (133.84 * hr11) + (173.14 * hr12) + (168.10 * hr13) + (152.25 * hr14) + (161.71 * hr15) + (223.83 * hr16) + (377.54 * hr17) + (170.42 * hr7) + (345.59 * hr18) + (236.91 * hr19) + (157.29 * hr20) + (107.84 * hr21) + (70.91 * hr22) + (32.11 * hr23) + (127.97 * atemp) + (85.43 * yr1) + (-10.41 * weathersit2) + (-65.19 * weathersit3) + (-62.58 * weathersit4) + (38.18 * season2) + (32.05 * season3) + (67.99 * season4) + (3.43 * mnth2) + (14.30 * mnth3) + (6.23 * mnth4) + (20.65 * mnth5) + (6.24 * mnth6) + (-13.27 * mnth7) + (7.90 * mnth8) + (32.27 * mnth9) + (15.84 * mnth10) + (-9.84 * mnth11) + (-6.26 * mnth12) + (-82.80 * hum) + (-16.95 * weekday1) + (-15.38 * weekday2) + (-12.60 * weekday3) + (-13.08 * weekday4) + (-8.78 * weekday5) + (16.09 * weekday6) + (26.23 * workingday1) + (-29.17 * windspeed) + (116.38 * temp)$$

3-)

Luego de crear el modelo una primera vez utilizando la selección hacia adelante con todas las variables predictoras se determinó que había una variable que no era necesarias ya que aportaban poca información en presencia de las demás. Este es el caso para la variable: “holiday”. Al eliminar esta variable el valor de R cuadrado ajustado no varía, se mantiene en 0.6864, y el estadístico F también se mantiene en 729.1, por estos motivos se consideró apropiado retirar esta variable del modelo final.

4-)

El valor final del R cuadrado ajustado para este modelo propuesto es de 0.6864 lo que es igual a un 68.64%, esto significa que el modelo explica el 68.64% de la variabilidad de “cnt” que es la variable dependiente. Por lo que se puede decir que el poder de predicción de este modelo es medianamente alto.

El valor del error cuadrático medio es de 10318.7, este valor es elevado debido a que este modelo no es capaz de predecir con un alto nivel de confianza a la variable “cnt”.

Segunda parte del Proyecto 1. Validación del modelo

1-)

a-)

El código R utilizado para resolver este ejercicio es el siguiente. También se adjuntó junto con el documento el archivo de Script de R que contiene este mismo código.

```
library(MASS)
library(ISLR2)
library("PerformanceAnalytics")
library(ggplot2)
library(magrittr)
library(interactions)
library(dplyr)
library(Metrics)
library(caret)

datosHour <-read.csv("hour.csv", header = T)
datosHourReduced <- select(datosHour, -instant, -dteday)
datosHourReduced <-na.omit(datosHourReduced)

attach(datosHourReduced)
head(datosHourReduced)
str(datosHourReduced)

datosHourReduced$season <- as.factor(datosHourReduced$season)
datosHourReduced$yr <- as.factor(datosHourReduced$yr)
datosHourReduced$mnth <- as.factor(datosHourReduced$mnth)
datosHourReduced$hr <- as.factor(datosHourReduced$hr)
datosHourReduced$holiday <- as.factor(datosHourReduced$holiday)
datosHourReduced$weekday <- as.factor(datosHourReduced$weekday)
datosHourReduced$workingday <- as.factor(datosHourReduced$workingday)
```

```
datosHourReduced$weathersit <- as.factor(datosHourReduced$weathersit)

boxplot(datosHourReduced$cnt ~ datosHourReduced$season, xlab = "season", ylab = "cnt")

boxplot(datosHourReduced$cnt ~ datosHourReduced$yr, xlab = "yr", ylab = "cnt")

boxplot(datosHourReduced$cnt ~ datosHourReduced$mnth, xlab = "mnth", ylab = "cnt")

boxplot(datosHourReduced$cnt ~ datosHourReduced$hr, xlab = "hr", ylab = "cnt")

boxplot(datosHourReduced$cnt ~ datosHourReduced$holiday, xlab = "holiday", ylab = "cnt")

boxplot(datosHourReduced$cnt ~ datosHourReduced$weekday, xlab = "weekday",
        ylab = "cnt")

boxplot(datosHourReduced$cnt ~ datosHourReduced$workingday, xlab = "workingday",
        ylab = "cnt")

boxplot(datosHourReduced$cnt ~ datosHourReduced$weathersit, xlab = "weathersit",
        ylab = "cnt")


chart.Correlation(select(datosHourReduced, temp, atemp, hum, windspeed, cnt),
        histogram = TRUE, pch = 19)


#############################################################################

###############Construyendo el modelo###################################


min.model <- lm(cnt~1, data = datosHourReduced)

summary(min.model)


fwd.model <- step(min.model, direction = "forward", scope
        = (~season + yr + mnth + hr + holiday + weekday + workingday
        + weathersit + temp + atemp + hum + windspeed))


summary(fwd.model)


fwd.model <- step(min.model, direction = "forward", scope
        = (~hr + atemp + yr + weathersit + season + mnth + hum + weekday
        + workingday + windspeed + temp))

summary(fwd.model)

fwd.model$coefficients
```

```
prediccion <- -fitted.values(fwd.model)

mse(cnt, prediccion)

############################################################

##############Validación del modelo mediante k - folds###############


folds <- createFolds(datosHourReduced$cnt, k = 10)

ctrl <- trainControl(method = "cv", index = folds)

model <- train(cnt ~ hr + atemp + yr + weathersit + season + mnth + hum + weekday
               + workingday + windspeed + temp, data = datosHourReduced,
               method = "lm", trControl = ctrl)

model$resample

print(model)


############################################################

########Validación del modelo mediante Split train - test##############

indice_entrenamiento <- sample(nrow(datosHourReduced),
               round(nrow(datosHourReduced) * 0.8), replace = FALSE)


datos_entrenamiento <- datosHourReduced[indice_entrenamiento, ]

datos_prueba <- datosHourReduced[-indice_entrenamiento, ]


modelo <- lm(cnt ~ hr + atemp + yr + weathersit + season + mnth + hum + weekday
               + workingday + windspeed + temp, data = datos_entrenamiento)


predicciones <- predict(modelo, newdata = datos_prueba)

MSE <- mean((predicciones - datos_prueba$cnt)^2)

R2 <- summary(modelo)$r.squared

MSE

R2
```

b-)

Los valores de RMSE y Rsquared para cada uno de los 10-folds se presentan en la tabla siguiente:

Fold	RMSE	Rsquared
1	142.2833	0.3865674
2	142.7336	0.3828851
3	142.3938	0.3831816
4	141.3432	0.3907630
5	142.5037	0.3848725
6	142.4667	0.3826995
7	142.3352	0.3859863
8	142.5303	0.3837663
9	142.1308	0.3863810
10	141.8380	0.3863052

c-) Los valores de RMSE y Rsquared que arroja la validación cruzada considerando el 10-fold son de 142.2101 y 0.3859 respectivamente.

2-)

Utilizando el método de Split train-test sobre el conjunto de datos y dividiendo la muestra en 80% para entrenar y 20% para probar se obtienen unos valores de MSE y Rsquared de 19735.08 y 0.3891 respectivamente.

Por tanto, la diferencia entre los resultados obtenidos con ambos métodos es la siguiente:

$$MSE = 19735.08 - 142.2101 = 19{,}592.87$$

$$\text{Rsquared} = 0.3891 - 0.3859 = 0.0032$$