BASE DE DATOS MASIVAS

11088

TRABAJO INTEGRADOR FINAL

Tonin Monzón Francisco
121461

A. Objetivo

B. Descripción

C. Alcance

a. Fuente de los Datos

b. Exploración de los datos

c. Preprocesamiento

d.1.Tratamiento de datos faltantes

d. Transformación

e. Algoritmos

f.1. Comprobación de supuestos

f.2. Ajuste del modelo

D. Bibliografía

A. Objetivo

El presente trabajo consiste en utilizar el dataset público que expone la empresa Rossman a través del sitio web Kaggle. El Dataset provisto contiene datos sobre ventas históricas de 1.115 farmacias pertenecientes a esta empresa. Para realizar el trabajo serán seguidos los pasos del descubrimiento de conocimiento y asi lograr predecir las ventas de seis semanas de cada una de las farmacias con el menor error posible.
Además se deberán encontrar la mejor forma de trabajar con el dataset para correr los algoritmos en un computadora con recursos básicos, ya que el mismo tiene un millón de instancias. Por lo que, se deben buscar estrategias para lograr una predicción precisa y rápida.

B. Descripción

En esta sección del trabajo serán descritas las variables que contiene el dataset. El lector comprenderá con cuales se cuenta inicialmente y pueda, posteriormente, entender las operaciones que se le aplicaran a las mismas.

Se tienen registros de ventas de los años:

- 2013
- 2014
- 2015
-

Los atributos con los que cuentan el dataset son:

- **Id** Representa la tupla Negocio y fecha dentro del set de entrenamiento.
- **Store** Id único para cada negocio
- **Sales** La ganancia de un dia. Esta variable es la que se debe predecir.
- **Customers** El número de clientes en un determinado día.
- **Open** Un indicador del estado del negocio:
-

- 1 Abierto
- 0 Cerrado

- **StateHoliday** indica un feriado estatal. Los tipos de feriado serán:
  - a = public
  - b = Easter holiday
  - c = Christmas
  - 0 = Ninguno
- **SchoolHoliday** indica si el negocio fue afectado por el cierre de las escuelas públicas.
- **StoreType** Diferencia entre 4 tipos de modelos de negocio. Los diferentes tipos son a,b,c y d.
- **Assortment** describe el nivel de variedad: a = basic, b = extra, c = extended
- **CompetitionDistance** distancia en metros de la competencia más cercana
- **CompetitionOpenSince[Month/Year]** da una estimación de la fecha en que abrió la competencia
- **Promo** Indica si la tienda tiene una promoción en ese dia.
- **Promo2** Es una promoción continua y consecutiva para algunas tiendas
  - 0 = el negocio no participa
  - 1 = el negocio participa
- **Promo2Since[Year/Week]** describe el año y la semana de calendario cuando el negocio empezó a participar de la promoción.
- **PromoInterval** describe el intervalo consecutivo en que empezo la promocion. Se usa el nombre del mes en el que empezó. Ejemplo: Feb, May, Aug.

C. Alcance

En la siguiente sección se expondrá el trabajo desde el análisis hasta llegar al objetivo final del mismo.

a. Fuente de los Datos

Serán listados en este apartado los archivos que entrega el sitio Kaggle para realizar la predicción.

El dataset está compuesto de cuatro archivos:

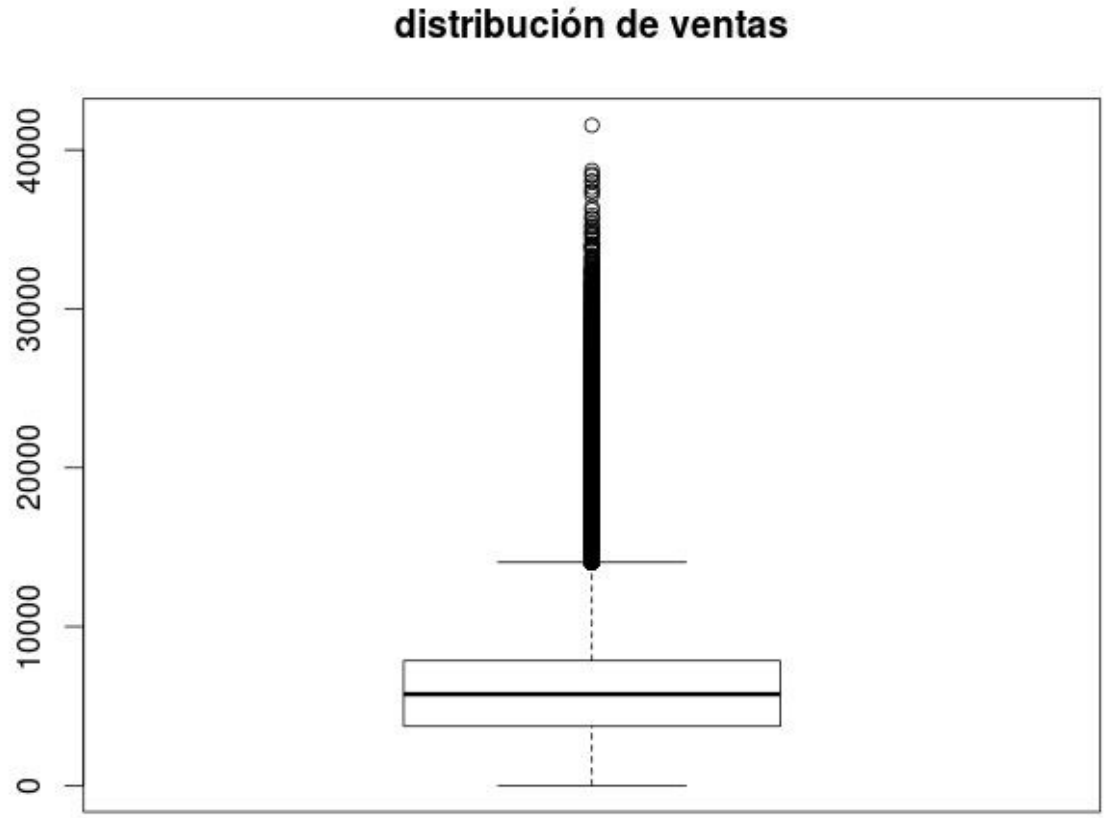
- Datos de entrenamiento
- Datos de testing

- Un ejemplo de cómo deben ser devueltas las predicciones
- Información adicional de cada negocio.

b. Exploración de los datos

En actual sección se analizarán todas las variables que posee el dataset, siempre en relación con la variable objetivo (Ventas). Se realiza este paso con el objetivo de encontrar patrones en los datos y documentar las mismos para poder luego realizar acciones al respecto.

A continuación se encuentra un gráfico caja que representa la distribución de la variable Ventas. El objetivo graficarlo es encontrar valores outliers y tener una visión de cómo se comporta la misma.



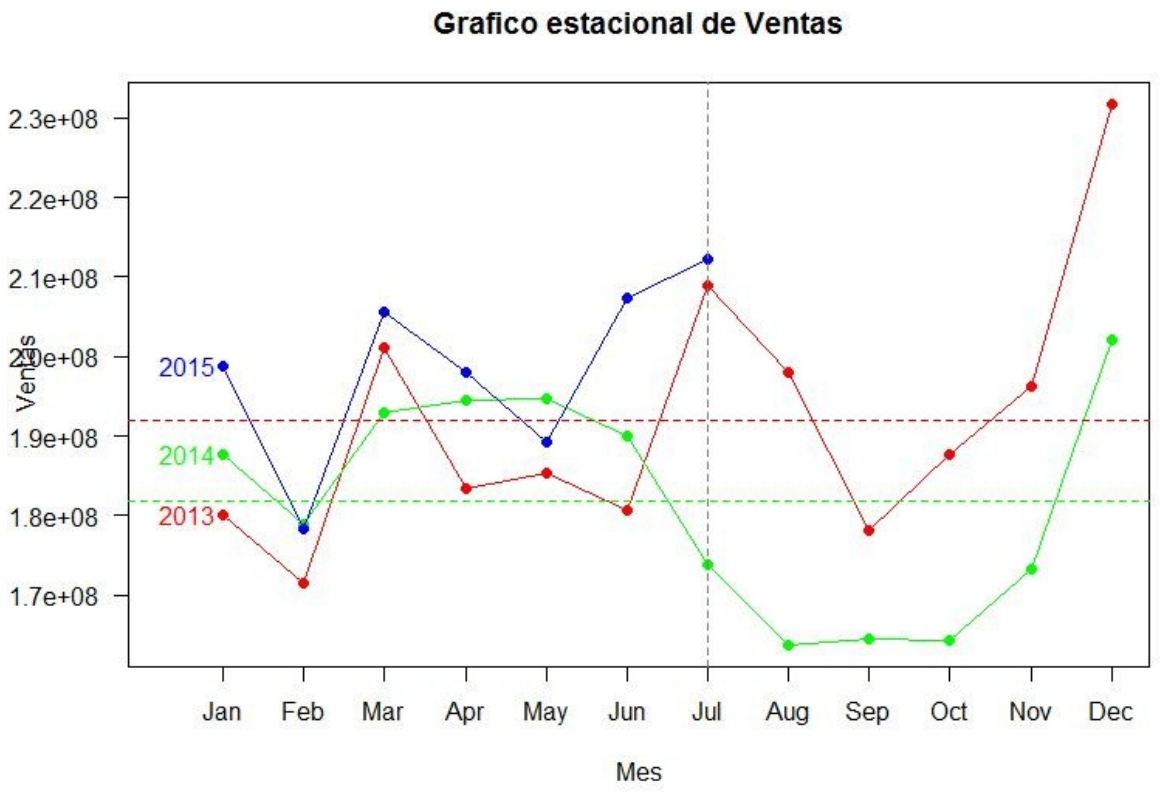
El gráfico expone que la media de las ventas está por debajo de los 10.000 euros y que se registraron días sin ventas. Además, se puede ver que son menos frecuentes las ventas mayores a 15.0000 euros y que se registra una venta única mayor a 40.000 euros (podría considerarse un valor outlier).

Ahora se muestra la ejecución de un comando clásico de R para ver las medidas tendenciales de un variable. El gráfico da una idea de cómo se comporta la variable, pero tener los valores numéricos aumenta la información sobre esta.

```
> summary(train$Sales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    3727    5744    5774    7856   41550
```

El valor extremo que es se podía detectar en el gráfico es 41550 euros y el mínimo es cero. Las ventas se encuentran muy concentradas entre los valores 3727 y 7856.

El siguiente gráfico de líneas intenta mostrar el comportamiento estacional de las ventas. Para graficarlo se calcula el promedio de ventas por cada mes y se representarán líneas diferentes por cada año. A partir del mismo se podrán sacar conclusiones sobre las ventas en distintos lapsos de tiempo.

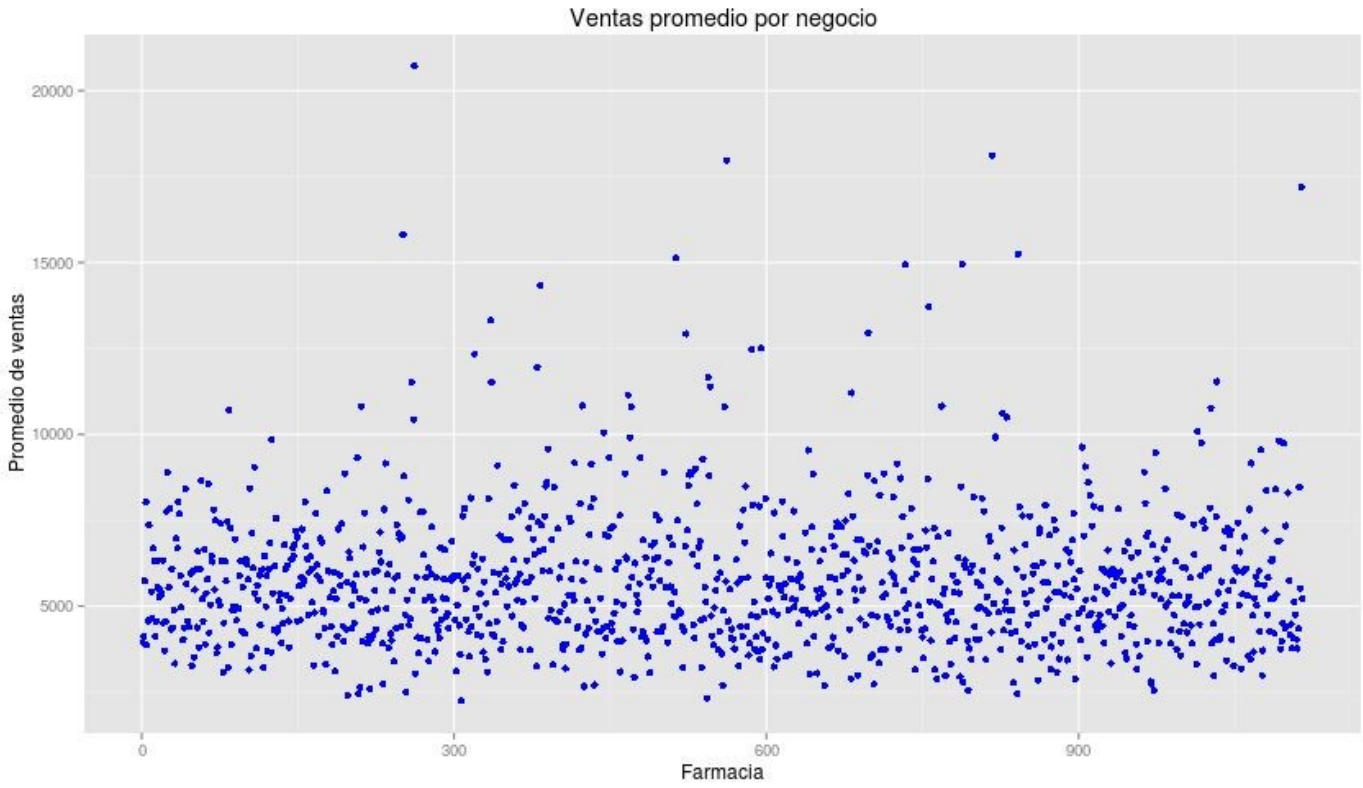


Este gráfico nos da una idea de la complejidad de predecir la ventas debido a la variabilidad que estas tienen. Queda claro que existe un comportamiento asociado a la temporalidad. Respecto al mismo se puede hacer las siguientes afirmaciones:

- Las ventas se redujeron durante el segundo mes de todos los años y aumentaron en Marzo.
- Las ventas tienden a aumentar durante los últimos 3 meses de los años registrados.
- El mayor valor de ventas promedio se registró en el mes de Diciembre para los años 2013 y 2014.
- El año 2013 registró las ventas promedio más altas y el 2014 las más bajas.

Si tuviéramos que predecir las ventas utilizando como base los años anteriores, nos encontraríamos con un gran reto, debido a que los años 13' y 14' tienen valores muy diferentes.

El siguiente diagrama de puntos refleja las ventas promedio de cada una de las 1.115 farmacias.

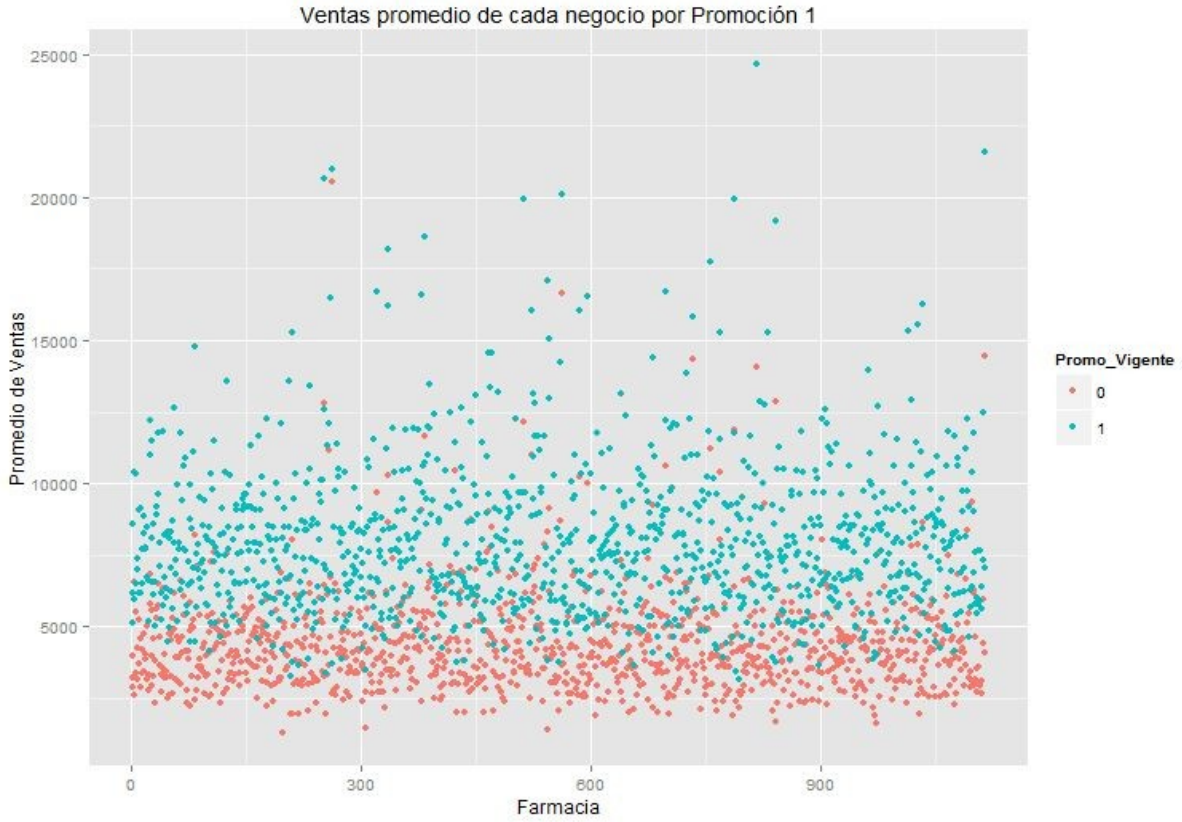


Claramente aparecen algunos promedios muy extremos en comparación con toda la distribución. Se debe averiguar a qué farmacias pertenecen esos valores y para detectar posibles anomalías.

ANÁLISIS DE PROMOCIONES

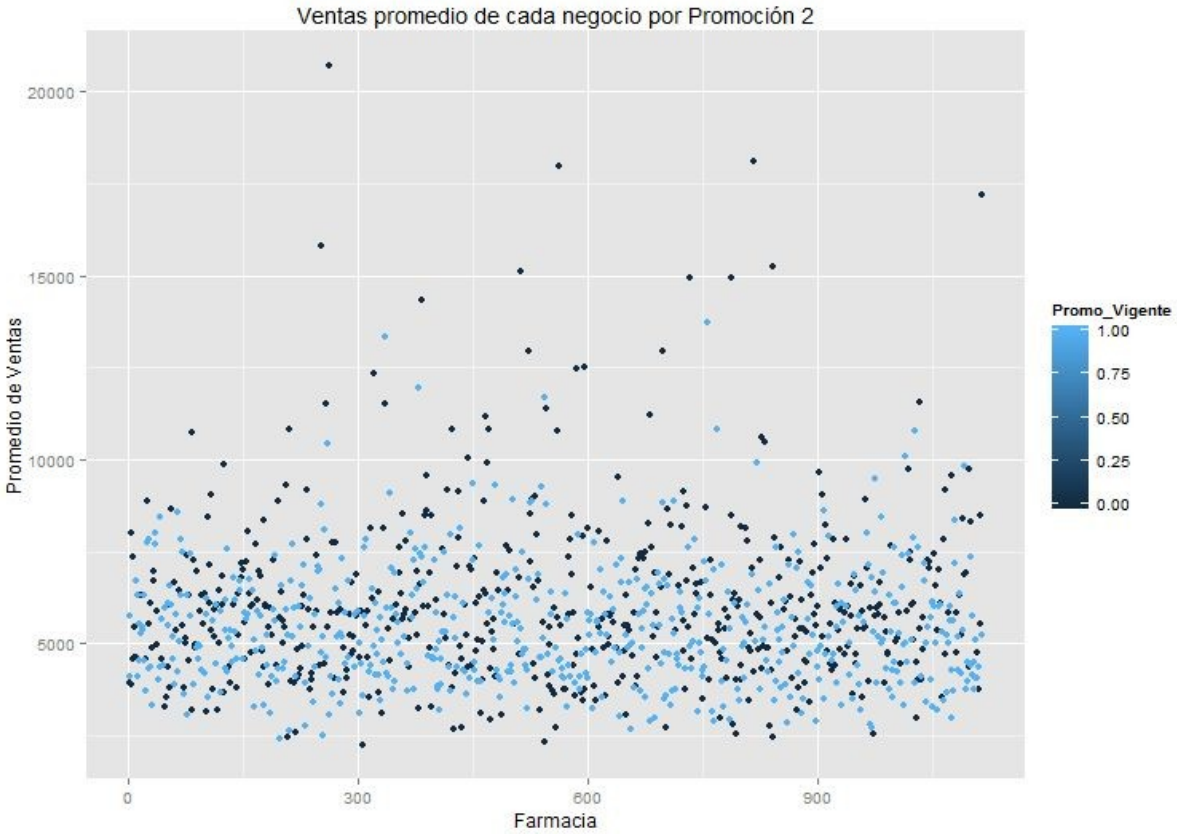
PROMOCIÓN I

Ahora estudiaremos el efecto de cada promoción en las ventas por separado.
Los dos diagramas de puntos que siguen representan la ventas promedio de cada farmacia cuando la promoción está vigente (puntos azules) y cuando no lo está (puntos rojos).

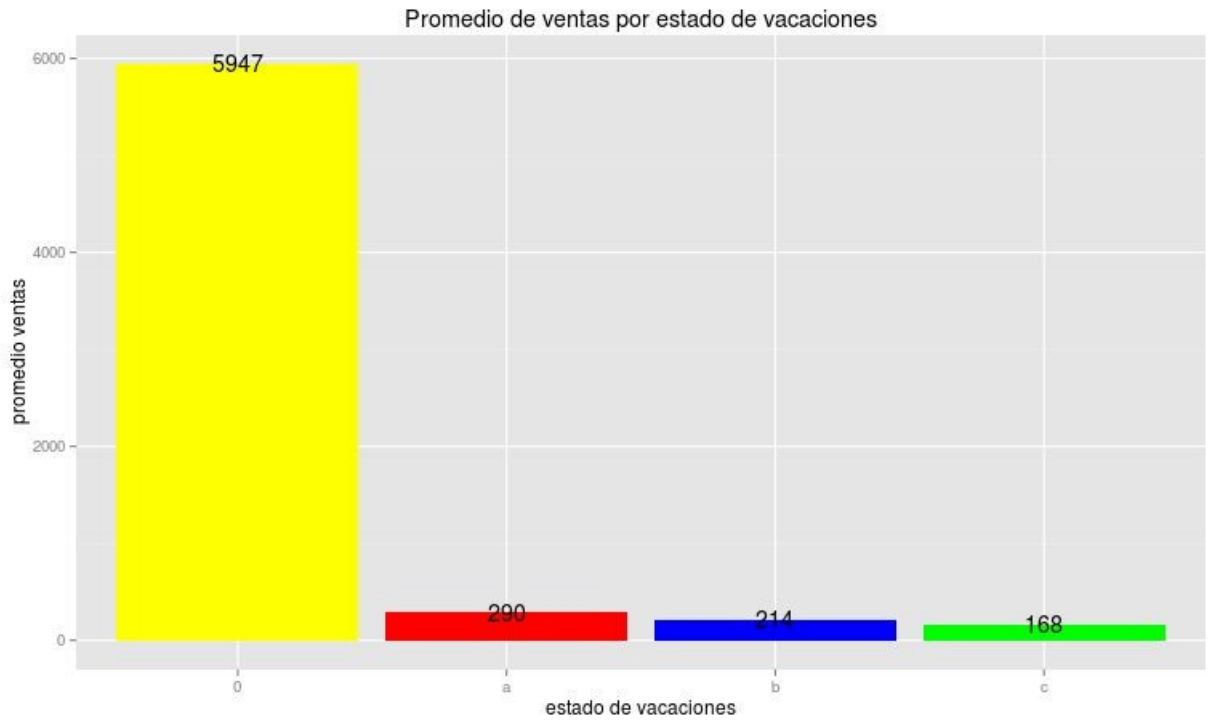


El gráfico indica que las ventas promedios por negocio son bajas cuando la promoción no esta vigente en relación a cuando si lo esta. Ambos puntos se encuentran bien agrupados, por lo que se ve claramente el patrón mencionado.

PROMOCIÓN II

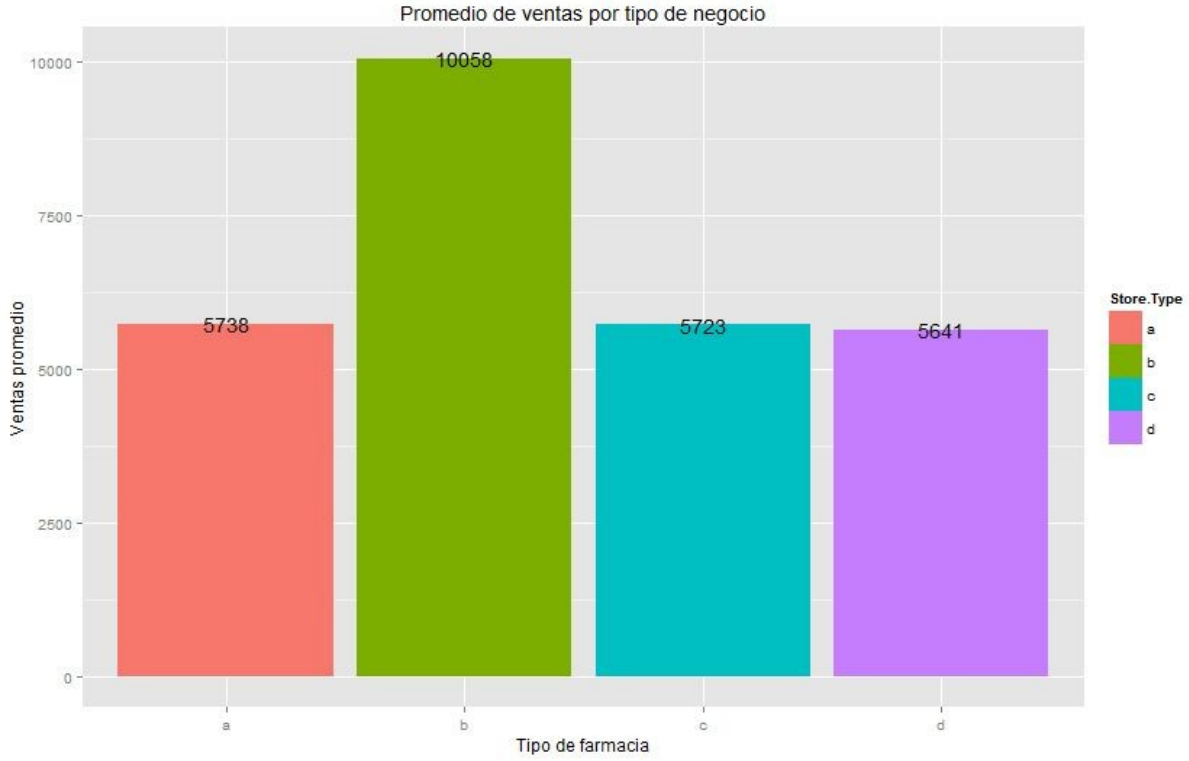


Comparando este gráfico con el de la promoción I, la promoción II no registra un patrón claro. Es decir, las ventas promedio con y sin promoción vigente ocupan un rango de ventas muy amplio y disperso.
El gráfico de barras a continuación representa las ventas promedio por feriados.



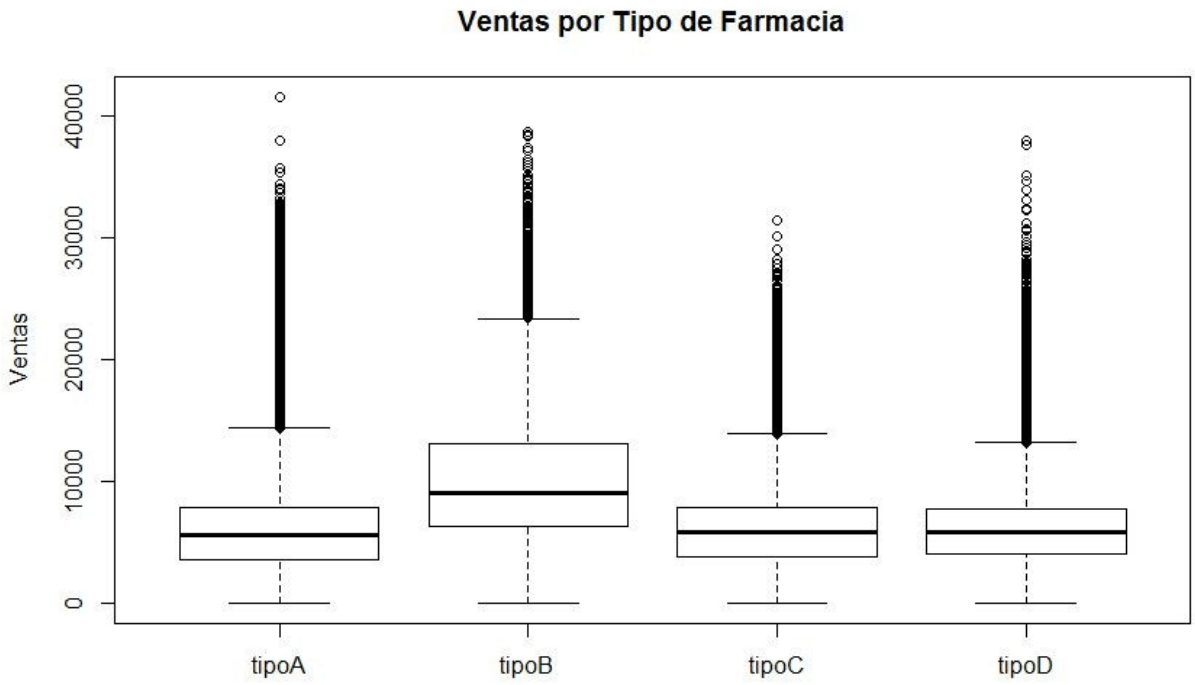
Se puede concluir que las ventas se reducen en días que son feriados (de cualquier tipo).

El siguiente gráfico de barras es similar al anterior, pero ahora, se busca conocer el comportamiento de las ventas por tipo de negocio.

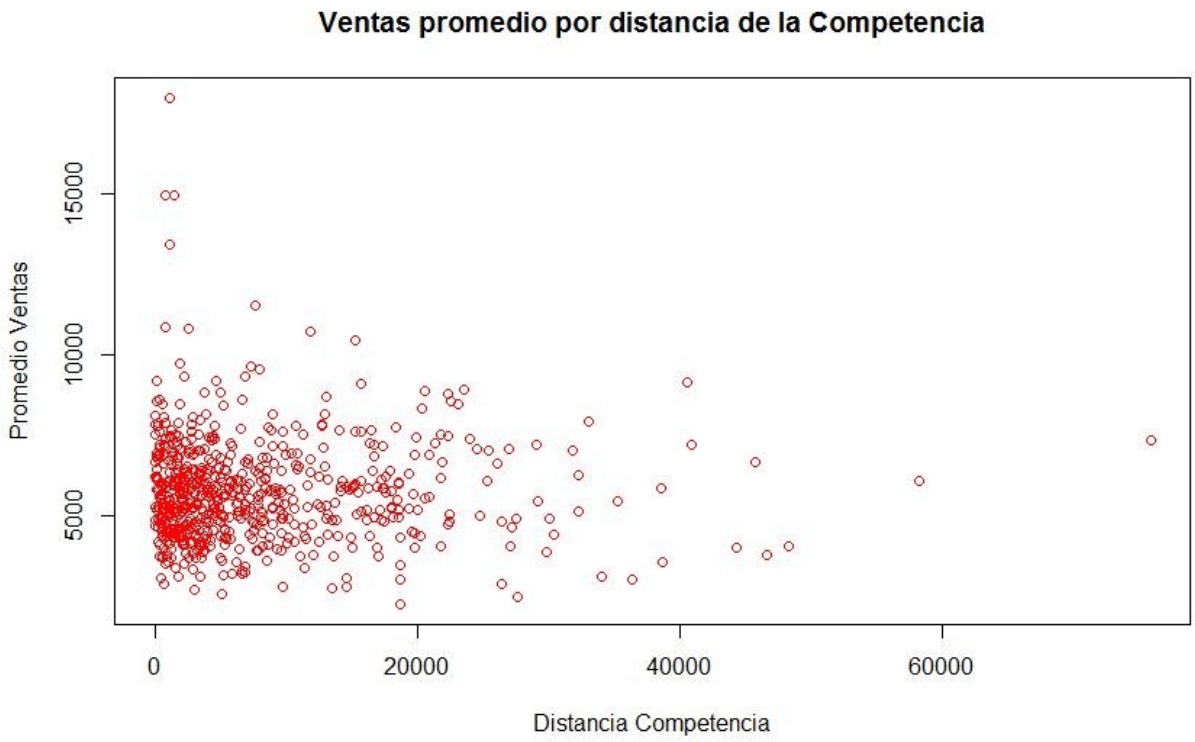


Se puede ver que, las farmacias de tipo B tienen un valor promedio de ventas más alto al de los demás tipos de negocio. Es una posibilidad, convertir los tipos de negocio a, c y d en una sola categoría por su similitud. Lo mismo podría hacerse, con la variable que representa el **estado de las vacaciones**.

Para corroborar que las farmacias de tipo B registran valores promedio de ventas, más altos y conocer en detalle la distribución, utilizaremos el gráfico de caja.



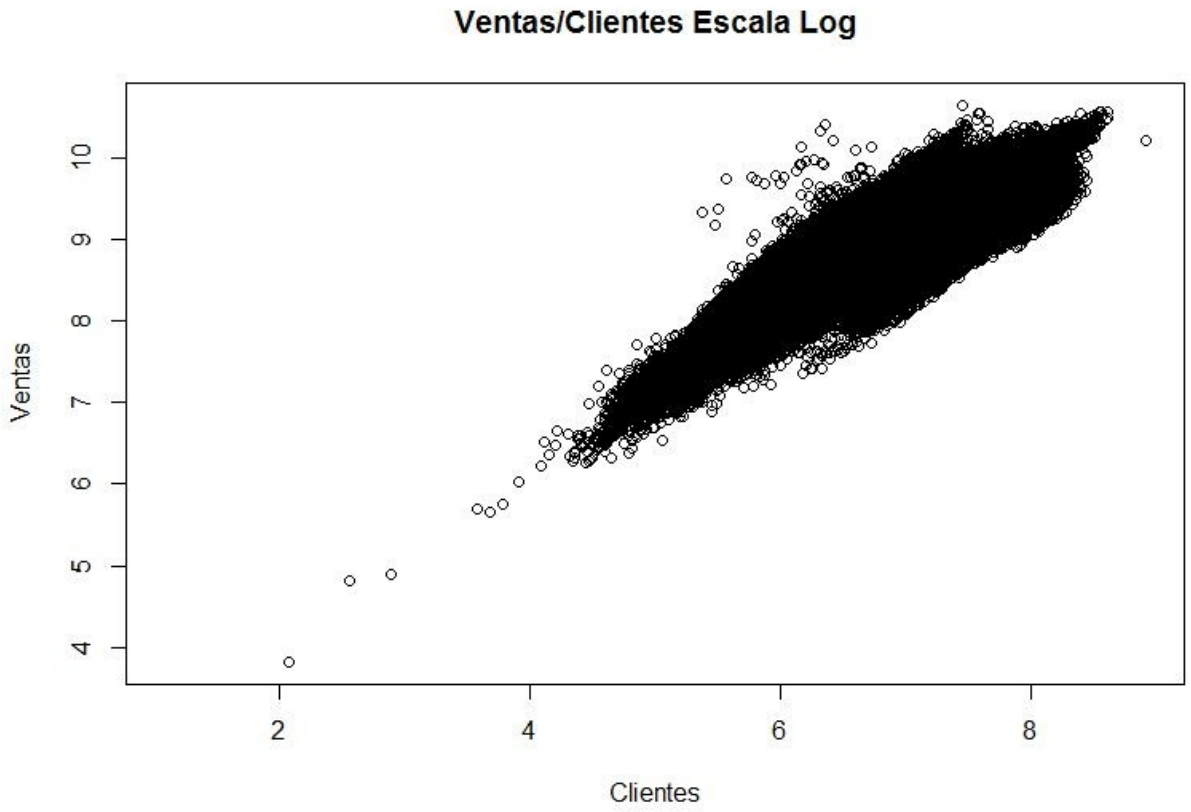
Como se esperaba, el gráfico de caja muestra que las ventas de las farmacias de tipo B son mayores al de los demás tipos de farmacias. Esto se puede afirmar, debido a que el rango intercuartil está por encima del de todos los demás.



Es sabido que mientras más lejos se encuentre la competencia de una farmacia, mejor deberían ser las ventas. En este caso, se registraron las ventas promedio más altas con farmacias que tienen competencia muy cercana. Esto podría deberse, a que, las farmacias que tienen competencia más cercana, se encuentran en ciudades de mayor población (mayor volumen de consumo).

También se puede ver que, mientras mayor son las distancias, las ventas promedio bajas son más escasas.

Uno de los atributos que a priori se puede considerar muy correlacionado con las ventas es el número de clientes.



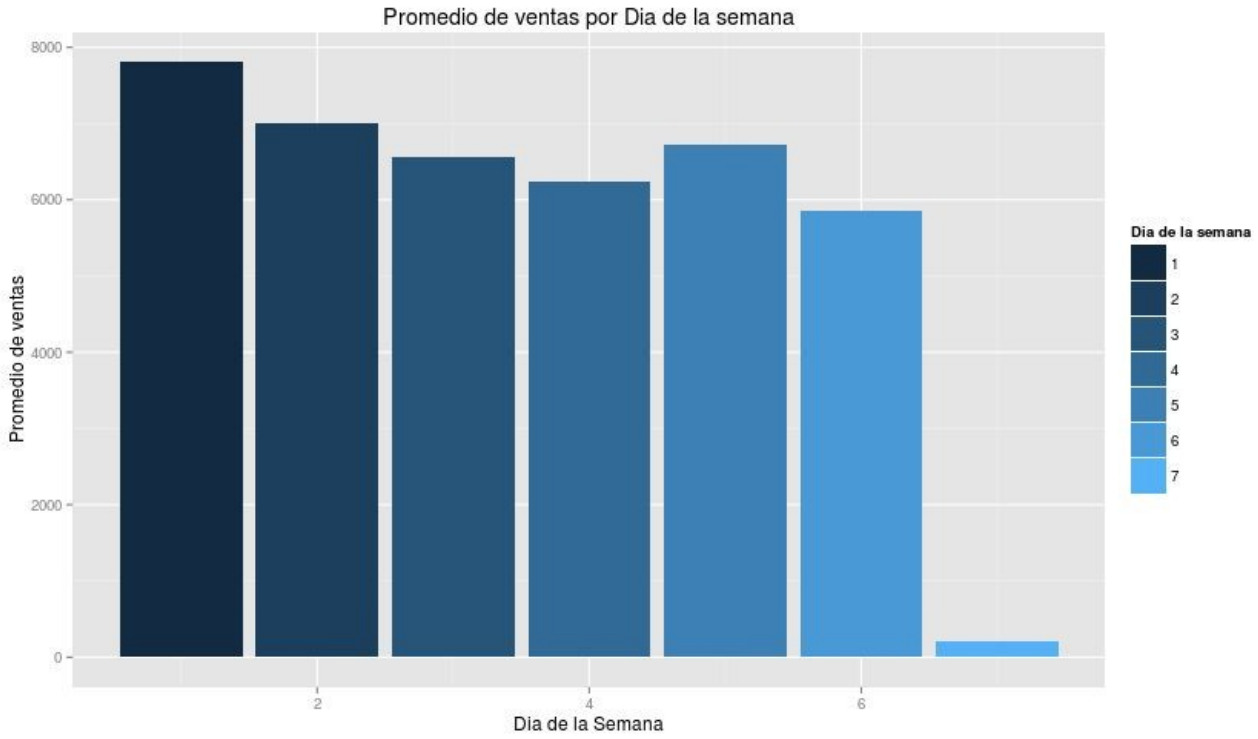
Se normalizaron ambas variables utilizando escala logarítmica. Se puede ver una correlación muy fuerte entre ambas variables. Por lo tanto, se puede inferir que la cantidad de clientes cumplira un rol importante en el modelo que predecirá las ventas.

```
 cor(n.train$Customers, n.train$Sales)
[1] 0.8947108
```

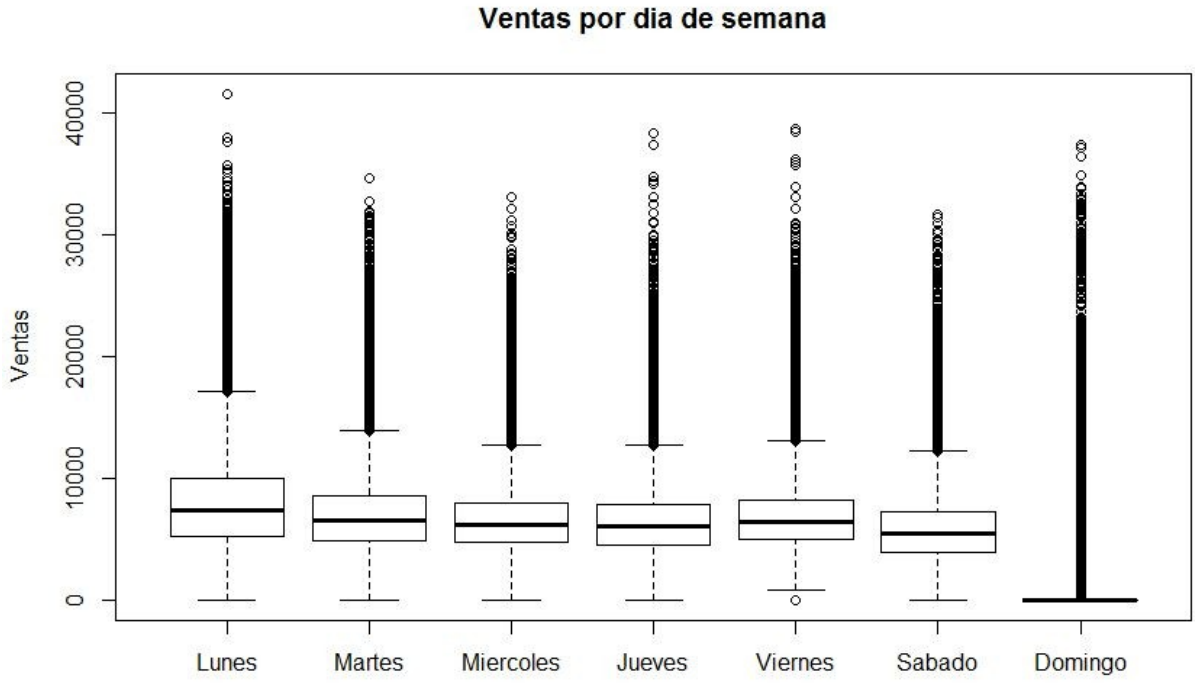
El cálculo de correlación en R confirma lo que el gráfico mostraba.

Análisis de ventas en relación con el tiempo

●Dia de la Semana



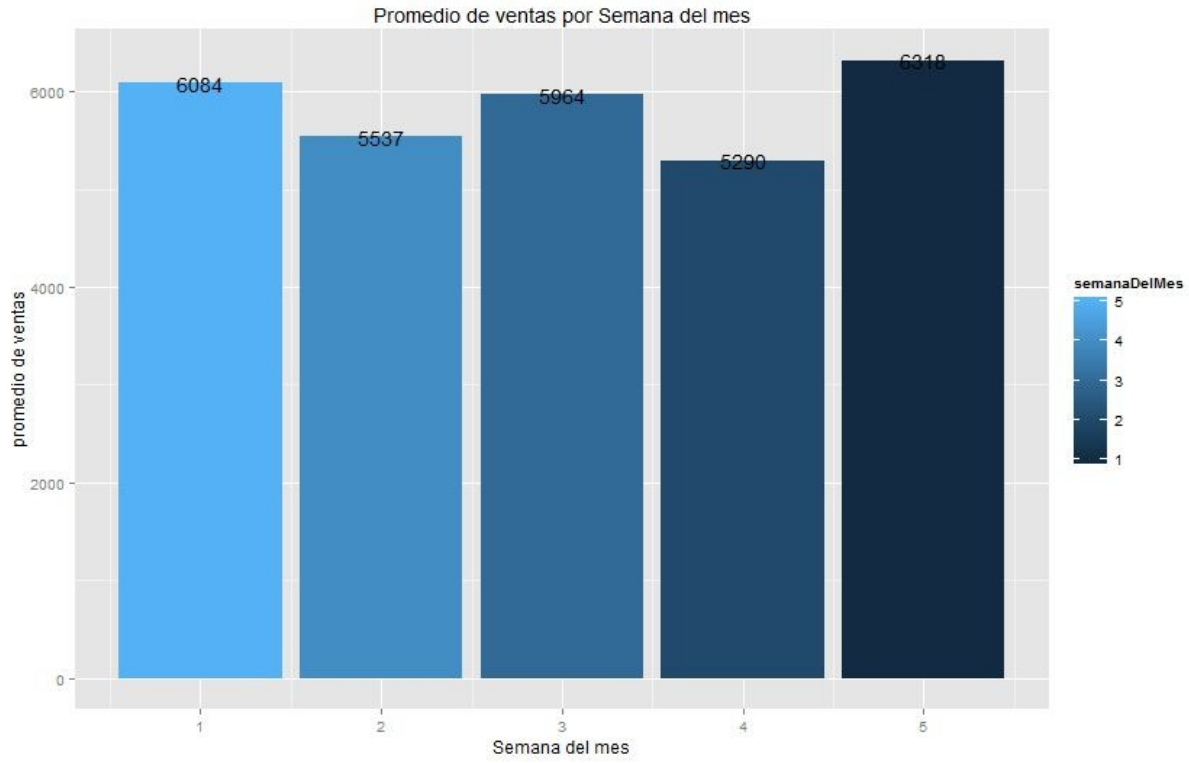
El gráfico anterior, representa las ventas promedio por día de la semana. El dia domingo registra ventas promedio muy bajas, en comparación con los demás días de la semana.



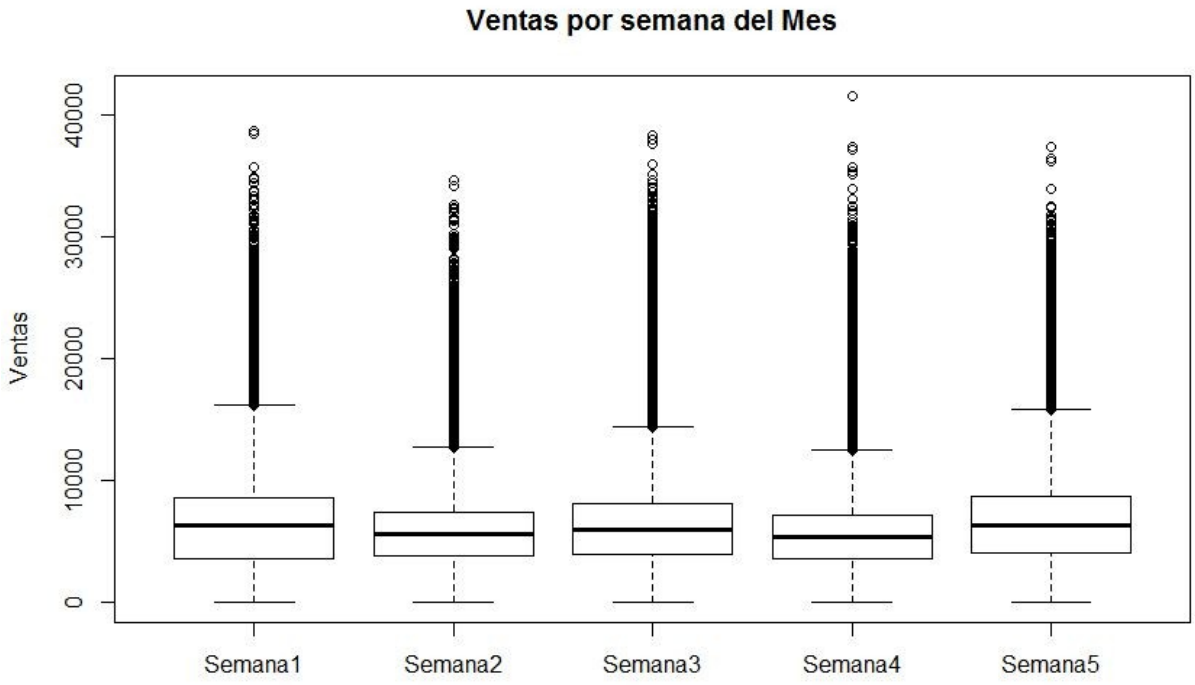
El gráfico de caja anterior, confirma que las ventas son más bajas los días domingos. Aunque este gráfico también explicita que hay excepciones en las que los domingos se realizaron ventas con valores altos.

●Semana del mes

En el siguiente gráfico, se representa las ventas en relación con la semana del mes. Se realiza este análisis en búsqueda de alguna estacionalidad con relación al cobro de sueldos.



No se encontró ningún patrón, como para incluir en el modelo la variable generada ‘Semana del mes’. Al igual que como se hizo antes, se utilizara el diagrama de caja para cotejar esta idea.



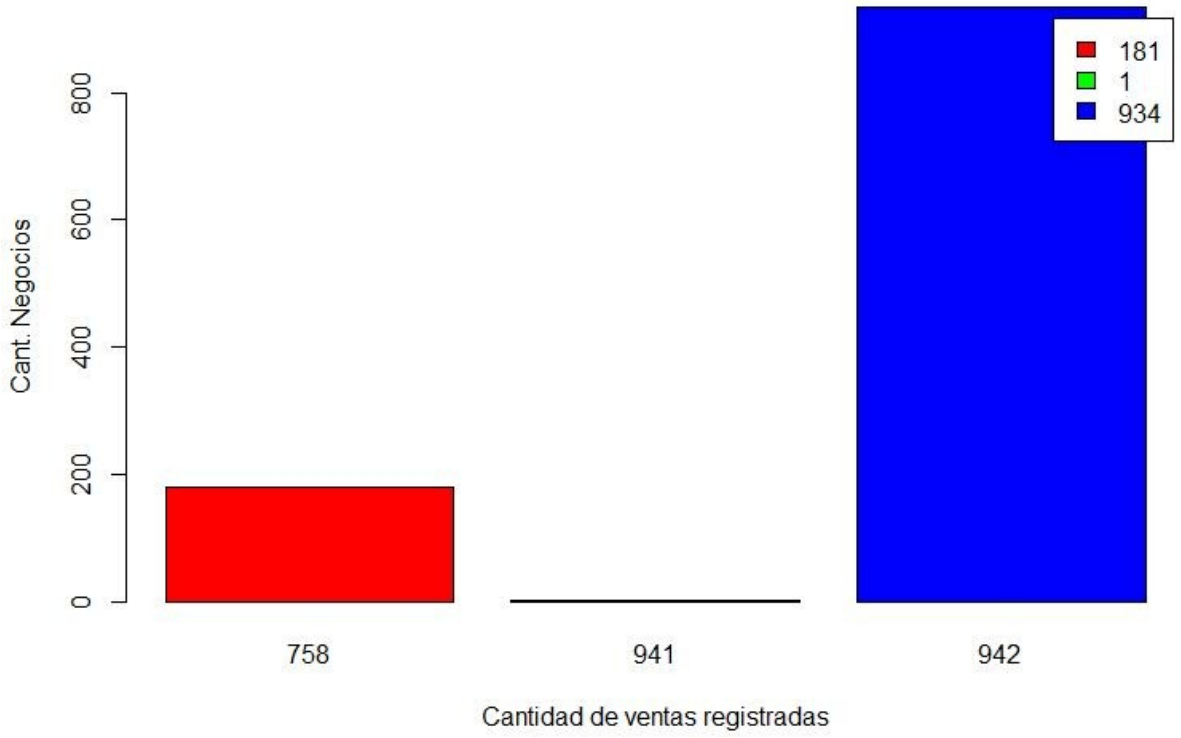
Como se esperaba, las ventas por semana del mes se comportan de forma muy similar.

Mediante la siguiente función, realizó un conteo de la cantidad de ventas que tiene registrado cada negocio. Esto es realizado con el objetivo de determinar si todos los negocios tienen la misma cantidad de ventas registradas, como es esperable.

```
cantidad_ventas < count(new.train, vars = "Store")
summary(cantidad ventas$freq)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  758.0   942.0   942.0   912.3   942.0   942.0
```

El dataset cuenta con una mediana de 942 registros de ventas diarias por cada negocio. Algunas farmacias tienen una cantidad menor de ventas registradas. La causa de esto es que algunos negocios estuvieron cerrados durante lapsos de tiempo.

El siguiente gráfico muestra el recuento de negocios que tienen la misma cantidad de ventas registradas



c. Preprocesamiento

En esta etapa se explicaran los procesos que se realizaron a las variables para mejorar la calidad de los mismos a fin de que permitan generar un mejor modelo de predicción.
Entre los procesos que se realizaron se encuentran:

Limpieza de datos

- Integración de los datos (unión entre los Dataset's).
- Agregación de los datos
- Generación de nuevas variables derivadas de otras.
- Tratamiento de datos faltantes.
-

Las variables cualitativas no pueden ser usadas para directamente para ajustar el modelo.
Dos de las acciones posibles son:
Tomar un atributo completo y convertirlo a numérico. Con el dataset de las farmacias sería equivalente a este ejemplo:

- **StoreType**
  - Original { a,b,c ,d }
  - Transformado { 1,2,3 ,4 }

Seleccionar algunos valores del atributo y binarizarlos (convertirlo en variable dummy). Esta opción es correcta, cuando se detecta que existen valores del atributo que explican mejor la variabilidad de la variable objetivo que otros valores. Es el caso de esta variable y se puede ver en el gráfico que refleja los promedios de farmacia por tipo de negocio.

Este tarea, antes comentada, se realizó mediante los siguientes comandos en R:

```
levels(new.train$StoreType)[match(c('a','c','d'),levels(new.train$StoreType))] < "o"
```

De esta forma, conseguimos que la variable tome solo dos valores y al aplicar el siguiente comando la variable queda en el formato que se requiere:

```
new.train$StoreType < as.numeric(new.train$StoreType)
> unique(new.train$StoreType)
[1] 0 1
```

- **StateHoliday**
  - Original { a,b,c, 0 }
  - Transformado: Elijo el valor a para ajustar mi modelo. 'A' tomará los valores {0, 1}.

Con esta variable sucede lo mismo que con la variable 'StoreType', debido a esto se usaron los mismos comandos:

```
levels(new.train$StateHoliday)[match(c('a','b','c'),levels(new.train$StateHoliday))] < "f"
new.train$StateHoliday < as.numeric(new.train$StateHoliday)
```

Tratamiento de Datos faltantes

El dataset de entrenamiento no posee valores nulos. En cambio, el dataset “Store” si posee valores nulos en las variables:

- CompetitionDistance
- CompetitionOpenSinceMonth
- CompetitionOpenSinceYear
- Promo2SinceWeak
- Promo2SinceYear
-

Variables referidas a la competencia

Respecto a los atributos que se refieren a la competencia de las farmacia, se quiere hacer alguna inferencia, para decidir cuál es la mejor imputación que se puede hacer para cada farmacia.

Existen 2 tipos de tuplas:

- No registran distancia de la competencia

Por ejemplo:

Store	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear
291	NA	NA	NA

Tres farmacias tienen valor nulo en la distancia con la competencia (id = 291, 419, 593). Se podría pensar entonces que si existiese una competencia, lo suficientemente lejos, deja de ser considerada como tal. Si esto se cumpliese, el mes de apertura y el año de apertura tomaron valores nulos, como sucede en el dataset.

La distancia de la competencia es una valor difícil de reemplazar,debido a que, si se lo coloca en cero o con otro valor bajo, el modelo tenderá a reducir el valor de las ventas. Por el contrario, si elije una distancia grande, el modelo tenderá a predecir un aumento en las ventas (en base al análisis). Imputar el valor se convierte en una situación de compromiso.

Eliminar las tuplas que tienen datos faltantes, no es una posibilidad en este dataset, debido a que se perdería información importante (Ej: tipo de negocio, variedad del negocio, etc). También, teniendo en cuenta que el número de instancias que tienen datos faltantes es muy elevado.

```
istore < store[!complete.cases(store),]
length(istore[,1])
750
```

Es decir poco más del 67% de los registros que contienen información sobre los negocios se encuentran incompletos. Un número para nada despreciable.

En base a esto, tomó la decisión de reemplazar la distancia de la competencia por el valor máximo que toma la variable.

- Registran distancia de la competencia pero no tiempo desde su apertura

Store	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear
152	1780	NA	NA

Lo más lógico que se podría pensar en estos casos, es que la competencia tiene fecha de apertura previa a la farmacia.
Esta variable, será imputada posteriormente, una vez que se combinen las variables que involucran la fecha de apertura de la competencia por el valor de la media de la misma.

Variables referidas a la promoción 2

~~Las variables que hacen referencia a la~~ promoción 2 son:

- Promo2
- Promo2SinceWeek
- Promo2SinceYear

Cuando la promoción 2 toma valor cero (la promo no está vigente) las variables Promo2SinceWeek y Promo2SinceYear toman valor nulo. Esto es totalmente lógico, pero como sucede con las demás variables, no es posible omitir las tuplas que tienen valores nulos porque se perdería demasiada información valiosa. Por lo que, se debe buscar un método de imputación.
En este caso, se considera que el mejor valor de imputación será cero.
Ejemplo:
Si tuviéramos la siguiente fórmula de regresión:

$$Y = \beta + \beta_1 X_1$$

Si $X_1 = 0$ entonces

$$Y = \beta$$

Se puede ver que, se logra el objetivo deseado, cuando la promoción tome valores iguales a cero la predicción no será afectada por la variable, más que por su efecto en la constante β (producto de la creación del modelo). La explicación puede resultar trivial, pero demuestra la efectividad de la imputación.

Unión de datasets

Para poder utilizar los datasets provistos, será necesario hacer una unión de ambos mediante la variable ‘Store’. De esta, se obtiene un dataset de entrenamiento con variables que, como se mostró en la fase de análisis, resultan muy útiles para la predicción de las ventas.

Store	DayOfWeek	Year	Month	Day	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
Store	StoreType	Assortment	Competitio11nDistance	tiempoCompetencia	Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval		

d. Transformación

En esta sección se desarrollarán las transformaciones realizadas a la variables del dataset para ajustar el modelo de predicción.

Variable Date

~~Esta variable~~ tenía el formato “yyyymmdd”.

De la misma se obtuvieron las variables:

- Dia del mes
- Mes
- Año

La misma variable fue removida en su formato original, porque no es posible utilizarla para ajustar el modelo, solo es usada para generar las demás variables.

Variable CompetitionOpenSinceYear y CompetitionOpenSinceMonth

~~Ambas variables, corresponden a tiempo desde la apertura de la compet~~encia. Por lo tanto, tiene sentido combinar ambas variables en una sola.

Para hacerlo se utilizará la siguiente función:

tiempo_compentencia = (año AñoMinimo) * 12) + mes

Una vez aplicada, esta transformación obtendrá una única variable que representa perfectamente las otras dos. Como se comentó antes, los valores faltantes serán reemplazados por la media de este valor.

e. Algoritmos

Se eligió como método de aprendizaje la regresión lineal múltiple. Se utiliza R para lograr un modelo que con el que mejor se pueda predecir las ventas.

f.1. COMPROBACIÓN DE SUPUESTOS

Para poder obtener un modelo robusto, se deben cumplir con los supuestos de regresión. Cuando no se cumple con los supuestos de regresión, se puede incurrir en modelos que generan grandes errores de predicción.

Supuestos de regresión:

- Linealidad
- Independencia
- Homocedasticidad
- Normalidad
- Nocolinealidad
-

Todos los supuestos serán evaluados con el modelo completo:
Para ello se usa la siguiente fórmula:

Sales ~.

Esto implica que la variable dependiente Y es Ventas y las predictoras son todas las demás (Clientes, Promocion, Dia de la Seman, etc.)

SUPUESTO DE LINEALIDAD

El modelo relaciona la variable Y (variable a predecir, en este caso las ventas) y los predictores (Todas las demás variables utilizadas para hacer fórmula de regresión)

, .., . que son asumidos como lineales en los parámetros de regresión

$x_1, x_2 \ldots x_p$) .

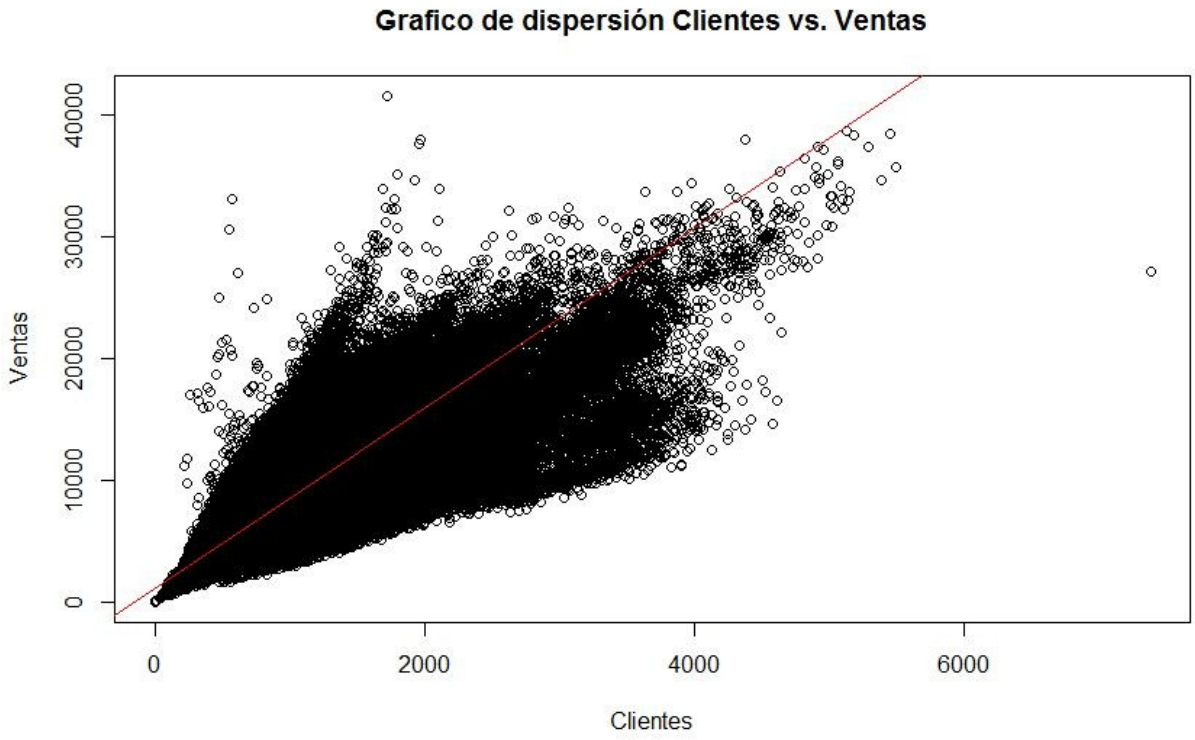
($\beta \quad \beta_1 \quad \beta_p$

$Y = \beta + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \varepsilon$

El supuesto de linealidad de los atributos se comprueba fácilmente, utilizando un gráfico de dispersión de Y contra cada unos de los X.

El gráfico es utilizable con dos variables cuantitativas. En este dataset, se cuenta con mayoría de variables cualitativas, por lo cual hacer el gráfico de dispersión resulta trivial.

La único gráfico que tendría sentido hacer seria Cantidad de clientes contra Ventas.

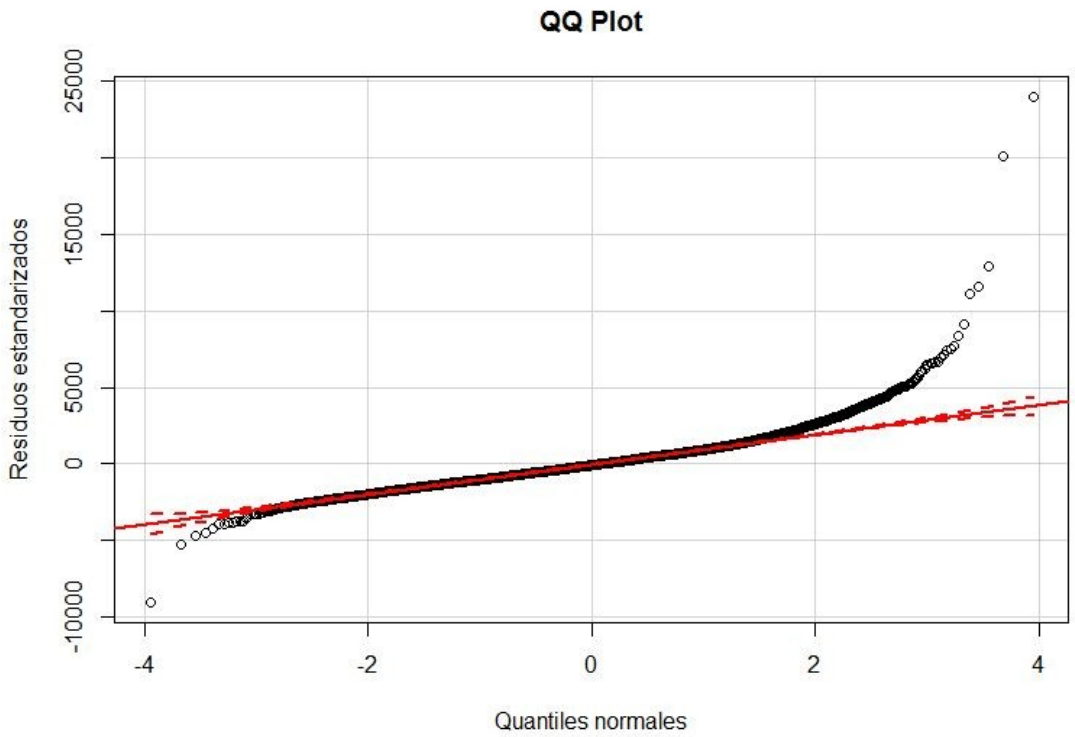


Se considera que está cumpliendo con este supuesto, ya que en el gráfico no se puede ver otra función que se ajuste mejor a la figura que producen las dos variables.

SUPUESTO DE NORMALIDAD DE LOS RESIDUOS

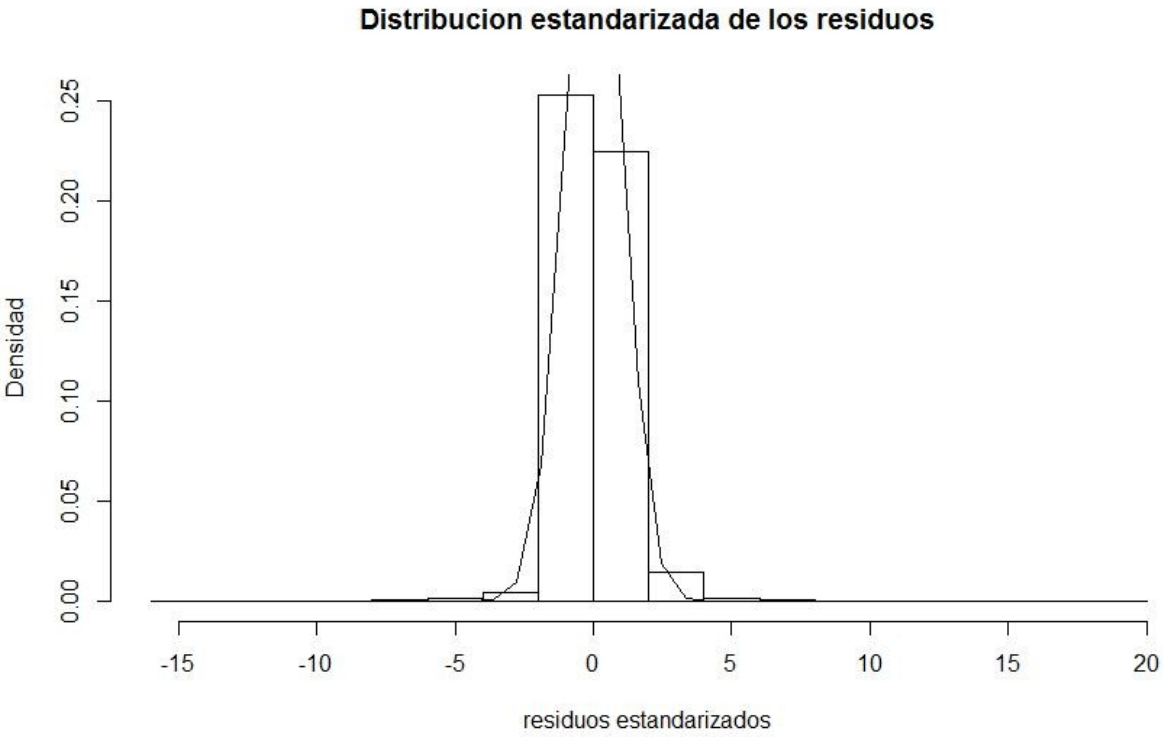
$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i,\ i = 1, 2, \ldots, n.$$

Este supuesto asume que los errores i= 1,2,..n tienen una distribución normal.
La forma más fácil de comprobar este supuesto es graficando la distribución de la variable para verificar que se está cumpliendo.
A continuación, utilizaremos un gráfico QQ para comprobar este supuesto.
El gráfico QQ es un gráfico que sirve para la comparación de la distribución de una variable con una distribución teórica (En este caso será la distribución normal).



Los errores se encuentran representados por los círculos negros y la distribución teórica por la línea roja. Se puede ver que los residuos se aproximan considerablemente al comportamiento de la distribución normal teórica..

Para ganar validez en la comprobación de este supuesto, agregaremos otro gráfico que muestra la distribución de la variable pero mediante su histograma.



Ahora se utilizan algoritmos conocidos para probar que los residuos siguen una distribución normal.

TEST ANDERSONDARLING

Funciona para probar normalidad en muestras de cualquier tamaño.

La prueba de hipótesis es la siguiente:

H_0: Los datos siguen una distribución normal.

H_1: Los datos no siguen una distribución normal.

```
ad.test(fit$residuals)
```

```
        AndersonDarling normality test

data:  fit$residuals
A = 8417.8, pvalue < 2.2e16
```

El valor de P debe estar entre 0,05 y 0,01 para no rechazar la hipótesis nula, por lo tanto,no hay evidencias suficientes para decir que los residuos siguen un distribución normal. Otra afirmación a esto es que el valor de A (valor resultante de la prueba) es muy grande para ser corresponderse con los valores que toman las distribuciones normales.

TEST SHAPIROWILK

Este test es muy parecido al de AndersonDarling solo que sirve para muestras pequeñas (3 5000 observaciones). En este caso tenemos una muestra mucho más grande por lo que se repetirá varias veces las prueba con el máximo permitido para eliminar el sesgo muestral.

```
p.valor < array(data = NA, dim = 10000)
df.boostrap< data.frame(p.valor)
df.boostrap['estadistico'] < 0
for(i in 1:10000) {
        set.seed(100)
        f.res < sample(fit$residuals, 5000, replace = FALSE)
        stest< shapiro.test(f.res)
        df.boostrap$estadistico[i] < stest$statistic
        df.boostrap$p.valor[i] < stest$p.value
}
  mp < mean(df.boostrap$estadistico)
  me < mean(df.boostrap$p.valor)


> mp
```

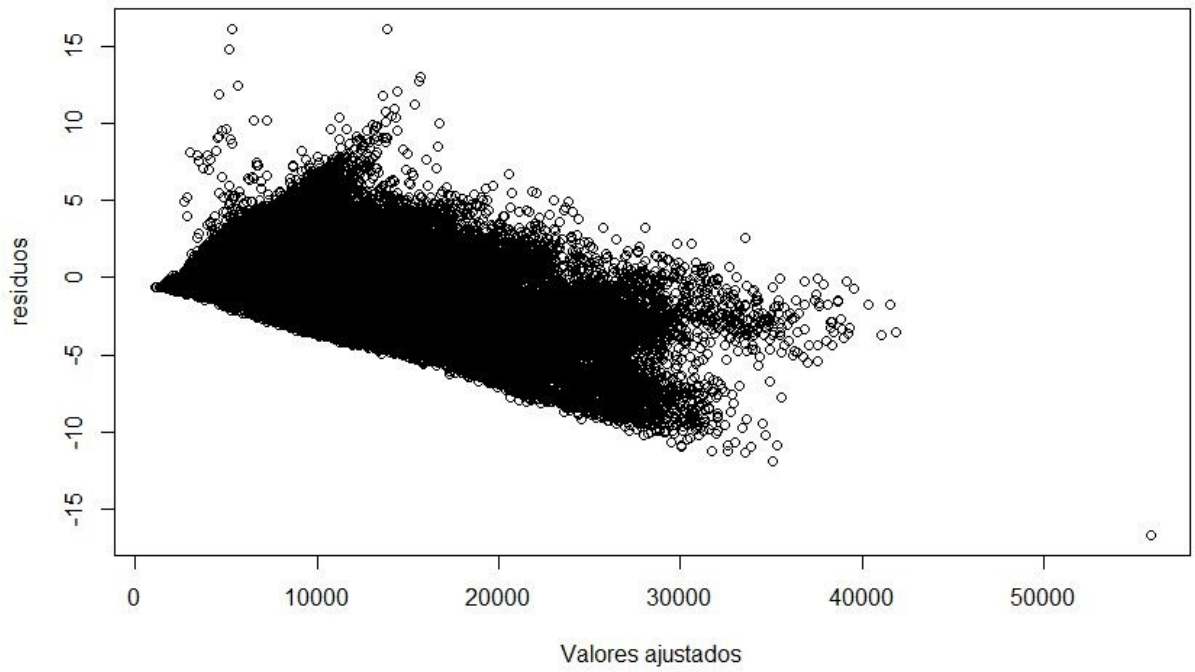
[1] 0.9866453
> me
[1] 2.340313e21

El resultado sigue siendo el mismo al del test anterior, el valor de p es menor a 0,01 por lo tanto no hay evidencia suficiente para decir que los residuos siguen una distribución normal.

En base a las pruebas realizadas se concluye que no hay evidencia suficiente para decir que los residuos siguen una distribución normal.

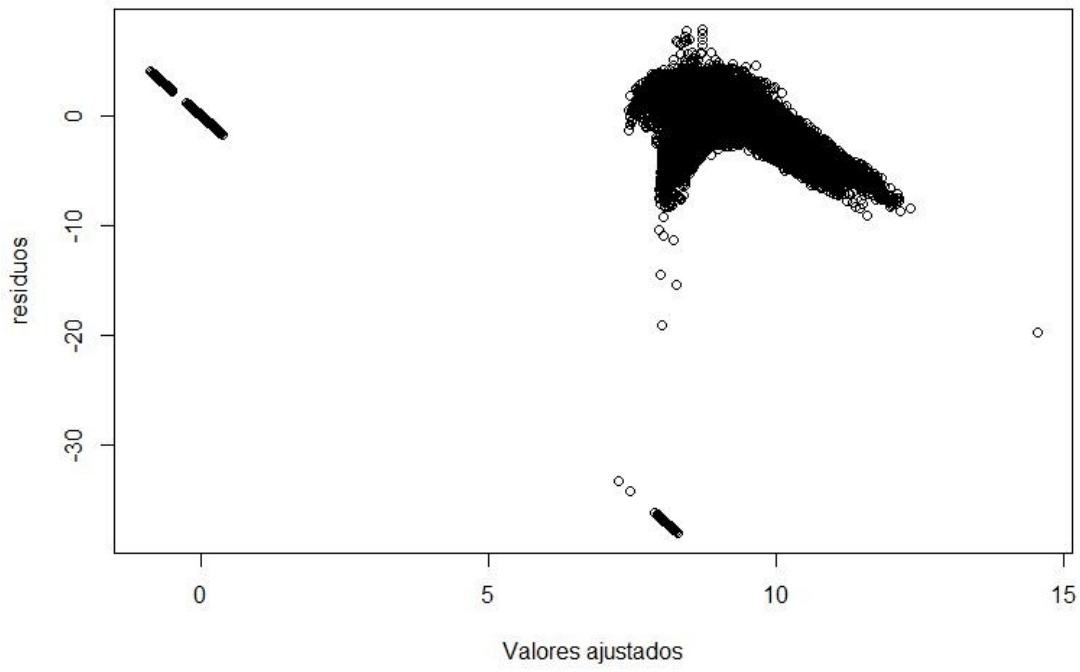
SUPUESTO DE HOMOCEDASTICIDAD

Es llamado también supuesto de varianza constante. Este supuesto asume que todos los errores ($\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$) tienen la misma varianza.



Existen algunos patrones que son demasiado claros. Por lo tanto, considerar que **no** se está

cumpliendo con este supuesto.



El gráfico anterior muestra los residuos después de aplicadas todas las transformaciones a las variables. Ahora se ve muy claro un comportamiento heterocedástico.

Ahora realizamos otra prueba para demostrar que los errores no se comportan de forma constante.
Utilizando regresión lineal intentaremos predecir los errores en base a los valores predichos. La prueba se hizo mediante los siguientes comandos:

```
 fit1 < lm(abs(residuals(fit)) ~ fitted(fit))
summary(fit1)

Call:
lm(formula = abs(residuals(fit)) ~ fitted(fit))

Residuals:
    Min      1Q  Median      3Q     Max
0.1910 0.0943 0.0232  0.0519  8.1147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

(Intercept) 5.390e02 3.555e04   151.6   <2e16 ***
fitted(fit) 1.364e02 4.452e05   306.4   <2e16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1484 on 1017207 degrees of freedom
Multiple Rsquared: **0.08452** **Adjusted Rsquared: 0.08452**
Fstatistic: **9.391e+04 on 1 and 1017207 DF, pvalue: < 2.2e16**

Se puede ver que el valor de R cuadrado ajustado es muy pequeño (las variables predictoras no explican gran medida de la variabilidad) , la prueba estadística devuelve un P muy pequeño y un F muy grande. Se rechaza la hipótesis nula de la variable predictora (valor de ventas predicho) sirve para predecir el valor de los residuos. Por lo tanto, la predicción del error **No** es buena debido a que los residuos no se comportan de forma constante.

SUPUESTO DE AUTOCORRELACIÓN DE LOS RESIDUOS
Los errores son independientes entre ellos (la covarianza entre las variables es cero). Cuando no se cumple con este supuesto se tiene un problema de autocorrelación.
Para comprobar que no existe autocorrelación se usará el test DurbinWatson.

> dwtest(fit, alt="two.sided")

DurbinWatson test

data: fit
DW = 0.7757, pvalue < 2.2e16
alternative hypothesis: true correlation is not 0

El valor de test de DurbinWatson debería ser mayor a 2. Por lo tanto se comprueba que existe correlación entre los errores. Por lo tanto, este supuesto no está siendo cumplido.
Una vez que se normalizan todas las variables y hacen todas las transformaciones comentadas se vuelve a hacer el análisis.
El resultado es:
dwtest(fit, alt="two.sided")

DurbinWatson test

data: fit
DW = 1.9939, pvalue = 0.00452
alternative hypothesis: true autocorrelation is not 0

El valor de DW se encuentra muy cerca de lo esperado (debería estar en 2 y 2.5), lo mismo sucede con el valor de p. De todas formas, se sigue aceptando la hipótesis nula (Los residuos están correlacionados).

SUPUESTO DE NOCOLINEALIDAD

Este supuesto consiste en que las variables predictoras x_1, x_2 , .., x_p son linealmente independientes.

Se realizó un análisis de correlación utilizando los siguientes comandos en R:
Este supuesto refiere a que los residuos o errores tomados a partir del cálculo

```
correlaciones < cor(new.train)
correlaciones[(abs(correlaciones) < 0.5)]< NA
```

A continuación la tablas que muestran las correlaciones entre las variables.
Serán marcadas con verde las celdas que presenten un valor absoluto de correlación mayor a 0.5.

	Store	DayOfWeek	Sales	Customers	Open	Promo	StateHoliday School	Holiday
Store	1	8,48E+00	6,68E+01	0.024324869	4,67E+01	5,79E+01	0.0005422012	0.0006407393
DayOfWeek	8,48E+00	1	0.5432181	0.386444721	5,29E+05	3,93E+05	0.0528887688	0.2053882508
Sales	6,68E+01	0.5432181	1	0.682318255	0.6167683	3,37E+05	0.3755555241	0.0904098626
Customers	2,43E+04	3,86E+05	0.6823183	1	0.9929465	3,16E+05	0.2266075746	0.0715678421
Open	4,67E+01	0.5289625	0.9929465	0.616768288	1,00E+00	2,95E+05	0.3783779582	0.0861706038
Promo	5,79E+01	3,93E+05	3,37E+05	0.316169477	2,95E+05	1,00E+00	0.0123530837	0.0674828123
StateHoliday	5,42E+02	5,29E+04	3,76E+05	0.226607575	3,78E+05	1,24E+04	1	0.1486510166
SchoolHoliday	6,41E+02	2,05E+05	9,04E+04	0.071567842	8,62E+04	6,75E+04	0.1486510166	1
StoreType	1,41E+03	1,88E+01	6,39E+04	0.366725636	5,12E+04	1,08E+02	0.0011637784	0.0018251354
Assortment	4,42E+03	5,17E+01	1,60E+04	0.007044375	2,97E+03	2,94E+02	0.0027551957	0.0025302259
CompetitionDistance	2,26E+04	1,66E+01	2,71E+03	0.101331229	7,08E+03	9,23E+01	0.0005521233	0.0037014975

Promo2	8,49E+03	1,68E+02	2,07E+04	0.150158665	8,31E+03	9,83E+02	0.0089320613	0.0069085639
Promo2SinceWeek	1,11E+04	2,08E+02	1,20E+04	0.098323096	7,44E+03	1,21E+03	0.0087955207	0.0066792505
Promo2SinceYear	8,52E+03	1,68E+02	2,08E+04	0.150138414	8,31E+03	9,83E+02	0.0089332353	0.0069115886
tiempoCompetencia	5,70E+03	2,08E+01	4,01E+03	0.006131529	2,34E+03	1,25E+02	0.0034385309	0.0012332193
Month	1,47E+03	5,36E+03	6,71E+03	0.038178955	6,81E+02	1,17E+04	0.0007937206	0.1032815116
Day	2,26E+01	5,14E+03	2,70E+04	0.004473140	3,34E+04	1,08E+05	0.0662941691	0.0305381849
Year	2,90E+02	1,94E+03	3,98E+03	0.001212180	1,01E+03	2,43E+04	0.0060744586	0.0365353455

	StoreType	Assortment	CompetitionDistance	Promo	Promo2SinceWeek	Promo2SinceYear	tiempoCompetencia
Store	1,4E+03	4,4E+03	2,3E+04	0.0084877142	0.0110531345	0.0085203925	5,7E+03
DayOfWeek	1,9E+01	5,2E+01	1,7E+01	0.0001682793	0.0002077195	0.0001683195	2,1E+01
Sales	6,4E+04	1,6E+04	2,7E+03	0.0207494382	0.0120259323	0.0207512101	4,0E+03
Customers	3,7E+05	7,0E+03	1,0E+05	0.1501580652	0.0983230957	0.1501384137	8,1E+03
Open	5,1E+04	3,0E+03	7,1E+03	0.0083092156	0.0074356301	0.0083078458	2,3E+03
Promo	1,1E+02	2,9E+02	9,2E+01	0.0009827565	0.0012100074	0.0009829896	1,3E+02
StateHoliday	1,2E+03	2,8E+03	5,5E+02	0.0089320613	0.0087955207	0.0089332353	3,4E+03
SchoolHoliday	1,8E+03	2,5E+03	3,7E+03	0.0069085639	0.0066792505	0.0069115886	1,2E+03
StoreType	1	3,7E+04	6,7E+04	0.0539764543	0.0433134562	0.0539392822	1,3E+04
Assortment	3,7E+04	1	1,3E+05	0.0082197788	0.0301820021	0.0082071222	5,0E+04
CompetitionDistance	6,7E+04	1,3E+05	1	0.1358650162	0.1264547739	0.1359145638	1,7E+04
Promo2	5,4E+04	8,2E+03	1,4E+05	1	0.7592403485	0.9999993160	6,4E+04
Promo2SinceWeek	4,3E+04	3,0E+04	1,3E+05	0.7592403485	1	0.7590539792	8,1E+04
Promo2SinceYear	5,4E+04	8,3E+03	1,4E+05	0.9999993160	0.7590539792	1	6,4E+04
tiempoCompetencia	1,3E+04	5,0E+04	1,7E+04	0.0043860976	0.0809920098	0.0043780346	1
Month	2,8E+03	7,6E+03	2,4E+03	0.0253233344	0.0311877894	0.0253293466	3,2E+03
Day	3,9E+01	1,0E+02	3,2E+01	0.0003541953	0.0004354525	0.0003542790	4,6E+01
Year	5,5E+02	1,5E+03	4,7E+02	0.0049820083	0.0061353925	0.0049831909	6,3E+02

	Month	Day	Year
Store	0.0014074067	2,26E+01	0.0002896753
DayOfWeek	0.0053017254	5,14E+03	0.0019372844
Sales	0.0007131721	2,70E+04	0.0039830690
Customers	0.0381789551	4,47E+03	0.0012121801
Open	0.0006808885	3,34E+04	0.0010094099
Promo	0.0117472641	1,08E+05	0.0242996108
StateHoliday	0.0007937206	6,63E+04	0.0000744586
SchoolHoliday	0.1032815116	3,05E+04	0.0305353455
StoreType	0.0027957106	3,87E+01	0.0005498345
Assortment	0.0070012487	1,04E+02	0.0014944815
CompetitionDistance	0.0023918883	3,22E+01	0.0004699910
Promo2	0.0253233344	3,54E+02	0.0049820083
Promo2SinceWeek	0.0311877894	4,35E+02	0.0061353925
Promo2SinceYear	0.0253293466	3,54E+02	0.0049831909
tiempoCompetencia	0.0032143950	4,59E+01	0.0006328553
Month	1	1,24E+04	0.2693823775
Day	0.0124415988	1	0.0024848497
Year	0.2693823775	2,48E+03	1

Existe correlación entre las variables predictoras, pero no se considera que haya alguna que deba ser omitida porque incorpora información redundante al modelo.

f.2. AJUSTE DEL MODELO

Existen varias métricas para selección de un modelo de regresión linear múltiple. Algunas de ellas serán descritas a continuación:

R cuadrado (R^2)

Debe ser interpretado como la proporción del total de la variabilidad de la variable dependiente Y (Ventas) que puede ser explicada por las variables independientes o predictoras (Ej: cantidad de clientes, dia del mes, etc.)

Cuando un modelo se ajusta perfectamente a los datos R^2 es igual a uno.

R cuadrado ajustado (R_a^2)

Es usado para comparar la calidad de modelos que tienen diferente número de variables predictoras. Al contrario de R^2 , R_a^2 no puede ser interpretado como la proporción total de la variación de Y que es explicado por las variables predictoras.

La prueba de hipótesis es realizada junto con el modelo:

Hipótesis Nula: Las variables predictoras **NO** tienen efecto sobre la variable dependiente.
Hipótesis Alternativa: Las variables predictoras tienen efecto sobre la variable dependiente.

Se rechaza la hipótesis nula si el valor p asociado al resultado observado es igual o menor que el nivel de significación establecido, convencionalmente 0,05 ó 0,01. Para este trabajo utilizaré un 0.05 de confianza.

También utilizaremos las métricas derivadas de los residuos para evaluar nuestro modelo. Los residuos de cada instancia se obtienen como la diferencia entre el valor real y valor predecido por el modelo.

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n.$$

Utilizaremos el modelo con todas las variables que tiene nuestro dataset. Este servirá como punto de partida, para comenzar a ajustarlo hasta lograr el modelo que permita predecir las ventas con el menor error posible. En este caso, el mismo que se utilizó para probar los supuestos de regresión.

```
fit < lm(Sales~.,data=new.train)
summary(fit)
```

Residuals

Min	1Q	Median	3Q	Max
9461.3	709.7	67.1	597.1	26396.6

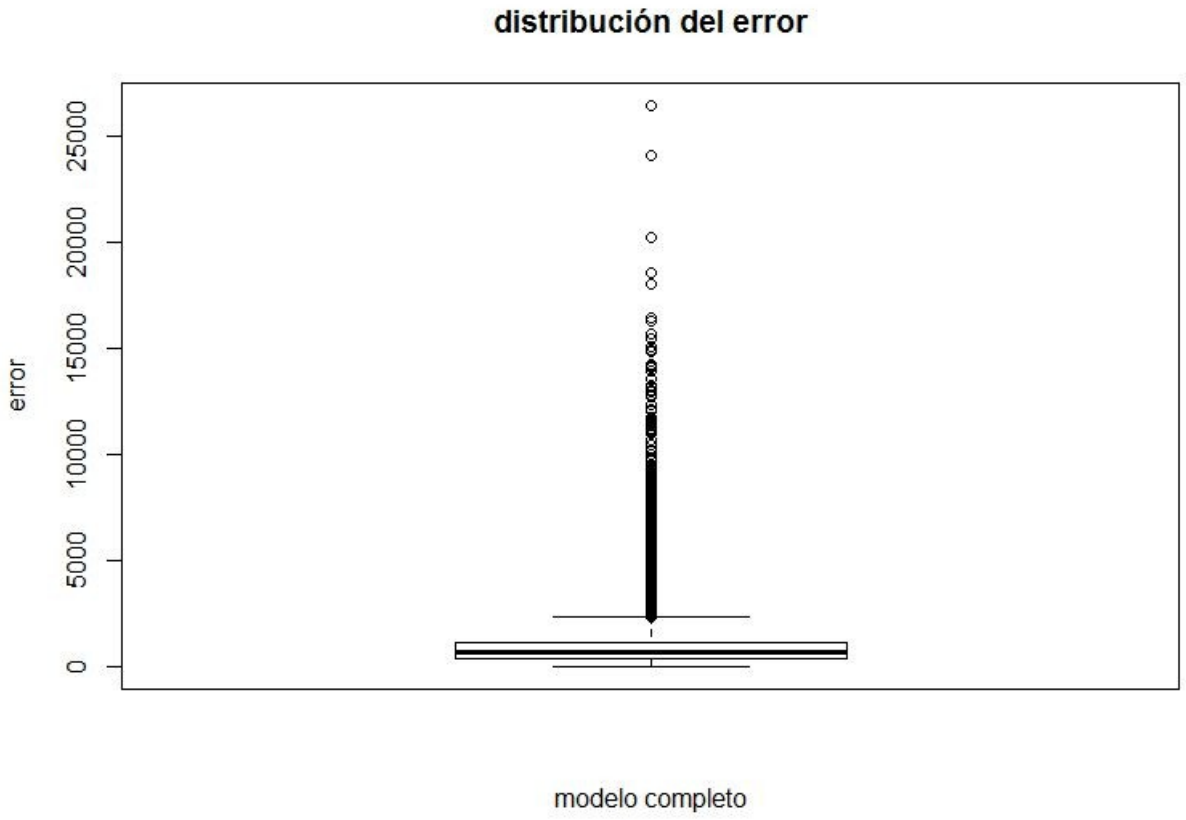
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2,59E+08	5,88E+06	44.022	< 2e16 ***
Store	1,83E+02	6,15E+00	29.744	< 2e16 ***
DayOfWeek	3,70E+04	1,30E+03	28.513	< 2e16 ***
Customers	7,82E+03	8,35E+00	936.907	< 2e16 ***
Open	6,89E+05	8,48E+03	81.284	< 2e16 ***
Promo	1,11E+06	4,65E+03	239.849	< 2e16 ***
StateHoliday	2,79E+05	8,00E+03	34.879	< 2e16 ***
SchoolHoliday	3,67E+04	5,50E+03	6.675	2.48e11 ***
StoreType	2,44E+05	1,57E+03	155.746	< 2e16 ***
Assortment	1,89E+05	2,14E+03	88.286	< 2e16 ***

CompetitionDistance	5,55E+01	3,96E01	140.012	< 2e16 ***
CompetitionOpenSince Month	9,82E+03	6,15E+02	15.965	< 2e16 ***
CompetitionOpenSince Year	7,69E+02	2,70E+02	2.853	0.00434 **
Promo2	NA	NA	NA	NA
Promo2SinceWeek	1,48E+04	1,49E+02	99.122	< 2e16 ***
Promo2SinceYear	2,01E+04	1,21E+03	16.646	< 2e16 ***
Month	2,85E+04	6,33E+02	45.065	< 2e16 ***
Day	3,45E+03	2,28E+02	15.083	< 2e16 ***
Year	1,47E+05	2,64E+03	55.939	< 2e16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1133 on 324308 degrees of freedom
(692883 observations deleted due to missingness)
Multiple Rsquared: 0.895, Adjusted Rsquared: 0.895
Fstatistic: 1.627e+05 on 17 and 324308 DF, pvalue: < 2.2e16

La variable Promo2 devuelve valores Nulos debido a que no se puede estimar el valor de su coeficiente. Por lo tanto, esta variable será retirada del dataset porque no aporta nada al modelo de predicción.



median(abs(fit$residuals))
656.7464

Este valor medio de de error es considerablemente grande, ya que, en todas la distribución podemos tener errores muy grandes.

max(abs(fit3$residuals))
26396.61

Se intentará reducir estos tipos de errores que son bastante graves.

SELECCIÓN DE MODELO

La selección de variables puede ser vista como la selección de modelos. Para elegir nuestro modelo utilizaremos principalmente AIC (Akaike Information Criteria). Este criterio hace un balance entre la necesidad de accuracy (fit) y la simplicidad (número pequeño de variables). AIC para una ecuación de ptérminos (una constante y p1 variables) está dado por:

$$\mathrm{AIC}_p = n \ln(\mathrm{SSE}_p/n) + 2p.$$

Los modelos con menor AIC son preferibles.
Por lo tanto AIC es usado para rankear los diferentes modelos en base a los criterios de ajuste y simpleza.
Grandes diferencias entre los valores de AIC indican una diferencia significativa entre la calidad de los modelos. El modelo con menor valor de AIC debe ser escogido.
Los modelos que sean comparados mediante este método deben tener todas sus instancias completas, las que lo estén serán eliminadas. No se tendrá ningún problema con esto, debido a que ya se realizó un detallado tratamiento de datos faltantes.

sw < stepwise(step.fit, direction="forward/backward", criterion = "AIC")

Direction: forward/backward
Criterion: AIC

Start: AIC=16795768
Sales ~ 1

		Df	Sum of Sq	RSS	AIC
+	Customers	1	1,21E+17	3,01E+16	15156051
+	Open	1	6,94E+16	8,14E+16	16168368
+	DayOfWeek	1	3,22E+16	1,19E+17	16551397
+	Promo	1	3,09E+16	1,20E+17	16562900
+	StateHoliday	1	7,91E+15	1,43E+17	16740963
+	Promo2SinceYear	1	1,25E+15	1,50E+17	16787301
+	Promo2	1	1,25E+15	1,50E+17	16787304
+	SchoolHoliday	1	1,09E+15	1,50E+17	16788372
+	Assortment	1	8,47E+14	1,50E+17	16790041
+	Month	1	3,59E+14	1,50E+17	16793348
+	Promo2SinceWeek	1	2,94E+14	1,50E+17	16793786
+	CompetitionDistance	1	8,93E+13	1,51E+17	16795167
+	Year	1	8,34E+13	1,51E+17	16795207
+	StoreType	1	2,57E+13	1,51E+17	16795597
+	Day	1	2,03E+13	1,51E+17	16795633
+	tiempoCompetencia	1	1,36E+13	1,51E+17	16795678
+	Store	1	3,96E+12	1,51E+17	16795743
<	none>			1,51E+17	16795768

Ahora se mostrará como el algoritmo agrega variables iterativamente a la fórmula en búsqueda del modelo con menor valor de la métrica AIC. Por cada uno de los modelos se devuelve la tabla ANOVA (Analysis of Variance) correspondiente.

		Df	Sum of Sq	RSS	AIC
<none>				1,99E+16	14736561
	tiempoCompetencia	1	7,73E+11	1,99E+16	14736598
	Day	1	1,97E+12	1,99E+16	14736659
	SchoolHoliday	1	3,26E+12	1,99E+16	14736725
	Store	1	1,76E+13	1,99E+16	14737460
	Promo2	1	4,17E+13	2,00E+16	14738689
	Promo2SinceYear	1	4,18E+13	2,00E+16	14738691
	Promo2SinceWeek	1	6,94E+13	2,00E+16	14740098
	StateHoliday	1	9,05E+13	2,00E+16	14741173
	DayOfWeek	1	1,14E+14	2,00E+16	14742363
	Year	1	1,21E+14	2,00E+16	14742723
	Month	1	1,33E+14	2,00E+16	14743324
	CompetitionDistance	1	2,50E+14	2,02E+16	14749270
	Assortment	1	3,08E+14	2,02E+16	14752150
	StoreType	1	8,25E+14	2,07E+16	14777842
	Open	1	1,20E+15	2,11E+16	14796290
	Promo	1	2,99E+15	2,29E+16	14879014
	Customers	1	4,72E+16	6,71E+16	15972978

Step: AIC=14736561

Modelo:

```
Sales ~ Customers + Promo + Open + StoreType + Assortment + CompetitionDistance +
    Promo2SinceWeek + Month + Year + DayOfWeek + StateHoliday +
    Store + SchoolHoliday + Promo2SinceYear + Promo2 + Day +
    tiempoCompetencia
```

Hasta ahora, el mejor modelo encontrado, para predecir las ventas sería el descrito anteriormente, el cual es el modelo completo (incluye todas las variable predictoras).
Una vez aplicadas todas las transformaciones mencionadas, en la fase de preprocesamiento y transformación, se vuelve a correr el algoritmo:

	Variables	Df	Sum of Sq	RSS	AIC
	<none>			48310	3099574
	tiempoCompetencia	1	5	48315	3099481
	SchoolHoliday	1	15	48326	3099252
	Promo2	1	82	48392	3097859
	Promo2SinceYear	1	82	48392	3097857
	Store	1	83	48393	3097837
	Month	1	212	48522	3095125
	Promo2SinceWeek	1	221	48532	3094929
	DayOfWeek	1	236	48546	3094626
	StateHoliday	1	270	48580	3093908
	Year	1	335	48645	3092550
	CompetitionDistance	1	386	48697	3091471
	Assortment	1	1298	49608	3072612
	StoreType	1	6008	54319	2980338
	Promo	1	7529	55840	2952242
	Customers	1	75004	123315	2146357
	Open	1	3779136	3827447	1347972

Step: AIC=3099574
Sales ~ Open + Customers + Promo + StoreType + Assortment + Promo2SinceWeek + CompetitionDistance + Year + Month + StateHoliday + DayOfWeek + Store + SchoolHoliday + Promo2SinceYear + Promo2 + tiempoCompetencia

El menor valor de AIC ahora es negativo pero sigue siendo el mejor modelo el completo, si utilizamos esta métrica para seleccionar el modelo.

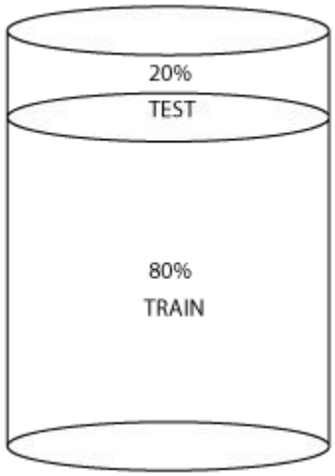
Aparece un problema en el momento de realizar la predicción. La variable ‘Customers’ (la segunda más importante para la predicción según el valor de AIC) que estaba presente en el dataset de entrenamiento no lo está en el dataset de prueba. Por lo tanto no contaremos con la misma para predecir las ventas. Deberíamos realizar la selección del modelo nuevamente sin incluir la variables ‘Customers’.

	Df Sum	of	Sq	R	SS	AIC
	<none>				123315	2146359
	tiempoCompetencia	1	7		123322	2146300
	SchoolHoliday	1	34		123349	2146076
	Promo2	1	40		123355	2146028
	Promo2SinceYear	1	41		123356	2146019
	Year	1	419		123734	2142913
	CompetitionDistance	1	475		123790	2142452
	Promo2SinceWeek	1	644		123959	2141062
	StateHoliday	1	787		124102	2139887
	Month	1	836		124151	2139490
	DayOfWeek	1	1165		124480	2136792
	StoreType	1	2101		125416	2129177
	Assortment	1	2426		125741	2126547
	Promo	1	19740		143055	1995315
	Open	1	5685864		5809179	1772389

Step: AIC=2146359
Sales ~ Open + Promo + StoreType + Assortment + Promo2SinceYear +
DayOfWeek + Promo2SinceWeek + StateHoliday + Month + CompetitionDistance +
Year + Promo2 + SchoolHoliday + tiempoCompetencia

Este es entonces, el mejor modelo que se puede utilizar para predecir las variables utilizando como métrica AIC.

Para probar el modelo el dataset será partido en dos, seleccionando instancias aleatoriamente para cada unos de ellos.



El dataset fue dividido en una 80% para generar el modelo y el otro 20% para probarlo.
La métrica que será usada para evaluar nuestro modelo es RMSPE (Root Mean Square Percentage Error). El valor producto de la prueba con el 20% es:

```
 rmspe(s.test$Sales, pred1[,1] , includeSE = TRUE)
$rmspe
[1] 0.3483751
```

Es decir, un error medio cuadrado promedio de más del de 3%.

PREDICCIÓN

Para hacer la predicción utilizaremos el dataset de testing.
El mismo contiene las variables:

- Id
- Store
- DayOfWeek
- Date
- Open
- Promo
- StateHoliday
- SchoolHoliday
-

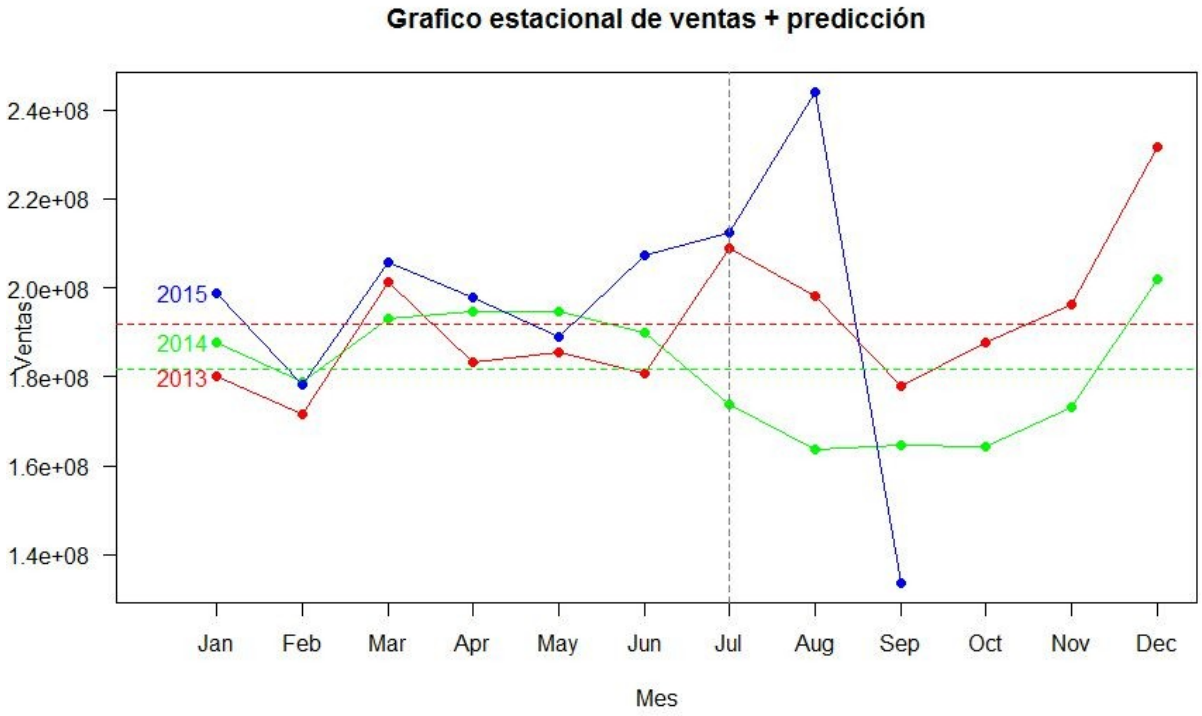
Para hacer la normalización de las variables en el caso de

- Normalización MinMax
- Normalización ZScore
-

Se deben utilizar los mínimos, máximos, media y desvío estándar de las distribución de entrenamiento.
Además al obtener el resultado de las ventas se le debe aplicar la función inversa para obtener el valor real.

$$Ventas = e^{valores\ predichos}$$

Utilizando el mismo gráfico que se empleó en el análisis para mostrar las ventas promedio por mes de cada año, se reflejara el resultado de las predicciones.



El gráfico muestra que durante el mes de agosto aumentaron las ventas. Durante septiembre se puede ver una caída abrupta de las ventas, se debe a que, solo se realizaron predicciones de pocos días de ese mes y por lo tanto el promedio es bajo en relación a los demás meses.

```
d < as.Date(test$Date)
 min(d)
[1] "20150801"
 max(d)
[1] "20150917"
```

La ejecución de los comandos previos muestran que se realizó una predicción desde el 18 hasta el 179 del año 2015. Si se hubiera realizado una predicción del mes completo de septiembre para todos los negocios, se podría compararlo con los demás.

Para probar el resultado, lo haremos con los datos reales, mediante el sitio que propone el desafío de predecir las ventas. Se creó un archivo de tipo ‘csv’ con las estructura indicada y fue subido al mismo. La captura de pantalla a continuación expone el resultado obtenido.

3295	—	WGuan	1.00000	1	Mon, 07 Dec 2015 17:37:23
3296	—	Ivan Iordanov	1.00000	2	Mon, 07 Dec 2015 17:44:19 (-0h)
3297	—	EHans	1.00000	1	Mon, 07 Dec 2015 17:57:10
3298	—	Petr Ermakov	1.00000	2	Mon, 07 Dec 2015 19:17:46 (-0h)
-		**FranciscoToninMonzon**	**1.26783**	-	**Mon, 21 Dec 2015 02:05:46** Post-Deadline

Post-Deadline Entry
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

3299	—	Karl, Adel & Tarek	1.62094	2	Sun, 06 Dec 2015 19:13:31
3300	↑2	vinotharun	1.72850	1	Wed, 18 Nov 2015 03:00:15

Conclusión

En base a los resultados obtenidos, con e
el ranking de la página Kaggle, se concluye
vez que participa en la competencia.

Más que por ser la primera vez en participar, diría que los resultados son aceptables porque trabajaste con una técnica que desconocías y que no es para nada trivial y que se trata de un problema extremadamente complejo que solo con regresiones no se encuentra una solución buena.

Son muchas las acciones que se podrían realizar para mejorar la performance del modelo de predicción. A continuación se listan algunas de ellas.

- Generar modelos independientes para cada farmacia o para algunas de ellas. El gráfico que muestra los promedios por cada farmacia revela que hay algunas de ellas que registran ventas mucho más elevadas a las demás. El error cometido en la predicción es mayor si se quiere predecir las ventas de farmacias con volúmenes de ventas tan diferentes. La precisión aumentará si se agruparan negocios con comportamiento similares (utilizando clustering) y se usará regresión de forma disjunta con los grupos resultantes.

Es probable que mejore segmentando las farmacias utilizando clustering o trabajando con modelos que no sean lineales.

- Crear modelos independientes para diferentes lapsos de tiempo. El gráfico de estacionalidad expuesto en el apartado de exploración de datos, muestra la variabilidad que existe entre meses del año. La predicción sea más precisa, si crearan modelos que se ajusten al comportamiento estacional de las ventas.
- Al utilizar un modelo lineal, este trabajo queda atado a la limitación de que las variables predictoras se relacionan linealmente con la variable objetivo. Si esto no sucede, asi el error de predicción aumenta.Otra acción que se podría tomar, es utilizar regresión no lineal y probar si esta se ajusta mejor a los datos.
- La regresión lineal múltiple es un método de predicción de costo computacional reducido en comparación con otros (Redes neuronales, Algoritmos genéticos). Esta fue la mejor opción por simplicidad y los recursos de hardware con los que se cuentan en el entorno doméstico. Por lo tanto, se puede utilizar cualquiera de los métodos mencionadas para probar si obtienen mejores resultados.

Claro, hay otras alternativas de regresiones (spline, bispline, bicubicas, etc) que se adaptan mejor. También hay redes neuronales, que permiten capturar esa variación estacional y hay series de tiempo (esto es para las ligas mayores! :D)

Mejor aún, se podría utilizar el método de ensamble como boosting. El método de boosting consiste en generar N modelos (con las técnicas antes mencionadas) que se complementen para reducir el error de predicción al mínimo. Emplear Boosting permitiría explotar las mejores cualidades de cada método.

Mas haya de la predicción, este trabajo genera conocimiento general sobre cómo se comportan diversos factores que afectan a las ventas de las farmacias, como el uso de promociones, la repercusión de los días feriados, el comportamiento semanal, mensual y anual de las ventas, etc. En base a este se puede dar un soporte muy robusto a la toma decisiones a la empresa Rossman, si se lo necesitara.

D. Bibliografía

Samprit Chatterjee, Ali S. Hadi. Regression Analysis by Example. ta Edición.
Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques. da Edición.
Comandos y librerias Rbase http://www.statmethods.net/
Implementación de RLM http://www.rbloggers.com/rtutorialseriesmultiplelinearregression/
Implementación de RLM https://cran.rproject.org/web/packages/lmtest/lmtest.pdf
Métricas de evaluación https://cran.rproject.org/web/packages/hydroGOF/hydroGOF.pdf
CrossValidation https://cran.rproject.org/web/packages/cvTools/cvTools.pdf
CrossValidation https://cran.rproject.org/web/packages/DAAG/DAAG.pdf
Pruebas de Normalidad https://cran.rproject.org/web/packages/nortest/nortest.pdf
Pruebas de Normalidad https://cran.rproject.org/web/packages/ggplot2/ggplot2.pdf
Supuesto de Regresión http://apuntesr.blogspot.com.ar/2015/04/supuestosenregresionlineal.html