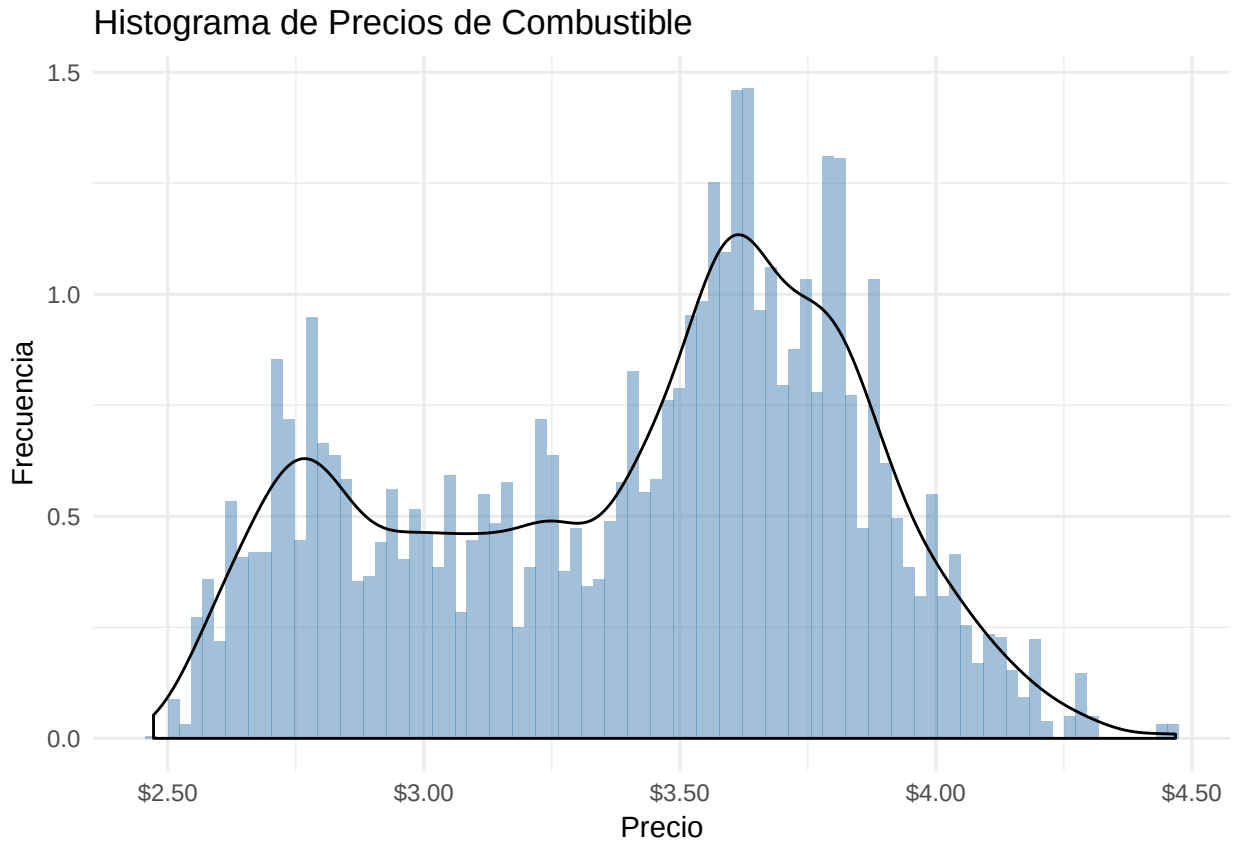
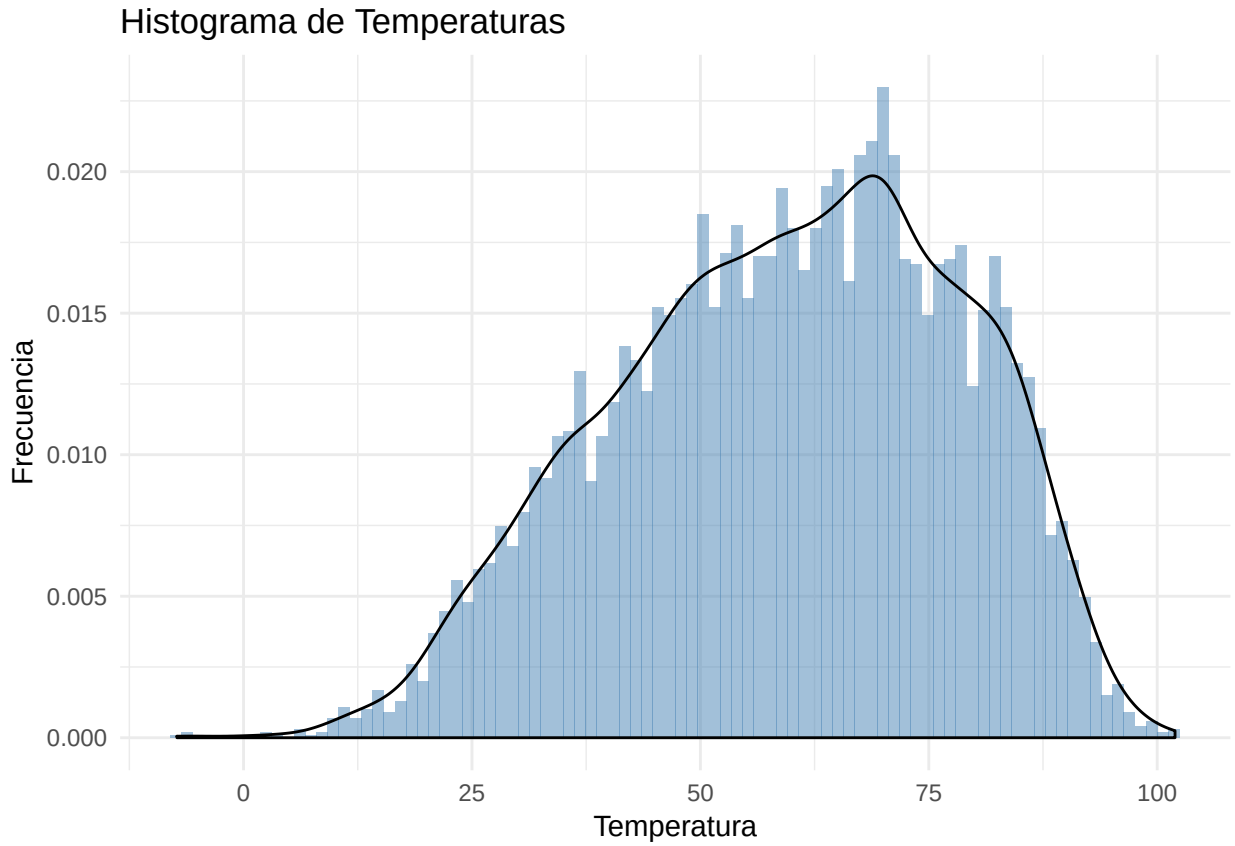
Ciencia de Datos aplicada en la Industria Retail

Como primer paso vamos a analizar cómo se comporta la variable de precio de combustible.

FP_moda	FP_media	FP_mediana	FP_std	T_moda	T_media	T_mediana	T_std
3.638	3.405992	3.513	0.4313366	70.28	59.3562	60.71	18.67861

Ahora graficamos los histogramas correspondientes para visualizar si se aproxima a alguna distribución.





La distribución de la variable temperatura se aproxima a una normal pero la variable precio de combustible no pareciera ser similar a ninguna. Planteamos un test KS donde: H0:los datos proceden de una distribución normal. H1:los datos no proceden de una distribución normal.

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  features$Fuel_Price
## D = 0.10373, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

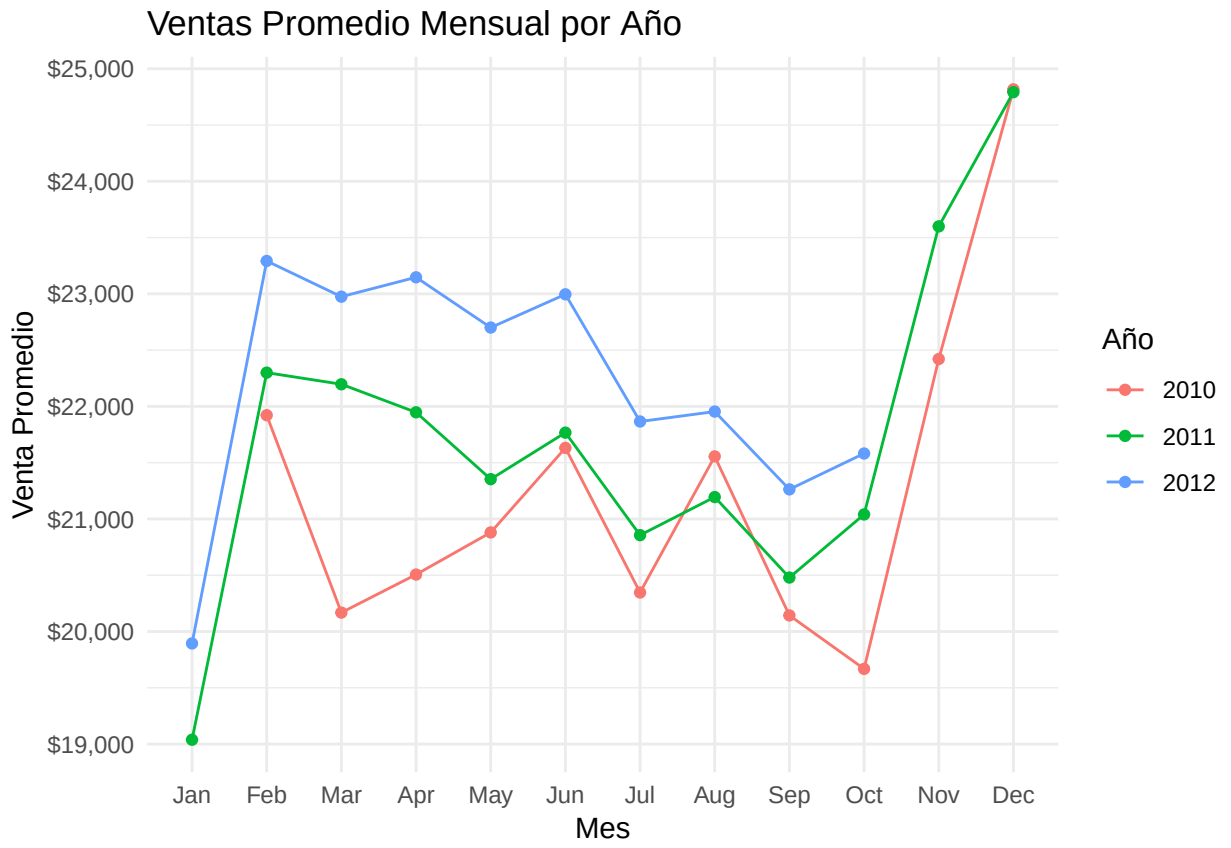
P-value = 2.2e-16 por lo que no es posible afirmar que la variable Fuel_Price proviene de una distribución normal.

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  features$Temperature
## D = 0.046466, p-value = 8.882e-16
## alternative hypothesis: two-sided
```

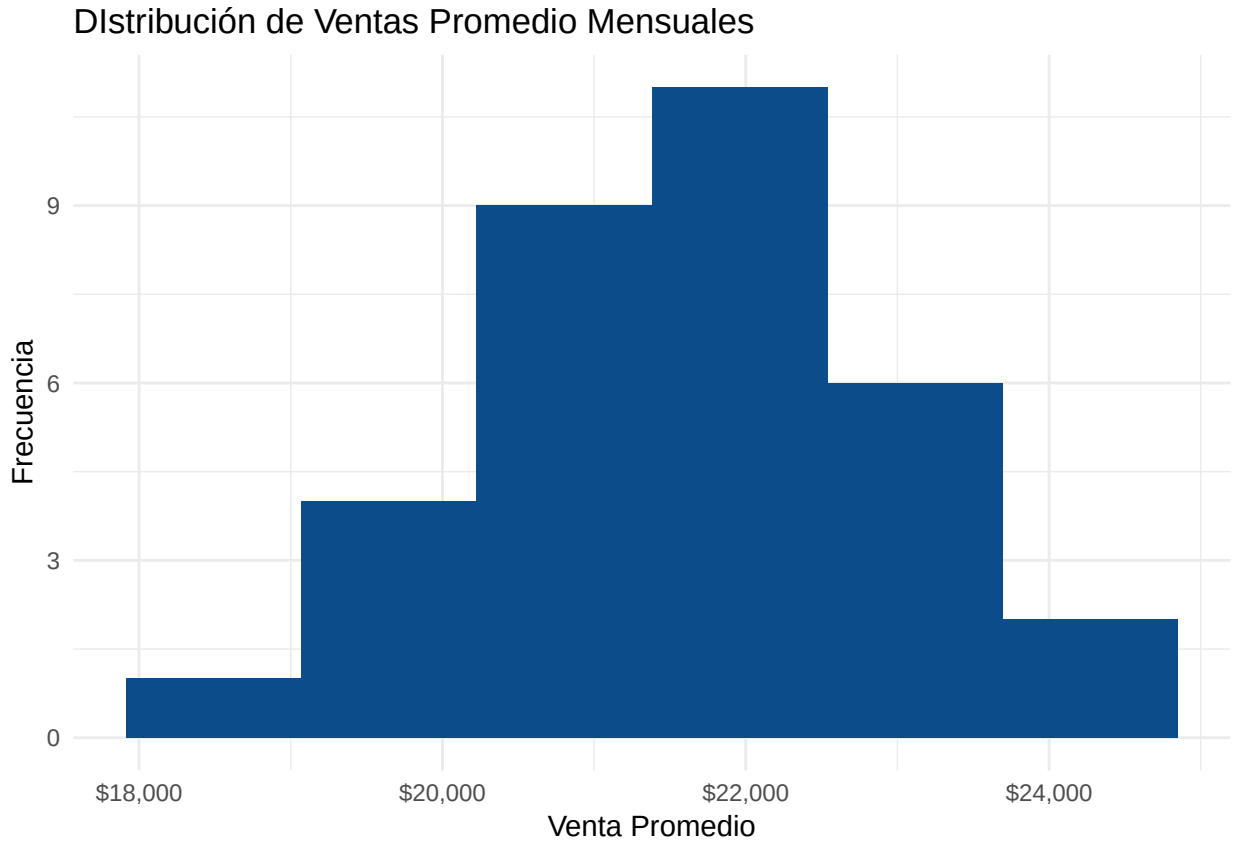
Vemos que p-value = 8.882e-16 por lo que no es posible afirmar que la variable Temperature proviene de una distribución normal.

Seleccionemos una tienda particular y hagamos un análisis de distribución de las ventas.

Para analizar un caso en particular elegimos la tienda nro 1:



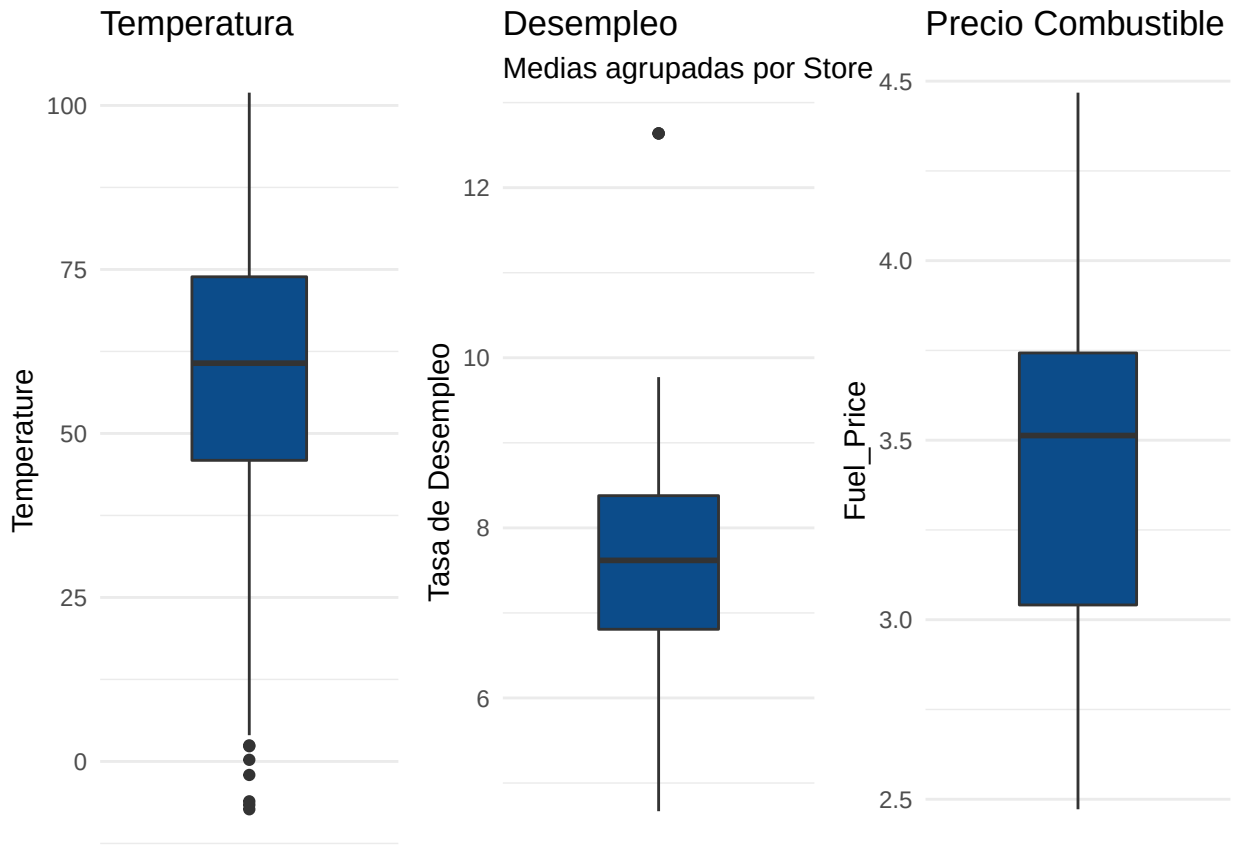
Vemos un indicio de que la mayoría de las ventas promedio por mes están entre los $20 mil y los $24 mil. Vamos a crear un histograma para corroborar la asunción.



Ahora sí vemos que la distribución del promedio de ventas mensuales para el store 1 se aproxima a una normal.

Ahora nos preguntamos si existen outliers para las variables más importantes de la tabla features.

Vamos a analizar las siguientes variables: tempreatura, desempleo y precio de combustible.



Vemos que la variable temperatura presenta outliers en valores cercanod a 0 grados y que la tasa de desempleo tiene outliers para la tasa media de algún/os stores en particular con tasas mayores al 10%. En cambio el precio del combustible no presenta valores atípicos.

Vamos a analizar la variable temperatura:

Store	Fecha	Temperature
7	2011-02-04	-2.06
7	2012-12-28	2.32
7	2013-01-04	-6.08
7	2013-01-11	-6.61
7	2013-01-18	-7.29
17	2013-01-04	2.45
17	2013-01-18	0.25

La tabla nos muestra que los outliers ocurrieron por un invierno muy frio alrededor de Enero 2013 en los stores nro 7 y 17.

Ahora analizamos las tasa de desempleo media por store:

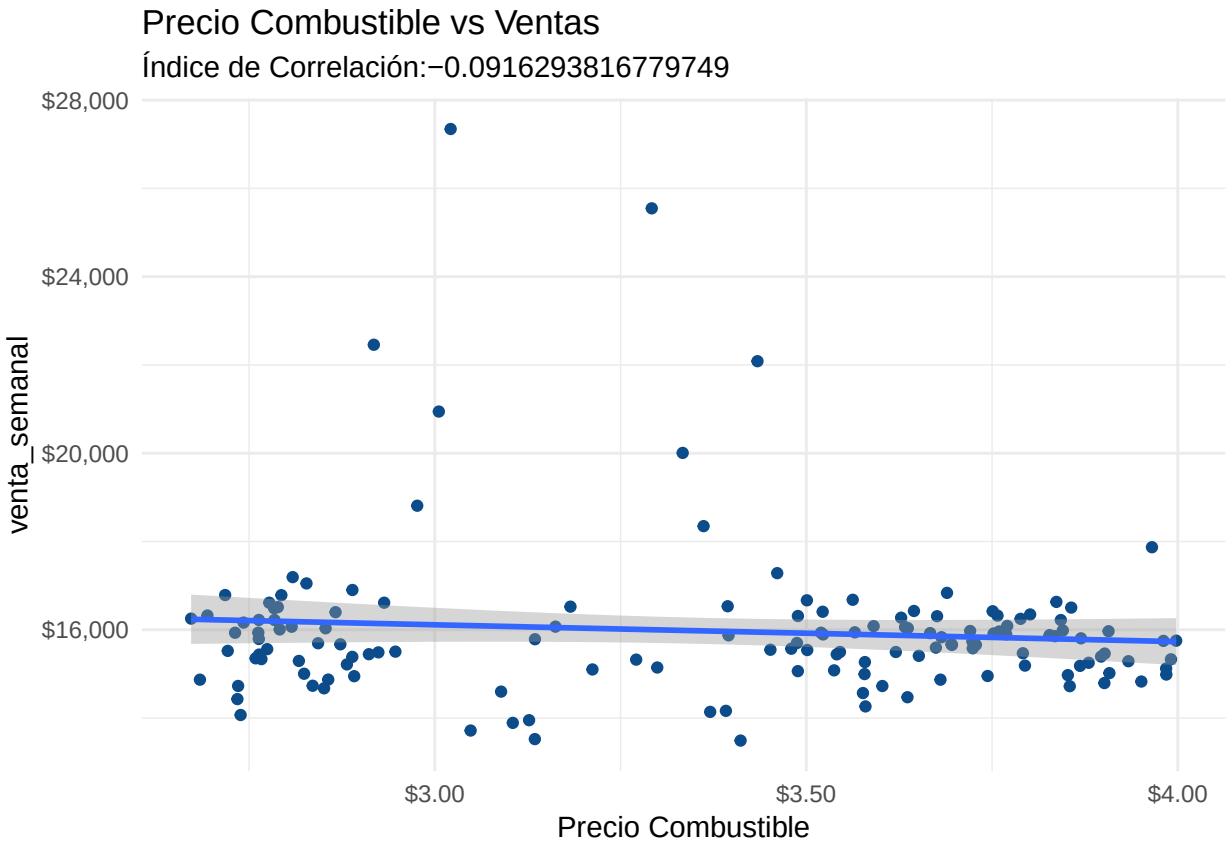
Store	media_desempleo
12	12.63772
28	12.63772
38	12.63772

Si filtramos las tasas de desempleo medias mayores a 10% obtenemos los Stores que tiene valores atípicos de desempleo (stores: 12, 28 y 38)

Para continuar con el análisis queremos saber cómo se relacionan las variables entre sí y cómo afectan los feriados y el mes del año al monto de venta.

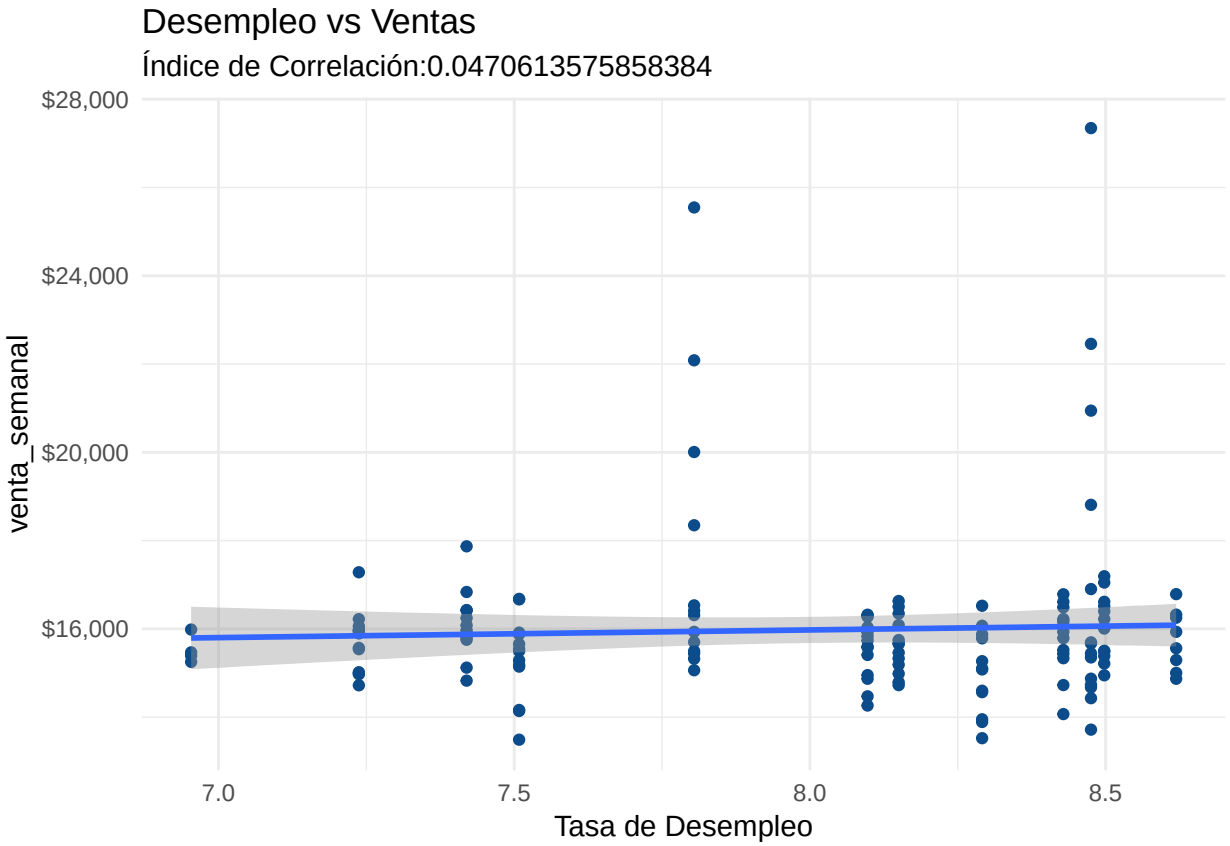
Primero vamos apreparar los datos y para eso necesitamos agrupar por semana y obtener las medias de ventas, precio de combustible y tasa de desempleo:

Ahora vamos a realizar el análisis de correlación de las ventas y el precio del combustible:



La correlación es negativa pero no muy signinficante (-0.09) como para aseverar que las ventas podrían aumentar con la disminución del precio del combustible.

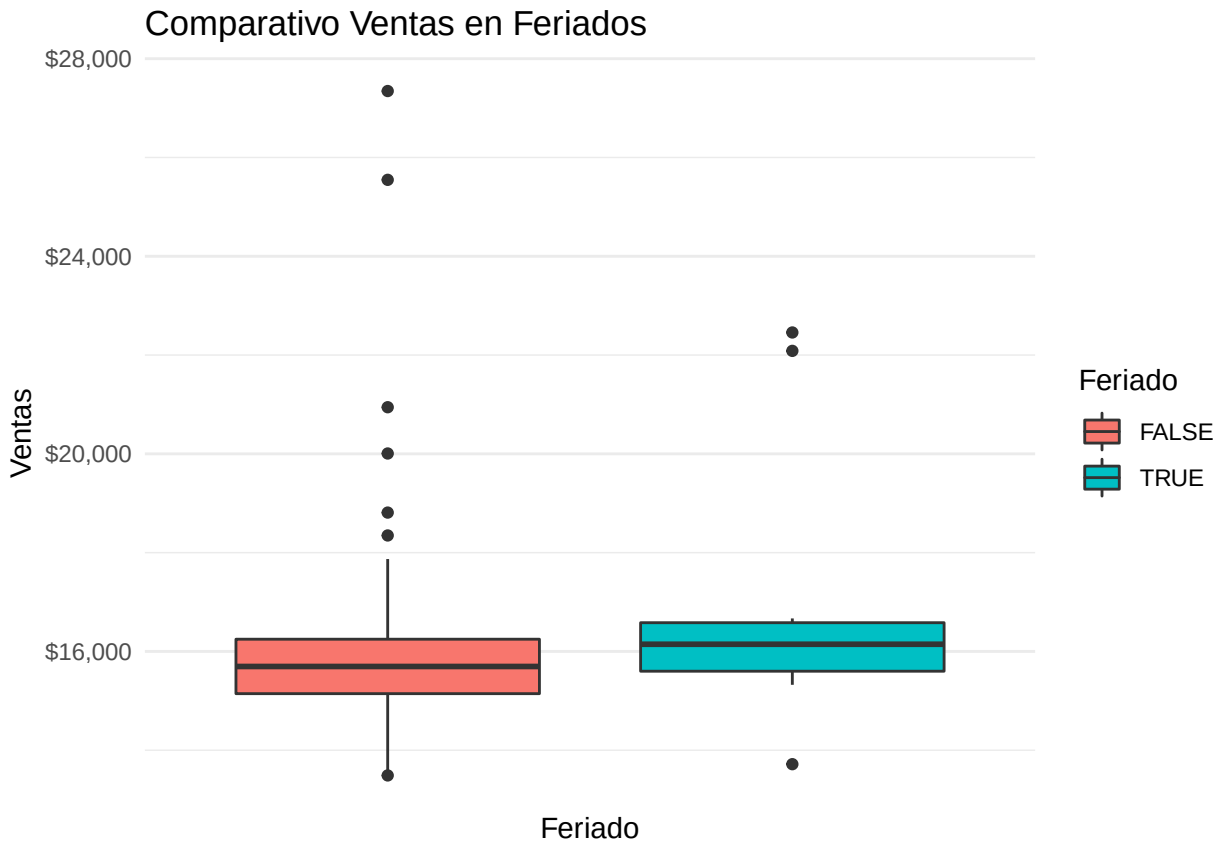
Ahora analizamos la correlación de las ventas con la tasa de desempleo:



La relación de la tasa de desempleo con las ventas es positiva (0.04) pero es insignificativa para aseverar que hay relación.

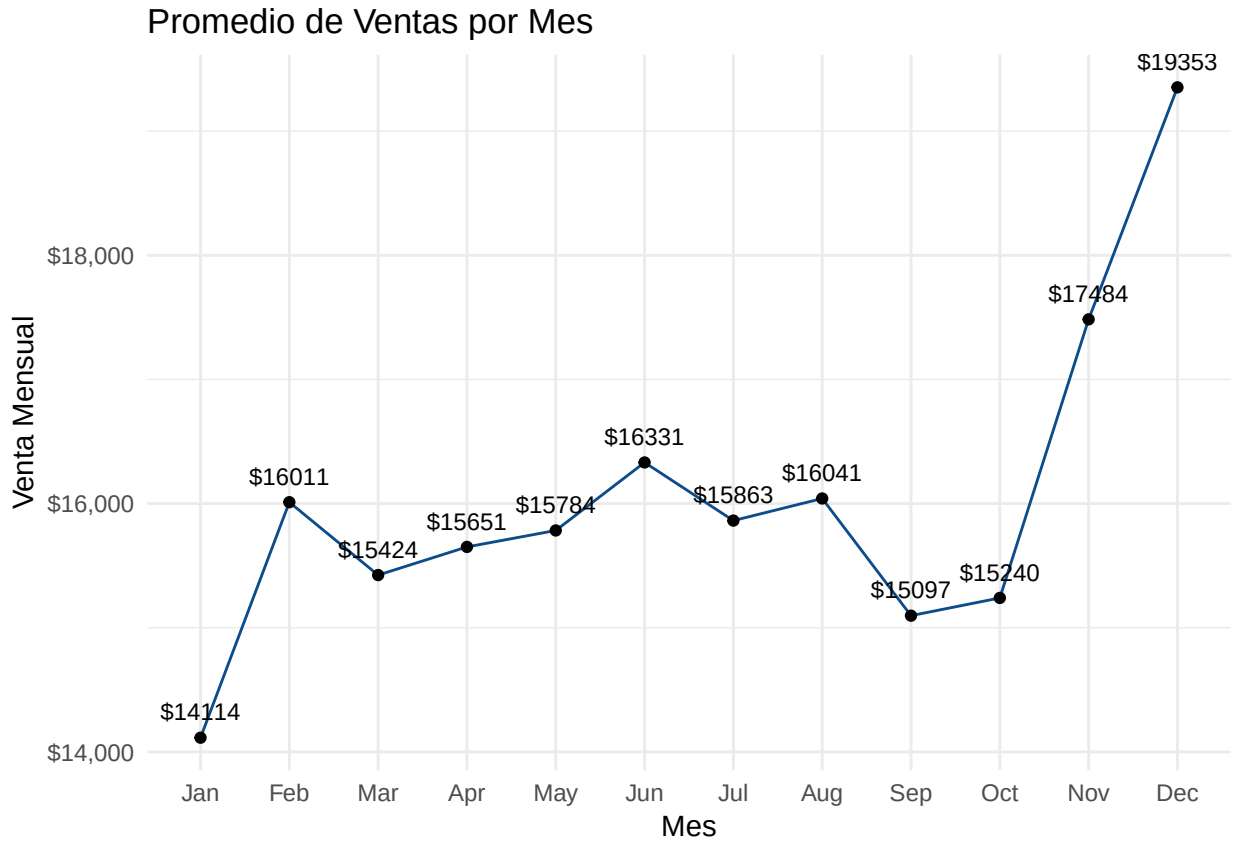
Ahora vamos a analizar el comportamiento de las ventas cuando la semana tiene un día feriado. Primero preparamos los datos y obtenemos los feriados para cada semana:

Ahora juntamos los feriados con las ventas y graficamos:



Efectivamente vemos que la distribución de los promedios de las ventas son relativamente superiores en días feriados.

Ahora para saber si existe una relación entre las ventas y el mes del año agrupamos todas las ventas por mes, computamos las medias y graficamos.



El gráfico nos muestra que las ventas son muy bajas durante el mes de enero, se mantienen en torno a los $15 mil de febrero a octubre y aumentan mucho para la temporada noviembre y diciembre seguramente por navidad.

Como siguiente paso del análisis queremos saber cómo son la probabilidades marginales y conjuntas para los días en que hay feriados y descuentos.

Primero vamos a preparar los datos creando variables binarias para saber cuándo es feriado, cuándo hay descuento y cuándo ambas.

Ahora que tenemos los campos vamos a calcular las probabilidades marginales de cada variable y la probabilidad conjunta.

```
##           IsMarkDown
## IsHoliday    0    1  Sum
##       0   3887 3718 7605
##       1    271  314  585
##       Sum 4158 4032 8190
```

```
## [1] "La probabilidad de que haya un feriado es: 0.07"
```

```
## [1] "La probabilidad de que haya un descuento es: 0.49"
```

```
## [1] "La probabilidad de conjunta de los eventos es: 0.04"
```

Utilicemos el teorema de Bayes para analizar cómo evoluciona la probabilidad de que haya rebajas en una semana que sabemos es feriado.

Primero vamos a crear numeros aleatorios entre 0 y 4 con distribución uniforme para poder separa el data set.

Agregamos la columna como índice y separamos los grupos

```
##           IsMarkDown
## IsHoliday    0    1  Sum
##       0    926  986 1912
##       1     61   73  134
##       Sum  987 1059 2046
```

Probabilidad de que que haya descuento (A) dado que es feriado(B) se calcula como P(A/B) = P(A)* P(B/A) / P(B) Entonces para la iteración 1 tenemos que :

```
## [1] "La probabilidad de que haya descuento dado que es feriado es: 0.52"
```

Vamos a trabajar con el segundo set de datos:

```
##           IsMarkDown
## IsHoliday    0    1  Sum
##       0    979  898 1877
##       1     68   78  146
##       Sum 1047  976 2023
```

Para la segunda iteración contamos con la probabilidad a priori recién calculada (0.56) y la vamos a introducir en la segunda iteración de cálculo de bayes:

```
## [1] "La probabilidad de que haya descuento dado que es feriado para la segunda iteración es: 0.59"
```

Ahora vamos a trabajar con el tercer set de datos:

```
##           IsMarkDown
## IsHoliday    0    1  Sum
##       0   1005  955 1960
##       1     63   88  151
##       Sum 1068 1043 2111
```

Para la tercera iteración realizamos el mismo cálculo:

```
## [1] "La probabilidad de que haya descuento dado que es feriado para la tercera iteración es: 0.67"
```

Ahora vamos a trabajar con el cuarto set de datos:

```
##           IsMarkDown
## IsHoliday    0    1  Sum
##       0   1005  955 1960
##       1     63   88  151
##       Sum 1068 1043 2111
```

Finalmente para la cuarta iteración

```
## [1] "La probabilidad de que haya descuento dado que es feriado para la cuarta iteración es: 0.77"
```