Análisis de datos a dataset de campañas **de ventas en medios de comunicación en lenguaje R**

Integrantes

Castillo Reyes Eduardo Armando

Maravilla Pérez Vianey

Vázquez Portuguez José Antonio

Docente

M. en C. Cristal Karina Galindo Durán

27 de octubre del 2021

Tabla de contenido

Introducción

Se analizará un dataset sobre campañas de ventas en distintos medios de comunicación como la televisión, radio y redes sociales para encontrar comportamientos y características significantes que pueden impactar en las ventas, así como la correlación entre los medios de comunicación y sus ventas y alguna predicción utilizando métodos estadísticos como la regresión lineal.

Así podremos responder a la pregunta ¿Cuál es el mejor medio de comunicación para realizar campañas de ventas? ¿Qué campaña me conviene dependiendo de mis recursos?

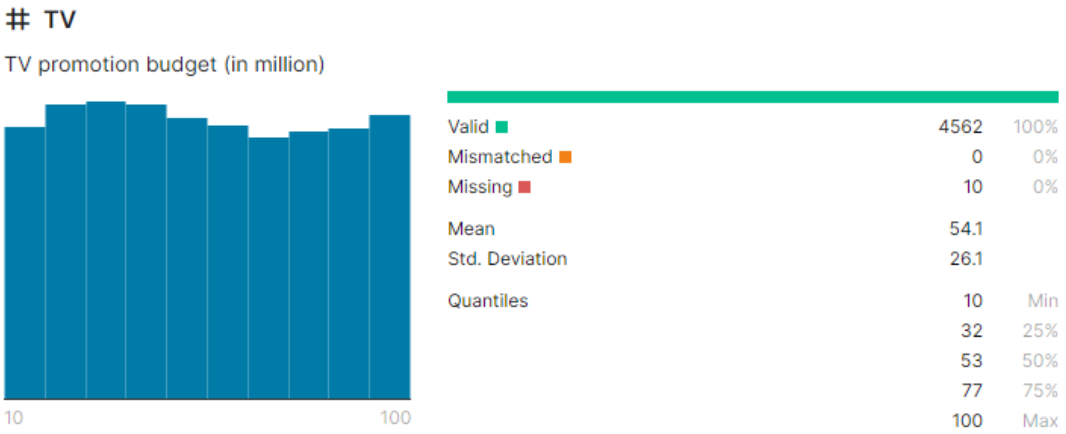
Descripción de los datos

Nuestro dataset lo obtuvimos de la plataforma Kaggle en formato csv, lo que nos facilita la lectura de datos para poder manipularnos en una lista de objetos en R, su estructura es la siguiente:

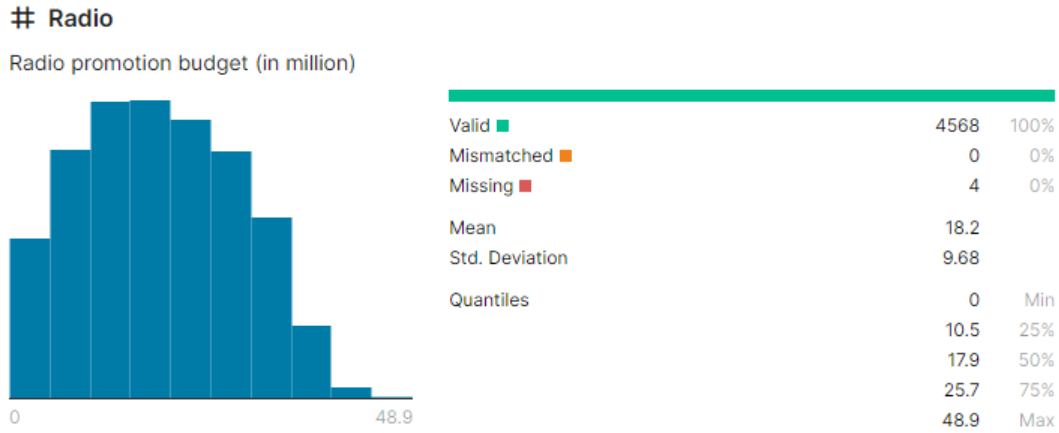
# TV promotion budget (in million)	# Radio promotion budget (in million)	# Social Media promotion budget (in million)	Aa Type of Influencers	# Sales (in million)

Podemos obtener un pequeño resumen de cada columna de nuestro dataset para ver la calidad de datos con la que estaremos manejando.

Resumen y limpieza de datos



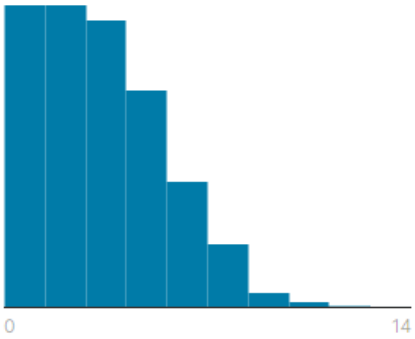
1. Resumen de la columna TV



2. Resumen de la columna Radio

Social Media

Social Media promotion budget (in million)



Valid	4566	100%
Mismatched	0	0%
Missing	6	0%
Mean	3.32	
Std. Deviation	2.21	
Quantiles	0	Min
	1.53	25%
	3.06	50%
	4.81	75%
	14	Max

3. Resumen de la columna Social Media

A Influencer

Type of Influencers

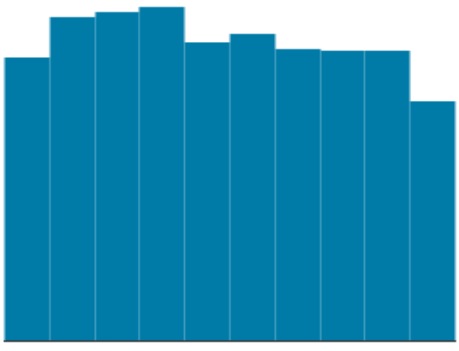
Mega	25%
Micro	25%
Other (2261)	49%

Valid	4572	100%
Mismatched	0	0%
Missing	0	0%
Unique	4	
Most Common	Mega	25%

4. Resumen de la columna Influencer

Sales

Sales (in million)



Valid	4566	100%
Mismatched	0	0%
Missing	6	0%
Mean	192	
Std. Deviation	93.1	
Quantiles	31.2	Min
	112	25%
	189	50%
	273	75%
	364	Max

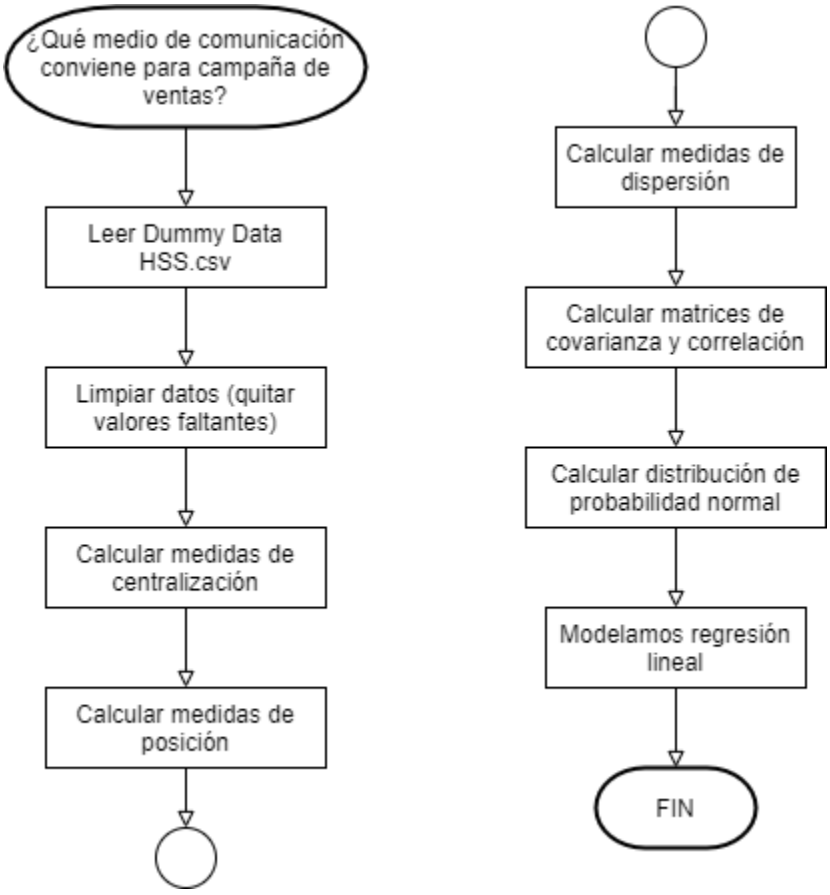
5. Resumen de la columna Sales

Podemos observar que la data de nuestras columnas, cuentan con calidad de datos debido a la poca cantidad de valores faltantes y la distribución que vemos en sus histogramas.

Desarrollo

Diagrama de flujo

A continuación se muestra el flujo del análisis que se llevará a cabo en los datos desde R, donde ocuparemos fórmulas y técnicas estadísticas para describir nuestros datos, los fenómenos sobre nuestra data.



Se realizó una limpieza de datos, donde quitaremos la columna de “Type of influencers” de nuestro dataset debido a que es de tipo caracteres, y solo utilizaremos las otras columnas con valores numéricos para realizar nuestros cálculos, además de que quitaremos los registros que cuenten con valor nulos.

Nos quedaría de la siguiente forma el dataset para poder analizarlo con un tamaño de 4614 registros con calidad de datos

# TV promotion budget (in million)	# Radio promotion budget (in million)	# Social Media promotion budget (in million)	# Sales (in million)

Con los datos ya limpios podremos empezar con nuestro flujo de análisis y descripción de nuestra data.

Medidas de centralización

Para las medidas de centralización (media, moda) utilizamos funciones como mean(), mlv() los cuales nos ayudan a calcular estas medidas con valores muy grandes como nuestro dataset.

Medidas de posición

Utilizamos la función median() para calcular la mediana de nuestros datos, para poder separar a la parte con mayor y menor valor de nuestros datos, además del uso de la función quantile() la cual nos ayuda a posicionar nuestros valores en cuartiles, deciles y percentiles, medidas que se derivan de la mediana.

Medidas de dispersión

Para las medidas de dispersión (desviación, varianza, rango de variación y coeficiente de variación) se utilizaron las funciones sd(), var(), range() y la razón de sd()/mean() para el coeficiente de variación.

Al modelar y obtener las medidas de dispersión se pudo ver la calidad de las regresiones, al igual que el comportamiento de los datos en los mismos rangos de datos.

Matriz de covarianza y correlación

Para calcular la covarianza entre alguna columna de medios y las ventas utilizamos la función cov(), y para modelar y calcular la matriz de covarianza utilizamos las funciones rcorr(), cor() y corrplot() de la librería corrplot.

Para el coeficiente de correlación lineal utilizamos las funciones cor() y plot() para realizar cálculos y poder visualizar la dispersión de la correlación además de la función Correlation() debido a que utilizamos la correlación lineal de Pearson y este nos entrega una gran visualización de esta métrica para los datos a analizar.

Distribución de los datos

La distribución de los datos para nuestro grupo de Radio se utiliza una distribución Normal principalmente por lo que utilizamos funciones como rnorm(), dnorm(), pnorm() y ggplot() para poder visualizar la distribución.

Nuestro conjunto de datos de TV y Sales son muy parecidos a una distribución Uniforme pero no cumple con las condiciones para que sea uniforme por lo que será considerada como distribución Normal, y para nuestra de Redes Sociales se utilizó una distribución Beta por la forma de sus datos, por lo que utilizamos funciones como rbeta(), dbeta(), pbeta() y ggplot()

Regresión lineal

Para el algoritmo de regresión lineal utilizamos las funciones lm(), plot() y abline() para poder obtener y visualizar nuestra línea de tendencia de datos además de obtener los coeficientes de nuestra ecuación de regresión lineal.

Además de que podemos calcular el error de estimación con la función std.error() para checar el nivel de dispersión de los datos, y con la función de Correlacion() podemos juntar las distribuciones, la dispersión, regresión lineal y el coeficiente de correlación con ayuda de un cuadrante y graficas.

Resultados

Medidas de centralización

Para las medidas de centralización obtuvimos los siguientes resultados:

Concepto	Media	Moda
T.V	54.06291	43
Radio	18.15753	0.000683948-9.556845464 (993)
Redes Sociales	3.323473	0.000031300-1.353228723 (993)
Ventas	192.4133	31.19941-102.9680 (993)

Se puede observar que se ha distribuido exponencialmente los recursos para cada medio de comunicación, teniendo una gran desventaja las redes sociales contra la Radio o TV, así que se esperan mejores resultados para la TV o Radio que las Redes Sociales.

Se tiene una moda multimodal de 993 elementos para la Radio, Redes Sociales y las Ventas por lo que se espera una distribución distinta a la normal, tal vez una uniforme debido a la cantidad de modas.

Medidas de posición

Para las medidas de posición se obtuvo la siguiente tabla:

Concepto	Mediana	Cuartiles					Deciles			Percentiles		
		0%	25%	50%	75%	100%	0%	50%	100%	0%	50%	100%
T.V	53	10	32	53	77	100	10	53	100	10	53	100
Radio	17.85951	0%	25%	50%	75%	100%	0%	50%	100%	0%	50%	100%
		0.0006	10.55	17.85	25.64	48.97	0.0006	17.8595	48.8711	0.006	17.85	96.90
Redes Sociales	3.055565	0%	25%	50%	75%	100%	0%	50%	100%	0%	50%	100%
		0.00003	1.53	3.05	4.80	13.98	0.00003	3.0555	13.9816	0.00003	3.0555	13.9816
Ventas	188.9637	0%	25%	50%	75%	100%	0%	50%	100%	0%	50%	100%
		31.19	112.43	188.96	272.32	364.07	31.19	188.96	364.07	31.19	188.96	364.0797

Como vemos en las medidas de posición, podemos ubicar el máximo (100%) y mínimo (0%) y se puede ver una gran diferencia en los presupuestos otorgados para cada campaña, por lo que se espera mejores resultados en ventas por parte de la TV a comparación de las Redes Sociales

Podemos ver que para las Redes Sociales se otorgó a muy pocas campañas un gran presupuesto debido a que solo el último cuartil cuenta con una gran cantidad de dinero, y los otros 3 cuartiles cuentan con muy poco presupuesto, teniendo como máximo 4.8M el 75% y solo otro 25% entre 4.8M - 13.98M

A comparación de las campañas de TV, la distribución de su presupuesto fue algo más uniforme teniendo aumentos lineales por cuartil.

Medidas de dispersión

Para las medidas de dispersión encontramos los siguientes resultados:

Concepto	Desviación	Varianza	Rango de Variación	Coeficiente de variación
T.V	26.10494	681.468	90	0.4828623
Radio	9.66326	93.37859	48.87048	0.5321901
Redes Sociales	2.211254	4.889644	13.98163	0.6653444
Ventas	93.01987	8652.697	364.0797	0.4834378

Podemos confirmar los enunciados que inferimos sobre la cantidad de dinero otorgada para cada campaña, con el rango de variación observamos que las campañas de TV tuvieron mayor presupuesto a comparación de las Redes Sociales, por lo que se espera una diferencia exponencial a nivel de ventas de la TV sobre las Redes Sociales.

Además de que la varianza y la desviación estándar nos indica que en las Redes Sociales se encuentran muy cercanos a su media, por lo que podremos ver el impacto en campañas de Redes Sociales con muy poco presupuesto, cercano a su media de 3.32M.

Mientras que en la TV podremos tener presupuestos en un mayor rango y podremos analizar el comportamiento en pequeñas, medianas y grandes campañas.

Matriz de covarianza y correlación

Concepto	Covarianza	Correlación
T.V	2427.058	0.9994974
Radio	780.797	0.8686378
Redes Sociales	108.4907	0.5274464

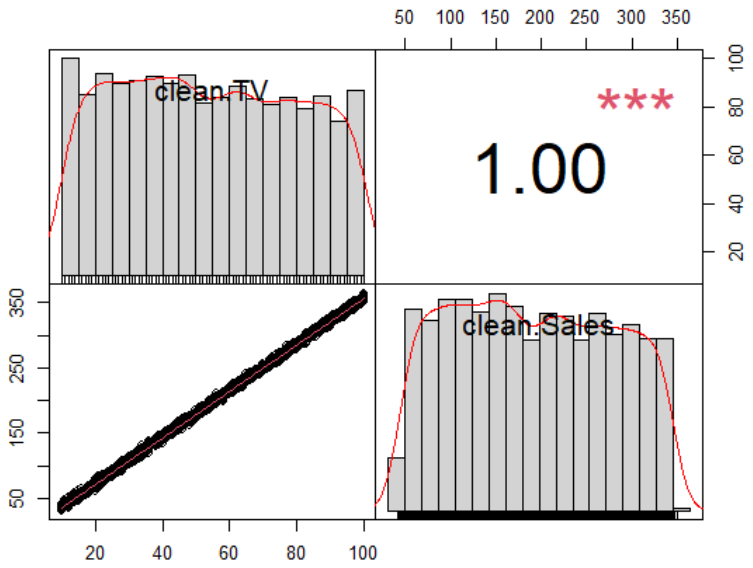
Observando los resultados de la covarianza y la correlación, vemos que existe una gran correlación entre la campaña de TV y sus ventas, esto tal vez debido a la gran desviación de datos que abarcan mucha más población que la de Redes Sociales o Radio.

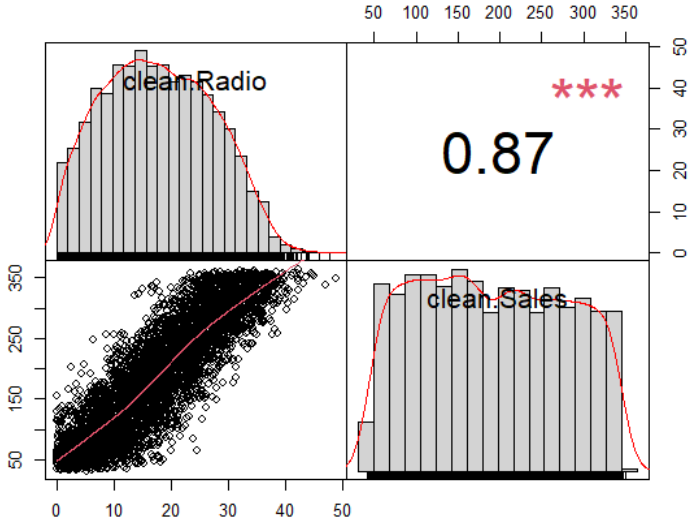
Se observa que la correlación de las campañas de TV (0.99) y Radio (0.86) son fuertes y perfectas para generar nuestro modelo de regresión lineal, mientras que para Redes Sociales es menor la covarianza y moderada su correlación para obtener un buen modelo.

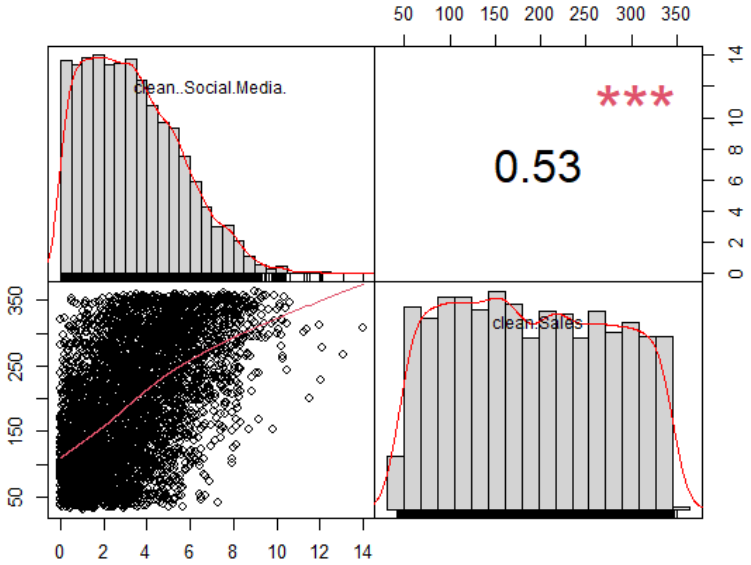
En las siguiente imagen mostramos una matriz donde se observan las correlaciones de los datos para ver qué campañas tienen mayor relación con las ventas.

Como podemos observar en la imagen, la correlación es mucho mejor entre TV y Ventas que con Redes Sociales, obteniendo una correlación perfecta positiva entre TV y Ventas. Además la Radio se acerca mucho a esa correlación perfecta mientras que las Redes Sociales siguen estando en último lugar para poder analizar sus campañas y obtener todas sus posibilidades.

Nuestro coeficiente de correlación lineal sería de Pearson debido a su distribución Normal como veremos en las siguientes imágenes se observan las gráficas de dispersión, con su línea de tendencia o regresión lineal que veremos en otro apartado, sus distribuciones y su medida de correlación para poder comprender mejor la correlación entre los datos:







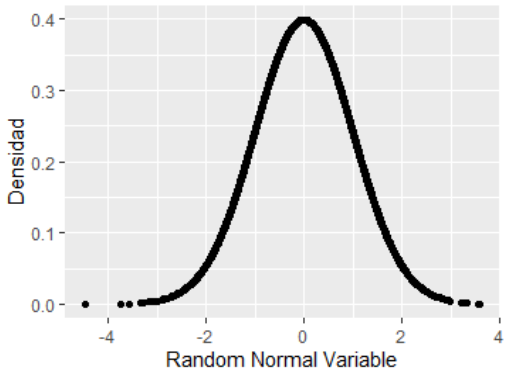
El único anormal como esperábamos es el de redes sociales, esto debido a que su distribución es muy distinta.

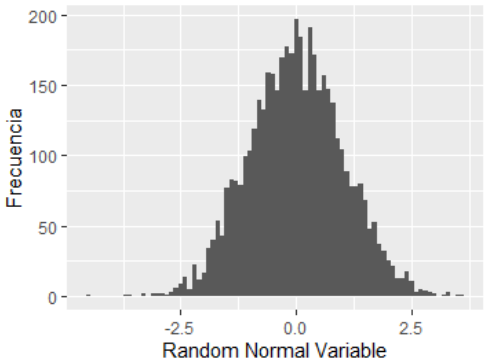
Distribución de los datos

La distribución de los datos la podemos observar en el análisis anterior, como vemos la de ventas y TV son muy parecidas, que se aproximan mucho a Uniforme pero siguen siendo Normal, y al ver sus distribuciones podemos ver fácilmente su comportamiento parecido, y la distribución Beta de la campaña de Redes Sociales.

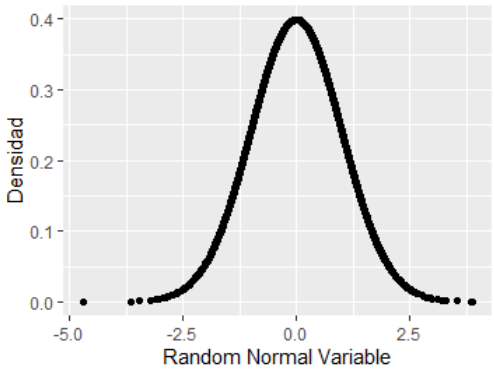
Utilizando las fórmulas de distribución normal y beta con *alpha = 1* y *betha = 5* obtenemos lo siguiente:

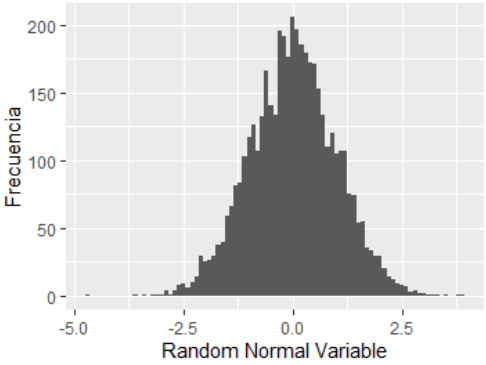
Para **TV**



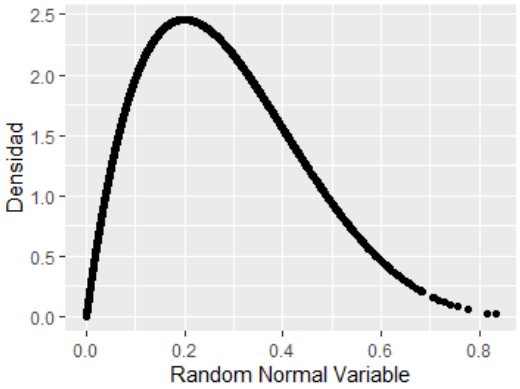


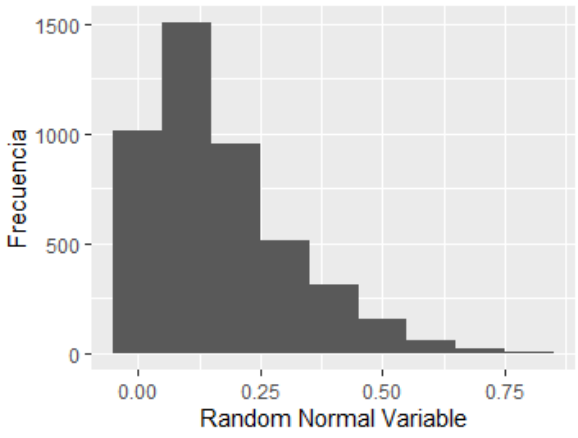
Para **Radio**





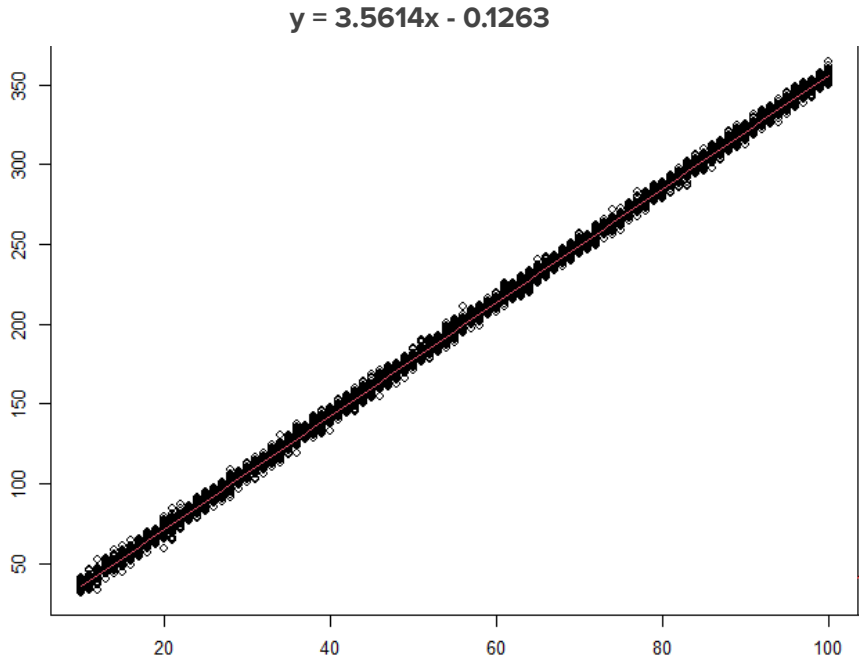
Para **Redes Sociales**



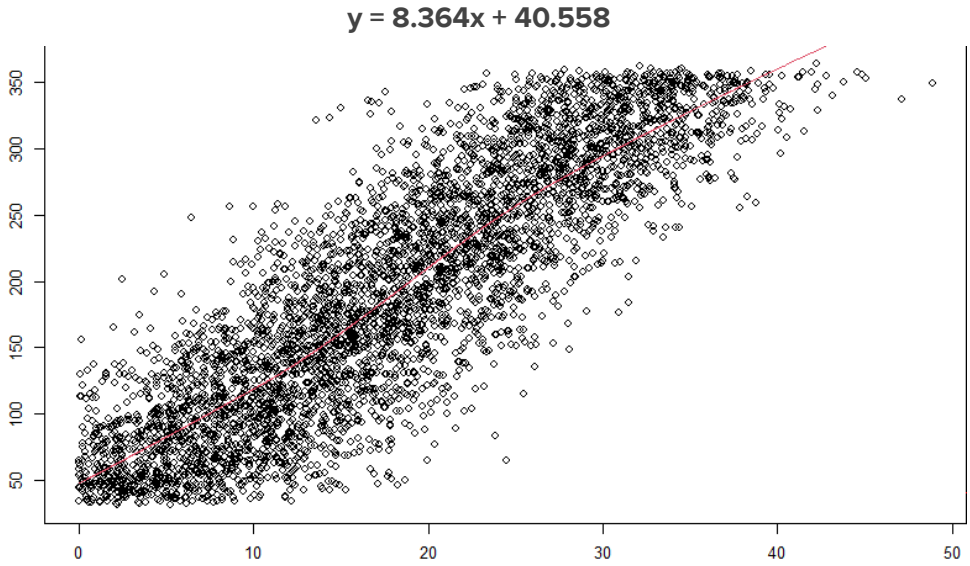


Regresión lineal

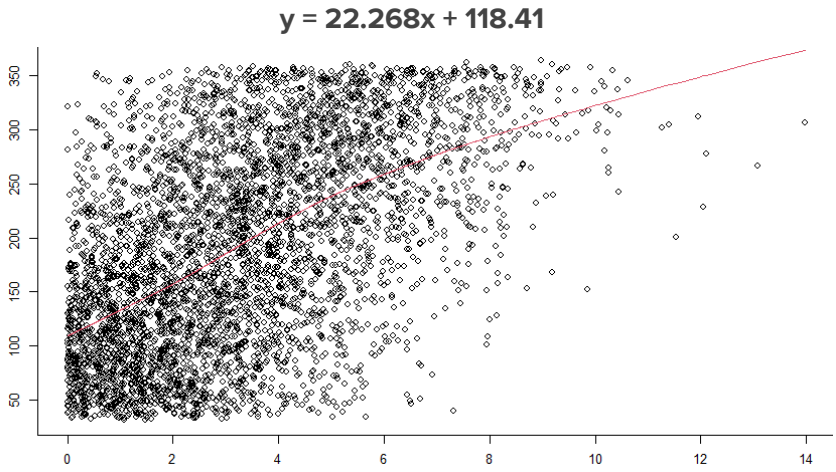
Obtuvimos la línea de tendencia central para la campaña de TV y Ventas con el modelo de regresión lineal y obtuvimos el siguiente modelo:



Obtuvimos la línea de tendencia central para la campaña de Radio y Ventas con el modelo de regresión lineal y obtuvimos el siguiente modelo:



Obtuvimos la línea de tendencia central para la campaña de Redes Sociales y Ventas con el modelo de regresión lineal y obtuvimos el siguiente modelo:



Conclusión

Al terminar los procesos en el código y mostrar los resultados podemos concluir que por cada aumento de 1 en anuncios de televisión, las ventas aumentarán en 3,5

Por cada aumento de 1 en anuncios de Soc Media, las ventas aumentarán solo 0,06.

Por ello se prevé que las ventas disminuyan en 0,4.

Los datos dentro del dataset sugieren que los anuncios de televisión son los que más impulsan las ventas, seguido de los anuncios en la radio. Si la empresa decide continuar invirtiendo en anuncios de redes sociales solo le convendría a una escala muy grande en comparación con los otros dos medios .

Se necesitaría realizar un estudio de campañas con mayor financiamiento en las redes sociales para checar un mejor comportamiento de su impacto en las ventas.

Referencia

Saragih, H. S. (2021, 11 marzo). Datos de marketing y ventas. Kaggle. Recuperado 20 de octubre de 2021, de https://www.kaggle.com/harrimansaragih/dummy-advertising-and-sales-data