Análisis exploratorio de datos

Una forma de aprender cosas nuevas es a través del descubrimiento. El análisis exploratorio de datos es un término atribuido al estadístico John Tukey en un libro del mismo nombre (Tukey, 1977).

El **análisis exploratorio de datos** implica examinar un conjunto de datos para descubrir sus características subyacentes con énfasis en la visualización. Te ayuda durante el diseño del análisis a determinar si debes recopilar más datos, sugerir hipótesis para probar e identificar modelos para desarrollar. En este documento, cubriremos los siguientes cuatro temas relacionados con el análisis exploratorio de datos:

- Entender el análisis exploratorio de datos
- Análisis de una sola variable de datos
- Análisis de dos variables juntas
- Exploración de múltiples variables simultáneamente

Aprenderás técnicas comunes que los estadísticos y analistas utilizan para caracterizar los datos. Estas incluyen métodos tabulares y gráficos para explorar el conjunto de datos. Hay muchas cosas interesantes para descubrir en un conjunto de datos, pero en los negocios o la ciencia, tú estás explorando para determinar los aspectos de tus datos que ayudarán a construir modelos analíticos.

Caso de uso: Preguntas sobre Bike Sharing, LLC

Has estado trabajando con datos de Bike Sharing durante el tiempo suficiente como para sentirte cómodo con su contenido y calidad. Ahora, el gerente de marketing de la empresa ha organizado una reunión para hablar contigo. Quiere obtener más información de los datos que le ayudarán a impulsar algunas decisiones comerciales que su grupo está considerando.

Aunque los modelos son un aspecto importante de la analítica, el proceso en realidad comienza con una buena pregunta. Descubriste que hacer la pregunta correcta suele ser la parte más difícil de resolver un problema. No puedes simplemente preguntarle al gerente de marketing: “¿Qué preguntas tiene para mí?” No es tan fácil.

Por lo tanto, pasas de ser un analista de negocios a asumir la personalidad de un consultor y facilitador.

Deberás trabajar con otras personas que tengan experiencia en el campo del marketing e interactuarás con ellas para ayudarlas a formular buenas preguntas. Los datos que explorarás en este documento son datos de marketing que utilizarás para el análisis en documentos posteriores. El archivo marketing.csv está disponible para ti.

Bike Sharing LLC adquirió un informe de la industria que contenía información comercial sobre otras 172 operaciones de alquiler de bicicletas recopilada durante un análisis de tendencias de la industria. El conjunto de datos consta de 172 observaciones y 7 variables:

- Gastos en publicidad en Google AdWords
- Gastos en publicidad en Facebook
- Gastos en publicidad en Twitter (X)
- Presupuesto total de marketing
- Ingresos asociados a esa instalación específica
- Número de empleados
- Mercado según la densidad de población (bajo, medio, alto)

Este tercer documento te ofrece la oportunidad de aprender nuevas habilidades, así como una nueva forma de analizar los problemas de datos. Buena suerte con tus exploraciones.

Comprensión del análisis exploratorio de datos

"Una respuesta aproximada a la pregunta correcta vale mucho más que una respuesta precisa a la pregunta equivocada".

– John Tukey

El análisis exploratorio de datos es un área fascinante, ya que combina el arte de la conversación, las habilidades de la ciencia de datos y los aspectos del dominio que se está estudiando. Es un proceso estructurado en el que se descubre información sobre las características de los datos y las relaciones entre dos o más variables.

Las preguntas importan

El bioestadístico Roger Peng ha dicho que desarrollar preguntas es una forma práctica de reducir la cantidad exponencial de formas en las que se puede explorar un conjunto de datos. En particular, una pregunta o hipótesis aguda puede servir como una herramienta de reducción de dimensión que puede eliminar variables que no son inmediatamente relevantes para la pregunta.

El caso de uso destaca la importancia de hacer buenas preguntas. Tu éxito en la realización del análisis exploratorio de datos dependerá de tus habilidades como investigador. Por un momento, asume el personaje de un periodista o investigador. Viven en un mundo de preguntas y buscan datos para responderlas. Puedes pensar en la ciencia de datos de manera similar.

Es posible que ya tengas una cantidad considerable de datos y sea tentador sumergirse en ellos y comenzar a responder preguntas. ¿Cómo sabes que los datos son apropiados para responder las preguntas?

Al comenzar este documento, consideramos esta historia. Un filósofo tenía algunos estudiantes y era el final de su tiempo juntos. Les preguntó a sus estudiantes: "Entonces, ¿cuál es la respuesta?" Sus estudiantes se miraron entre sí, pero nadie dijo nada. El filósofo luego hizo la pregunta más importante: "Bueno, si nadie tiene la respuesta, ¿al menos alguien puede decirme cuál es la pregunta?" El objetivo final del análisis es comunicar una respuesta a una pregunta. La siguiente figura muestra un proceso generalizado de ciencia de datos de principio a fin:

Los modelos y el análisis son útiles en la medida en que responden a una pregunta bien desarrollada. Un objetivo clave en la inteligencia de negocios es proporcionar **respuestas**, a través de **modelos**, a la **pregunta**.

Escalas de medición

A veces, es importante volver a los fundamentos de una disciplina y recordarnos cosas que creemos que entendemos muy bien. En este documento, te resultará importante comprender la escala de los datos con los que estás trabajando. Los datos pueden ser de diferentes tipos, y cada tipo tiene ciertas estadísticas asociadas a ellos, denominadas **escalas de medición**.

Stanley Stevens fue un psicólogo que acuñó el término escalas de medición. Destacó este tema en un artículo que culminó un comité de siete años de la Asociación Británica para el Avance de la Ciencia. Su tarea era agregar significado a la medición. El comité ideó cuatro tipos de escalas de medición: **nominal**, **ordinal**, de **intervalo** y de **razón**. Puede resultarte difícil de creer, pero la gente no pudo señalar un estándar sobre este tema hasta mediados del siglo pasado.

La siguiente tabla muestra cada escala, así como las operaciones que permiten y las estadísticas permitidas para cada escala. Tómate un poco de tiempo para revisar estas escalas, ya que haremos referencia a ellas en el documento. En el análisis exploratorio de datos, es importante comprender las escalas de medición al generar estadísticas y gráficos (plots) de resumen:

Escala	Operaciones empíricas básicas	Estadísticas permitidas
Nominal	Determinación de igualdad o pertenencia	Número de casos Moda Correlación de contingencia
Ordinal	Determinación de mayor o menor que	Mediana Percentiles
Intervalo	Determinación de igualdad de intervalo o diferencia	Media Desviación estándar Correlación de orden de rango Correlación producto-momento
Relación	Determinación de igualdad de relaciones	Coeficiente de variación

A continuación, se ofrece un ejemplo que ayuda a explicar las diferentes escalas. Los siguientes datos proceden de una lista de alumnos de una escuela secundaria que muestra el género, el nivel, la edad y el promedio de calificaciones (GPA, grade point average):

Gender	Level	Age	GPA
Male	Sophomore	15	3.6
Female	Freshman	14	3.2
Female	Sophomore	14	3.3
Male	Junior	16	3.7
Female	Senior	18	3.1
Male	Senior	17	2.8

Los datos anteriores se describen a continuación:

- El género (gender) es **nominal**. Un estudiante puede ser hombre o mujer. Puedes calcular el recuento (count) y la moda (mode) de los datos de escala nominal.
- El nivel (level) es **ordinal**. Los estudiantes de último año están por encima de los de tercer año, los de tercer año están por encima de los de segundo año y los de segundo año están por encima de los de primer año. Puedes calcular la mediana y los percentiles de datos de escala ordinal cuando se representan numéricamente.

- La edad (age) es **intervalo**. Los valores son continuos con un incremento entendido entre ellos. En este ejemplo, un número entero representa la edad de un estudiante y difieren en incrementos de un año. Puedes calcular la media y la desviación estándar de los datos de la escala de intervalo.
- El GPA es una **proporción** (**ratio**). Es una combinación de otros dos números expresados como número racional.

Lo importante que hay que recordar es que los datos pertenecen a diferentes escalas. Estas escalas permiten cierto tipo de operaciones y estadísticas, y no son aptas para otro tipo.

Tipos de datos R

Cuando los datos se cargan en R, se les asigna un tipo de datos. Los siguientes son los cinco tipos de datos más comunes que encontrarás en tus análisis:

- **Numérico (numeric)**: Un número con un valor decimal (por ejemplo, 10.1 y 2.0). R usa numérico como tipo de datos predeterminado para los números, a menos que los definas como otro tipo de datos. Numérico utiliza una escala de intervalo. La mayoría de las proporciones son numéricas.

- **Entero (integer)**: Un número sin valor decimal (por ejemplo, 10 y 2). Los números enteros también utilizan una escala de intervalo. Numérico es el tipo de datos predeterminado en R, pero puedes convertir un número en un número entero usando la función as.integer():

```
> as.integer(10.0)
[1] 10
```

- **Carácter (character)**: Una representación de un valor de cadena en R. Se pueden usar caracteres individuales o se pueden concatenar varios caracteres con la función cat() para formar cadenas más largas. Estos objetos de cadena utilizan una escala nominal u ordinal. Puedes forzar un número a un carácter usando la función as.character():

```
> as.character(10.1)
[1] "10.1"
```

- **Lógico (logical)**: Un valor almacenado como TRUE o FALSE. Los valores lógicos suelen ser el resultado de una operación que compara otros dos valores. Considera lo siguiente como ejemplo:

```
> 10 == 2
[1] FALSE
```

- **Factor**: Un factor se aplica a variables categóricas, tanto nominales como ordinales. Según la información de ayuda de R (disponible escribiendo ?factor() en la consola), un factor codifica un vector de categorías (carácter o números) de tal manera que se consideran niveles de esta variable. Estas etiquetas discretas son bastante útiles para agrupar y filtrar datos según categorías.

R tiene otros dos tipos de datos, complejos y sin procesar, pero no se analizarán por este momento.

Si deseas determinar en qué tipo de datos está almacenada tu variable, puedes aplicar la función class() a la variable y R generará su tipo de datos.

Consejo R: Es importante comprender los tipos de datos. Las bibliotecas y funciones de R a menudo requieren que los datos sean de un determinado tipo para poder ejecutarse. Crecerás como analista a medida que dediques tiempo a conocer los tipos de datos en su conjunto de datos y convertirlos cuando sea necesario.

Analizar una única variable de datos

Si tu conjunto de datos tiene una sola variable, tienes datos univariados. Al examinar datos univariados, puedes describir la distribución de los datos en términos de su valor y extensión. Un buen lugar para comenzar la exploración de datos univariados es con la función str() que aprendiste en el documento anterior, Limpieza de datos. Carga el conjunto de datos de marketing en R y ejecuta la función str():

```
> marketing <- read.csv("marketing.csv", stringsAsFactors = TRUE)
> str(marketing)
'data.frame':   172 obs. of  7 variables:
 $ google_adwords : num  65.7 39.1 174.8 34.4 78.2 ...
 $ facebook       : num  47.9 55.2 52 62 40.9 ...
 $ twitter        : num  52.5 77.4 68 86.9 30.4 ...
 $ marketing_total: num  166 172 295 183 150 ...
 $ revenues       : num  39.3 38.9 49.5 40.6 40.2 ...
 $ employees      : int  5 7 11 7 9 3 10 6 6 4 ...
 $ pop_density    : Factor w/ 3 levels "High","Low","Medium": 1 3 3 1 2 1 2 1 3 2 ...
```

Verás que contiene 172 observaciones de 7 variables. Las primeras seis variables son numéricas, mientras que la última es un factor que tiene tres niveles: Alto (High), Bajo (Low) y Medio (Medium). Si deseas que el factor tenga orden, debes definir el factor ordenado.

Creaste un código similar en el documento anterior. Puedes utilizar este conocimiento para convertir pop_density en un factor ordenado:

```
> marketing$pop_density <- factor(marketing$pop_density,
+   ordered = TRUE,
+   levels = c("Low", "Medium", "High"))
```

Concéntrate en dos variables, google_adwords (interval, numeric) y pop_density (ordinal, factor). Puedes aprender tus distribuciones utilizando un enfoque tabular o gráfico.

Exploración tabular

Este tipo de análisis exploratorio de datos es relativamente rápido y consiste en imprimir números en tu consola. La función summary() produce un resumen de cinco números (más la media) para las variables continuas. También produce un recuento de las variables categóricas:

```
> summary(marketing$google_adwords)
   Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
  23.65   97.25   169.47   169.87  243.10   321.00
```

Para cualquier variable de escala de intervalo en el conjunto de datos de marketing, la función summary() te proporciona lo que se conoce como el resumen de cinco números de Tukey más la media. Se proporciona la siguiente información:

- **Mínimo (minimum)**: Esta es la observación más pequeña en el conjunto de datos.
- **Primer cuartil (first quartile)**: El 25 % de los datos se encuentra por debajo de este valor.
- **Mediana (median)**: Esta es la observación del medio.
- **Media (mean)**: El promedio de las observaciones.
- **Tercer cuartil (third quartile)**: El 75 % de los datos se encuentra por debajo de este valor.
- **Máximo (maximum)**: Esta es la observación más grande en el conjunto de datos.

El resumen de cinco números proporciona un resumen fácil de entender de cómo se distribuyen los datos. Proporciona información sobre la tendencia central (la mediana), la dispersión (los cuartiles) y el rango (el mínimo y el máximo). Al utilizar la mediana en lugar de la media, el resumen de cinco números es aplicable a las mediciones ordinales, así como a las mediciones de intervalo y de razón. R también proporciona el valor medio de la variable. Si solo deseas el resumen de cinco números, puede utilizar la función fivenum().

Como se mencionó, R proporciona la media junto con el resumen de cinco números. Puede obtener la media por separado utilizando la función mean():

```
> mean(marketing$google_adwords)
[1] 169.8685
```

También puedes obtener dos medidas de dispersión que suelen estar asociadas con la media (desviación estándar y varianza) utilizando las funciones sd() y var() respectivamente:

```
> sd(marketing$google_adwords)
[1] 87.47228
```

Si utilizamos la siguiente entrada:

```
> var(marketing$google_adwords)
```

Obtendremos la varianza:

```
[1] 7651.4
```

Ahora sabes bastante sobre tus datos. Por ejemplo, en el caso de google_adwords, sabes:

- Los valores varían de un mínimo de 23.65 a un máximo de 321
- El 25 % de las observaciones están por debajo de 97.25
- El 50 % de las observaciones están por debajo de 169.50 (la mediana)
- El 75 % de las observaciones están por debajo de 243.10

Para la variable categórica pop_density, la función summary() solo enumera los tres niveles de factor y la cantidad de observaciones en cada nivel. Observa que los factores están ordenados según los niveles que estableciste, no alfabéticamente como lo estaban antes de definir un orden:

```
> summary(marketing$pop_density)
```

El resultado es el siguiente:

```
Low   Medium    High
  68     52       52
```

La exploración tabular te dice bastante. ¿Sientes que sabes todo lo que podrías? Algunas cosas permanecen ocultas hasta que las representas gráficamente

Exploración gráfica

"El mayor valor de una imagen es cuando nos obliga a notar lo que nunca esperamos ver".
– John Tukey

Obtener estadísticas resumidas es importante, pero ver la forma de los datos es revelador. El Cuarteto de Anscombe es un ejemplo clásico desarrollado específicamente para ilustrar el valor de la visualización de datos (Anscombe, 1973). Es tan clásico que R viene con los datos cargados:

```
> data("anscombe")
> anscombe
```

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

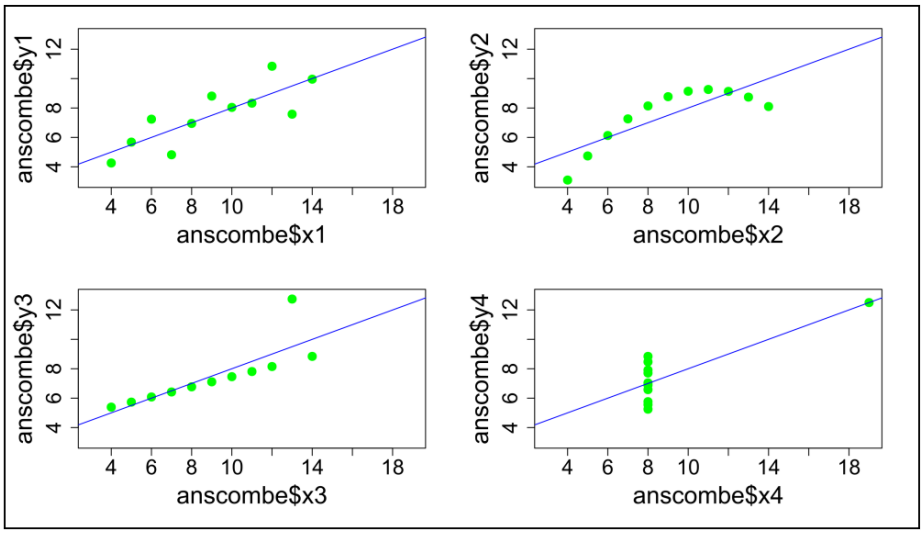
Lo interesante de estos cuatro conjuntos de datos es que las estadísticas resumidas (como la media, la desviación estándar y la varianza) para cada una de las variables son casi idénticas. A continuación, se muestra la salida de la consola de estas estadísticas descriptivas utilizando la función sapply(), que aplica otra función, como mean(), en todas las columnas:

```
> sapply(anscombe, mean)
    x1      x2      x3      x4      y1      y2      y3      y4
9.000000 9.000000 9.000000 9.000000 7.500909 7.500909 7.500000 7.500909
> sapply(anscombe, sd)
    x1      x2      x3      x4      y1      y2      y3      y4
3.316625 3.316625 3.316625 3.316625 2.031568 2.031657 2.030424 2.030579
> sapply(anscombe, var)
     x1       x2       x3       x4       y1       y2       y3       y4
11.000000 11.000000 11.000000 11.000000  4.127269  4.127629  4.122620  4.123249
```

A partir de esto, se puede concluir que estos conjuntos de datos son de alguna manera similares, o incluso idénticos. Al representar gráficamente los cuatro conjuntos de datos, se presenta una imagen muy diferente, como se muestra aquí:

Se deja de tarea el cómo agregar personalizaciones a los gráficos que se generan.
Algunos ejemplos:

```
> plot(anscombe)
> plot(anscombe$y1,anscombe$x1)
> plot(anscombe$y1,anscombe$x1, col="green", pch=19, cex=1.5)
```

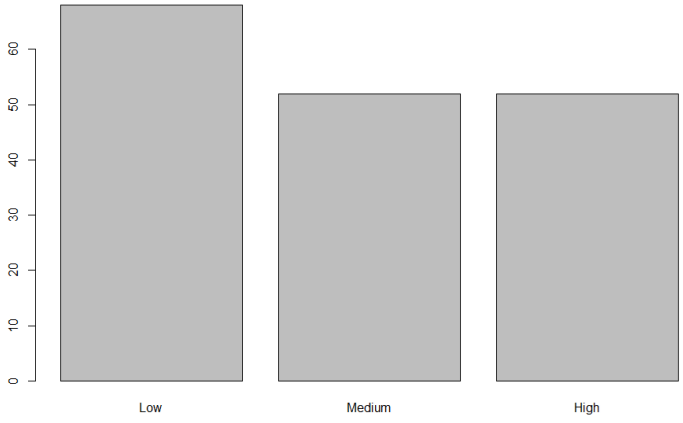


En la sección anterior, exploraste tus datos mediante una inspección tabular. Ahora, inspeccionarás tus datos utilizando técnicas gráficas. Utilizarás gráficos para inspeccionar factores ordinales, como pop_density. Los diagramas de caja y los histogramas presentan datos de escala de intervalo, como google_adwords.

Dibuja pop_density utilizando la función plot() para visualizar los valores 68, 52 y 52:

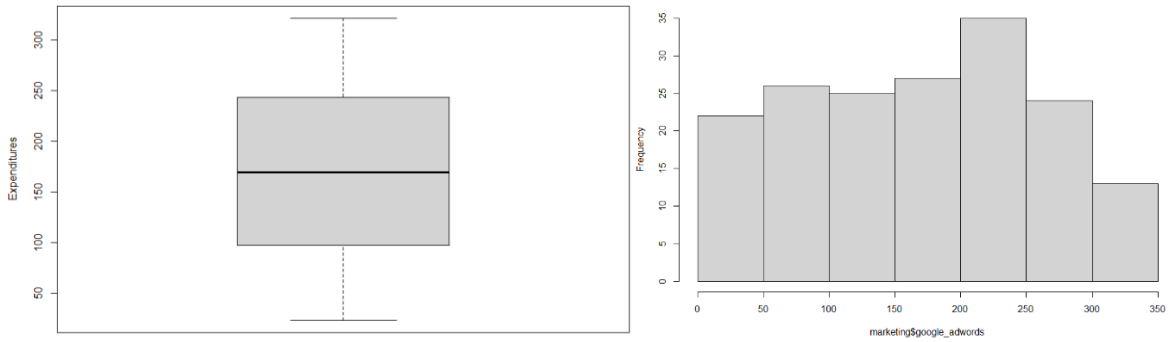
```
> plot(marketing$pop_density)
```

Devolverá el siguiente resultado:



Ahora, explora la variable google_adwords utilizando las funciones boxplot() y hist():

```
> boxplot(marketing$google_adwords, ylab = "Expenditures")
> hist(marketing$google_adwords, main = NULL)
```



El panel izquierdo muestra un diagrama de caja (*boxplot*) con una línea horizontal gruesa en el medio que representa la mediana. Verás que se alinea con el valor 169, que encontraste con la exploración tabular. Los límites superior e inferior del cuadro corresponden a los cuartiles superior e inferior, 243.10 y 97.25. Los bigotes se extienden desde la parte superior e inferior del cuadro.

En este caso, representan el máximo (321,00) y el mínimo (23,65). Este no es siempre el caso, y la información en el siguiente cuadro de información de bigotes del diagrama de caja (*boxplot whiskers*) lo explica. El diagrama de caja es coherente con el resumen de cinco números que extrajiste anteriormente.

El panel derecho es un histograma. Revela lo que los datos tabulares y el diagrama de caja no revelaron: la variable google_adwords está distribuida uniformemente. No es una curva de campana (curva normal).

Esta información sigue siendo coherente con el diagrama de caja y las estadísticas de resumen, pero una curva normal podría haber mostrado información similar. Observa el etiquetado sin procesar en el eje x del histograma. Esto está bien en el análisis exploratorio de datos para crear gráficos rápidamente.

Boxplot whiskers (opcionales e intermedios)

Los boxplot whiskers son complicados. No necesariamente representan los valores mínimo y máximo en el conjunto de datos, aunque pueden hacerlo. El bigote superior representa el que sea menor:

- (a), el valor máximo, (321)

- (b), el valor del cuartil superior, (243.10) + (1.5 x IQR), donde IQR es el rango intercuartil, o la diferencia entre el cuartil superior, (243), y el cuartil inferior, (97.25). Esto es 243 – 97.25 = 73.53.

En este caso, el valor máximo, (a), es 321. El valor alternativo es (b) = 243 + (1.5 x 73.53) = 243 + 113.29 = 356.29. El valor máximo, (a), del rango es menor, por lo tanto, esto establece el bigote superior.

El bigote superior representa el que sea mayor:

- (a), el valor mínimo, (23.65)
- (b), el valor del cuartil inferior, (97.25) – (1.5 x IQR)

En este caso, el valor mínimo, (a), es 23.65. El valor alternativo es (b) = 97.25 – (1.5 x 73.53) = 97.25 – 109.87 = -12.62. El valor mínimo, (a), del rango es mayor, por lo tanto, esto establece el bigote inferior.

La variable google_adwords es uniforme. ¿Cómo se vería una variable sesgada en formato tabular y gráfico? Resulta que la variable twitter está sesgada. Continua y observa el resultado tabular utilizando summary(marketing$twitter):

```
> summary(marketing$twitter)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
   5.89   20.94    34.59    38.98   52.94   122.19
```

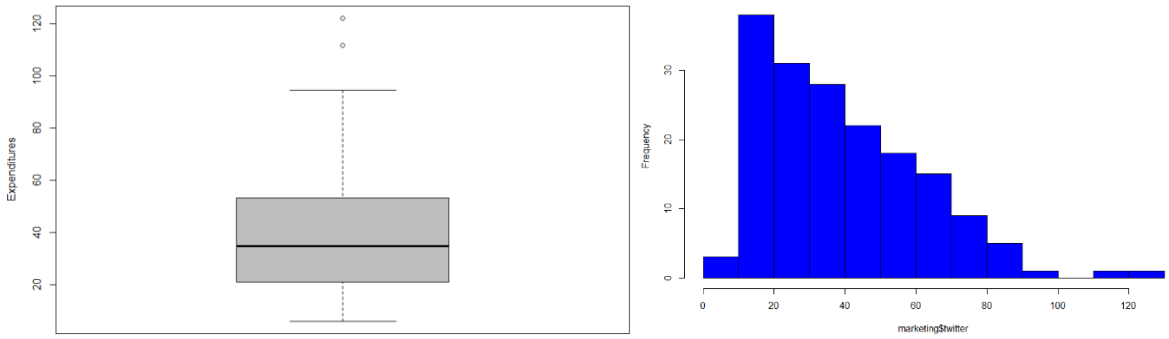
Observa la diferencia entre la mediana y la media. Esta diferencia indica datos sesgados. Dibuja el diagrama de caja y el histograma para ver el sesgo mediante la exploración gráfica:

```
> boxplot(marketing$twitter, ylab = "Expenditures", col = "gray")
> hist(marketing$twitter, main = NULL, col = "blue")
```

Obtendremos el siguiente resultado (ver las figuras en la siguiente página):

¿Qué notas diferente en estas inspecciones gráficas? Algunos puntos de datos en el diagrama de caja aparecen como pequeñas burbujas sobre el bigote superior. Esto significa que el bigote superior representa el valor del RIQ y no el valor máximo.

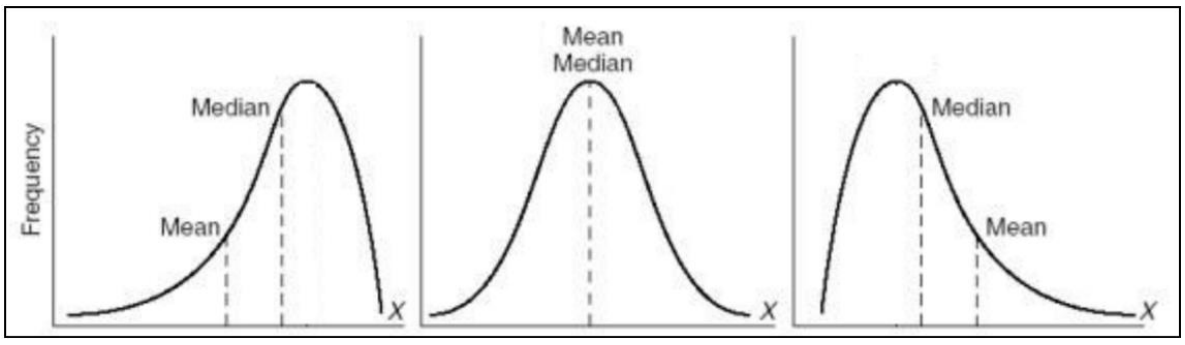
Esta es una indicación visual de los valores atípicos. Las dos burbujas en la parte superior del diagrama de caja aparecen en el histograma como los dos valores en el extremo derecho, más o menos entre 120.



Aprendizaje de gráficos: Verás que se agregó color a los gráficos mediante el parámetro col. Recuerda que no tienes que hacer que tus gráficos exploratorios sean bonitos. Crearás muchos gráficos en esta fase.
Concéntrate en la ganancia de información y profesionalizarás los gráficos finales más adelante. Sin embargo, este documento mostrará algunos parámetros gráficos para que comiences a aprender los aspectos clave de la representación gráfica en R.

El histograma en el panel derecho muestra la naturaleza sesgada hacia la derecha (positiva) de la variable Twitter. La asimetría también es visible en el diagrama de caja. Observa que no hay burbujas en el bigote inferior. Cuando veas esta combinación, puede ser una indicación de asimetría. Sabrás que este es el caso según tus datos tabulares.

Los siguientes paneles (adaptados de Johnson) te muestran la relación entre la media y la mediana en datos normales y sesgados. De izquierda a derecha, los paneles indican datos sesgados hacia la izquierda (negativos), datos distribuidos normalmente y datos sesgados hacia la derecha (positivos). Observa cómo la asimetría tira de la media hacia la izquierda o hacia la derecha y tiene menos efecto en la mediana:



Con una sola variable, eso es prácticamente todo lo que se puede hacer con los datos desde el punto de vista analítico. Si se tienen dos variables, las cosas se vuelven mucho más interesantes.

Análisis de dos variables juntas

Si tu conjunto de datos tiene dos variables, tienes datos bivariados. Al examinar datos bivariados, deseas explorar posibles relaciones. Aquí es donde vale la pena hacer buenas preguntas. Puedes utilizar las siguientes cuatro preguntas para guiar tu análisis exploratorio de datos:

- ¿Cómo se ven los datos?
- ¿Existe alguna relación entre dos variables?
- ¿Existe alguna correlación entre las dos?
- ¿Es significativa la correlación?

Cuatro palabras resumen estas preguntas: Observa-Relaciones-Correlación-Significación. Explorarás pares de variables del conjunto de datos de marketing para investigar estas preguntas utilizando métodos de exploración tanto tabulares como gráficos.

¿Cómo se ven los datos?

Esto es algo que ya sabes muy bien en este momento. Resume los datos en tu consola como se muestra a continuación:

```
> summary(marketing)
```

El siguiente es el resultado:

```
google_adwords      facebook        twitter       marketing_total     revenues       employees      pop_density
Min.   : 23.65   Min.   : 8.00   Min.   :  5.89   Min.   : 53.65   Min.   :30.45   Min.   : 3.000   Low   :68
1st Qu.: 97.25   1st Qu.:19.37   1st Qu.: 20.94   1st Qu.:158.41   1st Qu.:40.33   1st Qu.: 6.000   Medium:52
Median :169.47   Median :33.66   Median : 34.59   Median :245.56   Median :43.99   Median : 8.000   High  :52
Mean   :169.87   Mean   :33.87   Mean   : 38.98   Mean   :242.72   Mean   :44.61   Mean   : 7.866
3rd Qu.:243.10   3rd Qu.:47.80   3rd Qu.: 52.94   3rd Qu.:322.62   3rd Qu.:48.61   3rd Qu.:10.000
Max.   :321.00   Max.   :62.17   Max.   :122.19   Max.   :481.00   Max.   :58.38   Max.   :12.000
```

Agregar y eliminar variables: Es posible que hayas notado una columna adicional en estos datos. Agregamos la variable emp_factor para mostrarte gráficos entre dos variables que son factores. Así es como puedes agregar variables a un frame de datos:

```
> marketing$emp_factor <- cut(marketing$employees, 2)
```

Hay dos elementos en esta línea de código. Mira lo que está pasando aquí:

- La función cut(): Convierte un número en un factor dividiendo los valores en una cantidad de intervalos que elijas. En este caso, pasarás marketing$employees y le dirás a R que lo divida en dos (2) factores. Divide los datos en el punto medio entre

el valor mínimo (3) y el valor máximo (12), lo que da como resultado un factor de 3 a 7 y de 8 a 12 empleados.

- Crea una nueva variable: Puedes agregar nuevas columnas (variables) a un frame de datos asignando el resultado de alguna operación a un nuevo nombre de variable. En este caso, usaste la función cut() y asignaste el resultado a una nueva variable emp_factor usando este nuevo nombre como variable.

La nueva variable solo está ahí temporalmente para mostrarte algunas funciones en esta sección. La eliminarás al final de esta sección usando el siguiente código:

```
> marketing$emp_factor <- NULL
```
Eliminar variables

¿Existe alguna relación entre dos variables?

Para contextualizar, imagina que tienes dos variables: la edad y la altura de un grupo de personas.
Se esperaría una relación entre la edad y la altura. Esto se puede mostrar con exploración tabular usando la función table(). Crea una tabla de los dos factores, emp_factor y pop_density:

```
> table(marketing$emp_factor, marketing$pop_density)
```

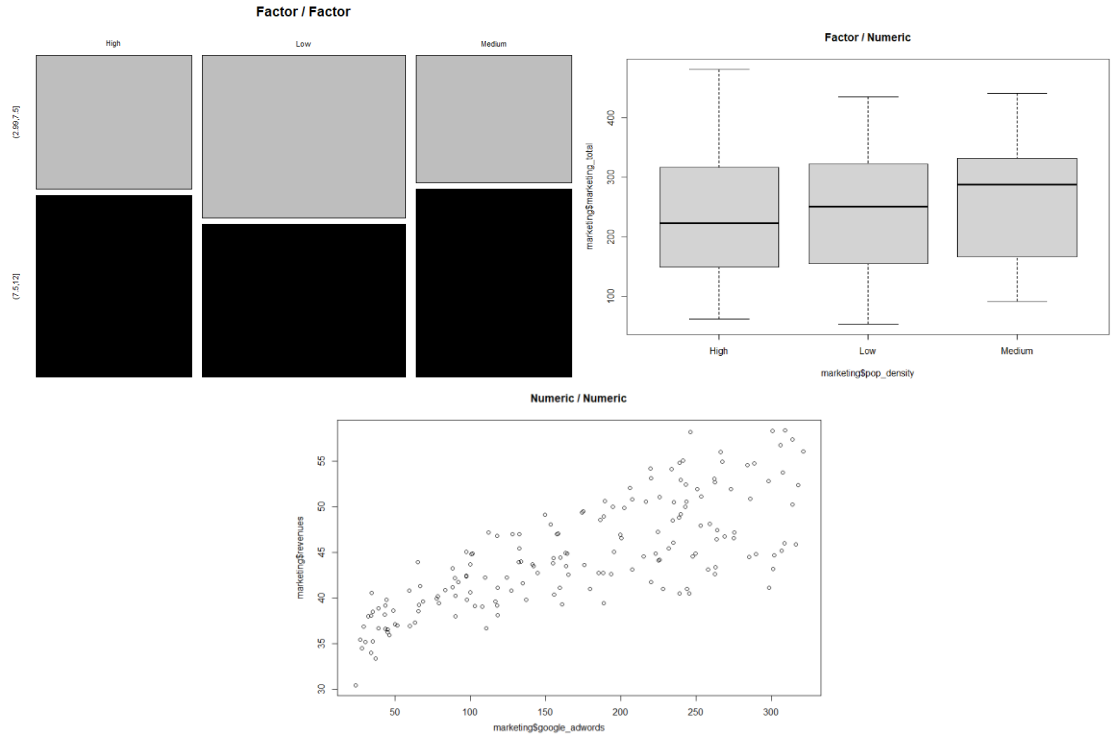
Devolverá el siguiente resultado:

	High	Low	Medium
(2.99,7.5]	22	35	21
(7.5,12]	30	33	31

Esto te da una idea de la relación entre estos dos factores. Parecen equilibrados en ambas dimensiones y tienes una idea de la relación de las variables factoriales entre sí. La función table() funciona muy bien para factores, pero no tan bien para datos numéricos. Pueden llegar a ser bastante grandes. La exploración gráfica es útil en estos casos. Hay tres combinaciones de gráficos al explorar variables bivariadas (se generan de forma separada!):

```
> mosaicplot(table(marketing$pop_density, marketing$emp_factor),
+   col = c("gray","black"), main = "Factor / Factor")
> boxplot(marketing$marketing_total ~ marketing$pop_density,
+   main = "Factor / Numeric")
> plot(marketing$google_adwords, marketing$revenues,
+   main = "Numeric / Numeric")
```

Las tres combinaciones de gráficos se muestran aquí:



Las tres combinaciones se describen a continuación:

- **Factor/Factor**: Trazarás dos factores categóricos utilizando la función mosaicplot(). Los datos primero deben colocarse en un objeto table(). En este caso, graficaste los datos igual que en la tabla anterior de 3 x 2. Ves la conexion?
- **Factor/Numeric**: Utilizarás la función boxplot() que utilizaste anteriormente en este documento para trazar un factor categórico con una variable numérica. Verás en este panel central que el monto total gastado en marketing varía ligeramente de ubicaciones con baja a alta densidad de población. Este gráfico demuestra por qué convertiste este factor en un factor ordenado. Si no hubieras ordenado el factor, se habrían leído alfabéticamente en la parte inferior (Alto, Bajo y Medio). Tu trabajo al ordenar el factor lo hace más legible.
- **Numeric/Numeric**: Se trazan dos variables numéricas en un diagrama de dispersión utilizando la función plot(). Primero se pasa la dimensión *x* a la función seguida de la dimensión *y*, separadas por una coma. Los diagramas de dispersión son útiles al visualizar relaciones.

Estos métodos tabulares y gráficos te dan una idea de las relaciones entre dos variables. Es posible que desees explorar otras variables ahora, especialmente aquellas que tienen sentido

para las preguntas comerciales. Ahora te concentras en algunas relaciones y haces la siguiente pregunta.

¿Existe alguna correlación entre las dos?

En dos dimensiones, puedes responder esta pregunta rápidamente usando la función cor():

```
> cor(marketing$google_adwords, marketing$revenues)
[1] 0.7662461
```

Si utilizamos la siguiente entrada:

```
> cor(marketing$google_adwords, marketing$facebook)
[1] 0.07643216
```

Como ya habrás adivinado, esta función te proporciona la correlación entre dos variables. El resultado numérico tiene dos componentes, signo y valor:

- El **signo** será positivo (+) o negativo (-). Correlación positiva significa que a medida que aumenta la primera variable, también aumenta la segunda variable. Correlación negativa significa que cuando la primera variable aumenta, la segunda variable disminuye. Los dos ejemplos anteriores están correlacionados positivamente.
- El **valor** del resultado de la correlación varía de cero (0) a uno (1). El valor aumenta a medida que aumenta la fuerza de la correlación. El primer ejemplo tiene una correlación de 0.766, que es mucho más fuerte que la del segundo ejemplo (0.076). Un valor de cero indica que no hay correlación entre dos variables.

Con dos variables, una simple exploración tabular puede ser suficiente. Verás los gráficos de estos dos ejemplos más adelante en esta sección. Ten en cuenta estos conceptos mientras pruebas la importancia de las correlaciones que encuentres.

¿Es significativa la correlación?

Puedes encontrar correlaciones fuertes o débiles. Ahora puedes determinar si una correlación es significativa. En un sentido estadístico, significativo significa que su correlación se debe a circunstancias distintas al azar.

En la última pregunta, viste que la correlación entre google_adwords y los ingresos (revenues) era de 0.766. Esto parece fuerte, especialmente en relación con la otra correlación de 0.076.

No se puede determinar la importancia mirando únicamente un número. Hay una prueba estadística incluida en R que te dirá si una correlación es significativa.

Ejecuta la prueba usando la función cor.test() en las variables google_adwords y revenues:

```
> cor.test(marketing$google_adwords, marketing$revenues)

        Pearson's product-moment correlation

data:  marketing$google_adwords and marketing$revenues
t = 15.548, df = 170, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6964662 0.8216704
sample estimates:
      cor
0.7662461
```

Están sucediendo muchas cosas en esta salida de consola. Cuando usas cor.test() en dos variables, realizas una prueba t para examinar algo llamado hipótesis nula.

Obtén más información: si no has aprendido sobre la hipótesis nula y la prueba de hipótesis, deberías considerar aprender sobre ellas para comprender esta técnica. La hipótesis nula está en el centro de muchas técnicas analíticas que tienes disponibles en R. Puede sobtener más información sobre las pruebas de hipótesis en línea en Stat Trek o Khan Academy:

https://stattrek.com/hypothesis-test/hypothesis-testing
https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample/more-significance-testing-videos/v/hypothesis-testing-and-p-values

Una **hipótesis nula** normalmente representa el status quo, lo que significa que no sucede nada. En nuestro ejemplo, la hipótesis nula es que la verdadera correlación entre google_adwords y revenues es igual a cero. En otras palabras, las variables no tienen correlación. La hipótesis alternativa es que la correlación entre ambos no es igual a cero y que la correlación es significativa.

Una **prueba t** aborda la pregunta: ¿qué tan sorprendente es ver este grado de correlación si las dos variables realmente no están correlacionadas? Entrar en detalles de la prueba t está fuera del alcance de este curso, pero el resultado de la prueba t es 15.548. ¿Qué tan sorprendente es este resultado? El valor p asociado con este valor de t es 2.2e-16. Básicamente, esto dice que, si la hipótesis nula fuera cierta, la probabilidad de obtener t = 15.548 sería 0.00000000000000022.

En otras palabras, la probabilidad de obtener este resultado, si la hipótesis nula fuera cierta, sería esencialmente cero. Por lo tanto, rechazarás la hipótesis nula afirmando que la correlación es cero. Un valor típico utilizado para rechazar una hipótesis nula es un valor p inferior a 0.05.

Consejo de BI: Mantén todas estas estadísticas en perspectiva. Tu objetivo no es convertirte en estadístico. Más bien, tu objetivo es volverte más consciente de lo que significa tu análisis y cuáles son sus limitaciones. Las decisiones comerciales no son lo mismo que los ensayos clínicos de medicamentos, pero deseas saber cuándo puedes tener fe en el análisis que proporciona a los líderes empresariales que informa sus decisiones.

Al volver a ponerse el sombrero de los negocios, decides determinar si los otros dos canales de marketing (Twitter y Facebook) están correlacionados y son significativos. Ejecuta cor.test() en estas variables contra los revenues. El siguiente es un resumen de los resultados de la prueba:

```
> cor.test(marketing$twitter, marketing$revenues)
> cor.test(marketing$facebook, marketing$revenues)
```

Relaciones	twitter and revenues	facebook and revenues
t	3.6516	9.2308
p-value	0.0003467	2.2e-16
correlación	0.2696854	0.5778213
significativa?	Si	Si

Estos resultados muestran que ambos pares están correlacionados positivamente y las correlaciones son significativas. Parece que Bike Sharing LLC ha encontrado buenos canales de marketing vinculados a los ingresos. ¿Eso significa que la publicidad en estos canales hace que los ingresos aparezcan en el resultado final? No necesariamente.

Existe una percepción errónea común de que la correlación implica causalidad. Aquí hay una mirada divertida a dos variables altamente correlacionadas:
Checar más correlaciones en: https://tylervigen.com/spurious-correlations

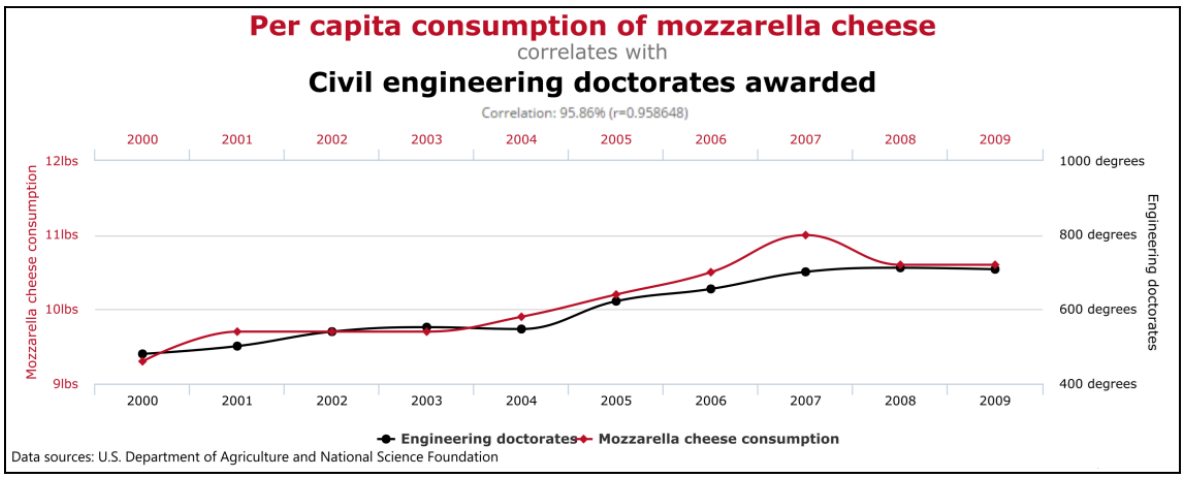
¡Guau! Una correlación de 0.959. Esto te genera curiosidad e ingresa los datos en R para un análisis rápido. Tienes las habilidades para ingresar estos datos y verificarlos por ti mismo:

```
> cheese <- c(9.3, 9.7, 9.7, 9.7, 9.9, 10.2, 10.5, 11, 10.6, 10.6)
> degrees <- c(480, 501, 540, 552, 547, 622, 655, 701, 712, 708)
> cor(cheese, degrees)
[1] 0.9586478
```

```
> cor.test(cheese, degrees)

Pearson's product-moment correlation

data:  cheese and degrees
t = 9.5274, df = 8, p-value = 1.217e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8300027 0.9904491
sample estimates:
      cor
0.9586478
```



Esto verifica una correlación de 0.959, así como t = 9.5274 y un valor p de 1.217e-05.

¿Mmm? ¡Esta correlación es fuerte y significativa! Sin embargo, sabes que correlación no implica causalidad. De hecho, este gráfico (y muchos otros similares) aparecen en el sitio web Spurious Correlations para ilustrar este punto de una manera divertida:

"Es una condición del espíritu humano encontrar la causa de un fenómeno para poder explicarlo y responder preguntas. La búsqueda es noble si el viaje es sólido".

–Jay Gendron

Hay casos en los que la correlación no es significativa. Al ejecutar cor.test() en google_adwords y facebook se obtiene un resultado que no muestra importancia:

```
> cor.test(marketing$google_adwords, marketing$facebook)

        Pearson's product-moment correlation
```

```
data:  marketing$google_adwords and marketing$facebook
t = 0.99948, df = 170, p-value = 0.319
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.07404915  0.22351032
sample estimates:
      cor
0.07643216
```

Una última prueba de correlación antes de explorar gráficamente algunos de estos resultados. Tu sentido empresarial se pregunta si existe alguna correlación significativa entre los gastos de marketing (marketing_total) y los revenues (ingresos) asociados con el marketing. Esta es una buena pregunta comercial que puede anticipar del grupo de marketing.

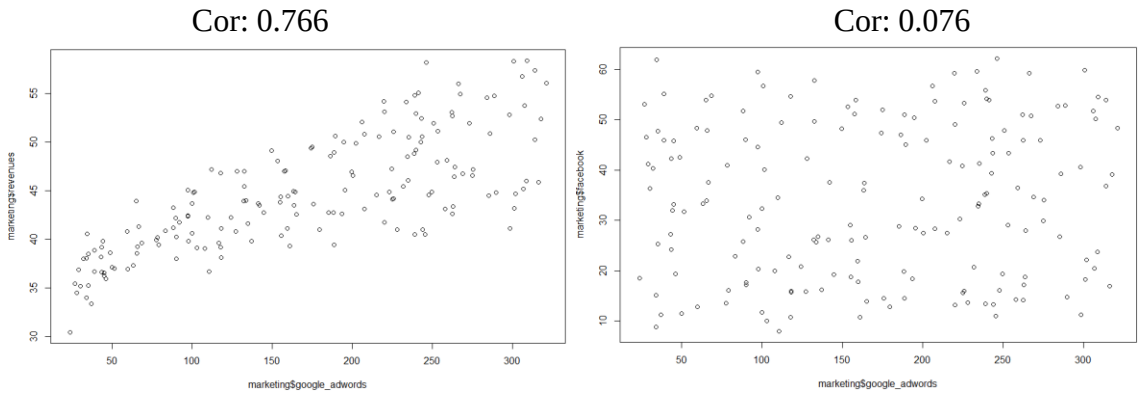
Ejecuta cor.test(marketing$revenues, marketing$marketing_total) y los resultados parecen prometedores: correlación = 0.853, t = 21.313 y valor de p = 2.2e-16.

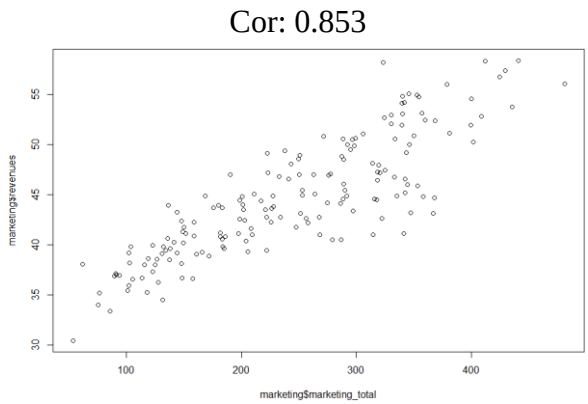
El intervalo de confianza es muy ajustado (0.806, 0.89), lo que indica poca variación en los datos. Un resultado como éste sugiere que vale la pena modelar esta relación en análisis posteriores. Generarás un modelo de regresión lineal sobre este par de variables más adelante.

Hasta este punto, has estado explorando la correlación utilizando resultados tabulares. También hay un papel para la exploración gráfica. El siguiente es el código para trazar tres correlaciones bivariadas que ya has visto en forma tabular:

```
> plot(marketing$google_adwords, marketing$revenues)
> plot(marketing$google_adwords, marketing$facebook)
> plot(marketing$marketing_total, marketing$revenues
```

El resultado se muestra en la siguiente figura:





Los plots refuerzan el significado detrás de los valores de correlación. Visualmente, puedes ver una correlación relativamente fuerte en el panel izquierdo entre google_adwords y los ingresos. La correlación es aún más fuerte en el panel central abajo entre ingresos y marketing_total.
Sabes que estas correlaciones también son significativas.

El panel a la derecha arriba muestra una distribución aleatoria de observaciones sin un patrón discernible.

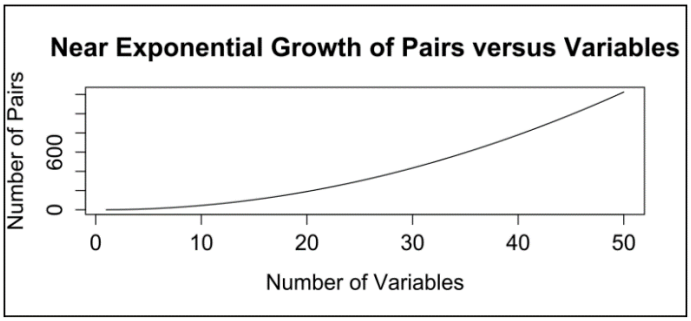
Esto indica una relación muy débil con una correlación pobre e insignificante entre las variables google_adwords y facebook.

Esto finaliza el análisis de dos variables. Buen trabajo. Era mucho material. Recuerda eliminar la variable temporal emp_factor ejecutando la siguiente línea de código presentada anteriormente:

```
> marketing$emp_factor <- NULL
```

Explorando múltiples variables simultáneamente

Está bien. Has llegado a la última sección del análisis exploratorio de datos. Ahora ampliarás tu exploración a múltiples variables a la vez. Los conjuntos de datos típicos tienen muchas variables, pero un análisis bivariado lo limita a comparaciones por pares. Explorar cinco variables, dos a la vez crea 10 pares, 10 variables crean 45, 20 variables crean 190, 40 variables crean 780, y así sucesivamente. El impacto en el flujo de trabajo es casi exponencial, como se muestra en el siguiente diagrama:



A medida que crece la cantidad de características (variables) en tu conjunto de datos, tu estrategia para el análisis de datos exploratorios debe escalar junto con tus datos. Tu conocimiento del análisis de datos exploratorios bivariados te proporciona los dos beneficios siguientes:

- Tienes las bases para explorar múltiples variables simultáneamente.
- Puedes utilizar el análisis bivariado para explorar más a fondo cualquier par interesante.

Seguirá utilizando el enfoque de cuatro preguntas de Mirada-Relaciones-Correlación-Importancia.

Observa

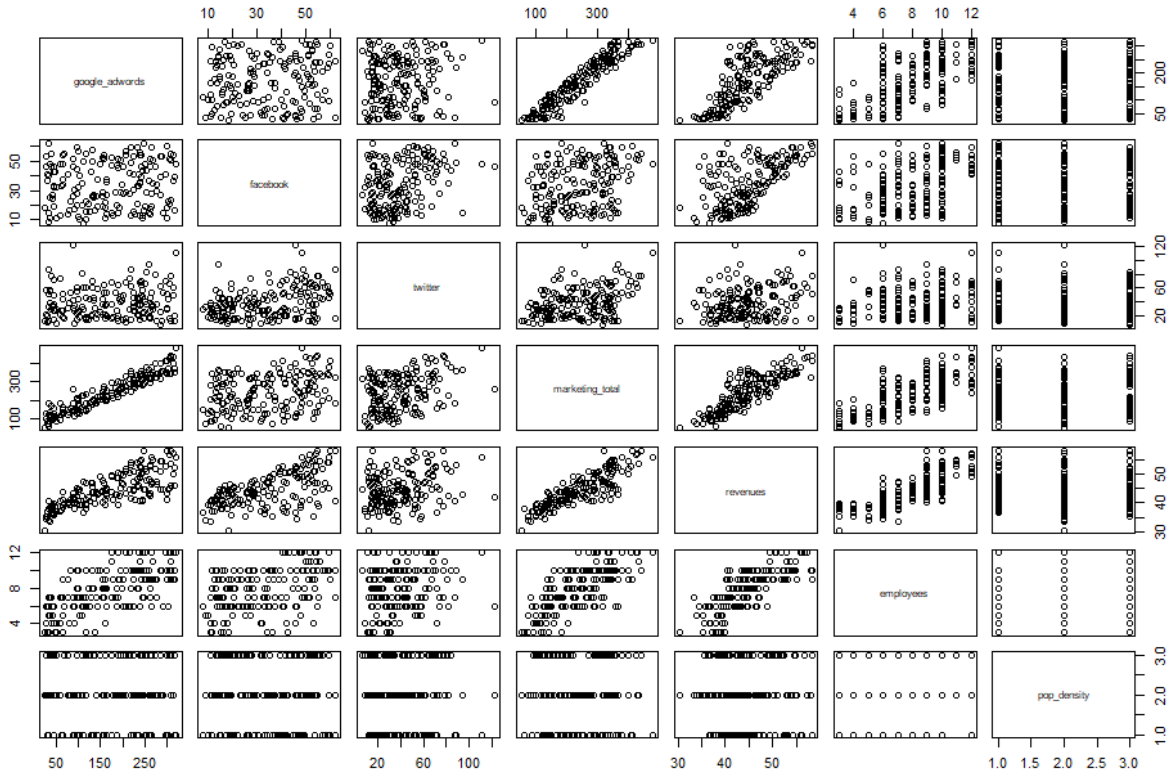
La primera pregunta es, ¿cómo son los datos? Utilizarás el mismo conjunto de datos de marketing para esta sección, pero esta vez explorarás todo el conjunto de datos. Mirar los datos se logra fácilmente usando la función summary(). Continúa y mira el conjunto de datos nuevamente para refrescar tu memoria de las 7 variables y 172 observaciones.

Relaciones

A continuación, deseas saber: ¿existe alguna relación entre las variables? En la última sección, aprendiste que la mejor manera de hacerlo es mediante exploración gráfica. Espera

un minuto... acabamos de decir que trazar variables en pares es muy ineficiente para muchas características. Esto es cierto. ¿Y si hubiera una manera de trazar todos los pares a la vez?

Continúa y escribe pairs(marketing) en tu consola. ¡Asombroso! Obtendremos el resultado de la siguiente manera:



Consejo R: Has creado gráficos de pares para 37 variables y es sorprendente lo que se ve en las miniaturas. Obtienes mucha información en muy poco código.

El resultado de pairs(marketing) es una cuadrícula de gráficos por pares que muestran la relación de cada variable con las demás. La diagonal de la cuadrícula te ayuda a navegar por el gráfico e identificar los pares que ves. Aquí hay dos ejemplos para mostrarle la navegación:

- Ingresos versus Google AdWords: Coloca el dedo en el panel google_adwords en la parte superior izquierda. Mueve el dedo cuatro paneles hacia abajo hasta la fila que incluye los ingresos. Esta es la misma trama que creaste en la última sección y muestra una fuerte relación positiva.
- Facebook versus Google AdWords: Comienza con el panel de Facebook y mueve un panel hacia la izquierda hasta la columna google_adwords. Esta es la misma trama que creaste en la última sección y casi no muestra ninguna relación.

Ahora tienes una capacidad de visualización poderosa y rápida para encontrar relaciones.

Correlación

Para saber si existe alguna correlación entre las variables, puedes utilizar la función cor() en el conjunto de datos. A diferencia de la función pairs(), solo puedes encontrar correlaciones entre variables numéricas. Subconjunto de marketing en las primeras seis variables (numéricas) y páselo a cor():

```
> cor(marketing[,1:6])
```

El resultado es como se muestra aquí:

```
                google_adwords   facebook   twitter marketing_total  revenues employees
google_adwords      1.00000000 0.07643216 0.0989750       0.9473566 0.7662461 0.6610312
facebook            0.07643216 1.00000000 0.3543410       0.3102232 0.5778213 0.4101966
twitter             0.09897500 0.35434096 1.0000000       0.3758691 0.2696854 0.2290618
marketing_total     0.94735659 0.31022316 0.3758691       1.0000000 0.8530354 0.7210171
revenues            0.76624608 0.57782131 0.2696854       0.8530354 1.0000000 0.7656857
employees           0.66103123 0.41019661 0.2290618       0.7210171 0.7656857 1.0000000
```

Esta función le brinda información sobre todas las correlaciones por pares a la vez. Al observar el resultado de la consola anterior, ¿puedes encontrar las dos correlaciones que calculaste como pairs? Por supuesto que puedes. Los siguientes son los resultados de tus cálculos por pares anteriores:

```
> cor(marketing$google_adwords, marketing$revenues)
[1] 0.7662461
```

El segundo par se muestra en el siguiente código:

```
> cor(marketing$google_adwords, marketing$facebook)
[1] 0.07643216
```

Significación

El análisis de datos exploratorio avanza un poco más rápido cuando se exploran múltiples variables.
También verás que los fundamentos aprendidos en la última sección te permiten encontrar características y relaciones de interés rápidamente. La última pregunta que tienes para este conjunto de datos antes de comenzar a modelar es: ¿existen correlaciones significativas en los datos?

Llegados a este punto, es posible que no te sorprenda saber que existe una manera de ejecutar todas las pruebas de correlación y significancia a la vez. El paquete psych proporciona una función corr.test() que combina la salida de cor() junto con los valores p. Revisa la sección ¿Es significativa la correlación? si necesitas solidificar alguno de estos conceptos:

```
> install.packages("psych")
> library(psych)
> corr.test (marketing[ ,1:6])
```

El resultado es el siguiente:

```
Call:corr.test(x = marketing[, 1:6])
Correlation matrix
                google_adwords facebook twitter marketing_total revenues employees
google_adwords            1.00     0.08    0.10            0.95     0.77      0.66
facebook                  0.08     1.00    0.35            0.31     0.58      0.41
twitter                   0.10     0.35    1.00            0.38     0.27      0.23
marketing_total           0.95     0.31    0.38            1.00     0.85      0.72
revenues                  0.77     0.58    0.27            0.85     1.00      0.77
employees                 0.66     0.41    0.23            0.72     0.77      1.00
Sample Size
[1] 172
Probability values (Entries above the diagonal are adjusted for multiple tests.)
                google_adwords facebook twitter marketing_total revenues employees
google_adwords            0.00     0.39    0.39               0        0      0.00
facebook                  0.32     0.00    0.00               0        0      0.00
twitter                   0.20     0.00    0.00               0        0      0.01
marketing_total           0.00     0.00    0.00               0        0      0.00
revenues                  0.00     0.00    0.00               0        0      0.00
employees                 0.00     0.00    0.00               0        0      0.00

 To see confidence intervals of the correlations, print with the short=FALSE option
```

Puedes ver que todas las correlaciones en la matriz de correlación son las mismas que los resultados cuando llamaste a cor() en el conjunto de datos de marketing. Además, la función corr.test() proporciona los valores de probabilidad (valores p) para todos los pares. Al realizar el análisis bivariado, encontraste resultados para algunas relaciones. Estos se comparan bien con el resultado tabular anterior:

Relación	Correlación	p-value	Significación
adwords and revenues	0.76625	0.00	Si
facebook and revenues	0.57782	0.00	Si
twitter and revenues	0.26968	0.00	Si
adwords and facebook	0.07643	0.319	No
revenue and marketing_total	0.853	0.00	Si

También puedes obtener información sobre todo el conjunto de datos con la exploración gráfica. El paquete corrgram contiene una función corrgram(), que es una versión mejorada del gráfico de pares y también incorpora elementos de la función corr.test().

Aquí está el código que produce un 'correlograma'. Utiliza un subconjunto del conjunto de datos de marketing porque las matemáticas sólo permiten valores numéricos. El orden se establece en FALSE, conservando la posición de las variables tal como se encuentran en el conjunto de datos:

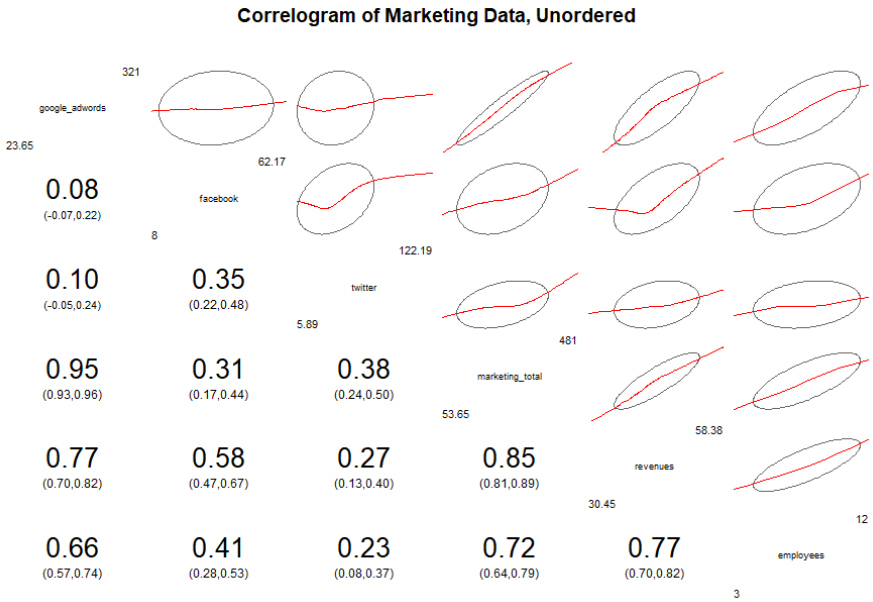
```
> install.packages("corrgram")
> library(corrgram)
> corrgram(marketing[ ,1:6], order = FALSE,
+   main = "Correlogram of Marketing Data, Unordered",
+   lower.panel = panel.conf, upper.panel = panel.ellipse,
+   diag.panel = panel.minmax, text.panel = panel.txt)
```

Cuatro parámetros aparecen en el código. Controlan el contenido de las regiones del correlograma. Estos parámetros en este ejemplo se explican a continuación:

- **Panel inferior (lower panel)**: La mitad inferior izquierda del gráfico se configuró para mostrar coeficientes de correlación e intervalos de confianza usando panel.conf.
- **Panel superior (upper panel)**: La mitad superior derecha del gráfico se configuró para mostrar elipses y líneas suaves usando panel.ellipse
- **Diagonal y Texto (diagonal and text)**: La diagonal contiene el nombre de la variable y sus valores mínimo y máximo usando panel.minmax y panel.txt, respectivamente

Puedes obtener más información sobre estos parámetros y otras opciones escribiendo ?corrgram() en la consola y viendo la ayuda disponible en el entorno R.

Esto es lo que genera la función corrgram():



Este simple gráfico incorpora mucha información que has visto con otras funciones. La siguiente es una explicación de tres características clave del correlograma:

- **Variables**: Las variables aparecen a lo largo de la diagonal con los valores mínimo y máximo de cada una.

- **Elipses y línea**: El panel superior (arriba a la derecha) se codificó para mostrar elipses de confianza de cada par de variables. La elipse captura la forma esencial del diagrama de dispersión de correlación utilizando un intervalo de confianza, así como una representación lineal suave de la relación entre las dos variables:
  - Las elipses delgadas representan una fuerte correlación.
  - Las elipses más redondas representan correlaciones débiles
  - La pendiente de la línea suave representa una correlación positiva o negativa.

- **Coeficiente de correlación e intervalo de confianza**: El panel inferior (abajo a la izquierda) se codificó para mostrar el coeficiente de correlación y el intervalo de confianza. Esto es similar a los resultados que obtienes de la función cor() o corr.test().

Si deseas adaptar las imágenes a tu audiencia, puedes ajustar los parámetros. Este orden de tiempo se establece en TRUE.

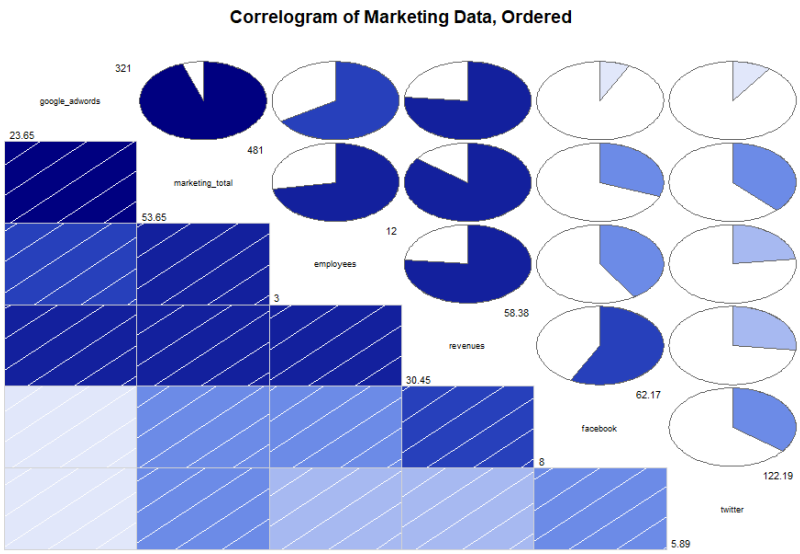
Esto reordena las variables para presentar las correlaciones más fuertes hacia la parte superior izquierda, con una correlación que se debilita gradualmente hacia la parte inferior derecha.

Diferentes parámetros enviados a la función dan como resultado un gráfico diferente, como se muestra a continuación:

```
> corrgram(marketing[ ,1:6], order = TRUE,
+   main = "Correlogram of Marketing Data, Ordered",
+   lower.panel = panel.shade, upper.panel = panel.pie,
+   diag.panel = panel.minmax, text.panel = panel.txt)
```

El resultado se muestra en la siguiente figura:

Práctica adicional: Regresa y explora el conjunto de datos que escribiste en un archivo en el Documento 2, Limpieza de datos, y responde las siguientes preguntas. Mientras lees estos datos de un archivo, se perderán los factores ordenados y el formato de fecha. Ya creaste el código en el último documento. Puedes usarlo nuevamente para convertir variables en los tipos de datos correctos:




```
> bike<-read.csv("clean_bike_sharing_data.csv", stringsAsFactors = TRUE)
> bike$season <- factor(bike$season, ordered = TRUE,
+   levels = c("spring","summer",
+   "fall","winter"))
> bike$weather <- factor(bike$weather, ordered = TRUE,
+   levels = c("clr_part_cloud",
+   "mist_cloudy",
+   "lt_rain_snow",
+   "hvy_rain_snow"))
> library(lubridate)
> bike$datetime <- ymd_hms(bike$datetime)
Warning message:
```

Preguntas: Utiliza tus nuevas habilidades para explorar los datos y responder estas preguntas:

¿Pudiste cargar los datos y reconvertir los tipos de datos? Solo el ultimo no
¿Cuántas variables hay en el conjunto de datos? ¿Cuántas observaciones? 17379 obs. of 13 variables:
¿Cuántas observaciones hay para cada estación?
¿Cuál es la media y la desviación estándar de la variable temporal?
¿Qué variable se distribuye casi normalmente?
¿Qué variable está más sesgada hacia la izquierda y más hacia la derecha?
¿Qué par de variables tiene la mayor correlación positiva y la mayor correlación negativa?
¿Qué variable(s) muestra una correlación significativa?
Respuestas: Las soluciones para estos ejercicios de práctica las deberás realizar por tu cuenta como parte de una tarea para casa (se responden algunas preguntas que deberás comprobar con tu trabajo):

- Hay 13 variables y 17.379 observaciones:
  primavera verano Otoño Invierno
  4242 4409 4496 4232
- Para la variable temp, media = 20.37647 y desviación estándar = 7.894801.
- La variable temp es la que más se distribuye normalmente.
- La variable workingday es la más sesgada a la izquierda y casual es la más sesgada a la derecha.
- La mayor correlación positiva es temp y atemp (correlación = 0.99) y la mayor correlación negativa es humedad y casual (correlación = -0.37).
- Todas las variables muestran correlación significativa. Esto se debe en parte al gran tamaño de la muestra.

Resumen

Bien hecho, aventurero de los datos. Exploraste un conjunto de datos bastante a fondo. El análisis de datos exploratorio es el paso de análisis preliminar, ya que te brinda información sobre las características o relaciones que son mejores para el modelado. Aprendiste que el análisis exploratorio de datos tiene una estructura: no es una serie de gráficos aleatorios. La estructura comienza teniendo preguntas generales en mente y luego familiarizándose con los datos abordando cuatro preguntas estructuradas: Observa-Relaciones-Correlación-Significación.

Observa te proporciona una idea general de lo que hay en los datos, qué tipo de escalas y tipos de datos se utilizan, y la verificación de que los datos son adecuados. Esta pregunta se relaciona principalmente con la salida tabular. Pueden existir relaciones entre variables y la exploración gráfica es muy adecuada para detectarlas. Las correlaciones son formas de describir relaciones numéricamente, utilizando el signo y el valor de los coeficientes de correlación. Hay muchas formas de calcularlos y mostrarlos.

La significancia ayuda a determinar si las correlaciones visibles en la salida tabular o gráfica también son significativas y no se deben al azar.

Ahora que tienes una idea de algunas posibles características influyentes, puedes aprovechar este conocimiento y crear mejores modelos para predecir los resultados. Eso es lo que harás en el próximo documento utilizando el conjunto de datos de marketing.