PUCE Virtual

Sistemas de Big Data

Exploración y preparación de datos

Clase 7

MAESTRÍA EN
SISTEMAS DE INFORMACIÓN
Mención Data Science

La excelencia no se improvisa

1. INTRODUCCIÓN DE LA CLASE

La **exploración y la preparación de datos** son etapas fundamentales en el proceso de análisis de datos. Antes de aplicar cualquier algoritmo o modelo, es crucial comprender la naturaleza del **conjunto de datos** y **prepararlo** adecuadamente para su posterior análisis. Durante la exploración de datos, se realizan tareas como inspección de la calidad de los datos, **identificación** de **valores faltantes** o **atípicos** y comprensión de la distribución y la relación entre las **variables**. Esta fase permite a los analistas identificar patrones, anomalías y errores que pueden afectar los **resultados** de los **análisis** (Kuhn & Johnson, 2013). La **preparación** de **datos**, por otro lado, implica transformar los datos brutos en un formato **adecuado** para su análisis. Esto puede incluir la normalización o estandarización de variables, la codificación de variables categóricas y la creación de nuevas variables que puedan ser más relevantes para los modelos (Han, Kamber & Pei, 2012).

El proceso de preparación de datos también abarca la **limpieza** y **reducción** de **ruido**, lo cual es esencial para garantizar que los modelos de análisis de datos sean precisos y confiables. El **preprocesamiento de datos** es una etapa crítica para evitar que los resultados estén sesgados, debido a datos sucios o mal estructurados. Las técnicas utilizadas incluyen la imputación de datos faltantes, la detección y corrección de errores en los datos y la **transformación** de las **variables** para mejorar su interpretabilidad y escalabilidad (Chawla & Davis, 2013). Sin una exploración y preparación adecuadas, incluso los modelos más sofisticados pueden producir resultados inexactos o engañosos, lo que resalta la importancia de estas etapas en el flujo de trabajo de **análisis** de **datos**.

RDA 3: Evaluar los fundamentos del Big Data para el análisis de grandes volúmenes de datos

Clase 7. Exploración y preparación de datos

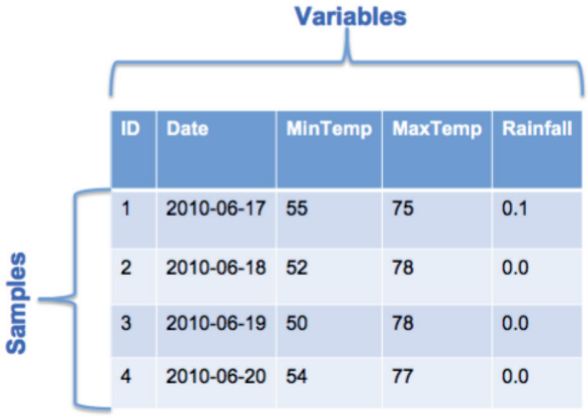
7.1. Exploración de datos

7.1.1. Terminología de los datos

Los **objetivos** de esta sección son los siguientes:

- **Entender** qué es una ***feature*** y su relación en la instancia o registro
- **Nombrar** algunos términos alternativos para ***feature***
- **Sintetizar** las diferencias entre las **variables categóricas y numéricas**

Una **instancia** o **registro** es un ejemplo de una entidad en tus datos. Esto es típicamente una fila en tu ***dataset***. La Figura 1 muestra una parte de un ***dataset*** de valores relacionados con el clima. Cada fila es una instancia que representa los datos del clima para un día en particular. Se muestran cuatro instancias del clima, cada una para un día diferente.



ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Figura 1. *Ejemplo de dataset.*

En esta figura, cada **instancia** tiene cinco valores asociados a ella. Estos valores son diferentes piezas de información sobre la instancia, tales como el **ID** de la instancia, la **fecha**, **temperatura mínima**, **temperatura máxima y probabilidad de lluvia de ese día**. Nosotros llamaremos a estos diferentes valores **variables de la instancia.**

Hay muchos nombres para referirse a **instancia** y a **variable. A**lgunos de los nombres para instancia son: **registro, ejemplo, fila y observación**. Es importante reconocer que todos estos términos hacen referencia a los mismo en ***machine learning;*** esto es, que todos se refieren a un ejemplo específico de una entidad en tu ***dataset***. También hay varios términos para referirse a variable, tales como feature, columna, dimensión, atributo y campo. Todos esos términos se refieren a características específicas de una fila en tu *dataset*.

Un punto importante para enfatizar en **relación** con las variables, es que tienen un tipo de dato asociado a ellas. Los **tipos** de **datos** más comunes son numéricos y categóricos. Hay otros tipos de datos también, tales como **STRING** y **DATE,** pero nosotros nos enfocaremos en dos de los tipos de datos más comunes, los numéricos y los categóricos.

Como el nombre lo indica, las variables numéricas son las que toman un valor numérico. Las variables **numéricas** pueden ser medidas y sus valores pueden ser ordenados de alguna manera. **Nótese** que las variables numéricas pueden tener valores enteros o valores continuos. También pueden tener números positivos, números negativos o ambos.

Algunos ejemplos de variables numéricas:

- **Altura**
- **Calificación** en un examen
- **Número** de transacciones por hora
- **Cambio** de precio en el stock de un producto (que puede ser positivo o negativo)

Una variable con ***labels***, nombres o categorías son llamadas variables categóricas. Por ejemplo, una variable puede describir el color de un ítem: el color de un auto puede tener valores tales como rojo, azul, blanco, negro. Estos son valores no numéricos, que describen una **cualidad** o **característica** de una entidad.

Estos valores pueden ser pensados como nombres o *labels* que pueden ser guardados en categorías. Además, las variables categóricas también son referidas como variables cualitativas o variables nominales. Algunos ejemplos de variables categóricas:

- **Género**
- **Estado civil**
- **Tipo** de cliente
- **Categorías** de producto
- **Color** de un ítem

7.1.2. Introducción a la exploración de datos

Los **objetivos** de esta sección son:

- **Explicar** por qué es necesaria la exploración de datos
- **Articular** los objetivos de la exploración de datos
- **Listar** las categorías y técnicas para la exploración de datos

La **exploración** de datos implica realizar una exploración preliminar del ***dataset***. El **objetivo** es lograr un mejor entendimiento de los datos, con los cuales vamos a trabajar. Si entiendes las características de tus datos, puedes usarlas de la manera más optimizada posible, en cualquier tipo de **procesamiento** y **análisis** subsecuentes. A la **exploración** de datos también se la conoce como *Exploratory Data Analysis*

(EDA).

Formas de explorar los datos:

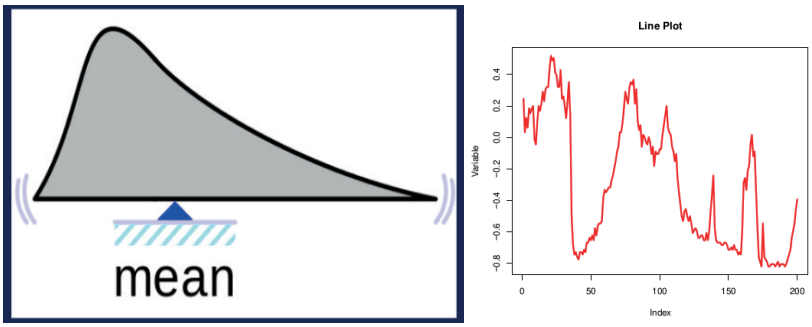


Figura 2. *Formas de explorar los datos.*

Hay principalmente dos maneras de explorar los datos, la una está basada en ***summary statistics*** y la otra en métodos de **visualización (*Data Visualization*)** .

Summary Statistics. Brinda información importante que resume lo que tiene un conjunto de datos. Hay varias métricas estadísticas tale como:

- Promedio (*mean*)
- Mediana
- Desviación estándar

Una *summary statistic* brinda un valor único que resumen un aspecto del *dataset*. Por ejemplo, el *mean* es un valor que describe el promedio de valores de un *dataset* sin importar cuan grande sea. Se puede pensar en el **promedio** como un indicador de esta **centralmente** ubicado tu *dataset* en una línea numérica.

Data Visualization. Permite dar un **vistazo** a tus datos de manera gráfica. Hay varios tipos de **gráficos** que puedes usar para visualizar datos. Por ejemplo, tenemos histograma, *line plot* y gráfico de dispersión. Cada tipo de gráfico tiene un propósito distinto, dependiendo de lo que se esté buscando. Algunas de las cosas que se pueden buscar son:

- *Correlaciones*
- *Tendencias*
- *Outliers*

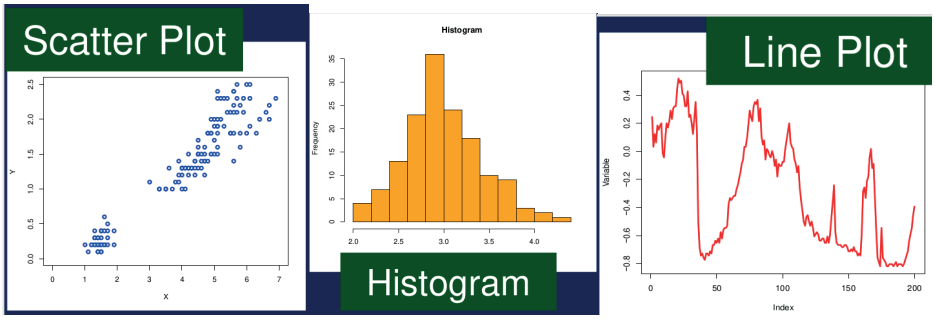


Figura 3. *Visualización de datos.*

Las **correlaciones** brindan información acerca de la relación entre las variables de tu *dataset;* analizando correlaciones, tú podrías ser capaz de determinar si dos variables se encuentran altamente **correlacionadas**, esto implica que estas variables brindan la misma información o información parecida en tu ***dataset***. Dado que esto también implica que existe información redundante, en estos casos se recomienda eliminar una de las variables para hacer el análisis más **simple**.

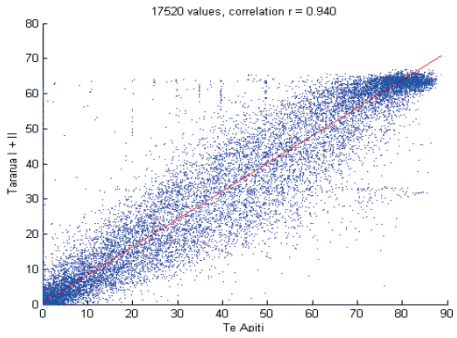


Figura 4. *Correlaciones.*

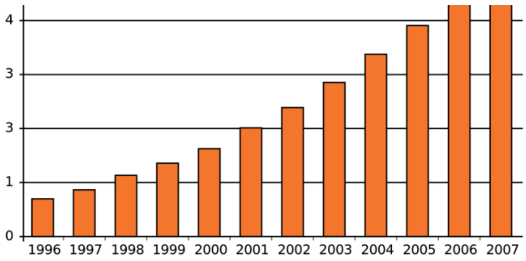


Figura 5. *Tendencias.*

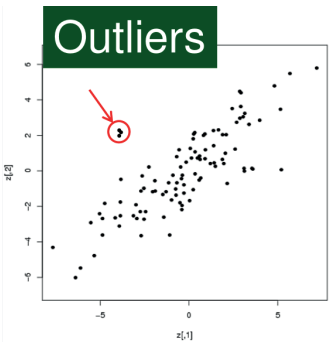


Figura 6. *Outliers.*

Las tendencias en el *dataset* revelan características en tus datos, por ejemplo, puedes ver dónde se encuentra la mayor cantidad de tus datos, dónde hay pocos datos, cuál es el valor o valores más **frecuentes** en tu *dataset*.

Mediante este análisis, también se puede llegar a la conclusión de que la **variable** se está moviendo a una dirección específica; por ejemplo, el crecimiento o decrecimiento de los ingresos por ventas a través de los **años**.

Calcular el mínimo, máximo y el rango de los datos son pasos básicos en el análisis **exploratorio** de datos. Encontrar ***outliers*** es también muy importante. Los *outliers* indican problemas potenciales en tus datos y probablemente tengan que ser eliminados en algunas aplicaciones. En otras aplicaciones, los *outliers* representan puntos importantes que necesitaran ser analizados con mayor detenimiento. En esos casos, los *outliers* usualmente requieren un análisis más exhaustivo.

7.1.3. Exploración de datos a través de gráficos

Los objetivos de esta sección son:

- **Entender** la utilidad de los gráficos en la exploración de datos
- **Describir** cómo se puede usar un gráfico de dispersión
- **Sumarizar** que es lo que muestra un *boxplot*

La visualización de datos significa mostrar los datos gráficamente, lo cual representa una excelente forma de explorar tu *dataset*. Visualización de datos es un gran complemento para los ***summary statistics*** en el momento de explorar los datos.

Existen varios tipos de gráficos que se pueden usar para visualizar los datos:

- Histogramas
- Gráficos de líneas
- Gráficos de dispersión
- Gráficos de barras
- Boxplots, entre otros.

Estos son los gráficos más comúnmente **usados**, pero existen mucho otros más también.

Introducción al análisis exploratorio con Python [https://datos.gob.es/es/documentacion/guia-practica-de-introduccion-al-analisis-exploratorio-de-datos-en-python]

Histogramas. Se usa para mostrar la distribución de una variable. El rango de valores para la variable es dividido en el número de bines, y se cuenta la cantidad de valores que caen en cada uno de los bines, mismo que determina la altura de cada *bin*. Un histograma puede revelar muchas cosas acerca de una variable en tu *dataset*.

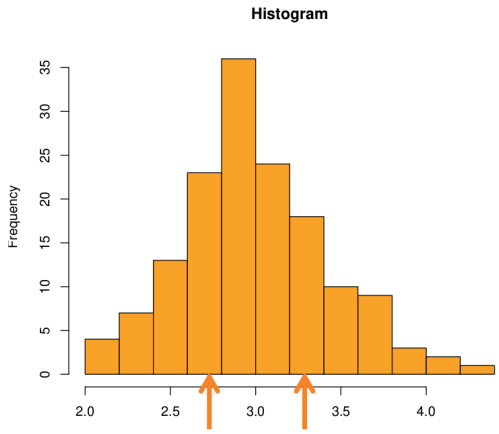


Figura 7. *Histograma.*

Por ejemplo, se puede determinar si los datos siguen una **distribución normal** o si están concentrados en **rangos bajos** o en **rangos altos**. También se puede ver el valor más frecuente que tiene la variable. También es posible determinar la existencia de ***ouliters.***

Figura 8. *Distribuciones con histograma.*

Gráfico de líneas. Muestra cómo cambian los datos a través del tiempo. Los valores de la variable se muestran en el eje Y; el eje X muestra alguna **unidad** de **tiempo**. La línea resultante muestra los valores distribuidos en función del tiempo. Un gráfico de líneas puede mostrar patrones en las variables.

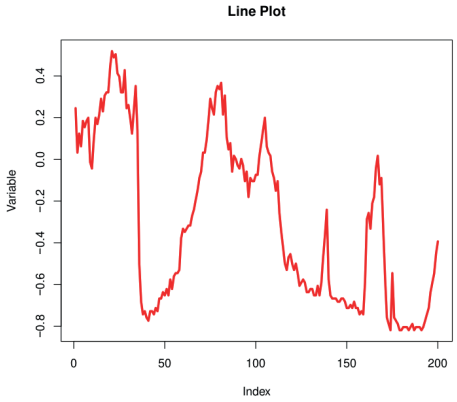


Figura 9. *Gráfico de líneas.*

Por ejemplo, un patrón cíclico puede ser detectado, como en la Figura 9, en donde los valores empiezan siendo altos y luego decrecen, pero finalmente vuelven a ser valores altos. Estos gráficos también pueden ayudar a mostrar tendencias, como en la Figura 9, en donde los valores muestran una **fluctuación**, pero en general tienen una tendencia creciente a lo largo del tiempo. También es fácil comparar como múltiples variables pueden cambiar a través del tiempo en un **mismo gráfico**.

Gráfico de dispersión. El gráfico de dispersión es una excelente forma de visualizar **relaciones** entre dos variables. Una variable en el eje X y otra variable en el eje Y. Cada punto es un producto y representa los valores que tiene una instancia para las dos variables. El gráfico resultante muestra cómo una variable **cambia** en **función** de la otra. Los gráficos de dispersión son muy útiles porque muestran la correlación entre dos variables.

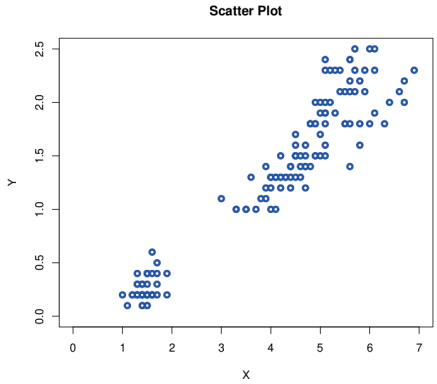


Figura 10. *Gráfico de dispersión.*

Por ejemplo, dos variables tales como la temperatura más alta del día y la temperatura más baja pueden tener una correlación positiva, tal como se muestra en la Figura 11:

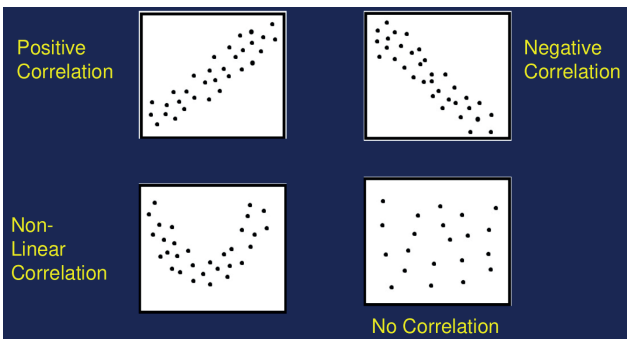


Figura 11. *Correlaciones con gráficos de dispersión.*

Una correlación positiva significa que cuando una variable crece la otra también lo hace en una canti-

dad o proporción **similar**. Por el contrario, también puede existir una **correlación negativa** entre las variables, esto significa que cuando una de las variables incrementa su valor, la otra variable decrece. Dos variables también pueden tener una relación no lineal; esto significa que el cambio en unas de las variables no siempre corresponde al mismo cambio en la otra variable. Esto se muestra gráficamente como una curva en un gráfico de dispersión en lugar de una línea recta para el caso de la **correlación lineal**. También puede que no exista una correlación entre las variables. En este caso se verá algo similar a puntos dispersos en el gráfico de dispersión, indicando que no hay una relación en el cambio de una de las variables respecto de la otra.

Gráfico de barras. El gráfico de barras es usado para mostrar la distribución de **variables categóricas**. Recordemos que el histograma también es usado para mostrar distribuciones, pero aplica solo para variables numéricas.

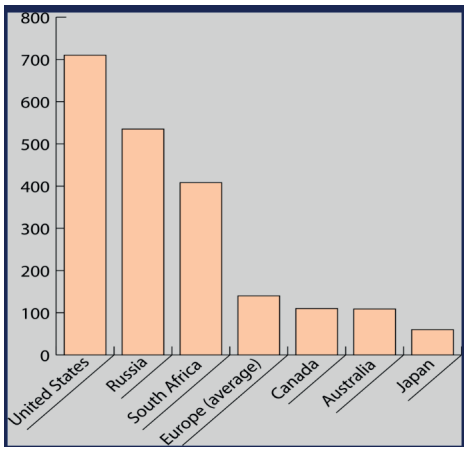


Figura 12. *Gráficos de barras.*

En un gráfico de barras las diferentes categorías de la variable categórica se muestran a lo largo del eje X, y el total de instancias por cada categoría en el eje Y. Esta es una forma efectiva de comparar las **diferentes categorías** de la variable. Por ejemplo, la categoría más frecuente puede ser encontrada fácilmente. Un gráfico de barras es una muy buena opción cuando se necesita comparar dos variables categóricas.

Por ejemplo, la Figura 13 compra dos variables categóricas. Una en **azul** y la otra en **anaranjado**, cada una con tres categorías diferentes. Esta figura muestra que la primera categoría tiene un valor más alto en la variable azul que en la variable anaranjada. A este tipo de gráfico también se lo conoce cómo gráfico de barras agrupado.

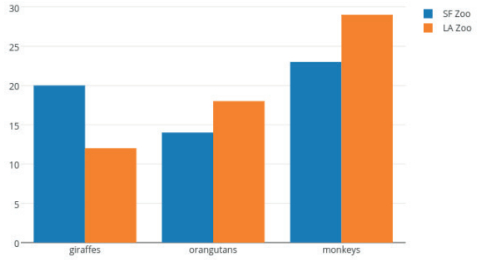


Figura 13. *Barras agrupadas.*

Otro tipo diferente de comparación puede ser realizado usando ***stacked bar chat,*** tal como se muestra Figura 14. En este caso, las cantidades de las variables se ponen una sobre otra por cada categoría. Mediante este tipo de gráfico se puede determinar que la suma combinada de la primera categoría es igual a la suma combinada de la segunda categoría, mientras que la suma combinada de la tercera es mucho mayor.

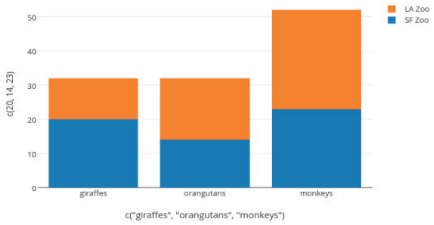


Figura 14. Stack Bar Chart.

Boxplot. Es otro **gráfico** que **muestra** la distribución de una variable **numérica**, exhibe la distribución en un formato distinto al del histograma. La Figura 15 permite ver la distribución de valores de una variable: la porción gris es la parte de caja (box), los límites inferior y superior de la caja representan los percentiles 25 y 75, respectivamente.

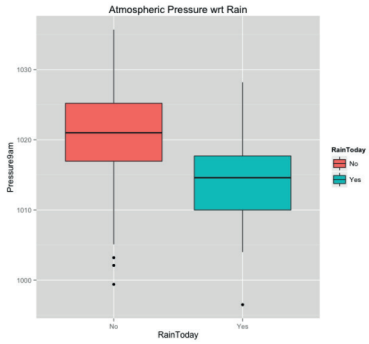


Figura 15. *Boxplot.*

Esto quiere decir que la caja **representa** el 50 % de los datos, la mediana es el percentil 50 **indicando** que el 50 % de los datos se encuentra por encima de ese valor y el otro 50 % por debajo de él. Las líneas superiores e inferiores del ***boxplot*** son los 'bigotes' y representan el percentil 10 y el percentil 90 respectivamente. Cualquier valor que se encuentre por fuera de los límites superiores e inferiores se los conoce como *outliers*.

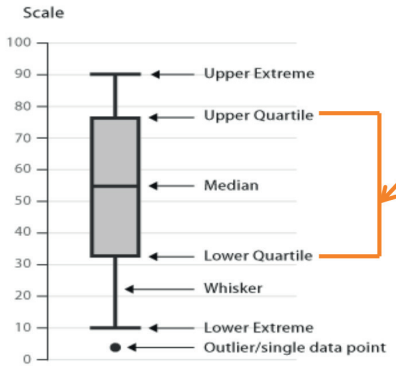


Figura 16. *Componentes del boxplot.*

Los ***boxplot*** también proveen una forma compacta de mostrar la **distribución** de las variables, esos son usados frecuentemente para comparar variables. El *boxplot* de la Figura 17 compara los niveles de salario en relación con dos roles distintos. Este ***boxplot*** brinda información rápidamente respecto de la mediana, el rango intercuartílico y la distancia entre las dos variables. Se puede ver rápidamente que la mediana del salario de marketing es más alta que la mediana del ***Reseacher***.

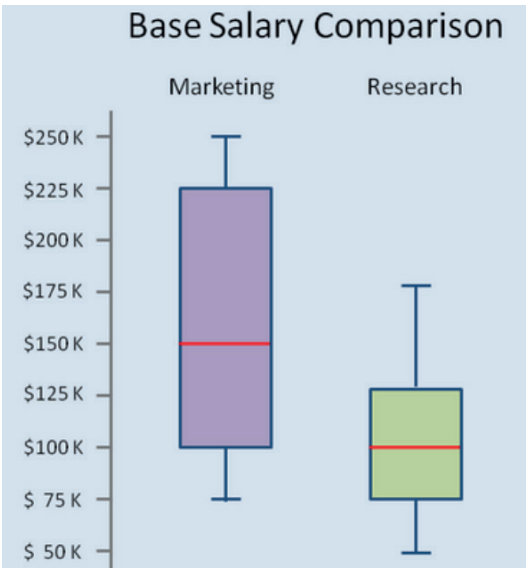


Figura 17. *Comparar distribuciones con boxplots.*

Los ***boxplot*** también se usan para mostrar si la distribución de una variable es **simétrica**, o tiene una distribución negativa o una positiva, tal como se muestra en la Figura 18.

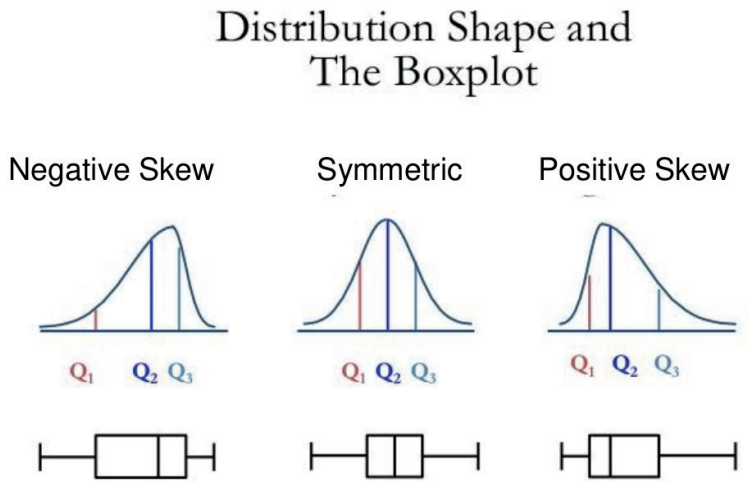


Figura 18. *Tipos de distribución con boxplots.* Obtenido de: http://www.slideshare.net/mido02/chap-3gbu

7.2. Preparación de datos

7.2.1. Preparación de datos para machine learning

Los **objetivos** de esta sección son:

- Articular la importancia de la preparación de datos
- Definir los objetivos de la preparación de datos
- Listar algunas actividades para la preparación de datos

La **información** cruda, que **obtenemos** desde las fuentes, rara vez está en el formato necesario para ser

usada para su análisis. El objetivo principal de la preparación de datos es crear los datos para el análisis, esto significa poner los datos en el formato adecuado para su análisis. Este proceso, generalmente implica seleccionar los *features* y transformar los datos tal y como se los necesita.

Es probable que los datos que se adquieren tengan algunos problemas; por lo tanto, una de las partes más importantes de la preparación de datos es **limpiar** los datos con los que vamos a trabajar, para lidiar con los que se conoce cómo ***issues* en la calidad de datos.** Hay muchos tipos de *issues* en la calidad de datos, entre ellos podemos citar:

- Valores faltantes
- Valores duplicados
- Datos inconsistentes
- Ruido
- Outliers

Los problemas citados anteriormente pueden afectar negativamente a la calidad de los datos y **comprometer** el proceso de análisis y los resultados, así que es altamente importante detectar y atacar estos *issues*. Dos de las técnicas que existen para lidiar con los *issues* en la calidad de datos son:

- **Eliminar** los registros con valores faltantes
- **Generar** el mejor estimador para estimar valores faltantes

Dado que el **objetivo** principal de la preparación de los datos es limpiarlos, a este proceso también se lo conoce como *data cleaning* o *data cleansing.* Una vez que los datos están limpios, otro de los objetivos de la preparación de los datos es llevarlos al formato necesario para el análisis. Este proceso es conocido con varios nombres: *data muging, data wranling*, *data preprocessing*. Las dos categorías principales del *data wrangling* son:

- ***Feature selection.*** Implica seleccionar los *features* que se usarán en función de las existentes, los *features* pueden ser removidos, añadidos o eliminados.
- ***Feature transformation.*** Implica cambiar el formato de los datos de alguna manera, para reducir el ruido o variabilidad y hacer que los datos sean más fáciles de analizar.

La preparación de los datos es un aspecto muy importante en el proceso del *machine learning*

Problemas más comunes en los datos [https://dataladder.com/es/los-12-problemas-de-calidad-de-datos-mas-comunes-y-su-origen/]

7.2.2. Calidad de datos

Los **objetivos** de esta sección son:

- **Describir** tres *issues* en la calidad de datos
- **Nombrar** tres razones por las que se puede tener una pobre calidad de datos
- **Explicar** por qué los *issues* en la calidad de datos deben ser atendidos

Los datos en el mundo real suelen estar usualmente muy desordenados. Por lo tanto, es necesario estar preparado para poder ordenar los datos. Un ***issue*** muy frecuente son los datos faltantes.

Los **datos duplicados** ocurren cuando el conjunto de datos tiene objetos que están duplicados, un ejemplo puede ser cuando un cliente dado tiene dos direcciones distintas; en este caso, puede ser que cuando el cliente cambió de dirección el registro se insertó en lugar de ser actualizada la dirección inicial. Por otro lado, los datos **inválidos** ocurren cuando la variable muestra un valor imposible, un ejemplo común es por ejemplo tener valores alfanuméricos en el campo de código postal cuando en realidad solo es posible tener valores numéricos.

Finalmente, los *outliers* son valores considerablemente diferentes al resto de casos, por ejemplo, cuan-

do tenemos valores de ingresos extremadamente altos o bajos en relación con el resto de los ingresos de nuestro ***dataset***.

Referencias citadas en la Clase 7

- Chawla, N. V., & Davis, D. A. (2013). *Data mining for imbalanced datasets: An overview*. In P. Perner (Ed.), *Advances in data mining. Applications and theoretical aspects* (pp. 278-300). Springer.
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
- Mido02. (s.f.). *Chap 3gbu*. SlideShare. https://www.slideshare.net/mido02/chap-3gbu

Definición de los términos citados en la Clase 7

Calidad de datos.
La calidad de los datos se refiere a la medida en que los datos son precisos, completos, consistentes, relevantes y actualizados; esto permite que sean utilizados de manera efectiva y eficiente para tomar decisiones, realizar análisis y generar valor. Una alta calidad de los datos asegura que la información sea confiable y útil, mientras que una baja calidad puede llevar a errores, malentendidos y decisiones incorrectas, afectando negativamente los resultados de cualquier proceso o actividad basada en esos datos.

Datos duplicados.
Los datos duplicados son registros repetidos o redundantes dentro de un conjunto de datos que contienen la misma información o información muy similar, lo que puede ocurrir por errores en la entrada de datos o por problemas en el proceso de integración de sistemas. Estos duplicados pueden generar confusión, distorsionar los análisis, aumentar el uso de almacenamiento innecesario y afectar la calidad de los datos, por lo que es crucial identificarlos y eliminarlos, para garantizar la precisión y eficiencia de los procesos basados en esos datos.

Profundización Clase 7

Problemas comunes en la calidad de datos:

1. Datos Incompletos

- **Descripción**: Los registros faltan información esencial, como valores en campos clave.
- **Impacto**: Resultados sesgados o decisiones erróneas debido a la falta de datos.
- **Solución**: Asegurar la recopilación completa de datos a través de procesos más rigurosos y controles de calidad.

2. Datos Duplicados

- **Descripción**: Existen registros repetidos que contienen la misma información.
- **Impacto**: Consumo innecesario de espacio de almacenamiento, distorsión de análisis y aumento del tiempo de procesamiento.
- **Solución**: Implementar procedimientos para detectar y eliminar duplicados de manera automática.

3. Datos Inexactos

- **Descripción**: Los datos no coinciden con la realidad o contienen errores.
- **Impacto**: Desinformación, decisiones incorrectas y análisis poco fiables.
- **Solución**: Validar los datos antes de su entrada y realizar auditorías periódicas para detectar y corregir errores.

4. Formato Inconsistente

- **Descripción**: Los datos están en diferentes formatos o estructuras, lo que dificulta su análisis y comparación.
- **Impacto**: Dificulta la integración de datos de diversas fuentes y afecta la precisión de los análisis.
- **Solución**: Establecer y seguir estándares de formato uniforme para la recopilación, almacenamiento y procesamiento de datos.

Cuando usar un *pie chart*

1. Mostrar Proporciones o Participaciones Relativas

- **Descripción**: Es ideal para mostrar cómo se distribuye un total entre varias categorías.
- **Ejemplo**: Porcentaje de ventas de diferentes productos dentro de una empresa.

2. Visualizar Datos Categóricos

- **Descripción**: Útil cuando se tiene un número limitado de categorías con diferencias claras entre ellas.
- **Ejemplo**: Distribución de respuestas en una encuesta de opinión.

3. Comparar Partes de un Todo

- **Descripción**: Cuando se quiere comparar las contribuciones de cada parte a un todo.
- **Ejemplo**: Porcentaje de uso de diferentes canales de marketing.

PUCE Virtual

La excelencia no se improvisa

síguenos