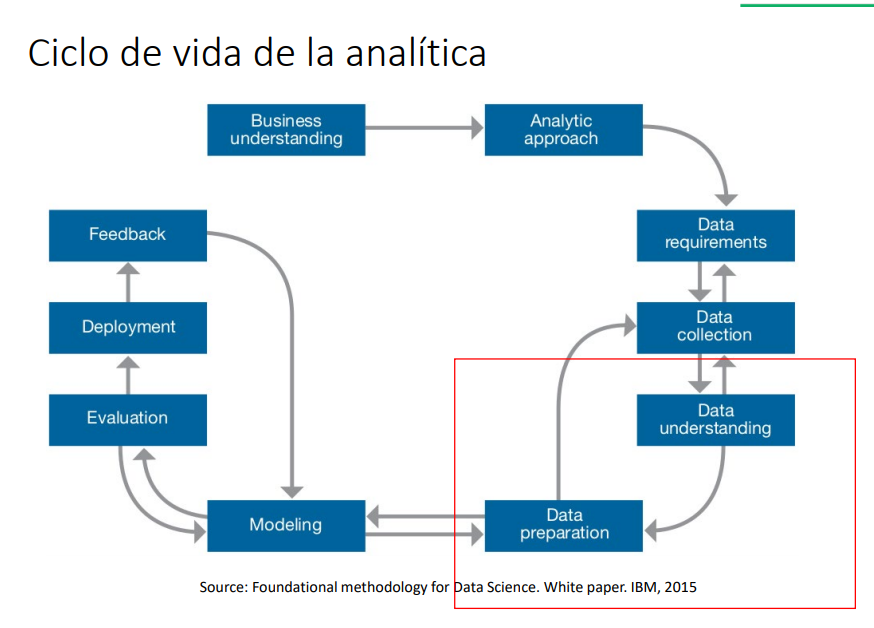
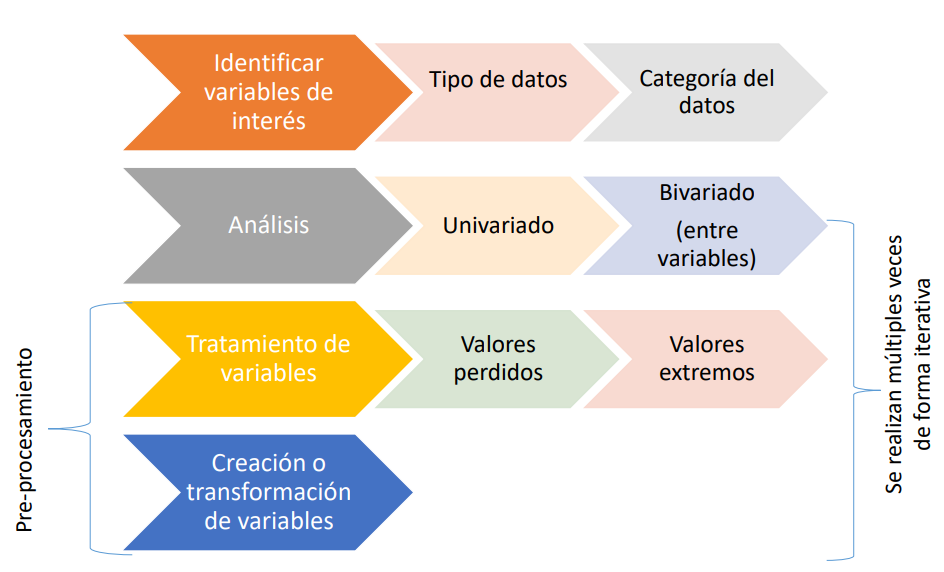
# Análisis exploratorio de datos





Tipos de preguntas en analítica (tipos de análisis) en orden de menor a mayor dificultad

•Análisis exploratorio

•Análisis inferencial

•Análisis predictivo

•Análisis causal

•Análisis mecánico

# **Análisis exploratorio** (EDA) ¡Entendiendo los datos!

# Objetivos:

# • Entender la estructura del conjunto de datos

# • Detectar errores y deficiencias en los datos

# • Encontrar relaciones entre las variables

# • Identificar modelos candidatos para el análisis de los datos

# Características:

# • Primer tipo de análisis que se realiza

# • Diferentes abstracciones de los datos

# • Incluye el análisis descriptivo (e.j., BI tradicional), pero no modelos estadísticos formales, inferencias o predicciones

# • Lo que se observa no puede generalizarse

**Análisis inferencial** ¡Concluyamos acerca de la población!

Objetivos:

• Analizar una muestra relativamente pequeña para decir algo acerca de la población a la que pertenece la muestra

Características:

• Se basa en modelos estadísticos

• Verificar una hipótesis y concluir sobre la población con base en el análisis de la muestra

• Determinar la probabilidad de que se presenten características en la población con base en las características de la muestra

• La selección de la muestra es crucial

**Análisis predictivo**

Objetivo:

• Predecir hechos futuros o desconocidos a partir del análisis de datos actuales e históricos

Características

• Aplican técnicas de la estadística, minería de datos, aprendizaje automático (machine learning), inteligencia artificial

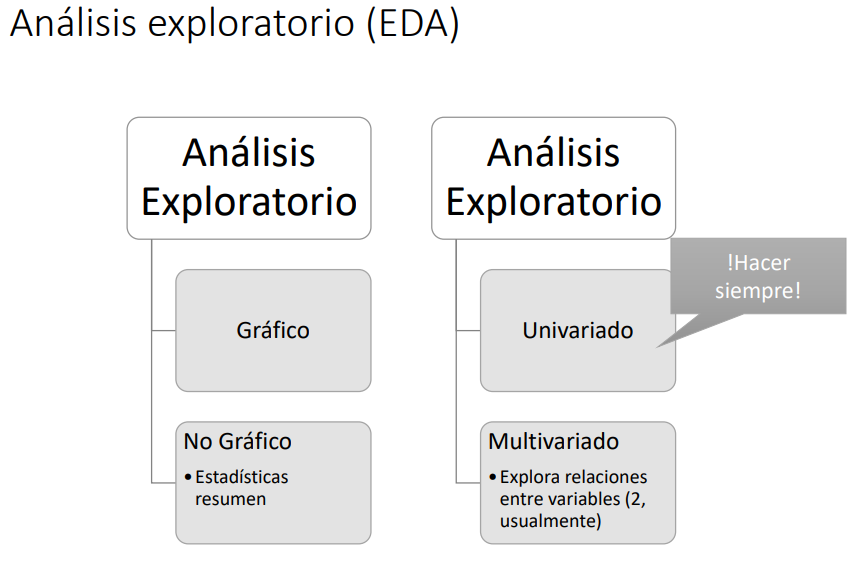
**Análisis causal**

Objetivo: Identificar qué sucede con una variable cuando se cambia otra

**Análisis mecánico**

Objetivo: Entender qué cambios en qué variables dan lugar a cambios en otras variables

Más usado en el análisis de fenómenos físicos



Exploración de variables: estadísticas resumen

**Numéricas**: funciones estadísticas de resumen:

• Medidas de tendencia central

- Media

-Mediana

• Medidas de dispersión de los datos ¿cómo son los rangos de las variables?

-desviación estandar

• Cuartiles y el resumen de 5 números (mínimo, máximo y los cuartiles)

-Los tres valores que dividen un conjunto de datos ordenado en cuatro partes

-Nos ayudan a entender qué tan dispersos están los datos, qué tan diversos son

-Entender si la mayoría de los datos se parecen o no a las medidas centrales (promedio, mediana)

• Útiles para:

• Entender la distribución de la muestra para aproximarse a la distribución de la población

• Detectar datos atípicos (outliers, valores fuera de las distribuciones más comunes)

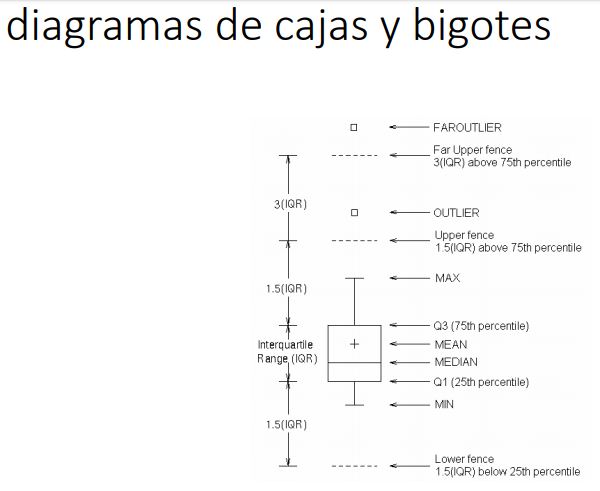
**Categóricas**: visualizaciones

• Útiles para entender la distribución de los datos y las relaciones entre variables

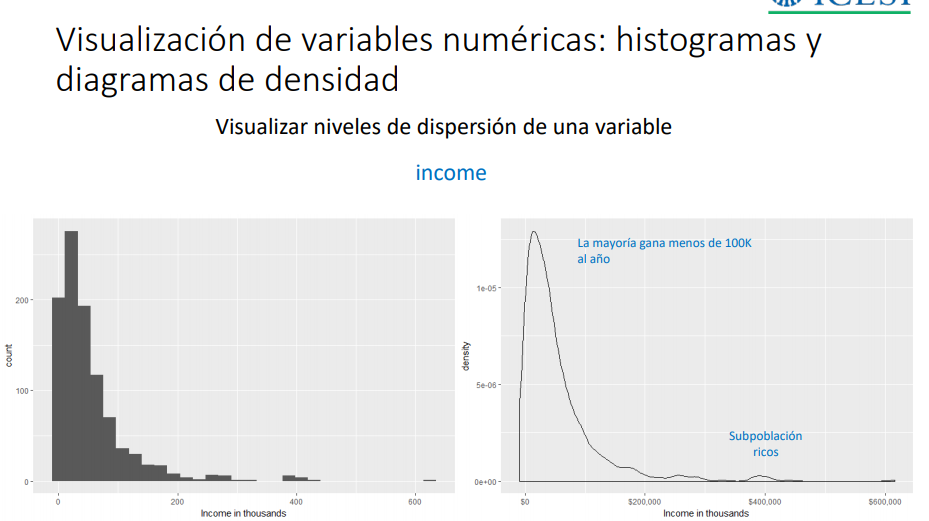
• Identificar el valor pico de las distribuciones

• Identificar si la distribución es unimodal o multimodal

• Identificar qué tan normal o lognormal son los datos

• Identificar la variación de los datos, ¿se concentran en cierto intervalo o categoría (en el caso de las categóricas)?

# 



# 

Exploración no gráfica de variables categóricas

• Este tipo de variables se examinan generalmente usando tablas de frecuencia para las categorías en lugar de estadísticas resumen.

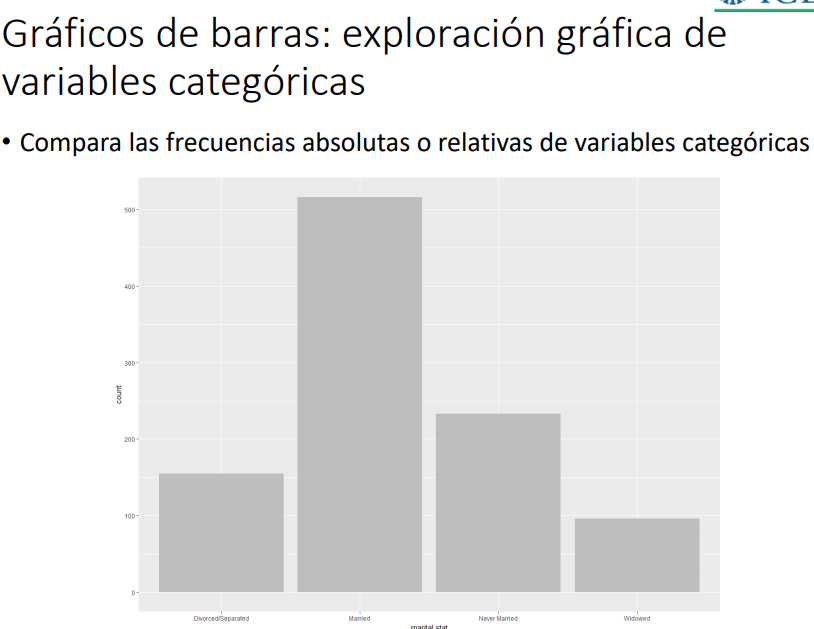
# Resultado de imagen para tablas de frecuencia

**La Moda**

• Medida central usualmente aplicada a variables categóricas

• Nos permite entender valores importantes de las variables categóricas. Dependiendo del negocio hay categorías más importantes que otras

• Las variables pueden ser unimodales, bimodales o multimodales



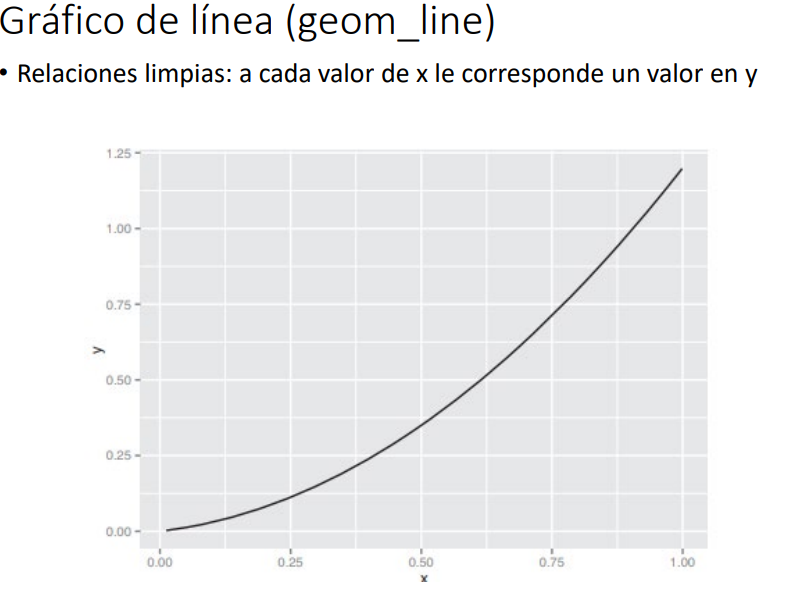
# Análisis multivariado:

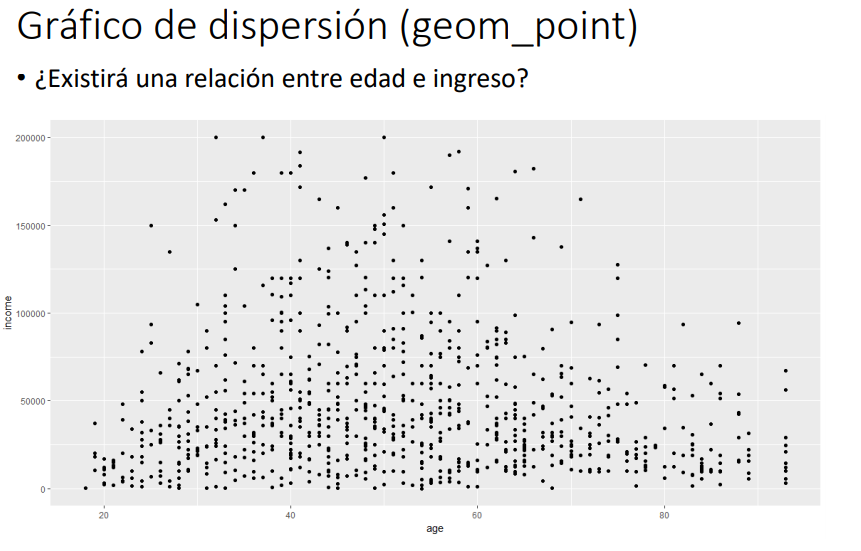
• Nos permite explorar relaciones entre dos o más variables para ir determinando variables explicativas que sean candidatas para ser incluidas en el modelo

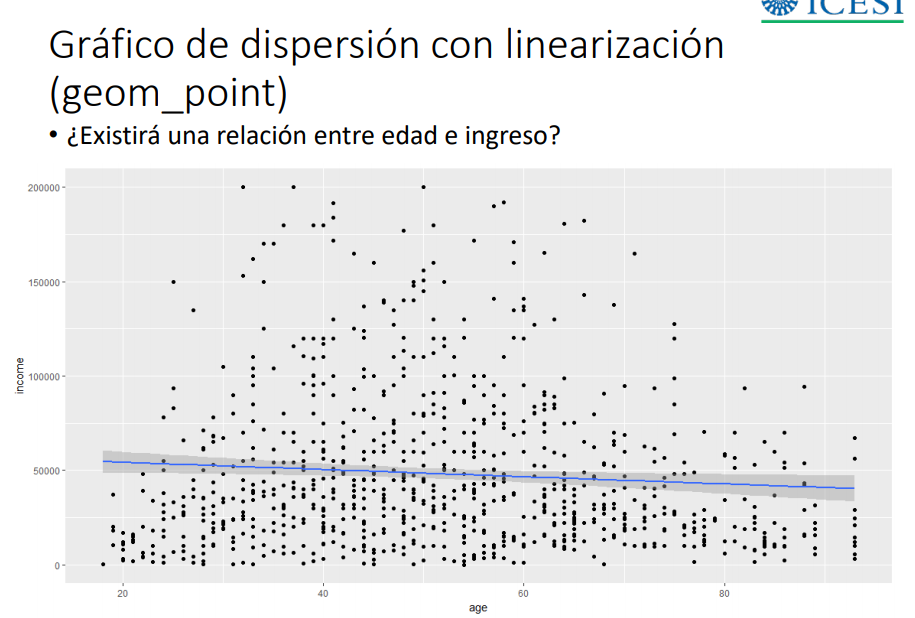
# **Tipos de preguntas**

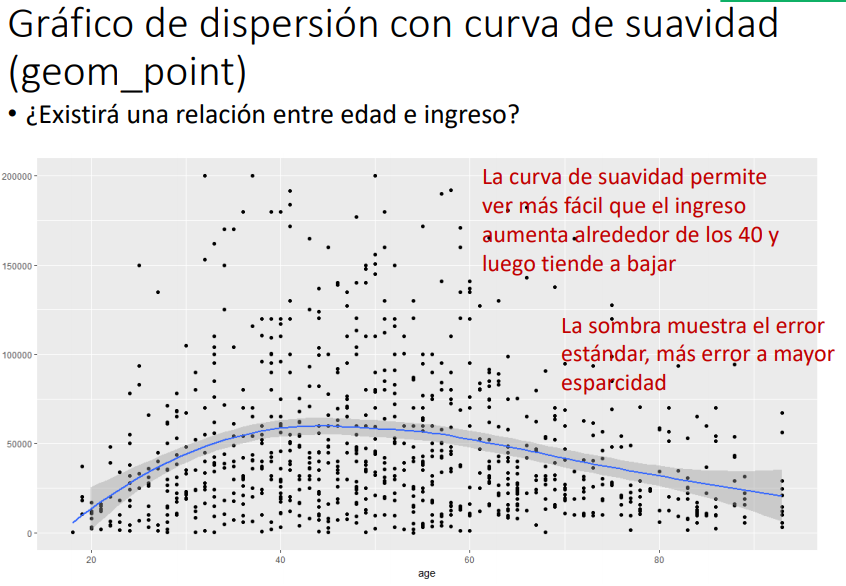
• ¿Existe correlación entre edad e ingreso? ¿Qué tipo de relación? ¿Qué tan fuerte es?

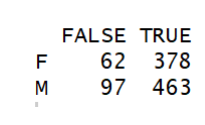
• ¿Existe alguna relación entre estado civil y el hecho de que la persona tenga o no un seguro de salud?



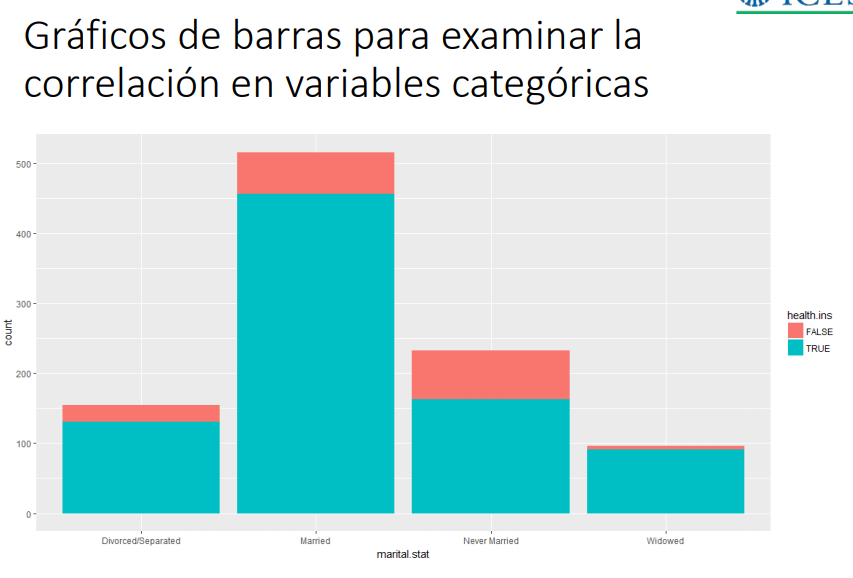






Correlación de variables categóricas

• Tablas cruzadas



Datos brutos vs. datos preparados

• Datos brutos (raw data)

• Son la fuente original de los datos

• Difíciles de analizar

• Usualmente se procesan solo una vez, para prepararlos

• No se ha ejecutado ningún software sobre los datos

• No ha habido manipulación de los datos

• No se han eliminado datos

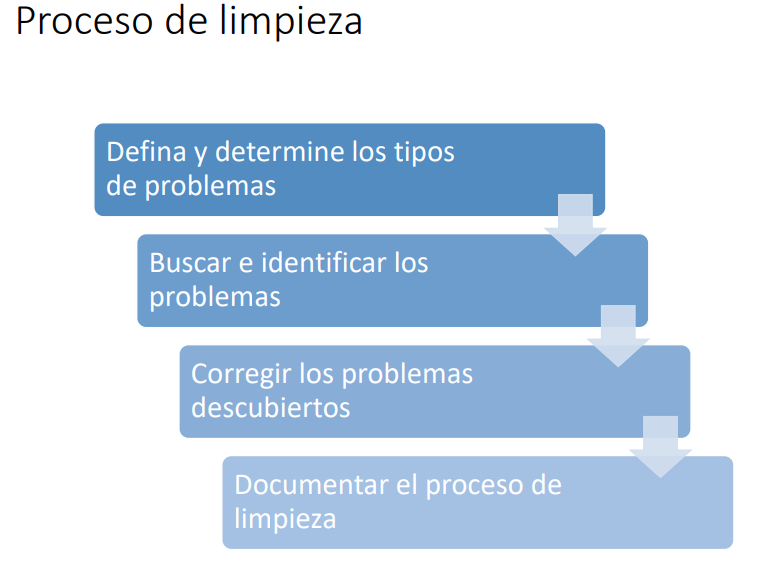
• No se han resumido los datos

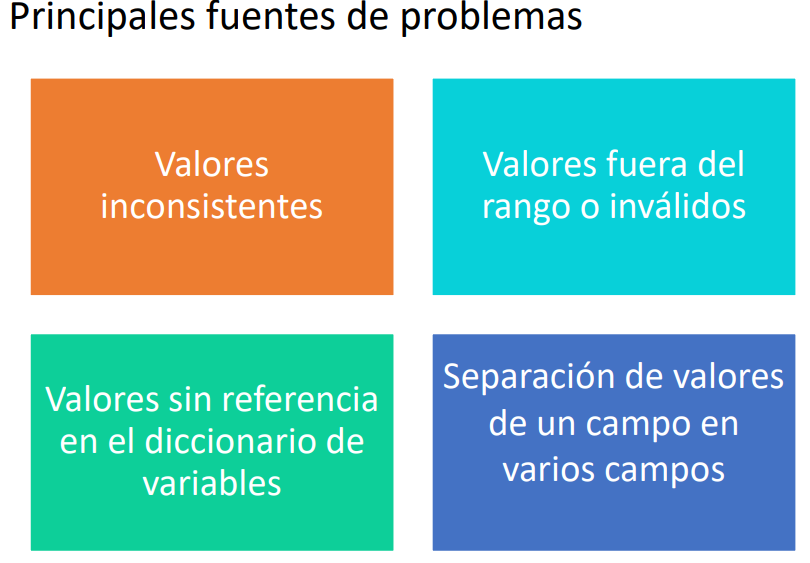
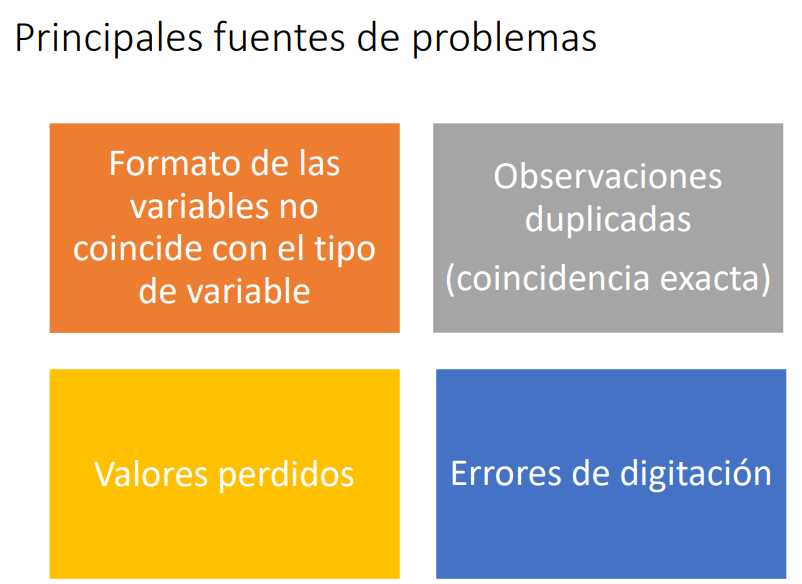
• Datos preparados (tidy data)

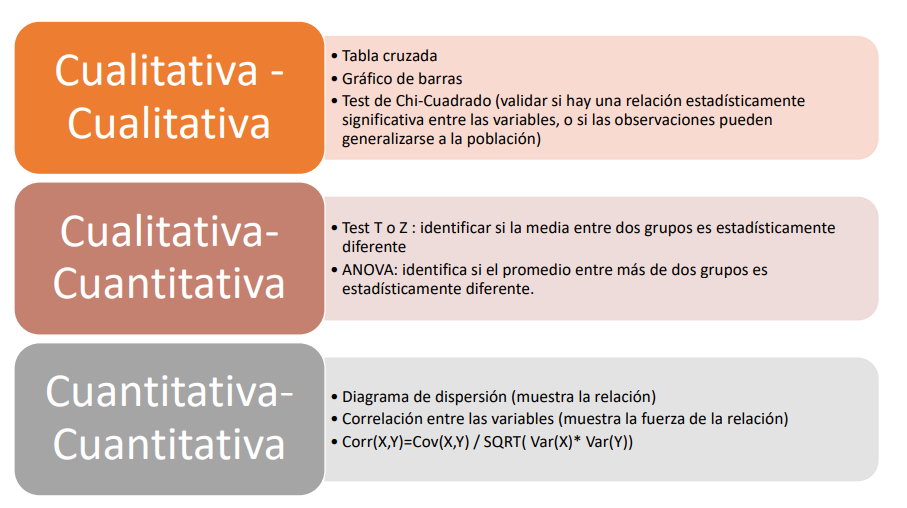
• Listos para ser analizados

• Su preparación incluye: limpieza, transformación, fusiones (merging), extracción de subconjuntos (subsetting)

• ¡Todos los pasos ejecutados en la preparación deben documentarse para que sean repetibles!







Pruebas de Hipótesis Ejemplo

• Formular hipótesis Hipótesis nula (Ho): No hay diferencia en el puntaje promedio de matemáticas entre hombres y mujeres.

Hipótesis alternativa(Ha): No Ho (Implica que sí hay una diferencia entre hombres y mujeres)

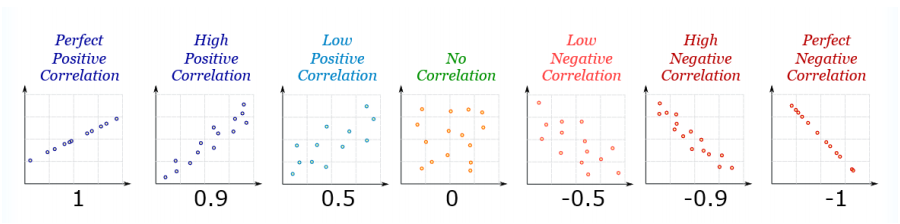
• Seleccionar la prueba estadística adecuada y evaluar

• Tomar una decisión Si el p valor < 0.05 se rechaza la hipótesis nula con un nivel de confianza del 95%.

• Interpretar decisión Rechazar Ho: Los resultados sustentan que hay una diferencia en el puntaje promedio de matemáticas entre hombres y mujeres. No rechazar Ho: Los datos no proveen suficiente evidencia que demuestre que hay una diferencia significativa en el puntaje promedio de matemáticas entre hombres y mujeres.

# Correlación Fuente:

# Una medida estadística que indica la fuerza de la relación lineal que existe entre dos variables. El signo y el valor absoluto indican la dirección y la magnitud de esa relación, respectivamente.



Podemos utilizar otros métodos que nos permiten ver como un todo la relación entre cada variable, entre estos podemos encontrar:

• Matriz de correlación

• Matriz de correlación para saleprice

• Mapas de dispersión entre las variables más correlacionadas

# Limpieza

Eliminar signos extraños de los nombres de las columnas.

Poner todos los nombres de las columnas en lower o en upper.

Quitar tildes de los nombres de las columnas

## Formato de variables no coincide con el tipo de variable

## Observaciones duplicadas

## Valores perdidos

## Errores de digitación

## Valores inconsistentes

## Valores inválidos o fuera del rango

## Valores sin referencia en el diccionario de variables

## Separación de variables de un campo en varios campos

# Exploración

## Tipo de datos

## Categoría de datos

## Análisis univariado

SUMMARY

STDDEV

Histogramas y Cajas de bigotes

Tablas de frecuencia absoluta y relativa (Cualitativas)

## Análisis bivariado

### Cualitativa – Cualitativa

Tabla cruzada

Gráfico de barras

Test Chi Cuad

### Cualitativa – Cuantitativa

Test t o Z

Anova (Recordar los supuestos)

### Cuantitativa – Cuantitativa

Diagrama de dispersión

Correlación entre variables

## Valores perdidos

Eliminar

Transformar

Tratar por separado

## Valores extremos

## Creación o transformación de variables

Estandarizar con cuadrática o logarítmica. Esto sesga entonces normaliza