**1. Define Ciencia de Datos**

Es la obtención de conocimiento proveniente de grandes volúmenes de datos estructurados y/o no estructurados. Además, permite la creación de productos de datos: una aplicación que adquiere su valor a partir de datos, y crea más datos como resultado.

**2. Define Big Data y las 3 V’s**

El término Big data se refiere a información que no puede ser procesada o analizada utilizando métodos tradicionales

– Una base relacional sería muy lenta y costosa

• Se describe en términos de:

– Volumen: enormes cantidades de datos estructurados, no estructurados y semiestructurados

– Variedad: gran diversidad en el Apo de datos, tales como correos, tweets, audio, videos, etc.

– Velocidad: respuestas rápidas para obtener la información necesaria en el Aempo preciso

• Big data requiere de la ciencia de datos para transformar la información en conocimiento

**3. Menciona y describe los tipos de preguntas que se pueden responder en un proyecto de ciencia de datos**

**Descriptiva**

• Recolecta, presenta y caracteriza un conjunto de datos

**Exploratoria**

• InvesAgación realizada para ganar familiaridad con un fenómeno

**Predictiva**

• Análisis de hechos actuales o pasados con el fin de predecir eventos futuros desconocidos

**Causal**

• Determinar las causas detrás de un evento

**Deductiva**

• Inferir a parAr de una muestra aleatoria el comportamiento de la población.

**Mecanicista**

• Comprender los mecanismos detrás de un proceso

• ¿Cuál es el nivel de ingreso de mis clientes?

• ¿Qué clientes están en riesgo de irse con la competencia y como puedo retenerlos?

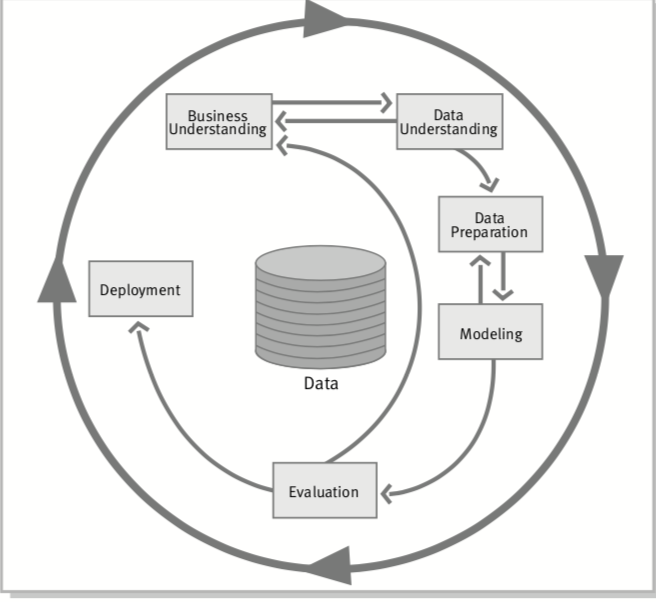
• ¿Qué productos debo tener en existencia en cada una de mis Aendas?

• ¿A qué equipo debo darle mantenimiento?

• ¿Cuál es el nivel adquisiAvo de mis clientes?

• ¿Qué opinan los clientes en redes sociales de nuestros productos o servicios?

**4. Menciona y describe las etapas de la metodología CRISP-DM**

****

The life cycle of a data mining project consists of six phases, shown in Figure 2. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. The outcome of each phase determines which phase, or particular task of a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases.

The outer circle in Figure 2 symbolizes the cyclical nature of data mining itself. Data mining does not end once a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more-focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones. In the following, we briefly outline each phase:

**Business understanding**

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

**Data understanding**

The data understanding phase starts with initial data collection and proceeds with activities that enable you to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information.

**Data preparation**

The data preparation phase covers all activities needed to construct the final dataset [data that will be fed into the modeling tool(s)] from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools.

**Modeling**

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary.

**Evaluation**

At this stage in the project, you have built a model (or models) that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain the model properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

**Deployment**

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying “live” models within an organization’s decision making processes—for example, real-time personalization of Web pages or repeated scoring of marketing databases. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

**5. Contrasta los métodos estadísticos tradicionales con el aprendizaje de máquina**

• **Inferencia**: el proceso de sacar conclusiones sobre la población a parAr de muestras

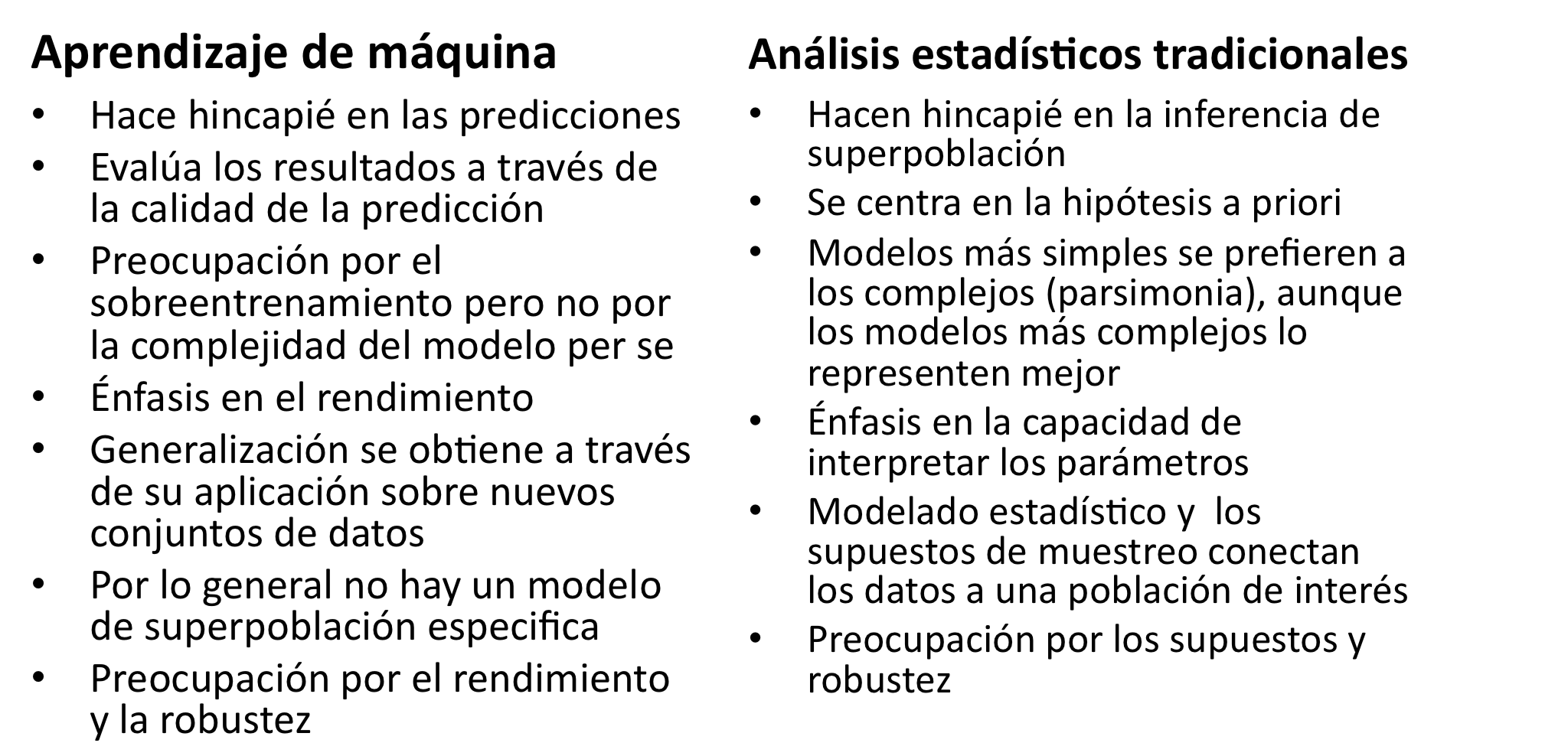
– estimación, muestreo, variabilidad, definición de poblaciones

La estadística busca modelos simples que expliquen el por qué de los fenómenos

**• Predicción:** el proceso de adivinar un resultado a partir de predictores

– clasificación, regresión y más recientemente aprendizaje de máquina, aprendizaje supervisado, deep learning, boosAng, árboles de decisión, bosques aleatorios, etc.

El aprendizaje de máquina busca que las predicciones sean lo más certero possible



**6. ¿Cómo se determina el éxito de un proyecto de ciencia de datos?**

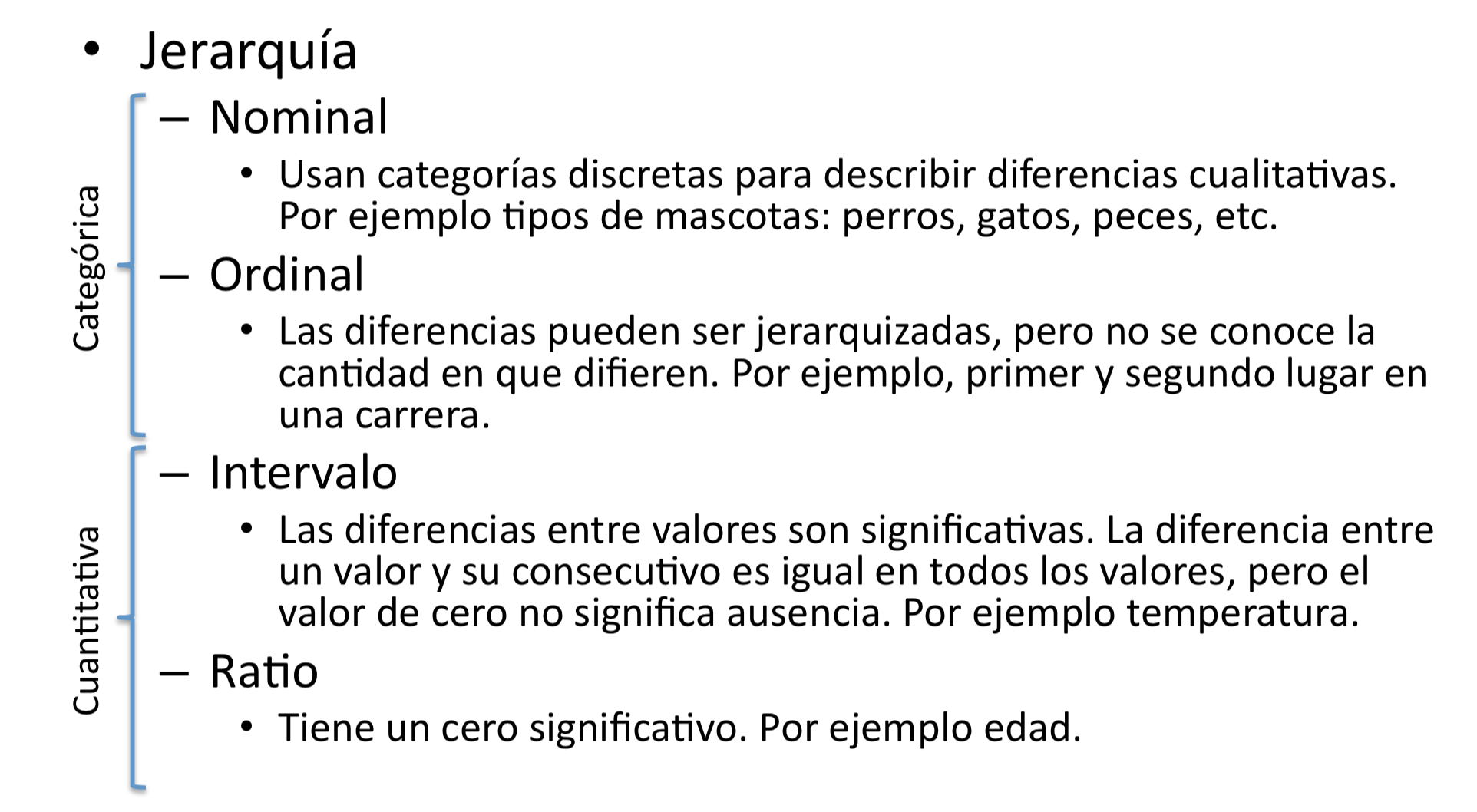
• Si se crea nuevo conocimiento.

• Si decisiones se toman o políAcas se determinan en función de los resultados del experimento

• Si se crea un informe, presentación o aplicación con impacto

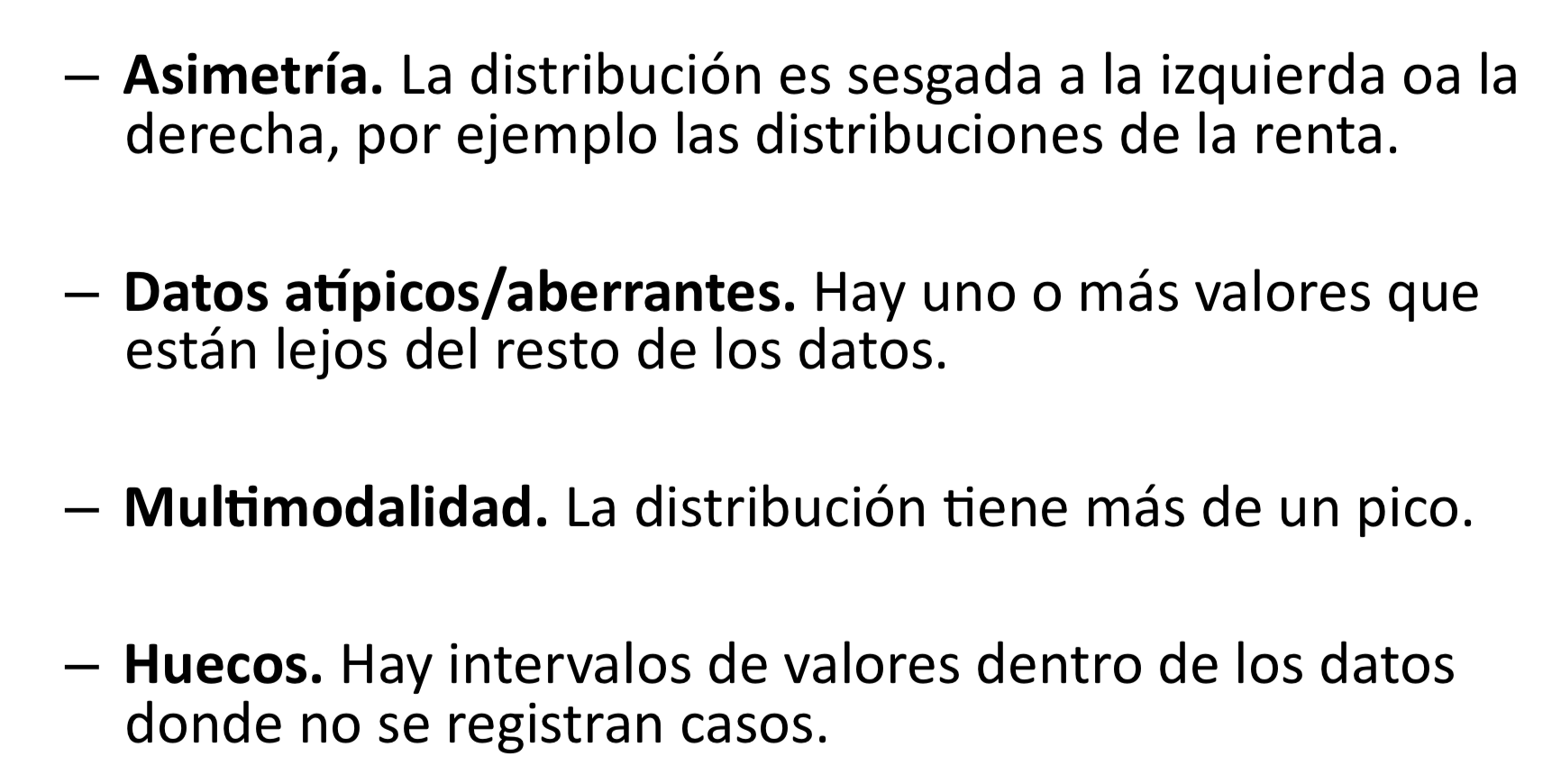
• Si se determina que los datos no pueden responder a la pregunta planteada

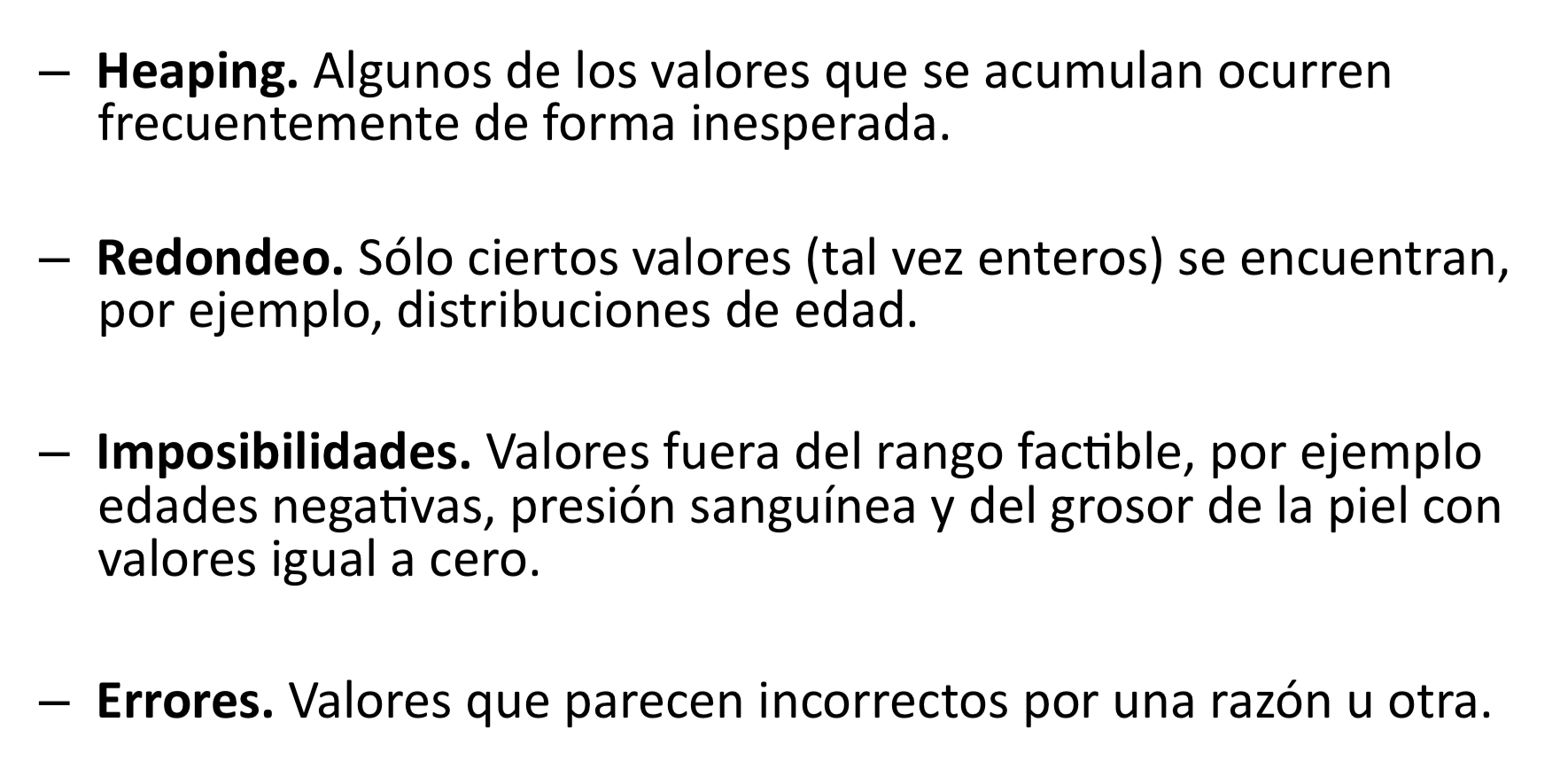
**7. Menciona y describe la jerarquía de escalas de medida**

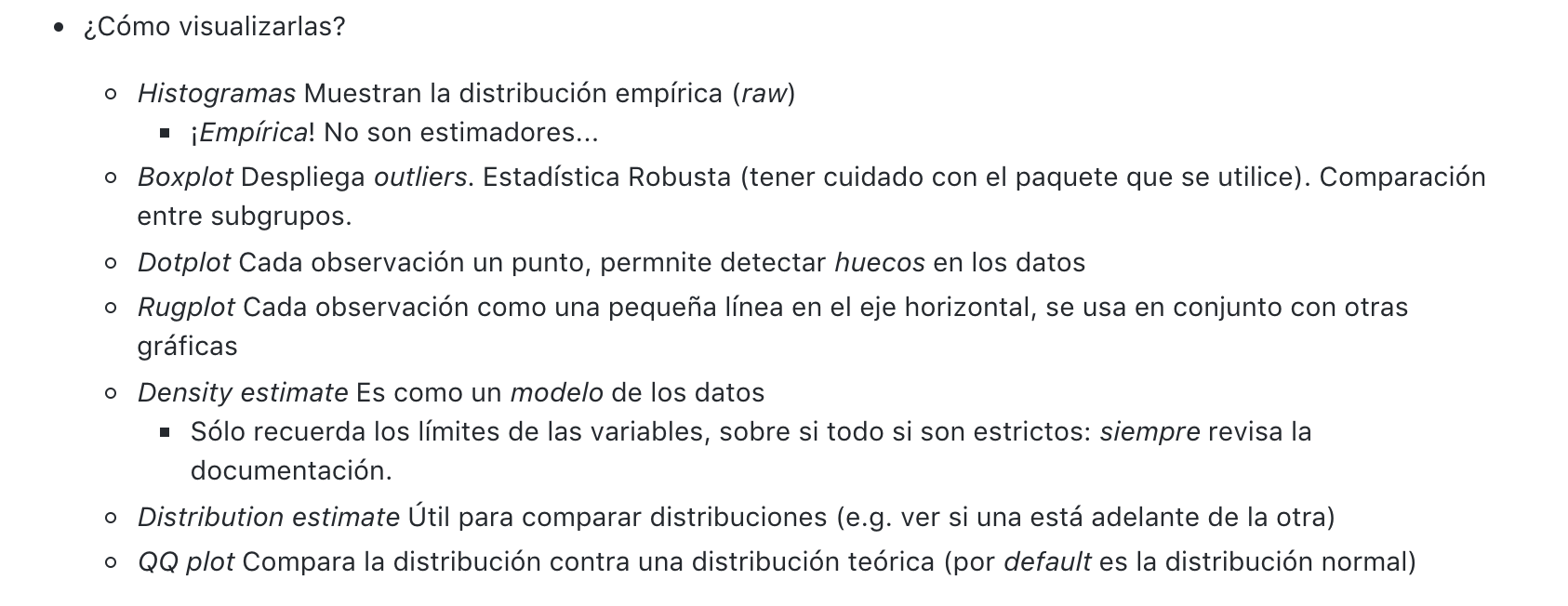
****

**8. ¿Qué podemos encontrar durante el análisis univariado cuantitativo y en el categórico?**

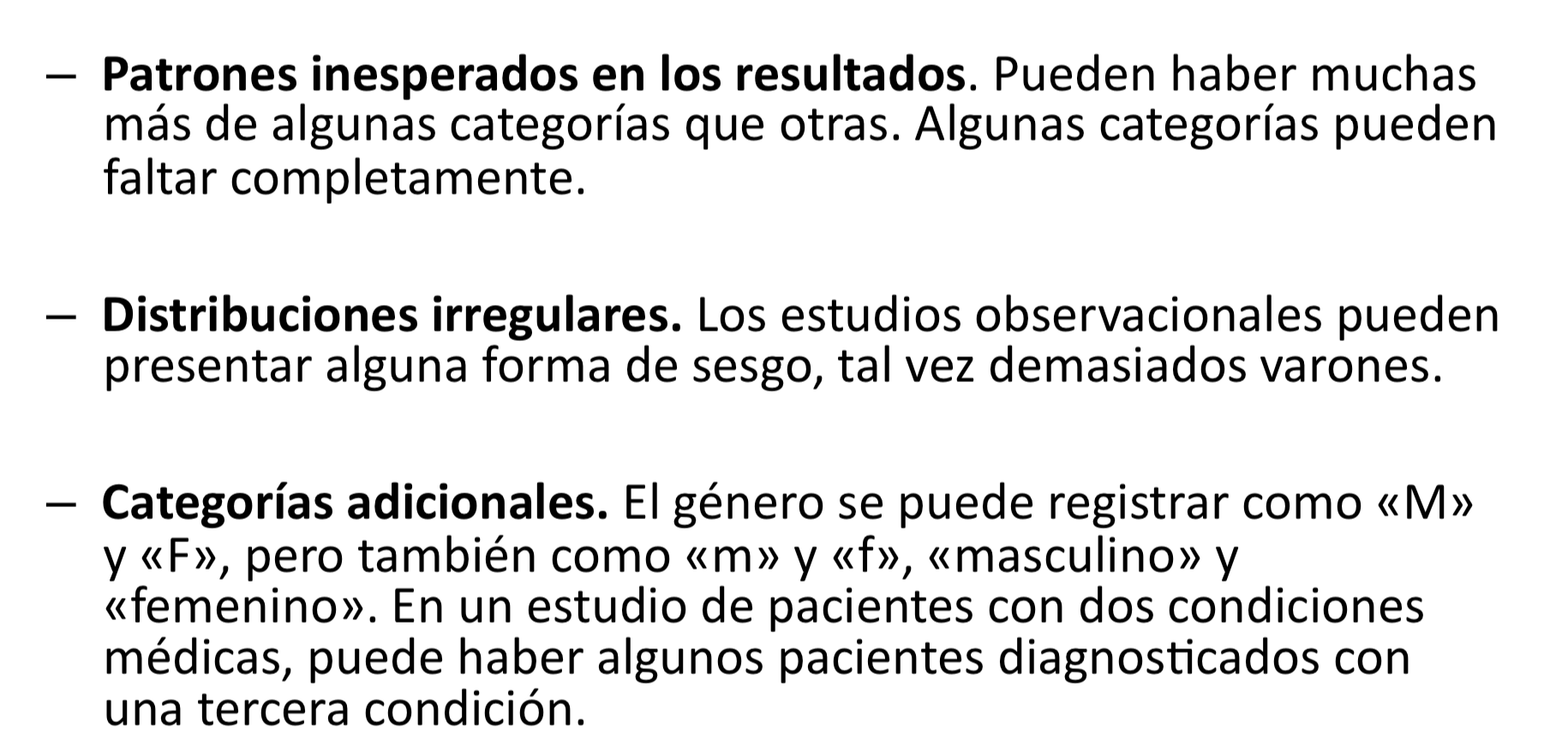
**Cuantitativas**

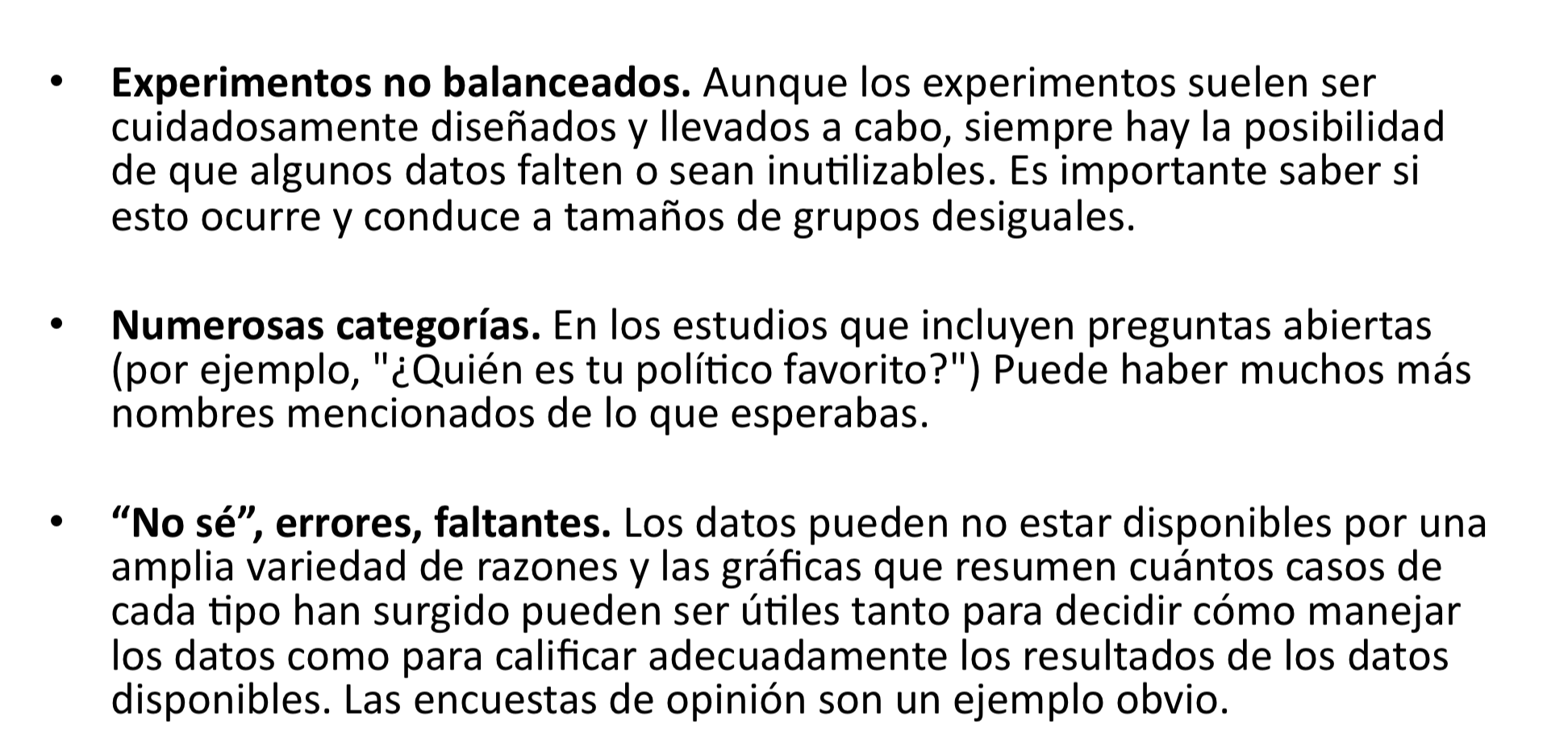
****

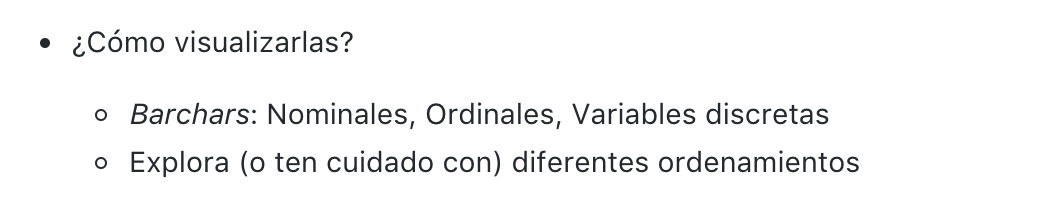
****

****

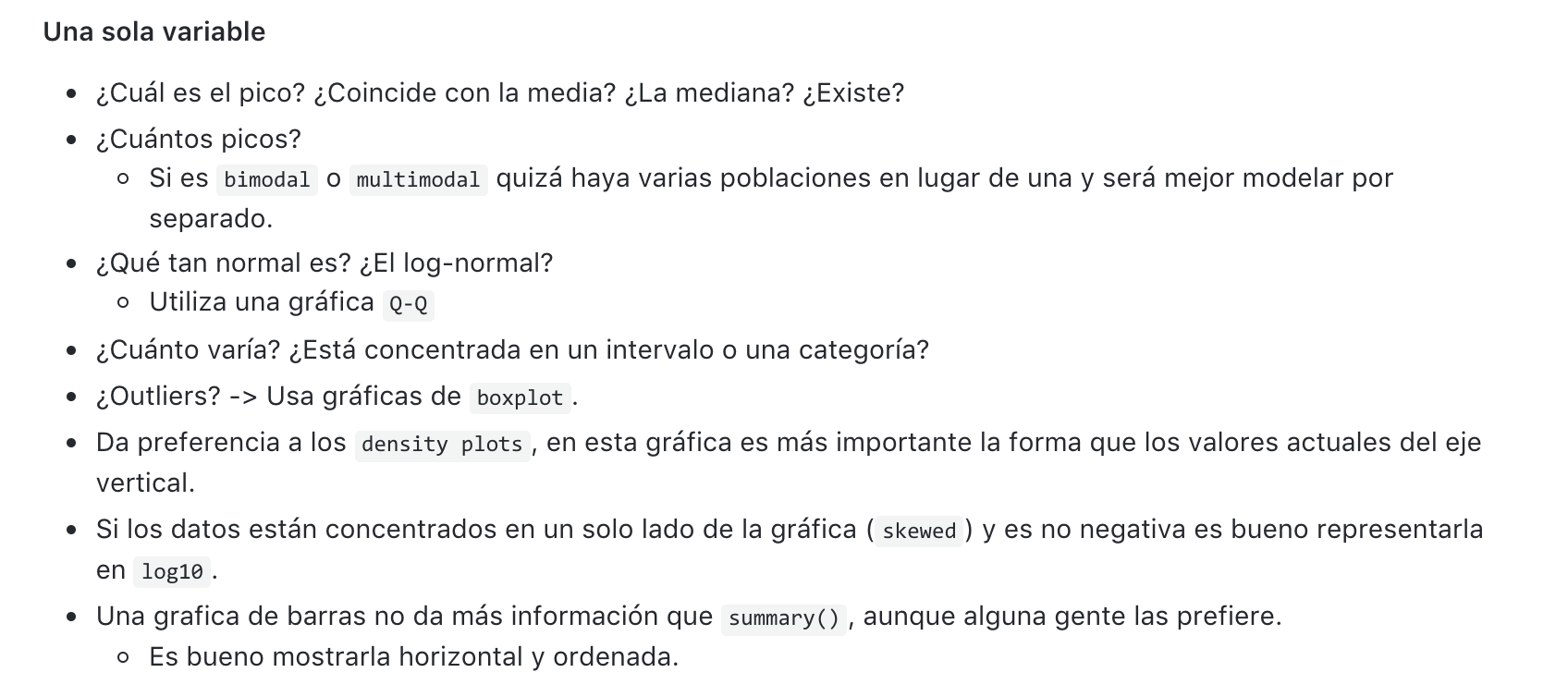
**Categoricas**

****

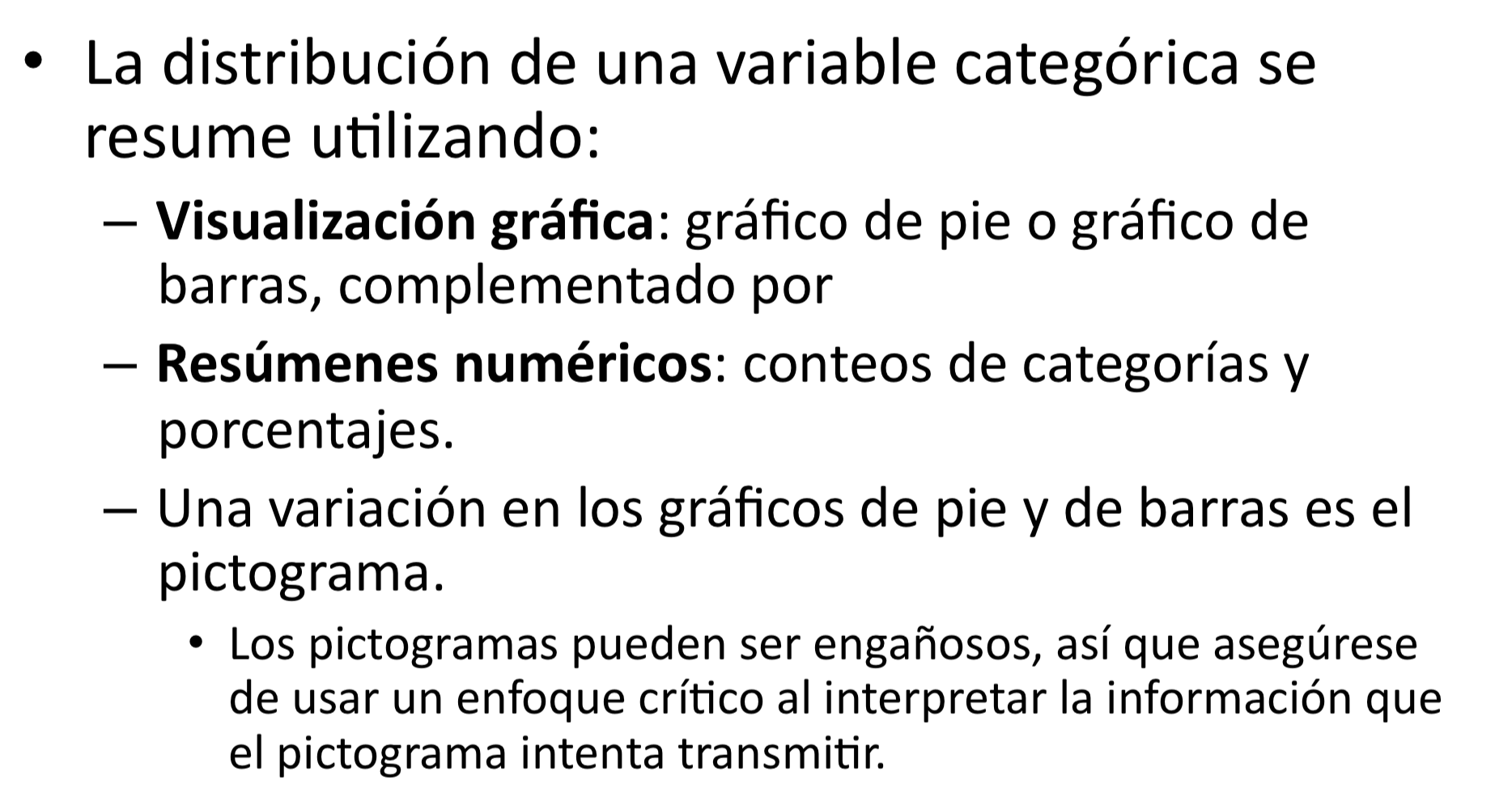
****

****

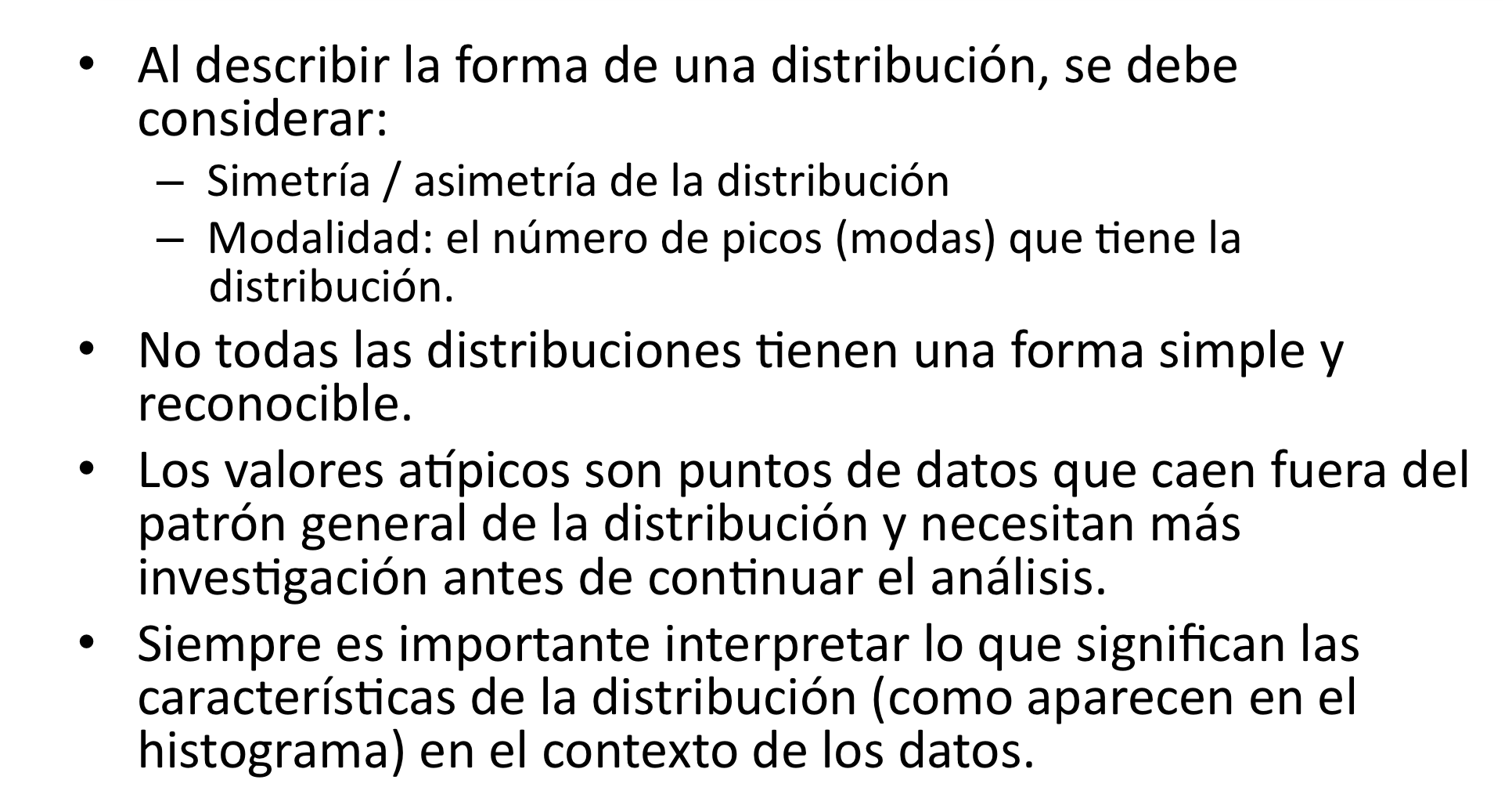
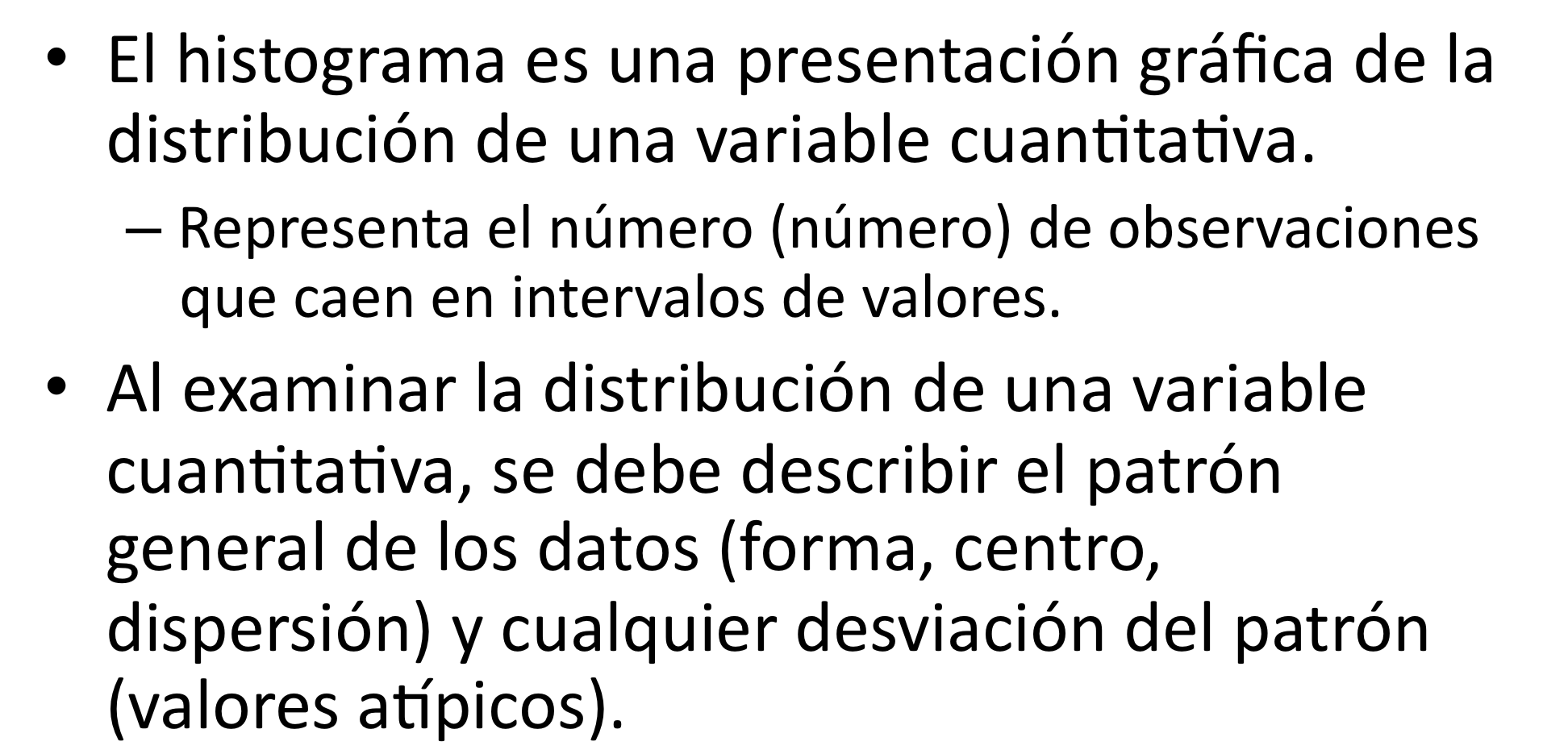
**9. ¿Qué tipo de gráficas se utilizan comúnmente en el análisis univariado cuantitativo y categórico y con qué fin?**

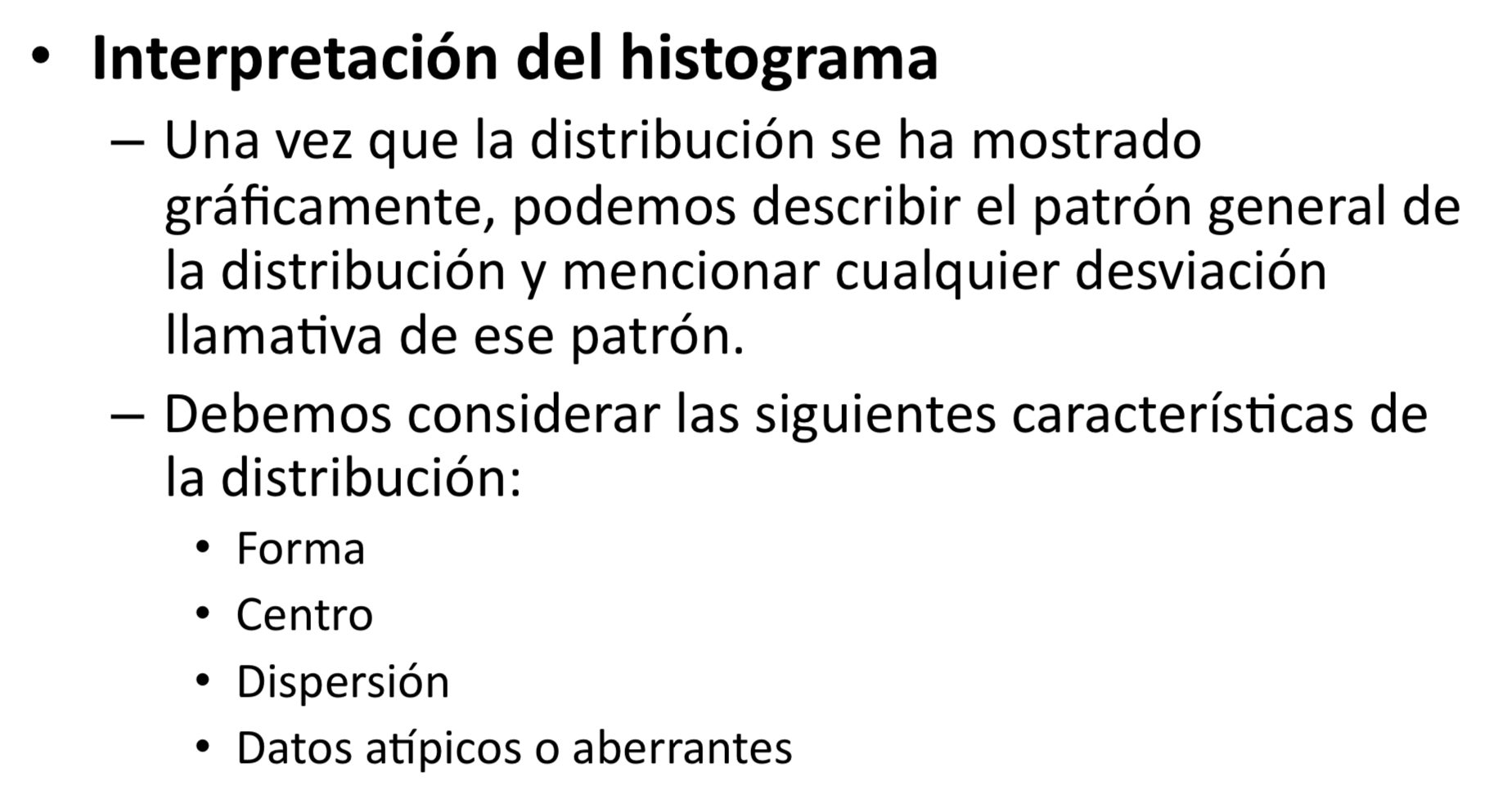
****

**Categóricas**

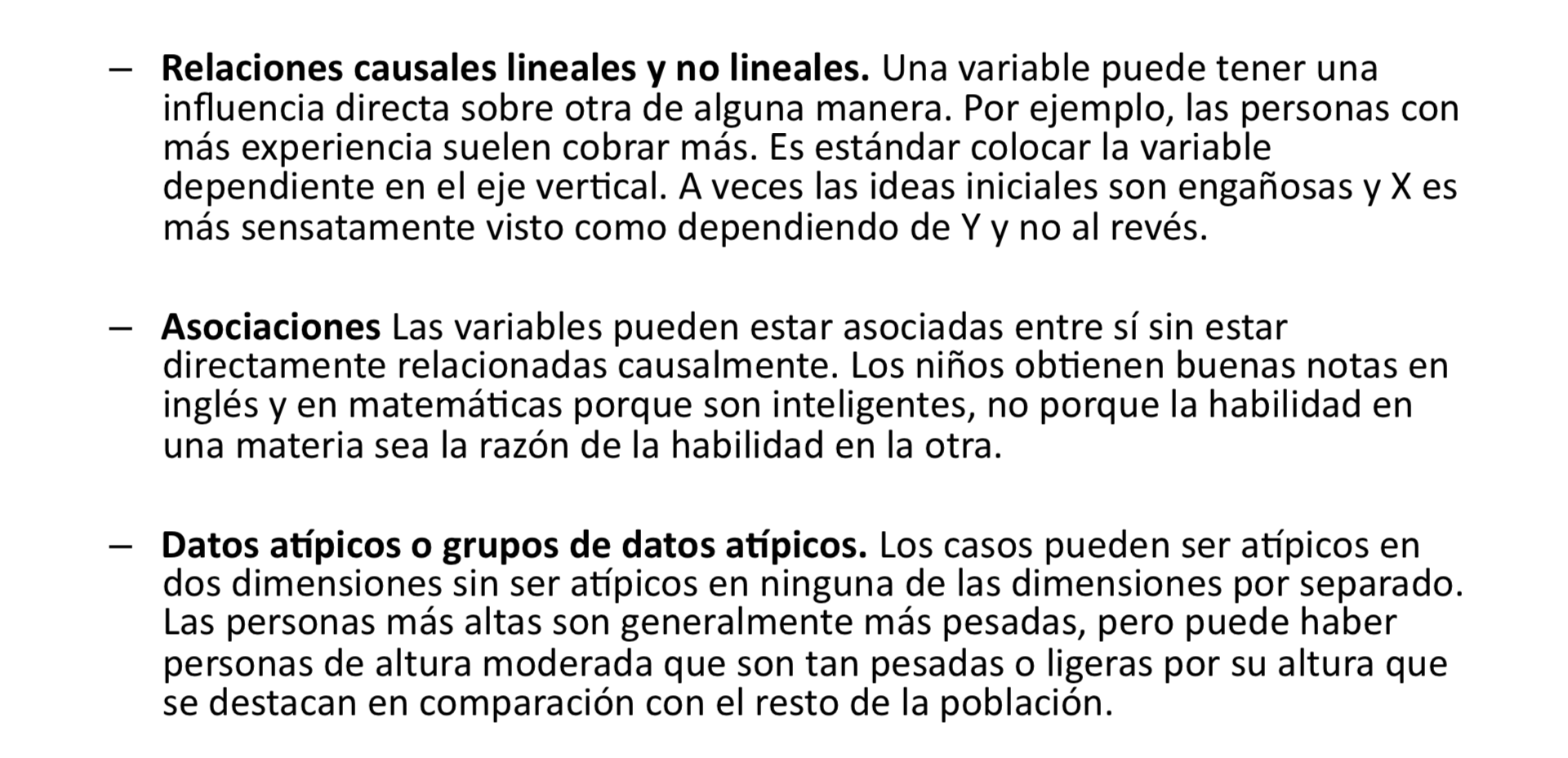
****

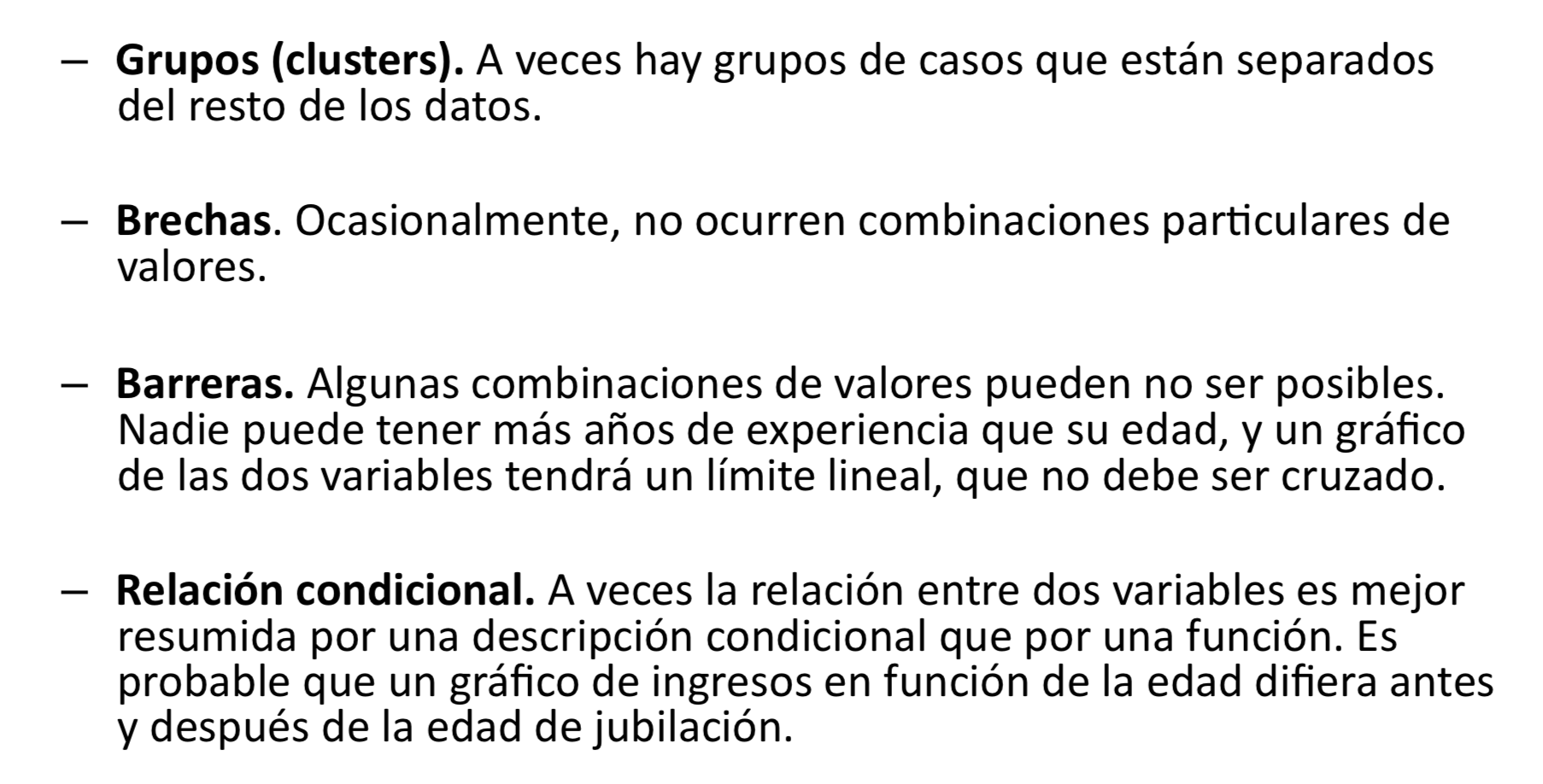
**Cuantitativas**

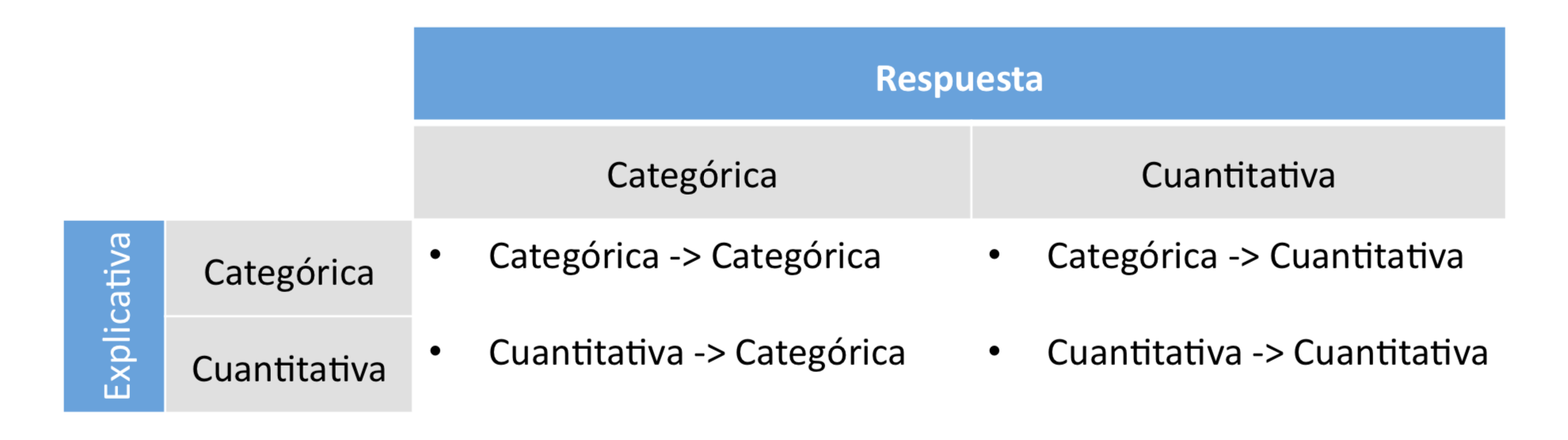
****

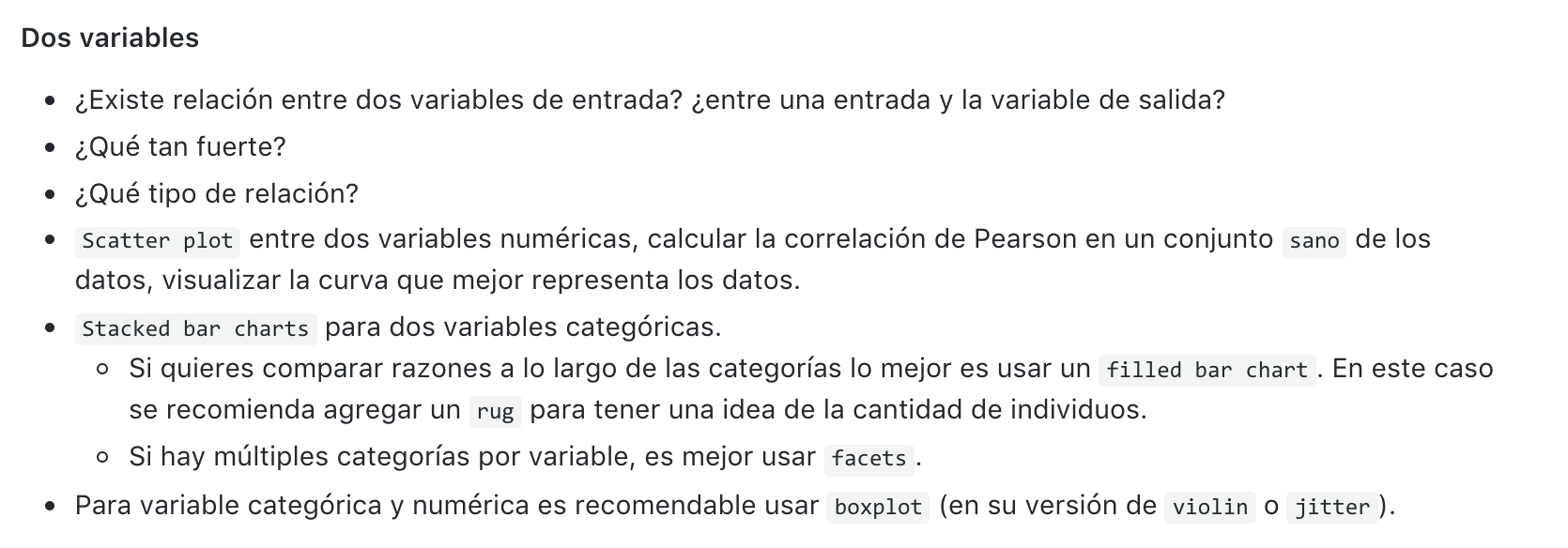
****

**10. ¿Qué podemos encontrar durante el análisis bivariado cuantitativo y en el categórico?**

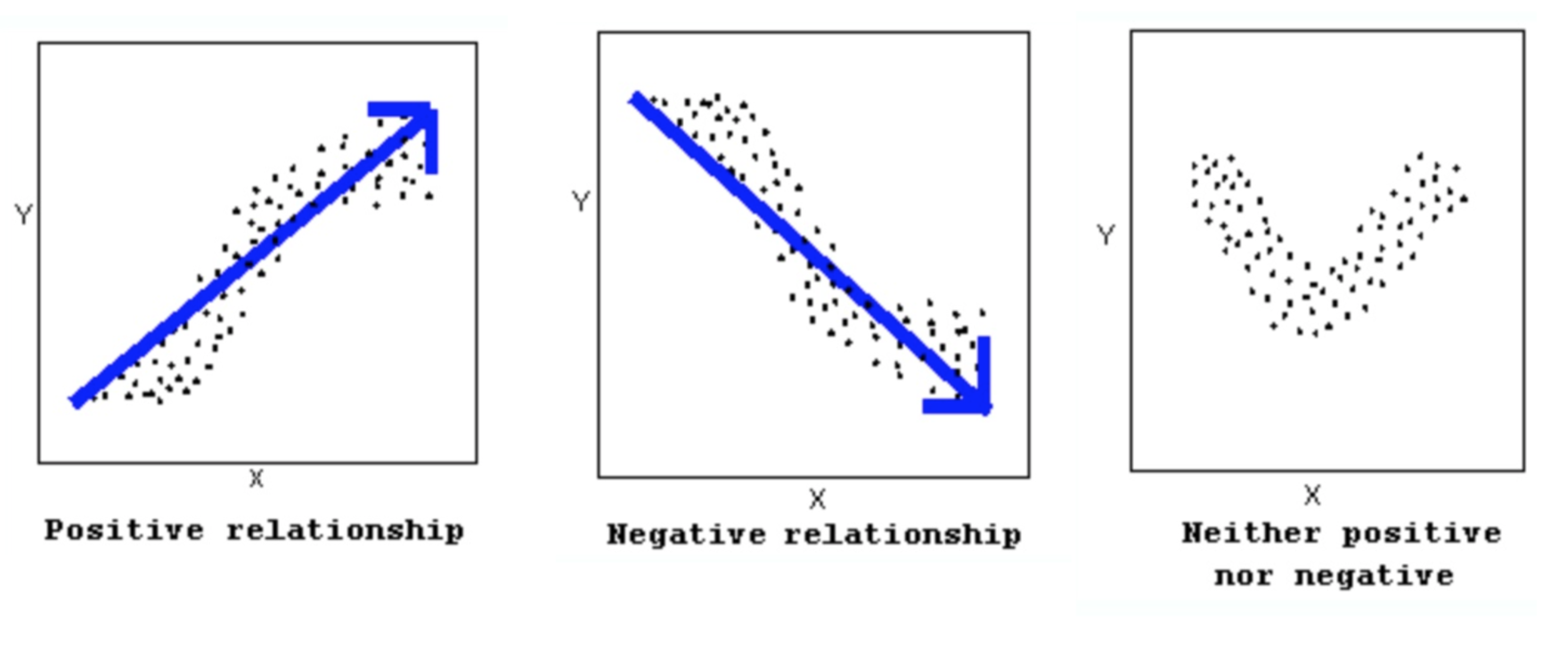
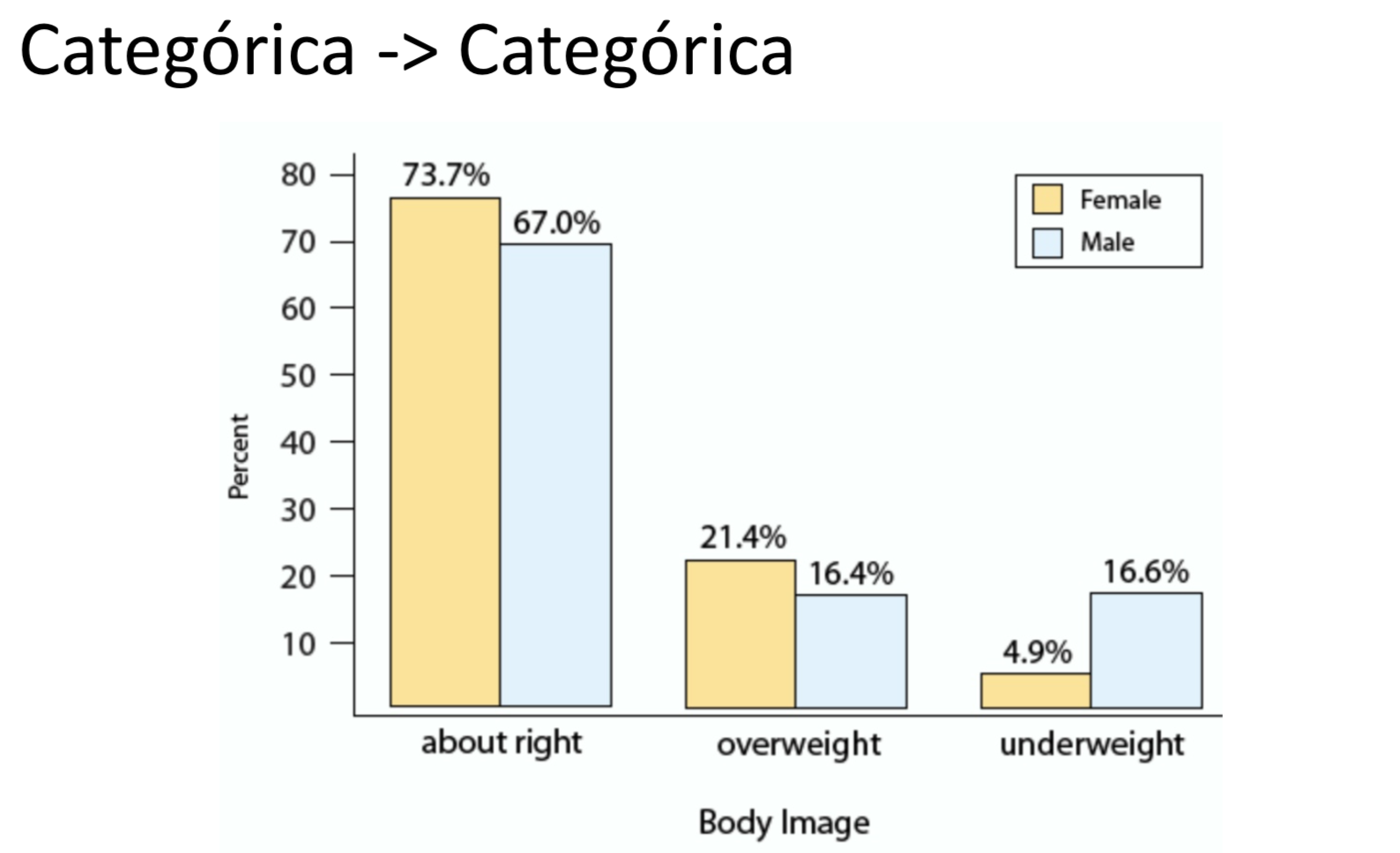
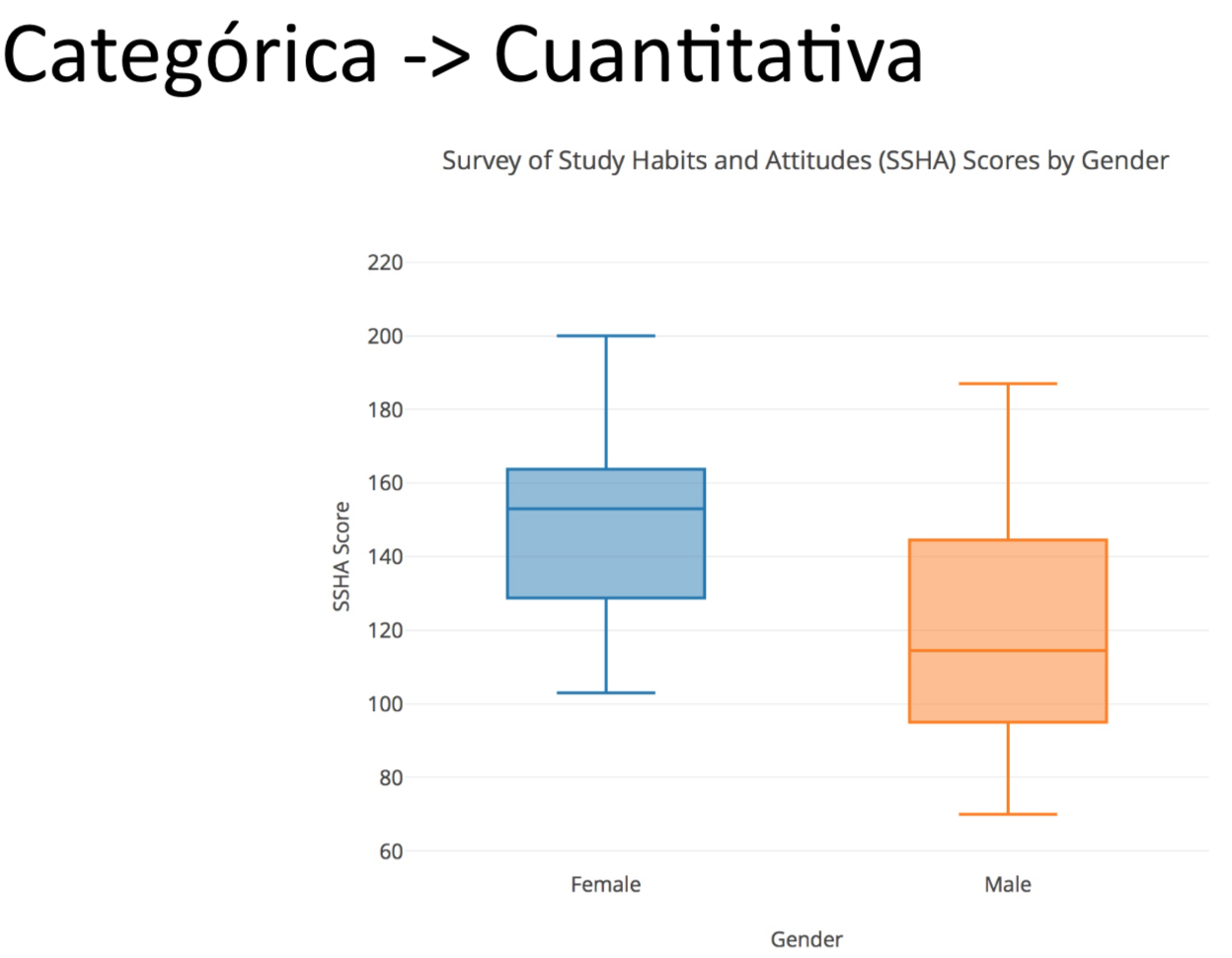
****

****

****

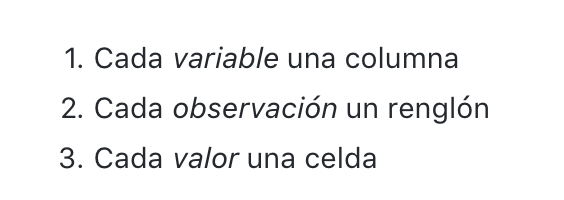
****

**11. ¿Qué tipo de gráficas se utilizan comúnmente en el análisis bivariado cuantitativo y categórico y con qué fin?**

****

**12. Describe las características de un conjunto de datos tidy**

Los datos Tidy, tienen tres caracteristicas:



**13. Describe el proceso de validación cruzada. ¿Para qué sirve?**

El proceso de validación cruzada consiste en tomar de un mismo set de datos, muestras aleatorias de n cantidad de registros cada una, y de esta manera hacer las pruebas de los modelos con combinaciones diferentes de esos sets de datos.

Por Ejemplo: si mi set de datos X = {x1, x2, x3, x4}, donde cada subconjunto xi tiene por ejemplo 5,000 registros. Voy a armar mi set de datos entrenamiento con combinaciones de 3 de esos sets, y el restante seria el set de pruebas. Se toman métricas del o de los modelos ajustando con cada set de datos, y de esta manera podemos sacar el mayor provecho de un modelo. Este proceso además, ayuda a que no sobre ajustemos o sub ajustemos un modelo. Se hace uso de los datos de la mejor manera posible, a mi criterio, ya que siempre los datos son un bien escaso.

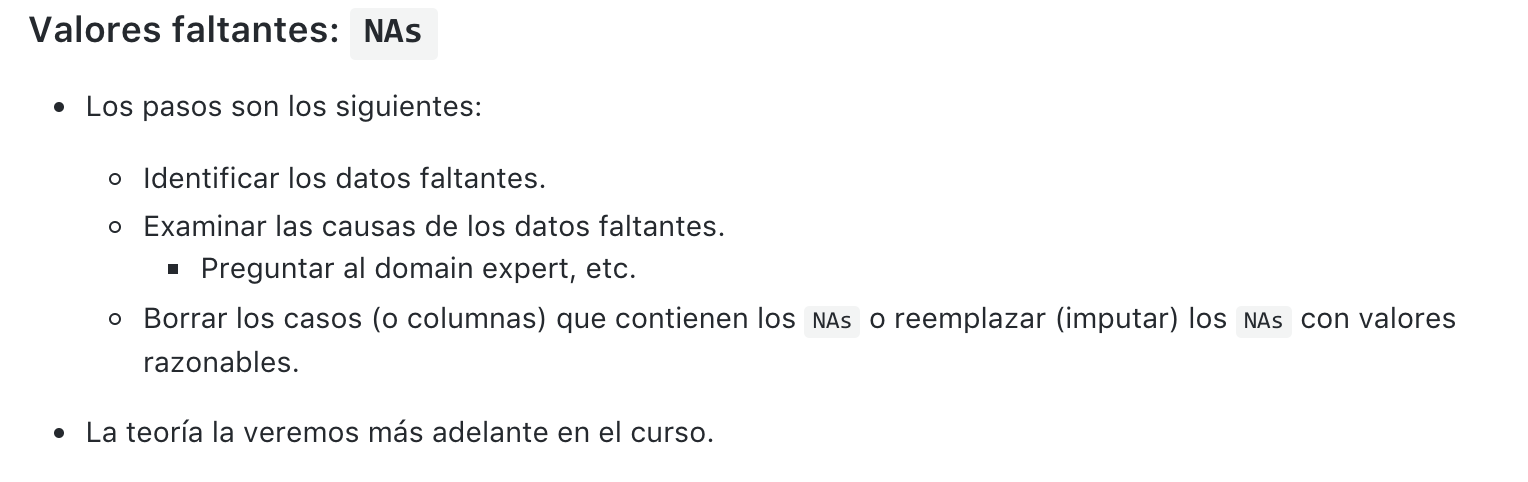
**14. ¿Qué es la maldición de la dimensionalidad?**

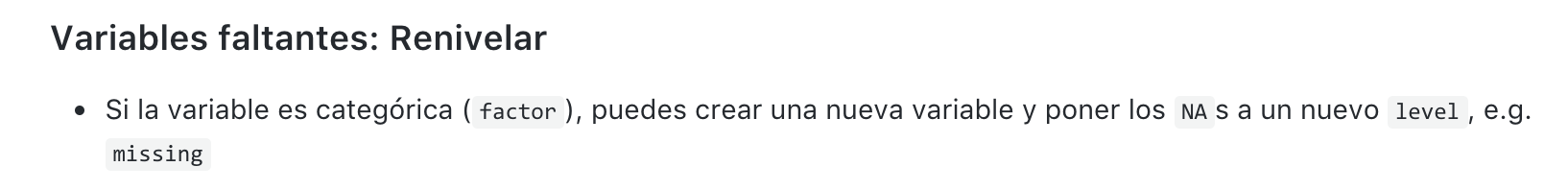
La maldición de la dimensionalidad se refiere a que en muchas oportunidades, debemos evaluar con mucho cuidado la cantidad de variables que vamos a utilizar en un modelo de datos, ya que a veces nos vemos tentados a usar más de las necesarias, y esto trae aparejado varios problemas, como por ejemplo, el conjunto de datos representativo de cada “unico” ejemplo de cada registro. Esto a su vez tiene grandes consecuencias en el tamaño del set de datos y por ende, en su procesamiento.

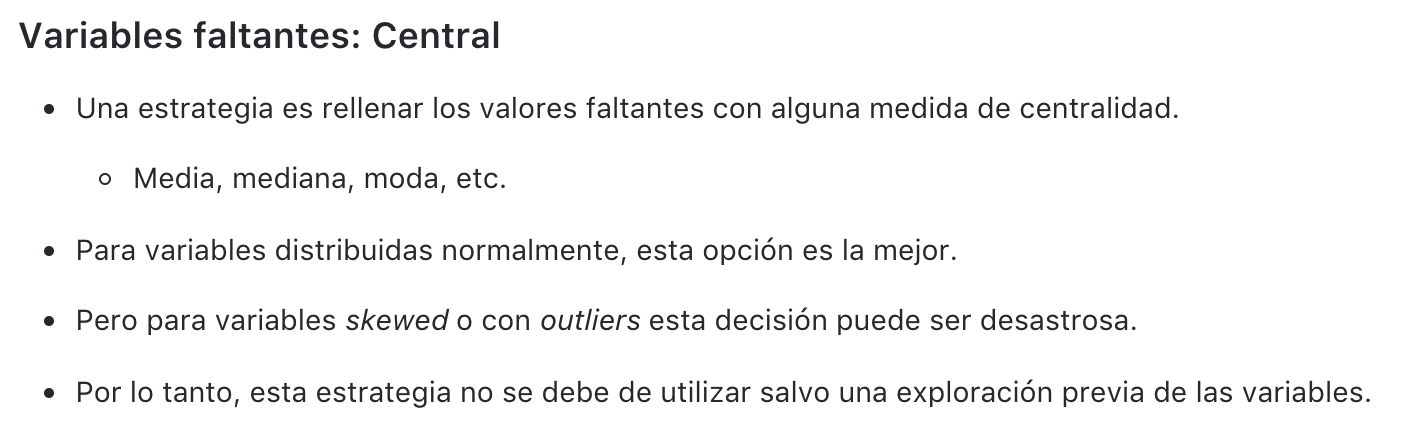
**15. ¿En qué consiste el efecto Rashomon? ¿Qué método puede evaluar de mejor forma la importancia de las variables?**

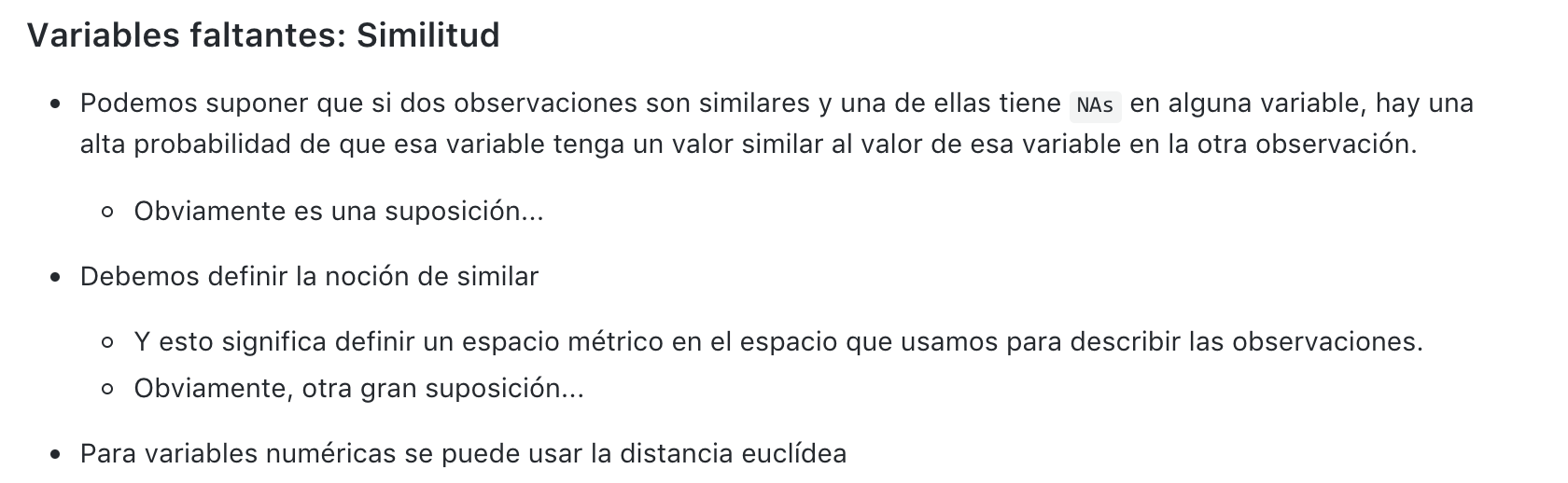
Many different models describe a situation equally accurate.

**16. Define tres formas de resolver el problema de variables faltantes indicando cuándo es conveniente utilizar cada una de ellas.**

****

****

****

****

**17. ¿Qué es aprendizaje supervisado, no supervisado y por refuerzo? ¿qué diferencia existe entre cada uno de ellos? ¿cuándo los uso?**

El aprendizaje se cataloga como supervisado o no supervisado dependiendo de si el algoritmo debe contar o no con información especifica de datos satisfactorios para el objetivo del aprendizaje

**No Supervisado**

Encargado de detectar factores no fácilmente observables dentro de los datos. Se utiliza principalmente en minería de datos. Sospechamos que existe algún tipo de relación entre los datos, pero es muy complejo intentar adivinarla.

Algoritmos

– Agrupamiento (clustering) – Componentes principales

– Modelo de mezclas (mixture models)

**Supervisado**

Utilizan una serie de variables y resultados observables asociados para predecir los resultados cuando estos no se conocen. Se utilizan ampliamente en analítica predictiva

Algoritmos

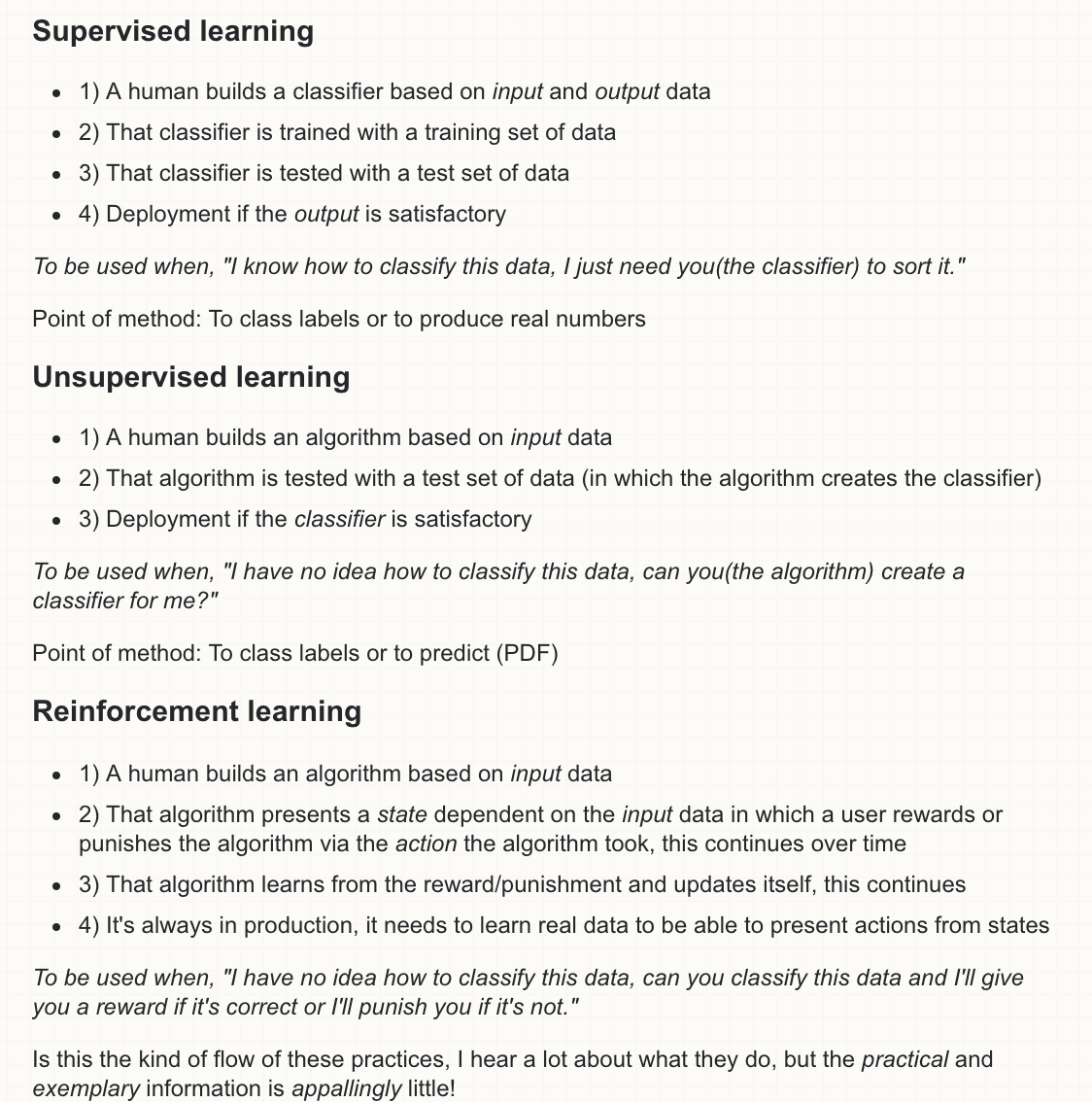
– Método de Bayes básico

– Regresiones lineales y no lineales – Redes neuronales

– Máquinas de soporte vectorial

– Árboles de decisión

– K-vecinos cercanos



**18. ¿Qué diferencia existe entre PCA y t-SNE? ¿Cuándo los utilizo?**

Ambos son algoritmos para reducer la dimensionalidad de los datos que voy a modelar.

PCA es una técnica matemática de reducción de dimensionalidad, mientras que TSNE convierte afinidad de puntos en probabilidades:

* PCA (linear)
* t-SNE (non-parametric/ nonlinear) usa hiperparámetros

The best way to use the algorithm is to use it for exploratory data analysis. It will give you a very good sense of patterns hidden inside the data. It can also be used as an input parameter for other classification & clustering algorithms.

**19. ¿Cómo puedo utilizar singular value decomposition para realizar clustering?**

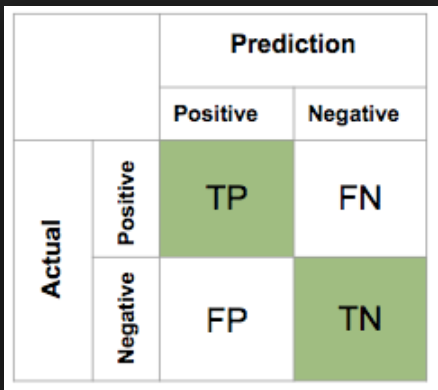
Clustering with Singular Value Decomposition (CSVD) method, which combines clustering and singular value decomposition (SVD) to reduce the number of index dimensions, while maintaining a reasonably high precision for a given value of recall. In the proposed CSVD method, homogeneous points are grouped into clusters such that the points in each cluster are more amenable to dimensionality reduction than the original dataset.

Reducir dimensiones de la matriz manteniendo las distamcias.

**20. ¿Para qué se utiliza una matriz de confusión? ¿Sólo sirve para clasificación binaria?**

Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases.

No, se puede extender para más de dos clases.

****

**21. ¿Qué es un micro-average?¿un macro-average? ¿Qué diferencia hay entre ellos? ¿Cuándo se utilizan?**

Evaluación de modelos de clasificación: **Accuracy**

¿Qué tan frecuente el clasificadorhace la predicción correcta?

El accuracy es un ejemplo de **micro-average**

La Matriz de confusion permite calcular el **accuracy de cada clase,** un ejemplo de un **macro-average**

A macro-average will compute the metric independently for each class and then take the average (hence treating all classes equally), whereas a micro-average will aggregate the contributions of all classes to compute the average metric. In a multi-class classification setup, micro-average is preferable if you suspect there might be class imbalance (i.e you may have many more examples of one class than of other classes). If your dataset varies in size, micro-average is the useful tool.

**22. ¿Por qué es importante separar los datos en entrenamiento y prueba antes de realizar el modelo? ¿Qué es el conjunto de validación?**

Para evitar el Overfitting. The models become very specific to the training data, so far as they even get trained on the occasional erroneous labels present in the training data. As a result, over fitted models don't work at well on the other data outside those in your actual data.

Training dataset

A training dataset is a dataset of examples used for learning, that is to fit the parameters (e.g., weights) of, for example, a classifier. Most approaches that search through training data for empirical relationships tend to over fit the data, meaning that they can identify apparent relationships in the training data that do not hold in general.

Test dataset

A test dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset. If a model fit to the training dataset also fits the test dataset well, minimal overfitting has taken place. A better fitting of the training dataset as opposed to the test dataset usually points to overfitting.

Training Dataset: The sample of data used to fit the model.

Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset **while tuning model hyperparameters.** The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

**23. ¿Por qué se dice que algoritmos como random forest, redes neuronales y máquina de vectores de soporte son cajas negras?**

Model interpretation is one of the key aspects of the model evaluation process. The explanation of the relationship between model variables and outputs is easy for statistical models, such as linear regressions, thanks to the availability of model parameters and their statistical significance.

For “black box” models, such as random forest, this information is hidden inside the model structure

**24. ¿Cómo puedo comprender los modelos generados por este tipo de algoritmos?**

Como una perturbacion en los datos de entrada afecta las salidas o analisis de importancia relative

A traves de un analisis de la importancia de variables var inplot R para árboles

**25. ¿Qué tipo de modelo usarías si tienes pocas observaciones y un número mayor de variables? ¿Por qué? ¿Cómo sé que puede funcionar?**

SVM reduce la maldición de la dimensionalidad y con kernels que relajan el requerimiento de la separabilidad lineal

**26. ¿Qué es el mean decrease accuracy? ¿Para que sirve?**

Es una medida de la importancia de cada variable en el modelo, y se prueba excluyendo cada variable y viendo cuanto se ve afectada el accuracy. Cuanto más afecta, mas importante es.

**27. ¿Qué es Caret? ¿Para qué funciona?**

**28. ¿Qué es doSnow? ¿Para qué funciona?**

**29. ¿Qué es un método de ensamble? ¿por qué y cómo funciona? ¿Sólo pueden hacerse ensambles del mismo tipo de algoritmo? Si/no por qué**

**30. ¿Cómo se resuelve un ensamble de regresión? ¿Cómo uno de clasificación?**

Considerando multiples hipotesis, es decir varios modelos de regression que tomen diferentes variables y con ellas ajusten major, para finalmente desembocar todos esos modelos intermedios en uno final que explique todo el modelo inicial

De regression cuando es una variable numerica continua, el de clasificación usa variables categoricas maximizar accuracy.

**31. ¿Qué es ICE? ¿Cómo funciona?**

**32. ¿Qué es LOCO? ¿Cómo funciona? ¿tiene algo en común con random forest?**

**Leave one covariate out**

**33. ¿Cuándo usaría Bayes Ingenuo? ¿por qué?**

**34. ¿Para qué sirve el truco del kernel en las máquinas de vectores de soporte?**

**35. Compara la cantidad de parámetros requerida para realizar predicciones con una regresión lineal, una red neuronal y K vecinos cercanos**

**36. Menciona dos métodos para representar numéricamente las variables categóricas. ¿Cuándo uso cada uno de ellos?**

**37. ¿Cuándo es conveniente convertir variables continuas a categóricas?**

**38. ¿En qué consiste la ingeniería de características?**

**39. Describe el proceso manual para generar características.**

**40. Describe el proceso automático para generar características.**

**41. ¿Cómo puedo crear una variable binaria a partir de una continua? ¿Cómo decido el punto de corte?**

**42. ¿Para qué sirve la selección de características?**

**43. ¿Qué es el filtrado basado en las propiedades de la distribución? Da un ejemplo**

**44. ¿En qué consiste el filtrado por correlación?**

**45. ¿En qué consiste el Fast-correlation filtering? Describe el algoritmo**

**46. ¿En qué consiste el Forward Selection?**

**47. ¿En qué consiste el Backward Selection?**

**48. ¿Cómo podemos usar los filtros ANOVA para seleccionar variables?**

**49. Describe el método de selección Epsilon tanto para variables categóricas como numéricas**

**50. ¿Cómo puedo utilizar los métodos de clustering para realizar ingeniería de características.**

**51. ¿Qué son precision, recall, f1 score? ¿Cómo se calculan? ¿Cuándo utilizo cada una de esas medidas?**

**52. ¿Qué es el área bajo la curva AUC? ¿Qué mide?**

**53. ¿Cómo se construye la curva ROC?**

**54. ¿Cuándo es importante utilizar distintos puntos de corte para clasificar?  da dos ejemplos.**

**55. ¿Cómo interpreto la gráfica de doble densidad?**

**56. ¿Qué mide la R^2? ¿Qué problemas tiene**

**57. ¿Qué mide el criterio de Akaike?**

**58. ¿Cómo puedo determinar si un modelo de clustering es adecuado?**

**59. ¿En qué consiste el K-fold cross-validation? ¿Para qué sirve?**

**60. ¿Se crean los modelos finales con cross-validation? Si/no y por qué.**

No, cross validation me sirve para calcular los hiperparametros (lambda) y luego, con todo el set de entrenamiento completo - no como en cross-validation, habiendo calculado la lambda - calculamos las betas de los parametros.

**61. ¿En qué consiste el leakage? Da un ejemplo**

**62. Da un ejemplo de cómo Caret puede ayudar en la selección de modelos**

**63. ¿Cuál es la filosofía UNIX para procesamiento de información?**

**64. Principales comando en UNIX**

**65. ¿En que se parecen los “|” en UNIX con “%>%” en R**

**66. ¿Qué diferencia hay entre una contenedor y una imagen en docker?**

**67. ¿Para qué sirve docker?**

**68.   ¿Cómo deep learning puede apoyar en el proceso de ingeniería de características?**

**69.   ¿Qué es lo que realizan las capas intermedias de las redes neuronales?**

**70.   Revisar el contenido de los artículos para realizar presentaciones ejecutivas y cómo contar una historia en ellas**

**71.   ¿Puedo utilizar un q-q plot para determinar si las distribuciones de dos variables continuas son semejantes? ¿Cómo?**