

Módulo 3 – Análisis Exploratorio y Programación Estadística

# Análisis Exploratorio de Datos

Ciencia de Datos

# Objetivos



- Reconocer el ciclo de vida de los problemas analíticos.
- Describir el concepto de Análisis Exploratorio.
- Reconocer qué es un Insight.

# Ciclo de vida de un problema de Ciencia de Datos



Comprender las necesidades del negocio

Formular la pregunta

Generar las hipótesis del trabajo

Recolección de datos

Procesamiento y limpieza de los datos (Data Wrangling)

Comprensión de los datos (Exploratory Data Analysis)

Feature Extraction

Feature Engineering

Entrenamiento del modelo

Evaluación y selección del modelo

Optimización del modelo

Presentación al negocio del modelo

Visualización de resultados del modelo

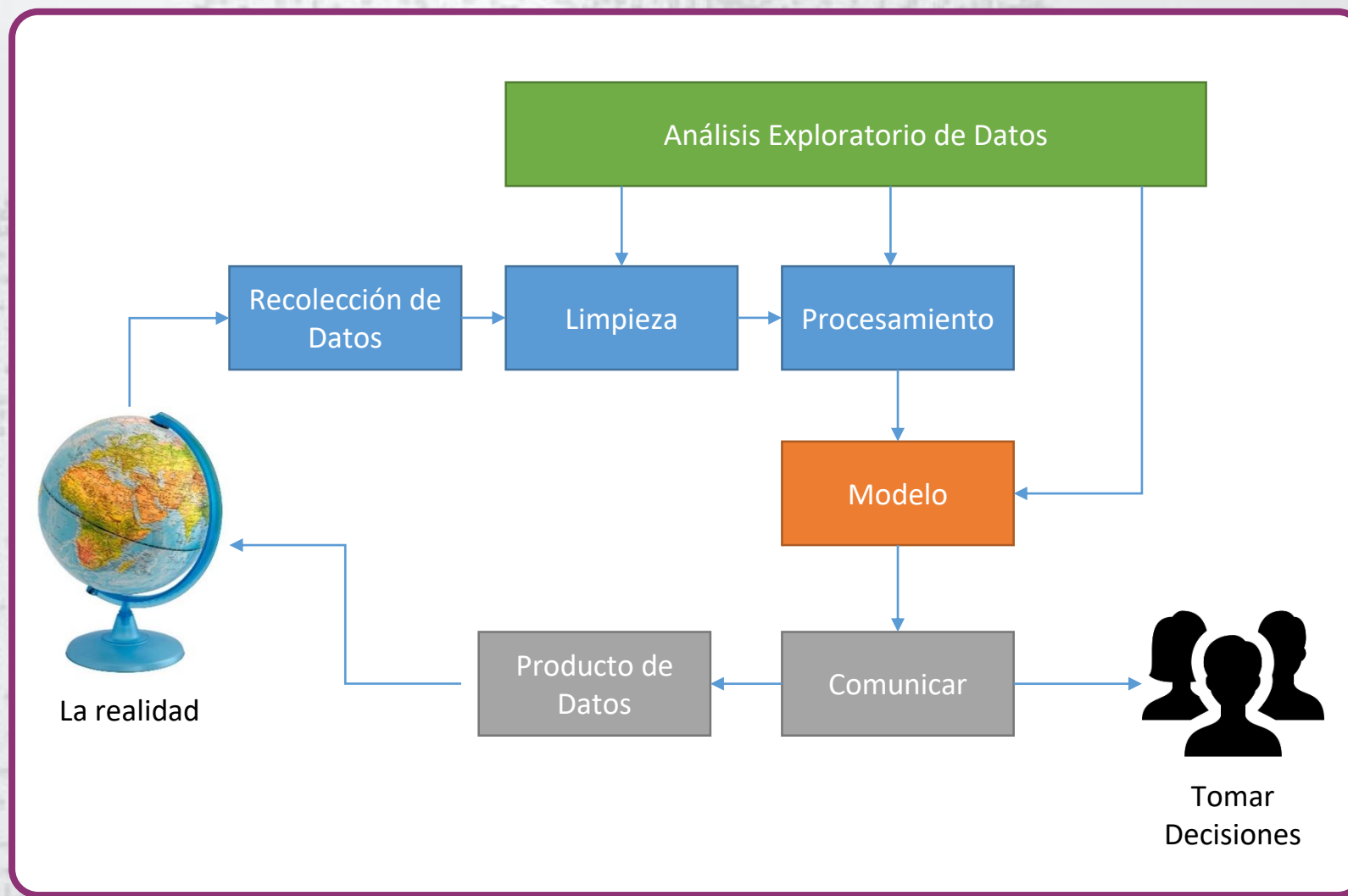
Despliegue productivo del modelo

Reimplementación de procesos

Predicciones



# Análisis Exploratorio de Datos



- En palabras simples, es una aproximación en el análisis de sets de datos para sumarizar sus características, a menudo valiéndose de métodos visuales.
- Utilizado también en etapas de preparación de datos en labores principalmente de reconocimiento y limpieza
- Utilizado para chequear hipótesis de trabajo en las etapas de modelamiento

# Análisis Exploratorio de Datos

# Análisis Exploratorio de Datos

*“Procedimientos para el análisis de datos, técnicas para la interpretación de los resultados de dichos procedimientos, vías para planear la recolección de datos que hagan el análisis más fácil, más preciso y exacto, y toda la batería de herramientas estadísticas que nos permitan analizar los datos”*

*John Tukey - 1961 – “The future of data analysis”*

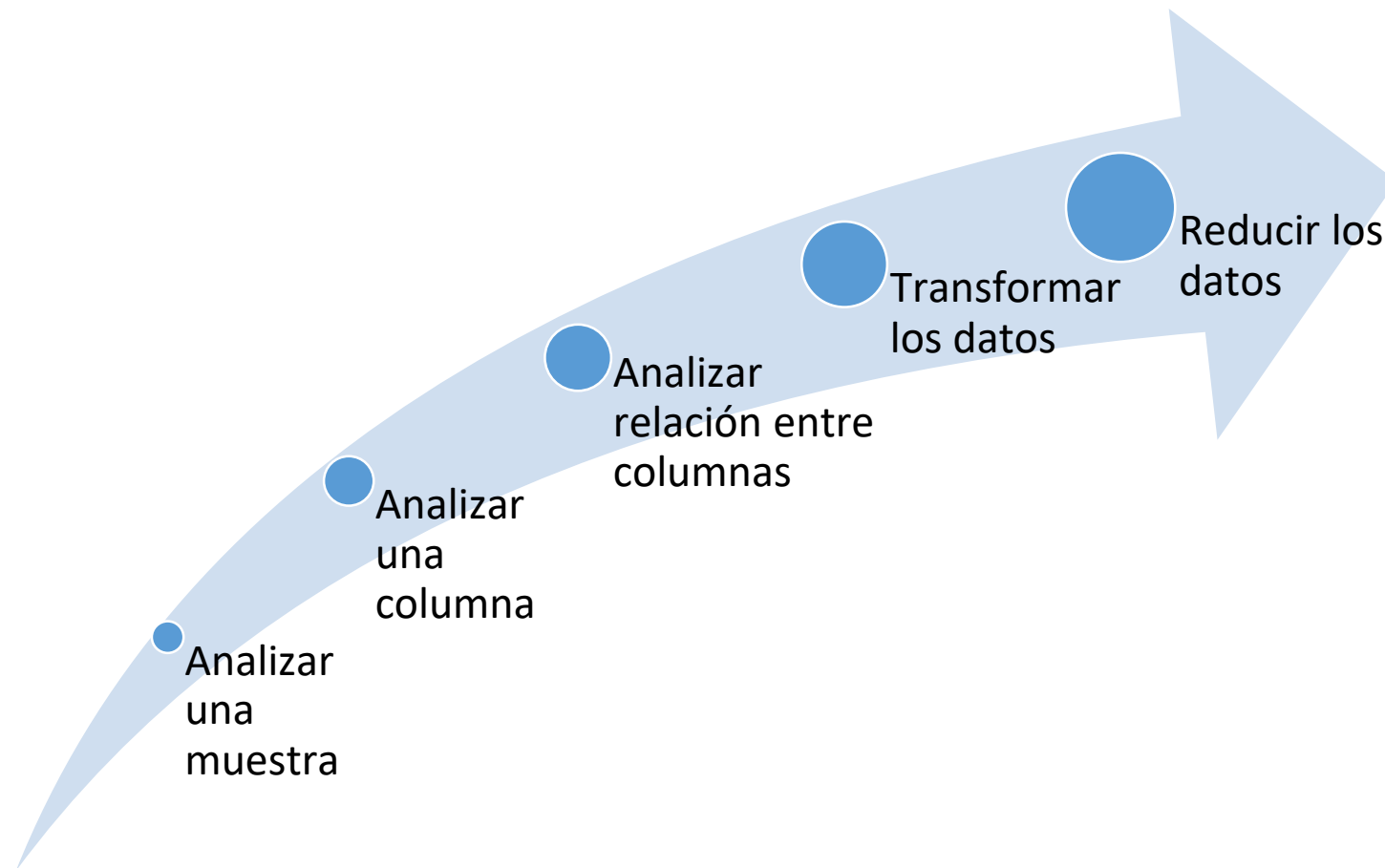
# Análisis Exploratorio de Datos

## EDA – Exploratory Data Analysis

### IDA – Initial Data Analysis

- Manejar valores perdidos y data errónea.
  - Detectar anomalías y outliers en los datos.
  - Identificar las variables más importantes.
  - Mapear la estructura de los datos.
  - Chequear la completitud.
  - Asegurar que los datos son apropiados para el caso de uso.
- Sugerir hipótesis acerca de la causa de un fenómeno observado.
  - Realizar/checkear supuestos para una posterior inferencia estadística.
  - Apoyar la selección de herramientas y técnicas estadísticas.
  - Proveer la base para futuros procesos de recolección de datos.

# Técnicas y Herramientas





# En búsqueda de Insights

En este ejemplo, se está analizando el sueldo de los empleados del ayuntamiento de San Francisco entre los años 2010 y 2014.

A partir de esta información:

- ¿Es posible hacerse una idea de cómo son los sueldos en dicho lugar?
- ¿Hay algo que llame la atención?.

```
# resumen de estadísticas  
df['BasePay'].describe()
```

Python

count	148045.000000
mean	66325.448841
std	42764.635495
min	-166.010000
25%	33588.200000
50%	65007.450000
75%	94691.050000
max	319275.010000
Name: BasePay, dtype: float64	

# En búsqueda de Insights

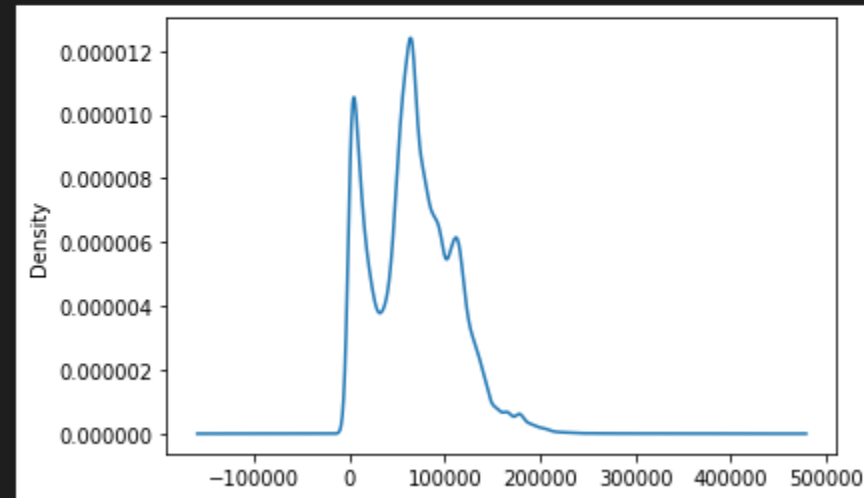
Ahora revisemos la distribución de sueldos para el mismo caso.

- ¿Es consistente con la información anterior?
- ¿Hay algo que llame la atención?.

```
df['BasePay'].plot(kind='kde')
```

Python

<matplotlib.axes.\_subplots.AxesSubplot at 0x19d25a18248>



# En búsqueda de Insights



¿Qué es un insight en el contexto de análisis de datos?



En el contexto del análisis de datos, un insight se refiere a una observación o conclusión valiosa y significativa que se obtiene al analizar y procesar grandes cantidades de datos. Es una percepción inesperada que proporciona una comprensión más profunda de los datos y puede ser utilizada para tomar decisiones informadas.

Por ejemplo, en el análisis de datos de ventas de una empresa, un insight podría ser que un producto en particular tiene un aumento en las ventas durante un cierto período de tiempo, y se puede atribuir a un evento o campaña publicitaria específica. Este insight podría ser utilizado para mejorar la estrategia de marketing de la empresa y generar más ventas en el futuro.

En el análisis de datos, un insight puede ser obtenido a través de técnicas como la minería de datos, el aprendizaje automático y la visualización de datos. Estas técnicas ayudan a identificar patrones, tendencias y correlaciones en los datos que de otra manera podrían pasar desapercibidos. El objetivo final de descubrir insights en el análisis de datos es ayudar a las organizaciones a tomar decisiones más informadas y mejorar su rendimiento.

```
# resumen de estadísticas  
df['BasePay'].describe()
```

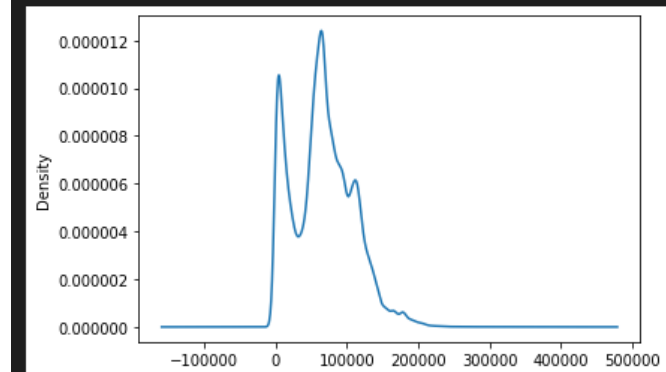
Python

```
count    148045.000000  
mean      66325.448841  
std       42764.635495  
min       -166.010000  
25%       33588.200000  
50%       65007.450000  
75%       94691.050000  
max       319275.010000  
Name: BasePay, dtype: float64
```

```
df['BasePay'].plot(kind='kde')
```

Python

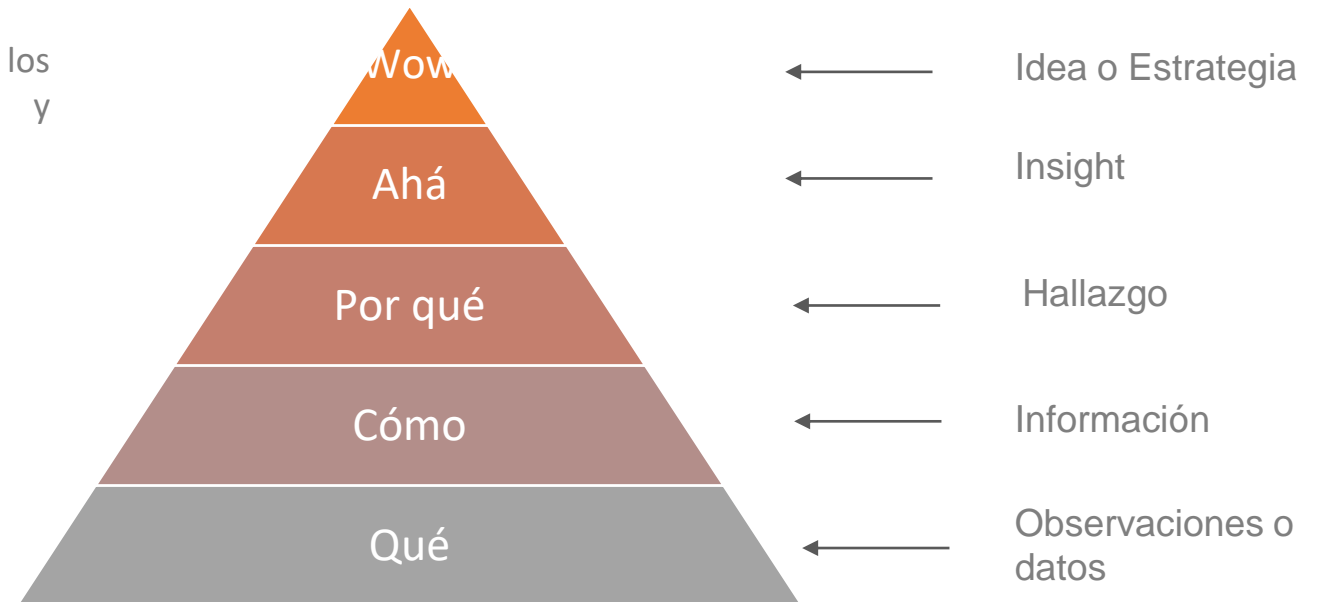
<matplotlib.axes.\_subplots.AxesSubplot at 0x19d25a18248>



# Importancia de encontrar Insights

Un ejemplo importante es la obtención de insights respecto al **comportamiento de los clientes**. Esto proporciona a la empresa una comprensión valiosa que permite adaptar sus estrategias y operaciones para aprovechar al máximo este patrón de compra específico.

La base del éxito está en la capacidad de entender a los clientes, sus verdaderas necesidades, expectativas, y traducirlas en estrategias.





# Importancia de encontrar Insights

Supongamos que una empresa de comercio electrónico está analizando los datos de sus clientes y descubre que **el 80% de las compras se realizan los viernes por la tarde**. Este descubrimiento podría llevar a la empresa a tomar varias acciones estratégicas:

- Programar campañas de marketing específicamente dirigidas a los clientes los viernes por la tarde, aprovechando el momento en que están más propensos a realizar compras.
- Ofrecer descuentos especiales o promociones exclusivas los viernes por la tarde para incentivar aún más las compras.
- Ajustar los horarios de atención al cliente o aumentar la disponibilidad de personal los viernes por la tarde para garantizar un servicio de alta calidad durante los momentos de mayor actividad.
- Optimizar el inventario y la logística para garantizar que los productos estén disponibles y puedan entregarse rápidamente a los clientes que realizan compras los viernes por la tarde.



# Dudas y consultas

Fin de la Presentación