CIBERTEC

# PYTHON FUNDAMENTALS FOR DATA SCIENCE

## Capítulo 4: Preprocesamiento de datos en Python

# OBJETIVOS

- Utilizar la librería Pandas.
- Aplicar el preprocesamiento de datos, previo a llevar a cabo actividades de machine learning.

# AGENDA

1. El Data Scientist.
2. Metodología Data Science.
3. NumPy.
4. Pandas.
5. Matplotlib.

# 1. EL DATA SCIENTIST

- Es sexy ser un Científico de Datos -

# Data Scientist:

## *The Sexiest Job of the 21st Century*

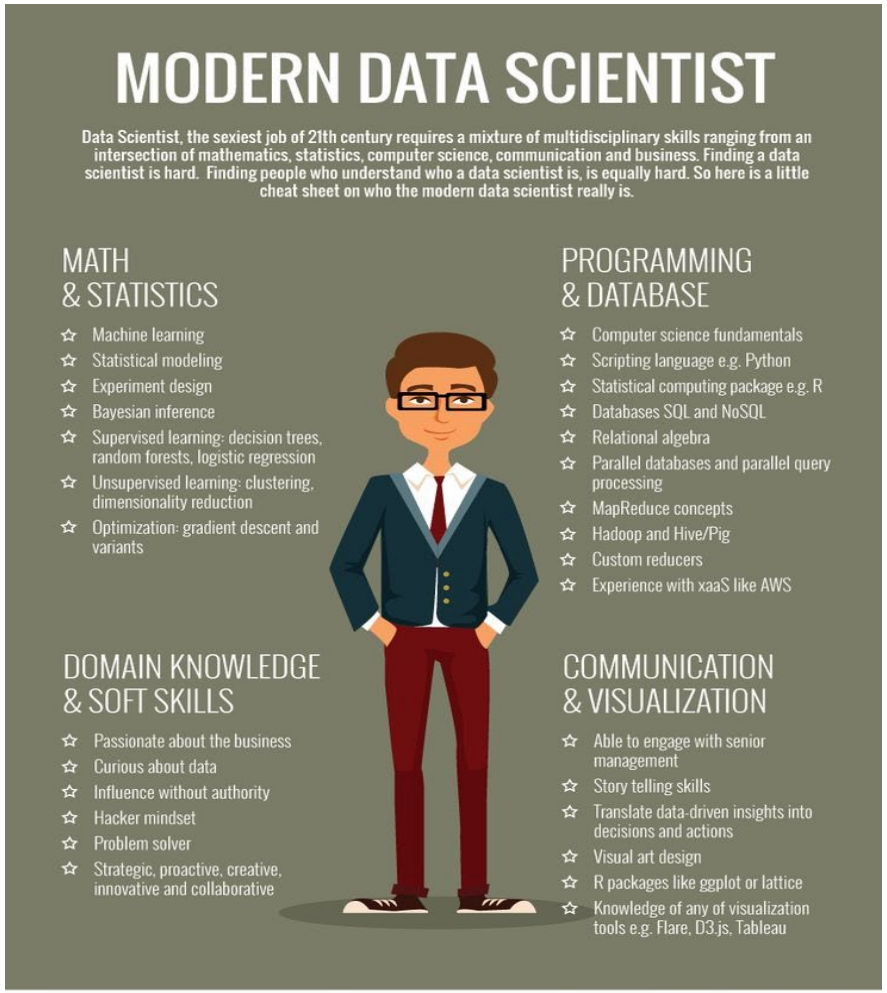
***Meet the people who can coax treasure out of messy, unstructured data.***

*by Thomas H. Davenport and D.J. Patil*

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

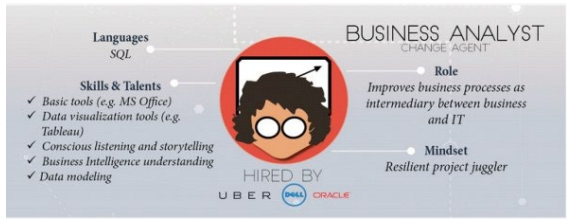
# 1. EL DATA SCIENTIST

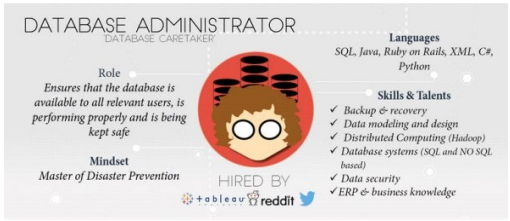
- Competencias de un Científico de Datos -



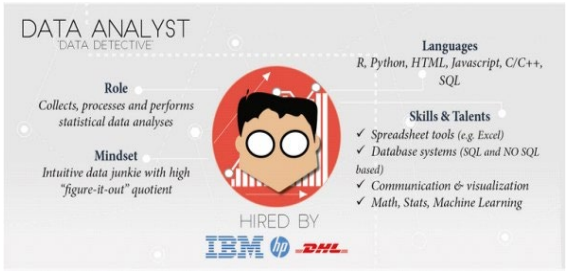
# 1. EL DATA SCIENTIST

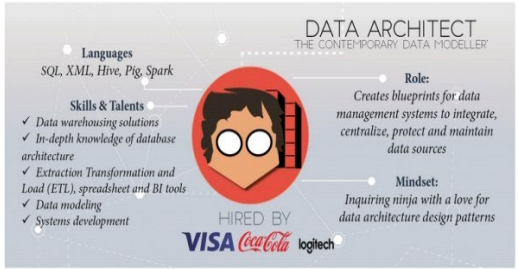
- Existen muchos Roles -

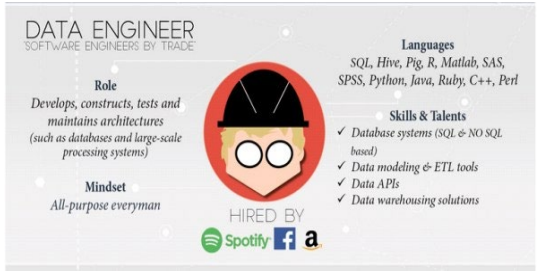


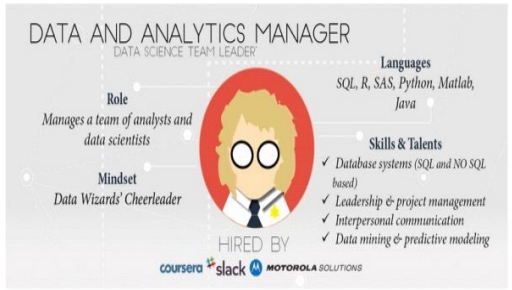







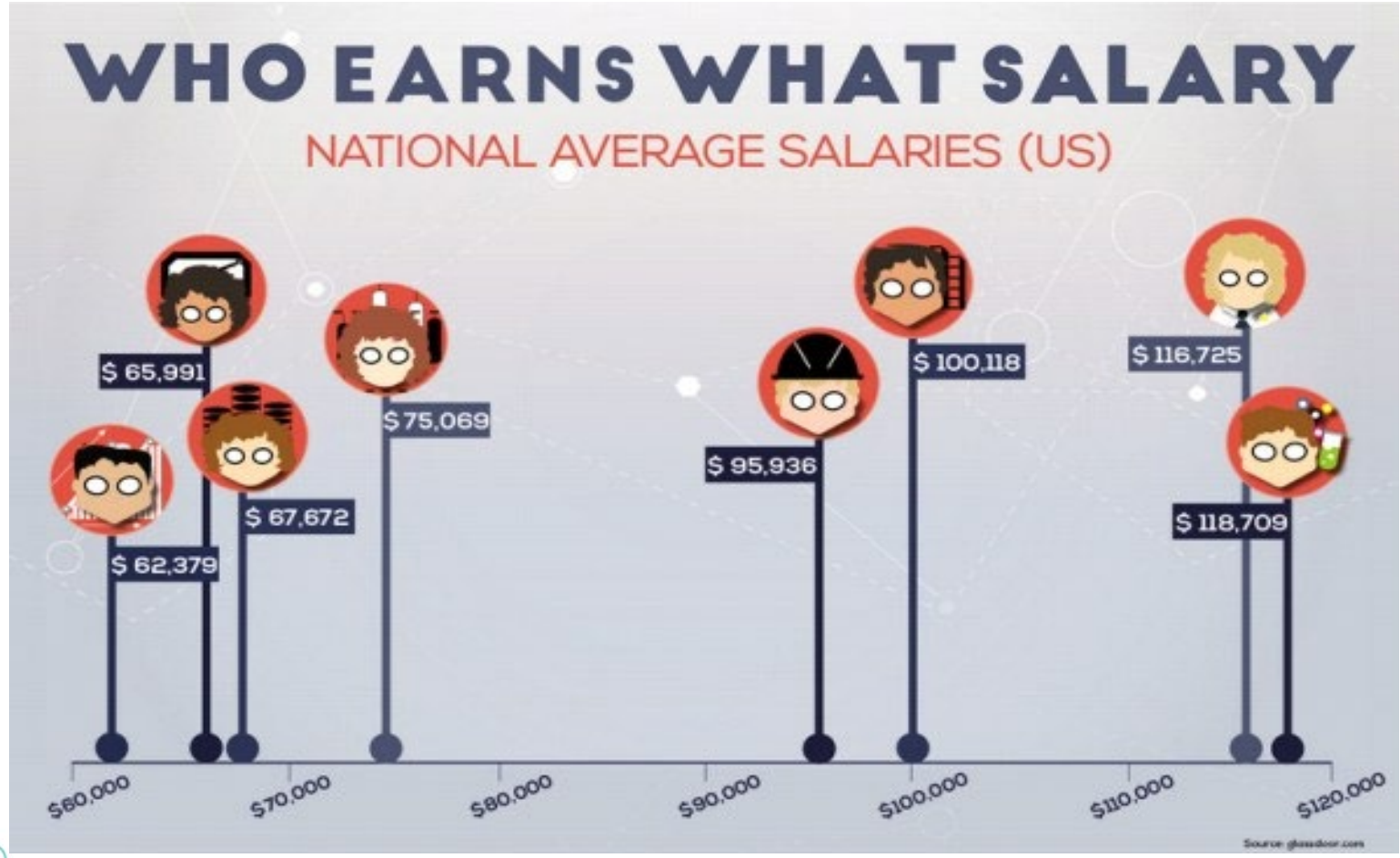





Fuente: kdnuggets

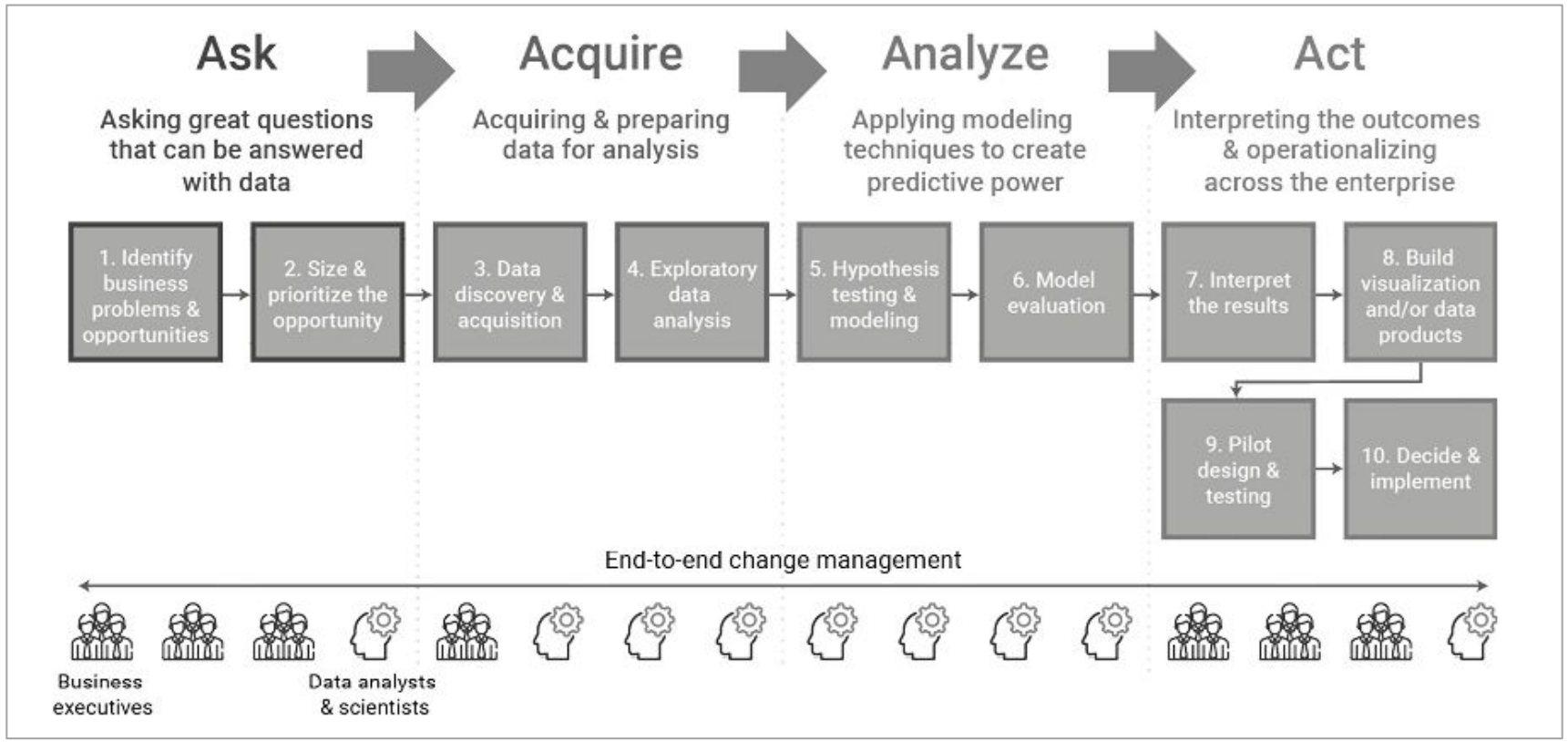
# 1. EL DATA SCIENTIST

- Sueldos Promedio (EEUU) -




Fuente: kdnuggets

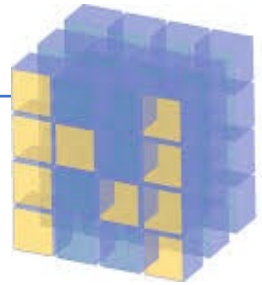
# 2. METODOLOGÍA DATA SCIENCE




Fuente: Kaldero (2018). *Data Science for Executives.*

# 3. NUMPY

- Una de las librerías principales de Data Science en Python.
- Prerrequisito para Pandas.
- Procesamiento de datos y operaciones de algebra lineal.



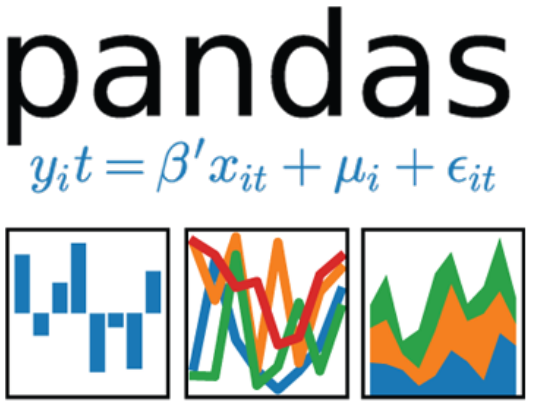
NumPy

# 4. PANDAS

- Fundamental para la exploración de datos.
- Construido encima de NumPy.
- Soporte para diversas fuentes de datos.
- Se crea una especie de hoja de cálculo en memoria llamada DataFrame.

Tareas:

- Limpieza de datos.
- Ingeniería de datos.
- Aplicar funciones a los datos.
- Creación de otras estructuras.



# 5. MATPLOTLIB

- Fundamental para la visualización de datos
- Integración con Pandas y otras librerías como Seaborn


Produce los siguientes tipos de gráficos:

- Líneas
- Barras
- Histogramas
- Scatterplot
- Piechart
- Boxplots

# LABORATORIO Nº 1: PANDAS

Al finalizar el laboratorio, el alumno logrará:

- Aplicar los fundamentos de NumPy.
- Aplicar los fundamentos de Pandas.

# LABORATORIO Nº 2: TITANIC

Al finalizar el laboratorio, el alumno logrará:

- Analizar la exploración de datos.
- Aplicar limpieza de datos.
- Aplicar transformaciones de datos.
- Aplicar estadísticas a los datos.
- Aplicar visualizaciones de datos.

# TAREA Nº 5: PANDAS

- Resolver los ejercicios en el Notebook Jupyter compartido.
- Enviar por **Notebook Jupyter** al correo del instructor.

# RESUMEN

En este capítulo, usted aprendió:

- Que Pandas es una herramienta fundamental para diversas tareas de preprocesamiento de datos, como lo es la limpieza de datos.
- Que el preprocesamiento de datos supone una actividad importante previa al machine learning.

# BIBLIOGRAFÍA


- Python. Python for beginners. https://www.python.org/doc/
- Scikit-learn. Biblioteca de aprendizaje automático. https://scikit-learn.org/stable/
- TensorFlow. Crea modelos de aprendizaje automático. https://www.tensorflow.org/?hl=es-419
- Kaggle. Comunidad de científicos de datos del aprendizaje automático. https://www.kaggle.com/


CIBERTEC