Curso Análisis exploratorio de datos en Python y R

Introducción y objetivos

Juan Camilo Perdomo

juan.perdomor@urosario.edu.co

Universidad del Rosario

Bogotá, Colombia

27 de julio de 2021

Tabla de contenidos

Análisis exploratorio de datos

El Análisis Exploratorio de Datos, conocido por sus siglas en inglés EDA (Exploratory Data Analysis), es considerado como el conjunto de procedimientos cuyo objetivo general es proporcionar una visión más detallada y precisa de la información (variables, indicadores, índices, entre otros) almacenada en un dataset, previo al uso de técnicas estadísticas inferenciales.

Se apoya en un planteamiento descriptivo y se realiza con una mentalidad “exploratoria”. Desde un punto de vista técnico, el EDA se caracteriza por el empleo de procedimientos analíticos y descriptivos de carácter gráfico o semigráfico, que muestren todas las particularidades y carácteres de las variables sacando a la luz las estructuras ocultas de los datos.(Rodríguez Jaume y Mora, 2001).

Análisis exploratorio de datos, una vista gráfica

De acuerdo con el portal "Towards Data Science", el análisis de datos luce de la siguiente manera:

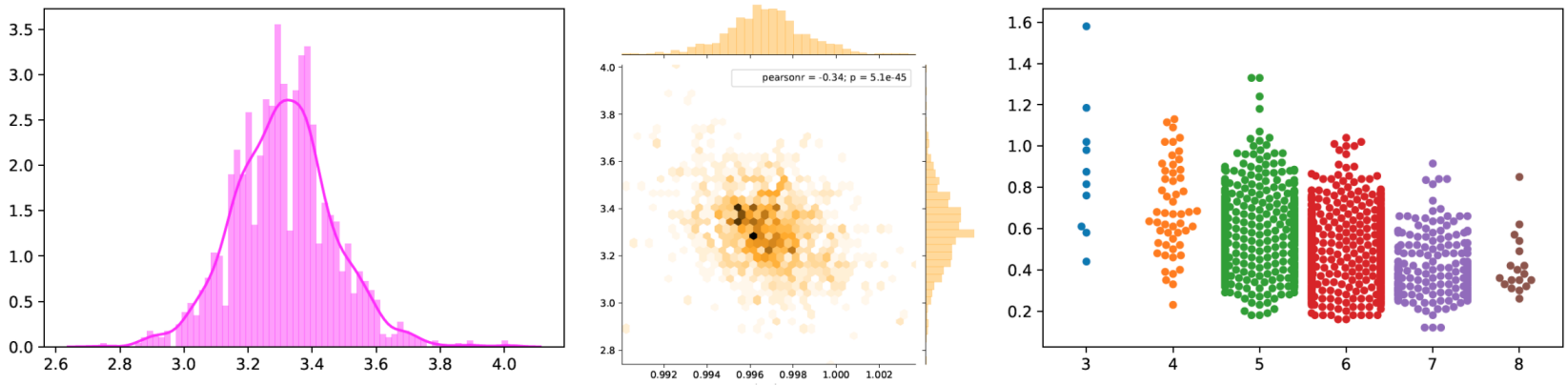

Figura: Tipos de gráficos para el EDA (Towards Data Science, 2018)

Análisis exploratorio de datos

En otras palabras, el análisis exploratorio de datos es un primer encuentro con los datos, donde se explora el número de estos, unidades de análisis, periodicidades, distribuciones, relaciones entre variables e información que va mucho más allá. Esto es posible hacerlo de una manera "manual", a través de Excel; sin embargo, existen softwares que permiten automatizar estos procesos y desarrollarlos de una manera más confiable y sencilla.

Además, es posible hacer uso de la estadística descriptiva (en el curso se verá la diferencia con la estadística inferencial) para conocer mejor nuestros datos: medidas de tendencia central, de dispersión y sus distribuciones.

Software estadístico para el análisis de datos

Existen diversos programas,aplicaciones, paquetes, softwares e, incluso, lenguajes de programación que permiten realizar exploración, análisis, cálculos, inferencia, ilustración, entre otros, de los datos.



PYTHON


Software estadístico para el análisis de datos

Este curso estará basado en los dos últimos (Python y R), dado que son de uso libre (gratuito) y, además, porque son los dos que están tomando, no solo mayor popularidad e importancia, sino que son las que más han crecido en los últimos años en términos de paquetes y librerías en el mundo del análisis y, en general, de la ciencia de datos.

PYTHON

Python / Jupyter Notebook

De acuerdo con el portal de Crehana, Python es un lenguaje de programación interpretado, multiparadigma y multiplataforma usado, principalmente, en Big Data, AI (Inteligencia Artificial), Data Science, frameworks de pruebas y desarrollo web. Esto lo convierte en un lenguaje de propósito general de gran nivel debido a su extensa biblioteca, cuya colección ofrece una amplia gama de instalaciones.

Su uso se da principalmente en una herramienta llamada Jupyter Notebook, cuya funcionalidad es interpretar el código en lenguaje Python (iPython) y mostrar sus resultados a pie de línea de una manera interactiva.

R / R-Studio

De acuerdo con su propio portal, R es un entorno de software libre para gráficos y computación estadística. Se compila y se ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOS, es también un lenguaje de programación, utilizado principalmente en la ciencia de datos y el análisis de datos.

Tal como en Python, R tiene un intérprete (R-Studio) que permite programar en este lenguaje y mostrar sus resultados en una interfaz de manera interactiva. Aunque, R también está siendo usado en otros sofware para edición de texto, como Visual Studio Code.

Bibliografía

Rodríguez-Jaume, María-José y Mora Catalá, Rafael (2001)

Análisis de regresión múltiple

Universidad de Alicante, Departamento de Sociología.

http://rua.ua.es/dspace/handle/10045/8143

Towards Data Science (2018)

What is Exploratory Data Analysis?

https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Crehana (2021)

¿Qué es Python? El lenguaje de programación más popular para aprender en 2021

https://www.crehana.com/co/blog/web/que-es-python/

r-project Org (2021)

The R Project for Statistical Computing

https://www.r-project.org/