Análisis exploratorio de datos

Sesión 02

Ing. Gómez Marín, Jaime[1]

Módulo 3 : Análisis de Datos con Python
Departamento de TdG

September 2019

Índice

Introducción

En este sesión se cubre los conceptos básicos del análisis de datos exploratorios con Python.
El análisis exploratorio de datos, (EDA), es un enfoque para analizar datos con el fin de:

- Resumir las características principales de los datos
- Obtener una mejor comprensión del conjunto de datos,
- Descubrir relaciones entre diferentes variables
- Extraer variables importantes para el problema que estamos tratando de resolver.

Teorema de Limite Central

La suma o promedio de cualquier conjunto de variable independientes generadas al azar se aproxima a la distribución normal

Estadistica Descriptiva

Cuando se empieza a analizar los datos, es importante explorar los datos antes de gastar mucho tiempo haciendo modelos. Una forma facil de hacer esto es calculando alguos valores de estadistica descriptiva para los datos.
El anáisis estadístico descriptivo ayuda a describir las caracteristicas básicas de los datos y obtener un breve resumen hacerca de la muestra y la mediciones de los datos.

Variables Númericas

En Pandas, los datagrama tiene incorporado la función describe() que nos permite obtener los principales estadisticos de los datos, para el cálculo se omiten los datos perdidos (NA).
Los estadisticos que calcula son: la media, el número total de datos, la desviación estandard ,los quantiles y los valores extremos.

Variables Categoricas

Los valores categoricos pueden ser agrupados en diferentes grupos y asignarles valores discretos.
En Pandas, los dataframen tiene incorporado la función value_counts() que nos permite asignar valores a las categorias

Boxplots

Los boxplots son una forma de visualizar los datos numericos, en el boxplot se pueden apreciar lo siguiente:

- Se visualiza la mediana
- El quantil mayor (Q3 = 75%)
- El quantil menor (Q1 = 25%)
- EL rango interquartilico (IQR) : Q3 - Q1
- Los outliers que estan localizados 1.5 veces el IQR encima o debajo del Q3 y Q1 respectvamente.
- Permite la comparación entre grupos.

Scatter plot

Las relaciones entre 2 variables puede ser representada en un diagrama de Scatter Plot.
La variable predictora es la variable que se usa para predecir una salida.

ANOVA : Análisis de la varianza

Se usa para analizar variables categoricas y ver las correlaciones para diferentes categorias.
ANOVA es una prueba estadistica que analiza las varianza entre varias categorias de una variable categorica.
ANOVA tambien puede ser usado para encontrar la correlación entre diferentes grupos de categorias de un variable categorica.
La pruebas de ANOVA devuelve el valor del estadistico de la prueba F-test y su p-value.
El estadistico de la prueba devuelve la tasa de variación entre la media del grupo que se analiza.
El p.valor muestra si el resultado es estadisticamente signitifcativo
La F-test calcula la razón de variación entre la media de los grupos sobre la variación en cada uno de los grupos de muestra.

Correlación

La correlación es un medida estadistica que nos permite medir la relación entre variables que son supuestamente independientes. Es decir si una variable cambia , como se ven afectadas las otras?

Correlación de Pearson

El método se aplica a variables numericas continuas y la prueba te devuelve 2 valores:

- Coeficiente de Correlación de Pearson:
  - Si el valor es cercano a 1, implica una correlación positiva .
  - Mientras que un valor cercano a -1 indica una correlacion negativa
  - Un valor cercano a 0 implica que no hay correlacion entre las variables
- p-value :
  - Un valor cercano a 0,001 da una fuerte correlación
  - Un valor entre 0.01 y 0.05 da una moderada correlación
  - Un valor entre 0.05 y 1 da una débil correlación

Conclusión

En esta sesión se ha visto los conceptos al análisis exploratorio de los datos donde se usan diferentes técnicas estadisticas que nos ayudaran a discenir o inferir sobre los datos.

Bibliography

 Naomi Ceder. The Quick Python Book - Manning Publications, 2018.