Universidad Nacional de Costa Rica


Trabajo presentado para el curso:

Programación Python Básico

Proyecto Final

Miércoles 5 de agosto

Profesor: Ing. Luis Diego Gamboa Chaverr

Estudiante:

Marco Vinicio Zamora Hernández

2020

La presente investigación tiene como objetivo abarcar de manera documental el conocimiento que se va a adquirir a lo largo del desarrollo del proyecto final del curso python básico.

*Los puntos a desarrollar en el proyecto, son los siguientes:*

1. Importar los datos por medio de la librería Pandas

2. Desplegar la información del dataframe mediante la función específica , en donde únicamente muestre las 10 primeras líneas.

3. Muestre las principales estadísticas del dataframe mediante el uso de la función específica.

4. Mediante la función específica muestre el tipo de cada variable en el dataframe

5. Con la información anterior, dado que posee los nombres de la columnas, analice que columnas poseen valores nulos (NaN)

6. Muestre mediante una tabla la información de la cantidad de valores nulos por variable.

7. Elimine todas las columnas que contiene valores nulos (NaN) y asigne a un nuevo dataframe la salida con los valores sin nulos

8. Verifique que efectivamente los valores NaN fueron removidos.

9. Renombre los títulos de columnas para que sea más explicativo. Para ello utilice el siguiente mapeo

10. En el nuevo dataframe muestre la correlación entre las variables. Además, indique e investigue para que es útil poseer la correlación entre los datos.

11. En el nuevo dataframe calcule la covarianza de las variables, Además, indique e investigue para que es útil poseer la covarianza entre los datos.

12. Elimine las columnas que por su correlación no aportan valor . Para este caso deberá eliminar las relaciones con valores de correlación mayores al valor absoluto de 0.9.

13. Por medio la libreria Matplotlib realice graficación que muestren
    a. La distribución por edad de los clientes
    b. La distribución por sexo de los clientes
    c. La distribución por educación de los clientes
    d. La distribución por Monto del crédito otorgado a los clientes

*Desarrollo del proyecto:*

1. **Importando los datos utilizando pandas:**

   Referencia utilizada:https://www.shanelynn.ie/python-pandas-read_csv-load-data-from-csv-files/

   como se pudo ver en la referencia brindada arriba. El comando para abrir .csv files en python es:

   ```
   datos = pd.read_csv('credit card clients.csv')
   ```

2. **Mostrar las diez primeras lineas del documento importado:**

   Referencia utilizada:https://pandas.pydata.org/pandas-docs/version/0.23.1/generated/pandas.DataFrame.head.html

   como se pudo ver en la referencia brindada arriba. El comando para mostrar las 'n' primeras filas es:

   ```
   datos.head(10)
   ```

3. **Mostrar las principales estadisticas del data-frame con el comando especifico:**

   El comando describe():

   Referencia utilizada:https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html

4. **Mostrar el tipo de dato de cada una de las columnas del documento :**

   Referencia utilizada:https://datacarpentry.org/python-ecology-lesson-es/04-data-types-and-format/

   Se ha utilizado el comando dtypes y se modifico el import del csv file de la siguiente manera:

   ```
   pd.read_csv('credit card clients.csv',header=1)
   ```

5. **Deteccion de valores NaN:**

   Referencia utilizada:https://datascience.stackexchange.com/questions/37878/difference-between-isna-and-isnull-in-pandas

6. **Mostrar la cantidad de valores nulos por variable:**

   Referencia utilizada:https://stackoverflow.com/questions/36226083/how-to-find-which-columns-contain-any-nan-value-in-pandas-dataframe

Referencia utilizada:https://stackoverflow.com/questions/26266362/how-to-count-the-nan-values-in-a-column-in-pandas-dataframe

7. **Cambiar los valores NaN por otro valor:**

   Referencia utilizda:https://kite.com/python/answers/how-to-replace-nan-values-with-zeros-in-a-column-of-a-pandas-dataframe-in-python

   He decidido cambiar el valor Nan por el número -126 principalmente porque en ninguna de las columnas sería un dato común, por lo tanto sería fácil de detectar y ajustar los calculos.

8. **Verificar que los valores Nan hayan sido cambiados por el valor deseado:**

   En este punto se ha utilizado el mismo codigo que se utilizo en el punto 6

9. **Renombrar las columnas:**

   Referencia utilizada:https://stackoverflow.com/questions/51507315/python-pandas-im-unable-to-set-second-row-as-column-headers

   Al inicio ubieron problemas porque los nombres de las columnas estaban en la segunda fila

   Referencia utilizada:https://www.geeksforgeeks.org/how-to-get-column-names-in-pandas-dataframe/

   Al final se ha encontrado una solucion definitiva

   Referencia utilizada:https://note.nkmk.me/en/python-pandas-dataframe-rename/

   y se ha usado por facilidad el comnado dtypes

10. **¿Qué es la correlacion de los datos?**

    Referencia utilizada:https://www.youtube.com/watch?v=rRZDaEPI04A

    Matemáticamente la correlacion se refiere a la relación lineal entre dos variables;

    Las variables tienen una correlacion positiva perfecta = 1

    Las variables no poseen relacion lineal = 0

    Las variables tienen una correlacion negativa o inversa perfecta = -1

    En python la matriz de correlación se refiere a una tabla tabulada que representa la correlacion entre pares de variables de ciertos datos y en el analisis de datos, los valores fuertemente correlacionados no generar ningun valor.

Referencia utilizada:https://stackoverflow.com/questions/29432629/plot-correlation-matrix-using-pandas

## 11. Covarianza de las variables

Referencia utilizada:https://www.pythonfordatascience.org/variance-covariance-correlation/

"La covarianza es una medida de la relación entre 2 variables que depende de la escala, es decir, cuánto cambiará una variable cuando cambie otra variable. Esto se puede representar con la siguiente ecuación:"

$$\text{Covariance}\,(x,y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Where,

- $x_i$ is the $i^{th}$ observation in variable x,
- $\bar{x}$ is the mean for variable x,
- $y_i$ is the $i^{th}$ observation in variable y,
- $\bar{y}$ is the mean for variable y, and
- $N$ is the number of observations

al igual que la correlacion, en python se refiere a una tabla tabulada que va de -1 a 1.

Las variables tienen una covarianza positiva perfecta = 1

Las variables no poseen relacion lineal = 0

Las variables tienen una covarianza negativa o inversa perfecta = -1

## 12. Eliminar las columnas que por su correlación no aportan valor

Referencia utilizada:https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop.html

Referencia utilizada:https://www.w3resource.com/pandas/dataframe/dataframe-drop.php

## 13. Graficación con matplotlib

Referencias utilizadas:

1. https://matplotlib.org/3.1.0/tutorials/introductory/sample_plots.html
2. https://matplotlib.org/3.1.0/gallery/lines_bars_and_markers/simple_plot.html#sphx-glr-gallery-lines-bars-and-markers-simple-plot-py
3. https://pandas.pydata.org/docs/getting_started/intro_tutorials/06_calculate_s

tatistics.html

4. https://www.tutorialspoint.com/python_data_science/python_normal_distribution.htm
5. https://stackoverflow.com/questions/51988691/how-to-put-parameters-obtained-through-pandas-describe-in-a-plot-in-one-go
6. https://stackoverflow.com/questions/40647396/hide-histogram-plot
7. https://statisticsbyjim.com/basics/normal-distribution/

Este por mucho fue el punto a desarrollar en el que se tuvo que hacer una mayor investigación de cómo graficar y un poco de cómo funcionan las dsitribuciones en estadistica.

Al final, luego de lo visto en la ultima semana de lecciones se ha logrado realizar una función que es capaz de graficar cualquier columna del dataframe, generarndo varios tipos de grafica, según se ajuste al tipo de dato que se desea procesar. Se han usado gráficos de barra, de puntos, de pie y de la curva normal para mostrar cada una de las distribuciones solicitadas.