INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA ACADÉMICA

DIRECCIÓN DE EDUCACIÓN SUPERIOR

Carmona Bartolome Aldo Armando

1AVI

Introducción a Ciencia de Datos

Práctica No. 2: Estadística descriptiva

Repo del codigo: Link

Introducción

La estadística descriptiva es la técnica matemática que obtiene, organiza, presenta y describe un conjunto de datos con el propósito de facilitar el uso, generalmente con el apoyo de tablas, medidas numéricas o gráficas. La estadística es uno de los puntos más importantes dentro de la ciencia de datos ya que nos ayuda a entender de mejor manera los datos mostrados asi como la utilización de estas para la predicción de temas referentes a dichos datos.

El siguiente programa utiliza un CSV (data.csv) para la obtención de datos previamente instanciados en el archivo acerca de dos diferentes tipos de empleos que actualmente son las carreras que la ESCOM esta ofertando. El programa utiliza pandas para el manejo de los datos, puesto que esto puede dar una mayor organización si se tienen muchos datos en un solo archivo y por ultimo se uso matplotlib.pyplot para el ploteo de las graficas para el entendimiento de los datos que arrojan los salarios de estas dos carreras (ciencia de datos e ingeniería en sistemas)

Desarrollo

```
# -*- coding: utf-8 -*-
"""
Editor de Spyder

Este es un archivo temporal.
"""

import matplotlib.pyplot as plot
import statistics as st
import seaborn as sns
import numpy as np
import os
from tkinter import *
import pandas as pd


dataset = pd.read_csv('data1.csv')
cd = dataset['Cientifico de Datos']
cd1 = dataset.sort_values('Cientifico de Datos')
"""
print(dataset)
print('<<<<>>>>')
print(dataset.keys())
```

```
print(cd)
print('<<<<<<<<<<<<<<<<<<<<>>>>>>>>>>>>>>>>>>>')
print(cd1)
"""

q1,q3 = np.percentile(cd,[25 , 75])

iqr = q3 - q1

os.system('cls')
print('Analisis de salarios de Cientificos de Datos')
print("<<Datos de entrada>>")
print(cd)
print('\n <<Resultados>>')

#medidas de centralidad
print('Media = ',cd.mean())
print('Mediana = ',cd.median())


#Mediddas  de Dispersion
print('Desviacion Estandar = ',cd.std())
print('Q1 y Q3 = ',q1, q3)
print('Distancia intercuartil = ',iqr )
print('\n')

#Grafica de caja y vigote
sns.boxplot(cd);
plot.xlabel('Salarios')
sns.swarmplot(cd, color='r')

# Histograma y curva de densidad
sns.displot(cd, kde='true', rug='true')
plot.xlabel('Salarios')
plot.ylabel('Frecuencia')
plot.axvline(cd.mean(), color='red', linestyle='--')
plot.axvline(cd.median(), color='green', linestyle='--')

plot.show

#####################################################
```

```
igs = dataset['Ing Sistemas']
igs1 = dataset.sort_values('Ing Sistemas')
"""
print(dataset)
print('<<<<>>>>')
print(dataset.keys())
print(cd)
print('<<<<<<<<<<<<<<<<<<<<>>>>>>>>>>>>>>>>>>>')
print(cd1)
"""


q1,q3 = np.percentile(igs,[25 , 75])

iqr = q3 - q1

os.system('cls')
print('Analisis de salarios de Ing de Sistemas')
print("<<Datos de entrada>>")
print(igs)
print('\n <<Resultados>>')

#medidas de centralidad
print('Media = ',igs.mean())
print('Mediana = ',igs.median())


#Mediddas  de Dispersion
print('Desviacion Estandar = ',igs.std())
print('Q1 y Q3 = ',q1, q3)
print('Distancia intercuartil = ',iqr )
print('\n')


#Grafica de caja y vigote
sns.displot(cd, kde='true', rug='true')
sns.boxplot(igs);
plot.xlabel('Salarios')
sns.swarmplot(igs, color='r')

# Histograma y curva de densidad
sns.displot(igs, kde='true', rug='true')
```

```
plot.xlabel('Salarios')
plot.ylabel('Frecuencia')
plot.axvline(igs.mean(), color='red', linestyle='--')
plot.axvline(igs.median(), color='green', linestyle='--')

plot.show
```

Cabe destacar que el uso de pandas como gestor de archivos puede hacer mas complicado el entendimiento del programa limitando hasta cierto punto algunas otras librerías que en un principio son mas sencillas de utilizar, sin embargo el beneficio que ofrece de gestión de varios datos es mayor, claro teniendo en cuenta de que la cantidad de datos que se este usando sea inmensa y mucho mayor a la presentada en este programa.

Ejecución

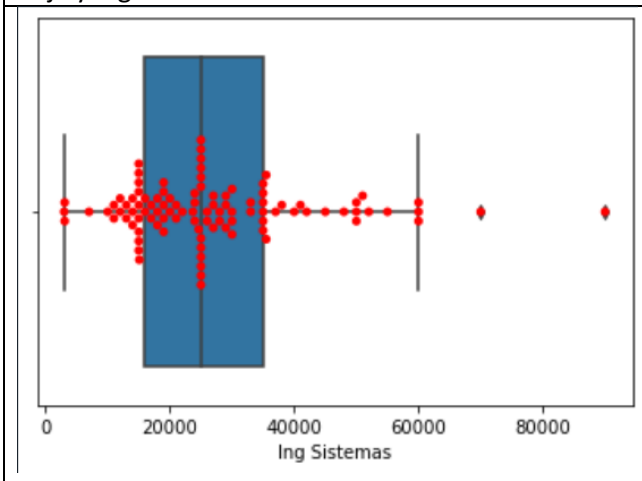
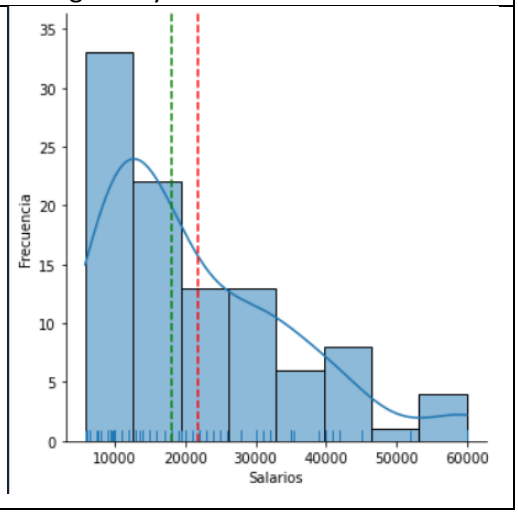
Primeramente, tomemos en cuenta los datos de entrada para el calculo de las estadísticas solicitadas para los salarios de los Científicos de Datos mostrando el dataframe referente a este.

```
In [94]: runfile('C:/Users/waldo/OneDrive/Escritorio/practica02.py', wdir='C:/Users/waldo/
OneDrive/Escritorio')
Analisis de salarios de Cientificos de Datos
<<Datos de entrada>>
0      6487
1      9400
2     60000
3     32000
4     45000
      ...
95     9000
96    35500
97    32000
98    18000
99    10000
Name: Cientifico de Datos, Length: 100, dtype: int64
```

Posteriormente mostraremos los resultados obtenidos después de calcular cada uno de los datos estadísticos referentes a este primer empleo.

```
 <<Resultados>>
Media =  21814.27
Mediana =  18000.0
Desviacion Estandar =  13552.64638204166
Q1 y Q3 =  12000.0 30000.0
Distancia intercuartil =  18000.0
```

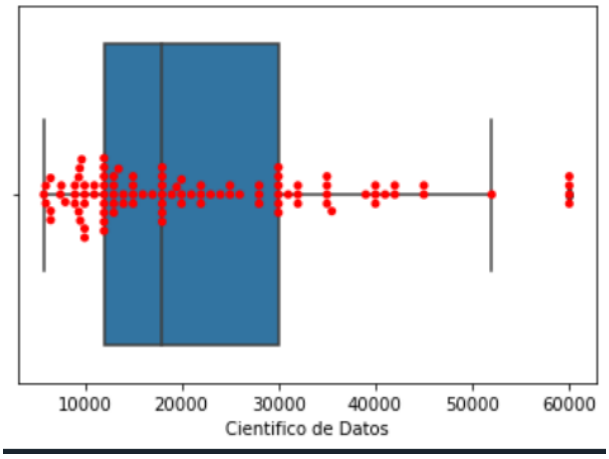
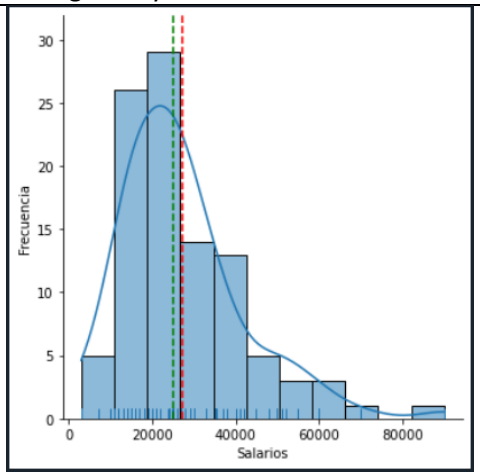
Por ultimo se muestran las dos graficas solicitadas (caja y bigote, Histograma y curva de densidad)

Caja y bigote	Histograma y curva de densidad
	

Todos estos pasos se repiten con los siguientes datos

```
Analisis de salarios de Ing de Sistemas
<<Datos de entrada>>
0     35000
1     30000
2     29000
3      3080
4     15000
      ...
95    20000
96    18000
97    70000
98    35000
99    12000
Name: Ing Sistemas, Length: 100, dtype: int64
```

```
 <<Resultados>>
Media =  27115.56
Mediana =  25000.0
Desviacion Estandar =  14867.674058354072
Q1 y Q3 =  16000.0 35000.0
Distancia intercuartil =  19000.0
```

Caja y bigote	Histograma y curva de densidad
	

Conclusión

Comparando las graficas puedo deducir que actualmente la bolsa de trabajo es mucho mejor para los ingenieros en sistemas computaciones en México teniendo picos mucho mas altos llegando a los 80k por mes a diferencia del científico de datos el cual llega a los 60k por mes, sin embargo estos trabajos excluyen trabajos transnacionales y debido a que ciencia de datos es una nueva carrera en México y actualmente el países se encuentra en una transición hacia la industria 4.0, donde sin duda alguna el científico de datos es de suma importancia para el uso de esta industria, por ende concluyo que durante los próximos años el salió base aumente para esta carrera y la cantidad de trabajos sea mucho mayor.

Por ultimo la importancia que tienen las estadísticas descriptivas es esencial para la ciencia de datos ya que esta área de la computación se dedica al entendimiento de la información y el uso de graficas y datos que pueden describir, de forma clara, una enorme cantidad datos que en un principio podrían tomar mucho su análisis y por ende el uso de los mismo para la predicción.

Referencias de Apoyo

Alberca, A. S. (2021, 14 mayo). *La librería Pandas*. Aprende con Alf. Recuperado 8 de noviembre de 2021, de https://aprendeconalf.es/docencia/python/manual/pandas/

Estadística en Python: media, mediana, varianza, percentiles (Parte III). (2017, 4 noviembre). Adrianistán. Recuperado 8 de noviembre de 2021, de https://blog.adrianistan.eu/estadistica-python-media-mediana-varianza-percentiles-parte-iii