Laboratorio 11

Sesión # 11 Componente Práctico

Título del Laboratorio: Aplicación del uso de la herramienta de visualización en Python.
Duración: 2 horas
Objetivos del Laboratorio: *Afianzar los conocimientos y manejo básico en Python sobre las visualizaciones con ejercicios prácticos planteados.*

Materiales Necesarios:

1. *Computador con acceso a internet.*
2. *Colocarlo en el repositorio de Github*
3. *Ampliar el conocimiento con el curso de datos en AWS y Cisco.*
4. *Python en línea: Google colab.*

Estructura del Laboratorio:
Realizar el laboratorio del curso de Cisco del análisis de los datos el cual solicitan la información detalla en la sesión a continuación, utilizar Excel o Python para la solución, realizar código y captura de pantalla.

Los ejercicios de práctica son escenarios reales que se deberás realizar el código usando las bibliotecas en Python para la creación de gráficos básico de los datos, pegar la captura de pantalla del resultado.

Parte 1

1. **Ejercicio de práctica 1.**

   El siguiente ejercicio es tomado del curso de Cisco (Bootcamp de Análisis de Datos- Lab - Interpretación de visualizaciones con respecto a valores atípicos), donde se desarrollará lo siguiente:

   Interpretación de visualizaciones con respecto a valores atípicos, en este ejercicio práctico deberás utilizar gráficos y funciones para detectar valores atípicos en los datos.

   Examinar un conjunto de datos en busca de valores atípicos, un valor atípico es un valor o punto de datos que varía significativamente de otros en el mismo conjunto de datos.

   Un valor atípico puede ser resultado de la variabilidad en las mediciones, errores experimentales o errores humanos al ingresar los datos.
   Para garantizar que cualquier análisis de datos sea correcto, es necesario identificar los valores atípicos y luego determinar cuál es la mejor manera de tratarlos.

Instrucciones:

Examinar un conjunto de datos en busca de valores atípicos

- **Paso 1: Abra el conjunto de datos.**

  - Descargar el archivo Bike Sales_Outlier_Lab.xlsx
  - Sube el archivo a GitHub, desarrollo en Excel o Python en línea.

- **Paso 2: Utilice una tabla dinámica para seleccionar datos para el análisis**

  a. Haga clic en cualquier celda de la hoja de cálculo Ventas de bicicletas.
  b. Inserte una tabla dinámica haciendo clic en Insertar > Tabla dinámica. Compruebe que Nueva hoja de cálculo esté seleccionada en el cuadro de diálogo Crear tabla dinámica y haga clic en Aceptar.
  Esto agrega una nueva hoja de trabajo para la tabla dinámica.
  c. En el cuadro de diálogo Campos de tabla dinámica, marque los campos Fecha y Cantidad de pedido.
  La tabla dinámica se crea con dos columnas Fecha y Suma de Order_Quantity .

Etiquetas de fila	Suma de Order_Quantity
1/12/2021	5
2/12/2021	3
3/12/2021	4
4/12/2021	4
5/12/2021	10
6/12/2021	6
7/12/2021	6
8/12/2021	11
9/12/2021	3
10/12/2021	8
11/12/2021	8
12/12/2021	12
13/12/2021	6
14/12/2021	4
15/12/2021	1
16/12/2021	5
17/12/2021	4
18/12/2021	19
19/12/2021	43
20/12/2021	13
21/12/2021	5
22/12/2021	10
23/12/2021	3
24/12/2021	4

- **Paso 3: Ordenar los datos para encontrar valores atípicos**

Una forma de identificar valores atípicos es simplemente ordenar los datos. Este método funciona con conjuntos de datos pequeños en los que los datos se pueden analizar fácilmente.

a. Ordenar la suma de la columna Qrder Quantity de mayor a menor

- Seleccione los puntos de datos en la columna Suma de cantidad de pedido. (No seleccione el Total de subvención ni el encabezado de la columna).
- Haga clic en Ordenar y filtrar > Ordenar descendente.

Esto ordena los puntos de datos de **Order Quantity** del mayor al menor.

Etiquetas de fila	Suma de Order_Quantity
19/12/2021	43
18/12/2021	19
20/12/2021	13
12/12/2021	12
8/12/2021	11
5/12/2021	10
22/12/2021	10
11/12/2021	8
10/12/2021	8
7/12/2021	6
6/12/2021	6
13/12/2021	6
21/12/2021	5
1/12/2021	5
16/12/2021	5
4/12/2021	4
24/12/2021	4
17/12/2021	4
3/12/2021	4
14/12/2021	4
2/12/2021	3
23/12/2021	3
9/12/2021	3
15/12/2021	1
Total general	**197**

- ¿En qué fecha de diciembre se registró la mayor cantidad de ventas?: El 19 de Diciembre de 2021
- ¿Cuál fue la cantidad de ventas?: 43 ventas
- Revise los datos de la hoja de cálculo **de ventas de bicicletas** del 19 de diciembre.
- ¿Qué entrada contribuye más a la suma de Order Quantity en la tabla dinámica?:

Numero de compra: 000261765, con cantidad: 11, hecha en Washington.

- En otras palabras, ¿qué número de pedido es el más responsable del valor atípico?: El pedido 000261765

FORMAS DE FILTRAR

Etiquetas de fila	Suma de Order_Quantity
19/12/2021	**43**
000261756	4
000261757	4
000261758	4
000261759	4
000261760	4
000261761	4
000261762	4
000261763	2
000261764	2
000261765	**11**
18/12/2021	**19**
000261749	4
000261750	4
000261751	3
000261752	3
000261753	3
000261754	1
000261755	1
20/12/2021	**13**

Date	19/12/2021

Etiquetas de fila	Suma de Order_Quantity
000261756	4
000261757	4
000261758	4
000261759	4
000261760	4
000261761	4
000261762	4
000261763	2
000261764	2
000261765	**11**
Total general	**43**

Paso 4: Utilice un gráfico de dispersión para encontrar valores atípicos

Un gráfico de dispersión puede ayudar a identificar valores atípicos, especialmente en conjuntos de datos grandes.

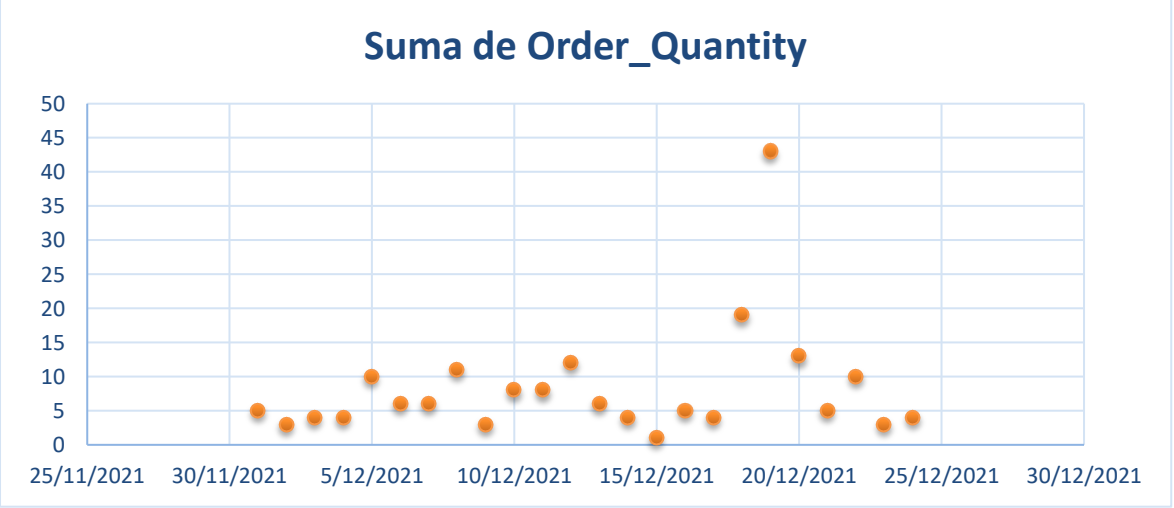
a. Regrese a la hoja de trabajo que contiene la tabla dinámica (Hoja1).

b. Copie y pegue los datos de la tabla dinámica en dos columnas en blanco (D y E). Copie la fila del encabezado con los datos, pero no copie la fila del total general. Excel no permite crear un gráfico de dispersión a partir de los datos de una tabla dinámica. Por lo tanto, los datos deben trasladarse a otras columnas.

c. Insertar diagrama de dispersión.

- Seleccione todas las celdas en los datos copiados y utilice Ordenar y filtrar para ordenarlos en orden ascendente.
- Resalte la columna **Suma de Order_Quantity** en los datos copiados.
- Haga clic en **Insertar** > **Dispersión** y luego seleccione el gráfico de dispersión superior izquierdo en la lista desplegable.

Tenga en cuenta que la imagen del gráfico de dispersión hace que las ventas del 19 de diciembre se destaquen fácilmente *como un valor atípico de los demás puntos* de datos de cantidad de pedidos, como se muestra a continuación.



- Eliminar el diagrama de dispersión.

- **Paso 5: Uso de las funciones GRANDE y PEQUEÑO para encontrar valores atípicos.**

Si hay muchos datos, se pueden usar las funciones GRANDE y PEQUEÑO para extraer los valores más grandes y pequeños, lo que puede ayudar a ver si hay valores atípicos.

Para este ejemplo, la columna Fecha es la columna D y la columna Suma de Cantidad_de_pedido es la columna E.

Las columnas en su hoja de cálculo pueden ser diferentes, así que ajuste las referencias de las celdas de función en consecuencia.

a. En una celda vacía ingrese la función =GRANDE($E$4:$E27,1).
Esta función examina las entradas de la celda E4 a la E27 y devuelve el valor más alto.

- ¿Qué valor fue devuelto?=43

PARTE 5.	
Pedidos Grande	43

b. Para obtener los 5 valores más altos, modifique las funciones a **=LARGE($E$4:$E27, ROW($1:5))** .

Esto devuelve los cinco valores más altos. Para devolver más valores, cambie el "5" al final de la función por la cantidad de valores que desea devolver.

Función: =K.ESIMO.MAYOR(E2:E25; {1;2;3;4;5})

5 Valores altos	43
	19
	13
	12
	11

- ¿Qué función devolvería los 6 valores más bajos?
=K.ESIMO.MENOR(E2:E25; {1;2;3;4;5;6})

	11
6 valores menores	1
	3
	3
	3
	4
	4

Una vez identificados los valores atípicos, el siguiente desafío es qué hacer con ellos. Los valores atípicos pueden indicar errores en los datos o pueden ser datos válidos que deben investigarse para determinar por qué parecen ser una anomalía.

Hay un par de formas en las que un analista de datos puede lidiar con los valores atípicos.

1. Elimínelos. En un conjunto de datos grande, eliminar algunos valores atípicos probablemente no afecte el análisis general. Sin embargo, es importante crear una copia de los datos para poder investigar qué causó los valores atípicos en primer lugar.

2. En este ejemplo, se podría eliminar la fila 72 del conjunto de datos Bike Sales.

3. Normalizarlos (Ajustar su valor). El valor de los valores atípicos se modifica para que esté ligeramente por encima del valor máximo en el conjunto de datos. Este es un buen método si no sesgará los datos.

Hay varios métodos estadísticos para normalizar los datos. Investigue los distintos métodos antes de ajustar aleatoriamente los valores de los datos.

La profe nos enseñó el método del IQR.

1er cuartil (Q1:0,25*24): posición 6	4
3er cuartil (Q3:0,75*24): posición 18	10
IQR:Q3-Q1	6
LIMITES VALORES ATIPICOS	
Limite inferior (Q1-1.5*IQR)	-5
Limite superior (Q3+1.5*IQR)	19

En el conjunto de datos de ejemplo de Ventas de bicicletas, la Cantidad_pedido del 19 de diciembre se podría cambiar de 43 a 20 para que esté justo por encima del valor máximo de 19.

Preguntas de reflexión

Enumere los factores que podrían determinar si los valores atípicos de los datos deben o no considerarse en el análisis final de un conjunto de datos.

Los valores atípicos pueden o no incluirse en el análisis final dependiendo de varios factores.

1. Es importante considerar el contexto, ya que algunos valores pueden ser inusuales pero válidos si tienen una explicación lógica.
2. Debe evaluarse la fuente del dato, pues si el outlier proviene de un error de medición o registro, es razonable descartarlo.
3. Evaluar impacto en los resultados estadísticos, ya que pueden distorsionar medidas como la media, desviación estándar o afectar modelos de regresión.
4. El tipo de análisis y el tamaño de la muestra influyen en esta decisión, pues en conjuntos pequeños los outliers pueden tener un efecto mayor.
5. Si hay muchos valores atípicos, podría tratarse de una tendencia válida o la presencia de un subgrupo, por lo que no siempre deben eliminarse.
6. También es fundamental tener claro el objetivo del análisis; si se buscan comportamientos extremos, los outliers deben conservarse.

Parte 2

DESARROLLO EN PYTHON

2. Ejercicio de práctica 2.

Deberás realizar el código y visualización correspondiente al escenario, dar una breve interpretación de los datos.

1. **Comparación de calificaciones de estudiantes en diferentes asignaturas**
   Se quiere comparar las calificaciones promedio de los estudiantes en distintas materias para detectar diferencias de rendimiento.

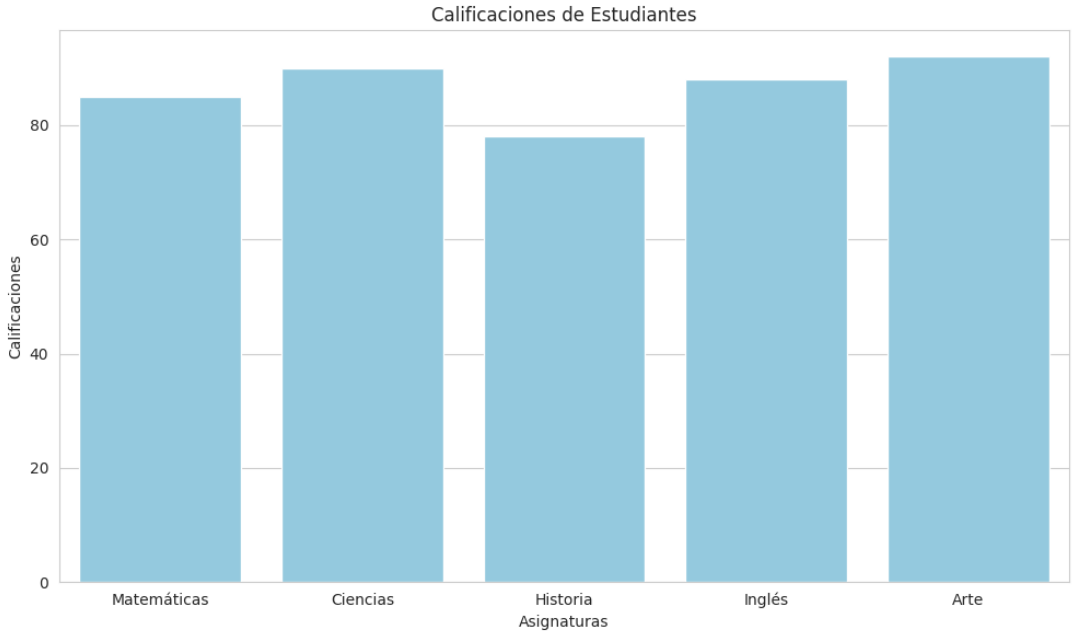
- **asignaturas** =Matemáticas,Ciencias,Historia,Inglés,Arte
- **calificaciones** = 85, 90, 78, 88, 92

- **Código**:

```
[1] import numpy as np
    import matplotlib.pyplot as plt
    import pandas as pd
    import seaborn as sns

    asignaturas = ['Matemáticas','Ciencias','Historia','Inglés','Arte']
    calificaciones = [85, 90, 78, 88, 92]
    df=pd.DataFrame({'Asignaturas':asignaturas,'Calificaciones':calificaciones})
    sns.set_style("whitegrid")
    plt.figure(figsize=(12, 6))
    sns.barplot(data=df, x='Asignaturas', y='Calificaciones', color='Skyblue')
    plt.title('Calificaciones de Estudiantes')
    plt.xlabel('Asignaturas')
    plt.ylabel('Calificaciones')
    plt.tight_layout()
    plt.show()
```

- **Interpretación**:



Los estudiantes tienen un promedio menor en la calificación de historia, mientras que en arte y ciencias tienen el mejor desempeño.

2. Comparación del tiempo de carga de diferentes páginas web

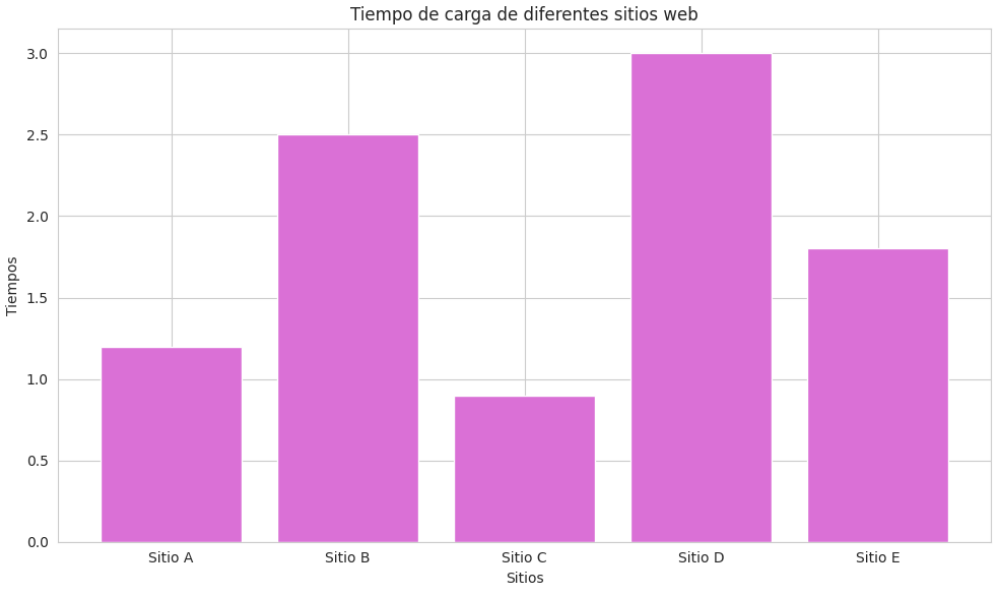
Se quiere visualizar el tiempo promedio de carga de diferentes sitios web para identificar cuál es más rápido o lento.

- **sitios** = Sitio A, Sitio B,Sitio C,Sitio D,Sitio E
- **tiempos** = 1.2, 2.5, 0.9, 3.0, 1.8

- **Código**:

```
Sitios = ['Sitio A', 'Sitio B','Sitio C','Sitio D','Sitio E']
Tiempos = [1.2, 2.5, 0.9, 3.0, 1.8]
plt.figure(figsize=(12, 6))
plt.bar(Sitios, Tiempos, color='Orchid')
plt.title('Tiempo de carga de diferentes sitios web')
plt.xlabel('Sitios')
plt.ylabel('Tiempos')
plt.tight_layout()
plt.show()
```

- **Interpretación**:



El sitio D y B tienen los mayores tiempos de carga, mientras el sitio C es el mas rápido.

3. Relación entre el número de horas de estudio y el rendimiento académico

Se quiere ver si existe una correlación entre las horas de estudio y el rendimiento en los exámenes.

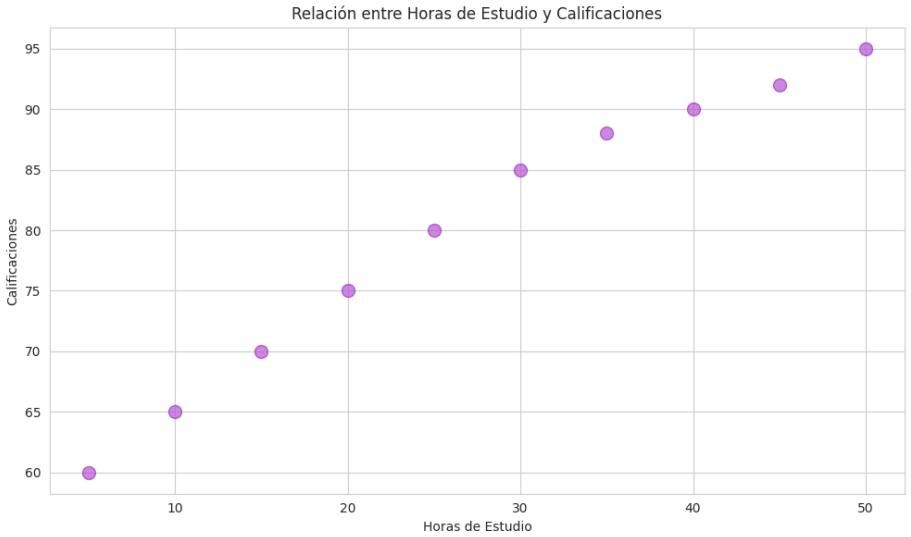
- **horas_estudio** = [5, 10, 15, 20, 25, 30, 35, 40, 45, 50]
- **calificaciones** = [60, 65, 70, 75, 80, 85, 88, 90, 92, 95]

- **Código**:

```
horas_estudio = [5, 10, 15, 20, 25, 30, 35, 40, 45, 50]
calificaciones = [60, 65, 70, 75, 80, 85, 88, 90, 92, 95]

plt.figure(figsize=(10, 6))
plt.scatter(horas_estudio, calificaciones, color='mediumorchid', s=100, edgecolor='darkorchid',
alpha=0.7)
plt.title('Relación entre Horas de Estudio y Calificaciones')
plt.xlabel('Horas de Estudio')
plt.ylabel('Calificaciones')
plt.tight_layout()
plt.show()
```

- **Interpretación**:



Entre mayor cantidad de horas de estudio, mejores calificaciones

4. Distribución de salarios en tres departamentos distintos

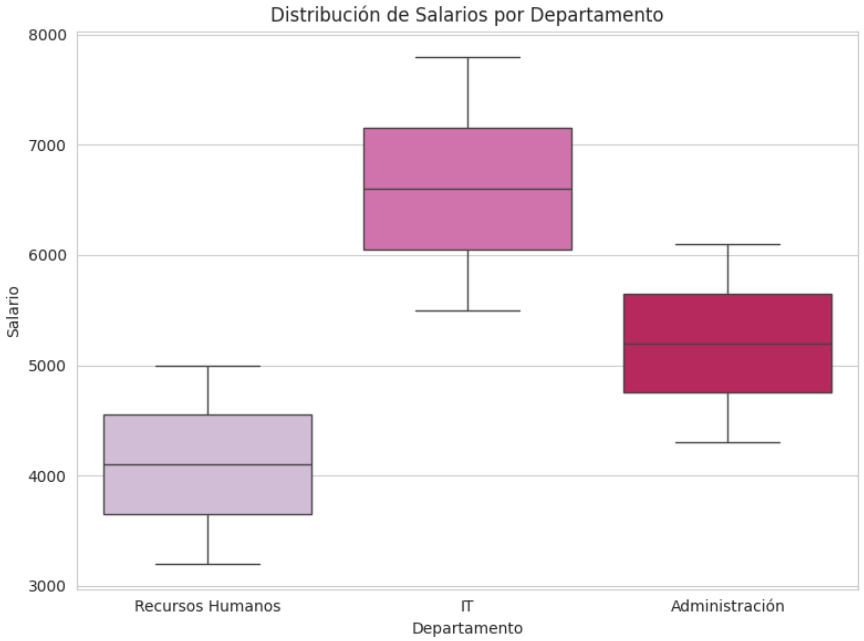
Se tiene un conjunto de datos de salarios en diferentes departamentos y quieres visualizar la variabilidad y mediana de los salarios en cada uno.

```
datos_salarios = {
  'Departamento': ['Recursos Humanos']*10 + ['IT']*10 + ['Administración']*10,
  'Salario': [3200, 3400, 3600, 3800, 4000, 4200, 4400, 4600, 4800, 5000,
              5500, 5700, 6000, 6200, 6500, 6700, 7000, 7200, 7500, 7800,
              4300, 4500, 4700, 4900, 5100, 5300, 5500, 5700, 5900, 6100]
}
```

- **Código**:

```
datos_salarios = {'Departamento': ['Recursos Humanos']*10 + ['IT']*10 + ['Administración']*10,
                  'Salario': [3200, 3400, 3600, 3800, 4000, 4200, 4400, 4600, 4800, 5000,
5500, 5700, 6000, 6200, 6500, 6700, 7000, 7200, 7500, 7800,
4300, 4500, 4700, 4900, 5100, 5300, 5500, 5700, 5900, 6100]
}
dfsalarios = pd.DataFrame(datos_salarios)
plt.figure(figsize=(8, 6))
sns.boxplot(data=dfsalarios, x='Departamento', y='Salario', hue='Departamento', palette='PuRd')
plt.title('Distribución de Salarios por Departamento')
plt.xlabel('Departamento')
plt.ylabel('Salario')
plt.tight_layout()
plt.show()
```

- **Interpretación**:



La mediana del salario de recursos humanos es de 4100 aproximadamente, para IT es de 6600 y administración de 5100-5200, según la gráfica, en cuanto a la variabilidad, las cajas no son anchas, lo que muestra que la mayoría de los datos se concentran para recursos humanos entre 3800 y 4500, para IT entre 6000 y 7100 y para administración entre 4800 y 5600, además, los bigotes son cortos, indicando que el resto de los datos tampoco son variables.

Link github: https://github.com/lmerasoc/Analisis-de-datos-Bootcamp/blob/main/Lab_11_Lina_Maria_Eraso.ipynb