sunlight_vs_park_attendance_solution

February 6, 2025

1 Sunlight and Park Attendance

1.1 1. Import the necessary modules:

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

1.2 2. Load sunlight.csv into a Pandas DataFrame:

[2]:
```
df = pd.read_csv("sunlight.csv")
```

1.3 3. Explore the data:

[3]:
```
df.sample(10)
```

[3]:

	Sunlight Hours	Park Attendance
24	4.560700	97.087996
15	1.834045	46.432239
28	5.924146	91.196551
11	9.699099	193.692585
16	3.042422	49.551113
10	0.205845	-18.560807
12	8.324426	138.551839
0	3.745401	81.360715
7	8.661761	149.095556
4	1.560186	59.141650

1.4 4. Plot the data using a Scatterplot

The scatterplot should conform to the following: - Where the number of hours of sunlight is less than 5, the point should be blue - Where the number of hours of sunlight is 5 or more, the point should be red - The plot should include a linear trend line - The trend line should be orange - The plot should include title and axis labels - The plot should display a grid background - The plot should have a width of 10 inches and a height of 6 inches

[4]:
```
plt.figure(figsize=(10, 6))
```

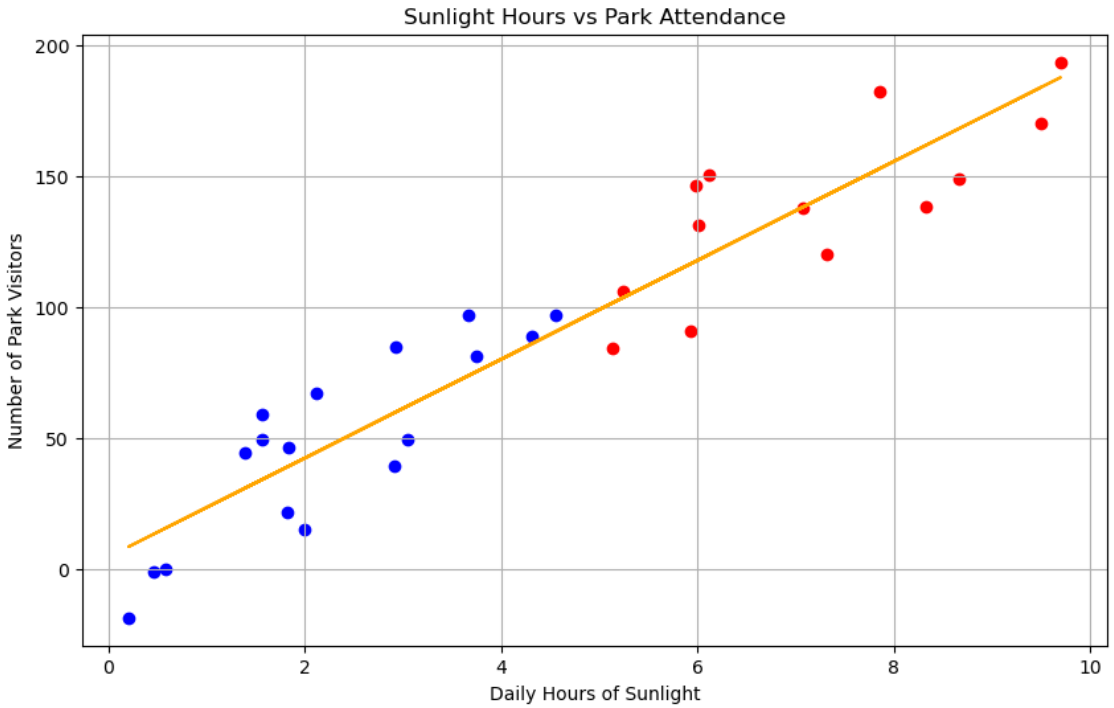
```
plt.scatter(x=df[df['Sunlight Hours'] < 5]['Sunlight Hours'],
            y=df[df['Sunlight Hours'] < 5]['Park Attendance'],
            color='blue')

plt.scatter(x=df[df['Sunlight Hours'] >= 5]['Sunlight Hours'],
            y=df[df['Sunlight Hours'] >= 5]['Park Attendance'],
            color='red')

m, c = np.polyfit(df['Sunlight Hours'], df['Park Attendance'], 1)
plt.plot(df['Sunlight Hours'],
         m * df['Sunlight Hours'] + c,
         color='orange')

plt.title("Sunlight Hours vs Park Attendance")
plt.xlabel("Daily Hours of Sunlight")
plt.ylabel("Number of Park Visitors")
plt.grid(True)

plt.show()
```



1.5 5. Analysis:

The orange line represents the trend. It suggests a positive correlation between hours of sunlight and park attendance. This trend is logical as people are more inclined to visit parks on sunnier

days. The scatter of points around the trend line indicates variability, which could be due to other factors not accounted for in this simple analysis.

1.6 Questions

Do you notice anything unexpected about the data? What is the lowest number of park visitors? What could this value possibly mean? It definitely looks like an error. Are the numbers realistic? We always need to keep a critial eye on the data we're investigating.