"In what ways can data be misleading or misinterpreted, even when statistical concepts are applied correctly? How can Python be used to identify and address these issues?"

Misleading or Misinterpreted Data

Sampling Bias: If the dataset does not represent the entire population of dog breeds accurately, any analysis might be biased. For example, if the dataset only includes popular breeds, rarer breeds might be underrepresented, skewing results towards the characteristics of more common breeds. Correlation vs. Causation: Statistical analysis might reveal correlations between variables (e.g., larger breeds having shorter life spans), but this does not imply causation. Other factors, like genetics or environment, could influence these traits. Outliers: Extreme values can distort statistical summaries like means and medians. For instance, a few breeds with exceptionally long or short lifespans could skew the average life expectancy. Overfitting in Predictive Modeling: When creating predictive models (e.g., predicting a breed's popularity based on its traits), using too many variables or overly complex models can lead to overfitting. The model might perform well on the dataset but poorly in real-world scenarios. Cherry-Picking Data: Selectively presenting data that supports a specific argument, while ignoring data that contradicts it, can lead to misleading conclusions.

Python can be used to:

Analyze Sampling: Use Python to check the representativeness of the dataset. Statistical Testing: Perform hypothesis testing to differentiate between correlation and causation. Identify and Manage Outliers: Use statistical methods to detect and handle outliers. Cross-Validation in Modeling: Implement cross-validation techniques to avoid overfitting. Comprehensive Data Analysis: Encourage comprehensive exploratory data analysis to avoid cherry-picking. Visual Prop I will create a scatter plot comparing two variables (e.g., life expectancy vs. weight) from the dataset. This plot will illustrate how correlation does not imply causation. For example, a visual trend might suggest that heavier breeds have shorter lifespans, but this doesn't necessarily mean that weight causes shorter life expectancy; other factors could be at play.

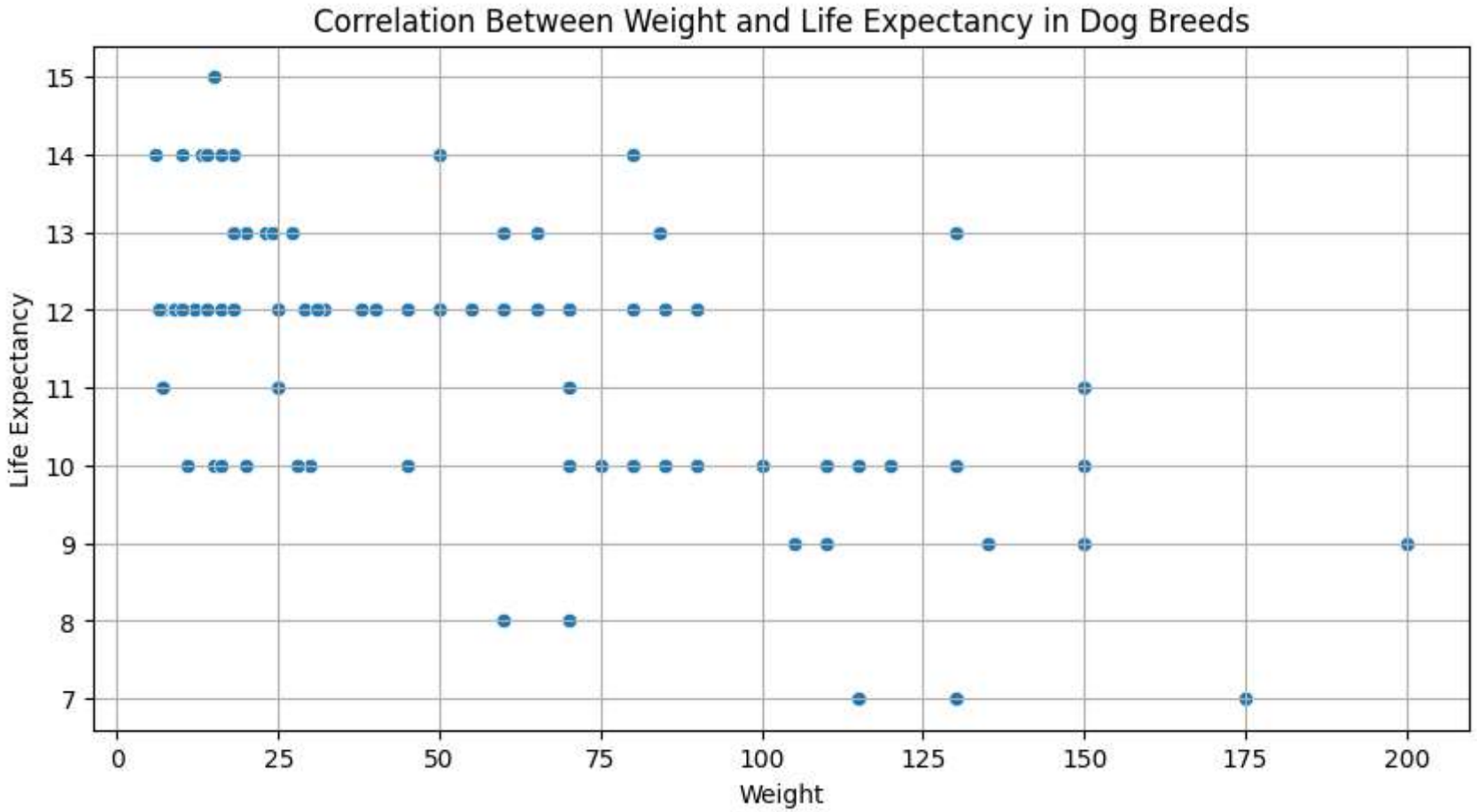
In [ ]:
```
#import libraries
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns

#import data
dog_data = pd.read_csv("C:/Users/khbil/Documents/GCU/DSC-510/Week_1/dog_breeds.csv")
data_for_plot = dog_data[['min_life_expectancy', 'max_weight_male']].dropna()
print(data_for_plot.head)
```

```
<bound method NDFrame.head of     min_life_expectancy  max_weight_male
0                    10             75.0
1                    12             32.0
2                    10             80.0
3                     7            175.0
4                    10             80.0
..                  ...              ...
92                   12             70.0
93                   12             65.0
94                   12             31.0
95                   11              7.0
96                   10             28.0

[97 rows x 2 columns]>
```

In [ ]:
```python
# Create the scatter plot
plt.figure(figsize=(10, 5))
sns.scatterplot(x='max_weight_male', y='min_life_expectancy', data=data_for_plot)
plt.title('Correlation Between Weight and Life Expectancy in Dog Breeds')
plt.xlabel('Weight')
plt.ylabel('Life Expectancy')
plt.grid(True)
plt.show()
```



This scatter plot illustrates the relationship between life expectancy and weight in different dog breeds. It shows a trend where heavier breeds tend to have shorter lifespans. However, this correlation does not imply causation, as other factors might influence both weight and life expectancy.

References

J. Singh, J. Singh, G. Singh and N. Kaur, "Exploratory Data Analysis for Interpreting Model Prediction using Python," 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/SMARTGENCON56628.2022.10083533.

Rogel-Salazar, J. (2023). Snakes, Bears & Other Numerical Beasts: NumPy, SciPy & pandas. In Statistics and data visualisation with Python (1st ed.). CRC Press.