seaborn

- What is seaborn , and how it compares to matplotlib
- Correlations with seaborn
- Useful statistical seaborn plots
- Easly combined data

Seaborn is a Python library for creating beautiful and informative data visualizations. Built on top of Matplotlib, it offers a simpler interface for drawing statistical plots. It works well with data frames and numerical arrays, making it ideal for visualizing relationships in large datasets. Seaborn supports various plot types like heatmaps, violin plots, and pair plots, with customizable themes and color palettes. It's a powerful tool for visualizing complex statistical relationships in data science.

seaborn

What is seaborn?

Feature	Seaborn	Matplotlib
Ease of Use	High-level, simple syntax	Low-level, more complex syntax
Customization	Limited but sufficient for most needs	Highly customizable
Integration	Designed for pandas DataFrames	Supports multiple data types
Flexibility	Limited to specific plot types	Extremely flexible for custom plots
Data Input	Pandas DataFrames	Lists, arrays, or DataFrames
Performance	Slightly slower for complex plots	Generally faster
Plot Types	Focus on statistical plots	Covers all basic and advanced plots

What is seaborn?

Seaborn has it’s own dataset for practice

Listing all the available datasets:

```
sns.get_dataset_names()
```

```
['anagrams',
 'anscombe',
 'attention',
 'brain_networks',
 'car_crashes',
 'diamonds',
 'dots',
 'dowjones',
 'exercise',
 'flights',
```

loading a dataset into a pandas dataframe:

```
df = sns.load_dataset('penguinss')
```

Correlation

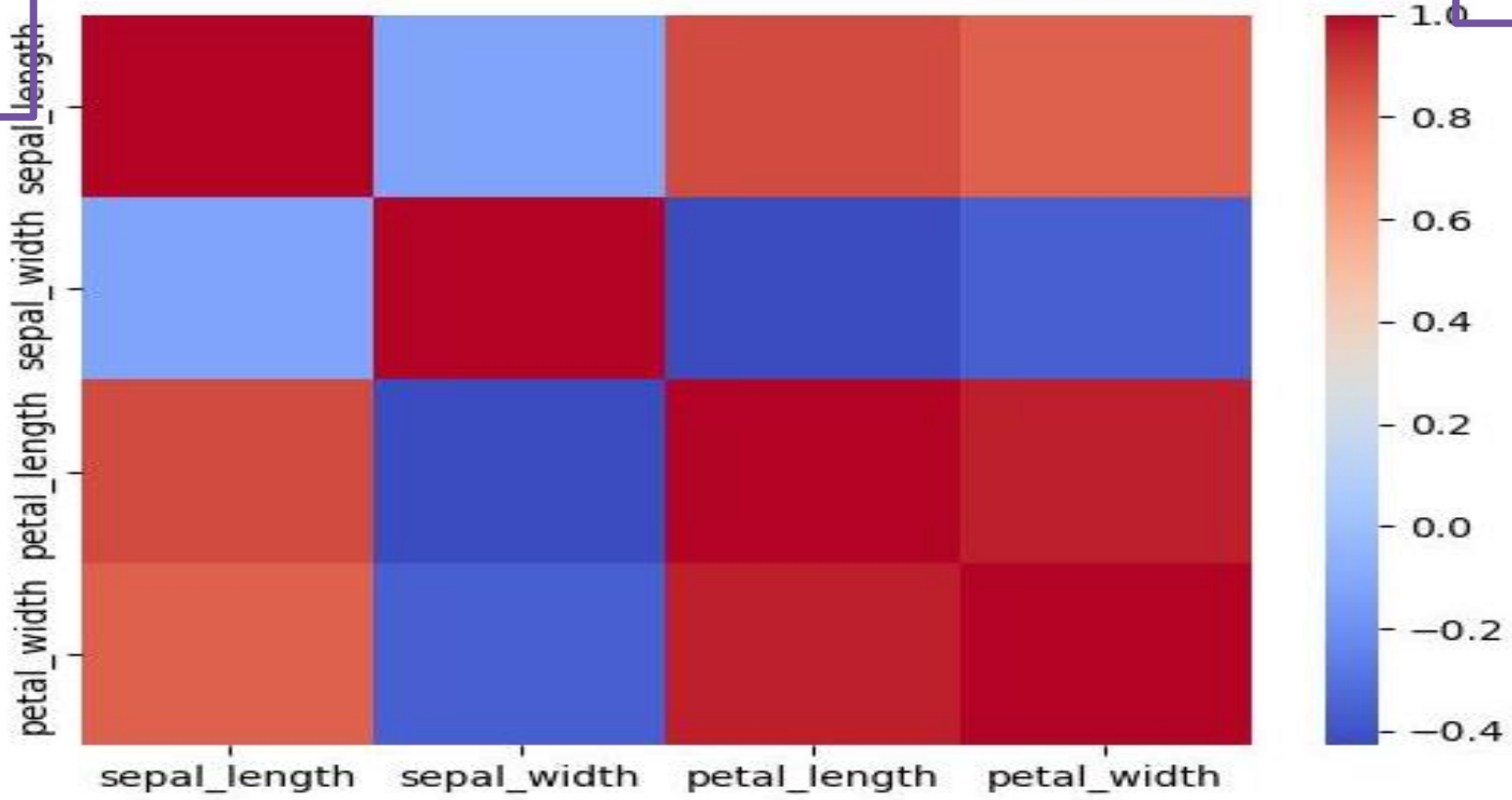
Correlation measures the strength and direction of a relationship between two variables. A positive correlation means that as one variable increases, the other tends to increase as well, while a negative correlation means that as one increases, the other tends to decrease. The value of correlation ranges from -1 (perfect negative) to 1 (perfect positive), with 0 indicating no relationship. It’s a key concept in understanding patterns in data and is often visualized using tools like scatterplots or heatmaps for clarity.

Correlation heatmap

```
sns.heatmap(df.select_dtypes(['number']).corr()
            ,   cmap ='coolwarm')
```

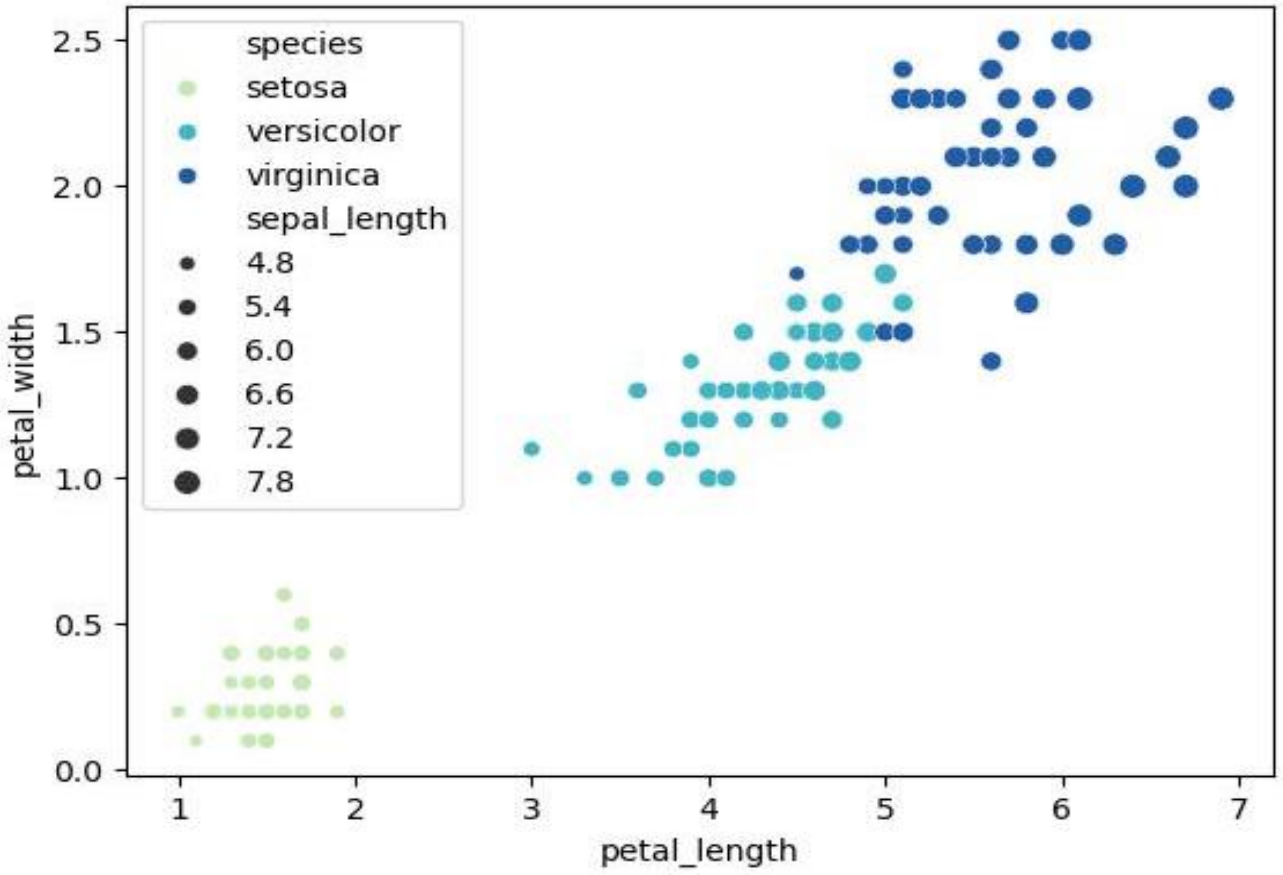
This way you take only the numeric columns, otherwise – you will get an error

Makesure you use the heatmap on correlated data



Correlation scatter plot

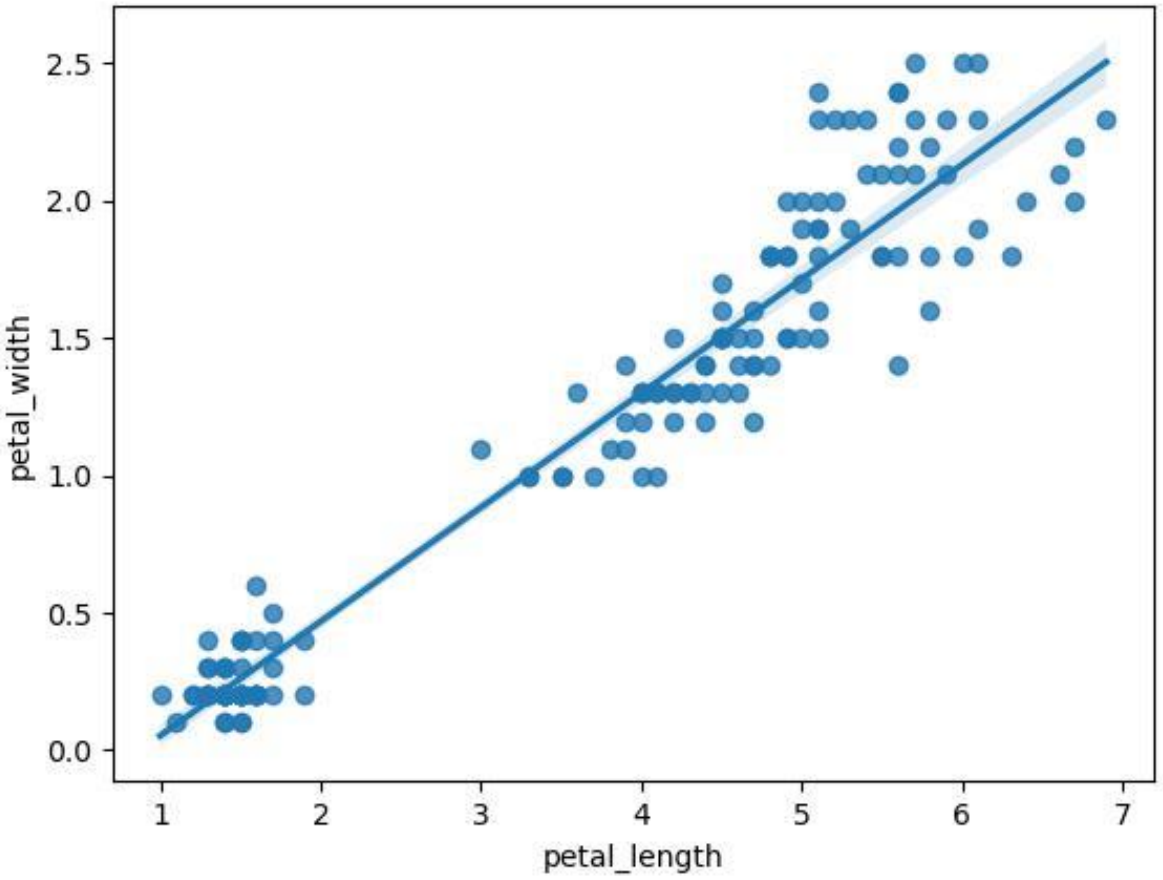
```
sns.scatterplot(x ='petal_length' ,y='petal_width', data =df ,#must
                hue ='species'  ,#for categorial data
                size ='sepal_length' ,
                palette = 'YlGnBu')
```



Correlation regration plot

```
sns.regplot(x='petal_length',y='petal_width',data=df)
```

Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It helps predict values and understand trends, often visualized as a line of best fit in a scatterplot.

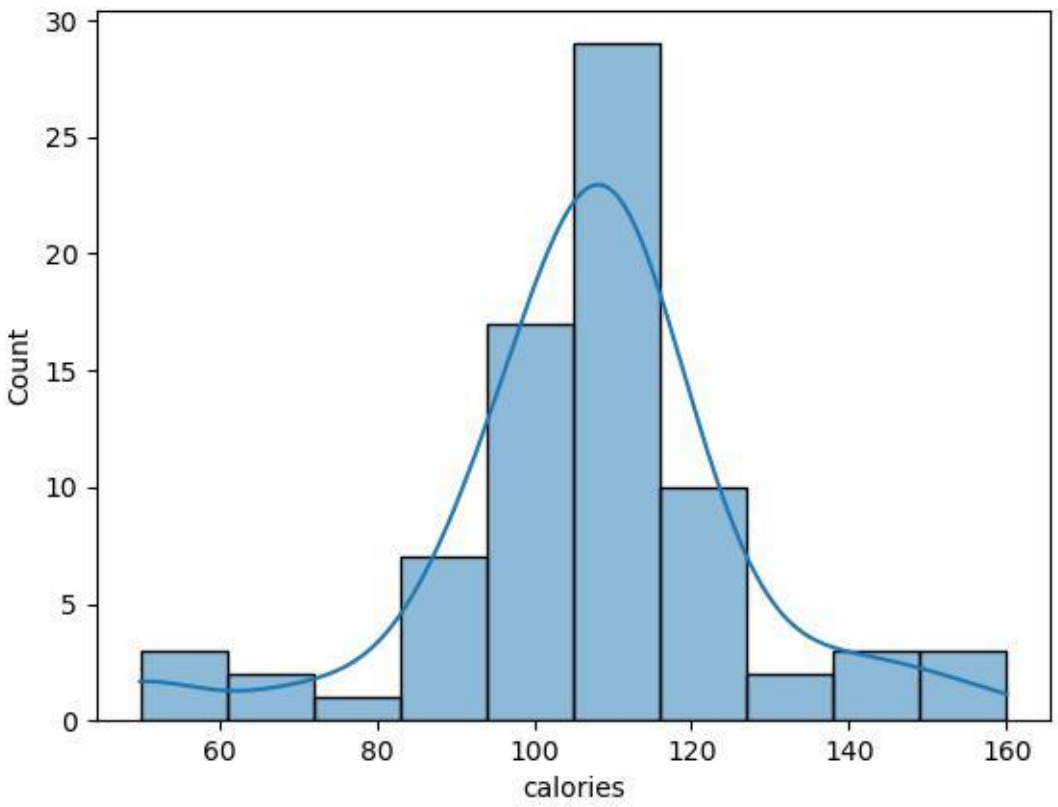


Statistical analysis histogram

```
cereal = pd.read_csv('cereal.csv')
sns.histplot(cereal['calories'],
             bins = 10,
            kde = True)
```

optional

KDE (Kernel Density Estimation) is a technique for visualizing the probability distribution of a dataset. It smooths data points into a continuous curve, making patterns and density easier to interpret.



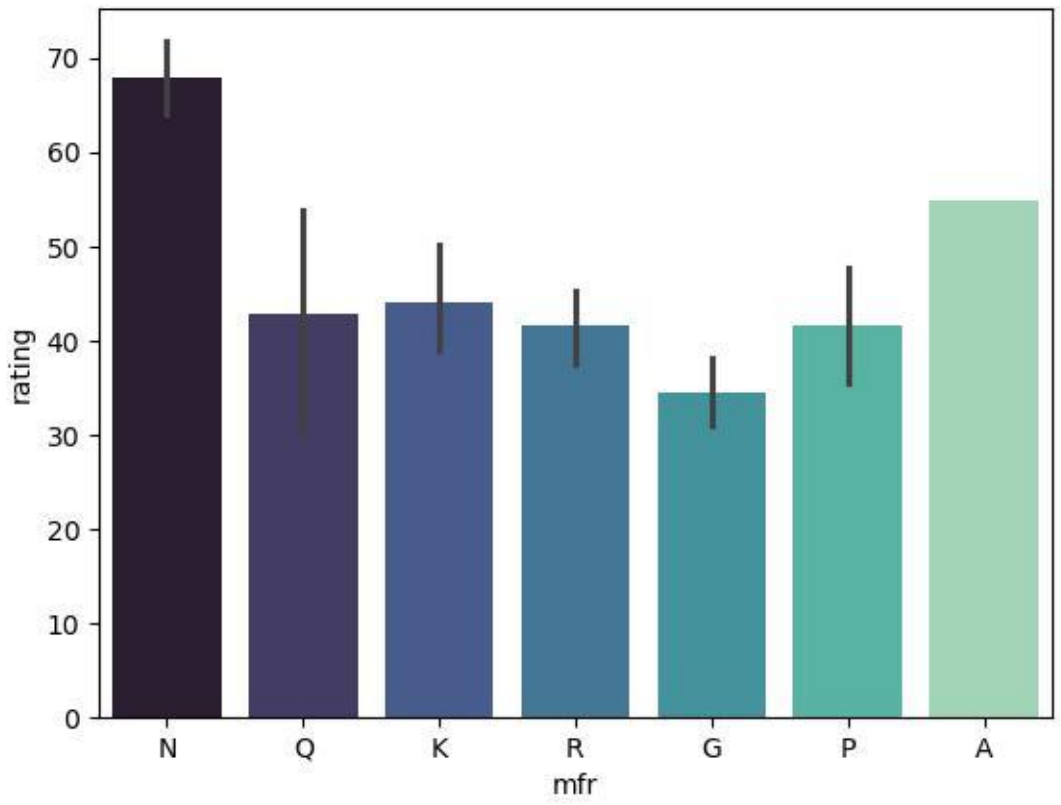
Statistical analysis bar

```
sns.barplot(x ='mfr', y = 'rating' ,data = cereal)
```

$$\text{Confidence Interval} = \text{Mean} \pm Z \cdot \frac{\sigma}{\sqrt{n}}$$

The black line represents the confidence interval .by default we asses the mean, with estimator we can asses other values.

```
sns.barplot(x ='mfr', y = 'rating' ,data = cereal,
            estimator = np.median)
```



Statistical analysis bar

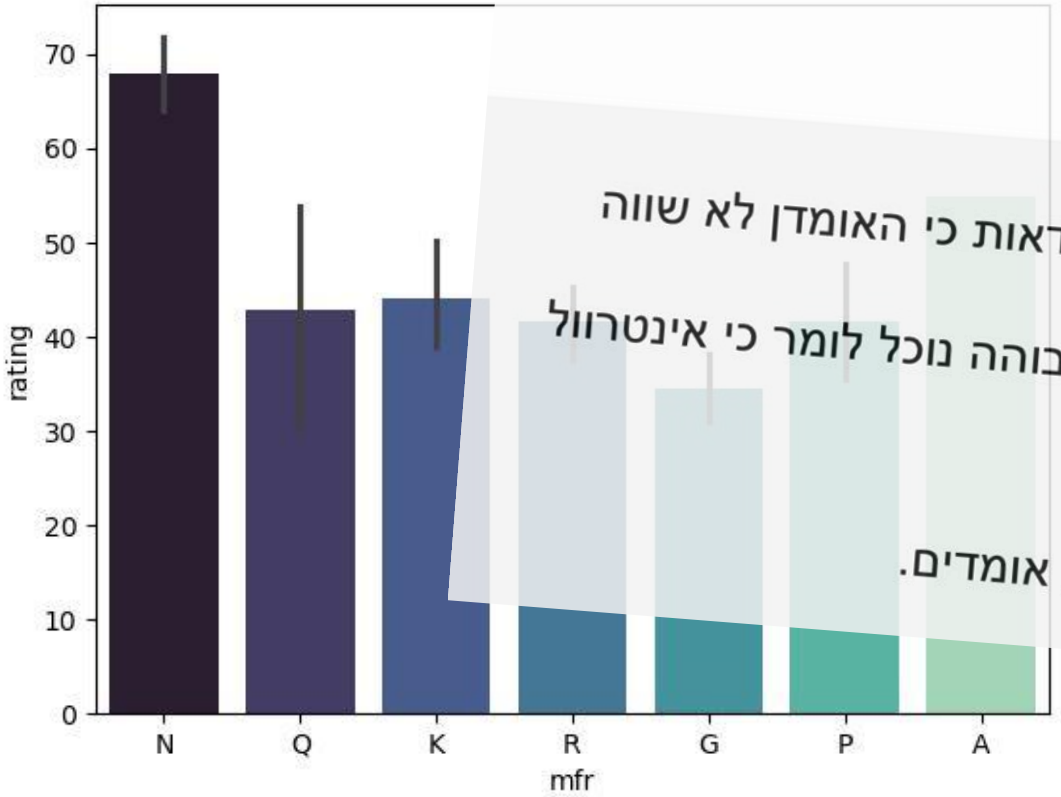
$$\text{Confidence Interval} = \text{Mean} \pm Z \cdot \frac{\sigma}{\sqrt{n}}$$

רווח סמך Confidance Interval

רווח סמך- הוא אינטרוול שבהסתברות מסוימת יכיל את הפרמטר הלא ידוע שאותו הוא אומד.
ההסתברות שרווח הסמך יכיל את הפרמטר הלא ידוע נקראית רמת הסמך, או רמת הביטחון.

אמד נקודתי ± טווח טעות

כאשר אנו אומדים פרמטר לא ידוע באמצעות אמד נקודתי, אנו יודעים בוודאות כי האומדן לא שווה לפרמטר שאותו אנו מנסים לאמוד.
ברצוננו למצוא אינטרוול (רווח) סביב האמד הנקודתי, כך שבהסתברות גבוהה נוכל לומר כי אינטרוול זה מכיל את הפרמטר הלא ידוע.
אינטרוול כזה נקרא בסטטיסטיקה **רווח סמך**.
רווחי הסמך, כמו האמדים הנקודתיים, ישתנו בהתאם לפרמטר אותו אנו אומדים.



רווח סמך Confidance Interval

רווח סמך- הוא אינטרוול שבהסתברות מסוימת יכיל את הפרמטר הלא ידוע שאותו הוא אומד.
ההסתברות שרווח הסמך יכיל את הפרמטר הלא ידוע נקראית רמת הסמך, או רמת הביטחון.

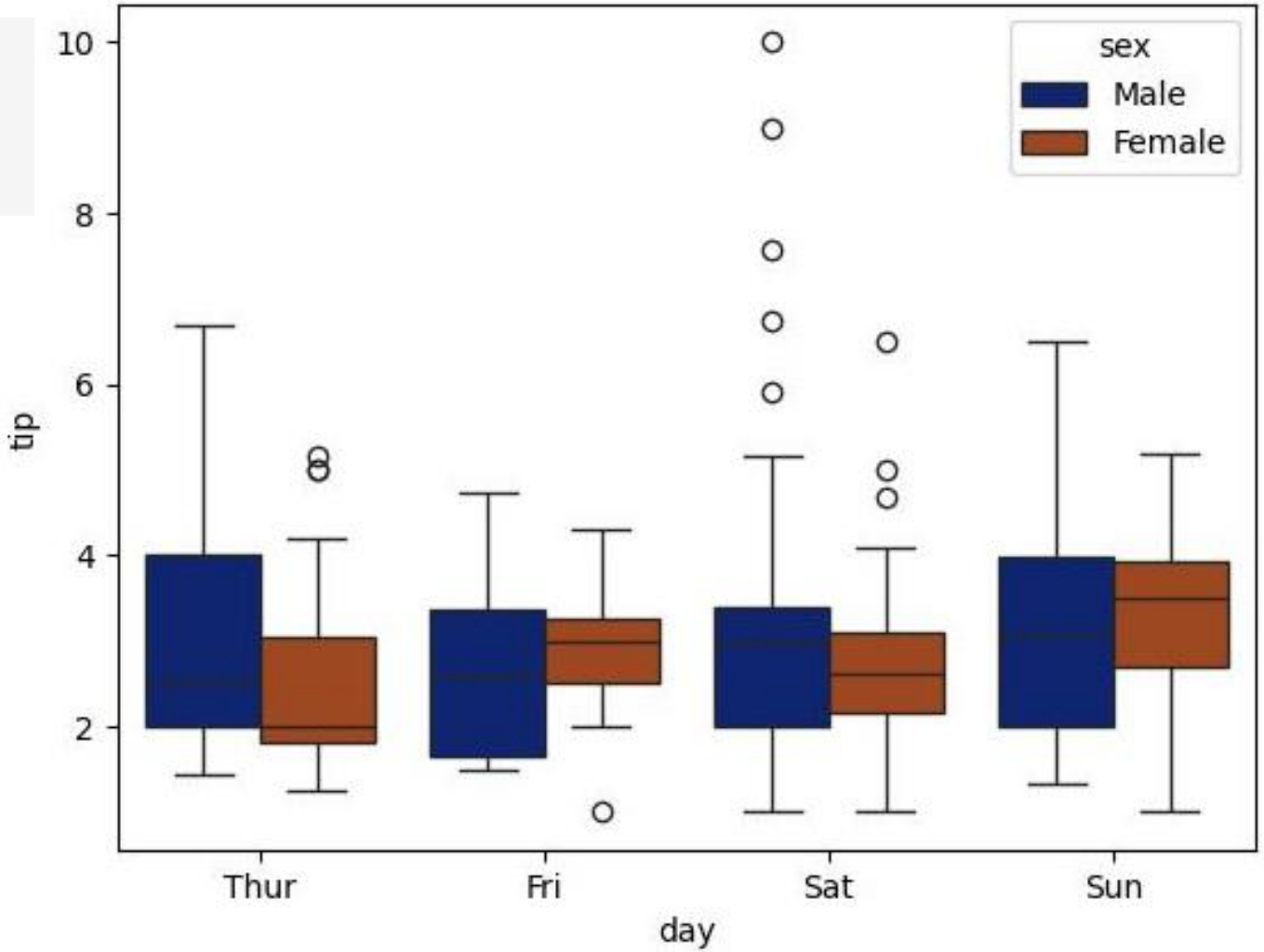
אמד נקודתי ± טווח טעות

כאשר אנו אומדים פרמטר לא ידוע באמצעות אמד נקודתי, אנו יודעים בוודאות כי האומדן לא שווה לפרמטר שאותו אנו מנסים לאמוד.
ברצוננו למצוא אינטרוול (רווח) סביב האמד הנקודתי, כך שבהסתברות גבוהה נוכל לומר כי אינטרוול זה מכיל את הפרמטר הלא ידוע.

אינטרוול כזה נקרא בסטטיסטיקה **רווח סמך**.
רווחי הסמך, כמו האמדים הנקודתיים, ישתנו בהתאם לפרמטר אותו אנו אומדים.

Statistical analysis box

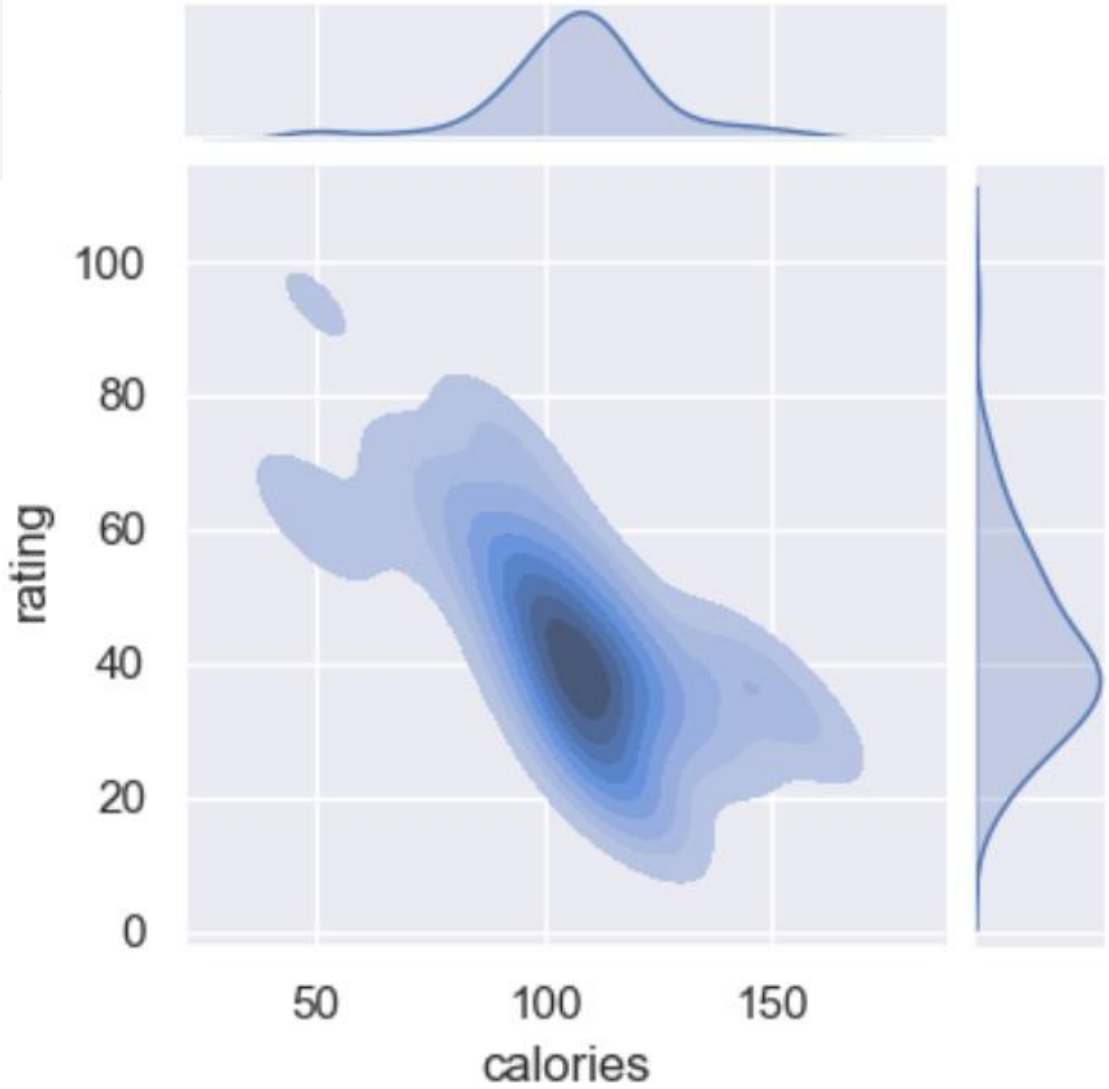
```
sns.boxplot(x='day', y='tip', data = tips,
            hue = 'sex',
            palette = 'dark')
```



A box plot is a graphical representation of data distribution using a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. It highlights outliers and variability in a dataset.

Joint plot

```
sns.jointplot(data = cereal, x = 'calories', y = 'rating'
              ,height = 4, #לא משפיע  figsize
              #להראות סוגים
             kind = 'kde',#hex , reg, kde,scatter...
             fill =True) #with kde
```



pairplot

```
sns.pairplot(titanic.select_dtypes (['number']))
```