

Do not distribute with the  
written consent of Professor  
Arup R. Das

# DS 520/CS 520 – Lecture 2

**DS-520-50: Data Analy: Conc/Tech**  
**2024 Fall**  
**MONMOUTH CAMPUS**

**T 7:30 PM - 10:20 PM**  
**9/3/2024 - 12/9/2024**  
**Howard Hall, 306 LECTURE**

**Professor Arup Das**  
[adas@Monmouth.edu](mailto:adas@Monmouth.edu)

**Disclaimer:**

- The views expressed are solely those of the presenter and not affiliated with any other party.
- This presentation is free of copyright violations, and external sources have been appropriately credited.
- **The content within this presentation is legally protected; unauthorized reproduction, including photography, will result in legal action.**
- **This material is not intended for distribution and must remain solely within the confines of this class.**  
**Do not distribute slides or assignments to other students**
- **Using cameras to take screenshots or photographs of the slides is strictly prohibited.**

# Course Logistics

Date	Week	Class Format/Location/Time	Topics	Readings Required (Due before class)	Assignment/Quiz
September 3,2024	Week_1	On-Premise/Howard Hall, 206 LECTURE/7:30 PM-10:20PM	Python for Machine Learning Refresher		
September 10,2024	Week_2	On-Premise/Howard Hall, 206 LECTURE/7:30 PM-10:20PM	Machine Learning Workflow and Exploratory Data Analysis (EDA)	Book 1 – Chapter 1, Chapter 2 (Pages 22 – Page 29)	
September 17,2024	Week_3	On-Premise/Howard Hall, 206 LECTURE/7:30 PM-10:20PM	Machine Learning Workflow and Exploratory Data Analysis (EDA)	Book 1 – Chapter 2 (Page 32- 63)	<b>Project 1 Distributed - Due Sep 27,2024</b>
September 24, 2024	Week_4	On-Premise/Howard Hall, 206 LECTURE/7:30 PM-10:20 PM	Machine Learning for Regression	Book 1 – Chapter 2 (Page 32- 63)	
October 1, 2024	Week_5	On-Premise/Howard Hall, 206 LECTURE/7:30 PM-10:20 PM	Machine Learning for Regression	Book 1 – Chapter 3 (Page 65- 110)	<b>Project 2 Distributed - Due Oct 12, 2024</b>
October 6, 2024	Week_6	On-Premise/Howard Hall, 206 LECTURE/7:30 PM-10:20 PM	Machine Learning for Classification	Book 1 – Chapter 4 (Page 113- 145)/Chapter 6	
October 22, 2024	Week_7	Zoom remote/8:00pm – 10:30 pm	Machine Learning for Classification	Book 1 – Chapter 4 (Page 113- 145), /Chapter 6	<b>Project 3 Distributed – Due Nov 1, 2024</b>
October 29,2024	Week_8	Zoom remote/8:00pm – 10:30 pm	Quiz 1 (Cover materials from Week 1-7)	Book 1 – Chapter 3 ( Pages 88- 92)	<b>Quiz 1 – Open Book/Open Notes</b>
November 5, 2024	Week_9	Zoom remote/8:00pm – 10:30 pm	Feature Selection, Model Selection and Tuning	Book 1 – Chapter 4 ( Pages 147- 151)	
November 12, 2024	Week_10	Zoom remote/8:00pm – 10:30 pm	Feature Selection, Model Selection, and Tuning	Book 1 – Chapter 4 ( Pages 147- 151)	<b>Project 4 Distributed – Due Nov 22, 2024</b>
November 19,2024	Week_11	Zoom remote/8:00pm – 10:30 pm	Unsupervised Learning	Professor Lecture Notes	
November 26, 2024	Week_12	Zoom remote/8:00 pm – 10:30 pm	Unsupervised Learning	Professor Lecture Notes	<b>Project 5 Distributed – Due Dec 6, 2024</b>
December 3, 2024	Week_13	Zoom remote/8:00 pm – 10:30 pm	AI Certifications Overview or additional topics spill over from preceding weeks	Professor Lecture Notes	<b>Quiz 2 Distributed</b>
<b>December 9, 2024 – Last day of class</b>	Week_14	Quiz 2 Due	Quiz 2 (Covers materials from Weeks 9 -12) and Course wrap-up		<b>Quiz 2 Due Dec 8, 2024 before midnight EST</b>

# Course Logistics

---

1. OneDrive link for professor notes and assignments/quiz
2. Check your Monmouth email for announcements
3. Check your Monmouth calendar for Zoom links for office hours and remote lectures
4. My contact information: [adas@monmouth.edu](mailto:adas@monmouth.edu), Cell # 917-523-7683
4. Office hours (zoom only) – Friday (EST) 7-7:30 pm EST
5. Assignment submission to [professoraruprdas@gmail.com](mailto:professoraruprdas@gmail.com) ( Notation for files: Assignment\_1\_Name\_of\_Student), Colab notebooks ipynb file and html file, all presentation in ppt format.
6. Quiz submission to [professoraruprdas@gmail.com](mailto:professoraruprdas@gmail.com) (Notation for file : Quiz\_1\_Name\_of\_Student. doc , Quiz\_2\_Name\_of\_Student.doc)

# Topics

---

1. Python for Data Sciences Pandas
2. Python for Data Sciences EDA
3. EDA Analysis

# Study Groups

---

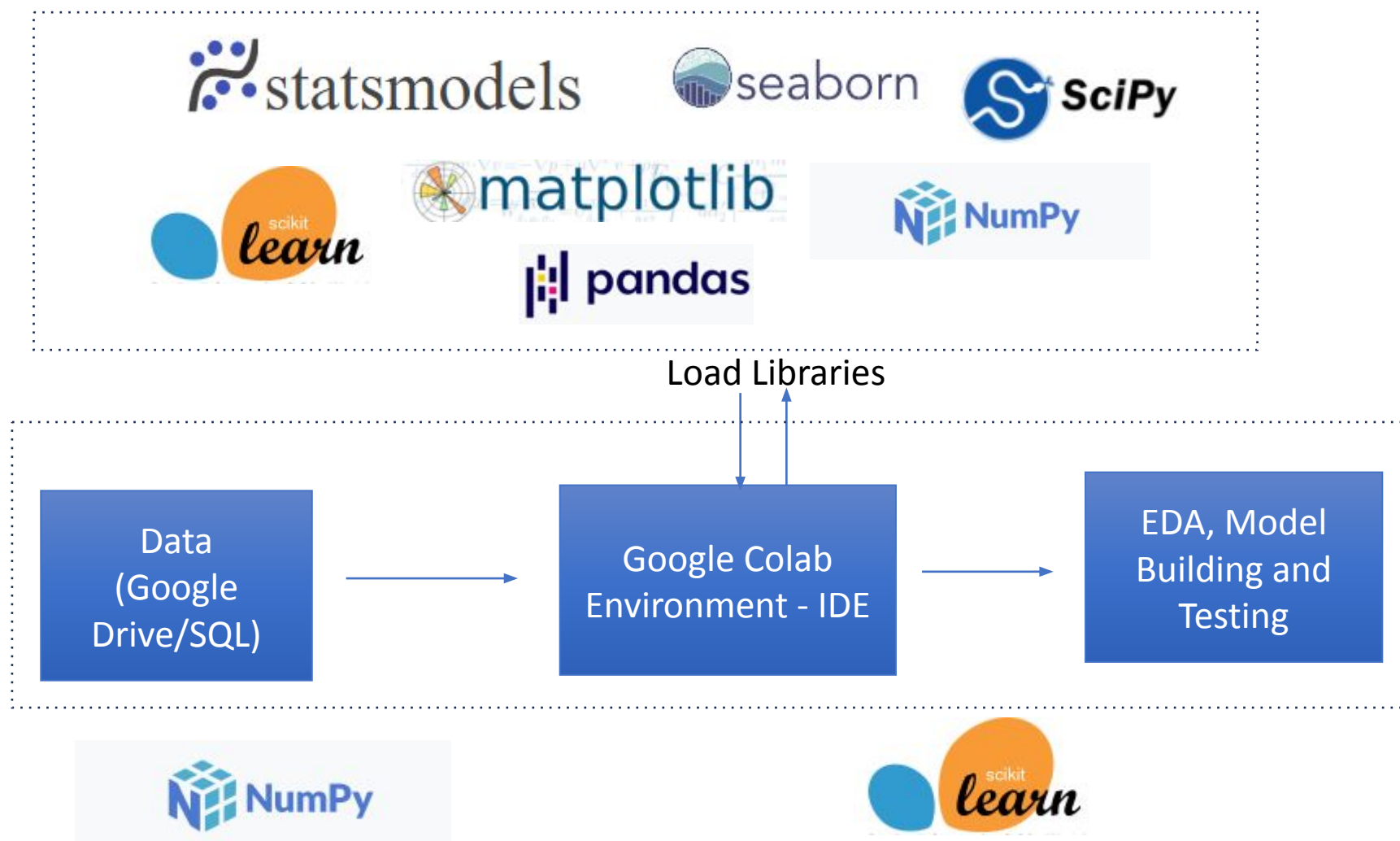
Group 1	Group 2	Group 3	Group 4
Andrew Captano - Group Leader	Kiran Ramjisingh - Group Leader	Shobharani Polasa- Group Leader	Sashank Vaddiparty- Group Leader
Ryan Sonn	Killariben Limbachiya	John Costa	Katia Bravo Bendezu
Alonso Aguilera	Ezzine Ndumnwere	Wang Dongyang	Amulya Konduru
Noah Ferker		Vibushan Raju Guduri	

Do not distribute with the  
written consent of Professor  
Arup R. Das

# Python for Data Science - Pandas



# ML Libraries



**NumPy:** NumPy stands for Numerical Python and it is a core scientific computing library in Python. It provides efficient multi-dimensional array objects and various operations to work with these array objects.

**Scikit-Learn:** Also known as Sklearn provides advanced analytics tools combined with complex machine learning capabilities. This allows you to build more sophisticated models, performing more complex and multivariate regressions, as well as data preprocessing.



**Pandas:** It offers a plenty of tools to manipulate, analyze, and even represent data structures and complex datasets. This includes time series and more complex data structures such as merging, pivoting, and slicing tables to create new views and perspectives on existing



**Matplotlib:** It is a comprehensive library for creating static, animated, and interactive visualizations in Python.



**Seaborn:** Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

# Important Numpy Functions

`np.array()` - To create an array

`np.zeros()` - To create an array of zeros

`np.ones()` - To create an array of ones

`np.random.randn()` - To create an array of specified shape filled with random values

`np.dot()` - Dot product

`np.transpose()` - Permute array dimensions

`np.concatenate()` - Concatenate two arrays

<https://numpy.org/>

# Important Pandas Functions

<https://www.kaggle.com/learn/pandas>

<https://pandas.pydata.org/>

[https://pandas.pydata.org/docs/user\\_guide/index.html](https://pandas.pydata.org/docs/user_guide/index.html)

df.head() - To get the top rows

df.tail() - To get the last rows

df.describe() - To get the quick statistic summary

pd.concat() - To concatenate two pandas objects

pd.merge() - To merge the pandas dataframes

df.groupby() - To split, apply or combine the data structures

df.astype() - To convert into some other data types

df.value\_counts() - To get count of some attributes

df.unique() - To get unique values

df.dtypes - To get the data types

df.shape - To get the shape (number of rows and columns)

# Working with Pandas\_ Notebook\_1

olympics\_dataset.csv

11 columns

0	1									
player_id	Name	Sex	Team	NOC	Year	Season	City	Sport	Event	Medal
0										
1										

252,565, Rows

Data is loaded into pandas dataframe for analysis



# **Python for Data Science - Exploratory Data Analysis (EDA)\_Part\_1 using Pandas, Matplotlib and Seaborn**

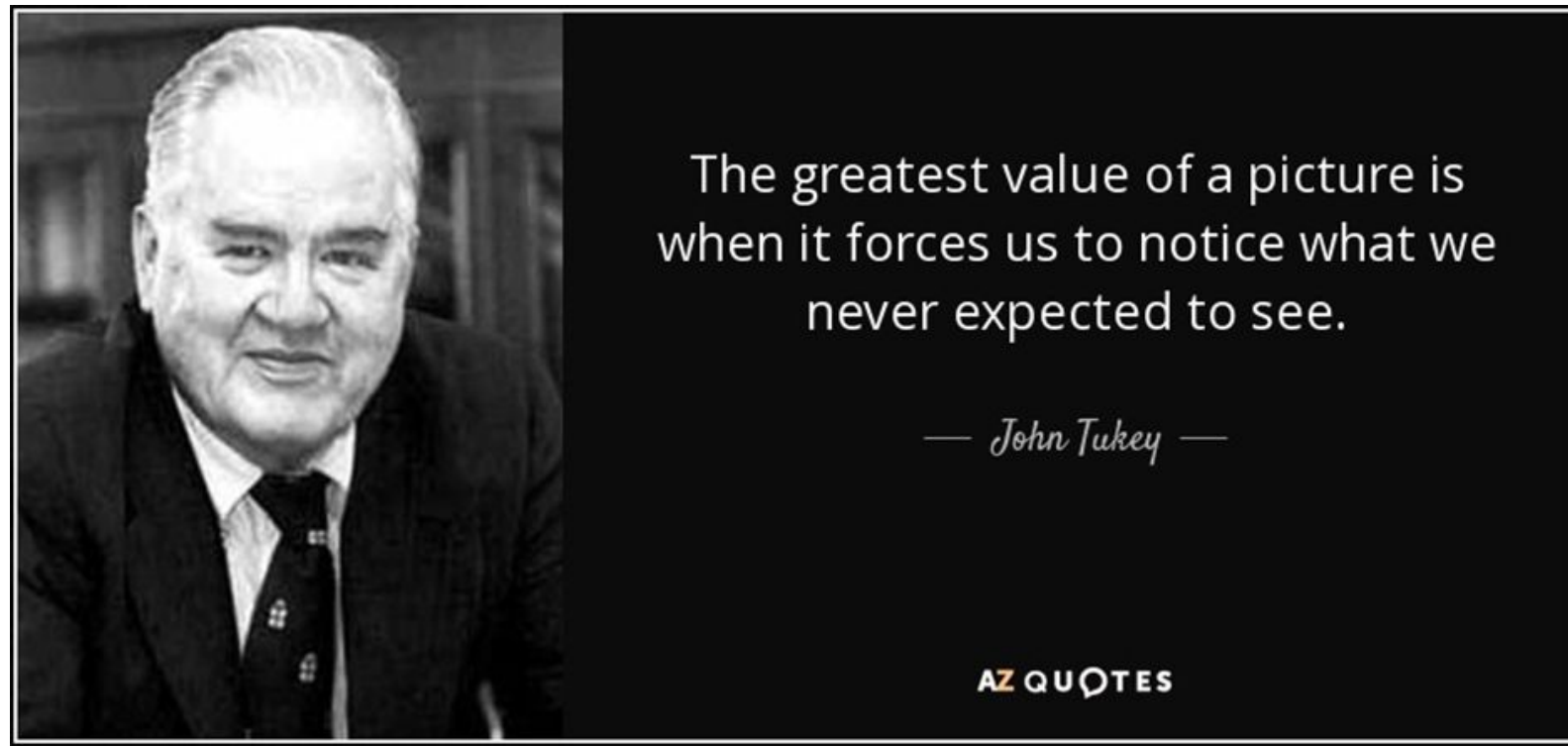
# **Python for Data Science - Exploratory Data Analysis (EDA)\_Part\_1 using Pandas, Matplotlib and Seaborn**

## **Notebook\_2**

## What is EDA

---

Exploratory Data Analysis is an integral part of working with data to to understand the trends, patterns, and relationships among various entities present in the data set. EDA can be carried out using one of three main techniques: univariate, bivariate, or multivariate analysis.



**The goal of the EDA is to get the data to life.  
to understand the underlying story**



# Python Visualization Libraries



[https://matplotlib.org/stable/plot\\_types/index.html](https://matplotlib.org/stable/plot_types/index.html)



<https://seaborn.pydata.org/>

[https://seaborn.pydata.org/tutorial/color\\_palettes.html](https://seaborn.pydata.org/tutorial/color_palettes.html)

[https://www.practicalpythonfordatascience.com/ap\\_seaborn\\_palette](https://www.practicalpythonfordatascience.com/ap_seaborn_palette)

We will be using seaborn which is build on matplotlib , For HW and quiz you need to use seaborn and matplotlib

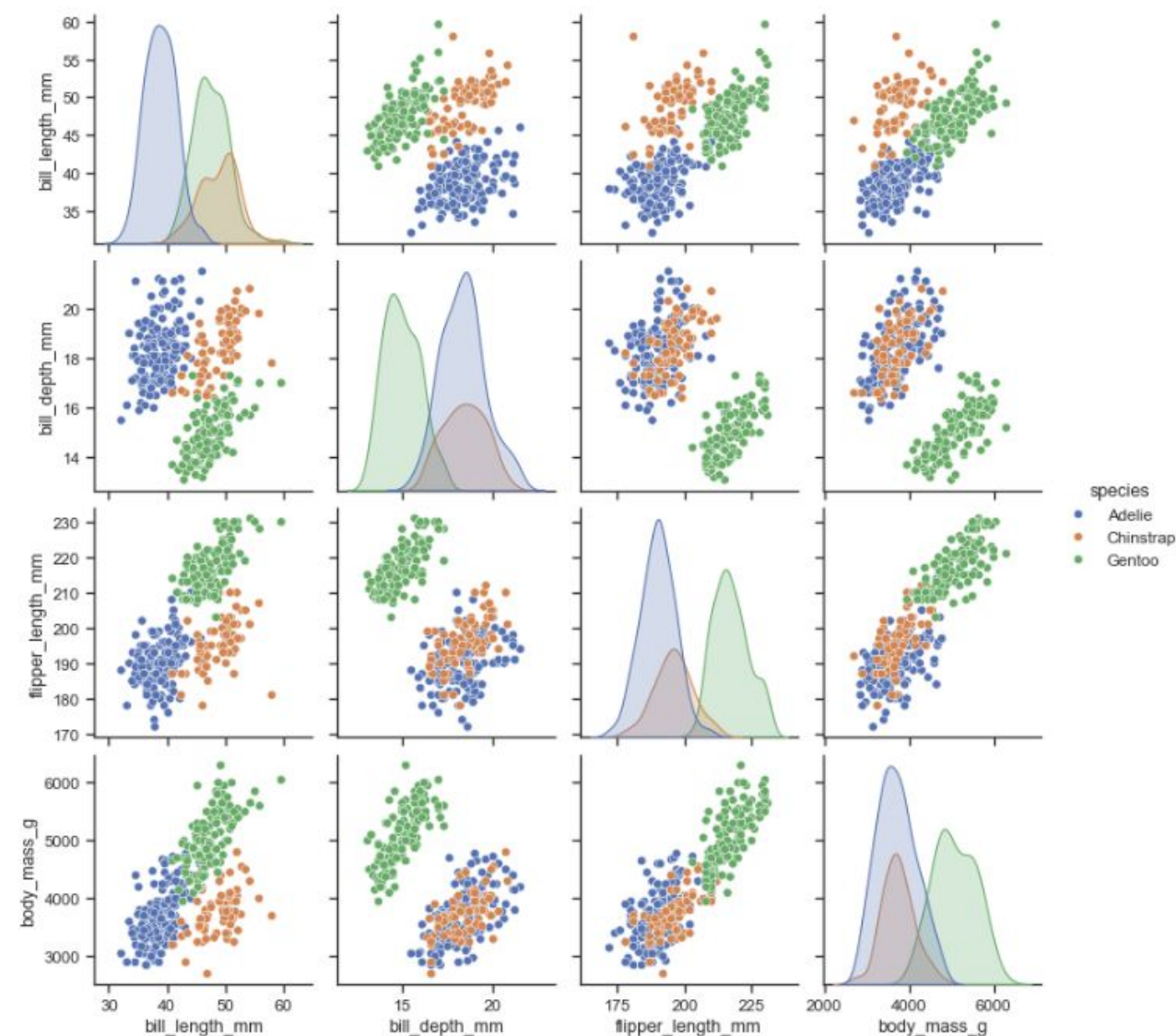


Optional interactive production tool

<https://plotly.com/>

<https://plotly.com/examples/>

## Scatterplot Matrix



seaborn components used: `set_theme()`, `load_dataset()`, `pairplot()`

```
import seaborn as sns
sns.set_theme(style="ticks")


df = sns.load_dataset("penguins")
sns.pairplot(df, hue="species")
```




# Data Types in Python

Machine learning models rely on four primary data types.

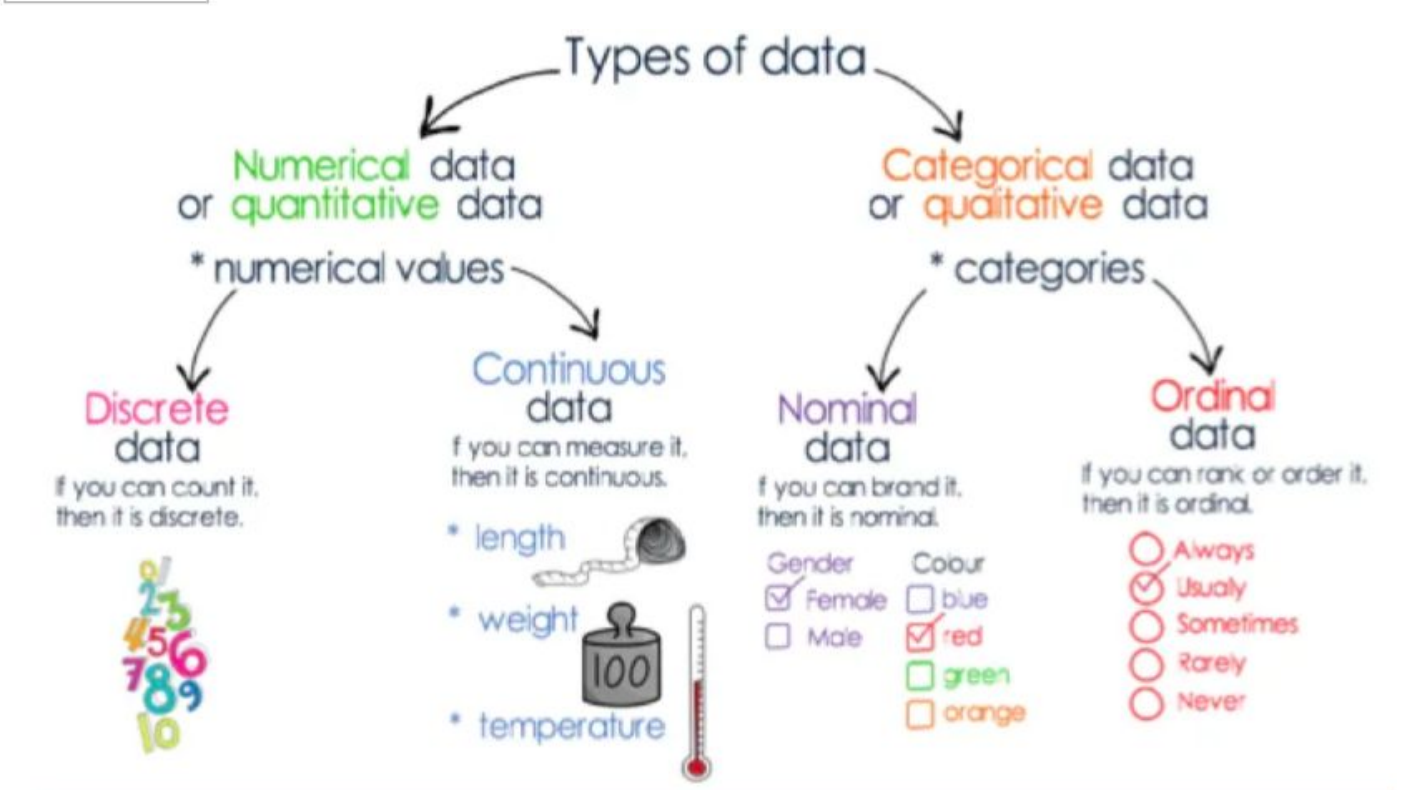
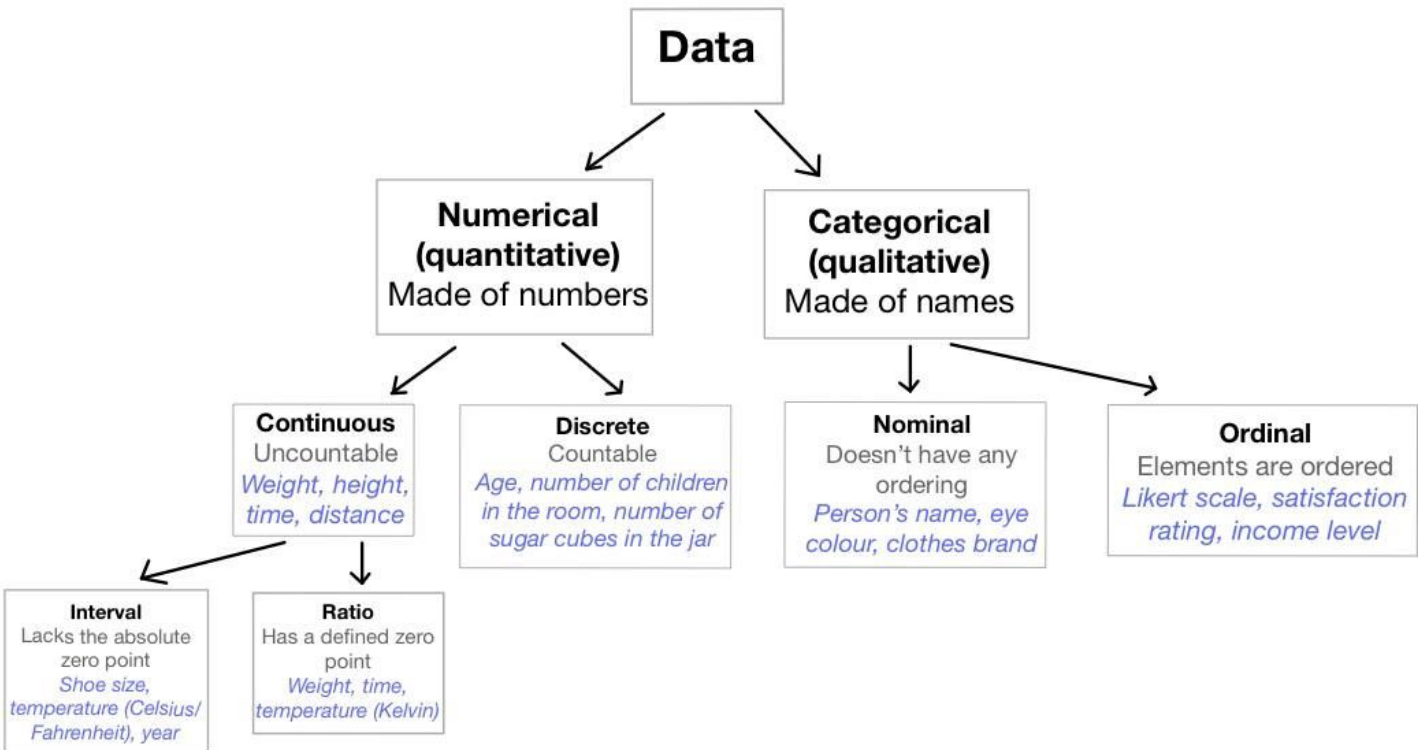
123  
Numerical Data

  
Categorical Data

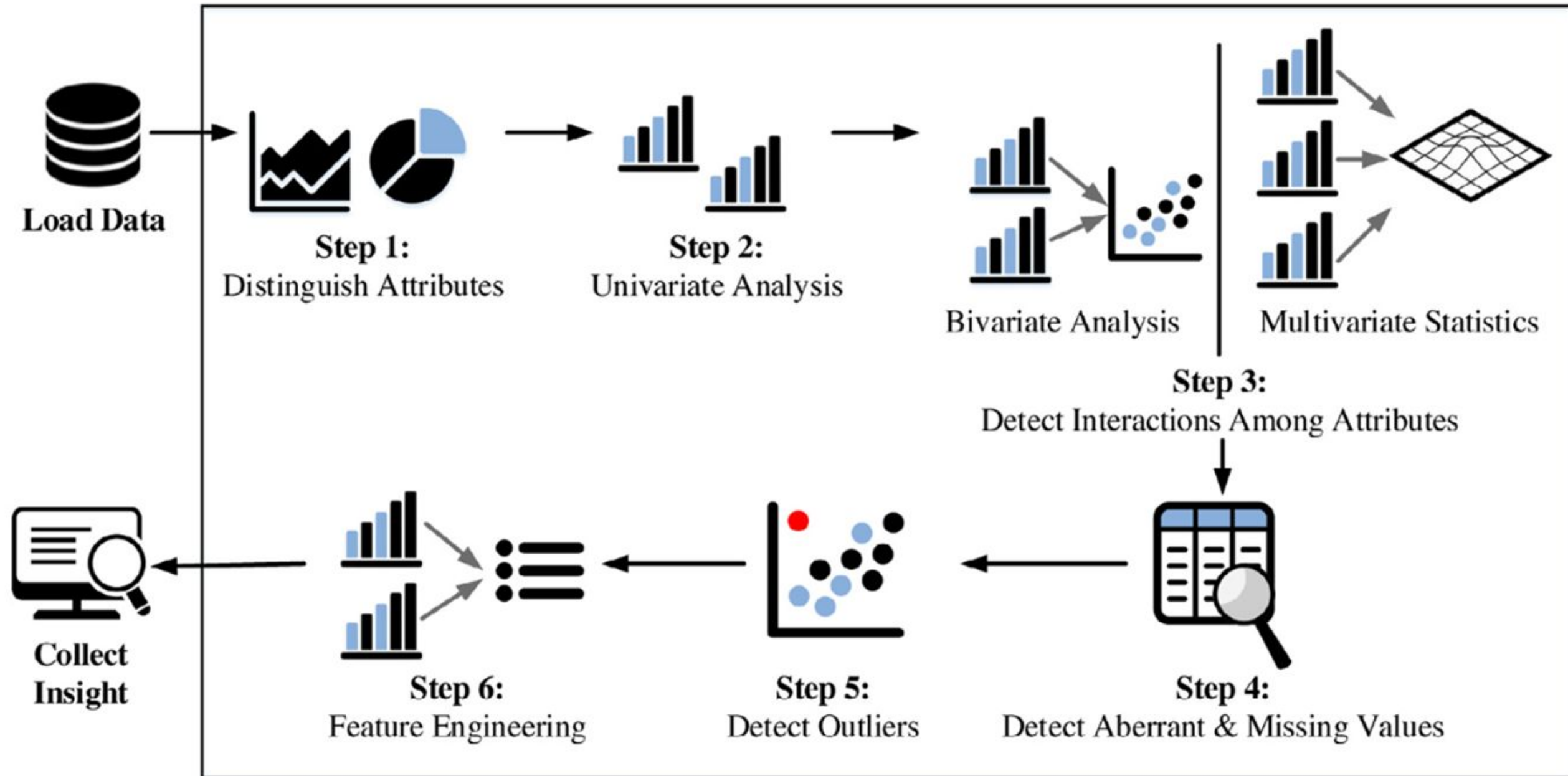
  
Time Series Data

[ text ]  
Text Data

Pandas dtype	Python type	NumPy type	Usage
object	str	string_, unicode_	Text
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float_, float16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values



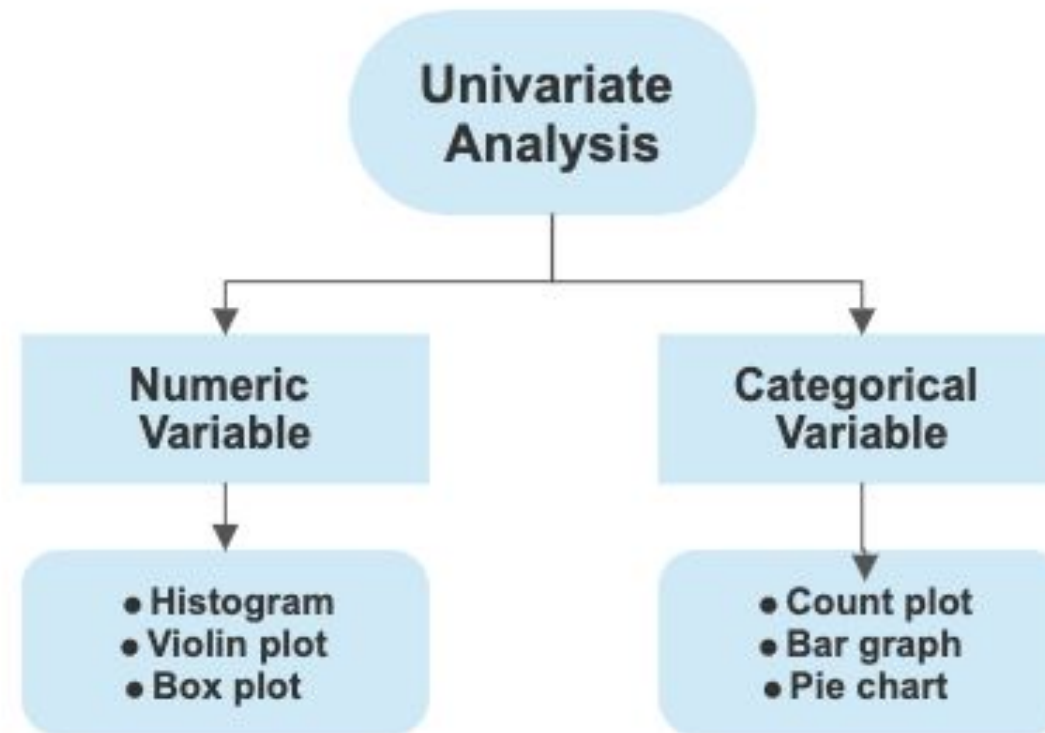
# EDA Process



# Univariate Analysis

---

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words, your data has only one variable. The goal of the univariate analysis is to derive, define, and summarize data and analyze patterns within it. This is done by looking at the mean, median, mode, spread, variance, range, standard deviation, etc. Since Univariate analysis is the analysis of single variables



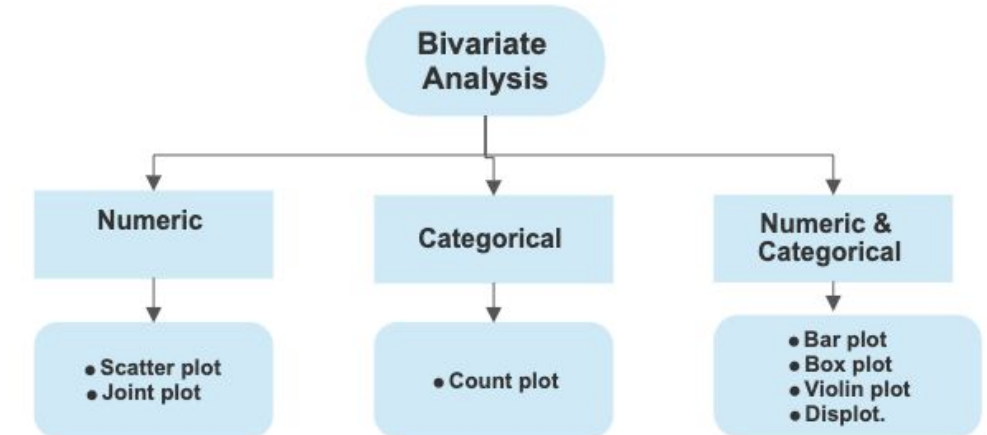


# Bi-variate Analysis

Bivariate analysis implies analyzing the relationship between two variables. These variables are usually denoted by X, and Y. Bivariate studies are used to examine whether there is a statistical relationship between two variables, how strong that relationship is, and whether one variable can be predicted from another.

The kind of bivariate analysis depends on the kind of attributes and variables used to analyze the data. Generally, we have two types of data, numerical and Categorical, and hence we can perform data analysis on the below combination of data

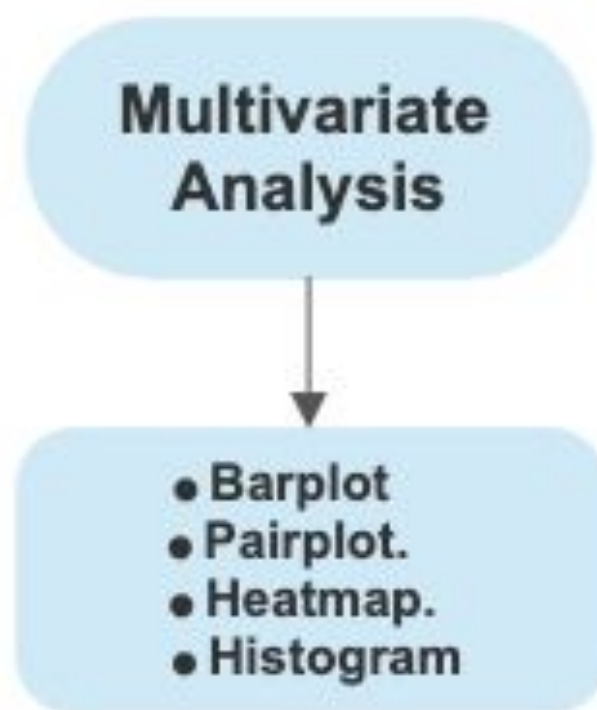
1. **Numerical and numerical** — Both variables have a numerical value in this bivariate correlation.
2. **Categorical and categorical** — In this, both the variables have categorical value
3. **Numerical and categorical** — One variable is numerical, and the other is categorical.



## Multivariate Analysis

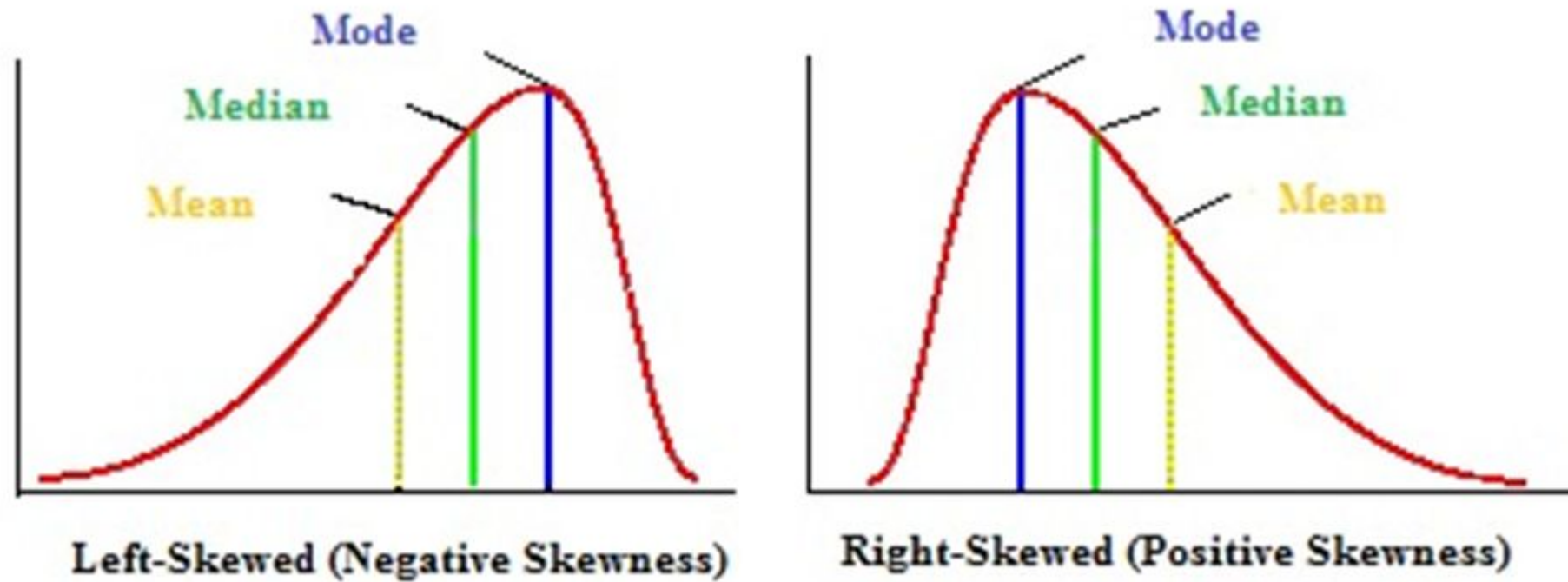
---

Multivariate analysis implies analyzing the relationship between numerous variables (more than two). We can use pair plots and heatmaps to visualize more than two variables.



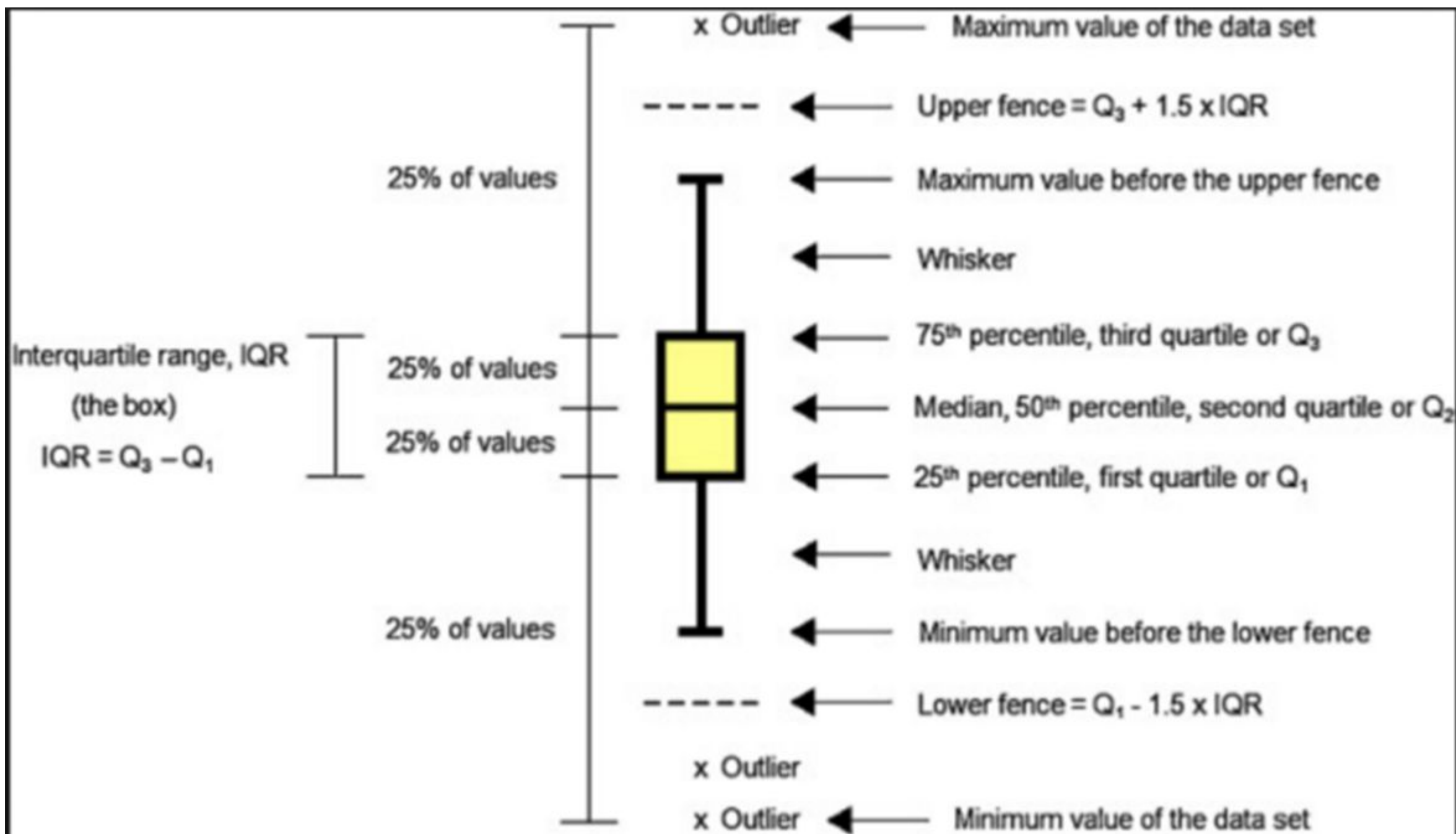
## Plot types - Distribution

---

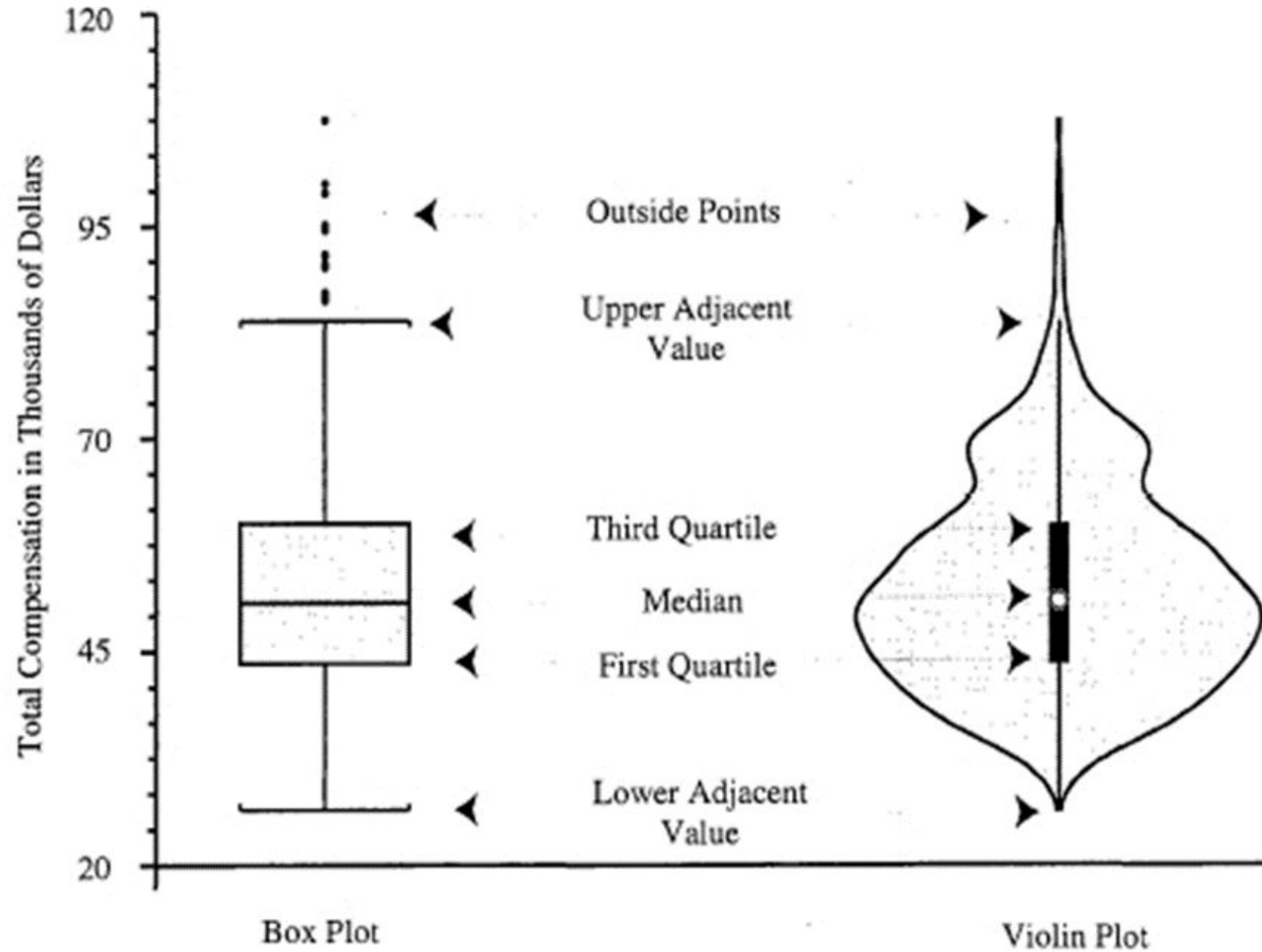


<https://www.statisticshowto.com/probability-and-statistics/skewed-distribution/>

## Plot types - Box Plot



## Plot types - Violin

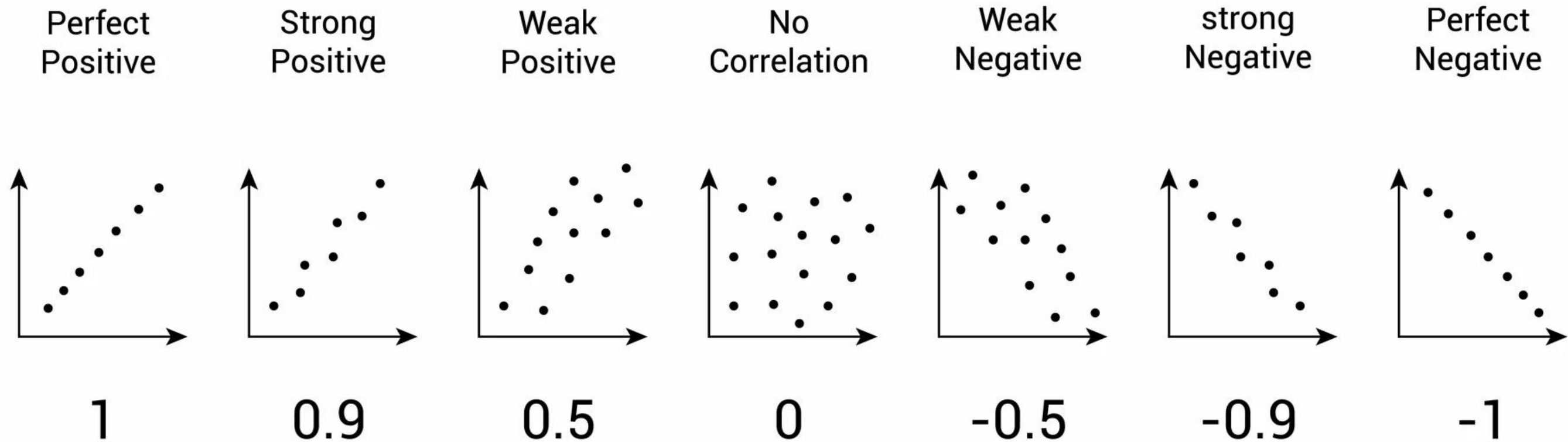


*Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.*



## Plot types - Correlation ( Multivariate Numerical)- Heatmap

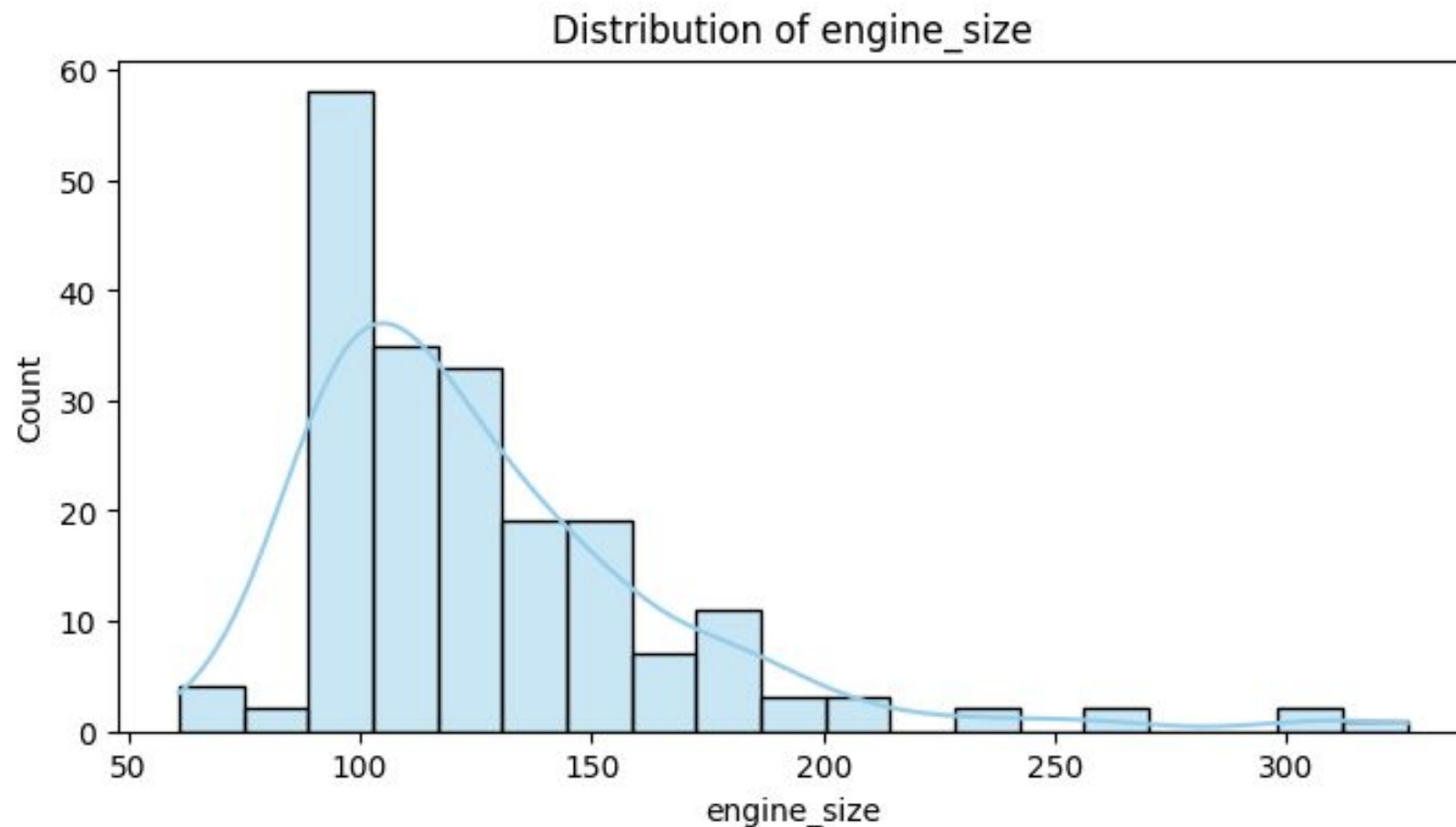
---



# **Python for Data Science - Exploratory Data Analysis (EDA)\_Part\_2 - *Analysis of Graphs (Very Critical) - Your Assignment and Quiz is going to have a lot of question on the analysis***

## **Notebook\_2**

# Univariate - Numerical

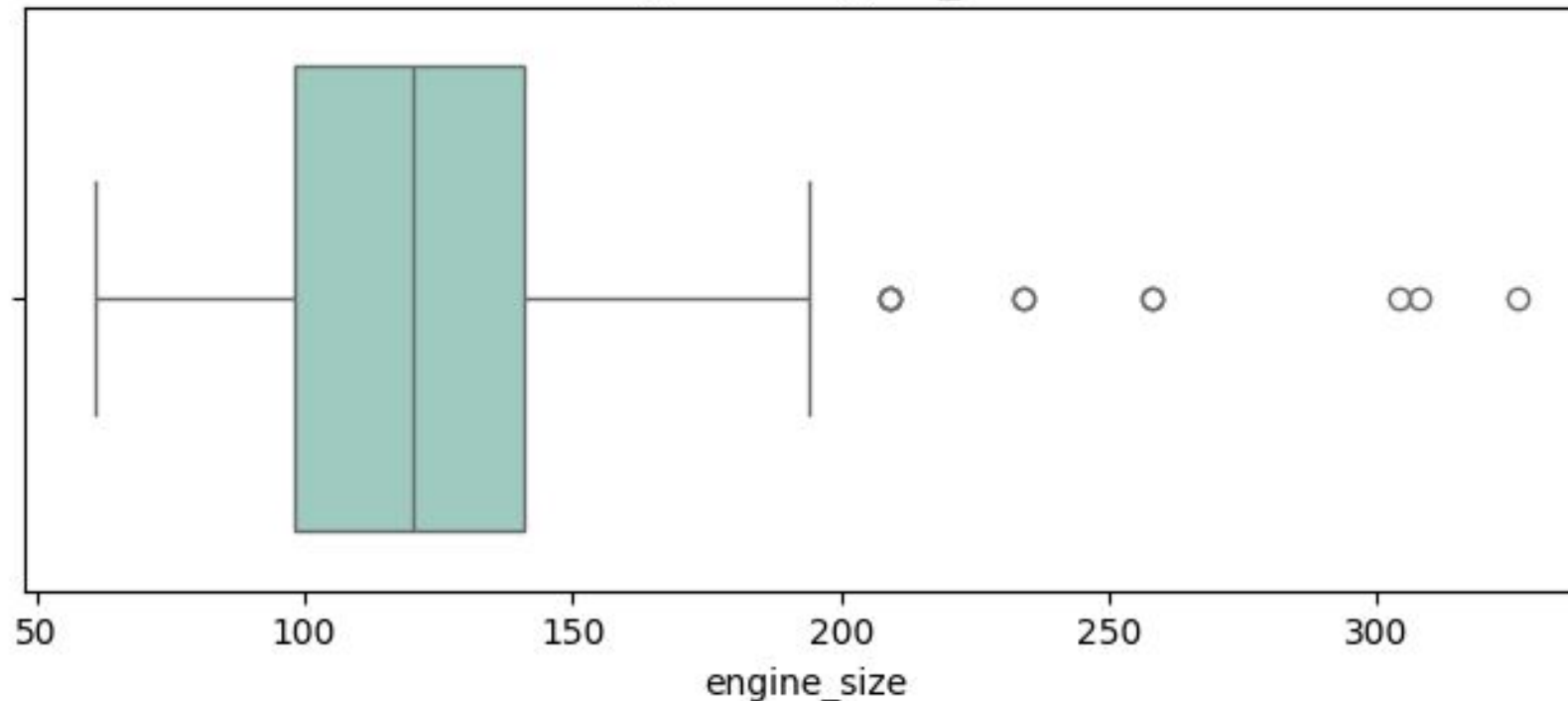


The graph shows the **distribution of engine sizes** with the x-axis representing engine size and the y-axis representing the count of observations (frequency). The distribution is right-skewed, meaning most engine sizes are concentrated on the lower end, around 100, and gradually taper off as the engine size increases.

## Key Business Insights:

- Most products in the dataset likely have **smaller engine sizes**, indicating a focus on fuel-efficient or compact engines.
- The **larger engines** are less frequent, possibly pointing to a niche market for high-performance or specialized vehicles.
- Businesses can tailor their offerings or marketing strategies to the more popular smaller engine sizes while catering to high-end clients with larger engines.

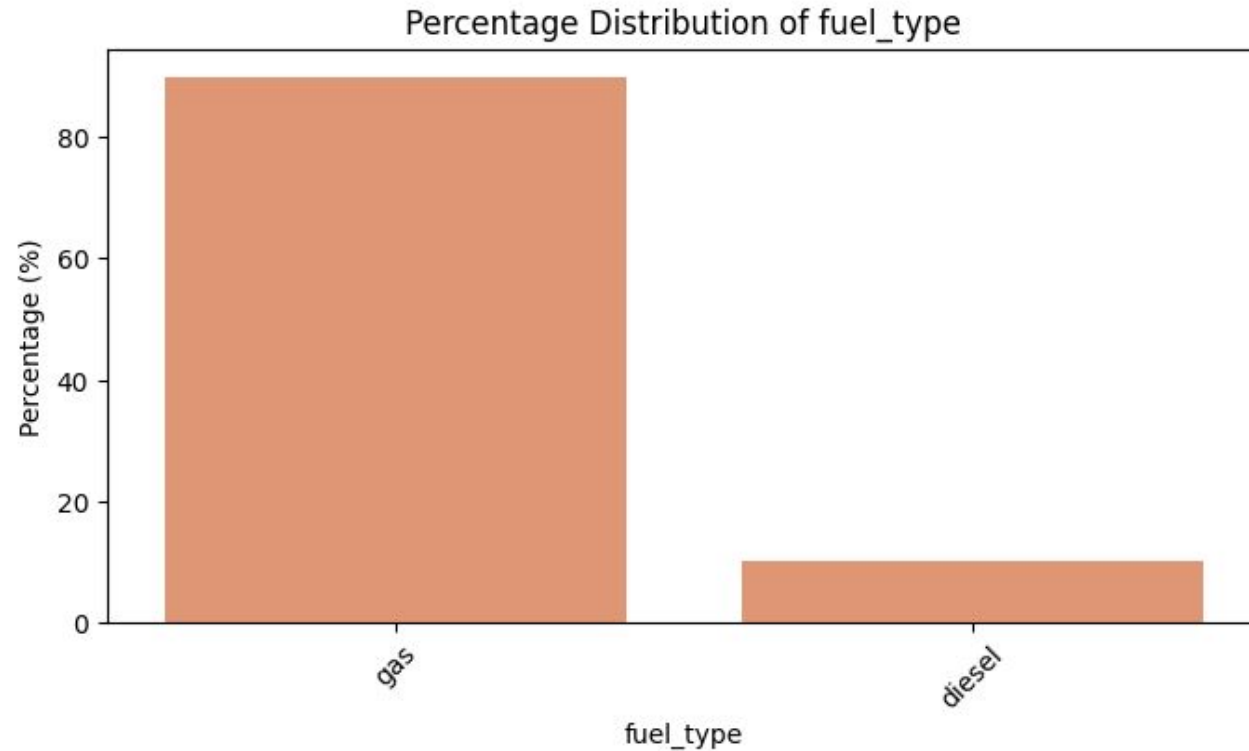
Boxplot of engine\_size



## Key Business Insights:

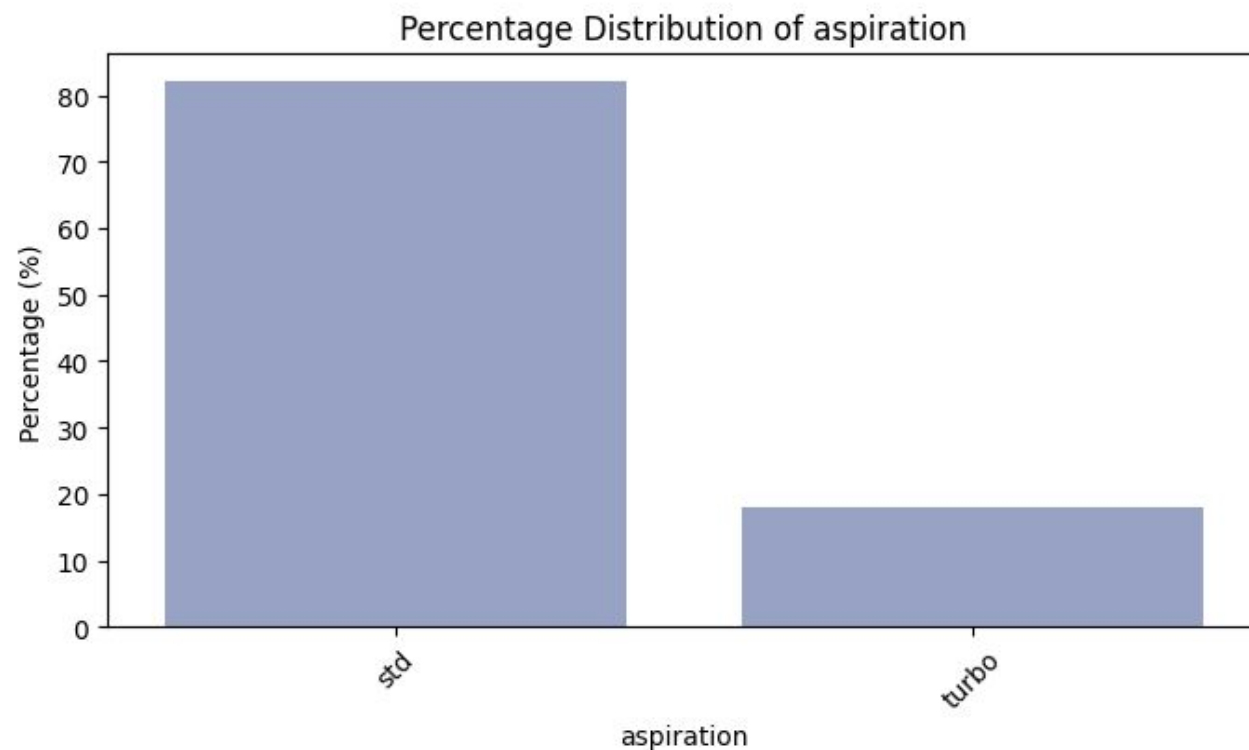
- Majority of Vehicles Have Moderate Engine Sizes:**
  - The bulk of the vehicles in this dataset have engine sizes between approximately 100 and 175 units. This suggests that most vehicles being analyzed are likely in the mid-range category, which could include standard sedans, compact cars, or family vehicles. For a business focusing on this segment, these are likely the vehicles to cater to in terms of products or services (e.g., parts, servicing, fuel efficiency solutions).
- Outliers Represent Larger Engines (High-End or Performance Vehicles):**
  - There are several vehicles with much larger engine sizes (above 200), representing outliers in the data. These outliers could correspond to high-performance or luxury vehicles, which may require special attention from a business perspective.
  - Opportunities:** If your business is targeting high-end customers, these outliers signal potential opportunities in providing premium products, aftermarket parts, or specialized services for these vehicles.
  - Risks:** However, if these high-end vehicles are not your target market, they could skew your overall analysis and mislead forecasting or inventory planning. These outliers could inflate demand expectations if not carefully accounted for, especially in a market where mid-sized vehicles dominate.

# Univariate - Categorical



## Top Plot: Percentage Distribution of Fuel Type

- The majority of vehicles in the dataset run on **gas** (over 85%), while **diesel** makes up a small portion (less than 15%).
- For businesses, this suggests that products or services should focus primarily on gas-powered vehicles, which dominate the market.
- Diesel vehicles represent a niche segment, offering potential for specialized products or services if targeting diesel engines.

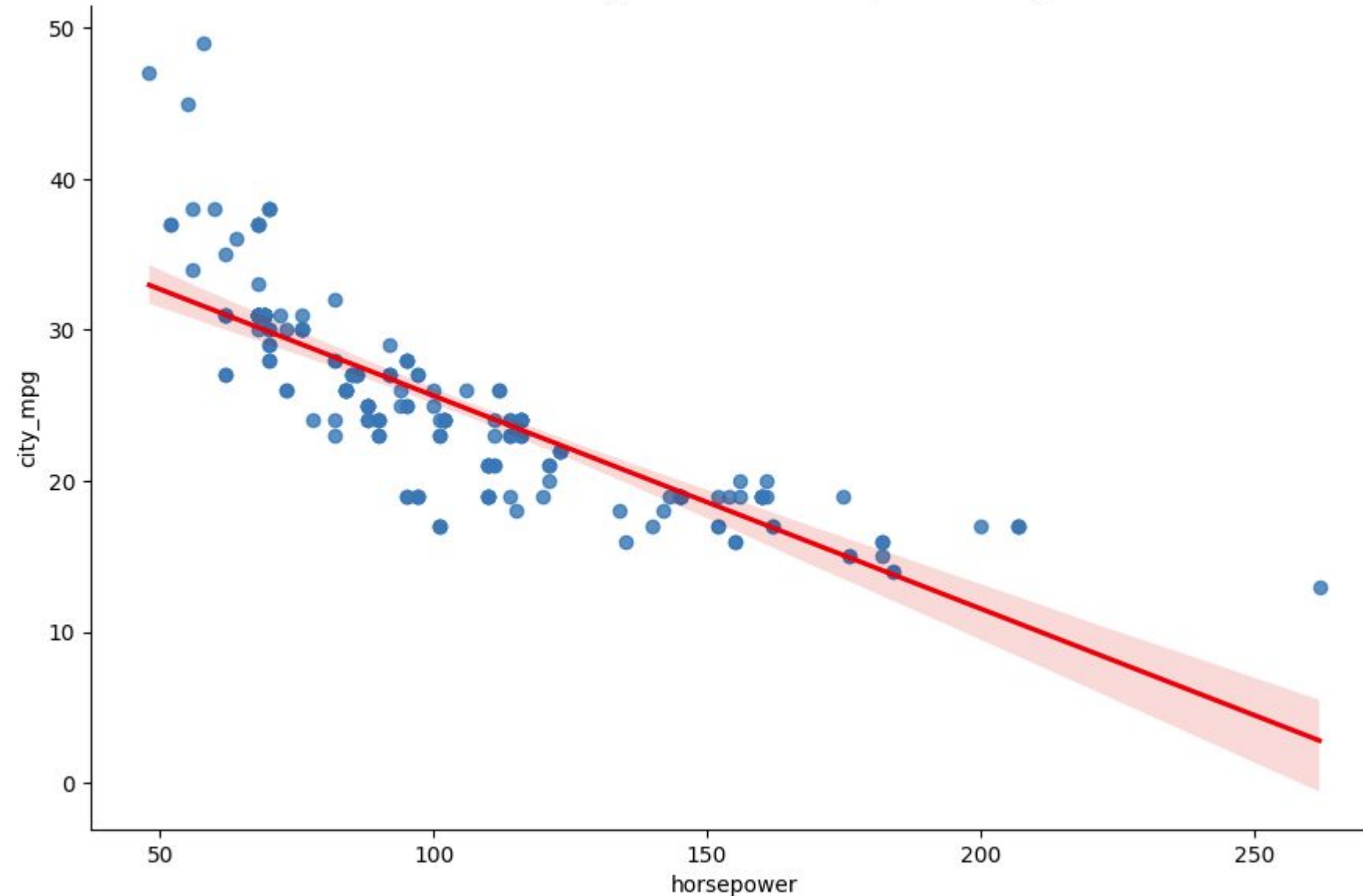


## Bottom Plot: Percentage Distribution of Aspiration Type

- A large proportion of vehicles have **standard (std)** aspiration systems (~80%), with a smaller segment using **turbo** (~20%).
- Businesses catering to automotive needs should prioritize offerings for standard aspiration systems, given their prevalence.
- The turbo segment, while smaller, may represent a higher-end or performance-oriented market, offering opportunities for specialized, premium products.

# Bi-variate - Numerical-Numerical

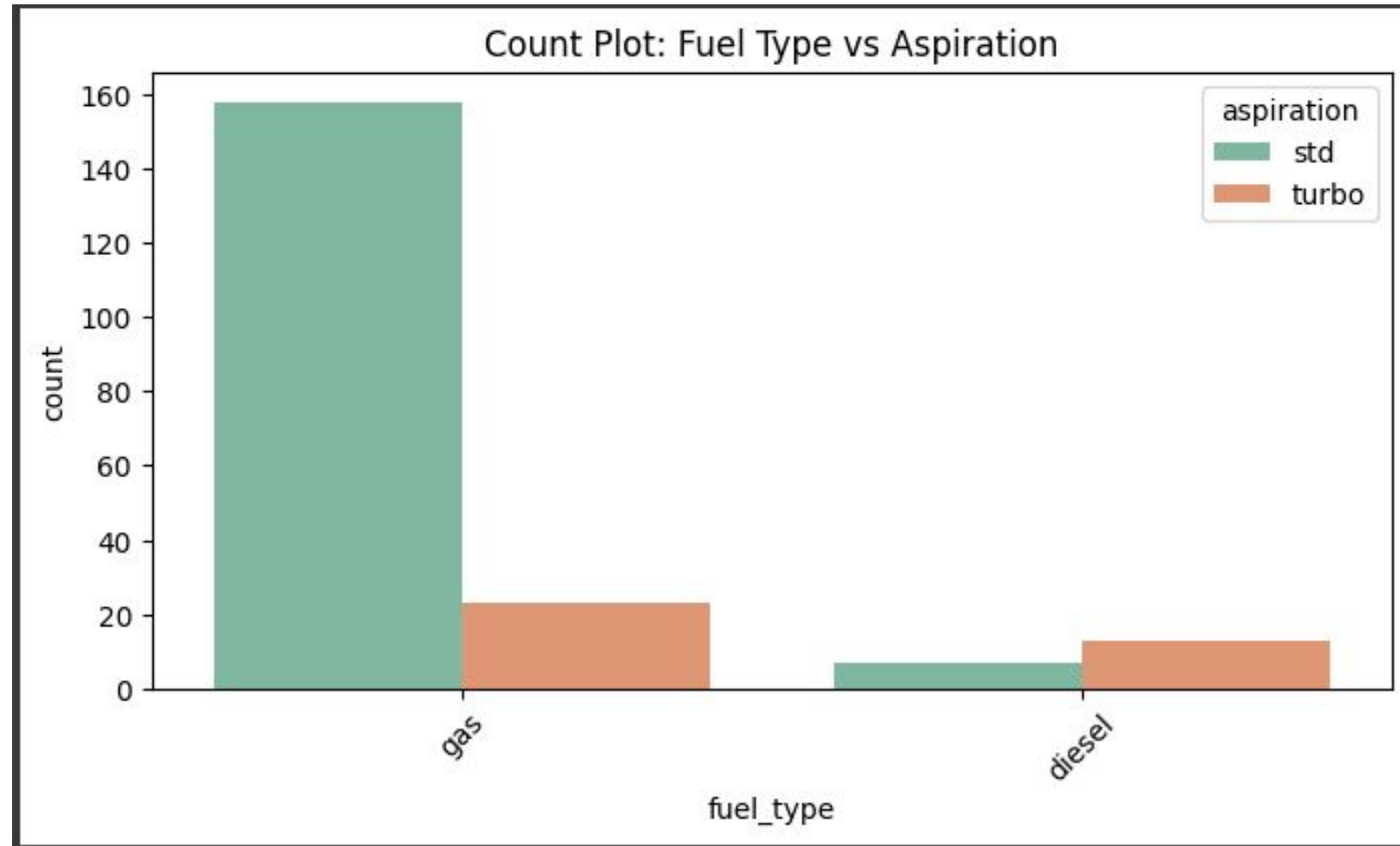
Scatter Plot with Regression Line: Horsepower vs City MPG



## Scatter Plot: Horsepower vs City MPG with Regression Line

- There is a clear negative relationship between **horsepower** and **city MPG**—as horsepower increases, city fuel efficiency decreases.
- This trend suggests that vehicles with higher horsepower are less fuel-efficient in city driving, aligning with the common trade-off between performance and fuel economy.
- Businesses targeting fuel-efficient markets should focus on lower-horsepower vehicles, while performance-focused brands might emphasize features other than fuel efficiency to appeal to consumers.

# Bi-variate - Categorical-Categorical with Hue



## Count Plot: Fuel Type vs Aspiration

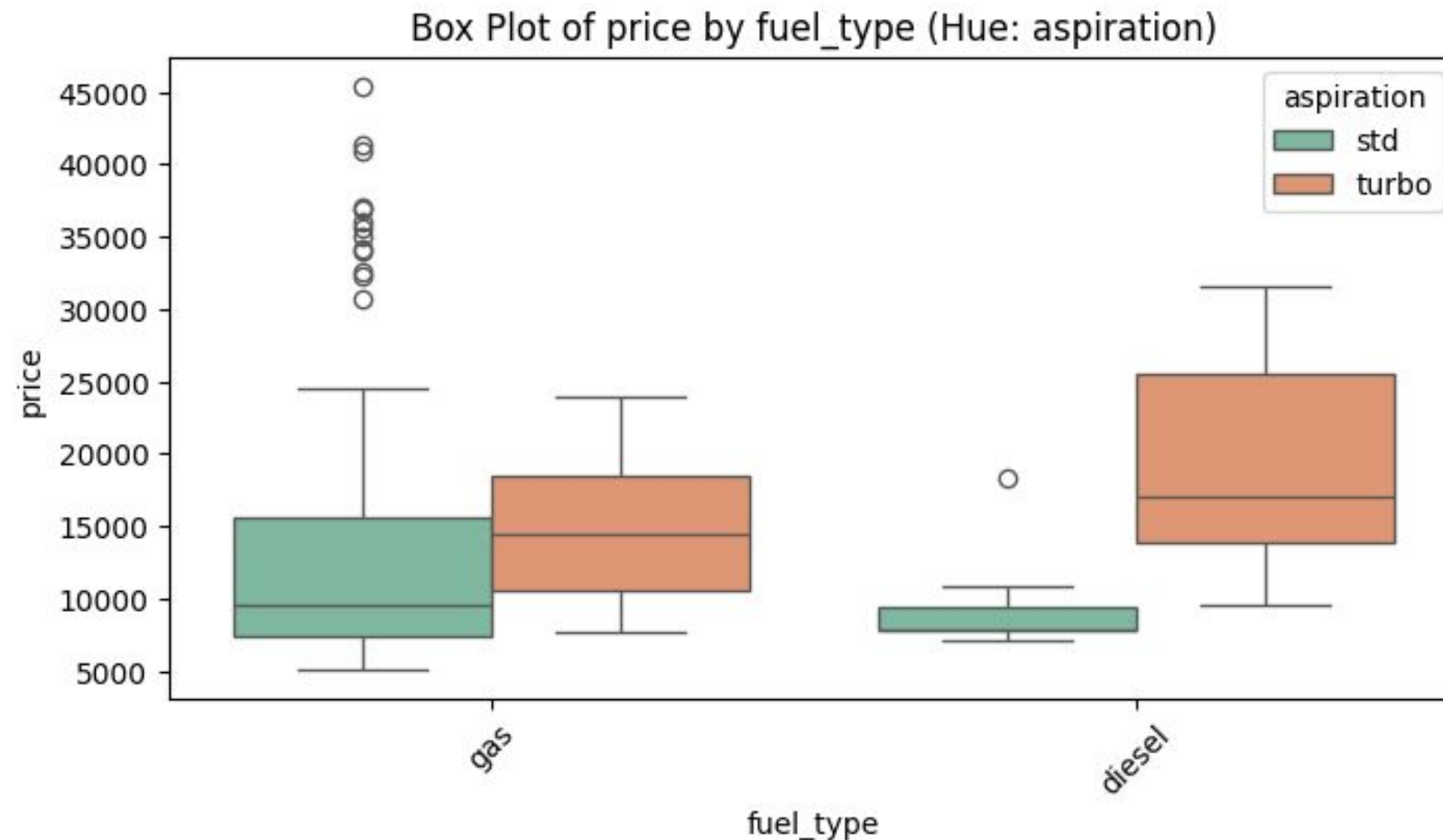
- The majority of vehicles run on **gas**, with most of them using **standard (std)** aspiration, while a smaller portion uses **turbo** aspiration.
- **Diesel** vehicles are rare, but there is a more balanced split between standard and turbo aspiration systems among them.
- For businesses, the focus should be on gas-powered vehicles with standard aspiration, as they dominate the market, while the turbo segment presents a smaller, performance-focused niche.

## What is the Hue?

- The **hue** in this plot represents the **aspiration** type, distinguishing between **standard (std)** and **turbo** aspiration for both gas and diesel vehicles. This helps break down the distribution further, showing not only fuel type but also how aspiration differs within each fuel category.



# Bi-variate - Categorical-Categorical with Hue

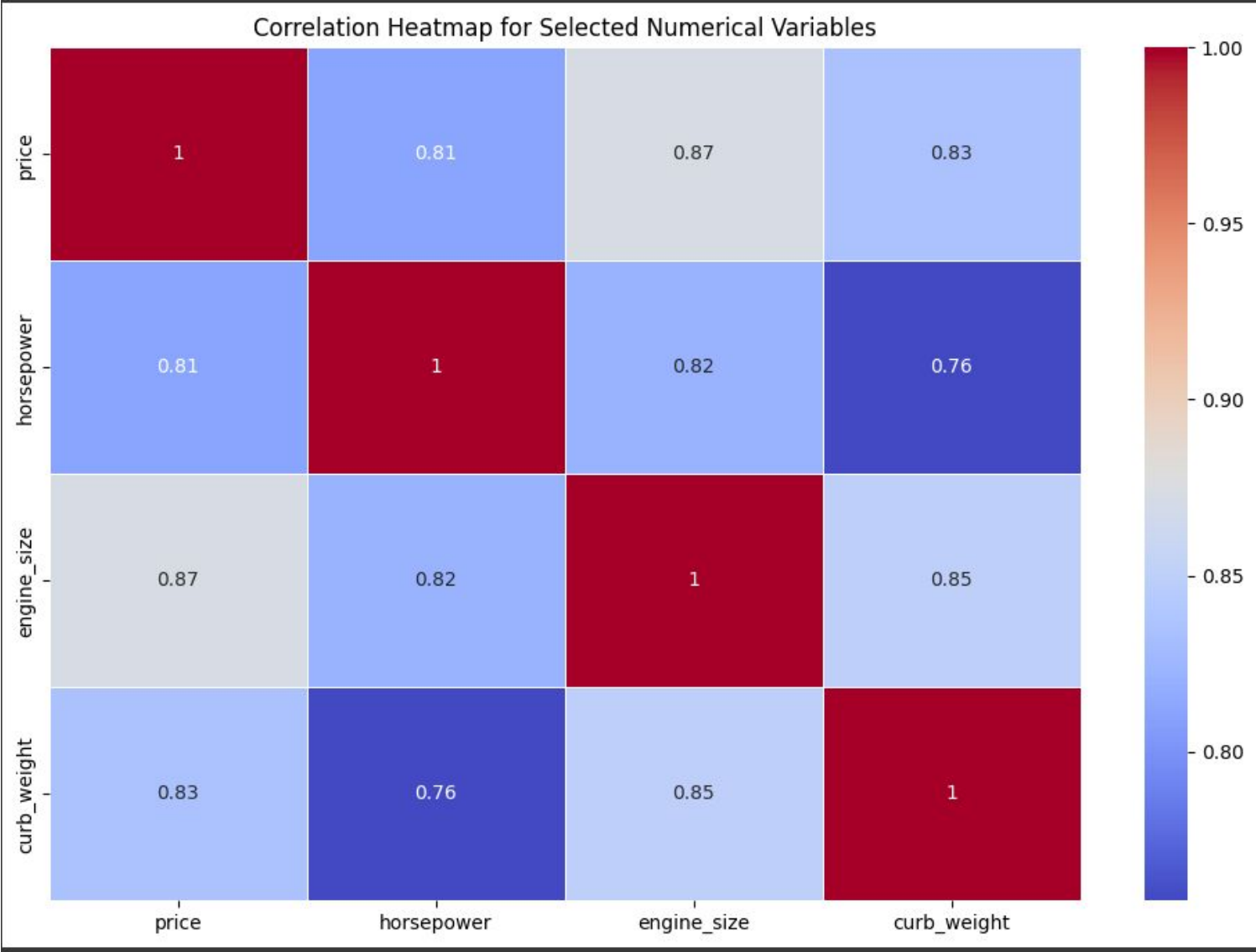


## Box Plot: Price by Fuel Type with Aspiration as Hue

- **Gas-powered vehicles** show a wider range of prices, with **turbo** models generally having higher prices than **standard (std)** ones. However, there are several outliers in the gas vehicles with standard aspiration that reach higher price levels.
- **Diesel vehicles** are priced higher on average, especially those with **turbo** aspiration, which tend to have a wider price distribution compared to diesel vehicles with standard aspiration.
- For businesses, this suggests that **turbo** vehicles, regardless of fuel type, tend to be priced higher, indicating that turbo models may target more premium market segments, while **standard** aspiration vehicles are generally more cost-effective, especially for gas. Diesel vehicles appear to cater to a higher-end niche overall.



# Multivariate - Numerical (Heatmap) - ( selected variables)



## Correlation Heatmap for Selected Numerical Variables

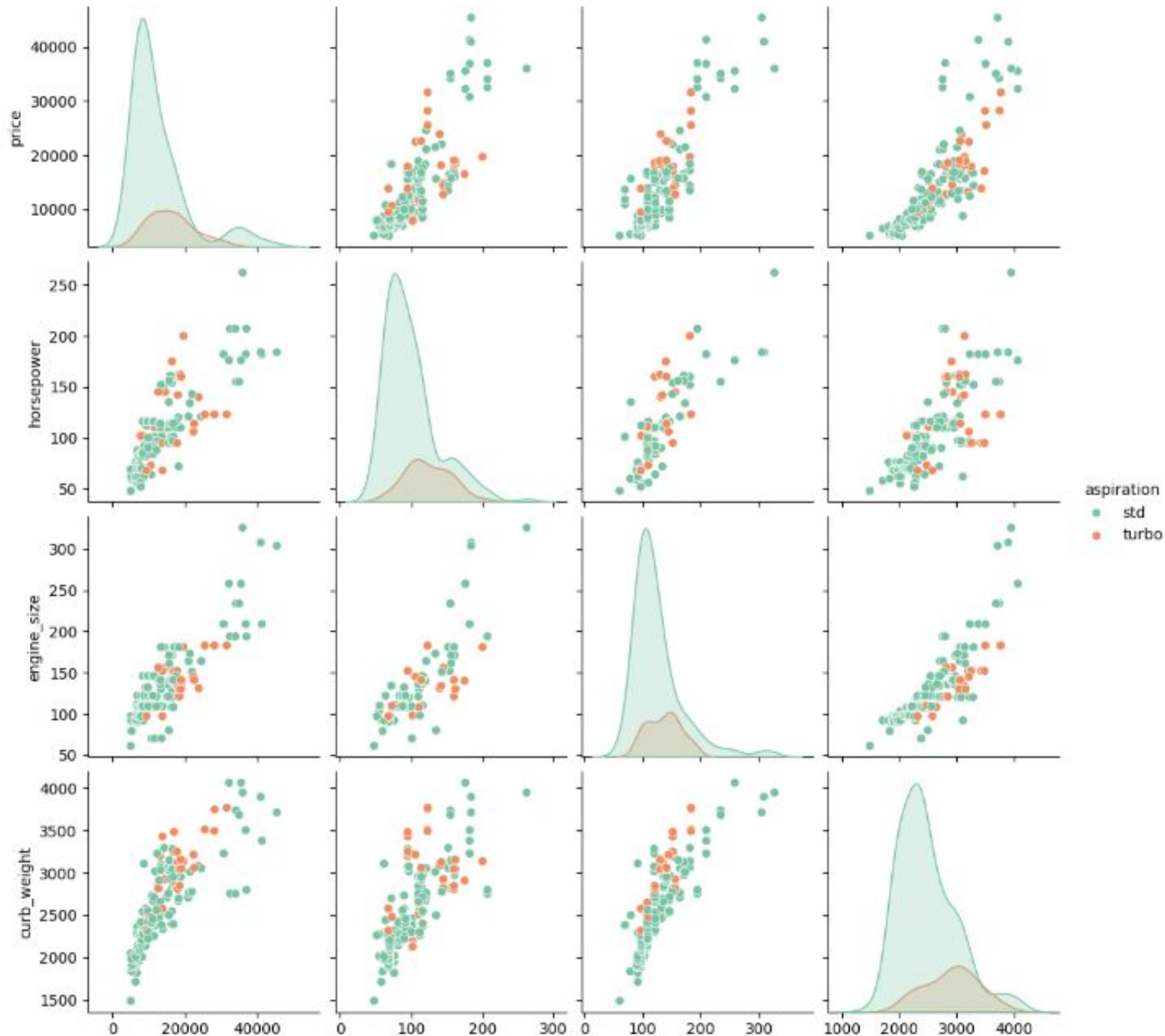
- Price** has a strong positive correlation with:
  - Engine size (0.87)**: Larger engine size is strongly linked to higher vehicle prices.
  - Horsepower (0.81)**: Vehicles with more horsepower tend to have higher prices.
  - Curb weight (0.83)**: Heavier vehicles are generally more expensive, possibly due to larger engines and additional features.
- Horsepower** shows a strong positive correlation with:
  - Engine size (0.82)**: Higher horsepower is associated with larger engine sizes.
  - Curb weight (0.76)**: Heavier vehicles tend to have higher horsepower, which might be due to the need for more powerful engines to move the extra weight.
- Engine size** correlates highly with:
  - Curb weight (0.85)**: Larger engines are found in heavier vehicles, indicating these may be larger or more robust models.

## Summary:

- Vehicles with larger engines, more horsepower, and greater curb weight tend to have higher prices.
- There is a strong relationship between engine size, horsepower, and curb weight, which likely indicates that larger, more powerful vehicles also tend to be heavier and costlier. This can inform businesses focused on premium or high-performance vehicle segments.

# Multivariate - Numerical (Pair plot ) - Selected variables with Hue

Pair Plot for Selected Variables with Hue = Aspiration



## Pair Plot for Selected Variables with Hue = Aspiration

### 1. Price vs Other Variables:

- **Price** has a clear positive relationship with **horsepower**, **engine size**, and **curb weight**, meaning that vehicles with higher horsepower, larger engines, or greater weight tend to be more expensive.
- **Turbo-aspiration vehicles** (orange dots) generally cluster at higher price levels, especially at the higher end of engine size, horsepower, and curb weight, indicating they are often higher-end models.

### 2. Horsepower vs Other Variables:

- **Horsepower** correlates positively with **engine size** and **curb weight**, with turbocharged vehicles having higher horsepower values, indicating a trend where turbo engines are linked to more powerful vehicles.
- The **turbo** (orange) vehicles have higher horsepower relative to **standard (std)** vehicles (green), even within the same weight or engine size range.

### 3. Engine Size and Curb Weight:

- **Curb weight** and **engine size** show a strong correlation, with larger engines typically found in heavier vehicles. Turbocharged vehicles often have larger engines and, as a result, higher curb weight.
- Standard aspiration vehicles are more evenly distributed, while turbocharged vehicles tend to cluster towards higher engine sizes and curb weights.