**Graphs**

# 1. How many variables are involved in answering/plotting?

# 2. What are their datatypes?

# 1. One variable - Univariate Data Viz.

# 2. Two variable - Bi-variate Data Viz. (Multi-variate)

# 3. 3 or more var – Multi-variate Data Viz.

# Case1: Univariate

# Continuous -

# Categorical - Bar chart, Pie chart

# Case2: Bi-variate

# Numerical-Numerical

# Numerical-Categorical

# Categorical-Categorical

# Case3: Multi-variate

# Numerical-Numerical-Categorical

# Categorical-Categorical-Numerical

# Categorical-Categorical-Categorical

# Numerical-Numerical-Numerical

**Univariate:**

**Categorical -** Bar graph or Pie chart

**import seaborn as sns**

sns.countplot(data=data, x="Genre", color= "orange", order = data["Genre"].value\_counts().index)

plt.xticks(rotation=90)

plt.title("No. of games per Genre")

plt.show()

**Continuous –** Histogram or Kernel Density Estimation(KDE) or Boxplot

**import seaborn as sns**

sns.histplot(data=data, x="Year", bins=15) # Histogram

sns.kdeplot(data=data, x="Year") # Kernel Density Estimation(KDE)

sns.boxplot(data=data, y="Year") # Boxplot

sns.violinplot(data=data, y="Year") # it’s a mix of kdeplot and boxplot

**Bi variate:**

**Continuous, Continuous –** line plot or scatter plot

1. Relationship between 2 features.
2. Trends of both features.
3. +ve or -ve corelation between both features.

**import seaborn as sns**

sns.scatterplot(data = ih, x="Year", y="NA\_Sales") # not the best way | used to find co-relation

sns.lineplot(data = ih, x="Year", y="NA\_Sales", color="red")

plt.title('Ice Hockey Sales Trend')

plt.xlim(left=1990, right=2010)

plt.show()

sns.lineplot(data = ih, x="Year", y="NA\_Sales")

sns.lineplot(data = bb, x="Year", y="NA\_Sales")

plt.title('Ice Hockey Sales Trend')

plt.legend(["Ice Hockey", "Baseball"], loc=(1.1, 0.5))

plt.grid()

plt.text(2002, 0.9, "Maximum IH Sales")

plt.show()

**Continuous, Categorical –** KDE plot, box plot, bar plot

Categorical always in x & continuous in y

sns.kdeplot(data = top3\_data, x="Global\_Sales", hue="Publisher") # Kernel Density Estimation(KDE)

sns.boxplot(data = top3\_data, x="Publisher", y="Global\_Sales") # Boxplot

# Top-3 Publisher and their Average Global Sales

sns.barplot(data = top3\_data, x="Publisher", y="Global\_Sales", estimator=np.mean) # bar plot

**Categorical, Categorical -** count plot

Find distribution of one feature in other

sns.countplot(data = top3\_data, x="Publisher", hue="Platform") # Dodget Bar Chart

**Multi-variate:**

**Numerical-Numerical-Categorical –– scatter plot**

**scatter plot –– X- numerical1 Y- numerical2 | hue – categorical feature**

**Categorical-Categorical-Numerical –– count plot, box plot**

**count plot –– X= Category1, hue= Category2 , Y = numerical feature**

**box plot –– X= Category1, Y = numerical feature, hue= Category2**

**Categorical-Categorical-Categorical  –– No Solution**

**Numerical-Numerical-Numerical –– scatter plot**

**scatter plot –– X = numerical1 Y = numerical2 size = numerical3**

**Sub-plots**

import matplotlib.pyplot as plt

import numpy as np

plt.figure(figsize=(20,12)).suptitle("NA Sales vs regions",fontsize=20)

# Using a 2x3 subplot

plt.subplot(2, 3, 1)

sns.scatterplot(x='NA\_Sales', y='EU\_Sales', data=top3\_data)

plt.subplot(2, 3, 3)

sns.scatterplot(x='NA\_Sales', y='JP\_Sales', data=top3\_data, color='red')

plt.subplot(2, 3, 4)

sns.scatterplot(x='NA\_Sales', y='Other\_Sales', data=top3\_data, color='green')

plt.subplot(2, 3, 6)

sns.scatterplot(x='NA\_Sales', y='Global\_Sales', data=top3\_data, color='orange')

plt.subplot(1, 3, 2)

sns.countplot(x='Publisher', data=top3\_data)

plt.show()

**Probability & statistics**

**Heat Map**

sns.heatmap(data=df1.corr(), annot=True)

**Crosstab**

pd.crosstab(**index**=df1['catagorical\_data1'],  
            **columns**=df1["catagorical\_data2"],  
            **margins**=True)