Artificial Intelligence (Machine Learning & Deep Learning) [Course]

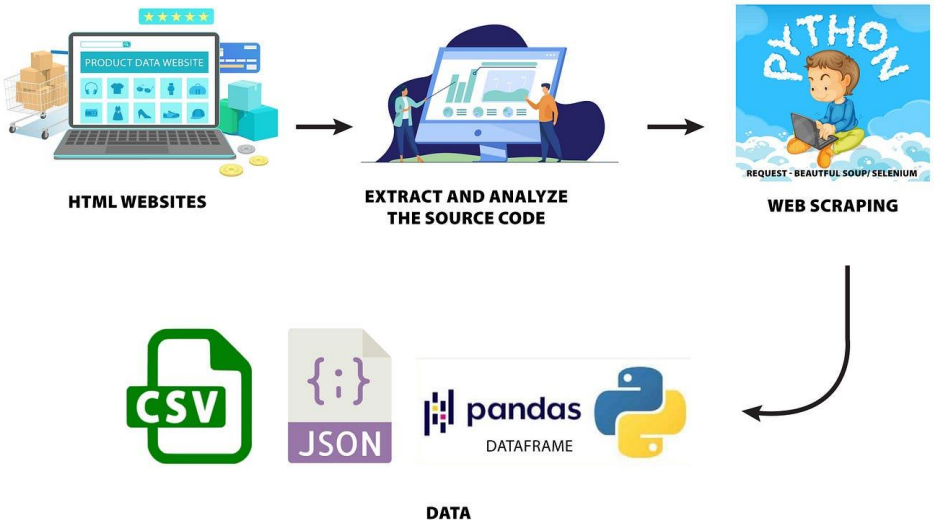
Week 3 – Web Scraping –Descriptive Statistics - SeaBorn

[See examples / code in GitHub code repository]

It is not about Theory, it is 20% Theory and 80% Practical – Technical/Development/Programming [Mostly Python based]

Data Collection & Web Scraping

Selenium and BeautifulSoup are powerful tools for web scraping in Python. Selenium automates browser interactions, making it ideal for handling JavaScript-heavy websites, while BeautifulSoup parses HTML content for data extraction.



References for BeautifulSoup :

https://www.geeksforgeeks.org/python/implementing-web-scraping-python-beautiful-soup/
https://toxigon.com/web-scraping-with-python-using-selenium-and-beautifulsoup
https://www.freecodecamp.org/news/better-web-scraping-in-python-with-selenium-beautiful-soup-and-pandas-d6390592e251/

References for Selenium

https://builtin.com/articles/selenium-web-scraping
https://medium.com/@datajournal/web-scraping-with-selenium-955fbaae3421

CODE – 2 METHODS :

https://github.com/ShahzadSarwar10/FULLSTACK-WITH-AI-BOOTCAMP-B1-MonToFri-2.5Month-Explorer/tree/main/Week3

Theoretical Background

What Are Descriptive Statistics?

Descriptive statistics are brief informational coefficients that summarize a given dataset, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables.

References:

https://www.investopedia.com/terms/d/descriptive_statistics.asp
https://corporatefinanceinstitute.com/resources/data-science/descriptive-statistics/

Theoretical Background

Probability

Probability is simply how likely something is to happen.

Probability means possibility. It is a branch of mathematics that deals with the occurrence of a random event.

Example 1:
There are 6 pillows in a bed, 3 are red, 2 are yellow and 1 is blue. What is the probability of picking a yellow pillow?

Ans: The probability is equal to the number of yellow pillows in the bed divided by the total number of pillows, i.e. 2/6 = 1/3.

Example 2:Flipping a coin:

$$P(H) = \frac{1}{2} = 50\%$$

References:

https://www.khanacademy.org/math/statistics-probability/prob
library/basic-theoretical-probability/a/probability-the-basics
https://byjus.com/maths/probability/
https://www.cuemath.com/data/probability/

Theoretical Background

Data and its types (structured, Unstructured)

Properties	Structured data	Unstructured data
Format examples	• CSV • Excel	• audio files (WAV, MP3, OGG) • PDF documents • images (JPEG, PNG, etc.)
Sources examples	• online forms • point-of-sale (POS) systems • online transaction processing (OLTP) systems	• emails • social media posts • multimedia files • IoT outputs
Nature of data	Quantitative	Qualitative
Databases	Relational (SQL)	Non-relational (NoSQL)
Storage for analytics use	Warehouses and data lakehouses	Data lakes and data lakehouses
Specialists to handle data	Business analysts, software engineers, data analysts	Data scientists, data engineers, data analysts
Main benefits	Easy to search and analyze, doesn't require much space	Easy to collect and store
Main challenges	All data must fit predefined schema	Difficult to search and analyze

References: https://www.altexsoft.com/blog/structured-unstructured-da
https://www.ibm.com/think/topics/structured-vs-unstructur

Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.

Mean (Arithmetic)

The mean is equal to the sum of all the values in the data set divided by the number of values in the data set.

Median

The middle score for a set of data that has been arranged in order of magnitude.

Mode

The most frequent score in our data set.

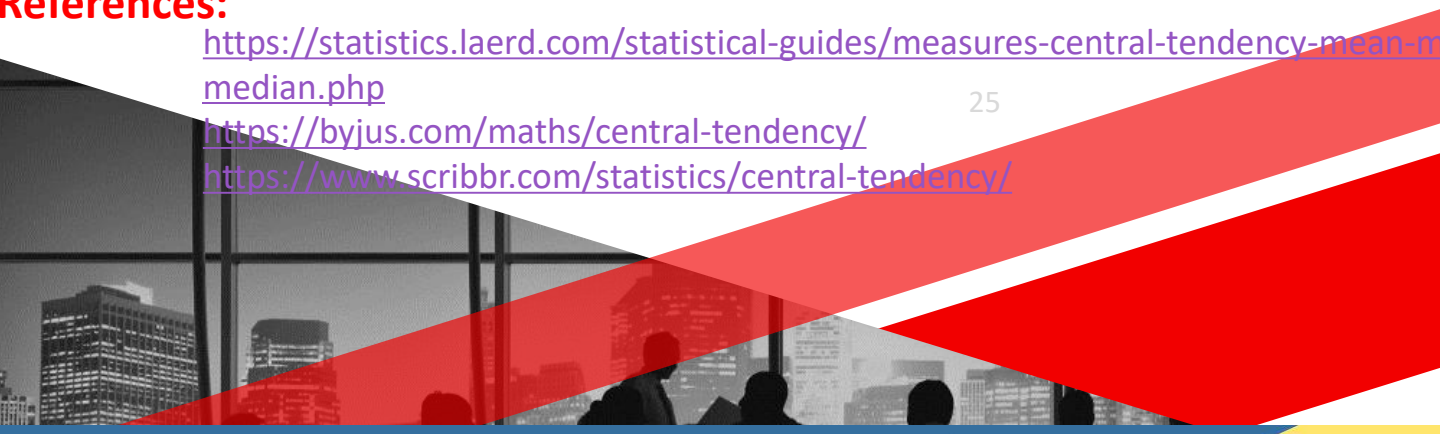
Example:

65 55 89 56 35 14 56 55 87 45 92

Mean:59 , Median: 56 , Mode: 56 , 55

References:

https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-m
median.php
https://byjus.com/maths/central-tendency/
https://www.scribbr.com/statistics/central-tendency/



Measures of Position

Measures of position give a range where a certain percentage of the data fall. The measures we consider here are percentiles and quartiles.

Variance

The average squared deviation from the mean of the given data set

$$\sum_1^n \frac{(X_i - \bar{X})^2}{n}$$

Standard Deviation

The square root of the variance gives the ”standard deviation”

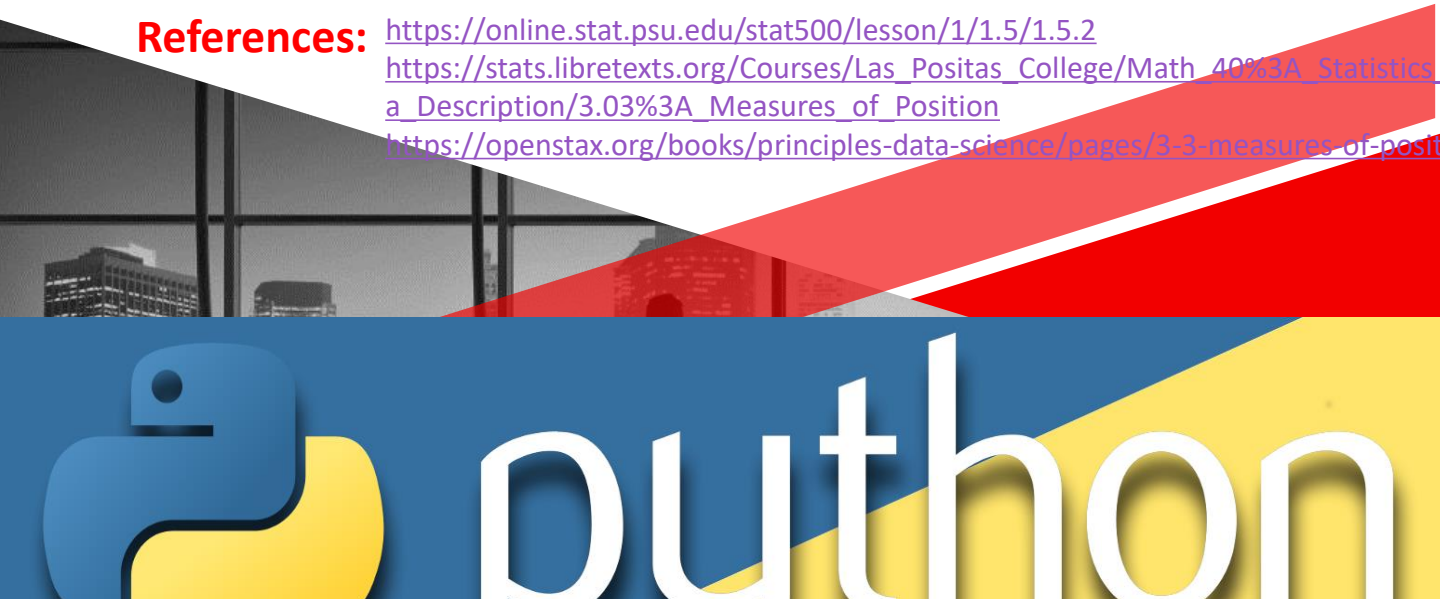
$$S.D. = \sqrt{Variance} = \sigma$$

Coefficient of Variation

The ratio of the standard deviation to the mean of the data set

$$(S.D. / Mean) * 100$$

References: https://online.stat.psu.edu/stat500/lesson/1/1.5/1.5.2
https://stats.libretexts.org/Courses/Las_Positas_College/Math_40%3A_Statistics_a_Description/3.03%3A_Measures_of_Position
https://openstax.org/books/principles-data-science/pages/3-3-measures-of-posit



Measures of Dispersion

Measures of dispersion are non-negative real numbers that help to gauge the spread of data about a central value.

Quartiles

Quartiles are numbers that separate the data into quarters.

first quartile is at position (n+1)/4, second quartile (i.e. the median) is at position 2(n+1)/4, and the third quartile is at position 3(n+1)/4.

Percentiles

Percentiles provide a way to assess and compare the distribution of values and the position of a specific data point in relation to the entire dataset by indicating the percentage of data points that fall below it.

$$\text{Percentile} = \frac{\text{number of data values below the measurement}}{\text{total number of data values}} \times 100\% = \frac{n}{N} \times 100\%$$

z-score

The z-score is a measure of the position of an entry in a dataset that makes use of the mean and standard deviation of the data.

$$z = \frac{x - \mu}{\sigma}$$

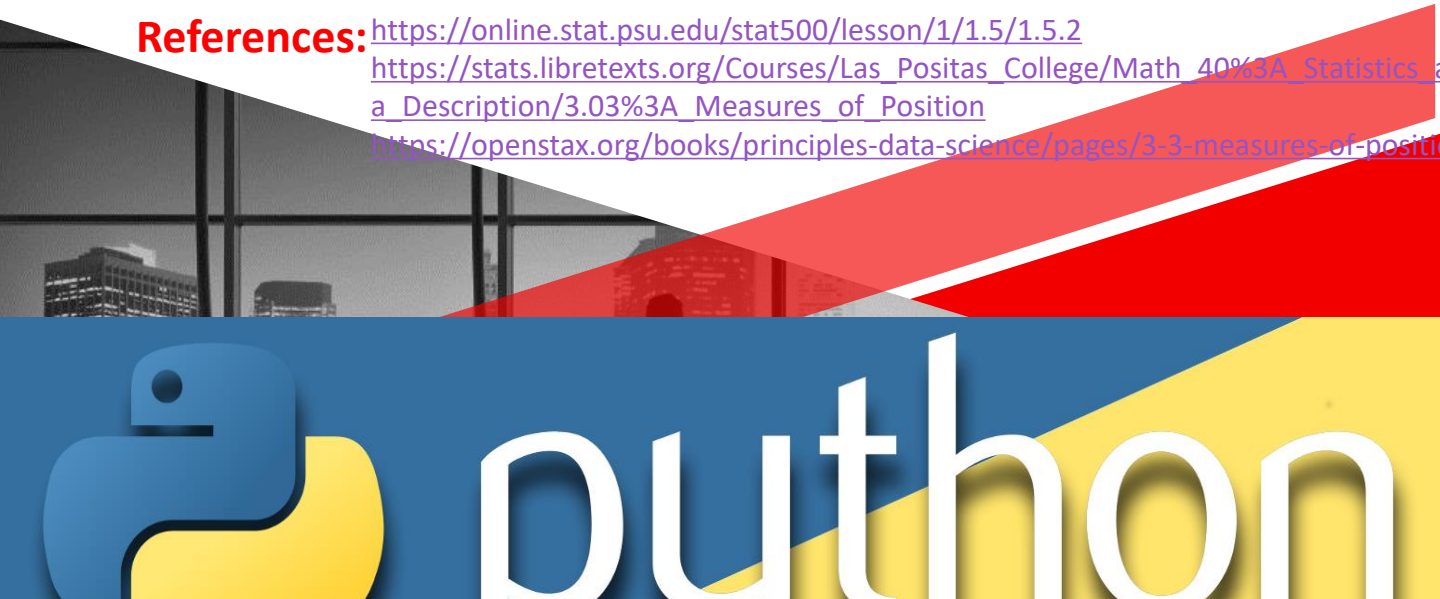
Where:

x is the measurement

μ is the mean

σ is the standard deviation

References: https://online.stat.psu.edu/stat500/lesson/1/1.5/1.5.2
https://stats.libretexts.org/Courses/Las_Positas_College/Math_40%3A_Statistics_a_Description/3.03%3A_Measures_of_Position
https://openstax.org/books/principles-data-science/pages/3-3-measures-of-positi



SeaBorn - Overview

seaborn

Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.

Python Seaborn Module

- **Data visualization** is considered as the best way to depict and analyze the data
- Python Seaborn module basically serves the purpose of Data Visualization at an ease with higher efficiency.
- It supports **NumPy** and **Pandas** data structure to represent the data sets.
- Seaborn stands out to have a better set of functions to carry out data visualization than **Matplotlib** in an optimized and efficient manner.

SeaBorn – Key Features

seaborn

Key Features of Seaborn

Aesthetic Themes

Provides appealing themes and palettes

Simplified Plot Creation

Reduces code complexity for creating plots

DataFrame Handling

Directly works with data frames for analysis

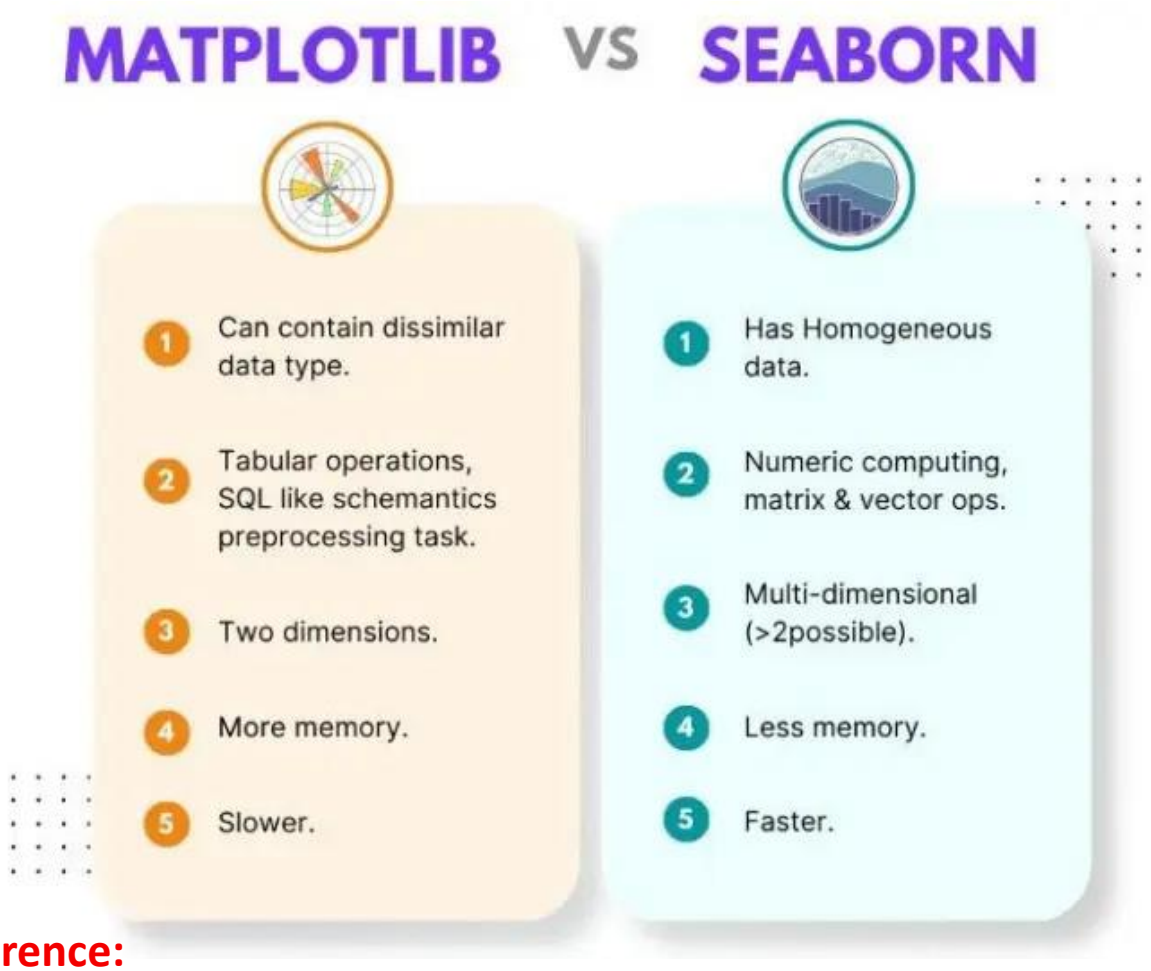
Statistical Visualization Focus

Emphasizes visualizing statistical data

python

MatPlotLib vs SeaBorn

seaborn

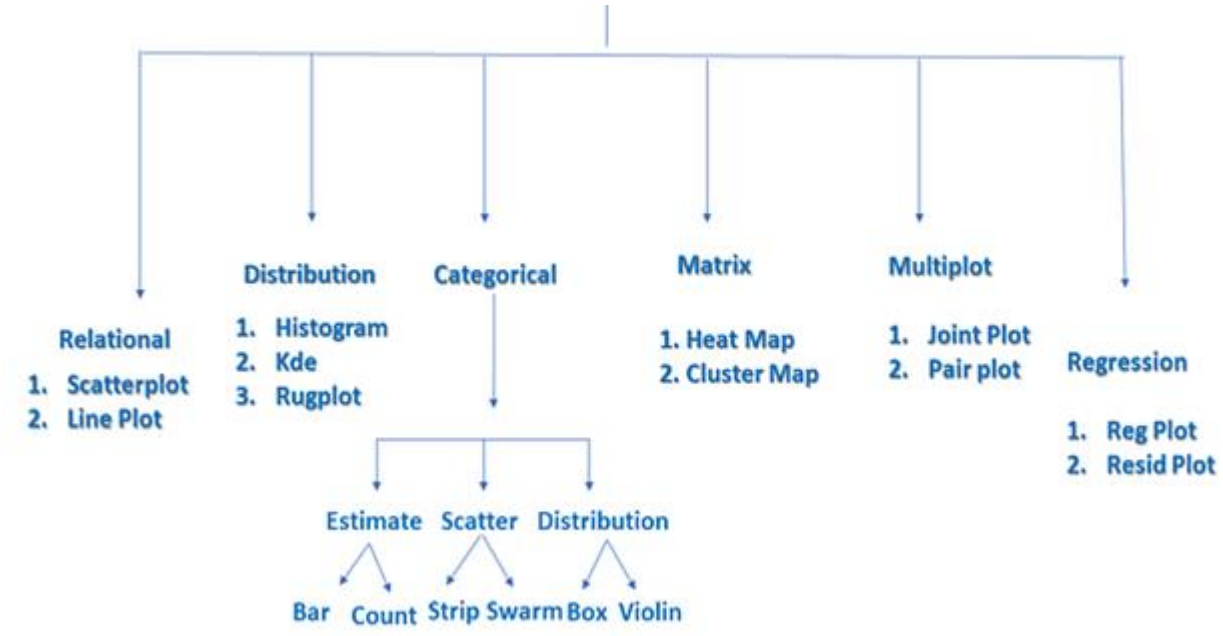


Reference:

https://docs.kanaries.net/topics/Seaborn/seaborn-vs-matplotlib

https://www.newhorizons.com/resources/blog/how-to-choose-between-seaborn-vs-matplotlib

SeaBorn - Plot Types



Reference:

https://medium.com/womenintechnology/complete-seaborn-tutorial-a5e184089c76
https://towardsdatascience.com/14-data-visualization-plots-of-seaborn-14a7bdd16cd7/
https://www.analyticsvidhya.com/blog/2021/12/12-data-plot-types-for-visualization/

Seaborn – cheat sheets

seaborn

3 Plotting With Seaborn

Axis Grids

```
>>> g = sns.FacetGrid(titanic, #Subplot grid for plotting conditional relationships
                      col="survived",
                      row="sex")
>>> g = g.map(plt.hist,"age")
>>> sns.factorplot(x="pclass", #Draw a categorical plot onto a Facetgrid
                   y="survived",
                   hue="sex",
                   data=titanic)
>>> sns.lmplot(x="sepal_width", #Plot data and regression model fits across a FacetGrid
               y="sepal_length",
               hue="species",
               data=iris)
>>> h = sns.PairGrid(iris) #Subplot grid for plotting pairwise relationships
>>> h = h.map(plt.scatter)
>>> sns.pairplot(iris) #Plot pairwise bivariate distributions
>>> i = sns.JointGrid(x="x", #Grid for bivariate plot with marginal univariate plots
                      y="y",
                      data=data)
>>> i = i.plot(sns.regplot,
               sns.distplot)
>>> sns.jointplot("sepal_length", #Plot bivariate distribution
                  "sepal_width",
                  data=iris,
                  kind='kde')
```

Regression Plots

```
>>> sns.regplot(x="sepal_width", #Plot data and a linear regression model fit
                y="sepal_length",
                data=iris,
                ax=ax)
```

Distribution Plots

```
>>> plot = sns.distplot(data.y, #Plot univariate distribution
                        kde=False,
                        color="b")
```

Matrix Plots

```
>>> sns.heatmap(uniform_data,vmin=0,vmax=1) #Heatmap
```

Categorical Plots

Scatterplot

```
>>> sns.stripplot(x="species", #Scatterplot with one categorical variable
                  y="petal_length",
                  data=iris)
>>> sns.swarmplot(x="species", #Categorical scatterplot with non-overlapping points
                  y="petal_length",
                  data=iris)
```

Bar Chart

```
>>> sns.barplot(x="sex", #Show point estimates & confidence intervals with scatterplot glyphs
                y="survived",
                hue="class",
                data=titanic)
```

Count Plot

```
>>> sns.countplot(x="deck", #Show count of observations
                  data=titanic,
                  palette="Greens_d")
```

Point Plot

4 Further Customizations

Also see Matplotlib

Axisgrid Objects

```
>>> g.despine(left=True) #Remove left spine
>>> g.set_ylabels("Survived") #Set the labels of the y-axis
>>> g.set_xticklabels(rotation=45) #Set the tick labels for x
>>> g.set_axis_labels("Survived", #Set the axis labels
                      "Sex")
>>> h.set(xlim=(0,5), #Set the limit and ticks of the x-and y-axis
          ylim=(0,5),
          xticks=[0,2.5,5],
```

Python Seaborn Cheat Sheet

https://www.datacamp.com/cheat-sheet/python-seaborn-cheat-sheet

https://cheatography.com/justin1209/cheat-sheets/seaborn/

https://book-of-gehn.github.io/articles/2021/06/05/Seaborn-Cheatsheet.html

python

SeaBorn - Exercises

seaborn

See code here: https://github.com/ShahzadSarwar10/FULLSTACK-WITH-AI-BOOTCAMP-B1-MonToFri-2.5Month-Explorer/blob/main/Week3/Case3-1-Seaborn-Zameencom-property-data-by-Kaggle.py

You should be able to analyze – each code statement, you should be able to see trace information – at each step of debugging. "DEBUGGING IS BEST STRATEGY TO LEARN A LANAGUAGE." So debug code files, line by line, analyze the values of variable – changing at each code statement. BEST STRATEGY TO LEARN DEEP.

Let's put best efforts.
Thanks.
Shahzad – Your AI – ML Instructor

Exercises

Covariance vs Correlation

Difference Between Correlation And Covariance	
Correlation	**Covariance**
1. Mathematical concept used to measure the relationship between two variables	1. Mathematical concept used to measure the variation between two variables
2. Shows the connection between the variables	2. Shows the variability between the variables
3. Its value lies between -1 and +1	3. Its value lies between – infinity and + infinity
4. Not influenced by the change in the scale	4. Easily influenced by the change in the scale
maindifferences.blogspot.com	

References:

https://www.coursera.org/in/articles/covariance-vs-correlation

https://builtin.com/data-science/covariance-vs-correlation

https://www.excelmojo.com/covariance-vs-correlation/

Outlier Detection, and Data Interpretation

References:

https://www.geeksforgeeks.org/data-science/detect-and-remove-the-outliers-using-python/

Thank you - for listening and participating

- ❑**Questions / Queries**
- ❑**Suggestions/Recommendation**
- ❑**Ideas…..?**

Shahzad Sarwar
Cognitive Convergence
https://cognitiveconvergence.com
shahzad@cognitiveconvergence.com
voice: +1 4242530744 (USA) +92-3004762901 (Pak)