**Data Visualization with Python**

**Week 1**

**Data Visualization :**

Data visualization is the graphical representation of data and information. It involves the process of creating visual representations of data. Data visualization can take many different forms, from basic charts and graphs to more complex interactive dashboards, maps, and infographics.

**Darkhorse Analytics** is a company that was created out of a research lab at the University of Alberta in 2008.

The company has done fascinating work on data visualization. Darkhorse Analytics specializes in quantitative consulting in several areas, including data visualization and geospatial analysis. Their approach when creating a visual revolves around the idea that **less is more effective**, more attractive, and more impactful

Source; <https://www.darkhorseanalytics.com/>

<https://www.tracktherecovery.org/>

**DATA BINING Example : How to Bin to Win**

<https://arize.com/blog-course/data-binning-production/>

<https://www.wallstreetmojo.com/data-binning/>

**Data Binning:**

* Data binning is a pre-processing method for data smoothing whereby the large set of original data is segregated into intervals called bins, and the discrete values in every bin are worked out to derive a representative value.
* It categorizes complex, continuous, and extensive data to decrease noise and reduce the effect of small observation errors on the analysis.
* Some widely used bucketing methods are equal-width binning, equal-frequency binning, and custom binning.
* Discretization may lead to information loss, over-smoothing, or under-smoothing of datasets, which can further result in misinterpretation and inaccurate outcomes.

Data binning is a way of pre-processing, summarizing, and analyzing data used to group continuous data into discrete bins or categories. It offers several benefits, such as simplifying [**data analysis**](https://www.wallstreetmojo.com/data-analysis/) and mitigating the impact of outliers in datasets. The process involves dividing the range of values into intervals and assigning each data point to an appropriate bin.

### ****Techniques****

Listed below are some prominent methods of data binning employed by analysts.

* **Equal-Width Binning**: This technique divides the data range into predetermined equal-width intervals or bins. The bin width can be computed by dividing the data range by the selected number of bins. While this method is simple and intuitive, it cannot be applied for skewed data distribution.
* **Equal-Frequency Binning**: In this method, the data is distributed into bins ensuring each bin has roughly the same number of data points. The data is first sorted, and then an equal number of data points is assigned to each bin. This approach is useful when it is essential to maintain similar frequencies or distributions across bins. This binning method can effectively tackle outliers and [**skewed**](https://www.wallstreetmojo.com/skewness/) data.
* **Entropy-Based Binning**: Under this type of discretization, continuous numerical values are categorized such that the clubbed variables represent the same class label. It analyzes the target class label and computes entropy, i.e., data impurities, and categorizes the split based on the level of information gain achievable.
* **Custom Binning**: This method allows users to set bin boundaries based on specific criteria or domain knowledge. Custom binning offers greater flexibility and control over data grouping. For example, bins can be created based on specific value ranges or required categories.
* **Quantile Binning**: A percentile binning technique applies to equal data distribution. It divides the data into bins based on percentiles. Thus, the number of bins is predetermined, and each bin comprises an equal number of data points. The bin boundaries are ascertained by the values at specific percentiles (e.g., 25th, 50th, and 75th percentiles).
* **Optimal Binning**: This bucketing technique aims to identify the most suitable set of bin boundaries based on specific optimization criteria. These methods employ statistical or machine learning algorithms to determine bin boundaries that minimize information loss or maximize desired objectives. For instance, it determines bin boundaries based on a [**decision tree**](https://www.wallstreetmojo.com/decision-tree/), [**chi-square**](https://www.wallstreetmojo.com/chi-square-test/), and Maximum Likelihood Estimation (MLE)

**What are the different types of binning?**

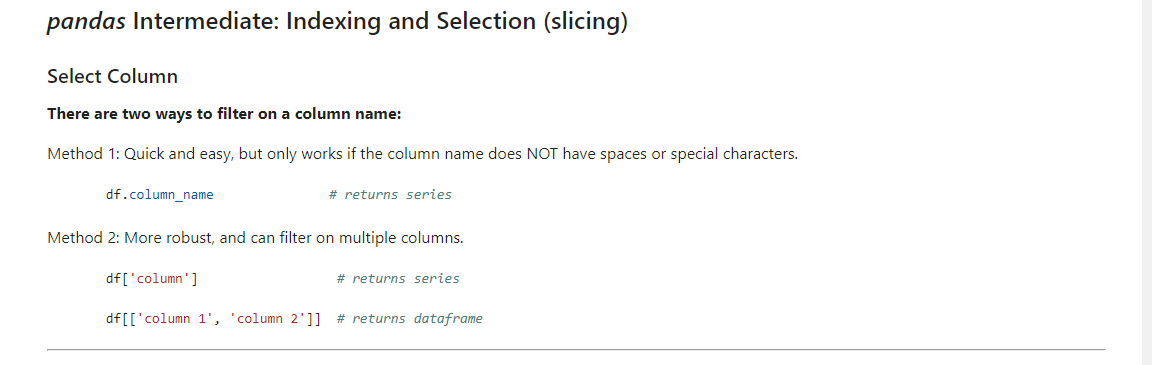
There are two types of data binning:  
● **Supervised binning**: Supervised bucketing uses the target class label to convert a numerical or continuous variable into a categorical value through the entropy-based binning technique.  
● **Unsupervised binning**: Unlike supervised binning, this bucketing does not depend on the target class label for categorizing continuous or numerical variables. It includes equal-width binning and equal-frequency binning.

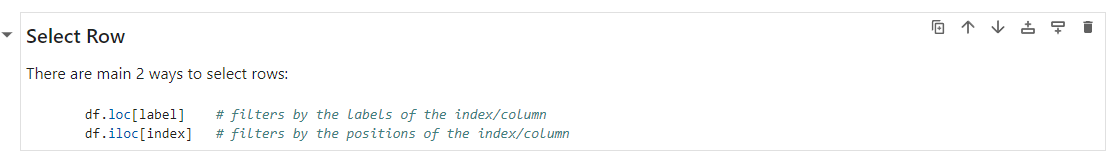
**Plotly:**

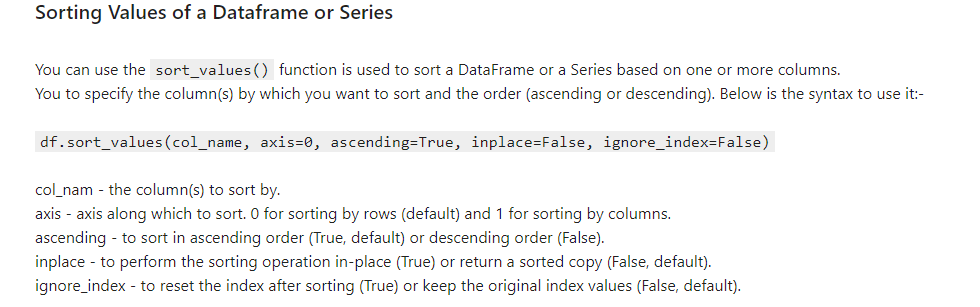
If you want to present your data in an interactive way, plotly is a great choice to explore. It offers highly interactive plots and dashboards. With Plotly, you can create line plots, scatter plots, bar charts, pie charts, 3D plots, and Choropleth maps, to name a few.

Its Plotly dash framework allows you to build interactive dashboards with rich visualizations and controls. Since Plotly is web based, it enables the rendering and viewing of plots in web browsers. This makes it convenient for sharing visualizations online, embedding them in web applications or dashboards, and collaborating with others.

**Pandas Stuff:**







**Matplotlib:**

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits.

One of the core aspects of Matplotlib is matplotlib.pyplot. It is Matplotlib's **scripting layer** which we studied in details in the videos about Matplotlib. Recall that it is a collection of command style functions that make Matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc. In this lab, we will work with the scripting layer to learn how to generate line plots. In future labs, we will get to work with the Artist layer as well to experiment first hand how it differs from the scripting layer.

**What is a line plot and why use it?**

A line chart or line plot is a type of plot which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields. Use line plot when you have a continuous data set. These are best suited for trend-based visualizations of data over a period of time.

A screenshot of a computer

Description automatically generated

# **Summary: Introduction to Data Visualization Tools**

Congratulations! You have completed this module. At this point in the course, you know:

* Data visualization is the process of presenting data in a visual format, such as charts, graphs, and maps, to help people understand and analyze data easily.
* Data visualization has diverse use cases, such as in business, science, healthcare, and finance.
* It is important to follow best practices, such as selecting appropriate visualizations for the data being presented, choosing colors and fonts that are easy to read and interpret, and minimizing clutter.
* There are various types of plots commonly used in data visualization.
* Line plots capture trends and changes over time, allowing us to see patterns and fluctuations.
* Bar plots compare categories or groups, providing a visual comparison of their values.
* Scatter plots explore relationships between variables, helping us identify correlations or trends.
* Box plots display the distribution of data, showcasing the median, quartiles, and outliers.
* Histograms illustrate the distribution of data within specific intervals, allowing us to understand its shape and concentration.
* Matplotlib is a plotting library that offers a wide range of plotting capabilities.
* Pandas is a plotting library that provides Integrated plotting functionalities for data analysis.
* Seaborn is a specialized library for statistical visualizations, offering attractive default aesthetics and color palettes.
* Folium is a Python library that allows you to create interactive and customizable maps.
* Plotly is an interactive and dynamic library for data visualization that supports a wide range of plot types and interactive features.
* PyWaffle enables you to visualize proportional representation using squares or rectangles.
* Matplotlib is one of the most widely used data visualization libraries in Python.
* Matplotlib was initially developed as an EEG/ECoG visualization tool.
* Matplotlib’s architecture is composed of three main layers: Backend layer, Artist layer, and the Scripting layer.
* The anatomy of a plot refers to the different components and elements that make up a visual representation of data.
* Matplotlib is a well-established data visualization library that can be integrated in different environments.
* Jupyter Notebook is an open-source web application that allows you to create and share documents.
* Matplotlib has a number of different backends available.
* You can easily include the label and title to your plot with plt.
* In order to start creating different types of plots of the data, you will need to import the data into a Pandas DataFrame.
* A line plot is a plot in the form of a series of data points connected by straight line segments.
* Line plot is one of the most basic type of chart and is common in many fields.
* You can generate a line plot by assigning "line" to 'Kind' parameter in the plot() function.