greatlearning
Learning for Life

Python - 2: Introduction to Visualization

Topics of the week

- Introduction to Visualization
- Matlplotlib, Seaborn and Plotly
- Hands-on visualization techniques
- Numpy, Pandas lab exercises
- Pandas lab exercises with visualizations
- Data visualization using Seaborn
- Case Study

Introduction to Visualization

Data visualization is an important skill in applied statistics and machine learning.

- It provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more.

- Visualisation is the most important aspect of exploratory data analysis (EDA)

Matplotlib, Seaborn and Plotly

greatlearning
Learning for Life

Matplotlib

- The matplotlib is a popular graphical subroutine and is used widely for data visualization applications.
- The matplotlib provides a context, one in which one or more plots can be drawn before the image is shown or saved to file. The context can be accessed via functions on *pyplot*.

There is some convention to import this context and name it plt; for example:

import matplotlib.pyplot as plt

Matplotlib, Seaborn and Plotly

Seaborn

Seaborn is complementary to Matplotlib and it specifically targets statistical data visualization. But it goes even further than that: Seaborn extends Matplotlib and that's why it can address the frustrations of working with Matplotlib.

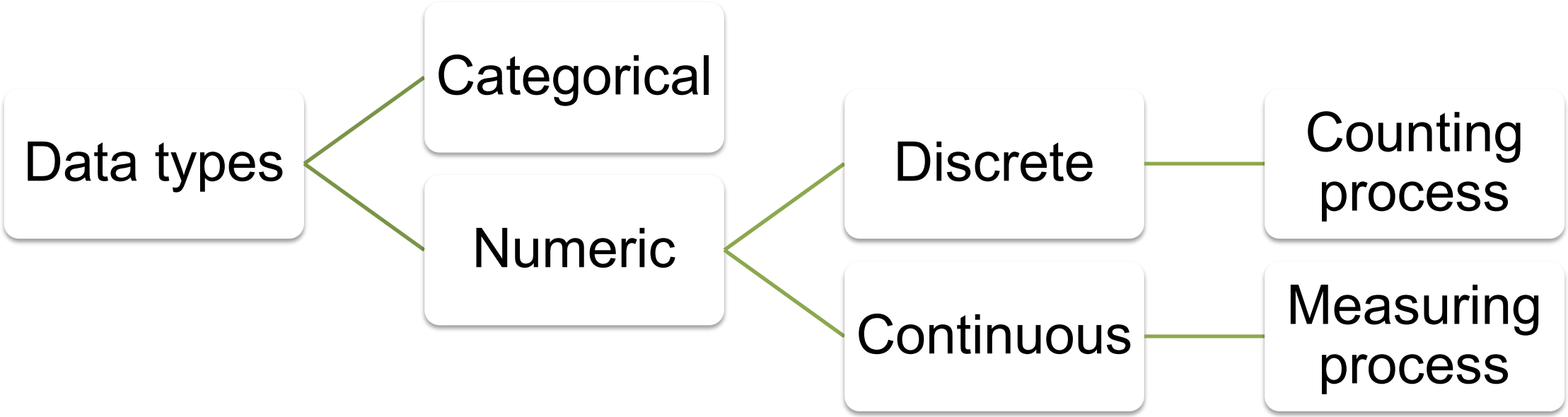
A saying around the matplotlib and the seaborn is, "matplotlib tries to make easy things easy and hard things possible, seaborn tries to make a well-defined set of hard things easy too."

Matplotlib, Seaborn and Plotly

Plotly

Plotly provides a web-service for hosting graphs. Plotly for Python can be configured to render locally inside Jupyter (IPython) notebooks, locally inside your web browser, or remotely in your online Plotly account.

Types of data



Different types of plots

- There are five key plots that you need to know well for basic data visualization. They are:
  - Line Plot
  - Bar Chart
  - Histogram Plot
  - Box and Whisker Plot
  - Scatter Plot

Chart selection

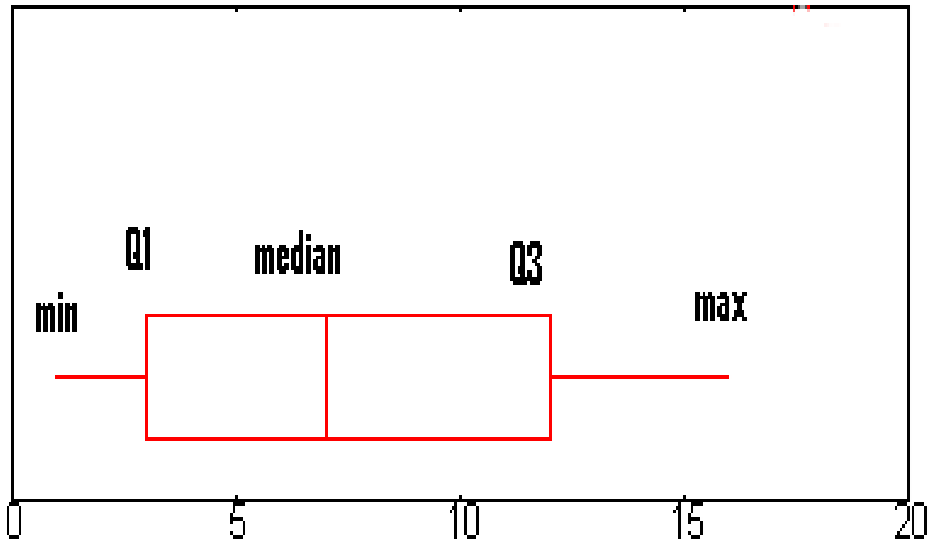
greatlearning
Learning for Life

X Variable	Y Variable	Purpose of analysis	Type of chart	Example
Continuous (numerical)	Continuous (numerical)	How Y changes with X	Scatter plot	How cholesterol varies with Age?
Continuous (numerical)	Categorical	How range of X varies for various category levels	Box plot	Cholesterol variation with Men and Women
Categorical	Categorical	What is the number or % of records of X which falls under each category	Stacked bar	How many men have heart disease compared to women?
Continuous	-	Look at the distribution of the values of the X variable	Histogram, boxplot	Distribution of cholesterol ranges
Impact of 2 X variables on a Y variable			Facet_grid()	Distribution of chol across mean and women – compared for people who have and don't have heart disease

Practical use cases of various visualizati

Boxplot

- Comparison of incomes of customers who leave and stay with an organisation in a customer churn problem
- Comparison of years of experience of people who leave and stay in an organisation in an attrition dataset



A bar plot helps in understanding the distribution of the data at hand. It gives us an understanding of the skewness of the data and also provides five point summary of the data.

Practical use cases of various visualizati

Scatterplot

- Relationship between customer age and average call duration in a telecom customer churn dataset
- How sales of a product varies with total minutes of advertisement aired
- How interest revenue of a customer varies with his annual income in a banks customer dataset
- Do cholesterol levels increase / decrease with a person's blood sugar value?

Practical use cases of various visualizati

greatlearning
Learning for Life

Stacked barplot

- Comparison of how '% attrition' varies between male and female.
- Comparison of how customer churn varies between 3 different customer plans.

Different Plots - Syntax

Different types of plotting functions between categorical and continuous variables:

1. sns.stripplot(auto['fuel_type'], auto['horsepower'])
2. sns.boxplot(auto['number_of_doors'], auto['horsepower'])
3. sns.barplot(auto['body_style'], auto['horsepower'])
4. sns.countplot(auto['body_style'])
5. sns.pointplot(auto['fuel_system'],auto['horsepower'])

Factor plot between multiple categorical variables:

sns.factorplot(x="fuel_type", y="horsepower", col="engine_location", data= auto, kind="swarm")

Case Study

Honey production data set visualization-

This dataset provides insight into honey production supply and demand in America by state from 1998 to 2012.

Dataset -

The dataset contains numcol, yieldprod, totalprod, stocks , priceperlb, prodvalue, and other useful information likeCertain states are excluded every year (ex. CT) to avoid disclosing data for individual operations.

For Reference: https://www.kaggle.com/arthurpaulino/honey-production/data

Steps

- Import pandas, numpy, seaborn, matplotlib.pyplot packages
- Get the data
- Explore the data for non-null and extreme values and try to answer the questions.
- How many States are included in the dataset?
- Which are he States that are included in this dataset?
- Calculate the average production for each state across all years
- How many years data is provided in the dataset? And what is the starting and ending year?

greatlearning
Learning for Life

HappyLearning

Questions?