

**Visualisation**

**Analytics with**

**Python**

**Lecture 3 Week 6**



# Welcome to lecture 3 of week 6.

|  |
| --- |
| **Python ecosystem for Data Science**   * Most popular Python Libraries * Python Analytical libraries: * NumPy * SciPy * Pandas * Python Visualisation libraries: * Matplotlib * *Seaborn* * Pygal, * Ploty, * … |
|  |

We continue our discussion about Python ecosystem for Data Science but with a focus on Visulisation

Python provides numerous methods for data visualization.

Various Python libraries, as we’ve mentioned, can be used for data visualization,

Python Pandas is the simplest method for basic plotting. Python Seaborn is great for creating visually appealing statistical charts that include color.

Python Bokeh works great for more complicated visualizations, especially for webbased interactive presentations.

Python Pygal works well for generating vector and interactive files.

However, it does not have the flexibility that other methods do.

Python Plotly is the most useful and easiest option for creating highly interactive web-based visualizations.

|  |
| --- |
| **Starting with Pandas Visulisation**  •  You have already seen reporting using Pandas:  •  Is any plot for data in a  dataframe  ?  plots only numerical data, ignores other attributes  Issues: no scaling!  phd  & service<< salary  the full syntax: |
|  |

You have already seen reporting using Pandas: what we discussed in the last video.

It is possible to plot numerical data in a dataframe

Pandas is a Python library with data frame features that supplies built in options for plotting visualizations in a two-dimensional tabular style.

|  |
| --- |
| **Visulisation with Pandas**  •  More plotting in Pandas: |
|  |

We can visualize more comparable units, as you can see on screen, such as the phd and services information, as shown in codesnip.

You can observe the correlation between phd and services over the years, except from age 55 up to 80, where services decline, which means that some people left the service at the age of 55 and older.

In the second codesnip you are grouping data by phd and summarizing the salaries per service category. Then you sort the derived data set in descending order according to the salaries. Finally, you plot the sorted data set using a bar chart.

|  |
| --- |
| **Visulisation with Matplotlib**  •  Matplotlib  is a 2D plotting library  •  High Quality and more control  •  Line Plot, scatter plot with log scale: |
|  |

Matplotlib is a Python 2D plotting library that produces high-quality figures in a variety of hard-copy formats and interactive environments across platforms. In Matplotlib, you can add features one by one, such as adding a title, labels, legends, and more.

In inline plotting, you should determine the x- and y-axes, and then you can add more features such as a title, a legend.

|  |
| --- |
| **Visulisation with Pandas/ Matplotlib**  •  Bar Plots  Pandas:  Matplotlib  : |
|  |

Bar charts are an essential visualization tool used to compare values in a variety of categories.

A bar chart can be vertically or horizontally oriented by adjusting the x- and y-axes, depending on what kind of information or categories the chart needs to present.

|  |
| --- |
| **Visulisation with Pandas/ Matplotlib**  • Histogram Plot  Pandas: Matplotlib: bin=10 bin=50 |
| STFC Visualisation 2014 University of Leeds |

A histogram can be used to represent a specific variable or set of variables. The Codesnip on screen plots 20 records of the salaries variables; it shows that salary packages of about 135,000 are the most frequent in this data set.

|  |
| --- |
| **Visulisation with Pandas/ Matplotlib**  •  Scatter Plot  Matplotlib  :    •  Pandas: |
| STFC Visualisation 2014 University of Leeds |

Scatter plot shows the relationship between two factors of an experiment (e.g. phd and service). A trend line is used to determine positive, negative, or no correlation.

|  |
| --- |
| **Visulisation with Pandas/ Matplotlib**  •  Pie Chart  •  Pandas:  Matplotlib  : |
|  |

Pie charts are useful for comparing parts of a whole.

They do not show changes over time. Bar graphs are used to compare different groups or to track changes over time.

However, when trying to measure change over time, bar graphs are best when the changes are larger.

In addition, a pie chart is useful for comparing small variables, but when it comes to a large number of variables, it falls short.

|  |
| --- |
| **Visulisation with Pandas/Seaborn**  •  Box Plot  Seaborn  :    •  Pandas: |
| STFC Visualisation 2014 University of Leeds |

Box plotting is used to compare variables using some statistical values.

The comparable variables should be of the same data units; the codesnip shows that when you compare phd and salary, it produces improper figures and does not provide real comparison information since the salary numerical units are much higher than the phd numerical values.

Plotting phd and services shows that the median and quantiles of phd are higher than the median and quantiles of the service information; in addition, the range of phd is wider than the range of service information.

The Python Seaborn library provides various plotting representations for visualizing data. A strip plot is a scatter plot where one of the variables is categorical. Strip plots can be combined with other plots to provide additional information. For example, a box plot with an overlaid strip plot is similar to a violin plot because some additional information about how the underlying data is distributed becomes visible.

Seaborn’s swarm plot is virtually identical to a strip plot except that it prevents data points from overlapping.

|  |
| --- |
| **Visulisation with Seaborn**   * Swarm Plot * Distribution of a variable against other variables. * It is to visualize different categories |
| STFC Visualisation 2014 University of Leeds |

A swarm plot is used to visualize different categories; it gives a clear picture of a variable distribution against other variables.

For instance, the salary distribution per gender and per profession indicates that the male professors have the highest salary range.

Most of the males are full professors, then associate, and then assistant professors. There are more male professors than female professors, but there are more female associate professors than male associate professors.

|  |
| --- |
| **Visulisation with Seaborn**  •  Joint Plot:  combines many plot to visualize the selected patterns |
| STFC Visualisation 2014 University of Leeds |

A joint plot combines more than one plot to visualize the selected patterns. The first plot combines a linear recreation with it band shown as cones with distribution of points in x and y axes

The last plot uses KDE or kernel density estimate, which is a method for visualizing the distribution of observations in a dataset, analagous to a histogram. KDE represents the data using a continuous probability density curve in one or more dimensions. This is similar to contours plot.

|  |
| --- |
| **Visulisation with Seaborn: Joint Plot** |
| STFC Visualisation 2014 University of Leeds |

Can you see any difference?

The first plot is normal scatter plot with KDE with 6 levels of contours plot.

The last plot is also a normal scatter plot with KDE with 6 levels of contours plot.

But it is also has the extra plot\_marginals(sns.rugplot) function, it can be used to add rugs on the side of any other kind of plot.

It shows marginal distributions uses a “rug” plot, which adds a small tick on the edge of the plot to represent each individual observation.

|  |
| --- |
| **Visual Analysis: An Example**  •  Regression  •  Discover if there’s a relationship  •  E.g.: Salary Vs. Rank  •  Useful in forecasting  •  Code for  salaries.cvs  •  Training set  •  Considered here  •  Testing set  •  For ML, not now  •  Visually can be solved |
|  |

Basically, Regression Analysis allows us to discover if there’s a relationship between an independent variable/s and a dependent variable (the target). For example, in a Simple Linear Regression we want to know if there’s a relationship between x and y.

This is very useful in forecasting (e.g. where is the trend going) and time series modelling (e.g. temperature levels by year and if global warming is true).