# Programming and Scripting Project (notes and references)

## Summary

* Sample summary: <http://rstudio-pubs-static.s3.amazonaws.com/450733_9a472ce9632f4ffbb2d6175aaaee5be6.html>
* Pandas basics: <https://www.geeksforgeeks.org/python-pandas-dataframe-info/>
* summary() function: <https://stackoverflow.com/questions/33889310/r-summary-equivalent-in-numpy>
* 5 number summary: <https://medium.com/@bramtunggala/5-number-summary-with-python-8aad81863b73>
* 5 number summary in python: <https://en.wikipedia.org/wiki/Five-number_summary>
* First and third quartiles: <https://stackoverflow.com/questions/45926230/how-to-calculate-1st-and-3rd-quartiles>
* Array basics: <https://www.geeksforgeeks.org/numpy-ndarray/>
* Table formatting: <https://stackoverflow.com/questions/9535954/printing-lists-as-tabular-data>
* Video on five number summary (bit of history): <https://www.youtube.com/watch?v=ZVm83jpabSw>
* Descriptive statistics realpython: <https://realpython.com/python-statistics/>
* Combine separate arrays into one 2d array: <https://cmdlinetips.com/2018/04/how-to-concatenate-arrays-in-numpy/>
* Assign each list to a variable: Week 5 additional exercises
* Importing mixed data: <https://campus.datacamp.com/courses/introduction-to-importing-data-in-python/introduction-and-flat-files-1?ex=13>
* Can’t access column when dtype=None. .shape gives different output to default dtype: <https://stackoverflow.com/questions/47733704/numpy-array-indexerror-too-many-indices-for-array>
* Better to import with pandas?: <https://stackoverflow.com/questions/35259650/numpy-structured-arrays-by-name-and-index>
* Structured array not working, mixed data types: <https://docs.scipy.org/doc/numpy/user/basics.rec.html>
* Pandas basic info: <https://www.shanelynn.ie/using-pandas-dataframe-creating-editing-viewing-data-in-python/>
* List unique values in df column: <https://chrisalbon.com/python/data_wrangling/pandas_list_unique_values_in_column/>
* String summary convert 1D array to dataframe: <https://stackoverflow.com/questions/11885503/numpy-transpose-not-giving-expected-result/11885962>
* Array to dataframe (string data): <https://stackoverflow.com/questions/20763012/creating-a-pandas-dataframe-from-a-numpy-array-how-do-i-specify-the-index-colum>
* ***Potential fixes: convert ‘count’ row values to integers (150.000)***
* Write df to text file: <https://stackoverflow.com/questions/51829923/write-a-pandas-dataframe-to-a-txt-file>
* Writing newlines to text file: <https://stackoverflow.com/questions/2918362/writing-string-to-a-file-on-a-new-line-every-time>

## For readme (summary)

* UCI mentions errors when compared to original dataset (see file downloaded)
* Finding correct headers for data (seem to be different in online examples e.g. <http://rstudio-pubs-static.s3.amazonaws.com/450733_9a472ce9632f4ffbb2d6175aaaee5be6.html> and <https://medium.com/@harimittapalli/exploratory-data-analysis-iris-dataset-9920ea439a3e> )
* I did find correct labels however in this example: <https://www.geeksforgeeks.org/box-plot-and-histogram-exploration-on-iris-data/>

## Histogram

* Basics: <https://realpython.com/python-histograms/>
* Saving histogram to png files: <https://stackoverflow.com/questions/46411533/how-can-i-save-histogram-plot-in-python>
* Output separate figures for each plot: <https://stackoverflow.com/questions/41978812/python-use-a-function-to-draw-multiple-plots>
* Y-axis values:
* Should I get a dataset that is more true?
* Histogram parameters: <https://matplotlib.org/3.2.1/api/_as_gen/matplotlib.pyplot.hist.html>

## Scatterplot

* Colour map: <https://matplotlib.org/tutorials/colors/colormaps.html>
* Solution to categorical variable colours (matplotlib): <https://stackoverflow.com/questions/14885895/color-by-column-values-in-matplotlib>
* Matplotlib explanation: <https://realpython.com/python-matplotlib-guide/>
* Matplotlib vs Seaborn vs Plotly: <https://towardsdatascience.com/matplotlib-vs-seaborn-vs-plotly-f2b79f5bddb>
* Pair plot for each pair of variables? Recommended by <https://blog.magrathealabs.com/choosing-one-of-many-python-visualization-tools-7eb36fa5855f>
* From data to vz. Great website: <https://python-graph-gallery.com/scatter-plot/>

## Beginning analysis

* Import + analyse + plot: <https://www.youtube.com/watch?v=pQv6zMlYJ0A>
* Source code for above: <http://apmonitor.com/che263/index.php/Main/PythonDataAnalysis>

## General stuff

* Getting the mean in numpy: <https://www.geeksforgeeks.org/numpy-mean-in-python/>
* Iris analysis overview: <https://www.kaggle.com/jchen2186/machine-learning-with-iris-dataset>
* Terminology for analysis/machine learning: <https://www.dummies.com/programming/big-data/data-science/how-to-prepare-data-for-predictive-analysis/>
* Terminology: Categorical variable, group, dependent variable, class, target variable, response

## Research README

* Used .genfromtxt instead of .loadtxt as loadtxt was not able to handle strings: <https://stackoverflow.com/questions/20245593/difference-between-numpy-genfromtxt-and-numpy-loadtxt-and-unpack>
* Add to git – LF > CRLF warning: <https://stackoverflow.com/questions/5834014/lf-will-be-replaced-by-crlf-in-git-what-is-that-and-is-it-important>
* Referencing in readme: <https://stackoverflow.com/questions/26587527/cite-a-paper-using-github-markdown-syntax>

## Example projects

* <https://www.kaggle.com/abhishekkrg/python-iris-data-visualization-and-explanation>
* GMIT 2018: <https://github.com/RitRa/Project2018-iris>
* Kaggle Indian dude: <https://www.kaggle.com/biphili/seaborn-matplotlib-plot-to-visualize-iris-data>

## Don’t forget!!

* Box plots for summary
* Justification for summary used (different functions in different libraries)
* Matplotlib as *verbose* relative to other visualisation tools
* Matplotlib as the fundamental tool for plotting, the basis of the others (quote article from above)

## Readme references:

* Fisher, R.A
* UCI website (classic in the field to this day)