20.2.23

Data Science Laboratory 1

Lab 5 Practical Challenges

1. Start simple and create a simple line plot which is a straight line starting at 0, ending at 8. ✓

2. Use `pd.read_csv()` to load the US presidential heights data set. Then, print the maximum height, minimum height, mean height and standard deviation. Next print the 25th percentile, median and 75th percentile. Finally, create a histogram of presidential heights. ✓

3. While we have seen how to use matplotlib, it is a fairly low level tool. Better options include using pandas (as we have seen) and seaborn. In particular, the latter is a statistical graphics library created by Michael Easkom (`https://seaborn.pydata.org/`) which simplifies many common visualisation types. Importing seaborn modifies the default matplotlib colour schemes and plot styles to improve the readability and aesthetics. Even if we do not use the seaborn API, it is sometimes preferable to import seaborn (we often use the alias `sb`) as a simple way to improve the visual aesthetics of general matplotlib plots.

   (a) Use DataFrame's `plot` method to plot five different lines (labelled A, B, C and D) on the same subplot (note the creation of of a legend automatically). The data should consist of 10 points and be randomly generated between 0 and 100. As an extra step, try cumulatively summing along the columns. ✓

   (b) Use `plot.bar` and `plot.barh()` to create vertical and horizontal bar plots. In particular, create a Series with 16 randomly generated values of float type from a uniform distribution over [0,1] (label them a, b, $c\ldots$) to use as the data. Ensure both are in the same figure, one on top of the other.

   (c) Use `pd.read_csv()` to load the restaurant tipping data set on canvas. Create a stacked bar plot showing the percentage of data points for each party size on each day.

   (d) Take the tipping data set again and use the seaborn package to create a bar plot which has the tipping percentage by day, along with error bars which represent the 95% confidence interval. The x-axis should have the tip percentage, while the y-axis should have the day.

   (e) By once again using the tipping data set, use seaborn to create a histogram of tip percentages of the total bill. On the same plot, add a density plot (using seaborn). A density plot is formed by computing an estimate of a continuous probability distribition that might have generated the observed data. The usual procedure is to approximate this distribution as a mixture of "kernels" (i.e. simpler distributions like the Normal distribution). Thus, density plots are also known as kernel density estimate (KDE) plots.

Remember: use GitHub to maintain your code!