# Applied Data Science: Weather visualization

## Task 2

### The statistics in different weather components

For task 2 of this lab, I would be attempting to plot the weather component data using pie charts and bar charts to present the first page of the infographic. The reason for choosing these charts to present the data is because both charts can summarise my justifications effectively and efficiently without too many details added. The charts are separated into both seasonal and monthly values, where the highest and the lowest mean value is chosen from the chart. I would not be presenting all monthly values on the pie chart, since it would be too confusing to read, and it doesn’t convey the summary as thorough as I want it to be. With the assist of figures, figure sizes and colour, I will present my 6 conclusions in the infographic about the outdoor temperature, rainfall, and humidity. The rest of the data, such as the indoor temperatures, will not be used, as the data are heavily influenced by the outdoor temperature, and it would make more sense to focus on the main factor that contributes to changes in temperature.

The colour theme for rainfall, outdoor temperature and humidity are grey, purple, and blue, so that they are visually separated. For rainfall, it is easier to mention percentages rather than the actual raw values. This is because the values in rainfall fluctuate significantly and there would be times that would rain a lot in a few days, but other times, it doesn’t for many days. This makes the average value significantly smaller than expected. Both the other 2 weather components would be described using bar charts, and to make the graphs that distinguish from each other, I decided to use a horizontal bar chart on outdoor temperature, and a vertical bar chart on humidity. The seasonal data are 3 months averages combined, and the monthly data are 30 days averages. The words and figures in this infographic are coloured to match the ones in the charts, hence it would be easier to follow.

### The timeline summary of the weather data

All the highest and the lowest recorded event would be highlighted. The information in the coloured boxes would display in the order of the category of the weather data, the highest/lowest value of the weather data and the time when the event occurred. Every node represents a month, and the colour on the timeline represents a season. There are 5 colours (seasons) on the timeline since the recording starts on the 9th of October 2016 and ended on the same date in 2017. This is because I wish to display the entire recording and describe some of the major events that happened at the start of the season (such as the highest recorded rainfall and the lowest outdoor temperature) and do not wish to mix these events with autumn 2017. All the events recorded are also colour-coded with the season colours, and the recorded data and date are both coloured separately so that it is easier to read. The timeline is split into a seasonal infographic and a monthly infographic.

### The NaN values in the dataset

There are NaN values spotted on rows 73, 218 and 284, in which both 73 and 218 display the NaN value in rainfall. For row 284, the humidity and all indoor temperature contain the NaN value. To deal with this problem, I decided to replace the NaN values with zeros. This is because I do not wish to create any bias by manipulating the data that has the potential to skew the ending results. Therefore, all the NaN values will be replaced with zeros instead of averages.