Data Science – Assignment #2: Data Exploration of Cyclist Dataset (65%)

**Problem:** The cycling shops association of the city of Cambridge would like you to analyze cyclist data from the years of 2016 and 2017, to assist them in their work schedule and equipment ordering. The dataset can be found at this website: <https://data.cambridgema.gov/dataset/Eco-Totem-Broadway-Bicycle-Count/q8v9-mcfg> . Analyze this dataset and determine if there are any particular:

1. Months in the year that are substantially busier than the rest. Create two plots showing the total cyclist per month for each year. Then draw conclusions from additional plots of the percentage difference for each month obtained by: %month diff. = 100\*(total cyclist in month – average total cyclist per month) / average total cyclist per month i.e. calculated per year.
2. Days in the week that are substantially busier than the rest during any particular quarter of the year. Create eight plots showing the total cyclist per day for the first week in: January, April, July, October, for each year. Then draw conclusions from additional plots of the percentage difference for each day obtained by: %day diff. = 100\*(total cyclist in day – average total cyclist per day) / average total cyclist per day i.e. calculated per month.
3. Time periods during a working day (Mon-Sat) that are substantially busier than the rest of the day. Create twelve plots, each showing the total cyclist per hour for each working day in the first week of April for each year. Draw conclusions from additional plots of the percentage difference for each hour obtained by: %hour diff. = 100\*(total cyclist in hour – average total cyclist per hour) / average total cyclist per hour i.e. calculated per week.

Submit this assignment as a pdf document with the following sections:

1. **Problem:** The problem as stated above.
2. **Methodology:** which includes three sub-sections outlining how you are going to solve each of the three problems outlined above. Each of these sub-sections should have: a) the major steps that has to be done to get the data, b) flow chart(s) indicating how the major steps would be performed. **(20%)**
3. **Results:** Plots of the data obtained for each task along with a discussion of any trends observed (use % differences of the total as a criterion). **(20%)**
4. **Conclusion:** answering the three questions given above. **(10%)**
5. **Appendix:** containing a copy of your code with comments corresponding to the major sections outlined in your methodology **(15%)**