Stat 437 HW1

Your Name (Your student ID)

General rule

Please show your work and submit your computer codes in order to get points. Providing correct answers without supporting details does not receive full credits. This HW covers:

- The basics of `dplyr`
- Creating scatter plot using `ggplot2`
- Elementary Visualizations (via ggplot2): density plot, histogram, boxplot, barplot, pie chart
- Advanced Visualizations via ggplot2: faceting, annotation

You DO NOT have to submit your HW answers using typesetting software. However, your answers must be legible for grading. Please upload your answers to the course space.

Problem 1

Please refer to the NYC flight data `nycflights13` that has been discussed in the lecture notes and whose manual can be found at https://cran.r-project.org/web/packages/nycflights13/index.html. We will use `flights`, a tibble from `nycflights13`.

You are interested in looking into the average `arr_delay` for 6 different `month` 12, 1, 2, 6, 7 and 8, for 3 different `carrier` "UA", "AA" and "DL", and for `distance` that are greater than 700 miles, since you suspect that colder months and longer distances may result in longer average arrival delays. Note that you need to extract observations from `flights` and obtain the needed sample means for `arr_delay`, and that you are requird to use `dplyr` for this purpose.

The following tasks and questions are based on the extracted observations.

(1.a) In a single plot, create a density plot for `arr_delay` for each of the 6 months with `color` aesthetic designated by `month`. Note that you need to convert `month` into a factor in order to create the plot. What can you say about the average `arr_delay` across the 6 months?

(1.b) In a single plot, create a boxplot for `arr_delay` for each of the 3 carriers. What can you say about the average `arr_delay` for the 3 carriers?

(1.c) Create a pie chart for the 3 carriers where the percentages are the proportions of observations for each carrier and where percentages are superimposed on the sectors of the pie chart disc.

(1.d) Plot `arr_delay` against `distance` with `facet_grid` designated by `month` and `carrier`.

(1.e) For each feasible combination of values of `month` and `carrier`, compute the sample average of `arr_delay` and save them into the variable `mean_arr_delay`, and compute the sample average of `distance` and save these averages into the variable `mean_distance`. Plot `month` against `mean_arr_delay` with `shape` designated by `carrier` and `color` by `mean_distance`, and plot `month` against `mean_arr_delay` with `shape` designated by `carrier` and `color` by `mean_distance` and annotate each point by its associated `carrier` name.

Problem 2

Please refer to the data set `mpg` that is available from the `ggplot2` package. Plot `displ` against `hwy` with faceting by `drv` and `cyl`, `color` disgnated by `class`, and `shape` by `trans`. This illustrates visualization with 4 factors.