Problem Set 2

The second problem set focuses on data transformation and visual exploration. Your solution should be composed of a well-structured R script which should provide the required analyses. Besides the functions the code should be directly runnable or at least sufficiently well documented (working directory, path settings) to be executed.

Furthermore, provide a documentation (in text format or powerpoint) where you document and illustrate your general approach, document the qualitative tasks and provide instantiations of your graphs. Provide some intuition of what your plots show and what you hypothesize.

Grading will reflect your performance on both the coding as well as the documentation and interpretation tasks – however, there is no fixed grading scheme between the two categories.

This problem set is **due on May 16th by 12.00** through the wuecampus group functionality.

Cooperation with other groups is not permitted and will lead to severe credit deductions.

1. Use the nycflights13 package and the flights and planes tables to answer the following questions:
    a. Which are the five oldest planes that flew from New York City airports in 2013? How far did they fly in comparison to all planes?
    b. How many airplanes (that flew from New York City) are included in the planes table? How many have missing date of manufacture?
    c. Display and interpret the distribution of the date of manufacture.
    d. Consider the following manufacturers: AIRBUS, AIRBUS INDUSTRIE, BOEING, BOMBARDIER INC, EMBRAER, MCDONNELL DOUGLAS, MCDONNELL DOUGLAS AIRCRAFT CO, MCDONNELL DOUGLAS CORPORATION (the most common manufacturers). Characterize and interpret the distribution of manufacturer. Has the distribution of each individual manufacturer changed over time as reflected by the airplanes flying from NYC in 2013? [Provide a plot and a table, Merge the two AIRBUS variants as discussed in the lecture, similarly merge the three MCDONNELL variants.]
    e. Using the same manufacturers as above, provide a graphical representation to display the arrival delays broken down by manufacturer.

2. Use the nycflights13 package and the weather table to answer the following questions:
    a. Plot the temperature distribution for the different months.
    b. Identify any important outliers in terms of the wind speed variable.
    c. What is the relationship between dewp and humid? [Provide a plot and comment.]
    d. What is the relationship between precip and visib? [Provide a plot and comment.]
    e. On how many days was there precipitation in the New York area in 2013?
    f. Provide a graphical representation of the relationships between precipitation, wind speed and arrival delay. Explain your insights.

3. In this exercise, you will apply your data wrangling and exploratory data analysis skills to perform sports analytics. To be precise we want to investigate whether or not the Oakland Athletics were as successful as the movie "Moneyball" suggested. To this end, you will use data from Sean Lahman's baseball database (http://www.seanlahman.com/baseball-archive/statistics/.) Luckily, the database is available as an R package [library(Lahman)]. After loading the package, you can access two data frames, salaries and teams. The former identifies players' salaries for each season and provides a team identifier. The teams table provides gameplay stats for each team over various seasons.
    a. Calculate the yearly payroll for all teams. Visualize the development in an appropriate manner.
    b. Visualize Oakland's performance by means of (a) scatterplot(s) displaying wins vs. payroll. Try to incorporate the time dimension as best as possible.
    c. Calculate for each team the number of wins per dollar for each year between 1990 and 2013. Visualize the data and include the league average.
    d. Using the league average wins per dollar you can approximate to what extent a team outperformed relative to its budget. Visualize your outperformance indicator for Oakland, the Yankees, the Red Sox, Blue Jays and the Kansas City Chiefs.
    e. Your results may be influence by the chosen indicators (wins and salary). Check the robustness of your analysis in d) by replacing wins by runs and total salary by median salary and rerunning the analysis.