## Problem Set 4

The third problem set focuses on data scraping and simple predictive modeling using linear regression. 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.

**Additionally, please save your scraped data and provide as a csv file.**

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 June 16th by 12.00** through the wuecampus submission functionality.

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

1. Clearly, you are taking data science to get a good job later on. Let's find out who the key employers in this field are. To this end, we leverage the public linked in website:

   https://www.linkedin.com/jobs/search

   Use the search terms "Data Analyst", "Data Science", "Analytics", "Business Intelligence" for Germany. Use the summary pane on the left side to identify the key firms, industries and regions where the job offers are located. Furthermore, extract the career tier of these listings.

   a. Visualize and discuss your results.
   b. Compare the different search terms – are the results similar?
   c. Using the summary statistic is a direct and easy way for scraping. However, what information do we lose this way? Argue whether or not this is a problem depending on the use case.

2. The vine village is over but we want to learn a bit about wine ratings and prices. The following website allows you to explore reviews of Rieslings from around the world:

   http://www.winemag.com/varietals/riesling/

   Use this overview website with different search filters to extract non-dessert / non-sparkling white Riesling prices, ratings, awards (Best buy, Editor's Choice, Cellar Selection), vintage (2012-2015) and origin (country level, six largest).

   Hint 1: For some attributes it may be easier to run extra searches

   Hint 2: After running a search the URL will include a fragment like this: &page=3&sort_by=pub_date_web&sort_dir=desc You can drop the sorting terms and then leverage explicit page numbering to effectively crawl different subpages.

   a. Visualize the price-rating relationship. Use colors / facets to incorporate your other data elements.
      - How would characterize the price-rating relationship?
      - Do the different origins differ in price and rating level?
      - Do the different vintages behave differently?
   b. We want to predict wine price based on the rating. Develop a simple linear regression model. Report and interpret your key results.
   c. Considering your visualization from a), pick a variable transformation which better captures the price-rating relationship and rerun the simple linear regression model. Compare the model performance.
   d. Develop a multiple regression model using the backward selection technique. Discuss the results.

3. You are planning to buy a new (used) car and this time want to do it in a data science fashion. Pick a vehicle of your choice (you should not opt for a too obscure make to find sufficient data), go to a used car website of your choice (e.g., autoscout24 or mobile.de) and retrieve at least 1,000 vehicles offers. To minimize the number of calls to the website you should avoid scraping from the individual offer site but rather extract as much information as possible from the overview site. [hint: again you may not have to extract the navigation links but can simply increment the page counter of the results url]

    a. Visualize the main value components (e.g., power, fuel type, mileage, age, color) in an appropriate manner. Try to combine multiple components where possible in your plots.
    b. Randomly choose 800 of your vehicles to train different linear regression models and subsequently use this model to forecast the prices of the remaining vehicles.
    c. Report the mean squared error of your models and compare this metric with the models' $R^2$ values.
    d. List the cars your models had most problems to predict. Can you explain the underlying problem?