**Exercise 1**

In the file **Insurance.csv** you can find some data about the insurance charges of several patients in the United States taken from the book written by Lantz (2013). Your task is to identify which factors play an important role in the level of the charges.

1. Create a new variable east\_west which measures whether a patient is from the east side or the west side of the United States. Hint: use the where() function from numpy
2. Create dummy variables of sex and smoker, add dtype = int to the get\_dummies command to turn the dummies into 0s and 1s (which is needed for calculating Vif). Check if there is multicollinearity between the variables (leave out the reference category when you do so).
3. Run a regression model with only east\_west as the independent variable. What do you see?
4. Test what the other variables do in combination with the east\_west variable. Which of them explains away the effect of the east\_west variable?

**Exercise 2**

In the file AB\_NYC\_2019 you can find a large list of Airbnb houses that were made publicly available by the Inside Airbnb (2020) website. All the houses are in New York from the year 2019. Your job is to build a regression model in which you try to predict the price of a listing based on the following variables: neighborhood group, latitude, longitude, room type, minimum number of nights, number of reviews, calculated host listings (how many houses the host has on Airbnb in total), availability (number of days per year that it is available), and the number of reviews per month.

Make sure to take these steps:

1. Make dummy variables of neighborhood group and room type
2. Check for multicollinearity (Neighbourhood\_group\_Manhattan and room\_type\_Private\_room as reference categories)
3. Standardize the variables
4. Present the results in an APA style table

Answer these questions:

1. What is the interpretation of all of the regression effects?
2. Which variable is the most important in predicting price?

**References**

Lantz, B. (2013). *Machine learning with R*. Packt publishing ltd.

Jost, S. D. (2020). *CSC 423 Data Analysis and Regression*. Retrieved from: http://facweb.cs.depaul.edu/sjost/csc423/