# Lab01: Basic tidyverse operations, simple modeling approach and data engineering

**Handed out:** Thursday, February 25, 2021

**Return date:** Thursday, March 11, 2021 by midnight in the *eLearning* assignment folder **Lab01Submit**.

**Grades:** Lab01 counts 16 % towards your final grade.

**Objectives:** Basic tidyverse operations, simple modeling approach and data engineering.

**Format of answer:** Your answers (statistical figures and verbal description) should be submitted as ***hardcopy***. Add a running title with the following information: Lab01, your name and page numbers. You may use this document as template. Copy the requested statistical figures into your document. Trial and error answers will lead to a deduction of points. Label each answer properly with the bold task and sub-task headings.

You are expected to hand in professionally formatted answers: use a fixed pitch font, like **Courier New**, for any  code and output (use single line-spacing). Copy and paste figures into your document. Make sure that each figure has a proper ***caption*** describing its content.

**Task 1:** Based on Chapter 5 *Data Transformation with dplyr* answer the following questions and show your code. Make sure to attach the libraries **tidyverse** and **nycflights13** to your session. [5 points]

[a] What is the difference between **==** and **near( )**? Give examples using numbers derived from floating point operations. [0.25 points]

[b] Find all flights that: [2 points]

Had an arrival delay of three or more hours.

Flew to Dallas (DFW and DAL).

Were operated by Southwest (WN) and American (AA).

Arrived more than two hours late but made up over 30 minutes in flight.

Were delayed by at least an hour but made up over 30 minutes in flight.

[c] How could you use **arrange( )** to sort all missing values to the start? (Hint: use **is.na( )**). Show your sample code and use a short artificial tibble. [0.5 points]

[d] What is the statement **select(tibble, var1, var2, everything())** achieving? [0.25 points]

[e] What is the difference between **mutate( )** and **transmute( )**? [0.25 points]

[f] Find the 10 most delayed flights using a ranking function. How do you handle ties? Read the documentation of **min\_rank( )**. [0.5 points]

[g] Which carrier has the worst delays? Challenge: can you disentangle the effects of bad destination airports versus bad carriers? Why or why not? Hint: think about **flights %>% group\_by(carrier, dest) %>% summarize(n( ))**. [0.75 point]

[h] What time of day should you fly if you want to avoid delays as much as possible? [0.5 points]

**Task 2:** Based on *Chapter 18 Pipes with magrittr* show the code how the piping operator **%>%** combines the three individual function calls **functA( )**, **functB( )** and **functC( )** internally into a joint function [2 points]

Sequence of function calls:

**x <- functA(w)**

**y <- functB(x)**

**z <- functC(y)**

**Task 3:** Follow the modeling approach outlined in Boehmke and Greenwell’s Chapter 2 *Modeling Process*. Your objective is to predict the probability of default in the dataset **credit.csv**. [6 points]

[a] Discuss your data splitting and resampling methods.[2 points]

[b] Show the code and output of your analysis. [3 points]

You can import the dataset with the statement

**credit <-read.csv("Drive:\\Path\\credit.csv",**

**header = TRUE, stringsAsFactors = TRUE)**

**sapply(credit, is.factor)**

Make sure to use proper a range to evaluate the hyperparameter of your nearest neighbor algorithm.

Because your prediction focuses on a classified data, the evaluation criterium needs to change from **metric = "RMSE"** to **metric = "Accuracy"**.

[c] Discuss the predictive quality of your model. [1 point]

You can evaluate the predicted probabilities with

**defaultPred <- predict(knn\_fit, default\_test, type="prob")**

**plot(defaultPred$yes~default, data=default\_test)**

**Task 4:** Follow the modeling approach outlined in Boehmke and Greenwell’s Chapter 3 *Feature and Target Engineering*. Use the dataset **tractShp** in the package **TexMix**. [3 points]

You can extract the attribute data-frame from the shape file with the statement **tractDf <- as.data.frame(tractShp)**.

Show all code and output.

[a] Evaluate the missingness of all variables in the dataset and impute their missing values. Should there be any census tracts that need to be excluded from the analysis on logical grounds? [2 points]

[b] Standardize all metric variables. [1 point]