Bitte frei lassen

Matrikelnummer:	
Sitzplatznummer:	

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Modulprüfung der Wirtschaftswissenschaftlichen Fakultät im Sommersemeste 2019

Applied Data Science (10608100)

Prüfer: Prof. Dr. Christoph M. Flath

- Die Bearbeitungszeit beträgt 60 Minuten.
- Bearbeiten Sie insgesamt 3 der 4 gestellten Aufgaben. Sollten Sie an mehr als 3 Aufgaben arbeiten, markieren Sie bitte deutlich welche Aufgaben gewertet werden sollen.
- Bearbeitung Sie die Aufgaben in den jeweils auf Github zur Verfügung gestellten R Skripten. Diese folgen jeweils der Namenskonvetion „ADS_SS19_q*.R“.
- Sie finden die benötigten Skripte und Dateien im Github Repository im Ordner „Exam“
- Alle von uns über Github bereitgestellten Unterlagen sind als Hilfsmittel zugelassen. Des Weiteren ist es erlaubt während der Klausur im Internet zu recherchieren. Ausdrücklich verboten ist jedoch die (digitale) Kommunikation während der Klausur.

Note:	
Punkte:	

Overview

Question	Points	
Functions in R	20	
Data Wrangling and Visualization	20	
Modelling	20	
Webscraping	20	
Total:	80	

Question 1 Functions in R . . . . *20 Points*

You recently started a new job as a student assistant at your university's IT department. Your first task is to generate email addresses as well as passwords for all incoming students.

To this end, you are provided with a data frame "names" of the following type:

firstName	**lastName**
Steve	Young
Natalia	Guerrero
. . .	. . .

(a) (5 Points) Write a function assigning an email address to a student given a first and a last name. Email addresses have to follow the pattern "firstname.lastname@uni-wuerzburg.de" (Note: You do not have to account for students with identical names).

(b) (5 Points) Write a function assigning a random initial password to a student. Initial passwords are 6 digit numbers (Note: You can use the function rdunif() to generate a random sample from a discrete uniform distribution).

(c) (5 Points) Write a function determining a user name for each student. User names consist of the first two letters of the first name (lowercase) followed by the first letter of the second name (lowercase) and 3 random integers between 0 and 9. For example, the user name of "Natalia Guerrero" could be 'nag317'.

(d) (5 Points) Write a function that returns a data frame with 5 columns (first name, last name, email address, password, user name) combining the functions from a), b), and c). Apply this function to all 100 students in the data frame "names".

Question 2 Data Wrangling and Visualization .... *20 Points*

The data frame “economics” provides information on economic indicators. The indicators are summarized in the following table:

Column Name	**Description**
date	Date of data collection
psavert	Personal savings rate
pce	Personal consumption expenditures
unemploy	Number of unemployed in thousands
uempmed	Median duration of unemployment
pop	Total population in thousands

(a) (3 Points) Use your data wrangling skills to calculate the unemployment rate and add it as an additional column (Note: You can assume that the whole population belongs to the labor force).

(b) (7 Points) Visualize the number of unemployed people. Subsequently, change the axis labels as well as the theme to recreate the following diagram:

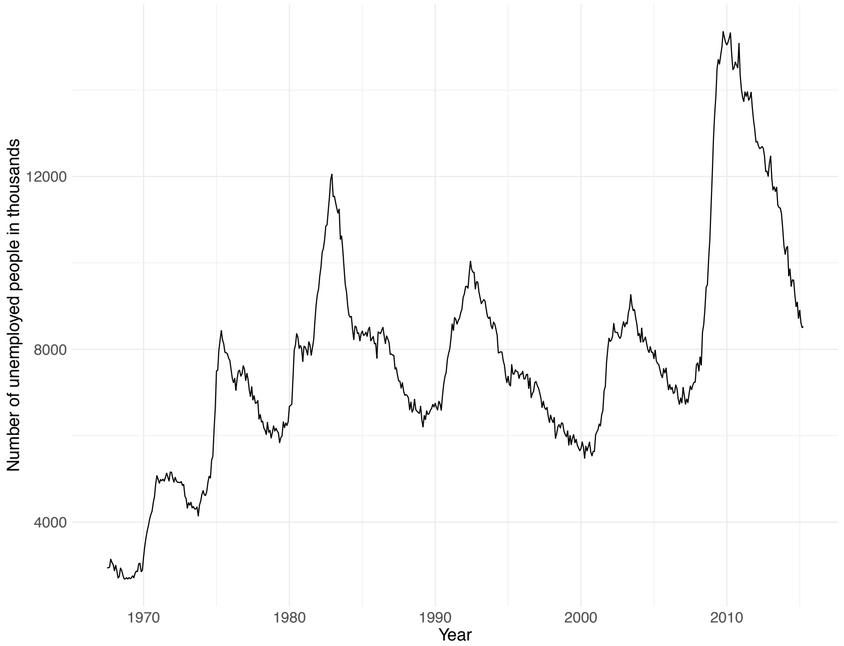

Abbildung 1: Caption

(c) (10 Points) Expand your code to visualize all indicators as facets (Note: You have to rearrange the data frame to the long format).

Question 3 Modelling .... *20 Points*

The data frame “titanic” provides information on the fate of passengers on the fatal maiden voyage of the ocean linear Titanic. The variables are summarized in the following table:

Column Name	**Description**
Survived	Passenger survival indicator
Pclass	Passenger Class
Sex	Gender
Age	Age
SibSp	Number of siblings/spouses aboard
Parch	Number of Parents/Children Aboard

Your task is to develop a machine learning model predicting the fate of individual passengers (Survived) based on the remaining variables in the following steps:

(a) (2 Points) Split the initial data set into 75% train and 25% test set. Use stratified sampling.

(b) Write the modelling recipe in the following steps:

i. (2 Points) Define the model recipe.

ii. (2 Points) Impute the missing values in all numeric predictors using mean imputation.

iii. (3 Points) Analyze the data types required for the different variables and transform them accordingly (Note: You will need 3 factor and 3 integer variables).

iv. (2 Points) Convert the factor variables to dummy variables.

(c) (3 Points) Prepare the recipe and apply it to the training as well as the test data.

(d) (3 Points) Use the preprocessed training data to train a boosting model (Note: Use the xgboost engine with 1000 trees and a tree depth of 5).

(e) (3 Points) Predict the survival of the passengers in the test set and evaluate the boosting model by reporting the confusion matrix.

Question 4 Webscraping . . . . *20 Points*

Your task is to scrape unanswered questions related to R from Stack Overflow. To this end, you are provided with the link to the first 50 questions (`https://stackoverflow.com/questions/tagged/r?tab=Unanswered&pagesize=50&page=1`).

Build your scraper according to the following steps:

(a) (3 Points) Create a vector of the urls for the first 3 pages (150 questions).

(b) (6 Points) Write a function to extract the title as well as the URL to the question page, all tags, and the time of the posting for a *single* question. The function should return a data frame with one row for *each tag* and repeated values for the other variables.

(c) (6 Points) Write a function that extracts all questions from one URL and subsequently applies the function from part b) to them.

(d) (5 Points) Apply your function to the first 3 pages. Analyze the resulting data frame to find the 10 most common tags.

The following CSS selectors should be helpful:

```
.summary

.post-tag

.question-hyperlink

.relativetime
```