Klausur Empirisches Arbeiten

Teil Explorative Datenanalyse mit R

Prüfer	Prof. Dr. Nicolas Meseth
Semester	Sommer 2023
Max. Punktzahl	40 (oder $\frac{1}{3}$ der Gesamtpunktzahl)
Erlaubte Hilfsmittel	alle

Hinweise zu diesem Klausurteil

- Bitte nutzt die Datei `nachname_vorname_lösungen.R` als Vorlage für die Beantwortung der Fragen und fügt euren R-Code jeweils unter der Frage ein. Bitte entfernt am Ende alle Codereste, die nicht zur Antwort gehören.
- Denkt daran, eure Matrikelnummer und Namen vor der Bearbeitung in die ersten beiden Zeilen einzutragen.
- Ersetzt vor der Abgabe eure Vor- und Nachnamen im Dateinamen. Als Beispiel: `max_mustermann_lösungen.R`
- Ladet die Datei über den Abgabeordner "Nachschreibeklausur Teil Meseth" im ILIAS-Lernraum der Veranstaltung hoch! Die Abgabe muss vor dem offiziellen Ende der Bearbeitungszeit erfolgen!

Teil 1: Datensatz "Campusbier-Bestellungen"

Im ersten von zwei Teilen könnt ihr insgesamt **20 Punkte** erreichen.

Bevor ihr mit der Bearbeitung der Aufgaben beginnt, kopiert die Datei `orders.csv` in euer Arbeitsverzeichnis und ladet den Datensatz als Tibble mit dem Namen `orders`.

```
library(tidyverse)
orders <- read_csv("orders.csv")
```

Aufgabe 1.1: Datentransformation

Beantwortet die folgenden Fragen mit R und dem Tidyverse. Das Ergebnis soll in diesem Teil als Tabelle (Tibble) und *nicht* als Visualisierung dargestellt werden.

a) Gebt alle Spaltennamen des Datensatzes aus, die mit dem Prefix "customer" beginnen enthalten! (1 Punkt)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

b) Der April welchen Jahres war vom Umsatz gesehen her der beste bisher? (3 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

c) Erstellt eine neue Spalte, in der ihr die Bestellungen anhand des Umsatzes in 3 Klassen einteilt. Die erste Klasse "small" soll für Umsätze unter 20 EUR vergeben werden, die zweite Klasse "medium" für Umsätze zwischen 20 und 40 EUR, und die dritte Klasse "large" für alle Bestellungen mit einem Umsatz über 40 EUR! (6 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

Aufgabe 1.2: Datenvisualisierung

Findet eine passende Visualisierungsform für die folgenden Fragen und erstellt diese mit R und `ggplot2`!

a) Tätigen Männer oder Frauen im Durchschnitt Bestellungen mit höherem Umsatz? (5 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

b) Erstellt ein Balkendiagramm, in dem jeder Balken den Umsatz eines Monats für den gesamten Zeitraum der Daten darstellt? Verwendet für jedes Jahr eine andere Farbe! (5 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

Teil 2: Datensatz "Titanic"

Im zweiten Teil könnt ihr insgesamt **20 Punkte** erreichen!

Bevor ihr mit der Bearbeitung der Aufgaben beginnt, kopiert die Datei `titanic.rds` in euer Arbeitsverzeichnis und ladet den Datensatz als Tibble mit dem Namen `titanic`. Der Datensatz beinhaltet Informationen zu den Passagieren der 1912 gesunkenen Titanic und stammen von der Datenplattform Kaggle.

```
library(tidyverse)
titanic <- readRDS("titanic.rds")
```

Aufgabe 2.1: Datentransformation

Beantwortet die folgenden Fragen mit R und dem Tidyverse. Das Ergebnis soll in diesem Teil als Tabelle (Tibble) und *nicht* als Visualisierung ausgegeben werden.

a) Wie viele Männer und Frauen waren an Bord der Titanic? (1 Punkt)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

b) Wer sind die zehn jüngsten überlebenden Passagiere? (2 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

c) Die Körper wie vieler Passagiere, die nicht überlegt haben, konnten später geborgen werden? (5 Punkte)

Hinweis: Die Spalte `body` beinhaltet eine Nummer, wenn ein Körper gefunden und identifiziert wurde. Sie beinhaltet `NA`, wenn kein Körper gefunden wurde.

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

d) Wie alt wäre der oder die älteste Überlebende heute? (2 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

Aufgabe 2.2: Datenvisualisierung

a) Erstellt eine passende Visualisierung für die folgende Frage: Gibt es Unterschiede bei den Überlebenschancen, wenn man die Ticketklassen (1. Klasse, 2. Klasse, 3. Klasse) betrachtet? (4 Punkte)

Hinweis: Die Ticketklasse findet ihr in der Variable `pclass`.

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

b) Erstellt den R-Code für die Histogramme unten und versucht dabei, das Diagramm möglichst exakt nachzubilden! (6 Punkte)

Hinweis: Das im Diagramm verwendete Theme ist `theme_bw` .