Übung Workshop Datenanalyse

31.1.2020

Einführung

Dieses Notebook können Sie im Rahmen der Übung zum Workshop Datenanalyse nutzen, um es mit Ihren eigenen Ergebnissen zu ergänzen. Zu Beginn werden die benötigten R-Pakete geladen. Sie werden Fehlermeldungen erhalten, wenn die Pakete noch nicht installiert sind - was für gewöhnlich der Fall sein wird.

Der folgende Codeblock wird wegen `eval=FALSE` nicht ausgeführt. Nur wenn Sie diesen *Chunk* direkt ausführen, findet die Installation statt.

Laden der Pakete

```
library(tidyverse)
```

Laden der Daten 1

Im ersten Teil arbeiten Sie mit einem Datensatz zu US-Präsidenten. Sie können die Daten folgendermaßen laden:

```
loc <- "https://raw.githubusercontent.com/anagistics/workshop_dataanalytics/master/presidents.csv"
presidents <- readr::read_csv2(loc)
```

Um einen ersten Blick auf die Daten zu werfen, können Sie das `glimpse`-Kommando verwenden:

```
glimpse(presidents)
```

Ein weiterer hilfreicher Befehl ist `summary(tab)`, dabei ist `tab` der Bezeichner der Datentabelle, für die man eine kompakte Zusammenfassung der Ausprägungen haben möchte.

Aufgaben Teil 1

Technische Validierung

1. Prüfen Sie, ob die Daten korrekt eingelesen wurden

- Stimmt das Zeichenencoding?
- Sind die Datentypen korrekt?

- Falls nicht: Was muss geändert werden?

1. Prüfen Sie auf fehlende Werte

- Sind fehlende Werte erklärbar?

1. Gibt es weitere Auffälligkeiten?

Fachliche Validierung

1. Datumsfelder sind oftmals eine Fehlerquelle. Prüfen Sie bitte die Datumsfelder, ob die dortigen Werte richtig sein können.
2. Nutzen Sie bitte Histogramme, um die Verteilung der Präsidentschaftsdauer, das maximale Alter und des Geburtsstaats zu untersuchen.

Laden der Daten 2

Der nächste Chunk lädt Daten zu KfZ-Haftpflichtversicherungen:

```
loc <- "https://raw.githubusercontent.com/anagistics/workshop_dataanalytics/master/motorinsurance.csv"
mid <- readr::read_csv2(loc)
```

Achten Sie in jedem Fall auf die Ausgabe beim Einlesen: Verstehen Sie die Angaben zu den Datentypen? Erscheinen Ihnen diese richtig?

Hier ist eine Erklärung der Datenfelder aus dem Dokument CAS datasets manual (Seite 62).

IDpol The policy ID (used to link with the claims dataset).

ClaimNb Number of claims during the exposure period.

Exposure The period of exposure for a policy, in years.

VehPower The power of the car (ordered values).

VehAge The vehicle age, in years.

DrivAge The driver age, in years (in France, people can drive a car at 18).

BonusMalus Bonus/malus, between 50 and 350: <100 means bonus, >100 means malus in France.

VehBrand The car brand (unknown categories).

VehGas The car gas, Diesel or regular.

Area The density value of the city community where the car driver lives in: from “A” for rural area *to “F” for urban centre.

Density The density of inhabitants (number of inhabitants per square-kilometer) of the city where *the car driver lives in.

Region The policy region in France (based on the 1970-2015 classification).

Daten verstehen

Dieser Datensatz ist erheblich umfangreicher als der erste, d.h. es ist wesentlich schwieriger, durch *Hingucken* ungewöhnliche Ausprägungen zu erkennen.

Auch hier sollten Sie sich zunächst einen Überblick über die Daten verschaffen:

```
glimpse(mid)
```

Aufgaben Teil 2

Wenn Sie eine große Zahl von Instanzen visualisieren wollen, kann das erstens etwas dauern und zweitens sieht man evtl. gar nicht mehr so viel. In diesem Fall kann es hilfreich sein, mit dem Kommando `sample_n` eine Stichprobe zu bilden.

1. Gemäß der Erklärung der Daten handelt es sich dabei um Daten aus einem Jahr.

- Wie beurteilen Sie mit diesem Wissen die Variable *Exposure*?

Einzelne Variablen

1. Untersuchen Sie die quantitativen Variablen aus dem Datensatz mittels *Boxplots*.

Mehrere Variablen

Den im Vorlesungsteil angesprochenen *Scatterplot* können Sie in R mit der `ggplot`-Bibliothek folgendermaßen erzeugen:

```
mtcars %>% ggplot() + geom_point(aes(x = wt, y = mpg, color = hp))
```

Mittels *Contourplots* kann man bivariate Verteilungsfunktionen (also von zwei Variablen) in einer 2D-Darstellung zeigen. Sie können dazu in R folgendermaßen vorgehen:

```
diamonds %>% ggplot(aes(x, y)) + stat_density_2d(aes(fill = stat(level)), geom = "polygon")
```

1. Erkennen Sie einen Zusammenhang zwischen
   - *DrivAge* und *Area*?
   - *Density* und *Area*?

Die Bevölkerungsdichte nimmt einen großen Wertebereich ein.

- Was könnte man tun, um diesen Wertebereich zu verkleinern ohne relevante Informationen zu verlieren?
- Wie sieht dann der Zusammenhang zwischen *Density* (Neu) und *Area* aus?
  - Wie könnte das den weiteren Verlauf Ihrer Analyse beeinflussen?