Fragensammlung Projekt

Welcher Datensatz soll verwendet werden?

Es soll der cancer_data.csv-Datensatz verwendet werden.

Dürfen im Ausgangsdatensatz Spalten auch komplett gelöscht werden, oder lediglich neu befüllt?

Wenn eine Spalte absolut nicht zum gewählten Krebstyp passt, dann kann diese auch gelöscht werden. Dies dann in der Präsentation und im Bericht explizit angeben.

Wie wird die Variable Geschlecht dekodiert?

Da dies ein simulierter Datensatz ist, dürfen Sie das frei entscheiden. Üblich ist 0 = männlich und 1 = weiblich. Auch diese Entscheidung bei der Präsentation und im Bericht angeben.

Was ist in der Spalte diet kodiert?

Hier ist die Ernährung kodiert – z.B. kommt es vor, dass manche Testgruppen fettreiches Essen vor der Medikamentengabe bekommen. Auch hier gilt: da es ein rein simulierter Datensatz ist, können Sie die Spalte auch nach Ihren Wünschen anpassen oder komplett löschen. Dies dann in der Präsentation und im Bericht explizit angeben.

Es wurde eine neue, binäre Variable erstellt, die mit 0 und 1 kodiert ist. Für eine Grafik soll diese zum Einfärben der Datenpunkte verwendet werden – allerdings ist die Variable nicht diskret, sondern stetig, und es wird ein Farbverlauf dargestellt. Woran liegt das?

Da für die Kodierung 0 und 1 verwendet wurde, ist der Datentyp der Variable integer (kann mit class(datensatzname$variablenname)). Bei der Verwendung in einem ggplot2-Plot wird R integer immer als stetige Variablen behandeln. Eine schnelle Lösung ist es, die Variable einfach in den Datentyp character umzuwandeln:

datensatzname$variablenname <- as.character(datensatzname$variablenname)

und schon ist die Variable diskret ;) Falls Sie die Variable als Integer noch weiterverwenden möchten, z.B. um Auszuzählen, wie häufig welche Ausprägung vorkommt, dann speichern Sie die character-Version einfach unter einem anderen Variablennamen:

datensatzname$anderer_variablenname <- as.character(datensatzname$variablenname)

Gibt es eine Möglichkeit, die deskriptiven Statistiken aus summary() in eine schöne Tabelle zu schreiben?

Sie können sich die Daten in einen Dataframe, hier test.data genannt, schreiben lassen mit einer do.call-Funktion aus baseR (mehr zu do.call hier:

https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/do.call - im Prinzip wenden Sie eine Funktion, hier summary(), auf alle Spalten Ihres Datensatzes df an):
test.data <- do.call(cbind, lapply(df, summary))
Achtung: R für die "summary statistics" bei character-Variablen nur die Gesamtzahl der Beobachtungen aus; mit der oben genannten Funktion wird die Anzahl Observationen in min, max und mean geschrieben und der Rest mit "character "aufgefüllt. Mein Tipp wäre daher, für Ihre zusammenfassende Tabelle zuvor alle character-Variablen temporär zu entfernen.

Aus der schönen summary-Tabelle sollen die NA-Einträge in Zeile 7 gelöscht werden. Wie geht das?

Am einfachsten, in dem man die tabelle (hier: test.data) überschreibt, ohne Zeile 7:
test.data <- test.data[-7,]

Für eine neue Spalte im cancer_data-Datensatz werden Werte via runif() zufällig generiert. Gibt es eine Möglichkeit, hier einen Bias für die niedrigen oder hohen Werte reinzubringen?

Nein, leider geht das aktuell nur mit sample() – die Ziehungen aus den Verteilungen erlauben keinen Bias. Sie können sich überlegen, nach dem Ziehen der Werte manche Werte gezielt mit anderen zu ersetzen, eine andere Möglichkeit gibt es leider nicht. Mehr zum Simulieren von Werten finden Sie z.B. hier: https://towardsdatascience.com/statistical-simulation-in-r-part-1-d9cb4dc393c9 (towardsdatascience.com ist allgemein eine sehr gute Seite, die ich Ihnen als Lesetipp empfehlen kann).

Nachtrag: ggf. geht es, wenn Sie das ganze mit einem Faktor multiplizieren, um die Daten hin zu großen oder kleinen Werten zu neigen.

Es sollen ausgewählte bzw eingetragene Werte als neue Zeile an den Studiendatensatz angehängt werden, wenn ein Knopf gedrückt wird. Wie geht das?

Da es recht umständlich ist, hier eine ausreichende Erklärung aufzuschreiben, ist stattdessen ein Beispiel mit guter Lösung angegeben: https://stackoverflow.com/questions/50259288/shiny-inputs-append-data-frame

Es soll ausgewählt werden, welcher von 3 Plots gerade angezeigt wird im mainpanel. Wie soll der Code dann am besten aufgesetzt werden?

Angenommen, Ihr Inputfeld für die Plotauswahl heißt plotinput in ui.R und hat die Werte plot1, plot2 und plot3. Dann schreiben Sie im server.R-Skript:

```
output$plot <- renderPlot({
    if (input$plotinput == "plot1")  { … Code für Plot 1 … }
   else if (input$plotinput == "plot2")  { … Code für Plot 2 … }
   else if (input$plotinput == "plot3")  { … Code für Plot 3 … }
 })
```

In ui.R rufen Sie dann einfach output$plot im mainpanel auf – die Grafik wird dann an die Auswahl angepasst!

Zum Einfärben eines Scatterplots wird eine RColowBrewer-Farbpalette verwendet. Hier tritt das Problem auf, dass R die hellen Farben zuerst verwendet werden, und man die Punkte so gut wie nicht sieht. Was kann man tun?

Zwei Lösungsansätze (Nummer 2 wurde verwendet):

1. Man dreht die Farbpalette um, so dass zuerst die dunklen Farben zuerst angezeigt werden. Dies geht mit dem Statement direction = -1. Beispielverwendung: scale_color_brewer(palette="Set3", direction = -1), und auch zum Beispiel hier, siehe Lösung 72: https://stackoverflow.com/questions/8750871/ggplot2-reverse-order-of-scale-brewer
2. Statt einer Farbpalette definiert und verwendet man gezielt die Anzahl Farben, die man braucht. Hier wurde dies realisiert mit dem Statement scale_fill_manual(values=c("#A9F0FF", "#EF5884", "#B8A9FF"))

Es wurde für Daten zuerst versucht, einen Grafiktyp anzuwenden, der leider dafür ungeeignet war – es kamen nur Fehlermeldungen. Gibt es eine Übersicht, wann welcher Grafiktyp verwendet werden kann?

Dazu gibt es u.a. hier einen guten Artikel: https://infogram.com/de/seite/waehlen-sie-das-richtige-diagramm-fuer-ihre-datenvisualisierung . Es lohnt sich generell vorab, eine Skizze seiner App zu machen und sich schon genau zu überlegen, welche Grafiken mit welchen Daten dargestellt werden sollen. Im Berufsleben bei Unsicherheiten am besten mit Kollegen oder Experten reden, die sich auf Datenvisualisierung spezialisiert haben – diese können einschätzen, ob der Grafiktyp sinnvoll ist und ggf. Alternativen vorschlagen.