Klausur Empirisches Arbeiten

Teil Explorative Datenanalyse mit R

Prüfer	Prof. Dr. Nicolas Meseth
Semester	WS 23/24
Max. Punktzahl	40 (oder $\frac{1}{3}$ der Gesamtpunktzahl)
Erlaubte Hilfsmittel	alle

Hinweise zu diesem Klausurteil

- Bitte nutzt die Datei `nachname_vorname_lösungen.R` 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 im ILIAS-Lernraum der Veranstaltung hoch! Die Abgabe muss vor dem offiziellen Ende der Bearbeitungszeit erfolgen!

Teil 1: Datensatz "REWE-Produkte"

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

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

```
library(tidyverse)
rewe <- read_csv("data/rewe_products.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 bool'sche Werte enthalten! (1 Punkt)

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

rewe |>
  select(where(is.logical)) |>
  colnames()
```

```
[1] "bio"              "vegan"            "vegetarian"       "manufacturerName"
```

b) Listet alle Produkte, denen Salz *zugesetzt* wurde! (2 Punkte)

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

rewe |>
  filter(str_detect(str_to_lower(ingredientStatement), "salz")) |>
  select(ingredientStatement, productName)
```

```
# A tibble: 3,828 x 2
   ingredientStatement                                                  productName
   <chr>                                                                <chr>
 1 Zutaten: KÄSEREIMILCH* Speisesalz mikrobielles Lab Säuerungsmitt~ REWE Bio M~
 2 Weizenmehl Wasser Natursauerteig (Weizenmehl Wasser) pflanzliche~ ja! Americ~
 3 Zutaten: KÄSEREIMILCH Speisesalz Käsereikulturen mikrobielles Lab ja! Gouda ~
 4 Schweinefleisch jodiertes Speisesalz (Speisesalz Kaliumjodat) Ko~ ja! Delika~
 5 pasteurisierte MILCH Kochsalz Milchsäurekulturen Calciumchlorid ~ Bergader B~
 6 Tomaten Speisesalz Säuerungsmittel Citronensäure                  ja! Tomate~
 7 Zutaten: Gouda (48 % Fett i Tr) [KÄSEREIMILCH Speisesalz Käserei~ ja! Gerieb~
 8 Zutaten: KÄSEREIMILCH Speisesalz Säuerungskulturen (enthalten MI~ ja! Mozzar~
 9 Zutaten: FRISCHKÄSE Speisesalz                                     ja! Frisch~
10 Putenbrust jodiertes Tafelsalz (Tafelsalz Kaliumjodat) Dextrose ~ Herta Fine~
# i 3,818 more rows
```

c) Welche fünf Produkte bieten den besten Preis pro Gramm enthaltenem Protein? Nutzt als Startpunkt die neue Spalte `grams`, die das Produktgewicht in Gramm aus dem Feld `grammage` extrahiert! (7 Punkte)

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

rewe |>
  mutate(grams = as.numeric(str_match(grammage, "^(\\d+)g")[,2])) |>
  filter(proteinInGram > 0) |>
  drop_na(grams, price) |>
  select(productName, grams, proteinInGram, price) |>
  mutate(total_protein = grams / 100 * proteinInGram) |>
  mutate(price_per_gram_protein = price / total_protein) |>
  select(productName, price_per_gram_protein) |>
  arrange(price_per_gram_protein) |>
  head(5)
```

```
# A tibble: 5 x 2
  productName                    price_per_gram_protein
  <chr>                                           <dbl>
1 ja! Zarte Haferflocken 500g                   0.00578
2 ja! Kernige Haferflocken 500g                 0.00578
3 ja! Paniermehl 1kg                            0.00608
4 ja! Spaghetti 500g                            0.0065
5 ja! Gemelli 500g                              0.0065
```

Aufgabe 1.2: Datenvisualisierung

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

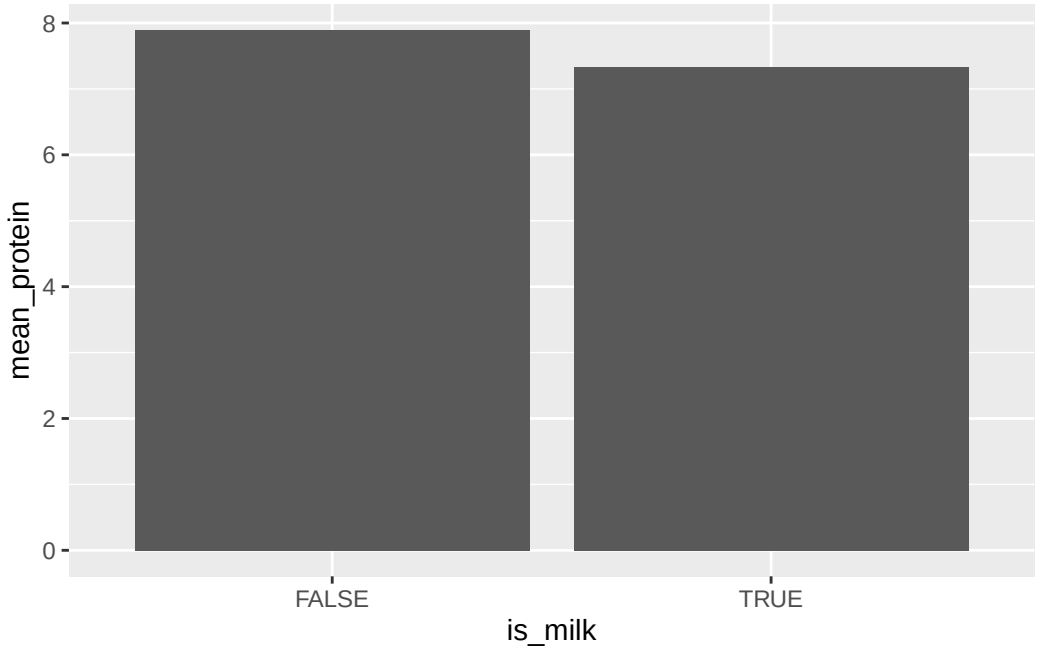
a) Enthalten Milchprodukte durchschnittlich mehr Eiweiß als andere Produkte? (5 Punkte)

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

rewe |>
  select(productName, allergenStatement, proteinInGram) |>
  mutate(allergenStatement = str_to_lower(allergenStatement)) |>
  mutate(is_milk = str_detect(allergenStatement, "milch")) |>
  drop_na() |>
  group_by(is_milk) |>
  summarise(mean_protein = mean(proteinInGram)) |>

  ggplot() +
```

```
  aes(x = is_milk, y = mean_protein) +
  geom_col()
```



b) Wie ist die Verteilung des Salzgehaltes für jede Unterkategorie der Kategorie "Nahrungsmittel"? Wählt eine sinnvolle Visualisierungsform, um die Verteilungen gut miteinander vergleichen zu können und zeigt nur den relevanten Bereich der Daten! (5 Punkte)

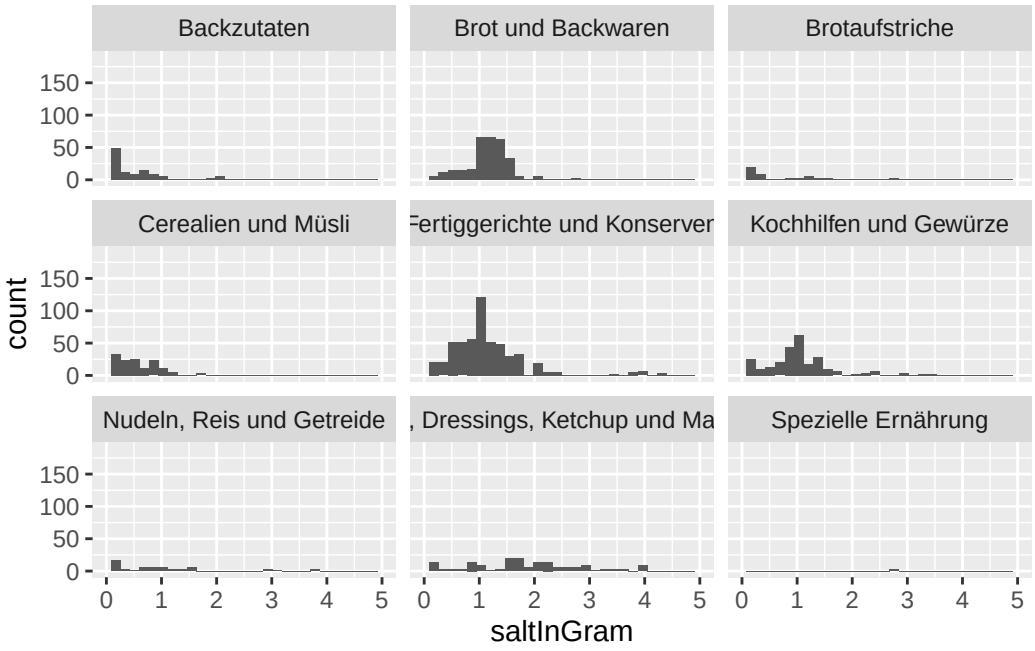
```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein

rewe |>
  filter(productCategory == "Nahrungsmittel") |>
  ggplot() +
  aes(x = saltInGram) +
  geom_histogram() +
  xlim(0, 5) +
  facet_wrap(~productSubCategory)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 229 rows containing non-finite values (`stat_bin()`).
```

```
Warning: Removed 18 rows containing missing values (`geom_bar()`).
```



Teil 2: Datensatz "Energie"

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

Bevor ihr mit der Bearbeitung der Aufgaben beginnt, kopiert die Datei `owid-energy-data.csv` in euer Arbeitsverzeichnis und ladet den Datensatz als Tibble mit dem Namen `energy`. Die Daten stammen aus Ritchie, Rosado, and Roser (2023). Ein Codebuch für die Spalten findet ihr unter diesem Link (wird für diese Klausur nicht benötigt).

```
library(tidyverse)
energy <- read_csv("data/owid-energy-data.csv")
```

Der Datensatz enthält Informationen zur Energie- und Elektrizitätserzeugung und zum Verbrauch aller Länder der Welt im Zeitverlauf.

Aufgabe 2.1: Datentransformation

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

a) Wie viele Variablen und Beobachtungen enthält der Datensatz? (1 Punkt)

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

energy |>
  dim()
```

```
[1] 21590   129
```

```
# Oder:

energy |>
  ncol()
```

```
[1] 129
```

```
energy |>
  nrow()
```

```
[1] 21590
```

b) Wählt alle Variablen aus, die einen Wert "pro Kopf" enthalten! (3 Punkte)

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

energy |>
  select(contains("per_capita"))
```

```
# A tibble: 21,590 x 30
   biofuel_cons_per_capita biofuel_elec_per_capita coal_cons_per_capita
                     <dbl>                   <dbl>                <dbl>
 1                      NA                      NA                   NA
 2                      NA                      NA                   NA
 3                      NA                      NA                   NA
 4                      NA                      NA                   NA
 5                      NA                      NA                   NA
 6                      NA                      NA                   NA
 7                      NA                      NA                   NA
 8                      NA                      NA                   NA
 9                      NA                      NA                   NA
10                      NA                      NA                   NA
# i 21,580 more rows
# i 27 more variables: coal_elec_per_capita <dbl>, coal_prod_per_capita <dbl>,
#   energy_per_capita <dbl>, fossil_elec_per_capita <dbl>,
#   fossil_energy_per_capita <dbl>, gas_elec_per_capita <dbl>,
#   gas_energy_per_capita <dbl>, gas_prod_per_capita <dbl>,
#   hydro_elec_per_capita <dbl>, hydro_energy_per_capita <dbl>,
#   low_carbon_elec_per_capita <dbl>, low_carbon_energy_per_capita <dbl>, ...
```

c) Welche 10 Länder hatten im Jahr 2021 den höchsten Energiebedarf (`energy_per_capita`) pro Einwohner? (3 Punkte)

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

energy |>
  filter(year == 2021) |>
  select(country, energy_per_capita) |>
  arrange(desc(energy_per_capita)) |>
  head(10)
```

```
# A tibble: 10 x 2
   country              energy_per_capita
```

```
   <chr>                          <dbl>
 1 Qatar                        199419.
 2 Bahrain                      161111.
 3 Iceland                      156924.
 4 Singapore                    153295.
 5 United Arab Emirates         139829.
 6 Trinidad and Tobago          111051.
 7 Norway                       105328.
 8 Brunei                       103268.
 9 Canada                       100739.
10 Kuwait                        98066.
```

d) Welches Land hatte 2019 die prozentual größte Reduktion beim Konsum fossiler Brennstoffe im Vergleich zum Vorjahr? (`fossil_cons_change_pct`)? Auf welchem Platz liegt Deutschland? (3 Punkte)

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

energy |>
  filter(year == 2019) |>
  select(country, fossil_cons_change_pct) |>
  arrange(fossil_cons_change_pct) |>
  head(10)
```

```
# A tibble: 10 x 2
   country             fossil_cons_change_pct
   <chr>                                <dbl>
 1 Estonia                             -21.7
 2 Venezuela                           -17.6
 3 Kuwait                               -9.00
 4 Denmark                              -7.64
 5 Slovakia                             -7.15
 6 Argentina                            -6.99
 7 Ukraine                              -6.22
 8 Hong Kong                            -5.24
 9 Finland                              -5.18
10 Trinidad and Tobago                  -5.08
```

```r
# Deutschland liegt auf Platz 11
```

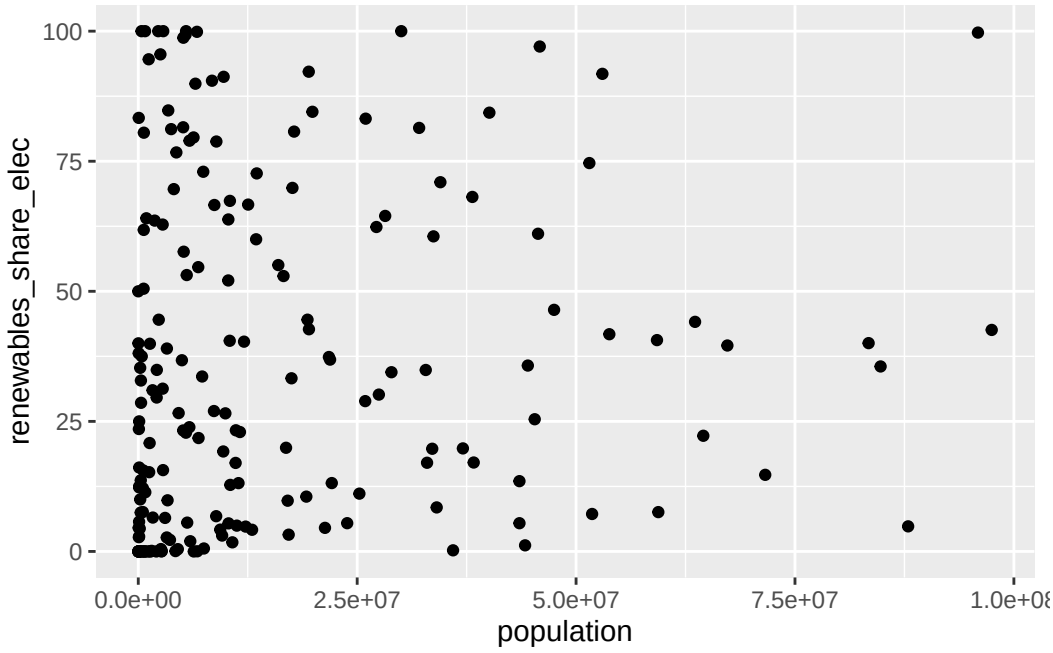
Aufgabe 2.2: Datenvisualisierung

a) Überprüft einen möglichen Zusammenhang zwischen der Einwohnerzahl (`population`) und dem Anteil an erneuerbaren Energien bei der Erzeugung von Strom (`renewables_share_elec`)? Betrachtet dabei das Jahr 2021 und nur Länder mit weniger als 100 Mio. Einwohner! (5 Punkte)

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

energy |>
  filter(year == 2021) |>
  filter(population < 100000000) |>
  ggplot() +
  aes(x = population, y = renewables_share_elec) +
  geom_point()
```

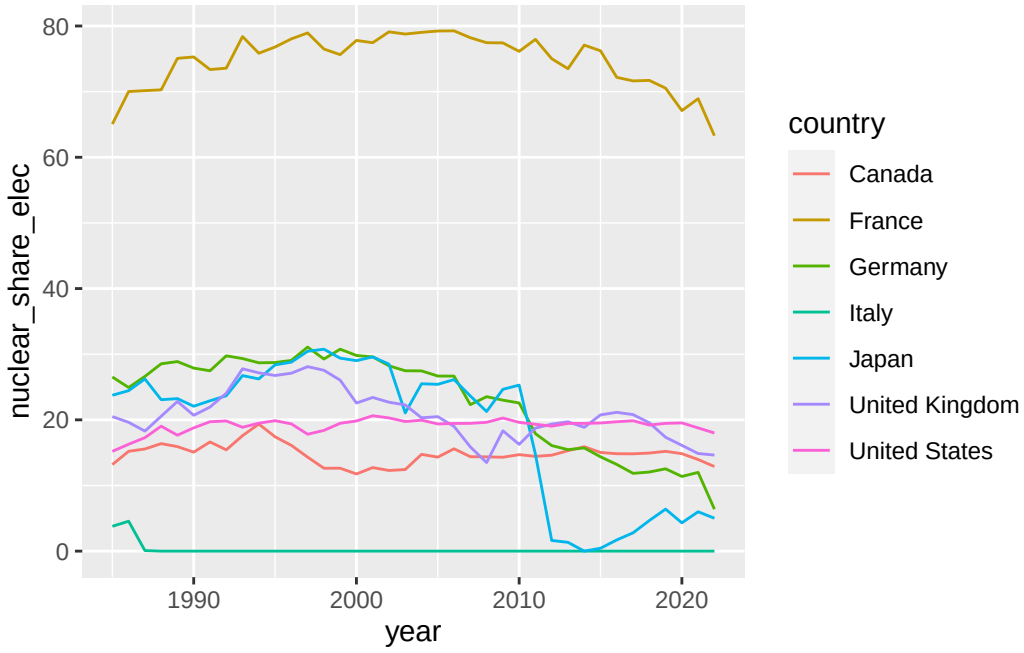
```
Warning: Removed 13 rows containing missing values (`geom_point()`).
```



b) Visualisiert die Entwicklung des Anteils der Stromerzeugung aus Atomkraftwerken (`nuclear_share_elec`) für die G7-Länder über die Zeit! (5 Punkte)

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

energy |>
  drop_na(nuclear_share_elec) |>
  filter(country %in% c("Canada", "France", "Germany", "Italy", "Japan", "United Kingdom",
  ggplot() +
  aes(x = year, y = nuclear_share_elec, color = country) +
  geom_line()
```



Quellen

Ritchie, Hannah, Pablo Rosado, and Max Roser. 2023. "Energy." *Our World in Data.*