Ćwiczenia 6 & 7 (Wizualizacja i przetwarzanie danych)

1. Zbiór danych `airquality` zawiera informacje o warunkach pogodowych w Nowym Yorku od maja do września 1973 roku. Wykonaj poniższy fragment kodu (pamiętaj o odpowiednich bibliotekach) aby przygotować dane do dalszych ćwiczeń.

```
airquality %>%
  select(Temp, Month, Day) %>%
  as_tibble() -> df
```

   (a) Przekształć dane w postać szeroką. Kluczem powinien być miesiąc, a wartościami temperatura.
   (b) Dane szerokie z poprzedniego punktu przywróć do postaci długiej.
   (c) Połącz zmienne `Day` oraz `Month` w nową zmienną `Date` o formacie %d.%m.
   (d) Podziel uprzednio utworzoną zmienną `Date` na dwie zmienne: `Day` oraz `Month`.
   (e) Wygeneruj pięć braków w danych za pomocą poniższego kodu.

```
df[sample(nrow(df), 5, replace = FALSE), 'Temp'] <- NA
```

   Zastąp braki danych (NA) przez Unknown.
   (f) Zastąp braki w danych za pomocą uzupełniania przez ostatnią zaobserwowaną wartość.

2. **(S)** Wszystkie polecenia w tym zadaniu dotyczą zbioru `auta2012` z pakietu `PogromcyDanych`.
   (a) Ile cech jest cechami jakościowymi?
   (b) Która marka samochodów jest najpopularniejsza?
   (c) Ile procent samochodów jest napędzane benzyną?
   (d) Ile aut ma cenę wyższą niż 2000 PLN?
   (e) Ile procent aut ma pojemność silnika większą bądź równą 1500 cm^3?
   (f) Ile aut zostało zarejestrowanych w Polsce i jest tańsze od 2000 PLN?
   (g) Ile procent aut ma pojemność silnika większą od 1500 cm^3 i jest dieslem.
   (h) Wybierz jedynie auta marki `Volkswagen`. Dla tak wybranych danych utwórz tablicę kontyngencji dla zmiennej `Type.of.fuel`.
   (i) Wybierz jedynie auta marki `Volkswagen`. Dla tak wybranych danych wyznacz średnią cenę i średni przebieg.
   (j) Wyznacz średnią cenę dla każdej marki.
   (k) Wybierz jedynie auta Toyota Corolla. Dla tak wybranych danych wyznacz pierwszy i trzeci kwartyl ceny.
   (l) Wybierz jedynie auta marki `Toyota`. Dla tak wybranych danych, dla każdego modelu wyznacz średnią cenę. Wyniki przedstaw posortowane w kolejności malejącej.
   (m) Wybierz auta Volkswagen Passat z roku 2006. Dla tak wybranych danych wyznacz średnią cenę. Ile spośród wybranych aut jest tańsze od 35 000 PLN?
   (n) Wybierz jedynie auta z roku 2007. Dla tak wybranych danych ile mamy aut każdej marki? Przedstaw wyniki w postaci posortowanej (kolejność rosnąca) po wielkości każdej grupy.

3. Zbiór danych `Fertility` z pakietu `AER` zawiera informacje na temat zamężnych kobiet w wieku 21-35 lat, które posiadają dwoje lub więcej dzieci (spis z roku 1980 w USA).

(a) Przyjrzyj się danym wykorzystując np. polecenie **glimpse()**.
(b) Wybierz wiersze od 35 do 50 i kolumny `age` oraz `work`.
(c) Wybierz ostatni wiersz danych.
(d) Ile kobiet miało trzecie dziecko?
(e) Która z kombinacji płci (4 możliwości) dla pierwszej dwójki dzieci jest najpopularniejsza?
(f) Wyznacz procent kobiet pracujących 4 tygodnie lub mniej biorąc pod uwagę czynnik rasowy.
(g) Wyznacz procent kobiet w wieku 22-24 lat, których pierwszym dzieckiem był chłopiec.
(h) Dla jakiej rasy proporcja chłopców jako pierwsze dziecko jest najmniejsza. Ile jest takich kobiet?
(i) Wyznacz procent kobiet posiadających trzecie dziecko z podziałem na płeć dwóch pierwszych dzieci.

4. Zbiór danych `Theoph` zawiera dane z eksperymentu dotyczącego farmakokinetyki teofiliny. Wykonaj poniższy fragment kodu aby uzyskać obiekt `df`.

```
df <- tibble::as_tibble(Theoph)
```

(a) Wybierz wszystkie kolumny pomiędzy (włącznie) `Subject` i `Dose`.
(b) Posortuje dane biorąc jako pierwsze kryterium wagę (rosnąco), a jako drugie czas (malejąco).
(c) Dodaj dodatkową zmienną `weight.cat`, która opsiuje klasyfikację osób według poniższego schematu:
  - Poniżej 66,8 kg – Welterweight,
  - 66,8 – 72,57 – Light-Middleweight,
  - 72,57 – 76,2 – Middleweight,
  - Powyżej 76,2 kg – Super-Middleweight.

(d) Pogrupuj dane ze względu na zmienną `weight.cat` i znajdź średni czas i sumę dawek dla każdej kategorii wagowej.