Домашняя работа 1

Лимонов А.А, Маркова Э.С, Гаврилова П.А

23 февраля 2022 г.

Содержание


```
library(tidyverse)
```

Домашняя работа 1

Для анализа был взят датасет с популярными для релокации городами. Предположим, что мы IT-специалист, который подбирает место для будущей жизни и его интересуют только часть переменных из датасета

```
best_cities_for_a_workation <-
  read_csv("best cities for a workation.csv") %>% select(-"Ranking") %>% rename(
  remote_connection_speed = "Remote connection: Average WiFi speed (Mbps per second)",
  coffee_price = "Caffeine: Average price of buying a coffee",
  appartment_price = "Accommodation: Average price of 1 bedroom apartment per month",
  drinks_price = "After-work drinks: Average price for 2 beers in a bar",
  restaurant_price = "Food: Average cost of a meal at a local, mid-level restaurant",
  city = "City",
  country = "Country",
  coworking_space = "Co-working spaces: Number of co-working spaces",
  taxi_price = "Travel: Average price of taxi (per km)",
  sunshine_hours = "Climate: Average number of sunshine hours",
  tripadvisor_stats="Tourist attractions: Number of 'Things to do' on Tripadvisor",
  instagram_photos="Instagramability: Number of photos with #"
)
```

На основании имеющихся переменных подсчитаем сколько примерно можно потратить за вечер, проведённый в городе:

```
best_cities_for_a_workation <- best_cities_for_a_workation %>% mutate(
  average_evening_spends = taxi_price * 5 + drinks_price + restaurant_price
)
```

После обработки датасет выглядит так:

```
## tibble [6 x 13] (S3: tbl_df/tbl/data.frame)
##  $ city                    : chr [1:6] "Bangkok" "New Delhi" "Lisbon" "Barcelona" ...
##  $ country                 : chr [1:6] "Thailand" "India" "Portugal" "Spain" ...
##  $ remote_connection_speed: num [1:6] 28 12 33 37 17 37
##  $ coworking_space         : num [1:6] 117 165 95 136 67 40
##  $ coffee_price            : num [1:6] 1.56 1.42 1.56 1.59 1.22 1.2
##  $ taxi_price              : num [1:6] 0.82 0.19 0.4 1.01 0.47 0.72
```

```
## $ drinks_price           : num [1:6] 3.08 2.9 3.42 5.12 2.16 2.4
## $ appartment_price       : num [1:6] 415 179 736 768 230 ...
## $ restaurant_price       : num [1:6] 1.54 2.9 7.69 10.25 5.15 ...
## $ sunshine_hours         : num [1:6] 2624 2685 2806 2591 2525 ...
## $ tripadvisor_stats      : num [1:6] 2262 2019 1969 2739 1660 ...
## $ instagram_photos       : num [1:6] 28386616 28528249 10205538 62894055 21293975 ...
## $ average_evening_spends : num [1:6] 8.72 6.75 13.11 20.42 9.66 ...
```

Сохраним полученный датасет в формат .rds

```
saveRDS(best_cities_for_a_workation, file="our_data.rds")
```

Теперь перейдём к разделению на группы, посмотрим на список стран, выберем из них несколько интересующих нас и выделим из датасета 5 стран с наибольшим количеством городов

```
top_five_countries <-
  best_cities_for_a_workation %>% group_by(country) %>%
  summarise(count=n()) %>% arrange(desc(count)) %>% slice(1:5)
usa_cities <- best_cities_for_a_workation %>% filter(
  country =="United States"
)
germany_cities <- best_cities_for_a_workation %>% filter(
  country == "Germany"
)
canada_cities <- best_cities_for_a_workation %>% filter(
  country == "Canada"
)
spain_cities <- best_cities_for_a_workation %>% filter(
  country == "Spain"
)
uk_cities <- best_cities_for_a_workation %>% filter(
  country == "United Kingdom"
)
countries_and_ctities <-
  bind_rows(usa_cities, germany_cities, canada_cities, spain_cities, uk_cities)
str(countries_and_ctities)
```

Затем посчитаем основные описательные статистики для всех наших стран

```
  countries_and_ctities %>%
  summarise(
    mean(average_evening_spends),
    median(average_evening_spends),
    sd(average_evening_spends),
    min(average_evening_spends),
    max(average_evening_spends)
    ) %>% str()
```

```
## tibble [1 x 5] (S3: tbl_df/tbl/data.frame)
## $ mean(average_evening_spends)  : num 25
## $ median(average_evening_spends): num 25.1
## $ sd(average_evening_spends)    : num 4.03
## $ min(average_evening_spends)   : num 15.1
## $ max(average_evening_spends)   : num 33.5
```

Затем посчитаем отдельно для Германии, Великобритании и США

```
germany_cities %>%
summarise(
  mean(average_evening_spends),
  median(average_evening_spends),
  sd(average_evening_spends),
  min(average_evening_spends),
  max(average_evening_spends)
  ) %>% str()
```

```
## tibble [1 x 5] (S3: tbl_df/tbl/data.frame)
##  $ mean(average_evening_spends)  : num 24.3
##  $ median(average_evening_spends): num 23.9
##  $ sd(average_evening_spends)    : num 1.53
##  $ min(average_evening_spends)   : num 22.1
##  $ max(average_evening_spends)   : num 27.3
```

```
uk_cities %>%
summarise(
  mean(average_evening_spends),
  median(average_evening_spends),
  sd(average_evening_spends),
  min(average_evening_spends),
  max(average_evening_spends)
  ) %>% str()
```

```
## tibble [1 x 5] (S3: tbl_df/tbl/data.frame)
##  $ mean(average_evening_spends)  : num 27.8
##  $ median(average_evening_spends): num 28.2
##  $ sd(average_evening_spends)    : num 4.54
##  $ min(average_evening_spends)   : num 20.1
##  $ max(average_evening_spends)   : num 33.5
```

```
usa_cities %>%
summarise(
  mean(average_evening_spends),
  median(average_evening_spends),
  sd(average_evening_spends),
  min(average_evening_spends),
  max(average_evening_spends)
  ) %>% str()
```

```
## tibble [1 x 5] (S3: tbl_df/tbl/data.frame)
##  $ mean(average_evening_spends)  : num 27.3
##  $ median(average_evening_spends): num 26.2
##  $ sd(average_evening_spends)    : num 3.01
##  $ min(average_evening_spends)   : num 23
##  $ max(average_evening_spends)   : num 31.7
```

Для сравнения посчитаем то же самое для всех городов из нашего датасета

```
best_cities_for_a_workation %>%
summarise(
  mean(average_evening_spends),
  median(average_evening_spends),
  sd(average_evening_spends),
  min(average_evening_spends),
```

```
  max(average_evening_spends)
  ) %>% str()
```

```
## tibble [1 x 5] (S3: tbl_df/tbl/data.frame)
##  $ mean(average_evening_spends)  : num 19.4
##  $ median(average_evening_spends): num 20.4
##  $ sd(average_evening_spends)    : num 9.06
##  $ min(average_evening_spends)   : num 4.41
##  $ max(average_evening_spends)   : num 45.8
```

Если взглянуть на итоговые данные, то можно заметить, что разброс средних трат среди всех стран довольно высокий: sd(average_evening_spends) = 9.06, что связано с малым кол-вом данных о городах и их расположении. В то же время отметим, что в Германии, например, довольно стабильные траты на вечер, в отличие от других интересующих нас стран. Также среднее среди трат внутри 5 наиболее представленных стран лежит в пределах 1 стандартного отклонения среди всех городов. Можно попробовать определить, что больше всего влияет на нашу перменную, посчитав статистики для всех городов и сравнив их с нашими

```
best_cities_for_a_workation %>%
  select(drinks_price, taxi_price, restaurant_price) %>%
  summary()
```

```
##   drinks_price      taxi_price     restaurant_price
##  Min.   : 1.080   Min.   :0.150   Min.   : 1.250
##  1st Qu.: 3.340   1st Qu.:0.545   1st Qu.: 4.800
##  Median : 5.740   Median :0.940   Median : 8.520
##  Mean   : 6.204   Mean   :1.005   Mean   : 8.159
##  3rd Qu.: 8.570   3rd Qu.:1.310   3rd Qu.:11.495
##  Max.   :17.800   Max.   :3.000   Max.   :19.760
```

Как можно заметить исходя из среднего значения, наибольшее влияние оказывают цена на напитки и ужин в ресторане, стоимость такси влияет меньше всего.