作业一

zetatech-bootcamp-101

截止日期：2021-05-15

1 准备工作

首先，请你下载我提供的 github repo。进入 github，点击绿色的 `code`，里面有 `Download Zip` 进行下载。下载后，进入文件夹，双击 `zetatech-bootcamp-101.Rproj` 进入这个 R project (Rporj)。

1. Loading packages 导入包

如果你没有按照 `here` 包的话，请你使用 `install.packages("here")` 进行安装。

```
library(tidyverse)
library(here)
```

2. 导入数据

- `rds` 是一组文件类型，专供 R 使用的
- `here` 函数让我们方便的找到文件的路径

```
dat <- read_rds(here("data", "fake_fulldata.rds"))
```

2 探索性分析 EDA

2.1 了解数据的基本情况

Use `glimpse` function to check the basic information about your data.

请运行下方 code,

```
glimpse(dat)
```

注意：日期类型 `<date>` 和字符串 `<chr>` 的区别！容易出错哦！

```
a <- class(as.Date("2021-05-15"))
a
```

```
## [1] "Date"
```

```r
b <- class("2021-05-15")
b
```

```
## [1] "character"
```

```r
a == b
```

```
## [1] FALSE
```

使用 summary 查看数据基本情况

```r
summary(dat)
```

2.1.1 Question

- dat 有多少行？多少列？
- 哪些变量是 <date> 类型的？
- 哪些变量是 <dbl> 类型的？
- age 的最大最小值、四分位点 quantiles，以及平均值是多少？
- 猜猜看这份数据是什么行业的？

3 数据清理

现在我们来清理 dat 数据集。

3.1 filter 函数

根据我们的条件，来筛选目标行，可以使用 filter 函数。

3.1.1 Question

1. 筛选 dat 并把结果保存到 subdat 数据集。条件包括：
    - date 介于 2018-01-17 至 2018-05-14 直接
    - cate_tag 为钻石
    - 最终成交价格 fnl_price 在 10000 以上

检测你的结果是否如下：

```r
subdat %>%
  summarize(
    min_date = min(date),
```

```
    max_date = max(date),
    mean_price = mean(fnl_price),
    cate_tag = unique(cate_tag),
    n_species_tag = n_distinct(species_tag),
    total_row = n()
  )
```

min_date	max_date	mean_price	cate_tag	n_species_tag	total_row
2018-01-17	2018-05-14	26876.45	钻石	5	118

2. 找到 `subdat` 中满足以下条件的行：
    - 种类标签 `species_tag` 为吊坠或耳饰
    - 主石大小 `mstone` 介于 30 到 40 之间
    - 把结果保存到 `temp` 中

检测你的结果：

```
temp %>%
  select(cardNum, mstone, species_tag)
```

cardNum	mstone	species_tag
uid03144	30.9	吊坠
uid12187	34.6	吊坠

3.2 Arrange 函数

3.2.1 Question

1. 对 subdat 的以下变量进行排序：
    - 对 `age` 升序，同时对 `fnl_price` 降序

3.3 Select 函数

3.3.1 Question

1. 对于 subdat，选择 cardNum, prodNum 两列
2. 对于 subdat，选择第一列，第二列，使用变量的位置
3. 对于 subdat，选择变量名中含有 `card` 的列
4. 对于 subdat，选择变量名中不包含 `_` 的列

5. 对于 subdat，选择变量名中以 `phone` 结尾，但不以 `cell` 开头的列。注意：不要直接选择 homephone，通过代码来判断
6. 对于 subdat，选择 cardNum, prodNum 两列，并将它们重新命名为：会员卡号，条码号
7. 函数 everything() 有什么作用，举例说明
8. 函数 rename() 和 select() 有什么相似处，有什么不同？举例说明

3.4 `group_by` 和 `summarize` 函数