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)
library(kableExtra)
```

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,

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

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
res <- 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()
  )

# this is only for formatting, you can ignore
res %>%
  kbl() %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

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：

```
res <- temp %>%
  select(cardNum, mstone, species_tag)

# this is only for formatting, you can ignore
res %>%
  kbl() %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

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

3.4.1 Question

下列问题基于 `dat` 数据集

1. 请你用 `summarize` 函数证明：`prodNum` 能够唯一确定一条销售记录，即：不存在两条销售记录拥有同样的 `prodNum`。提示：需要使用 `n_distinct()`, `n()`。不同的 `prodNum` 有多少？整个数据有多少行?
2. 使用 `group_by`, `summarize` 等函数证明：每个会员卡号 `cardNum`，唯一对应一个手机 `cellphone`，唯一对应一个会员姓名 `cardHolder` 。提示：使用 `arrange` 。
3. 计算每个会员的 RFM 统计量
   - 每个会员到店多少次 (freq)？注意：交易日期 (date) 相同，计作一次。每个会员有多少个不同的交易日期 (date)，这个量被计作次数 (freq)。
   - 每个会员最近一次消费发生在哪个日期 (recency)?
   - 每个会员的总消费 (monetary) 是多少？注意：fnl_price 为单次成交金额
   - 提示：结果是包含 4 列的 tibble：cardNum, freq, recency, monetary
4. 每个会员在：各个品类 `cate_tag` 下的平均消费是多少？各个品类 `cate_tag` 下买了多少件产品？各个品类 `cate_tag` 下购买产品的最高价格是多少？各个品类 `cate_tag` 下的总消费金额是多少？**针对每个会员**，请按照其购买各品类的总消费金额由大到小进行排序。注意：1. 不要忘记 `ungroup` 。2. 认真阅读 `arrange` 函数中的参数。结果的前 5 行应类似下方表格：

cardNum	cate_tag	n	mean_price	max_price	total_sale
uid00001	K 金	2	954	1344	1908
uid00001	足金	1	550	550	550
uid00001	银饰品	2	128	246	256
uid00002	K 金	1	1845	1845	1845
uid00003	钻石	1	15136	15136	15136

5. 请你使用：`group_by`, `summarize`, `arrange`, 探索每个品类 `cate_tag` 下，成交金额 `fnl_price` 的波动情况，即：计算标准差。按照波动情况由大到小将品类 `cate_tag` 进行排序，并用文字来简述你的发现。结果的前 3 行应如下表：

cate_tag	sd
其他	40239.312
钻石	11701.999
玉器	7797.937

6. 重复上面第 5 题，但是这次请你，仅使用 `count` 函数来得到同样结果，包括列名哦。请你思考 `count` 与 `group_by`, `summarize` 之间的关系，因为将来 `count` 将成为你的最爱。

3.5 Mutate 函数

3.5.1 Question

下列问题基于 `dat` 数据集

1. 仅选择 `cate_tag`, `fnl_price`, `weighted_price` 3 列, 增加一列 `wt`, `wt` 为 `fnl_price/weighted_price`。
2. 由于 `mstone` 的单位是 Ct，修改 `mstone` 这列，将其至转化至 Pt, 提示：1Ct=100Pt
3. 计算每个会员 `cardNum` 在各品类 `cate_tag` 的总消费金额，然后增加一列，该会员的总消费金额。然后增加一列，该会员各品类的消费占比。结果的前 5 行应如下表：

cardNum	cate_tag	sale	total_sale	pct
uid00001	K 金	1908	2714	0.7030214
uid00001	足金	550	2714	0.2026529
uid00001	银饰品	256	2714	0.0943257
uid00002	K 金	1845	1845	1.0000000
uid00003	钻石	15136	15136	1.0000000