数据分析与可视化 第二次作业

说明：

1. 作业上交日期为 2020 年 11 月 10 日中午 12 点，提交至助教邮箱。

2. 请将作业保存成 pdf（不要用 word）上传；文件名为 XXXX_YY_Ass2.pdf，其中 XXXX 为你的学号名，YY 为你的名字。

3. 作业中的每个问题，涉及到代码问题的，都需要在该题目位置附上相应的 R（Rstudio）源代码，不要用截屏记录代码，否则视为没有作答（助教将复制并运行你的源代码）。

4. 作业中涉及到简答问题的，请给出你的答案和理由。

5. 涉及到生成随机数的问题，请设定“种子”；其中 seed 的值请设定为你的 9 位数学号，请在合适的位置用如下代码： set.seed(XXXXXXXX)

利用 tidyverse 的框架，基于数据 dataset-cac-ma.xlsx 进行如下内容

（1）将数据读入 R，并保存成 tibble 类型的数据。展示该数据框的前 6 行。

（2）展示该数据的所有变量的变量类型和描述性统计。

（3）选择第 5-10 行在 list_price 和 cost 这两列上面的数据。

（4）分别选择 customer 为 175749 的所有数据, 以及 region 上取值为 Midwest 的所有数据

（5）该数据中有多少个不同的 region 取值？其中是否有错误的数据？如果有，请改正。

（6）计算 sales revenue， variable cost 和 contribution margin 的数值，其中：

sales revenue = quantity sold * list_price

variable cost = quantity sold * cost

contribution margin = (sales revenue - variable cost) / sales revenue

（7）根据 year 和 quarter 的信息，计算每个地区的 contribution margin 的平均值。

（8）展示每年中，contribution margin 平均值最高的前 3 个 region。这个名单随着年份变化而变化吗？

（9）每年中，最赚钱的 collection 前 3 名分别是什么？

（10）2018 年中，每个 brand 最赚钱的 collection 是什么？