Fall 2020 UTSG STA304 Final Report Guide Manual

1. 作业内容
    - 5 个选项（rohan 班 7 个）选一个
    - 除了 a,b 和 d 以外的其他所有选项都不推荐。
2. A 选项：
    - Data
        - Toronto Open Data Portal (https://open.toronto.ca/)
        - Canadian General Social Survey (GSS) (http://dc.chass.utoronto.ca/myaccess.html)
        - American Community Surveys (IPUMS) (https://usa.ipums.org/usa/cite.shtml)
        - Kaggle (https://www.kaggle.com/)

      选 data 的注意事项：
        - Categorical 和 numerical 都有
        - 不要选大段文字的 data
        - Observation 要够多（去掉 missing data 以后至少>500 行，越多越好）
        - 找的 data 合不合适直接决定了后面的步骤好不好写
    - 备选方案
      找出第一个 problem set 里当初研究的 dataset，用这门课后面学的知识重新做一遍当初的作业。或者用上一个 problem set 的 data 但是改变研究方向
    - 思路举例：[redacted] Data: 收入水平的高低(treatment)对于选举的偏好(variable of interest)真的有影响吗？
      可以用的 Analysis: Propensity Score Matching+ Logistic Regression → Causal Inference
      Sample code 在 OptionA_example.r 里
    - 注意：
        - 写文章的时候，要体现你所使用的 method 是哪一种？Question-driven, data-drive or method-drive?
        - 要提到 causal inference
3. B 选项：
    - Data

- Canadian Election Survey (CES)
  必须要用
  获取 data 的 R Code 可以在 problem set1 的 code 里面找到：

```
library(cesR)
library(labelled)

# call 2019 CES online survey
get_ces("ces2019_web")

# convert values to factor type
ces2019_web <- to_factor(ces2019_web)
head(ces2019_web)
```

  因为要做 MRP，最重要的是 demographic data 和 voting intention。
  demographic data ： cps19_gender, cps19_province, cps19_education
  voting intention： cps19_votechoice

- Post-Stratification Data (i.e. Census Data)
  -自己去找合适的 datasource
  -推荐：Stat Canada 2016 Education Census
  https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/edu-sco/index-eng.cfm
  -右下角下载 “Highest level of educational attainment (general) by sex and selected age groups”
  -解压以后用名字里有 “CANPR”的那个.csv 文件，这个的意思是 Canadian Province。其他的几个文件用不上。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Geographic c	Geographic r	Global non-r	Data quality	Age	Sex	Total - Highest certific	No certificat	Secondary (h	Apprenticesh	College, CEG	University ce	University ce	Total - Highe	No certifica
2	1	Canada	5.1	20000	All ages, 15-	Both sexes	28643015	5239580	7576400	2800265	5553830	813335	6659620	100	18.3
3	1	Canada	5.1	20000	All ages, 15-	Male	13990430	2667995	3686630	1906610	2326940	331940	3070320	100	19.1
4	1	Canada	5.1	20000	All ages, 15-	Female	14652585	2571585	3889765	893650	3226885	481395	3589300	100	17.6
5	1	Canada	5.1	20000	25 to 64	Both sexes	18931385	2169795	4494590	2042425	4241985	580885	5401710	100	11.5
6	1	Canada	5.1	20000	25 to 64	Male	9268560	1200105	2247020	1377775	1786065	240040	2417550	100	12.9
7	1	Canada	5.1	20000	25 to 64	Female	9662825	969685	2247565	664655	2455920	340850	2984155	100	10
8	1	Canada	5.1	20000	25 to 34	Both sexes	4576575	398475	1007695	457755	987585	125480	1599585	100	8.7
9	1	Canada	5.1	20000	25 to 34	Male	2264965	235565	573790	312840	432175	52135	658455	100	10.4
10	1	Canada	5.1	20000	25 to 34	Female	2311610	162905	433905	144915	555410	73350	941135	100	7
11	1	Canada	5.1	20000	35 to 44	Both sexes	4507775	395425	926885	467345	1065970	145130	1507020	100	8.8
12	1	Canada	5.1	20000	35 to 44	Male	2195070	225525	502430	310500	454280	59950	642380	100	10.3
13	1	Canada	5.1	20000	35 to 44	Female	2312700	169900	424450	156845	611690	85175	864640	100	7.3

  -这个 data 需要 pivot 了以后才能用，你可以选择名字里有 count 的 columns，也可以选择名字里有%的 columns，本质上都是同一种 data，但是不要都选. Pivot 的 code 可以这么写：

```
library(tidyverse)
data<-read_csv('98-402-X2016010-T1-CANPR-eng.csv')
educ_cols_count<-c("Total - Highest certificate, diploma or degree (2016 counts)"
        ,"No certificate, diploma or degree (2016 counts)"
        ,"Secondary (high) school diploma or equivalency certificate (2016 counts)"
```

```
          ,"Apprenticeship or trades certificate or diploma (2016 counts)"
          ,"College, CEGEP or other non-university certificate or diploma
(2016 counts)"
          ,"University certificate or diploma below bachelor level (2016
counts)")
data_pivot<-data %>% select(c("Age","Sex",educ_cols))%>%
  pivot_longer(cols=educ_cols_count,
names_to='education',values_to="total_count")
```

如果有别的方法 pivot，请用自己的方法

-Pivot 之后它长这样：

Filter

	Age	Sex	education	total_count
1	25 to 34	Both sexes	Total – Highest certificate, diploma or degree (2016 c...	4576575
2	25 to 34	Both sexes	No certificate, diploma or degree (2016 counts)	398475
3	25 to 34	Both sexes	Secondary (high) school diploma or equivalency certifi...	1007695
4	25 to 34	Both sexes	Apprenticeship or trades certificate or diploma (2016 ...	457755
5	25 to 34	Both sexes	College, CEGEP or other non-university certificate or ...	987585
6	25 to 34	Both sexes	University certificate or diploma below bachelor level ...	125480
7	25 to 34	Male	Total – Highest certificate, diploma or degree (2016 c...	2264965
8	25 to 34	Male	No certificate, diploma or degree (2016 counts)	235565
9	25 to 34	Male	Secondary (high) school diploma or equivalency certifi...	573790
10	25 to 34	Male	Apprenticeship or trades certificate or diploma (2016 ...	312840

-注意这个 dataset 里面是有重复性的 data 的，比如 age 里面有“All ages”开头的内容，Sex 里面有“Both sexes”. 这一类的 data 要记得用 filter() 删掉.

-**因为 stat Canada 最新的 census 是 2016 年的，所以 assumption 一定要写 assume 16 到 19 年 population 没有显著变化**

-从这一步开始往后就和上一个 problem set 的内容基本上一样了

-注意：这个 dataset 不是唯一的！只是一种可以用的 source 而已，如果有更合适的选项那么用自己的 dataset 更好.

4. D 选项

-Upworthy 是个新闻平台，常年做实验，把内容一模一样的文章稍微改一改标题/摘要/图片然后一起放在网上展示，然后看读者们的反应如何。他们把这么多年积累的数据攒在一起供人研究用。
https://upworthy.natematias.com/index

-Data access 需要发邮件给：cebersole@virginia.edu 他大概过 2-3 个工作日会回复你。

Sounds great—how do I get started?

For access to the Exploratory Dataset, email Charlie Ebersole (cebersole@virginia.edu).

-同一个 clickability_test_id 底下，文章的内容是一样的，但是其他信息（比如 headline，excerpt）可能不同。每一行代表一条文章，Upworthy 叫它 package。

-Data 里每一列的具体意义：https://upworthy.natematias.com/about-the-archive

-可以参考的写作思路：

1. 有没有图片(eyecatcher_id)会不会影响点击率(clicks/impression)？
2. 有没有摘要(excerpt)会不会影响点击率(clicks/impression)？
3. headline 的长度会不会影响点击率(clicks/impression)？
4. Headline 的长度，摘要和图片的有/无会对点击率造成影响吗？