Big Data and Social Analysis

Spring 2022

Midterm Exam

April 21, 2022

- There are 2 sections and 13 questions in the exam.
- The total points are 200 (+ 10 bonus points).
- You only need to write R codes and answers in your answer sheet.
- You don't need to paste plots and write comments for R codes.

Section 1:

The data files **result_2018.xlsx** and **result_2020.xlsx**[1] are li-level data of 2018 Taiwan local election and 2020 presidential election results.

In November 2018's election, Taiwan's ruling party, the Democratic Progressive Party (DPP), experienced dismal failure. The DPP lost several important counties and cities' mayors, including Kaohsiung where the DPP had dominated over 20 years. The substantial defeat for the DPP had significant impacts on Taiwan's politics (Templeman 2020). President Ing-wen Tsai announced her resignation as chairperson for the DPP. No one expected President Tsai to continue in office after the forthcoming presidential election in January 2020 (Sui 2018).

[1] To improve the data's explanation power, I omit offshore island counties' data, such as Jinmen, Matsu, Peng-hu and so on. In addition, I delete several observations because they have coding error.

However, President Tsai made a remarkable comeback in the presidential campaign in 2020. The DPP won the election by a landslide (Kuo 2020).

The dramatic change in results of the two elections is an important case in social science. To provide crucial information for future academic studies, policy making and election campaign consulting, we are going to discover new insights factors that affect the results.

Table 1 is to describe important variables in **result_2018.xlsx** and **result_2020.xlsx**.

Table 1:

Variable	Definition
county	The Chinese names of cities or counties
county_e	The English names of cities or counties
district	The Chinese names of districts or towns
district_e	The English names of districts or towns
TownID	The ID number of districts
Li	The Chinese names of lis
li_id	The ID number of lis
DPP	The number of votes of DPP candidates in the 2018 or 2020 elections
KMT	The number of votes of KMT candidates in the 2018 or 2020 elections
valid_vote	The number of valid votes in the 2018 or 2020 election

Questions:

1. Read **result_2018.xlsx** and **result_2020.xlsx** to assign them two objects called tw_2018 and tw_2020 in R. (5 points)

2. We attempt to create a plot to show the dramatic change in results of the two elections.

(1) Please calculate the Kuomintang (KMT) and DPP's percentage of votes at the national level in these two elections. (5 points)

kmt_rate is calculated by dividing KMT by valid_vote.
dpp_rate is calculated by dividing DPP by valid_vote.

(2) Please create a line plot with two lines to show changes in kmt_rate, and dpp_rate in these two elections (15 points).

3. We attempt to find the top 10 towns with the largest decrease in kmt_rate. Please (1) use group_by, summarise and other functions to calculate every town's kmt_rate in these two elections, (2) calculate all towns' gap of kmt_rate between the two elections, (3) arrange the descending order of all towns' gap of kmt_rate, (4) create a bar plot to show the top 10 towns with the largest decrease in kmt_rate. (20 points)

4. We attempt to find factors which are related to the dramatic change in results of the two elections. The zip file **popu_2018.zip** is 2018 li-level population data. Please use a loop to combine them into a dataframe object called popu_2018. (25 points)

Table 2 is to describe important variables in **popu_2018.zip:**

Table 2:

Variable	Definition
P_CNT	Li's population number in 2018
A15A64_CNT	Li's population number from age 15-64 in 2018
A60UP_CNT	Li's old population number in 2018
college	The number of population has college degree in lis in 2018
income_avg	The mean income of lis in 2017
income_mid	The median income of lis in 2017

5. Please merge tw_2018, tw_2020, and popu_2018 into a new dataframe

called election_2018_2020. (15 points)

6. Please use dplyr's mutate() to create the following variables in election_2018_2020: kmt_gap, college_rate, aged_rate, and working_rate. (10 points)

   kmt_gap is calculated by kmt_rate in 2020 - kmt_rate in 2018.
   college_rate is calculated by dividing college by P_CNT
   aged_rate is calculated by dividing A60UP_CNT by P_CNT
   working_rate is calculated by dividing A15A64_CNT by P_CNT

7. Please create 4 plots with regression lines to show the following four relationships: (1) college_rate (x) and kmt_gap (y), (2) aged_rate (x) and kmt_gap (y), (3) working_rate (x) and kmt_gap (y), and (4) income_mid (x) and kmt_gap (y). (15 points)

8. Please interpret the results of four plots. Which x has a tiny impact on y. Which x promotes the decrease in kmt_rate. Which x has positive effects on kmt_rate. (15 points)

9. A steeper line of a negative slope means the x (you find in Question 8) has a greater impact on kmt_rate's decrease. Please tell me how many towns' X slope is steeper than whole Taiwan's. (10 bonus points)

Section 2:

Question:

The data files **fake_tweets_election.xlsx** is misinformation tweets collected during and after the first 2020 US presidential debate. The debate took place on September 29, 2020. Many studies have demonstrated that misinformation is able to change citizens' voting behavior (Cantarella et al. 2020; Grinberg et al. 2019). We attempt to analyze this dataset in order to understand the misinformation characteristics in US elections.

Table 3 is to describe important variables in **fake_tweets_election.xlsx:**
Table 3:

Variable	Definition
content_id	Tweets' ID number
type	Regular tweets or retweets
text	The content of tweets
date	Tweets' created time

10. There are few observations before the debate. Please delete the observation on the date before September 27, 2020. (10 points)

11. Please create two new variables called biden and trump, If the misinformation tweets include Biden or Trump, please label 1 otherwise 0. (10 points)

12. We want to know who experienced larger misinformation's attacks. Please create and reshape your table in order to create a dodged side-to-side bar plot (similar to Week 5's 45th pages): Daily number of tweets which mentioned Trump and Biden. Then tell me Trump or Biden suffered from more misinformation attacks. (25 points)

13. Misinformation monitor centers usually use misinformation post time to

guess a set of fake news' creators. If they are almost posted at 9:00am – 5:00pm, we tend to argue that they are created by paid cyber warriors. Please create and reshape your table in order to create a two-line plot: hourly number of tweets which mentioned Trump and Biden. Then tell me Trump or Biden suffered from paid cyber warriors' misinformation attacks. (25 points)

Reference:

Cantarella, M., Fraccaroli, N. and Volpe, R., 2020. Does Fake News Affect Voting Behaviour?.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. and Lazer, D., 2019. Fake News on Twitter during the 2016 US Presidential Election. Science, 363(6425), pp.374-378.

Kuo, L., 2020. Taiwan Election: Tsai Ing-Wen Wins Landslide in Rebuke to China. The Guardian. January 11, 2020.

Templeman, K., 2020. How Taiwan Stands up to China. Journal of Democracy, 31(3), pp.85-99.

Sui, C., 2018. Taiwan's Political Earthquake: Does China Gain from Tsai Ing-wen's Losses? BBC News. November 26, 2018.