Data Appendix to Individual Charateristics' Influence on Public Transportation Planning

Iris Zhong

Contents

1 Raw data

1.1 Gwinnett County 2019 Referendum Dataset

Citation: Results—Gwinnett—Election Night Reporting. (n.d.). Retrieved March 11, 2020, from https://results.enr.clarityelections.com/GA/Gwinnett/94961/Web02.225391/#/
DOI: N/A
Date Downloaded: Mar 11, 2020
Filename: raw_data/vote_result.xls
Unit of observation: Precinct
Dates covered: Mar 19, 2019

1.1.1 To obtain a copy

Users can visit the website that displays election results at Gwinnett County at https://results.enr.clarityelections.com/GA/Gwinnett/94961/Web02.225391/#/ and choose the **Detail XLS** link at the bottom right corner.

The xls file contains three spreadsheets. I will be using the second and the third sheets.

1.1.2 Importable version

Filename: importable_data/vote_result_importable.xlsx

The raw dataset is hard to be imported to R directly because of the following reasons. First, it has three sheets. Second, the top two rows do not contain any useful information or should be incorporated to the next row. The file uses the extension .xls, which is incompatible with R. Therefore, an importable version of the dataset was created.

Here are the steps:

1. Open the original files in Excel.
2. Move the turnout rate for each precinct displayed in the second spreadsheet to the third spreadsheet.
3. Delete the first and the second spreadsheets.

4. Remove the top two rows of the third spreadsheet.
5. Rename the columns to reflect whether the vote was for or against the proposal.
6. Change the extension from .xls to .xlsx.

1.1.3 Variable descriptions

I cannot find any information that describes the variables in this dataset. Therefore, the description below is my understanding.

- **precinct:** The name of the precinct.
- **registered_voters:** The number of registered voters in the precinct.
- **total_votes:** The number of votes received in this referendum.
- **voter_turnout:** The percentage of voters who voted in this referendum. (*total_votes/reigstered_voters*)
- **election_day_yes:** The number of people who voted yes during the election day.
- **absentee_mail_yes:** The number of people who voted yes by mailing paper ballots prior to the election day.
- **advance_in_person_1_yes:** The number of people who voted yes prior to the election day.
- **advance_in_person_2_yes:** The number of people who voted yes prior to the election day. The difference between this variable from the previous one is not clear. My speculation is they record people voting on different days before election.
- **provisional_yes:** The number of people who voted yes but had questions in their eligibility.
- **votes_yes:** The number of people who voted yes in total. (the sum of the previous five variables)
- **election_day_no:** The number of people who voted no during the election day.
- **absentee_mail_no:** The number of people who voted no by mailing paper ballots prior to the election day.
- **advance_in_person_1_no:** The number of people who voted no prior to the election day.
- **advance_in_person_2_no:** The number of people who voted no prior to the election day. The difference between this variable from the previous one is not clear. My speculation is they record people voting on different days before election.
- **provisional_no:** The number of people who voted no but had questions in their eligibility.
- **votes_no:** The number of people who voted no in total. (the sum of the previous five variables)

1.1.4 Data import code and summary

```
vote_result <- read_excel("importable_data/vote_result_importable.xlsx",
    col_types = c("text", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric", "numeric", "numeric",
        "numeric"))
View(vote_result)
export_summary_table(dfSummary(vote_result))
```

1.2 Gwinnett County Census Data

Citation: U.S. Census Bureau. (2018). American Community Survey. https://data.census.gov/cedsci/table?d=ACS%205-Year%20Estimates%20Data%20Profiles&table=DP05&tid=ACSDP5Y2018.DP05&g=0500000US13135,13135.140000&hidePreview=true&t=Age%20and%20Sex%3AHousing%3AHousing%20Units%3ARace%20and%20Ethnicity
DOI: N/A
Date Downloaded: Mar 19, 2020
Filename: N/A
Unit of observation: Census tract
Dates covered: 2018 (5-year estimate)

1.2.1 To obtain a copy

Users can obtain a copy of the dataset from an R package `tidycensus`.

Citation: Walker, K., Eberwein, K., & Herman, M. (2020). tidycensus: Load US Census Boundary and Attribute Data as "tidyverse" and "sf"-Ready Data Frames (Version 0.9.6) [Computer software]. https://CRAN.R-project.org/package=tidycensus

Below are the steps to use `tidycensus` to obtain the data:

1. Install the package `tidycensus` in R.
2. If haven't, register to get an API key in order to download data from the package. The key can be acquired at http://api.census.gov/data/key_signup.html.
3. Load the key into R with the following code:
   `census_api_key(key, install = TRUE)`

4. Load the library in R. Execute the code chunk below to get the data frame `acs18`. `acs18` shows all of the variables present in ACS-5 2018 survey and their IDs.

```
library(tidycensus)
acs18 <- load_variables(2018, "acs5", cache = TRUE)
```

5. Search for desirable variables in acs18 and record their IDs. The selected variables are: population, median income, median age, white population, the number of people who work, the number of people who commute by car, the number of people who commute by public transportation, the number of people who commute by subway, the number of people who go to work by bike, the number of people who walk to work, the number of people who use other transportation means, and the number of people who work at home. Subway is a subcategory of public transportation, but since it is particularly important in this project, it is also selected. Besides, variables that reflect people's travel time to work are potentially useful.

```
cbdata <- get_acs(geography = "tract",
                  variables = c(total = "B01001_001",
                                medincome = "B19013_001",
                                medage = "B01002_001",
                                white = "B01001A_001",
                                transportation_total = "B08006_001",
                                car = "B08006_002",
                                public = "B08006_008",
                                subway = "B08006_011",
                                bike = "B08006_014",
                                walk = "B08006_015",
                                other_transport = "B08006_016",
                                no_transport = "B08006_017",
                                time_total = "B08012_001",
                                time_less_5 = "B08012_002",
                                time_5_9 = "B08012_003",
                                time_10_14 = "B08012_004",
                                time_15_19 = "B08012_005",
                                time_20_24 = "B08012_006",
                                time_25_29 = "B08012_007",
                                time_30_34 = "B08012_008",
                                time_35_39 = "B08012_009",
                                time_40_44 = "B08012_010",
                                time_45_59 = "B08012_011",
                                time_60_89 = "B08012_012",
                                time_more_90 = "B08012_013"),
```

```
                  state = "GA",
                  county = "Gwinnett",
                  year = 2018)
```

```
## Getting data from the 2014-2018 5-year ACS
```

The code above constructs a data frame called cbdata by calling the function `get_acs()`, which pulls the data from the American Community Survey. Inside the function, the unit of measurement is specified by the `geography` argument. In this case, select **tract** for census tract. Put all the chosen variables in the `variables` argument. Finally, address `state` (**GA**), `county` (**Gwinnett**), and `year` of survey (**2018**).

1.2.2 Data wrangling

The current data frame requires modifications. First, each row displays one variable from one tract. However, to make tract as the unit of measurement, each row should include all the variables of one tract. Second, instead of the actual number of people who are white, the percentage of white population is more informative. Similarly, the percentage of people who go to work by certain transportation should also be calculated. Finally, travel time data is mostly grouped by a 5-minute band, which is too detailed for this project. It will be categorized with wider range to reduce variables. After consolidation, the percentages will be calculated.

Here are the steps of data wrangling:

1. Remove the margin of error for each measurement (**moe**), because it is not useful in later analyses.

```
cbdata_moe <- cbdata %>%
  select (-moe)
```

2. Use `pivot_wider()` to transpose the data.

```
cbdata_wider <- cbdata_moe %>%
  pivot_wider(names_from = variable,
              values_from = estimate)
```

3. Calculate the percentage of white population (*white*/*total*).

```
cbdata_white <- cbdata_wider %>%
  mutate(white_pct = white / total) %>%
  select (-white)
```

4. Calculate the percentage of people using each transportation method.

```
cbdata_transport <- cbdata_white %>%
  mutate(car_pct = car / transportation_total,
         public_pct = public / transportation_total,
         subway_pct = subway / transportation_total,
         bike_pct = bike / transportation_total,
         walk_pct = walk / transportation_total,
         other_pct = other_transport / transportation_total,
         no_pct = no_transport / transportation_total) %>%
  select(-c(car, public, subway, bike, walk, other_transport, no_transport))
```

5. Combine the ranges of travel time data and calculate the percentages.

```
cbdata_tidy <- cbdata_transport %>%
  mutate(time_0_29_pct = (time_less_5 + time_10_14 + time_15_19 +
                            time_20_24 + time_25_29) / time_total,
         time_30_59_pct = (time_30_34 + time_35_39 + time_40_44 +
                             time_45_59) / time_total,
         time_60_89_pct = time_60_89 / time_total,
```

```
        time_more_90_pct = time_more_90 / time_total) %>%
  select(-c(time_less_5, time_5_9, time_10_14, time_15_19, time_20_24,
            time_25_29, time_30_34, time_35_39, time_40_44, time_45_59,
            time_60_89, time_more_90))
```

1.2.3 Variable descriptions

- **GEOID:** The geographic identifier of the census tract.
- **NAME:** The name of the census tract.
- **total:** The total population of the tract.
- **medage:** The median age of the population in the tract.
- **medincome:** The median income of the population in the tract.
- **white_pct:** The percentage of white population in the tract.
- **transportation_total:** The number of people who were sampled in the transportation survey.
- **car_pct:** The percentage of people who go to work by car, truck or van.
- **public_pct:** The percentage of people who go to work by public transportation (excluding taxi or cab).
- **subway_pct:** The percentage of people who go to work by subway or elevated.
- **bike_pct:** The percentage of people who go to work by bike.
- **walk_pct:** The percentage of people who go to work on foot.
- **other_pct:** The percentage of people who go to work by other transportation means such as taxi, cab and motorcycle.
- **no_pct:** The percentage of people who work at home (i.e. no transportation needed).
- **time_total:** The number of people who were sampled in the travel time to work survey.
- **time_0_29_pct:** The percentage of people who travel less than 30 minutes to work.
- **time_30_59_pct:** The percentage of people who travel between 30 and 59 minutes to work.
- **time_60_89_pct:** The percentage of people who travel between 60 and 89 minutes to work.
- **time_more_90_pct:** The percentage of people who travel more than 90 minutes to work.

1.2.4 Data summary

```
View(cbdata_tidy)
export_summary_table(dfSummary(cbdata_tidy))
```

1.3 2016 Presidential Election Data at Gwinnett County

Citation: MIT Election Data and Science Lab, 2018, U.S. President Precinct-Level Returns 2016, *Harvard Dataverse, V11,* UNF:6:hQyVqHW+vTFnAW2jYIOy/Q== [fileUNF]
DOI: doi:10.7910/DVN/LYWX3D
Date Downloaded: Mar 19, 2020
Filename(s): N/A
Unit of observation: Precinct
Dates covered: November 8, 2016

1.3.1 To obtain a copy

Users can obtain a copy of the dataset at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LYWX3D. Under the **Files** tab, select and click the download button of the first file **2016-precinct-president.tab**. The description of the variables can be found in the second file, **codebook-2016-precinct-president.md**.

However, the data file is not included in the raw_data folder, because it is too large to be uploaded to github.

1.3.2 Importable version

Filename: importable_data/election_result_importable.csv

The original file contains data from all the precincts in the United States. Therefore, it is too large to be loaded to R or github. Therefore, filtering is necessary before importing.

Here are the steps of filtering:

1. Open the file in excel.
2. Click on **Sort & Filter** button; click **Filter**.
3. Go to column H – **county_name** and click on the small triangle in cell H1.
4. Find and select **Gwinnett County** in the drop-down menu.
5. Copy and paste this subset of data into another file and put it into the importable data folder.

1.3.3 Data import and wrangling

```
election <- read_csv("importable_data/election_result_importable.csv")
```

The data needs more cleaning in R because firstly, it has unnecessary variables. Secondly, each row in this dataset represents the number of people voting for one particular candidate by one particular mode in a precinct. To make precinct as the unit of measurement, the ideal dataset will have the voting results for each candidate at one precinct in one row.

Here are the steps of data wrangling:

1. Select the useful variables: **precinct**, **candidate**, **votes**, **mode**. For more information about the removed variables, check out the codebook in the raw data folder.

```
election_variables <- election %>%
  select(precinct, candidate, votes, mode)
```

2. Summarize the number of votes for each candidate in a precinct. This is done by adding across different modes of votes (election day, absentee by mail, advance in person, and provisional). Then calculate the percentage of votes for each candidate.

```
election_mode <- election_variables %>%
  group_by(precinct, candidate) %>%
  summarize(votes = sum(votes))
```

3. Remove the write-in votes because they don't belong to any specific precincts; transpose the data; finally, calculate the precentage of votes for each candidate.

```
election_tidy <- as.data.frame(election_mode %>%
  filter(precinct != "Write-ins") %>%
  pivot_wider(names_from = candidate,
              values_from = votes) %>%
  mutate(total = `Donald Trump` + `Hillary Clinton` + `Gary Johnson`,
         trump_pct = `Donald Trump` / total,
         clinton_pct = `Hillary Clinton` / total,
         johnson_pct = `Gary Johnson` / total) %>%
  select(precinct, trump_pct, clinton_pct, johnson_pct))
```

1.3.4 Variable description

precinct: The name of the precinct.
trump_pct: The percentage of votes for Donald Trump in that precinct.
clinton_pct: The percentage of votes for Hillary Clinton in that precinct.
johnson_pct: The percentage of votes for Gary Johnson in that precinct.

1.3.5 Data summary

```
View(election_tidy)
export_summary_table(dfSummary(election_tidy))
```

1.4 Gwinnett County Precinct Map

Citation: District Maps | Gwinnett County. (n.d.). Retrieved March 27, 2020, from https://www.gwinnettcounty.com/web/gwinnett/Departments/Elections/ElectedOfficials/DistrictMaps
DOI: N/A
Date Downloaded: Mar 27, 2020
Filename(s): raw_data/precinct_map.pdf
Unit of observation: N/A
Dates covered: Unknown

1.4.1 To obtain a copy

Users can obtain a copy of the map by going to Gwinnett County website at https://www.gwinnettcounty.com/web/gwinnett/Home. In the navigation panel, select **Departments** > **Voting Registration and Election**. Click on **District Maps** in the blue panel. Click on the fifth result, **Precinct District (PDF)**. Download the PDF to get a copy.

1.5 Gwinnett County Census Tract Map

Citation: 2010 Census—Census Tract Reference Map. (2010). [Map].
DOI: N/A
Date Downloaded: Mar 27, 2020
Filename: raw_data/census_tract_map.pdf
Unit of observation: N/A
Dates covered: 2010

1.5.1 To obtain a copy

A copy of the map can be obtained with the following procedures:

1. Go to the website of the Census Bureau: https://www.census.gov/.
2. Type **2010 census tract reference maps** in the search bar.
3. Click on the first result.
4. Under Tract Maps, select **Georgia** as the state.
5. The returning page shows the folders of maps of all counties in Georgia. Find Gwinnett County: **c13135_gwinnett/**.
6. Open the folder; the first one **DC10CT_C13135_001.pdf** is the target file.
7. Open the file and save the PDF to get a copy.

1.6 Gwinnett Transportation Expansion Plan Map

Citation: Plan Documents | Gwinnett County. (n.d.). Retrieved March 28, 2020, from https://www.gwinnettcounty.com/web/gwinnett/departments/transportation/connectgwinnett/plandocuments
DOI: N/A
Date Downloaded: Mar 28, 2020
Filenames: raw_data/existing_map.pdf, raw_data/short_range_map.pdf, raw_data/mid_range_map.pdf, raw_data/long_range_1_map.pdf, raw_data/long_range_2_map.pdf
Unit of observation: N/A
Dates covered: N/A

1.6.1 To obtain a copy

A copy of the maps can be obtained at the official website of Gwinnett County with the following steps:

1. Go to https://www.gwinnettcounty.com/web/gwinnett/Home.
2. Select **Transportation** under the **Departments** tab.
3. Click on **Connect Gwinnett: Transit Plan** on the left panel.
4. Click on **Plan Documents** on the left panel.
5. Under the Reports title, there are detailed descriptions of the transit plan from short-term to long-term period. Under the **Maps** title, users can find the maps that are used in this research.
6. Click on the map that is interested and save the PDF file to get a copy.