Data Appendix to "What determines transit use in the Allston-Brighton neighborhood of Boston?"

Victoria Tse

Contents

```
## Warning: package 'summarytools' was built under R version 3.6.3
```

1 Appendix description

Your Data Appendix should begin with a brief statement explaining its purpose like the following one.

This Data Appendix documents the data used in "Paper Title". It was prepared in a Rmarkdown document that contains both the documentation and the R code used to prepare the data used in the final estimation. It also includes descriptive statistics for both the original data and the final dataset, with a discussion of any issues of note.

The datasets used directly by the final analysis are saved in `processed-data/` at the end of this file.

Note: this document structure will require you to re-run steps of your analysis multiple times. If your code takes a long time, please come talk with me about strategies to reduce run time or save earlier results.

2 Instructions for Use

This document includes instructions for how to create your Data Appendix. Outside of this section, instruction paragraphs are listed in *italics* (like the first paragraph above). Instructions should be removed before submission.

To start creating your own data appendix, follow these steps:

1. Replace the title and author in the section at the top of the file (called the YAML).
2. Commit your changes with a message like "customizing data appendix".
3. Delete this instruction section of the document.
4. Remove any other instructions in italics and examples from the completed sections of the document.

Remember that you will submit your assignment by committing and then pushing your versions to your repository. I encourage you to commit your changes often as you work, but there are three specific points at which you need to both submit and push changes, corresponding to course deadlines:

1. You must submit a version with the original data section completed by the Data Appendix 1 deadline. This will include the .Rmd file, the .pdf file, and the html data summary files stored in the output folder.
2. You must submit a version with all parts completed by the Data Appendix 2 deadline.

3. You must submit a final version of this document that is consistent with your final paper by the final project deadline.

While creating your data appendix, refer regularly to the assignment descriptions posted on Moodle.

A few tips:

- When creating a list like this one, be sure to put an empty line above the list. If you don't do this, your entries won't be formatted a list.
- Make sure you have empty lines above and below section and subsection headings.
- When creating numbered lists, you can number all items in your list with 1. Rmarkdown will number them sequentially when it creates your final document.

3 Raw data

*Each dataset you use will have its own documentation section. The next subsection in this document (Dataset description) is a template. You can copy this section and paste it into your document each time you need to add a section for a new dataset. Note that each line in the Dataset description section **must** end with two spaces.* This section documents the datasets used in this analysis.

3.1 Dataset description

Citation: Massachusetts Department of Transportation. (2012, June). Massachusetts Travel Survey. **DOI:** N/A **Date Downloaded:**
3/7/20 **Filename(s):**

- raw_data/filename.csv

- HH.txt
- PER.txt
- PLACE.txt
- VEH.txt **Unit of observation:**
  Each unit is an individual person **Dates covered:**
  6/14/10 - 10/24/11

3.1.1 To obtain a copy

To obtain a copy, be a part of the Smith College institution and ask Professor Sayre! Also, agree to the necessary privacy measures to ensure the safety of the data.

3.1.2 Importable version (if necessary)

N/A

Filename(s):

- raw_data/filename.csv

- HH.txt
- PER.txt
- PLACE.txt
- VEH.txt

3.1.3 Variable descriptions

- **RIBUS:** Use transit on a regular basis during the week

  1 - Yes

  2- No

- **HHVEH:** Number of household vehicles in working condition

  Range 1-7 8 - 8 or more

- **HISP:** Hipanic or Latino

  1- Yes

  2- No

- **RACE:** Ethnicity or race

  1 White alone

  2 Black or African American alone

  3 Indigenous/ Alaska Native alone

  4 Asian alone

  5 Native Hawaiian/ Pacific Islander alone

  6 Some other race alone

  7 Two or more races

- **OWN:** Homeownership

  1 Own/Mortgaged

  2 Renter

- **INCOME:** Household Income in 2009

  1 Less than $15,000

  2 $15,00-$24,999

  3 $25,000-$34,999

  4 $35,000- $49,999

    5 $50,000-$74,999

    6 $75,000-$99,999

    7 $100,00- $149,999

    8 $150,000 or more

- **AGE:** Age in years

- **LIC:** Valid Driver's License

  1- Yes

  2- No

- **TRANS:** Have a transit pass?

  1- Yes

  2- No

- **EMPLY:** Employed

  1- Yes

  2- No

- **WORKS:** Does this person work? [If EMPLY = 1]

  1- Yes

  2- No

- **WMODE:** Usual mode to work [WORKS=1]

  0- Works from home

  1- Walk

  2- Bike

  3- Auto/Van/Truck Driver

  4- Auto/Van/Truck Passenger

  5- Bus/ Public Transit

  6- Dial-A-Ride/ Paratransit

  7- Taxi

  8- Motorcycle Driver

  9- Motorcycle Passenger

- **EDUCA:** Level of Education

  1- Not a high school graduate, 12th or less (Includes very young children)

  2-High school diploma or GED

  3- Some college credit but no degree

  4- Associate or technical school degree

  5- Bachelor's or undergraduate degree

  6- Graduate Degree (Includes professional degrees, MD, DDs, JD)

- **ENROL:** Student

  1- Yes

  2- No

- **SMODE:** Usual mode to school [ENROL = 1]

    0- Homeschooled

    1- Walk

    2- Bike

    3- Auto/Van/Truck Driver

    4- Auto/Van/Truck Passenger

    5- Bus/ Public Transit

    6- Dial-A-Ride/ Paratransit

    7- Taxi

    8- Motorcycle Driver

    9- Motorcycle Passenger

- **BIKFA:** Bike facilities available at work/school? [ENROL = 1, WORK=1]

    1- Yes

    2- No

3.1.4 Data import code and summary

Once you've described the variables, enter an R chunk by selecting Code -> Insert Chunk, or Ctrl+Alt+I, give it a name to describe the dataset you are importing. After importing, export a dataframe summary using the command.

```r
hh_data <- read_csv(file.path("raw-data","mts-survey-data","HH.txt"), guess_max = 15000)
```

```
## Parsed with column specification:
## cols(
##    .default = col_double(),
##    MPO = col_character(),
##    O_RESTY = col_character(),
##    O_OWN = col_character(),
##    HCITY = col_character(),
##    HSTATE = col_character(),
##    HZIP = col_character(),
##    HTOWN = col_character(),
##    HPUMA10 = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```r
per_data <- read_csv(file.path("raw-data","mts-survey-data","PER.txt"), guess_max = 15000)
```

```
## Parsed with column specification:
## cols(
##    .default = col_double(),
##    O_PASST = col_character(),
##    O_WKSTAT = col_character(),
##    O_WMODE = col_character(),
##    WTRSB = col_character(),
```

```
##   O_FLEXP = col_character(),
##   O_EDUCA = col_character(),
##   O_SCHOL = col_character(),
##   O_SMODE = col_character(),
##   STRSB = col_character(),
##   O_TYPDY = col_character(),
##   O_TYPPL = col_character(),
##   O_NOGO = col_character(),
##   WCITY = col_character(),
##   WSTATE = col_character(),
##   WZIP = col_character(),
##   WTOWN = col_character(),
##   WPUMA10 = col_character(),
##   SCITY = col_character(),
##   SSTATE = col_character(),
##   SZIP = col_character()
##   # ... with 2 more columns
## )


## See spec(...) for full column specifications.
```

```
place_data <- read_csv(file.path("raw-data","mts-survey-data","PLACE.txt"), guess_max = 15000)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   O_TPURP = col_character(),
##   O_TPUR2 = col_logical(),
##   O_MODE = col_character(),
##   O_PRKTY = col_character(),
##   PRKLC = col_character(),
##   HOV = col_character(),
##   O_FARE = col_character(),
##   ROUTE = col_character(),
##   CITY = col_character(),
##   STATE = col_character(),
##   ZIP = col_character(),
##   TOWN = col_character(),
##   PUMA10 = col_character()
## )


## See spec(...) for full column specifications.


## Warning: 46 parsing failures.
##   row     col            expected       actual                                      file
## 36226 O_TPUR2 1/0/T/F/TRUE/FALSE VOLUNTEERING 'raw-data/mts-survey-data/PLACE.txt'
## 42538 O_PRKCS a double            DK/RF        'raw-data/mts-survey-data/PLACE.txt'
## 43711 O_PRKCS a double            DK/RF        'raw-data/mts-survey-data/PLACE.txt'
## 57718 O_PRKCS a double            DK/RF        'raw-data/mts-survey-data/PLACE.txt'
## 61029 O_PRKCS a double            DK/RF        'raw-data/mts-survey-data/PLACE.txt'
## .....  .......  ................. ............ .......................................
## See problems(...) for more details.
```

```
veh_data <- read_csv(file.path("raw-data","mts-survey-data","VEH.txt"), guess_max = 15000)
```

```
## Parsed with column specification:
## cols(
##   SAMPN = col_double(),
##   VEHNO = col_double(),
##   YEAR = col_double(),
##   MAKE = col_double(),
##   O_MAKE = col_character(),
##   MODEL = col_character(),
##   BODY = col_double(),
##   O_BODY = col_character(),
##   FUEL = col_double(),
##   O_FUEL = col_character(),
##   USER = col_double(),
##   CNTV = col_double(),
##   O_CNTV = col_character(),
##   HHWGT = col_double(),
##   EXPWGT = col_double()
## )
```

```
export_summary_table(dfSummary(dataset_name))
```

While it will make your resulting file long, you should not modify the chunk options to suppress printing of code and output. I would likely not include this in the documentation for an actual paper I was submitting, but including them here will let me read your code and the output message from R and may help identify data import concerns early in the process. Since these files will exist only electronically, their length is less of a concern. If you like to print out files to proofread and want me to help you shorten the printed versions, let me know. We can temporarily modify the chunk options for printing and restore them before you submit the assignment.

"`

4   Data Processing and Combination

This section should include a discussion of the processing and merging steps needed to create your basic data. The code to implement these steps should be included in chunks in this section. Once the final merged data has been created, you should use the dfSummary function again to summarize the data you will be using. You should also save a file containing all the objects you will use in your final analysis to the processed_data folder.

5   Analysis Variables

This section should include a description of all the variables that are used in your final analysis. At the end of the section, you should save all of these variables in the processed_data folder of your repository.

6   Discussion of Data

This section should include a discussion of any data patterns you notice based on the summaries created in the code above.