Sample Datasets Info Page - DACSS 601 Aug 2021

Sean Conway

8/8/2021

Datasets

This document summarizes the datasets that have been collected for use in DACSS 601 for the August 2021 session. All files can be found in the `_data` folder on the course blog. Note that some of these datasets require significant wrangling/cleaning. Also note that any .xls/.xlsx files may have multiple sheets, so it will be helpful to open these files in a spreadsheet software first, to examine the file you are reading in.

Hotel Bookings

This dataset contains hotel bookings from 2015-2017. Each row is an individual hotel booking. The file is named `hotel_bookings.csv`. Because the file format is .csv, we can use the function `read_csv()` from the `readr` package to read in the data to `R`.

```
hotels <- read_csv(here("_data","hotel_bookings.csv"))
hotels
```

```
## # A tibble: 119,390 x 32
##    hotel        is_canceled lead_time arrival_date_year arrival_date_month
##    <chr>              <dbl>     <dbl>             <dbl> <chr>
##  1 Resort Hotel           0       342              2015 July
##  2 Resort Hotel           0       737              2015 July
##  3 Resort Hotel           0         7              2015 July
##  4 Resort Hotel           0        13              2015 July
##  5 Resort Hotel           0        14              2015 July
##  6 Resort Hotel           0        14              2015 July
##  7 Resort Hotel           0         0              2015 July
##  8 Resort Hotel           0         9              2015 July
##  9 Resort Hotel           1        85              2015 July
## 10 Resort Hotel           1        75              2015 July
## # ... with 119,380 more rows, and 27 more variables:
## #   arrival_date_week_number <dbl>, arrival_date_day_of_month <dbl>,
## #   stays_in_weekend_nights <dbl>, stays_in_week_nights <dbl>, adults <dbl>,
## #   children <dbl>, babies <dbl>, meal <chr>, country <chr>,
## #   market_segment <chr>, distribution_channel <chr>, is_repeated_guest <dbl>,
## #   previous_cancellations <dbl>, previous_bookings_not_canceled <dbl>,
## #   reserved_room_type <chr>, assigned_room_type <chr>, booking_changes <dbl>,
## #   deposit_type <chr>, agent <chr>, company <chr>, days_in_waiting_list <dbl>,
## #   customer_type <chr>, adr <dbl>, required_car_parking_spaces <dbl>,
## #   total_of_special_requests <dbl>, reservation_status <chr>,
## #   reservation_status_date <date>
```

Source: https://www.kaggle.com/jessemostipak/hotel-booking-demand

Also see the link for a detailed key.

2019 New York City Air BnB Bookings

This dataset contains Air Bnb bookings from 2019 in New York City. Each row contains an individual Air Bnb listing, and each column contains information about it (e.g., number of reviews per month, price, data of last review). The file is named `AB_NYC_2019.csv`. Because the file format is .csv, we can use the function `read_csv()` from the `readr` package to read in the data to `R`.

```
air_bnb <- read_csv(here("_data","AB_NYC_2019.csv"))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   id = col_double(),
##   name = col_character(),
##   host_id = col_double(),
##   host_name = col_character(),
##   neighbourhood_group = col_character(),
##   neighbourhood = col_character(),
##   latitude = col_double(),
##   longitude = col_double(),
##   room_type = col_character(),
##   price = col_double(),
##   minimum_nights = col_double(),
##   number_of_reviews = col_double(),
##   last_review = col_date(format = ""),
##   reviews_per_month = col_double(),
##   calculated_host_listings_count = col_double(),
##   availability_365 = col_double()
## )
```

Source: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Also see the link for a detailed key.

2017 Austrailian Marriage Law

Data on public opinion of a proposed same sex marriage law in Australia in 2017. The file is called `australian_marriage_law_postal_survey_2017_-_response_final.xlsx`, so we can use the function `read_excel()` to read in the data. However, this dataset was designed as an Excel spreadsheet, and so will take some extra work to be read into `R`.

Source: https://www.abs.gov.au/ausstats/abs@.nsf/mf/1800.0

DOD Active Duty Marital Status

Count data on various demographic charasterics, notably marital status and child status, by pay grade, for multiple branches of the military (as well as DOD as a whole). This file is called

`ActiveDuty_MaritalStatus.xls`. However, this dataset was designed as an Excel spreadsheet, and so will take some extra work to be read into `R`.

Source: https://catalog.data.gov/dataset/active-duty-marital-status/resource/638cad03-b16c-48ac-8346-f858ff89d202

Public School Characteristics 2017-2018

Data on characteristics of every US public school from 2017-2018. File is called `Public_School_Characteristics_2017-18.c` Note that this file is fairly large, and if you aren't careful, you may encounter parsing errors when reading in the file.

Source: https://catalog.data.gov/dataset/public-school-characteristics-2017-18

2012 US Railroad Employment.

Data breaking down US railroad employment numbers in 2012 by state and county. File is `StateCounty2012.xls`.

Source: https://catalog.data.gov/dataset/total-railroad-employment-by-state-and-county-2012/resource/5a0b2831-23b9-4ce9-82e9-87a7d8f2c5d8

Organic Egg & Poultry Prices

Data on organic egg & poultry prices in the US from 2004-2013. File is `organiceggpoultry.xls`.

Source: https://www.ers.usda.gov/data-products/organic-prices.aspx