Lab 2 - Data Handling

Stats 20 - Summer 2020 - Section 1

Due Saturday July 11, 2020 by 11:59PM PDT via upload to CCLE

Contents

Introduction

This data is a subset of the original dataset used in the 2017 DataFest competition held at UCLA. The competition involved an analysis of Expedia data.

1. The Data

There are two datasets used in Lab 2, they are named

dataLab2.txt
destLab2.txt

A. Please read/import them into R (make them usable by R). The objects you create should have the names data and dest.

You may get warnings or errors when you try this, you can safely ignore them for this lab as long as you can demonstrate the following:

```
glimpse(data)
```

```
## Rows: 1,594,596
## Columns: 27
## $ date_time                <dttm> 2015-09-03 16:37:00, 2015-09-21 22:47:00...
## $ site_name                <chr> "EXPEDIA.COM", "EXPEDIA.COM", "EXPEDIA.CO...
## $ user_location_country    <chr> "UNITED STATES OF AMERICA", "UNITED STATE...
## $ user_location_region     <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA",...
## $ user_location_city       <chr> "BRENTWOOD", "BRENTWOOD", "BRENTWOOD", "B...
## $ user_location_latitude   <chr> "37.92381", "37.92381", "37.92381", "37.9...
## $ user_location_longitude  <chr> "-121.69622", "-121.69622", "-121.69622",...
## $ orig_destination_distance <chr> "5539.583", "5873.028", "5329.1407", "532...
## $ user_id                  <dbl> -2147479371, -2147479371, -2147479371, -2...
## $ is_mobile                <dbl> 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0,...
## $ is_package               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ channel                  <dbl> 324, 324, 541, 541, 541, 324, 293, 231, 5...
```

```
## $ srch_ci                   <date> 2016-05-19, 2016-05-12, 2015-11-29, 2016...
## $ srch_co                   <date> 2016-05-23, 2016-05-15, 2015-11-30, 2016...
## $ srch_adults_cnt           <dbl> 2, 2, 2, 2, 2, 2, 3, 2, 1, 1, 2, 2, 2, 3,...
## $ srch_children_cnt         <dbl> 0, 0, 0, 0, 0, 0, 3, 2, 0, 0, 0, 0, 0, 0,...
## $ srch_rm_cnt               <dbl> 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1,...
## $ srch_destination_id       <dbl> 24802975, 187569808, 5526772, 5526772, 55...
## $ hotel_country             <chr> "FRANCE", "FRANCE", "UNITED KINGDOM", "UN...
## $ is_booking                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ hotel_id                  <dbl> 17366540, 133800525, 439889, 41796865, 36...
## $ prop_is_branded           <dbl> 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1,...
## $ prop_starrating           <dbl> 3, 4, 4, 4, 4, 5, 4, 3, 3, 3, 2, 3, 4, 5,...
## $ distance_band             <chr> "M", "C", "M", "M", "M", "VC", "F", "VF",...
## $ hist_price_band           <chr> "L", "M", "H", "M", "H", "VH", "L", "H", ...
## $ popularity_band           <chr> "VH", "VH", "VH", "VH", "M", "H", "H", "H...
## $ cnt                       <dbl> 1, 2, 1, 1, 1, 1, 1, 5, 2, 1, 1, 1, 1, 2,...
```

and

```
glimpse(dest)
```

```
## Rows: 2,846
## Columns: 12
## $ srch_destination_id           <dbl> 8369, 8710, 8803, 8865, 8989, 9020, 9...
## $ srch_destination_name         <chr> "Abilene (and vicinity), Texas, Unite...
## $ srch_destination_type_id      <dbl> 464, 464, 464, 464, 464, 464, 464, 55...
## $ srch_destination_latitude     <dbl> 32.412406, 31.617903, 5.597013, 28.92...
## $ srch_destination_longitude    <dbl> -99.764760, -84.222600, -0.178161, -1...
## $ popular_naturalfeature_beach  <dbl> -2.201047, -2.189381, -2.005850, -1.7...
## $ popular_activity_dining       <dbl> -1.809095, -1.891543, -1.925625, -1.7...
## $ popular_social_bars           <dbl> -2.030768, -2.139417, -2.009178, -1.8...
## $ popular_shopping_shopping     <dbl> -1.969560, -2.025266, -2.133380, -2.0...
## $ popular_historical_historical <dbl> -2.150781, -2.189381, -2.139776, -2.1...
## $ popular_food_coffee           <dbl> -2.022664, -2.110232, -2.162235, -2.1...
## $ popular_field_business        <dbl> -2.046192, -2.018189, -1.992753, -2.2...
```

2. Cleaning/Verification

A. Turns out that the R object data **SHOULD** only have users from the UNITED STATES OF AMERICA and not from any other country.

If you determine that there are users from another country, they need to be removed from the data and a new version should be saved (please give it a new name like data2)

B. The dest data has 7 measures which start with the word "popular_" and they are a bit difficult for the typical person (think, a future supervisor) to interpret because they have ranges such as a minimum of -2.35912 and a maximum of -0.9772187.

To make each of these 7 more interpretable to the average person, please first compute its mean and then subtract this computed mean from each value that was used in the computing of the mean.

So for example, St. Pete Beach, Florida, United States of America has a score of -0.9772187 for popular_naturalfeature_beach and the mean of that measure is -2.045672. If we subtract (-0.9772187-2.045672) the result will be 1.0684533. We can save the result and more important, now values less than zero can be interpreted as "below average", a value at or near zero is "about average" and those values above zero are "above average".

Please show how to create these 7 new variables from the existing 7 popular, you can name them popular1, popular2… to popular7 if you want (or you can make them more descriptive), Save the srch_ variables plus your 7 new variables to a new data set named dest2 and run a summary(dest2).

3. Joining/Merging

In the original competition, many teams joined or merged data and dest (in your case, data2 and dest2) to create a richer dataset for analysis.

A. Please examine the variables in both datasets and determine which variable should be used to match the datasets together? (Hint it starts with srch_)

B. Please join or merge (they are the same to me) data2 and dest2 using the variable you identified in A as your by = variable.

C. (challenge yourself) This is totally optional, but try it, at least conceptually. How would you join/merge the third dataset named citiwiki.csv to your result in B? Which variable would you use as a by variable?

4. What kinds of questions?

This is something to discuss with your TA or classmates and then write up your thoughts (a sentence or two is fine, Stats is STEM). I'd like to see two questions from each of you, but I'll be happy with one.

The original descriptions of the data are at the end of this page, they may help you formulate questions that could be answered using the data.

5. What to Turn In

A .Rmd file following the format given to you in the next link on CCLE AND your knit result (.html only). Both of these should be uploaded to CCLE before the due date.

B. The programming code (when needed) and any answers (when requested) to the questions 1A, 2A, 2B, 3A, 3B, and 4. If you want to include the challenge in 3C, please go ahead, it won't be graded but I will look at them to see what you developed.

6. The original descriptions

Each row represents a site user's interaction with an Expedia site ("event"), which is either clicking on or booking a hotel, during a session.

A session is defined as a series of user actions that are not interrupted for more than 30 minutes.

Name of field	Description
date_time	Timestamp time, date, year in user's local time
site_name	Expedia point of sale, i.e., the site (Expedia.com is the US, Expedia.ca is Canada, Expedia.
user_location_country	Country the customer is located at the time of interaction with Expedia sites
user_location_region	Region the customer is located (State in the US, greater metro area for Europe, etc but the
user_location_city	City the customer is located at the time of interaction with Expedia sites
user_location_latitude	Latitude of the city where the customer is located, i.e. the city center lat-long
user_location_longitude	Longitude of the city where the customer is located, i.e. the city center lat-long
orig_destination_distance	Physical distance between a hotel and a customer at the time of search, in miles
user_id	ID of a user (unique)
is_mobile	1 when a user connected from a mobile device (whether app or not), 0 otherwise
is_package	1 if the click/booking was generated as a part of a package search (i.e. a hotel combined w
channel	ID of a marketing channel (anonymized), i.e. how a user may have arrived at an Expedia sit
srch_ci	Check-in date specified in the customer search
srch_co	Check-out date specified in the customer search
srch_adults_cnt	The number of adults specified to occupy the hotel room
srch_children_cnt	The number of (optional) children specified to occupy the hotel room
srch_rm_cnt	The number of hotel rooms specified in the search

Name of field	Description
srch_destination_id	ID of the destination where the hotel search was performed
hotel_country	Country the hotel is located in
is_booking	1 if a booking, 0 if a click
hotel_id	ID of the hotel (there is no particular significance to the ID number)
prop_is_branded	1 if the hotel is part of a major hotel chain (Hilton, Marriott, Sheraton, etc.), 0 if it is not
prop_starrating	The star rating of the hotel, from 1 to 5, in increments of 1. A 0 indicates the property has
distance_band	Banded distance of a hotel from the search destination center relative to other hotels in th
hist_price_band	Banded historical purchase price of a hotel relative to other hotels in the same destination
popularity_band	Banded hotel popularity relative to other hotels in the same destination, i.e. how often it i
cnt	Number of similar events (clicks or bookings) in the context of the same user session
Name of field	Description
srch_destination_id	ID of the destination where the hotel search was performed
srch_destination_name	Name of the destination where the hotel search was performed
srch_destination_type_id	Type of destination
srch_destination_latitude	Latitude of destination
srch_destination_longitude	Longitude of destination
popular_*	Popularity scores of travel related facets of destinations. The interpretation of the popular