

# Preparing the trip data

*Georgie Knight*

*19 August, 2016*

We'll now combine our data file on station data with the station information

## Dplyr and tidyr

Load the *dplyr* and *tidyr* packages which will help us wrangle the data:

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library("tidyr")
```

## Loading up the data

Load up our status data and station information:

```
trip_read <- read.csv("201508_trip_data.csv")
trip_df   <- data.frame(trip_read)
trip      <- dplyr::tbl_df(trip_df)
glimpse(trip)
```

```
## Observations: 354,152
## Variables: 11
## $ Trip.ID      <int> 913460, 913459, 913455, 913454, 913453, 913452...
## $ Duration     <int> 765, 1036, 307, 409, 789, 293, 896, 255, 126, ...
## $ Start.Date   <fctr> 8/31/2015 23:26, 8/31/2015 23:11, 8/31/2015 2...
## $ Start.Station <fctr> Harry Bridges Plaza (Ferry Building), San Ant...
## $ Start.Terminal <int> 50, 31, 47, 10, 51, 68, 51, 60, 56, 47, 60, 67...
## $ End.Date     <fctr> 8/31/2015 23:39, 8/31/2015 23:28, 8/31/2015 2...
## $ End.Station  <fctr> San Francisco Caltrain (Townsend at 4th), Mou...
## $ End.Terminal <int> 70, 27, 64, 8, 60, 70, 60, 74, 55, 66, 77, 70,...
## $ Bike..      <int> 288, 35, 468, 68, 487, 538, 363, 470, 439, 472...
## $ Subscriber.Type <fctr> Subscriber, Subscriber, Subscriber, Subscribe...
## $ Zip.Code     <fctr> 2139, 95032, 94107, 95113, 9069, 94118, 92562...
```

We note the following:

*-Trip ID: numeric ID of bike trip -Duration: time of trip in seconds -Start Date: start date of trip with date and time, in PST -Start Station: station name of start station -Start Terminal: numeric reference for start station -End Date: end date of trip with date and time, in PST -End Station: station name for end station -End Terminal: numeric reference for end station -Bike #: ID of bike used -Subscription Type: Subscriber = annual or 30-day member; Customer = 24-hour or 3-day member -Zip Code: Home zip code of subscriber (customers can choose to manually enter zip at kiosk however data is unreliable)*

Combine with the station location data and create a date column

```
station_read <- read.csv("station.csv")
station_df <- data.frame(station_read)
station <- dplyr::tbl_df(station_df)
station_short <-select(station, station_id, lat, long, landmark)

trip <- trip %>%
  mutate(Date = as.Date(trip$Start.Date[1], format = "%m/%d/%Y")) %>%
  left_join(station_short, by = c("Start.Terminal" = "station_id")) %>%
  mutate(start_lat =lat, start_long = long) %>%
  select(-lat, -long)

station_short <-select(station, station_id, lat, long)

trip <- trip %>%
  left_join(station_short, by = c("End.Terminal" = "station_id")) %>%
  mutate(end_lat =lat, end_long = long) %>%
  select(-lat, -long)

glimpse(trip)
```

```
## Observations: 354,152
## Variables: 17
## $ Trip.ID      <int> 913460, 913459, 913455, 913454, 913453, 913452...
## $ Duration     <int> 765, 1036, 307, 409, 789, 293, 896, 255, 126, ...
## $ Start.Date   <fctr> 8/31/2015 23:26, 8/31/2015 23:11, 8/31/2015 2...
## $ Start.Station <fctr> Harry Bridges Plaza (Ferry Building), San Ant...
## $ Start.Terminal <int> 50, 31, 47, 10, 51, 68, 51, 60, 56, 47, 60, 67...
## $ End.Date     <fctr> 8/31/2015 23:39, 8/31/2015 23:28, 8/31/2015 2...
## $ End.Station  <fctr> San Francisco Caltrain (Townsend at 4th), Mou...
## $ End.Terminal <int> 70, 27, 64, 8, 60, 70, 60, 74, 55, 66, 77, 70,...
## $ Bike..      <int> 288, 35, 468, 68, 487, 538, 363, 470, 439, 472...
## $ Subscriber.Type <fctr> Subscriber, Subscriber, Subscriber, Subscribe...
## $ Zip.Code     <fctr> 2139, 95032, 94107, 95113, 9069, 94118, 92562...
## $ Date        <date> 2015-08-31, 2015-08-31, 2015-08-31, 2015-08-3...
## $ landmark     <fctr> San Francisco, Mountain View, San Francisco, ...
## $ start_lat    <dbl> 37.79539, 37.40044, 37.78898, 37.33739, 37.791...
## $ start_long   <dbl> -122.3942, -122.1083, -122.4035, -121.8870, -1...
## $ end_lat      <dbl> 37.77662, 37.38922, 37.78226, 37.33017, 37.804...
## $ end_long     <dbl> -122.3953, -122.0819, -122.3927, -121.8858, -1...
```

Note we have only added the landmark once. We presume that start and end landamrks are the same.

```

weather_read <- read.csv("weather.csv")
weather_df <- data.frame(weather_read)
weather <- dplyr::tbl_df(weather_df)
weather <- weather %>%
  select(-X) %>%
  mutate(Date = as.Date(trip$Start.Date[1], format = "%m/%d/%Y"))
glimpse(weather)

```

```

## Observations: 1,825
## Variables: 25
## $ Max.TemperatureF      <int> 83, 72, 76, 74, 72, 72, 72, 68, 72, ...
## $ Mean.TemperatureF     <int> 70, 66, 69, 68, 66, 66, 66, 64, 65, ...
## $ Min.TemperatureF      <int> 57, 60, 61, 61, 60, 60, 60, 59, 57, ...
## $ Max.Dew.PointF        <int> 58, 58, 57, 57, 57, 55, 54, 52, 56, ...
## $ MeanDew.PointF        <int> 56, 57, 56, 57, 56, 54, 53, 51, 53, ...
## $ Min.DewpointF         <int> 52, 55, 55, 56, 54, 52, 45, 50, 52, ...
## $ Max.Humidity          <int> 86, 84, 84, 84, 84, 84, 78, 72, 87, ...
## $ Mean.Humidity         <int> 64, 73, 69, 71, 71, 69, 66, 64, 72, ...
## $ Min.Humidity          <int> 42, 61, 53, 57, 57, 53, 53, 55, 57, ...
## $ Max.Sea.Level.PressureIn <dbl> 29.86, 29.87, 29.81, 29.81, 29.92, 2...
## $ Mean.Sea.Level.PressureIn <dbl> 29.82, 29.82, 29.76, 29.76, 29.87, 2...
## $ Min.Sea.Level.PressureIn <dbl> 29.76, 29.79, 29.72, 29.72, 29.81, 2...
## $ Max.VisibilityMiles    <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
## $ Mean.VisibilityMiles   <int> 10, 10, 10, 10, 9, 9, 10, 10, 10, 9,...
## $ Min.VisibilityMiles    <int> 8, 7, 10, 8, 7, 7, 10, 10, 10, 5, 2,...
## $ Max.Wind.SpeedMPH      <int> 16, 21, 21, 22, 18, 17, 18, 18, 17, ...
## $ Mean.Wind.SpeedMPH     <int> 7, 8, 8, 8, 8, 9, 10, 12, 7, 5, 7, 8...
## $ Max.Gust.SpeedMPH      <int> 20, NA, 24, 25, 32, 30, 28, 22, 21, ...
## $ PrecipitationIn        <fctr> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ CloudCover             <int> 0, 5, 4, 5, 4, 4, 3, 4, 5, 1, 2, 1, ...
## $ Events                 <fctr> , , , , , , , , , , , , , , Rai...
## $ WindDirDegrees         <int> 290, 290, 276, 301, 309, 290, 293, 2...
## $ Zip                   <int> 94107, 94107, 94107, 94107, 94107, 9...
## $ landmark               <fctr> San Francisco, San Francisco, San F...
## $ Date                   <date> 2015-08-31, 2015-08-31, 2015-08-31,...

```

```

trip <-semi_join(trip, weather, by="Date")
glimpse(trip)

```

```

## Observations: 354,152
## Variables: 17
## $ Trip.ID              <int> 913460, 913459, 913455, 913454, 913453, 913452...
## $ Duration             <int> 765, 1036, 307, 409, 789, 293, 896, 255, 126, ...
## $ Start.Date           <fctr> 8/31/2015 23:26, 8/31/2015 23:11, 8/31/2015 2...
## $ Start.Station        <fctr> Harry Bridges Plaza (Ferry Building), San Ant...
## $ Start.Terminal       <int> 50, 31, 47, 10, 51, 68, 51, 60, 56, 47, 60, 67...
## $ End.Date             <fctr> 8/31/2015 23:39, 8/31/2015 23:28, 8/31/2015 2...
## $ End.Station          <fctr> San Francisco Caltrain (Townsend at 4th), Mou...
## $ End.Terminal         <int> 70, 27, 64, 8, 60, 70, 60, 74, 55, 66, 77, 70,...
## $ Bike..               <int> 288, 35, 468, 68, 487, 538, 363, 470, 439, 472...
## $ Subscriber.Type      <fctr> Subscriber, Subscriber, Subscriber, Subscribe...

```

```
## $ Zip.Code      <fctr> 2139, 95032, 94107, 95113, 9069, 94118, 92562...
## $ Date          <date> 2015-08-31, 2015-08-31, 2015-08-31, 2015-08-3...
## $ landmark      <fctr> San Francisco, Mountain View, San Francisco, ...
## $ start_lat     <dbl> 37.79539, 37.40044, 37.78898, 37.33739, 37.791...
## $ start_long    <dbl> -122.3942, -122.1083, -122.4035, -121.8870, -1...
## $ end_lat       <dbl> 37.77662, 37.38922, 37.78226, 37.33017, 37.804...
## $ end_long      <dbl> -122.3953, -122.0819, -122.3927, -121.8858, -1...
```

```
write.csv(trip, file="trip_full.csv")
```