

Data preprocessing notebook for Biketown PDX

Raymond Ford (raymond.anthony.ford@gmail.com)

2020-05-27

Contents

Bringing in the raw data	1
Cleaning the data	2
Final cleaning	9
Known issues	11
Future work	11
Session information	11

This notebook shows how the raw data was brought into R, transformed into a more usable form, and finally an interim data set is written to file. An in depth exploratory data analysis will be performed in a later notebook. It should be noted that our target variable is `Duration` since that seems to be where most of the revenue comes from.

Bringing in the raw data

We first begin by bringing our raw data into R. Since all of the files we need to bring in are stored in `/data/raw/` and no other files are located in that directory we will do the following:

1. Set our working directory to the folder containing the data for each of the 45 months.
2. Obtain the names of each of the files in `/data/raw/` ending in `.csv`.
3. Use these names to create a `list()` object from the files found in #1.
4. Make this `list()` object into a R dataframe object.
5. Create a data frame `dat.raw` which we will use for the remainder of this notebook.

```
setwd("~/GoogleDrive/pdxbikes/pdxbikes/data/raw/")
file.names <- list.files(pattern="*.csv")
make.df <- lapply(file.names, read.csv)
rm(file.names) # We do not need these anymore, so it's best to remove them.
dat.raw <- do.call(rbind.data.frame, make.df)
rm(make.df)
```

Since `/data/raw/` is hard coded, we obtain the above warning from R. We can move past this warning and continue on in our quest to better understand the data. Setting a relative path to remove this warning is something that we will work on at a later date.

Cleaning the data

Now we will perform some basic transformations to the data frame that we created in the previous section. We begin by first looking at the structure of the data frame using the `str()` function.

```
str(dat.raw)
```

```
## 'data.frame': 1236392 obs. of 19 variables:
## $ RouteID : int 1282087 1282113 1282118 1282120 1282123 1282125 1282127 1282131 1282134 1282137 ...
## $ PaymentPlan : Factor w/ 3 levels "Casual","Subscriber",...: 1 2 2 2 2 1 1 2 1 1 ...
## $ StartHub : Factor w/ 212 levels "", "N Failing at Williams",...: 23 1 43 1 1 55 55 31 15 15 ...
## $ StartLatitude : num 45.5 45.5 45.5 45.5 NA ...
## $ StartLongitude : num -123 -123 -123 -123 NA ...
## $ StartDate : Factor w/ 1353 levels "7/19/2016","7/20/2016",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ StartTime : Factor w/ 1441 levels "0:00","0:01",...: 138 144 146 147 148 148 148 149 150 151 ...
## $ EndHub : Factor w/ 212 levels "", "N Failing at Williams",...: 1 1 85 54 1 85 85 49 57 57 ...
## $ EndLatitude : num 45.5 45.5 45.5 45.5 NA ...
## $ EndLongitude : num -123 -123 -123 -123 NA ...
## $ EndDate : Factor w/ 1360 levels "", "7/19/2016",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ EndTime : Factor w/ 1441 levels "", "0:00","0:01",...: 166 165 275 155 155 179 178 164 167 167 ...
## $ TripType : Factor w/ 5 levels "", "commute","errand",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ BikeID : int 6083 6238 7271 6875 7160 6590 6582 6534 6573 6559 ...
## $ BikeName : Factor w/ 2152 levels "", "0001 BIKETOWN",...: 440 732 336 72 124 33 53 871 795 3 ...
## $ Distance_Miles : num 1.19 2.95 13.46 0.53 0 ...
## $ Duration : Factor w/ 17841 levels "", "0:01:00","0:01:01",...: 1488 1066 4862 263 260 1665 1665 1665 1665 1665 ...
## $ RentalAccessPath: Factor w/ 8 levels "admin","keypad",...: 2 3 3 2 2 2 2 2 2 2 ...
## $ MultipleRental : logi FALSE FALSE FALSE FALSE FALSE FALSE ...
```

To further see this data set at a snapshot we will use the `summary()` function.

```
summary(dat.raw)
```

```
##      RouteID      PaymentPlan
## Min.   : 1282087  Casual      :711282
## 1st Qu.: 3591394  Subscriber:521022
## Median : 7220719           : 4088
## Mean   : 7226320
## 3rd Qu.:11064799
## Max.   :13201070
## NA's   :4088
##
##           StartHub      StartLatitude      StartLongitude
##           :378982      Min.   :45.30      Min.   :-123.14
## SW Salmon at Waterfront Park : 27814      1st Qu.:45.52      1st Qu.: -122.68
## SW Moody at Aerial Tram Terminal: 20167      Median :45.52      Median : -122.67
## SW River at Montgomery       : 17924      Mean    :45.52      Mean    : -122.67
## NW Everett at 22nd           : 17776      3rd Qu.:45.53      3rd Qu.: -122.66
## NW 13th at Marshall          : 17093      Max.    :45.73      Max.    : 67.18
## (Other)                      :756636      NA's    :4680      NA's    :4680
##
##      StartDate      StartTime
## 5/27/2018: 4792      : 4088
##           : 4088      17:06 : 2223
## 5/26/2018: 3762      17:09 : 2205
## 5/28/2018: 3731      17:08 : 2189
## 5/19/2018: 3725      17:07 : 2182
## 5/12/2018: 3591      17:11 : 2175
## (Other) :1212703      (Other):1221330
```

```
##                               EndHub      EndLatitude  EndLongitude
##                               :319512  Min.      :34.26  Min.      :-134.4
## SW Salmon at Waterfront Park : 31329  1st Qu.:45.52  1st Qu.: -122.7
## NW 13th at Marshall          : 22421  Median   :45.52  Median   :-122.7
## SW Moody at Aerial Tram Terminal: 22373  Mean     :45.52  Mean     :-122.7
## NW Couch at 11th             : 21178  3rd Qu.:45.53  3rd Qu.: -122.7
## SW River at Montgomery       : 20746  Max.     :49.16  Max.     : 45.5
## (Other)                      :798833  NA's     :4730  NA's     :4730
##      EndDate      EndTime      TripType      BikeID
## 5/27/2018: 4804      : 4394      :1235063  Min.     : 5986
##      : 4394  17:27 : 2255  commute : 406  1st Qu.: 6302
## 5/26/2018: 3771  17:29 : 2191  errand  : 172  Median : 6574
## 5/28/2018: 3759  17:28 : 2155  recreation: 663  Mean   : 6960
## 5/19/2018: 3689  17:25 : 2142  work    : 88   3rd Qu.: 7155
## 5/20/2018: 3586  17:18 : 2135      Max.   :35537
## (Other) :1212389 (Other):1221120      NA's   :4088
##      BikeName      Distance_Miles      Duration
##      : 4873  Min.      : 0.000      : 8662
## 0090 BIKETOWN: 1562  1st Qu.: 0.720  0:06:37: 1225
## 0153 BIKETOWN: 1547  Median   : 1.300  0:06:39: 1216
## 0139 BIKETOWN: 1521  Mean     : 2.002  0:05:41: 1212
## 0646 BIKETOWN: 1518  3rd Qu.: 2.390  0:05:36: 1211
## 0584 BIKETOWN: 1508  Max.     :15527.180  0:06:07: 1210
## (Other) :1223863  NA's     :4088  (Other):1221656
##      RentalAccessPath MultipleRental
## keypad      :915062  Mode :logical
## mobile      :185136  FALSE:1120063
## keypad_rfid_card :126201 TRUE :112241
##      : 4088  NA's :4088
## keypad_phone_number: 2829
## web      : 2205
## (Other)  : 871
```

Using information obtained in the above two outputs—and combined with our wanting to understand trip duration since that determines the amount of revenue—we will remove several columns from the data. We remove the columns for `StartHub`, `EndHub`, `TripType` (due to the sheer number of missing values), `BikeID`, and `BikeName` from the data with the following code.

```
dat.raw <- dat.raw[, -c(3, 8, 13:15)]
```

With these columns removed, we now take another look at the data using the `summary()` command.

```
summary(dat.raw)
```

```
##      RouteID      PaymentPlan      StartLatitude  StartLongitude
## Min.      : 1282087  Casual      :711282  Min.      :45.30  Min.      :-123.14
## 1st Qu.: 3591394  Subscriber:521022  1st Qu.:45.52  1st Qu.: -122.68
## Median : 7220719      : 4088  Median   :45.52  Median   :-122.67
## Mean   : 7226320      Mean     :45.52  Mean     :-122.67
## 3rd Qu.:11064799      3rd Qu.:45.53  3rd Qu.: -122.66
## Max.   :13201070      Max.     :45.73  Max.     : 67.18
## NA's   :4088      NA's     :4680  NA's     :4680
##      StartDate      StartTime      EndLatitude  EndLongitude
## 5/27/2018: 4792      : 4088  Min.      :34.26  Min.      :-134.4
##      : 4088  17:06 : 2223  1st Qu.:45.52  1st Qu.: -122.7
## 5/26/2018: 3762  17:09 : 2205  Median   :45.52  Median   :-122.7
```

```
## 5/28/2018: 3731 17:08 : 2189 Mean :45.52 Mean :-122.7
## 5/19/2018: 3725 17:07 : 2182 3rd Qu.:45.53 3rd Qu.: -122.7
## 5/12/2018: 3591 17:11 : 2175 Max. :49.16 Max. : 45.5
## (Other) :1212703 (Other):1221330 NA's :4730 NA's :4730
##      EndDate      EndTime      Distance_Miles      Duration
## 5/27/2018: 4804 : 4394 Min. : 0.000 : 8662
## : 4394 17:27 : 2255 1st Qu.: 0.720 0:06:37: 1225
## 5/26/2018: 3771 17:29 : 2191 Median : 1.300 0:06:39: 1216
## 5/28/2018: 3759 17:28 : 2155 Mean : 2.002 0:05:41: 1212
## 5/19/2018: 3689 17:25 : 2142 3rd Qu.: 2.390 0:05:36: 1211
## 5/20/2018: 3586 17:18 : 2135 Max. :15527.180 0:06:07: 1210
## (Other) :1212389 (Other):1221120 NA's :4088 (Other):1221656
##      RentalAccessPath MultipleRental
## keypad :915062 Mode :logical
## mobile :185136 FALSE:1120063
## keypad_rfid_card :126201 TRUE :112241
## : 4088 NA's :4088
## keypad_phone_number: 2829
## web : 2205
## (Other) : 871
```

Since Duration is our target variable, and we see missing values, we will remove the observations containing missing values for Duration and output the updated summary with the following code.

```
dat.raw <- dat.raw[-which(dat.raw$Duration==""), ]
summary(dat.raw)
```

```
##      RouteID      PaymentPlan      StartLatitude      StartLongitude
## Min. : 1282087 Casual :708276 Min. :45.30 Min. : -123.14
## 1st Qu.: 3585289 Subscriber:519454 1st Qu.:45.52 1st Qu.: -122.68
## Median : 7231950 : 0 Median :45.52 Median : -122.68
## Mean : 7230198 Mean :45.52 Mean : -122.67
## 3rd Qu.:11069075 3rd Qu.:45.53 3rd Qu.: -122.66
## Max. :13201070 Max. :45.73 Max. : 67.18
## NA's :509 NA's :509
##      StartDate      StartTime      EndLatitude      EndLongitude
## 5/27/2018: 4740 17:06 : 2218 Min. :34.26 Min. : -134.4
## 5/26/2018: 3703 17:09 : 2200 1st Qu.:45.52 1st Qu.: -122.7
## 5/28/2018: 3690 17:08 : 2183 Median :45.52 Median : -122.7
## 5/19/2018: 3678 17:07 : 2176 Mean :45.52 Mean : -122.7
## 5/12/2018: 3546 17:11 : 2165 3rd Qu.:45.53 3rd Qu.: -122.7
## 5/20/2018: 3501 17:12 : 2138 Max. :49.16 Max. : 45.5
## (Other) :1204872 (Other):1214650 NA's :559 NA's :559
##      EndDate      EndTime      Distance_Miles      Duration
## 5/27/2018: 4752 17:27 : 2246 Min. : 0.00 0:06:37: 1225
## 5/28/2018: 3718 17:29 : 2186 1st Qu.: 0.72 0:06:39: 1216
## 5/26/2018: 3712 17:28 : 2146 Median : 1.30 0:05:41: 1212
## 5/19/2018: 3642 17:25 : 2134 Mean : 2.00 0:05:36: 1211
## 5/20/2018: 3532 17:18 : 2128 3rd Qu.: 2.39 0:06:07: 1210
## 5/12/2018: 3521 17:31 : 2121 Max. :15527.18 0:05:37: 1199
## (Other) :1204853 (Other):1214769 (Other):1220457
##      RentalAccessPath MultipleRental
## keypad :911536 Mode :logical
## mobile :184476 FALSE:1115832
## keypad_rfid_card :125863 TRUE :111898
```

```
## keypad_phone_number: 2802
## web : 2186
## unknown : 832
## (Other) : 35
```

Next we notice that `StartLongitude` has some very high values compared to where the Portland, OR metropolitan area is located, so we will take a closer look at these observations.

```
large.start.longitude <- which(dat.raw$StartLongitude >= -122)
dat.raw[large.start.longitude, ]
```

```
##      RouteID PaymentPlan StartLatitude StartLongitude StartDate StartTime
## 256402 3124476      Casual    45.30201    -121.74448 6/7/2017    12:59
## 894634 10753147 Subscriber    45.51789      67.18144 2/7/2019      7:51
## 988408 11468802 Subscriber    45.56278     33.92405 5/23/2019    13:01
##      EndLatitude EndLongitude   EndDate EndTime Distance_Miles Duration
## 256402    45.53087   -122.6659 6/7/2017   13:11      5248.97 0:12:02
## 894634    45.51750   -122.6926 2/7/2019    7:53      6125.21 0:02:24
## 988408    45.56269   -122.6750 5/23/2019   13:44      5990.42 0:43:25
##      RentalAccessPath MultipleRental
## 256402          keypad          FALSE
## 894634          mobile          FALSE
## 988408 keypad_rfid_card          FALSE
```

In the above output we have identified three problematic observations/trips: 256402, 894634, and 988408. Looking at these trips individually we note the following:

1. **256402:** The distance traveled in the given time frame is completely unreasonable, even if the trip based on starting and ending location are both within the Portland metro area. That is approximately 5200 miles in only 12 minutes. We will remove this observation from the data.
2. **894634:** This observation did not begin in the Portland area, and it's distance when compared to time is completely unreasonable. Again, we will remove this observation.
3. **988408:** Again this has the same issues as 2. We will also remove this observation.

We remove these observations with the following code and output the summary of our updated data set.

```
dat.raw <- dat.raw[-large.start.longitude, ]
rm(large.start.longitude) # We do not need this anymore.
summary(dat.raw)
```

```
##      RouteID      PaymentPlan StartLatitude StartLongitude
## Min.   : 1282087      Casual   :708275    Min.   :45.35    Min.   : -123.1
## 1st Qu.: 3585292      Subscriber:519452 1st Qu.:45.52    1st Qu.: -122.7
## Median : 7231950              :      0    Median :45.52    Median : -122.7
## Mean   : 7230195              :      0    Mean   :45.52    Mean   : -122.7
## 3rd Qu.:11069074              :      0    3rd Qu.:45.53    3rd Qu.: -122.7
## Max.   :13201070              :      0    Max.   :45.73    Max.   : -122.4
##                                     NA's   :509      NA's   :509
##      StartDate      StartTime EndLatitude EndLongitude
## 5/27/2018:    4740    17:06 :    2218    Min.   :34.26    Min.   : -134.4
## 5/26/2018:    3703    17:09 :    2200    1st Qu.:45.52    1st Qu.: -122.7
## 5/28/2018:    3690    17:08 :    2183    Median :45.52    Median : -122.7
## 5/19/2018:    3678    17:07 :    2176    Mean   :45.52    Mean   : -122.7
## 5/12/2018:    3546    17:11 :    2165    3rd Qu.:45.53    3rd Qu.: -122.7
## 5/20/2018:    3501    17:12 :    2138    Max.   :49.16    Max.   :  45.5
## (Other)   :1204869    (Other):1214647    NA's   :559      NA's   :559
##      EndDate      EndTime Distance_Miles Duration
```

```
## 5/27/2018: 4752 17:27 : 2246 Min. : 0.000 0:06:37: 1225
## 5/28/2018: 3718 17:29 : 2186 1st Qu.: 0.720 0:06:39: 1216
## 5/26/2018: 3712 17:28 : 2146 Median : 1.300 0:05:41: 1212
## 5/19/2018: 3642 17:25 : 2134 Mean : 1.986 0:05:36: 1211
## 5/20/2018: 3532 17:18 : 2128 3rd Qu.: 2.390 0:06:07: 1210
## 5/12/2018: 3521 17:31 : 2121 Max. :15527.180 0:05:37: 1199
## (Other) :1204850 (Other):1214766 (Other):1220454
## RentalAccessPath MultipleRental
## keypad :911535 Mode :logical
## mobile :184475 FALSE:1115829
## keypad_rfid_card :125862 TRUE :111898
## keypad_phone_number: 2802
## web : 2186
## unknown : 832
## (Other) : 35
```

Next we convert Duration into a decimal value recorded in minutes. The code to accomplish this task is below and inspired by this [StackOverflow](#) post.

```
dat.raw$Duration <- sapply(strsplit(as.character(dat.raw$Duration),":"),
  function(x) {
    x <- as.numeric(x)
    x[1]*60+x[2]+x[3]/60
  }
)
summary(dat.raw)
```

```
## RouteID PaymentPlan StartLatitude StartLongitude
## Min. : 1282087 Casual :708275 Min. :45.35 Min. : -123.1
## 1st Qu.: 3585292 Subscriber:519452 1st Qu.:45.52 1st Qu.: -122.7
## Median : 7231950 : 0 Median :45.52 Median : -122.7
## Mean : 7230195 Mean :45.52 Mean : -122.7
## 3rd Qu.:11069074 3rd Qu.:45.53 3rd Qu.: -122.7
## Max. :13201070 Max. :45.73 Max. : -122.4
## NA's :509 NA's :509
## StartDate StartTime EndLatitude EndLongitude
## 5/27/2018: 4740 17:06 : 2218 Min. :34.26 Min. : -134.4
## 5/26/2018: 3703 17:09 : 2200 1st Qu.:45.52 1st Qu.: -122.7
## 5/28/2018: 3690 17:08 : 2183 Median :45.52 Median : -122.7
## 5/19/2018: 3678 17:07 : 2176 Mean :45.52 Mean : -122.7
## 5/12/2018: 3546 17:11 : 2165 3rd Qu.:45.53 3rd Qu.: -122.7
## 5/20/2018: 3501 17:12 : 2138 Max. :49.16 Max. : 45.5
## (Other) :1204869 (Other):1214647 NA's :559 NA's :559
## EndDate EndTime Distance_Miles Duration
## 5/27/2018: 4752 17:27 : 2246 Min. : 0.000 Min. : 1
## 5/28/2018: 3718 17:29 : 2186 1st Qu.: 0.720 1st Qu.: 7
## 5/26/2018: 3712 17:28 : 2146 Median : 1.300 Median : 13
## 5/19/2018: 3642 17:25 : 2134 Mean : 1.986 Mean : 564
## 5/20/2018: 3532 17:18 : 2128 3rd Qu.: 2.390 3rd Qu.: 27
## 5/12/2018: 3521 17:31 : 2121 Max. :15527.180 Max. :33042053
## (Other) :1204850 (Other):1214766
## RentalAccessPath MultipleRental
## keypad :911535 Mode :logical
## mobile :184475 FALSE:1115829
## keypad_rfid_card :125862 TRUE :111898
```

```
## keypad_phone_number: 2802
## web : 2186
## unknown : 832
## (Other) : 35
```

In the above output we notice at least one particular value for Duration is quite large. So we will look at the details for this one observation.

```
dat.raw[which(dat.raw$Duration >= 33042053-1), ]
```

```
##      RouteID PaymentPlan StartLatitude StartLongitude StartDate StartTime
## 178744 2463277 Subscriber          NA          NA 3/10/2017      19:07
##      EndLatitude EndLongitude  EndDate  EndTime Distance_Miles Duration
## 178744          NA          NA 1/5/2080    16:00             0 33042053
##      RentalAccessPath MultipleRental
## 178744          keypad          FALSE
```

We notice that this observation's trip lasted until 1/5/2080, clearly in the future. Rather than remove just this one observation at this time, we will convert all of the dates into a date format and see how many other observations have start/end trips outside of a reasonable time frame.

```
dat.raw$StartDate <- as.Date(dat.raw$StartDate, format="%m/%d/%Y")
dat.raw$EndDate <- as.Date(dat.raw$EndDate, format="%m/%d/%Y")
```

Next we will eliminate all observations not found in our date time range, and output the summary.

```
known.dates <- as.factor(seq(as.Date("2016-07-19"), as.Date("2020-03-31"), "days"))
in.known.range <- which(as.factor(dat.raw$EndDate) %in% known.dates)
dat.raw <- dat.raw[in.known.range, ]
rm(known.dates, in.known.range) # We do not need these anymore.
summary(dat.raw)
```

```
##      RouteID      PaymentPlan  StartLatitude  StartLongitude
## Min.   : 1282087  Casual      :708274  Min.   :45.35  Min.   :-123.1
## 1st Qu.: 3585329  Subscriber:519431  1st Qu.:45.52  1st Qu.: -122.7
## Median : 7232030           :      0  Median :45.52  Median : -122.7
## Mean   : 7230244           :      0  Mean   :45.52  Mean   : -122.7
## 3rd Qu.:11069097           :      0  3rd Qu.:45.53  3rd Qu.: -122.7
## Max.   :13201070           :      0  Max.   :45.73  Max.   : -122.4
##                                     NA's   :499    NA's   :499
##      StartDate      StartTime      EndLatitude      EndLongitude
## Min.   :2016-07-19  17:06 : 2218  Min.   :34.26  Min.   :-134.4
## 1st Qu.:2017-07-13  17:09 : 2200  1st Qu.:45.52  1st Qu.: -122.7
## Median :2018-05-30  17:08 : 2183  Median :45.52  Median : -122.7
## Mean   :2018-05-02  17:07 : 2176  Mean   :45.52  Mean   : -122.7
## 3rd Qu.:2019-03-31  17:11 : 2165  3rd Qu.:45.53  3rd Qu.: -122.7
## Max.   :2020-03-31  17:12 : 2138  Max.   :49.16  Max.   : 45.5
##                                     (Other):1214625  NA's   :552    NA's   :552
##      EndDate      EndTime      Distance_Miles      Duration
## Min.   :2016-07-19  17:27 : 2246  Min.   : 0.000  Min.   : 1.00
## 1st Qu.:2017-07-13  17:29 : 2186  1st Qu.: 0.720  1st Qu.: 7.25
## Median :2018-05-30  17:28 : 2146  Median : 1.300  Median : 13.42
## Mean   :2018-05-02  17:25 : 2134  Mean   : 1.986  Mean   : 27.55
## 3rd Qu.:2019-03-31  17:18 : 2128  3rd Qu.: 2.390  3rd Qu.: 27.15
## Max.   :2020-03-31  17:31 : 2121  Max.   :15527.180  Max.   :57460.92
##                                     (Other):1214744
##      RentalAccessPath MultipleRental
```

```
## keypad          :911516   Mode :logical
## mobile          :184474   FALSE:1115809
## keypad_rfid_card :125860   TRUE :111896
## keypad_phone_number: 2802
## web             : 2186
## unknown         : 832
## (Other)         : 35
```

From the above output, we still see some values that are unreasonable: `Distance_Miles` and `Duration`. Since we're interested in `Duration` we will remove their extreme outliers.

```
ex.out <- 3 * IQR(dat.raw$Duration)
dur.out.in <- which(dat.raw$Duration >= ex.out)
dat.raw <- dat.raw[-dur.out.in, ]
rm(dur.out.in, ex.out) # We no longer need these values
summary(dat.raw)
```

```
##      RouteID          PaymentPlan    StartLatitude  StartLongitude
## Min.   : 1282087    Casual      :633712    Min.   :45.40    Min.   : -122.9
## 1st Qu.: 3606950    Subscriber:495962    1st Qu.:45.52    1st Qu.: -122.7
## Median : 7300984          :      0    Median :45.52    Median : -122.7
## Mean   : 7270893          Mean   :45.52    Mean   : -122.7
## 3rd Qu.:11090696          3rd Qu.:45.53    3rd Qu.: -122.7
## Max.   :13201070          Max.   :45.68    Max.   : -122.4
##                                     NA's   :476      NA's   :476
##      StartDate          StartTime      EndLatitude  EndLongitude
## Min.   :2016-07-19    17:06 : 2123    Min.   :45.40    Min.   : -122.8
## 1st Qu.:2017-07-14    17:09 : 2087    1st Qu.:45.52    1st Qu.: -122.7
## Median :2018-06-02    17:08 : 2074    Median :45.52    Median : -122.7
## Mean   :2018-05-06    17:07 : 2057    Mean   :45.52    Mean   : -122.7
## 3rd Qu.:2019-04-03    17:11 : 2053    3rd Qu.:45.53    3rd Qu.: -122.7
## Max.   :2020-03-31    17:10 : 2033    Max.   :45.63    Max.   : 45.5
##                                     (Other):1117247    NA's   :523      NA's   :523
##      EndDate          EndTime      Distance_Miles  Duration
## Min.   :2016-07-19    17:27 : 2073    Min.   : 0.000    Min.   : 1.000
## 1st Qu.:2017-07-14    17:29 : 2025    1st Qu.: 0.690    1st Qu.: 6.883
## Median :2018-06-02    17:28 : 1981    Median : 1.200    Median :12.167
## Mean   :2018-05-06    17:25 : 1980    Mean   : 1.657    Mean   :16.529
## 3rd Qu.:2019-04-03    17:21 : 1977    3rd Qu.: 2.090    3rd Qu.:22.350
## Max.   :2020-03-31    17:31 : 1976    Max.   :15527.180    Max.   :59.683
##                                     (Other):1117662
##      RentalAccessPath MultipleRental
## keypad          :830402   Mode :logical
## mobile          :174581   FALSE:1036726
## keypad_rfid_card :119698   TRUE :92948
## keypad_phone_number: 2500
## web             : 1832
## unknown         : 630
## (Other)         : 31
```

From the above output we notice that there still may exist some extreme outliers for `Distance_Miles` so we proceed to remove them too.

```
ex.out <- 3 * IQR(dat.raw$Distance_Miles)
dur.out.in <- which(dat.raw$Distance_Miles >= ex.out)
dat.raw <- dat.raw[-dur.out.in, ]
```



```
rm(dur.out.in, ex.out) # We no longer need these values
summary(dat.raw)
```

```
##      RouteID      PaymentPlan      StartLatitude      StartLongitude
## Min.   : 1282087      Casual      :602783      Min.   :45.40      Min.   : -122.8
## 1st Qu.: 3607394      Subscriber:486839      1st Qu.:45.52      1st Qu.: -122.7
## Median : 7304754              :      0      Median :45.52      Median : -122.7
## Mean   : 7274793              Mean   :45.52      Mean   : -122.7
## 3rd Qu.:11091024              3rd Qu.:45.53      3rd Qu.: -122.7
## Max.   :13201070              Max.   :45.62      Max.   : -122.4
##                                     NA's   :476      NA's   :476
##      StartDate      StartTime      EndLatitude      EndLongitude
## Min.   :2016-07-19      17:06 : 2053      Min.   :45.40      Min.   : -122.8
## 1st Qu.:2017-07-14      17:08 : 2008      1st Qu.:45.52      1st Qu.: -122.7
## Median :2018-06-02      17:09 : 2004      Median :45.52      Median : -122.7
## Mean   :2018-05-06      17:07 : 1994      Mean   :45.52      Mean   : -122.7
## 3rd Qu.:2019-04-03      17:11 : 1992      3rd Qu.:45.53      3rd Qu.: -122.7
## Max.   :2020-03-31      17:12 : 1969      Max.   :45.62      Max.   : -122.4
##                                     (Other):1077602      NA's   :523      NA's   :523
##      EndDate      EndTime      Distance_Miles      Duration
## Min.   :2016-07-19      17:27 : 2011      Min.   :0.000      Min.   : 1.000
## 1st Qu.:2017-07-14      17:29 : 1950      1st Qu.:0.670      1st Qu.: 6.717
## Median :2018-06-02      17:28 : 1926      Median :1.160      Median :11.700
## Mean   :2018-05-06      17:25 : 1912      Mean   :1.401      Mean   :15.515
## 3rd Qu.:2019-04-03      17:26 : 1910      3rd Qu.:1.960      3rd Qu.:20.817
## Max.   :2020-03-31      17:21 : 1908      Max.   :4.190      Max.   :59.683
##                                     (Other):1078005
##      RentalAccessPath      MultipleRental
## keypad      :799282      Mode :logical
## mobile      :168289      FALSE:1002169
## keypad_rfid_card :117377      TRUE :87453
## keypad_phone_number: 2396
## web         : 1673
## unknown     : 576
## (Other)     : 29
```

Final cleaning

Finally we will do some final cleaning before writing the data to file. We first check whether or not the values for RouteID are all unique.

```
length(unique(dat.raw$RouteID)) == dim(dat.raw)[1]
```

```
## [1] TRUE
```

Since TRUE (i.e. they are all unique) we remove RouteID as it adds no additional information.

```
dat.raw <- dat.raw[, -1]
summary(dat.raw)
```

```
##      PaymentPlan      StartLatitude      StartLongitude      StartDate
## Casual      :602783      Min.   :45.40      Min.   : -122.8      Min.   :2016-07-19
## Subscriber:486839      1st Qu.:45.52      1st Qu.: -122.7      1st Qu.:2017-07-14
##              :      0      Median :45.52      Median : -122.7      Median :2018-06-02
##              Mean   :45.52      Mean   : -122.7      Mean   :2018-05-06
```

```
##          3rd Qu.:45.53    3rd Qu.: -122.7    3rd Qu.:2019-04-03
##          Max.    :45.62    Max.    : -122.4    Max.    :2020-03-31
##          NA's    :476      NA's    :476
##      StartTime      EndLatitude      EndLongitude      EndDate
## 17:06 : 2053    Min.    :45.40    Min.    : -122.8    Min.    :2016-07-19
## 17:08 : 2008    1st Qu.:45.52    1st Qu.: -122.7    1st Qu.:2017-07-14
## 17:09 : 2004    Median :45.52    Median : -122.7    Median :2018-06-02
## 17:07 : 1994    Mean    :45.52    Mean    : -122.7    Mean    :2018-05-06
## 17:11 : 1992    3rd Qu.:45.53    3rd Qu.: -122.7    3rd Qu.:2019-04-03
## 17:12 : 1969    Max.    :45.62    Max.    : -122.4    Max.    :2020-03-31
## (Other):1077602    NA's    :523      NA's    :523
##      EndTime      Distance_Miles      Duration
## 17:27 : 2011    Min.    :0.000    Min.    : 1.000
## 17:29 : 1950    1st Qu.:0.670    1st Qu.: 6.717
## 17:28 : 1926    Median :1.160    Median :11.700
## 17:25 : 1912    Mean    :1.401    Mean    :15.515
## 17:26 : 1910    3rd Qu.:1.960    3rd Qu.:20.817
## 17:21 : 1908    Max.    :4.190    Max.    :59.683
## (Other):1078005
##      RentalAccessPath  MultipleRental
## keypad                :799282    Mode :logical
## mobile                :168289    FALSE:1002169
## keypad_rfid_card      :117377    TRUE :87453
## keypad_phone_number: 2396
## web                   : 1673
## unknown               : 576
## (Other)               : 29
```

Finally we will remove `StartLatitude`, `StartLongitude`, `EndLatitude`, `EndLongitude`, and `MultipleRental` as these values are not of interest to us in this project.

```
dat.raw <- dat.raw[, -c(2, 3, 6, 7, 13)]
summary(dat.raw)
```

```
##      PaymentPlan      StartDate      StartTime
## Casual :602783    Min.    :2016-07-19    17:06 : 2053
## Subscriber:486839    1st Qu.:2017-07-14    17:08 : 2008
##          : 0      Median :2018-06-02    17:09 : 2004
##          Mean    :2018-05-06    17:07 : 1994
##          3rd Qu.:2019-04-03    17:11 : 1992
##          Max.    :2020-03-31    17:12 : 1969
##          (Other):1077602
##      EndDate      EndTime      Distance_Miles      Duration
## Min.    :2016-07-19    17:27 : 2011    Min.    :0.000    Min.    : 1.000
## 1st Qu.:2017-07-14    17:29 : 1950    1st Qu.:0.670    1st Qu.: 6.717
## Median :2018-06-02    17:28 : 1926    Median :1.160    Median :11.700
## Mean    :2018-05-06    17:25 : 1912    Mean    :1.401    Mean    :15.515
## 3rd Qu.:2019-04-03    17:26 : 1910    3rd Qu.:1.960    3rd Qu.:20.817
## Max.    :2020-03-31    17:21 : 1908    Max.    :4.190    Max.    :59.683
##          (Other):1078005
##      RentalAccessPath
## keypad                :799282
## mobile                :168289
## keypad_rfid_card      :117377
## keypad_phone_number: 2396
```

```
## web          : 1673
## unknown      : 576
## (Other)      : 29
```

We then write this data set to file so we can use it in the future.

```
write.csv(dat.raw, file="~/GoogleDrive/pdxbikes/pdxbikes/data/interim/interim.csv",
          row.names=FALSE)
```

Known issues

1. File paths are not relative.
2. There are some hardcoded values used in cleaning the data.

Future work

1. Fix all of the problems listed in the “Known issues” section
2. Take another look at the raw data to find interesting observations. In particular observations riding at an unreasonable speed and trips that are either from the past or the future. This will be accomplished with another notebook
3. Write an R script that will create the file outputted from this notebook without having to run through this notebook. We will store it in a `/src/` directory.

Session information

Below you will find the output from `sessionInfo()` to assist in reproducing the work shown in this notebook.

```
sessionInfo()

## R version 3.5.2 (2018-12-20)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS 10.15.4
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## loaded via a namespace (and not attached):
## [1] compiler_3.5.2  magrittr_1.5    tools_3.5.2    htmltools_0.4.0
## [5] yaml_2.2.1      Rcpp_1.0.4.6    stringi_1.4.6  rmarkdown_2.1
## [9] knitr_1.28      stringr_1.4.0   xfun_0.13      digest_0.6.25
## [13] rlang_0.4.5     evaluate_0.14
```