Data aggregating notebook for Biketown PDX

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

2020-06-18

Contents

This notebook will take the interim data and transform it into a form that separates `Casual` and `Subscriber` users.

Preliminaries

Before we begin this notebook we will need the following function and `R` package.

Needed function

This function is based on a solution provided from Stack Overflow, and we have adapted it to accompish what we need it to do.

```
totals <- function(x){
  # This function will take a numerical vector x and output the sum of x and
  # the mean of x.
  # ----------
  # INPUT
  # x := a numerical vector.
  # ----------
  # OUTPUT
  # A data object with columns containing the sum and mean of x.
  c(sum=sum(x), mean=mean(x))
}
```

Needed packages

In order to aggregate the data we will take advantage of the syntactic sugar provided by the `zoo` package avialable in `R`.

```
require(zoo)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Bringing in the data

We begin by bringing the data into `R`, and output a summary of the data using `summary()`.

```
int.dat <- read.csv("https://raw.githubusercontent.com/raford/pdxbikes/master/data/interim/interim.csv"
                    header=TRUE, sep=",")
summary(int.dat)
```

```
##  PaymentPlan         StartDate          StartTime           EndDate
##  Length:1089622     Length:1089622     Length:1089622     Length:1089622
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    EndTime          Distance_Miles     Duration      RentalAccessPath
##  Length:1089622     Min.   :0.000   Min.   : 1.000   Length:1089622
##  Class :character   1st Qu.:0.670   1st Qu.: 6.717   Class :character
##  Mode  :character   Median :1.160   Median :11.700   Mode  :character
##                     Mean   :1.401   Mean   :15.515
##                     3rd Qu.:1.960   3rd Qu.:20.817
##                     Max.   :4.190   Max.   :59.683
```

We next subset the data into two distinct groups: one for `Casual` users and one for the `Subscriber` group—both based on the payment plan used.

```
cas.dat <- subset(int.dat, PaymentPlan == "Casual")[, -1] # Payment plan in first column
sub.dat <- subset(int.dat, PaymentPlan == "Subscriber")[, -1]
rm(int.dat)
```

Aggregate the data

We next construct two separate data sets: one containing the data for the `Casual` users and one containing the `Subscriber` users, and then write each to their own CSV file. The code used to create the datasets for each of these classes is below, and labeled appropriately.

Casual users

```
cas.dat$StartDate <- as.Date(cas.dat$StartDate)
Date <- seq.Date(from=min(cas.dat$StartDate), to=max(cas.dat$StartDate), "days")
cas.dat <- cas.dat[, -c(2, 3, 4, 7)]
cas.dat.agg <- read.zoo(cas.dat, header=TRUE, aggregate=totals)
cas.dat.agg <- cbind(Date, as.data.frame(cas.dat.agg))
write.csv(cas.dat.agg, file="~/GoogleDrive/pdxbikes/pdxbikes/data/aggregate/casual.csv",
          row.names=FALSE)
rm(cas.dat, cas.dat.agg)
```

Subscribers

```
sub.dat$StartDate <- as.Date(sub.dat$StartDate)
Date <- seq.Date(from=min(sub.dat$StartDate), to=max(sub.dat$StartDate), "days")
sub.dat <- sub.dat[, -c(2, 3, 4, 7)]
sub.dat.agg <- read.zoo(sub.dat, header=TRUE, aggregate=totals)
sub.dat.agg <- cbind(Date, as.data.frame(sub.dat.agg))
write.csv(sub.dat.agg, file="~/GoogleDrive/pdxbikes/pdxbikes/data/aggregate/subscriber.csv",
          row.names=FALSE)
rm(sub.dat, sub.dat.agg)
```

Future work

1. Perform some exploratory data analysis (EDA) for these data sets.
2. ~~Create R scripts that will create these data sets.~~ **Completed on 18 June 2020. These scripts are located in `/src/data/`**

Session information

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

```
sessionInfo()
```

```
## R version 4.0.1 (2020-06-06)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.4
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/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
##
## other attached packages:
```

```
## [1] zoo_1.8-8
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.0.1  magrittr_1.5    tools_4.0.1     htmltools_0.4.0
##  [5] yaml_2.2.1      Rcpp_1.0.4.6    stringi_1.4.6   rmarkdown_2.2
##  [9] grid_4.0.1      knitr_1.28      stringr_1.4.0   xfun_0.14
## [13] digest_0.6.25   rlang_0.4.6     lattice_0.20-41 evaluate_0.14
```