station-analysis

Julian DeGroot-Lutzner

12/14/2017

```r
station_432 <- read.csv("station-432.csv")
station_521 <- read.csv("station-521.csv")
Weather_NYC <- read_csv("~/Documents/math154/ma154-project24-teambike/station-analysis/Weather_NYC.csv"
```

HERE COULD BE MORE ANALYSIS ABOUT EACH STATION

How far does the average citibike move in a week? Change avg. trip duration to median.

```r
# Parsing all the start times into one format
mdy <- mdy_hms(station_432$starttime)
ymd <- ymd_hms(station_432$starttime)
f1 <- mdy_hm(station_432$starttime)
mdy[is.na(mdy)] <- ymd[is.na(mdy)]
station_432$starttime <- mdy
station_432$starttime[is.na(station_432$starttime)] <- f1[is.na(station_432$starttime)]

# Parsing all the start times into one format
mdy <- mdy_hms(station_521$starttime)
ymd <- ymd_hms(station_521$starttime)
f1 <- mdy_hm(station_521$starttime)
mdy[is.na(mdy)] <- ymd[is.na(mdy)]
station_521$starttime <- mdy
station_521$starttime[is.na(station_521$starttime)] <- f1[is.na(station_521$starttime)]
```

```r
# took out rides that looped to the same station because these
# rides don't impact our prediction model
median_ <- function(...) median(..., na.rm=T)

# hourly sums of station 521
hour_sums_521 <- station_521 %>%
  select(tripduration, starttime, start.station.id,
         end.station.id, usertype) %>%
  mutate(starttime = floor_date(starttime, "hour"),
         started.here = (start.station.id == 521),
         ended.here = (end.station.id == 521),
         subscriber = (usertype == "Subscriber"),
         customer = (usertype == "Customer")) %>%
  mutate(subscriber.started.here =
           (started.here & !ended.here & subscriber),
         subscriber.ended.here =
           (!started.here & ended.here & subscriber),
         customer.started.here =
           (started.here & !ended.here & subscriber),
         customer.ended.here =
           (!started.here & ended.here & subscriber)) %>%
  mutate_all(funs(ifelse(is.na(.), 0, .))) %>%
  mutate(duration.from.start =
          ifelse(started.here & !ended.here, tripduration, NA),
         duration.to.finish =
          ifelse(!started.here & ended.here, tripduration, NA)) %>%
  group_by(starttime) %>%
  summarize(median.trip.from.521 =
              median_(duration.from.start),
            median.trip.to.521 =
              median_(duration.to.finish),
            num.subscribers.started.521 =
              sum(subscriber.started.here),
            num.subscribers.ended.521 =
              sum(subscriber.ended.here),
            num.customers.started.521 =
              sum(customer.started.here),
            num.customers.ended.521 =
              sum(customer.started.here),
            total.trips.started.521 =
              sum(started.here & !ended.here),
            total.trips.ended.521 =
              sum(!started.here & ended.here))

# hourly sums for station 432
hour_sums_432 <- station_432 %>%
  select(tripduration, starttime, start.station.id,
         end.station.id, usertype) %>%
  mutate(starttime = floor_date(starttime, "hour"),
         started.here = (start.station.id == 432),
         ended.here = (end.station.id == 432),
         subscriber = (usertype == "Subscriber"),
         customer = (usertype == "Customer")) %>%
  mutate(subscriber.started.here =
           (started.here & !ended.here & subscriber),
         subscriber.ended.here =
           (!started.here & ended.here & subscriber),
         customer.started.here =
           (started.here & !ended.here & subscriber),
         customer.ended.here =
           (!started.here & ended.here & subscriber)) %>%
  mutate_all(funs(ifelse(is.na(.), 0, .))) %>%
  mutate(duration.from.start =
          ifelse(started.here & !ended.here, tripduration, NA),
         duration.to.finish =
          ifelse(!started.here & ended.here, tripduration, NA)) %>%
  group_by(starttime) %>%
  summarize(median.trip.from.432 =
              median_(duration.from.start),
            median.trip.to.432 =
              median_(duration.to.finish),
            num.subscribers.started.432 =
              sum(subscriber.started.here),
            num.subscribers.ended.432 =
              sum(subscriber.ended.here),
            num.customers.started.432 =
              sum(customer.started.here),
            num.customers.ended.432 =
              sum(customer.started.here),
            total.trips.started.432 =
              sum(started.here & !ended.here),
            total.trips.ended.432 =
              sum(!started.here & ended.here))
```

```
hourly_station_data <-
  left_join(hour_sums_432, hour_sums_521, by ="starttime")
```

Prepare the Weather Data

```
# Removing the extraneous variables
Weather_NYC <- Weather_NYC %>%
  select(-station,-drct,-p01i,-skyc1,
         -skyc2,-skyc3,-skyc4,skyl1,
         -skyl2,-skyl2,-skyl3,-skyl4,-metar)

Weather_NYC <- Weather_NYC %>%
  mutate(valid = ymd_hms(valid)) %>%
  filter(minute(valid)=="51")  %>%
  mutate(valid = ceiling_date(valid, unit = "hour"),
         Month=month(valid)) %>%
  mutate( summer=ifelse(Month=="6"|Month=="7"|Month=="8",1,0),
          spring=ifelse((Month=="3"|Month=="4"|Month=="5"),1,0),
          winter=ifelse((Month=="1"|Month=="2"|Month=="12"),1,0),
          fall=ifelse((Month=="9"|Month=="10"|Month=="11"),1,0),
          day.of.week=wday(valid)) %>%
  mutate(week.day=
           ifelse(day.of.week > 1 & day.of.week < 7,TRUE,FALSE),
         weekend.day=
           ifelse(day.of.week == 1 | day.of.week == 7,TRUE,FALSE),
         starttime = valid) %>%
  select(-valid)
```

```
bikes_weather_combined <-
  inner_join(Weather_NYC, hourly_station_data, by= "starttime" )
```

```
head(bikes_weather_combined)
```

```
## # A tibble: 6 x 35
##    tmpf  dwpf  relh  sknt  alti    mslp  vsby  gust   skyl1 presentwx
##   <dbl> <dbl> <dbl> <chr> <dbl>   <chr> <dbl> <chr>   <chr>     <chr>
## 1 75.02 69.98 84.34  0.00 29.95 1013.10     8     M 1500.00         M
## 2 73.94 69.98 87.45  3.00 29.94 1012.80     8     M 1400.00         M
## 3 73.94 69.98 87.45  0.00 29.94 1012.70     9     M 1100.00         M
## 4 73.94 69.98 87.45  0.00 29.96 1013.50     9     M 1200.00         M
## 5 75.02 69.98 84.34  4.00 29.96 1013.70     7     M 1000.00         M
## 6 75.02 71.06 87.50     M 29.97 1013.90     6     M 1000.00        BR
## # ... with 25 more variables: Month <dbl>, summer <dbl>, spring <dbl>,
## #   winter <dbl>, fall <dbl>, day.of.week <dbl>, week.day <lgl>,
## #   weekend.day <lgl>, starttime <dttm>, median.trip.from.432 <dbl>,
## #   median.trip.to.432 <dbl>, num.subscribers.started.432 <int>,
## #   num.subscribers.ended.432 <int>, num.customers.started.432 <int>,
## #   num.customers.ended.432 <int>, total.trips.started.432 <int>,
## #   total.trips.ended.432 <int>, median.trip.from.521 <dbl>,
## #   median.trip.to.521 <dbl>, num.subscribers.started.521 <int>,
## #   num.subscribers.ended.521 <int>, num.customers.started.521 <int>,
## #   num.customers.ended.521 <int>, total.trips.started.521 <int>,
## #   total.trips.ended.521 <int>
```

```
tbl_df(bikes_weather_combined[1,])
```

```
## # A tibble: 1 x 35
##    tmpf  dwpf  relh  sknt  alti    mslp  vsby  gust   skyl1 presentwx
##   <dbl> <dbl> <dbl> <chr> <dbl>   <chr> <dbl> <chr>   <chr>     <chr>
## 1 75.02 69.98 84.34  0.00 29.95 1013.10     8     M 1500.00         M
## # ... with 25 more variables: Month <dbl>, summer <dbl>, spring <dbl>,
## #   winter <dbl>, fall <dbl>, day.of.week <dbl>, week.day <lgl>,
## #   weekend.day <lgl>, starttime <dttm>, median.trip.from.432 <dbl>,
## #   median.trip.to.432 <dbl>, num.subscribers.started.432 <int>,
## #   num.subscribers.ended.432 <int>, num.customers.started.432 <int>,
## #   num.customers.ended.432 <int>, total.trips.started.432 <int>,
## #   total.trips.ended.432 <int>, median.trip.from.521 <dbl>,
## #   median.trip.to.521 <dbl>, num.subscribers.started.521 <int>,
## #   num.subscribers.ended.521 <int>, num.customers.started.521 <int>,
## #   num.customers.ended.521 <int>, total.trips.started.521 <int>,
## #   total.trips.ended.521 <int>
```