station-analysis

Julian DeGroot-Lutzner

12/14/2017

```
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.

```
# 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)]
```

```
# 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(duration.from.start =
          ifelse(started.here & !ended.here, tripduration, NA),
         duration.to.finish =
          ifelse(!started.here & ended.here, tripduration, NA),
         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)) %>%
  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(duration.from.start =
          ifelse(started.here & !ended.here, tripduration, NA),
         duration.to.finish =
          ifelse(!started.here & ended.here, tripduration, NA),
         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)) %>%
  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))
```

```
# combine hourly station data
head(hour_sums_432)
```

```
## # A tibble: 6 x 9
##             starttime median.trip.from.432 median.trip.to.432
##                <dttm>                <dbl>              <dbl>
## 1 2013-07-01 00:00:00                840.0                 NA
## 2 2013-07-01 01:00:00               1071.0                 NA
## 3 2013-07-01 02:00:00               1373.0                 NA
## 4 2013-07-01 03:00:00                927.0                 NA
## 5 2013-07-01 04:00:00                   NA                690
## 6 2013-07-01 06:00:00                425.5                 NA
## # ... with 6 more variables: 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>
```

```
hour_sums_521
```

```
## # A tibble: 21,919 x 9
##              starttime median.trip.from.521 median.trip.to.521
##                 <dttm>                <dbl>              <dbl>
##  1 2013-07-01 00:00:00               1257.0              843.0
##  2 2013-07-01 01:00:00                   NA              677.0
##  3 2013-07-01 02:00:00                   NA              247.0
##  4 2013-07-01 05:00:00                403.0              441.0
##  5 2013-07-01 06:00:00                740.0              429.5
##  6 2013-07-01 07:00:00                604.0              509.0
##  7 2013-07-01 08:00:00                998.5              671.0
##  8 2013-07-01 09:00:00                832.0              693.0
##  9 2013-07-01 11:00:00                225.0                 NA
## 10 2013-07-01 12:00:00               2001.0              493.0
## # ... with 21,909 more rows, and 6 more variables:
## #   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>
```

```
hour_sums_521 %>% select(total.trips.started.521, total.trips.ended.521) %>% head()
```

```
## # A tibble: 6 x 2
##   total.trips.started.521 total.trips.ended.521
##                     <int>                 <int>
## 1                       4                     3
## 2                       0                     1
## 3                       0                     1
## 4                       3                     3
## 5                      11                     4
## 6                      29                     9
```