Setup

```r
# libraries
library(sf)
```

```
## Linking to GEOS 3.10.2, GDAL 3.4.1, PROJ 8.2.1; sf_use_s2() is TRUE
```

```r
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
```

```r
library(ggplot2)
library(jpeg)

# data sets
restaurants = read.csv("restaurants.csv")
movements = read.csv("movements.csv")

# function for calculating buffer based on an input in sq. footage (circle->radius in feet->meters->coo
calc_buffer <- function(size)
{
  return((((size / pi)^(0.5))*0.3048)/111111)
}
```

Cleaning up movements

```r
# Sort by datetime
movements <- movements %>%
  mutate(datetime = as.POSIXct(datetime, format = "%Y-%m-%d %H:%M:%S")) %>%
  arrange(datetime)

# Filter duplicates
movements <- movements %>%
  mutate(err = if_else(datetime > lag(datetime, default = as.POSIXct("2019-12-31 23:59:59", format = "%
  filter(err == 0)
```

EDA

```r
# Categories of restaurants
rest_types <- restaurants %>%
  count(Category)
rest_types
```

```
##                                              Category n
## 1                                 Convenience Stores 3
## 2                          Fruit & Vegetable Markets 4
## 3                           Full-Service Restaurants 7
## 4                        Limited-Service Restaurants 5
## 5 Supermarkets/Other Grocery (Exc Convenience) Strs 2
## 6                     Warehouse Clubs & Supercenters 4
```

```r
restaurants
```

```
##    Restaurant.ID                    Name
## 1           R000     Pullman Fine Dine 0
## 2           R001    Pullman Quick Eats 1
## 3           R002  Pullman Fresh Market 2
## 4           R003     Pullman Fine Dine 3
## 5           R004     Pullman Fine Dine 4
## 6           R005     Pullman Groceries 5
## 7           R006   Pullman Supercenter 6
## 8           R007  Pullman Fresh Market 7
## 9           R008    Pullman Quick Eats 8
## 10          R009    Pullman Quick Stop 9
## 11          R010  Pullman Supercenter 10
## 12          R011    Pullman Groceries 11
## 13          R012 Pullman Fresh Market 12
## 14          R013   Pullman Quick Eats 13
## 15          R014    Pullman Fine Dine 14
## 16          R015   Pullman Quick Stop 15
## 17          R016 Pullman Fresh Market 16
## 18          R017  Pullman Supercenter 17
## 19          R018   Pullman Quick Stop 18
## 20          R019    Pullman Fine Dine 19
## 21          R020  Pullman Supercenter 20
## 22          R021    Pullman Fine Dine 21
## 23          R022   Pullman Quick Eats 22
## 24          R023   Pullman Quick Eats 23
## 25          R024    Pullman Fine Dine 24
##                                               Category Longitude Latitude
## 1                            Full-Service Restaurants -122.2725 47.65121
## 2                         Limited-Service Restaurants -122.2994 47.58553
## 3                           Fruit & Vegetable Markets -122.4015 47.60990
## 4                            Full-Service Restaurants -122.2828 47.64479
## 5                            Full-Service Restaurants -122.3048 47.60799
## 6  Supermarkets/Other Grocery (Exc Convenience) Strs -122.4267 47.57224
## 7                      Warehouse Clubs & Supercenters -122.3335 47.67040
## 8                           Fruit & Vegetable Markets -122.3818 47.54453
## 9                         Limited-Service Restaurants -122.3532 47.56739
## 10                                 Convenience Stores -122.3389 47.57299
## 11                    Warehouse Clubs & Supercenters -122.3201 47.56489
## 12 Supermarkets/Other Grocery (Exc Convenience) Strs -122.3717 47.68925
## 13                         Fruit & Vegetable Markets -122.3508 47.59793
## 14                      Limited-Service Restaurants -122.3179 47.59217
## 15                         Full-Service Restaurants -122.4096 47.59302
## 16                               Convenience Stores -122.3068 47.56951
## 17                         Fruit & Vegetable Markets -122.3586 47.58226
## 18                    Warehouse Clubs & Supercenters -122.3270 47.54416
## 19                               Convenience Stores -122.3164 47.61851
## 20                         Full-Service Restaurants -122.3521 47.60202
## 21                    Warehouse Clubs & Supercenters -122.3028 47.60313
## 22                         Full-Service Restaurants -122.3107 47.62365
## 23                      Limited-Service Restaurants -122.3747 47.59370
## 24                      Limited-Service Restaurants -122.2835 47.66092
## 25                         Full-Service Restaurants -122.4359 47.61980
```

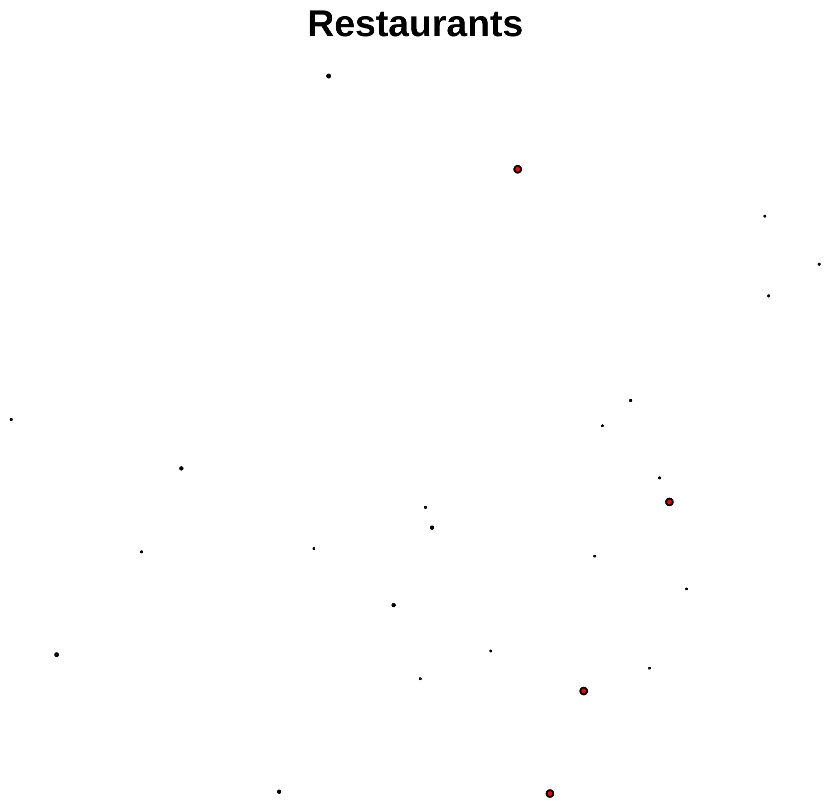
```
# Time range
#print(first(movements$datetime))
#print(last(movements$datetime))
#print(difftime(last(movements$datetime), first(movements$datetime)))
```

Parameters

```
# buffers for categories of restaurants (converted from ft. sq.)
buffer_full = calc_buffer(5000)
buffer_limited = calc_buffer(2000)
buffer_market = calc_buffer(22500)
buffer_supermarc = calc_buffer(33360)
buffer_warehouse = calc_buffer(187000)
buffer_conv = calc_buffer(2000)

# threshold for considering a new visit (in minutes)
time_gap_threshold <- 5
```

Convert restaurant set into sf object and add buffers (dependent on category)

```
# Convert restaurants to sf object. Keep "Name" and "Category" variables. Use Long and Lat for coordina
rest_sf <- st_as_sf(restaurants[, c("Name", "Category", "Longitude", "Latitude")], coords = c("Longitud

# Add column for appropriate buffers then apply them to the object
rest_sf <- rest_sf %>%
  mutate(buffer = case_when(
    Category == "Full-Service Restaurants" ~ buffer_full,
    Category == "Limited-Service Restaurants" ~ buffer_limited,
    Category == "Fruit & Vegetable Markets" ~ buffer_market,
    Category == "Supermarkets/Other Grocery (Exc Convenience) Strs" ~ buffer_supermarc,
    Category == "Warehouse Clubs & Supercenters" ~ buffer_warehouse,
    Category == "Convenience Stores" ~ buffer_conv
  ))
rest_sf <- st_buffer(rest_sf, dist = rest_sf$buffer)

plot(select(rest_sf, -Category, -buffer), col = "red", main = "Restaurants")
```



For each point in the movement set, find which (if any) restaurant's buffer it is within.

- Creates a data frame of TRUE and FALSE where movements are observations and each restaurant's buffer is a variable

```
# Use long and lat in movements data set against restaurant buffers using st_within
movements_within <- as.data.frame(st_within(st_as_sf(movements, coords = c("longitude", "latitude")), r
```

Reduce movements_within to only contain id of restaurant within or 0 if not within any. Bind the datetime as well.

```r
# If whole row is FALSE, set to 0, otherwise set to index of TRUE. Convert to a data frame.
movements_within <- data.frame(location = ifelse(rowSums(movements_within) == 0, 0, max.col(movements_wi

# Bind datetime from movements set (convert to POSIX as well)
movements_within <- cbind(movements_within, datetime = as.POSIXct(movements$datetime))
```

View visits

```r
head(movements_within)
```

```
##   location            datetime
## 1        0 2020-01-01 00:00:00
## 2        0 2020-01-01 00:01:07
## 3        0 2020-01-01 00:02:30
## 4        0 2020-01-01 00:03:40
## 5        0 2020-01-01 00:05:07
## 6        0 2020-01-01 00:05:54
```

Cleanup for memory

```
##             used  (Mb) gc trigger   (Mb)  max used   (Mb)
## Ncells   985125  52.7   48464152 2588.3  53591381 2862.1
## Vcells 20605188 157.3  299220747 2282.9 315877305 2410.0
```

Collect consecutive time spent in a location into a single "visit"

```r
# Create data frame for visits
visits <- data.frame(loc = numeric(0), start = character(0), end = character(0), stringsAsFactors = FALS

# New visit starts when prev location is different from current then use cumsum to give each an ID
movements_within <- movements_within %>%
  mutate(new_visit = location != lag(location, default = FALSE))
movements_within <- movements_within %>%
  mutate(visit_id = cumsum(new_visit))

# Filter out non-visits and reduce the group forming a visit (by visit_id) into location, start datetim
visits <- movements_within %>%
  filter(location != 0) %>%
  group_by(visit_id) %>%
  summarize(loc = first(location), start = first(datetime), end = last(datetime)) %>%
  ungroup() %>%
  select(-visit_id)
```

For consecutive visits to the same location, if the time between visits in under threshold (in minutes), merge into one.

- Could be stepping out to car, or to smoke, etc.

```
# Check if location of next visit is same location And if the gap in time is under threshold.
# If yes, time_gap = difftime, otherwise make it 0. If a gap is present, mark the leading row for remov
# and replace the end time with proper end. Then remove redundant (marked) and reduce to loc, start, en
visits <- visits %>%
  mutate (time_gap = ifelse(loc == lead(loc) & difftime(lead(start), end, units = "mins") <= time_gap_t
                      difftime(lead(start), end, units = "mins"), 0),
          rm = ifelse(lag(time_gap, default = 0) > 0, 1, 0),
          end = if_else(time_gap > 0, as.POSIXct(lead(end)), end)
         ) %>%
  filter(rm == 0) %>%
  select(-rm)
```

Visualize visits

```
# Column for visit length
visits <- mutate(visits, length = round(difftime(end, start, units = "mins"), 2))

# Replace loc with restaurant names and add categories from the data set
visits$Category <- restaurants$Category[visits$loc]
visits$loc <- restaurants$Name[visits$loc]

# Filter out visits under category dependent length
visits <- visits %>%
  filter(case_when(
    Category == "Full-Service Restaurants" ~ length > 10,
    Category == "Limited-Service Restaurants" ~ length > 5,
    Category == "Fruit & Vegetable Markets" ~ length > 5,
    Category == "Supermarkets/Other Grocery (Exc Convenience) Strs" ~ length > 5,
    Category == "Warehouse Clubs & Supercenters" ~ length > 5,
    Category == "Convenience Stores" ~ length > 2
  ))

# Tabular of min max mean stdev of visit by location, category, etc.
print("Visit Stats By Category")
```

```
## [1] "Visit Stats By Category"
```

```
category_stats <- visits %>%
  group_by(Category) %>%
  summarize(
    min = min(length),
    max = max(length),
    avg = mean(length),
    sdev = sd(length),
    total = n()
  )
print(category_stats)
```

```
## # A tibble: 6 x 6
##   Category                                          min   max   avg    sdev total
##   <chr>                                             <drt> <drt> <drt> <dbl> <int>
## 1 Convenience Stores                                 2.0~ 11.6~  3.6~ 1.21   8579
## 2 Fruit & Vegetable Markets                          5.0~ 23.7~  6.9~ 1.85   9604
## 3 Full-Service Restaurants                          10.0~ 64.5~ 12.0~ 8.32     42
## 4 Limited-Service Restaurants                        5.0~ 11.6~  5.8~ 0.837  1785
## 5 Supermarkets/Other Grocery (Exc Convenience) St~  5.0~ 25.4~  7.3~ 2.24   6290
## 6 Warehouse Clubs & Supercenters                     5.0~ 73.4~ 11.1~ 5.83  26500
```

```
# Tabular of min max mean stdev of visit by location, category, etc.
print("Visit Stats By Location")
```

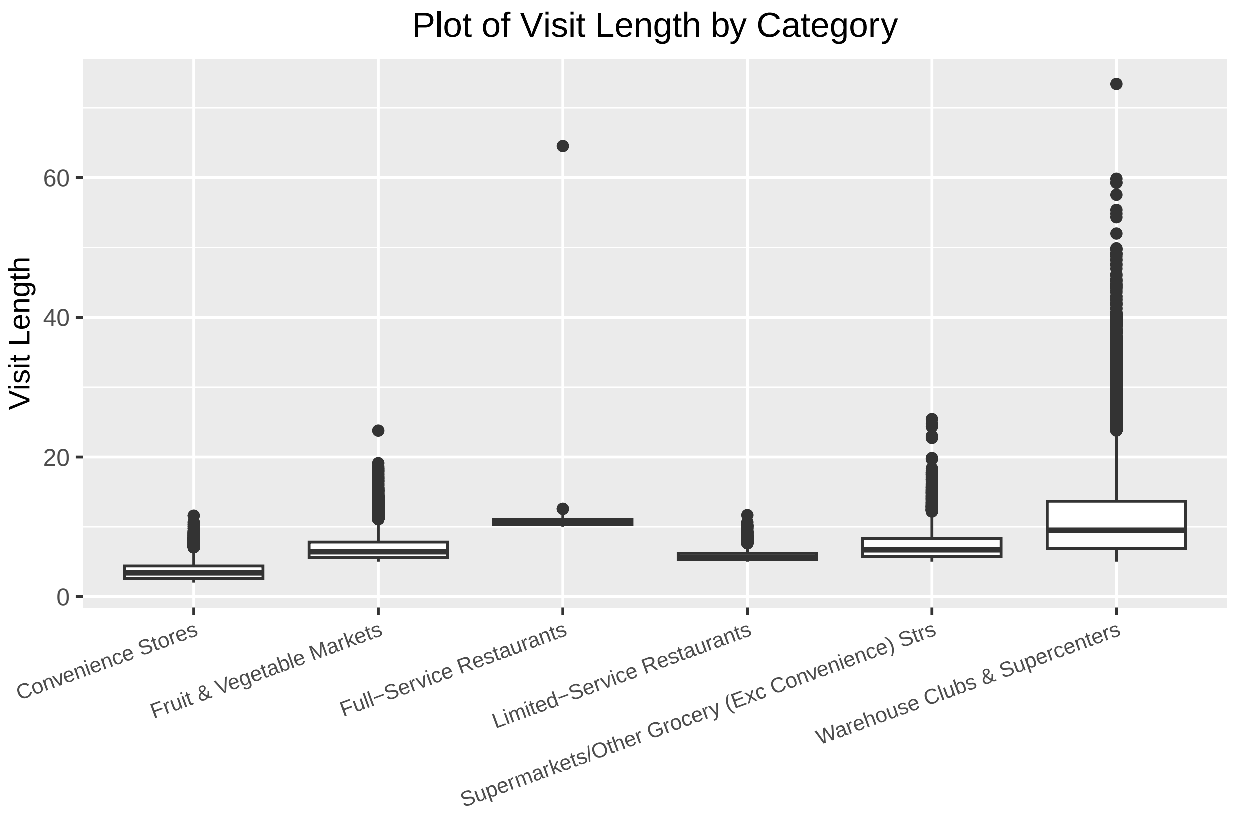
```
## [1] "Visit Stats By Location"
```

```
loc_stats <- visits %>%
  group_by(loc) %>%
  summarize(
    min = min(length),
    max = max(length),
    avg = mean(length),
    sdev = sd(length),
    total = n()
  )
print(loc_stats)
```

```
## # A tibble: 25 x 6
##    loc                    min        max        avg                sdev total
##    <chr>                  <drtn>     <drtn>     <drtn>            <dbl> <int>
##  1 Pullman Fine Dine 0    10.35 mins 11.08 mins 10.703333 mins  0.366     3
##  2 Pullman Fine Dine 14   10.13 mins 11.38 mins 10.811000 mins  0.453    10
##  3 Pullman Fine Dine 19   10.20 mins 11.70 mins 10.862500 mins  0.641     4
##  4 Pullman Fine Dine 21   10.20 mins 12.58 mins 11.044000 mins  0.935     5
##  5 Pullman Fine Dine 24   10.07 mins 11.03 mins 10.522500 mins  0.304     8
##  6 Pullman Fine Dine 3    10.02 mins 10.37 mins 10.241667 mins  0.137     6
##  7 Pullman Fine Dine 4    10.03 mins 64.53 mins 20.005000 mins 21.8       6
##  8 Pullman Fresh Market 12  5.02 mins 18.63 mins  7.004619 mins  1.86   2401
##  9 Pullman Fresh Market 16  5.02 mins 18.38 mins  6.964026 mins  1.79   2367
## 10 Pullman Fresh Market 2   5.02 mins 23.78 mins  7.017820 mins  1.85   2468
## # i 15 more rows
```

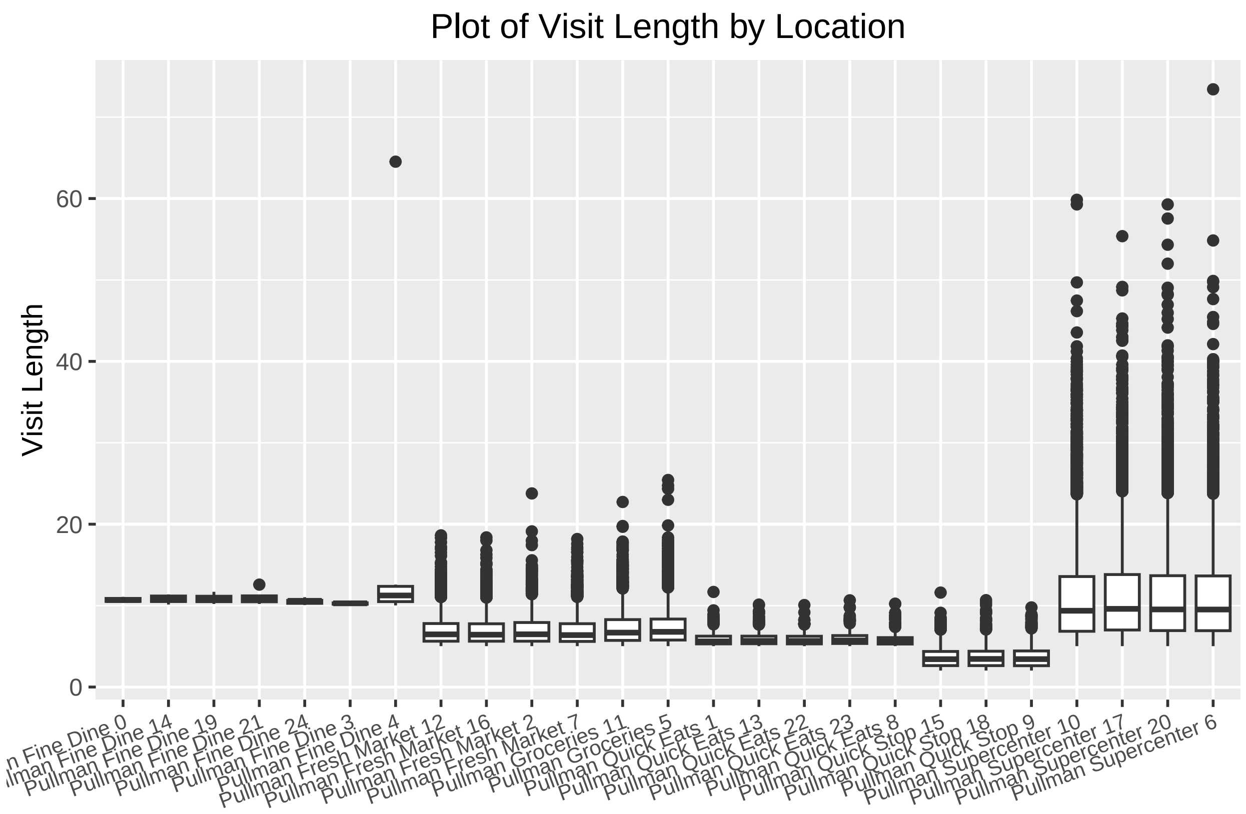
```
# Boxplot by Category
ggplot(visits, aes(x = Category, y = length)) +
  geom_boxplot() +
  labs(title = "Plot of Visit Length by Category", x="", y = "Visit Length") +
  theme(axis.text.x = element_text(angle = 20, hjust = 1, size = 8),
        plot.title = element_text(hjust = 0.5))
```

```
## Don't know how to automatically pick scale for object of type <difftime>.
## Defaulting to continuous.
```



```r
# Boxplot by Location
ggplot(visits, aes(x = loc, y = length)) +
  geom_boxplot() +
  labs(title = "Plot of Visit Length by Location", x="", y = "Visit Length") +
  theme(axis.text.x = element_text(angle = 20, hjust = 1, size = 8),
        plot.title = element_text(hjust = 0.5))
```

```
## Don't know how to automatically pick scale for object of type <difftime>.
## Defaulting to continuous.
```



```r
head(visits)
```

```
## # A tibble: 6 x 6
##   loc           start               end                 time_gap length Category
##   <chr>         <dttm>              <dttm>                 <dbl> <drtn> <chr>
## 1 Pullman Supe~ 2020-01-01 05:37:20 2020-01-01 05:50:30     0    13.17~ Warehou~
## 2 Pullman Supe~ 2020-01-01 12:39:18 2020-01-01 12:49:59     1.95 10.68~ Warehou~
## 3 Pullman Quic~ 2020-01-01 13:50:37 2020-01-01 13:53:53     2.17  3.27~ Conveni~
## 4 Pullman Supe~ 2020-01-01 16:08:03 2020-01-01 16:17:36     2.68  9.55~ Warehou~
## 5 Pullman Supe~ 2020-01-01 16:54:25 2020-01-01 17:12:16     2.97 17.85~ Warehou~
## 6 Pullman Supe~ 2020-01-01 18:52:39 2020-01-01 19:06:48     3.32 14.15~ Warehou~
```