hw4 : Bike Sharing in Goi창nia, Brazil

### Data 변수 선택

names(brz\_gyn)

## [1] "X1" "address" "available" "created\_at" "free"   
## [6] "lat" "lng" "name" "source" "status"

df <- brz\_gyn[,-c(1, 9)]

변수 중에서 X1과 source 제거

summary(df)

## address available created\_at   
## Length:99644 Min. : 0.000 Min. :2017-09-13 15:25:37   
## Class :character 1st Qu.: 3.000 1st Qu.:2017-09-25 04:10:10   
## Mode :character Median : 5.000 Median :2017-10-07 00:00:11   
## Mean : 4.987 Mean :2017-10-07 07:01:03   
## 3rd Qu.: 7.000 3rd Qu.:2017-10-19 09:30:05   
## Max. :12.000 Max. :2017-10-31 18:10:10   
## free lat lng name   
## Min. : 0.000 Min. :-16.71 Min. :-49.28 Length:99644   
## 1st Qu.: 5.000 1st Qu.:-16.70 1st Qu.:-49.27 Class :character   
## Median : 7.000 Median :-16.69 Median :-49.26 Mode :character   
## Mean : 7.013 Mean :-16.69 Mean :-49.26   
## 3rd Qu.: 9.000 3rd Qu.:-16.68 3rd Qu.:-49.26   
## Max. :12.000 Max. :-16.67 Max. :-49.25   
## status   
## Length:99644   
## Class :character   
## Mode :character   
##   
##   
##

length(unique(df$name))

## [1] 15

summary(df$created\_at)

## Min. 1st Qu. Median   
## "2017-09-13 15:25:37" "2017-09-25 04:10:10" "2017-10-07 00:00:11"   
## Mean 3rd Qu. Max.   
## "2017-10-07 07:01:03" "2017-10-19 09:30:05" "2017-10-31 18:10:10"

df <- df%>% select(name, address, status, lat, lng, free, available, created\_at)

15개의 station 존재

2017년 9월~ 10월 한달 반쯤 자료

### Data missing 살펴보기

# 라틴어 -> 영어로 바꾸기  
df$name <- stringi::stri\_trans\_general(df$name, "Latin-ASCII")  
  
# Bosque dos Buritis 데이터 개수 다름.  
table(df$name)

##   
## Areiao Bandeirante Bosque dos Buritis   
## 6922 6922 2736   
## Bougainville Lago das Rosas Marista   
## 6922 6922 6922   
## Paranaiba Parque Vaca Brava Praca Civica   
## 6922 6922 6922   
## Praca do Sol Praca Tamandare Praca Universitaria   
## 6922 6922 6922   
## Ricardo Paranhos T63 UNIMED   
## 6922 6922 6922

# 연 - 월 - 일 - 시간 - 분 - 초 쪼개기  
df <- df %>% mutate(time = as.character(created\_at))  
df <- df %>% separate(time, c("date", "hour"), sep=" ")  
df <- df %>% separate(date, c("year", "month", "day"), sep="-")  
df <- df %>% separate(hour, c("hour", "min", "sec"), sep=":")  
  
df %>% group\_by(year, month, day, hour, min, sec) %>% count() %>% filter(n!=15)

## # A tibble: 4,496 x 7  
## # Groups: year, month, day, hour, min, sec [4,496]  
## year month day hour min sec n  
## <chr> <chr> <chr> <chr> <chr> <chr> <int>  
## 1 2017 09 13 21 30 15 14  
## 2 2017 09 13 21 30 16 1  
## 3 2017 09 14 03 00 17 13  
## 4 2017 09 14 03 00 18 2  
## 5 2017 09 14 04 00 05 12  
## 6 2017 09 14 04 00 06 3  
## 7 2017 09 14 07 40 06 5  
## 8 2017 09 14 07 40 07 10  
## 9 2017 09 14 09 10 09 9  
## 10 2017 09 14 09 10 10 6  
## # ... with 4,486 more rows

df %>% group\_by(year, month, day, hour, min) %>% count() %>% filter(n!=15)

## # A tibble: 4,185 x 6  
## # Groups: year, month, day, hour, min [4,185]  
## year month day hour min n  
## <chr> <chr> <chr> <chr> <chr> <int>  
## 1 2017 10 02 15 30 14  
## 2 2017 10 02 15 40 14  
## 3 2017 10 02 15 50 14  
## 4 2017 10 02 16 00 14  
## 5 2017 10 02 16 10 14  
## 6 2017 10 02 16 20 14  
## 7 2017 10 02 16 30 14  
## 8 2017 10 02 16 40 14  
## 9 2017 10 02 16 50 14  
## 10 2017 10 02 17 00 14  
## # ... with 4,175 more rows

df %>% group\_by(year, month, day, hour, min) %>% count() %>% filter(n!=15) %>% filter(n!=14)

## # A tibble: 1 x 6  
## # Groups: year, month, day, hour, min [1]  
## year month day hour min n  
## <chr> <chr> <chr> <chr> <chr> <int>  
## 1 2017 10 02 20 20 28

데이터 설명과 달리 5분 마다 데이터가 수집된 것 아님.

측정 초 차이 존재

중복 측정 가능성.

### 데이터 삭제

df %>% filter(year == "2017", month == "10", day == "02", hour == "20", min == "20") %>% select(name, free, available, sec) %>% print(n=Inf)

## # A tibble: 28 x 4  
## name free available sec   
## <chr> <dbl> <dbl> <chr>  
## 1 Paranaiba 2 10 08   
## 2 Bandeirante 8 4 08   
## 3 Praca Universitaria 4 8 08   
## 4 Praca Civica 4 8 08   
## 5 Praca Tamandare 5 7 08   
## 6 Lago das Rosas 8 4 08   
## 7 UNIMED 6 6 08   
## 8 Praca do Sol 6 6 08   
## 9 Bougainville 10 2 08   
## 10 Marista 12 0 08   
## 11 Areiao 6 6 08   
## 12 Ricardo Paranhos 11 1 08   
## 13 Parque Vaca Brava 6 6 08   
## 14 T63 8 4 08   
## 15 Paranaiba 2 10 19   
## 16 Bandeirante 8 4 19   
## 17 Praca Universitaria 4 8 19   
## 18 Praca Civica 4 8 19   
## 19 Praca Tamandare 5 7 19   
## 20 Lago das Rosas 8 4 19   
## 21 UNIMED 6 6 19   
## 22 Praca do Sol 6 6 19   
## 23 Bougainville 10 2 19   
## 24 Marista 12 0 19   
## 25 Areiao 6 6 19   
## 26 Ricardo Paranhos 11 1 19   
## 27 Parque Vaca Brava 6 6 19   
## 28 T63 8 4 19

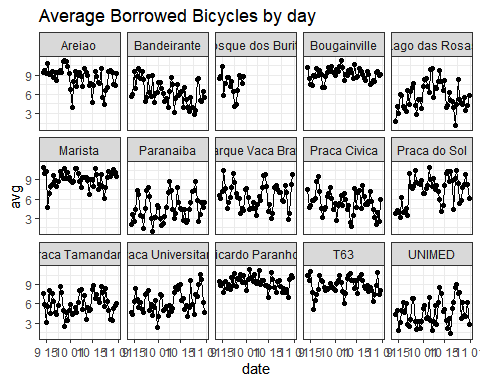
df <- df %>% filter(month != "10"| day != "02"| hour != "20"| min != "20"| sec != "19")

2017-10-02 20:20분에 기록이 2번 일어남

임의로 19초를 그냥 지우기로 결정.

### 1. Average Borrowed Bicycles by day

df <- df %>% mutate(weekday = weekdays(created\_at))  
df$weekday <- factor(df$weekday, levels = c("월요일", "화요일", "수요일", "목요일", "금요일", "토요일", "일요일"))  
  
temp <- df %>% unite("date", c(year, month, day), sep = "-")  
temp$date <- as.Date(temp$date)  
temp %>% filter(status == "A") %>% group\_by(name, date)%>% summarise(avg = mean(free)) %>% ggplot(aes(date, avg)) + geom\_path(aes(group=1)) + geom\_point() + facet\_wrap(~name, ncol = 5) + labs(title = "Average Borrowed Bicycles by day") + theme\_bw()



free = 자전거를 빌려가서 없는 곳 = 자전거를 반납할 수 있는 공간

temp[temp$name == "Bosque dos Buritis", ] %>% select(date) %>% tail()

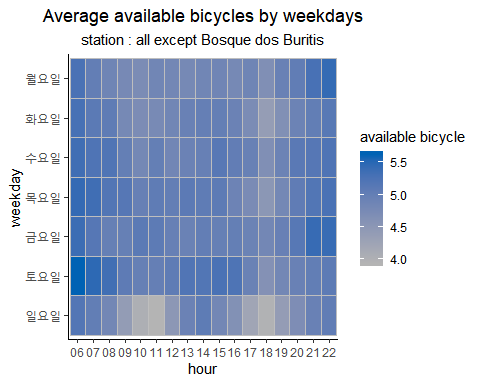
## # A tibble: 6 x 1  
## date   
## <date>   
## 1 2017-10-02  
## 2 2017-10-02  
## 3 2017-10-02  
## 4 2017-10-02  
## 5 2017-10-02  
## 6 2017-10-02

df$date <- temp$date   
rm(temp)

Bosque dos Buritis 10월 2일에서 관측중단.

### 2. Heat Map : Average Available Bicycles by weekdays by stations

df %>% filter(name != "Bosque dos Buritis", status == "A", hour %in% c("06", "07", "08", "09", "10", "11", "12", "13", "14", "15", "16", "17", "18", "19", "20", "21", "22")) %>% group\_by(hour, weekday) %>% summarise(available\_mean = mean(available)) %>% ggplot(aes(hour, weekday, fill = available\_mean)) + geom\_tile(col="gray") +  
 labs(title = "Average available bicycles by weekdays", subtitle = "station : all except Bosque dos Buritis") + scale\_fill\_gradient("available bicycle", low = "#B4B4B4FF", high = "#0062B4FF")+  
 theme\_classic()+theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5))+ scale\_y\_discrete(limits = c("일요일", "토요일", "금요일", "목요일", "수요일", "화요일", "월요일"))



available = 빌릴 있는 자전거 개수 = 빌릴 수 있는 시간 오전 6시 ~ 오후 10시 59분

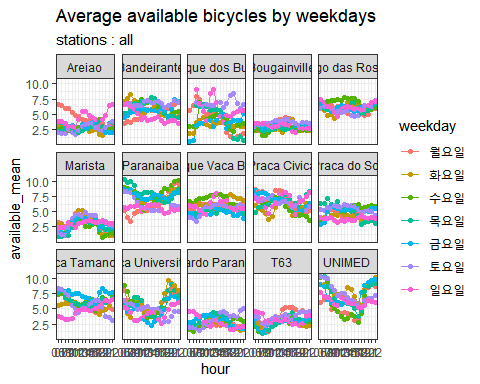
반납 가능 시간 24hours

Bosque dos Buritis 제외, status = Active만

일요일 아침 10시-11시/ 오후 17시-18시 빌릴 있는 자전거 가장 적을 때

### 3. Line graph : Average available bicycles by weekdays by stations

df%>% filter(status == "A" , hour %in% c("06", "07", "08", "09", "10", "11", "12", "13", "14", "15", "16", "17", "18", "19", "20", "21", "22")) %>% group\_by(name, hour, weekday) %>% summarise(available\_mean = mean(available)) %>%   
 ggplot(aes(hour, available\_mean)) + geom\_path(aes(group = weekday, col = weekday)) + geom\_point(aes(col = weekday)) + facet\_wrap(~name, ncol = 5) + labs(title = "Average available bicycles by weekdays", subtitle = "stations : all")+  
 theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5), axis.text.x = element\_text(angle = 90)) + theme\_bw()



Bosque dos Buritis 포함

각 station마다 일정한 패턴이 보인다.

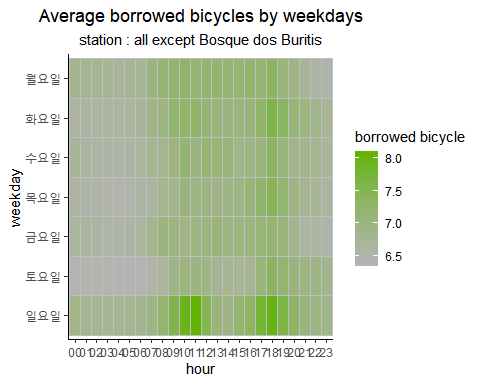
Bougainville, T63, Ricardo Paranhos, Areiao : 5개 미만을 유지, Areiao의 월요일 특이한 패턴. 일요일의 22시 선과 월요일의 6시 선이 연결됨.

Marista : 오전중에는 available bicycle이 없다가 오후에 증가.다시 퇴근시간에 줄어듦.

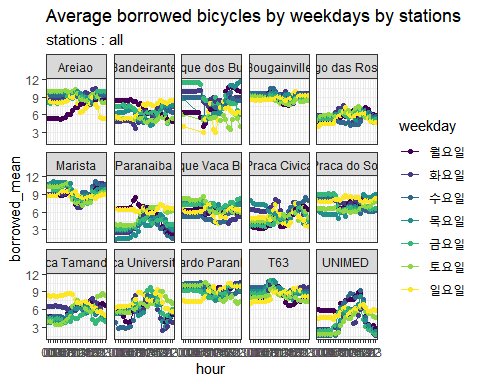
Parque Vaca Brava, Praca Civica, Lago das Rosas Bandeirante 5~7개 사이를 늘 오간다. 중상층에 위치

### 4. (상) Average borrowed bicycles by weekdays (하) Average borrowed bicycles by weekdays

df %>% filter(name != "Bosque dos Buritis", status == "A") %>% group\_by(hour, weekday) %>% summarise(borrowed\_mean = mean(free)) %>% ggplot(aes(hour, weekday, fill = borrowed\_mean)) + geom\_tile(col="gray") +  
 labs(title = "Average borrowed bicycles by weekdays", subtitle = "station : all except Bosque dos Buritis") + scale\_fill\_gradient("borrowed bicycle", low = "#B4B4B4FF", high = "#62B400FF")+  
 theme\_classic()+theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5))+ scale\_y\_discrete(limits = c("일요일", "토요일", "금요일", "목요일", "수요일", "화요일", "월요일"))



df%>% filter(status == "A") %>% group\_by(name, hour, weekday) %>% summarise(borrowed\_mean = mean(free)) %>%   
 ggplot(aes(hour, borrowed\_mean)) + geom\_path(aes(group = weekday, col = weekday)) + geom\_point(aes(col = weekday)) + facet\_wrap(~name, ncol = 5) + labs(title = "Average borrowed bicycles by weekdays by stations", subtitle = "stations : all")+  
 theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5), axis.text.x = element\_text(angle = 90)) + theme\_bw() + scale\_color\_viridis\_d()



(상), (하) 그래프는 available과 반대되는 free를 그린것으로 #2, #3의 그래프와 반대의 양상보여준다.

### free, available missing 없음. 항상 합 12

sum(rowSums(df[,c("free", "available")]) != 12)

## [1] 0

### preparation for Map

### 5. “Station Location”

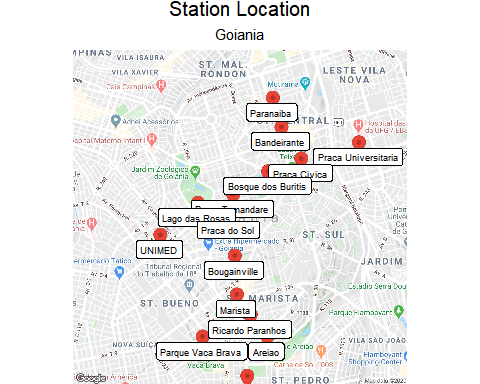
loc <- data.frame(long = unique(df$lng), lat = unique(df$lat))  
gyn <- get\_googlemap('Goiânia', markers = loc, scale = 2, zoom = 14)

## Source : https://maps.googleapis.com/maps/api/staticmap?center=Goiania&zoom=14&size=640x640&scale=2&maptype=terrain&markers=-16.670492,-49.259535%7C-16.675068,-49.258084%7C-16.677528,-49.245354%7C-16.680073,-49.254844%7C-16.682072,-49.260273%7C-16.685711,-49.266029%7C-16.687036,-49.271852%7C-16.692067,-49.278014%7C-16.688996,-49.266675%7C-16.695357,-49.265774%7C-16.701461,-49.265447%7C-16.708108,-49.260434%7C-16.704655,-49.263153%7C-16.708095,-49.27105%7C-16.714282,-49.263722&key=xxx-52vngY38jjKx6jV4iyDmFWuiw

## Source : https://maps.googleapis.com/maps/api/geocode/json?address=Goiania&key=xxx-52vngY38jjKx6jV4iyDmFWuiw

loc\_name <- cbind(unique(df$name), loc)  
names(loc\_name)[1] <- "name"  
gynmap <- ggmap(gyn, extent = "device", base\_layer = ggplot(aes(x=long, y = lat), data = loc\_name))  
gynmap + geom\_label(label = loc\_name$name, size = 2.5) + labs(title = "Station Location", subtitle = "Goiânia") + theme(plot.title = element\_text(hjust = 0.5, size = 15), plot.subtitle = element\_text(hjust = 0.5))

## Warning: Removed 1 rows containing missing values (geom\_label).



Zoom으로 인해 가장 아래에 위치하는 T63이 사라짐.

### 6. (가장 의미를 많이 함축한 그림)Availability Rate by weekday by station

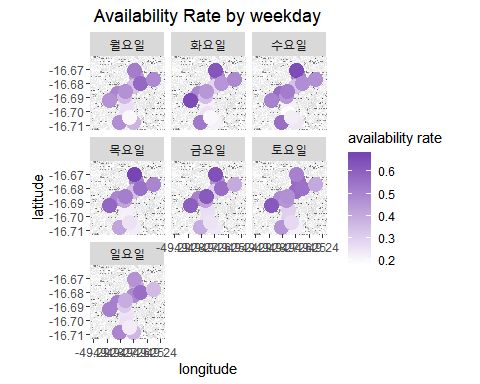
temp <- df %>%filter(status == "A", hour %in% c("06", "07", "08", "09", "10", "11", "12", "13", "14", "15", "16", "17", "18", "19", "20", "21", "22")) %>% group\_by(name, lat, lng, weekday) %>% summarise(avg = mean(available)) %>%  
 mutate(rate = avg/12)  
gyn\_temp <- get\_map(location = 'Goiânia', zoom = 14, color = "bw")

## Source : https://maps.googleapis.com/maps/api/staticmap?center=Goiania&zoom=14&size=640x640&scale=2&maptype=terrain&language=en-EN&key=xxx-52vngY38jjKx6jV4iyDmFWuiw

## Source : https://maps.googleapis.com/maps/api/geocode/json?address=Goiania&key=xxx-52vngY38jjKx6jV4iyDmFWuiw

gyn\_map\_temp <- ggmap(gyn\_temp, base\_layer = ggplot(aes(x = lng, y = lat), data = temp))  
gyn\_map\_temp + geom\_point(aes(color = rate), size = 5) + facet\_wrap(~weekday) + scale\_color\_gradient("availability rate", low = "white", high = "#7846B4") + labs(x = "longitude", y = "latitude", title = "Availability Rate by weekday")+ theme(plot.title = element\_text(hjust = 0.5))

## Warning: Removed 7 rows containing missing values (geom\_point).



zoom으로 인해 T63 스테이션 사라짐.

비교적 상단에 위치한 station이 availability가 높은편이고 아래에 위치한 station일수록 availability가 낮다.

특히 paraniba는 평일에 availability가장 높은편.

아래쪽에 위치한 Parque Vaca Brava도 비교적 높은 availability가 유지되는 편.

위의 line graph에서 살펴봤던것처럼 availability가 시간대별 그리고 요일별로 안정적인 station과 변동범위가 컸던 station을 고려해줘야한다.

=> 교통 편의를 향상시키기 위해 추가적인 데이터를 수집하여 자전거 station의 특징별로 clustering해줄 필요있다.