hw3-NYC

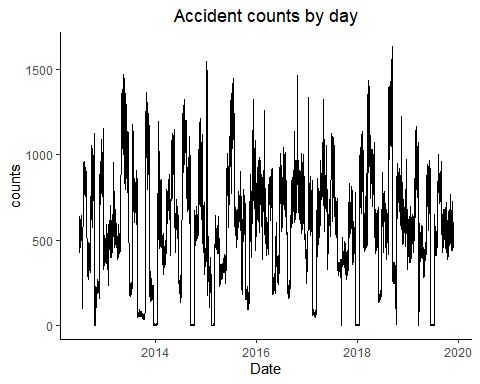
### 데이터 로드

## Parsed with column specification:  
## cols(  
## .default = col\_character(),  
## `ACCIDENT DATE` = col\_datetime(format = ""),  
## `ACCIDENT TIME` = col\_time(format = ""),  
## `ZIP CODE` = col\_double(),  
## LATITUDE = col\_double(),  
## LONGITUDE = col\_double(),  
## `NUMBER OF PERSONS INJURED` = col\_double(),  
## `NUMBER OF PERSONS KILLED` = col\_double(),  
## `NUMBER OF PEDESTRIANS INJURED` = col\_double(),  
## `NUMBER OF PEDESTRIANS KILLED` = col\_double(),  
## `NUMBER OF CYCLIST INJURED` = col\_double(),  
## `NUMBER OF CYCLIST KILLED` = col\_double(),  
## `NUMBER OF MOTORIST INJURED` = col\_double(),  
## `NUMBER OF MOTORIST KILLED` = col\_double(),  
## COLLISION\_ID = col\_double()  
## )

## See spec(...) for full column specifications.

### 1. 일별로 본 교통 사고 횟수

dat %>% group\_by(ACCIDENT.DATE) %>% count() %>% ggplot(aes(ACCIDENT.DATE, n)) + geom\_line() +theme\_classic()+  
 labs(x="Date", y = "counts", title = "Accident counts by day") + theme(plot.title = element\_text(hjust = 0.5))



일자별 교통 사고 건수를 살펴보았을 때 시계열 패턴이 보이는 듯하다. 그러나 직관적으로 이해하기 어렵다.

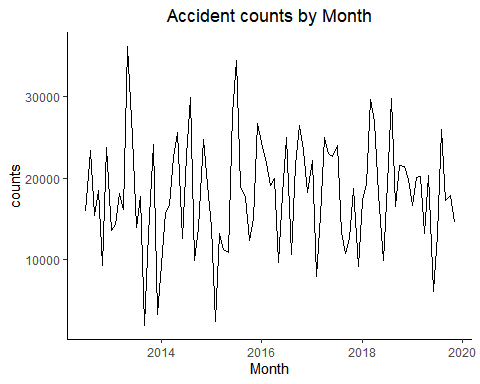
### 2. 달별로 본 교통 사고 횟수

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

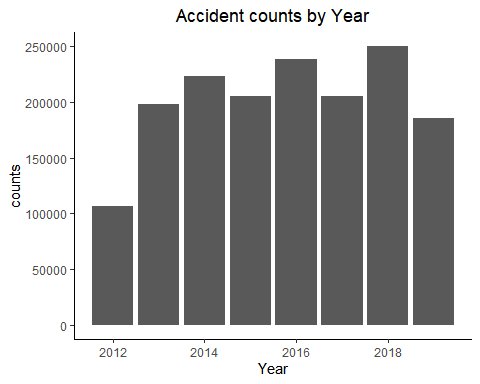
dat %>% mutate(Month = floor\_date(ACCIDENT.DATE, "month")) %>% group\_by(Month) %>% count() %>% ggplot(aes(Month, n)) + geom\_line()+theme\_classic()+ labs(x="Month", y = "counts", title = "Accident counts by Month") +theme(plot.title = element\_text(hjust = 0.5))



월별 교통 사고 건수를 보았을 때 시계열 패턴이 보이는 듯하다. 주기적이 peak가 보이지만 해석하기 어렵다.

### 3. 연별로 본 교통 사고 횟수

dat %>% mutate(Year = floor\_date(ACCIDENT.DATE, "year")) %>% group\_by(Year) %>% count() %>% ggplot(aes(Year, n)) + geom\_bar(stat = "identity")+theme\_classic()+ labs(x="Year", y = "counts", title = "Accident counts by Year") +theme(plot.title = element\_text(hjust = 0.5))



연별 교통 사고 건수를 보면 2012년과 2019년이 적어보이지만 이는 데이터가 완성형아니기 때문이다. 아래를 보자.

### 데이터의 시간

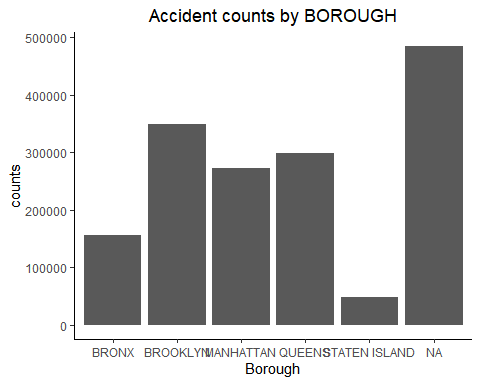
## Min. 1st Qu. Median   
## "2012-07-01 00:00:00" "2014-06-15 00:00:00" "2016-04-10 00:00:00"   
## Mean 3rd Qu. Max.   
## "2016-04-03 12:22:05" "2018-02-21 00:00:00" "2019-11-26 00:00:00"

2012년 7월 ~ 12월 6개월

2019년 11월 26일까지밖에 자료가 없음.

### 4. BOROUGH별로 본 교통 사고 횟수

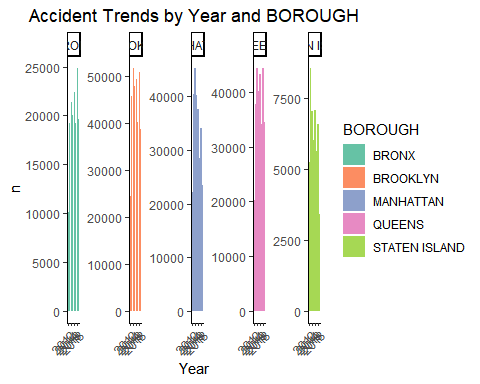
dat %>% group\_by(BOROUGH) %>% count() %>% ggplot(aes(BOROUGH, n)) + geom\_bar(stat="identity")+theme\_classic()+ labs(x="Borough", y = "counts", title = "Accident counts by BOROUGH") +theme(plot.title = element\_text(hjust = 0.5))



교통 사고 건수 순 : NA - BROOKLYN - QUEENS - MANHATTAN - BRONX - STATEN ISLAND

### 5. Year and BOROUGH별로 본 교통 사고 횟수 추이

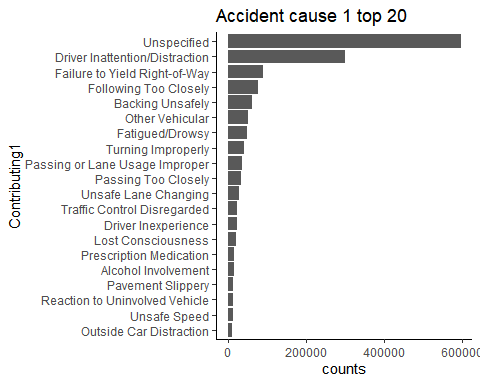
dat %>% filter(!is.na(BOROUGH)) %>% mutate(Year = floor\_date(ACCIDENT.DATE, "year")) %>% group\_by(Year, BOROUGH) %>% count() %>%  
 ggplot(aes(Year, n, fill = BOROUGH)) + geom\_bar(stat = "identity") + facet\_wrap(~BOROUGH, scales = "free\_y", ncol = 5)+  
 scale\_fill\_brewer(palette = "Set2") + theme\_classic() + labs(title = "Accident Trends by Year and BOROUGH")+  
 theme(plot.title = element\_text(hjust = 0.5), axis.text.x = element\_text(hjust = 1, angle = 45))



STATEN ISLAND과 MANHATTAN이 해가 감에 따라 감소 추세 인듯보인다.

### 6. CONTRIBUTION.VEHICLE.1 TOP20

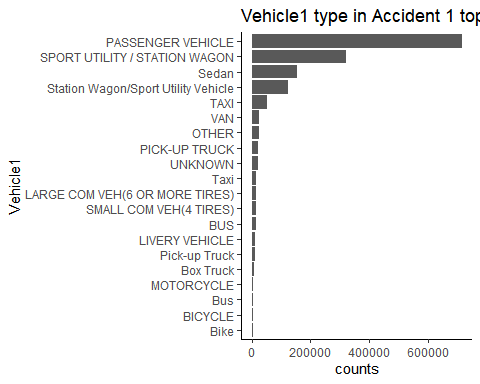
dat %>% filter(!is.na(CONTRIBUTING.FACTOR.VEHICLE.1)) %>% group\_by(CONTRIBUTING.FACTOR.VEHICLE.1) %>% count() %>% arrange(-n)%>% head(n =20) %>%   
 ggplot(aes(reorder(CONTRIBUTING.FACTOR.VEHICLE.1,n), n))+geom\_bar(stat="identity") + coord\_flip() + labs(y = "counts", x = "Contributing1", title = "Accident cause 1 top 20") + theme\_classic()



unspecified - Driver Inattention/distraction - Failure to Yield Right-of-Way 순이다.

### 7. VEHICLE.TYPE.1 TOP20

dat %>% filter(!is.na(VEHICLE.TYPE.CODE.1)) %>% group\_by(VEHICLE.TYPE.CODE.1) %>% count() %>% arrange(-n)%>% head(n =20) %>%   
 ggplot(aes(reorder(VEHICLE.TYPE.CODE.1,n), n))+geom\_bar(stat="identity") + coord\_flip() + labs(x = "Vehicle1", y = "counts", title = "Vehicle1 type in Accident 1 top 20") + theme\_classic()



Passenger vehicle - sport utility/station wagon - sedan 순이다. TAXI = Taxi, BUS = Bus 같은 것으로 보여 합쳐줘야함.

### 2012년과 2019년 데이터 삭제

dat3 <- dat  
dat3$ACCIDENT.DATE.2 <- dat3$ACCIDENT.DATE  
dat3$ACCIDENT.DATE.2 <- as.character(dat3$ACCIDENT.DATE.2)  
dat3 <- dat3 %>% separate(ACCIDENT.DATE.2, c("y", "m", "d"), sep="-")  
dat3 <- dat3[-which(dat3$y %in% c("2012", "2019") ), ]

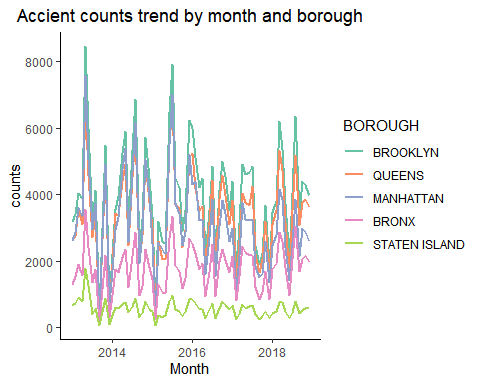
2013~2018년의 완성된 데이터로 5년 추이를 보기 위해 미완성의 2012년과 2019년 데이터 삭제 dat3 생성

### 교통사고 건수 합계로 BOROUGH level 조정

dat3$BOROUGH <- factor(dat3$BOROUGH, levels = c("BROOKLYN", "QUEENS", "MANHATTAN", "BRONX", "STATEN ISLAND"))

### 8. 교통사고 건수 구별 달별

dat3 %>% filter(!is.na(BOROUGH)) %>% mutate(Month = floor\_date(ACCIDENT.DATE, "month")) %>% group\_by(Month, BOROUGH) %>% count() %>% ggplot(aes(Month, n)) + geom\_line(aes(group = BOROUGH, col = BOROUGH), size = 1) +  
 scale\_color\_brewer(palette = "Set2") + theme\_classic() + labs(y="counts", title = "Accient counts trend by month and borough") +  
 theme(plot.title = element\_text(hjust = 0.5))



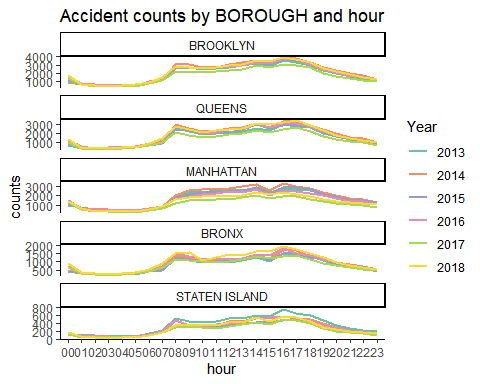
스테이튼 아일랜드를 제외한 나머지는 비슷한 패턴을 보이는 듯하다. 특히 브루클린 퀸즈 그리고 맨하탄은 수치도 비슷하여 그래프가 겹친다. 뉴욕 4개의 구가 달별로 비슷한 교통사고 건수 패턴을 보인다는 것을 확인할 수 있다. 매년 달별로 어떤 패턴이 보인다기보다는 뉴욕의 4개의 구가 비슷한 패턴으로 교통사고 건수가 증가감소한다. 뉴욕의 교통이 서로 연관있기 때문일지 생각해볼 수 있다.

### ACCIDENT.TIEM을 시간 - 분 - 초 분리

dat3$ACCIDENT.TIME2 <- as.character(dat3$ACCIDENT.TIME)  
dat3 <- dat3 %>% separate(ACCIDENT.TIME2, c("h", "mm", "ss"), sep=":") %>% select(-mm, -ss)

### 9. 교통사고 건수 구별 해별 시간별 건수 추이

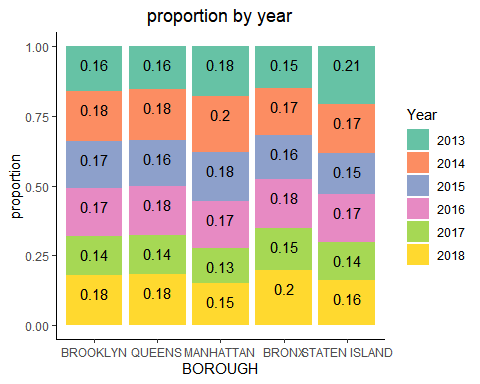
dat3 %>% filter(!is.na(BOROUGH)) %>% group\_by(BOROUGH, y, h) %>% count() %>% ggplot(aes(h, n)) + geom\_line(aes(group=y, col = y), size = 1)+  
 theme\_classic() + scale\_color\_brewer(name = "Year", palette = "Set2") + facet\_wrap(~BOROUGH, ncol = 1, scales = "free\_y") + labs(x = "hour", y ="counts", title = "Accident counts by BOROUGH and hour")+  
 theme(plot.title = element\_text(hjust = 0.5))



구별로 매해 시간대별 교통사고 건수는 비슷해 보인다. 특징적인 건 맨해탄의 2014년 스테이튼 아일랜드의 2013년 브롱스의 2018

### 10. 교통 사고 전체 발생 건수 중에서 년도별 차지 비중

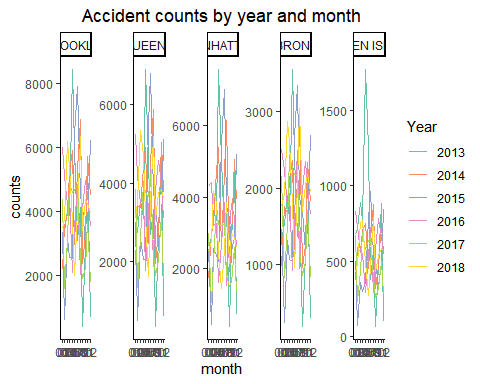
total <- dat3 %>% filter(!is.na(BOROUGH)) %>% group\_by(BOROUGH, y) %>% count() %>% group\_by(BOROUGH) %>% summarise(sum=sum(n))  
dat3 %>% filter(!is.na(BOROUGH)) %>% group\_by(BOROUGH, y) %>% count() %>% left\_join(total, by = "BOROUGH") %>% mutate(label = round(n/sum,2)) %>%  
 ggplot(aes(BOROUGH, n, fill = y, label = label, vjust =2)) + geom\_bar(stat='identity', position="fill") + geom\_text(position='fill')+scale\_fill\_brewer(name = "Year", palette = "Set2") +  
 labs(y = "proportion", title = "proportion by year") + theme\_classic() + theme(plot.title = element\_text(hjust = 0.5))



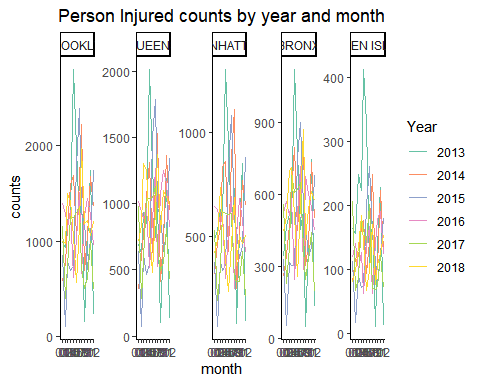
브루클린과 퀸즈는 비율이 똑같다. 맨해튼은 2014년이 가장 많다. 스테이튼 아일랜드는 2013년 브롱스는 2018년의 비율가 가장 크다.

### 11. 교통사고 발생 건수와 PERSON INJURED가 달별로 비슷한 추이를 보여주는 그래프

dat3 %>% filter(!is.na(BOROUGH)) %>% group\_by(BOROUGH, y, m) %>% count() %>% ggplot(aes(m, n, col = y, group = y)) + geom\_line() + facet\_wrap(~BOROUGH, ncol = 5, scales="free\_y") + labs(x="month", y="counts", title = "Accident counts by year and month")+theme\_classic() + theme(plot.title = element\_text(hjust = 0.5)) + scale\_color\_brewer(name = "Year", palette = "Set2")



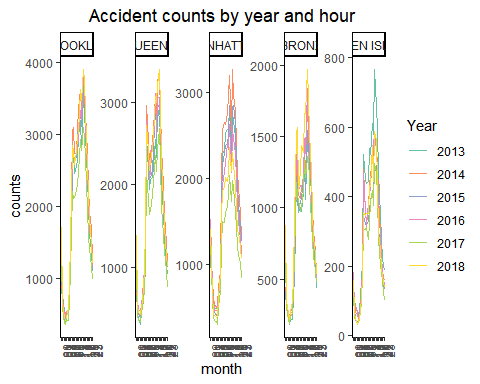
dat3 %>% filter(!is.na(BOROUGH)) %>% group\_by(BOROUGH, y, m) %>% summarise(sum = sum(NUMBER.OF.PERSONS.INJURED, na.rm=T), n = n()) %>%  
 ggplot(aes(m, sum, col = y, group = y)) + geom\_line() + facet\_wrap(~BOROUGH, ncol = 5, scales="free\_y") + labs(x="month", y="counts", title = "Person Injured counts by year and month")+theme\_classic() + theme(plot.title = element\_text(hjust = 0.5)) + scale\_color\_brewer(name = "Year", palette = "Set2")



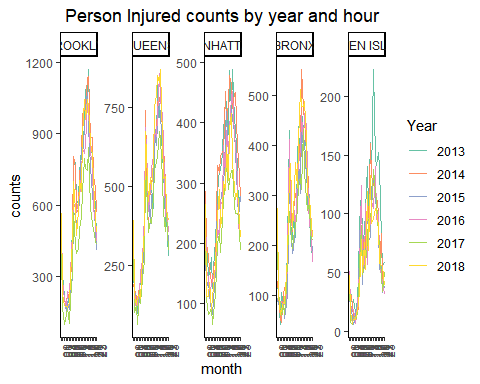
형태를 알기 어렵다. 두 그래프 모두 그렇다. 이 그래프만 봐서는 교통사고 발생 건수와 PERSON INJURED추이가 같은지 여부를 알기 어렵다.

### 12. 교통사고 발생 건수와 PERSON INJURED가 시간대별로 비슷한 추이를 보여주는 그래프

dat3 %>% filter(!is.na(BOROUGH)) %>% group\_by(BOROUGH, y, h) %>% count() %>% ggplot(aes(h, n, col = y, group = y)) + geom\_line() + facet\_wrap(~BOROUGH, ncol = 5, scales="free\_y") + labs(x="month", y="counts", title = "Accident counts by year and hour")+theme\_classic() + theme(plot.title = element\_text(hjust = 0.5),axis.text.x = element\_text(hjust = 1, angle = 90)) + scale\_color\_brewer(name = "Year", palette = "Set2")



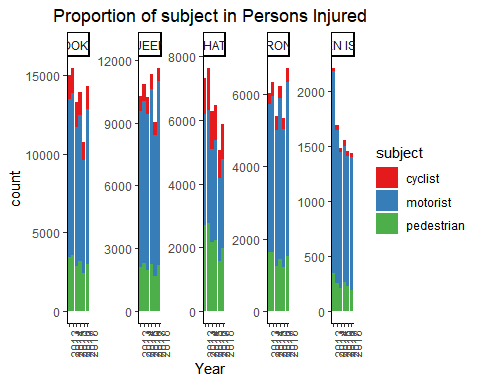
dat3 %>% filter(!is.na(BOROUGH)) %>% group\_by(BOROUGH, y, h) %>% summarise(sum = sum(NUMBER.OF.PERSONS.INJURED, na.rm=T), n = n()) %>%  
 ggplot(aes(h, sum, col = y, group = y)) + geom\_line() + facet\_wrap(~BOROUGH, ncol = 5, scales="free\_y")+ labs(x="month", y="counts", title = "Person Injured counts by year and hour")+theme\_classic() + theme(plot.title = element\_text(hjust = 0.5),axis.text.x = element\_text(hjust = 1, angle = 90)) + scale\_color\_brewer(name = "Year", palette = "Set2")



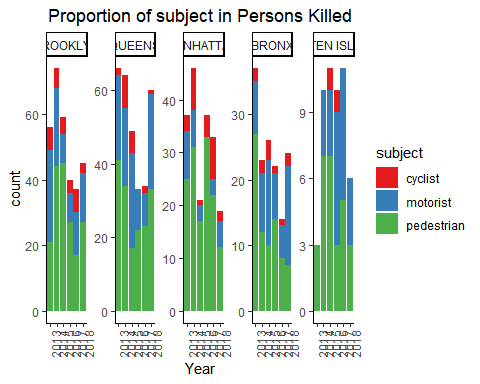
시간대별 교통사고 발생 건수와 PERSON INJUSRED는 비슷한 양상을 보인다.

### 13. (상) 다친 사람들 중 비율 (하) 사망한 사람들 비율

dat3 %>% filter(!is.na(BOROUGH))%>% group\_by(BOROUGH, y) %>% summarise( cyclist = sum(NUMBER.OF.CYCLIST.INJURED, na.rm=T), pedestrian= sum(NUMBER.OF.PEDESTRIANS.INJURED, na.rm=T), motorist = sum(NUMBER.OF.MOTORIST.INJURED, na.rm=T)) %>%   
 gather("total", "count",cyclist, pedestrian, motorist) %>% ggplot(aes(y, count)) + geom\_bar(stat = "identity", aes(fill=total)) + facet\_wrap(~BOROUGH, ncol = 5, scales = "free\_y")+theme\_classic() + scale\_fill\_brewer(name = "subject", palette = "Set1")+ labs(x="Year", y = "count", title = "Proportion of subject in Persons Injured") + theme(plot.title=element\_text(hjust = 0.5), axis.text.x = element\_text(hjust=1, angle = 90))



dat3 %>% filter(!is.na(BOROUGH))%>% group\_by(BOROUGH, y) %>% summarise( cyclist = sum(NUMBER.OF.CYCLIST.KILLED, na.rm=T), pedestrian= sum(NUMBER.OF.PEDESTRIANS.KILLED, na.rm=T), motorist = sum(NUMBER.OF.MOTORIST.KILLED, na.rm=T)) %>%   
 gather("total", "count",cyclist, pedestrian, motorist) %>% ggplot(aes(y, count)) + geom\_bar(stat = "identity", aes(fill=total)) + facet\_wrap(~BOROUGH, ncol = 5, scales = "free\_y")+theme\_classic() + scale\_fill\_brewer(name = "subject", palette = "Set1")+ labs(x="Year", y = "count", title = "Proportion of subject in Persons Killed") + theme(plot.title=element\_text(hjust = 0.5), axis.text.x = element\_text(hjust=1, angle = 90))

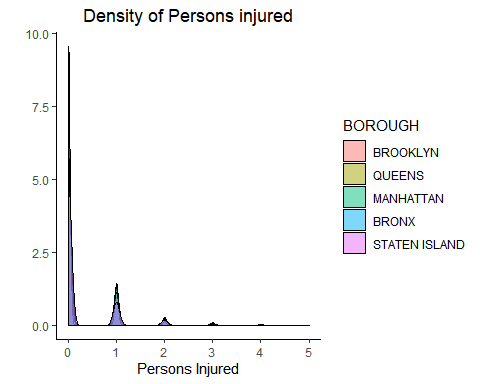


다친 사람들은 대부분의 구에서 Motorist가 많았고 사망한 사람들은 대부분의 구에서 보행자가 많았다. 브루클린과 퀸즈는 보행자 비중이 커지고 나머지구는 보행자 비중이 작아진다. 2018년 퀸즈는 사망자 합계가 증가했고 특히 motorist가 큰폭으로 증가했다. 맨하탄은 다른 구에 비해 cyclist 사망자가 많은편인데 2018년에는 적었다.

### 14. (상) density of person injured (하) density of persons killed

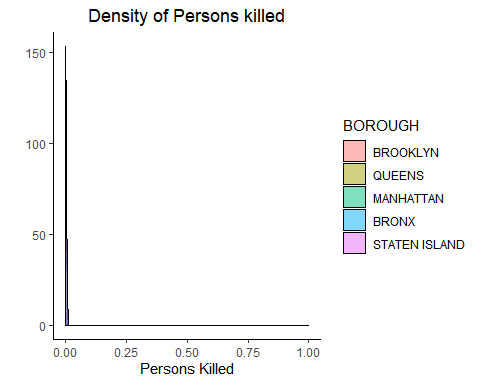
dat3 %>% filter(!is.na(BOROUGH)) %>% ggplot(aes(NUMBER.OF.PERSONS.INJURED, fill = BOROUGH)) + geom\_density(y="..density..", alpha = 0.5, col = "black") + xlim(c(0,5)) + theme\_classic() + labs(x = "Persons Injured", y = " ", title = "Density of Persons injured") + theme(plot.title = element\_text(hjust = 0.5))

## Warning: Removed 911 rows containing non-finite values (stat\_density).



dat3 %>% filter(!is.na(BOROUGH)) %>% ggplot(aes(NUMBER.OF.PERSONS.KILLED, fill = BOROUGH)) + geom\_density(y="..density..", alpha = 0.5, col = "black") + xlim(c(0, 1))+ theme\_classic() + labs(x = "Persons Killed", y = " ", title = "Density of Persons killed") + theme(plot.title = element\_text(hjust = 0.5))

## Warning: Removed 45 rows containing non-finite values (stat\_density).



injured 와 killed 모두 0 근처에 가장 많이 데이터가 분포해있다.

### 가장 많은 의미를 함축하는 같은 그림.

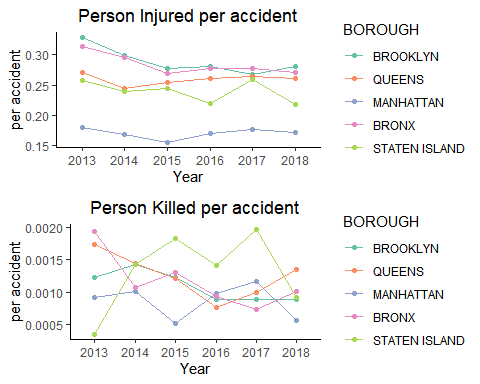
### 15. (상) 구별 사건당 상해자 수 (하) 구별 사건당 사망자 수

library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

p1 <- dat3 %>% filter(!is.na(BOROUGH))%>% group\_by(BOROUGH, y) %>% summarise(sum\_person = sum(NUMBER.OF.PERSONS.INJURED, na.rm=T), sum\_cyl = sum(NUMBER.OF.CYCLIST.INJURED, na.rm=T), sum\_ped= sum(NUMBER.OF.PEDESTRIANS.INJURED, na.rm=T), sum\_mot = sum(NUMBER.OF.MOTORIST.INJURED, na.rm=T), total = n()) %>%  
 mutate(rate = sum\_person/total) %>% ggplot(aes(y, rate, group=BOROUGH, col = BOROUGH)) + geom\_line() +geom\_point() + theme\_classic() + scale\_color\_brewer(palette = "Set2") + labs(x = "Year", y = "per accident", title = "Person Injured per accident") + theme(plot.title = element\_text(hjust = 0.5))  
  
p2 <- dat3 %>% filter(!is.na(BOROUGH))%>% group\_by(BOROUGH, y) %>% summarise(sum\_person = sum(NUMBER.OF.PERSONS.KILLED, na.rm=T),total = n()) %>%  
 mutate(rate = sum\_person/total) %>% ggplot(aes(y, rate, group=BOROUGH, col = BOROUGH)) + geom\_line() +geom\_point() + theme\_classic() + scale\_color\_brewer(palette = "Set2") + labs(x = "Year", y = "per accident", title = "Person Killed per accident") + theme(plot.title = element\_text(hjust = 0.5))  
grid.arrange(p1, p2, ncol = 1)



맨하탄은 교통사고 건수 3위 임에도 PERSON INJRED/per case 수가 가장 적다. 대부분 1000달러 이상 재산상 피해가 있는 교통사고가 발생하는 것으로 추측 가능함. 그에 반해 브롱스는 교통사고 건수 4위임에불 부구하고 person injured/per case가 높은편.

스테이튼 아일랜드의 Persons killed/per case는 해가 감에 따라 증가하다가 2018년 수치가 떨어짐 스테이튼 아일랜드를 제외한 4개구는 2013년에 비해 Persons killed/per case가 점점 감소하는 양상보임.