uk road safety

Amit Anchalia

25/05/2020

```
#install.packages('stats19')
library(stats19)
```

```
## Warning: package 'stats19' was built under R version 3.6.3
```

```
## Data provided under OGL v3.0. Cite the source and link to:
## www.nationalarchives.gov.uk/doc/open-government-licence/version/3/
```

```
library(ggplot2)
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
```

```
## Getting data for 3 years (2016 2017 & 2018) for accidents, vehicles  & casualities




# d16 = "casualtiestRoadSafetyData_Accidents_2016"
# dl_stats19(file_name = paste0(d17, ".zip"))
# crashes_2017_raw = read_accidents(year = 2017,
#                                   filename = "Acc.csv")
#
# dl_stats19(year = 2017, type = "vehicles", ask = FALSE)
# vehicles_2017_raw = read_vehicles(year = 2017)
#
#
# crashes = list()
# vehicles = list()
# casualties = list()
#
# for (i in seq(1:2))
# {
#   file = "casualtiestRoadSafetyData_Accidents_"
#   year = 2015 + i
#   file_name = paste0(file, year, '.zip')
#   filename = paste0(file, year, '.csv')
#   dl_stats19(file_name = file_name)
#   crashes_raw = read_accidents(year = year, filename = filename)
#   crashes[i] = format_accidents(crashes_raw)
#   dl_stats19(year = year, type = "vehicles", ask = FALSE)
#   vehicles_raw= read_vehicles(year = year)
#   vehicles[i]  = format_vehicles(vehicles_raw)
#   dl_stats19(year = year, type = "casualties", ask = FALSE)
#   casualties_raw= read_casualties(year = year)
#   casualties[i]= format_casualties(casualties_raw)
# }
#
#
# head(crashes[1])
```

```
casualties_2016 <- read.csv('../dataset/dftRoadSafetyData_Casualties_2016.csv')
casualties_2017 <- read.csv('../dataset/dftRoadSafetyData_Casualties_2017.csv')
casualties_2018 <- read.csv('../dataset/dftRoadSafetyData_Casualties_2018.csv')

#dim(casualties_2016)
#dim(casualties_2017)
#dim(casualties_2018)

colnames(casualties_2016) <- c("Accident_Index",
                               "Vehicle_Reference",
                               "Casualty_Reference",
                               "Casualty_Class",
                               "Sex_of_Casualty",
                               "Age_of_Casualty",
                               "Age_Band_of_Casualty",
                               "Casualty_Severity",
                               "Pedestrian_Location",
                               "Pedestrian_Movement",
                               "Car_Passenger",
                               "Bus_or_Coach_Passenger",
                               "Pedestrian_Road_Maintenance_Worker",
                               "Casualty_Type",
                               "Casualty_Home_Area_Type",
                               "Casualty_IMD_Decile")

colnames(casualties_2017) <- colnames(casualties_2016)
colnames(casualties_2018) <- colnames(casualties_2016)
```

```
casualties_2016$Year <- 2016
casualties_2017$Year <- 2017
casualties_2018$Year <- 2018

casualties <- rbind(casualties_2016, casualties_2017, casualties_2018)

#glimpse(casualties)

# casualties <- casualties[(casualties$Vehicle_Reference != -1 & casualties$Vehicle_Reference
!= 999 &
#               casualties$Casualty_Reference != -1 & casualties$Casualty_Reference != 991 &
#               casualties$Casualty_Class != -1 &
#               casualties$Sex_of_Casualty != -1 &
#               casualties$Age_of_Casualty != -1 &
#               casualties$Age_Band_of_Casualty != -1 &
#               casualties$Casualty_Severity != -1  &
#               casualties$Pedestrian_Location != -1  &
#               casualties$Pedestrian_Movement != -1 &
#               casualties$Car_Passenger != -1 &
#               casualties$Bus_or_Coach_Passenger != -1 &
#               casualties$Pedestrian_Road_Maintenance_Worker != -1 &
#               casualties$Casualty_Type != -1 &
#               casualties$Casualty_Home_Area_Type != -1 &
#               casualties$Casualty_IMD_Decile != -1), ]


#unique(casualties$Casualty_Type)

casualties[-6] <- lapply(casualties[-6], factor)

glimpse(casualties)
```

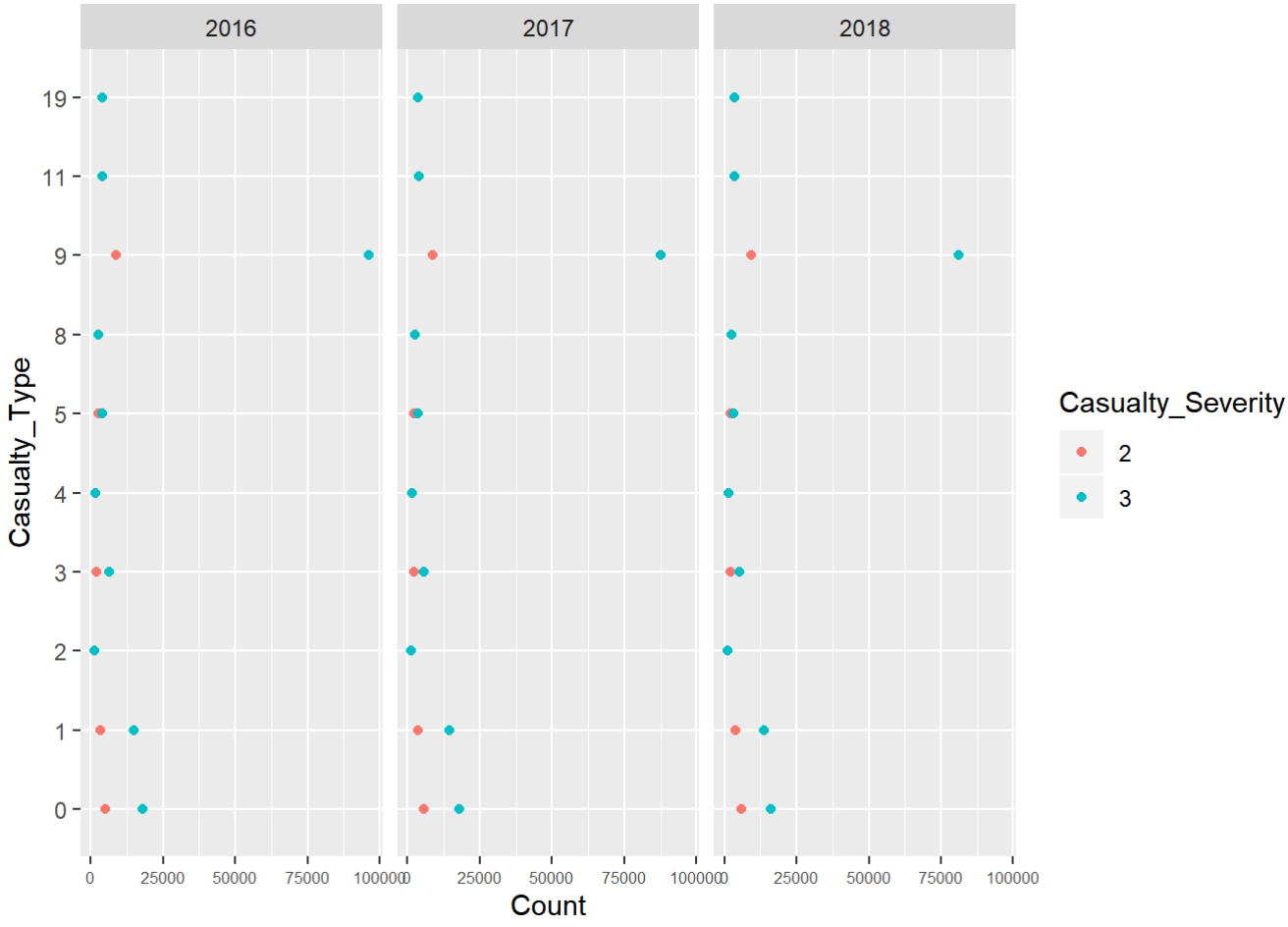
```
## Observations: 512,974
## Variables: 17
## $ Accident_Index                   <fct> 2016010000005, 201601000000...
## $ Vehicle_Reference                <fct> 2, 1, 1, 1, 2, 1, 1, 2, 1, ...
## $ Casualty_Reference               <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ Casualty_Class                   <fct> 1, 1, 1, 2, 1, 1, 3, 1, 1, ...
## $ Sex_of_Casualty                  <fct> 1, 2, 1, 2, 1, 2, 2, 2, 1, ...
## $ Age_of_Casualty                  <int> 23, 36, 24, 59, 28, 30, 33,...
## $ Age_Band_of_Casualty             <fct> 5, 7, 5, 9, 6, 6, 6, 6, 5, ...
## $ Casualty_Severity                <fct> 3, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ Pedestrian_Location              <fct> 0, 0, 0, 0, 0, 0, 5, 0, 0, ...
## $ Pedestrian_Movement              <fct> 0, 0, 0, 0, 0, 0, 1, 0, 0, ...
## $ Car_Passenger                    <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ Bus_or_Coach_Passenger           <fct> 0, 0, 0, 3, 0, 0, 0, 0, 0, ...
## $ Pedestrian_Road_Maintenance_Worker <fct> 0, 0, 0, 0, 0, 0, 2, 0, 0, ...
## $ Casualty_Type                    <fct> 2, 9, 9, 11, 1, 9, 0, 9, 4,...
## $ Casualty_Home_Area_Type          <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ Casualty_IMD_Decile              <fct> 4, 10, 8, 4, 6, 3, 1, 7, -1...
## $ Year                             <fct> 2016, 2016, 2016, 2016, 201...
```

```
#write.csv(casualties, '../dataset/dftRoadSafetyData_Casualties.csv')
```

```
table(casualties$Casualty_Type)
```

```
##
##     -1      0      1      2      3      4      5      8      9     10
##     10  69787  54348   4902  23695   6444  18247   8263 293844   1000
##     11     16     17     18     19     20     21     22     23     90
##  12283    254    290     34  12583    926   2097    599    156   1812
##     97     98
##    713    687
```

```
casualty_type <- casualties[casualties$Casualty_Type != -1, ] %>%
            group_by(Casualty_Type, Casualty_Severity, Year) %>%
            summarise(Count = n())

#sort(casualty_type$Count)

top_casualty_type <- casualty_type[casualty_type$Count > 1000, ]



ggplot(data=top_casualty_type, aes(x=Casualty_Type, y=Count, color=Casualty_Severity)) +
  geom_point() +
  facet_wrap(~Year) +
  coord_flip() +
  theme(axis.text.x = element_text(size=6))
```

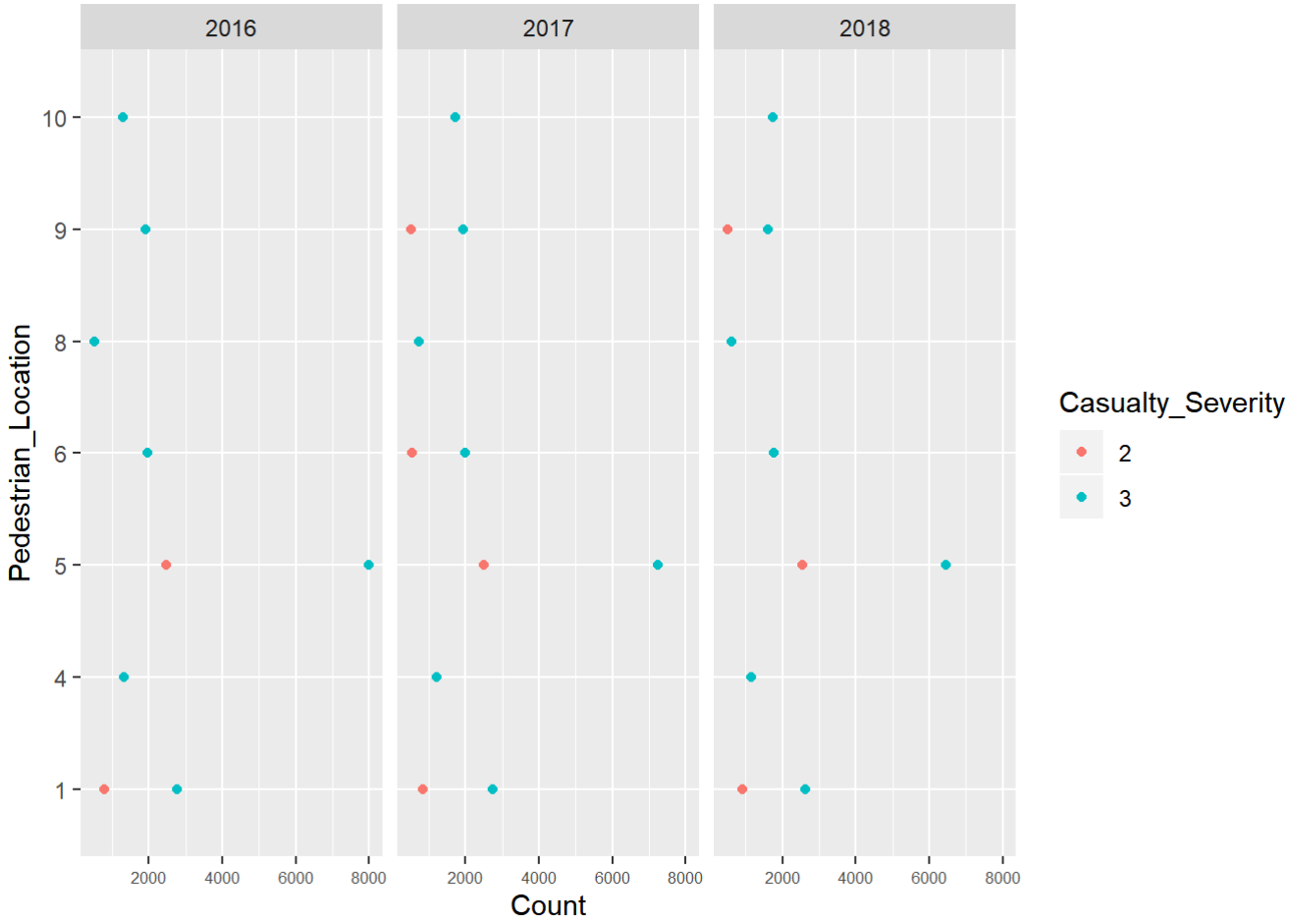


-We can see over the year most casualties are of type 9, 0 & 1 which represent Car occupant, Pedestrian & Cyclist respectively. -type 8 which represent taxi, have low casualties but again we are not aware of the actual number of taxi on roads.

```
## Pedestrain Casualties Cases

table(casualties$Pedestrian_Location)
```

```
##
##     -1      0      1      2      3      4      5      6      7      8
##      9 443185  10842    339    161   5154  29736   7312    356   2648
##      9     10
##   7141   6091
```

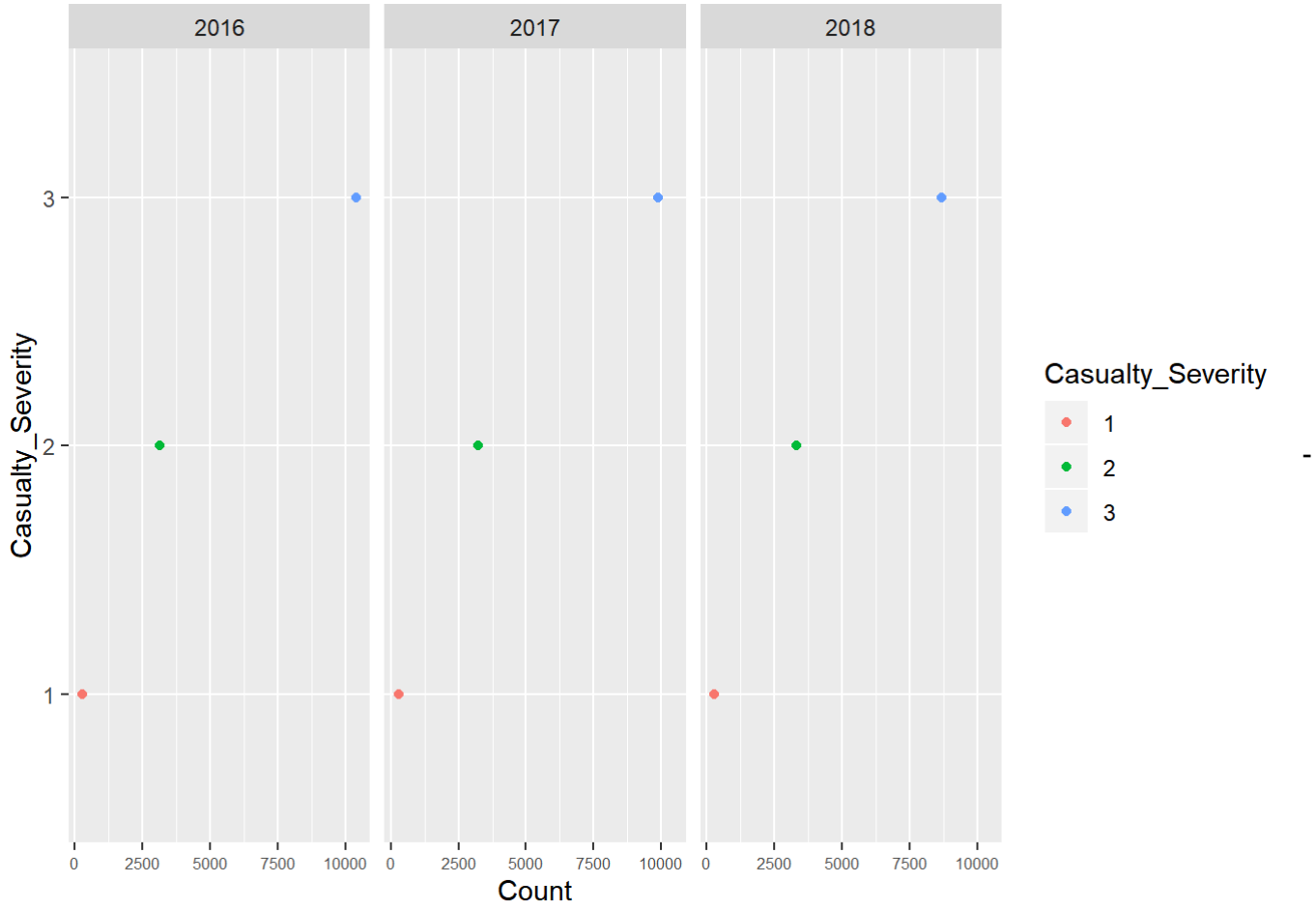
```
t = (casualties$Casualty_Type  == 0 & casualties$Pedestrian_Location != -1)

pedestrian_casualties <- casualties[t, ] %>%
            group_by(Pedestrian_Location, Casualty_Severity, Year) %>%
            summarise(Count = n())

#pedestrian_casualties[pedestrian_casualties$Casualty_Severity == 1, ]

ggplot(data=pedestrian_casualties[pedestrian_casualties$Count > 500, ], aes(x=Pedestrian_Loca
tion, y=Count, color=Casualty_Severity)) +
  geom_point() +
  facet_wrap(~Year) +
  coord_flip() +
  theme(axis.text.x = element_text(size=6))
```



-The below graph show the location with with pedestrain casualty count more than 500. -Most of the pedestrian have slight severity. -While cases for location at 1, 5 which is “Crossing on pedestrian crossing facility” & “In carriageway, crossing elsewhere” have some serious severity cases. -It may suggest that people are being irresponsible and not using pedestrain crossing for case 5.

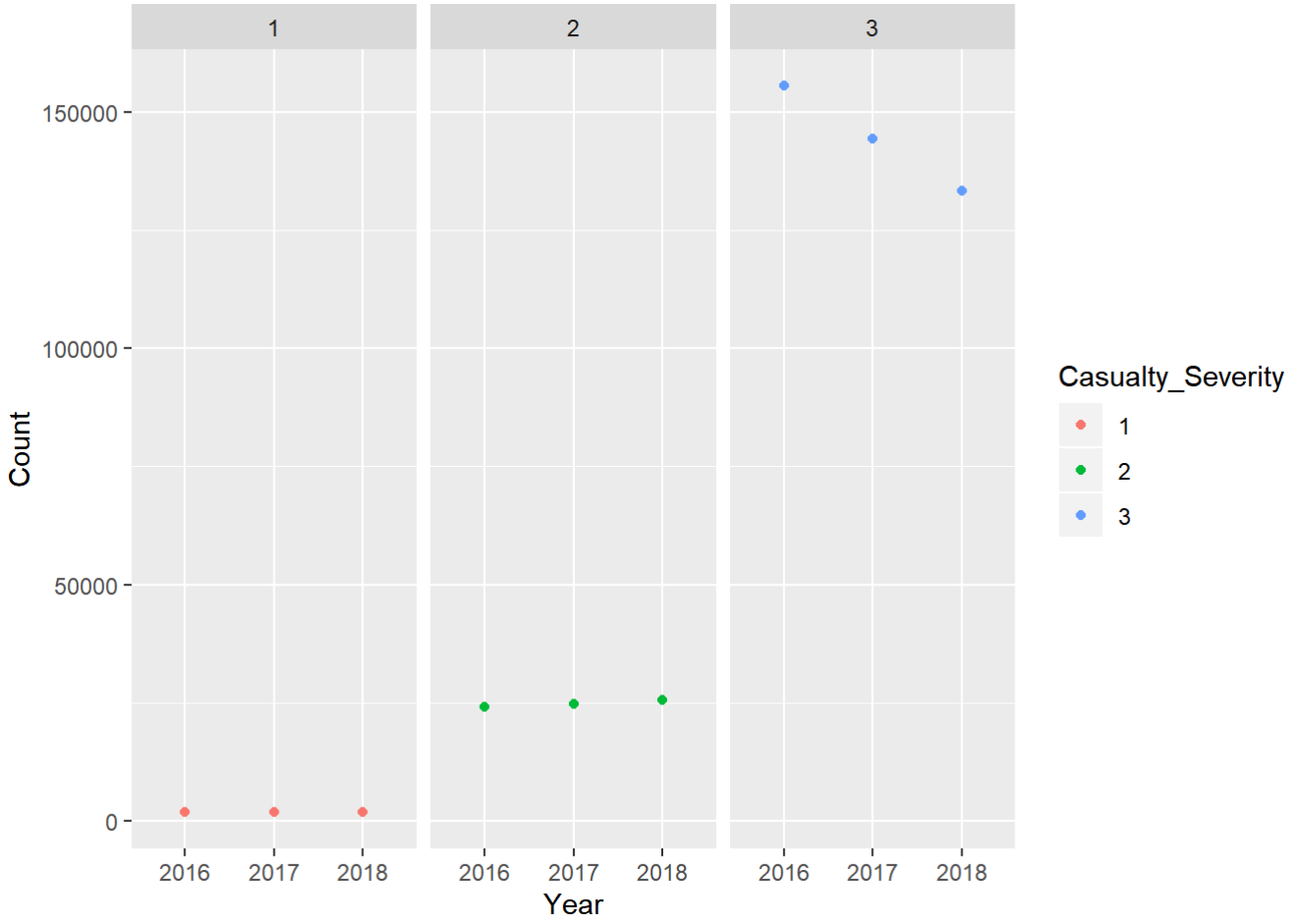
-Then there are some cases at location 6 which is “On footway or verge” suggest drivers are being irresponsible. -cases at location 9 which is In carriageway, not crossing. So it is similar to case 5.

```
## Pedestrain in Carriageway Casualties Cases
pedestrain_in_carriageway <- pedestrian_casualties[(pedestrian_casualties$Pedestrian_Location
== 9 |
                       pedestrian_casualties$Pedestrian_Location == 5 |
                       pedestrian_casualties$Pedestrian_Location == 8), ]


pedestrain_in_carriageway <- pedestrain_in_carriageway %>%
            group_by(Casualty_Severity, Year) %>%
            summarise(Count = sum(Count))

ggplot(data=pedestrain_in_carriageway, aes(x=Casualty_Severity, y=Count, color=Casualty_Sever
ity)) +
  geom_point() +
  facet_wrap(~Year) +
  coord_flip() +
  theme(axis.text.x = element_text(size=6))
```



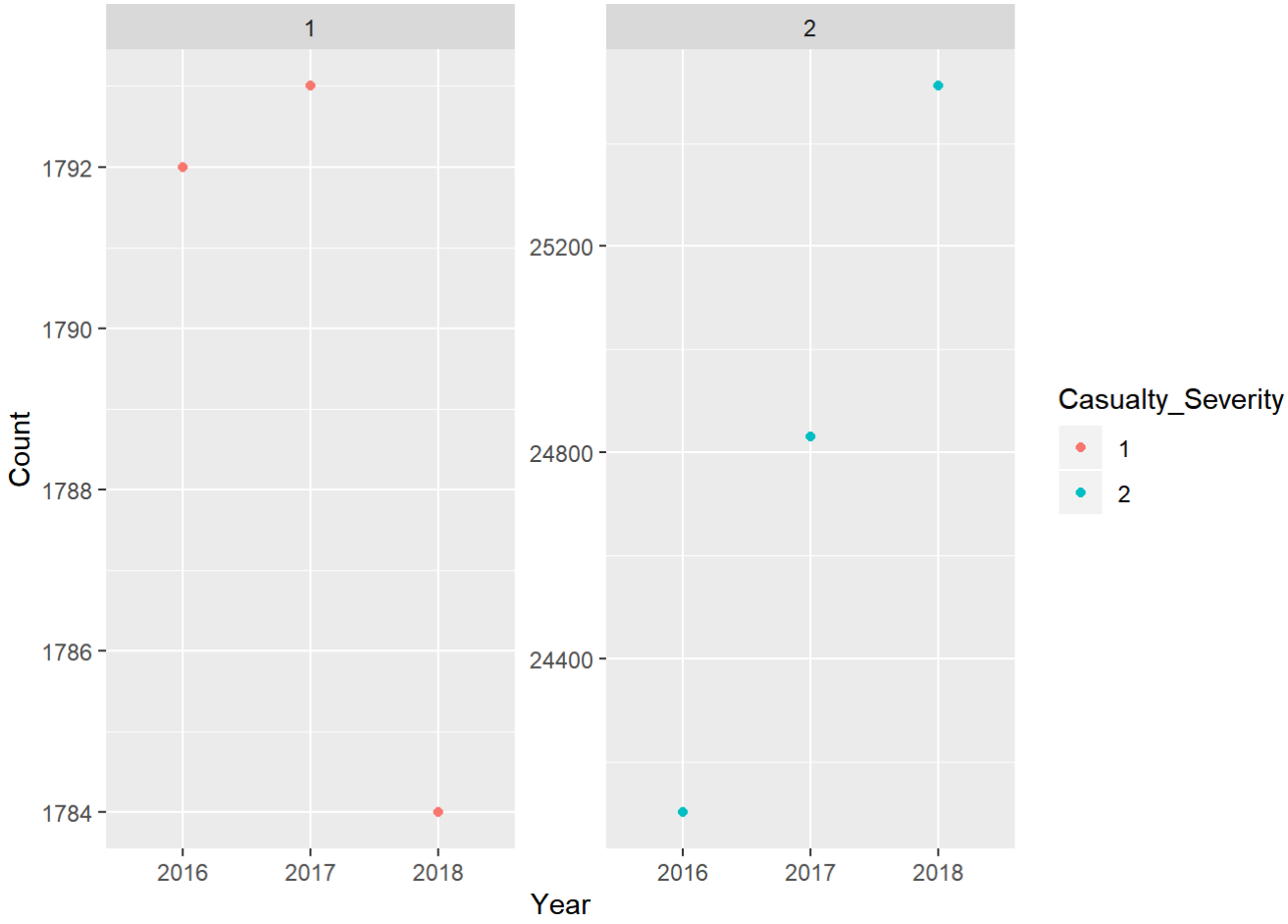
Fatal & serious Casualties count have not improved much perhaps serious cases have slightly increase over year. -Slight casualties case count have improved over year.

```
#unique(casualties$Casualty_Severity)

severity <- casualties %>%
            group_by(Casualty_Severity, Year) %>%
            summarise(Count = n())

ggplot(data=severity, aes(x=Year, y=Count, color=Casualty_Severity)) +
  geom_point() +
  facet_wrap(~Casualty_Severity)
```



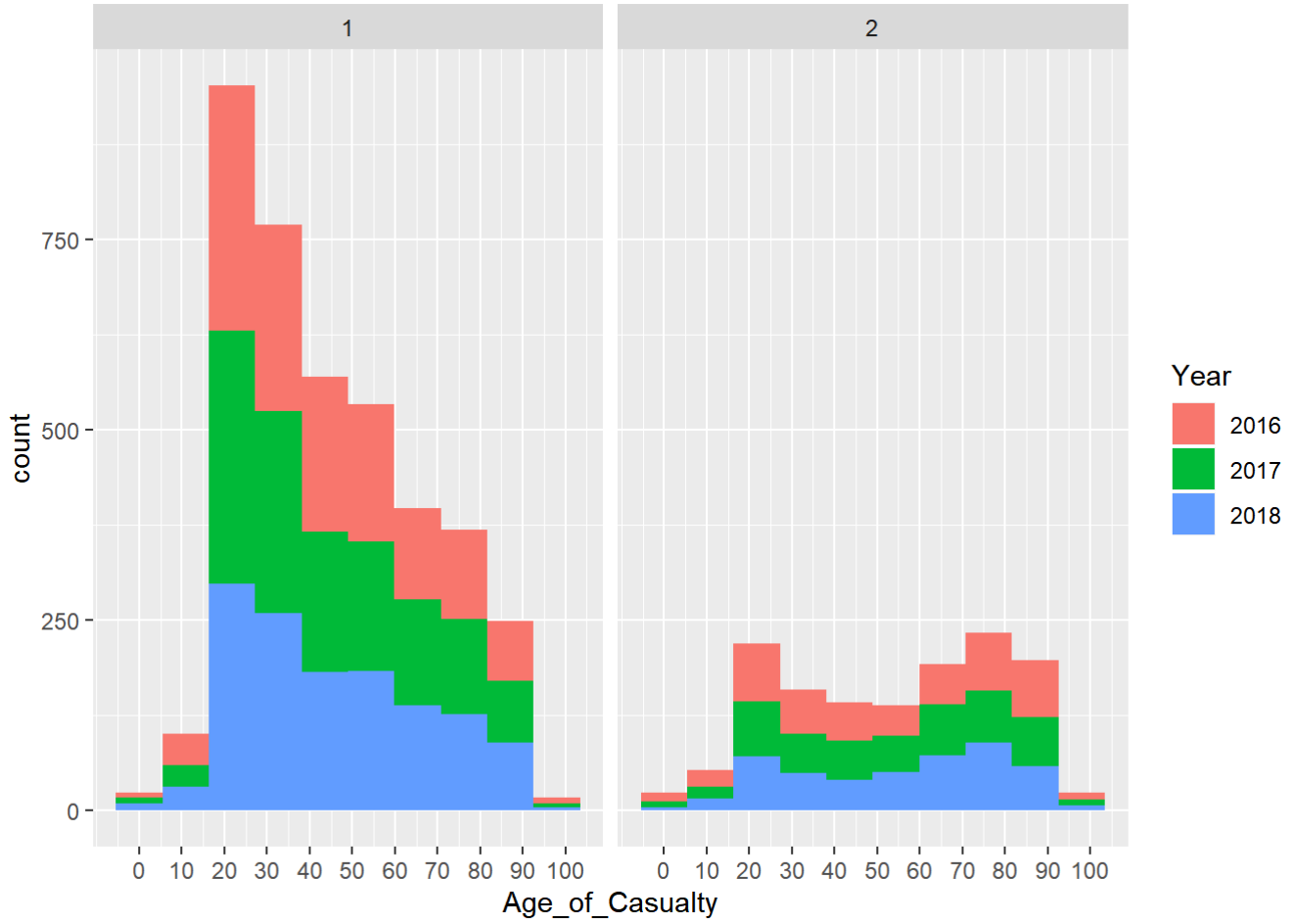
Severities type 1 & 2 are almost same over year 2016 to 2018 but there has been decrease in type 3 over the years. 1 - Fatal, 2 - Serious, 3 - Slight

```
## Considering fatal & serious cases
ggplot(data=severity[severity$Casualty_Severity != 3, ], aes(x=Year, y=Count, color=Casualty_
Severity)) +
  geom_point() +
  facet_wrap(~Casualty_Severity, scales = "free")
```



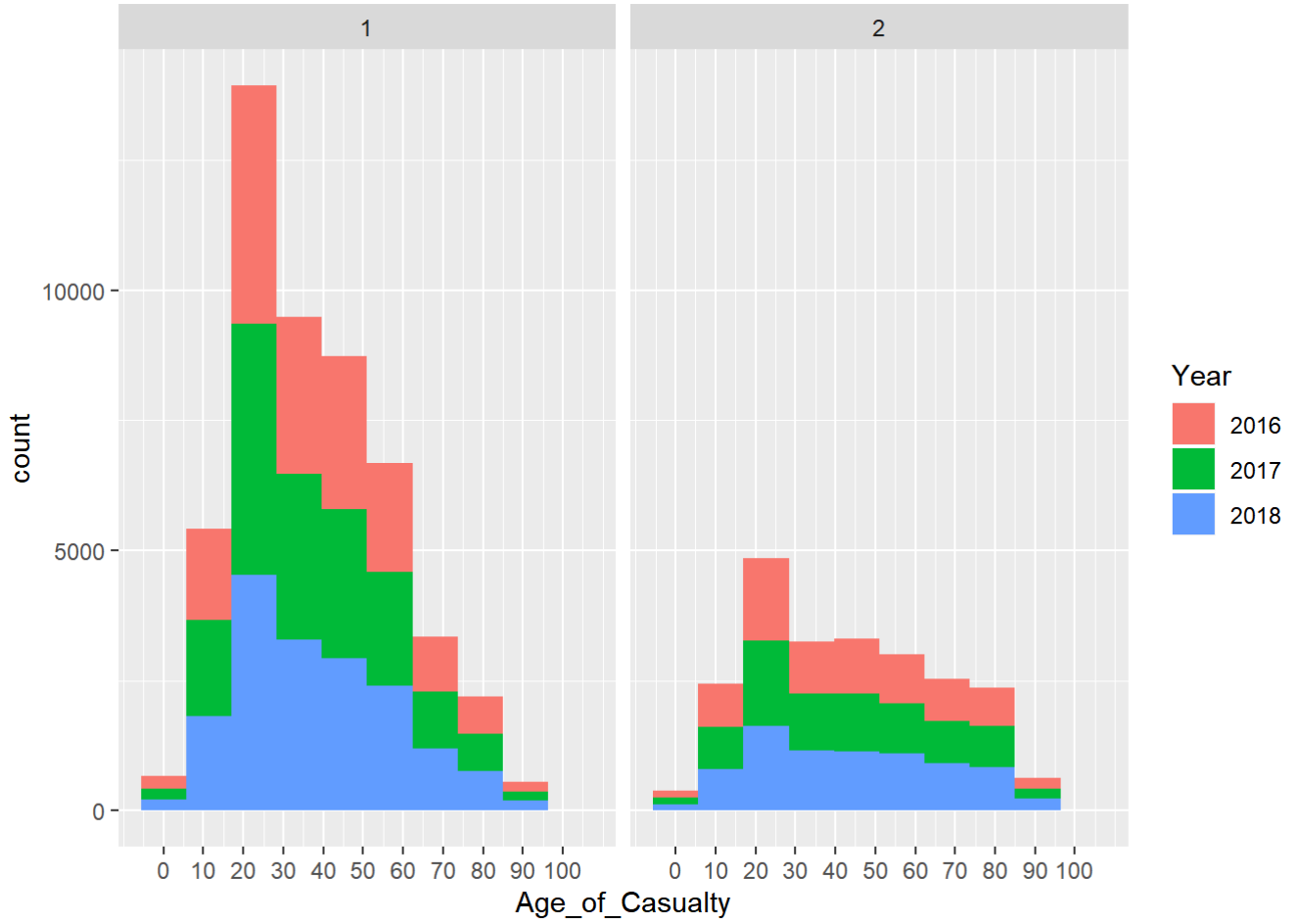
While consdering just the carriage way for pedesterian we notice similar trend, fatal cases have not imporved much and there is increase in serious cases over the year.

```
## Fatal Cases
fatalities <- casualties[casualties$Casualty_Severity == 1, ]

ggplot(data=fatalities[fatalities$Age_of_Casualty != -1, ], aes(x=Age_of_Casualty)) +
  geom_histogram(bins=10, aes(fill=Year)) +
  scale_x_continuous(breaks = seq(0,100,10)) +
  facet_wrap(~Sex_of_Casualty)
```



-Fatal casualties are higest among yonger male which reduces with age. -For female fatalties is high for yonger female which slight decrease for age group till 60, and then its again high for age 60 to aroud 90. -There has been a similar trend over the year.

```
## Serious Severity Cases
serious_severity <- casualties[casualties$Casualty_Severity == 2, ]

serious_severity <- serious_severity[serious_severity$Sex_of_Casualty != -1, ]

ggplot(data=serious_severity[serious_severity$Age_of_Casualty != -1, ], aes(x=Age_of_Casualt
y)) +
  geom_histogram(bins=10, aes(fill=Year)) +
  scale_x_continuous(breaks = seq(0,100,10)) +
  facet_wrap(~Sex_of_Casualty)
```




```
length(unique(casualties$Accident_Index))
```

```
## [1] 389238
```

-Similar to fatal casualties, serious severity case are higest among yonger male which reduces with age. -For female serious severity is high for yonger female, but after age 30 it reduces with age. -There has been a similar trend over the year.

-Either there are more number of younger population on the road or there is more casualties among this age.

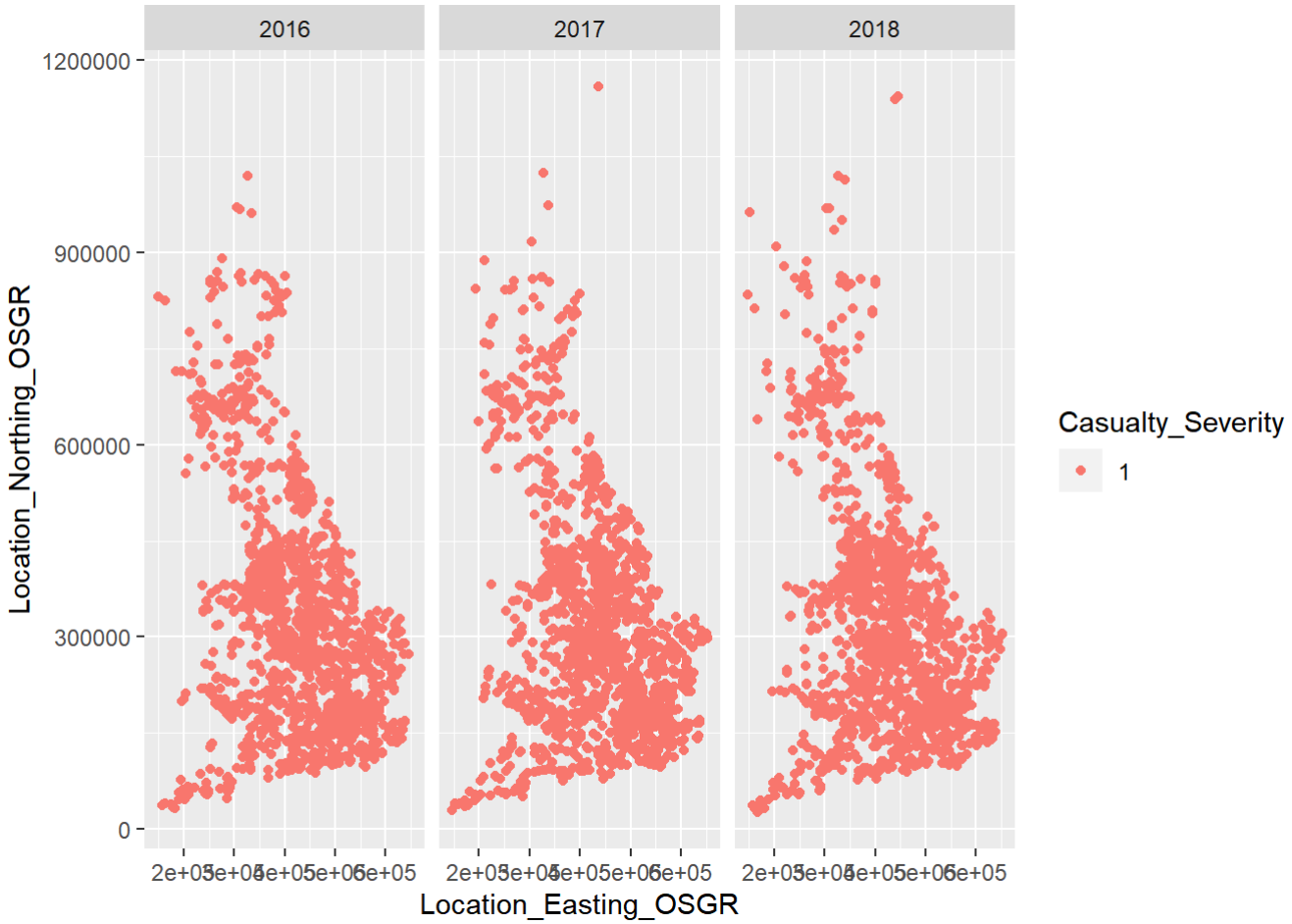
```
accidents_2016 <- read.csv('../dataset/dftRoadSafetyData_Accidents_2016.csv')
accidents_2017 <- read.csv('../dataset/dftRoadSafetyData_Accidents_2017.csv')
accidents_2018 <- read.csv('../dataset/dftRoadSafetyData_Accidents_2018.csv')


#dim(accidents_2016)
#dim(accidents_2017)
#dim(accidents_2018)

colnames(accidents_2017) <- colnames(accidents_2016)
colnames(accidents_2018) <- colnames(accidents_2016)


accidents_2016$Year <- 2016
accidents_2017$Year <- 2017
accidents_2018$Year <- 2018

accidents <- rbind(accidents_2016, accidents_2017, accidents_2018)
```

```
## Getting location of the accidents
accident_location <- accidents %>%
                        select(Accident_Index, Location_Easting_OSGR, Location_Northing_OSGR)



## Adding location to casualties data
casualties_location <- merge(casualties, accident_location, by="Accident_Index")

dim(casualties_location)
```

```
## [1] 512913     19
```

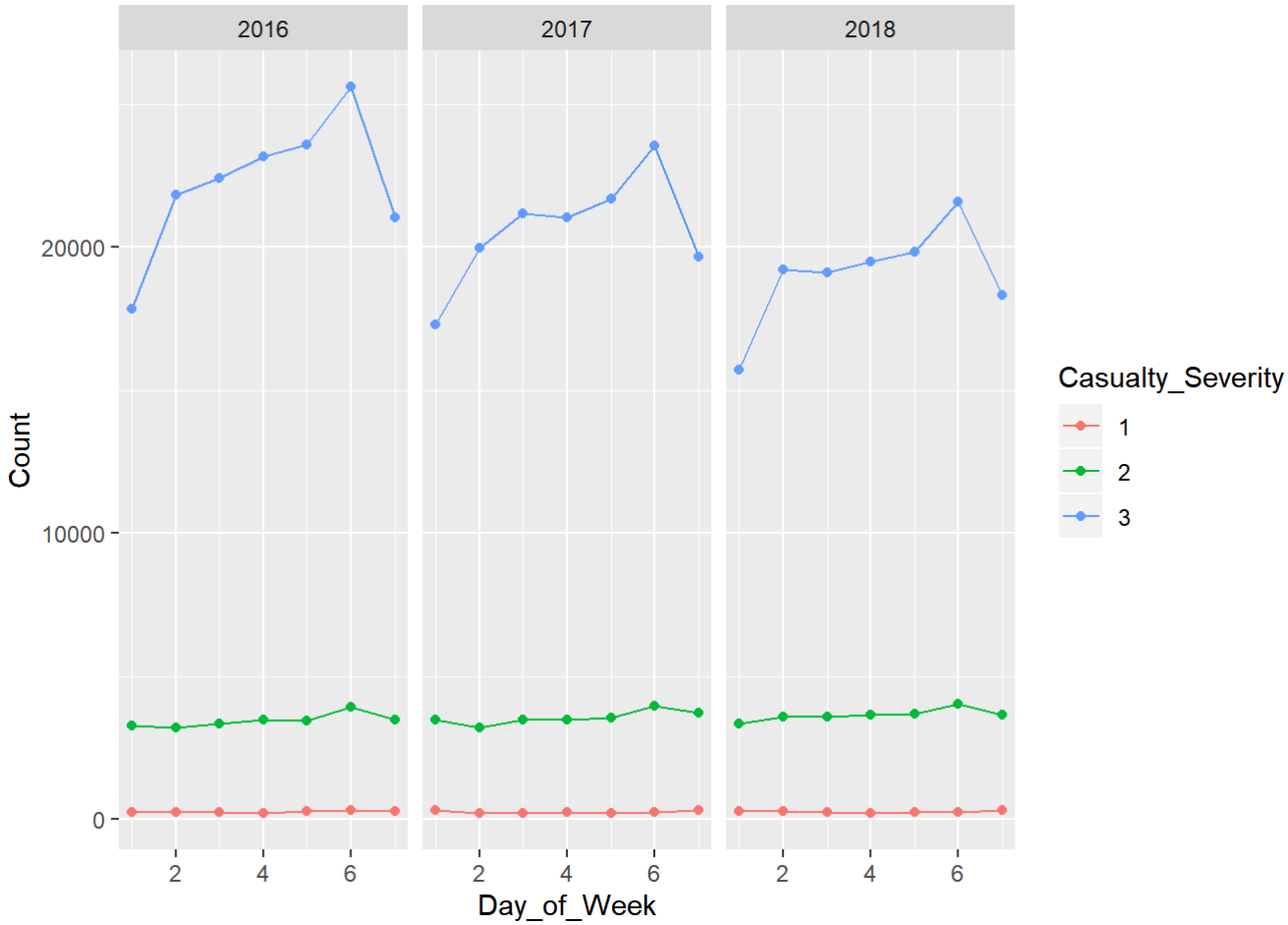
```
ggplot(data=casualties_location[casualties_location$Casualty_Severity == 1, ], aes(x=Location
_Easting_OSGR, y=Location_Northing_OSGR)) +
        geom_point(aes(color=Casualty_Severity)) +
          facet_wrap(~Year)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



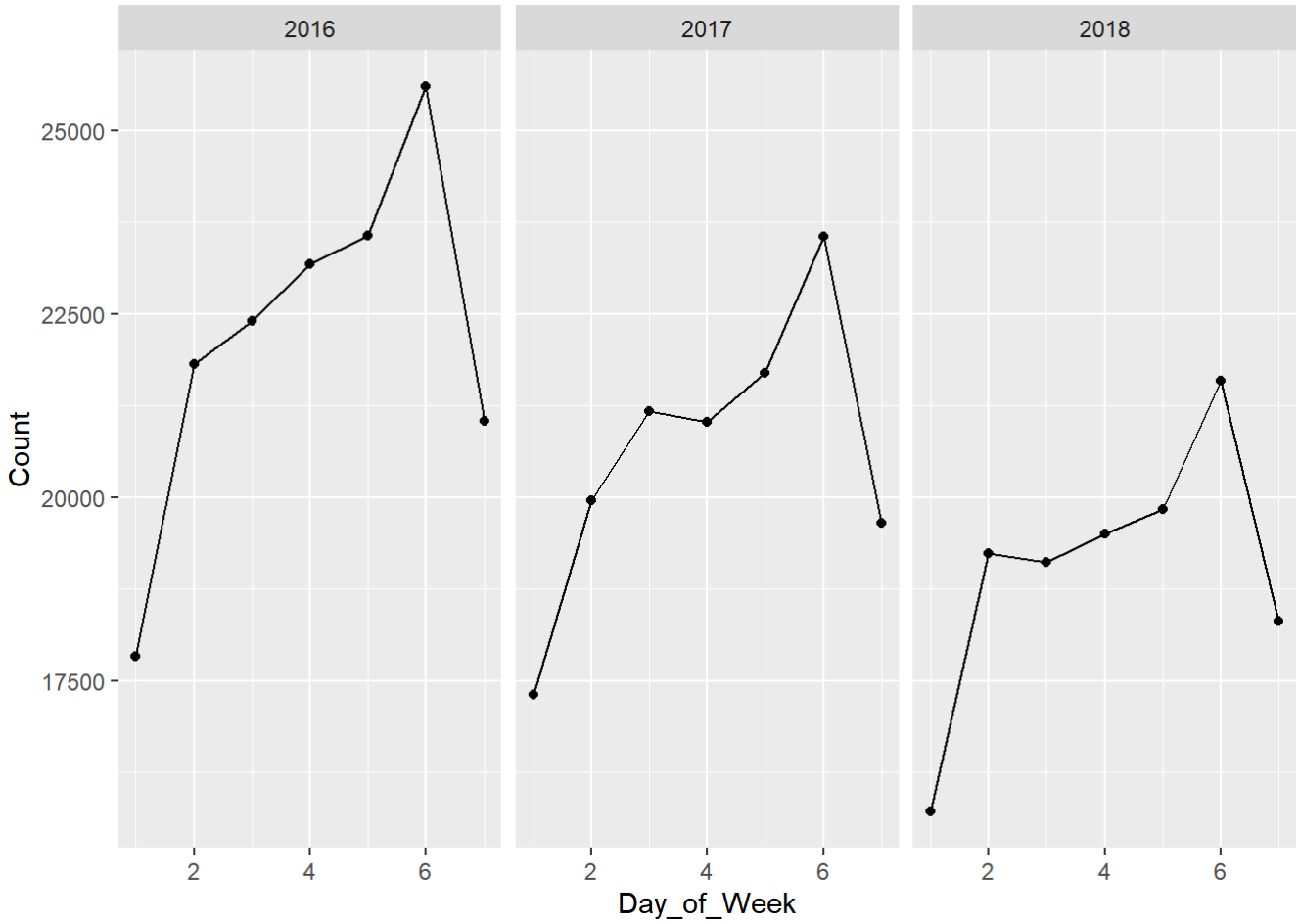
```
accident_time <- accidents %>%
                        select(Accident_Index, Date, Day_of_Week, Time)


accident_time$Hour <- format(strptime(accident_time$Time,"%H:%M"),'%H')
accident_time$Day <- format(as.Date(accident_time$Date, "%d/%m/%Y"), "%d")
accident_time$Month <- format(as.Date(accident_time$Date, "%d/%m/%Y"), "%m")

accident_time$Quarter <- as.factor(ifelse(as.integer(accident_time$Month) > 9, 4,
                                  ifelse(as.integer(accident_time$Month) > 6, 3,
                                         ifelse(as.integer(accident_time$Month) > 3, 2, 1
))))



#head(accident_time[20:30, ])

casualties_time <- merge(casualties, accident_time, by="Accident_Index")
```

```
## Casualties by week
casualties_week <- casualties_time %>%
                    group_by(Casualty_Severity, Day_of_Week, Year) %>%
                    summarise(Count=n())


ggplot(data=casualties_week, aes(x=Day_of_Week, y=Count, color=Casualty_Severity)) +
        geom_point() +
        geom_line() +
        facet_wrap(~Year)
```

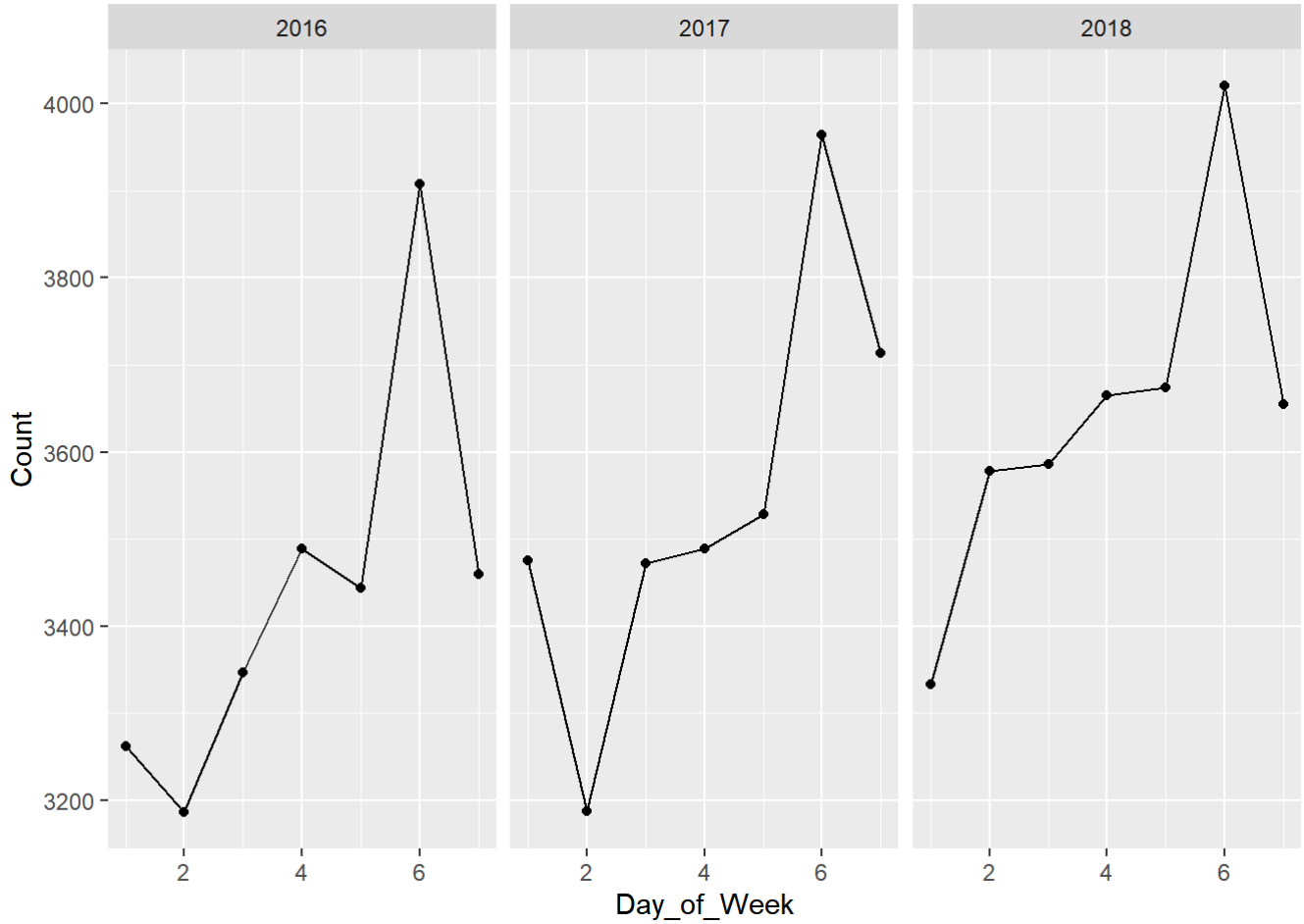


Severity 3 (slight) cases increase from day 1 (sunday) till day 6 (friday) which is highest and saturday have low such cases as compared to other days except sunday.

```
## Severity 3 - slight
ggplot(data=casualties_week[casualties_week$Casualty_Severity == 3, ], aes(x=Day_of_Week, y=C
ount)) +
        geom_point() +
        geom_line() +
        facet_wrap(~Year)
```

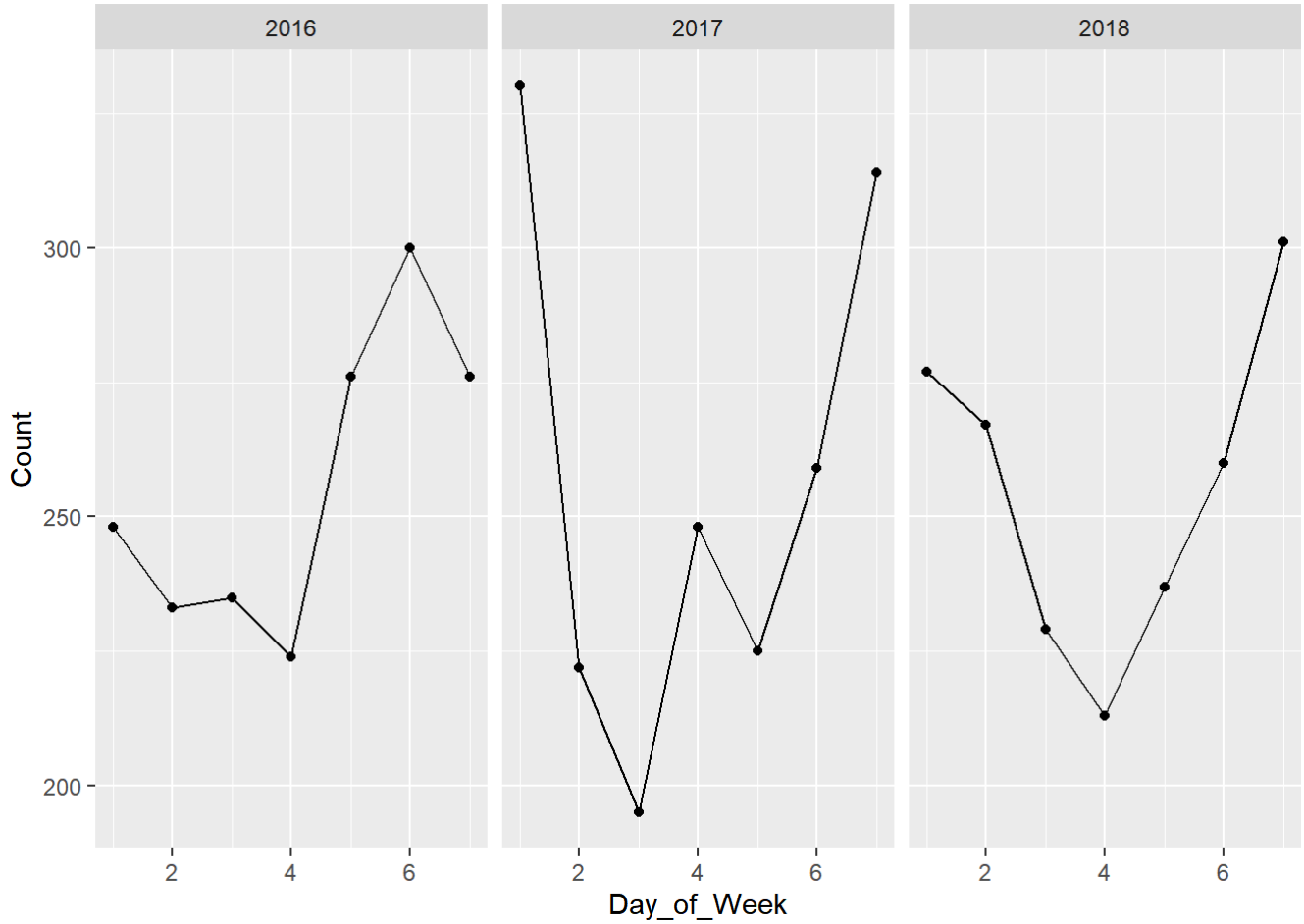



```
## Severity 2 - serious case
ggplot(data=casualties_week[casualties_week$Casualty_Severity == 2, ], aes(x=Day_of_Week, y=C
ount)) +
        geom_point() +
        geom_line() +
        facet_wrap(~Year)
```



Severity 2 (serious) have similar monday to friday increase trend. friday with higest count.

```
## Severity 1 - fatal
ggplot(data=casualties_week[casualties_week$Casualty_Severity == 1, ], aes(x=Day_of_Week, y=C
ount)) +
        geom_point() +
        geom_line() +
        facet_wrap(~Year)
```

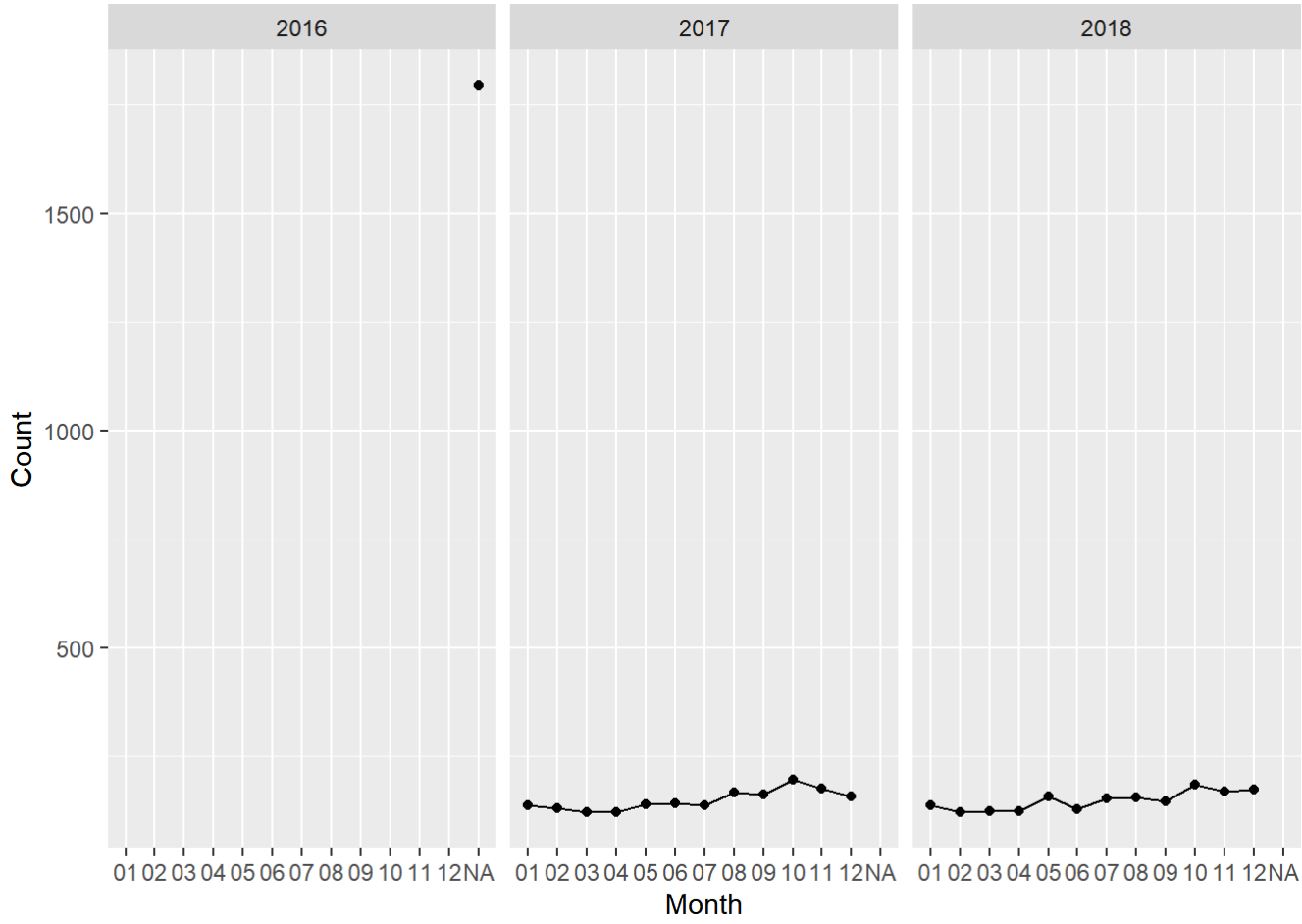


Severity 3 (fatal) cases are higher on sunday and saturday except 2016 year where thrusday, friday and saturday have higher count.

```
casualties_month <- casualties_time %>%
                    group_by(Casualty_Severity, Month, Year) %>%
                    summarise(Count=n())


casualties_month$Month <- as.factor(casualties_month$Month)

ggplot(data=casualties_month[casualties_month$Casualty_Severity == 1, ], aes(x=Month, y=Coun
t, group=Year)) +
        geom_point() +
        geom_line() +
        facet_wrap(~Year)
```

```
## Warning: Factor `Month` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



Over the year fatalities have lower count for month 2,3,4. For some reason there is a sudden increase in 5th & 8th month every year.

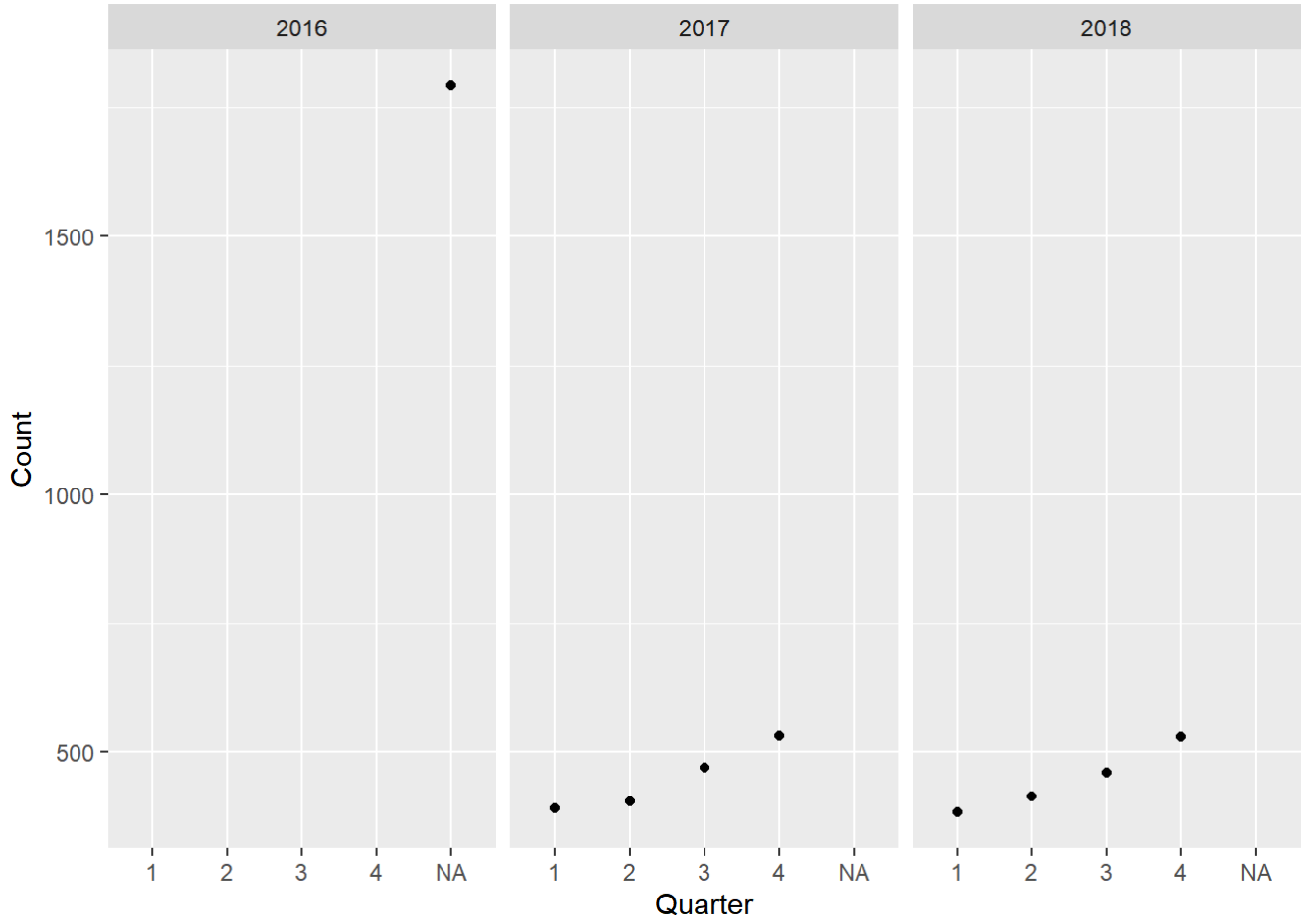
```
casualties_quarter <- casualties_time %>%
                    group_by(Casualty_Severity, Quarter, Year) %>%
                    summarise(Count=n())
```

```
## Warning: Factor `Quarter` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
ggplot(data=casualties_quarter[casualties_quarter$Casualty_Severity == 1, ], aes(x=Quarter, y
=Count)) +
        geom_point() +
        geom_line() +
        facet_wrap(~Year)
```

```
## Warning: Factor `Quarter` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```

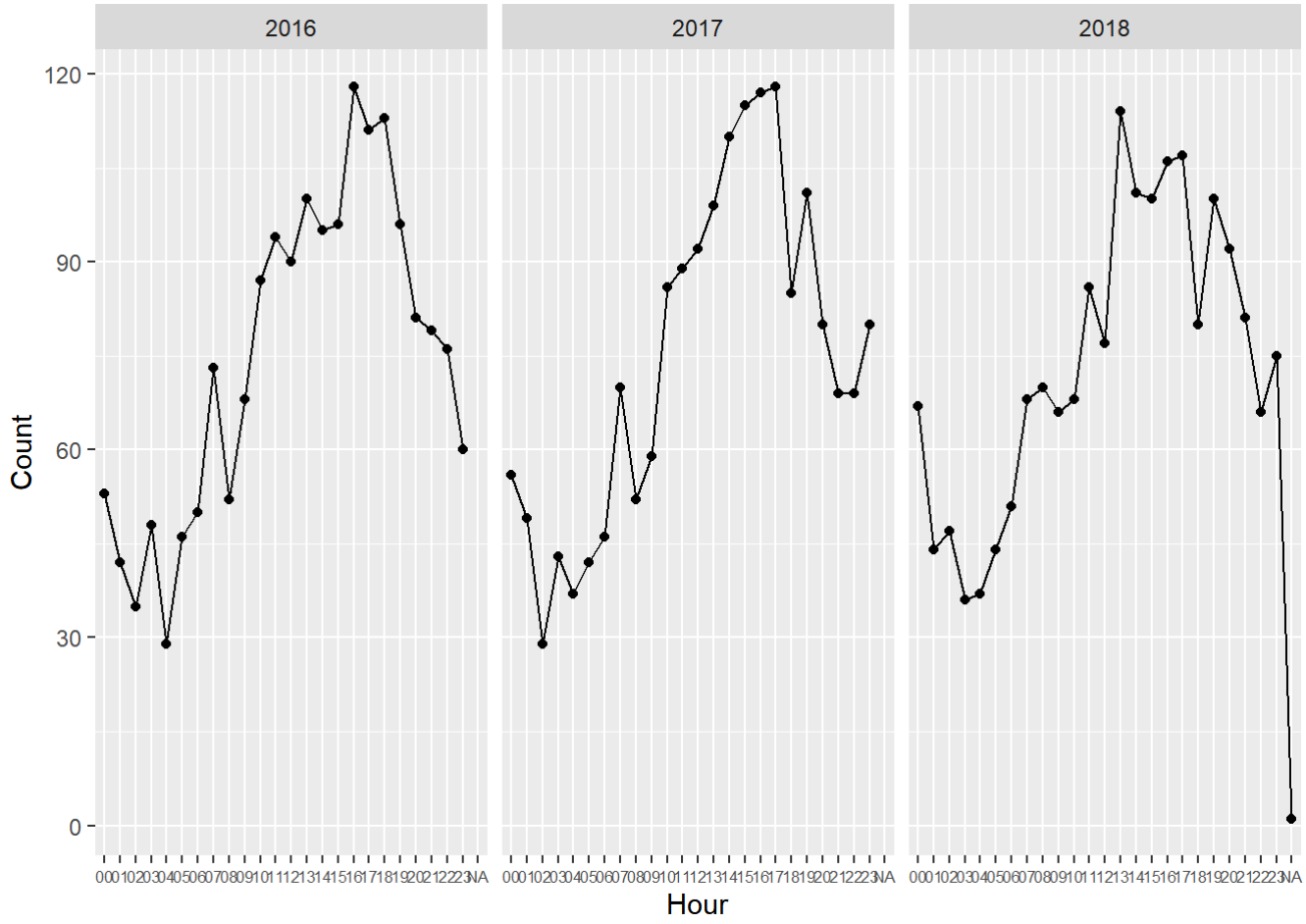


Over the year, There is an increasing trend for the fatalities case for the 1 to 4 quarter. For first quarter of year the count for fatalities have decreased but this is opposite for last 2 quarters.

```
casualties_hour <- casualties_time %>%
                   group_by(Casualty_Severity, Hour, Year) %>%
                   summarise(Count=n())

sum(is.na(casualties_hour))
```

```
## [1] 5
```

```
ggplot(data=casualties_hour[casualties_hour$Casualty_Severity == 1, ], aes(x=Hour, y=Count, g
roup=Year)) +
        geom_point() +
        geom_line() +
        facet_wrap(~Year) +
        theme(axis.text.x=element_text(size=6))
```



There is a increasing trend from around 7am to 7pm over the year. There is a peak from 2 to 7 pm. Night time after 7 till 12 have more count of fatalities than after 12 till 5-6 pm. May be there are less number of cars and pedestrian then.

There is a sudden rise at 7 am. May be because of increase in bus and cars on road.