Pedestrian - Biker Crash Data Set

Data Exploration

Working on exploratory data analysis (EDA) on Pedestrian and Bike Crash Data provided by the St. Paul Police Department in Minnesota. The work is done with an emphasis on extracting, transforming, and loading the data (ETL) to produce data visualizations using primarily tidyverse and ggplot whenever possible. The beauty in using this format as opposed to other EDA tools (e.g. Tableau) is that this procedure produces **reproducible work**, so the analysis can be continued or re-used over time as the data set is updated. The code to produce these plots will not change so long as the tidyverse and ggplot packages do not become deprecated (which is highly unlikely). Furthermore, once you have the code to produce these plots, it is a short step away from creating an interactive dashboard via the R Shiny package.

The data source is provided at the following url: https://www.stpaul.gov/departments/police/pedestrian-and-bike-crash-data-city-st-paul

They also have summaries of their data sets (e.g. at this url: https://www.stpaul.gov/sites/default/files/Media%20Root/Police/2018%20Ped%20Crash%20Data%20%28SPPD%29%20-%20%28PUBLIC%29%20-%20Summary.pdf) However, for the purposes of EDA, I will not be examining these too closely so I can generate my own questions and answers.

Shortcuts for working in markdown.

Since this document is an exercise in familiarizing myself with tidyverse and ggplot2, all code chunks will be shown and a list of short cuts for working with these packages in markdown are provided here.

- %>% = Cntrl + Shift + M
- r-chunk = Cntrl + Alt + I

Loading the data.

Cleaning the data.

There are a couple of fields where Crash type is 'Ped2' or a mispelled version of Pedestrian. While the cleaning data section appears above all else, the mistakes in the data set were discovered along the way of exploring the data, namely when I was taking the sum of crash counts for each crash type. You will learn that the data cleaning process is often times *not* the first step in analysis, but it is something you do along the way as you discover ways to tidy up the data set.

Exploring data set variables.

The first step in EDA is viewing the data set. It is omitted here due to length; but the column names are shown below. Looking at the data set, we can begin to ask some interesting questions and think about ways to exploring how to answer these questions using the data.

```
cleaned_data %>%
  #View()
  colnames()
```

```
##  [1] "Crash.Type"                      "Report.Made."
##  [3] "Date...Time"                     "Case.Number"
##  [5] "District"                        "Crash.Location"
##  [7] "Lanes.Of.Traffic"                "Signal.Present."
##  [9] "Speed.Limit.Of.Road..MPH."       "Road.Type"
## [11] "Synopsis"                        "Ticket.Arrest."
## [13] "Citation.To"                     "Pedestrian.Age"
## [15] "Pedestrian.Gender"               "Pedestrian.City.Of.Residence"
## [17] "Pedestrian.Zip.Code"             "Biker.Age"
## [19] "Biker.Gender"                    "Biker.City.of.Residence"
## [21] "Biker.Zip.Code"                  "Driver.Age"
## [23] "Driver.Gender"                   "Driver.City.Of.Residence"
## [25] "Driver.Zip.Code"                 "Injury.to.Pedestrian."
## [27] "Level.of.Injury.to.Pedestrian"   "Pedestrian.to.Hospital."
## [29] "Injury.to.Biker."                "Level.of.Injury.to.Biker"
## [31] "Biker.to.Hospital."              "Crash.Lat.Long.Location"
## [33] "Count"                           "District.Council...Map"
## [35] "Council.Ward"
```

Some initial EDA questions are…

1. How many crashes occurred for each crash type?
2. Which (police dept.) districts responded to the crashes and how many did they respond to?
3. How many of each type of crash did each district respond to?
4. How do incidents of each type of crash differ across time?
5. How many of each crash type occurred for each speed limit?
6. How are ages associated with crash counts and types?
7. How are genders associated with crash counts and types?
8. Are there areas where accidents are generally more clustered?
9. Does a signal being present have an effect on crash incidents?
10. What type of crash incident is associated with more severe injuries?

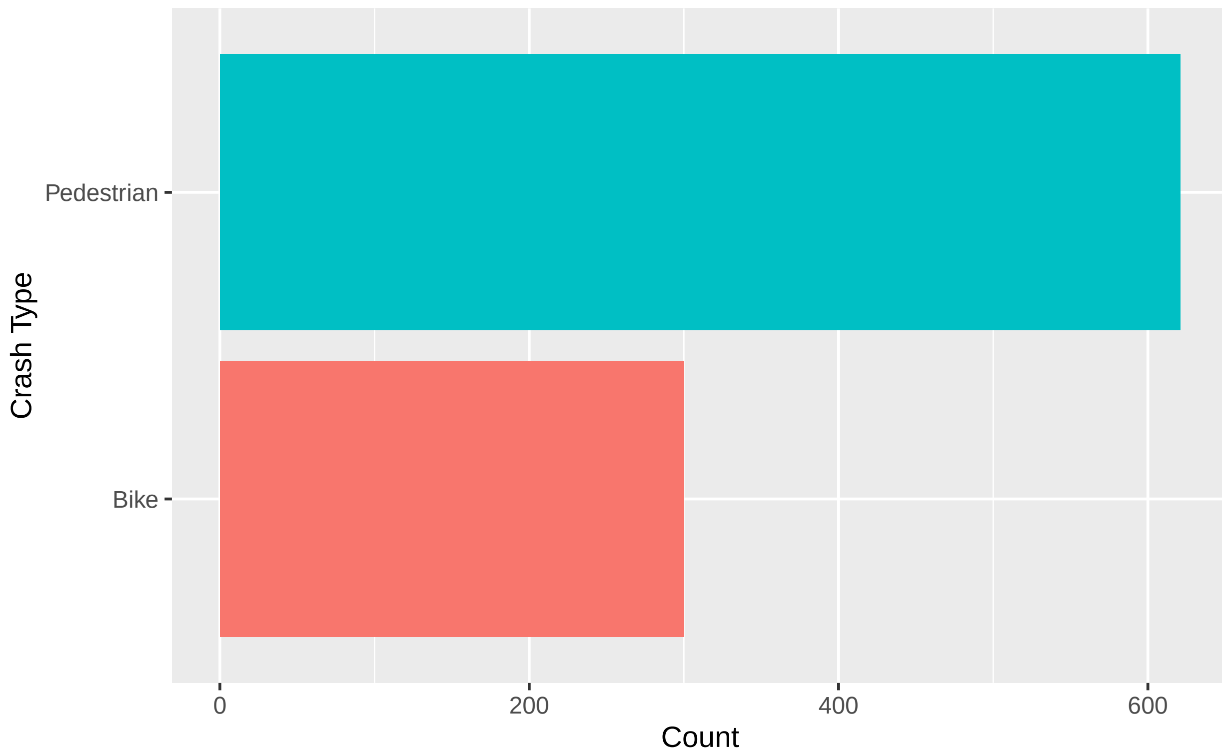
Exploratory Data Analysis Section

1. How many crashes occurred for each crash type?

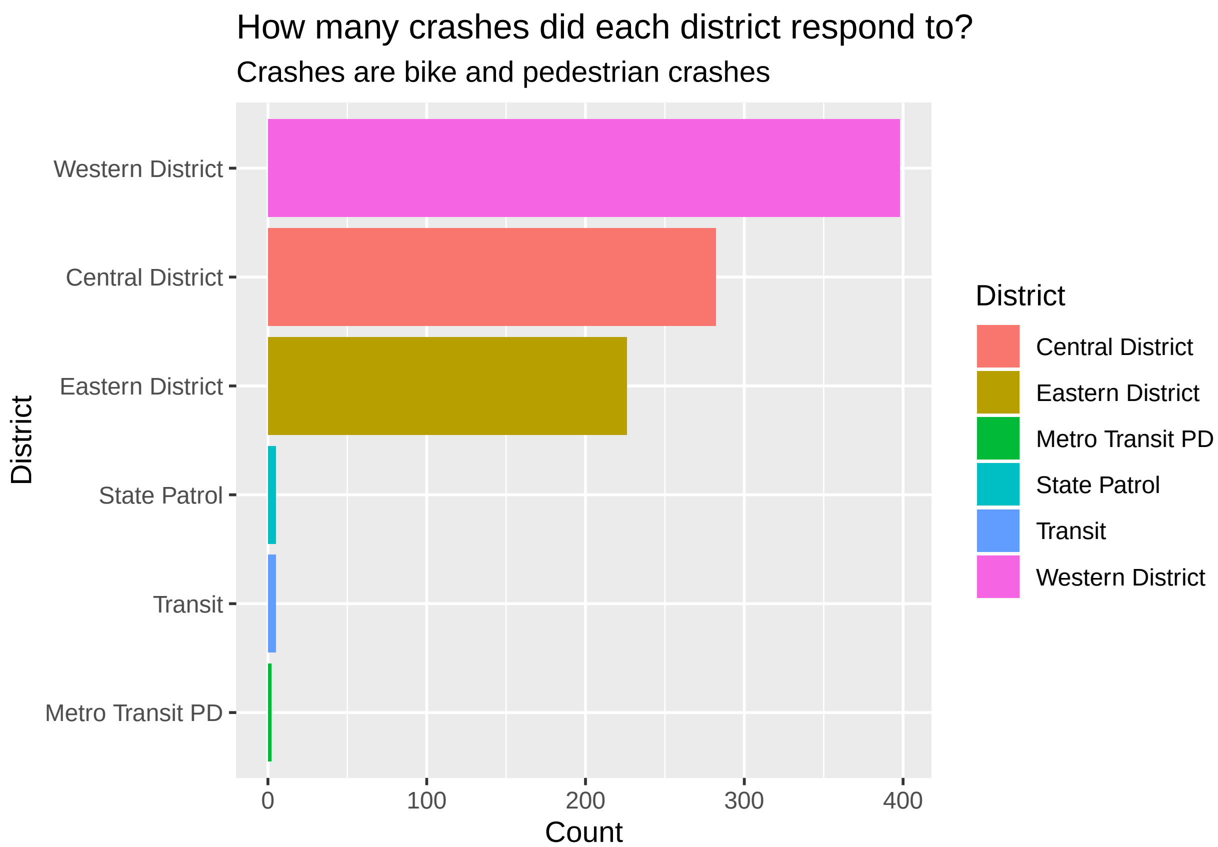
I could have asked how many crashes occurred in total, but it would be an uninteresting graph. It would be a simple calculation of sum(count), invariant of crash type. TO make it more interesting, I want to see how many of each crash types occurred, and the reader can infer for themselves the total count.

How many crashes occurred for each crash type?

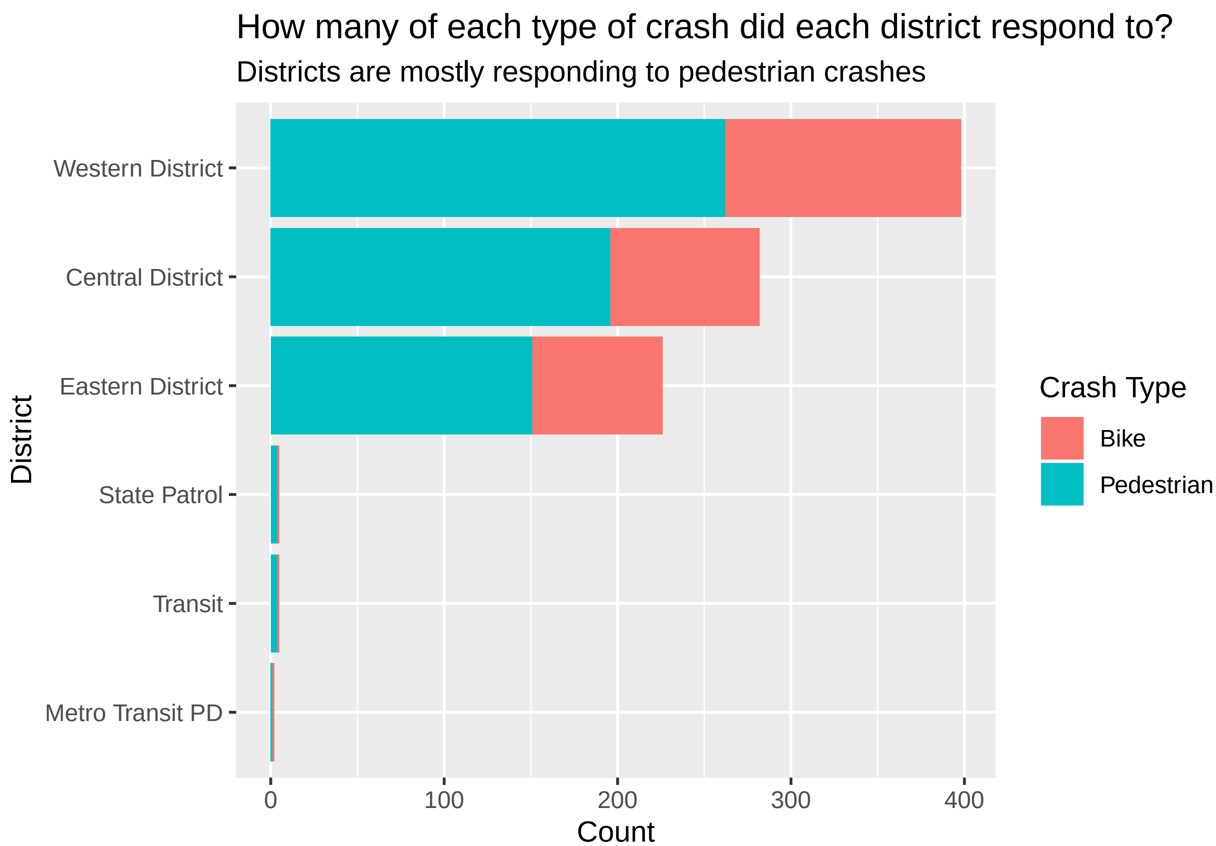
Crash types are bikes and pedestrians



2. Which districts responded to the crashes and how many did they respond to?



3. What were the most common accidents (among bikes and pedestrian crashes) for each district?

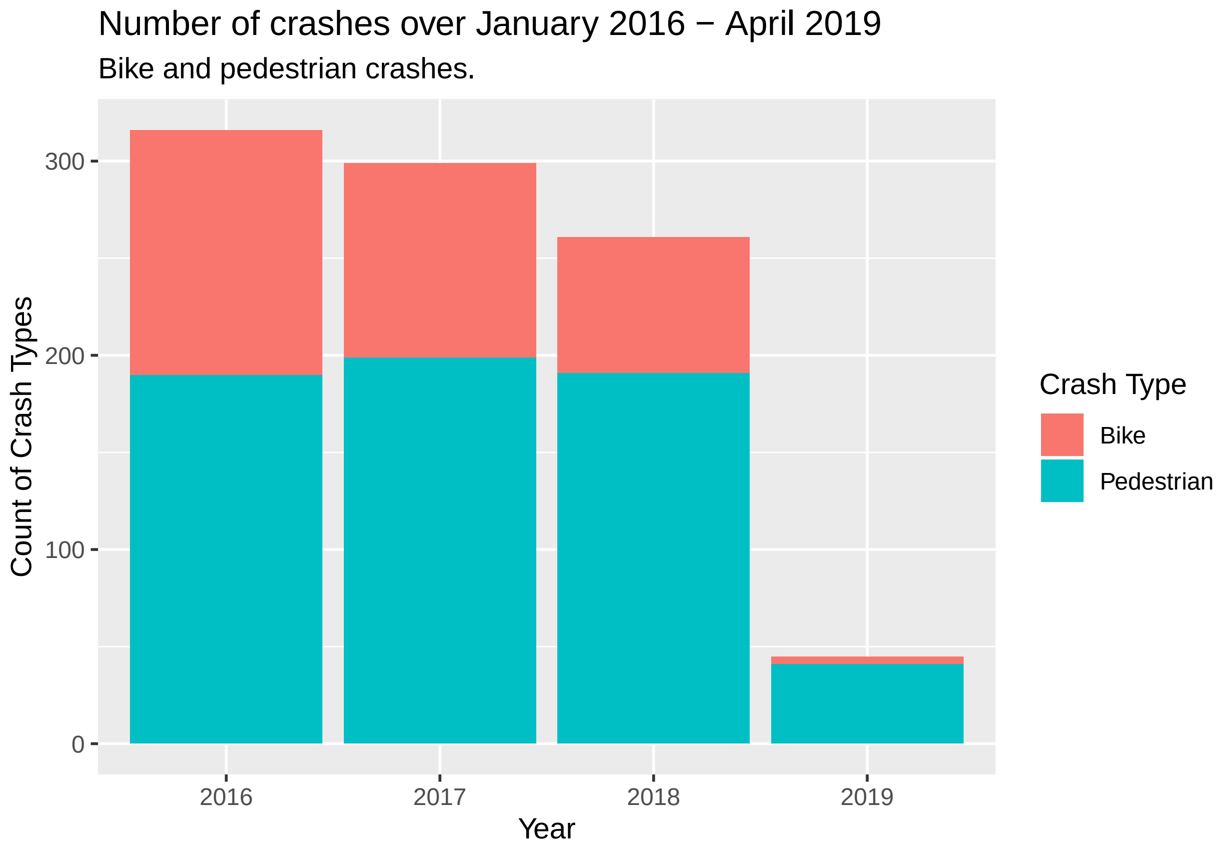


4. How do incidents of each type of crash differ across time?

The Date/Time variable contains a TON of information. We can filter by time hr. $\in$ (0-24) hr periods, days $\in$ (0-365), months $\in$ (1-12), and years $\in$ (2016-2018). The selection choice of date and time will depend on the relevant question of interest…

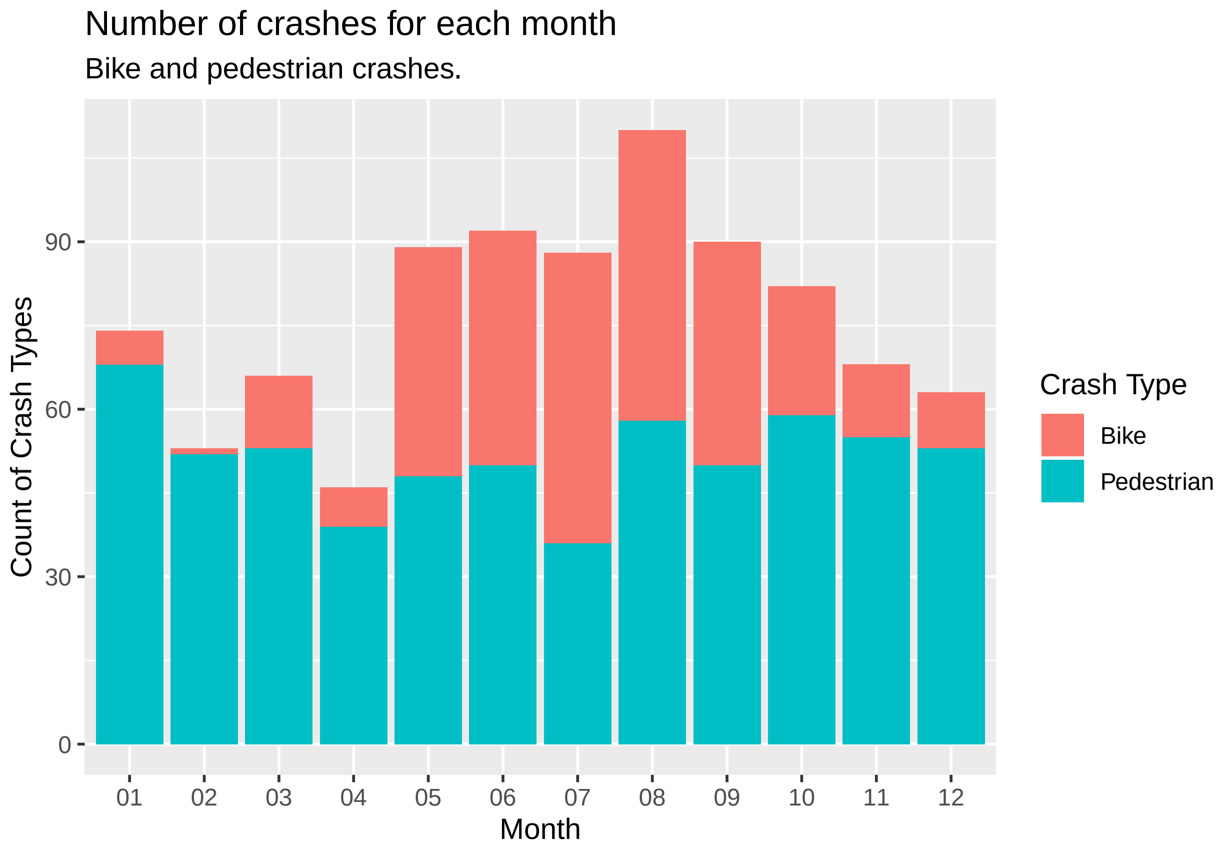
For the purpose of brevity, the question boils down to a time-series situation, where we will be choosing the period or seasonality of crash incidents. For this section, I will bin by years, months, hours, and month-day combinations. I didn't start out wanting to see all these graphs, but the more graphs I created, the more curious I became about what the data could tell me – very fortunate for the reader :)

Yearly data



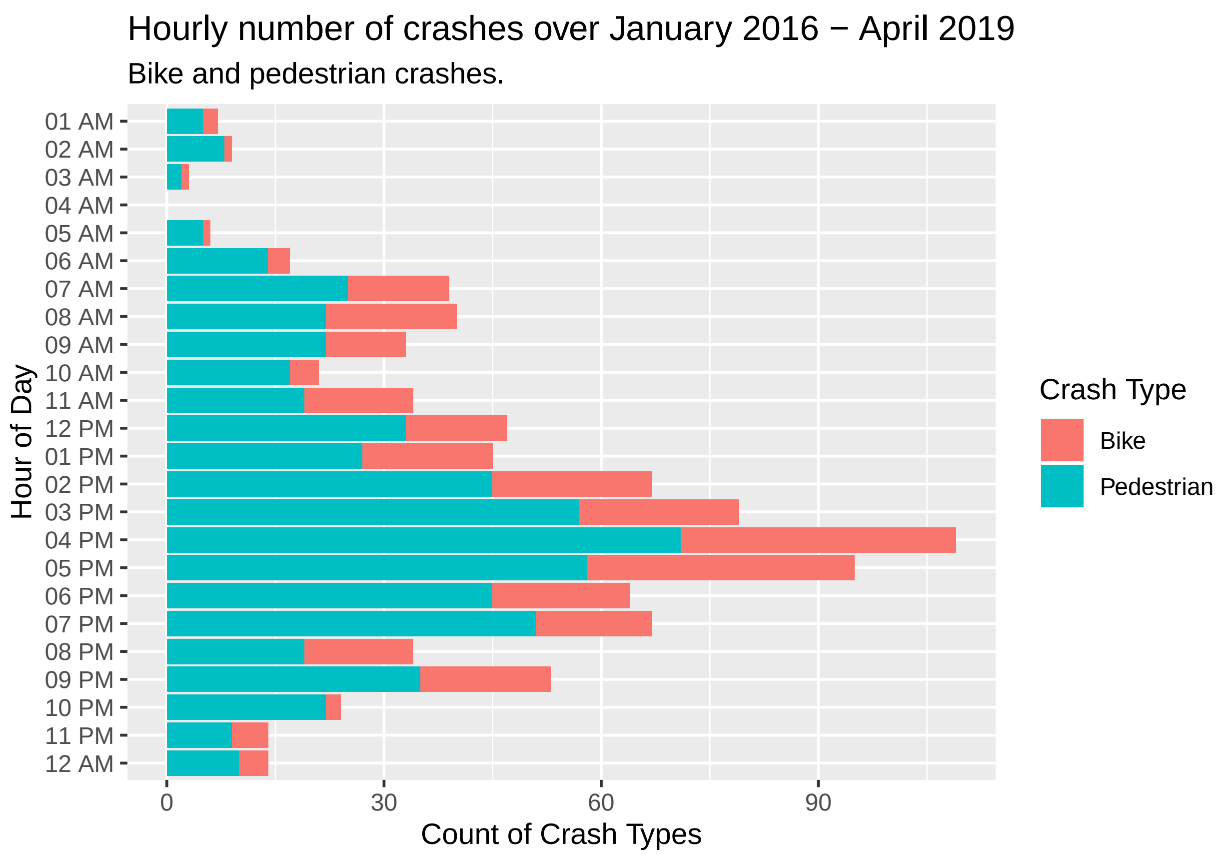
Clearly, we can now see that there is a general downward trend where the number of total crashes have been decreasing over the years. One conclusion that we can tentatively state is that crash incidents tend to peak up during the summer season, tapering off in other seasons; overall crash incidents have been decreasing over the years. In the summer months, there is a distinguishable increase in the number of bikes out on the streets, which is the main driver for more accidents occuring over the summer. However, there are more pedestrian accidents during the winter months, which suggest that cross-walks and sidewalks may be difficult to navigate due to ice and snow, which is abundant in Minnesota.

Monthly Data



It is interesting to see that many crashes occur during the summer months and it drops off as it enters into the winter to spring session, picking back up to summer. However, it doesn't capture how crashes have been changing over the years, so now I would like to see that.

Hourly Data

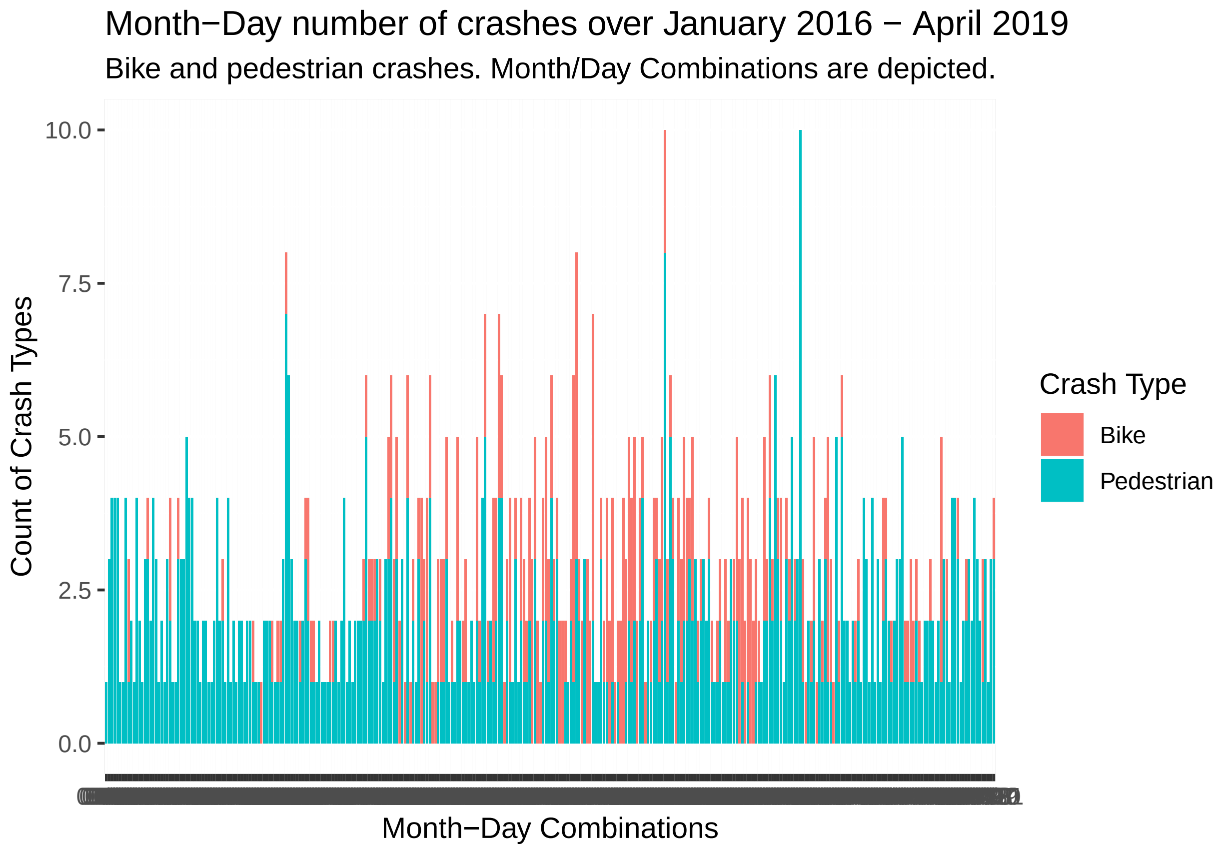


Kind of weird that there's nothing going on for 04 AM in the morning, but the data set has no data for the 04 AM time period, which is shown below, alongside with the 05 AM time chunk to verify that the code works.

```
## [1] Crash.Type Count      hourly
## <0 rows> (or 0-length row.names)
```

```
##     Crash.Type Count hourly
## 320 Pedestrian     1  05 AM
## 321 Pedestrian     1  05 AM
## 322 Pedestrian     1  05 AM
## 323 Pedestrian     1  05 AM
## 324 Pedestrian     1  05 AM
## 325       Bike     1  05 AM
```

Month-Day combination Data



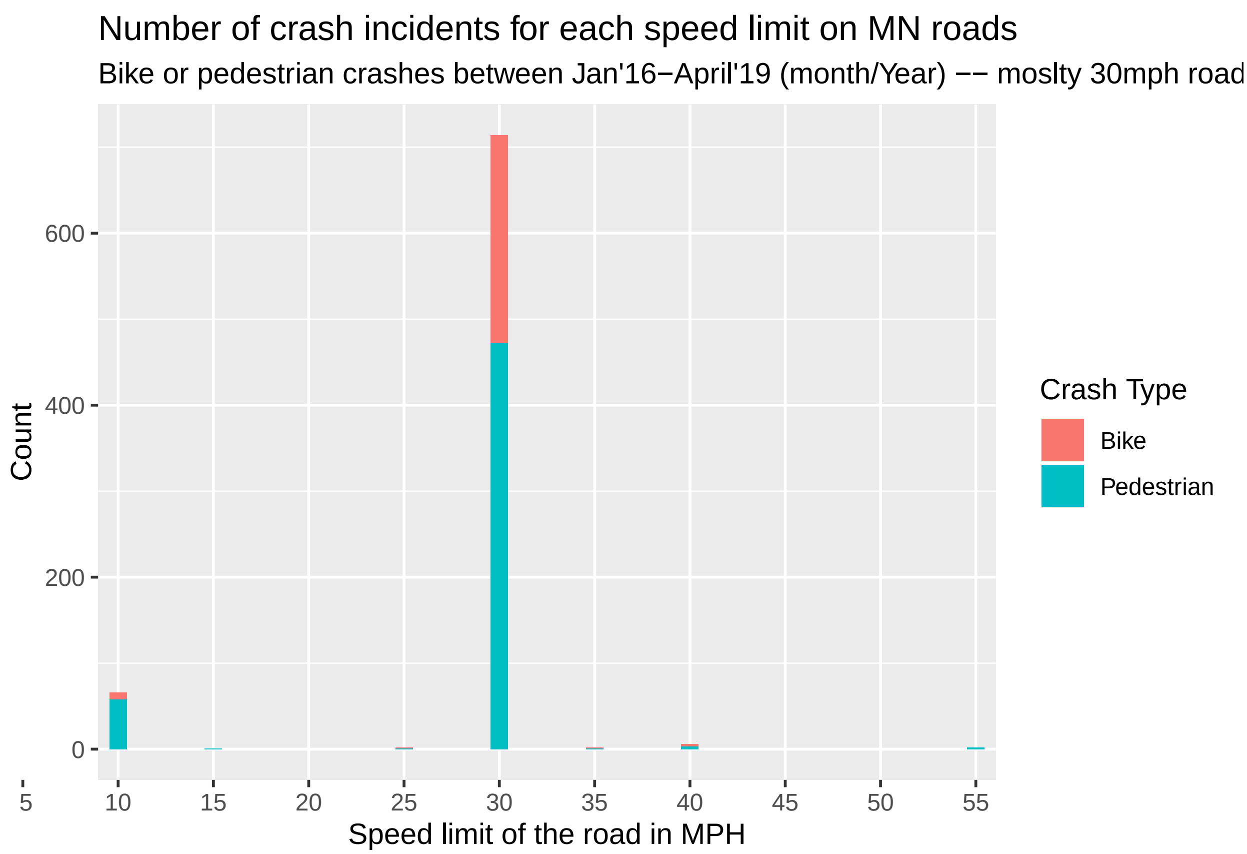
What I have done is selected month/day combinations the crash totals by crash type. Now this is evidently an awful graph to show to an audience based on the 'Month-Day' axis labelling – it is far too granular to depict every month-day combination clearly. While this graph is far from publish-ready, it shows some interesting peaks which may correspond to certain holidays or festivities. It might be worth investigating which days these are to find out when crashes will more likely occur. It may also show that some of these 'day' values are significant drivers to what makes a month appear to have a higher than average crash incident rate.

```
## # A tibble: 5 x 2
##   Date...Time highs
##   <chr>       <int>
## 1 08/16          10
## 2 10/06          10
## 3 03/14           8
## 4 07/13           8
## 5 06/08           7
```

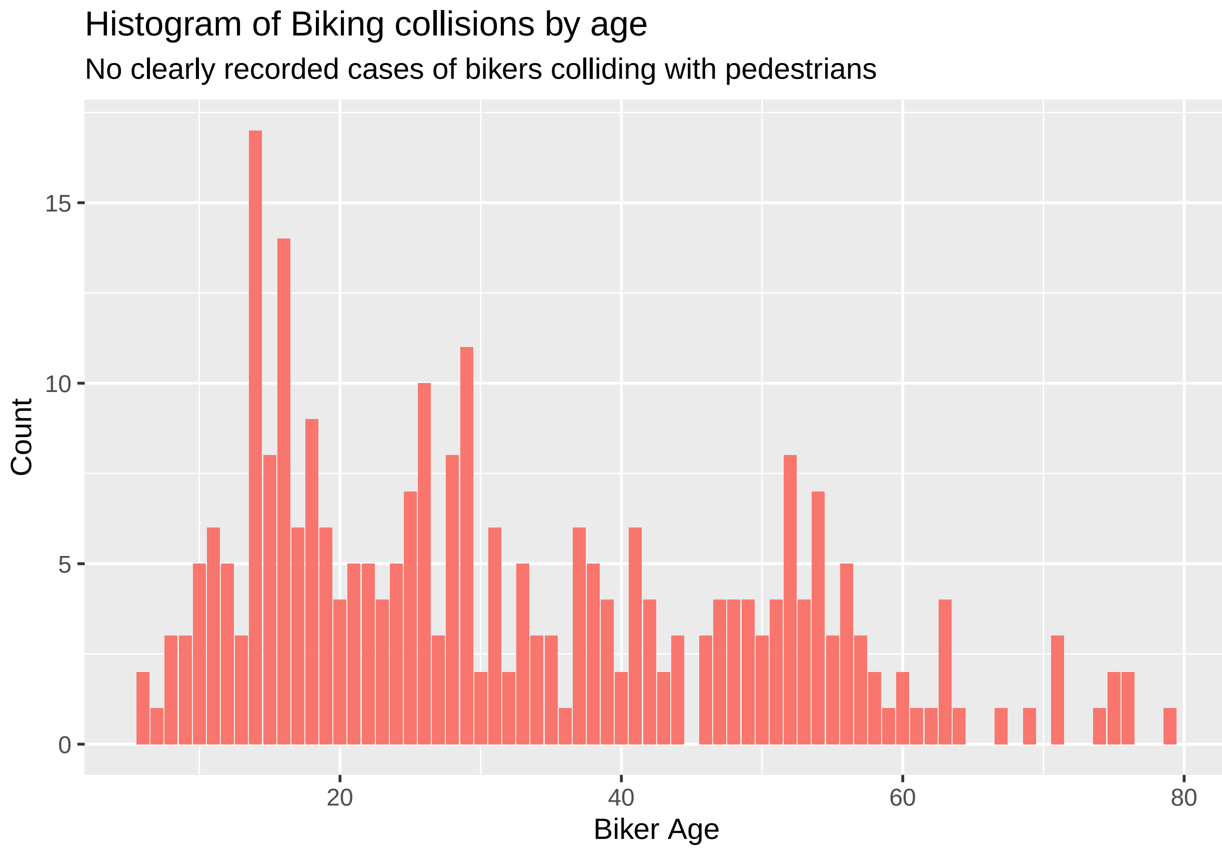
I don't recognize these dates as being anything of a significant holiday or festivity, so the month-daily combination incident rates may be random noise, or there may be some other latent variable that I am not accounting for at this time.

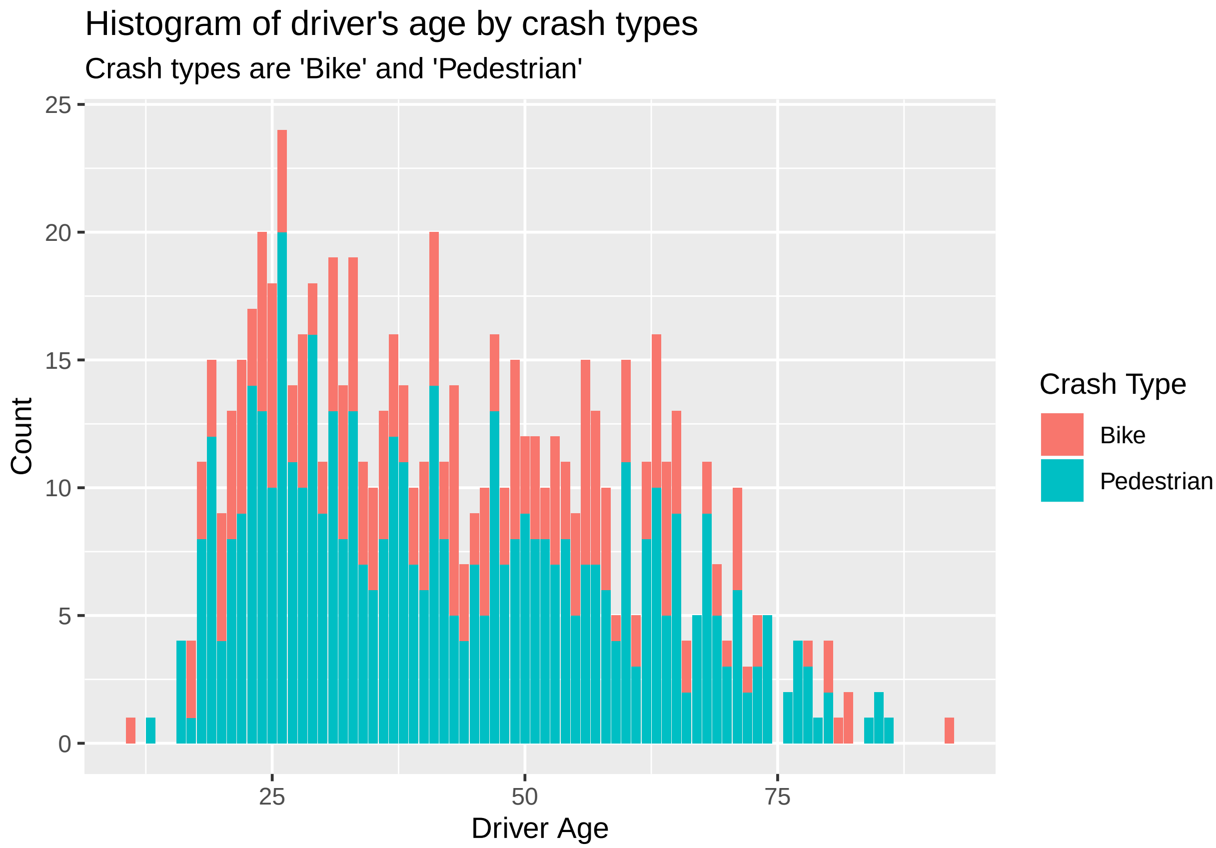
5. How many of each crash type occurred for each speed limit?

Unfortunately, there are a lot of NA values for speed limit. Furthermore, most of the data cotain speed limits of 30 MPH, so it is largely uninteresting. However, the gist of it is that if the data set *were* more interesting, we could get some insight into how speed limits play a factor in crash incident types and rates.


Number of crash incidents for each speed limit on MN roads
Bike or pedestrian crashes between Jan'16−April'19 (month/Year) −− moslty 30mph road
Count
600
400
200
0
5
10
15
20
25
30
35
40
45
50
55
Speed limit of the road in MPH
Crash Type
Bike
Pedestrian

6. How are ages associated with crash counts and types?

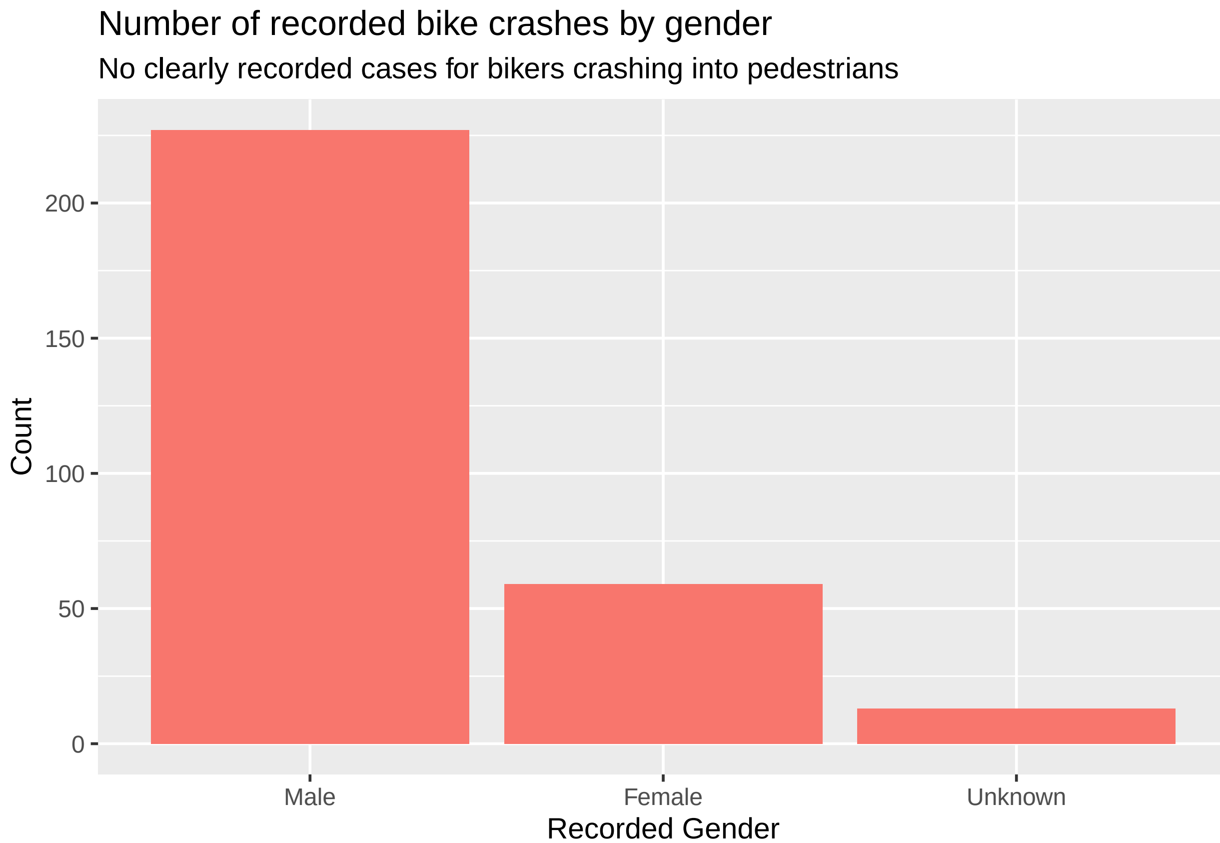


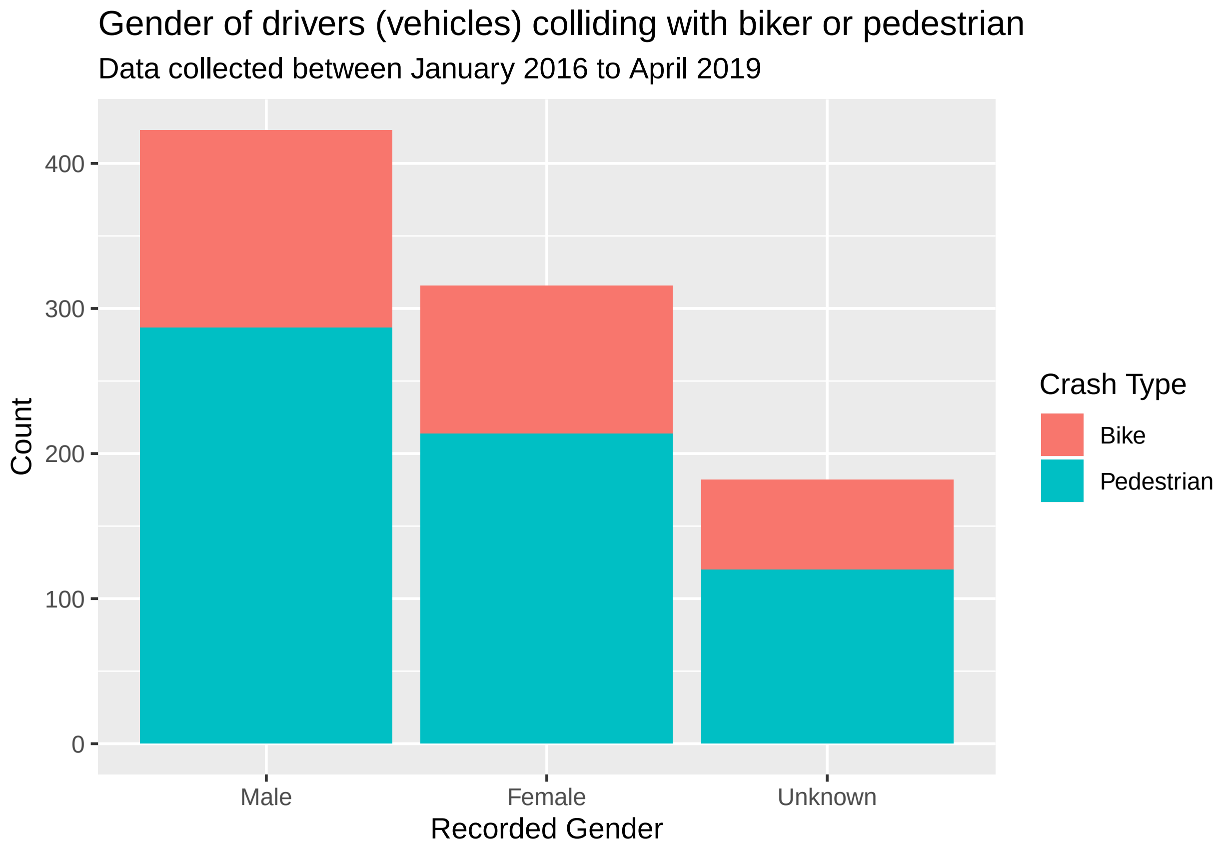


For some cross sections of age, there are more biker collissions than pedestrian collisions. Although, in general, there are more pedestrian collisions than biking collisions with drivers for most age groups. Younger drivers (e.g. around 25 years old) tend to get into collisions with bikers and pedestrians than older age groups.

7. How are genders associated with crash counts and types?

The data set is interesting for this one since it partitions the gender according to crash type. It would be interesting to introduce facets of graphs in this scenario, although I need to tidy up the data – some of the data entry is inconsistent (e.g most records state "Male", but there are some entries labelled 'ma', 'Ma', 'male' and 'Fe' for the more common 'Female'), so a lot of data needs to be fixed and aggregated.

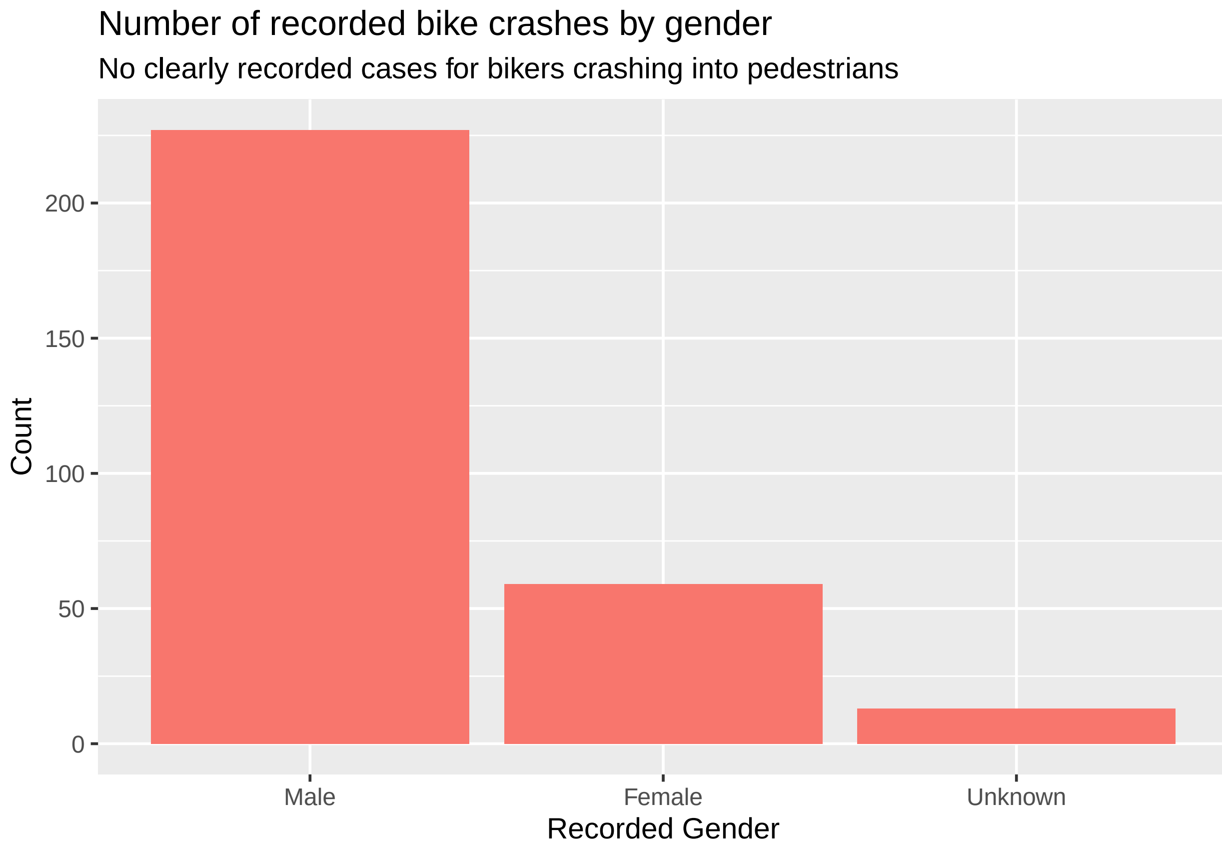

Number of recorded bike crashes by gender
No clearly recorded cases for bikers crashing into pedestrians
Count
200
150
100
50
0
Male
Female
Unknown
Recorded Gender

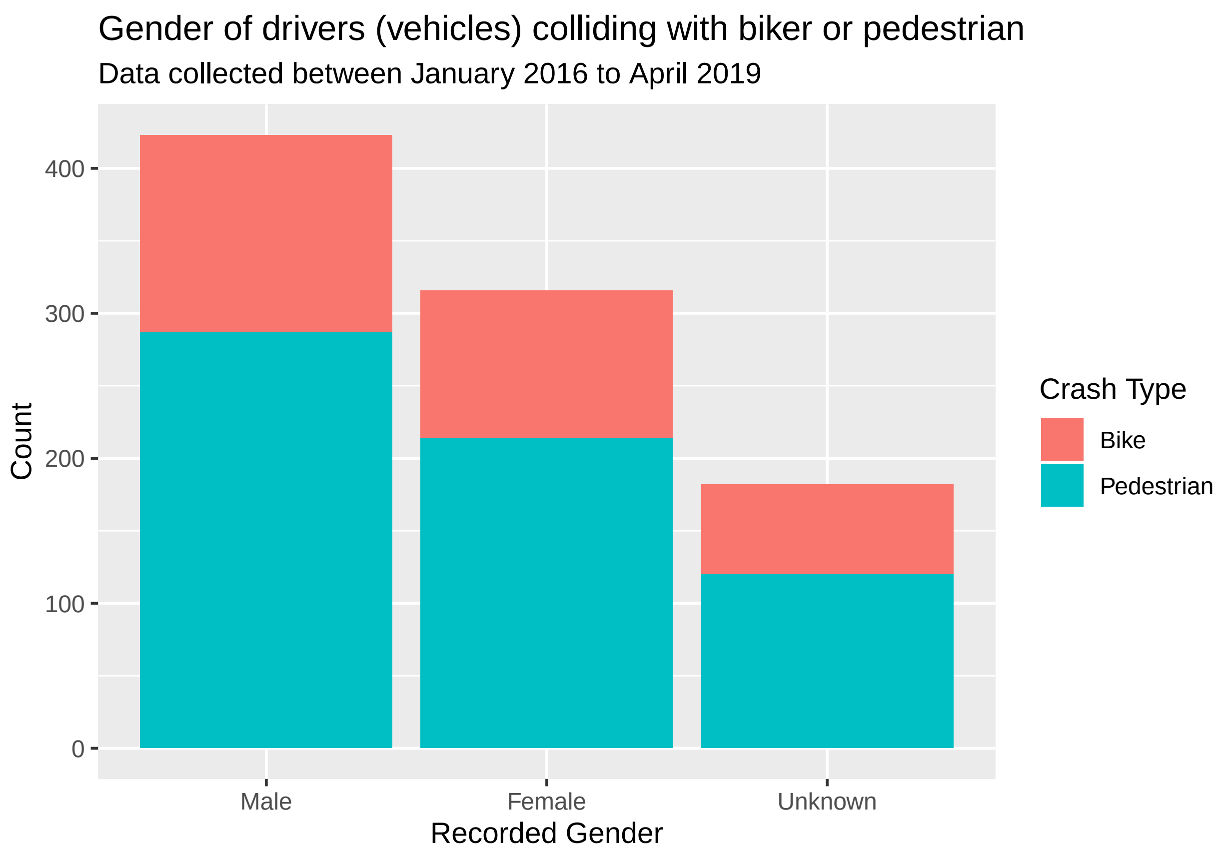


In general, males are more involved in bike and pedestrian crashes – although there is some latent variable here, and that is that men typically drive more miles than females so their chances are getting into an accident scales proportionally up with the time spent on the roads. There is no such variable to see it clearly here, but there are some documented studies of this – a short google search away show these studies.

In other words, there is difficulty in visualizations like these; there is an implicit bias to think that x-axis is an independent variable that effects the y-axis dependent variable. As much as I'd like to say no such thing is happening here, I would be wrong because it is very easy to think that men are more involved in accidents than women. A better way to think about these graphs are association/correlation as opposed to causation because graphs with few variables rarely paint a full picture.

The data set is interesting for this one since it partitions the gender according to crash type. It would be interesting to introduce facets of graphs in this scenario, although I need to tidy up the data – some of the data entry is inconsistent (e.g most records state "Male", but there are some entries labelled 'ma', 'Ma', 'male' and 'Fe' for the more common 'Female'), so a lot of data needs to be fixed and aggregated.


Number of recorded bike crashes by gender
No clearly recorded cases for bikers crashing into pedestrians
Count
200
150
100
50
0
Male
Female
Unknown
Recorded Gender



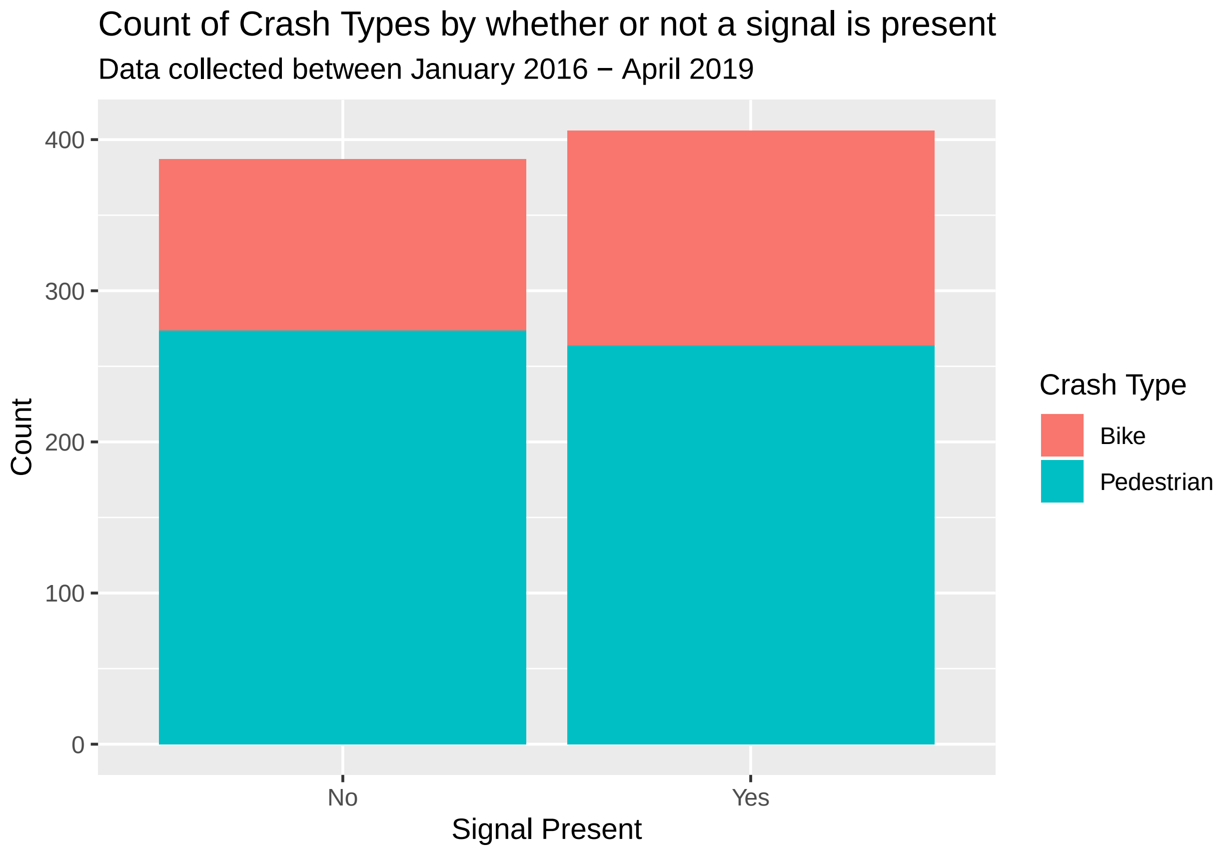
In general, males are more involved in bike and pedestrian crashes – although there is some latent variable here, and that is that men typically drive more miles than females so their chances are getting into an accident scales proportionally up with the time spent on the roads. There is no such variable to see it clearly here, but there are some documented studies of this – a short google search away show these studies.

In other words, there is difficulty in visualizations like these; there is an implicit bias to think that x-axis is an independent variable that effects the y-axis dependent variable. As much as I'd like to say no such thing is happening here, I would be wrong because it is very easy to think that men are more involved in accidents than women. A better way to think about these graphs are association/correlation as opposed to causation because graphs with few variables rarely paint a full picture.

8. Are there areas where accidents are generally more clustered?

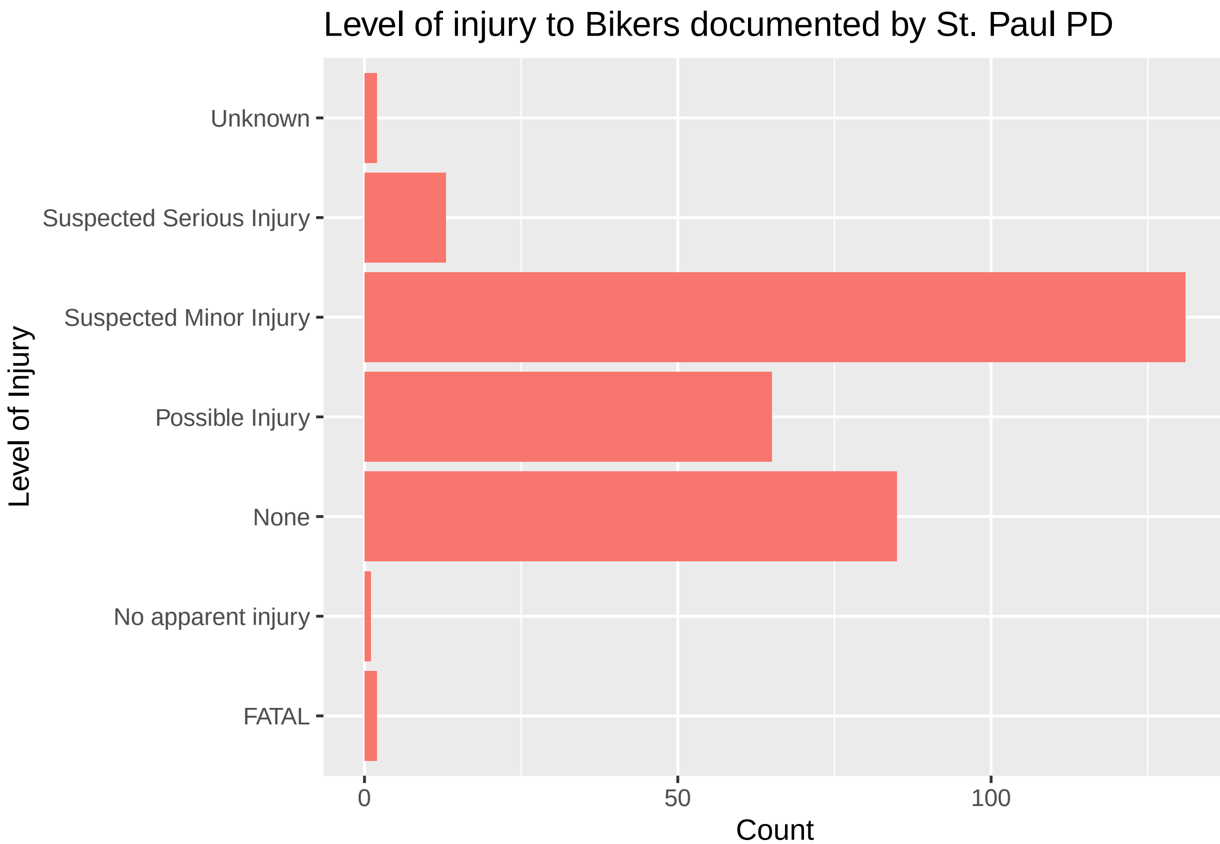
Spatial statistics can be very interesting to locate areas that contain high densities of crash incidents. A preliminary gesture in this direction uses the leaflet package. This will be shown in a different document.

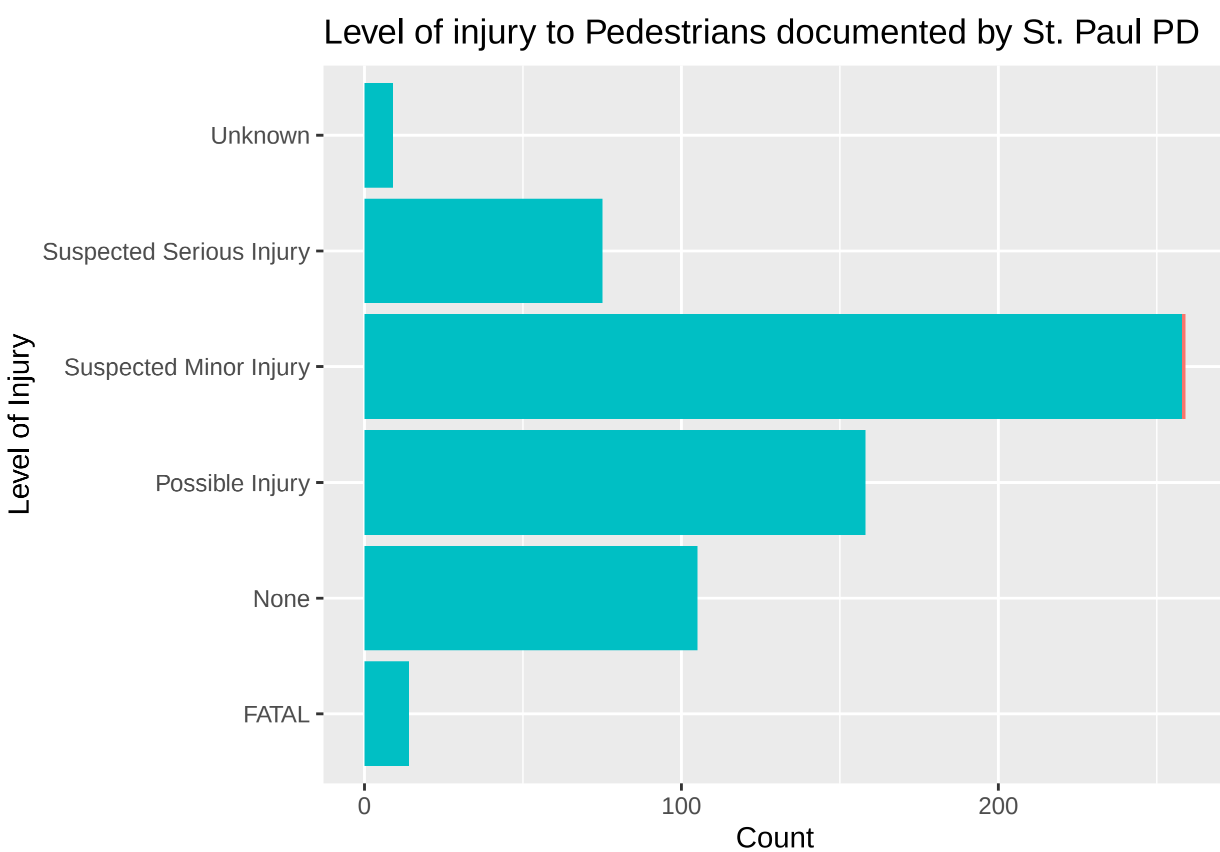
9. Does a signal being present have an effect on crash incidents?



It doesn't appear that having a signal present matters when it comes to crash incidents, which is surprising to me.

10. What type of crash incident is associated with more severe injuries?





Summary of EDA

From EDA, I can tentatively say the following:

1. More pedestrian crashes occur than bike crashes.
2. Most crashes occur under jurisdictions of the Western, Central, and Eastern Districts.
3. By year, crashes have been trending downward.
4. By month, crashes tend to occur more frequently in the summer season.
5. By hour, the 2:00pm - 7:00pm time slot seems to have the most frequent incidents of crashes.
6. Younger bikers and drivers tend to get more involved in crashes.
7. Males are associated with higher crash incidents.
8. Signals do not have an obvious effect on crash incidents.

These EDA insights are worth studying more in-depth using modelling and predictive techniques, but the purpose here was to work with some EDA techniques using ggplot2 and tidyverse.