Background

We're going to explore a large data set or traffic crashes to learn about what factors are connected with injuries. We will use data from the city of Chicago's open data portal. (This activity is derived from a blog post by Julia Silge)

```
years_ago <- mdy("01/01/2022") # data from last 2 years. May take time to load!
crash_url <- glue::glue("https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3i
crash_raw <- as_tibble(read.socrata(crash_url)) # a new way to read in data, don't worry about it!
```

This dataset is pretty crazy! Take a look at it in the viewer, and then let's do some data munging to get it into a nicer form.

-create a variable called `injuries` which indicates if the crash involved injuries or not. -create an unknown category for missing `report_type`s -decide which other variables to keep

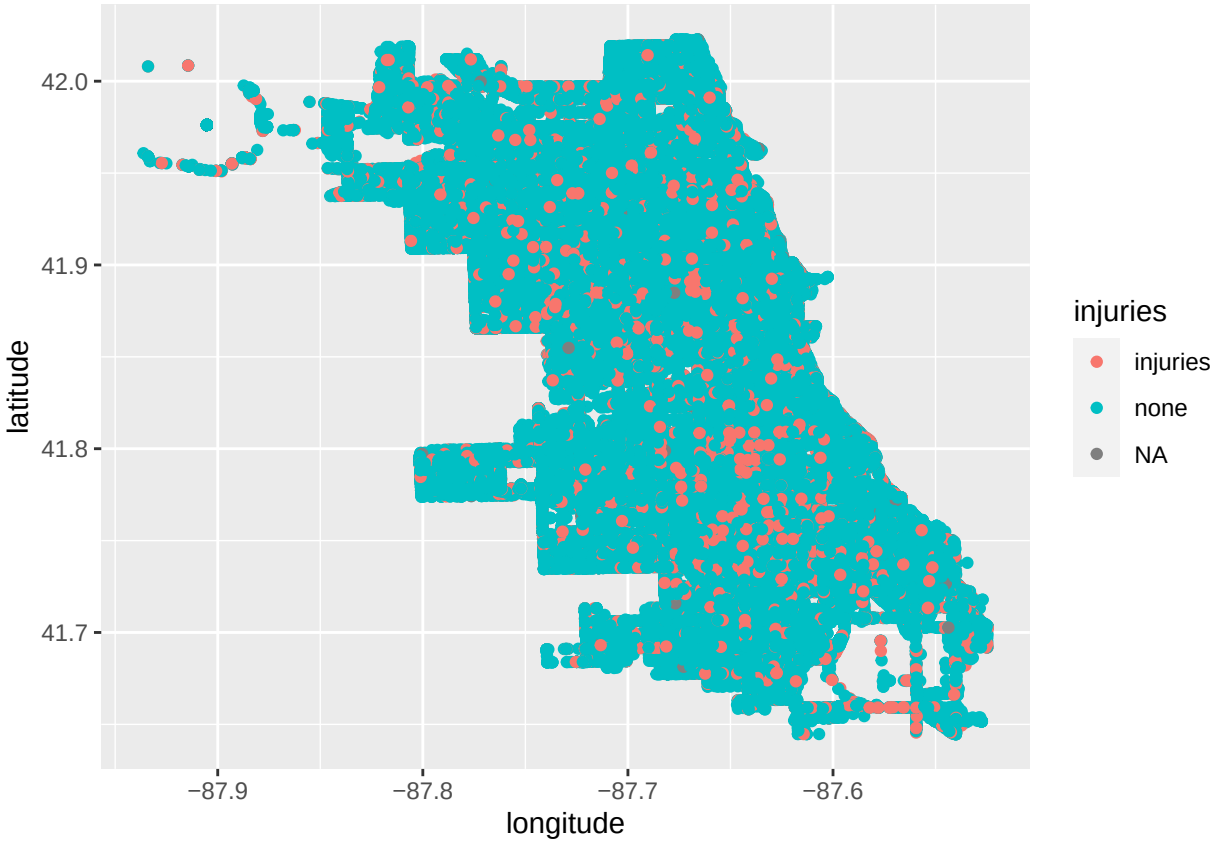
```
crash <- crash_raw %>%
  arrange(desc(crash_date)) %>%
  transmute(
    injuries = as.factor(if_else(injuries_total > 0, "injuries", "none")),
    report_type = replace(crash_raw$report_type, crash_raw$report_type == "", "Unknown"),
    # choose your variables here (use ?transmute to see what this does)
    latitude, longitude, crash_date
  )
```

Exploratory Data Analysis

Here's a few questions to get you started.

1. Take a look at crashes by latitude and longitude, colored by injuries. What do you notice?

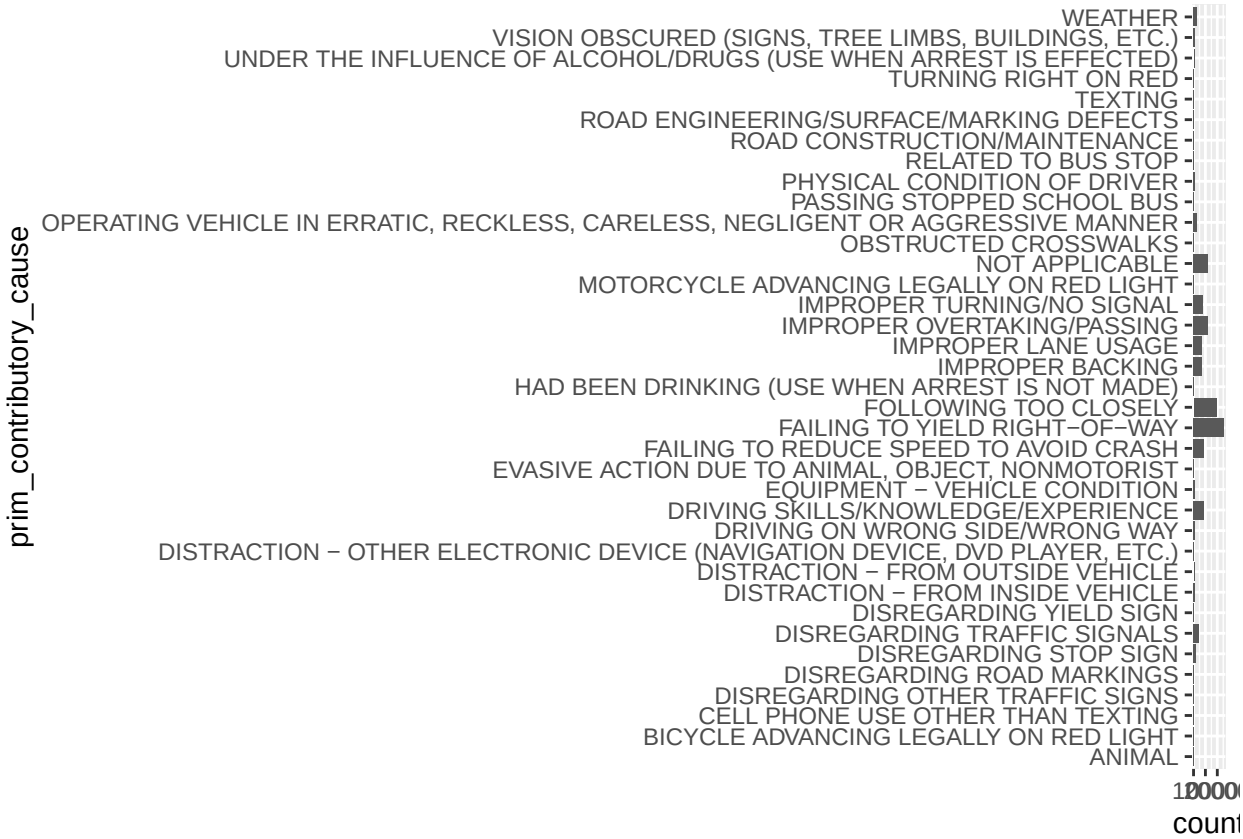
```
ggplot(
  crash %>% filter(latitude > 0),
  mapping=aes(x=longitude, y=latitude)
) +
  geom_point(mapping=aes(color=injuries))
```



One thing I notice is that the outline that forms looks like the city limits of Chicago. Along with this there seems to be a number of areas where injuries occur more frequently then they don't.

2. What are the most common contributing factors to a crash?

```
ggplot(
  crash_raw %>% filter(prim_contributory_cause != "UNABLE TO DETERMINE"),
  aes(x=prim_contributory_cause)
)+
  geom_bar()+
  coord_flip()
```




```
crash1 <- crash_raw %>%
  count(prim_contributory_cause) %>%
  arrange(desc(n)) %>%
  mutate(prim_contributory_cause = fct_reorder(prim_contributory_cause, n)) %>%
  slice(2:4,6:7) %>%
  ggplot( aes(y = factor(prim_contributory_cause), x=n)) +
  geom_bar(stat="identity") +
  labs(title="Primary Causes of Crash, top 5", y="")
```

The top 5 common factors to a crash are; 1) Failing to yield right of way, 2) Following too closely, 3) Improper overtaking/passing, 4) Failing to reduce speed to avoid crash, 5) driving skills/knowledge/experience.

3. How do crashes vary month by month? Compare crashes by month in 2022 to 2023.

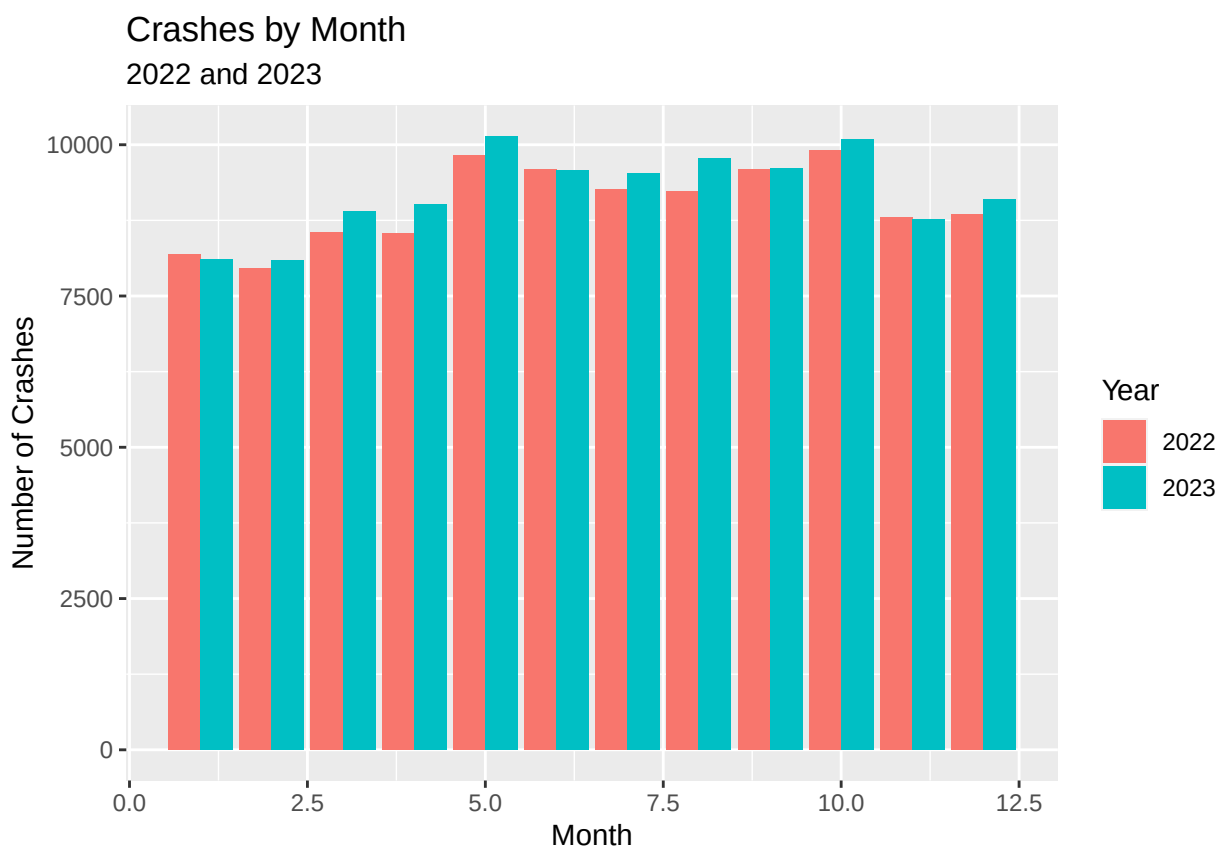
```
crash_2022 <- crash_raw %>% filter(as.Date(crash_date) >= as.Date('2022-01-01') & as.Date(crash_date) <

crash_2023 <- crash_raw %>% filter(as.Date(crash_date) >= as.Date('2023-01-01') & as.Date(crash_date) <

combined_years <- bind_rows(
  mutate(crash_2022, year = 2022),
  mutate(crash_2023, year = 2023)
)

ggplot(combined_years, aes(x = crash_month, fill = as.factor(year))) +
  geom_bar(position = "dodge", stat = "count") +
```

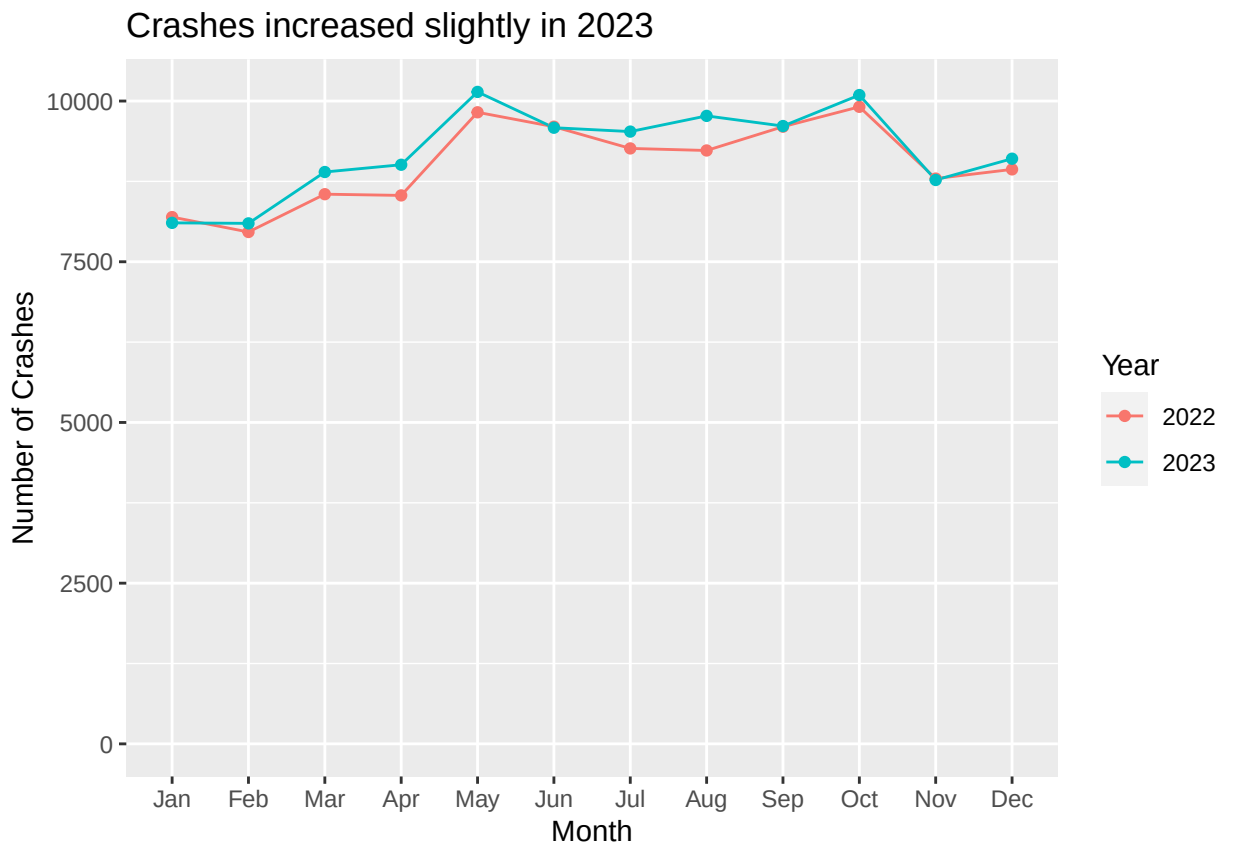
```
  labs(title = "Crashes by Month",
       subtitle = "2022 and 2023",
       x="Month",
       y="Number of Crashes",
       fill="Year")
```




```
q <- crash_raw %>%
  mutate(
    crash_year = as.factor(year(crash_date)),
    month_name = month(crash_month, label = TRUE, abbr = TRUE)
    ) %>%
  drop_na(crash_year) %>%
  filter(crash_year == 2022 | crash_year == 2023) %>%
  group_by(crash_year, month_name) %>%
  summarize(n = n()) %>%
  ggplot(aes(x=month_name, y = n, group=crash_year, color=crash_year)) +
  geom_line() + geom_point() +
  ylim(c(0,NA)) +
  labs(title="Crashes increased slightly in 2023",
       y="Number of Crashes", x="Month", color="Year")
```

```
## `summarise()` has grouped output by 'crash_year'. You can override using the
## `.groups` argument.
```

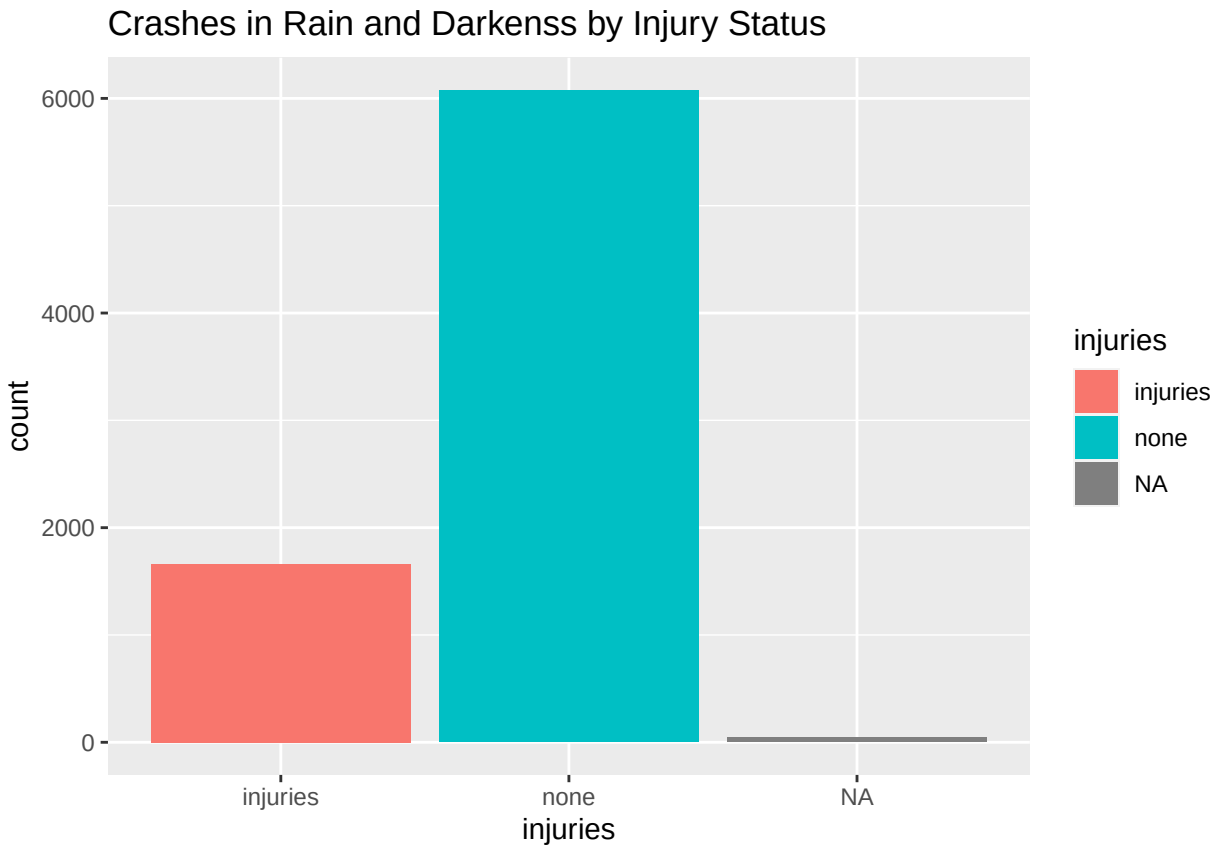
```
q
```



When looking at all the data, both May and October are the two months that have the most crashes. As you can see by looking at these graphs, there seems to be a very similar distribution of crashes throughout the months.

4. Are crashes more likely to cause injuries when it is rainy and dark? Use the variables `weather_condition` and `lighting_condition` to explore.

```
ggplot(
  crash_raw %>%
    filter(weather_condition == "RAIN" & (lighting_condition == "DARKNESS" | lighting_condition == "DARI
    mutate(injuries = if_else(injuries_total > 0, "injuries", "none")),
  aes(x = injuries)
) +
geom_bar(aes(fill=injuries)) +
labs(title = "Crashes in Rain and Darkenss by Injury Status")
```



When looking at injuries that occurred under the conditions when it was rainy and dark, with or without lighted roads, injuries do not occur more often then not. As you can see in the graph above, of the many crashes that occurred under these conditions injuries happened in roughly 1 out of every 5 crashes.

5. Choose a question you want to explore, and create an appropriate visual.

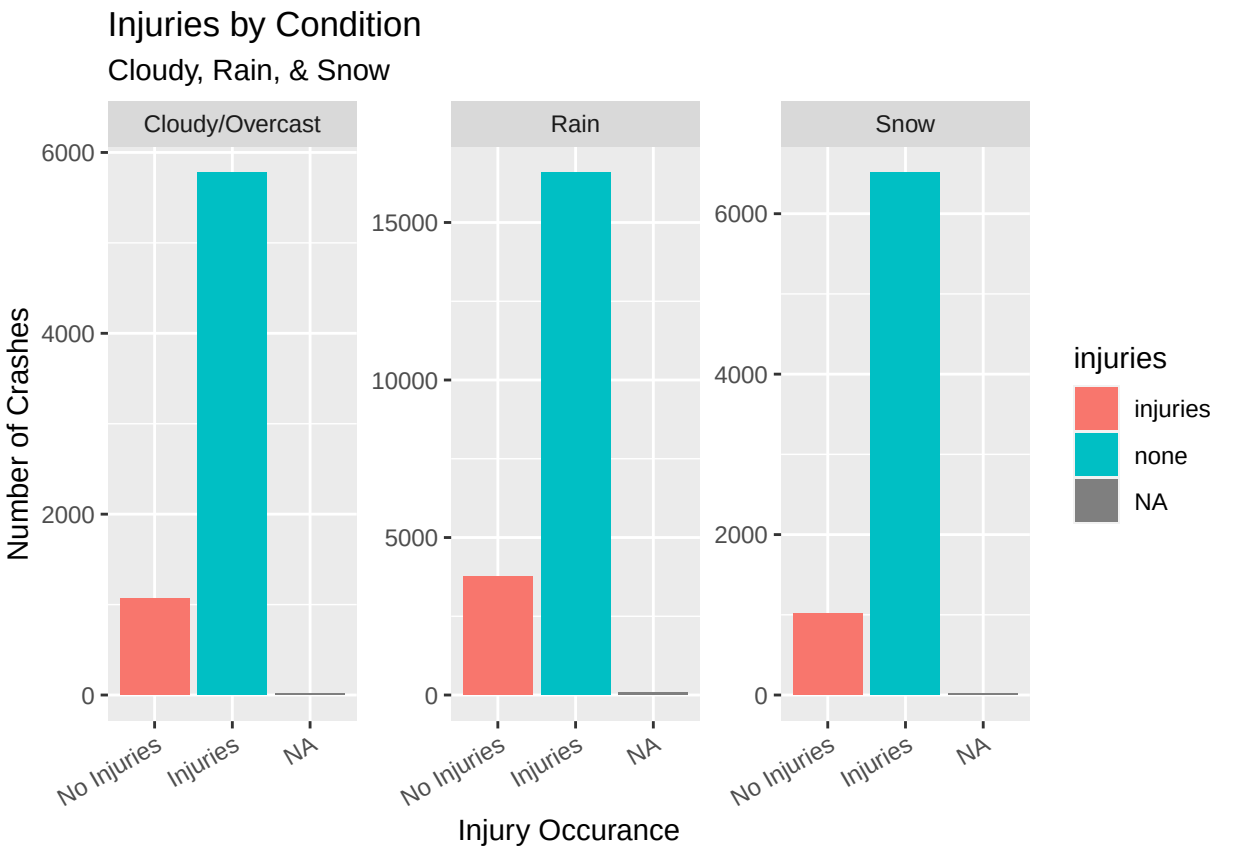
What conditions cause the highest percentage of injuries?

```
snow <- crash_raw %>% filter(weather_condition == "SNOW")
rain <- crash_raw %>% filter(weather_condition == "RAIN")
clear <- crash_raw %>% filter(weather_condition == "CLEAR")
cloudy <- crash_raw %>% filter(weather_condition == "CLOUDY/OVERCAST")

combined_conditions <- bind_rows(
  mutate(snow, condition = "Snow"),
  mutate(rain, condition = "Rain"),
  mutate(cloudy, condition = "Cloudy/Overcast")
)

ggplot(combined_conditions %>%
    mutate(injuries = if_else(injuries_total > 0, "injuries", "none")),
    aes(x = injuries)
) +
  geom_bar(aes(fill=injuries)) +
  labs(title = "Injuries by Condition",
       subtitle = "Cloudy, Rain, & Snow",
```

```
      y = "Number of Crashes",
      x = "Injury Occurance") +
facet_wrap(~condition, scales = "free_y") +
scale_x_discrete(labels = c("No Injuries", "Injuries"))  +
theme(axis.text.x = element_text(angle = 30, hjust = 1))
```



By looking at these even though we can see the y-axis are set to different scales, the bars that depict when there were no injuries are all at a equal height. Therefore we can then look at the proportions of each graph to determine which condition has the highest probability to have an injury in an accident. Out of these three conditions we can see that Rain has the highest probability, followed by Cloudy or Overcast and then lastly, Snow.