Amusement Park Injuries

Andrew Farina

null

TidyTuesday

These data are from the #TidyTuesday (10 Sept 2019) project. #TidyTuesday is a weekly data project aimed at the R ecosystem. As this project was borne out of the R4DS Online Learning Community and the R for Data Science textbook, an emphasis was placed on understanding how to summarize and arrange data to make meaningful charts with ggplot2, tidyr, dplyr, and other tools in the tidyverse ecosystem.

The intent of Tidy Tuesday is to provide a safe and supportive forum for individuals to practice their wrangling and data visualization skills independent of drawing conclusions. While we understand that the two are related, the focus of this practice is purely on building skills with real-world data.

Amusement Park Injuries

This particular dataset is from the SaferParks Database.

A lot of free text in these data, some inconsistent NAs (n/a, N/A) and dates (ymd, dmy). A good chance to do some data cleaning and then take a look at frequency, type of injury, and analyze free text.

```
safer_parks <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da
glimpse(safer_parks)
```

```
## Observations: 8,351
## Variables: 23
## $ acc_id              <dbl> 1005813, 1004032, 1007658, 1007098, 1000094, 1...
## $ acc_date            <chr> "6/12/2010", "6/12/2010", "7/10/2010", "7/10/2...
## $ acc_state           <chr> "OH", "OH", "CA", "CA", "CO", "WI", "WI", "CO"...
## $ acc_city            <chr> "Cleveland", "Cleveland", "Anaheim", "Carlsbad...
## $ fix_port            <chr> "F", "P", "F", "F", "F", "F", "P", "F", "P", "...
## $ source              <chr> "Ohio Dept. of Agriculture", "United States Co...
## $ bus_type            <chr> "Sports or recreation facility", "Sports or re...
## $ industry_sector     <chr> "recreation", "recreation", "amusement ride", ...
## $ device_category     <chr> "inflatable", "inflatable", "water ride", "flo...
## $ device_type         <chr> "Inflatable slide", "Inflatable slide", "Boat ...
## $ tradename_or_generic <chr> "inflatable slide", "inflatable slide", "boat ...
## $ manufacturer        <chr> "Scherba Industries / Inflatable Images", "Sch...
## $ num_injured         <dbl> 9, 8, 1, 1, 1, 1, 1, 20, 1, 1, 2, 1, 1, 1, 1, ...
## $ age_youngest        <dbl> NA, 54, 37, 37, NA, 12, 16, NA, 14, NA, 16, 36...
## $ gender              <chr> NA, "M", "F", "F", "M", "F", "F", NA, "M", NA,...
## $ acc_desc            <chr> "Inflatable slide tipped over while 7-9 patron...
## $ injury_desc         <chr> "The man who was crushed by the device died 9 ...
```

```
## $ report                  <chr> "https://saferparksdata.org/sites/default/file...
## $ category                <chr> "Device tipped over, blew away, or collapsed",...
## $ mechanical              <dbl> NA, NA, NA, NA, 1, NA, 1, NA, NA, NA, 1, NA, N...
## $ op_error                <dbl> 1, 1, NA, NA, NA, 1, NA, 1, NA, NA, NA, NA, NA...
## $ employee                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ notes                   <chr> "http://www.cleveland.com/metro/index.ssf/2012...
```

The safer_parks dataset contains 8351 incidents that were recorded from 17 different types of parks.

```r
safer_parks %>% dplyr::count(bus_type, sort = TRUE, name = "number_of_incidents")
```

```
## # A tibble: 17 x 2
##    bus_type                       number_of_incidents
##    <chr>                                        <int>
##  1 Amusement park                                3667
##  2 Water park                                    1767
##  3 Carnival or rental                             701
##  4 Trampoline park                                698
##  5 Family entertainment center                    484
##  6 Go kart track                                  228
##  7 Sports or recreation facility                  178
##  8 Mountain resort                                174
##  9 School or church                                95
## 10 City or county park                             90
## 11 Adventure course                                86
## 12 Zoo or museum                                   61
## 13 Mall, store or restaurant                       53
## 14 Pool waterslide                                 27
## 15 Unknown                                         19
## 16 Camp                                            14
## 17 Other                                            9
```

For our analysis, I wanted to focus on the amusement park industry. The first thing I did was filter these data to only include incidents from either an *Amusement park* or *Carnival or rental* when the industry sector was listed as *amusement ride*.

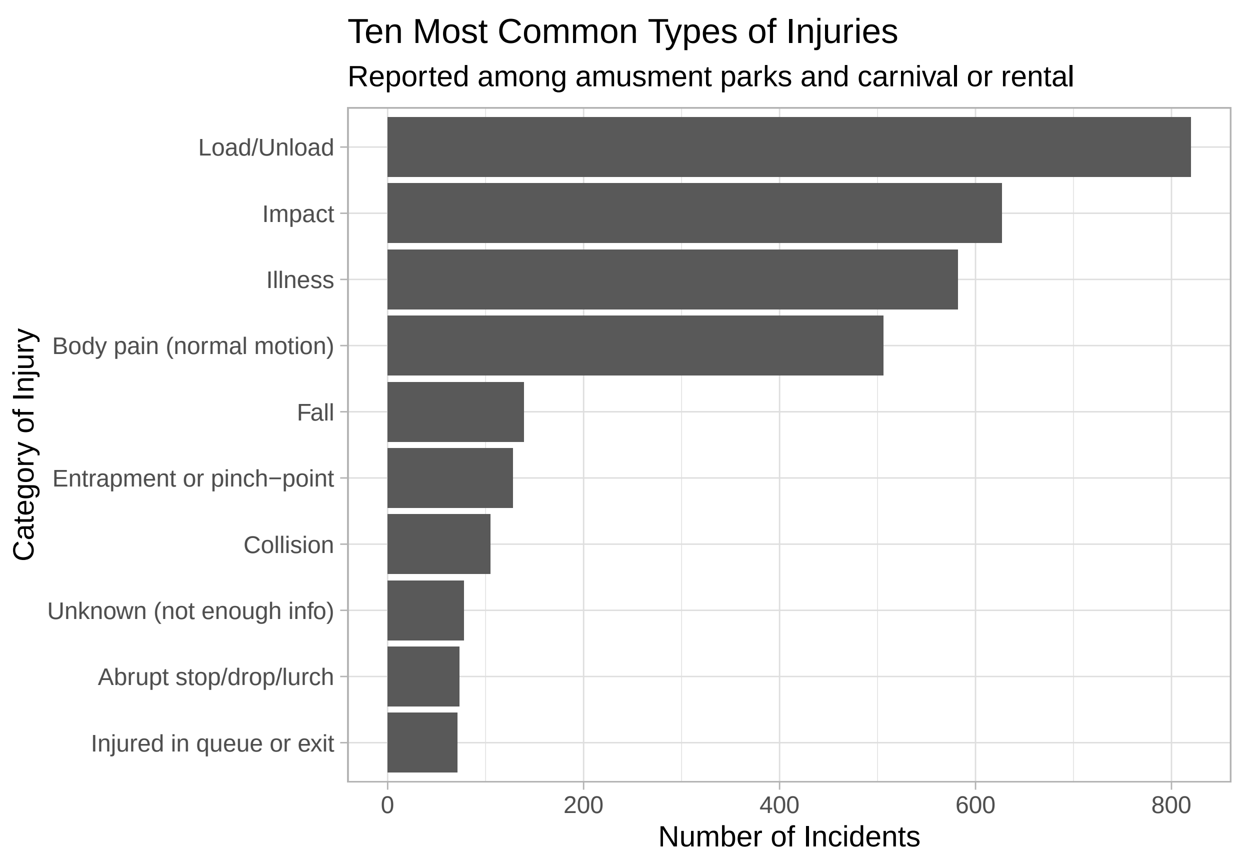
```r
dat <- safer_parks %>%
  filter(bus_type %in% c("Amusement park", "Carnival or rental") &
                             industry_sector == "amusement ride") %>%
  mutate(acc_date = mdy(acc_date),
         month = factor(month(acc_date, label = TRUE, abbr = TRUE)),
         manufacturer = ifelse(is.na(manufacturer), "Unknown", manufacturer),
         category2 = ifelse(str_detect(category, ":"),
                            str_extract(category, "^([^:])+"), category),
         category2 = ifelse(str_detect(category, "Illness"),
                            "Illness", category2))
```

The resulting dataset contains 3438 incidents that we will further analyze.

Type of Injury

Let's first look at the 10 most common types of injuries:

```r
dat %>%
  count(category2) %>%
  mutate(category2 = fct_reorder(category2, n)) %>%
  top_n(n=10) %>%
  ggplot(aes(category2, n)) +
  geom_col() +
  coord_flip() +
  labs(x = "Category of Injury",
       y = "Number of Incidents",
       title = "Ten Most Common Types of Injuries",
       subtitle = "Reported among amusment parks and carnival or rental")
```



Injury According to State

Now let's look at the most common states where injuries happen:

```r
dat %>% count(acc_state, sort = TRUE)
```

```
## # A tibble: 33 x 2
##    acc_state     n
##    <chr>     <int>
##  1 CA         2067
##  2 PA          494
##  3 NJ          143
```

```
##   4 TX          141
##   5 OK          138
##   6 FL          130
##   7 IL           43
##   8 NH           42
##   9 MI           39
##  10 KY           38
## # ... with 23 more rows
```

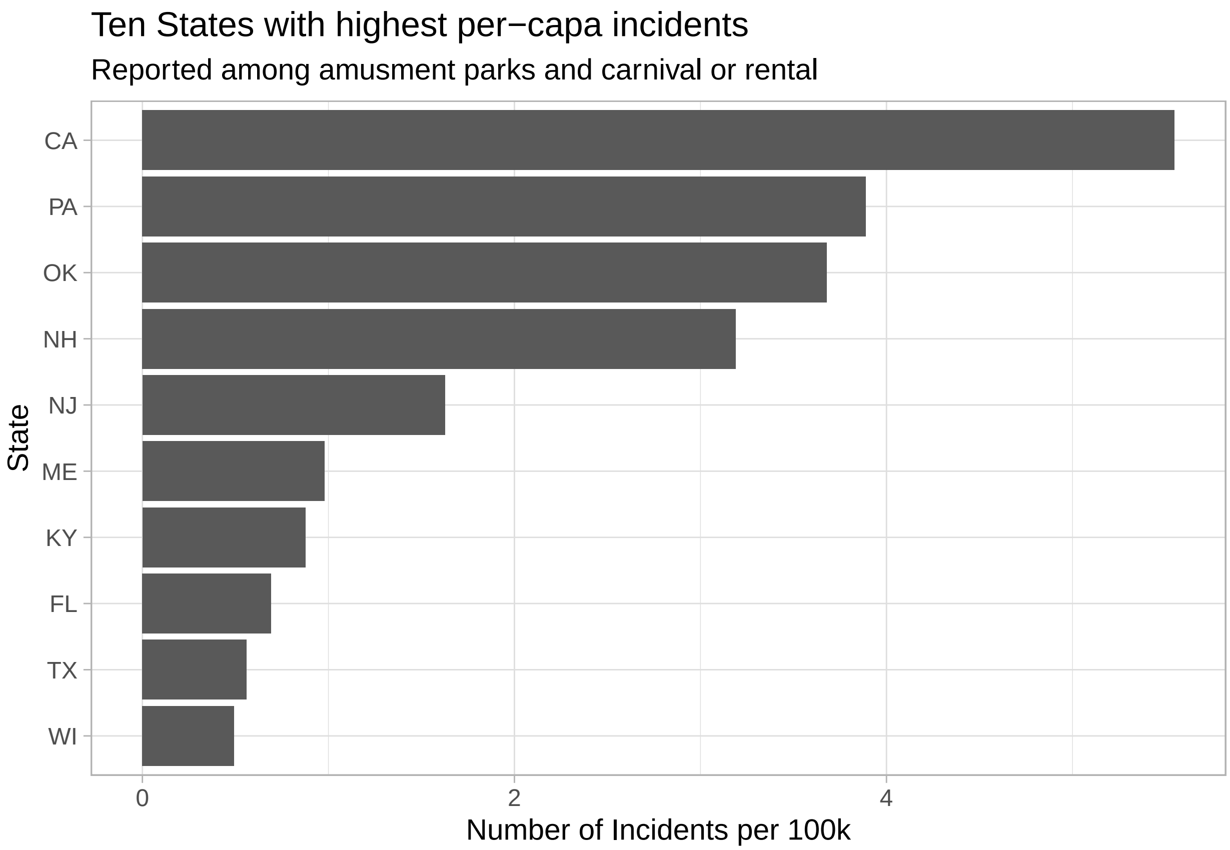
Although interesting, this may be misleading given the population differences between these states. For example, Pennsylvania is substantially smaller in population than California, yet it has almost 1/2 of the number of incidents. To better understand these data, we will pull in state population data from the 2010 decennial US Census and match state names with the built in state abbreviations.

```r
st_crosswalk <- tibble(state = state.name) %>%
  bind_cols(tibble(abb = state.abb))

state_pop <- as_tibble(tidycensus::get_decennial(geography = "state",
                           variables = "P001001") %>% select(state = NAME, pop = value)) %>%
  left_join(st_crosswalk, by = "state") %>% select(state = abb, pop)

state_dat <- left_join(dat%>% count(acc_state), state_pop, by = c("acc_state" = "state"))
```

```r
state_dat %>%
  mutate(incident_per_100k = ((n / pop)*100000),
         acc_state = fct_reorder(acc_state, incident_per_100k)) %>%
  top_n(n=10) %>%
  ggplot(aes(acc_state, incident_per_100k)) +
  geom_col() +
  coord_flip() +
  labs(x = "State",
       y = "Number of Incidents per 100k",
       title = "Ten States with highest per-capa incidents",
       subtitle = "Reported among amusment parks and carnival or rental")
```



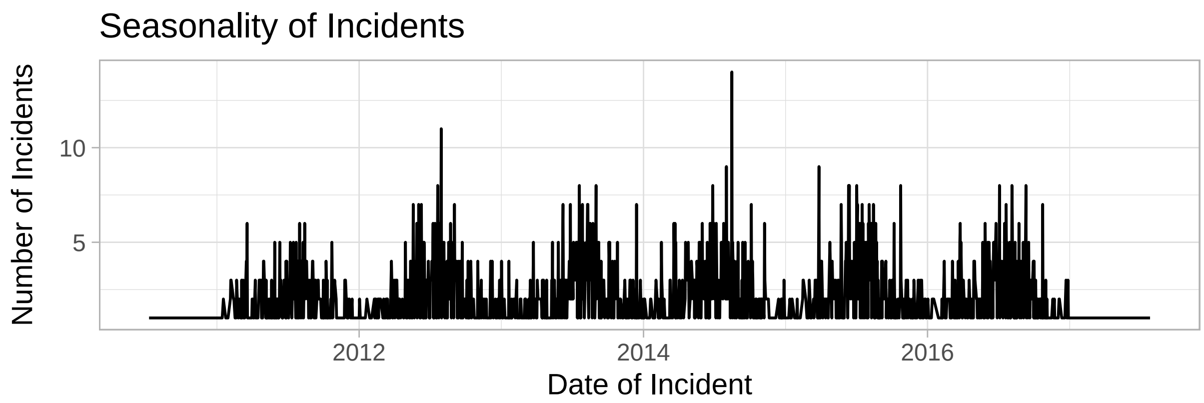
Injury According to Date

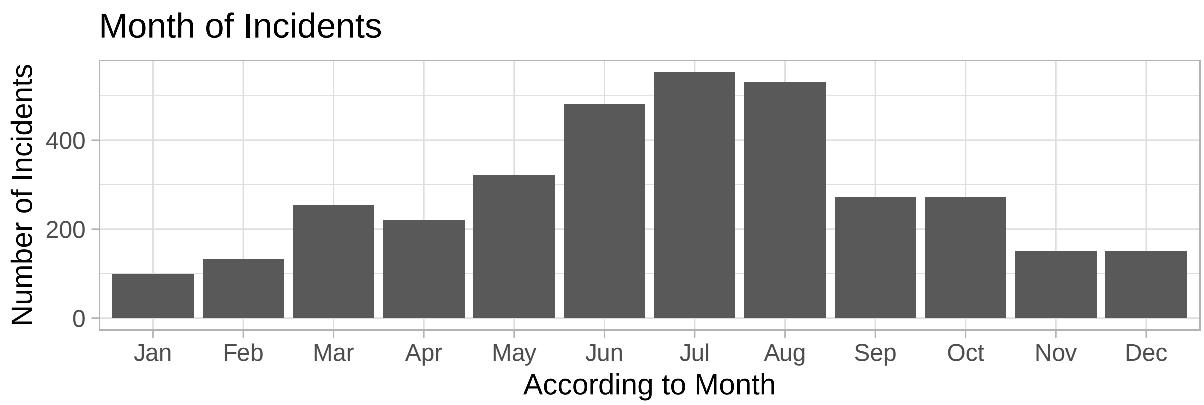
Next, we will look when the injuries tend to cluster. We would expect injuries to cluster around the summer time as it is the most likely time that people tend to go to amusement parks.

```
p1 <- dat %>%
  count(acc_date) %>%
  ggplot(aes(acc_date, n)) +
  geom_line() +
  labs(x = "Date of Incident",
       y = "Number of Incidents",
       title = "Seasonality of Incidents")

p2 <- dat %>%
  count(month) %>%
  ggplot(aes(month, n)) +
  geom_col() +
  labs(x = "According to Month",
       y = "Number of Incidents",
       title = "Month of Incidents")

(p1 / p2)
```





Injury by device category

```
dat %>% count(device_category, sort = TRUE)
```

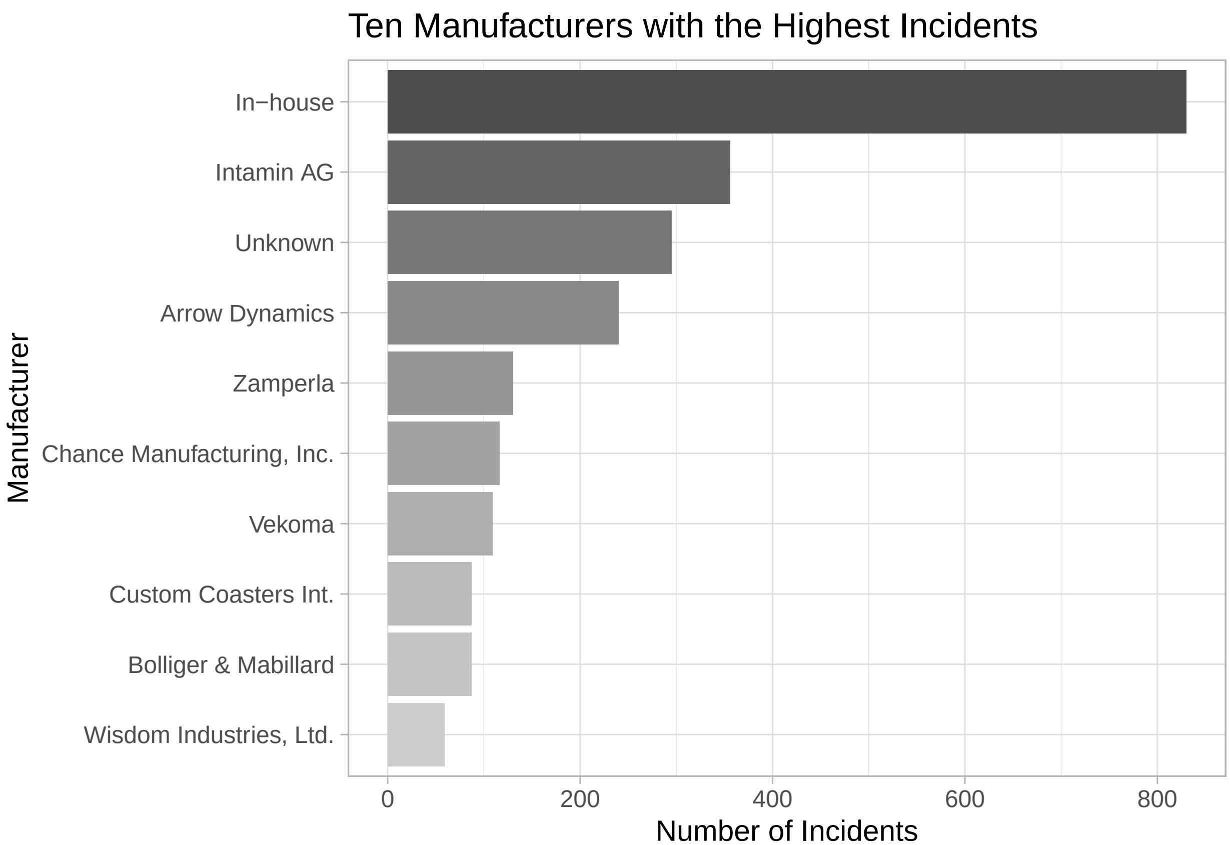
```
## # A tibble: 7 x 2
##   device_category         n
##   <chr>               <int>
## 1 coaster              1162
## 2 spinning              834
## 3 cars & track rides    554
## 4 water ride            507
## 5 other attraction      167
## 6 pendulum              122
## 7 vertical drop          92
```

Injury by manufacturer

Next we will look at the incidents by manufacturer.

```
dat %>%
  count(manufacturer, sort = TRUE) %>%
  mutate(manufacturer = fct_reorder(manufacturer, n)) %>%
  top_n(n = 10) %>%
  ggplot(aes(manufacturer, n, fill = manufacturer)) +
```

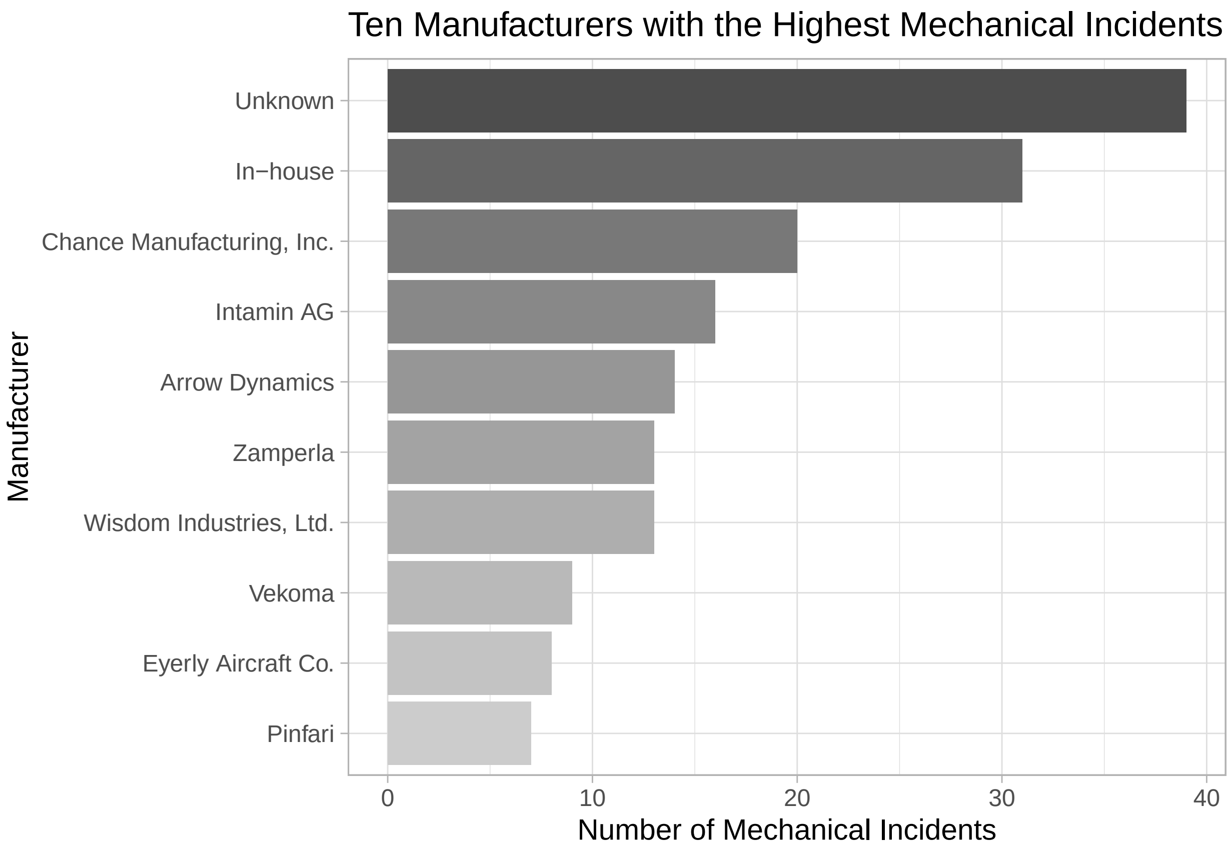
```r
  geom_col(show.legend = FALSE)+
  coord_flip() +
  scale_fill_grey(start = 0.8, end = 0.3) +
  labs(x = "Manufacturer",
       y = "Number of Incidents",
       title = "Ten Manufacturers with the Highest Incidents")
```



It appears that *in-house* made amusement rides have the highest reported number of incidents.

A Column also lists if an incident was related to a mechanical issue, let's look at the manufacturers with the highest reported mechanical issues.

```r
dat %>%
  filter(mechanical == 1) %>%
  count(manufacturer) %>%
  mutate(manufacturer = fct_reorder(manufacturer, n)) %>%
  top_n(n = 10) %>%
  ggplot(aes(manufacturer, n, fill = manufacturer)) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  scale_fill_grey(start = 0.8, end = 0.3) +
  labs(x = "Manufacturer",
       y = "Number of Mechanical Incidents",
       title = "Ten Manufacturers with the Highest Mechanical Incidents")
```



Injury Text Analysis

Types of Injury

Q: What types of injuries occur most often?

For text analysis, we will use the {tidytext} package. First, we need to isolate the injury text and *tokenize* it. This function splits each row so that one token (word) is in each row. Additionally, punctuation is removed and words are converted to lowercase. I have also removed numbers.

```
injury_tokens <- dat %>%
  select(acc_id, injury_desc) %>%
  unnest_tokens(word, injury_desc) %>%
  filter(is.na(as.numeric(word)))
```

Next, we will remove *stop words*, or words that are extremely common such as "the", "of", "to", etc. I have also removed the words *injury* and *pain* as they are not descriptive for our purposes and appear to be commonly used in the injury description.

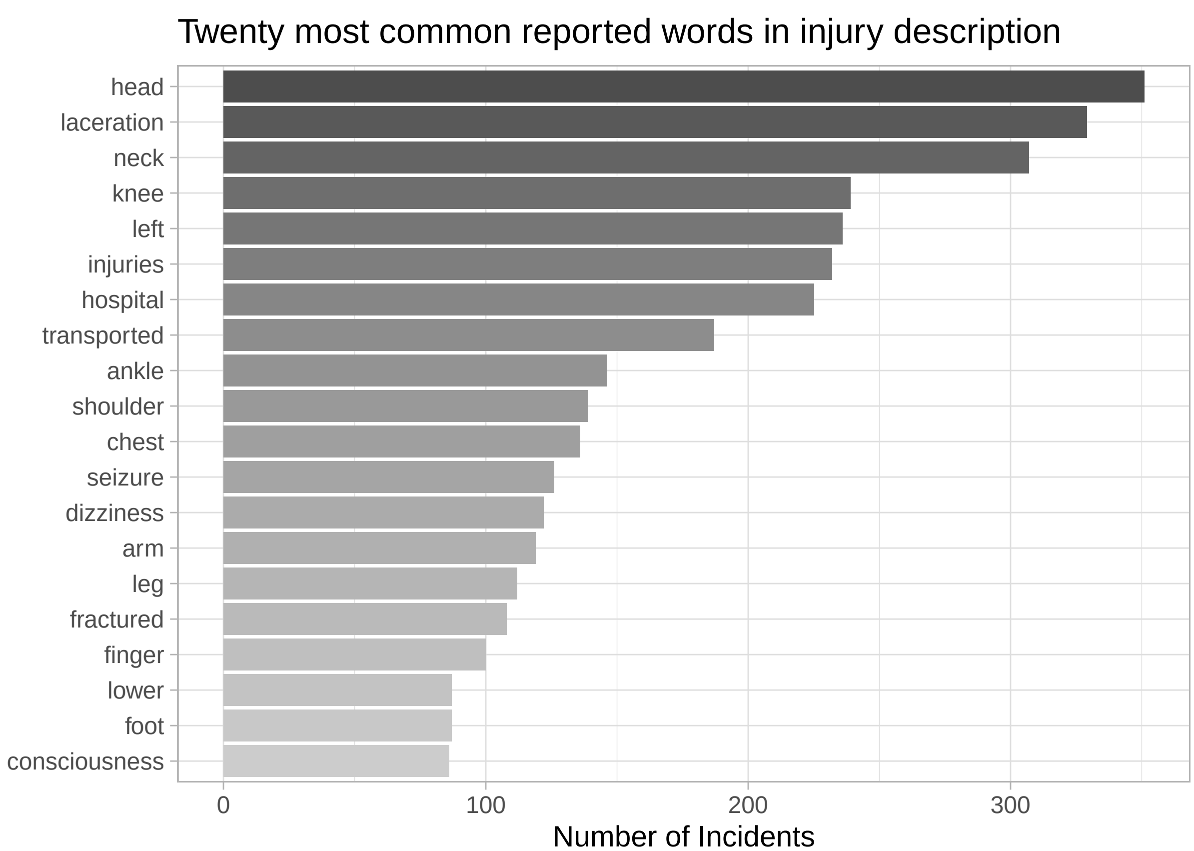
```
data(stop_words)
my_stop_words <- c("injury", "pain")

injury_tokens <- injury_tokens %>% anti_join(stop_words) %>% filter(!word %in% my_stop_words)
```

Barplot

We can use three visualizations to understand this a little easier. The first is a barplot showing the 20 most commonly used words when describing the injury.

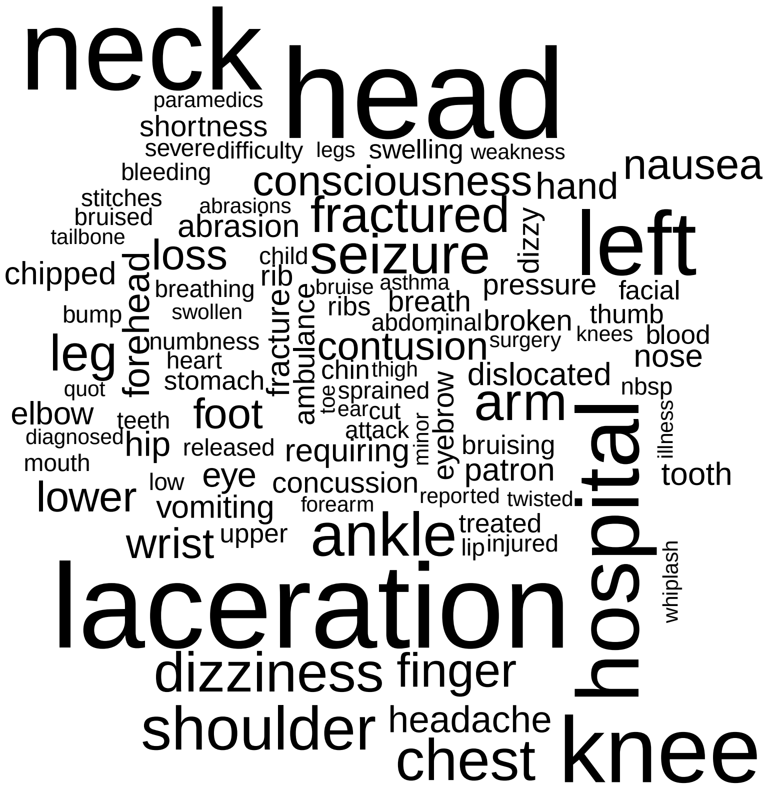
```
injury_tokens %>% count(word, sort = TRUE) %>%
  mutate(word = fct_reorder(word, n)) %>%
  slice(1:20) %>%
  ggplot(aes(x = word, y = n, fill = word)) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  scale_fill_grey(start = 0.8, end = 0.3) +
  labs(x = "", y = "Number of Incidents",
       title = "Twenty most common reported words in injury description")
```



Wordcloud

The second is a wordcloud showing the frequency that words appear in the injury description column.

```
injury_tokens %>% count(word) %>% with(wordcloud(word, n, max.words = 100))
```



Bi-gram graph

The third type of visualization is the relationships between words (n-grams). We are looking at how often words co-occur in these data.

```
injury_bigrams <- dat %>%
  unnest_tokens(bigram, injury_desc, token = "ngrams", n = 2)

injury_bigrams_sep <- injury_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

injury_bigrams_filtered <- injury_bigrams_sep %>%
  filter(!word1 %in% c(stop_words$word, my_stop_words),
         !word2 %in% c(stop_words$word, my_stop_words),
         !is.na(word1),
         is.na(as.numeric(word1)),
         is.na(as.numeric(word2)))

injury_bigrams_filtered %>% unite(bigram, word1, word2, sep = " ") %>% count(bigram, name = "number_of_i
```

```
## # A tibble: 1,604 x 2
##    bigram                number_of_incidents
##    <chr>                               <int>
##  1 chipped tooth                          35
##  2 left knee                              34
##  3 blood pressure                         26
##  4 fractured arm                          21
```

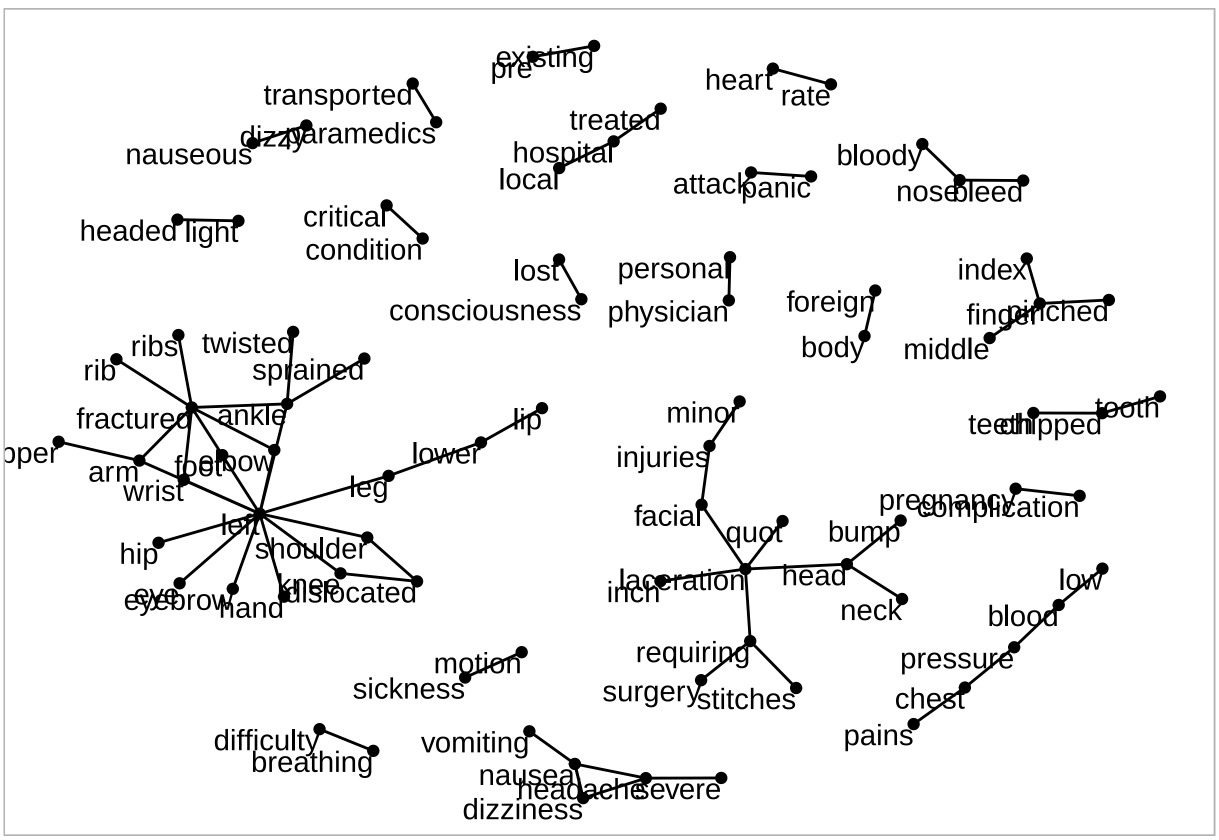
```
##  5 left shoulder                          20
##  6 difficulty breathing                   19
##  7 dislocated knee                        17
##  8 dislocated shoulder                    17
##  9 left arm                               16
## 10 requiring stitches                     16
## # ... with 1,594 more rows
```

We can visually look at how these words relate (cluster) together through the use of a bigram graph.

```r
bigram_graph <- injury_bigrams_filtered %>% count(word1, word2, sort = TRUE) %>%
  filter(n > 5)  %>%
  igraph::graph_from_data_frame()

set.seed(2020)

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1)
```



We see how these words tend to cluster together. If we had a specific research question, we could dig deeper into these.

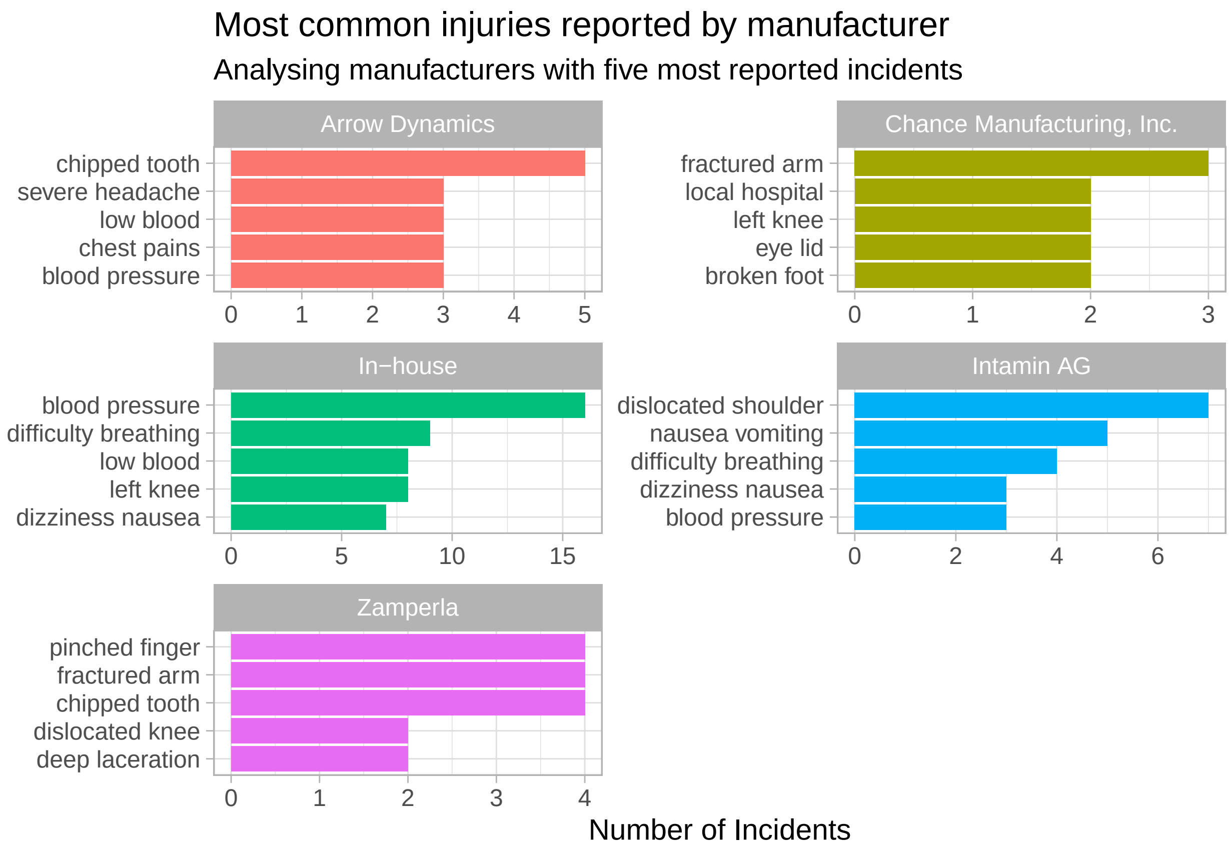
Types of Injury by Manufacturer

Q: What injuries are most often associated with what manufacturers? We can organize these data to look at the most common injuries reported by manufacturer as well.

```
injury_bigrams_unite <- injury_bigrams_filtered %>%
  unite(bigram, word1, word2, sep = " ")

manuf_of_interest <- injury_bigrams_unite %>%
  count(manufacturer, sort = TRUE) %>%
  filter(manufacturer != "Unknown") %>%
  slice(1:5)

injury_bigrams_unite %>%
  filter(manufacturer %in% manuf_of_interest$manufacturer) %>%
  count(manufacturer, bigram, sort = TRUE) %>%
  arrange(manufacturer, desc(n)) %>%
  group_by(manufacturer) %>%
  slice(1:5) %>%
  ungroup() %>%
  mutate(manufacturer = as.factor(manufacturer),
         bigram = reorder_within(bigram, n, manufacturer)) %>%
  ggplot(aes(bigram, n, order = -n, fill = manufacturer)) +
  geom_col(show.legend = FALSE) +
  scale_x_reordered() +
  labs(x = NULL, y = "Number of Incidents",
       title = "Most common injuries reported by manufacturer",
       subtitle = "Analysing manufacturers with five most reported incidents") +   facet_wrap(~manufactu
  coord_flip()
```



Types of Injury by Device Category

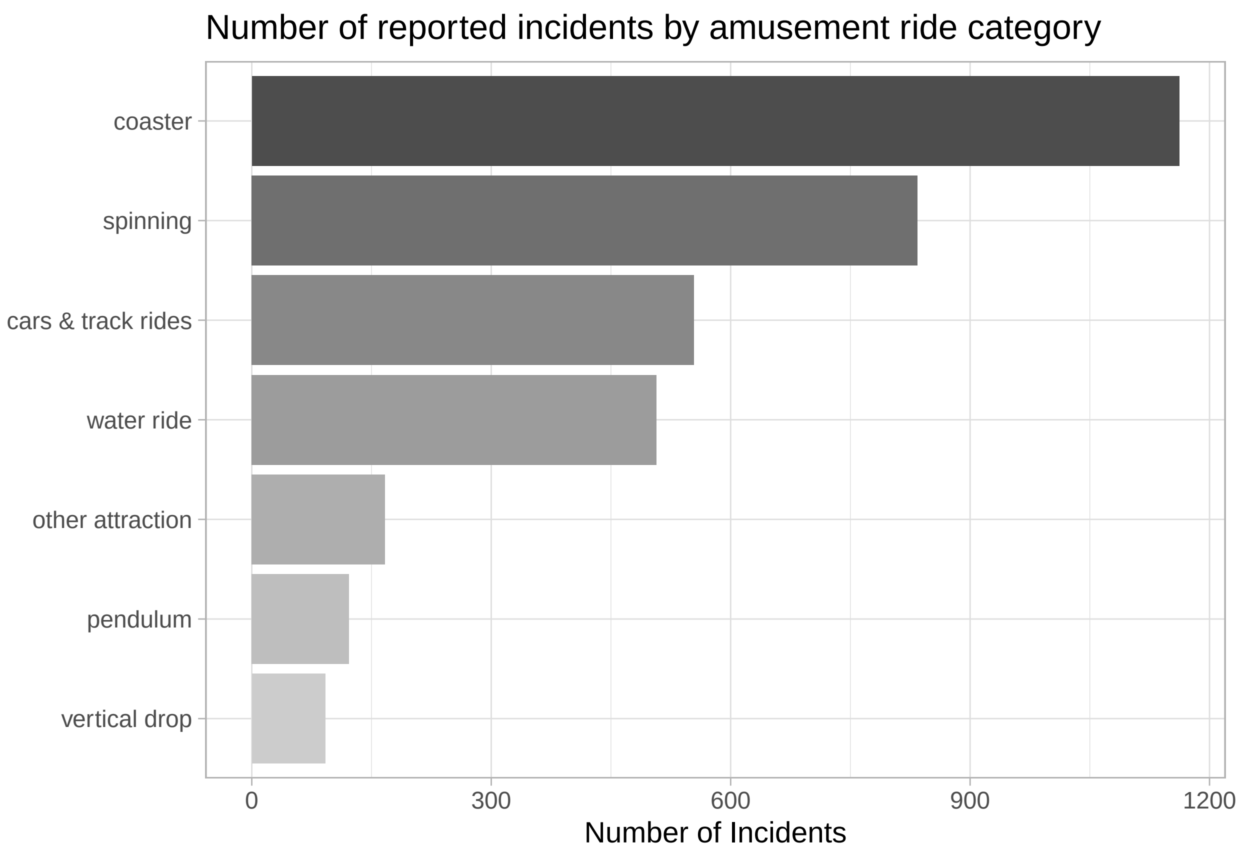
Q: Are certain types of rides more dangerous?

```
dat %>% count(device_type, name = "number_of_injuries", sort = TRUE)
```

```
## # A tibble: 52 x 2
##    device_type                 number_of_injuries
##    <chr>                                    <int>
##  1 Coaster - steel                            860
##  2 Track ride                                 302
##  3 Coaster - wooden                           208
##  4 Flume ride                                 170
##  5 Boat ride                                  156
##  6 Carousel                                   101
##  7 Rafting ride                                95
##  8 Coaster - family/kiddie                     92
##  9 Drop tower                                  90
## 10 Bumper car                                  87
## # ... with 42 more rows
```

```
dat %>%
  count(device_category) %>%
  mutate(device_category = fct_reorder(device_category, n)) %>%
  ggplot(aes(device_category, n, fill = device_category)) +
```

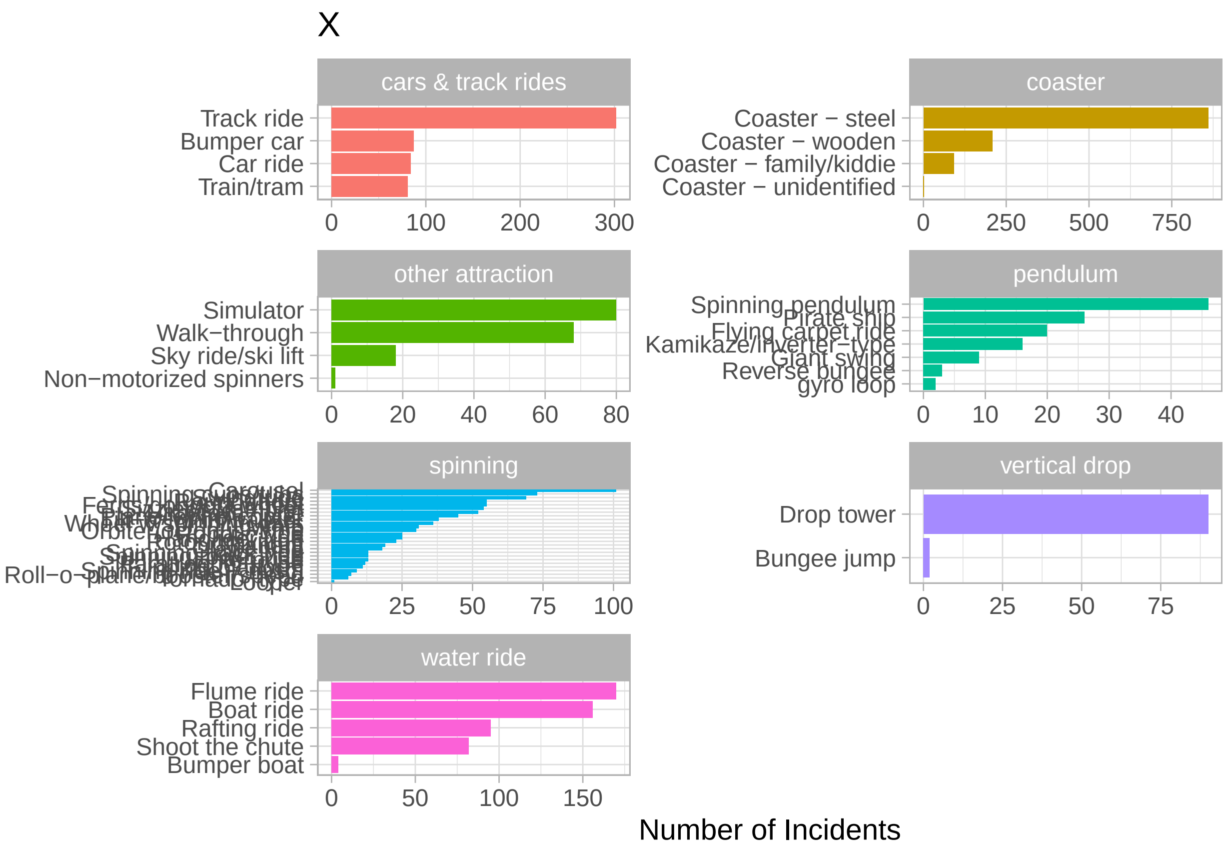
```r
  geom_col(show.legend = FALSE) +
  coord_flip() +
  scale_fill_grey(start = 0.8, end = 0.3) +
  labs(x = NULL, y = "Number of Incidents",
       title = "Number of reported incidents by amusement ride category")
```



Clearly, there are more incidents that involve *Coaster* rides. We could also expand the device categories to see what device types are creating the most number of incidents.

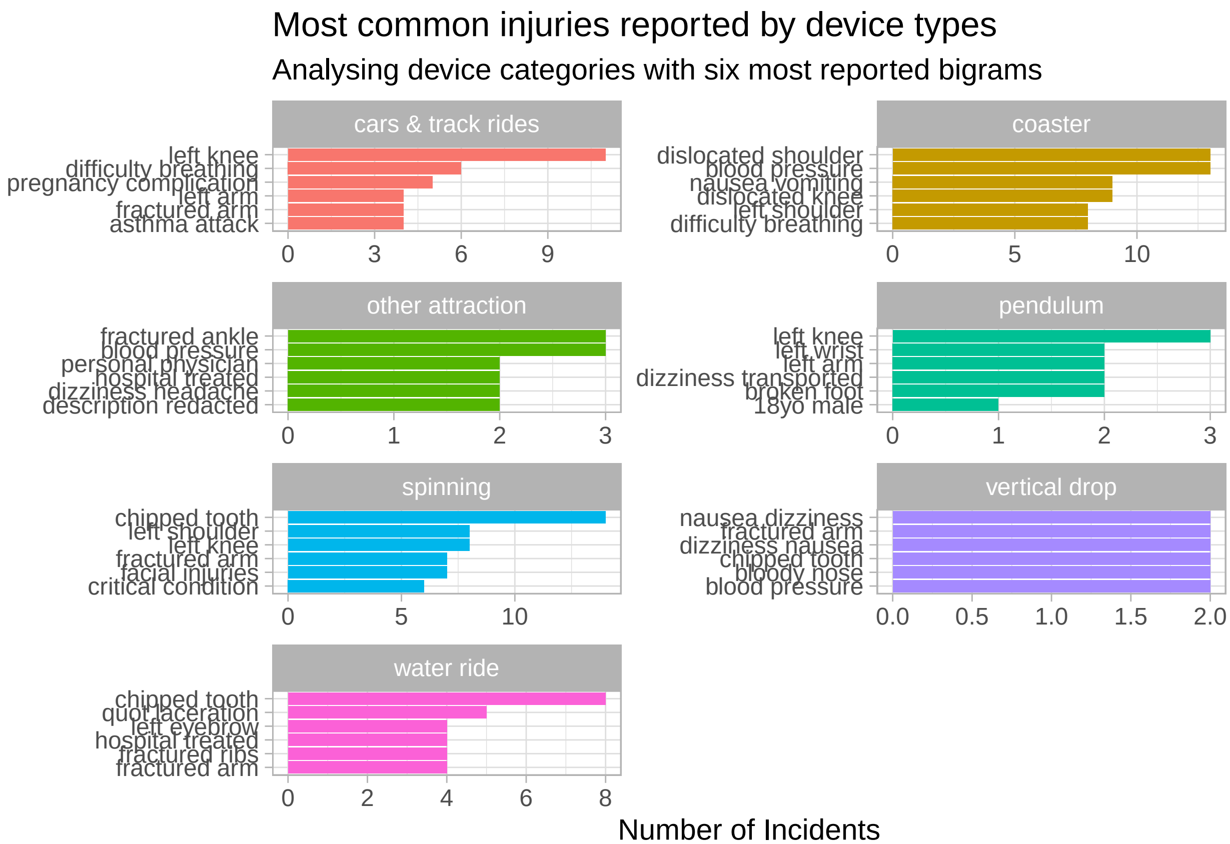
Types of Injury by Device Type

```r
dat %>%
  count(device_category, device_type, sort = TRUE) %>%
  arrange(device_category, desc(n)) %>%
  mutate(device_category = as.factor(device_category),
         device_type = reorder_within(device_type, n, device_category)) %>%
  ggplot(aes(device_type, n, order = -n, fill = device_category)) +
  geom_col(show.legend = FALSE) +
  scale_x_reordered() +
  labs(x = NULL, y = "Number of Incidents",
       title = "X") +
  facet_wrap(~device_category, ncol = 2, scales = "free") +
  coord_flip()
```



Injury Description by Device Category.

```r
injury_bigrams_unite %>%
  count(device_category, bigram, sort = TRUE) %>%
  arrange(device_category, desc(n)) %>%
  group_by(device_category) %>%
  slice(1:6) %>%
  ungroup() %>%
  mutate(device_category = as.factor(device_category),
         bigram = reorder_within(bigram, n, device_category)) %>%
  ggplot(aes(bigram, n, order = -n, fill = device_category)) +
  geom_col(show.legend = FALSE) +
  scale_x_reordered() +
  labs(x = NULL, y = "Number of Incidents",
       title = "Most common injuries reported by device types",
       subtitle = "Analysing device categories with six most reported bigrams") +
  facet_wrap(~device_category, ncol = 2, scales = "free") +
  coord_flip()
```



These make sense, it is interesting to see that the majority of the injuries occur on the left side of individuals. This could be due to the reporting practices (in fact, no 'right' side injuries were reported at all) or due to some other commonality of rides (perhaps they all swing in the same direction).

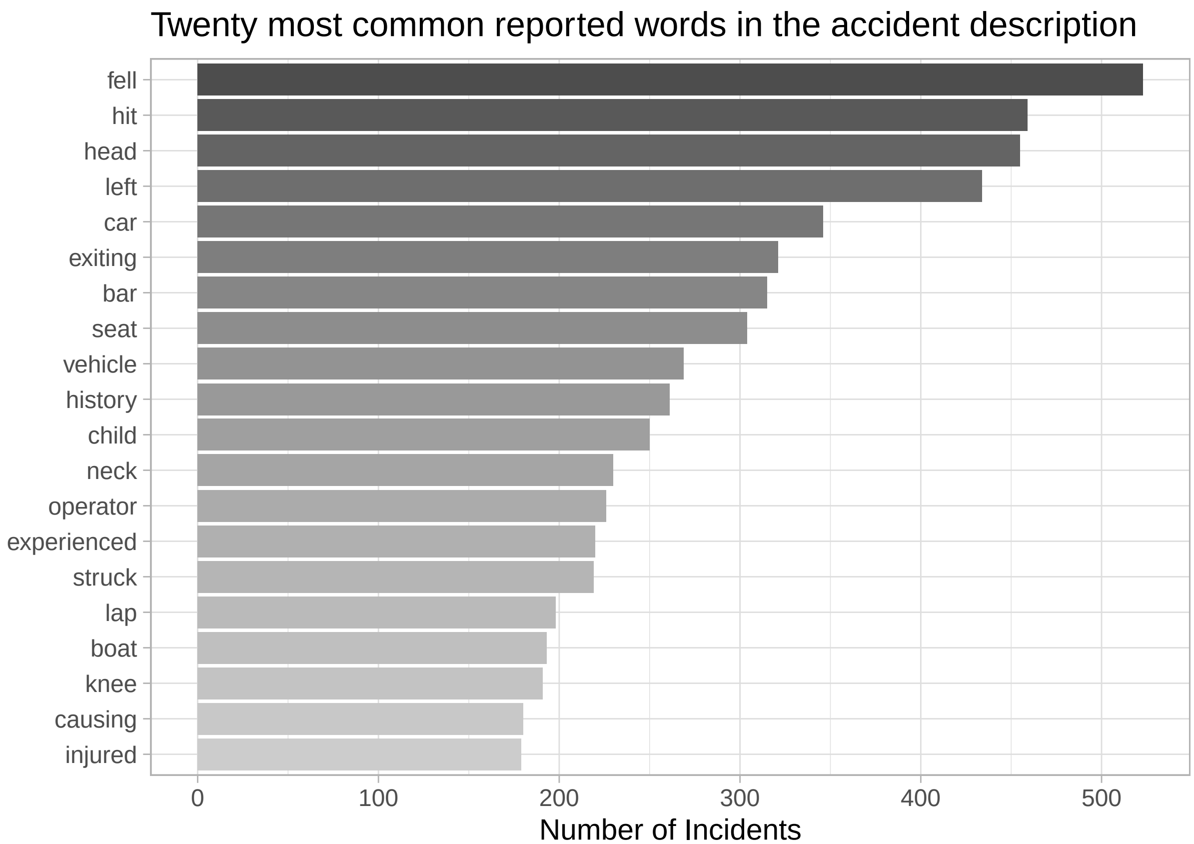
Accident Text Analyis

Q: What types of accidents occur most requently?

We will use the same text analysis process that we used previously to see if we can find patterns in the accident descriptions according to incident.

Barplot

```
acc_tokens %>% count(word, sort = TRUE) %>%
  mutate(word = fct_reorder(word, n)) %>%
  slice(1:20) %>%
  ggplot(aes(x = word, y = n, fill = word)) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  scale_fill_grey(start = 0.8, end = 0.3) +
  labs(x = "", y = "Number of Incidents",
       title = "Twenty most common reported words in the accident description")
```



Wordcloud

```
acc_tokens %>% count(word) %>% with(wordcloud(word, n, max.words = 100))
```

Bi-gram graph

```r
acc_bigrams <- dat %>%
  unnest_tokens(bigram, acc_desc, token = "ngrams", n = 2)

acc_bigrams_sep <- acc_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

acc_bigrams_filtered <- acc_bigrams_sep %>%
  filter(!word1 %in% c(stop_words$word, my_stop_words),
         !word2 %in% c(stop_words$word, my_stop_words),
         !is.na(word1),
         is.na(as.numeric(word1)),
         is.na(as.numeric(word2)))

acc_bigrams_filtered %>% unite(bigram, word1, word2, sep = " ") %>% count(bigram, name = "number_of_inc
```

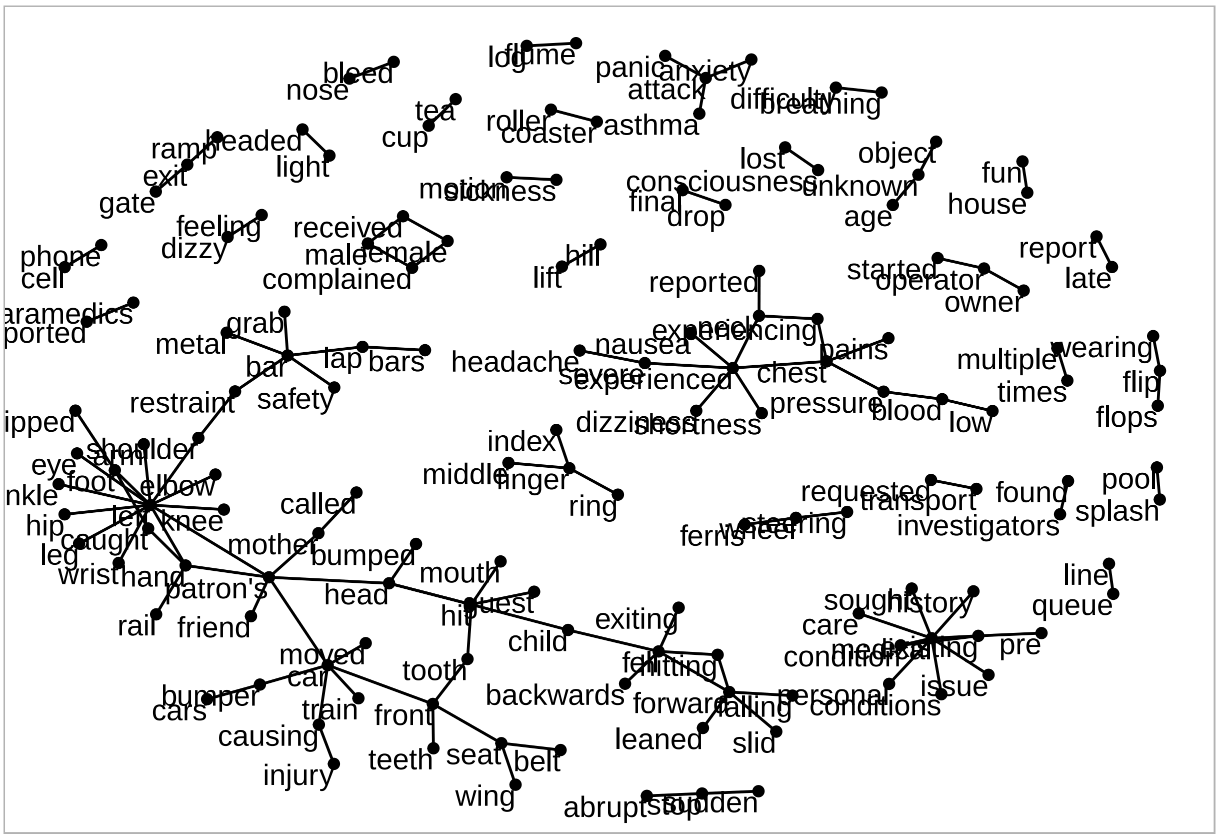
```
## # A tibble: 4,967 x 2
##    bigram           number_of_incidents
##    <chr>                          <int>
##  1 lap bar                          181
##  2 pre existing                      68
##  3 left knee                         60
##  4 blood pressure                    49
##  5 left shoulder                     35
##  6 existing medical                  33
```

```
##   7 medical condition              31
##   8 left arm                       29
##   9 left ankle                     27
## 10 left hand                       27
## # ... with 4,957 more rows
```

```r
acc_bigram_graph <- acc_bigrams_filtered %>% count(word1, word2, sort = TRUE) %>%
  filter(n > 5)  %>%
  igraph::graph_from_data_frame()

set.seed(2020)

ggraph(acc_bigram_graph, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1)
```



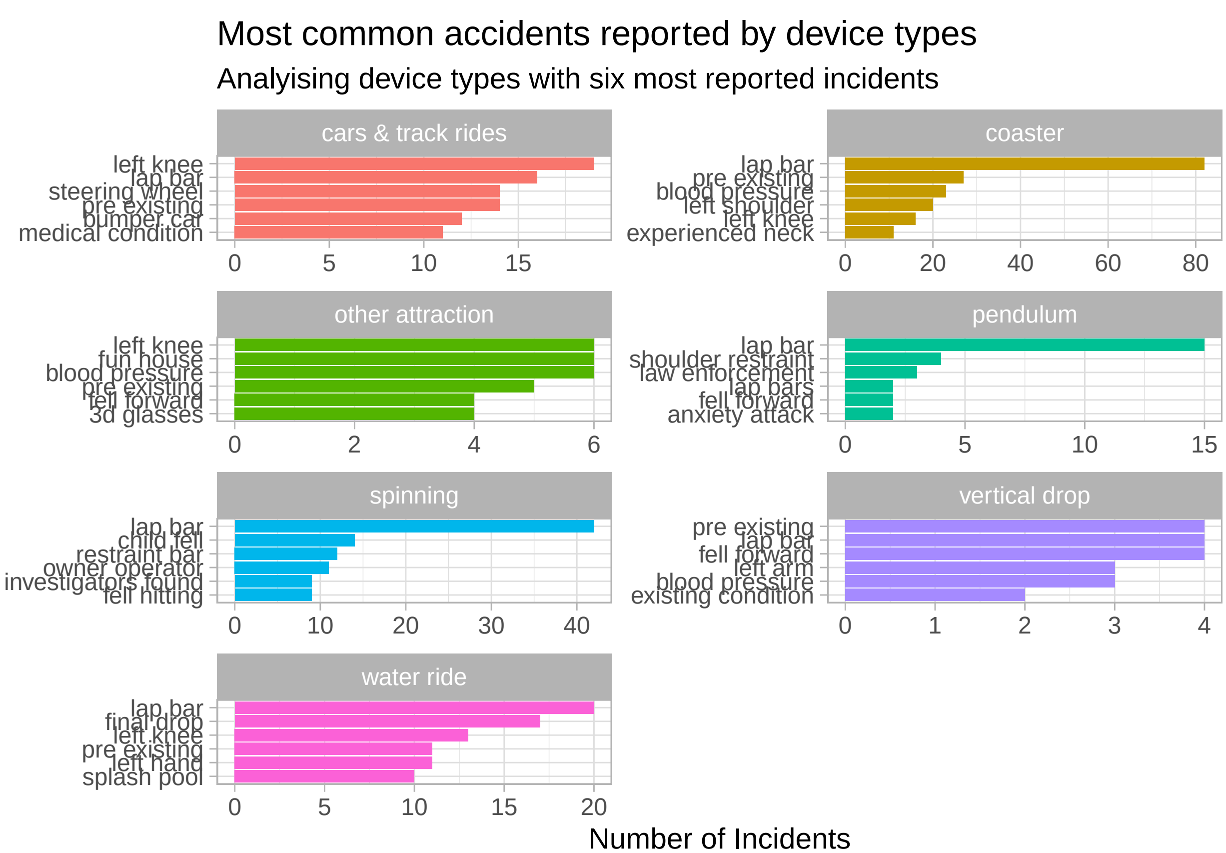
We see how these words tend to cluster together. If we had a specific research question, we could dig deeper into these.

Accident Description by Device Category.

```r
acc_bigrams_unite <- acc_bigrams_filtered %>%
  unite(bigram, word1, word2, sep = " ")
```

```r
manuf_of_interest <- acc_bigrams_unite %>%
  count(manufacturer, sort = TRUE) %>%
  filter(manufacturer != "Unknown") %>%
  slice(1:5)

acc_bigrams_unite %>%
  count(device_category, bigram, sort = TRUE) %>%
  arrange(device_category, desc(n)) %>%
  group_by(device_category) %>%
  slice(1:6) %>%
  ungroup() %>%
  mutate(device_category = as.factor(device_category),
         bigram = reorder_within(bigram, n, device_category)) %>%
  ggplot(aes(bigram, n, order = -n, fill = device_category)) +
  geom_col(show.legend = FALSE) +
  scale_x_reordered() +
  labs(x = NULL, y = "Number of Incidents",
       title = "Most common accidents reported by device types",
       subtitle = "Analyising device types with six most reported incidents") +
  facet_wrap(~device_category, ncol = 2, scales = "free") +
  coord_flip()
```



It appears that the lap bar tends to be the biggest cause of accidents among all of the device types (except *Other Attraction*).