Amusement Park

Introduction

This report extracts and analyse the data from tidytuesday 2019-09-10 exercise

Texas Injuries

Load packages

```
library(tidyverse)
library(lubridate)
```

Get the data from csv file

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

Data examination

```
head(tx_injuries)
```

```
## # A tibble: 6 x 13
##   injury_report_r~ name_of_operati~ city  st    injury_date ride_name
##              <dbl> <chr>            <chr> <chr> <chr>       <chr>
## 1             2032 Skygroup Invest~ Aust~ TX    2/12/2013   I Fly
## 2             1897 Willie G's Post~ Galv~ TX    3/2/2013    Gulf Gli~
## 3              837 Great Wolf Lodge Grap~ TX    3/3/2013    Howlin T~
## 4               99 Six Flags Fiest~ San ~ TX    3/3/2013    Scooby D~
## 5               55 Ray Cammack Sho~ Lave~ AZ    3/11/2013   Alien Ab~
## 6              780 ZDT's Amusement~ Segu~ TX    3/12/2013   Go Karts
## # ... with 7 more variables: serial_no <chr>, gender <chr>, age <chr>,
## #   body_part <chr>, alleged_injury <chr>, cause_of_injury <chr>,
## #   other <chr>
```

Check relative injuries to different body_part

```
head(unique(tx_injuries$body_part))
```

```
## [1] "Mouth"          "Knee"           "Right Shoulder" "Lower Leg"
## [5] "Head"           "Bottom of foot"
```

```
length(unique(tx_injuries$body_part))
```

```
## [1] 189
```

Check rides

```
head(unique(tx_injuries$ride_name))
```

```
## [1] "I Fly"                    "Gulf Glider"
## [3] "Howlin Tornado"           "Scooby Doo Ghost Blasters"
## [5] "Alien Abduction"          "Go Karts"
```

```
length(unique(tx_injuries$ride_name))
```

```
## [1] 252
```

I find that the columns body_part and ride_name do not have any generic discrete variables. It would be difficult to convert these columns into desired categorical variables.The column injuries by date seems to be an ideal candidate for analysis. In next steps, I select and rename required columns and plot injuries by date.

```
tx_injuries_selected <- tx_injuries %>%
     select(park = name_of_operation,
            city,
            state = st,
            ride = ride_name,
            body_part,
            injury_type = alleged_injury,
            date = injury_date)
head(tx_injuries_selected)
```

```
## # A tibble: 6 x 7
##   park          city    state ride       body_part  injury_type       date
##   <chr>         <chr>   <chr> <chr>      <chr>      <chr>             <chr>
## 1 Skygroup Inv~ Austin  TX    I Fly      Mouth      Student hit mout~ 2/12~
## 2 Willie G's P~ Galves~ TX    Gulf Glid~ Knee       Alleged arthrosc~ 3/2/~
## 3 Great Wolf L~ Grapev~ TX    Howlin To~ Right Sho~ Pain in shoulder  3/3/~
## 4 Six Flags Fi~ San An~ TX    Scooby Do~ Lower Leg  Contusion         3/3/~
## 5 Ray Cammack ~ Laveen  AZ    Alien Abd~ Head       Laceration        3/11~
## 6 ZDT's Amusem~ Seguin  TX    Go Karts   Bottom of~ cut requiring st~ 3/12~
```

Wrangle the date column

```
dates_formatted = mdy(tx_injuries_selected$date)
```

```
## Warning: 349 failed to parse.
```

```
dates_base_1899 = as.Date(as.numeric(tx_injuries_selected$date), origin = "1899-12-30")
```

```
## Warning in as.Date(as.numeric(tx_injuries_selected$date), origin =
## "1899-12-30"): NAs introduced by coercion
```

```
date_wrangled = if_else(is.na(dates_formatted), dates_base_1899, dates_formatted)
tx_injuries_wrangled <- tx_injuries_selected %>% mutate(date = date_wrangled)
head(tx_injuries_wrangled)
```

```
## # A tibble: 6 x 7
##   park         city    state ride     body_part injury_type      date
##   <chr>        <chr>   <chr> <chr>    <chr>     <chr>            <date>
## 1 Skygroup In~ Austin  TX    I Fly    Mouth     Student hit mou~ 2013-02-12
## 2 Willie G's ~ Galves~ TX    Gulf Gl~ Knee      Alleged arthros~ 2013-03-02
## 3 Great Wolf ~ Grapev~ TX    Howlin ~ Right Sh~ Pain in shoulder 2013-03-03
## 4 Six Flags F~ San An~ TX    Scooby ~ Lower Leg Contusion        2013-03-03
## 5 Ray Cammack~ Laveen  AZ    Alien A~ Head      Laceration       2013-03-11
## 6 ZDT's Amuse~ Seguin  TX    Go Karts Bottom o~ cut requiring s~ 2013-03-12
```

Find injuries by date

```
injuries_vs_date <- tx_injuries_wrangled %>% count(date)
head(injuries_vs_date)
```

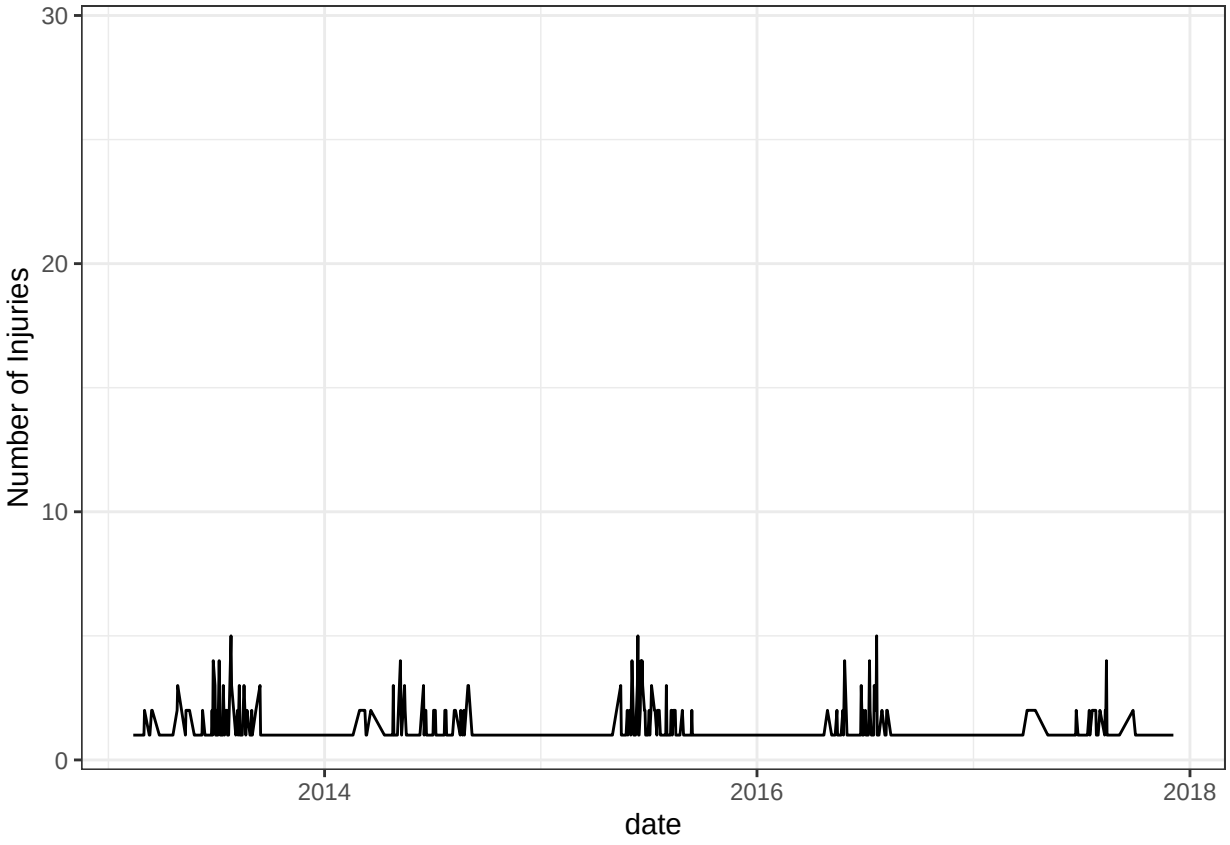
```
## # A tibble: 6 x 2
##   date           n
##   <date>     <int>
## 1 2013-02-12     1
## 2 2013-03-02     1
```

```
## 3 2013-03-03     2
## 4 2013-03-11     1
## 5 2013-03-12     1
## 6 2013-03-15     2
```

Plot injuries by date

```
injuries_vs_date %>% ggplot(aes(x = date, y = n)) +
      geom_line() +
      theme_bw() +
      labs(y = "Number of Injuries") +
      guides(color = "none")
```

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



It can be observed that during the middle of every year (late spring and summer), there is a spike in the number of injuries. # Safer parks The data set is required to be tranformed into a tidy format. The package dplyr is required for sorting, filtering and summarizing data in this set.

```
library("dplyr")
library("tidytext")
safer_parks <-
  readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-
head(safer_parks)
```

```
## # A tibble: 6 x 23
##   acc_id acc_date acc_state acc_city fix_port source bus_type
```

```
##    <dbl> <chr>    <chr>     <chr>    <chr>    <chr>  <chr>
## 1 1.01e6 6/12/20~ OH        Clevela~ F        Ohio ~ Sports ~
## 2 1.00e6 6/12/20~ OH        Clevela~ P        Unite~ Sports ~
## 3 1.01e6 7/10/20~ CA        Anaheim  F        Calif~ Amuseme~
## 4 1.01e6 7/10/20~ CA        Carlsbad F        Calif~ Water p~
## 5 1.00e6 7/29/20~ CO        Littlet~ F        Color~ Family ~
## 6 1.01e6 7/30/20~ WI        Wiscons~ F        Wisco~ Amuseme~
## # ... with 16 more variables: industry_sector <chr>,
## #   device_category <chr>, device_type <chr>, tradename_or_generic <chr>,
## #   manufacturer <chr>, num_injured <dbl>, age_youngest <dbl>,
## #   gender <chr>, acc_desc <chr>, injury_desc <chr>, report <chr>,
## #   category <chr>, mechanical <dbl>, op_error <dbl>, employee <dbl>,
## #   notes <chr>
```

I need to restructure injury description as one token per row format, removing rows with stopwords. Finally I filter them into six possible body part buckets.

```
body_parts <- tolower(c("HEAD", "BACK", "EAR", "HIP", "ARM", "LEG"))
body_parts_freq <- safer_parks %>%
  unnest_tokens(output=word, input=injury_desc) %>% #
  anti_join(get_stopwords()) %>% # Remove stopwords
  filter(word %in% body_parts) %>% # Filter to rows with body part
  mutate(word = toupper(word)) %>% # Convert all string variables in word to upper case
  distinct(word, .keep_all = TRUE) %>% # Remove duplicate rows based on word
  mutate(word=as.factor(word)) %>% # Creating a new word datacolumn and categorizing the data into leve
  group_by(word) %>% # Converts word to a grouped table
  summarise(total = sum(num_injured))
```

```
## Joining, by = "word"
```

```
head(body_parts_freq)
```

```
## # A tibble: 6 x 2
##   word  total
##   <fct> <dbl>
## 1 ARM       1
## 2 BACK      8
## 3 EAR       8
## 4 HEAD      8
## 5 HIP       1
## 6 LEG       1
```

It can be observed that head and back injuries are proportianately higher compared to other body part injuries.