Footballer Death Analysis

Owais Khan

11/30/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.4     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

###The goal is to analyse the data set which contains the death of footballers as per date and the type of incident which because the cause of their death.

```
football_deaths <- read.csv("~/STA 518/R-for-data-science/data/football_deaths.csv")
```

```
#data cleaning. Omitting all NA values:
football_deaths <- na.omit(football_deaths)
```

I can import data from a variety of sources.

I can restructure information to be in a "tidy" format. Using 5 Boolean column which represents the type of incident and replacing them with a single Incident type column.

```
data_by_incident <- football_deaths %>%
gather(Incident_Type,j,-row_id)%>%
filter(j==1)%>%
select(-j)
```

I can implement sampling methods to make conclusions about data Sampling out 5% random records of data_by_incident to help getting an idea of the data set's frequencies.

```
sample_5 <- sample_frac(data_by_incident,0.05)
print(sample_5)
```

```
##    row_id   Incident_Type
## 1      74       collision
## 2     143       collapsed
## 3      96       collapsed
## 4     125       collapsed
## 5      88   heart_related
## 6      89   heart_related
## 7     124       collapsed
## 8      81       collision
## 9     131 cardiac_related
## 10    177 cardiac_related
## 11     52   heart_related
```

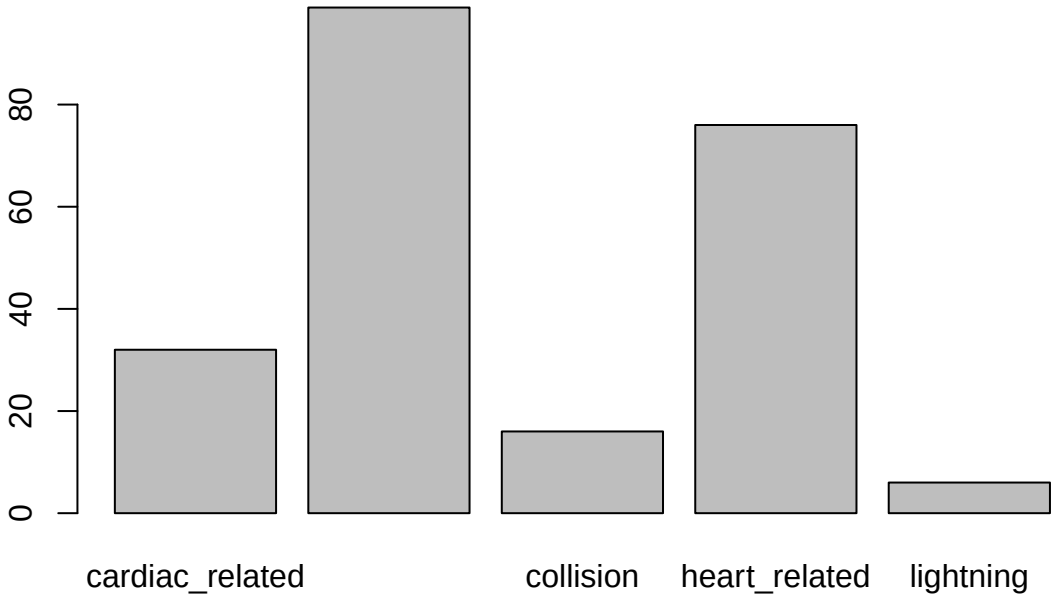
I can combine information from multiple data sources. Combining the two data sets (actual football_deaths & data_by_incident) and replacing the 5 Boolean columns with the single incident_type column.

```
combined <- merge(x=football_deaths,y=data_by_incident,by="row_id")

combined <- combined %>% select(-heart_related, -cardiac_related, -collapsed, -lightning, collision)
```

I can create graphical displays of data that highlight key features. Drawing a bar plots to analyze the frequency of the occurrence of the type of incident which becomes the cause of footballer's death.

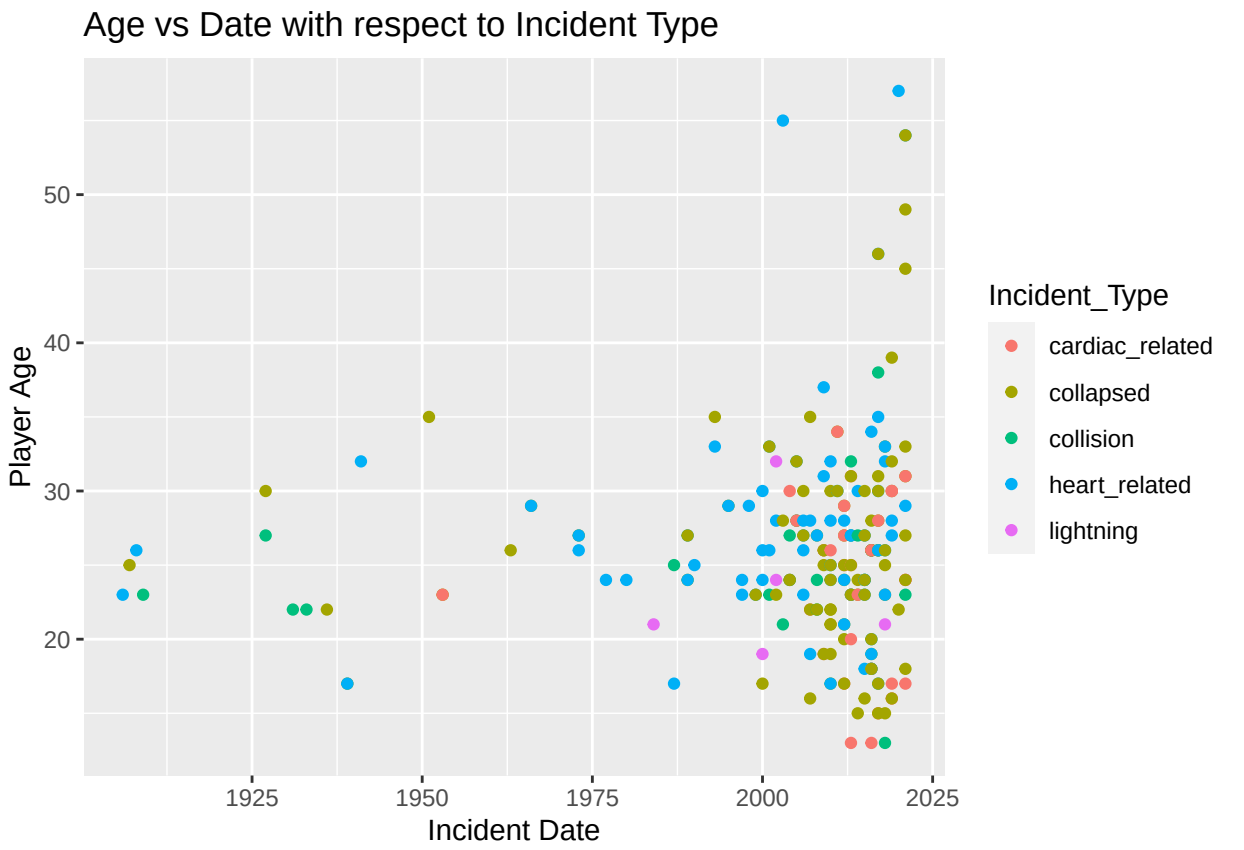
```
x <- table(combined$Incident_Type)
barplot(x)
```



From the above analysis, it can be clearly seen that most deaths are caused by the *collapsing* of the football players while *lightnining* shows the lowest number of occurrences.

Creating a plot which helps analyzing the death not just based on the incident, but also the age of the player. The incident date is also kept in order to understand in which duration, most of the death took place.

```
combined %>%  ggplot(aes(year(incident_date), player_age, group = Incident_Type, color = Incident_Type))
  geom_point() +
  xlab("Incident Date") +
  ylab("Player Age") +
  ggtitle(label = waiver()) +
  labs(title = "Age vs Date with respect to Incident Type")
```



Above plot helps us understand that most of the death caused (regardless of the incident type) are between 2000 and 2020 among which collapsed shows the highest rate. Although lightening seems to be very rare but there were no lightening deaths reported before the year 1975.