Report

Bryana Gutierrez, Lily Li, Erica Wong

December 8, 2015

Cleaning Data

This script starts off with two raw data files and we first worked on cleaning each individually by extracting the data we wanted and then concatenated both data sets based on the date of the accidents.

First we download the raw data onto R studio:

```
av_data_1 <- read.delim('../rawdata/aviation_data.txt', header = TRUE,
                        sep = '|' , stringsAsFactors = FALSE)
air_data_1 <- read.csv('../rawdata/airplane_crashes.csv',
                       stringsAsFactors = FALSE)
```

We will use a numbering convention so that at each modification the index increases by one. This takes only the airplane data from the data frame as opposed to all aircrafts.

```
av_data_2 <- subset(av_data_1, Aircraft.Category == ' Airplane '|
                      Aircraft.Category == '  ')
air_data_2 <- air_data_1[!grepl('airship', air_data_1$Type), ]
```

Next, we merge these two date tables so that air_data provides information from 1920-1981 and av_data information from 1982-2015. So now we extract only the years we want from each data frame.

```
acc_date <- av_data_2$Event.Date
av_data_2$year <- as.numeric(substr(acc_date, start = nchar(acc_date)-4,
                                    stop = nchar(acc_date)))
av_data_3 <- subset(av_data_2, year >= 1982)

ac_date <- air_data_2$Date
air_data_2$year <- as.numeric(substr(ac_date, start = nchar(ac_date)-3,
                                     stop = nchar(ac_date)))
air_data_3 <- subset(air_data_2, year <= 1981 & year >= 1920)
```

In av_data the columns 'Location' and 'Country' should be combined as well as 'Make' and 'Model'

```
av_data_3$new_location <- paste0(av_data_3$Location, av_data_3$Country)
av_data_3$Type <- paste0(av_data_3$Make, av_data_3$Model)
```

Now, we work on reorganizing the data so that it is in the format that we want. In air_data the order of data is currently in ascending order, but to match av_data we want it to be in descending order. We also extract the columns that we want and rename then so that the two data frames have the same column names. We do this, so that once we extract missing values we can concatenate the two data frames. Finally we make the type and location names uniform by changing all characters to lower case.

```
air_data_4 <- air_data_3[nrow(air_data_3):1,]

# Now to extract the information (columns) that we want
av_data_4 <- av_data_3[c('Event.Date', 'new_location', 'Type')]
air_data_5 <- air_data_4[c('Date', 'Location', 'Type')]

# We don't want the crash information for crashes missing 'Type' information.
av_data_5 <- subset(av_data_4, Type != '    ')
air_data_6 <- subset(air_data_5, Type != '')

# Rename column names so that they are the same for both data frames.
names(av_data_5) <- c('date', 'location', 'type')
names(air_data_6) <- c('date', 'location', 'type')

# Now we combine the two data frames.
data <- rbind(av_data_5, air_data_6)

# Changing the 'type' column to all lowercase
data$type <- tolower(data$type)

# Changing the 'location' column to all lowercase
data$location <- tolower(data$location)
```

This is our clean data

```
head(data)
```

```
##           date                          location
## 1  11/06/2015       gonzales, la  united states
## 2  11/04/2015   mount pocono, pa  united states
## 3  11/04/2015        wakeman, oh  united states
## 4  11/04/2015         haines, ak  united states
## 6  11/03/2015       tecumseh, mi  united states
## 7  11/03/2015   fayetteville, ar  united states
##                                type
## 1                      cessna  120
## 2               schweizer  269c 1
## 3  american autogyro  sparrowhawk
## 4                      cessna  180
## 6                piper  pa 18-150
## 7       cirrus design corp  sr22t
```

Exporting the `data` file into the data folder.

```
write.csv(data, file = '../data/data.csv')
```

Part 1: Monthly Data

In this part of the project, we were looking to see which of the months had the most plane crashes. In order to do this, we will need to first load the necessary code and packages.

```r
# packages needed
library(stringr)

# function needed
source('../code/num_name.R')
```

Our cleaned data set does not have a column that just contains the months when each accident happens, so we will had to make our own by doing the following.

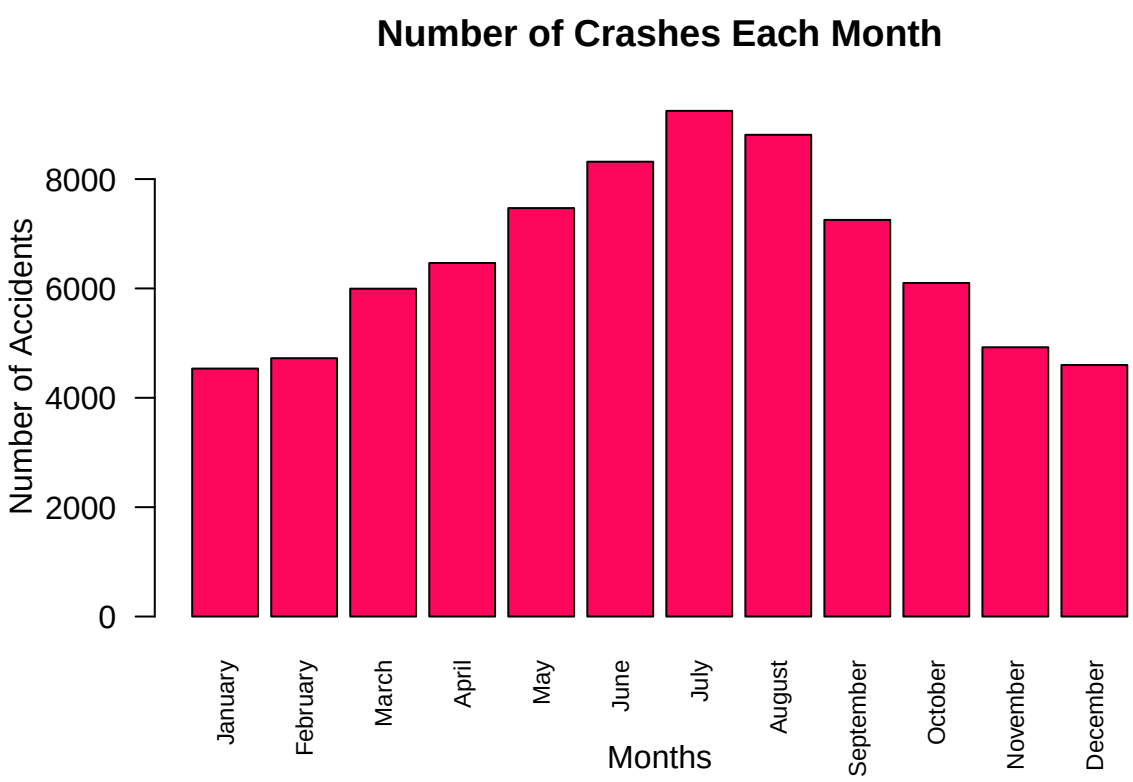
```r
# importing data
data <- read.csv('../data/data.csv', stringsAsFactors = FALSE)

# removing first column due to repeated index
data$X <- NULL

# extracting months
data$month <- as.numeric(str_extract(data$date, '[0-9]+'))

# labeling months with words using the num_name function.
dates <- data$month
month <- num_name(dates)

# seeing the new data frame
head(data)
```

```
##           date                          location
## 1  11/06/2015       gonzales, la  united states
## 2  11/04/2015   mount pocono, pa  united states
## 3  11/04/2015        wakeman, oh  united states
## 4  11/04/2015         haines, ak  united states
## 5  11/03/2015       tecumseh, mi  united states
## 6  11/03/2015   fayetteville, ar  united states
##                               type month
## 1                      cessna  120     11
## 2                schweizer  269c 1     11
## 3  american autogyro  sparrowhawk     11
## 4                      cessna  180     11
## 5                 piper  pa 18-150     11
## 6       cirrus design corp  sr22t     11
```

With this column, we are now ready to graph to see our findings.

```r
# Number of Entries Each Month
month_names <- c('January', 'February', 'March', 'April', 'May', 'June',
                 'July', 'August', 'September', 'October', 'November',
                 'December')

monthly_data = NULL
for (name in month_names){
  total_per_month <- sum(month == name)
  monthly_data <- c(monthly_data, total_per_month)
}
```

```r
# Graphing the months
barplot(monthly_data, xlab = 'Months',
        ylab = 'Number of Accidents',
        main = 'Number of Crashes Each Month',
        names.arg = month_names,
        las = 2,
        cex.names = .75,
        col = '#FD075E')
```



Our original belief was that the number of accidents every month would be relatively uniform, but instead, we found that there are more accidents that occur during the middle of the year, with the most occurring in July. We found this idea to be interesting and surprising. The number of accidents could have increased due to there being more traveling being done during the summer due to kids being on summer vacation. However, during the summer seasons, there are also a lot of tropical storms , thunderstorms, and changes in the weather that could affect how safe it is to travel and increases the chance of getting into a plane accident. According to a New York Times article, the number of airplane accidents could be greater in the summer months because that is typically when people fly their private airplanes. A smaller, private aircraft is more likely to get into an accident than a commercial plane. The article states that, "Statistics from the N.T.S.B. show that general aviation aircraft average nearly seven accidents per 100,000 flight hours, compared with an average of 0.16 accidents per 100,000 hours for commercial airlines" (Fowler).

Finally, we will export our graphics.

```r
# Exporting the Graphics
# PDF
pdf('../plots_and_graphics/number_of_crashes_each_month.pdf')
barplot(monthly_data, xlab = 'Months',
        ylab = 'Number of Accidents',
        main = 'Number of Crashes Each Month',
        names.arg = month_names,
```

```
        las = 2,
        cex.names = .75,
        col = '#FD075E')
dev.off()
```

```
## pdf
##   2
```

```
# PNG
png('../plots_and_graphics/number_of_crashes_each_month.png', res = 96)
barplot(monthly_data, xlab = 'Months',
        ylab = 'Number of Accidents',
        main = 'Number of Crashes Each Month',
        names.arg = month_names,
        las = 2,
        cex.names = .75,
        col = '#FD075E')
dev.off()
```

```
## pdf
##   2
```

Part 2: Data by Decades

```
# packages needed
library(stringr)

# function needed
source('../code/norm_decade.R')
```

We wanted to look at the number of accidents per decade to see if any particular decades differed from the rest. However, the data is not organized into decades. First, we used regular expressions to pull the year out of the date column and created a new column named "year".

```
# making a year only column
data$year <- as.numeric(str_extract(data$date, '[0-9]{4}'))
```

For graphing, we wanted the axis to display decades, for example "1980s", and the corresponding aggregated data for that decade. We created a loop that categorizes the year and would group all years in a certain decade into that decade.

```
# labeling by decades
decade = NULL
for(i in 1:length(data$year)){
  x <- data$year[i]
  if ((x %in% c(1920:1929))==TRUE) x = '1920s'
  if ((x %in% c(1930:1939))==TRUE) x = '1930s'
  if ((x %in% c(1940:1949))==TRUE) x = '1940s'
  if ((x %in% c(1950:1959))==TRUE) x = '1950s'
  if ((x %in% c(1960:1969))==TRUE) x = '1960s'
```

```
  if ((x %in% c(1970:1979))==TRUE) x = '1970s'
  if ((x %in% c(1980:1989))==TRUE) x = '1980s'
  if ((x %in% c(1990:1999))==TRUE) x = '1990s'
  if ((x %in% c(2000:2009))==TRUE) x = '2000s'
  if ((x %in% c(2010:2019))==TRUE) x = '2010s'
  decade = c(decade, x)
}
data$decade <- decade
```

Next, we looked at the frequency of accidents for each decade by calling upon the data through table().

```
# looking at freq of accidents
table(data$decade)
```

```
##
## 1920s 1930s 1940s 1950s 1960s 1970s 1980s 1990s 2000s 2010s
##   167   319   500   592   719   837 24422 22953 19419  8507
```

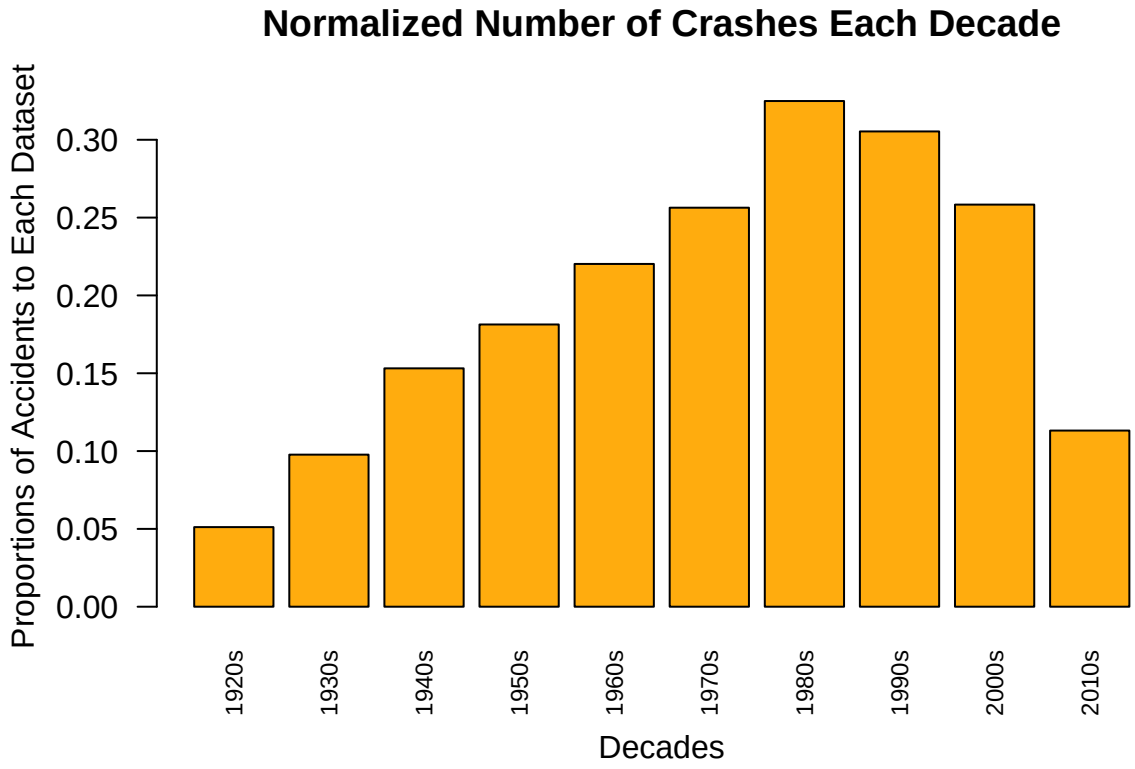
We started off by graphing a bar chart of the number of accidents per decade but found that the data was skewed due to the significantly larger size of the dataset for years 1980s and onwards. We then normalized the data by taking proportion of each decade's accidents to the total amount of accidents for the data set it came from. For example, 1920s proportion was: (# of accidents in 1920s) / (total # of accidents of dataset1)

```
# size of each dataset
amt_dataset1 <- nrow(air_data_6)
amt_dataset2 <- nrow(av_data_5)

norm_decade_1920 <- norm_decade('1920s')
norm_decade_1930 <- norm_decade('1930s')
norm_decade_1940 <- norm_decade('1940s')
norm_decade_1950 <- norm_decade('1950s')
norm_decade_1960 <- norm_decade('1960s')
norm_decade_1970 <- norm_decade('1970s')
norm_decade_1980 <- norm_decade('1980s')
norm_decade_1990 <- norm_decade('1990s')
norm_decade_2000 <- norm_decade('2000s')
norm_decade_2010 <- norm_decade('2010s')

vec_norm_decade <- c(norm_decade_1920, norm_decade_1930,  norm_decade_1940,
                     norm_decade_1950, norm_decade_1960, norm_decade_1970,
                     norm_decade_1980, norm_decade_1990, norm_decade_2000,
                     norm_decade_2010)
```

There seems to be a gradual increase in accidents till 1980s where it starts to drop off starting 1990s. Data from the 2000s show a decrease, which agrees with the FAA report.

```
# Graphing
decade_names <- c('1920s', '1930s', '1940s', '1950s', '1960s', '1970s',
                  '1980s', '1990s', '2000s', '2010s')

barplot(vec_norm_decade, xlab = 'Decades',
        ylab = 'Proportions of Accidents to Each Dataset',
```

```
        main = 'Normalized Number of Crashes Each Decade',
        names.arg = decade_names,
        las = 2,
        cex.names = .75,
        col = '#FFAC0E')
```



According to the FAA: "Between 2001 and 2007, aviation witnessed one of its safest periods for scheduled air carriers. Not counting the terrorist activities of September 11, 2001, there were only three fatal accidents in 2001; none in 2002; two in 2003; one in 2004; three in 2005; two in 2006; and none in 2007. Fatal accidents became rare events with only .01 accidents per 100,000 flight hours or .018 accidents per 100,000 departures."

Another explanation could be the improvements of air traffic control since after the 1980s (start of the digital age).

According to Wikipedia: As computers became more sophisticated in the 2000s, they began to take over routine aspects of the air traffic controller's task. Up until then all air traffic in nearby airspace was tracked and displayed, with the air traffic controller responsible for monitoring its position and assessing any need for action. Modern computerized systems are capable of monitoring the flight paths of many more aircraft at a given time, allowing the controller to manage more aircraft and to focus on the decision-making and follow-up processes.

Part 3: Type analysis

Here we are looking at how the most common types of airplanes to crash changed over different decades. We are parse out the data by decade so that we can plot by type. Specifically, we extract just the top three values.

```
# packages needed
library(stringr)
```

```r
library(ggplot2)

# functions needed
source('../code/fun_top3.R')
```

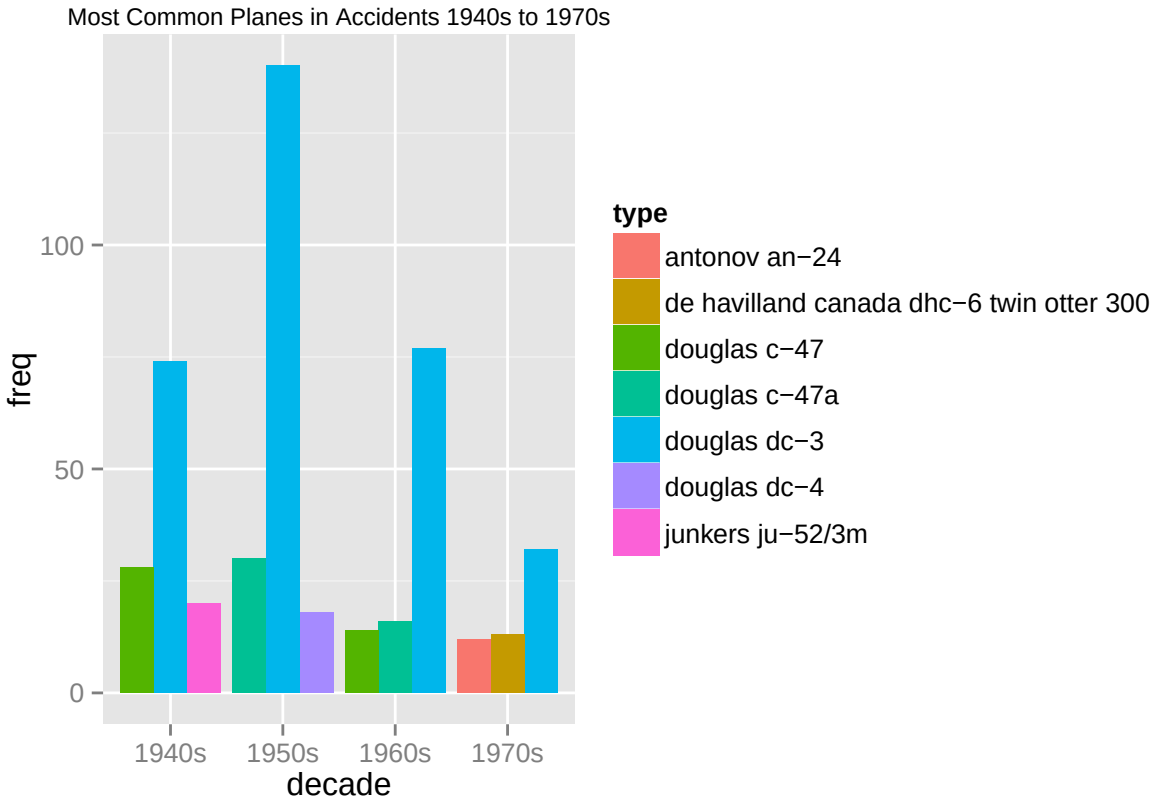
```r
top3_array <- fun_top3(data, decade_names)
```

Now we want to create a data frame that combines the data from each decade in a way that we can plot it efficiently.

```r
type <- names(top3_array)
freq <- as.vector(top3_array)
decade <- rep(decade_names, each = 3)

top3 <- data.frame(type = type, freq = freq, decade = decade)
```

These are the two sets of decades we want to focus on. Because of the differences among our data sets, we want to look at: - from 1940-1980

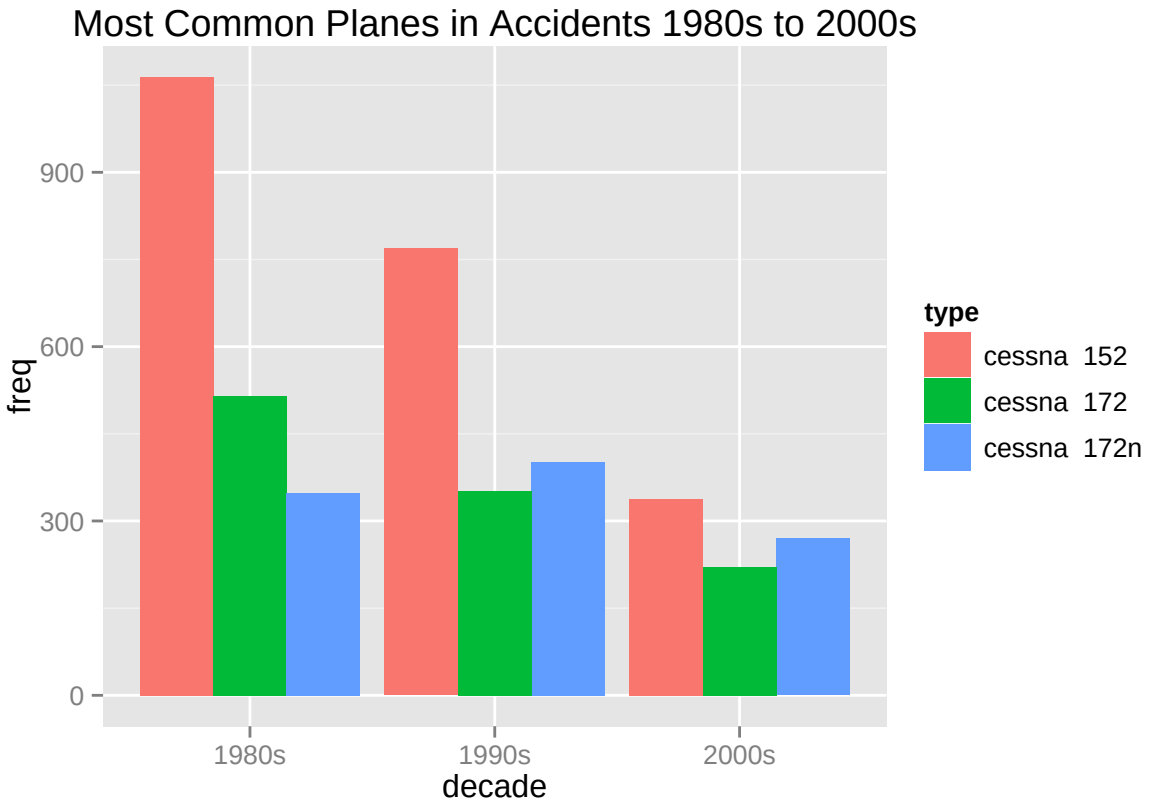
```r
decade_names_1 <- decade_names[3:6]
type_1 <- type[7:18]
freq_1 <- freq[7:18]
decade_1 <- rep(decade_names_1, each = 3)

top3_1 <- data.frame(type = type_1, freq = freq_1, decade = decade_1)

# plotting the trend
ggplot(top3_1, aes(x = decade, y = freq, fill = type))+
  geom_bar( stat = 'identity', position = position_dodge())+
  ggtitle('Most Common Planes in Accidents 1940s to 1970s')+
  theme(plot.title = element_text(size = rel(.75)))
```



From our graph, we can see that from 1940-1979, the Douglas DC-3, was the plane that got into the most accidents. However, this could be due to the fact that the Douglas DC-3 was the most produced airliner. So, because there were more of them in production, the chances that they got into accidents were higher for that reason.

- from 1980-2010

```
decade_names_2 <- decade_names[7:9]
type_2 <- type[19:27]
freq_2 <- freq[19:27]
decade_2 <- rep(decade_names_2, each = 3)

top3_2 <- data.frame(type = type_2, freq = freq_2, decade = decade_2)

# plotting the trend
ggplot(top3_2, aes(x = decade, y = freq, fill = type))+
  geom_bar( stat = 'identity', position = position_dodge())+
  ggtitle('Most Common Planes in Accidents 1980s to 2000s')
```



From our graph here, we can see that from 1980-2010 the plane that got into the most accidents was Cessina 152. In fact, all of the top 3 planes that got into accidents are Cessina. Cessina 152 is a private plane and according to an article from The Economist, commercial air travel has been a lot safer since the 1980s, which would explain why from 1980 forward, private planes, such as Cessina, have been getting into more accidents. There are also many other reasons as to why Cessina and other private planes have been getting into more accidents. For example, not a lot of private plane owners have a lot of experience, private planes are flown by one person and most accidents have come from them losing control or not seeing something and flying into it due to no one looking out for them, and having smaller fuel capacities so there is less room for error. Finally, we might have seen more Cessinas involved in accidents because there have been more Cessina produced in recent years than any other kind of plane.