Lab 2: data.gov and reproducibility

James Barbour

January 18, 2017

Aviation Accidents and Fatalities, 1975-2014

This data set is provided by the NTSB, found here. Initially, it was poorly formatted and unusable, but since it is a small dataset, I was able to quickly manually clean up the data.

The data set contains the number of aviation accidents (All) from 1975 to 2014. For each year, the number of fatal accidents (Fatal), the total number of fatalities (Total), number of fatalaties aboard flights (Aboard), and flight hours logged that year (Flight.Hours).

```
datagov <- read_csv("aviation_accidents-2014.csv")
```

```
## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   All = col_integer(),
##   Fatal = col_integer(),
##   Total = col_integer(),
##   Aboard = col_integer(),
##   Flight.Hours = col_integer()
## )
```

```
summary(datagov)
```

```
##       Year           All           Fatal           Total       
##  Min.   :1975   Min.   :1221   Min.   :222.0   Min.   : 391.0  
##  1st Qu.:1984   1st Qu.:1694   1st Qu.:323.0   1st Qu.: 572.0  
##  Median :1994   Median :2056   Median :404.0   Median : 734.0  
##  Mean   :1994   Mean   :2336   Mean   :428.4   Mean   : 799.9  
##  3rd Qu.:2004   3rd Qu.:2878   3rd Qu.:521.5   3rd Qu.:1004.5  
##  Max.   :2014   Max.   :4216   Max.   :719.0   Max.   :1556.0  
##      Aboard        Flight.Hours     
##  Min.   : 386.0   Min.   :18103000  
##  1st Qu.: 558.5   1st Qu.:23891000  
##  Median : 727.0   Median :25998000  
##  Mean   : 781.5   Mean   :26752641  
##  3rd Qu.: 983.0   3rd Qu.:28736000  
##  Max.   :1398.0   Max.   :38641000  
```

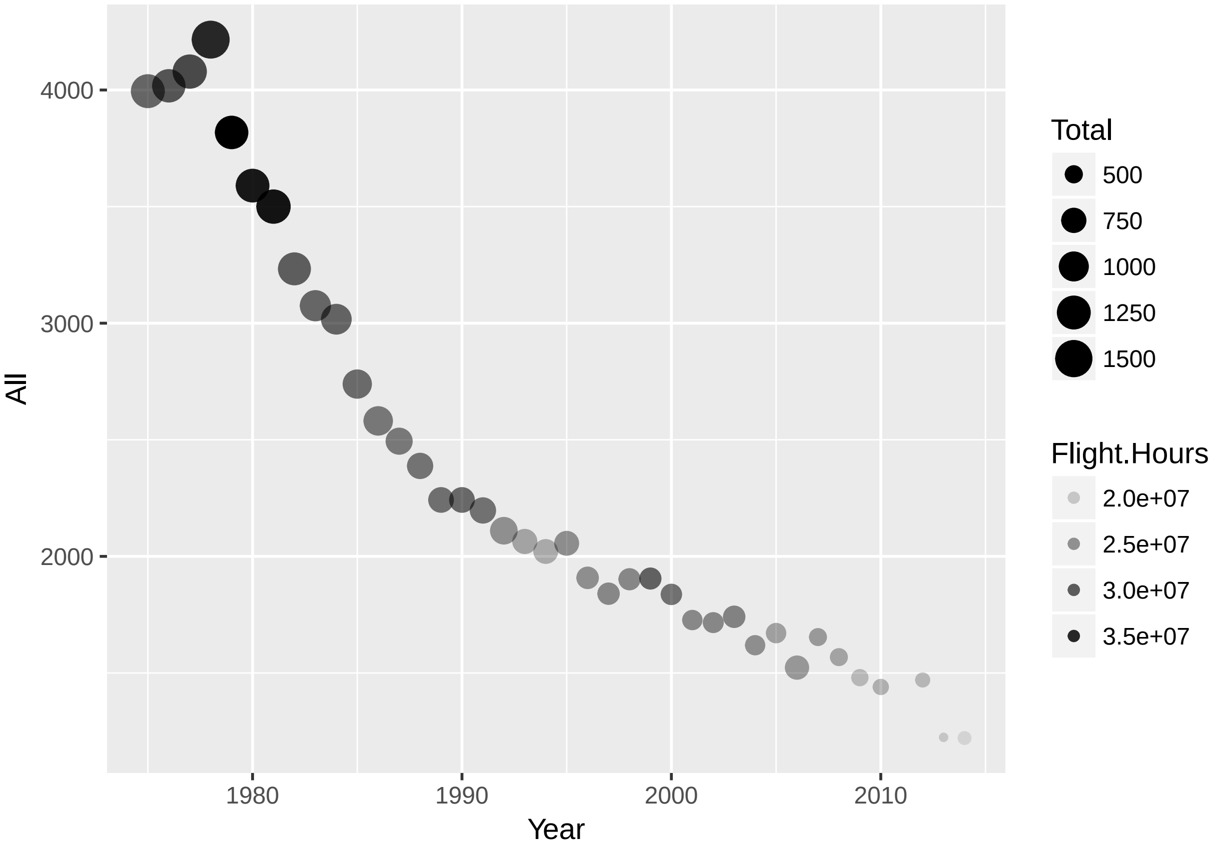
```
head(datagov)
```

```
## # A tibble: 6 × 6
##    Year   All Fatal Total Aboard Flight.Hours
##   <int> <int> <int> <int>  <int>        <int>
## 1  1975  3995   633  1252   1231     28799000
## 2  1976  4018   658  1216   1203     30476000
## 3  1977  4079   661  1276   1265     31578000
## 4  1978  4216   719  1556   1398     34887000
## 5  1979  3818   631  1221   1203     38641000
## 6  1980  3590   618  1239   1230     36402000
```

Total accidents per year

Points sized by number of fatalities, alpha is flight hours per year. This plot tells us there is a correlation between both year and flight hours and total number of annual accidents.

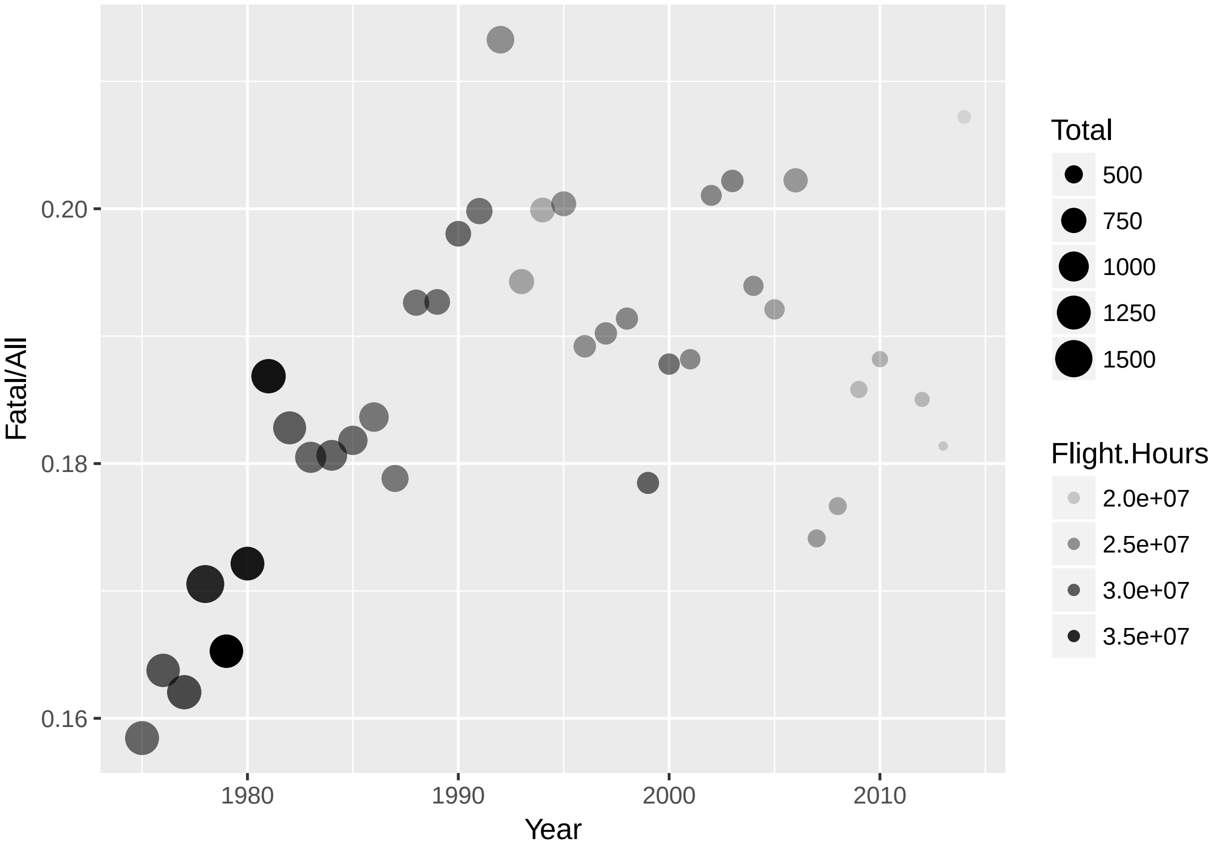
```
ggplot(data = datagov) + geom_point(mapping = aes(x = Year, y = All, size = Total, alpha = Flight.Hours)
```



Percentage of fatal accidents per year

Points sized by total number of fatalities, alpha is flight hours per year. This plot tells us there may be a weak correlation between of percentage of fatal accidents and year.

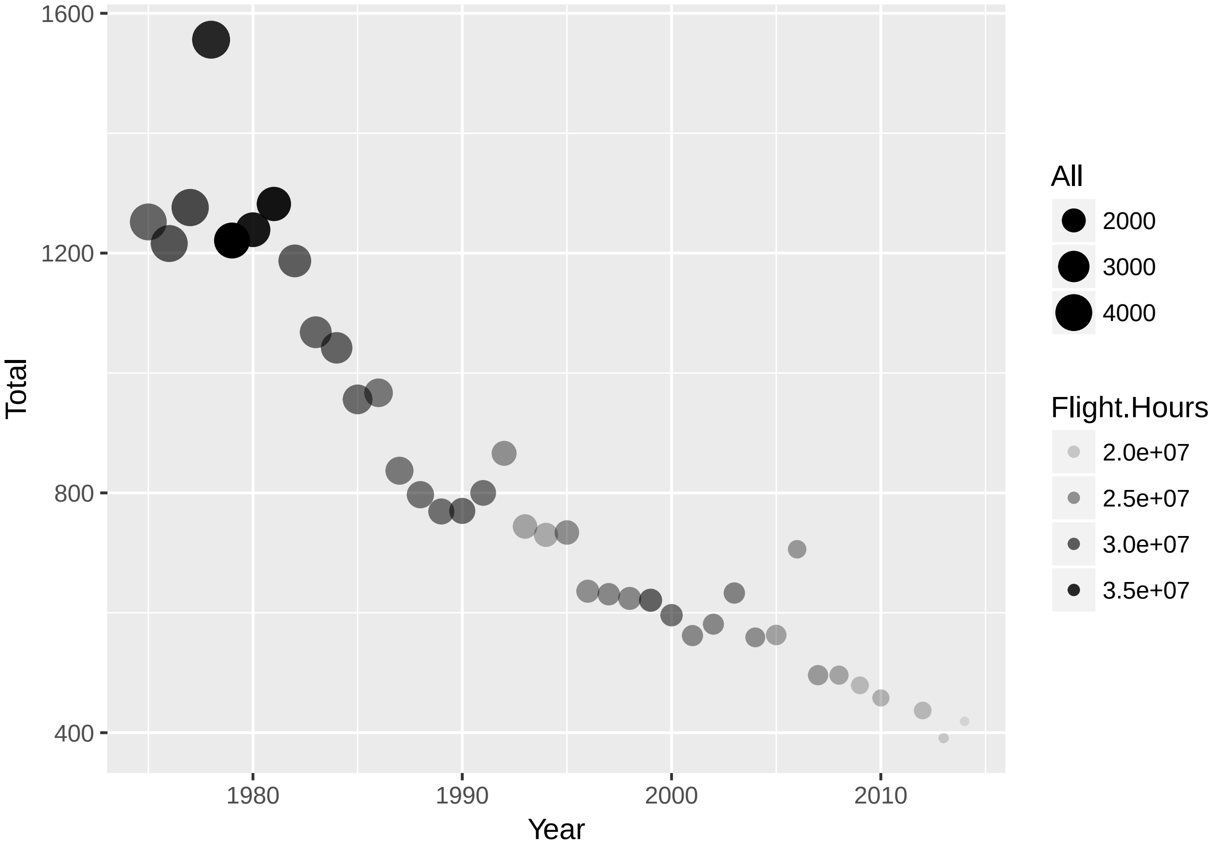
```
ggplot(data = datagov) + geom_point(mapping = aes(x = Year, y = Fatal/All, size = Total, alpha = Flight
```



Fatalities per year

Points sized by total accidents, alpha is flight hours per year. This plot tells us there is a correlation between both year and flight hours and total number of annual fatalities.

```
ggplot(data = datagov) + geom_point(mapping = aes(x = Year, y = Total, size = All, alpha = Flight.Hours
```


1600
1200
800
400
Total
1980
1990
2000
2010
Year
All
2000
3000
4000
Flight.Hours
2.0e+07
2.5e+07
3.0e+07
3.5e+07