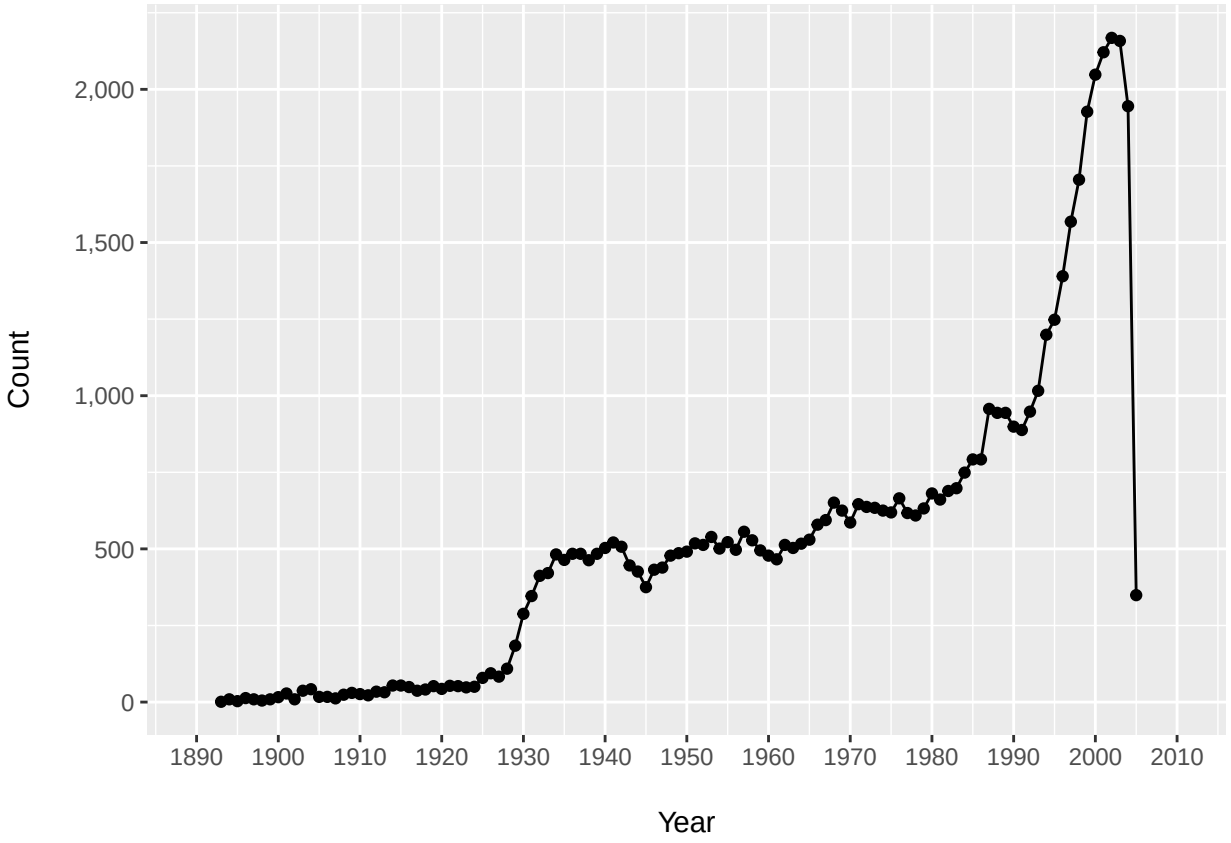
Case Study in `dplyr`: `ggplot2` Movies

Paul Intrevado

July 26, 2018

The `ggplot2movies` package contains a dataset named `movies`, which provides information on 58,788 movies, datging as far back as 1893! Here is a quick graphical summary of how many movies were realeased each year:



Aggregating by deacde, we observe that the numnber of movies produced by decade is monotonically non-decreaing (see Table below), save for the final deacde, perhaps because data was not collected fully through that decade.

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

Decade	Count
1890s	49
1900s	232
1910s	401
1920s	795
1930s	4,328
1940s	4,613
1950s	5,160
1960s	5,456
1970s	6,270
1980s	7,907
1990s	12,788

Decade	Count
2000s	10,789

Per decade, we can also track the total budget spent on movies.

Decade	Total Budget
1890s	0
1900s	2,250
1910s	2,871,593
1920s	47,962,944
1930s	183,711,680
1940s	252,771,487
1950s	412,471,220
1960s	1,139,290,959
1970s	1,511,260,277
1980s	5,684,252,230
1990s	27,630,455,573
2000s	33,081,206,384

As judicious and transparent analysts, we must also disclose that the `budget` variable contains 53,573, the majority of which are in earlier decades. The following table offers a holistic view of the data we have.

Decade	Total Budget	Total NAs	Total Movies	% NAs
1890s	0	49	49	100
1900s	2,250	229	232	99
1910s	2,871,593	378	401	94
1920s	47,962,944	719	795	90
1930s	183,711,680	4,070	4,328	94
1940s	252,771,487	4,446	4,613	96
1950s	412,471,220	4,911	5,160	95
1960s	1,139,290,959	5,119	5,456	94
1970s	1,511,260,277	5,912	6,270	94
1980s	5,684,252,230	7,342	7,907	93
1990s	27,630,455,573	11,471	12,788	90
2000s	33,081,206,384	8,927	10,789	83

It is very clear that the budget information we have is highly unrealiable as 91.13% our of data are `NA`s. Looking at movie length, we observe that our data is far more reliable (see proceeding table).

Decade	Total Movie Length (min)	Mean Length	Median Length	Total NAs	Total Movies	% NAs
1890s	66	1	1	0	49	0
1900s	1,264	5	3	0	232	0
1910s	16,307	41	24	0	401	0
1920s	53,196	67	71	0	795	0
1930s	270,355	62	70	0	4,328	0
1940s	318,004	69	76	0	4,613	0
1950s	403,624	78	85	0	5,160	0
1960s	481,796	88	92	0	5,456	0
1970s	592,273	94	94	0	6,270	0

Decade	Total Movie Length (min)	Mean Length	Median Length	Total NAs	Total Movies	% NAs
1980s	745,238	94	93	0	7,907	0
1990s	1,114,029	87	93	0	12,788	0
2000s	844,327	78	90	0	10,789	0