# Tacoma Business Licenses: Landlord Spike

*Joe Izenman*

*June 20, 2016*

## Summary

These are the steps to recreate my initial exploratory analysis of the City of Tacoma open business license data. Data was sourced from the city's open data portal, and exported to CSV.

Data Source: https://data.cityoftacoma.org/Business/TACOMA-BUSINESS-LICENSE/29fz-3eiz

## Loading and Recoding

First, load the data from the export

```
businesses <- read.csv("data/TACOMA_BUSINESS_LICENSE.csv", header=T)
```

Next, modify a few fields to work better with R.

```
# Convert BUSINESS.OPEN.DATE, currently a text string, to R's date format
businesses$BUSINESS.OPEN.DATE <- as.Date(businesses$BUSINESS.OPEN.DATE, format("%m/%d/%Y"))

# Convert the categorical variables to factors
businesses$NAICS.CODE <- factor(businesses$NAICS.CODE)
businesses$NAICS.CODE.DESCRIPTION <- factor(businesses$NAICS.CODE.DESCRIPTION)
```

## Businesses By Year

We want to view open businesses by year, to look for growth trend over time. To simplify later code, break the year out into its own column.

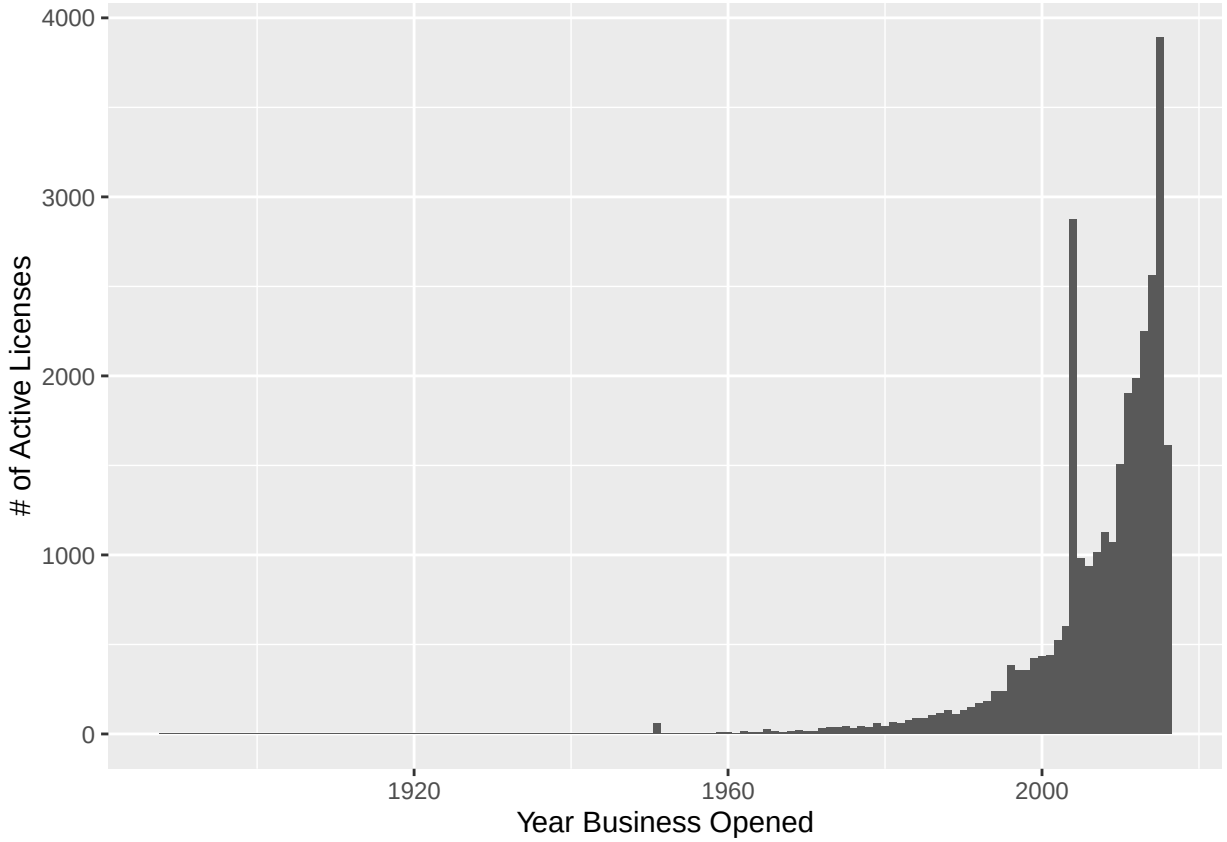
```
# Create a year column from the date
businesses$BUSINESS.OPEN.YEAR <- as.numeric(format(businesses$BUSINESS.OPEN.DATE, "%Y"))
```

A quick look at the data shows several rows with years in the future. It's not clear where this data came from, but it's not useful for the current analysis, so set up an alternate data set with only businesses up to the current year.

```
# Remove outliers whose year came out greater than the current one
open.businesses <- businesses[businesses$BUSINESS.OPEN.YEAR < 2017,]
```

Using the new BUSINESS.OPEN.YEAR column, plot a basic histogram of the data.

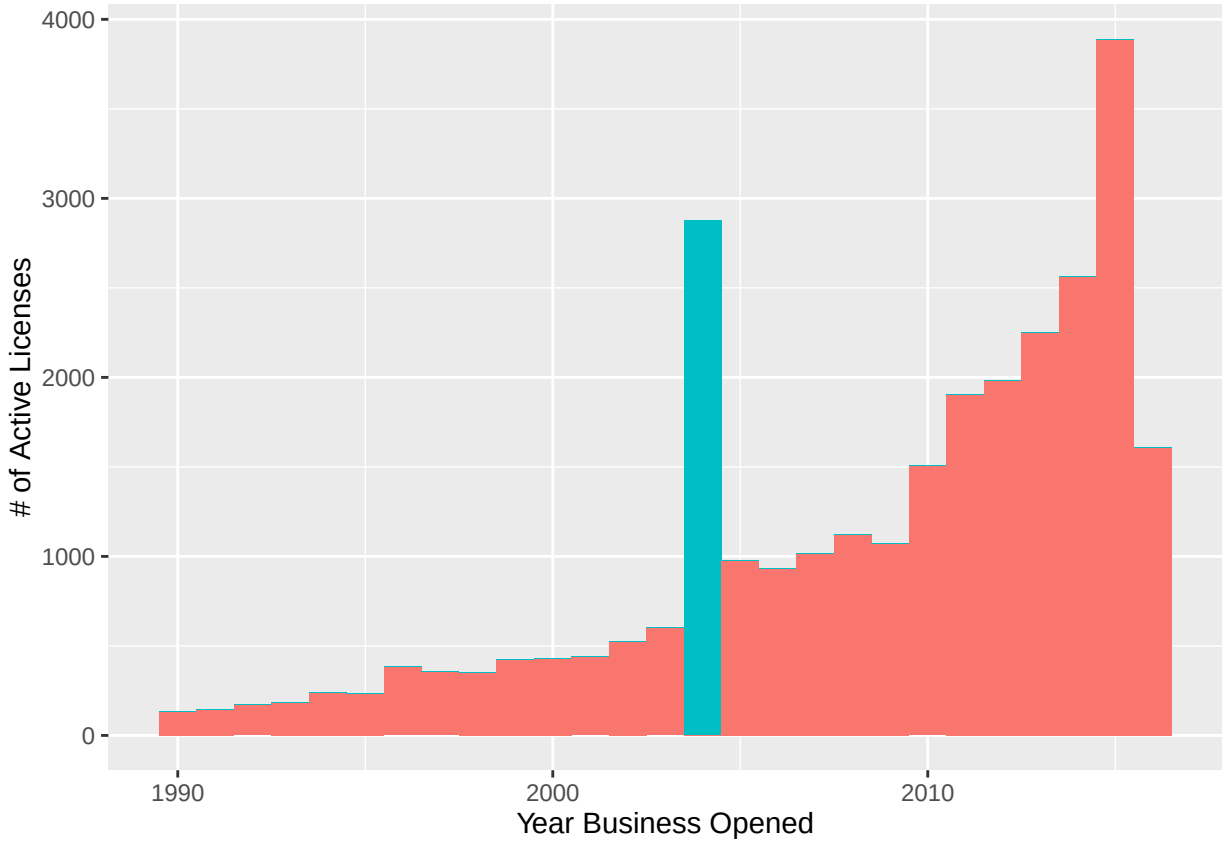
```
# Build the initial plot object
businesses.by.year <- ggplot(open.businesses, aes(BUSINESS.OPEN.YEAR)) +
  labs(legend.position = "none")
# Output a histogram, with one bin per year
businesses.by.year +
  geom_histogram(binwidth = 1, na.rm=TRUE) +
  labs(x = "Year Business Opened", y = "# of Active Licenses")
```



### What's With 2004?

Zoom in on the data, and highlight 2004.

```
# Build a plot object that starts at 1990, and highlights 2004
businesses.by.year.90 <- ggplot(
  open.businesses[open.businesses$BUSINESS.OPEN.YEAR >= 1990,],
  aes(BUSINESS.OPEN.YEAR, fill = BUSINESS.OPEN.YEAR == 2004)
  ) + labs(legend.position = "none")
# Output a similar histogram as above
businesses.by.year.90 +
  geom_histogram(binwidth = 1, na.rm=TRUE) +
  labs(x = "Year Business Opened", y = "# of Active Licenses") +
  guides(fill=F)
```



Summarize the NAICS.CODE.DESCRIPTION field. Since it was made a factor above, it will show counts of each category. (Only displaying the first 5 here.)

```
summary(open.businesses$NAICS.CODE.DESCRIPTION[open.businesses$BUSINESS.OPEN.YEAR == 2004 &
                                                 !is.na(open.businesses$BUSINESS.OPEN.YEAR)])
```

```
##               Lessors of Residential Buildings and Dwellings
##                                                         2084
## Lessors of Nonresidential Bildngs (ex Miniwrhss) (ES 0200 02
##                                                          199
## Commercial and Institutional Building Construction  (ES  020
##                                                           16
##                         Lessors of Other Real Estate Property
##                                                           15
##                           Beauty Salons  (ES 0200 0225    08)
##                                                           13
##  [ reached getOption("max.print") -- omitted 95 entries ]
```

Since the spike is pretty clearly due to the Lessors of Residential Buildings category, compare its percentage to neighboring years (code 531110 is the numeric equivalent to this category, observed in the dataset).

```
nrow(businesses[businesses$BUSINESS.OPEN.YEAR == 2003 & businesses$NAICS.CODE == 531110,]) /
  nrow(businesses[businesses$BUSINESS.OPEN.YEAR == 2003,])
```

```
## [1] 0.1221374
```

```
nrow(businesses[businesses$BUSINESS.OPEN.YEAR == 2004 & businesses$NAICS.CODE == 531110,]) /
  nrow(businesses[businesses$BUSINESS.OPEN.YEAR == 2004,])
```

```
## [1] 0.7243852
```

```
nrow(businesses[businesses$BUSINESS.OPEN.YEAR == 2005 & businesses$NAICS.CODE == 531110,]) /
  nrow(businesses[businesses$BUSINESS.OPEN.YEAR == 2005,])
```

```
## [1] 0.3734239
```

The value jumps from 12% of licenses issued, to 72%, back down to 37%, cementing our conclusion that this is the cause of the 2004 spike.