Exploratory Analysis on active buiness in San Francisco

Shuyu Sui

April 30, 2018

```
suppressPackageStartupMessages({
  library(ggplot2)
  library(dplyr)
  library(ggmap)
  library(tidyr)
  library(lubridate)
  library(stringr)
  library(VIM)
})
```

In this analysis I analysed data from DataSF. The Dataset contains registered business opening, ending times, location, neighbor hood and related information for all businesses that pay/paid taxes to City and County of San Francisco.

The dataset is updated weekly on the website. The dataset for this analysis was collected on 2016-09-30. So we are mainly concerned with business activate around that time.

In the following analysis, I cleaned the dataset to eliminate abnormal data points. I then identify active high-growth regions using `ggmap`. I also analyse the start dates and find interesting insights about industry trend.

1. Data Cleaning

```
SFbusiness <- read.csv("sf business dataset.csv")
```

1.1 Preprocessing

Preprocessing and data cleaning is very important for us to look at the quality of the data and assess risks of our assumptions.

1.1.1

First, we need to change date related data from text to date type in order to do further analysis

```
#First change the type of time related data into dates
SFbusiness$Business.Start.Date <- as.Date(SFbusiness$Business.Start.Date, format = '%m/%d/%Y')
SFbusiness$Business.End.Date <- as.Date(SFbusiness$Business.End.Date, format = '%m/%d/%Y')
SFbusiness$Location.Start.Date <- as.Date(SFbusiness$Location.Start.Date, format = '%m/%d/%Y')
SFbusiness$Location.End.Date <- as.Date(SFbusiness$Location.End.Date, format = '%m/%d/%Y')
```

1.1.2

As we are mainly focusing on business in San Francisco, we need to filter for shops that have primary City as San Francisco. There might be variations of how San Francisco is written. So we need to check for City that

contains'San' or 'Fran' or 'SF'.

```
head(unique(SFbusiness$City[str_detect(SFbusiness$City, 'Fran')]), 10)
```

```
##  [1] San Francisco       Franklin            S San Fran
##  [4] Franklin+lakes      San Francisco Co    San Francisoc
##  [7] San Franciso        South San Francisco S San Francisco
## [10] Sans Francisco
## 2337 Levels:  0 0akland 18773 1945+grant+ave+richmond 94122 94124 ... Yuma
```

```
#unique(SFbusiness$City[str_detect(SFbusiness$City, 'Fran')])
unique(SFbusiness$City[str_detect(SFbusiness$City, 'SF')])
```

```
## factor(0)
## 2337 Levels:  0 0akland 18773 1945+grant+ave+richmond 94122 94124 ... Yuma
```

As we could see from the result, there are many variations for 'San Francisco' within the City variable. we also notice that there's no abbreviation used for the city, so we only need to consider 'San Francisco' and its variations. I used `agrep` to filter those most likely to be San Francisco.

```
#Fuzzy Matching utilized Levenshtein to do fuzzy match
sf_cityname <- unique(SFbusiness$City[str_detect(SFbusiness$City,
                                                 'San')])[agrepl('San Francisco', unique(SFbusiness$City

#remove south SF
sf_cityname <- sf_cityname[!agrepl('south',sf_cityname)]
```

```
#create filterd frame for analysis focusing on SF city
sfbusiness_city <- SFbusiness[SFbusiness$City %in% sf_cityname, ]
sfbusiness_city <-sfbusiness_city[sfbusiness_city$City != 'Treasure Island, San Francisco', ]
```

1.2 Handling missing values and abnormal data points

1.2.1 Summary statistics

To get a general picture of the data we want to analyze. I start from general statistics to see if there's any abnormal that should be considered before jumping into business analysis.

1) There are many Blanks for business corridor, so it might not be a good indication of location of business.
2) Neighborhood variable contains about 20% of blanks but probably is a better indication of location.
3) We need to extract latitude and longitude and examine if this would be a better indication of location
4) Max business and location start/end dates have some outliers which are not reasonable. While one might argue that such abnormal start date might be due to advanced planning, due to uncertain factors in the future, we might want to leave these data points out for the current analysis.
5) For business and location end dates we need to make sure location end dates proceeds or equals to business end date. We also need to make sure that both end dates are later or equal to than start dates.

For further analysis, we will need to identify missing values and what to do with these entries. And then we could filtered out abnormal data points.

```
summary(sfbusiness_city %>%
          select(City, Business.Start.Date,
                 Business.End.Date,
                 Location.Start.Date,
                 Location.End.Date,
                 NAICS.Code.Description,
                 Neighborhoods...Analysis.Boundaries,
```

```
                Business.Corridor,
                Business.Location))
```

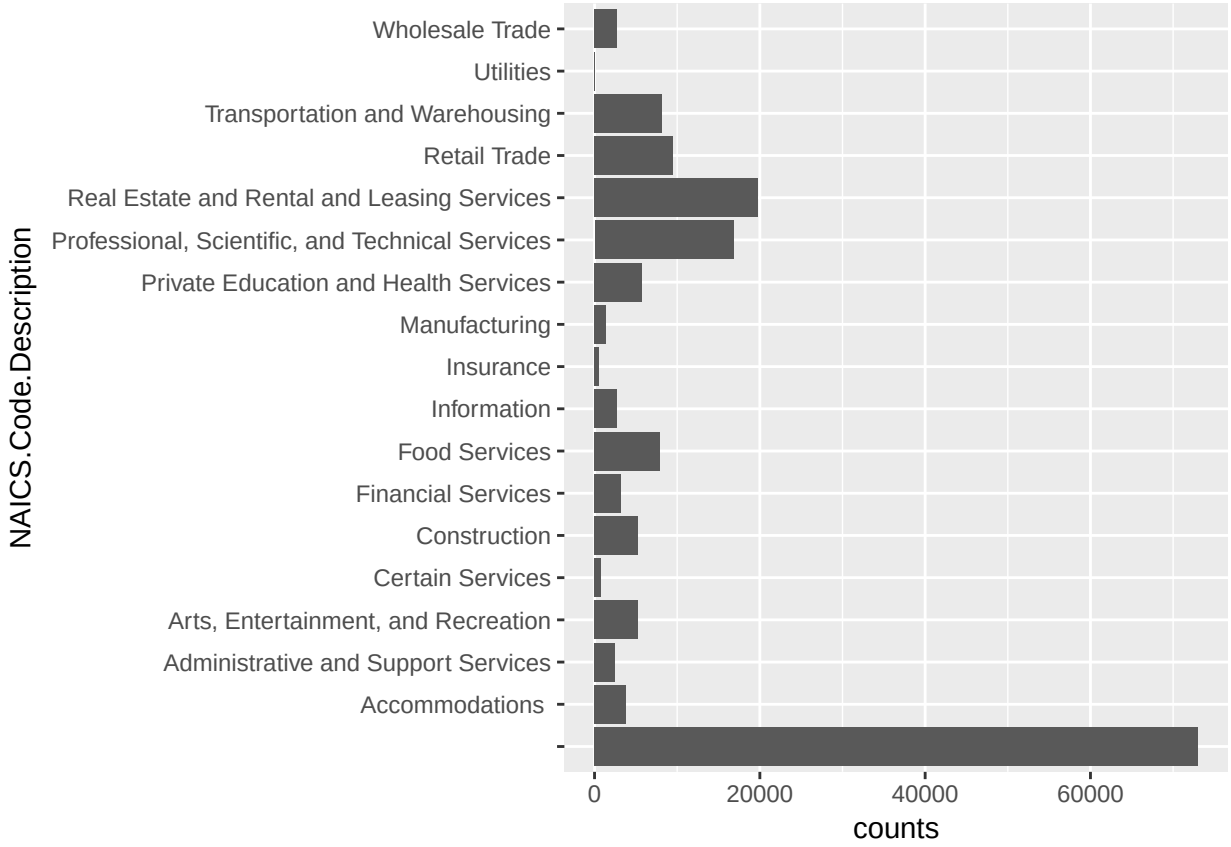
```
##              City        Business.Start.Date  Business.End.Date
##  San Francisco :154100   Min.   :1849-09-01   Min.   :1957-08-24
##  San+francisco : 14416   1st Qu.:2002-03-07   1st Qu.:2014-04-04
##  Sanfrancisco  :    71   Median :2010-01-01   Median :2015-05-15
##  San Francsico :    66   Mean   :2005-12-15   Mean   :2015-01-16
##  San Franciso  :    42   3rd Qu.:2014-02-10   3rd Qu.:2016-05-01
##  San Franciscco:    38   Max.   :2029-03-11   Max.   :2017-06-26
##  (Other)       :   330                        NA's   :135309
##  Location.Start.Date  Location.End.Date
##  Min.   :1849-09-01   Min.   :1957-08-24
##  1st Qu.:2006-06-01   1st Qu.:2014-03-31
##  Median :2011-11-22   Median :2015-05-01
##  Mean   :2008-07-26   Mean   :2015-01-05
##  3rd Qu.:2014-11-07   3rd Qu.:2016-04-19
##  Max.   :2029-03-11   Max.   :2017-06-26
##                       NA's   :120677
##                                        NAICS.Code.Description
##                                                      :73004
##  Real Estate and Rental and Leasing Services         :19826
##  Professional, Scientific, and Technical Services:16814
##  Retail Trade                                        : 9458
##  Transportation and Warehousing                      : 8139
##  Food Services                                       : 7934
##  (Other)                                             :33888
##              Neighborhoods...Analysis.Boundaries      Business.Corridor
##                                      :38370                      :168797
##  Financial District/South Beach:17759            Chinatown     :   101
##  Mission                        :10332           Central Market:    71
##  South of Market                : 7355           North Beach   :    47
##  Sunset/Parkside                : 7234           Lower Polk    :    20
##  Bayview Hunters Point          : 5621           Lombard Street:    10
##  (Other)                        :82392           (Other)       :    17
##                                                                Business.Location
##                                                                     : 22497
##  870 MARKET ST\nSan Francisco, CA 94102\n(37.784761, -122.407108)    :   374
##  450 SUTTER ST\nSan Francisco, CA 94108\n(37.789331, -122.407788)    :   243
##  220 MONTGOMERY ST\nSan Francisco, CA 94104\n(37.791218, -122.402449):   238
##  50 CALIFORNIA ST\nSan Francisco, CA 94111\n(37.793604, -122.397307) :   225
##  44 MONTGOMERY ST\nSan Francisco, CA 94104\n(37.789704, -122.402127) :   203
##  (Other)                                                            :145283
```

1.2.1 Missing values

Missing values could come from both NAICS category and business location. From the below graph we could clearly see that half of data doesn't have a category. Then I extracted latitude and longitude of business locations.

```
#view category
sfbusiness_city %>%
  group_by(NAICS.Code.Description)%>%
  summarise(counts = n())%>%
```

```
  ungroup() %>%
  ggplot() +
  geom_bar(aes(x = NAICS.Code.Description, y = counts), stat = 'identity') + coord_flip()
```




```
#extract latitudes and longitudes from business location
sfbusiness_city <- sfbusiness_city%>%
  separate(col = Business.Location, into = c("street", 'lat'), sep = "\\(") %>%
  separate(col = lat, into = c("lat", 'lon'), sep = ',') %>%
  separate(col = lon, into = c('lon', 'blank'), sep = '\\)') %>%
  select(-c(blank, street)) %>%
  mutate_if(is.character,funs(as.numeric))
```

```
## Warning: Too many values at 2 locations: 28892, 102433
```

```
## Warning: Too few values at 24477 locations: 25, 43, 187, 207, 372, 461,
## 462, 465, 468, 625, 705, 1139, 1183, 1348, 1548, 1634, 1643, 1644, 1739,
## 1767, ...
```

```
## Warning: Too few values at 11 locations: 11767, 18796, 28892, 31096, 31119,
## 92946, 102433, 126682, 127460, 143070, 146046
```

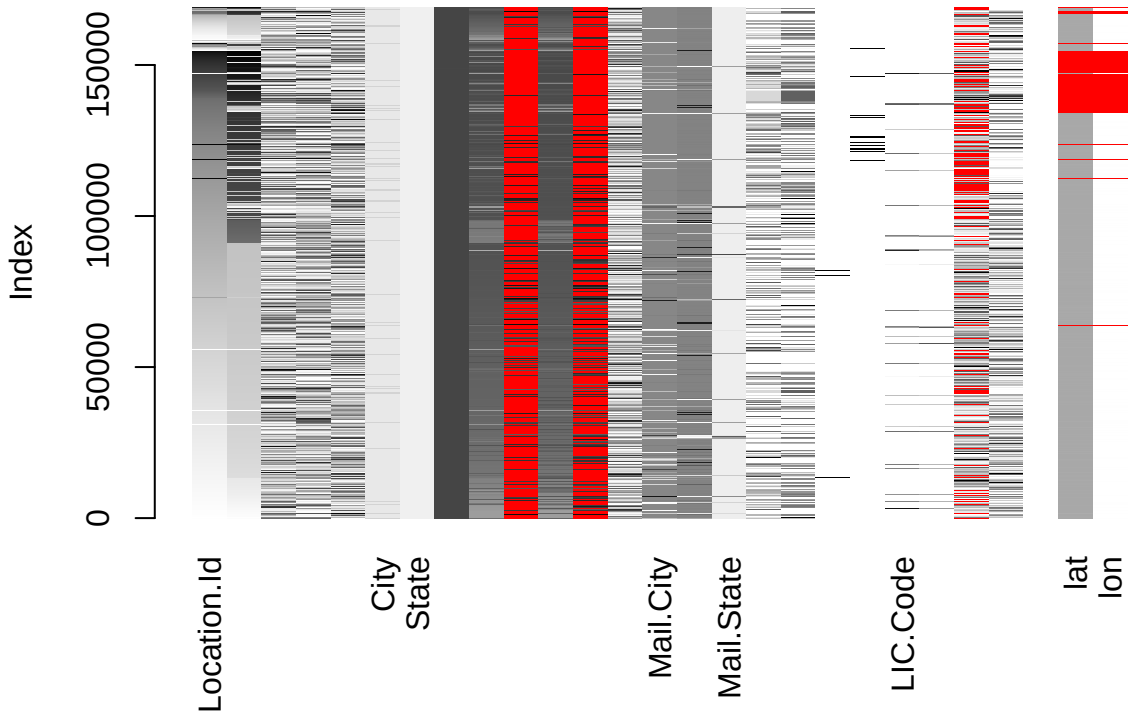
```
## Warning in evalq(as.numeric(lat), <environment>): NAs introduced by
## coercion
```

```
## Warning in evalq(as.numeric(lon), <environment>): NAs introduced by
## coercion
```

After the extraction, we could now examine the NAs for the whole dataset.As we could see from the below graph, there's high concentration of business location missing values for certain part of the data. To examine

whether there's correlation between missing values within location and within other variables, I examined density plots for time-related vairables and bar graph for categorical variables.

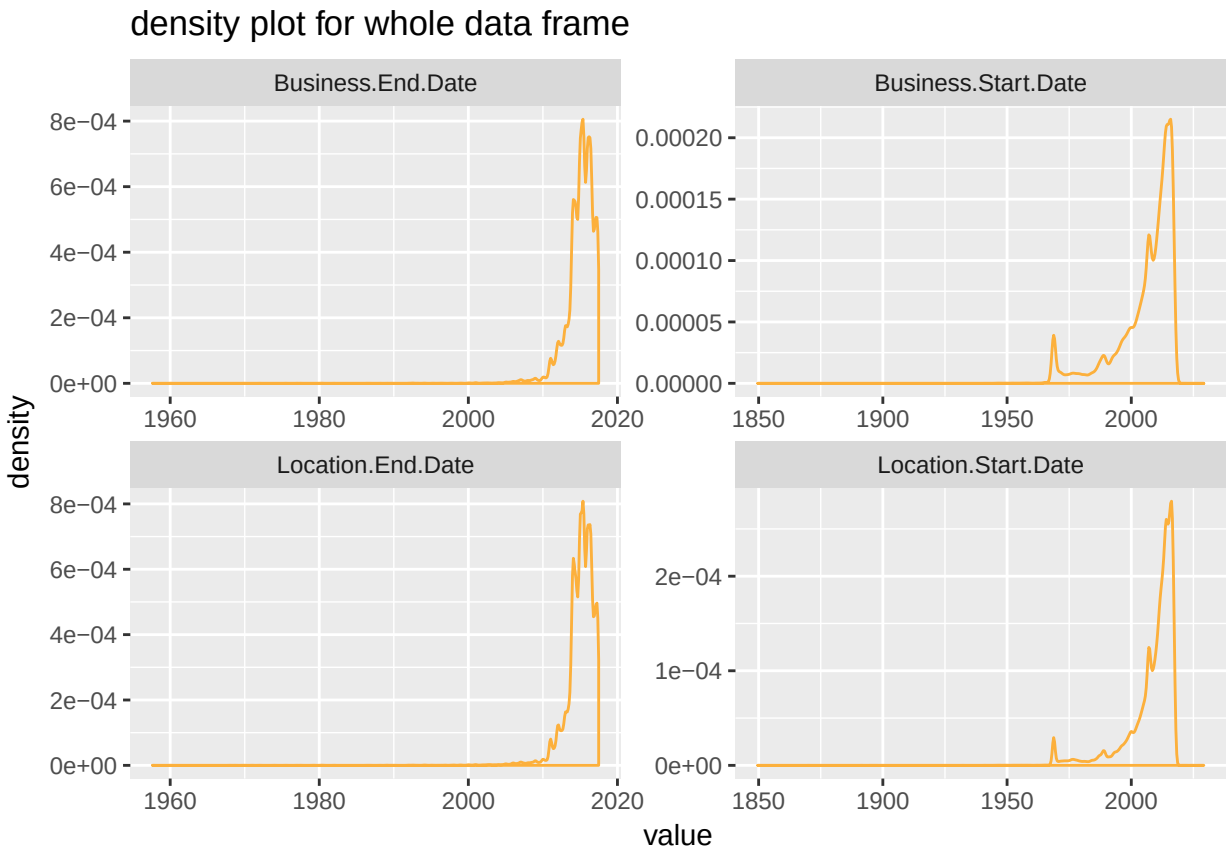
```
matrixplot(sfbusiness_city)
```



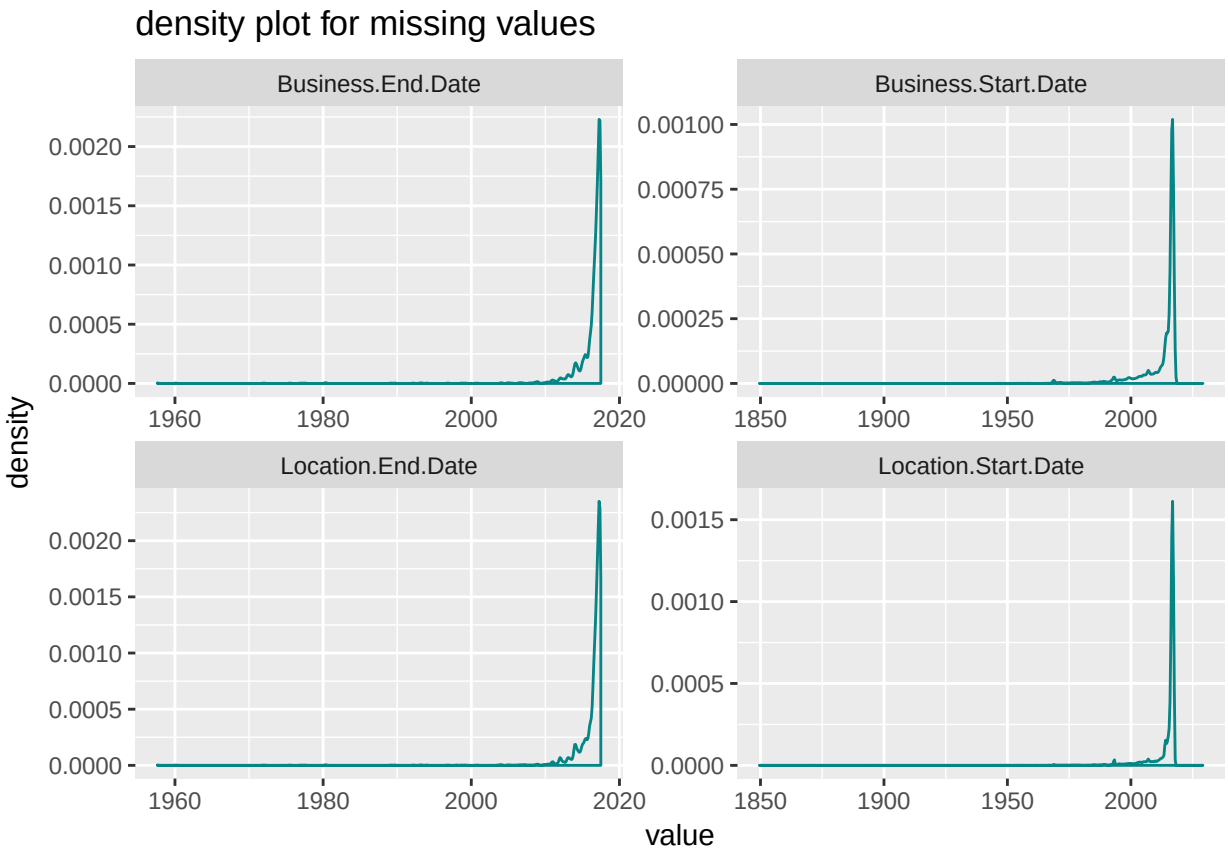
From the below two density plots we could see that the density plot of missing values are more skewed than the general population. The highly skewed parts for the missing values include those data points with abnormal dates in the future. Therefore, I decided to remove the abnormal start dates and end dates first.

```
#color pallete for better illustration
color_pal <- c('#1A4F63', '#068587', '#6FB07F', '#FCB03C', '#FC5B3F')

#plot for density plot for entire dataframe
sfbusiness_city %>% select(Location.Start.Date,
                           Location.End.Date, Business.Start.Date,
                           Business.End.Date)%>%
gather(key = "var", value = "value") %>%
  ggplot(aes(value)) +
    geom_density(color = color_pal[4], na.rm = TRUE) +
    facet_wrap(~ var, scales = "free") +
  ggtitle("density plot for whole data frame")
```




```
#plot for density plot for missing values
sfbusiness_city %>%
  filter(is.na(lat)) %>%
  select(Location.Start.Date,
                           Location.End.Date, Business.Start.Date,
                           Business.End.Date)%>%
gather(key = "var", value = "value") %>%
  ggplot(aes(value)) +
    geom_density(color = color_pal[2], na.rm = TRUE) +
    facet_wrap(~ var, scales = "free") +
  ggtitle("density plot for missing values")
```



When removing abnormal dates, we have to set a standard to remove. As businesses do not open just in one day, it is reasonable to allow for a three months window after the data collection day. And therefore, we consider all data with start/end dates as abnormal data points after 2017-01-01 for both Business and Location.

I also examined data with end dates earlier than start dates and with location start dates earlier than business start dates. As data points with these situations are less than 5% I will leave these data now to be.

```
#Remove business dates and location start/end dates that's after 2017-01-01
sfbusiness_2016 <- sfbusiness_city %>%
  filter(is.na(Business.Start.Date)|Business.Start.Date < '2017-01-01') %>%
  filter(is.na(Business.End.Date) | Business.End.Date < '2017-01-01') %>%
  filter(is.na(Location.Start.Date) | Location.Start.Date < '2017-01-01') %>%
  filter(is.na(Location.End.Date) | Location.End.Date < '2017-01-01')

#Examine whether there are business/location dates with end dates ealier then start dates
#sfbusiness_2016 %>%
#  filter(Business.Start.Date > Business.End.Date)

dim(sfbusiness_2016 %>%
  filter(Location.Start.Date > Location.End.Date))
```

```
## [1] 426  27
```

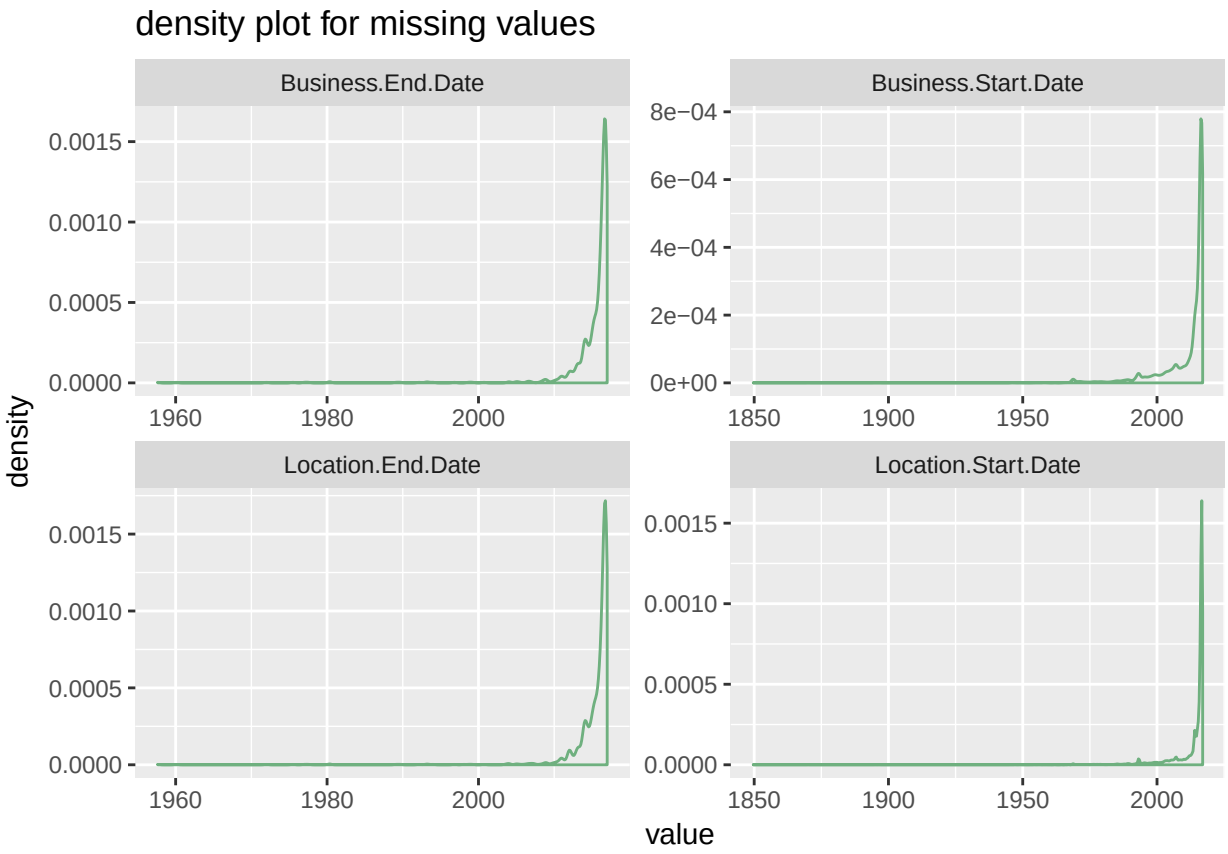
```
dim(sfbusiness_2016 %>%
  filter(Location.Start.Date < Business.Start.Date))
```

```
## [1] 458  27
```

After cleaning the dates, we could revisit the comparison and see how's the density looking now. As we could

see, the missing values are still concentrating on dates after 2010. However, it is very difficult for us to impute the location data. And therefore we need to proceed with our analysis keeping in mind that location data could produce biased results. when we are doing time and location combined analysis.

```
#plot for density plot for missing values
sfbusiness_2016 %>%
  filter(is.na(lat)) %>%
  select(Location.Start.Date,
                          Location.End.Date, Business.Start.Date,
                          Business.End.Date)%>%
gather(key = "var", value = "value") %>%
  ggplot(aes(value)) +
    geom_density(color = color_pal[3], na.rm = TRUE) +
    facet_wrap(~ var, scales = "free") +
  ggtitle("density plot for missing values")
```

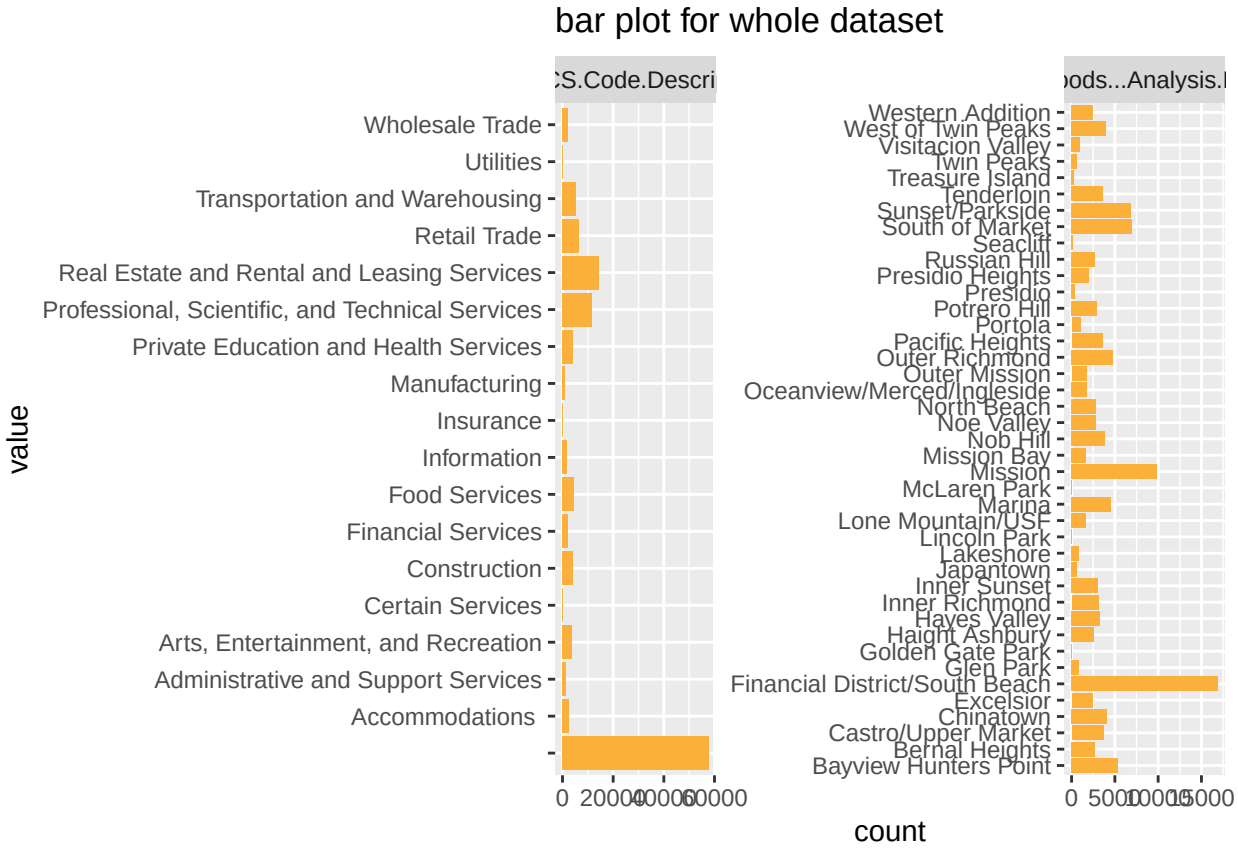


We also need to check the distribution of categorical variables to see if the missing values are related to business type or certain neighborhoods. We could see that even though the category type distributions are different, The Neighborhood distributions are almost the same. This could be a supportive indication that while the missing values might affect analysis related to time and business type, it will be relatively fair when it comes to location alone.

```
#bar plots for whole dataframe
sfbusiness_2016 %>%
  filter(Neighborhoods...Analysis.Boundaries != "") %>%
  select(NAICS.Code.Description, Neighborhoods...Analysis.Boundaries)%>%
gather(key = "var", value = "value") %>%
  ggplot(aes(value)) +
```

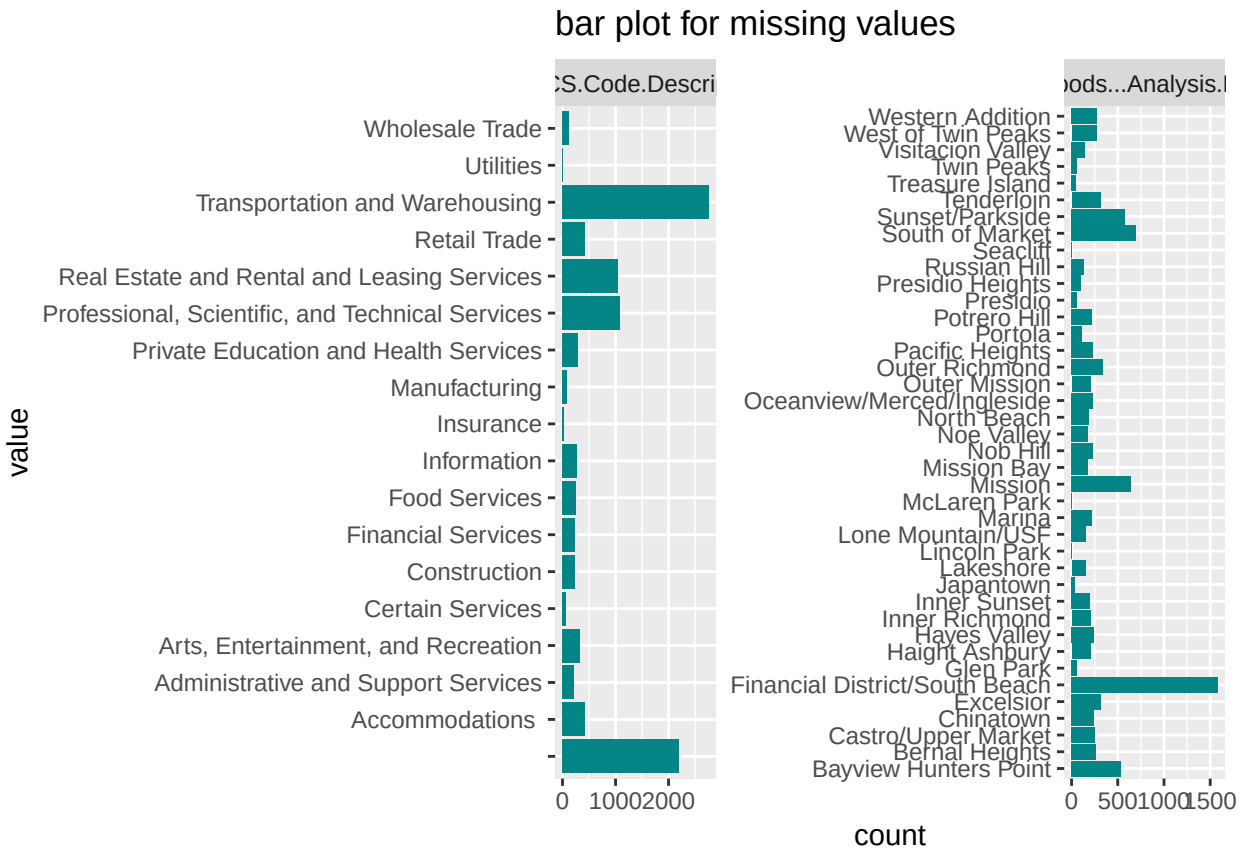
```
    geom_bar(fill= color_pal[4], na.rm = TRUE) +
    facet_wrap(~ var, scales = "free") +
  coord_flip() +
  ggtitle("bar plot for whole dataset")
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```




```
#bar plots for missing values
sfbusiness_2016 %>%
  filter(is.na(lat)) %>%
  filter(Neighborhoods...Analysis.Boundaries != "") %>%
  select(NAICS.Code.Description, Neighborhoods...Analysis.Boundaries)%>%
gather(key = "var", value = "value") %>%
  ggplot(aes(value)) +
    geom_bar(fill= color_pal[2], na.rm = TRUE) +
    facet_wrap(~ var, scales = "free") +
  coord_flip() +
  ggtitle("bar plot for missing values")
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```




```
SFbusiness %>%
  group_by(Business.Corridor) %>%
  summarise(countofc = n()) %>%
  ungroup() %>%
  arrange(desc(countofc))
```

```
## # A tibble: 11 x 2
##          Business.Corridor countofc
##                      <fctr>    <int>
##  1                          216383
##  2              Chinatown      101
##  3          Central Market       71
##  4             North Beach       47
##  5              Lower Polk       20
##  6          Lombard Street       10
##  7       Lower 24th Street        8
##  8          Mission Street        5
##  9 Fillmore Street (Lower)        2
## 10                 24th St        1
## 11           Market/Castro        1
```

Q1 Identify pockets of high concentration of active business.

Based on the assumptions I made, I identified that Downtown, Embarcadero, Financial District, South of Market and South Beach has high concentrations of active business. Also, Mission district and districts around Cow Hollow also have active business concentrations.

1) Define Pockets as location areas to identify active business

When it comes to pockets of active business, one could understand it as locations with significantly more business activities. One could also explain pockets as location in combination with business type. For example, Glen Park might be an active business area for Food Services. For this analysis, I will mainly focus on location wise analysis.

After identifying pocket as mainly location related pockets, one needs to develop metrics for active business. As we do not have the data for number of customers or daily transactions, one could only use dates to see if certain business is active, where is the business and when did it started. Based on the distribution of location start date and business operation durations(end dates minus start dates), we could develop a metric to filter out active businesses.

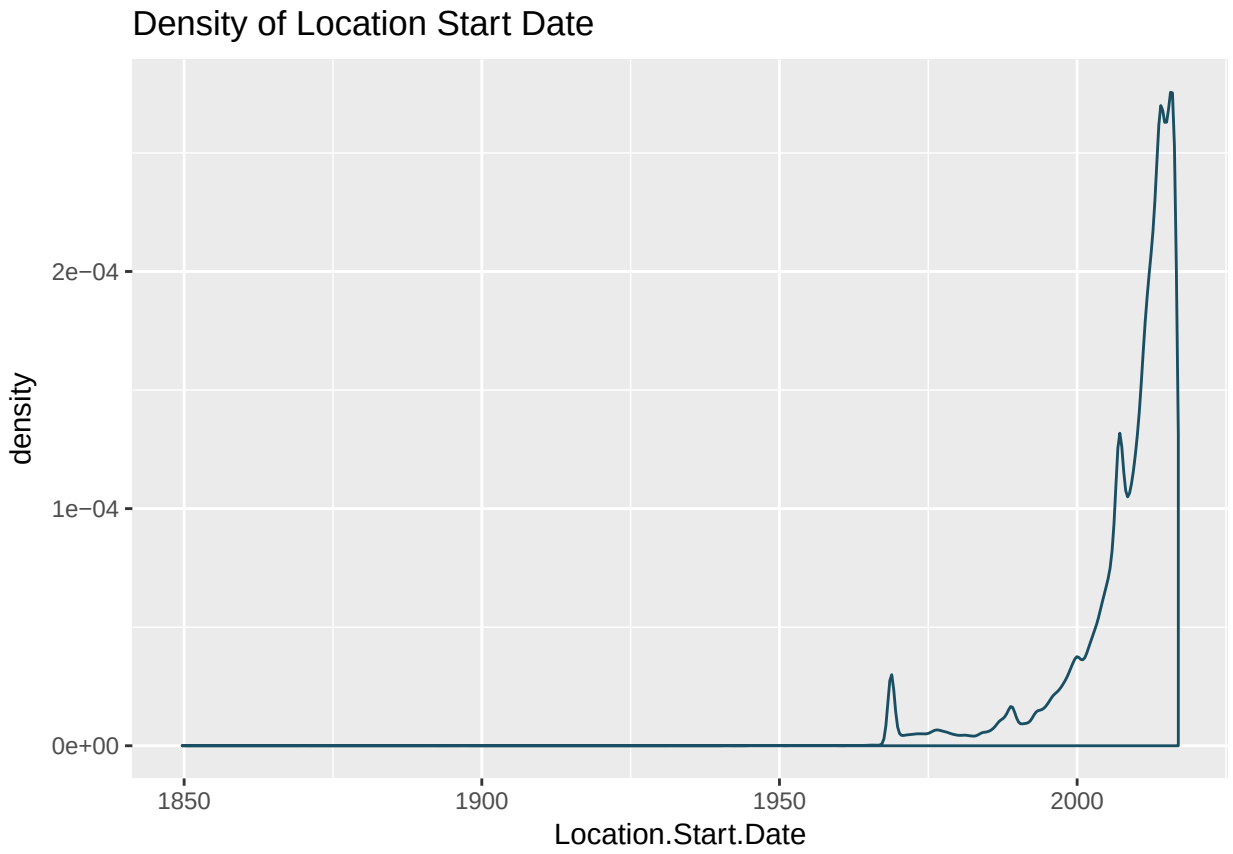
2) Utilize start dates and business duration to identify active business

I utilized start dates and business duration information within the dataset to identify active businesses as businesses opened after 2012 and not closed before 2015-09-13.

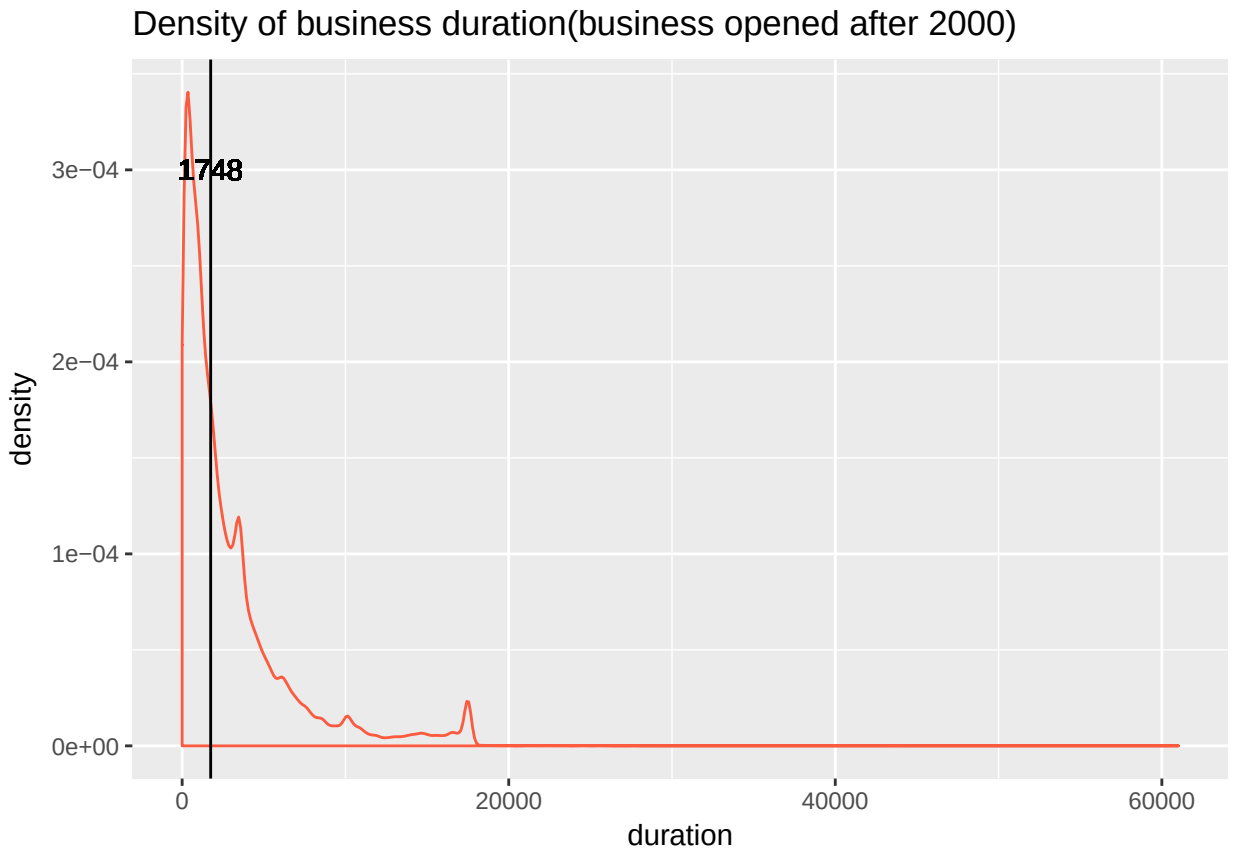
First, let's take a look at the distribution of location start date and business operation durations. As we could see from the density plot of start dates, business in SF rocketed since 2000. Also, I calculated the duration of businesses. We could see from the plot that the duration distribution is very skewed, and the median of business operation is 1748 days(three to four years).

```
#assuming no end date means open to 2016-09-13, calculate business duration
sfbusiness_2016$Location.End.Date[is.na(sfbusiness_2016$Location.End.Date)] <- ymd("2016-09-13")
sfbusiness_2016$duration <- as.numeric(sfbusiness_2016$Location.End.Date -
                                       sfbusiness_2016$Location.Start.Date)
```

```
#density of location start date
sfbusiness_2016 %>% ggplot()+
  geom_density(aes(Location.Start.Date), color = color_pal[1]) +
  ggtitle("Density of Location Start Date")
```




```
sfbusiness_2016 %>%
  filter(duration > 0) %>%
  ggplot() +
  geom_density(aes(duration), color = color_pal[5]) +
  geom_vline(aes(xintercept = median(duration))) +
  geom_text(aes(label = median(duration), y=3e-04, x = median(duration))) +
  ggtitle("Density of business duration(business opened after 2000)")
```



While businesses opened in 2012 with business duration equal to median days(1,748 days) could still indicate active business activities. Businesses opened in 2000 and closed four or five years afterwards might not be indicative because time has passed and these data would produce noise for the result.

My assumption is that business opened 1,748 days before the data collection date and closed previous to 2015-09-13 is not of interest to our analysis. Also, businesses that opened a long time ago whether standing or not also has limited power to indicate activeness of business operations. Shops that opened for many years and stayed at the same place indicates more about stability instead of growth or change.

Based on the above assumption, I filtered out businesses identified as non-active just to focus on high-growth areas. And then I plotted both active business versus active business & long-standing business on the map.

3) Visualization and Conclusion

As we could see from the visualization of active business, Downtown, Embarcadero, Financial District South of Market and South Beach has a high concentration of active business. Mission district and districts around Cow Hollow are relatively more active than other areas.

However, as we mentioned from previous missing values analysis, the latitudes and longitudes data has more missing values for years after 2010. So there are possibilities that we did not capture all active business areas. So I also visualize the active business with the long-standing ones. As we could see from the second visualizations with both active and long-standing businesses, areas like North Beach and Lower Height has more long-standing business. These areas could also be active, and we did not identify these as their location information are missing from the dataset.

```
activebusiness <- sfbusiness_2016 %>%
  filter(!is.na(lat))%>%
  filter(Location.End.Date > '2015-09-13' & Location.Start.Date > (ymd('2016-09-13') - 1748))
```

```
activebusiness2 <- sfbusiness_2016 %>%
  filter(!is.na(lat))%>%
  filter(Location.End.Date > '2015-09-13')
```

```
sf_road_map <- qmap("San Francisco", zoom=13, source = "google", maptype="roadmap")
```

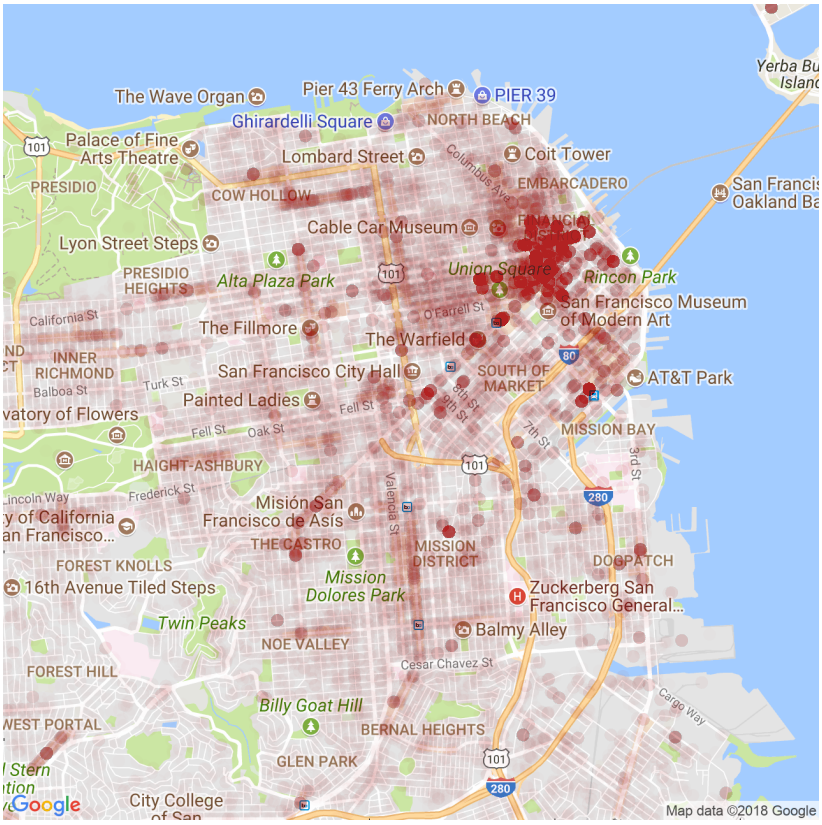
```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=San+Francisco&zoom=13&size=640x6
```

```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=San%20Francisco&sens
```

```
## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```

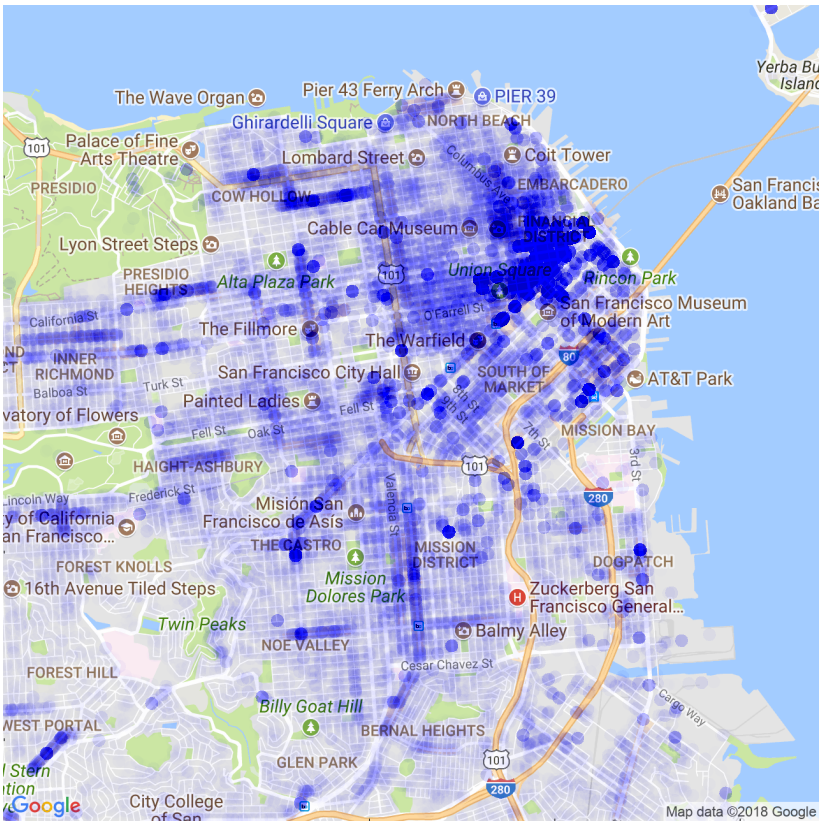
```
sf_road_map + geom_point(aes(x=lon, y=lat),
                         data = activebusiness,alpha = 0.01,color = "firebrick", na.rm = TRUE) +
  ggtitle("active business pockets in SF")
```

active business pockets in SF




```
sf_road_map + geom_point(aes(x=lon, y=lat),
                         data = activebusiness2,alpha = 0.007,color = "blue", na.rm = TRUE) +
  ggtitle("active and long-standing business in SF")
```

active and long−standing business in SF

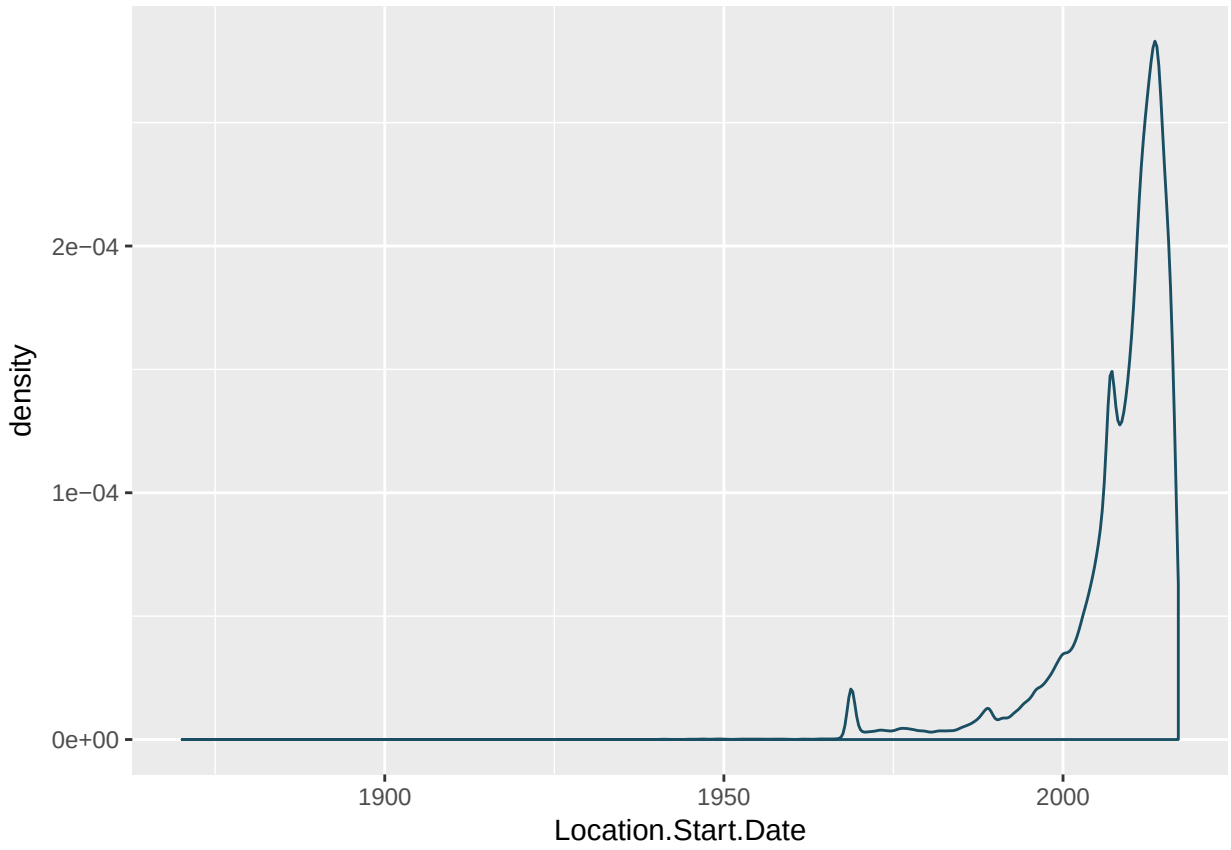


Industry Type trend over the Years

1) Assessing the impact of missing NACIS codes

As we could see from the previous analysis, lots of businesses do not have NAICS code. I used density plot again to examine the data points without a NACIS code. As we could see that the following graph is very similar compared to the ones we saw for the whole dataset. Therefore, we need not worry that the trend analysis might be biased when combined with location start dates.

```
sfbusiness_2016 %>%
  mutate(NAICS = paste(NAICS.Code, NAICS.Code.Description, sep = ",")) %>%
  select(NAICS, Location.Start.Date)%>%
  filter(NAICS == ",") %>%
  ggplot() +
  geom_density(aes(x = Location.Start.Date), color = color_pal[1])
```



2) Identifying business trends across time using location start dates

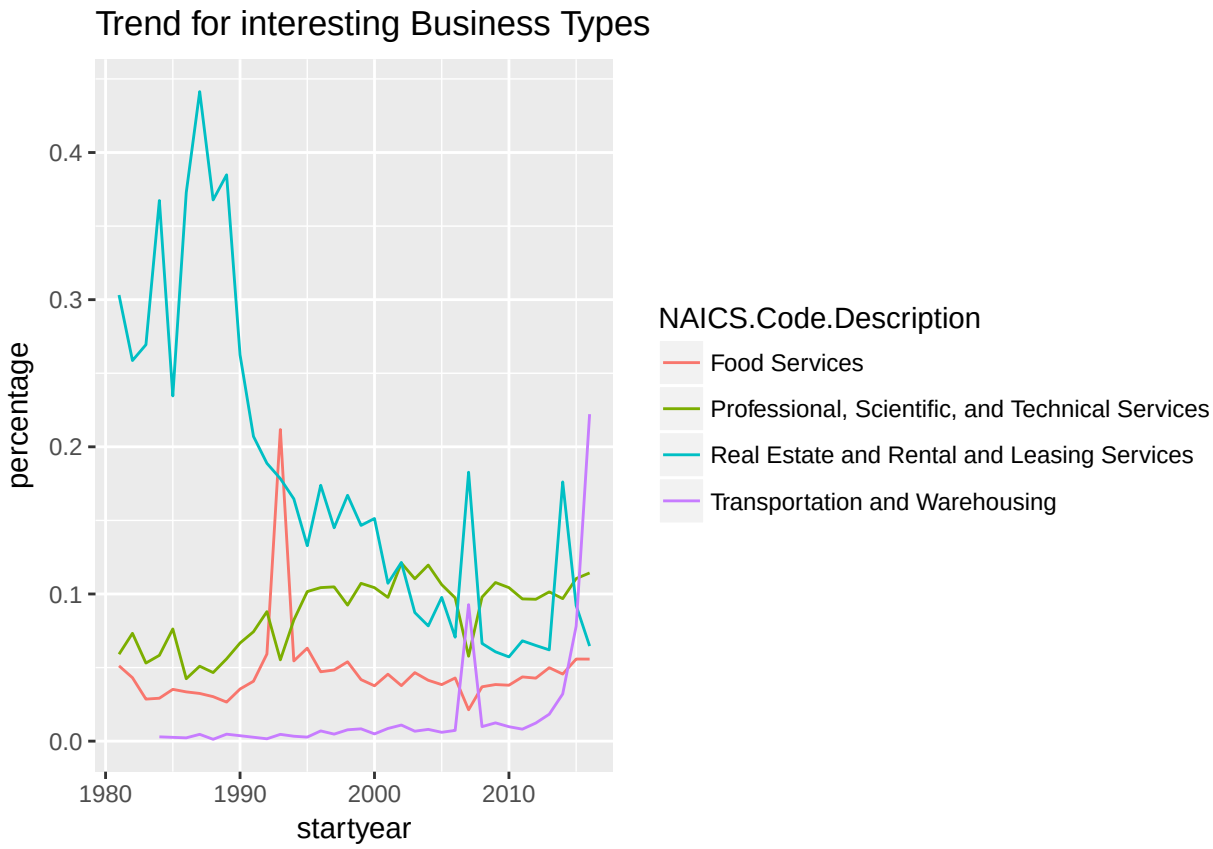
To study the trend of different business types, we could calculate the percentage of certain business type started within a specific year. For example, we could calculate the percentage of Food Services for 2010 by dividing the count of food services started that year by total business started. In this way, we could scale the data and see the trend of percentage for each business type.

We could see from the below graph that: a. New businesses in Real Estate has been declining over the years. However, this industry has two spikes at around 2007-2008 and 2013-2014 b. New businesses in Transportation and Warehousing has seen rapid growth in recent years. And it's latest spike correlates with real estate spike. c. Food Services saw a huge jump during early 1990 and seems to enjoy a steady growth for the past few years d. Professional, Scientific, and Technical Services have seen continous strong new businesses over the past then years.

```
sfbusiness_2016 %>%
  mutate(startyear = year(Location.Start.Date))%>%
  filter(startyear > 1980) %>%
  group_by(startyear, NAICS.Code.Description) %>%
  summarise(business_type = n()) %>%
  left_join(sfbusiness_2016 %>%
              mutate(startyear = year(Location.Start.Date)) %>%
              filter(startyear > 1980) %>%
              group_by(startyear) %>%
              summarise(ttl_business = n(), by = "startyear")) %>%
  mutate(percentage = business_type /ttl_business ) %>%
  select(startyear, NAICS.Code.Description, percentage) %>%
  filter(NAICS.Code.Description != "") %>%
```

```
  filter(NAICS.Code.Description == 'Transportation and Warehousing'
         | NAICS.Code.Description == 'Real Estate and Rental and Leasing Services'|
           NAICS.Code.Description == 'Food Services'|
           NAICS.Code.Description == 'Professional, Scientific, and Technical Services') %>%
  ggplot()+
  geom_line(aes(x = startyear, y = percentage, color = NAICS.Code.Description)) +
  ggtitle("Trend for interesting Business Types")
```

```
## Joining, by = "startyear"
```




```
sfbusiness_2016 %>%
  mutate(startyear = year(Location.Start.Date))%>%
  filter(startyear > 1980) %>%
  group_by(startyear, NAICS.Code.Description) %>%
  summarise(business_type = n()) %>%
  left_join(sfbusiness_2016 %>%
              mutate(startyear = year(Location.Start.Date)) %>%
              filter(startyear > 1980) %>%
              group_by(startyear) %>%
              summarise(ttl_business = n(), by = "startyear")) %>%
  mutate(percentage = business_type /ttl_business ) %>%
  select(startyear, NAICS.Code.Description, percentage) %>%
  filter(NAICS.Code.Description != "") %>%
  ggplot()+
  geom_line(aes(x = startyear, y = percentage, color = NAICS.Code.Description)) +
  ggtitle("Trend for all Business Types")
```

```
## Joining, by = "startyear"
```