Final Report

Tianying Xu

2018/12/15

I. Abstract

This report aims to find out whether lot area follows Benford's law or not, what factors may affect this deviation against benford's law. First, I test if lot area data follow Benford's law. Then, I do Exploratory Data Analysis (EDA) to find out what similar features "suspects" all have. Finally, I created plot to get a whole idea of suspects and get to conclusion that what may cause data to deviate against Benfrod's law.

II. Introduction

Houses are what we need to live every day, and the lot area of every house varies a lot. According to Benford's law, which is also called first-digit law, leading digits in many real-life data follow a distribution. Thus, I would like to test if lot area data follows Bonford's law, and if not, what factors may be associated with the deviation.

III. Method

A. Data Source

The data is published on kaggle.com, where it is a competition to predict house price using lot area, lot shape and other predictors (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data).

B. Analysis

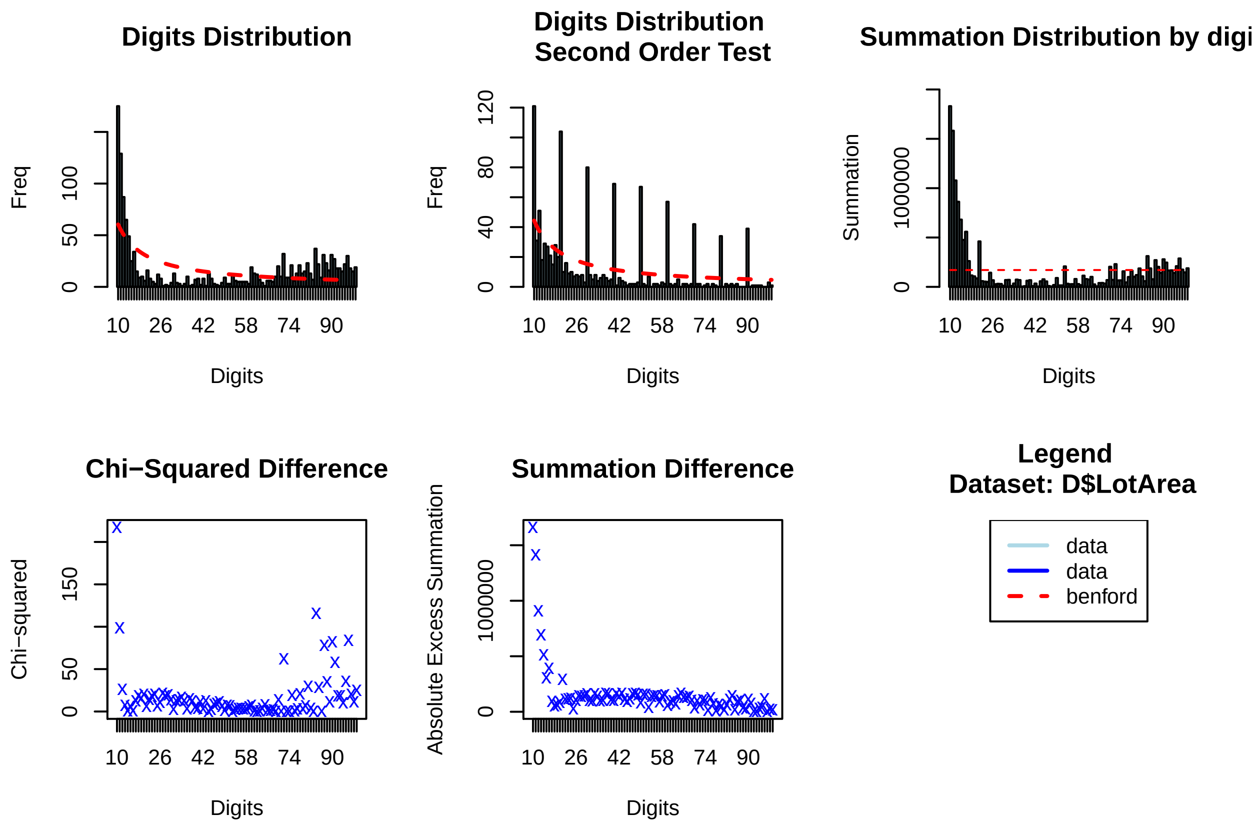
1. Benford's law

I conduct a test to lot area data to see if it follows Beonford's law. The result is as following:

```
D <- read.csv("train.csv",header=T)
#Lot Area
bfd_area <- benford(D$LotArea)
bfd_area
```

```
##
## Benford object:
##
## Data: D$LotArea
## Number of observations used = 1460
## Number of obs. for second order = 1072
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic Value
##         Mean  0.53
##          Var  0.15
##  Ex.Kurtosis -1.76
##     Skewness -0.14
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1      10         114.57
## 2      11          73.83
## 3      12          36.25
## 4      84          29.50
## 5      18          25.28
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  D$LotArea
## X-squared = 1588.5, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  D$LotArea
## L2 = 0.33858, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.009158604
## Distortion Factor: 21.85242
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

```r
plot(bfd_area)
```



From the plots of Benford analysis of the lot area data, we can discover several features. From the first plot and the benford analysis, it is evident that digits from "10"-"15" and after "70" have deviations from the Benford Law. Among which, digit "10" and "11" have the largest deviation against benford law. From the second plot, we can see that the structure of data is slightly deviated from the benford law. The last two plot shows similar result that digits "10" and "11" have the largest deviation against Benford's law.

2. Similar Features

First I created a table to get a whole idea of lot area of "suspects". Then, I classify variables into severall categories. First is outsider living conditions, including zoning, building type and linear feet of street connected to property. Then is inside living condition, which contains garage type, screen porch area, total rooms above grade and quality. Finally, I also check on the time and geography feature.

```
D_spts <- getSuspects(bfd_area,D)
summary(D_spts$LotArea)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10000   10400   10800   11230   11342  115149
```

From the table, we can see that lot area among "suspects" are within range from 10,000 to 12,000, since lot area which begins with digits "10" and "11" have most severe deviation against Benford's law.

1) Outside Living Conditions

a. Zoning

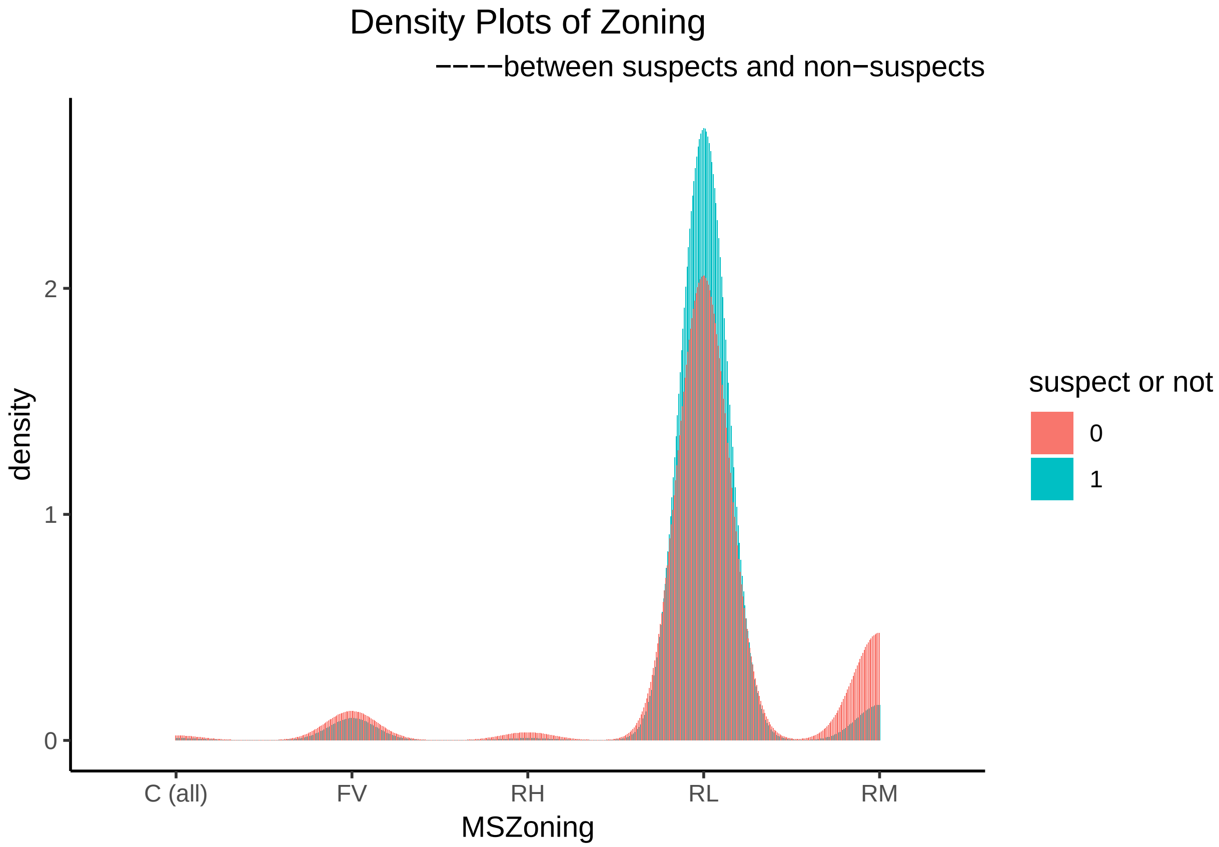
The variable "MSZoning" identifies the general zoning classification of the sale. The values are as following:

```
A    -- Agriculture
C    -- Commercial
FV   -- Floating Village Residential
I    -- Industrial
RH   -- Residential High Density
RL   -- Residential Low Density
RP   -- Residential Low Density Park
RM   -- Residential Medium Density
```

Since I want to find out th esimilar features of "suspects", I create density plots between suspects and non-suspects.

```
# get suspects
suspects_area <- getSuspects(bfd_area, D,how.many = 2)

# density plot of suspects & non-suspects
D1 <- D %>%
  mutate(ifsuspect=ifelse(Id %in% suspects_area$Id,1,0))
ggplot(D1, aes(x=MSZoning,fill=factor(ifsuspect),group=factor(ifsuspect)))+
  geom_histogram(stat="density",position="dodge") +
  guides(fill=guide_legend(title="suspect or not"))+
  theme_classic()+
  ggtitle("Density Plots of Zoning",subtitle = "----between suspects and non-suspects")+
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 1))
```

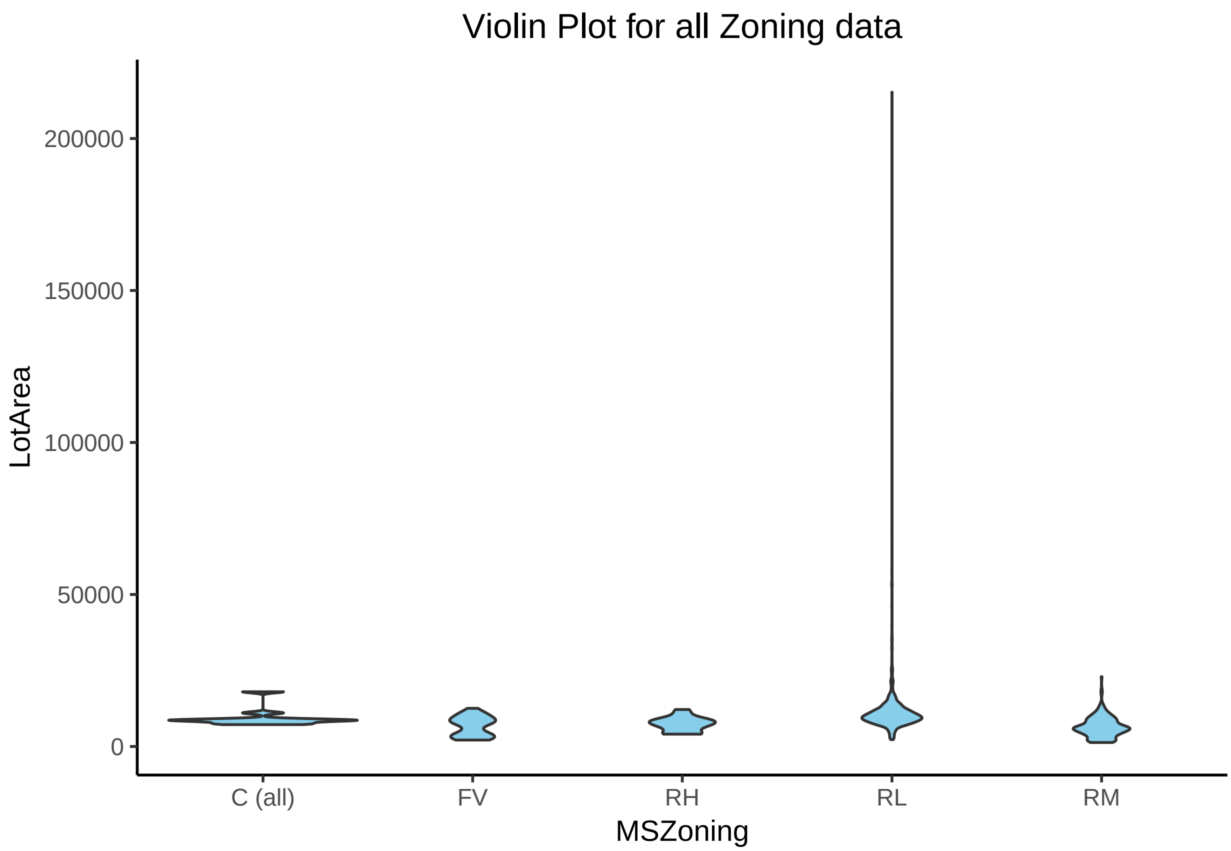
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



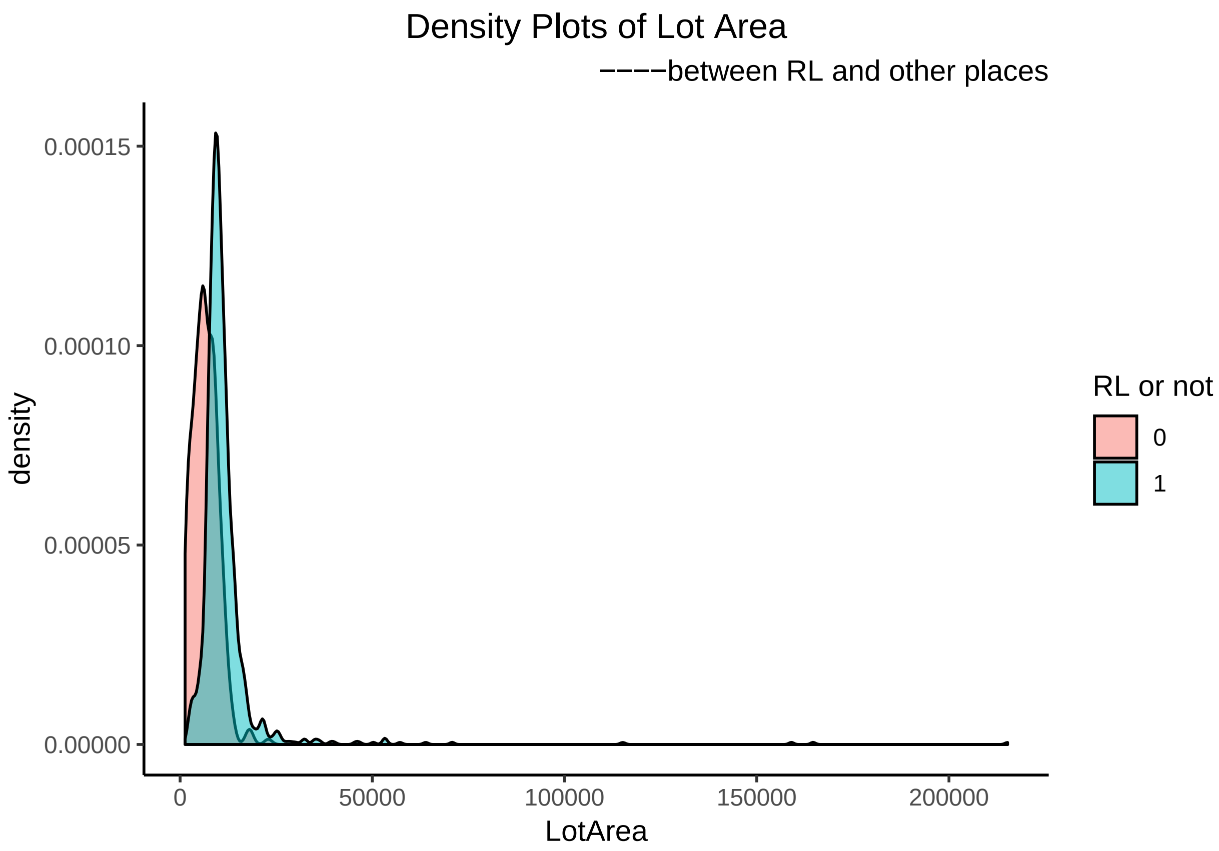
```r
# RL more dense
```

The density plot here sums area to 1, thus there may exist numbers on y axis that are larger than 1. From the plot, it is evident that "RL" has more density among "suspects" than non-suspects. "RL" means "Residential Low Density", therefore I guess lot area among suspects are larger than in other places. I draw a violin plot, a density plot and a bar plot for this:

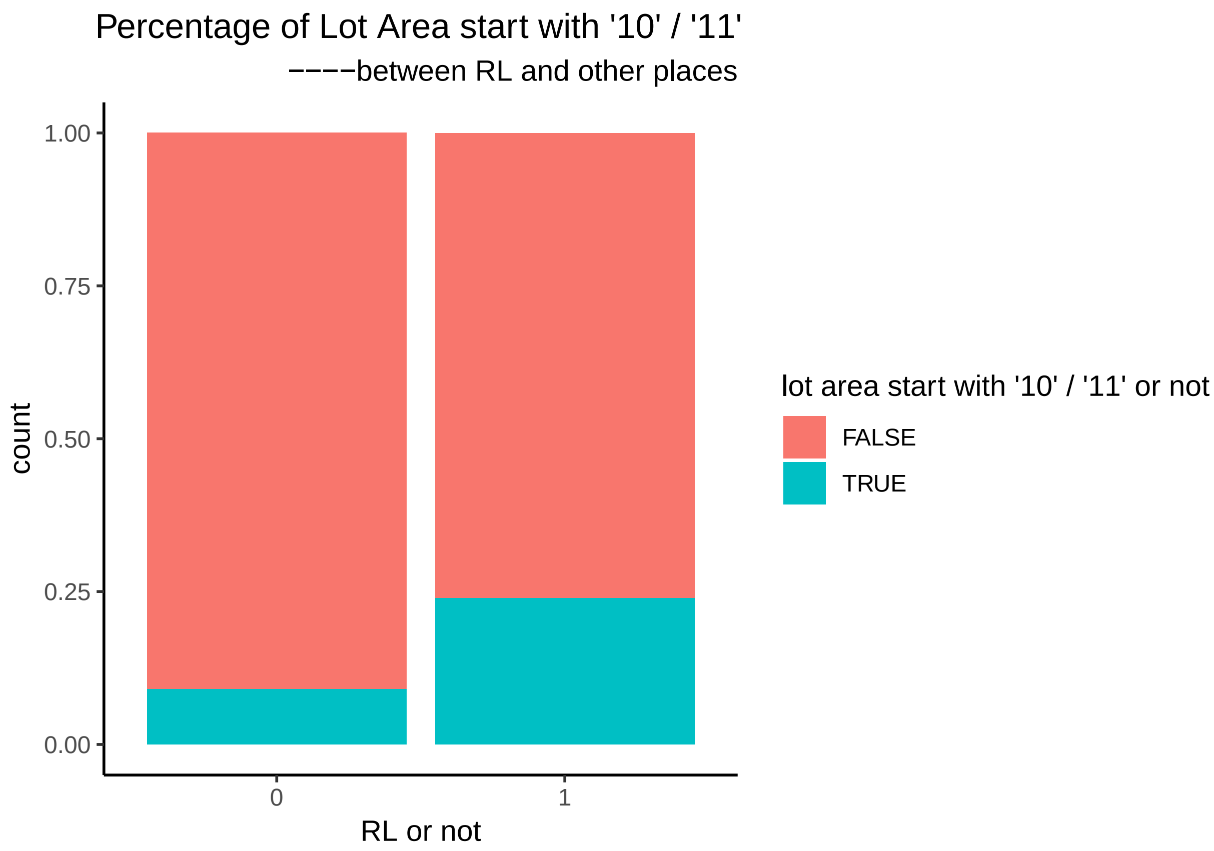
```r
# violin plot
ggplot(data = D) +
  aes(x = MSZoning, y = LotArea) +
  geom_violin(scale = "area", adjust = 1, fill = "skyblue")+
  theme_classic()+
  ggtitle("Violin Plot for all Zoning data")+
  theme(plot.title = element_text(hjust = 0.5))
```



```r
# density plot between RL and other places
D_RL <- D %>% mutate(RL=ifelse(MSZoning=="RL",1,0))
ggplot(D_RL,aes(x=LotArea,y=..density..,fill=factor(RL),group=factor(RL)))+
  geom_density(alpha=0.5)+
  guides(fill=guide_legend(title="RL or not"))+
  theme_classic()+
  ggtitle("Density Plots of Lot Area",subtitle = "----between RL and other places")+
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 1))
```



```
# percentage of lot area start with "10"/"11" between RL and others
D_RL_1011 <- D_RL %>%
  mutate(s1011=str_detect(LotArea,"^1[01]"))
ggplot(D_RL_1011,mapping=aes(x=factor(RL),fill=factor(s1011)))+
  geom_bar(position="fill")+
  guides(fill=guide_legend(title="lot area start with '10' / '11' or not"))+
  xlab("RL or not")+
  theme_classic()+
  ggtitle("Percentage of Lot Area start with '10' / '11'", subtitle = "----between RL and other places"
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 1))
```



From the violin plot, we can see that lot area in "RL" has really extreme data point, and it is clear in the density plot that lot area in "RL" is muach larger than those in other zonings in general. This makes sense since lot area tend to be larger in places of low density of residence. Also, the last plot tells us exactly that percentage of lot area beiginning with digits "10" or "11" is much higher in "RL" than in other places. Hence, the deviation against Benford's law may be because that places where these data is collected contain more "RL" than average, which lead to high proportion of "RL", and then lead to higher proportion of lot area start with digits "10" or "11", and finally lead to deviation.

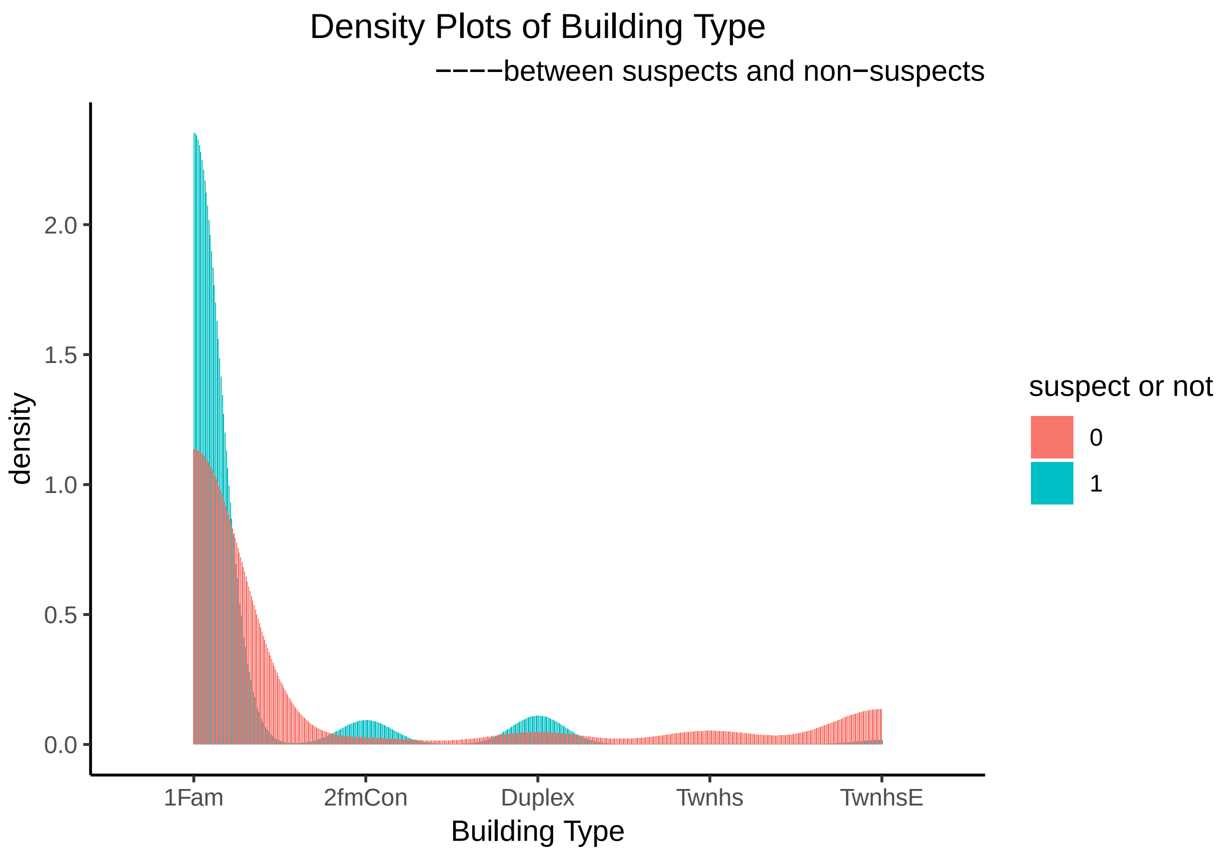
b. Building Type

The variable "BldgType" means the type of dwelling. The values are:

```
1Fam -- Single-family Detached
2FmCon   -- Two-family Conversion; originally built as one-family dwelling
Duplx    -- Duplex
TwnhsE   -- Townhouse End Unit
TwnhsI   -- Townhouse Inside Unit
```

```r
ggplot(D1, aes(x=BldgType,fill=factor(ifsuspect),group=factor(ifsuspect)))+
  geom_histogram(stat="density",position="dodge")+
  guides(fill=guide_legend(title="suspect or not"))+
  xlab("Building Type")+
  theme_classic()+
  ggtitle("Density Plots of Building Type",subtitle = "----between suspects and non-suspects")+
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 1))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



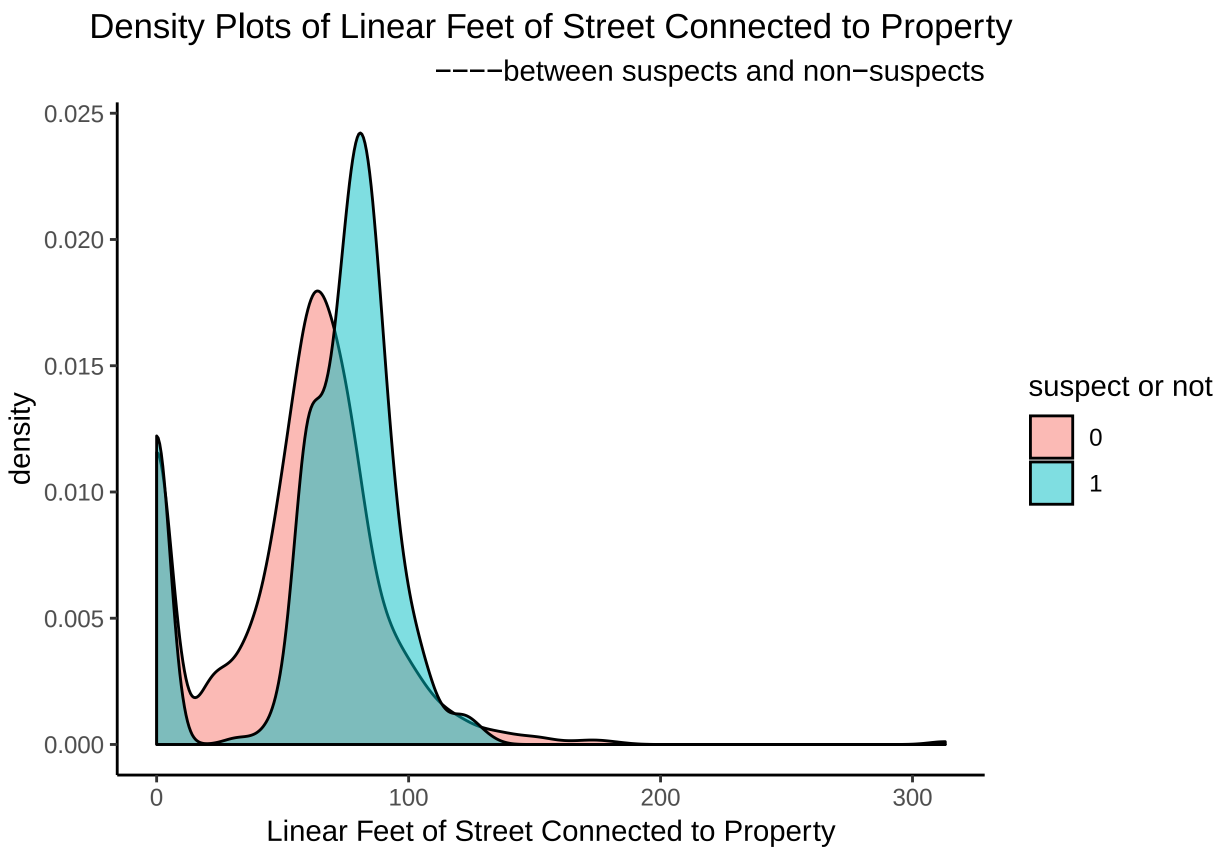
The plot shows evidently that "1Fam" is more popular among suspects than among non-suspects, which means another similar feature of suspects is that their building type is "Single-family Detached". This can be caused by previous result that there are high proportion of places with low residential density. Places with low residential density will have more houses where building type is "Single-family Detached", "Two-family Conversion; originally built as one-family dwelling" or "Duplex" than places with high residential density.

c. Linear Feet of Street Connected to Property

```
D2 <- D1
D2[is.na(D2)] <- 0
```

```
## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated
## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated

## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level,
## NA generated
```

```r
ggplot(D2,aes(x=LotFrontage,y=..density..,fill=factor(ifsuspect),group=factor(ifsuspect)))+
  geom_density(alpha=0.5)+
  guides(fill=guide_legend(title="suspect or not"))+
  xlab("Linear Feet of Street Connected to Property")+
  theme_classic()+
  ggtitle("Density Plots of Linear Feet of Street Connected to Property",subtitle = "----between suspec
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 1))
```



```
# D_lotf <- D %>%
#   mutate(s1011=str_detect(LotArea,"^1[01]"))
# ggplot(D_lotf,mapping=aes(x=factor(LotFrontage),fill=factor(s1011)))+
#   geom_bar(position="fill")+
#   theme(axis.text.x = element_text(angle=45))
```

The plot clearly shows that "suspects" have longer street connected to property in general than non-suspects. Most suspects has 100 feet long street connected to their property, which may also lead to houses with similar lot area. Thus, this could be a factor that there are many houses with lot area within a range from 10,000 to 12,000.

To sum up, places where the data is collected may contain high proportion of low residential density area, where houses are mostly single-family detached, have similar length of street connected and are with similar lot area within 10,000 to 12,000, thus lead to "suspects" with lot area number beginning with "10" or "11".

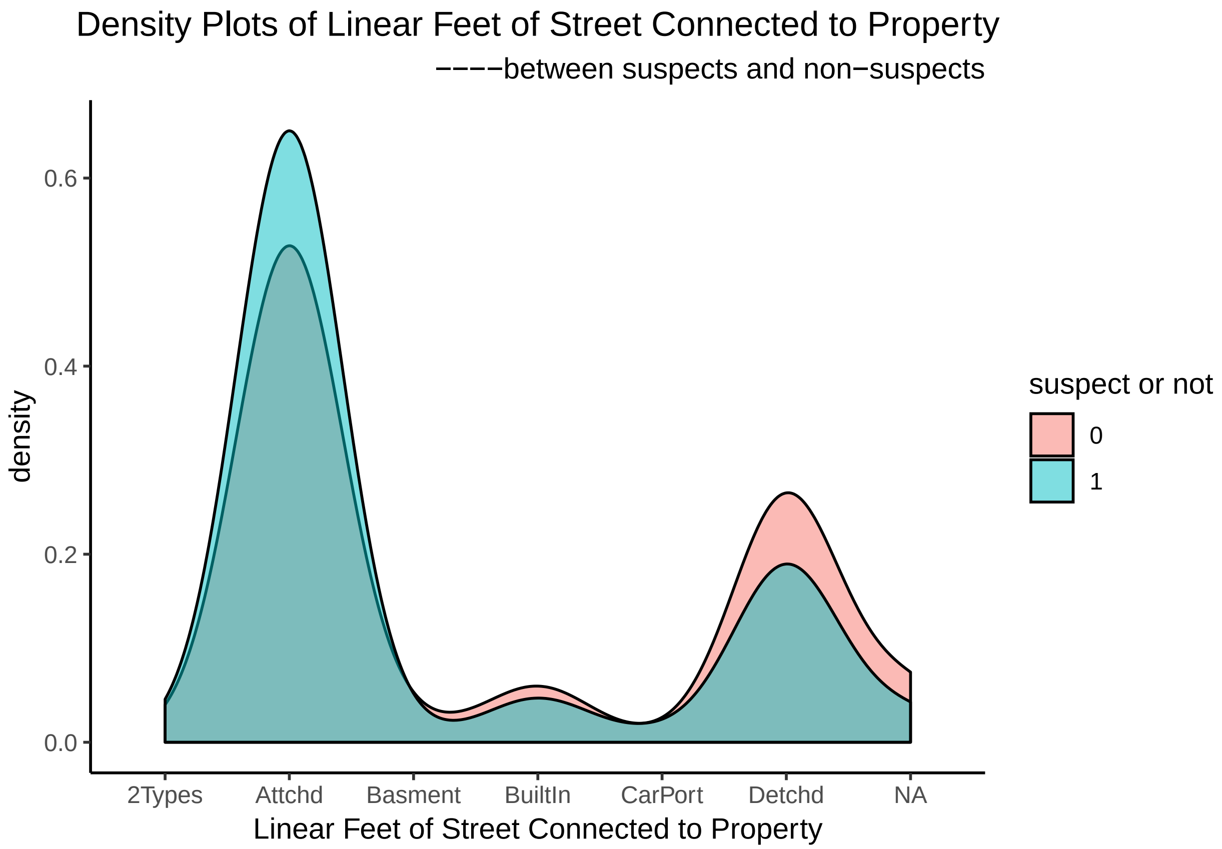
2) Inside Living Conditions

a. Garage Type

The variable "GarageType" represent the garage location. The values are:

```
2Types -- More than one type of garage
Attchd -- Attached to home
Basment -- Basement Garage
BuiltIn -- Built-In (Garage part of house - typically has room above garage)
CarPort -- Car Port
Detchd -- Detached from home
NA -- No Garage
```

```
ggplot(D1, aes(x=GarageType,y=..density..,fill=factor(ifsuspect),group=factor(ifsuspect)))+
  geom_density(alpha=0.5)+
  guides(fill=guide_legend(title="suspect or not"))+
  xlab("Linear Feet of Street Connected to Property")+
  theme_classic()+
  ggtitle("Density Plots of Linear Feet of Street Connected to Property",subtitle = "----between suspec
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 1))
```

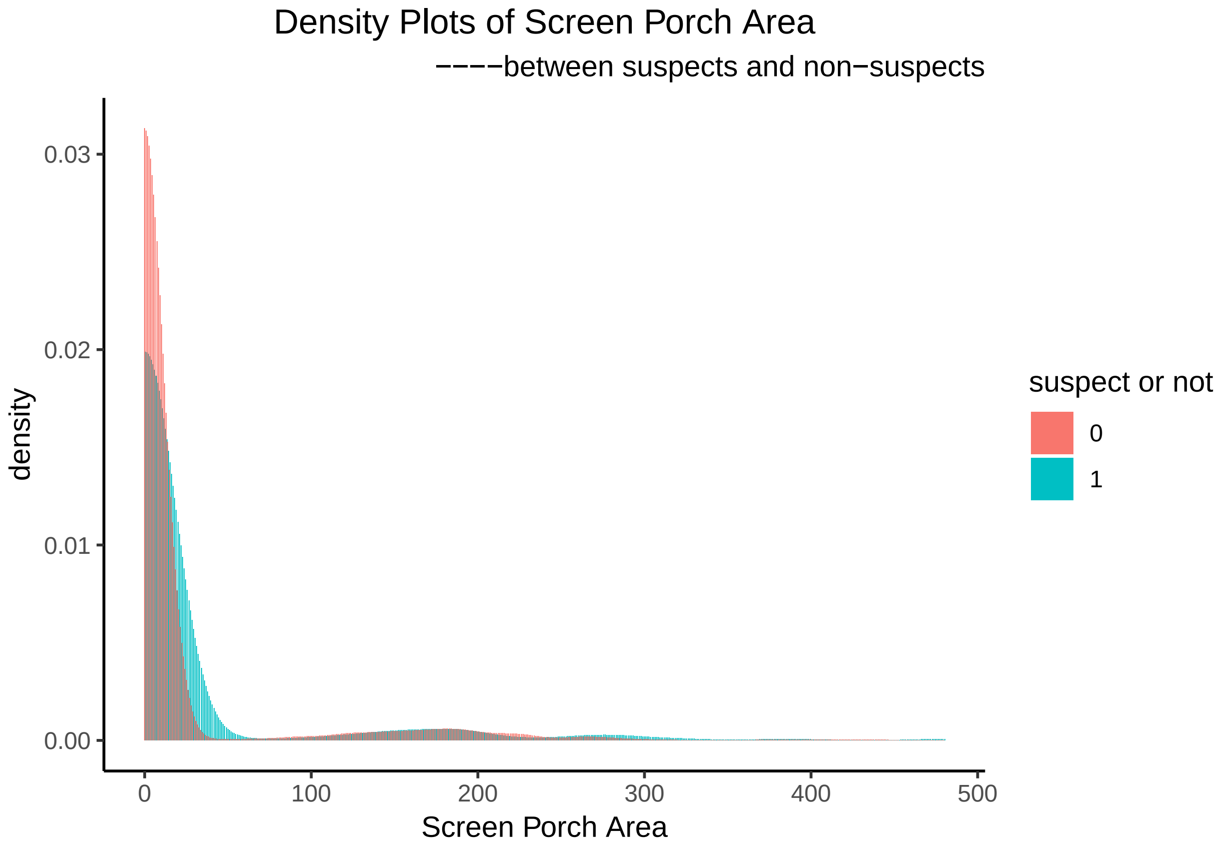


It is evident from the plot that "suspects" usually have more garage attached than others. This makes sense in that houses with average or larger lot area tend to have garage attached.

b. Screen Porch Area

```
ggplot(D1, aes(x=ScreenPorch,fill=factor(ifsuspect),group=factor(ifsuspect)))+
  geom_histogram(stat="density",position="dodge")+
  guides(fill=guide_legend(title="suspect or not"))+
  xlab("Screen Porch Area")+
  theme_classic()+
  ggtitle("Density Plots of Screen Porch Area",subtitle = "----between suspects and non-suspects")+
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 1))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

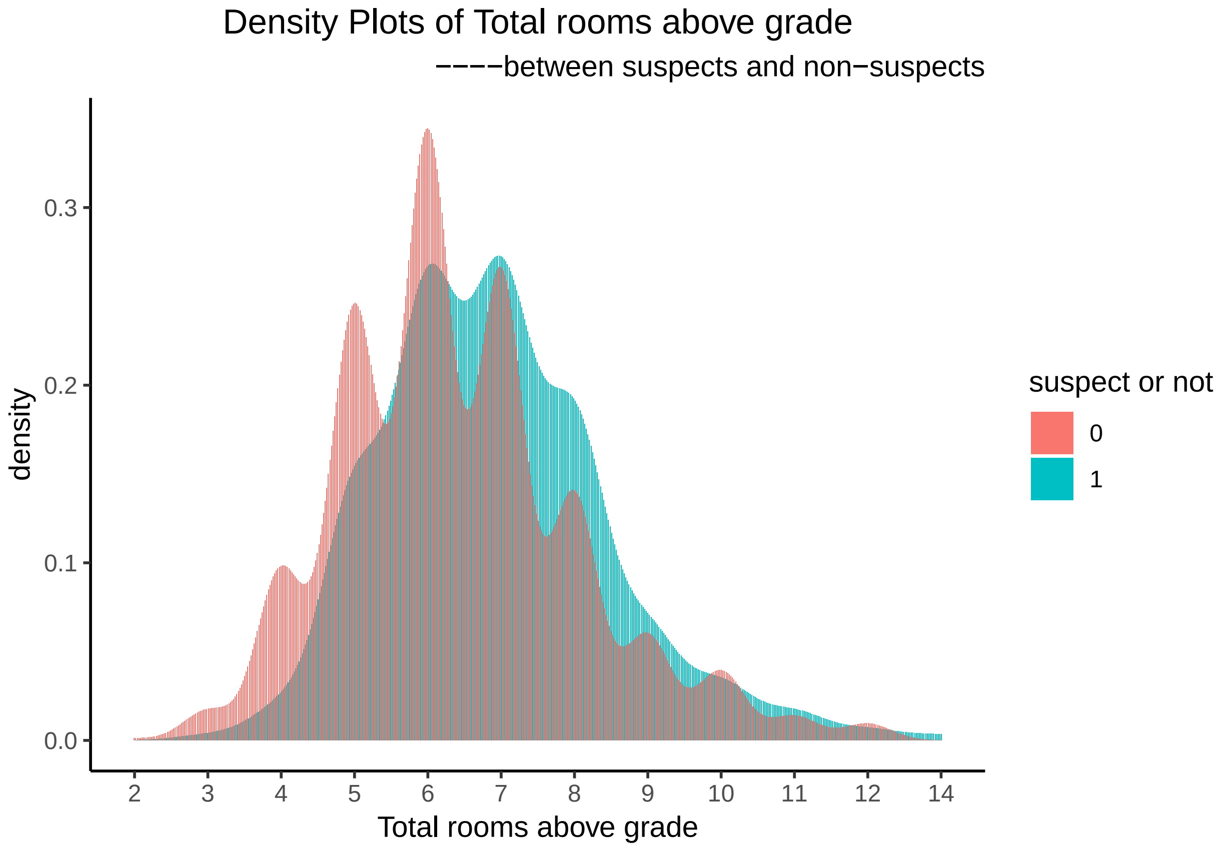


From the plot, we can see that screen porch area among “suspects” are more scattered, varing from 0 too almost 500 square feet, while among non-suspects, the screen porch area are all under 300 square feet. It is understandable that houses with lot area between 10,000 and 12,000 will have larger screen porch area.

c. Total Rooms Above Grade (does not include bathrooms)

```
ggplot(D1, aes(x=factor(TotRmsAbvGrd),fill=factor(ifsuspect),group=factor(ifsuspect)))+
  geom_histogram(stat="density",position="dodge")+
  guides(fill=guide_legend(title="suspect or not"))+
  xlab("Total rooms above grade")+
  theme_classic()+
  ggtitle("Density Plots of Total rooms above grade",subtitle = "----between suspects and non-suspects"
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 1))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

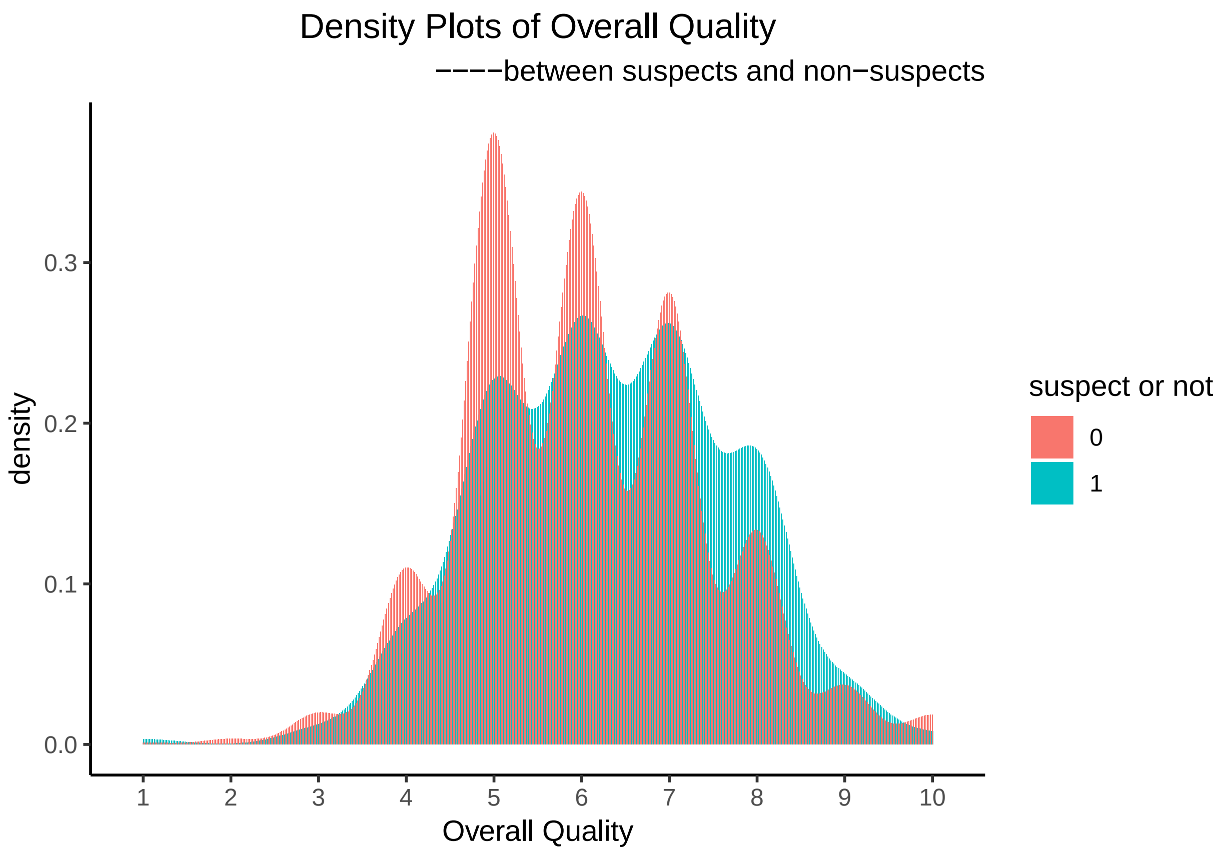


This plot shows that there are generally more rooms above grade among "suspects" than others. This makes sense since houses with lot area between 10,000 and 12,000 will tend to have more rooms than the others.

d. Quatlity

```
#quality
ggplot(D1, aes(x=factor(OverallQual),fill=factor(ifsuspect),group=factor(ifsuspect)))+
  geom_histogram(stat="density",position="dodge")+
  guides(fill=guide_legend(title="suspect or not"))+
  xlab("Overall Quality")+
  theme_classic()+
  ggtitle("Density Plots of Overall Quality",subtitle = "----between suspects and non-suspects")+
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 1))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



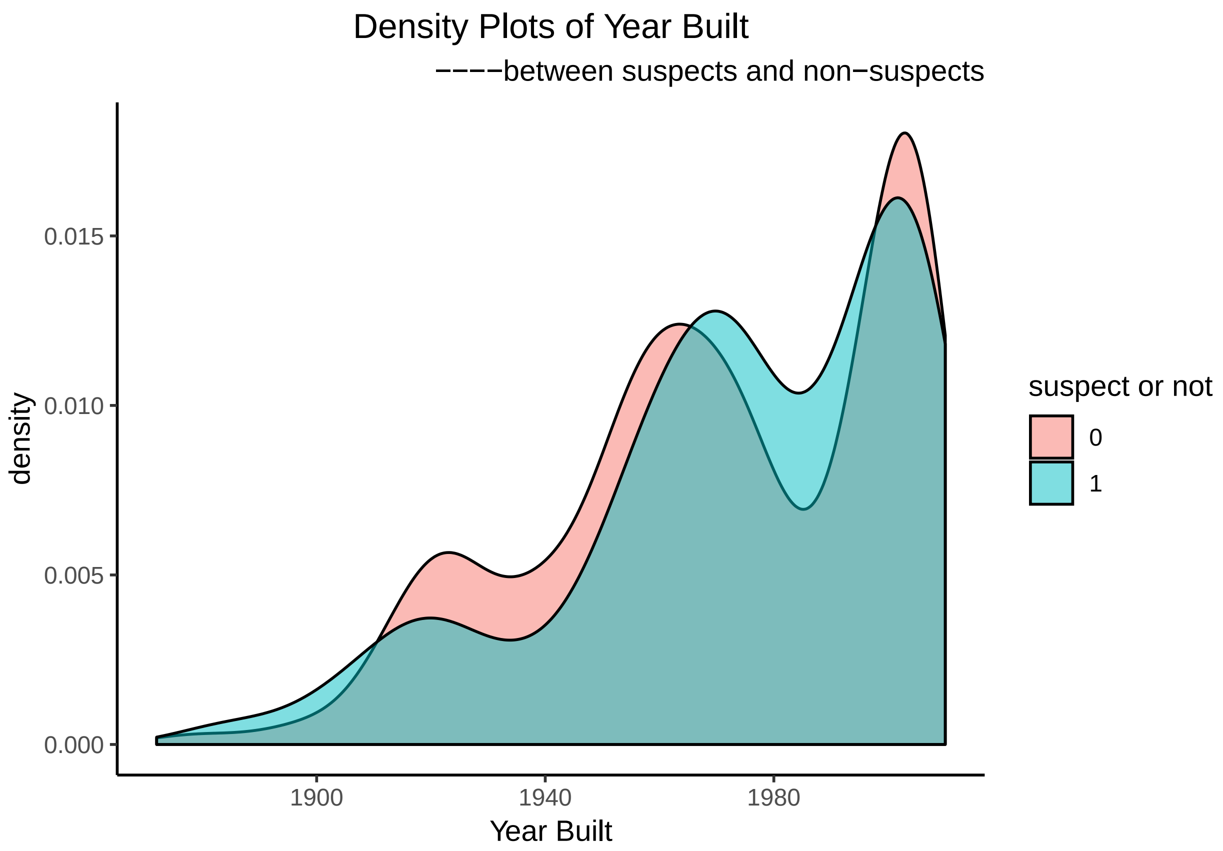
The plot of overall quality indicates that "suspects" tend to have higher overall quality than others. This may due to the large lot area, attached garage, larger creen porch and more rooms.

In terms of inside living conditions, "suspects" tend to have more attached garage, larger screen porch, more rooms and higher quality, and these may all dut to larger mean lot area of "suspects".

3) Time

I also check the time factor for "suspects".

```
# Density Plots of Year Built
ggplot(D1, aes(x=YearBuilt,y=..density..,fill=factor(ifsuspect),group=factor(ifsuspect)))+
  geom_density(alpha=0.5)+
  guides(fill=guide_legend(title="suspect or not"))+
  xlab("Year Built")+
  theme_classic()+
  ggtitle("Density Plots of Year Built",subtitle = "----between suspects and non-suspects")+
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 1))
```

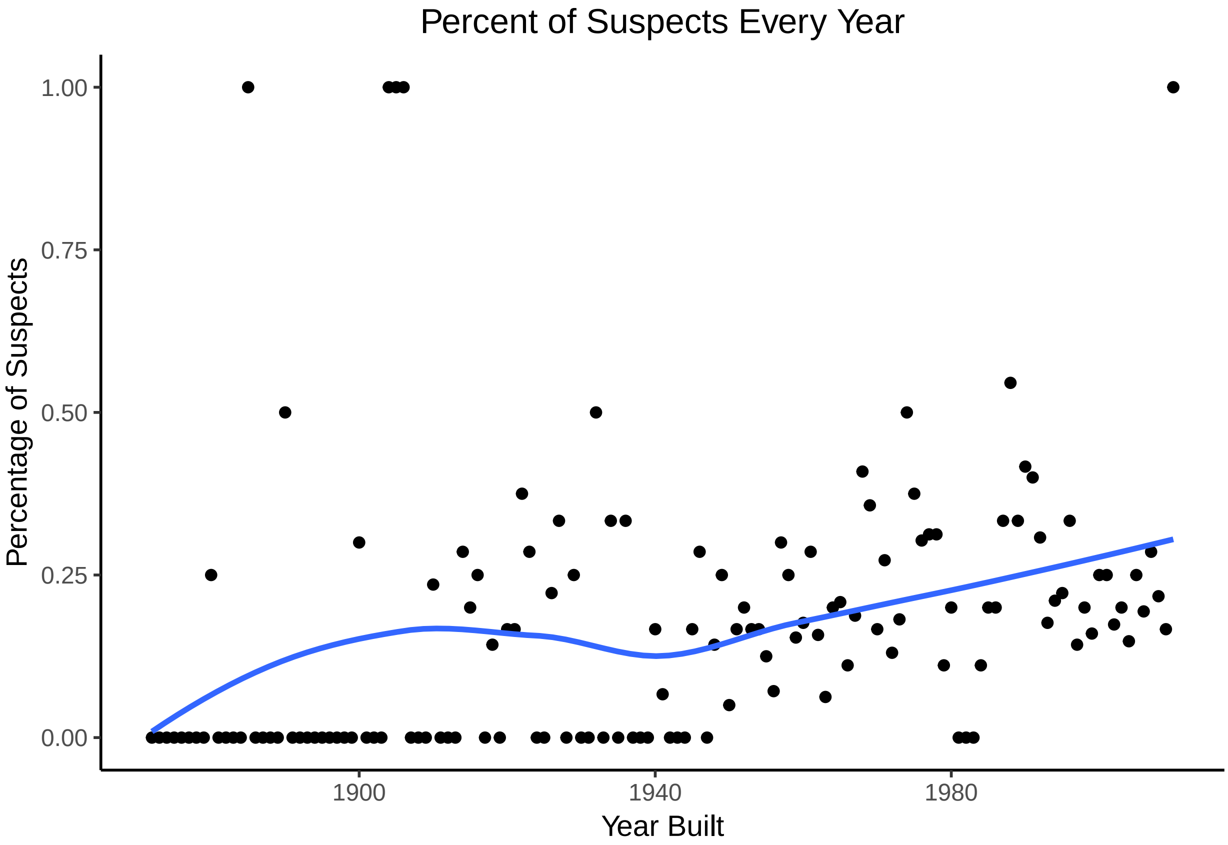


```r
# Data for plot
D_sus <- D1 %>%
  filter(ifsuspect=="1") %>%
  group_by(YearBuilt) %>%
  summarise(n=n())
D2 <- D1 %>%
  group_by(YearBuilt) %>%
  summarise(n=n())
D_ts <- data.frame(YearBuilt=1872:2010)
D_ts <- left_join(D_ts,D_sus)
```

```
## Joining, by = "YearBuilt"
```

```r
D_ts <- left_join(D_ts,D2,by="YearBuilt")
D_ts[is.na(D_ts)] <- 0
D_ts <- D_ts %>%
  mutate(per=n.x/n.y)
D_ts[is.na(D_ts)] <- 0

# Plot for Percent of Suspects Every Year
ggplot(D_ts, aes(x=YearBuilt,y=per))+
  geom_point()+
  geom_smooth(se=FALSE)+
  xlab("Year Built")+ylab("Percentage of Suspects")+
  theme_classic()+
  ggtitle("Percent of Suspects Every Year")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



According to these two plots, we can get the information that "suspects" are mostly built in late years, and percentage of "suspects" in all built houses every year begin to increase since 1940. Thus, deviation from Benford's law could also due to the reason of time, when lot area within 10,000 and 12,000 become popular as time goes by.

4) Geography

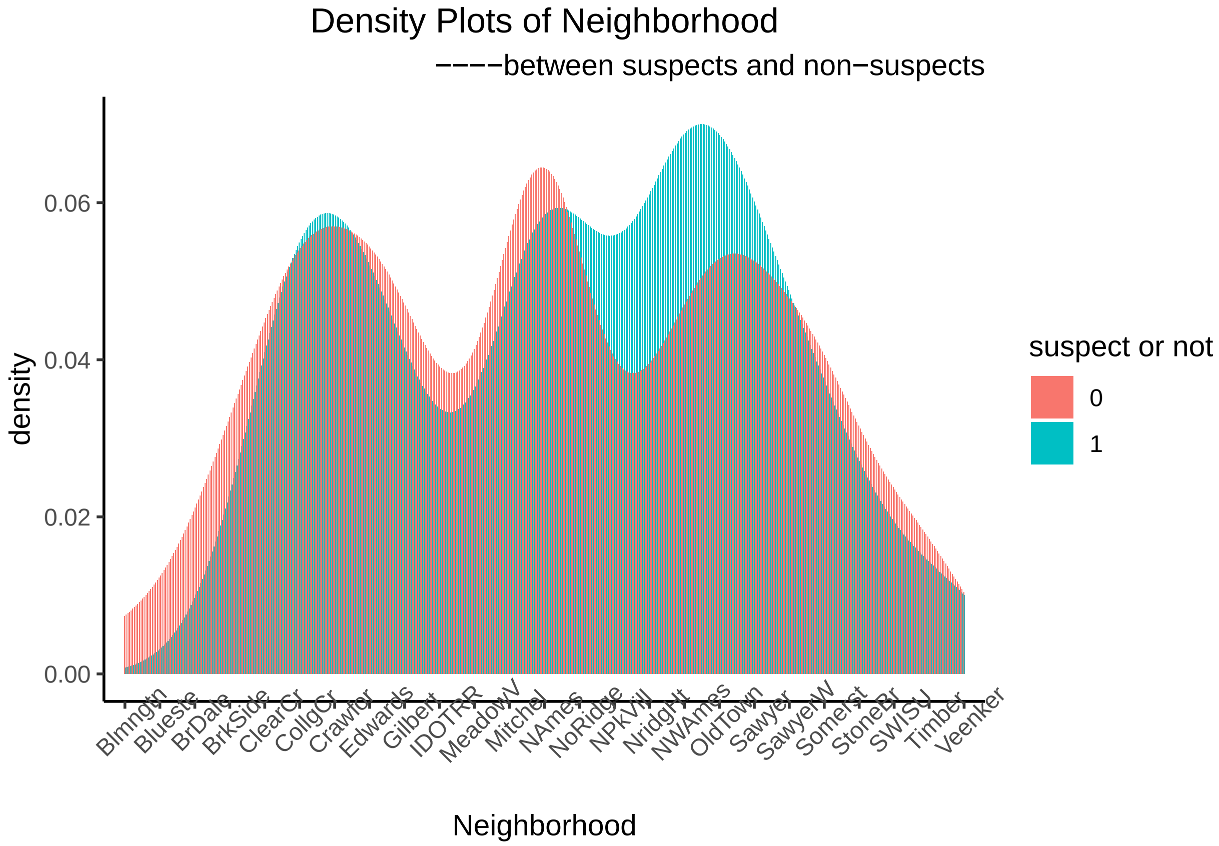
Finally I check the geography for "suspects". The values of variable "neighborhood" are:

```
Blmngtn -- Bloomington Heights
Blueste -- Bluestem
BrDale -- Briardale
BrkSide -- Brookside
ClearCr -- Clear Creek
CollgCr -- College Creek
Crawfor -- Crawford
Edwards -- Edwards
Gilbert -- Gilbert
IDOTRR -- Iowa DOT and Rail Road
MeadowV -- Meadow Village
Mitchel -- Mitchell
Names -- North Ames
NoRidge -- Northridge
NPkVill -- Northpark Villa
NridgHt -- Northridge Heights
NWAmes -- Northwest Ames
```

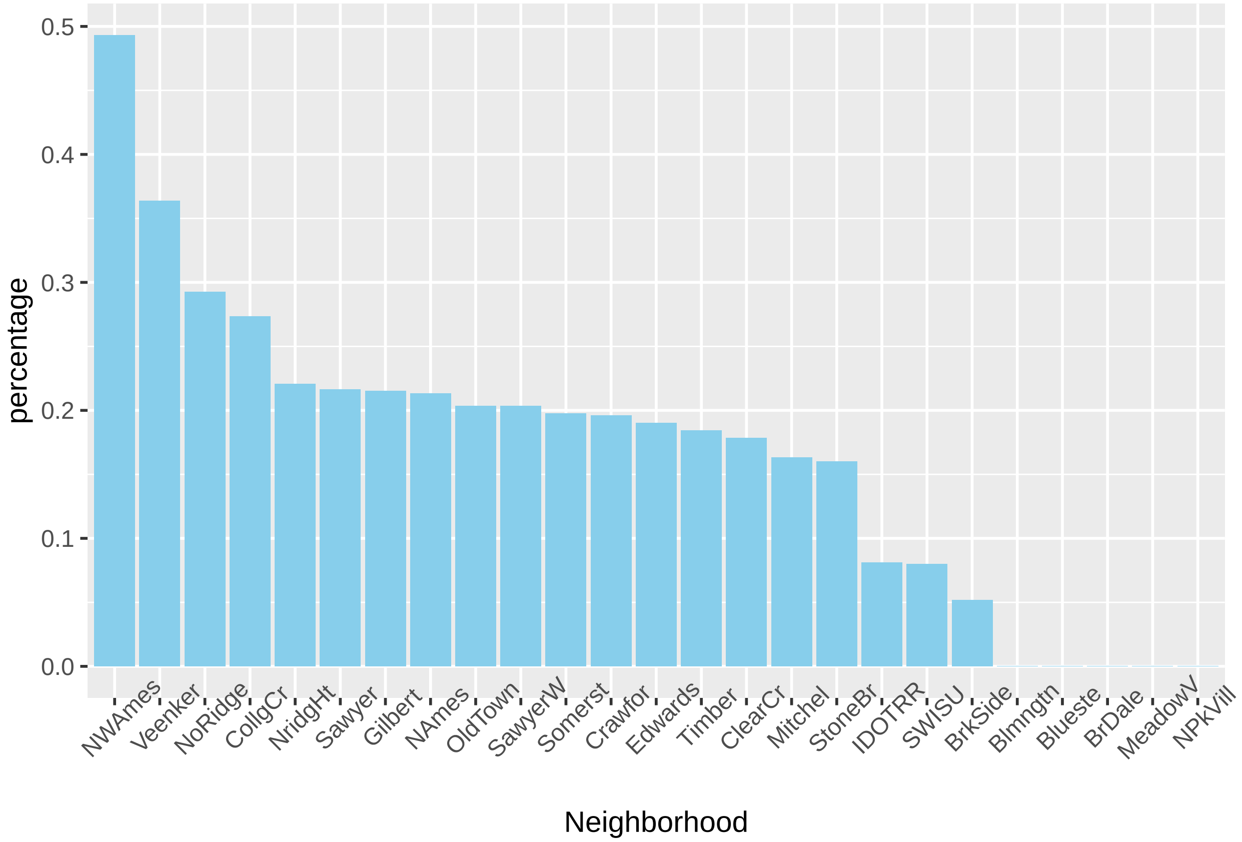
```
   OldTown -- Old Town
   SWISU -- South & West of Iowa State University
   Sawyer -- Sawyer
   SawyerW -- Sawyer West
   Somerst -- Somerset
   StoneBr -- Stone Brook
   Timber -- Timberland
   Veenker -- Veenker
```

```
# Density Plots of Neighborhood
ggplot(D1, aes(x=factor(Neighborhood),fill=factor(ifsuspect),group=factor(ifsuspect)))+
  geom_histogram(stat="density",position="dodge")+
  guides(fill=guide_legend(title="suspect or not"))+
  xlab("Neighborhood")+
  theme_classic()+
  ggtitle("Density Plots of Neighborhood",subtitle = "----between suspects and non-suspects")+
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 1),axis.text.x = ele
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



```
# Percentage of "suspects" in every neighborhood
D_geo <- D1 %>%
  group_by(Neighborhood) %>%
  summarise(per=sum(ifsuspect=="1")/n())
ggplot(D_geo)+
  geom_bar(mapping=aes(x=reorder(Neighborhood,desc(per)),y=per),stat="identity",fill="skyblue")+
  theme(axis.text.x = element_text(angle=45))+
ylab("percentage")+xlab("Neighborhood")
```



```
ggtitle("Percentage of Suspects Every Neighborhood")+
theme_classic()+
theme(plot.title = element_text(hjust = 0.5),axis.text.x = element_text(angle=45))
```

```
## NULL
```

As shown by these two plots, it is evident that neighborhood Northwest Ames and has the highest proportion of "suspects" and also a large portion of "suspects" are in this neighborhood. Almost half of the Neighborhood houses are "suspects", thus this may be the reason that lead to the deviation against Benford's law.

IV. Conclusion

Lot area in this dataset does not follow Benford's law. Also, the deviation from Benford's law may be due to the high proportion of low reidential density places, especially neighborhood Northwest Ames. Low reidential density places can lead to many houses with similar lot are between 10,000 and 12,000 and many similar features related to outside and inside living conditions, time and geography as showed before.