Airbnb Data Analysis

Zhaobin Liu

2018 12 12

Introduction

The airbnb information of three cities(Boston, Chicago, Seattle) dataset are from the Airbnb website: http://tomslee.net/airbnb-data-collection-get-the-data. I am using R to combine all of three separate csv files into one csv file called "total_data" to do one of the benford analysis.

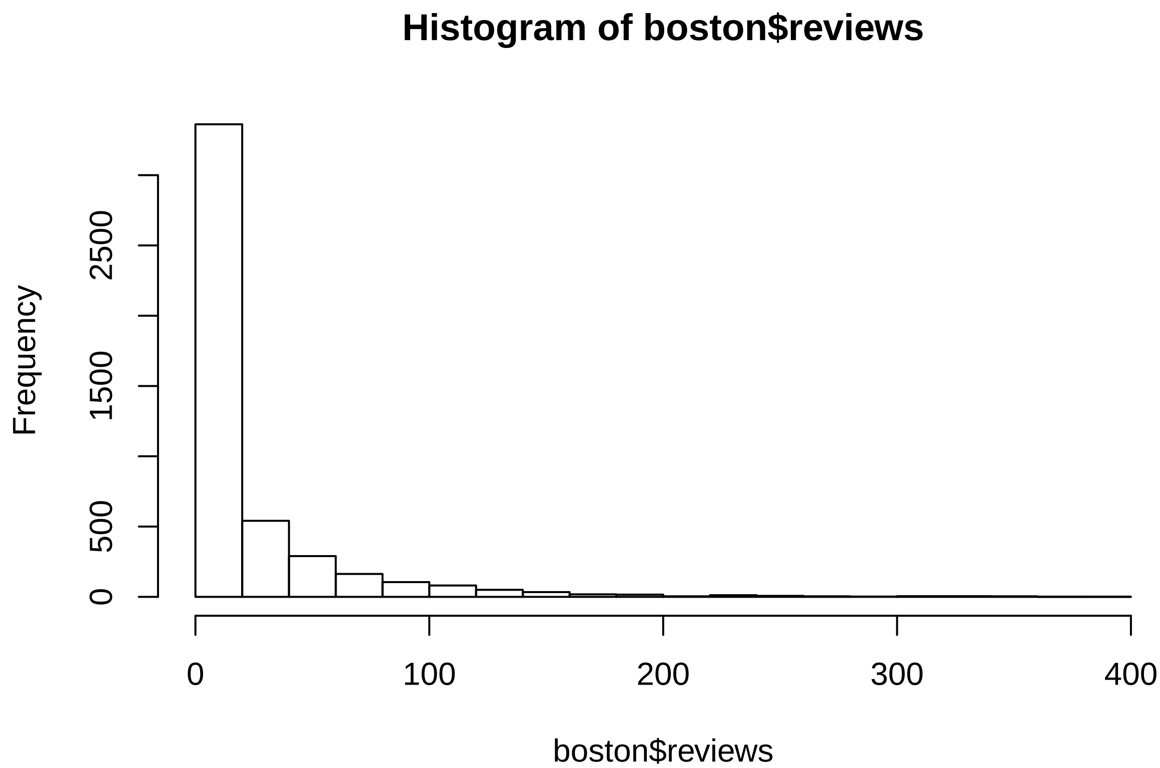
The interesting variables I am using in the dataset are: room_type, neighborhood, reviews, accommodates, bedrooms, price, latitude and longitude. I will do some EDA for these variables of three cities.

For Benford analysis, I will analyze three cities separately, and get conclusion by analyzing the total_data which is the combination of three cities.

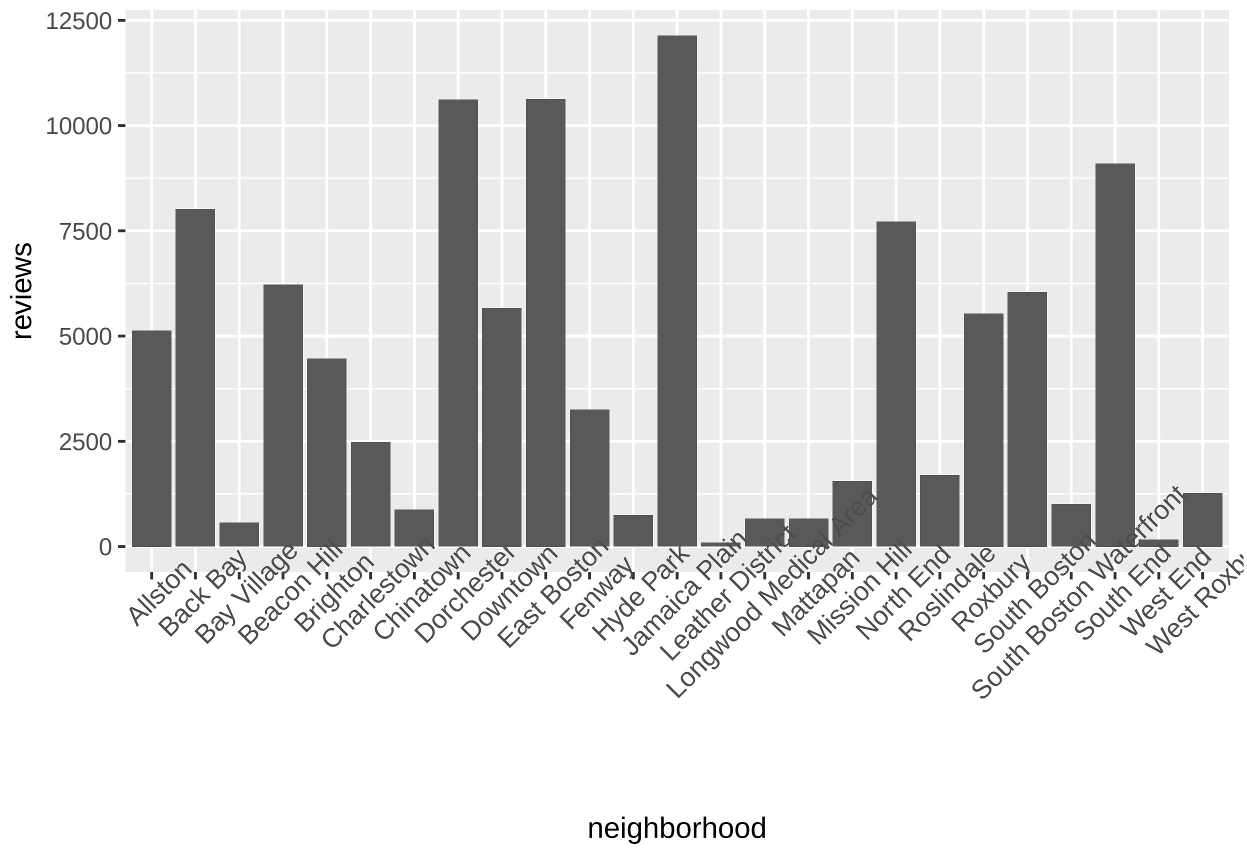
EDA

Boston

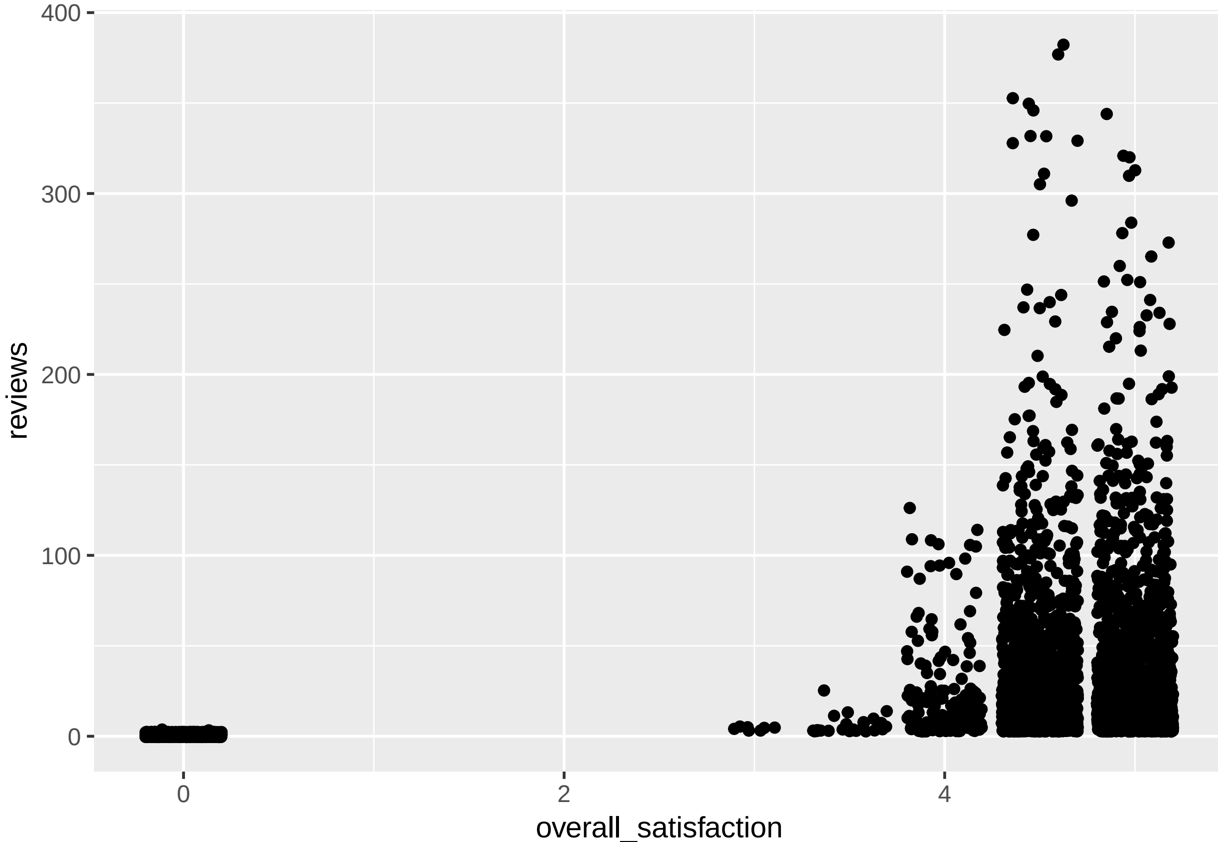
```
library(ggplot2)
library(readr)
boston <- read_csv("boston.csv", col_types = cols(borough = col_skip(),country = col_skip(),
                                                  location = col_skip()))

## Review Part

# reviews
# overall trend of reviews
hist(boston$reviews)
```



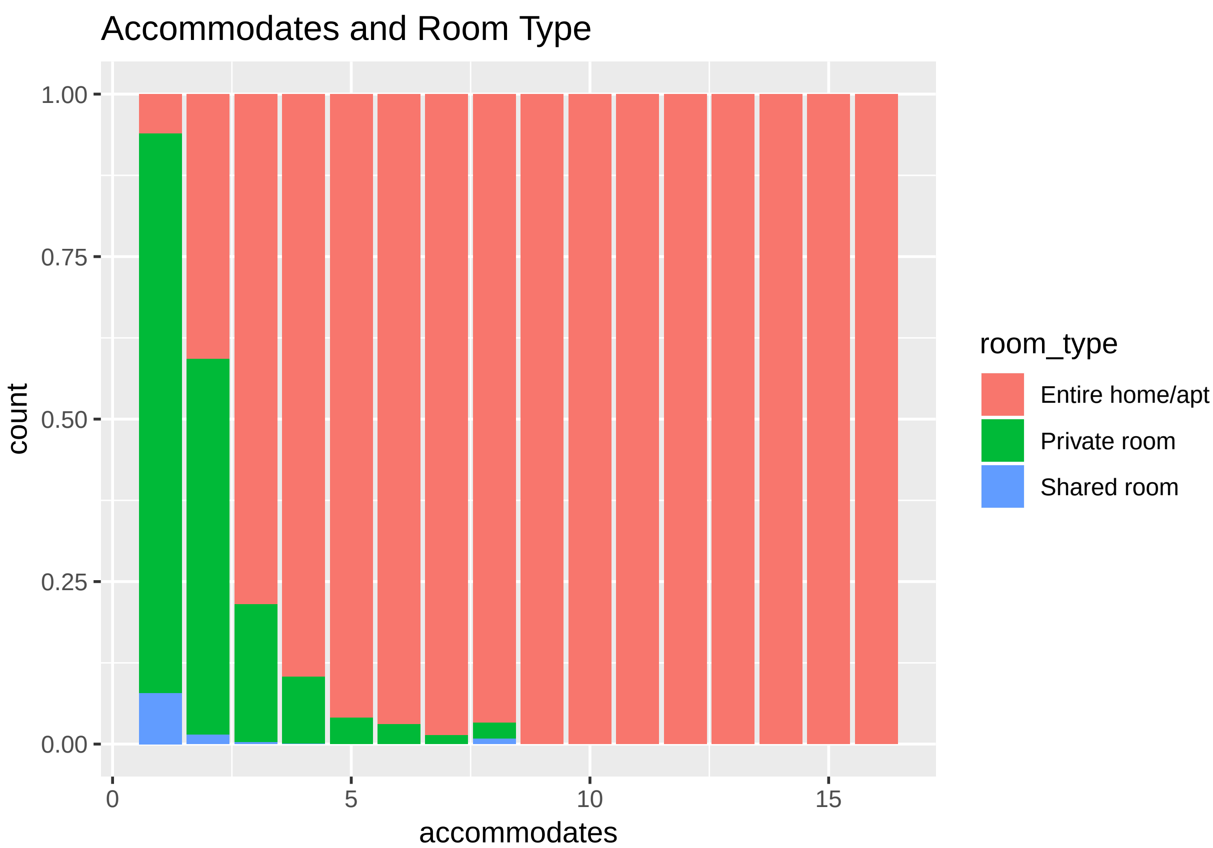

```r
# reviews with neighborhood
# total reviews from different neighborhood
ggplot(data=boston, aes(x=neighborhood, y=reviews))+geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(size=10, angle=45))
```



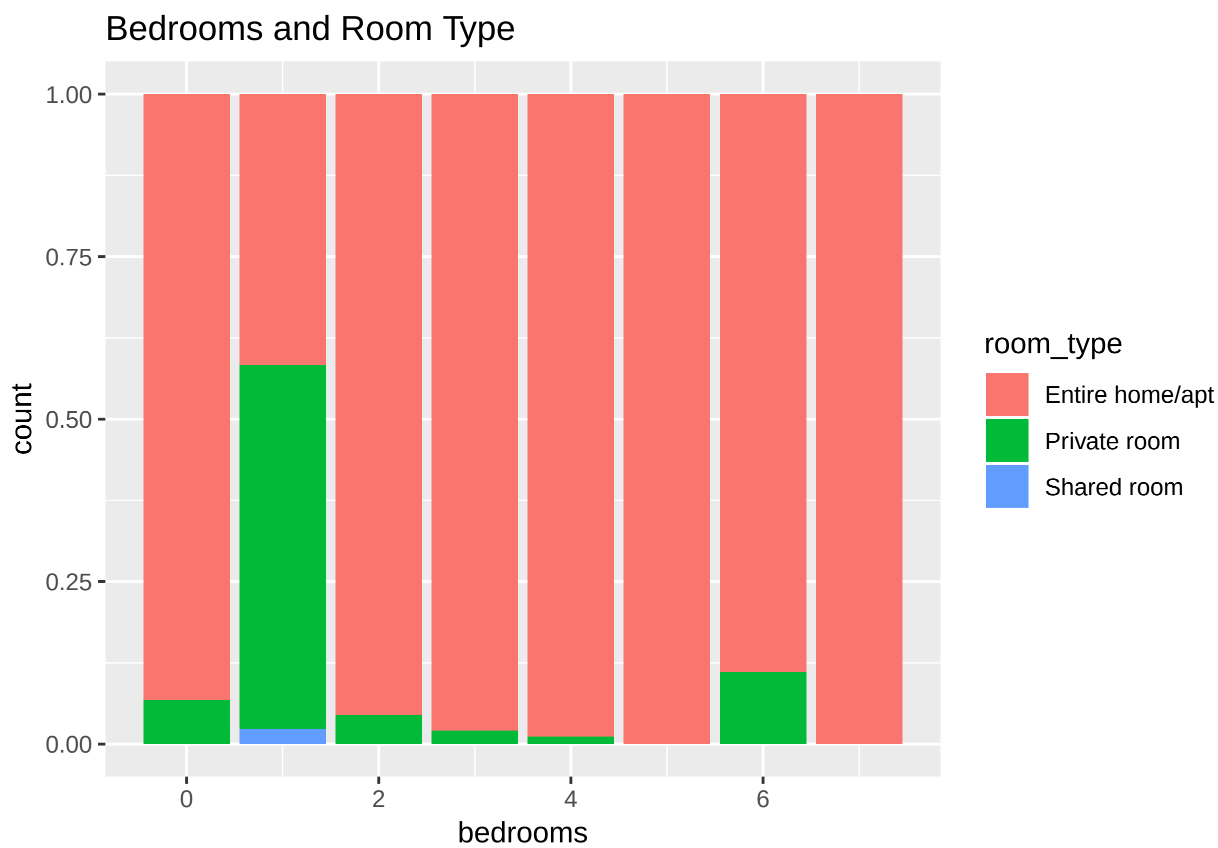

```r
# rating vs Review
# Higher rating with more reviews
ggplot(data=boston, aes(x=overall_satisfaction, y=reviews)) + geom_jitter()
```



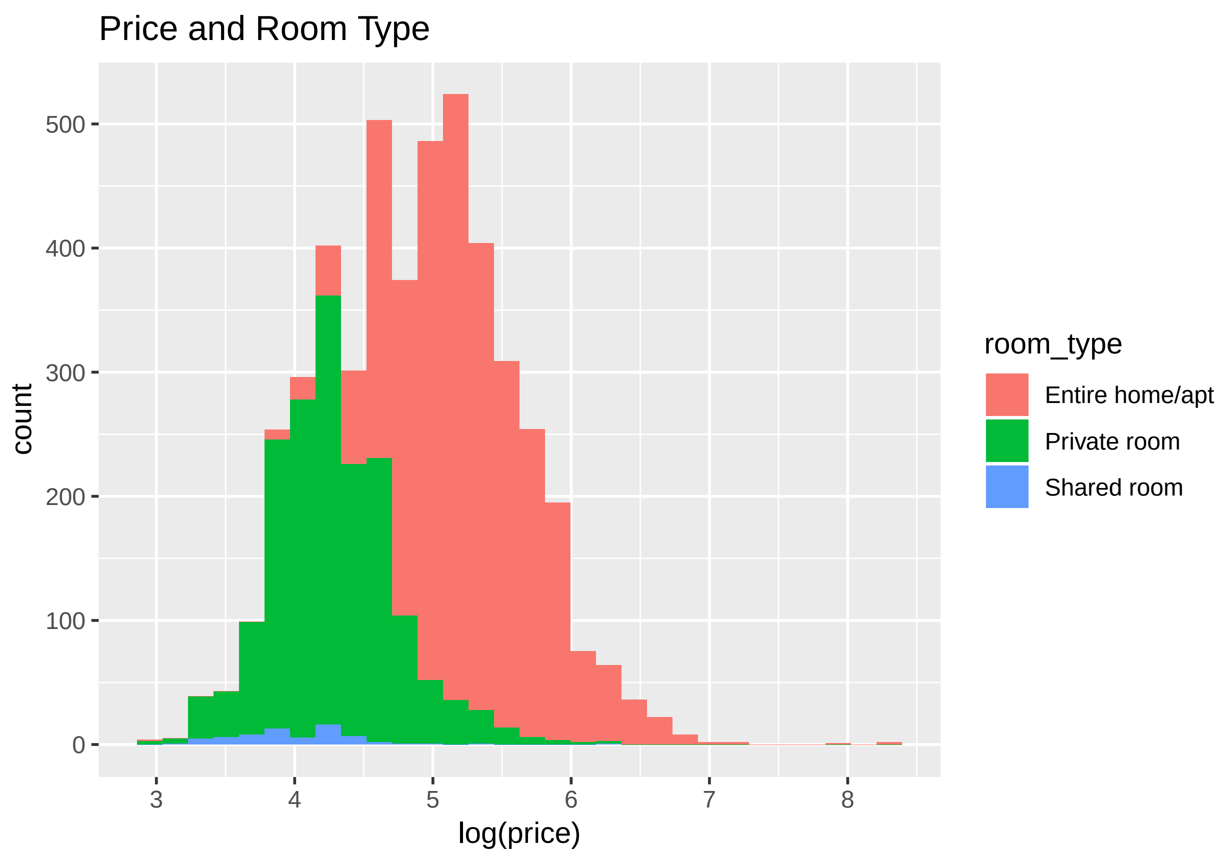

```
## Other Part

# room type and accommodates: entire home tends to allow more accommodates
ggplot(data=boston, aes(x=accommodates, fill=room_type))+geom_bar(position = "fill")+
  ggtitle("Accommodates and Room Type")
```



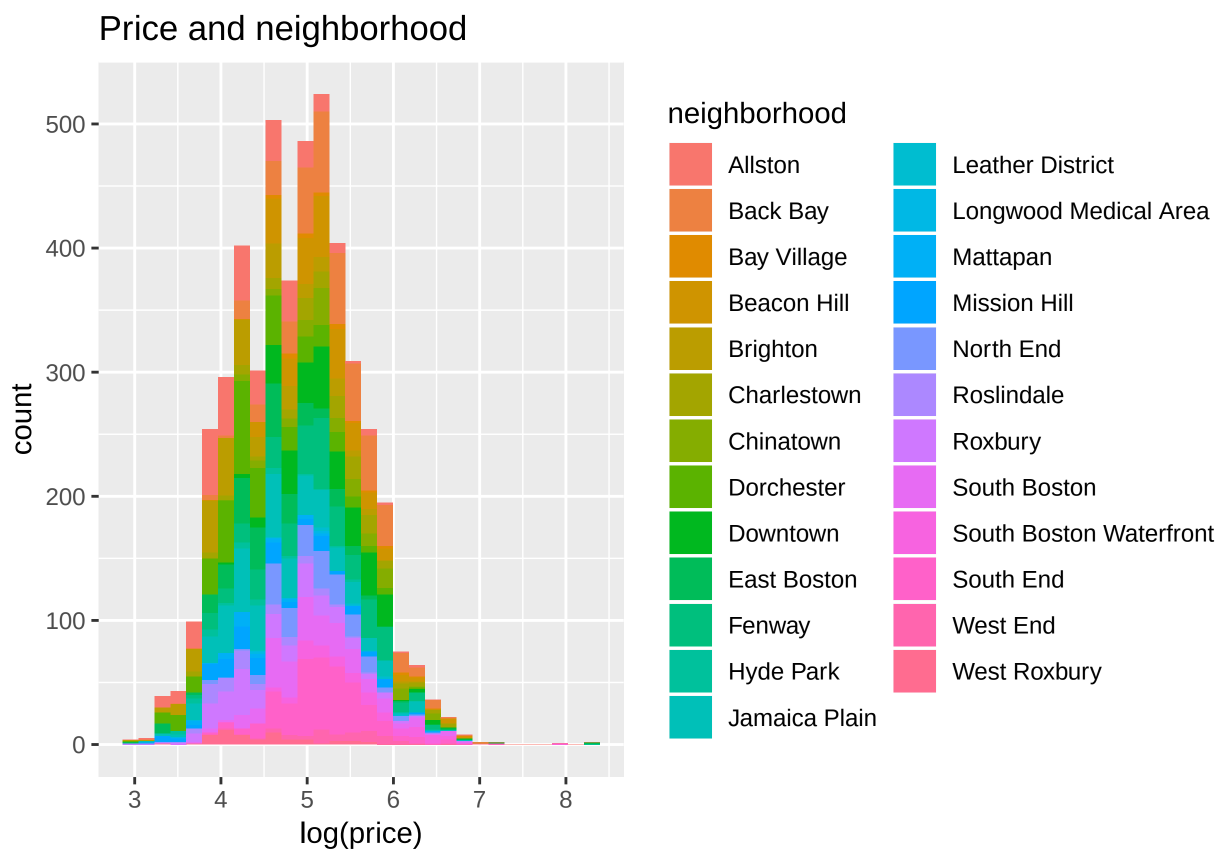

```r
#  room type and bedrooms: most of airbnb listes are entire home/apt or private room
ggplot(data=boston, aes(x=bedrooms, fill=room_type))+geom_bar(position = "fill")+
  ggtitle("Bedrooms and Room Type")
```



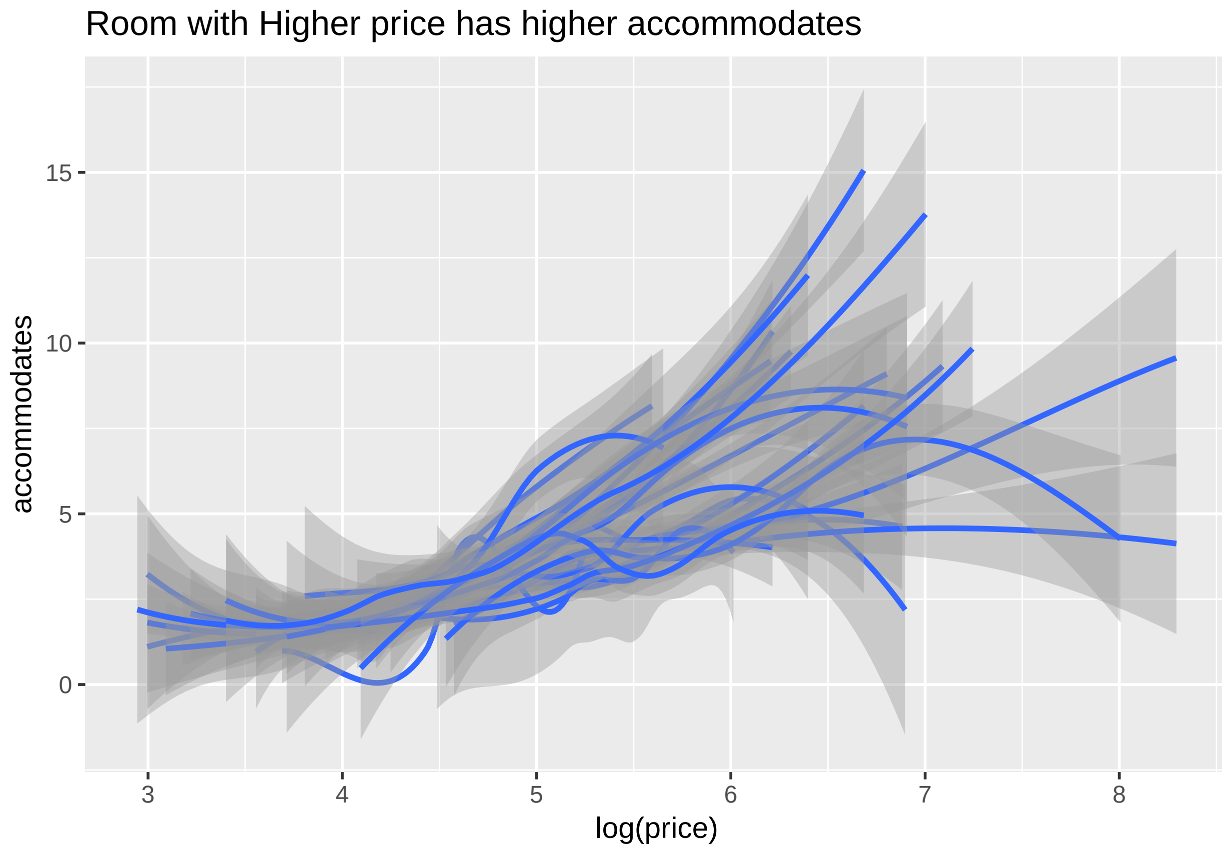

```
# Price part

#  price and property type: most of airbnb listed are entire home/apt and Private room
#  former has higher price overall
ggplot(data=boston, aes(x= log(price), fill=room_type))+geom_histogram()+
  ggtitle("Price and Room Type")
```



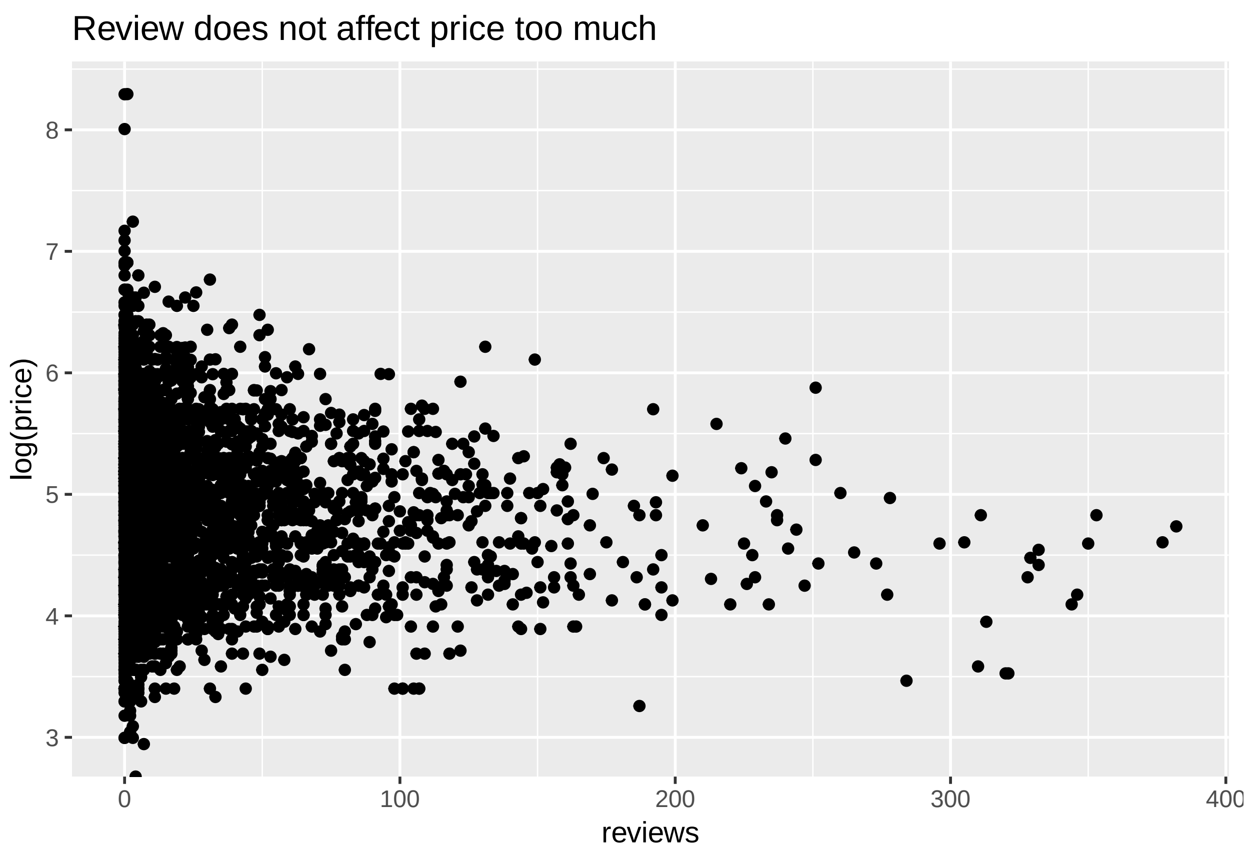

```
# price and neighborhood: variability between neighborhoods
ggplot(data=boston, aes(x=log(price), fill=neighborhood))+geom_histogram()+
  ggtitle("Price and neighborhood")
```




```r
# For majority part of neighborhood: higher price has higher accommodates
ggplot(boston, aes(x = log(price), y = accommodates, group = neighborhood)) + geom_smooth() +
  ggtitle("Room with Higher price has higher accommodates")
```

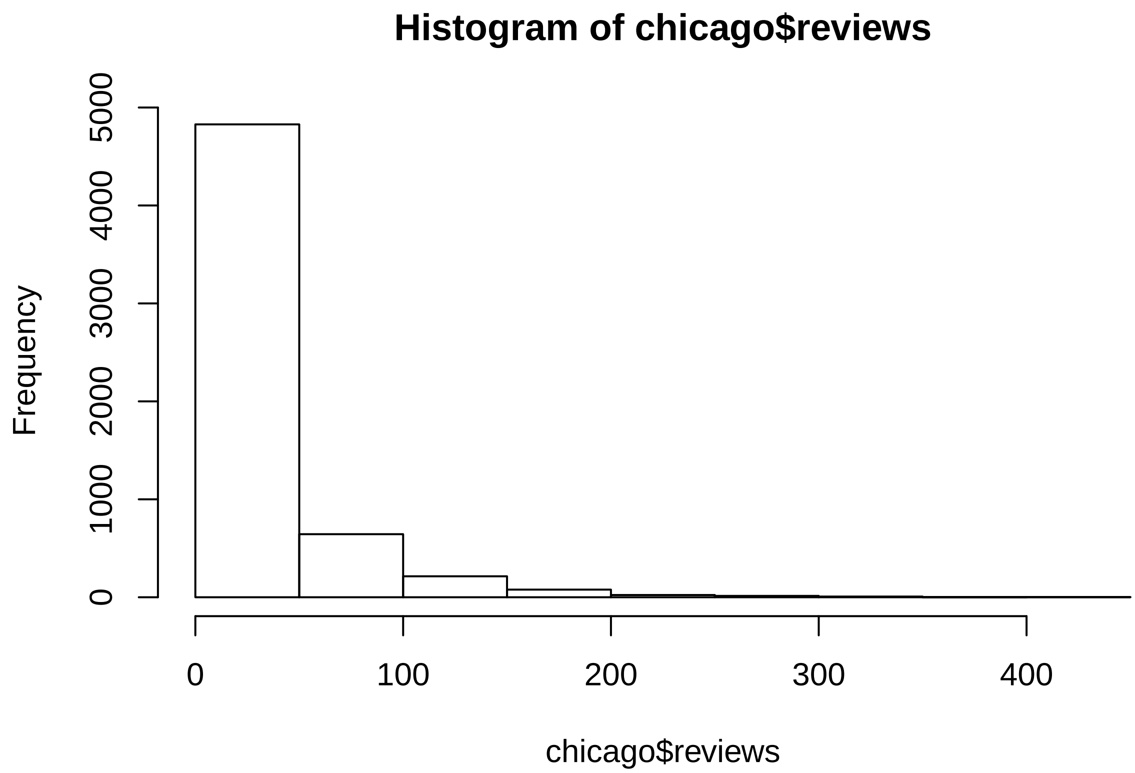



```r
# Review does not affect price too much
ggplot(boston) + aes(x = reviews, y = log(price)) + geom_point() +
  ggtitle("Review does not affect price too much")
```

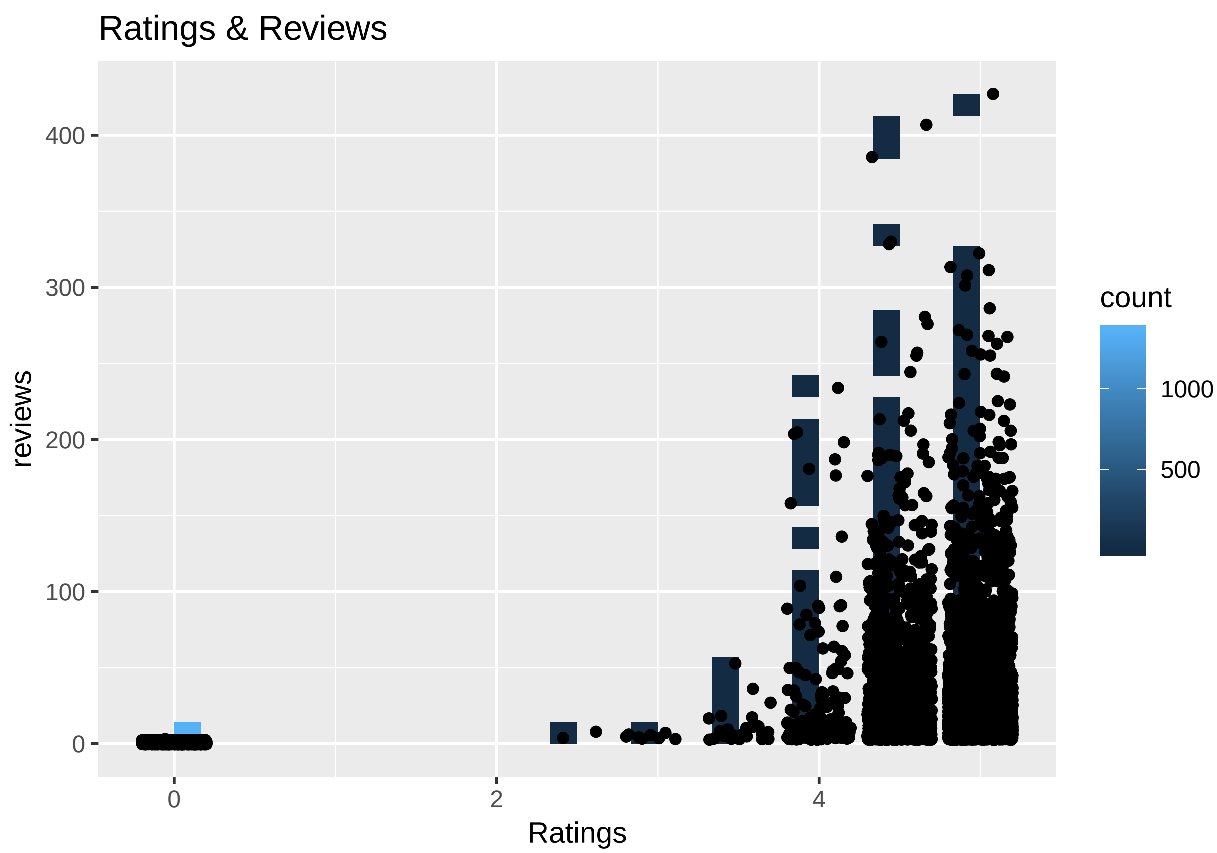


Chicago

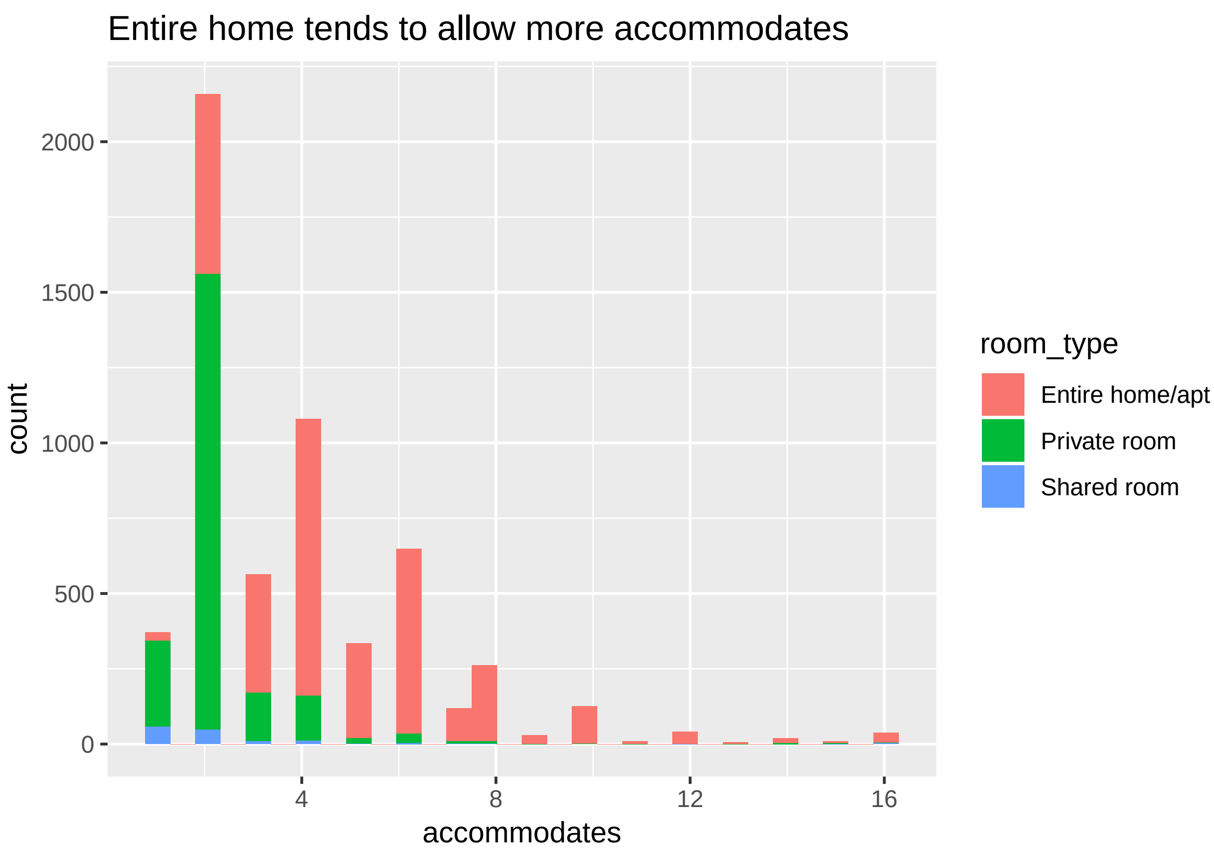
```
library(readr)
library(ggplot2)
chicago <- read_csv("chicago.csv", col_types = cols(borough = col_skip(),
     country = col_skip(), location = col_skip()))


## Review Part

# reviews
# overall trend of reviews
hist(chicago$reviews)
```



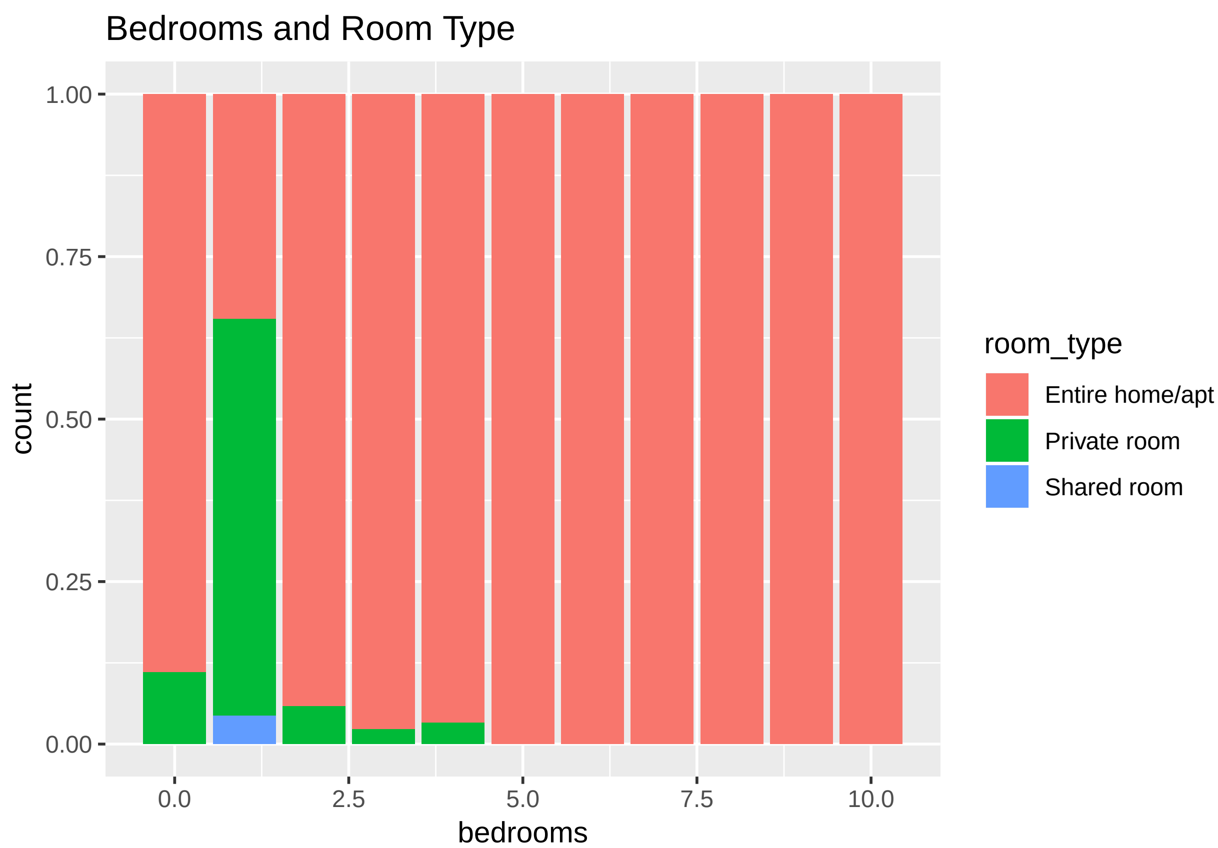

```r
# rating vs Review
# higher rating has more reviews
ggplot(data=chicago, aes(x=overall_satisfaction, y=reviews))+geom_bin2d()+xlab("Ratings")+
  ggtitle("Ratings & Reviews") + geom_jitter()
```



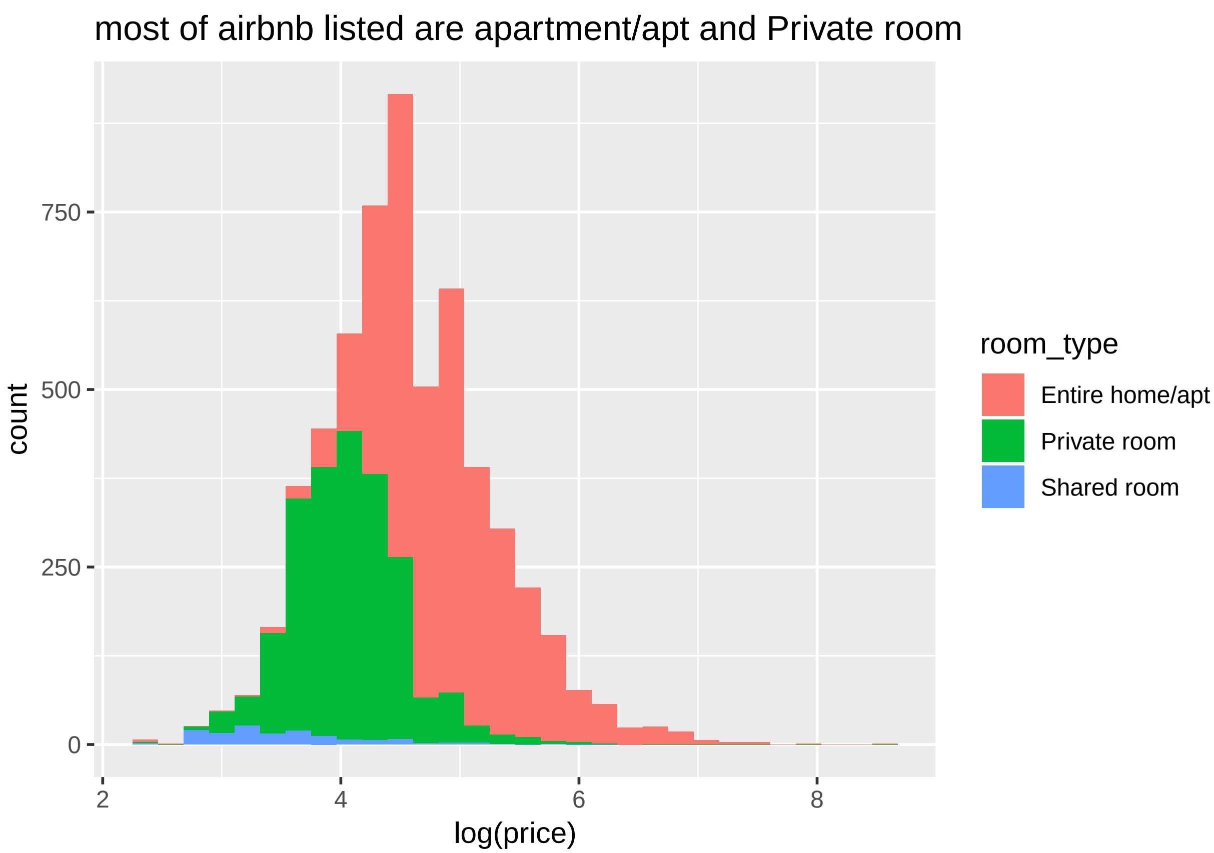

```
## Random Part

# room type and accommodates: entire home tends to allow more accommodates
ggplot(data=chicago, aes(x=accommodates, fill=room_type))+geom_histogram() +
  ggtitle("Entire home tends to allow more accommodates")
```



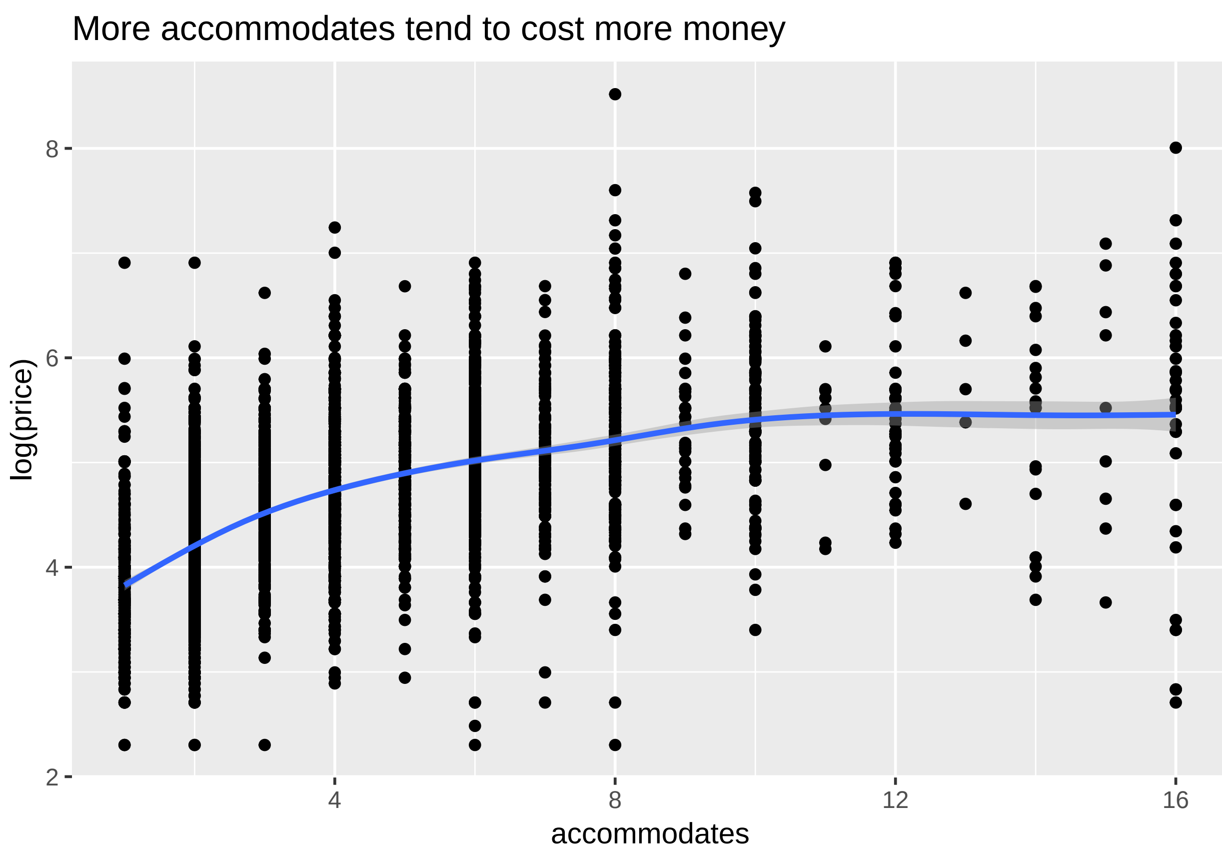

```r
# most of airbnb listes are entire home/apt
ggplot(data=chicago, aes(x=bedrooms, fill=room_type))+geom_bar(position = "fill")+
  ggtitle("Bedrooms and Room Type")
```



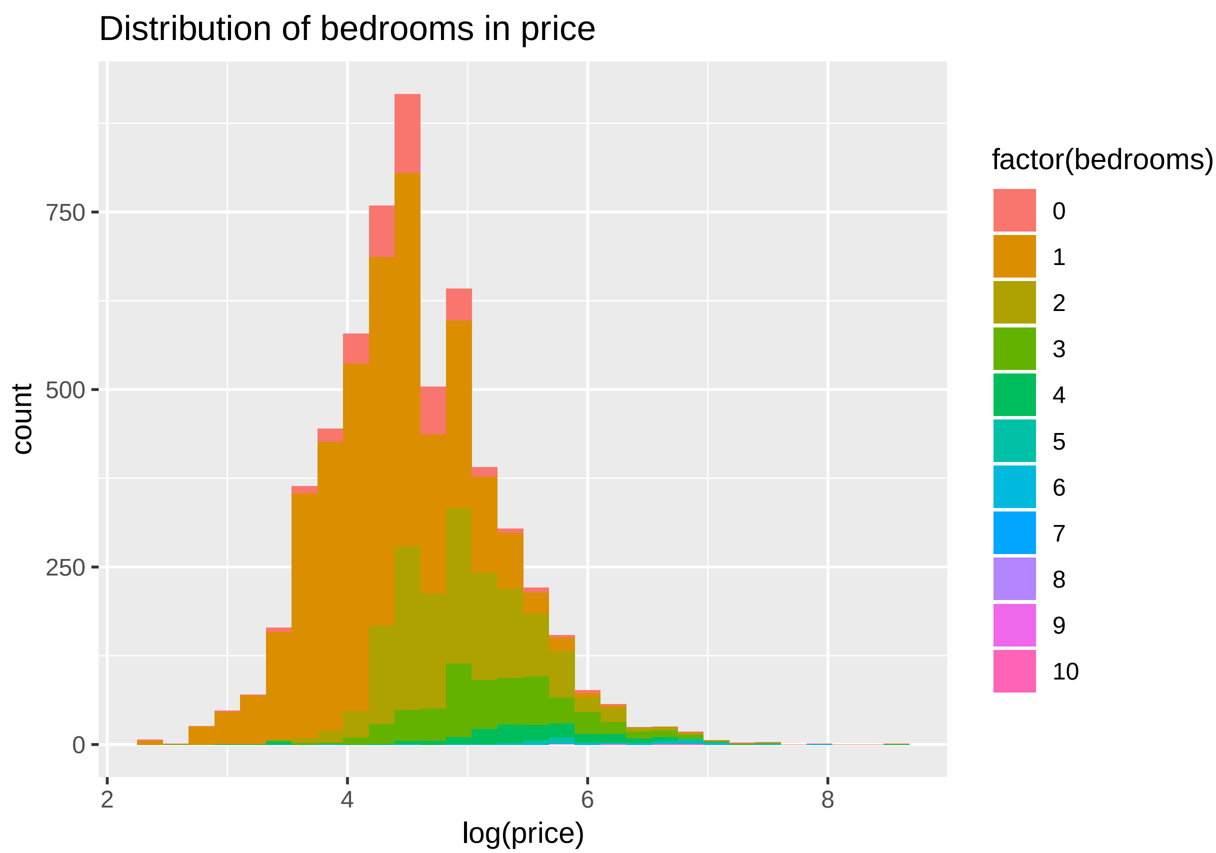

```r
# Price part

#  price and property type: most of airbnb listed are apartment/apt and Private room
ggplot(data=chicago, aes(x= log(price), fill=room_type))+geom_histogram()+
  ggtitle("most of airbnb listed are apartment/apt and Private room")
```




```r
# More accommodates tend to cost more money
ggplot(chicago, aes(x = accommodates,y = log(price))) +geom_point() +geom_smooth()+
  ggtitle("More accommodates tend to cost more money")
```



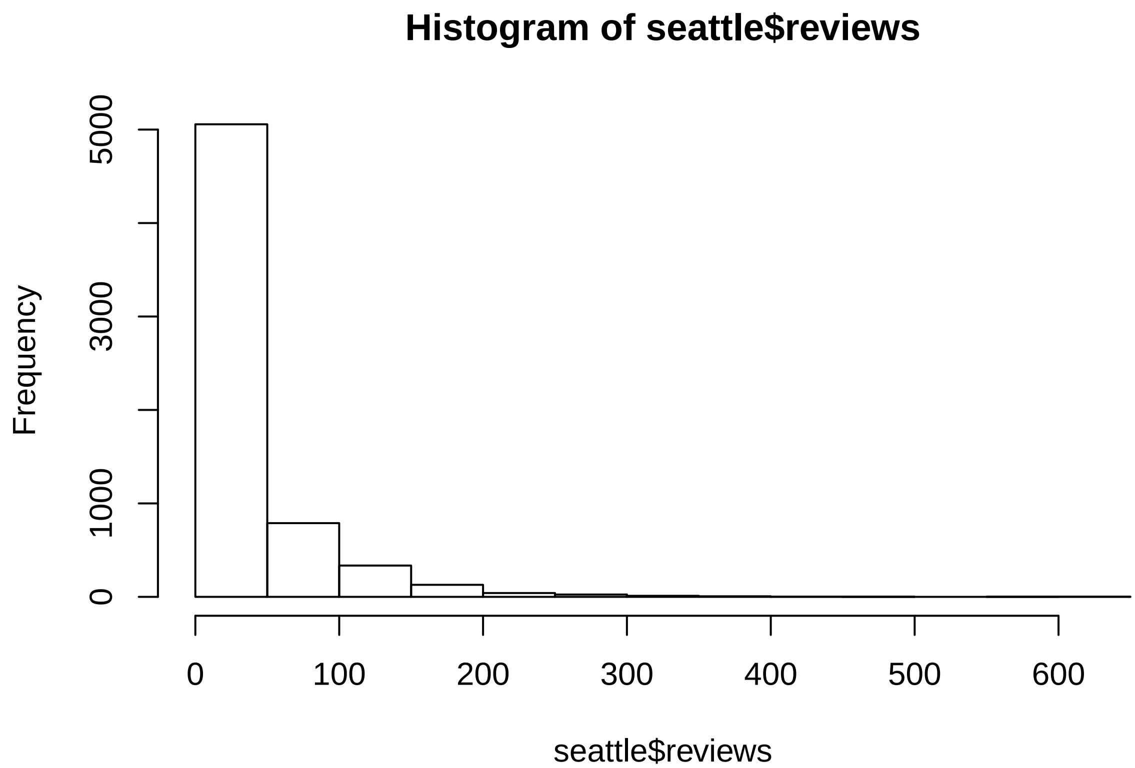

```r
# Treat bedrooms as categorical factors
# See the distribution of bedrooms in price
ggplot(data=chicago, aes(x= log(price), fill= factor(bedrooms)))+geom_histogram() +
  ggtitle("Distribution of bedrooms in price")
```



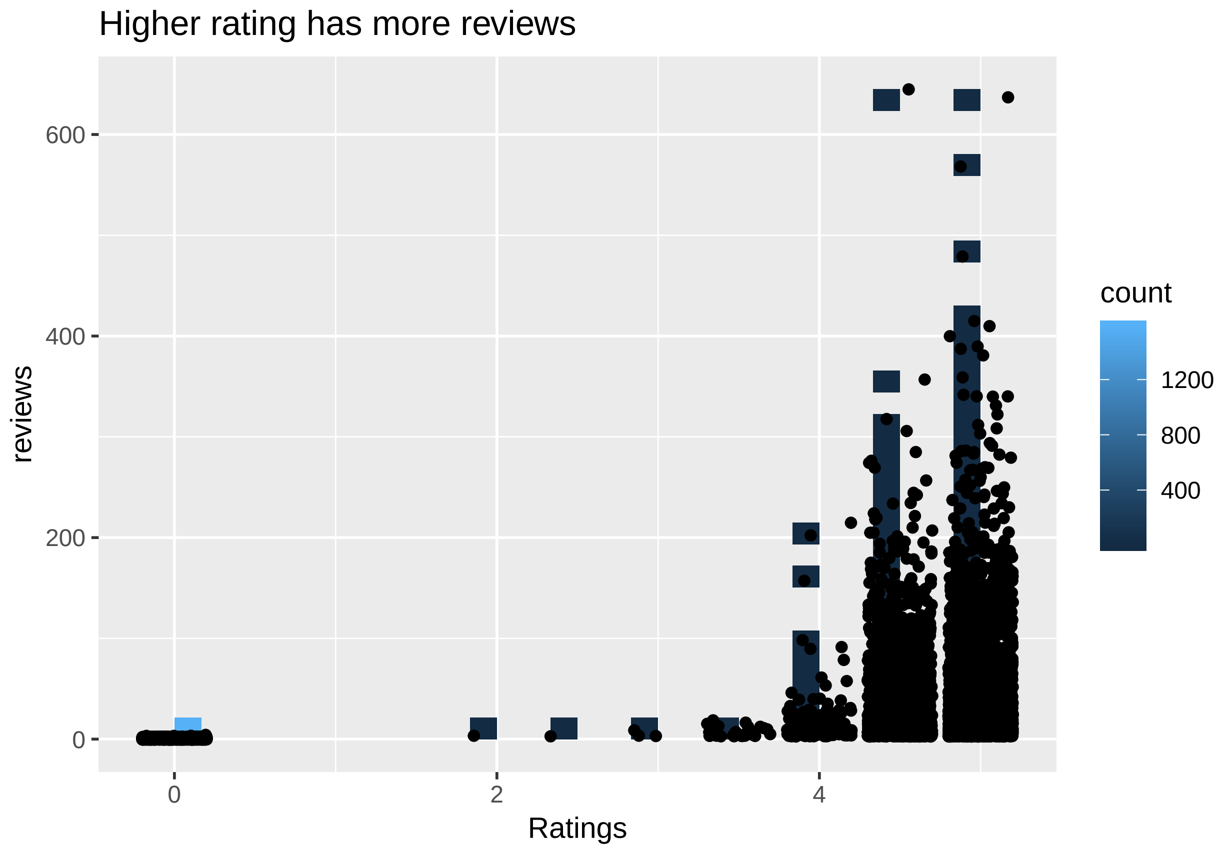

```r
#Reviews does not affect the price very much
ggplot(chicago) + aes(x = reviews, y = log(price)) + geom_point() +
  ggtitle("Reviews does not affect the price very much")
```

Seattle

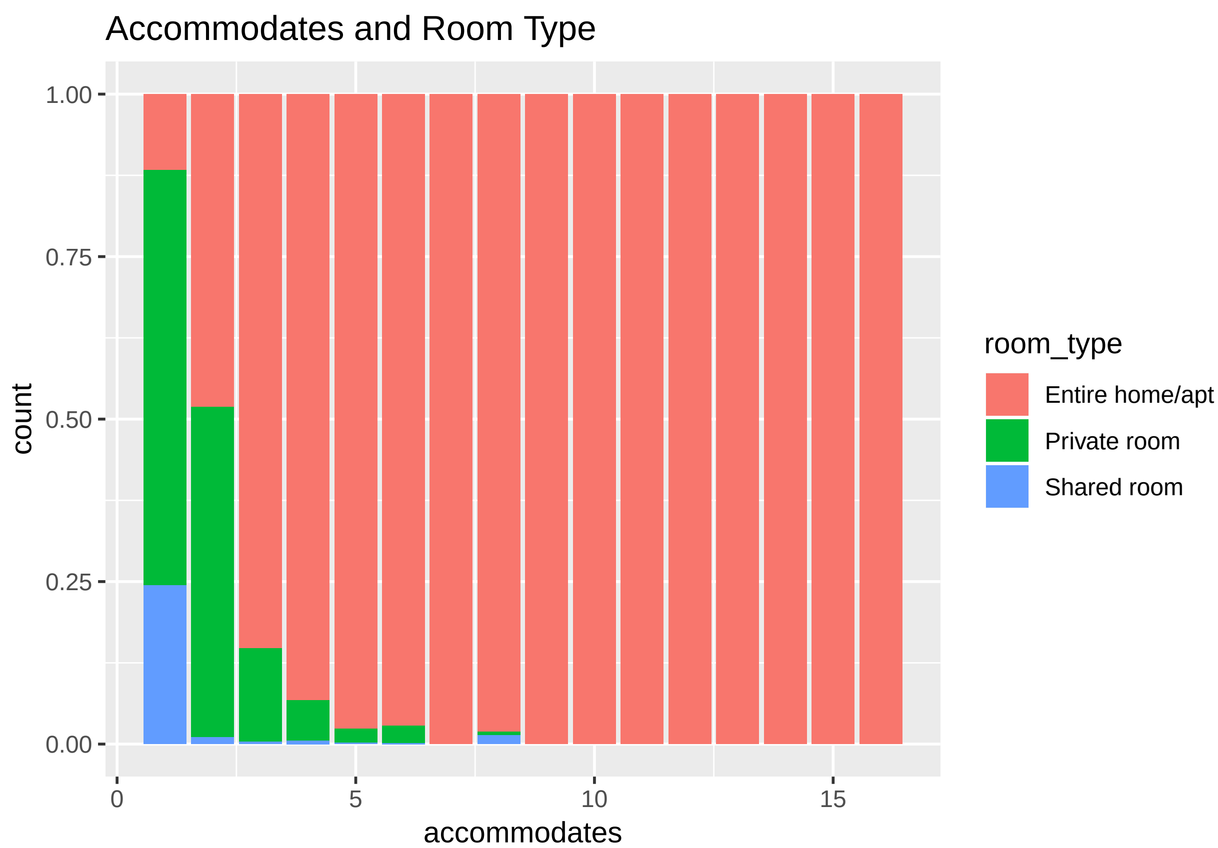
```
seattle <- read_csv("seattle.csv", col_types = cols(borough = col_skip(),
     country = col_skip(), location = col_skip()))


## Review Part

# reviews
# the trend of the reviews
hist(seattle$reviews)
```



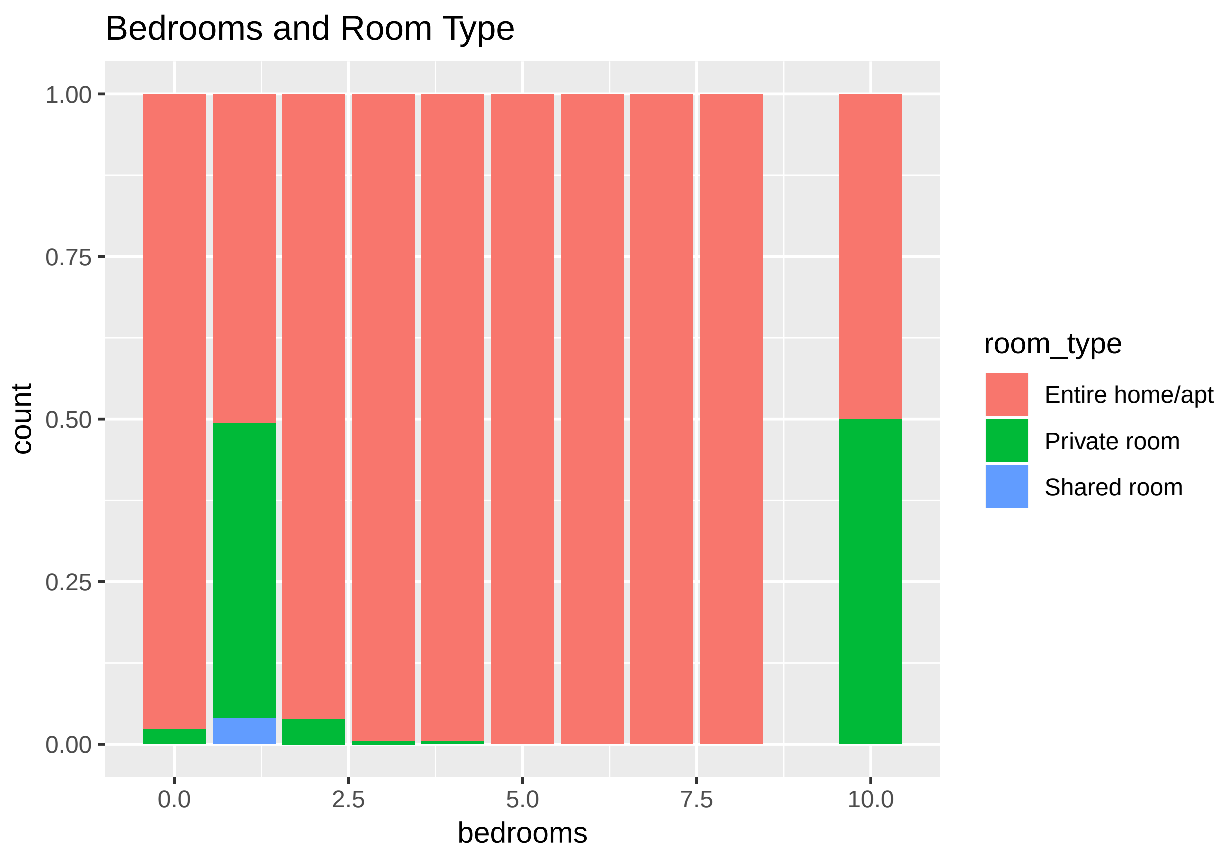

```
# rating vs Review
# higher rating has more reviews
ggplot(data=seattle, aes(x=overall_satisfaction, y=reviews))+geom_bin2d()+xlab("Ratings")+
  ggtitle("Higher rating has more reviews") + geom_jitter()
```



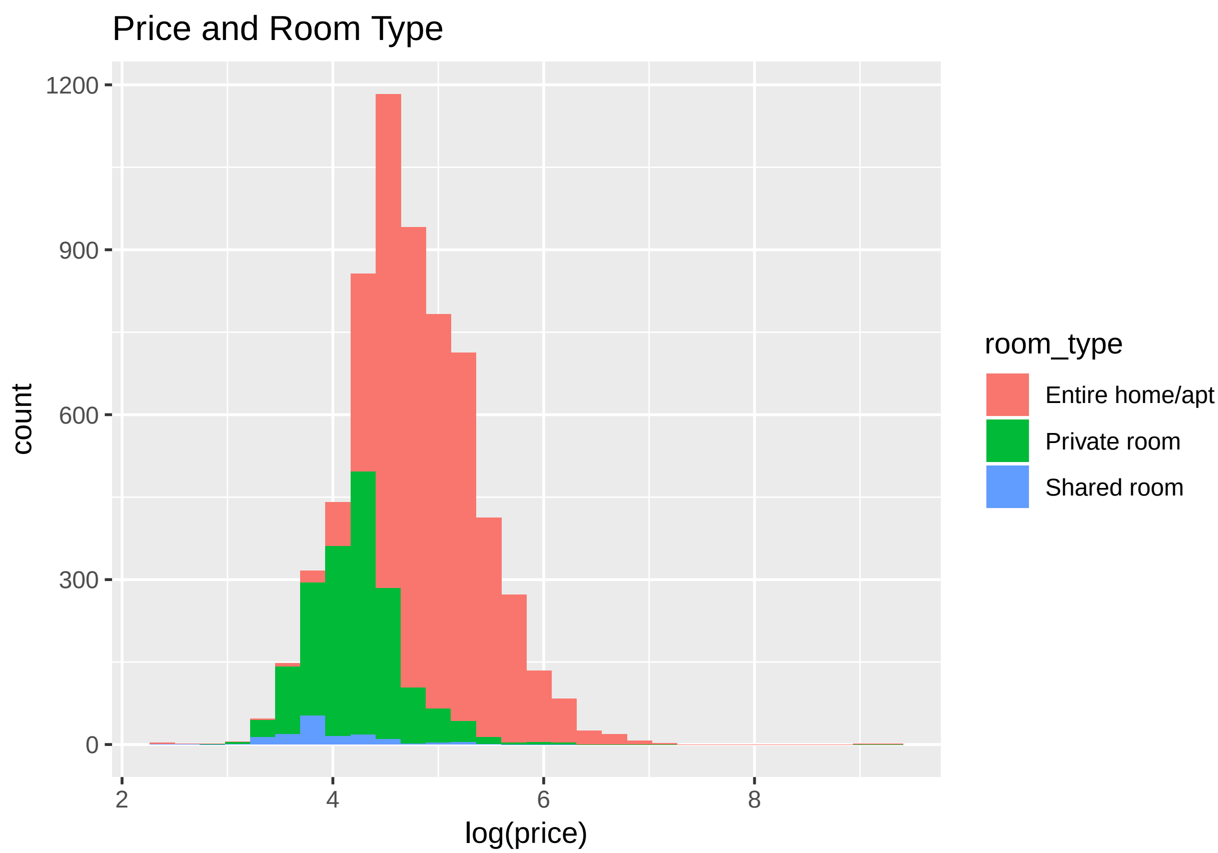

```
## Random Part

# room type and accommodates: entire home/apt tends to allow more accommodates
ggplot(data=seattle, aes(x=accommodates, fill=room_type))+geom_bar(position = "fill")+
  ggtitle("Accommodates and Room Type")
```



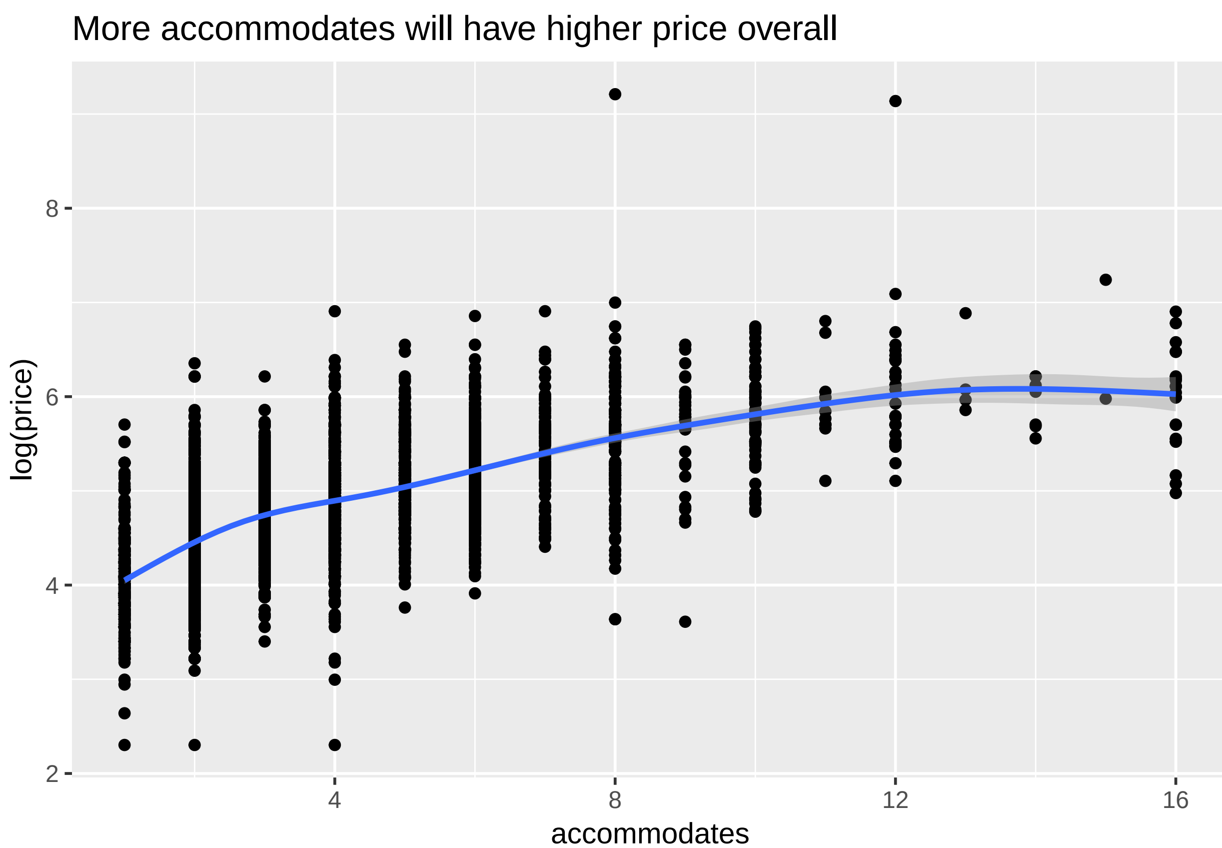

```r
#  room type and bedrooms: most of airbnb listes are entire home/apt or private room
ggplot(data=seattle, aes(x=bedrooms, fill=room_type))+geom_bar(position = "fill")+
  ggtitle("Bedrooms and Room Type")
```



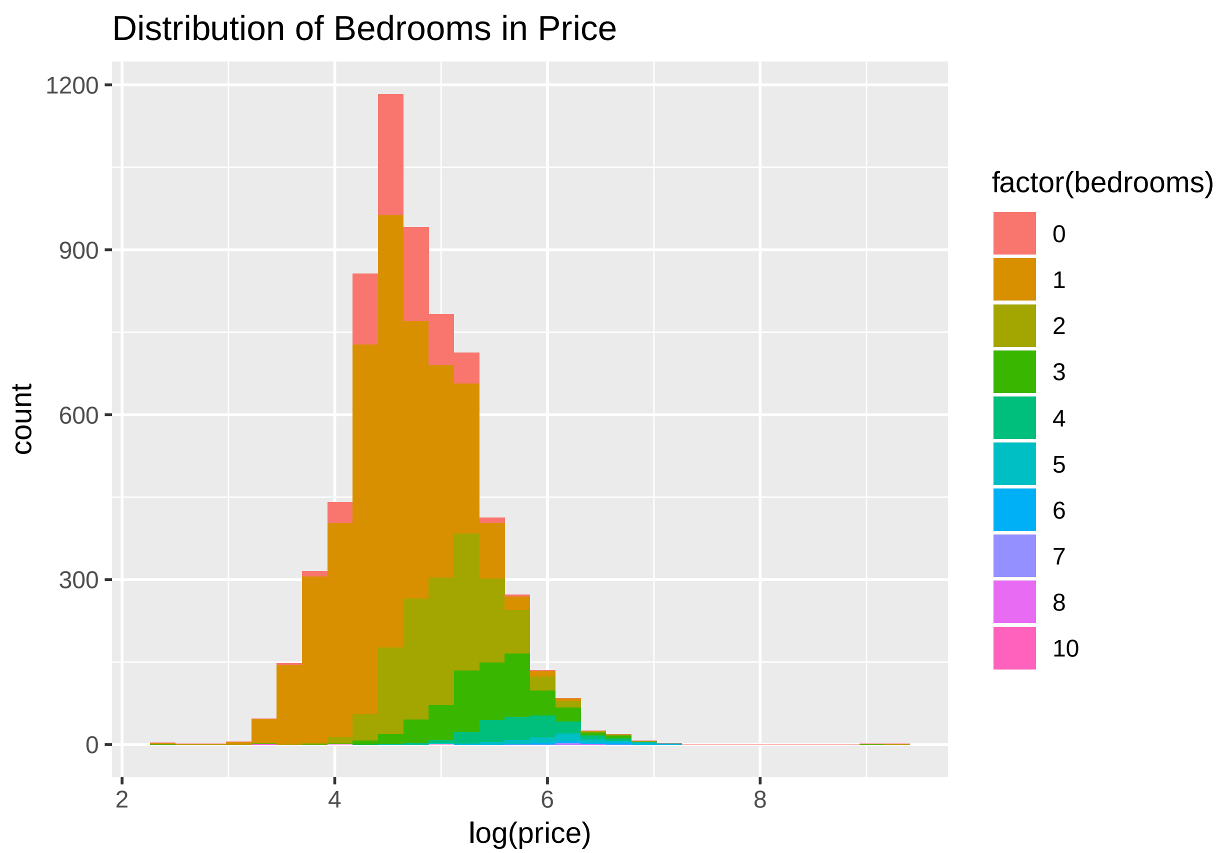

```r
# Price part

#  price and room type: most of airbnb listed are apartment/apt or private room
ggplot(data=seattle, aes(x= log(price), fill=room_type))+geom_histogram()+
  ggtitle("Price and Room Type")
```




```
#More accommodates will have higher price overall
ggplot(seattle, aes(x = accommodates,y = log(price))) +geom_point() +geom_smooth() +
  ggtitle("More accommodates will have higher price overall")
```




```
# Treat bedrooms as categorical factors
# See the distribution of bedrooms in price
ggplot(data=seattle, aes(x= log(price), fill= factor(bedrooms)))+geom_histogram() +
  ggtitle("Distribution of Bedrooms in Price")
```

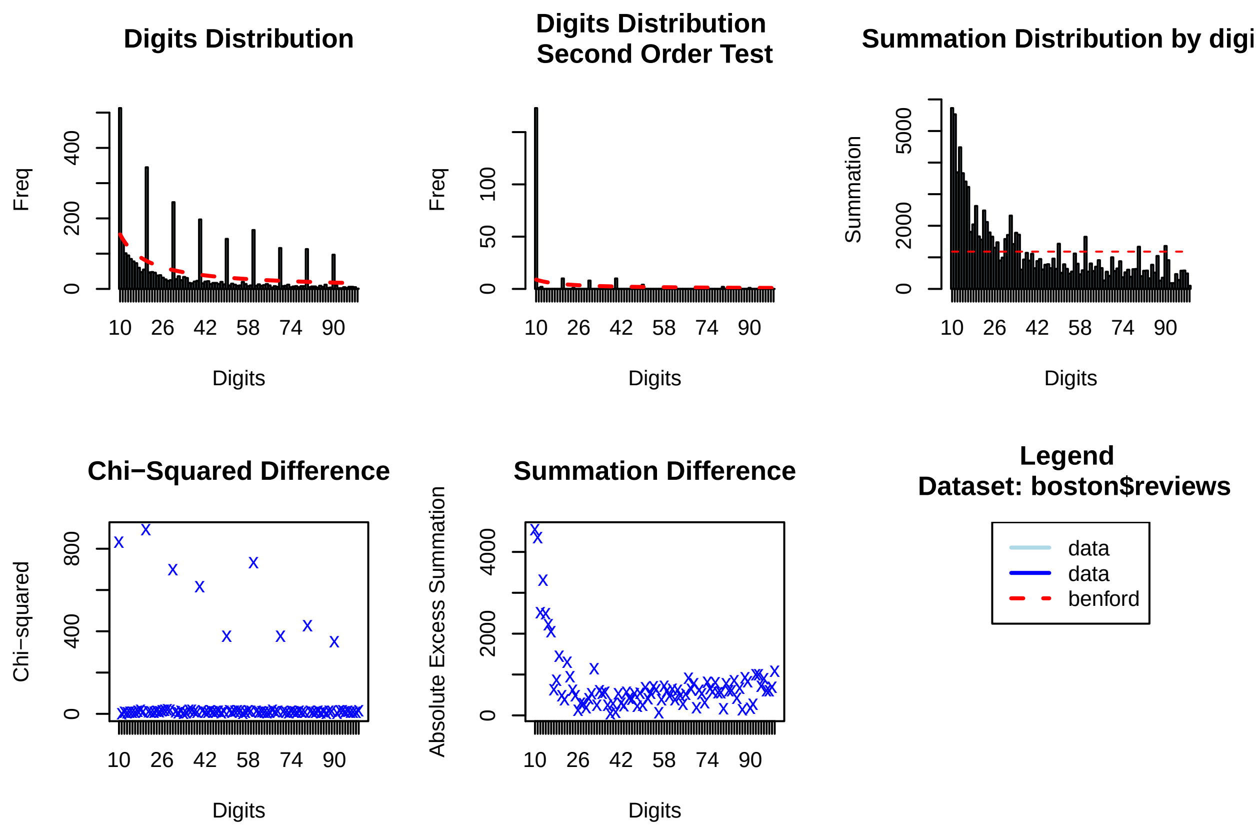



```
#Reviews does not affect the price too much
ggplot(seattle) + aes(x = reviews, y = log(price)) + geom_point() +
  ggtitle("Reviews does not affect the price too much")
```

Benford analysis

Boston reviews

```
#Benford Boston reviews
library(benford.analysis)
plot(benford(boston$reviews, number.of.digits = 2, sign = "positive", discrete=TRUE, round=3))
```



The original data is in blue and the expected frequency according to Benford's law is in red.

Benford's analysis of the first digits indicate the data basically follows Benford's Law.

Digit Distribution Second Order Test calculates the digit frequencies of the differences between the ordered (ranked) values in a data set. It shows that this dataset generally follows Benford's law, except some specific two digits.

The Chi-Square and Summation Difference plots almost fit Benford's law, but not good enough.

```r
benford(boston$reviews)
```

```
##
## Benford object:
##
## Data: boston$reviews
## Number of observations used = 3736
## Number of obs. for second order = 215
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.434
##          Var  0.092
##  Ex.Kurtosis -1.221
##     Skewness  0.091
##
```

```
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     10        358.36
## 2     20        265.84
## 3     30        192.80
## 4     40        156.94
## 5     60        140.18
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  boston$reviews
## X-squared = 6093.1, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  boston$reviews
## L2 = 0.004444, df = 2, p-value = 6.159e-08
##
## Mean Absolute Deviation: 0.008856045
## Distortion Factor: -19.5255
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

Above result shows 5 largest discrepancies. As we can see from the plot, the highest deviation is 10.

From the log mantissa of the data, we can tell that the data follows Benford's Law. Because Mean closes to 0.5, Variance closes to 0.092, Ex. Kurtosis closes to -1.2, and Skewness closes to 0.

Degree of freedom equals 89 and p-value is small enough that we are supposed to reject the benford's law. X-squared value equals 6093.1 and stays away to the value of degree of freedom. Thus the dataset might have some problems by looking at these two values. The distribution of this data set looks good. All in all, this dataset should follow Benford's law.

The distortion factor is -19.5255.

```r
library(tidyverse)
library(knitr)
#Gets the the statistics of the first Digits of a benford object.
Bfd_boston_reviews <- getBfd(benford(boston$reviews))
#From this table, we can get the distribution of dataset by first two digits.

kable(Bfd_boston_reviews[1:10, 1:6])
```

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
10	0.1373126	0.8046512	0.0413927	173	513
11	0.0364026	0.0046512	0.0377886	1	136
12	0.0270343	0.0093023	0.0347621	2	101
13	0.0254283	0.0000000	0.0321847	0	95
14	0.0227516	0.0000000	0.0299632	0	85
15	0.0208779	0.0000000	0.0280287	0	78
16	0.0195396	0.0000000	0.0263289	0	73

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
17	0.0160600	0.0000000	0.0248236	0	60
18	0.0128480	0.0000000	0.0234811	0	48
19	0.0147216	0.0000000	0.0222764	0	55

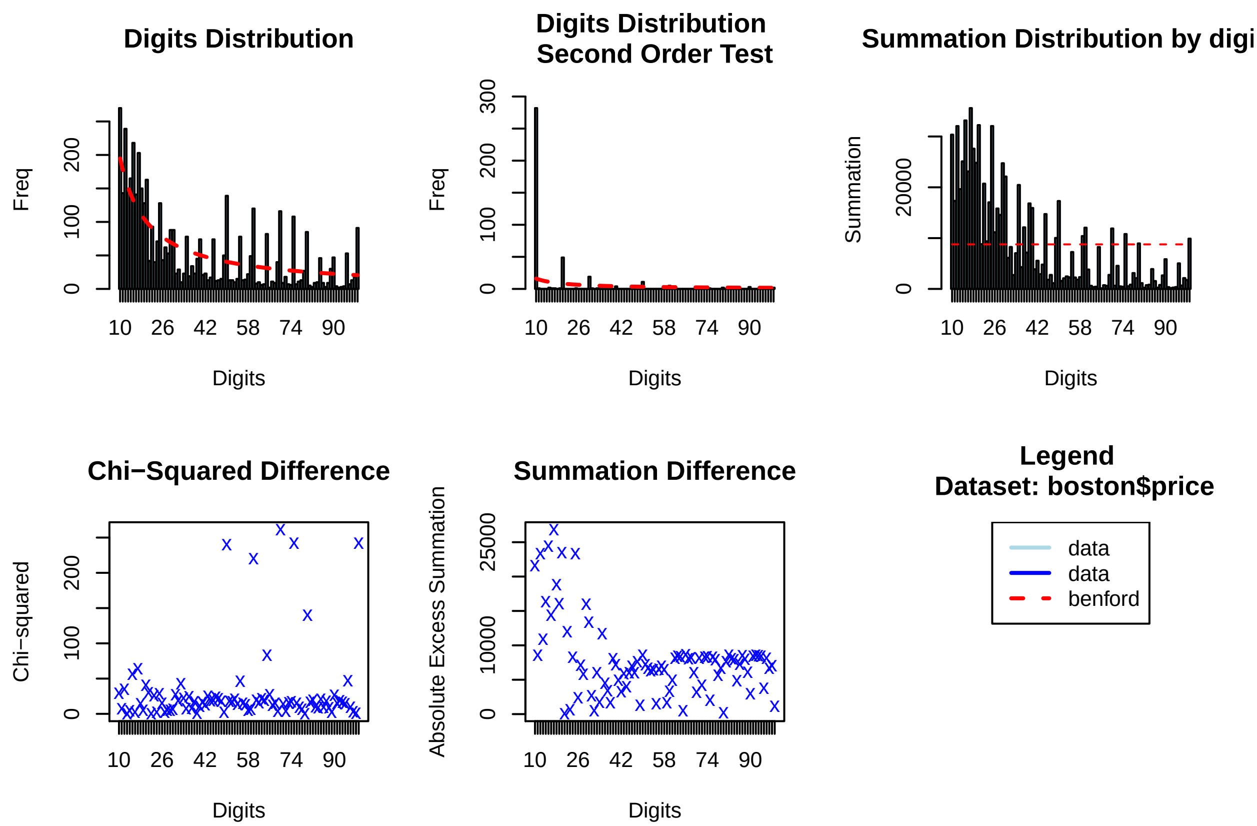
Table above shows the distribution of population data by first two digits.

digits	absolute.diff
10	358.35693
20	265.83678
30	192.79772
40	156.93564
60	140.18081
50	109.86976
70	92.98509
80	92.84416
90	79.07137
18	39.72537

Above table shows ten suspected two digits that contain most discrepancies from Benford's law.

Boston price

```
plot(benford(boston$price, number.of.digits = 2, sign = "positive", discrete=TRUE, round=3))
```



The original data is in blue and the expected frequency according to Benford's law is in red.

Benford's analysis of the first digits indicate the data basically follows Benford's Law.

Digit Distribution Second Order Test calculates the digit frequencies of the differences between the ordered (ranked) values in a data set. It shows that this dataset generally follows Benford's law, except some specific two digits.

The Chi-Square and Summation Difference plots almost fit Benford's law, but not good enough.

```r
benford(boston$price)
```

```
##
## Benford object:
##
## Data: boston$price
## Number of observations used = 4704
## Number of obs. for second order = 389
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.465
##          Var  0.091
##  Ex.Kurtosis -1.288
##     Skewness  0.189
##
```

```
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     50         98.54
## 2     70         87.02
## 3     60         86.23
## 4     17         86.23
## 5     15         86.15
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  boston$price
## X-squared = 2765.2, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  boston$price
## L2 = 0.017693, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.006371872
## Distortion Factor: -5.071143
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

Above result shows 5 largest discrepancies. As we can see from the plot, the highest deviation is 50.

From the log mantissa of the data, we can tell that the data follows Benford's Law. Because Mean closes to 0.5, Variance closes to 0.091, Ex. Kurtosis closes to -1.2, and Skewness closes to 0.

Degree of freedom equals 89 and p-value is small enough that we are supposed to reject the benford's law. X-squared value equals 2765.2 and stays away to the value of degree of freedom. Thus the dataset might have some problems by looking at these two values. Thus, the price seems not follow the Benford distribution very well.

The distortion factor is -5.071143.

```r
library(tidyverse)
library(knitr)
#Gets the the statistics of the first Digits of a benford object.
Bfd_boston_price <- getBfd(benford(boston$price))
#From this table, we can get the distribution of dataset by first two digits.
kable(Bfd_boston_price[1:10, 1:6])
```

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
10	0.0573980	0.7249357	0.0413927	282	270
11	0.0303997	0.0025707	0.0377886	1	143
12	0.0508078	0.0000000	0.0347621	0	239
13	0.0310374	0.0000000	0.0321847	0	146
14	0.0350765	0.0000000	0.0299632	0	165
15	0.0463435	0.0051414	0.0280287	2	218
16	0.0299745	0.0025707	0.0263289	1	141
17	0.0431548	0.0025707	0.0248236	1	203

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
18	0.0318878	0.0000000	0.0234811	0	150
19	0.0272109	0.0025707	0.0222764	1	128

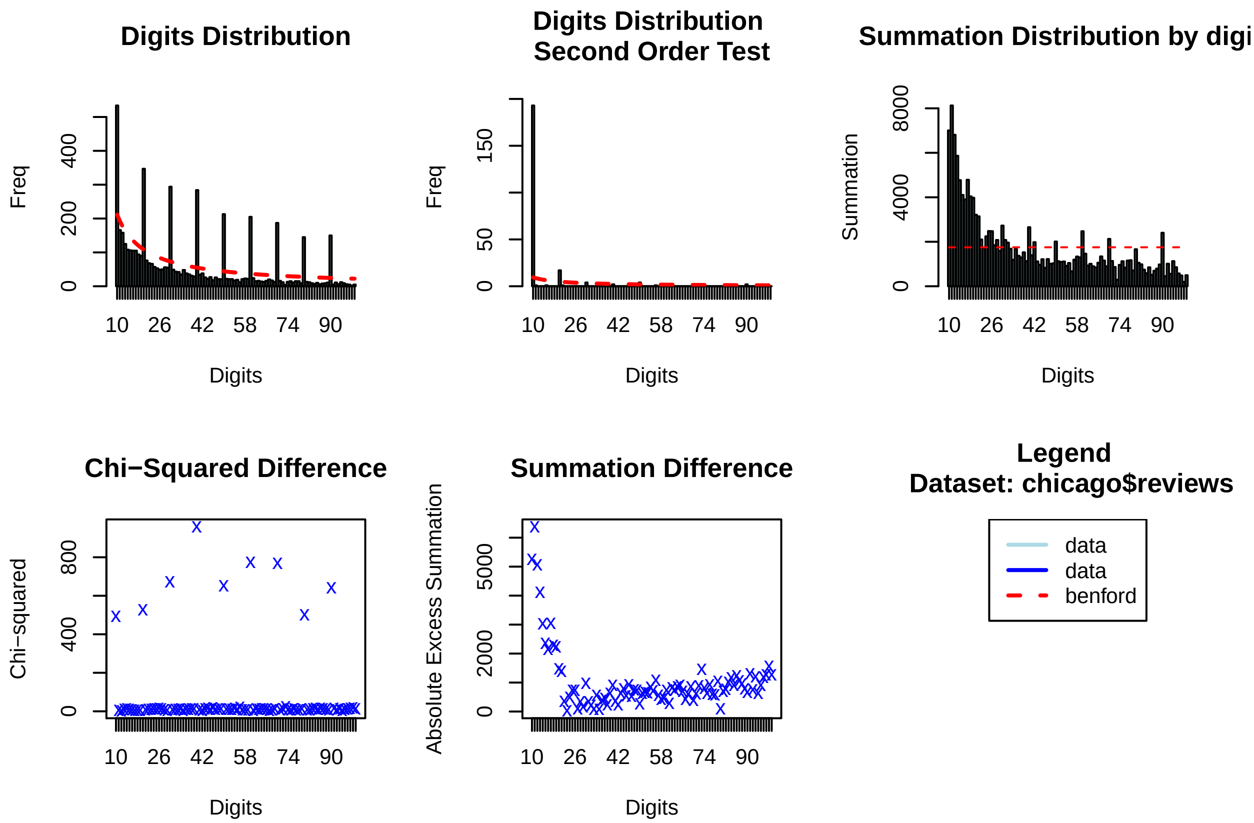
Table above shows the distribution of population data by first two digits.

digits	absolute.diff
50	98.54479
70	87.02191
60	86.23194
17	86.22986
15	86.15288
75	80.94104
12	75.47905
10	75.28881
99	70.46796
20	63.32554

Above table shows ten suspected two digits that contain most discrepancies from Benford's law.

Chicago reviews

```r
library(benford.analysis)
plot(benford(chicago$reviews, number.of.digits = 2, sign = "positive", discrete=TRUE, round=3))
```



The original data is in blue and the expected frequency according to Benford's law is in red.

Benford's analysis of the first digits indicate the data basically follows Benford's Law.

Digit Distribution Second Order Test calculates the digit frequencies of the differences between the ordered (ranked) values in a data set. It shows that this dataset generally follows Benford's law, except some specific two digits.

The Chi-Square and Summation Difference plots almost fit Benford's law, but not good enough.

```
benford(chicago$reviews)
```

```
##
## Benford object:
##
## Data: chicago$reviews
## Number of observations used = 5111
## Number of obs. for second order = 229
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic   Value
##         Mean  0.4610
##          Var  0.0893
##  Ex.Kurtosis -1.2024
##     Skewness -0.0051
##
##
```

```
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     10        322.44
## 2     20        238.70
## 3     40        229.19
## 4     30        221.22
## 5     50        169.04
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  chicago$reviews
## X-squared = 6795, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  chicago$reviews
## L2 = 0.0011289, df = 2, p-value = 0.00312
##
## Mean Absolute Deviation: 0.007597205
## Distortion Factor: -16.51803
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

Above result shows 5 largest discrepancies. As we can see from the plot, the highest deviation is 10.

From the log mantissa of the data, we can tell that the data follows Benford's Law. Because Mean closes to 0.5, Variance closes to 0.09, Ex. Kurtosis closes to -1.2, and Skewness closes to 0.

Degree of freedom equals 89 and p-value is small enough that we are supposed to reject the benford's law. X-squared value equals 6795 and stays away to the value of degree of freedom. Thus the dataset might have some problems by looking at these two values. The distribution of this data set looks good. Overall, this dataset should follow Benford's law.

The distortion factor is -16.51803.

```r
library(tidyverse)
library(knitr)
#Gets the the statistics of the first Digits of a benford object.
chicago_reviews <- getBfd(benford(chicago$reviews))
#From this table, we can get the distribution of dataset by first two digits.
kable(chicago_reviews[1:10, 1:6])
```

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
10	0.1044805	0.8427948	0.0413927	193	534
11	0.0324790	0.0043668	0.0377886	1	166
12	0.0309137	0.0000000	0.0347621	0	158
13	0.0244571	0.0000000	0.0321847	0	125
14	0.0211309	0.0000000	0.0299632	0	108
15	0.0207396	0.0043668	0.0280287	1	106
16	0.0205439	0.0000000	0.0263289	0	105
17	0.0205439	0.0000000	0.0248236	0	105
18	0.0183917	0.0000000	0.0234811	0	94

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
19	0.0176091	0.0000000	0.0222764	0	90

Table above shows the distribution of population data by first two digits.

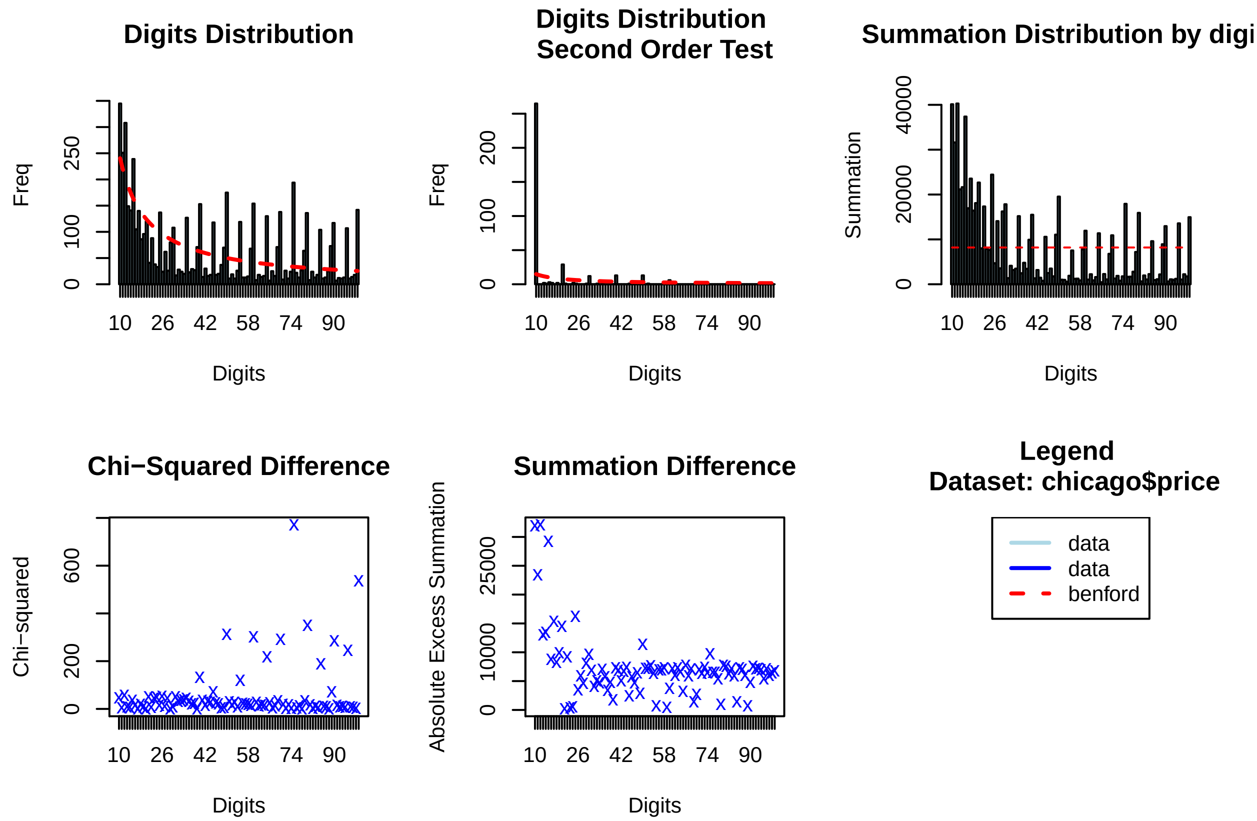
```
# Show ten suspected two digits that contain most discrepancies from Benford's law.
kable(head(suspectsTable(benford(chicago$reviews)),10))
```

digits	absolute.diff
10	322.44199
20	238.70149
40	229.19032
30	221.21712
50	169.04452
60	168.31025
70	155.51466
90	125.47291
80	117.42599
14	45.14203

Above table shows ten suspected two digits that contain most discrepancies from Benford's law.

Chicago price

```
plot(benford(chicago$price, number.of.digits = 2, sign = "positive", discrete=TRUE, round=3))
```



The original data is in blue and the expected frequency according to Benford's law is in red.

Benford's analysis of the first digits indicate the data basically follows Benford's Law.

Digit Distribution Second Order Test calculates the digit frequencies of the differences between the ordered (ranked) values in a data set. It shows that this dataset generally follows Benford's law, except some specific two digits.

The Chi-Square and Summation Difference plots almost fit Benford's law, but not good enough.

```
benford(chicago$price)
```

```
##
## Benford object:
##
## Data: chicago$price
## Number of observations used = 5811
## Number of obs. for second order = 359
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic Value
##         Mean  0.53
##          Var  0.10
##  Ex.Kurtosis -1.41
##     Skewness -0.17
##
##
```

```
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     75        160.57
## 2     50        125.02
## 3     99        116.64
## 4     60        112.29
## 5     12        106.00
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  chicago$price
## X-squared = 5310.9, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  chicago$price
## L2 = 0.046897, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.007206168
## Distortion Factor: 10.81249
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

Above result shows 5 largest discrepancies. As we can see from the plot, the highest deviation is 75.

From the log mantissa of the data, we can tell that the data follows Benford's Law. Because Mean closes to 0.5, Variance closes to 0.1, Ex. Kurtosis closes to -1.41, and Skewness closes to 0.

Degree of freedom equals 89 and p-value is small enough that we are supposed to reject the benford's law. X-squared value equals 5310.9 and stays away to the value of degree of freedom. Thus the dataset might have some problems by looking at these two values. The distribution of this data set looks good. Thus, the price seems not follow the Benford distribution very well.

The distortion factor is 10.81249.

```r
library(tidyverse)
library(knitr)
#Gets the the statistics of the first Digits of a benford object.
chicago_prices <- getBfd(benford(chicago$price))
#From this table, we can get the distribution of dataset by first two digits.
kable(chicago_prices[1:10, 1:6])
```

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
10	0.0593702	0.7381616	0.0413927	265	345
11	0.0431939	0.0000000	0.0377886	0	251
12	0.0530029	0.0000000	0.0347621	0	308
13	0.0256410	0.0055710	0.0321847	2	149
14	0.0242643	0.0027855	0.0299632	1	141
15	0.0411289	0.0083565	0.0280287	3	239
16	0.0180692	0.0055710	0.0263289	2	105
17	0.0240922	0.0000000	0.0248236	0	140
18	0.0147995	0.0055710	0.0234811	2	86

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
19	0.0165204	0.0000000	0.0222764	0	96

Table above shows the distribution of population data by first two digits.

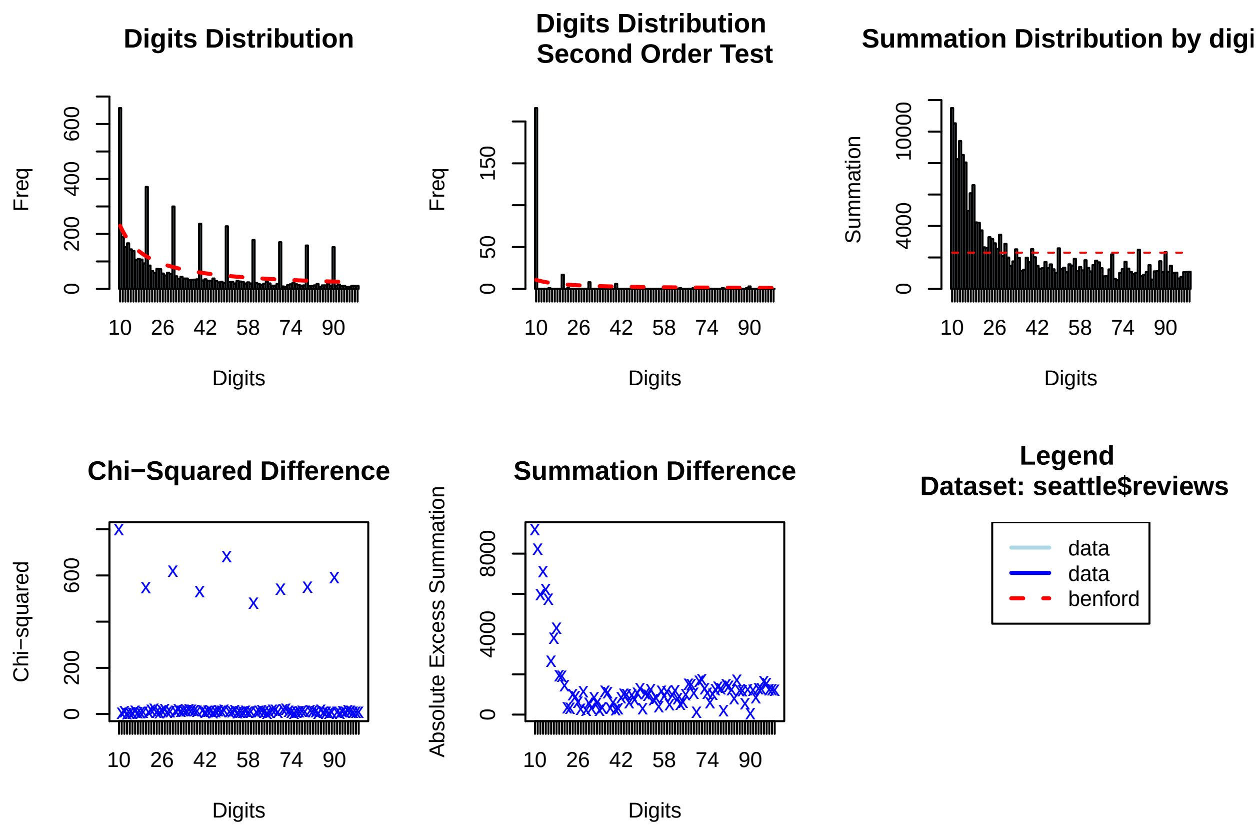
```
# Show ten suspected two digits that contain most discrepancies from Benford's law.
kable(head(suspectsTable(benford(chicago$price)),10))
```

digits	absolute.diff
75	160.57322
50	125.02440
99	116.63612
60	112.28524
12	105.99740
80	104.64947
10	104.46711
70	102.20245
65	91.46971
40	90.68362

Above table shows ten suspected two digits that contain most discrepancies from Benford's law.

Seattle review

```
library(benford.analysis)
plot(benford(seattle$reviews, number.of.digits = 2, sign = "positive", discrete=TRUE, round=3))
```



The original data is in blue and the expected frequency according to Benford's law is in red.

Benford's analysis of the first digits indicate the data basically follows Benford's Law.

Digit Distribution Second Order Test calculates the digit frequencies of the differences between the ordered (ranked) values in a data set. It shows that this dataset generally follows Benford's law, except some specific two digits.

The Chi-Square and Summation Difference plots almost fit Benford's law, but not good enough.

```
benford(seattle$reviews)
```

```
##
## Benford object:
##
## Data: seattle$reviews
## Number of observations used = 5549
## Number of obs. for second order = 262
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.451
##          Var  0.094
##  Ex.Kurtosis -1.246
##     Skewness  0.063
##
```

```
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     10        428.31
## 2     20        253.42
## 3     30        220.98
## 4     50        180.28
## 5     40        177.49
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  seattle$reviews
## X-squared = 6122.8, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  seattle$reviews
## L2 = 0.0056979, df = 2, p-value = 1.856e-14
##
## Mean Absolute Deviation: 0.007160048
## Distortion Factor: -14.72944
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

Above result shows 5 largest discrepancies. As we can see from the plot, the highest deviation is 10.

From the log mantissa of the data, we can tell that the data follows Benford's Law. Because Mean closes to 0.5, Variance closes to 0.09, Ex. Kurtosis closes to -1.2, and Skewness closes to 0.

Degree of freedom equals 89 and p-value is small enough that we are supposed to reject the benford's law. X-squared value equals 6122.8 and stays away to the value of degree of freedom. Thus the dataset might have some problems by looking at these two values. The distribution of this data set looks good. Overall, this dataset should follow Benford's law.

The distortion factor is -14.72944.

```r
library(tidyverse)
library(knitr)
#Gets the the statistics of the first Digits of a benford object.
seattle_reviews <- getBfd(benford(seattle$reviews))
#From this table, we can get the distribution of dataset by first two digits.
kable(seattle_reviews[1:10, 1:6])
```

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
10	0.1185799	0.8244275	0.0413927	216	658
11	0.0342404	0.0000000	0.0377886	0	190
12	0.0273923	0.0000000	0.0347621	0	152
13	0.0299153	0.0000000	0.0321847	0	166
14	0.0259506	0.0000000	0.0299632	0	144
15	0.0248693	0.0038168	0.0280287	1	138
16	0.0191025	0.0000000	0.0263289	0	106
17	0.0196432	0.0000000	0.0248236	0	109

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
18	0.0192828	0.0000000	0.0234811	0	107
19	0.0167598	0.0000000	0.0222764	0	93

Table above shows the distribution of population data by first two digits.

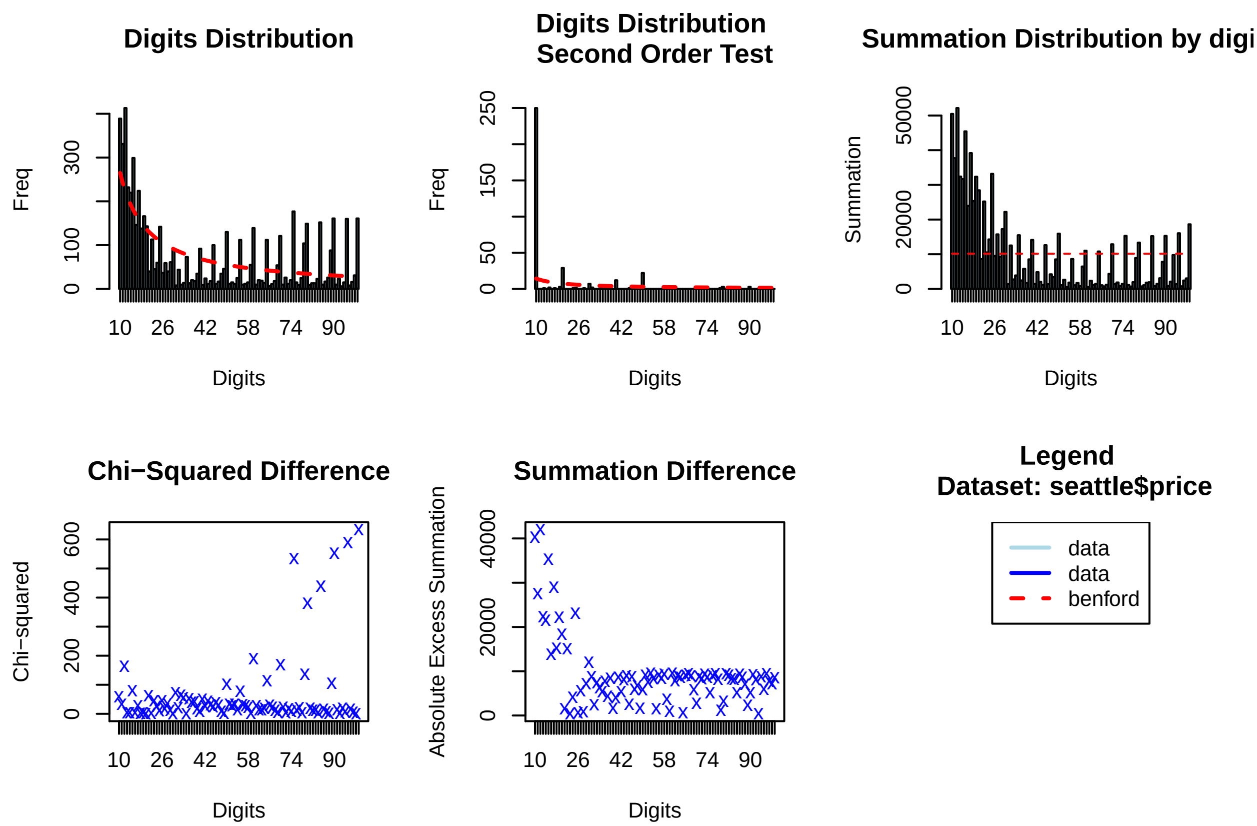
```
# Show ten suspected two digits that contain most discrepancies from Benford's law.
kable(head(suspectsTable(benford(seattle$reviews)),10))
```

digits	absolute.diff
10	428.31199
20	253.42058
30	220.97980
50	180.27765
40	177.49327
60	138.16603
70	135.81645
80	128.06297
90	125.37100
23	44.56442

Above table shows ten suspected two digits that contain most discrepancies from Benford's law.

Seattle price

```
plot(benford(seattle$price, number.of.digits = 2, sign = "positive", discrete=TRUE, round=3))
```



The original data is in blue and the expected frequency according to Benford's law is in red.

Benford's analysis of the first digits indicate the data basically follows Benford's Law.

Digit Distribution Second Order Test calculates the digit frequencies of the differences between the ordered (ranked) values in a data set. It shows that this dataset generally follows Benford's law, except some specific two digits.

The Chi-Square and Summation Difference plots almost fit Benford's law, but not good enough.

```
benford(seattle$price)
```

```
##
## Benford object:
##
## Data: seattle$price
## Number of observations used = 6399
## Number of obs. for second order = 346
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic Value
##         Mean  0.49
##          Var  0.11
##  Ex.Kurtosis -1.54
##     Skewness  0.08
##
```

```
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     12        190.56
## 2     75        140.19
## 3     99        133.07
## 4     95        130.90
## 5     90        130.29
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  seattle$price
## X-squared = 5819.3, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  seattle$price
## L2 = 0.086697, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.007187829
## Distortion Factor: 5.86225
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

Above result shows 5 largest discrepancies. As we can see from the plot, the highest deviation is 12.

From the log mantissa of the data, we can tell that the data follows Benford's Law. Because Mean closes to 0.5, Variance closes to 0.11, Ex. Kurtosis closes to -1.54, and Skewness closes to 0.

Degree of freedom equals 89 and p-value is small enough that we are supposed to reject the benford's law. X-squared value equals 5819.3 and stays away to the value of degree of freedom. Thus the dataset might have some problems by looking at these two values. The distribution of this data set looks good. Thus, the price seems not follow the Benford distribution very well.

The distortion factor is 5.86225.

```r
library(tidyverse)
library(knitr)
#Gets the the statistics of the first Digits of a benford object.
seattle_price <- getBfd(benford(seattle$price))
#From this table, we can get the distribution of dataset by first two digits.
kable(seattle_price[1:10, 1:6])
```

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
10	0.0607907	0.7225434	0.0413927	250	389
11	0.0517268	0.0000000	0.0377886	0	331
12	0.0645413	0.0000000	0.0347621	0	413
13	0.0362557	0.0028902	0.0321847	1	232
14	0.0343804	0.0000000	0.0299632	0	220
15	0.0467261	0.0057803	0.0280287	2	299
16	0.0228161	0.0000000	0.0263289	0	146
17	0.0350055	0.0028902	0.0248236	1	224

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
18	0.0215659	0.0000000	0.0234811	0	138
19	0.0259416	0.0086705	0.0222764	3	166

Table above shows the distribution of population data by first two digits.

```
# Show ten suspected two digits that contain most discrepancies from Benford's law.
kable(head(suspectsTable(benford(seattle$price)),10))
```

digits	absolute.diff
12	190.55728
75	140.19085
99	133.06961
95	130.89973
90	130.29195
10	124.12821
15	119.64420
85	119.49612
80	114.47719
60	93.06424

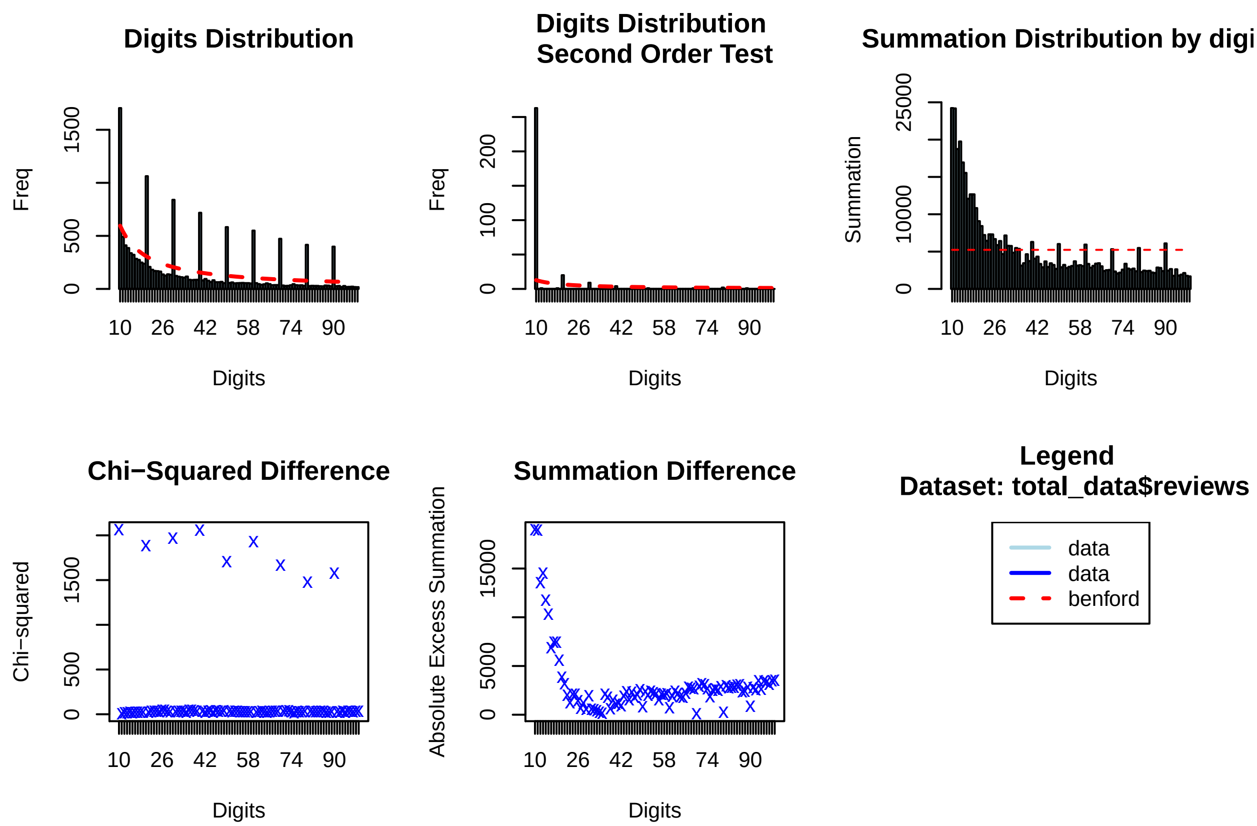
Above table shows ten suspected two digits that contain most discrepancies from Benford's law.

Total

```
boston_new <- boston[,-14]
total_data <- rbind(boston_new,chicago,seattle)
```

Review

```
plot(benford(total_data$reviews))
```




```
benford(total_data$reviews)
```

```
##
## Benford object:
##
## Data: total_data$reviews
## Number of observations used = 14396
## Number of obs. for second order = 309
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.450
##          Var  0.092
##  Ex.Kurtosis -1.226
##     Skewness  0.046
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     10       1109.11
## 2     20        757.96
## 3     30        634.99
## 4     40        563.62
```

```
## 5     50        459.19
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  total_data$reviews
## X-squared = 18631, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  total_data$reviews
## L2 = 0.0032038, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.007755391
## Distortion Factor: -16.49611
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

Above result shows 5 largest discrepancies. As we can see from the plot, the highest deviation is 10. The order looks like Benford analysis (10<20<30<40<50)

From the log mantissa of the data, we can tell that the data follows Benford's Law. Because Mean closes to 0.5, Variance closes to 0.11, Ex. Kurtosis closes to -1.3, and Skewness closes to 0.

Degree of freedom equals 89 and p-value is small enough that we are supposed to reject the benford's law. X-squared value equals 18631 and stays away to the value of degree of freedom. Thus the dataset might have some problems by looking at these two values. The distribution of this data set looks good. Overall, the reviews should follow the Benford distribution.

The distortion factor is -16.49611.

```r
library(tidyverse)
library(knitr)
#Gets the the statistics of the first Digits of a benford object.
total_reviews <- getBfd(benford(total_data$reviews))
#From this table, we can get the distribution of dataset by first two digits.
kable(total_reviews[1:10, 1:6])
```

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
10	0.1184357	0.8511327	0.0413927	263	1705
11	0.0341762	0.0000000	0.0377886	0	492
12	0.0285496	0.0032362	0.0347621	1	411
13	0.0268130	0.0000000	0.0321847	0	386
14	0.0234093	0.0000000	0.0299632	0	337
15	0.0223673	0.0000000	0.0280287	0	322
16	0.0197277	0.0000000	0.0263289	0	284
17	0.0190331	0.0000000	0.0248236	0	274
18	0.0172965	0.0032362	0.0234811	1	249
19	0.0165324	0.0000000	0.0222764	0	238

Table above shows the distribution of population data by first two digits.

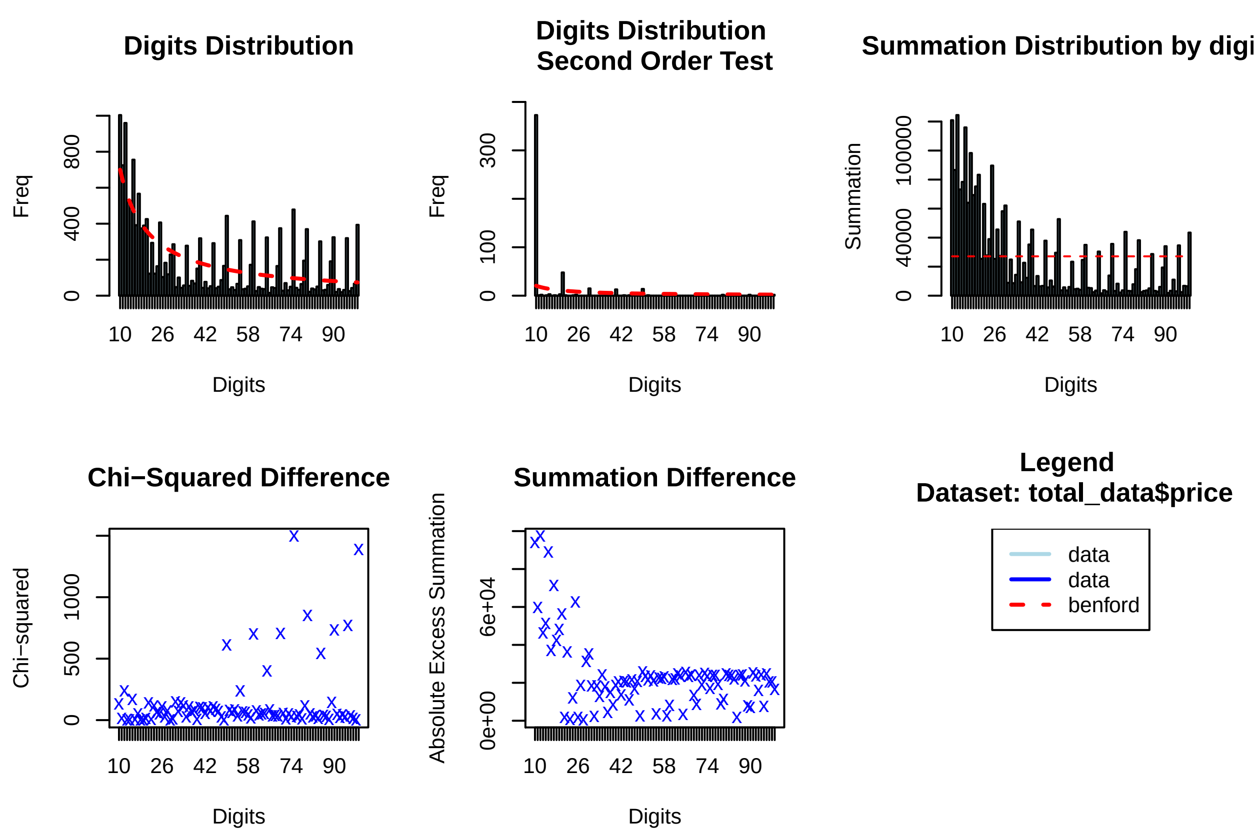
```r
# Show ten suspected two digits that contain most discrepancies from Benford's law.
kable(head(suspectsTable(benford(total_data$reviews)),10))
```

digits	absolute.diff
10	1109.1109
20	757.9589
30	634.9946
40	563.6192
50	459.1919
60	446.6571
70	384.3162
80	338.3331
90	329.9153
27	101.3743

Above table shows ten suspected two digits that contain most discrepancies from Benford's law.

Price

```r
library(benford.analysis)
plot(benford(total_data$price, number.of.digits = 2, sign = "positive", discrete=TRUE, round=3))
```



```r
benford(total_data$price)
```

```
##
## Benford object:
##
## Data: total_data$price
## Number of observations used = 16914
## Number of obs. for second order = 495
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.497
##          Var  0.105
##  Ex.Kurtosis -1.455
##     Skewness  0.028
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     75        381.71
## 2     12        372.03
## 3     99        320.17
## 4     10        303.88
## 5     50        298.54
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  total_data$price
## X-squared = 12826, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  total_data$price
## L2 = 0.04284, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.006653423
## Distortion Factor: 4.522246
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

Above result shows 5 largest discrepancies. As we can see from the plot, the highest deviation is 75. The order does not look like Benford distribution.

From the log mantissa of the data, we can tell that the data follows Benford's Law. Because Mean closes to 0.5, Variance closes to 0.1, Ex. Kurtosis closes to -1.4, and Skewness closes to 0.

Degree of freedom equals 89 and p-value is small enough that we are supposed to reject the benford's law. X-squared value equals 12826 and stays away to the value of degree of freedom. Thus the dataset might have some problems by looking at these two values. Overall, We conclude that the prices does not follow the

Benford distribution very well.

The distortion factor is 4.522246.

```
library(tidyverse)
library(knitr)
#Gets the the statistics of the first Digits of a benford object.
total_prices <- getBfd(benford(total_data$reviews))
#From this table, we can get the distribution of dataset by first two digits.
kable(total_prices[1:10, 1:6])
```

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
10	0.1184357	0.8511327	0.0413927	263	1705
11	0.0341762	0.0000000	0.0377886	0	492
12	0.0285496	0.0032362	0.0347621	1	411
13	0.0268130	0.0000000	0.0321847	0	386
14	0.0234093	0.0000000	0.0299632	0	337
15	0.0223673	0.0000000	0.0280287	0	322
16	0.0197277	0.0000000	0.0263289	0	284
17	0.0190331	0.0000000	0.0248236	0	274
18	0.0172965	0.0032362	0.0234811	1	249
19	0.0165324	0.0000000	0.0222764	0	238

Table above shows the distribution of population data by first two digits.

```
# Show ten suspected two digits that contain most discrepancies from Benford's law.
kable(head(suspectsTable(benford(total_data$price)),10))
```

digits	absolute.diff
75	381.7051
12	372.0337
99	320.1737
10	303.8841
50	298.5367
60	291.5814
15	281.9222
80	278.7484
70	270.8045
90	243.8317

Above table shows ten suspected two digits that contain most discrepancies from Benford's law.

Conclusion

By the EDA of all three cities, we can see that there are a large amount of zeros for reviews suggesting that host needs to find a way to encourage their guests to give the feedback on the website. The rating and reviews are correlated in a positive way. The Entire home/apt is a majority type of Airbnb house. Also, it concludes more bedrooms and higher price. On the other hand, the amount of reviews doese not affect price very much.

For the Benford analysis part, I analyze two variables in each of cities: Review and price.

By looking at the five basic graph, it looks like that they all follow the Benford distribution; however, when we look at the details of the data, we can find something different. The overall trend of the review is in a good shape. The largest deviation always start with the smallest number. On the other hand, the graph of the price looks a bit more away with the Benford distribution. The largest deviations always start with a much larger number suggesting that it does not follow the distribution very well. The results of the total data which combines three cities also support this conclusion. Thus, in this dataset, reviews are good to trust, but we need to think more while looking at the price