

BSE658A-Summary-05

Abin Thomas (Roll No. 21218261)

2022-09-05

```
library(lsr)
library(tidyverse)

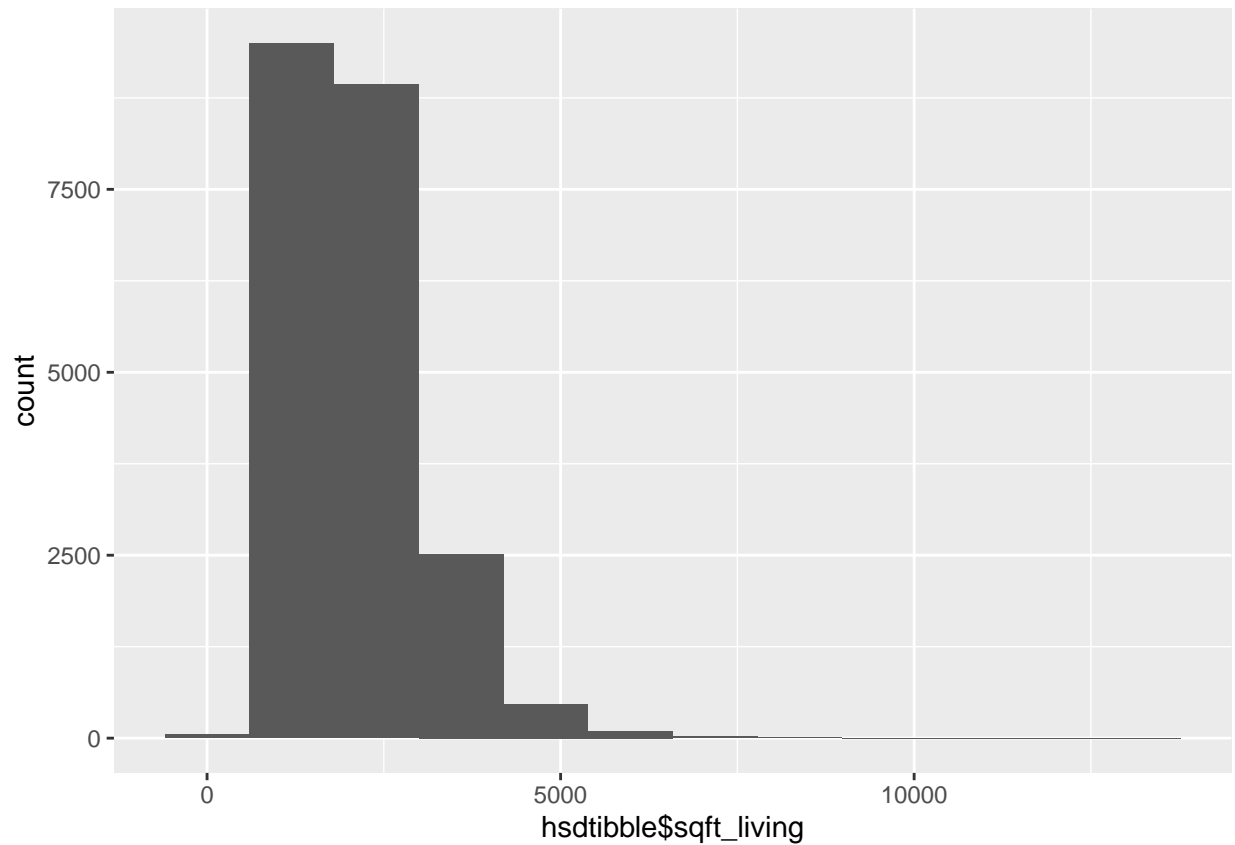
## -- Attaching packages ----- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr  0.3.4
## v tibble  3.1.8      v dplyr  1.0.9
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

path <- "D:/Abin/Documents/BSE658A_BioStats/kc_house_data.csv"
kc_housesales_data <- read.csv(path)
#kc_housesales_data

hsdtibble <- as_tibble(kc_housesales_data)
#hsdtibble

df <- data.frame(hsdtibble$sqft_living)

ggplot(df, aes(x = hsdtibble$sqft_living)) + geom_histogram(bins=12)
```



```
library(psych)
```

```
##  
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':  
##  
##    %+%, alpha
```

```
#to find skewness of the data  
skew(hsdtibble$sqft_living)
```

```
## [1] 1.473011
```

```
#Measures of Central Tendency
```

```
mean(hsdtibble$sqft_living)
```

```
## [1] 2080.322
```

```
median(hsdtibble$sqft_living)
```

```
## [1] 1910
```

```
modeOf(hsdtibble$sqft_living)
```

```
## [1] 1300
```

```
maxFreq(hsdtibble$sqft_living)
```

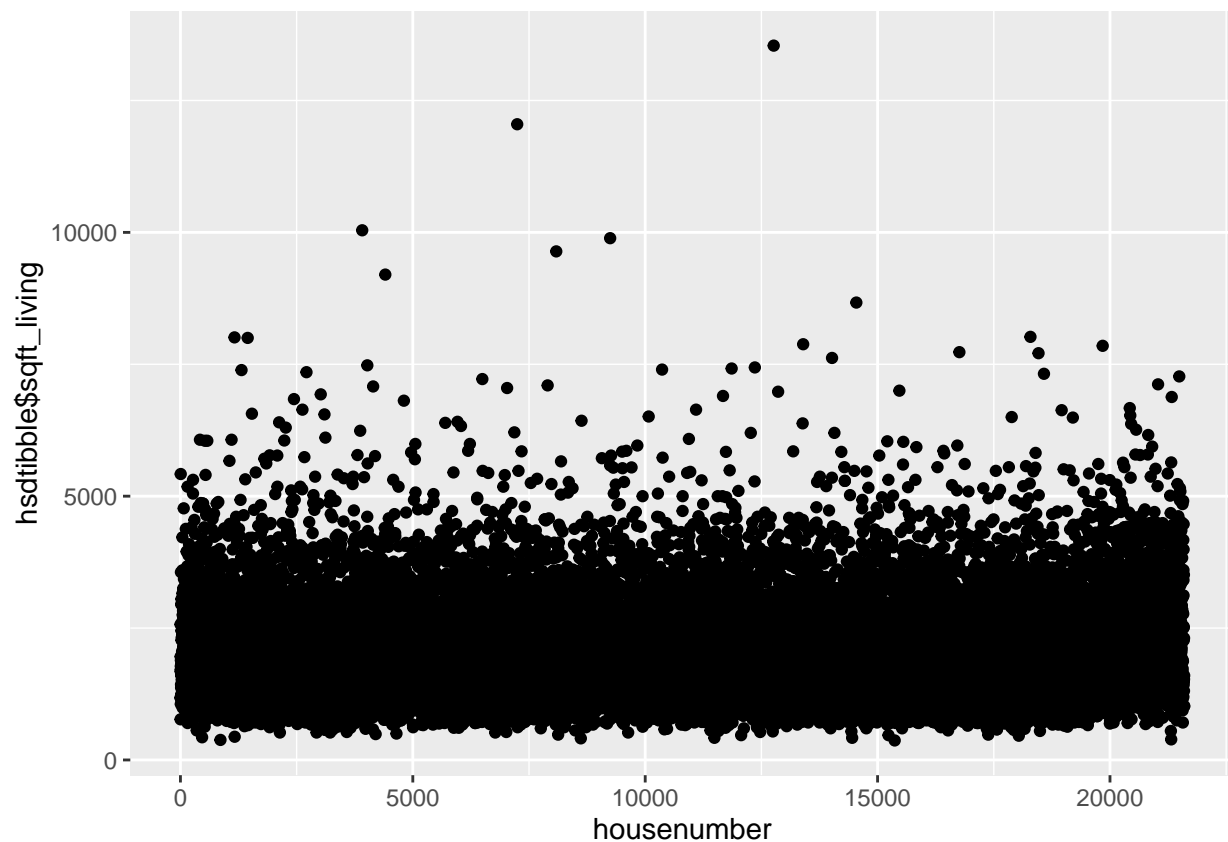
```
## [1] 138
```

```
#Measures of variability
```

```
library(dplyr)
```

```
df <- mutate(df, housenumber= 1:length(hsdtibble$sqft_living))
```

```
ggplot(df, aes(y = hsdtibble$sqft_living, x = housenumber)) + geom_point()
```



```
range(hsdtibble$sqft_living)
```

```
## [1] 370 13540
```

```
#iqr
```

```
quantile(x = hsdtibble$sqft_living, probs = c(.25,.75)) #second
```

```
## 25% 75%
```

```
## 1430 2550
```

```
#variance  
var(hsdtibble$sqft_living)
```

```
## [1] 842918.9
```

```
sd(hsdtibble$sqft_living)
```

```
## [1] 918.1061
```

```
#let's check ifessel's correction makes a difference
```

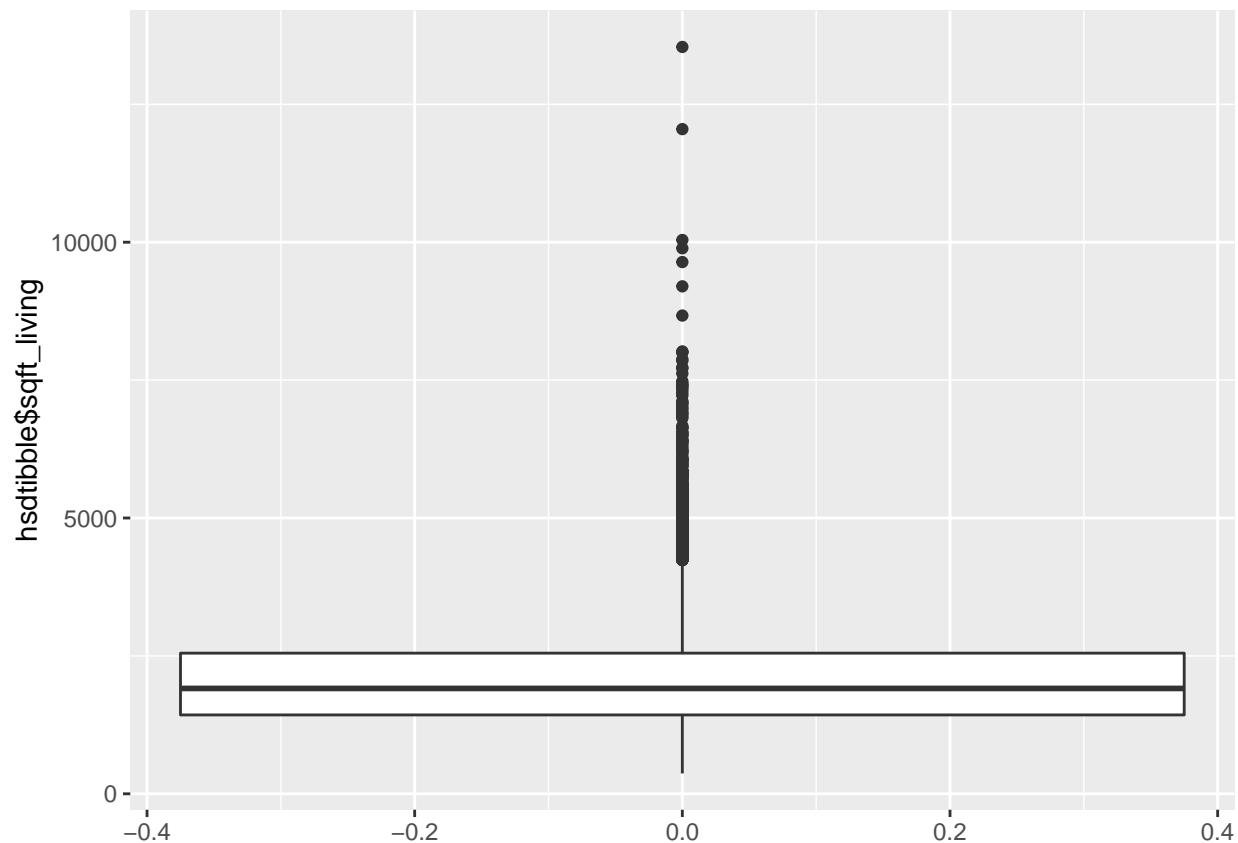
```
sdn <- function(x) {  
  return(sqrt(mean((x - mean(x))^2)))  
}  
sdn(hsdtibble$sqft_living)
```

```
## [1] 918.0849
```

```
summary(hsdtibble$sqft_living)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
##      370   1430   1910   2080   2550   13540
```

```
#boxplot?  
ggplot(df, aes(x=hsdtibble$sqft_living, )) +  
  geom_boxplot()+ coord_flip()
```



```
#let's try the summary of whole dataset
summary(hdsdtibble)
```

```
##      id          date          price          bedrooms
## Min.   :1.000e+06  Length:21597   Min.    : 78000   Min.    : 1.000
## 1st Qu.:2.123e+09   Class :character  1st Qu.: 322000   1st Qu.: 3.000
## Median :3.905e+09   Mode  :character  Median : 450000   Median : 3.000
## Mean   :4.580e+09                      Mean   : 540297   Mean   : 3.373
## 3rd Qu.:7.309e+09                      3rd Qu.: 645000   3rd Qu.: 4.000
## Max.   :9.900e+09                      Max.    :7700000   Max.    :33.000
##  bathrooms    sqft_living    sqft_lot    floors
## Min.    :0.500   Min.    : 370   Min.    : 520   Min.    :1.000
## 1st Qu.:1.750   1st Qu.: 1430   1st Qu.: 5040   1st Qu.:1.000
## Median :2.250   Median : 1910   Median : 7618   Median :1.500
## Mean    :2.116   Mean    : 2080   Mean    : 15099   Mean    :1.494
## 3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.: 10685   3rd Qu.:2.000
## Max.    :8.000   Max.    :13540   Max.    :1651359   Max.    :3.500
##  waterfront    view          condition    grade
## Min.    :0.000000   Min.    :0.0000   Min.    :1.00   Min.    : 3.000
## 1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.00   1st Qu.: 7.000
## Median :0.000000   Median :0.0000   Median :3.00   Median : 7.000
## Mean    :0.007547   Mean    :0.2343   Mean    :3.41   Mean    : 7.658
## 3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.00   3rd Qu.: 8.000
## Max.    :1.000000   Max.    :4.0000   Max.    :5.00   Max.    :13.000
##  sqft_above    sqft_basement    yr_built    yr_renovated
```

```
## Min. : 370 Min. : 0.0 Min. :1900 Min. : 0.00
## 1st Qu.:1190 1st Qu.: 0.0 1st Qu.:1951 1st Qu.: 0.00
## Median :1560 Median : 0.0 Median :1975 Median : 0.00
## Mean :1789 Mean : 291.7 Mean :1971 Mean : 84.46
## 3rd Qu.:2210 3rd Qu.: 560.0 3rd Qu.:1997 3rd Qu.: 0.00
## Max. :9410 Max. :4820.0 Max. :2015 Max. :2015.00
## zipcode lat long sqft_living15
## Min. :98001 Min. :47.16 Min. : -122.5 Min. : 399
## 1st Qu.:98033 1st Qu.:47.47 1st Qu.: -122.3 1st Qu.:1490
## Median :98065 Median :47.57 Median : -122.2 Median :1840
## Mean :98078 Mean :47.56 Mean : -122.2 Mean :1987
## 3rd Qu.:98118 3rd Qu.:47.68 3rd Qu.: -122.1 3rd Qu.:2360
## Max. :98199 Max. :47.78 Max. : -121.3 Max. :6210
## sqft_lot15
## Min. : 651
## 1st Qu.: 5100
## Median : 7620
## Mean : 12758
## 3rd Qu.: 10083
## Max. :871200
```

```
describe(hsdtibble)
```

```
## vars n mean sd median trimmed
## id 1 21597 4580474287.77 2.876736e+09 3.90493e+09 4500243579.12
## date* 2 21597 197.73 1.046800e+02 2.01000e+02 199.56
## price 3 21597 540296.57 3.673681e+05 4.50000e+05 481824.89
## bedrooms 4 21597 3.37 9.300000e-01 3.00000e+00 3.34
## bathrooms 5 21597 2.12 7.700000e-01 2.25000e+00 2.08
## sqft_living 6 21597 2080.32 9.181100e+02 1.91000e+03 1984.75
## sqft_lot 7 21597 15099.41 4.141264e+04 7.61800e+03 8257.51
## floors 8 21597 1.49 5.400000e-01 1.50000e+00 1.45
## waterfront 9 21597 0.01 9.000000e-02 0.00000e+00 0.00
## view 10 21597 0.23 7.700000e-01 0.00000e+00 0.00
## condition 11 21597 3.41 6.500000e-01 3.00000e+00 3.30
## grade 12 21597 7.66 1.170000e+00 7.00000e+00 7.58
## sqft_above 13 21597 1788.60 8.277600e+02 1.56000e+03 1683.11
## sqft_basement 14 21597 291.73 4.426700e+02 0.00000e+00 205.50
## yr_built 15 21597 1971.00 2.938000e+01 1.97500e+03 1973.09
## yr_renovated 16 21597 84.46 4.018200e+02 0.00000e+00 0.00
## zipcode 17 21597 98077.95 5.351000e+01 9.80650e+04 98074.73
## lat 18 21597 47.56 1.400000e-01 4.75700e+01 47.57
## long 19 21597 -122.21 1.400000e-01 -1.22230e+02 -122.23
## sqft_living15 20 21597 1986.62 6.852300e+02 1.84000e+03 1914.20
## sqft_lot15 21 21597 12758.28 2.727444e+04 7.62000e+03 7901.45
## mad min max range skew kurtosis
## id 3.561991e+09 1000102.00 9900000190.00 9.899000e+09 0.24 -1.26
## date* 1.304700e+02 1.00 372.00 3.710000e+02 -0.15 -1.16
## price 2.223900e+05 78000.00 7700000.00 7.622000e+06 4.02 34.53
## bedrooms 1.480000e+00 1.00 33.00 3.200000e+01 2.02 49.81
## bathrooms 7.400000e-01 0.50 8.00 7.500000e+00 0.52 1.28
## sqft_living 8.006000e+02 370.00 13540.00 1.317000e+04 1.47 5.25
## sqft_lot 3.881450e+03 520.00 1651359.00 1.650839e+06 13.07 285.40
## floors 7.400000e-01 1.00 3.50 2.500000e+00 0.61 -0.49
```

```
## waterfront      0.000000e+00      0.00      1.00 1.000000e+00 11.38 127.49
## view            0.000000e+00      0.00      4.00 4.000000e+00  3.40  10.89
## condition       0.000000e+00      1.00      5.00 4.000000e+00  1.04   0.52
## grade           1.480000e+00      3.00     13.00 1.000000e+01  0.79   1.13
## sqft_above      6.671700e+02     370.00    9410.00 9.040000e+03  1.45   3.40
## sqft_basement   0.000000e+00      0.00     4820.00 4.820000e+03  1.58   2.71
## yr_built        3.410000e+01    1900.00    2015.00 1.150000e+02 -0.47  -0.66
## yr_renovated    0.000000e+00      0.00     2015.00 2.015000e+03  4.55  18.68
## zipcode         6.227000e+01   98001.00    98199.00 1.980000e+02  0.41  -0.85
## lat             1.600000e-01     47.16     47.78 6.200000e-01 -0.49  -0.68
## long            1.500000e-01    -122.52    -121.32 1.200000e+00  0.88   1.05
## sqft_living15   6.078700e+02     399.00    6210.00 5.811000e+03  1.11   1.59
## sqft_lot15      3.713910e+03     651.00   871200.00 8.705490e+05  9.52 151.35
##
## se
## id              19575066.73
## date*           0.71
## price           2499.80
## bedrooms        0.01
## bathrooms       0.01
## sqft_living     6.25
## sqft_lot        281.80
## floors          0.00
## waterfront      0.00
## view            0.01
## condition       0.00
## grade           0.01
## sqft_above      5.63
## sqft_basement   3.01
## yr_built        0.20
## yr_renovated    2.73
## zipcode         0.36
## lat             0.00
## long            0.00
## sqft_living15   4.66
## sqft_lot15      185.59
```

```
library(dplyr)
samplesz = c(4,40,400,800,1000,1600,2000)
sno = c(1:7)
sz = as_tibble_col(sno,column_name = "Sample Number")
sz <- mutate(sz,'Sample Size' = samplesz)
sz
```

```
## # A tibble: 7 x 2
##   'Sample Number' 'Sample Size'
##           <int>         <dbl>
## 1             1             4
## 2             2            40
## 3             3           400
## 4             4           800
## 5             5          1000
## 6             6          1600
## 7             7          2000
```

```
smean = c()
for (i in sz$`Sample Size`) {
  smean1 <- mean((sample_n(hsdtibble,i))$sqft_living)
  smean = c(smean,smean1)
}
```

```
Rsampletib <- sz %>% mutate('Sample Mean' = smean )
Popmean = c(rep(mean(hsdtibble$sqft_living),7))
Rsampletib <- Rsampletib %>% mutate('Population Mean' = Popmean )
Rsampletib
```

```
## # A tibble: 7 x 4
##   'Sample Number' 'Sample Size' 'Sample Mean' 'Population Mean'
##           <int>         <dbl>         <dbl>         <dbl>
## 1             1             4         1812.         2080.
## 2             2            40         1944         2080.
## 3             3           400         2141.         2080.
## 4             4           800         2130.         2080.
## 5             5          1000         2126.         2080.
## 6             6          1600         2046.         2080.
## 7             7          2000         2077.         2080.
```

```
samsd = c()
for (i in sz$`Sample Size`) {
  samsd1 <- sd((sample_n(hsdtibble,i))$sqft_living)
  samsd = c(samsd,samsd1)
}
```

```
Rsampletib <- Rsampletib %>% mutate('Sample Standard Deviation' = samsd )
Popsd = c(rep(sd(hsdtibble$sqft_living),7))
Rsampletib <- Rsampletib %>% mutate('Population Standard Deviation' = Popsd )
Rsampletib
```

```
## # A tibble: 7 x 6
##   'Sample Number' 'Sample Size' 'Sample Mean' 'Population Mean' Sample-1 Popul-2
##           <int>         <dbl>         <dbl>         <dbl>         <dbl>   <dbl>
## 1             1             4         1812.         2080.          605.    918.
## 2             2            40         1944         2080.          932.    918.
## 3             3           400         2141.         2080.          940.    918.
## 4             4           800         2130.         2080.         1028.    918.
## 5             5          1000         2126.         2080.          951.    918.
## 6             6          1600         2046.         2080.          945.    918.
## 7             7          2000         2077.         2080.          882.    918.
## # ... with abbreviated variable names 1: 'Sample Standard Deviation',
## # 2: 'Population Standard Deviation'
```