# ProjectEDA

*My Dinh*

*11/13/2017*

```
library(data.table)
library(ggplot2)
library(ggforce)
library(GGally)
library(car)
library(nortest)
library(MASS)
library(forecast)
```

```
## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2017c.
## 1.0/zoneinfo/America/Los_Angeles'
```

```
##
## Attaching package: 'forecast'
```

```
## The following object is masked from 'package:ggplot2':
##
##     autolayer
```

Nov 27: finish EDA: 2pm - 4:30 pm. baseline Dec1: after 2PM

when meet up; feature enginering: feature selection

Dec 4:

```
#rm(list = ls())
dir = "/Users/MyDinh/Downloads/Stat154/Projects"
setwd(dir)
df  = read.table("data/adult.data", sep = ",")
summary(df)
```

```
##        V1                         V2              V3
##  Min.   :17.00    Private         :22696   Min.   :  12285
##  1st Qu.:28.00    Self-emp-not-inc: 2541   1st Qu.: 117827
##  Median :37.00    Local-gov       : 2093   Median : 178356
##  Mean   :38.58    ?               : 1836   Mean   : 189778
##  3rd Qu.:48.00    State-gov       : 1298   3rd Qu.: 237051
##  Max.   :90.00    Self-emp-inc    : 1116   Max.   :1484705
##                   (Other)         :  981
##              V4              V5                              V6
##   HS-grad     :10501   Min.   : 1.00    Divorced             : 4443
##   Some-college: 7291   1st Qu.: 9.00    Married-AF-spouse    :   23
##   Bachelors   : 5355   Median :10.00    Married-civ-spouse   :14976
##   Masters     : 1723   Mean   :10.08    Married-spouse-absent:  418
##   Assoc-voc   : 1382   3rd Qu.:12.00    Never-married        :10683
##   11th        : 1175   Max.   :16.00    Separated            : 1025
##  (Other)      : 5134                    Widowed              :  993
##                 V7                     V8
```

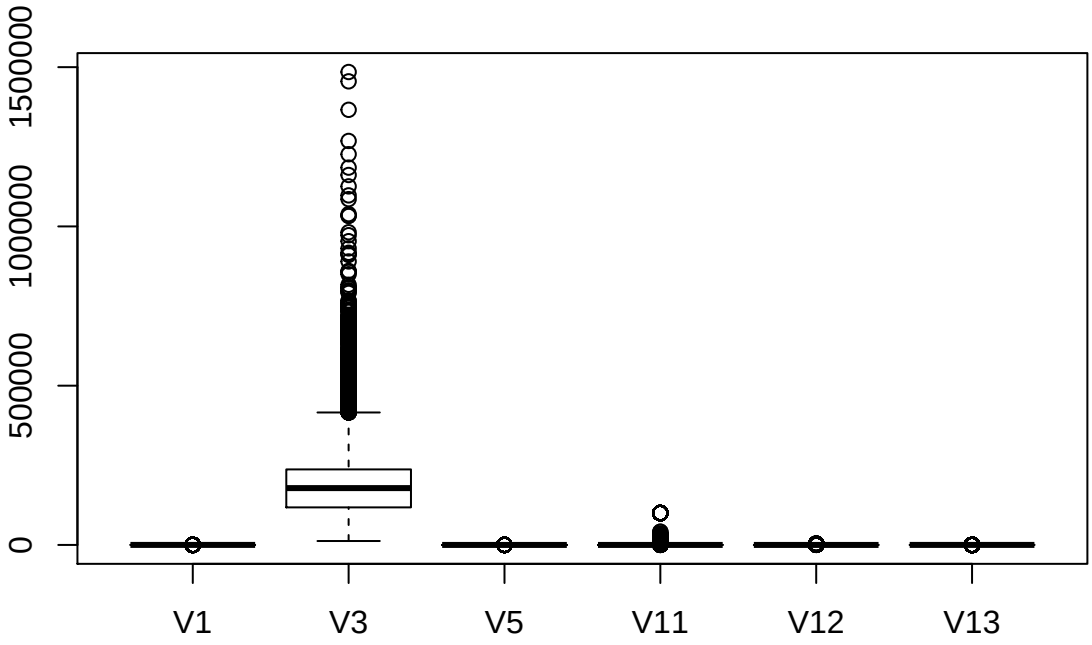
```
##   Prof-specialty :4140    Husband       :13193
##   Craft-repair   :4099    Not-in-family : 8305
##   Exec-managerial:4066    Other-relative:  981
##   Adm-clerical   :3770    Own-child     : 5068
##   Sales          :3650    Unmarried     : 3446
##   Other-service  :3295    Wife          : 1568
##  (Other)         :9541
##                    V9             V10             V11
##   Amer-Indian-Eskimo:  311    Female:10771   Min.   :    0
##   Asian-Pac-Islander: 1039    Male  :21790   1st Qu.:    0
##   Black             : 3124                   Median :    0
##   Other             :  271                   Mean   : 1078
##   White             :27816                   3rd Qu.:    0
##                                              Max.   :99999
##
##       V12              V13                    V14            V15
##  Min.   :   0.0   Min.   : 1.00    United-States:29170    <=50K:24720
##  1st Qu.:   0.0   1st Qu.:40.00    Mexico       :  643    >50K : 7841
##  Median :   0.0   Median :40.00    ?            :  583
##  Mean   :  87.3   Mean   :40.44    Philippines  :  198
##  3rd Qu.:   0.0   3rd Qu.:45.00    Germany      :  137
##  Max.   :4356.0   Max.   :99.00    Canada       :  121
##                                   (Other)      : 1709
```

```
str(df)
```

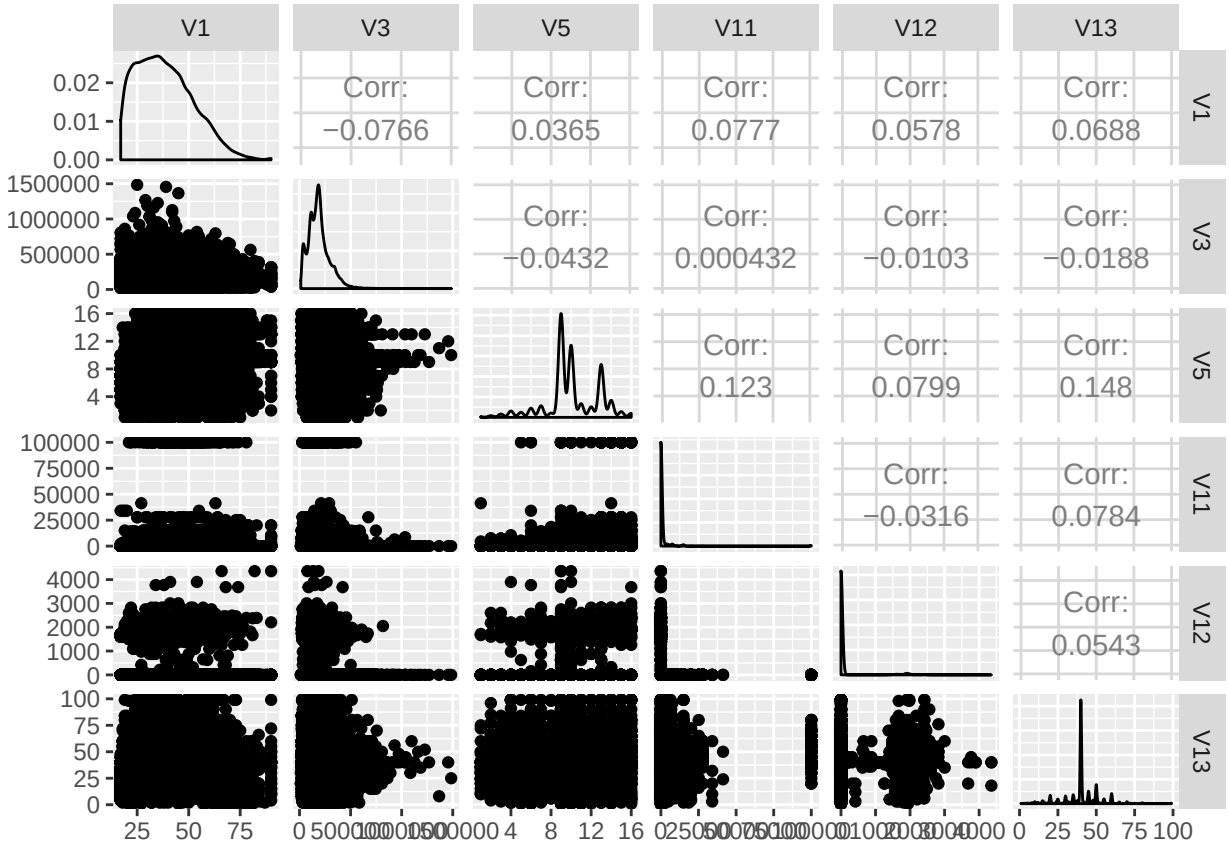
```
## 'data.frame':    32561 obs. of  15 variables:
##  $ V1 : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ V2 : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ V3 : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ V4 : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ V5 : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ V6 : Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ V7 : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ V8 : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
##  $ V9 : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ V10: Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ V11: int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ V12: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V13: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ V14: Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
##  $ V15: Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

### Data Transformation

```
### Let's check the box plot:
df_cont = df[, c(1,3, 5, 11:13)]
boxplot(df_cont)
```




```
### check the distribution plot


ggpairs(df_cont)
```




```
### V1 and V3 and right-skewed distribution
### V11 and V12 hav heavy tail distribution
```

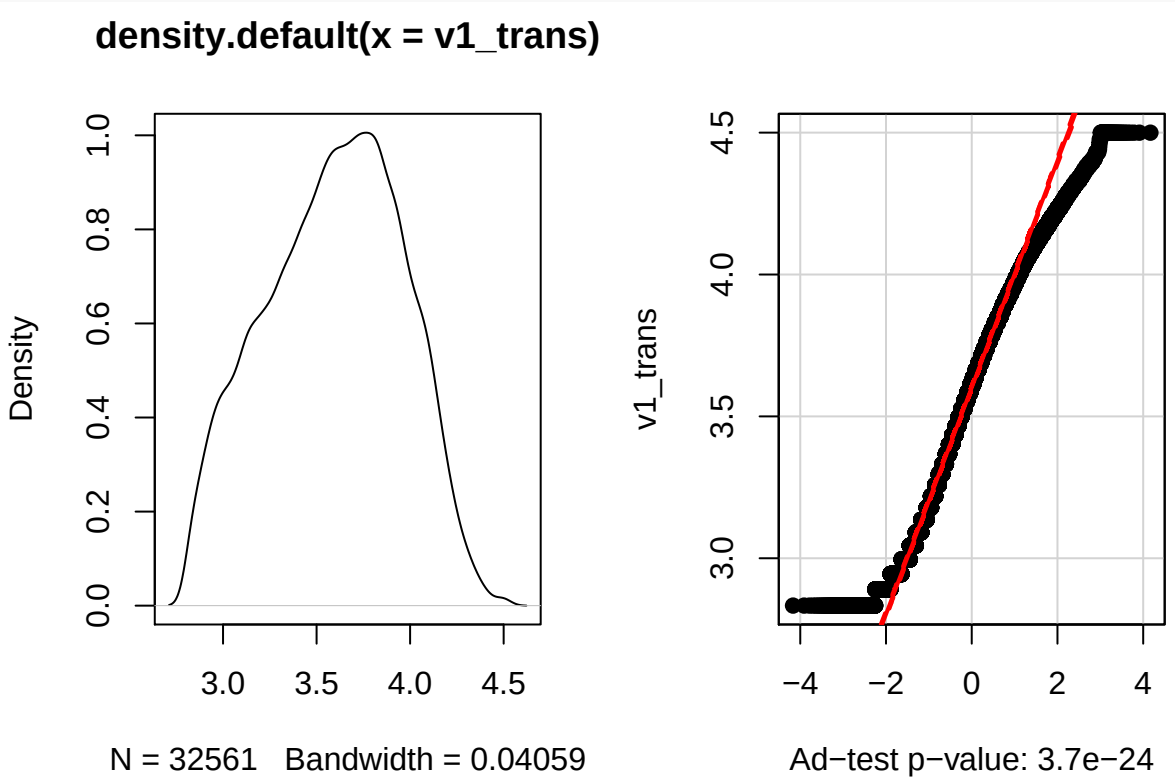
```
##
```

We observed:

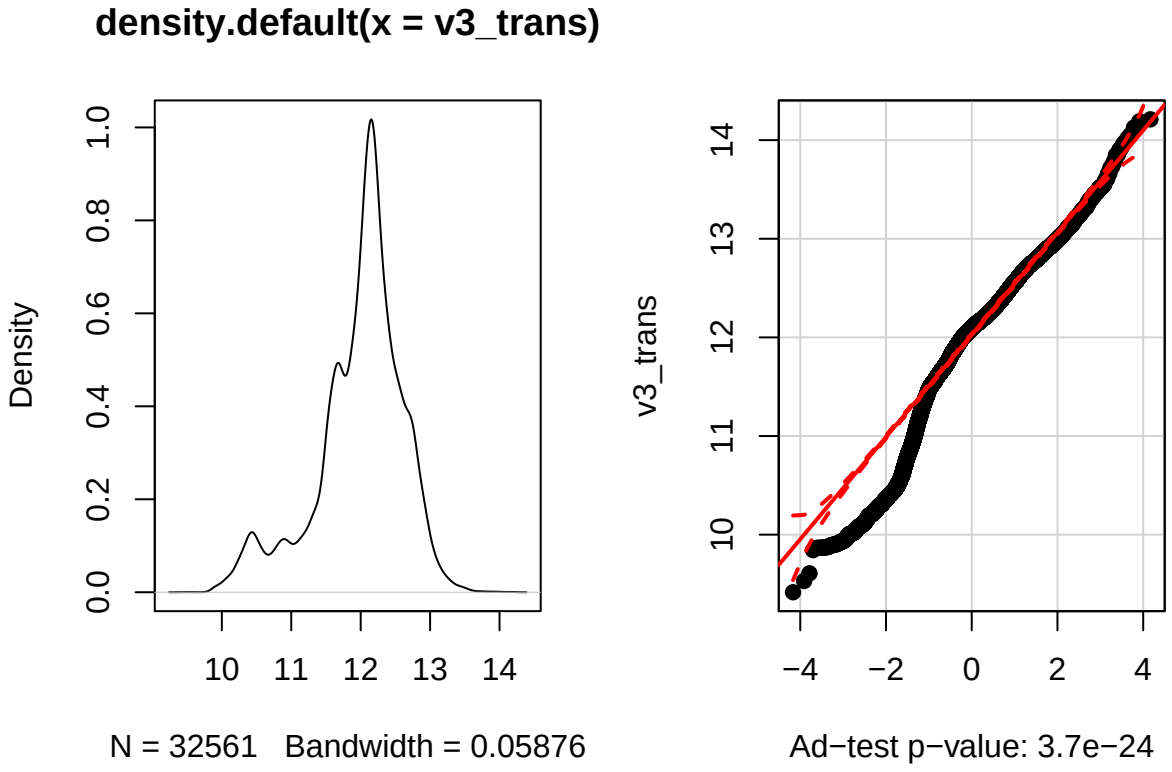
- V1 and V3 can achieve normality by performaing log transformation.
- V5 and V13 may need more work than regular log transformation. We can try power transfrmation family.
- V11 and V12 have heavy tails, which requires more research. One suggestion is to use inverse hyperbolic sine transformation which takes care of zero.

Let's try to transform the distribution:

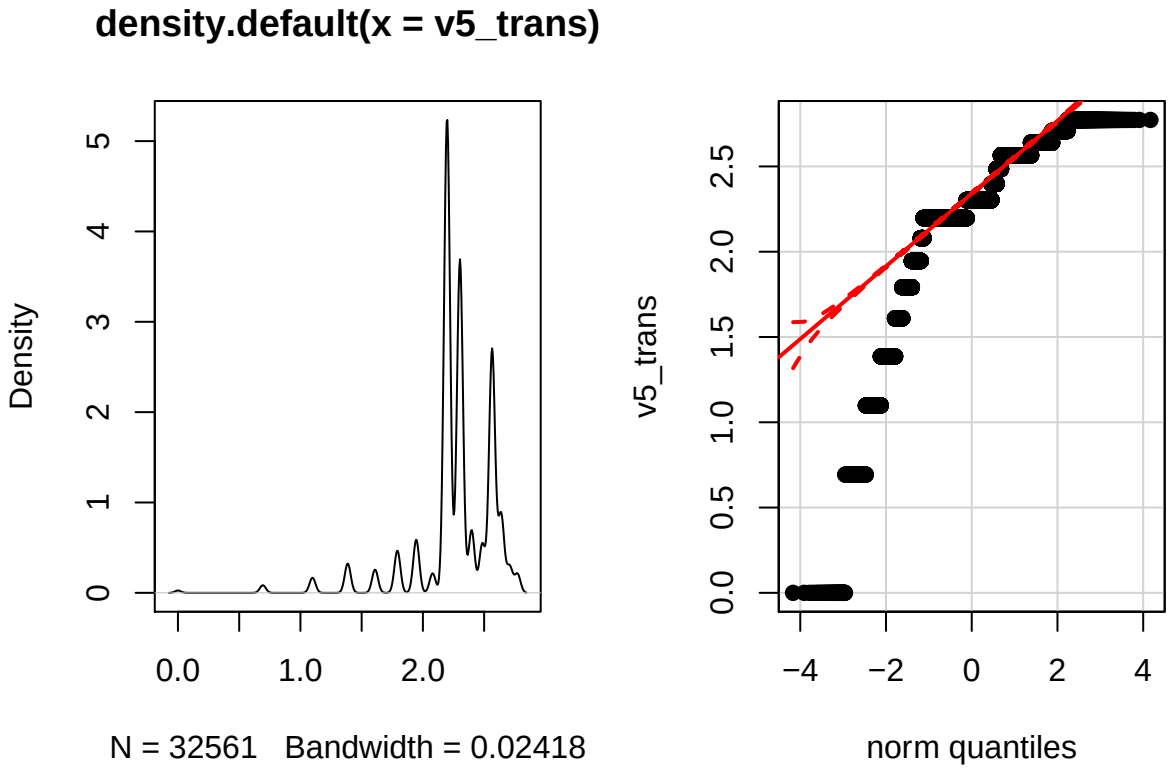
```
df_transform = cbind(df)


par(mfrow= c(1,2))
## try log transform for V1
v1_trans = log(df_transform$V1)
plot(density(v1_trans))
### qqplot with normality test
qqPlot(v1_trans, dist= "norm", col=palette()[1],xlab = paste0("Ad-test p-value: ", ad.test(v1_trans)$p.v
```



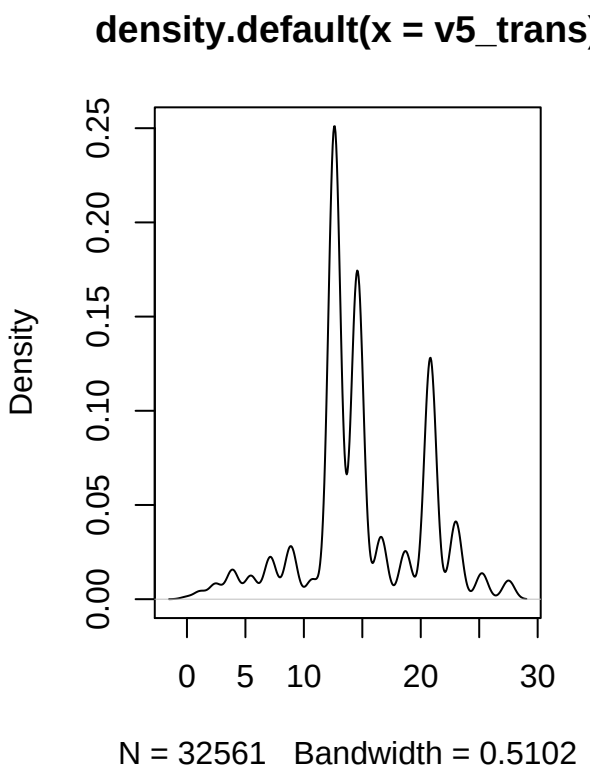

```
v3_trans = log(df_transform$V3)
plot(density(v3_trans))
qqPlot(v3_trans, dist= "norm", col=palette()[1],xlab = paste0("Ad-test p-value: ", ad.test(v3_trans)$p.v
```

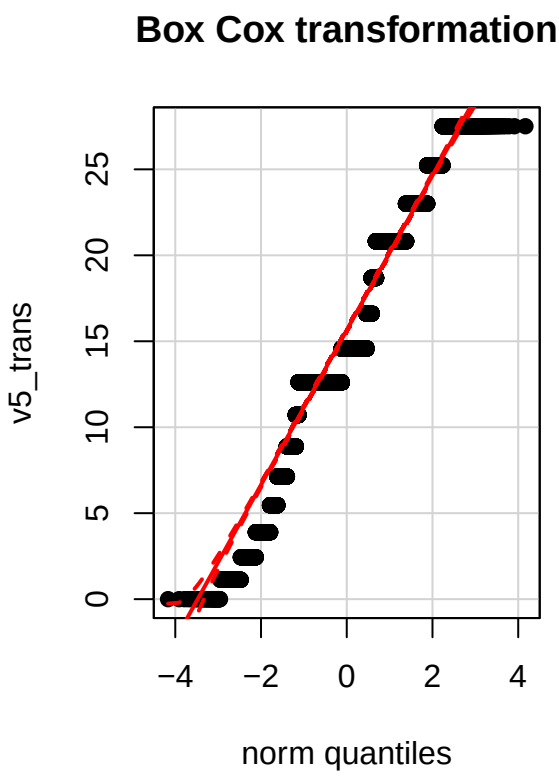



```
v5_trans = log(df_transform$V5)
plot(density(v5_trans))
qqPlot(v5_trans, dist= "norm", col=palette()[1], pch = 19)
```



```
### Let's try box cox transformation:
## find optimal lambda
lam_5= BoxCox.lambda(df_transform$V5, method = "loglik")
v5_trans = BoxCox(df$V5, lam_5)
plot(density(v5_trans))
qqPlot(v5_trans, dist= "norm", col=palette()[1], pch = 19, main = "Box Cox transformation")
```

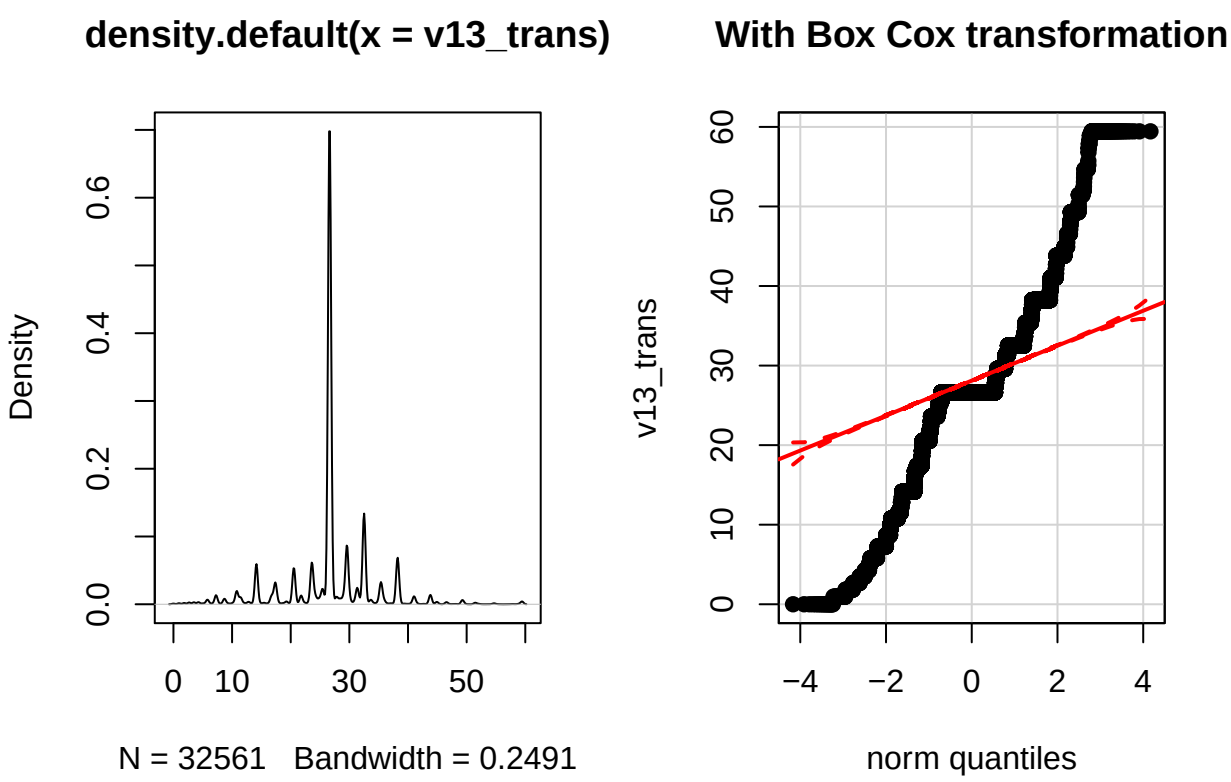


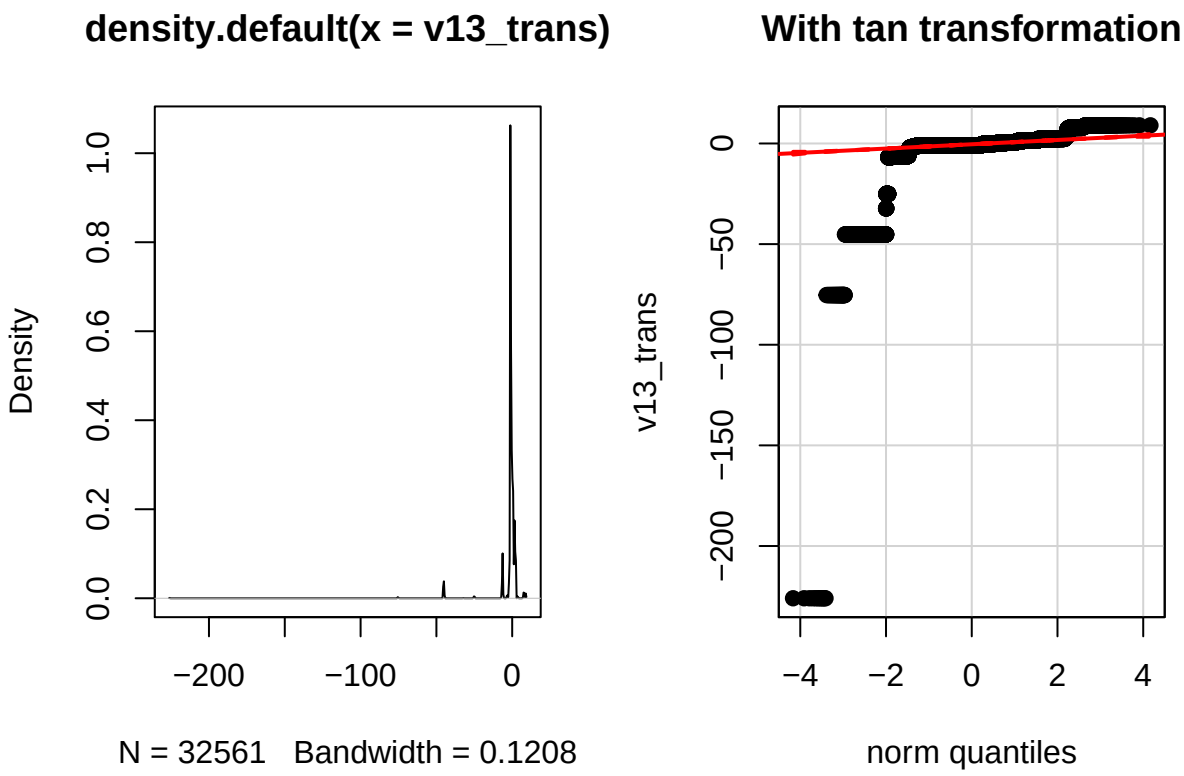

```
# it works perfectly for v5



lam_13 = BoxCox.lambda(df_transform$V13)
lam_13
```

```
## [1] 0.8605036
```

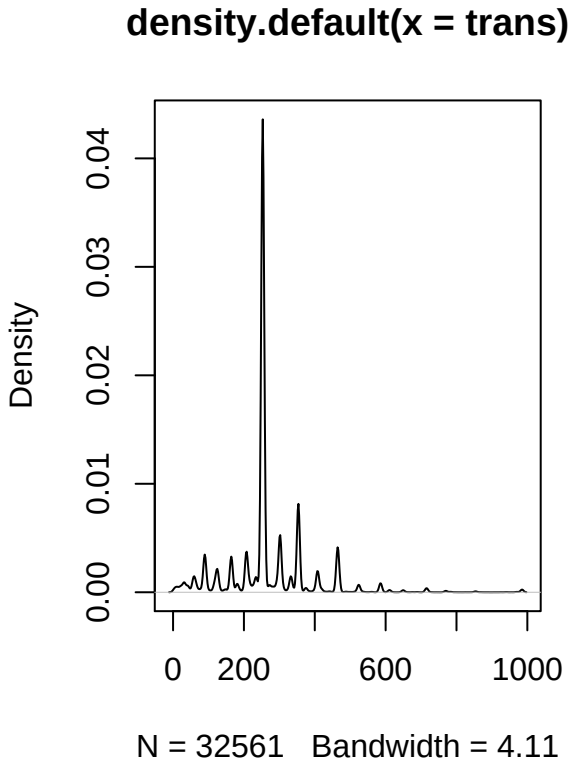
```
v13_trans = BoxCox(df$V13, lam_13)
plot(density(v13_trans))
qqPlot(v13_trans, dist= "norm", col=palette()[1], pch = 19, main = "With Box Cox transformation")
```

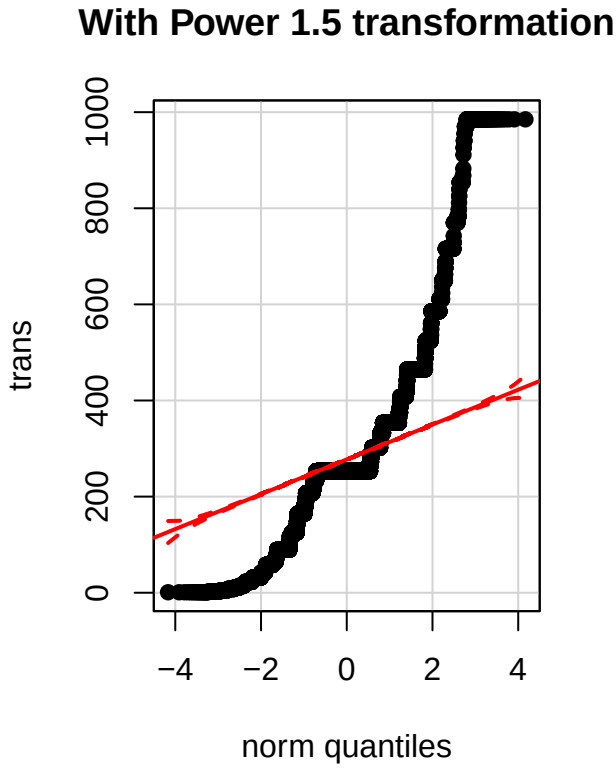


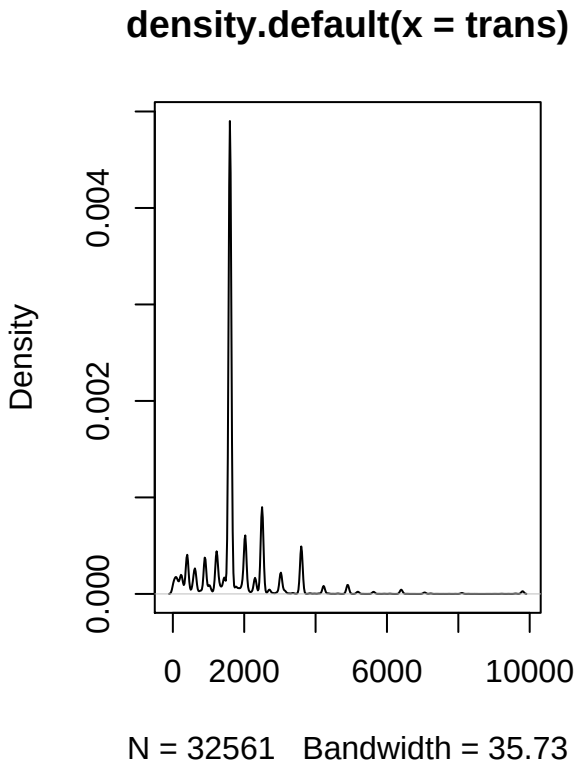

```
#### Doesn't work for v13



v13_trans = tan(df_transform$V13)
plot(density(v13_trans))
qqPlot(v13_trans, dist= "norm", col=palette()[1], pch = 19, main = "With tan transformation")
```

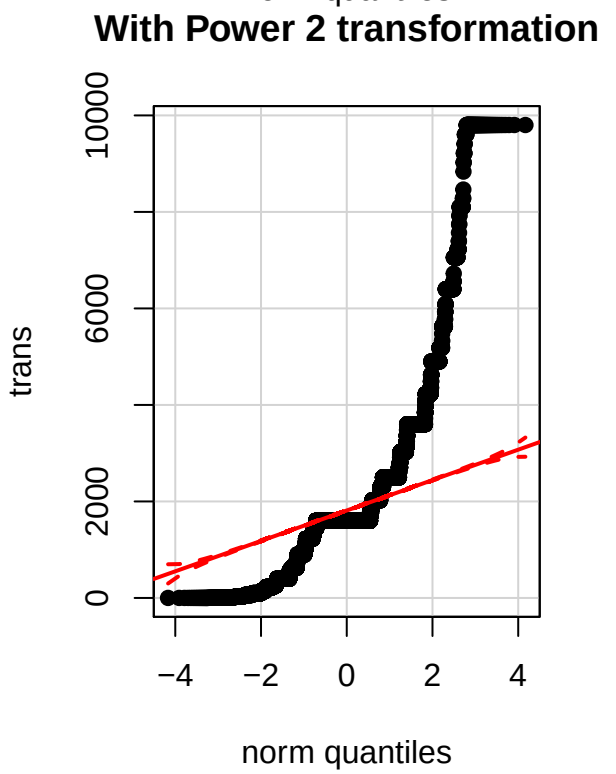


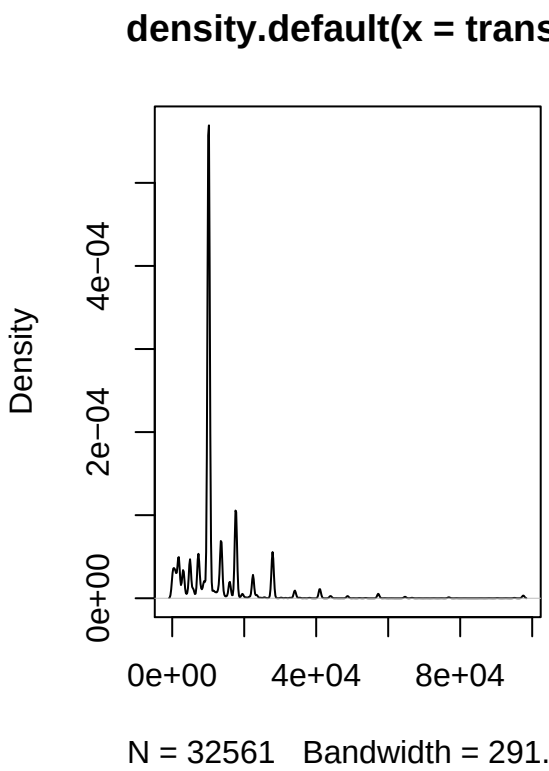

```
#### Looks like this is reasonable transformation but let's check other transformation

power = seq(1.5, 3, 0.5)
for (p in power){
  trans = (df_transform$V13)**p
  plot(density(trans))
  qqPlot(trans, dist= "norm", col=palette()[1], pch = 19, main = paste("With Power",p, "transformation")
}
```

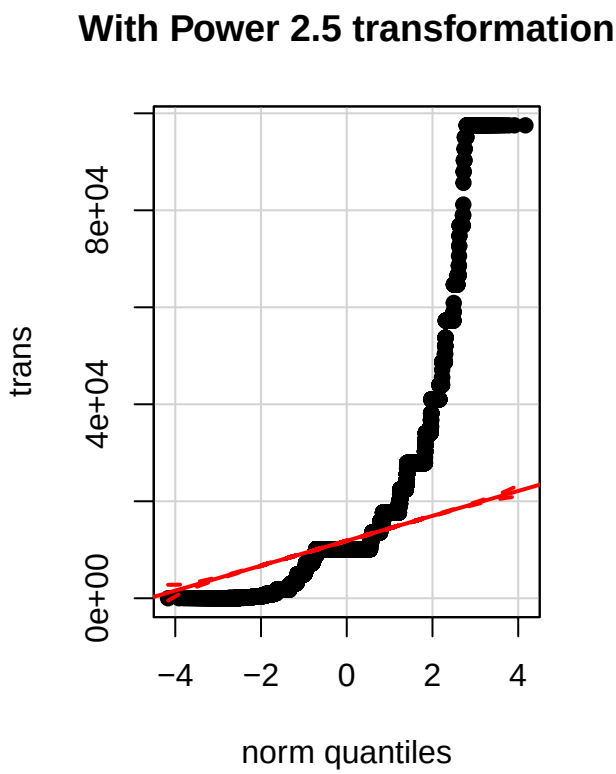

density.default(x = trans)
Density
0.00
0.01
0.02
0.03
0.04
0
200
600
1000
N = 32561 Bandwidth = 4.11

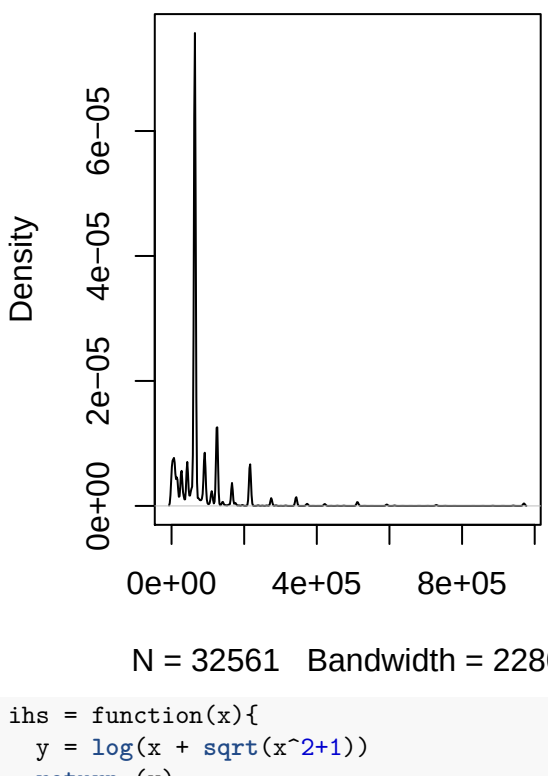

With Power 1.5 transformation
trans
0
200
400
600
800
1000
−4
−2
0
2
4
norm quantiles

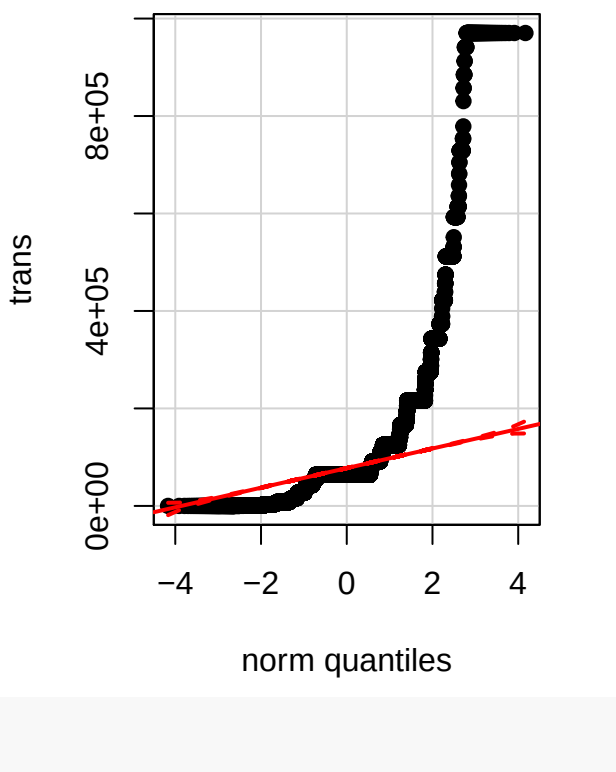

density.default(x = trans)
Density
0.000
0.002
0.004
0
2000
6000
10000
N = 32561 Bandwidth = 35.73


With Power 2 transformation
trans
0
2000
6000
10000
−4
−2
0
2
4
norm quantiles



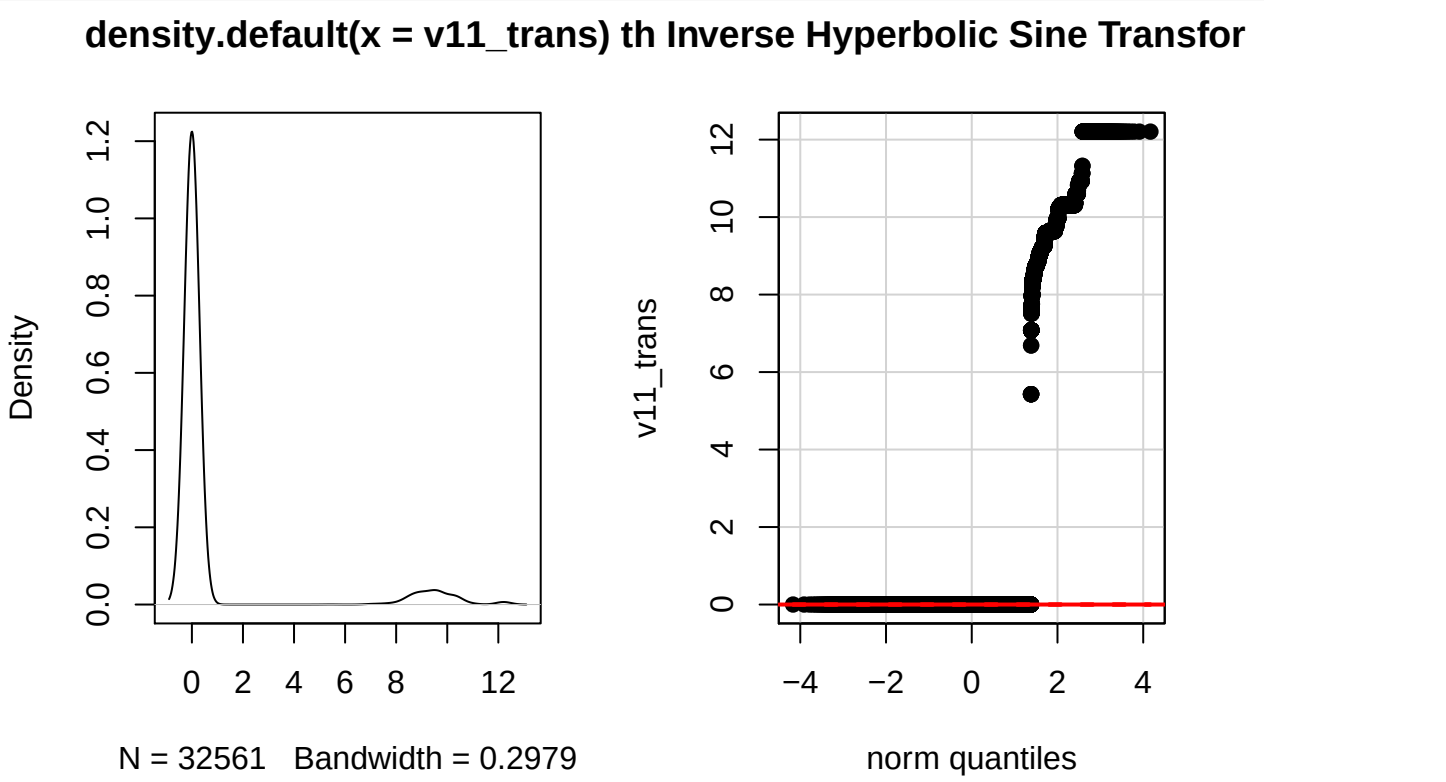



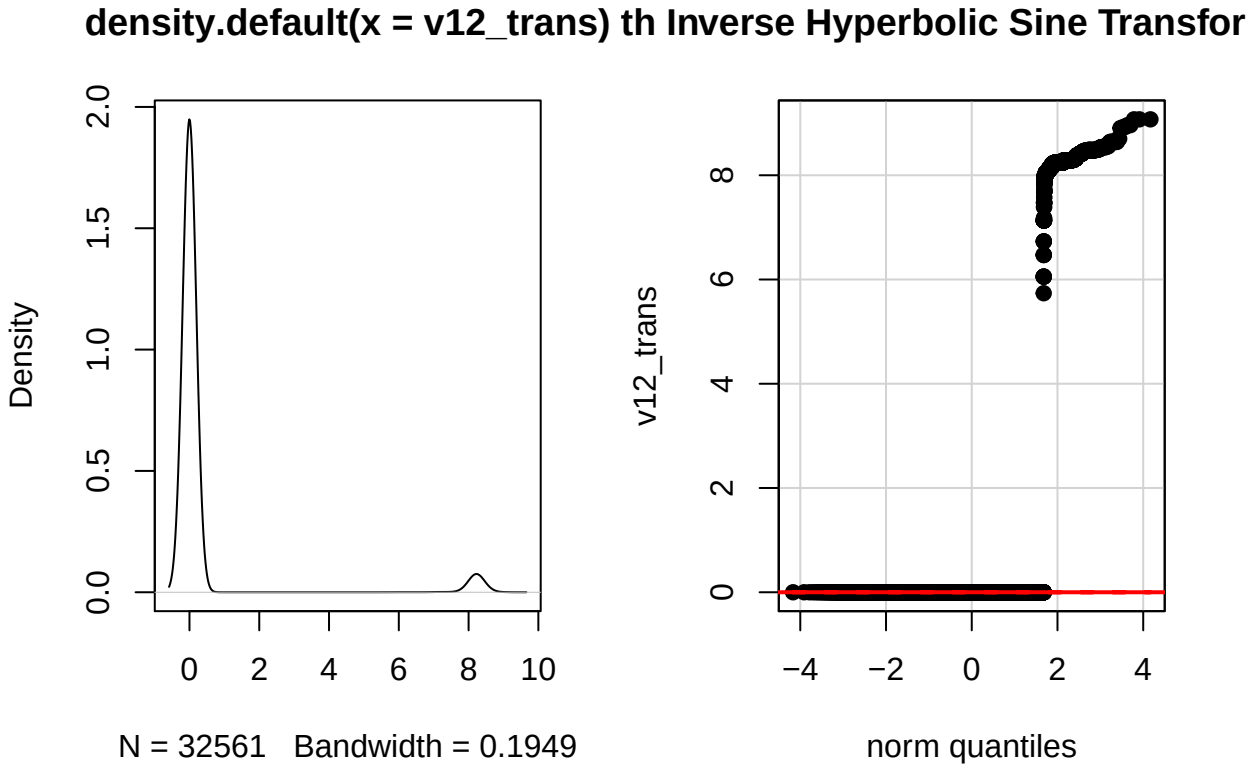




```
ihs = function(x){
  y = log(x + sqrt(x^2+1))
  return (y)
}
```

```
v11_trans = ihs(df_transform$V11)
plot(density(v11_trans))
qqPlot(v11_trans, dist= "norm", main = "With Inverse Hyperbolic Sine Transformation",col=palette()[1], p
```




```
v12_trans = ihs(df_transform$V12)
plot(density(v12_trans))
qqPlot(v12_trans, dist= "norm", col=palette()[1], main = "With Inverse Hyperbolic Sine Transformation",
```




Let check how many missing values for each variable:

```
sapply(df, function(x) sum(is.na(x)))
```

```
##  V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15
##   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
```

```
### from the data structure, there're "?" represents as missing values
### Lets do a recount again:
### convert factor columns into charactor:
factor_vect = paste0("V", c(2, 4, 6:10, 14))
df = data.table(df)
df_fac = df[,..factor_vect]
sapply(df_fac, function(x) table(x)[1])
```

```
##                 V2. ?              V4. 10th          V6. Divorced
##                  1836                   933                  4443
##                 V7. ?           V8. Husband V9. Amer-Indian-Eskimo
##                  1843                 13193                   311
##           V10. Female               V14. ?
##                 10771                   583
```

There are about 4000 mising value. We will need to deal with later.