Project Report on the Data set 'Loss given Default'

Ashwathy Joshy (ashwathyjoshy@gmail.com (mailto:ashwathyjoshy@gmail.com))

30 January 2018

1. Introduction

The data set has been kindly provided by a European bank and has been slightly modified and anonymized. It includes 2,545 observations on loans and LGDs. Key variables are:

1. LTV: Loan-to-value ratio, in %
2. Recovery_rate: Recovery rate, in %
3. lgd_time: Loss rate given default (LGD), in %
4. y_logistic: Logistic transformation of the LGD
5. lnrr: Natural logarithm of the recovery rate
6. Y_probit: Probit transformation of the LGD
7. purpose1: Indicator variable for the purpose of the loan; 1 = renting purpose, 0 = other
8. event: Indicator variable for a default or cure event; 1 = event, 0 = no event

2. Hypothesis:

We study the relationship between the LGD and the other variables in the data set and investigate if we can fit the data into the linear model:

$$LGD = (\alpha \times LTV) + (\beta \times recoveryrate) + (\gamma \times event) + (\delta \times purpose) + \epsilon$$

```
fit<-lm(Y_probit~  Recovery_rate + LTV + event +purpose1 , data =  lgd.df)
summary(fit)
```

```
##
## Call:
## lm(formula = Y_probit ~ Recovery_rate + LTV + event + purpose1,
##     data = lgd.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90410 -0.20377  0.00277  0.24170  1.40942
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)    0.89910    0.04840   18.577   <2e-16 ***
## Recovery_rate -5.14840    0.03663 -140.563   <2e-16 ***
## LTV           -0.03355    0.03105   -1.081    0.280
## event          2.02258    0.02523   80.176   <2e-16 ***
## purpose1       0.03673    0.03943    0.932    0.352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.511 on 2540 degrees of freedom
## Multiple R-squared:  0.9509, Adjusted R-squared:  0.9508
## F-statistic: 1.23e+04 on 4 and 2540 DF,  p-value: < 2.2e-16
```

```
fit<-lm(lgd_time ~  Recovery_rate + LTV + event +purpose1 , data = lgd.df)
summary(fit)
```

```
##
## Call:
## lm(formula = lgd_time ~ Recovery_rate + LTV + event + purpose1,
##     data = lgd.df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.839e-10 -9.210e-11 -7.053e-11 -4.270e-12  9.672e-10
##
## Coefficients:
##                 Estimate Std. Error    t value Pr(>|t|)
## (Intercept)    1.000e+00  2.214e-11  4.518e+10  < 2e-16 ***
## Recovery_rate -1.000e+00  1.675e-11 -5.969e+10  < 2e-16 ***
## LTV            3.555e-11  1.420e-11  2.504e+00   0.0124 *
## event          9.465e-11  1.154e-11  8.204e+00 3.66e-16 ***
## purpose1       1.531e-11  1.803e-11  8.490e-01   0.3958
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.337e-10 on 2540 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.261e+21 on 4 and 2540 DF,  p-value: < 2.2e-16
```

```
fit<-lm(y_logistic ~  Recovery_rate + LTV + event +purpose1 , data =  lgd.df)
summary(fit)
```

```
##
## Call:
## lm(formula = y_logistic ~ Recovery_rate + LTV + event + purpose1,
##     data = lgd.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4894 -0.5768  0.0110  0.7087  4.8873
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.04443    0.15511  -0.286   0.7746
## Recovery_rate -11.38723    0.11738 -97.008   <2e-16 ***
## LTV            -0.16855    0.09951  -1.694   0.0904 .
## event           7.00240    0.08085  86.611   <2e-16 ***
## purpose1        0.09940    0.12635   0.787   0.4316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.638 on 2540 degrees of freedom
## Multiple R-squared:  0.9274, Adjusted R-squared:  0.9273
## F-statistic:  8112 on 4 and 2540 DF,  p-value: < 2.2e-16
```

Conclusion:

From the regression analysis, we can conclude that the data fits into the linear model:

1. $$logisticLGD = -0.04443 - 5.14840recoveryrate - 0.03355LTV + 7.00240event + 0.03673purpose$$

   with 92.73 % of data being accounted for by this model.

2. $$ProbitLGD = 0.89910 - 11.38723recoveryrate - 0.16855LTV + 2.02258event + 0.09940purpose$$

   with 95.08 % of data being accounted for by this model.

Analysis of the data set

```
lgd.df <- read.csv(paste("lgd.csv", sep= ""))
library(psych)
describe(lgd.df)
```

```
##               vars    n  mean   sd median trimmed  mad    min   max range
## LTV              1 2545  0.68 0.36   0.66    0.66 0.39   0.00  1.98  1.98
## Recovery_rate    2 2545  0.77 0.33   0.97    0.84 0.05   0.00  1.00  1.00
## lgd_time         3 2545  0.23 0.33   0.03    0.16 0.05   0.00  1.00  1.00
## y_logistic       4 2545 -3.94 6.07  -3.41   -4.49 5.09 -11.51 11.51 23.03
## lnrr             5 2545 -1.00 2.70  -0.03   -0.24 0.05 -11.51  0.00 11.51
## Y_probit         6 2545 -1.65 2.30  -1.85   -1.93 2.68  -4.26  4.26  8.53
## purpose1         7 2545  0.07 0.26   0.00    0.00 0.00   0.00  1.00  1.00
## event            8 2545  0.71 0.45   1.00    0.77 0.00   0.00  1.00  1.00
##                skew kurtosis   se
## LTV            0.48     0.16 0.01
## Recovery_rate -1.31     0.27 0.01
## lgd_time       1.31     0.27 0.01
## y_logistic     0.53     0.23 0.12
## lnrr          -3.37    10.09 0.05
## Y_probit       0.79     0.29 0.05
## purpose1       3.29     8.83 0.01
## event         -0.95    -1.10 0.01
```

```
str(lgd.df)
```

```
## 'data.frame':    2545 obs. of  8 variables:
##  $ LTV          : num  0.214 0.214 0.214 0.214 0.214 ...
##  $ Recovery_rate: num  0.698 0.78 0.702 0.754 0.803 ...
##  $ lgd_time     : num  0.302 0.22 0.298 0.246 0.197 ...
##  $ y_logistic   : num  -0.838 -1.266 -0.858 -1.12 -1.404 ...
##  $ lnrr         : num  -0.36 -0.248 -0.353 -0.282 -0.22 ...
##  $ Y_probit     : num  -0.519 -0.772 -0.531 -0.687 -0.852 ...
##  $ purpose1     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ event        : int  1 1 1 1 1 1 0 1 1 0 ...
```

```
length(lgd.df)
```

```
## [1] 8
```

```
xtabs(~purpose1, data= lgd.df)
```

```
## purpose1
##    0    1
## 2360  185
```

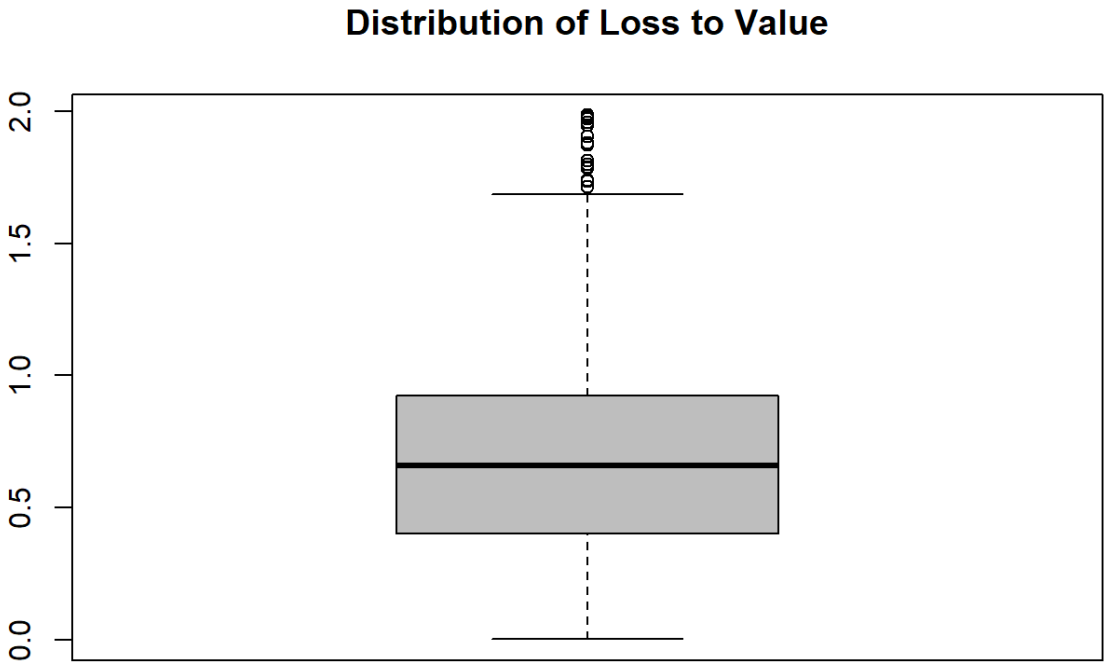
```
xtabs(~event, data= lgd.df)
```

```
## event
##    0    1
##  728 1817
```

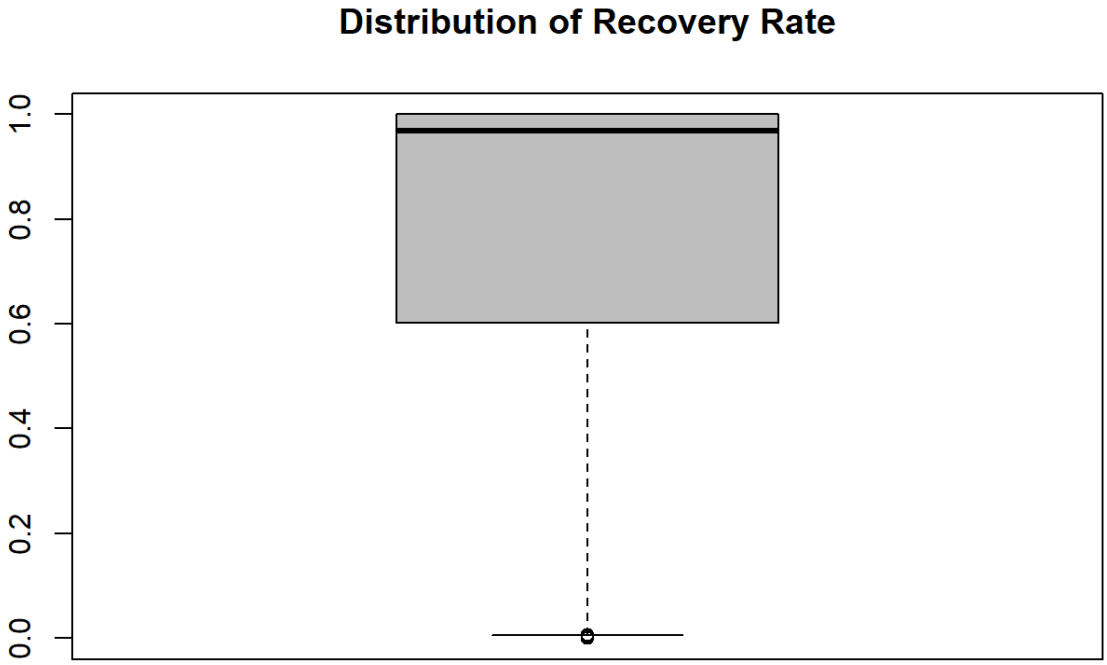
```
xtabs(~purpose1+event, data= lgd.df)
```

```
##         event
## purpose1    0    1
##        0  704 1656
##        1   24  161
```

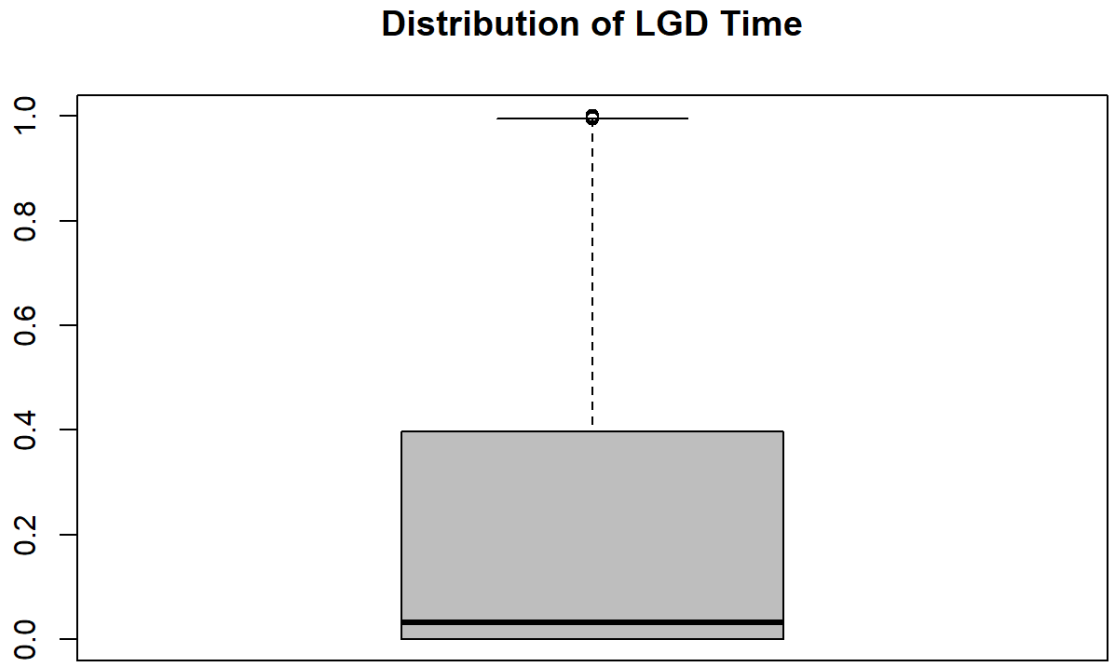
```
boxplot(lgd.df$LTV,  main= "Distribution of Loss to Value", col = "grey", vertical=TRUE)
```



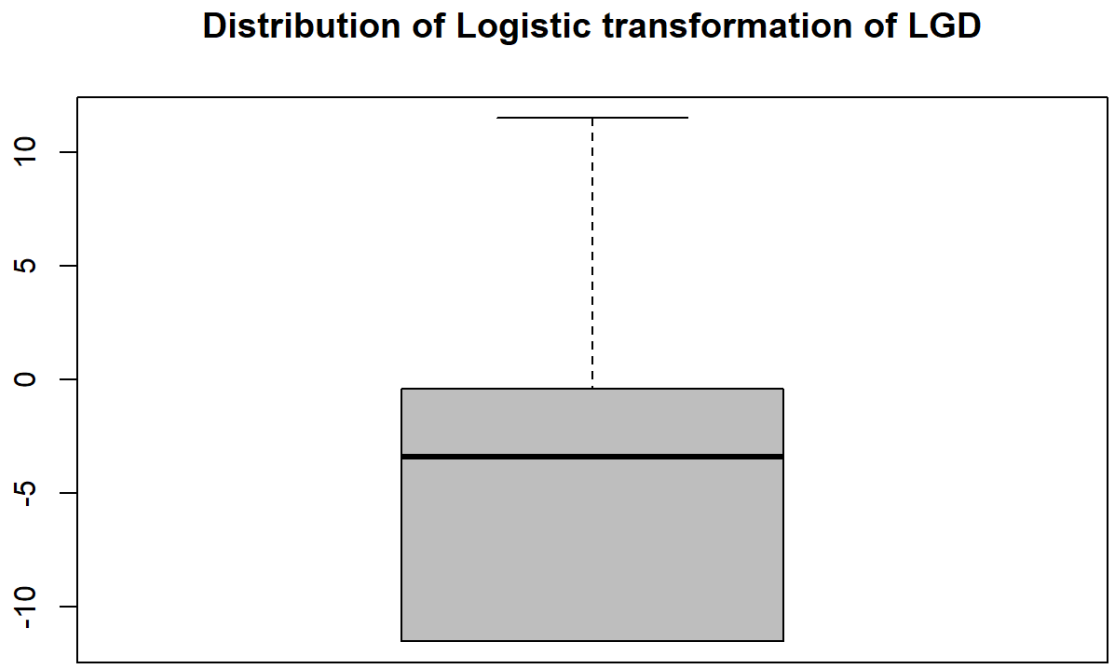

```
boxplot(lgd.df$Recovery_rate,  main= "Distribution of Recovery Rate", col = "grey", vertical=TRUE)
```



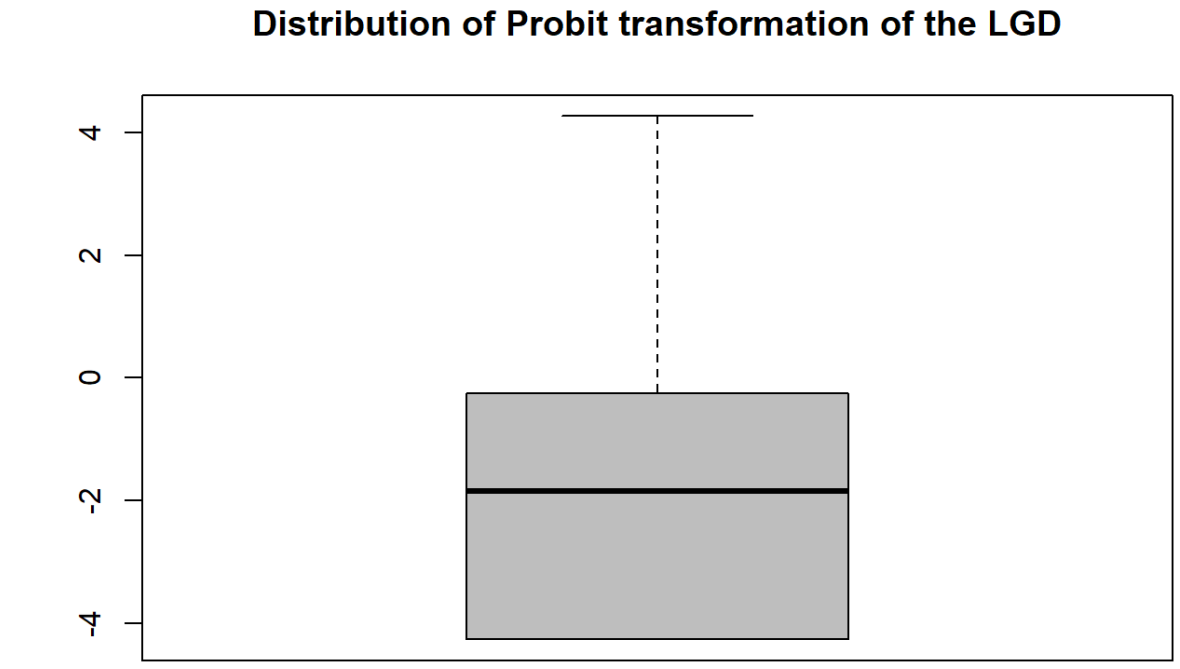

```
boxplot(lgd.df$lgd_time ,  main= "Distribution of LGD Time", col = "grey", vertical=TRUE)
```



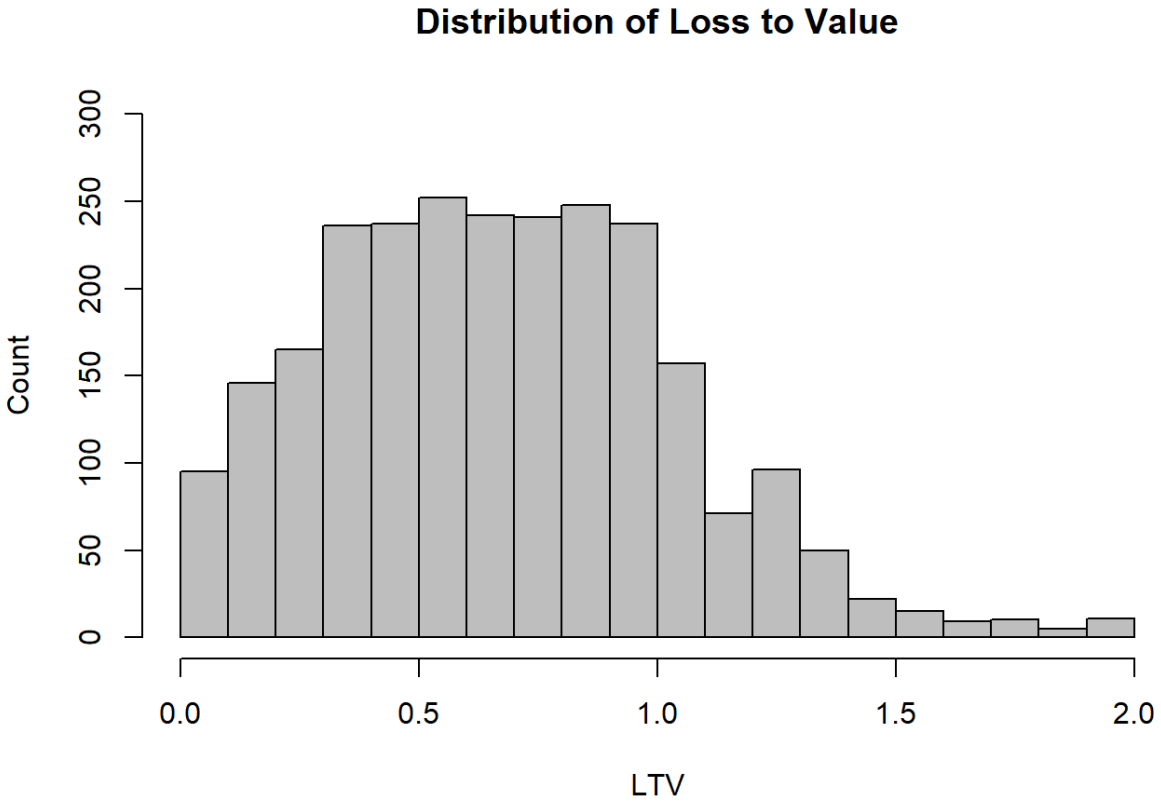

```
boxplot(lgd.df$y_logistic ,  main= "Distribution of Logistic transformation of LGD", col = "grey", v
ertical=TRUE)
```



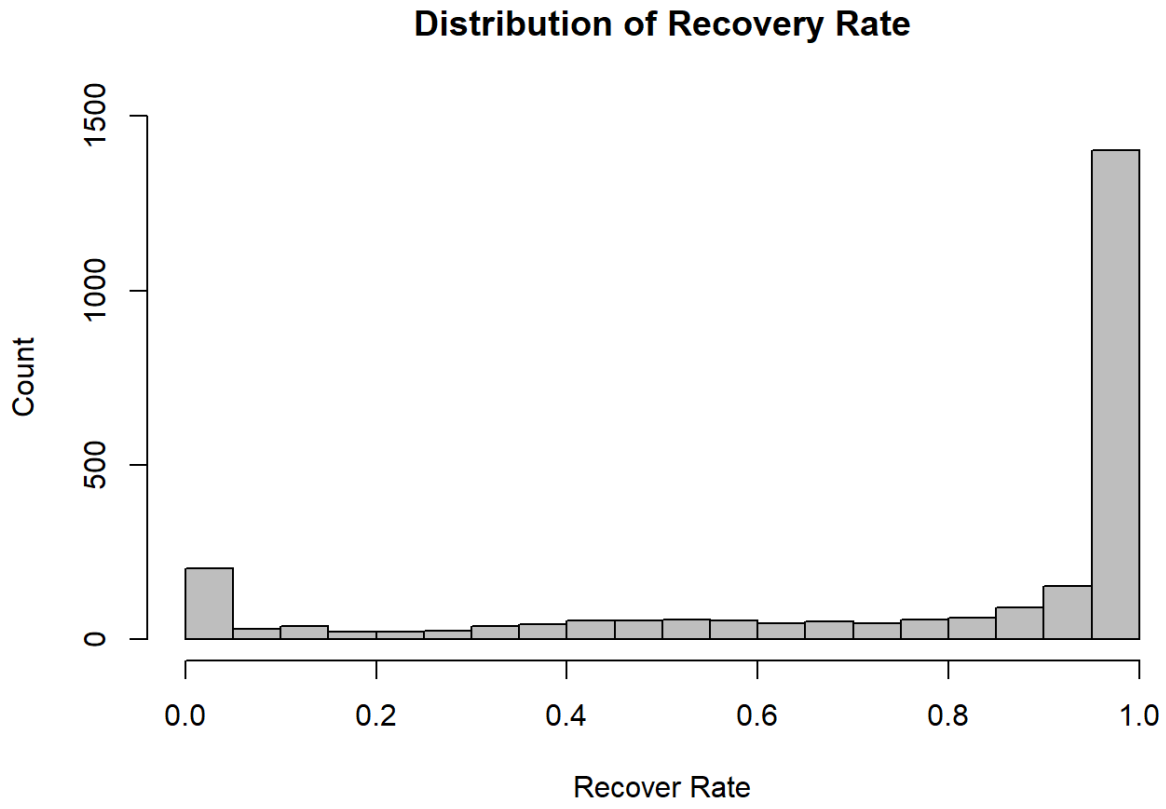

```
boxplot(lgd.df$Y_probit,  main= "Distribution of Probit transformation of the LGD", col = "grey", ve
rtical=TRUE)
```



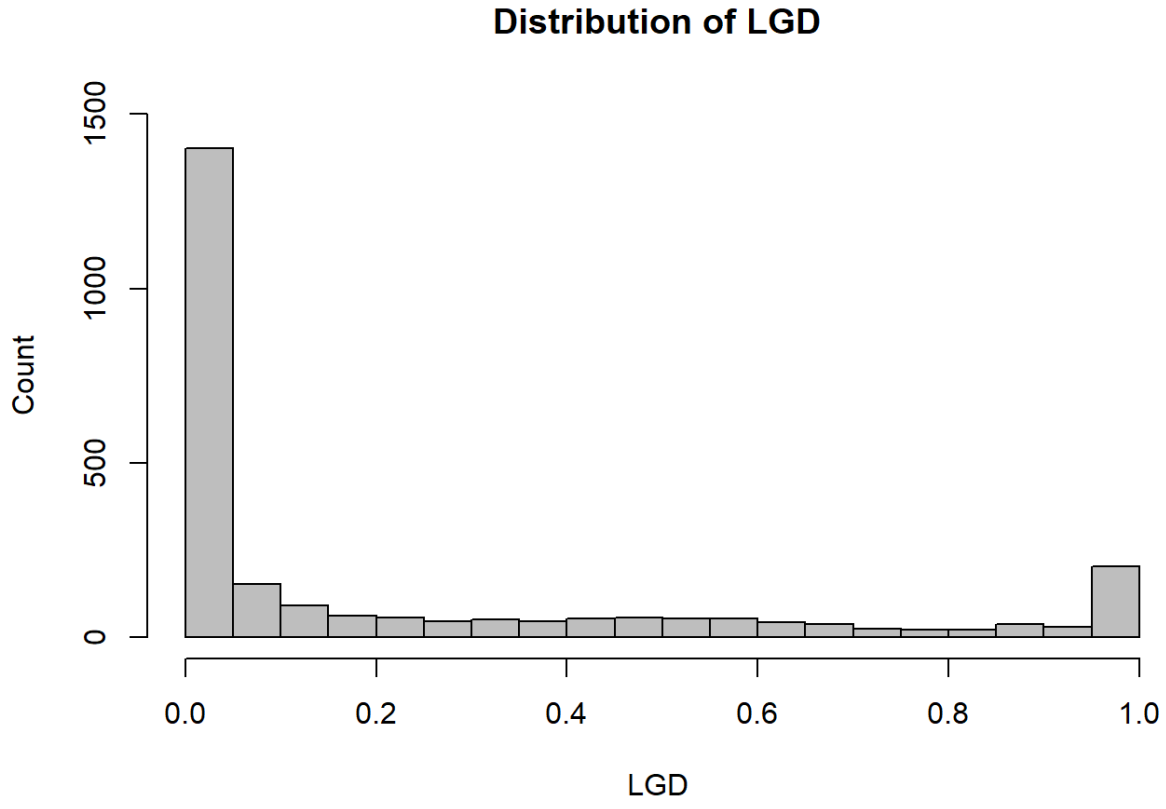

```
hist(lgd.df$LTV,  main=" Distribution of Loss to Value ",  xlab = "LTV",  ylab = "Count", breaks = 2
0, col = "grey", xlim= c(0,2), ylim=c(0,300))
```



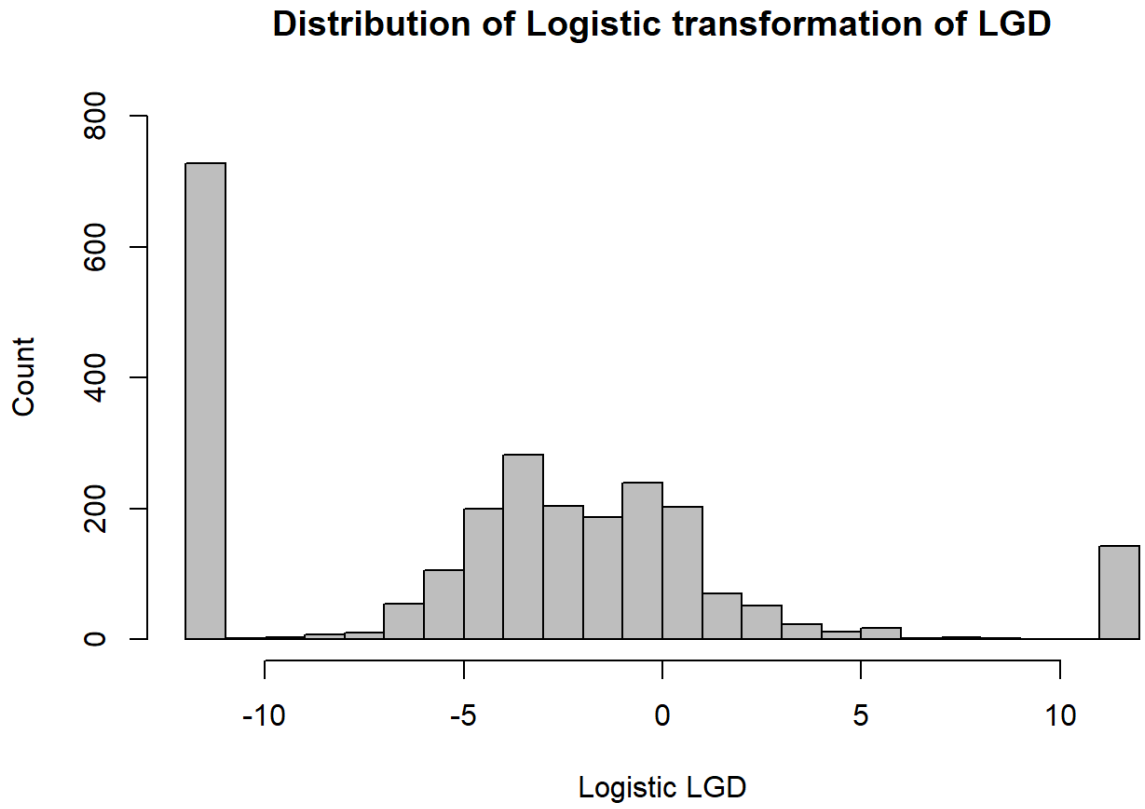

```
hist(lgd.df$Recovery_rate,  main=" Distribution of Recovery Rate ",  xlab = "Recover Rate",  ylab =
"Count", breaks = 20, col = "grey", xlim= c(0,1), ylim=c(0,1500))
```



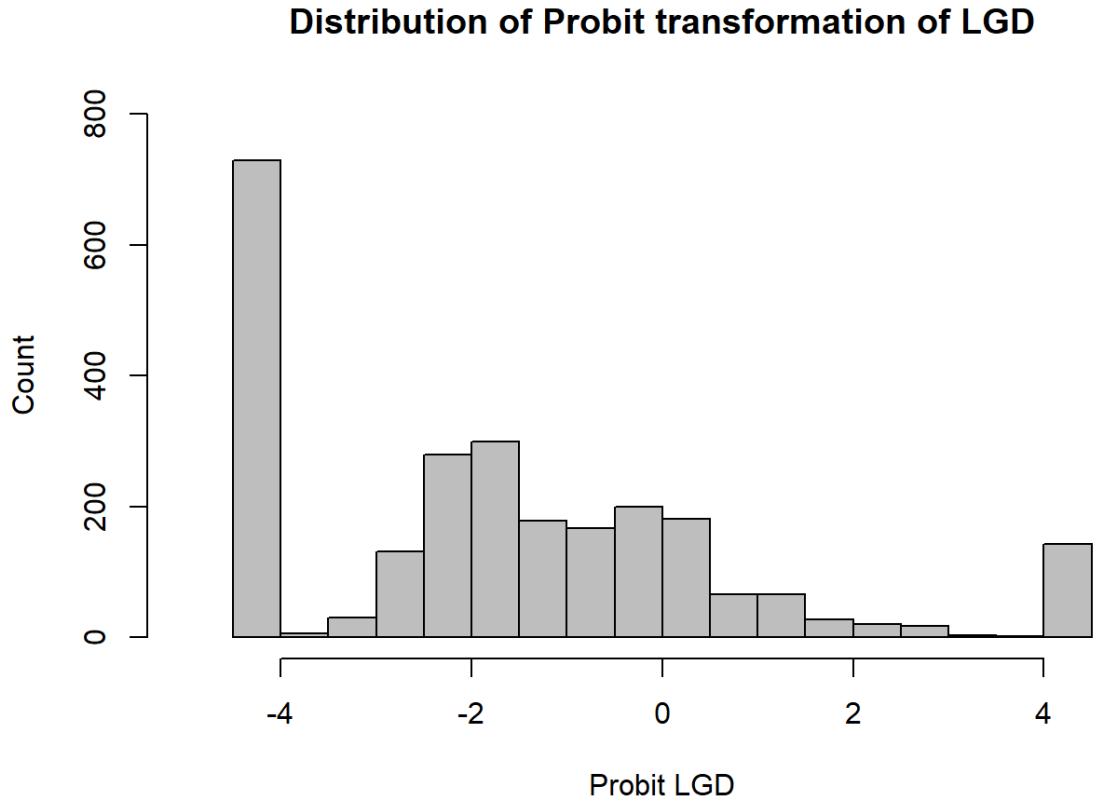

```
hist(lgd.df$lgd_time ,  main=" Distribution of LGD  ",  xlab = "LGD ",  ylab = "Count", breaks = 20,
col = "grey", xlim= c(0,1), ylim=c(0,1500))
```



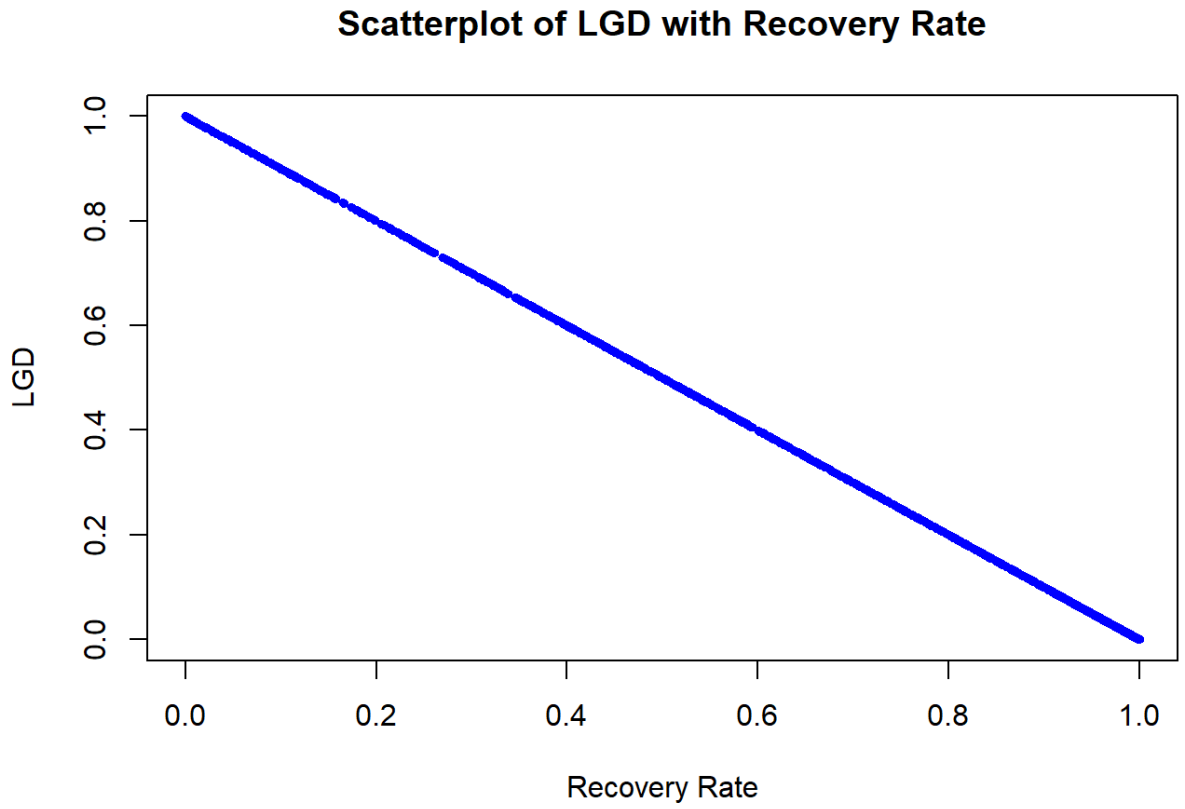

```
hist(lgd.df$y_logistic ,  main=" Distribution of Logistic transformation of LGD ",  xlab = "Logistic
LGD",  ylab = "Count", breaks = 20, col = "grey", xlim= c(-12,12), ylim=c(0,800))
```



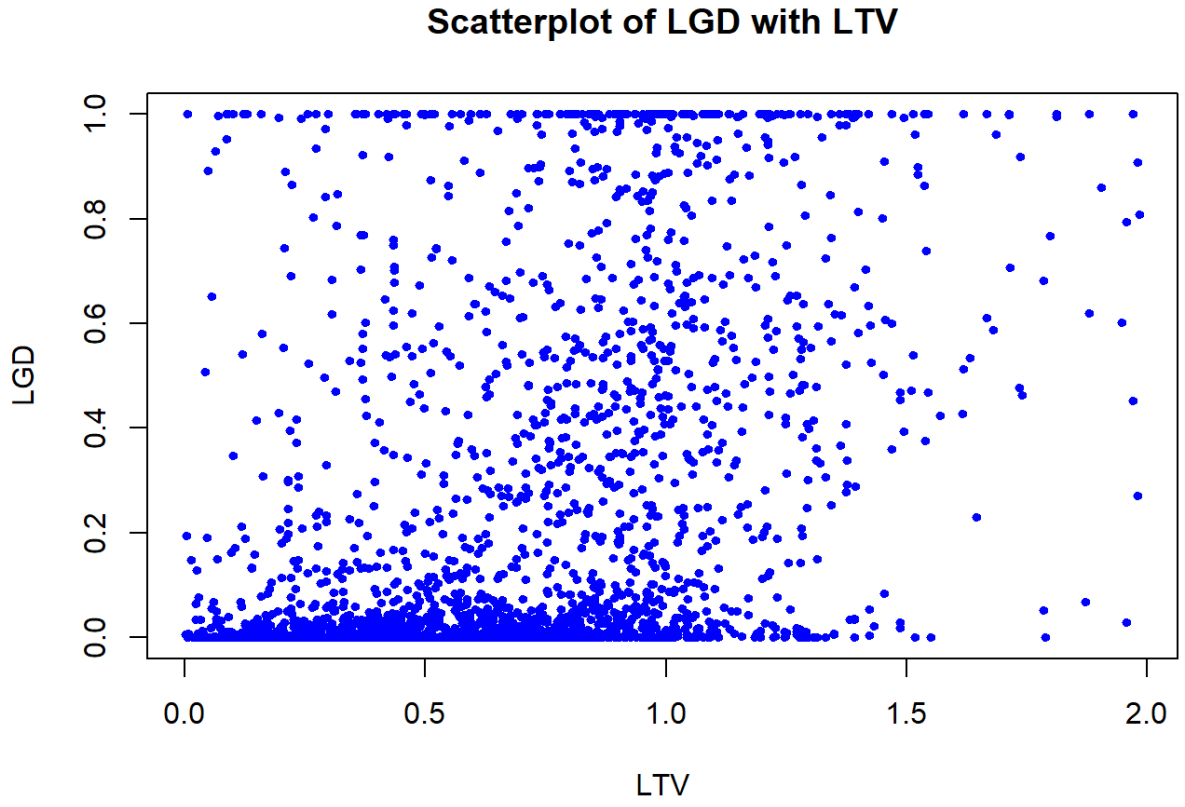

```
hist(lgd.df$Y_probit ,  main=" Distribution of Probit transformation of LGD ",  xlab = "Probit LGD",
ylab = "Count", breaks = 20, col = "grey", xlim= c(-5,5), ylim=c(0,800))
```



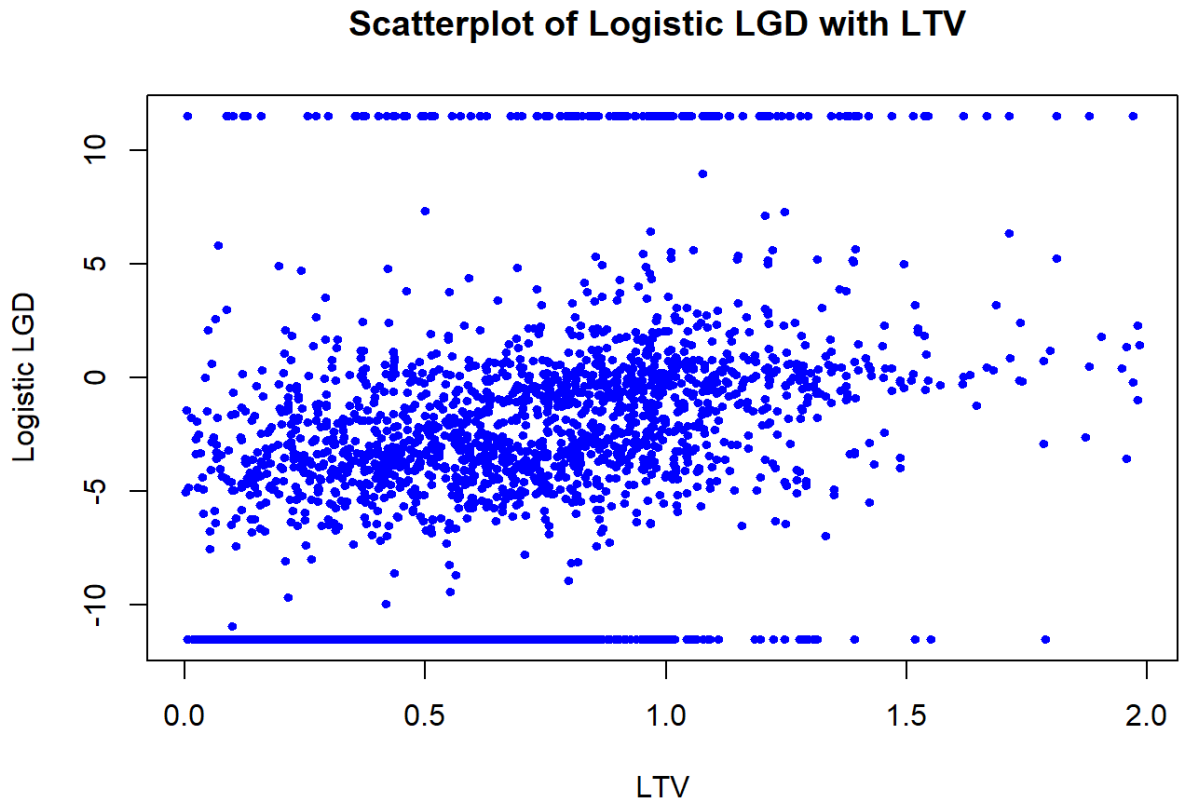

```
plot(lgd.df$Recovery_rate,lgd.df$lgd_time, col="blue",main="Scatterplot of LGD with Recovery Rate",p
ch=19,cex=0.6, xlab="Recovery Rate", ylab="LGD")
```



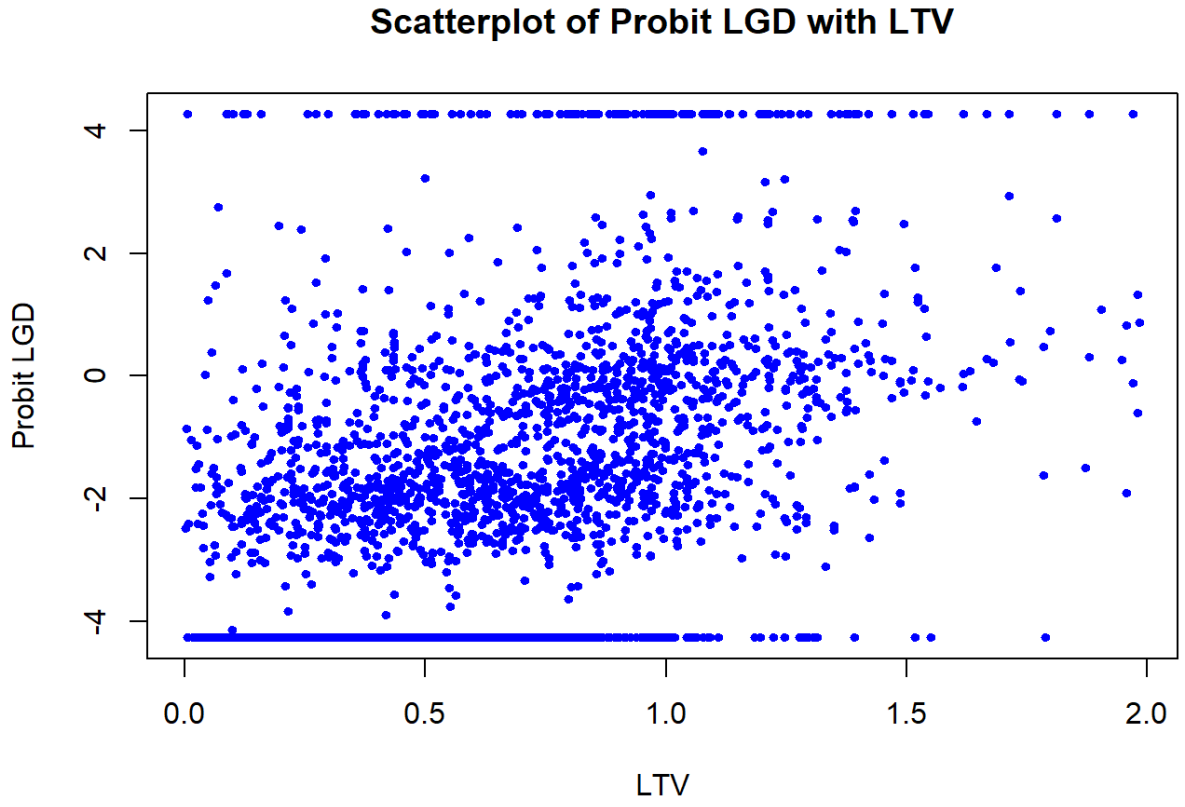

```
plot(lgd.df$LTV,lgd.df$lgd_time, col="blue",main="Scatterplot of LGD with LTV",pch=19,cex=0.6, xlab=
"LTV", ylab="LGD")
```



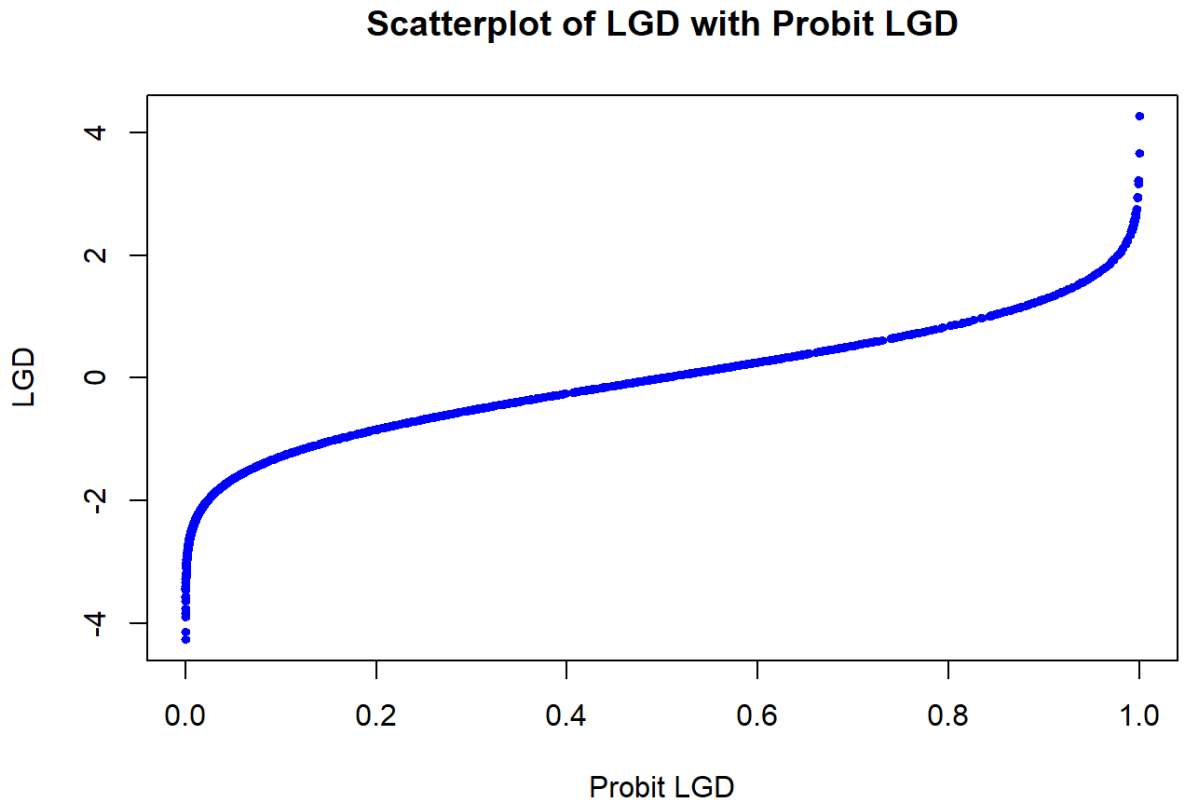

```
plot(lgd.df$LTV,lgd.df$y_logistic, col="blue",main="Scatterplot of Logistic LGD with LTV ",pch=19,ce
x=0.6, xlab="LTV", ylab="Logistic LGD")
```



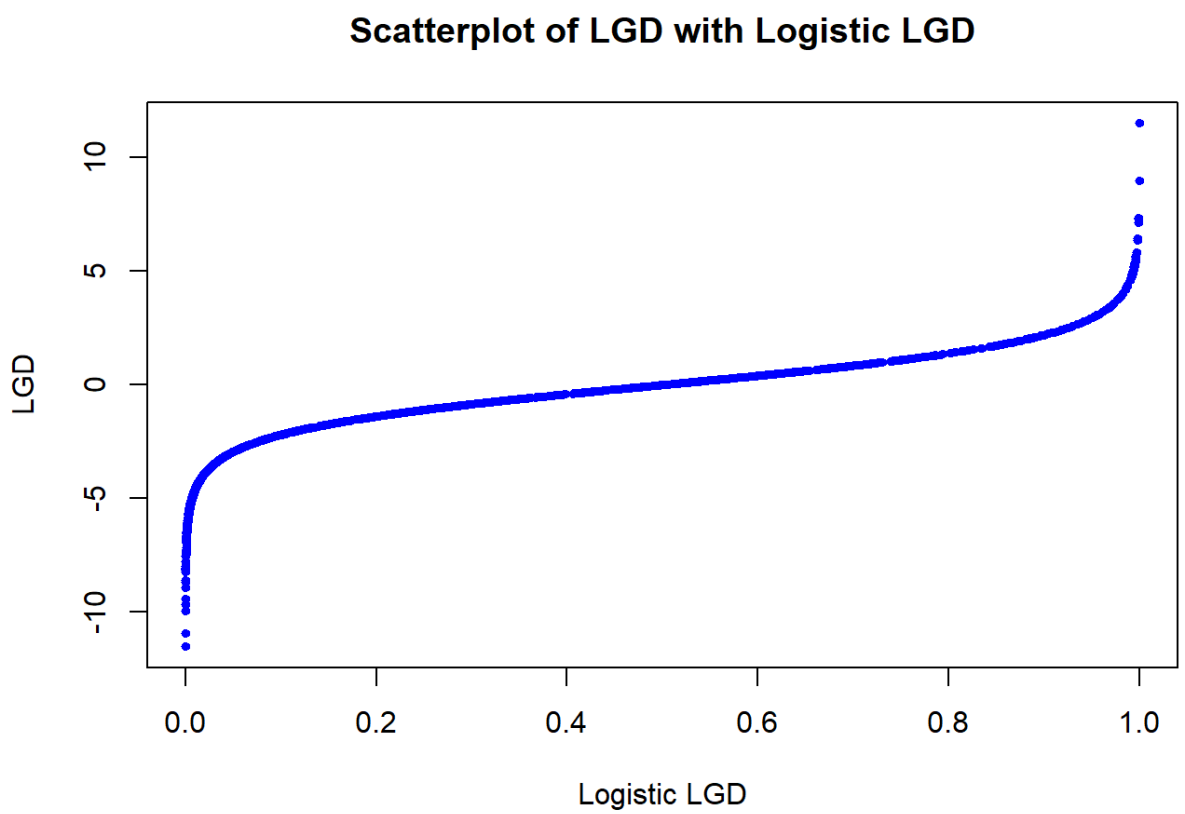

```
plot(lgd.df$LTV,lgd.df$Y_probit, col="blue",main="Scatterplot of Probit LGD with LTV",pch=19,cex=0.6
, xlab="LTV", ylab="Probit LGD")
```




```
plot(lgd.df$lgd_time,lgd.df$Y_probit, col="blue",main="Scatterplot of LGD with Probit LGD",pch=19,ce
x=0.6, xlab="Probit LGD", ylab="LGD")
```



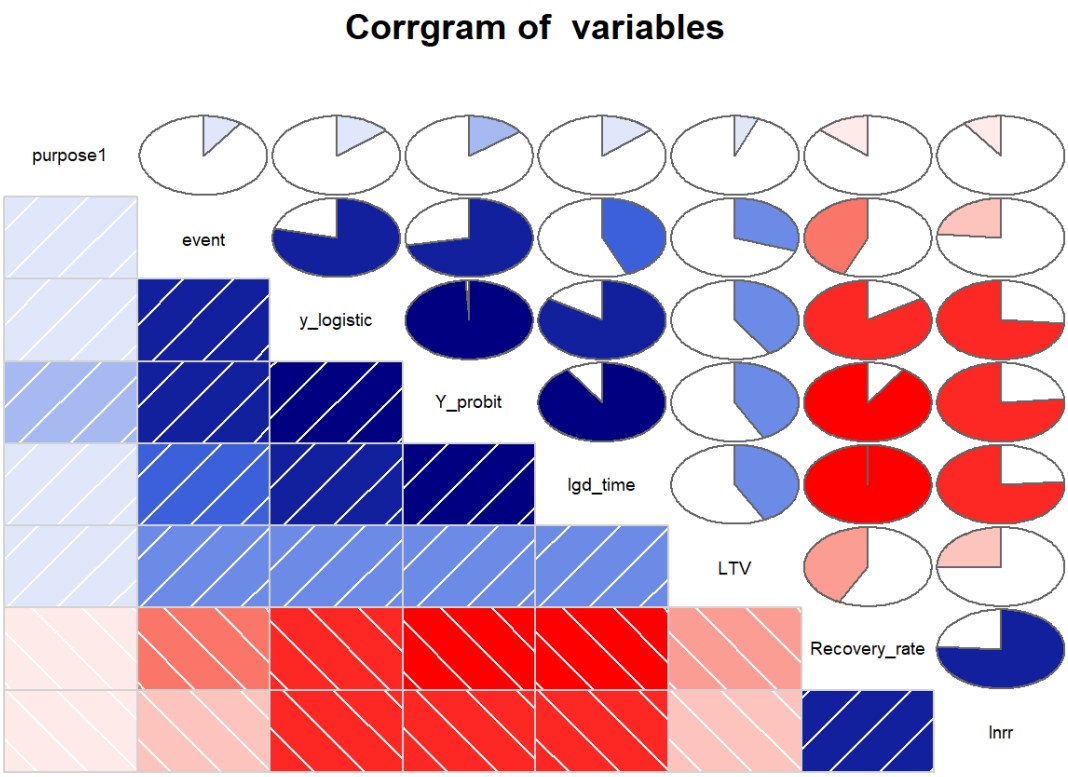

```
plot(lgd.df$lgd_time,lgd.df$y_logistic, col="blue",main="Scatterplot of LGD with Logistic LGD",pch=1
9,cex=0.6, xlab="Logistic LGD", ylab="LGD")
```




```
cor(lgd.df,y= NULL)
```

```
##                      LTV Recovery_rate   lgd_time y_logistic        lnrr
## LTV            1.0000000    -0.4243962  0.4243962  0.4100271 -0.25153723
## Recovery_rate -0.4243962     1.0000000 -1.0000000 -0.8420673  0.76084748
## lgd_time       0.4243962    -1.0000000  1.0000000  0.8420673 -0.76084748
## y_logistic     0.4100271    -0.8420673  0.8420673  1.0000000 -0.73542845
## lnrr          -0.2515372     0.7608475 -0.7608475 -0.7354284  1.00000000
## Y_probit       0.4273198    -0.9078090  0.9078090  0.9901022 -0.76413429
## purpose1       0.0578538    -0.1383490  0.1383490  0.1395092 -0.09929195
## event          0.3031827    -0.4388312  0.4388312  0.7892896 -0.23369946
##                 Y_probit    purpose1       event
## LTV            0.4273198  0.05785380  0.30318272
## Recovery_rate -0.9078090 -0.13834902 -0.43883122
## lgd_time       0.9078090  0.13834902  0.43883122
## y_logistic     0.9901022  0.13950922  0.78928960
## lnrr          -0.7641343 -0.09929195 -0.23369946
## Y_probit       1.0000000  0.14399780  0.71826833
## purpose1       0.1439978  1.00000000  0.09684848
## event          0.7182683  0.09684848  1.00000000
```

```
library(corrgram)
corrgram(lgd.df, order=TRUE, lower.panel= panel.shade, upper.panel= panel.pie,text.panel = panel.tx
t, main= "Corrgram of  variables")
```

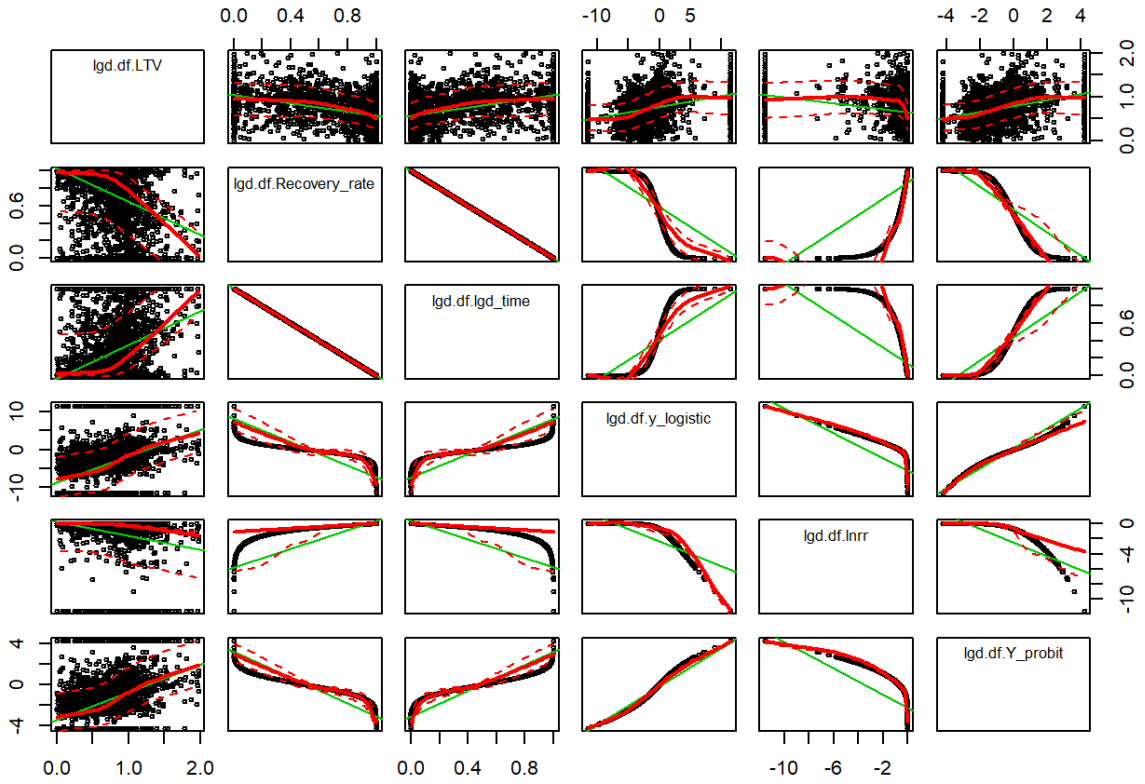



```
library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
scatterplotMatrix(formula = ~ lgd.df$LTV + lgd.df$Recovery_rate + lgd.df$lgd_time + lgd.df$y_logisti
c + lgd.df$lnrr+ lgd.df$Y_probit , data=lgd.df,cex=0.5, diagonal="none" )
```




```
cor.test(lgd.df$LTV , lgd.df$lgd_time )
```

```
##
##  Pearson's product-moment correlation
##
## data:  lgd.df$LTV and lgd.df$lgd_time
## t = 23.636, df = 2543, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3920057 0.4557358
## sample estimates:
##       cor
## 0.4243962
```

```
t.test(lgd.df$y_logistic ~ lgd.df$purpose1, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  lgd.df$y_logistic by lgd.df$purpose1
## t = -7.1047, df = 2543, p-value = 1.559e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.163301 -2.362244
## sample estimates:
## mean in group 0 mean in group 1
##       -4.178519       -0.915746
```

```
t.test(lgd.df$y_logistic ~ lgd.df$event, var.equal= TRUE)
```

```
##
##  Two Sample t-test
##
## data:  lgd.df$y_logistic by lgd.df$event
## t = -64.823, df = 2543, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.92601 -10.28439
## sample estimates:
## mean in group 0 mean in group 1
##     -11.5129155      -0.9077129
```