Math156 Final Project Neural Network Report

Group 5

7/22/2020

PART I. Loading data and Define helper functions

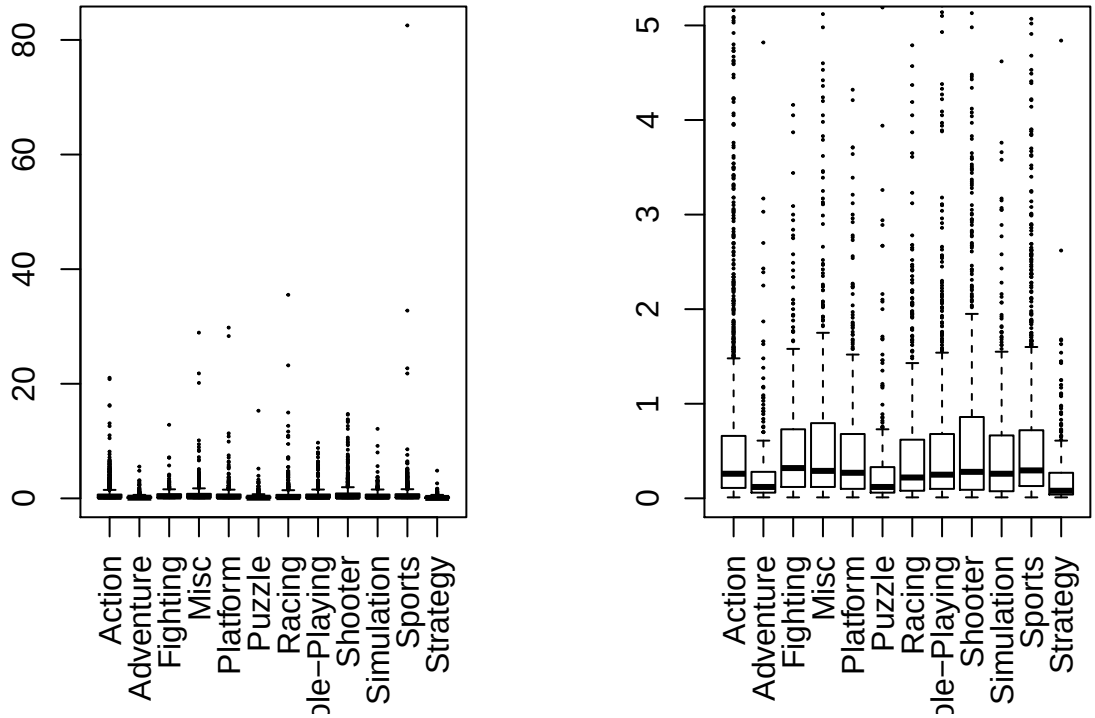
```
Vdata <- read.csv("../data/videogames.csv")
#Vdata <- Vdata[as.numeric(as.character(Vdata$Year_of_Release)) > 2009, ] # only game after 2010
Vdata <- Vdata[(!is.na(Vdata$Critic_Score)),] # remove missing values

Anova_test <- function(classifier, data = Vdata) {
  summary(aov(data$Global_Sales ~ data[[classifier]]))
}

get_freq <- function(x, breaks) {
  len <- length(breaks)
  res <- numeric(len)
  for (i in seq_len(len)) {
    res[i] <- sum(breaks[i] < x & x <= breaks[i+1])
  }
  res
}

get_density <- function(x, breaks, by) {
  res <- get_freq(x, breaks) / (length(x) * by)
  res
}

get_density_idx <- function(gs) {
  ceiling(20 * gs)
}

testclass <- function(classifier, data=Vdata){
  table <- tapply(Vdata$Global_Sales, Vdata[[classifier]], function(x){x})
  table[[1]] <- NULL
  par(mfrow = c(1, 2))
  boxplot(table, las = 3, cex=0.2, pch=20)
  boxplot(table, ylim=c(0,5), las = 3, cex=0.2,pch=20)
  Anova_test(classifier)
}

rangef <- function(x, divider=5) {
  x %/% divider * divider + divider/2
}
```

Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Coun
Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76	51	8	322
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	*NA*	*NA*		*NA*
Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82	73	8.3	709
Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80	73	8	192
Pokemon Red/Po...	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	*NA*	*NA*		*NA*
Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30.26	*NA*	*NA*		*NA*
New Super Mario ...	DS	2006	Platform	Nintendo	11.28	9.14	6.50	2.88	29.80	89	65	8.5	431
Wii Play	Wii	2006	Misc	Nintendo	13.96	9.18	2.93	2.84	28.92	58	41	6.6	129
New Super Mario ...	Wii	2009	Platform	Nintendo	14.44	6.94	4.70	2.24	28.32	87	80	8.4	594
Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31	*NA*	*NA*		*NA*
Nintendogs	DS	2005	Simulation	Nintendo	9.05	10.95	1.93	2.74	24.67	*NA*	*NA*		*NA*
Mario Kart DS	DS	2005	Racing	Nintendo	9.71	7.47	4.13	1.90	23.21	91	64	8.6	464
Pokemon Gold/P...	GB	1999	Role-Playing	Nintendo	9.00	6.18	7.20	0.71	23.10	*NA*	*NA*		*NA*
Wii Fit	Wii	2007	Sports	Nintendo	8.92	8.03	3.60	2.15	22.70	80	63	7.7	146

Figure 1: Snapshot of Data

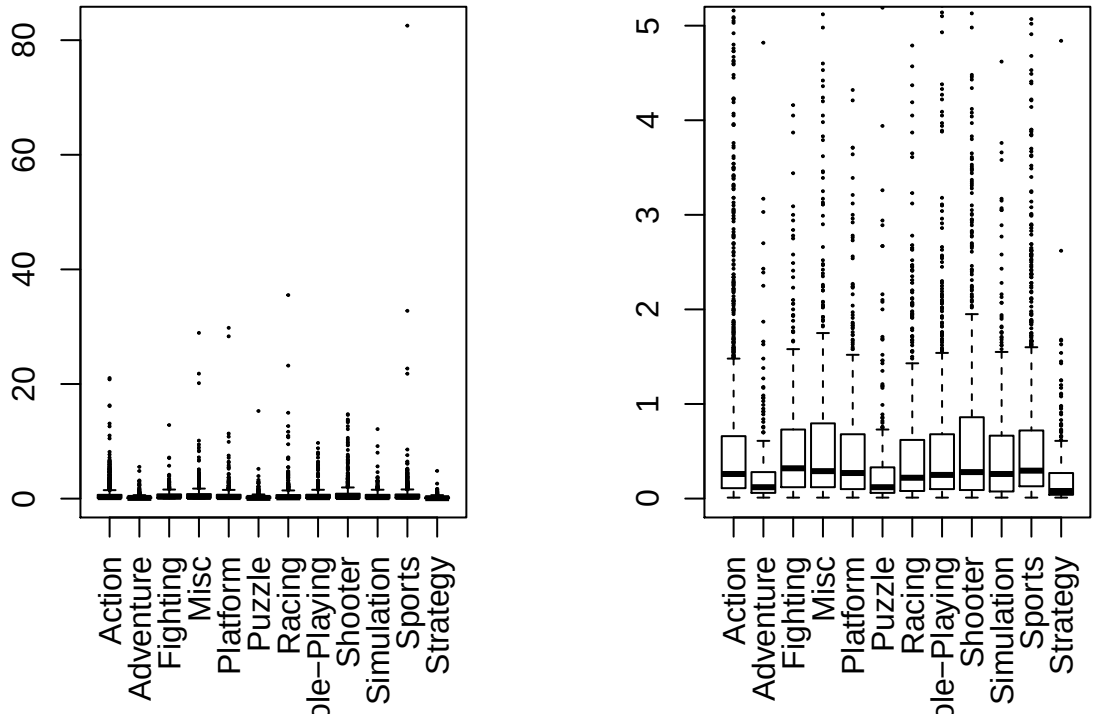
PART II. Heuristically explore the effect of each paramemter

Anova test on Genre



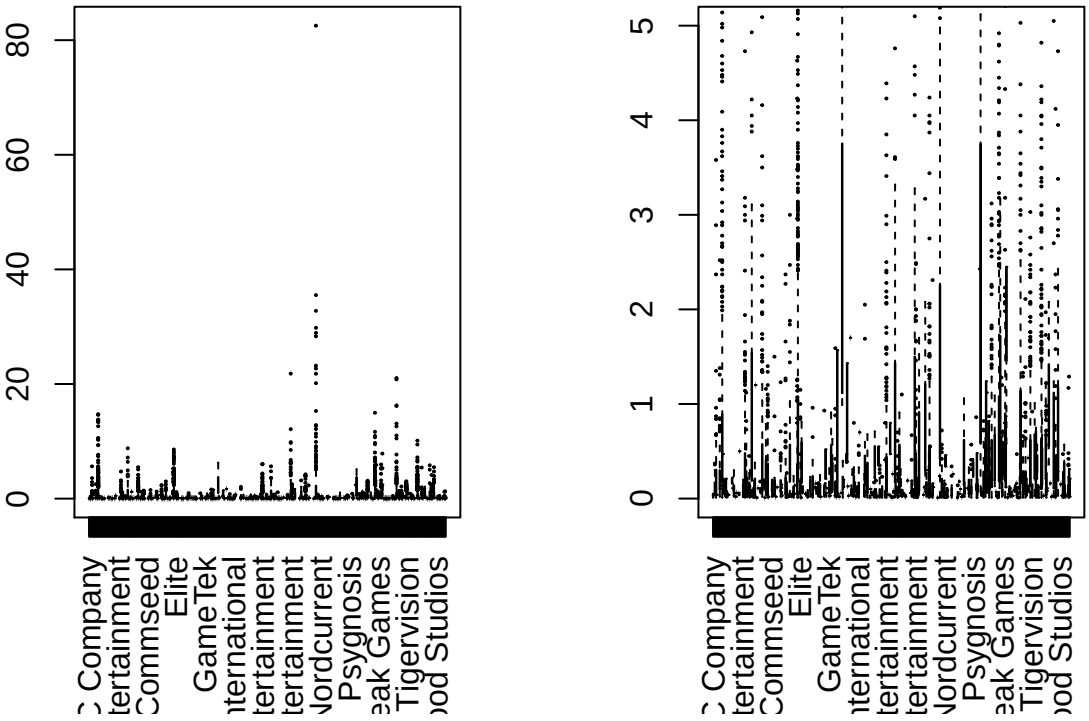

```
##                       Df Sum Sq Mean Sq F value   Pr(>F)
## data[[classifier]]    11    190  17.294    5.27 2.37e-08 ***
## Residuals           8125  26662   3.281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Avona test on Platform




```
##                       Df Sum Sq Mean Sq F value   Pr(>F)
## data[[classifier]]    11    190  17.294    5.27 2.37e-08 ***
## Residuals           8125  26662   3.281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

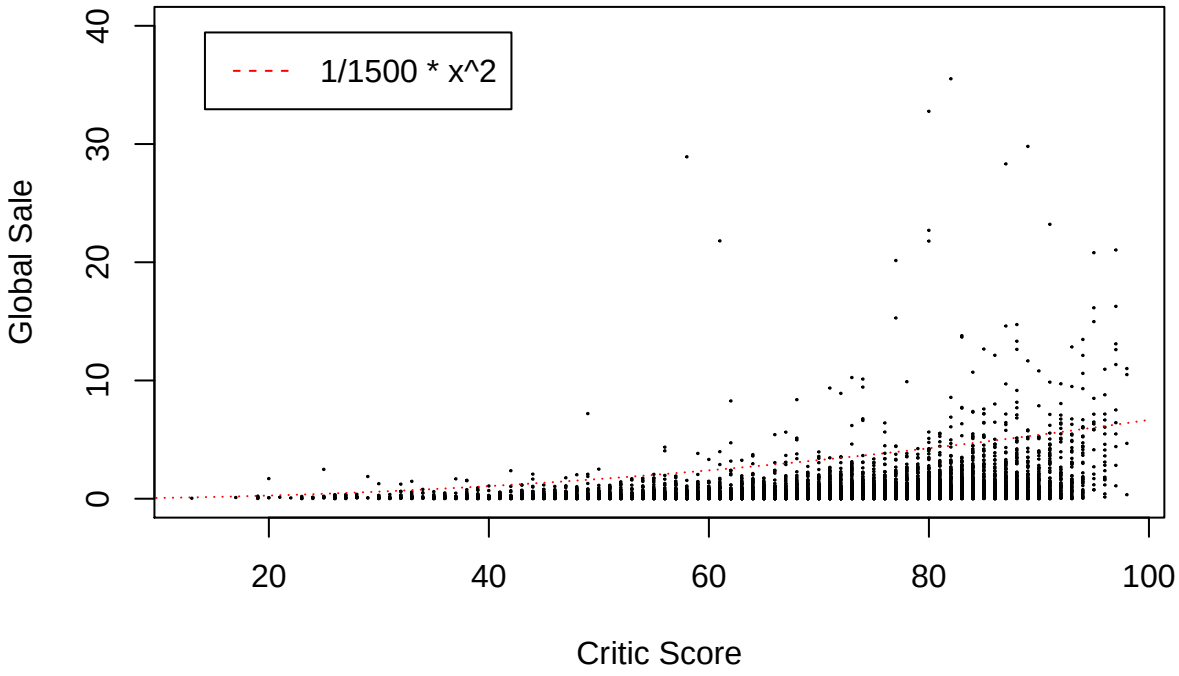
Anova Publisher




```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## data[[classifier]]  303   2447   8.077   2.592 <2e-16 ***
## Residuals          7833  24405   3.116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

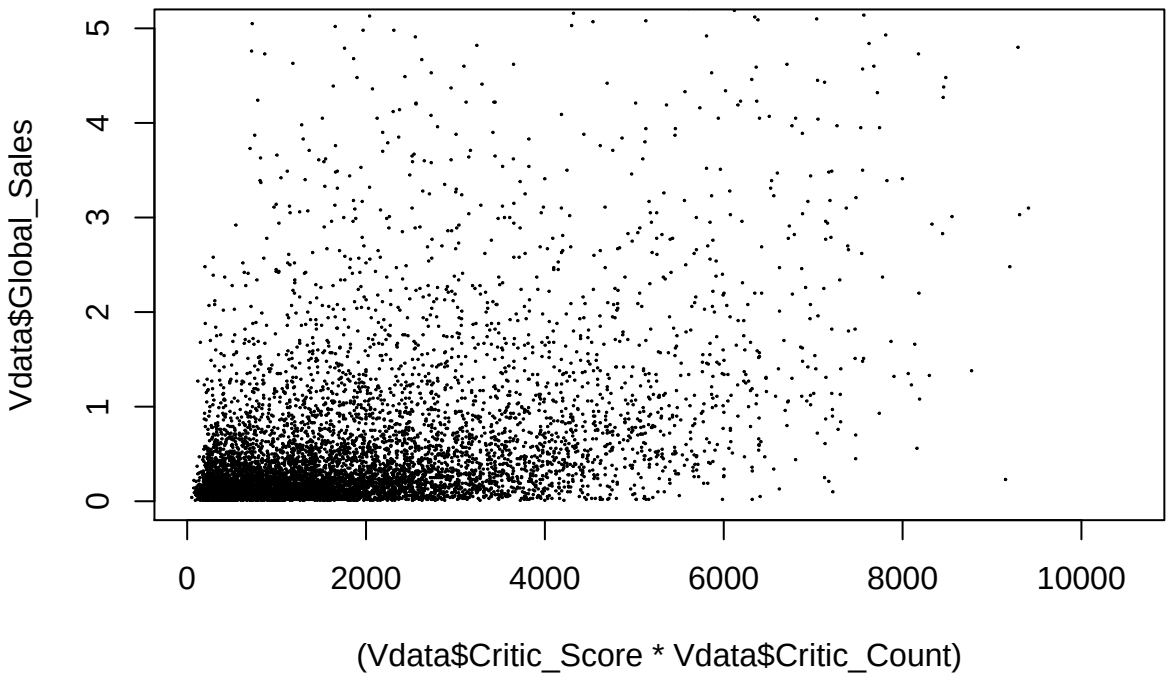
Plot Global Sale ~ Paramemter

Critic_Score defines an upperbound

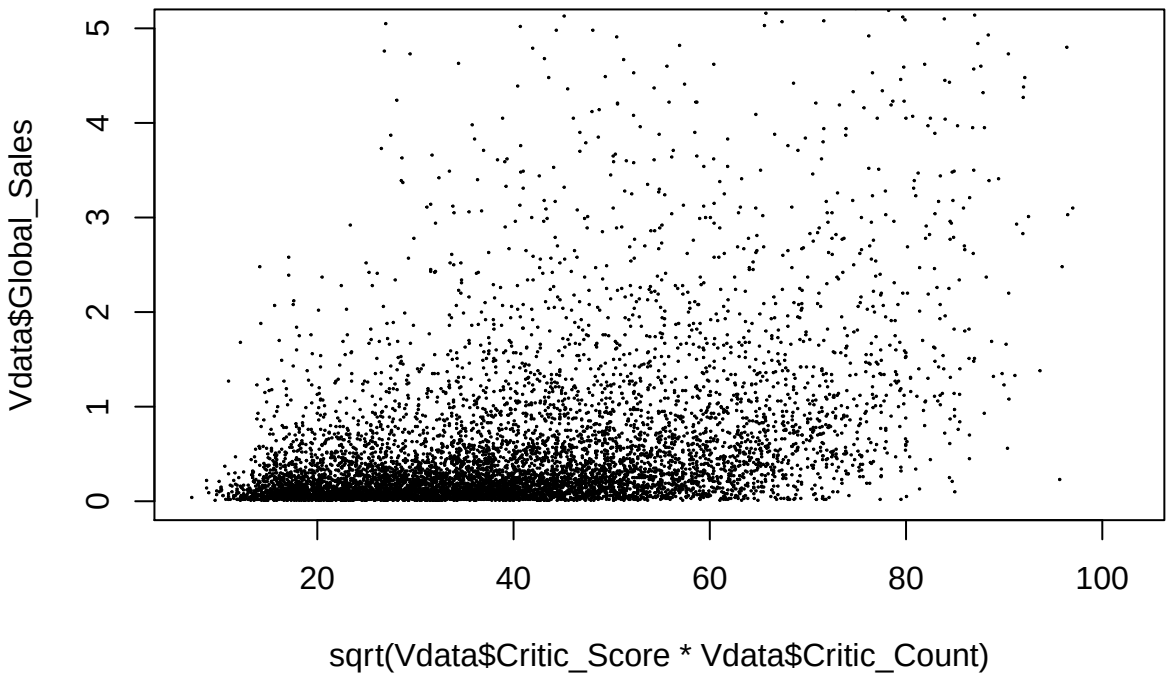


Critic_Score × Critic_count no useful finding

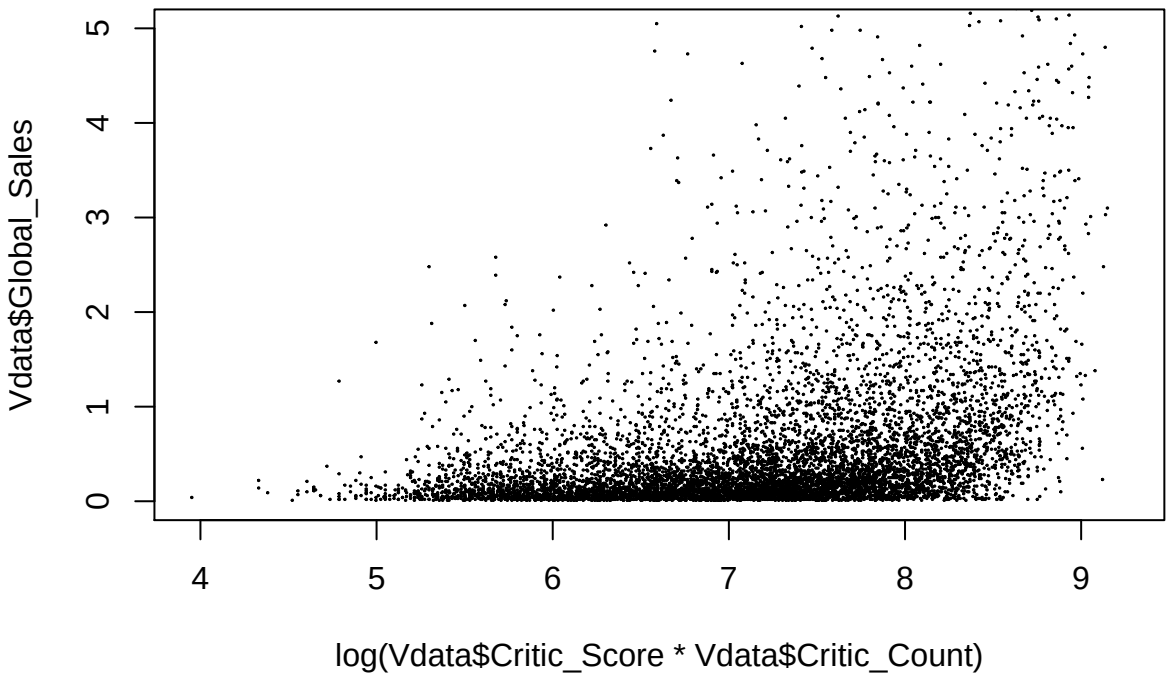
```
plot((Vdata$Critic_Score*Vdata$Critic_Count), ylim=c(0, 5),Vdata$Global_Sales, cex=0.1)
```




```
plot(sqrt(Vdata$Critic_Score*Vdata$Critic_Count), ylim=c(0, 5),Vdata$Global_Sales, cex=0.1)
```




```
plot(log(Vdata$Critic_Score*Vdata$Critic_Count), ylim=c(0, 5),Vdata$Global_Sales, cex=0.1)
```



Test multicolineality

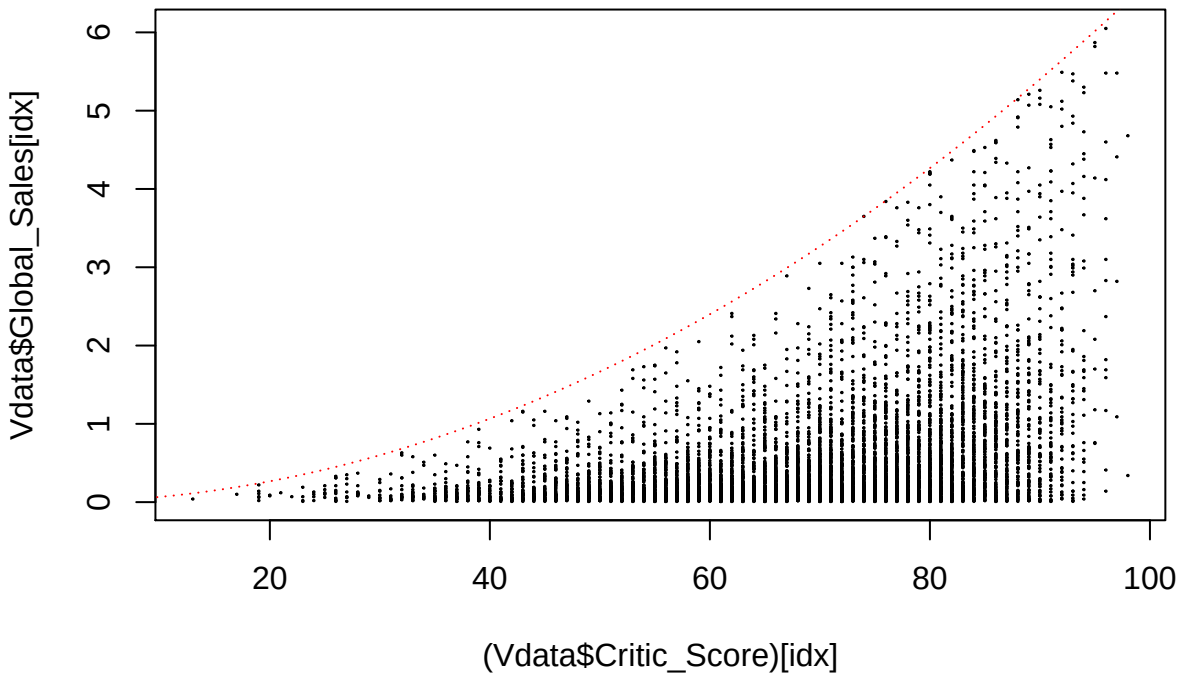
```
cor.test(as.numeric(as.character(Vdata$User_Score)), Vdata$Critic_Score)
```

```
## Warning in cor.test(as.numeric(as.character(Vdata$User_Score)),
## Vdata$Critic_Score): NAs introduced by coercion
```

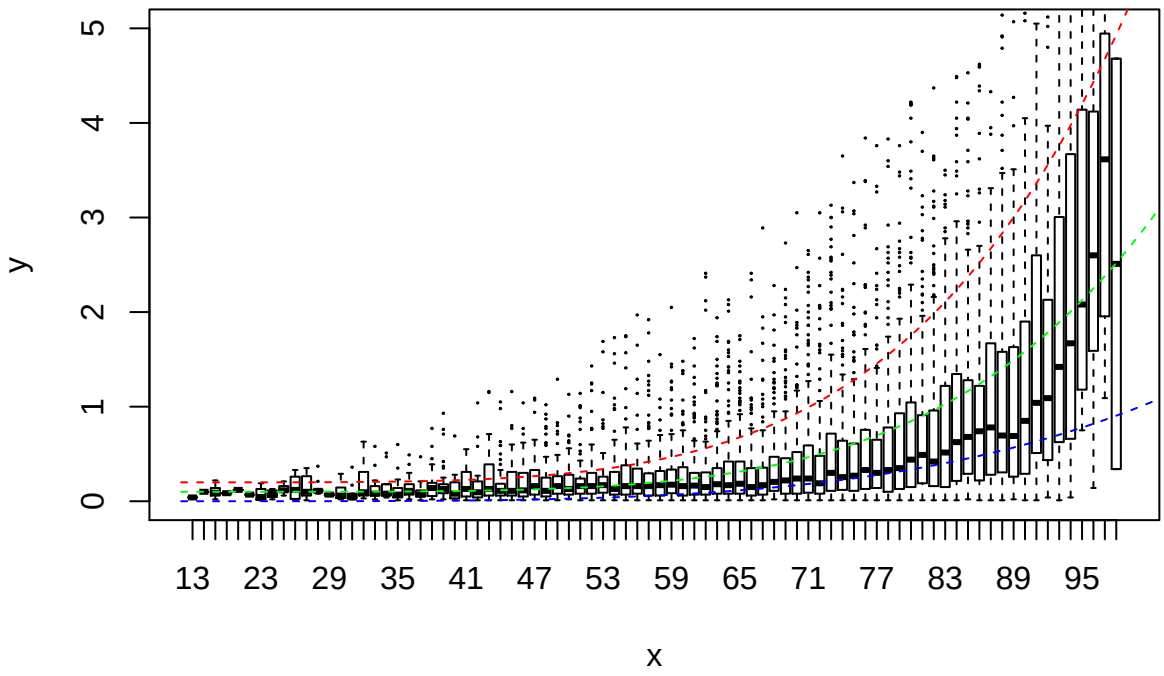
```
##
##  Pearson's product-moment correlation
##
## data:  as.numeric(as.character(Vdata$User_Score)) and Vdata$Critic_Score
## t = 59.769, df = 7015, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5651609 0.5961733
## sample estimates:
##       cor
## 0.5808778
```

Focus on Critic score

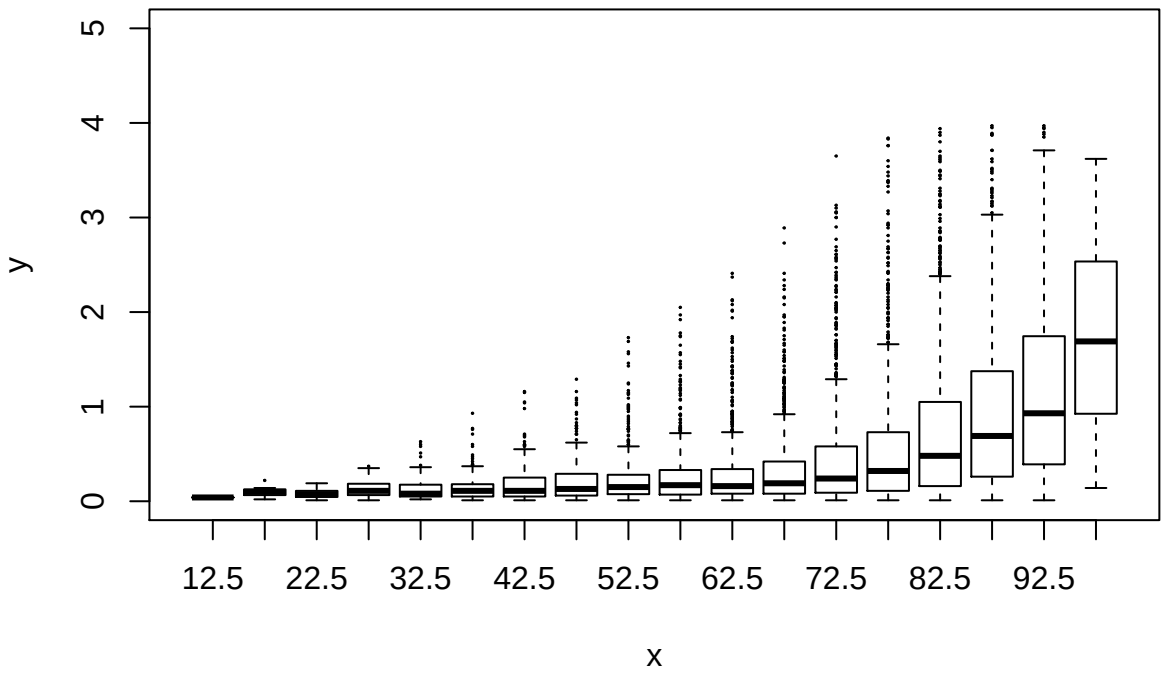
```r
#plot((Vdata$Critic_Score),Vdata$Global_Sales, cex=0.1, ylim=c(0,20))
x <- seq(0,100,len=1000)
y <- 1/1500*x^2
idx <- Vdata$Global_Sales <= (Vdata$Critic_Score)^2/1500
plot((Vdata$Critic_Score)[idx],Vdata$Global_Sales[idx], cex=0.1)
lines(x,y,lty=3,col="red")
```



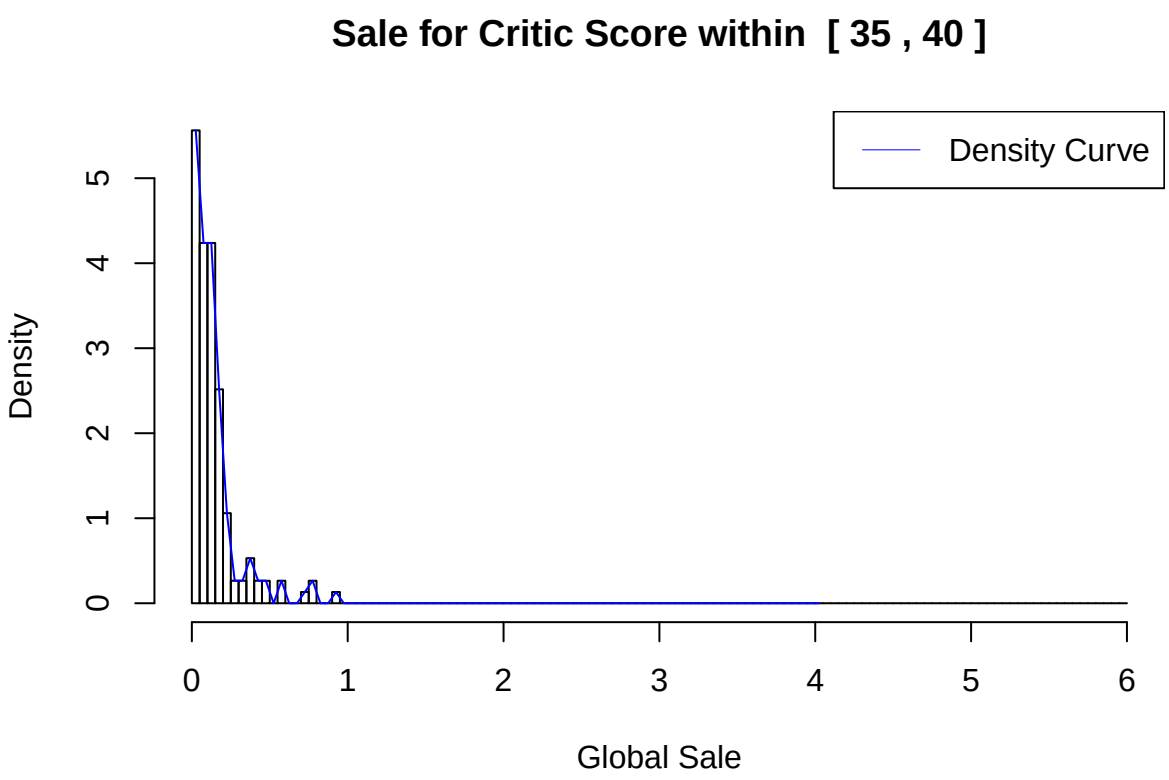

```r
plot(factor((Vdata$Critic_Score)[idx]), ylim=c(0,5),Vdata$Global_Sales[idx], cex=0.1)
f <- 1/20000000*x^4+1/1500000000*x^5+0.2
t <- 1/50000000*x^4
m <- 1/80000000*x^4+1/2000000000*x^5+0.1
lines(x,f,lty=2,col="red")
lines(x,t,lty=2,col="blue")
lines(x,m,lty=2,col="green")
```

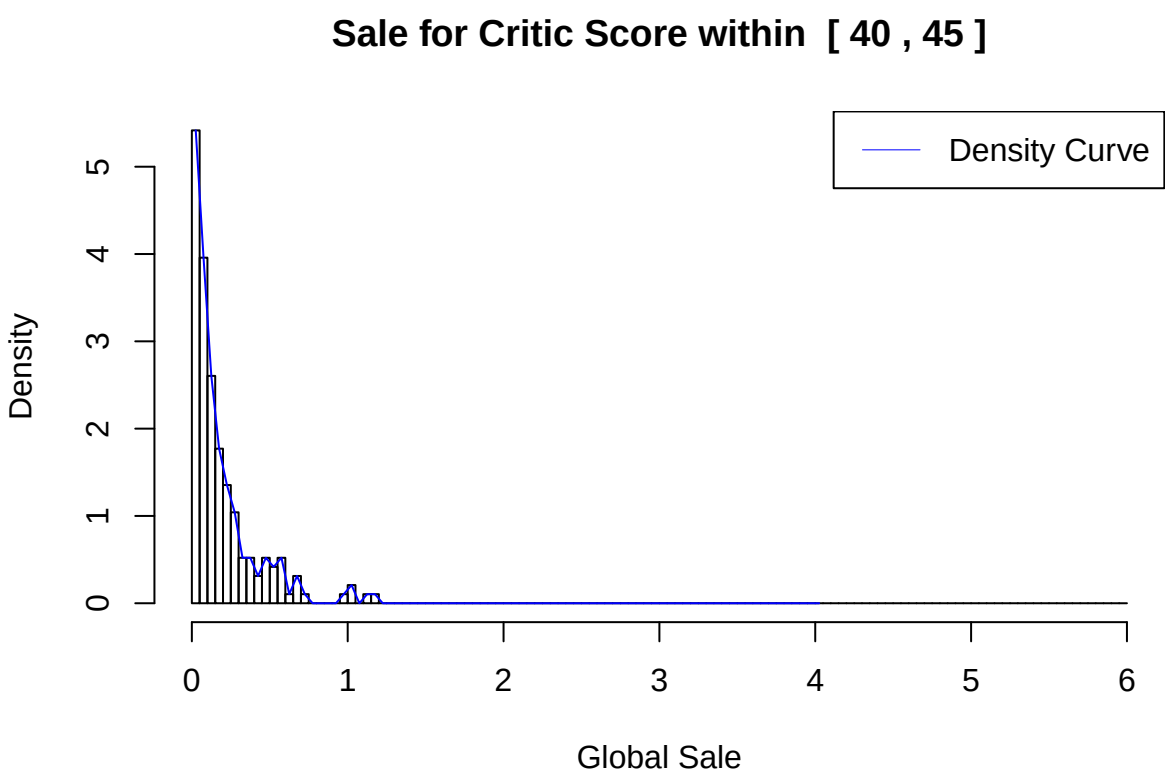


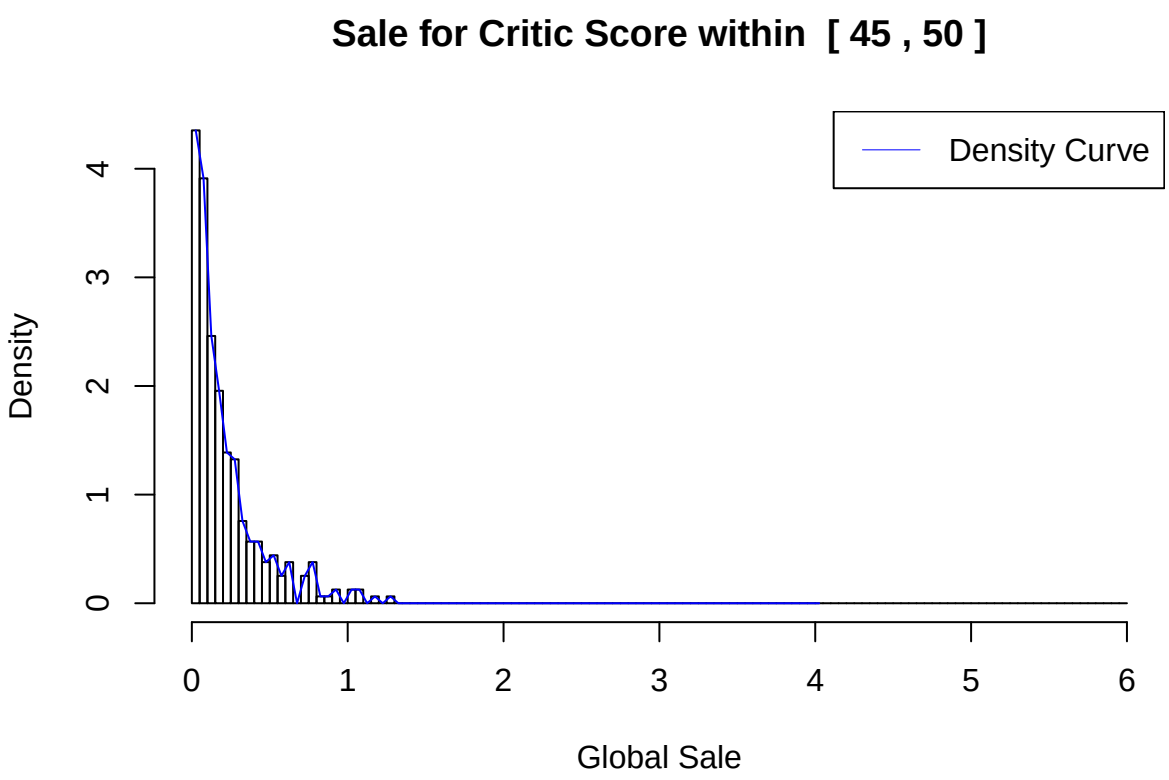

```
# 5 range
Vdata <- Vdata[idx, ]
Vdata <- Vdata[Vdata$Global_Sales <= 4, ]
ran <- rangef(Vdata$Critic_Score)
plot(factor(ran), ylim=c(0,5),Vdata$Global_Sales, cex=0.1)
```

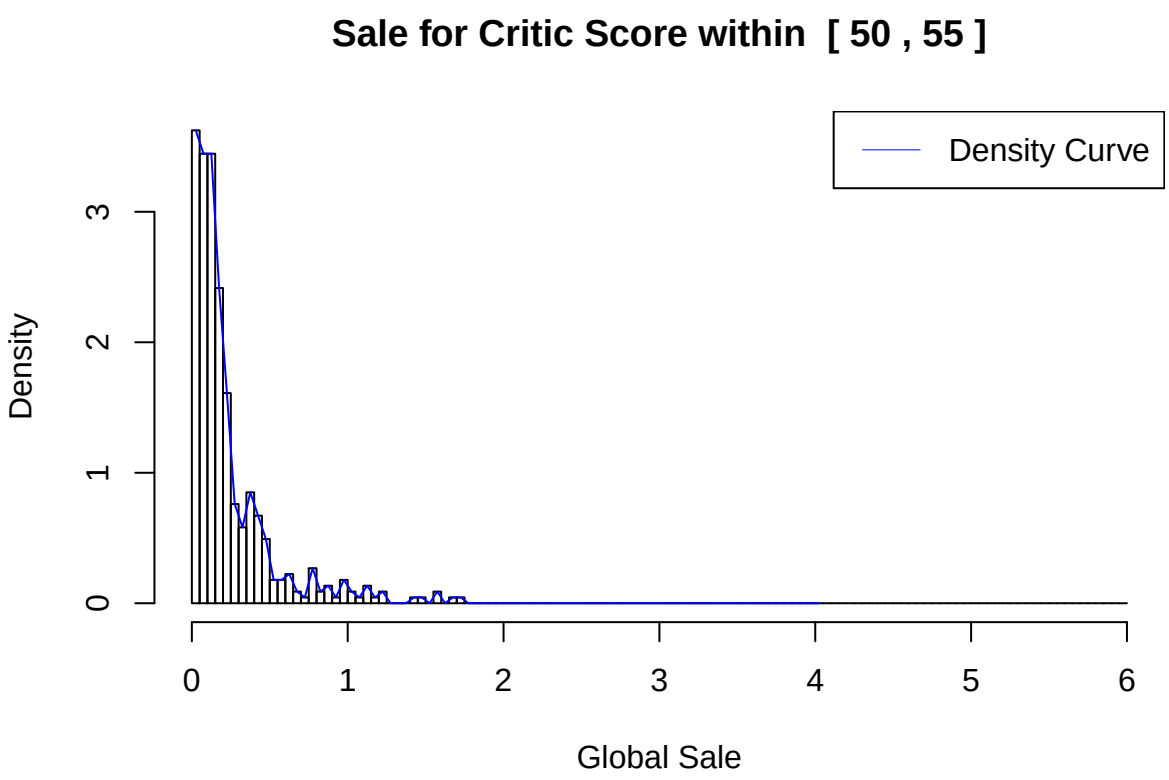


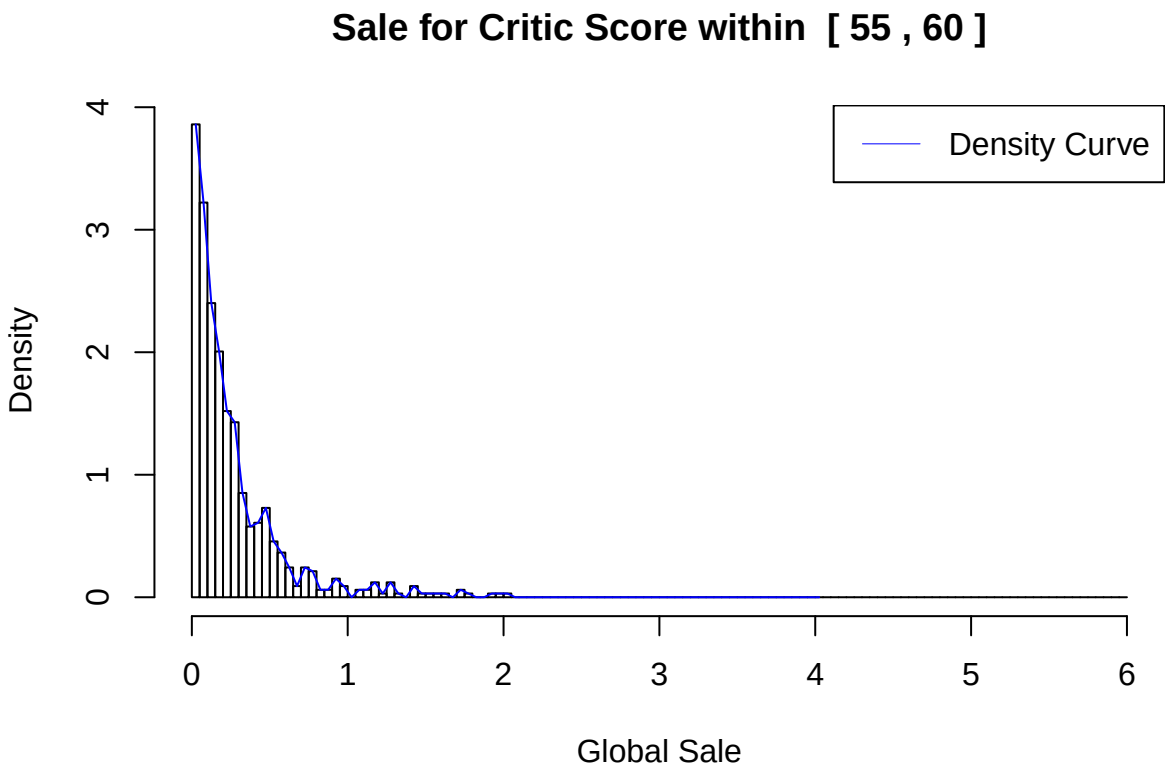

```
for (i in seq(11)) {
  hist(Vdata$Global_Sales[ran == (32.5+5*i)], breaks=seq(0,6,by=0.05), main =
         paste("Sale for Critic Score within ", "[",as.character(32.5+5*i-2.5), ",",
        as.character(32.5+5*i+2.5),"]"), xlab="Global Sale", probability = TRUE)
  lines(seq(0,4,by=0.05)+0.025, get_density(Vdata$Global_Sales[ran == (32.5+5*i)], seq(0,4,by=0.05), 0.
  legend("topright", "Density Curve", lwd = 0.5, col = "blue")
}
```

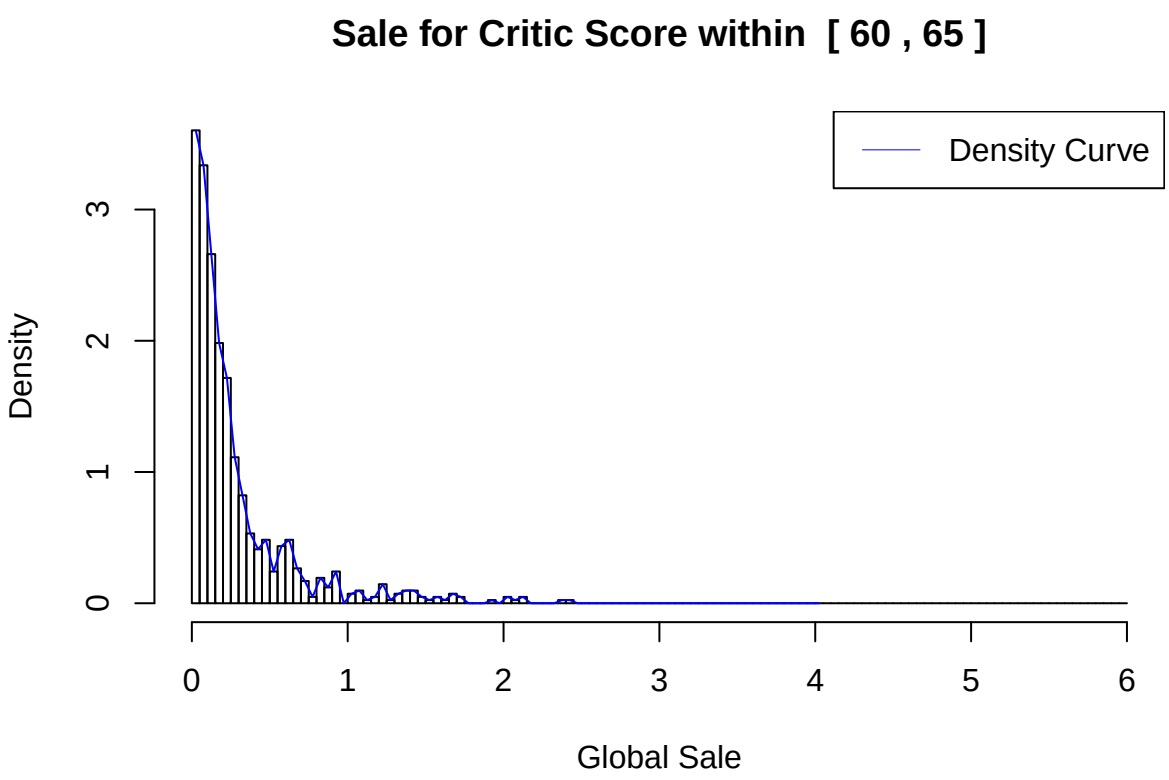

Sale for Critic Score within [35 , 40]
Density Curve
Density
0
1
2
3
4
5
0
1
2
3
4
5
6
Global Sale

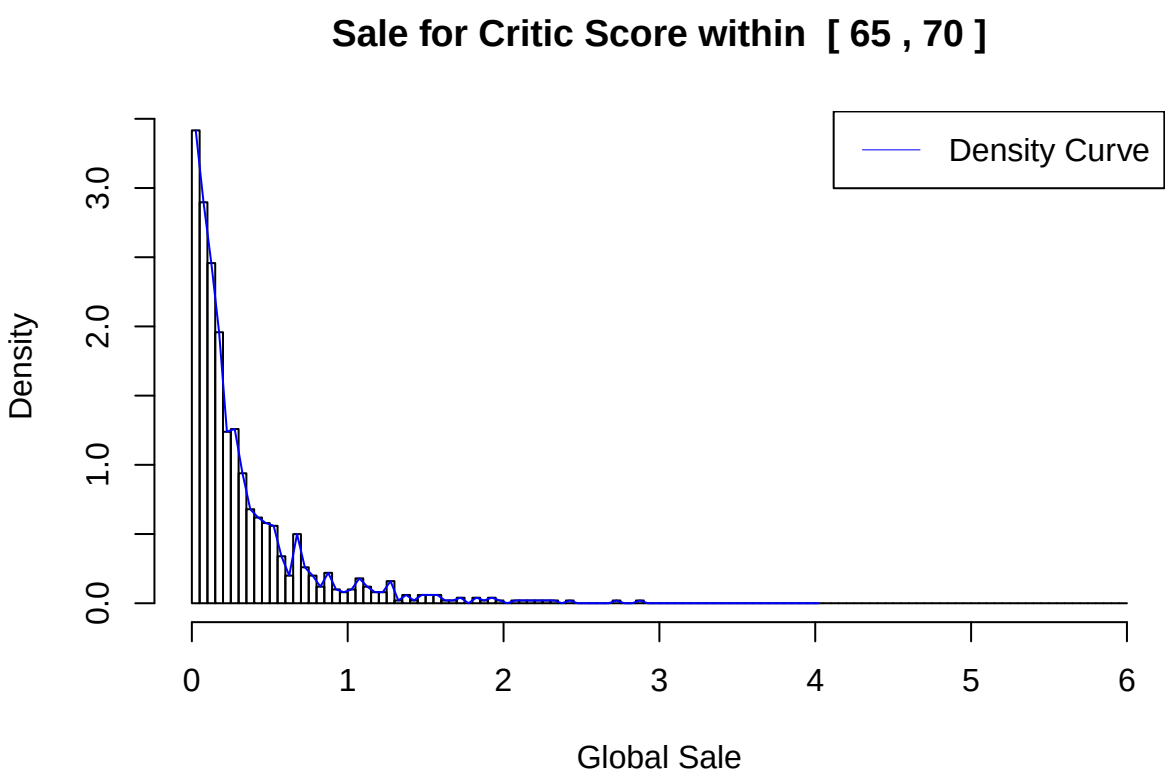

Sale for Critic Score within [40 , 45]
Density Curve
Density
0
1
2
3
4
5
0
1
2
3
4
5
6
Global Sale

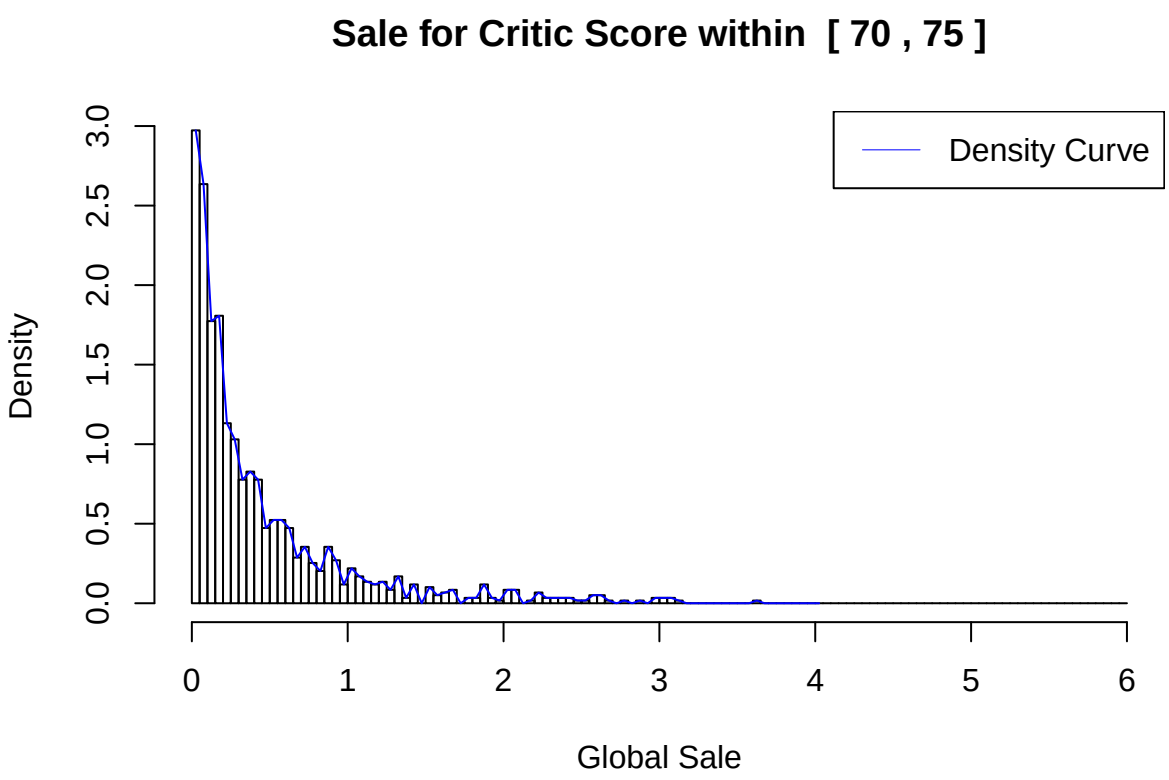

Sale for Critic Score within [45 , 50]
Density Curve
Density
0
1
2
3
4
0
1
2
3
4
5
6
Global Sale

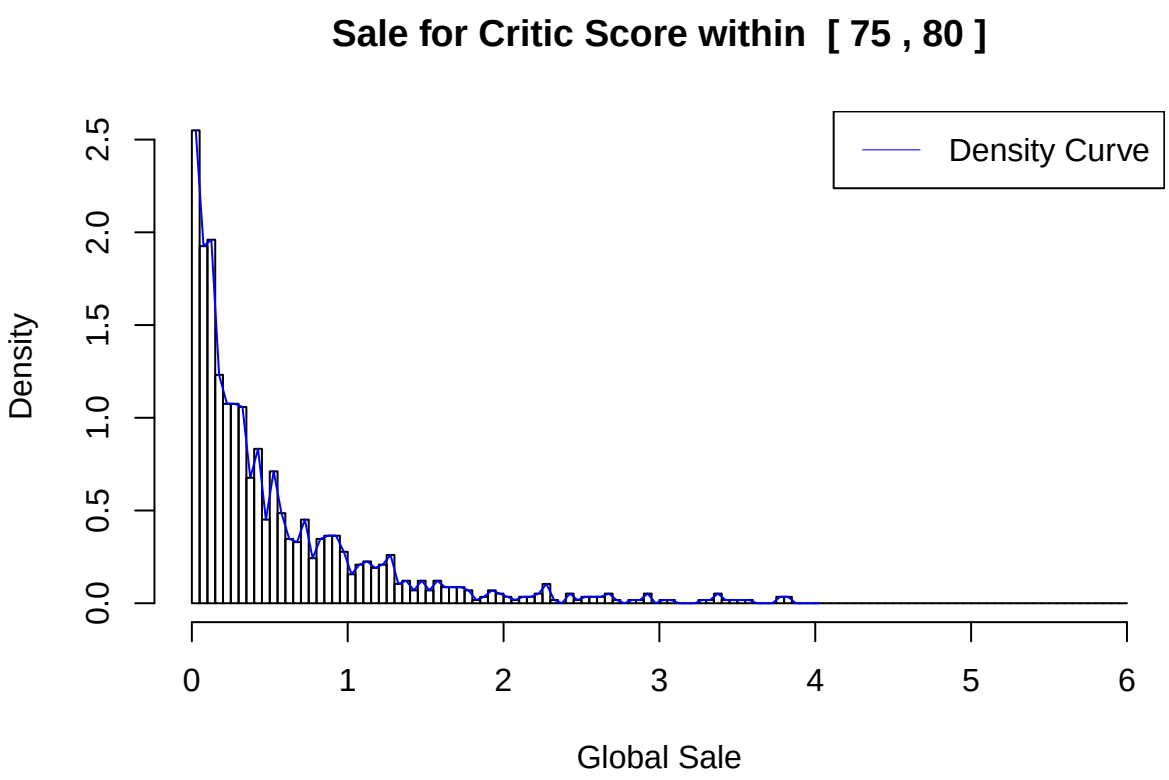

Sale for Critic Score within [50 , 55]
Density Curve
Density
0
1
2
3
0
1
2
3
4
5
6
Global Sale

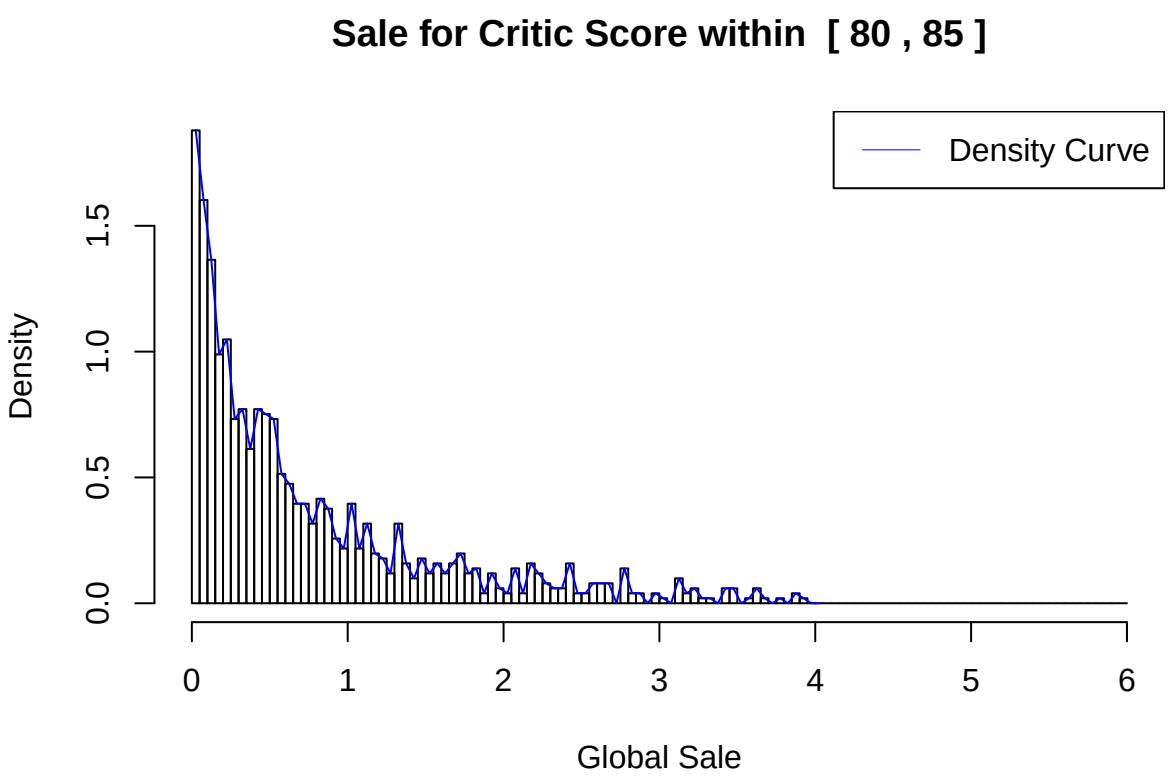

Sale for Critic Score within [55 , 60]
Density Curve
Density
0
1
2
3
4
0
1
2
3
4
5
6
Global Sale

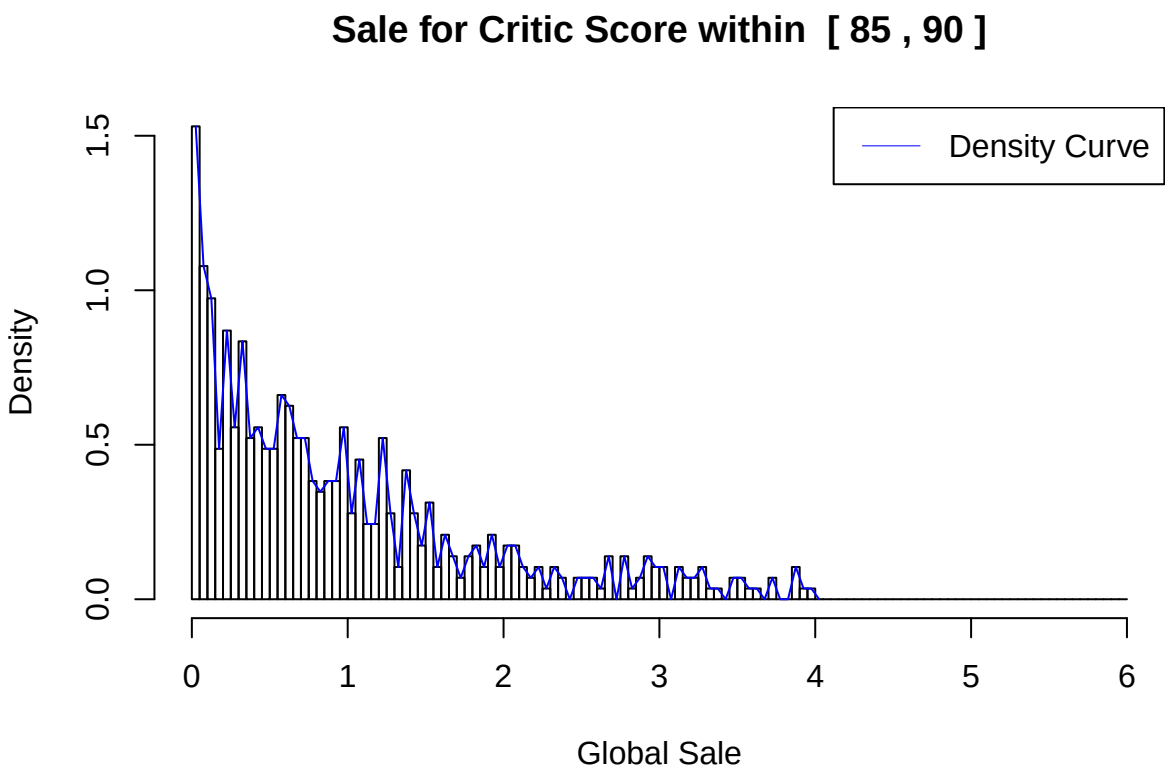

Sale for Critic Score within [60 , 65]
Density Curve
Density
0
1
2
3
0
1
2
3
4
5
6
Global Sale


Sale for Critic Score within [65 , 70]
Density Curve
Density
0.0
1.0
2.0
3.0
0
1
2
3
4
5
6
Global Sale


Sale for Critic Score within [70 , 75]
Density Curve
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
6
Global Sale


Sale for Critic Score within [75 , 80]
Density Curve
Density
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3
4
5
6
Global Sale


Sale for Critic Score within [80 , 85]
Density Curve
Density
0.0
0.5
1.0
1.5
0
1
2
3
4
5
6
Global Sale



The goal is to approximate the blue density curve given the value of Critic Score

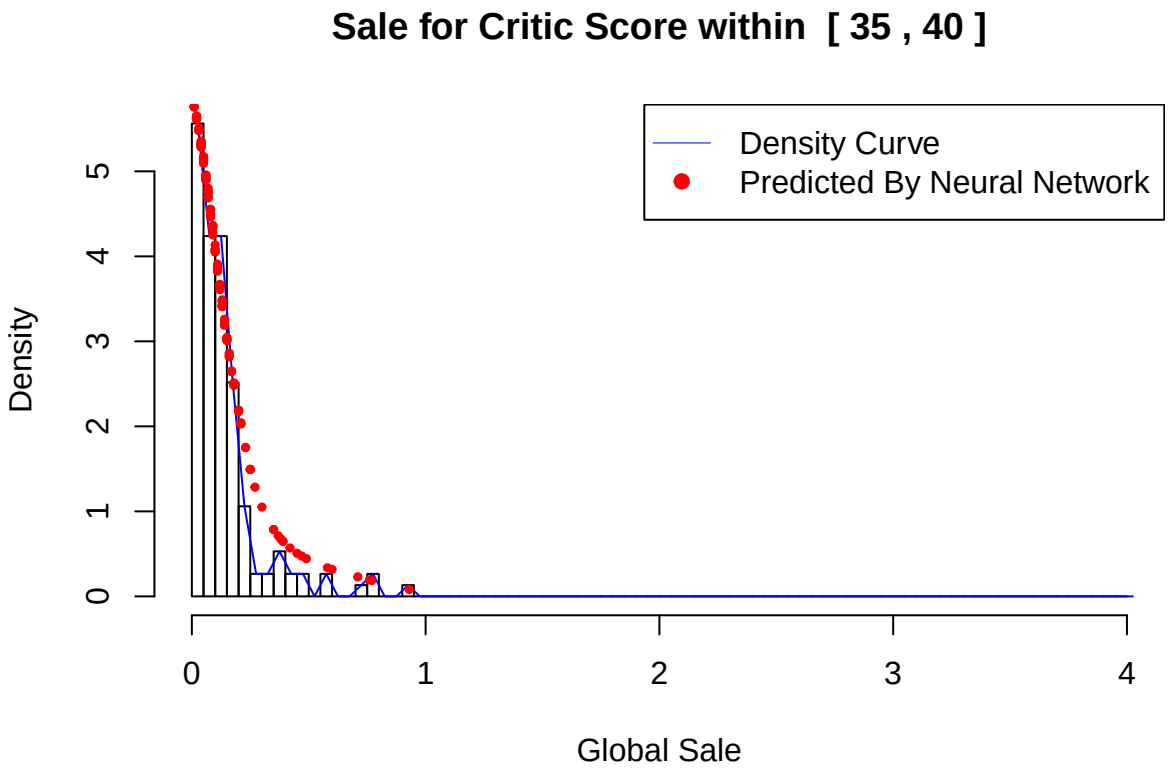
Objective function: $\sum(f(CS, GS) - \hat{f}(CS, GS))^2$

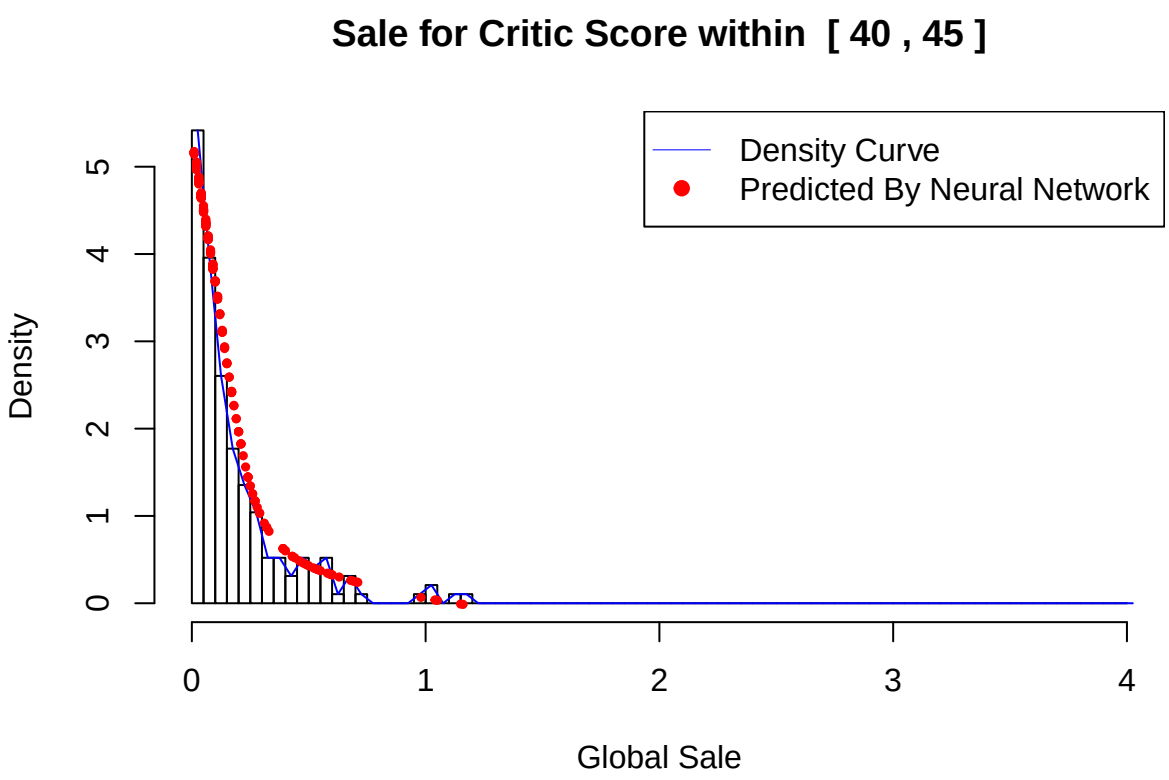
```
ran <- rangef(Vdata$Critic_Score)
dataf <- data.frame("Global_Sale" = numeric(0), "Critic_Score" = numeric(0), "Density" = numeric(0))
for (i in seq(11)) {
  gs <- Vdata$Global_Sales[ran == (32.5+5*i)]
  cs <- Vdata$Critic_Score[ran == (32.5+5*i)]
  ccount <- log10(Vdata$Critic_Count[ran == (32.5+5*i)])
  ds <- get_density(Vdata$Global_Sales[ran == (32.5+5*i)], seq(0,4,by=0.05), 0.05)
  ds_idx <- get_density_idx(gs)
  cs_range <- 32.5+5*i
  ds <- ds[ds_idx]
  dataf <- rbind(dataf, cbind(gs, cs_range,ccount, cs,ds))
}
write.csv(dataf, "CS_GL_FQ.csv")
```

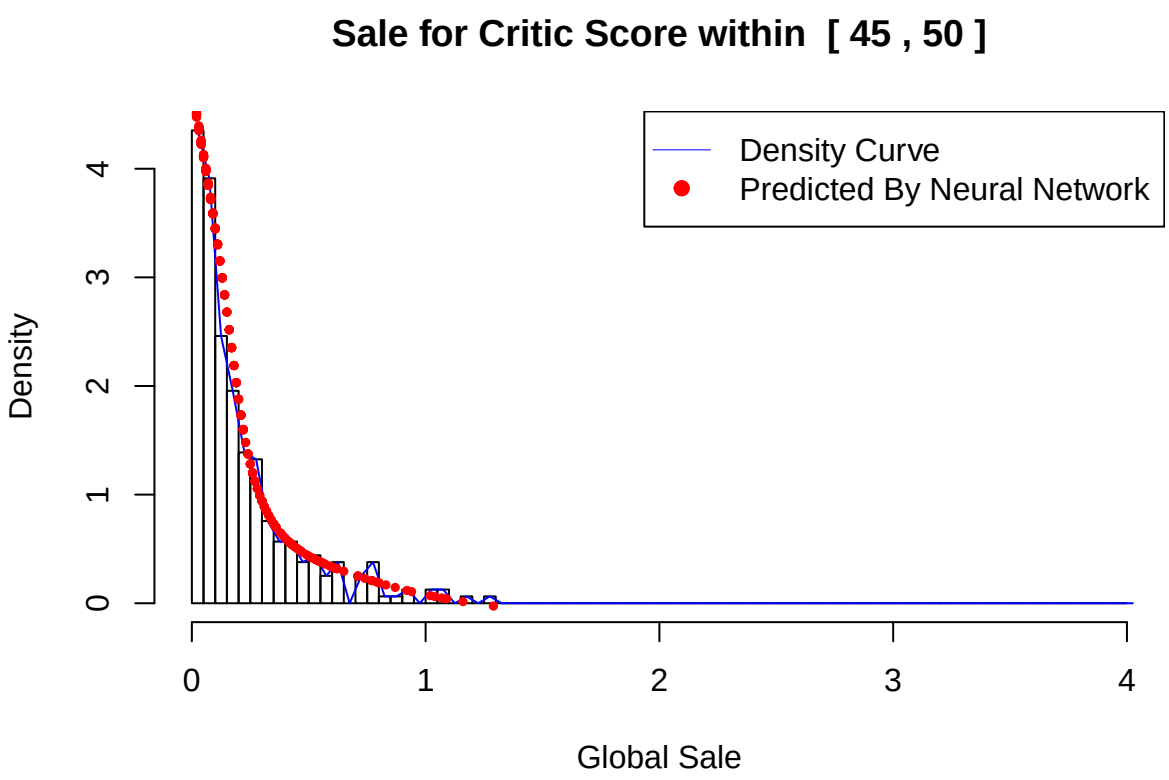
Use elu 4-6-2-1 network Adam MSE:0.0397418439388827515

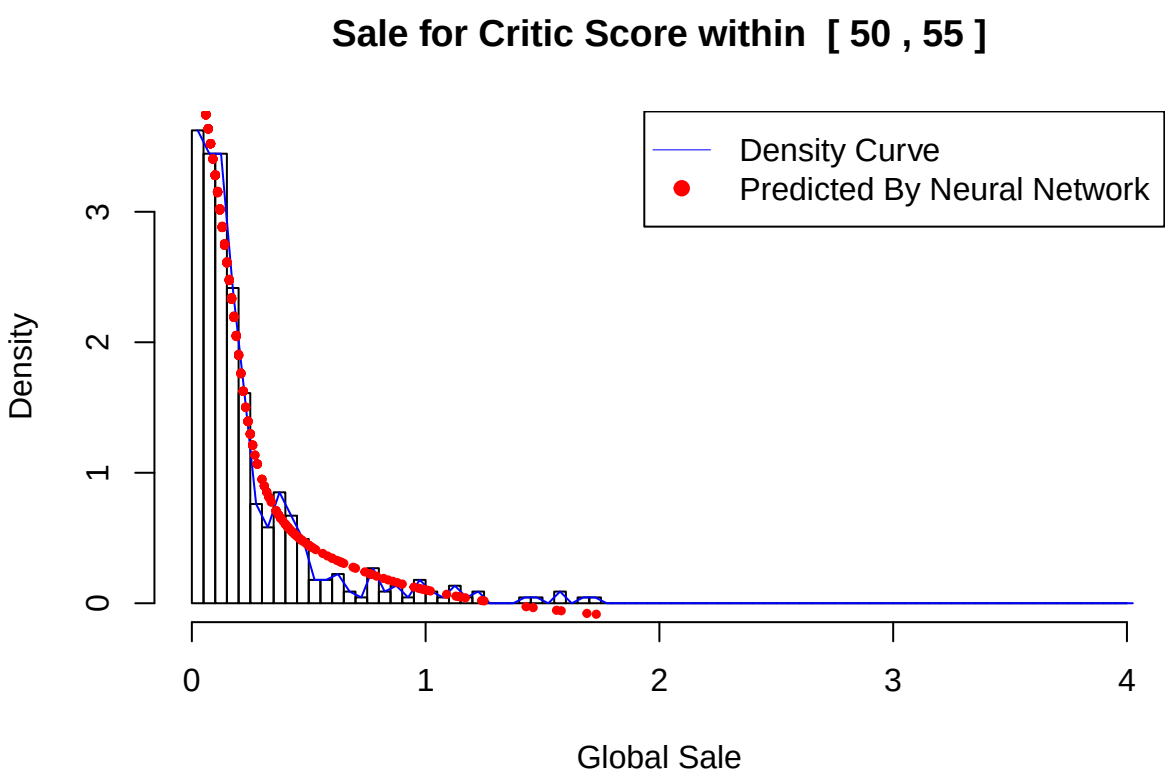
```
Pdata <- read.csv('predictedData.csv')
for (i in seq(11)) {
```

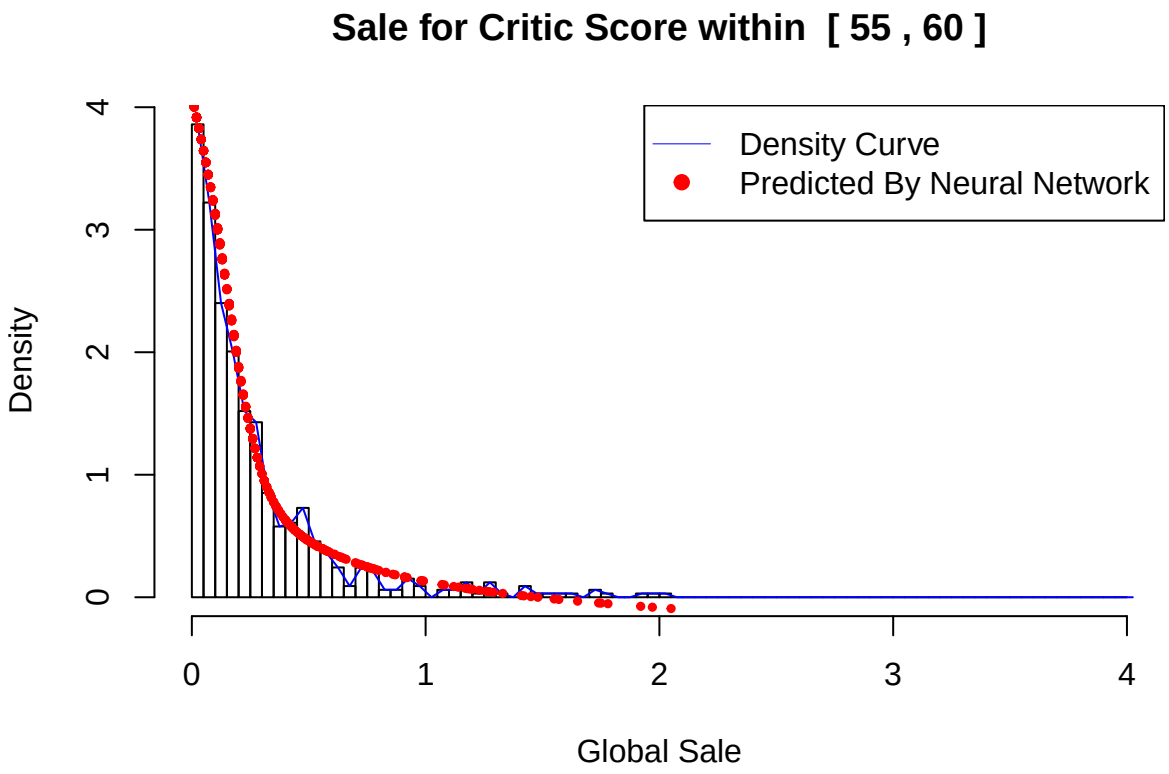
```
  hist(Vdata$Global_Sales[ran == (32.5+5*i)], breaks=seq(0,4,by=0.05), main =
         paste("Sale for Critic Score within ", "[",as.character(32.5+5*i-2.5), ",",
        as.character(32.5+5*i+2.5),"]"), xlab="Global Sale", probability = TRUE)
  lines(seq(0,4,by=0.05)+0.025, get_density(Vdata$Global_Sales[ran == (32.5+5*i)], seq(0,4,by=0.05), 0.
  gs <- Pdata$gs[Pdata$cs_range == (32.5+5*i)]
  y <- Pdata[[6]][Pdata$cs_range == (32.5+5*i)]
  points(gs, y, col='red', pch=19,cex=0.5)
  legend("topright", c("Density Curve", "Predicted By Neural Network"), lwd = c(0.5, NA), col = c("blue
}
```

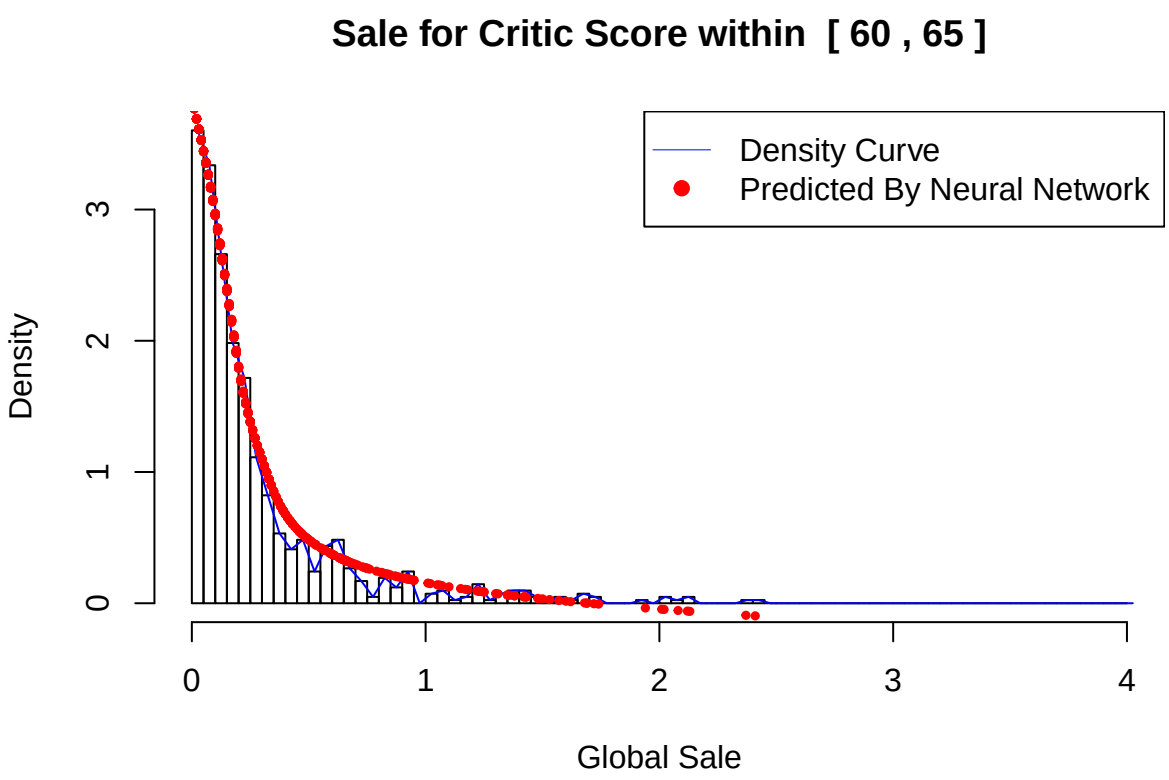


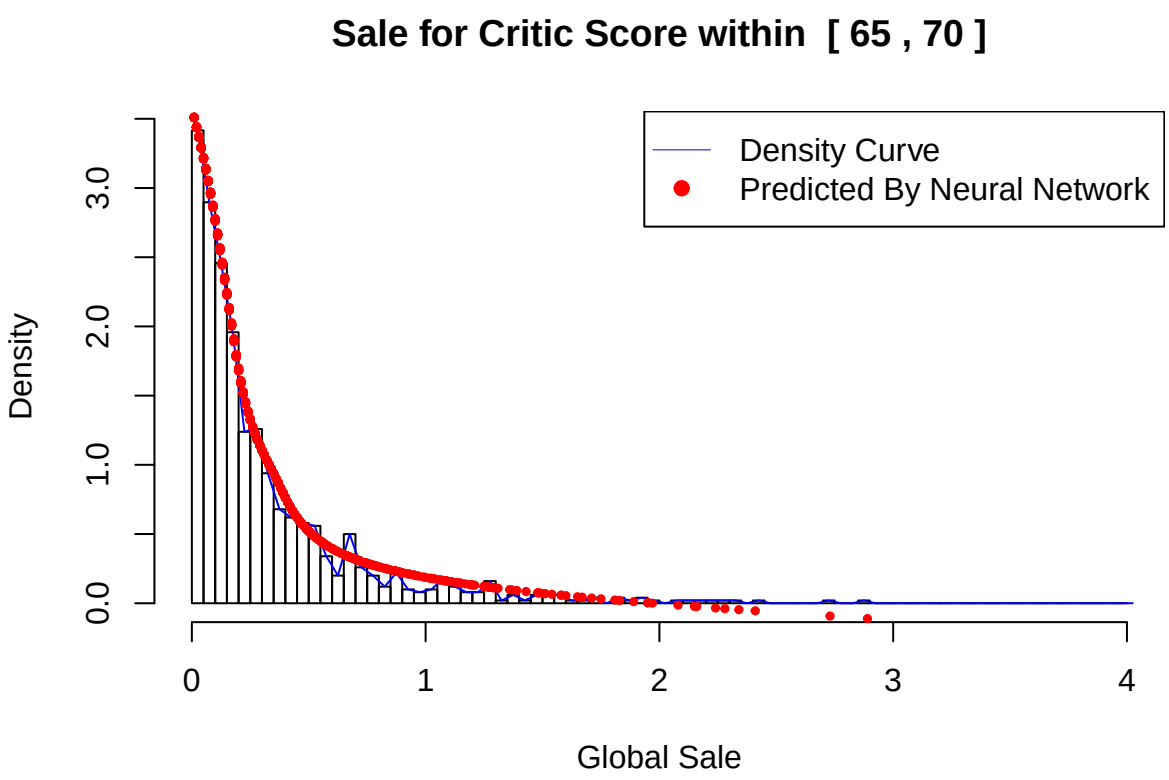

Sale for Critic Score within [40 , 45]
Density Curve
Predicted By Neural Network
Density
0
1
2
3
4
5
0
1
2
3
4
Global Sale

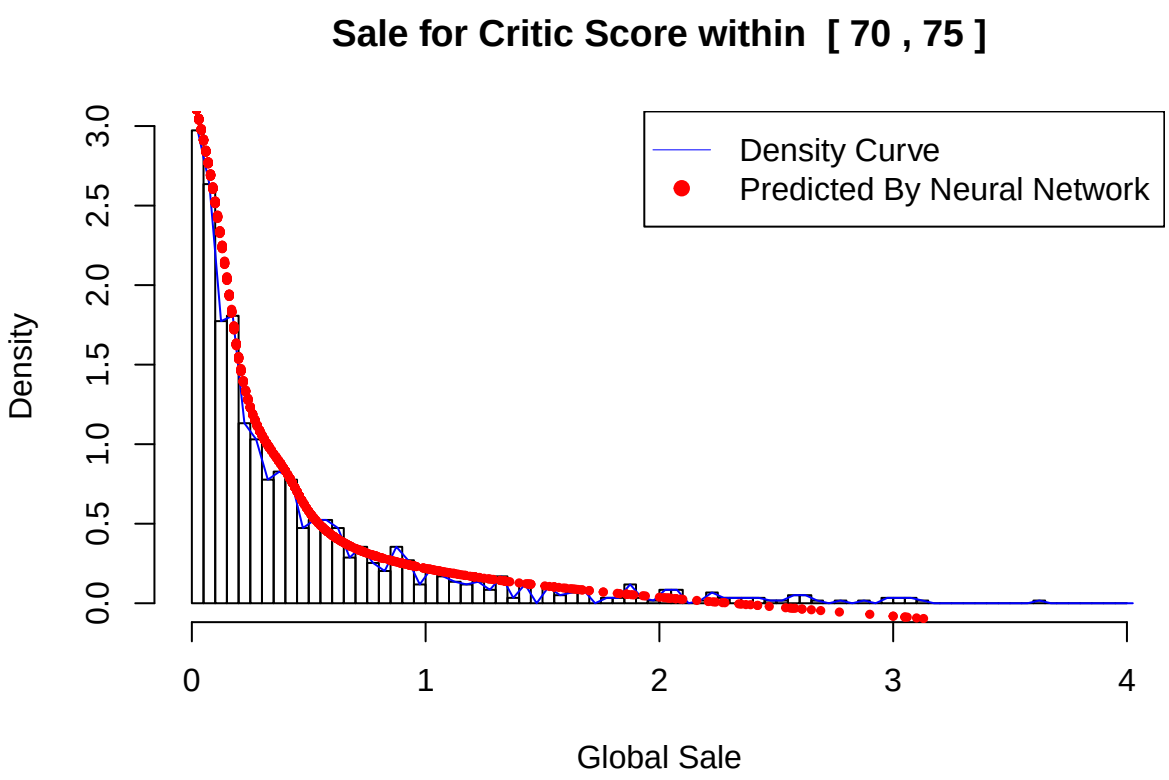

Sale for Critic Score within [45 , 50]
Density Curve
Predicted By Neural Network
Density
0
1
2
3
4
Global Sale

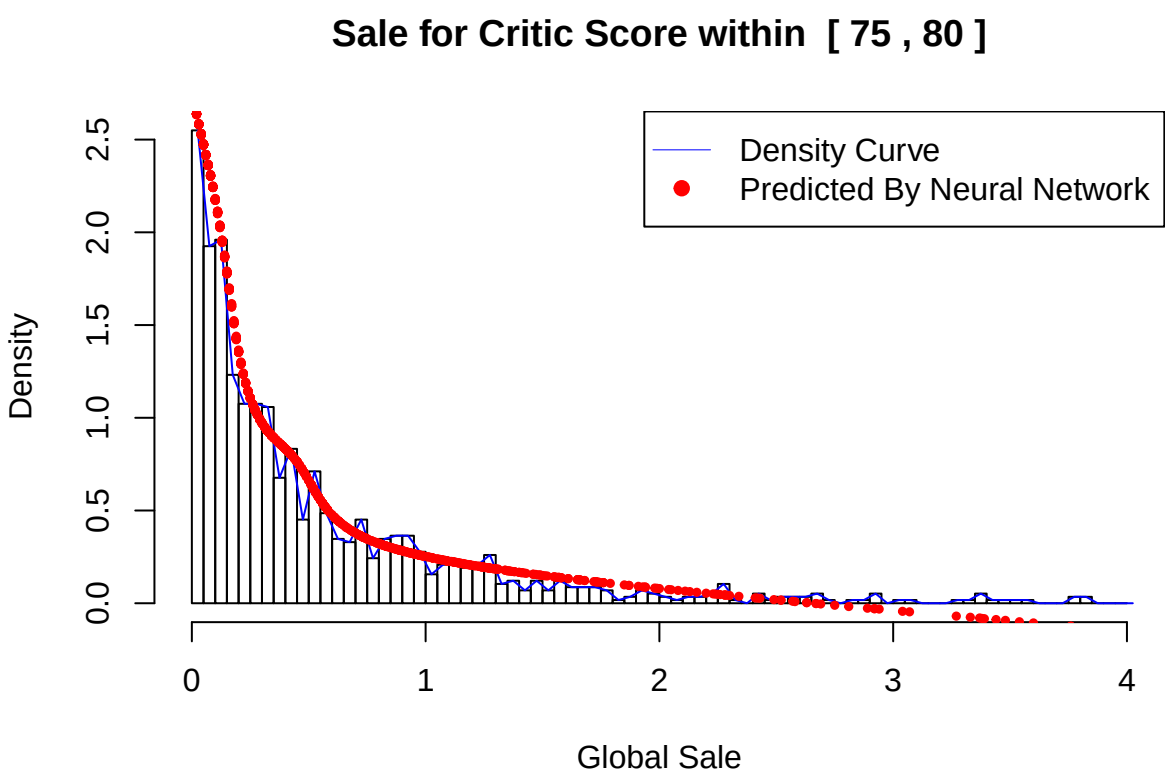

Sale for Critic Score within [50 , 55]
Density Curve
Predicted By Neural Network
Density
0
1
2
3
0
1
2
3
4
Global Sale

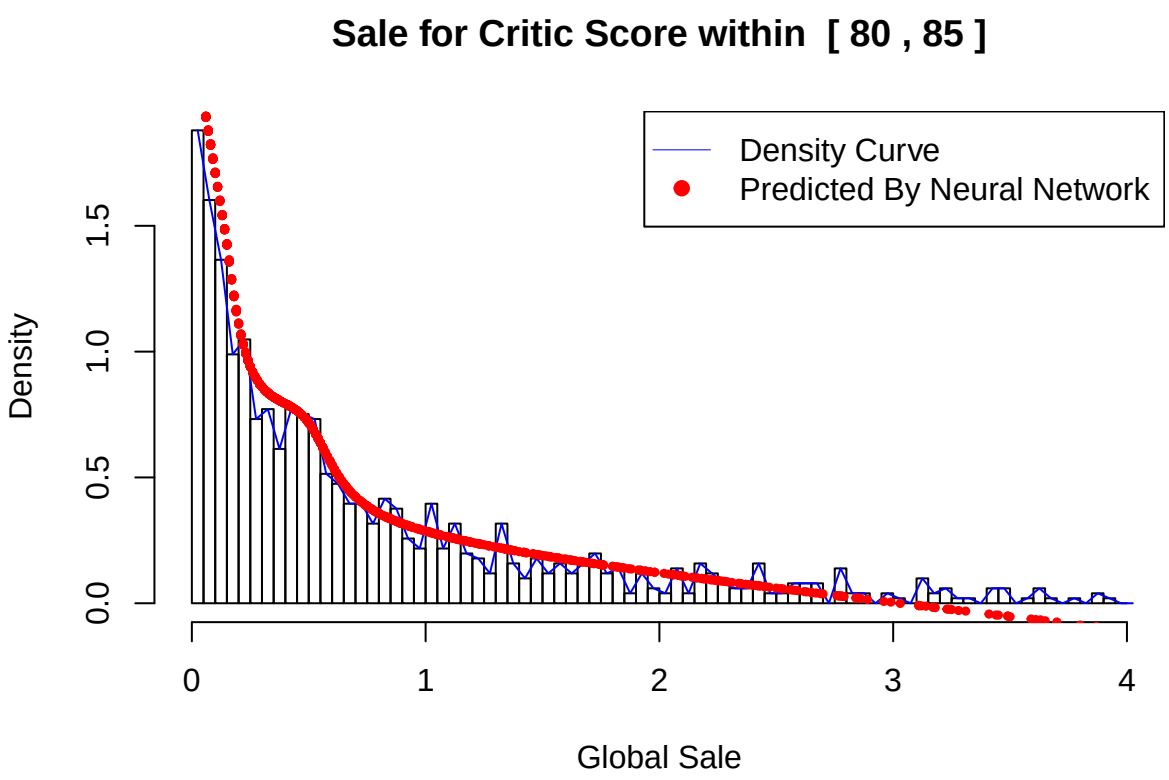

Sale for Critic Score within [55 , 60]
Density Curve
Predicted By Neural Network
Density
0
1
2
3
4
Global Sale

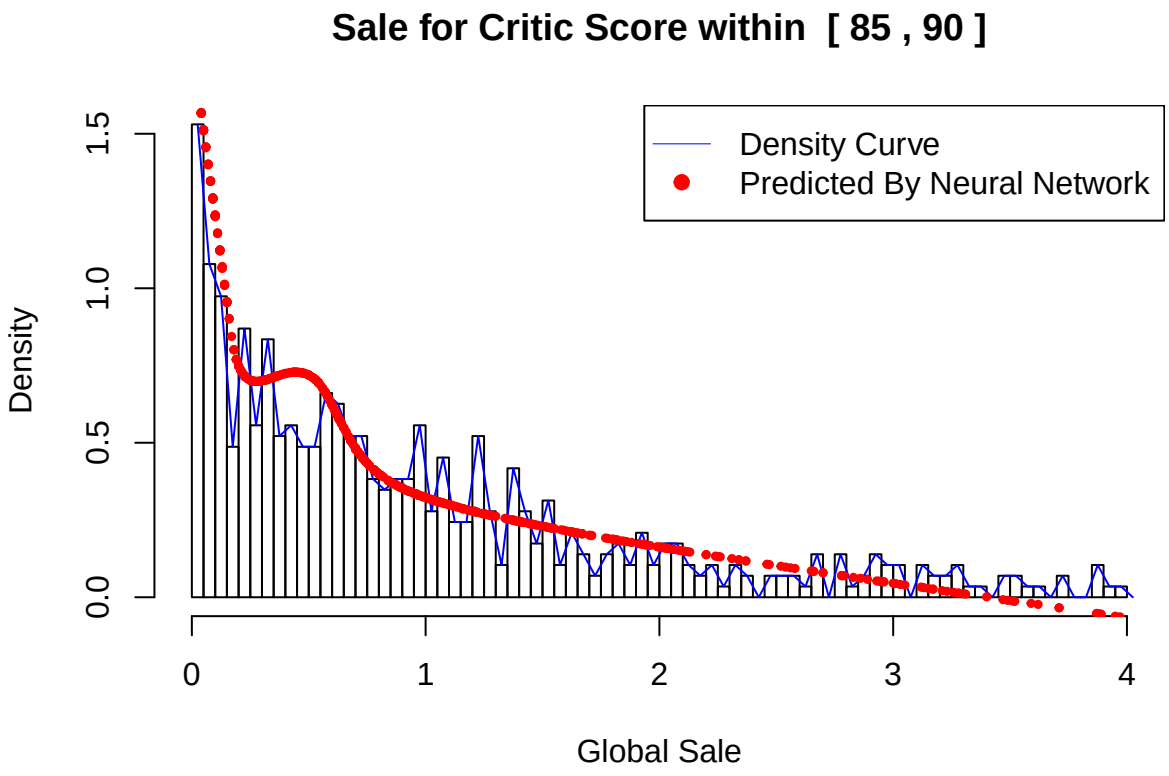

Sale for Critic Score within [60 , 65]
Density Curve
Predicted By Neural Network
Density
0
1
2
3
0
1
2
3
4
Global Sale


Sale for Critic Score within [65 , 70]
Density Curve
Predicted By Neural Network
Density
0.0
1.0
2.0
3.0
0
1
2
3
4
Global Sale

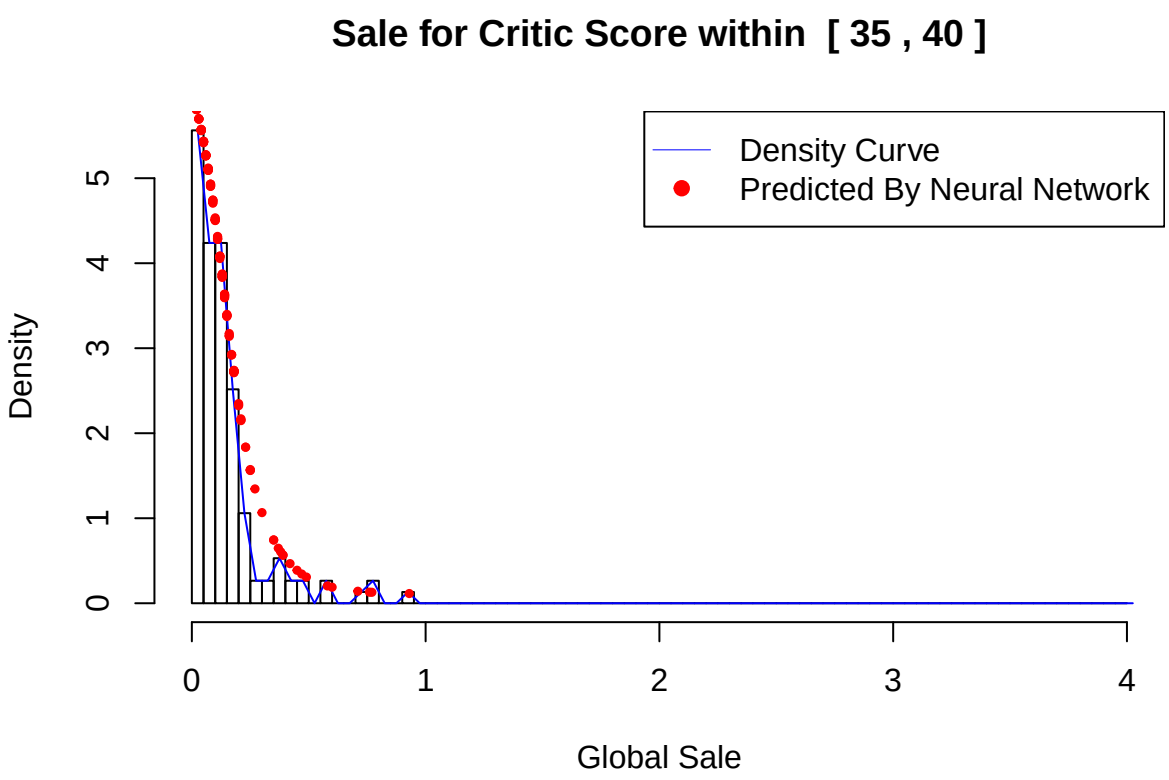

Sale for Critic Score within [70 , 75]
Density Curve
Predicted By Neural Network
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
Global Sale

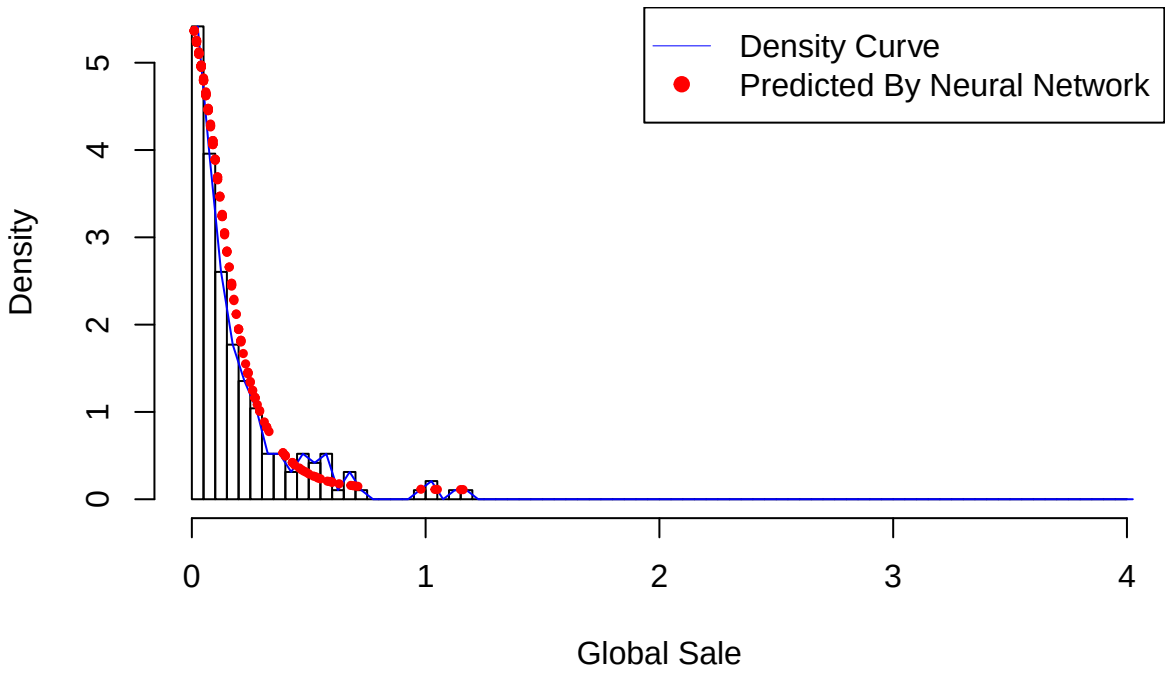

Sale for Critic Score within [75 , 80]
Density Curve
Predicted By Neural Network
Density
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3
4
Global Sale

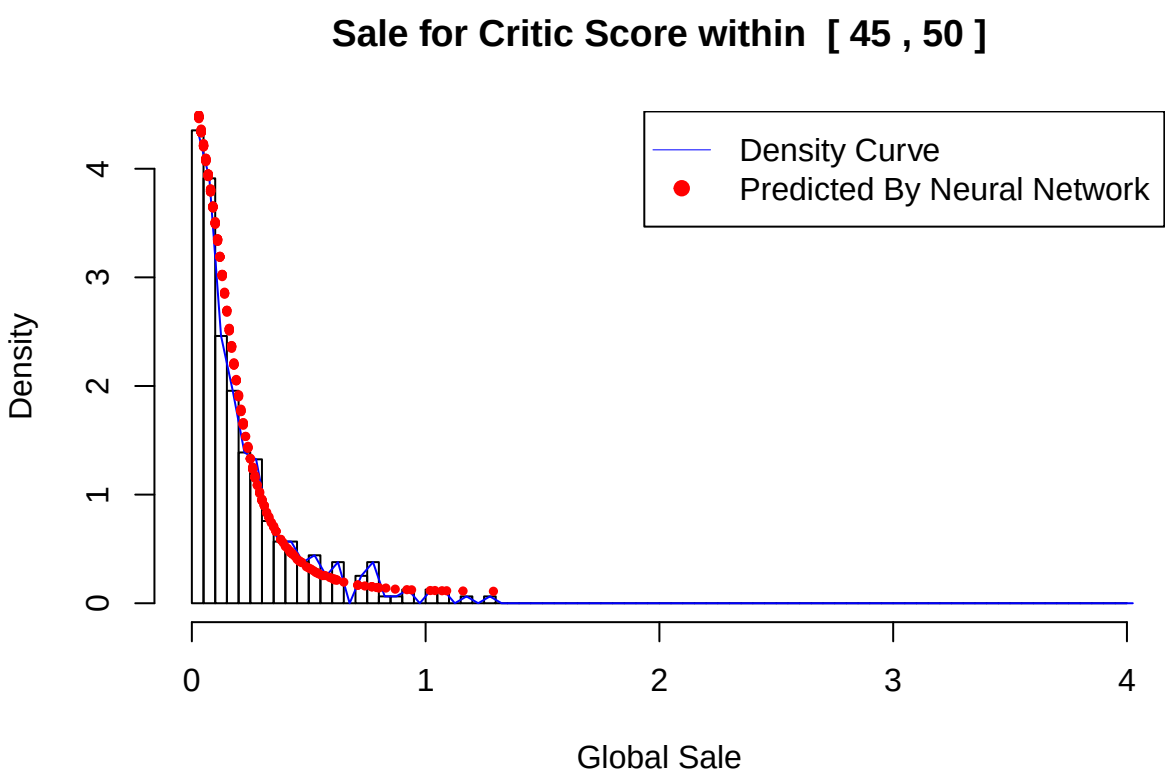

Sale for Critic Score within [80 , 85]
Density Curve
Predicted By Neural Network
Density
0.0
0.5
1.0
1.5
0
1
2
3
4
Global Sale

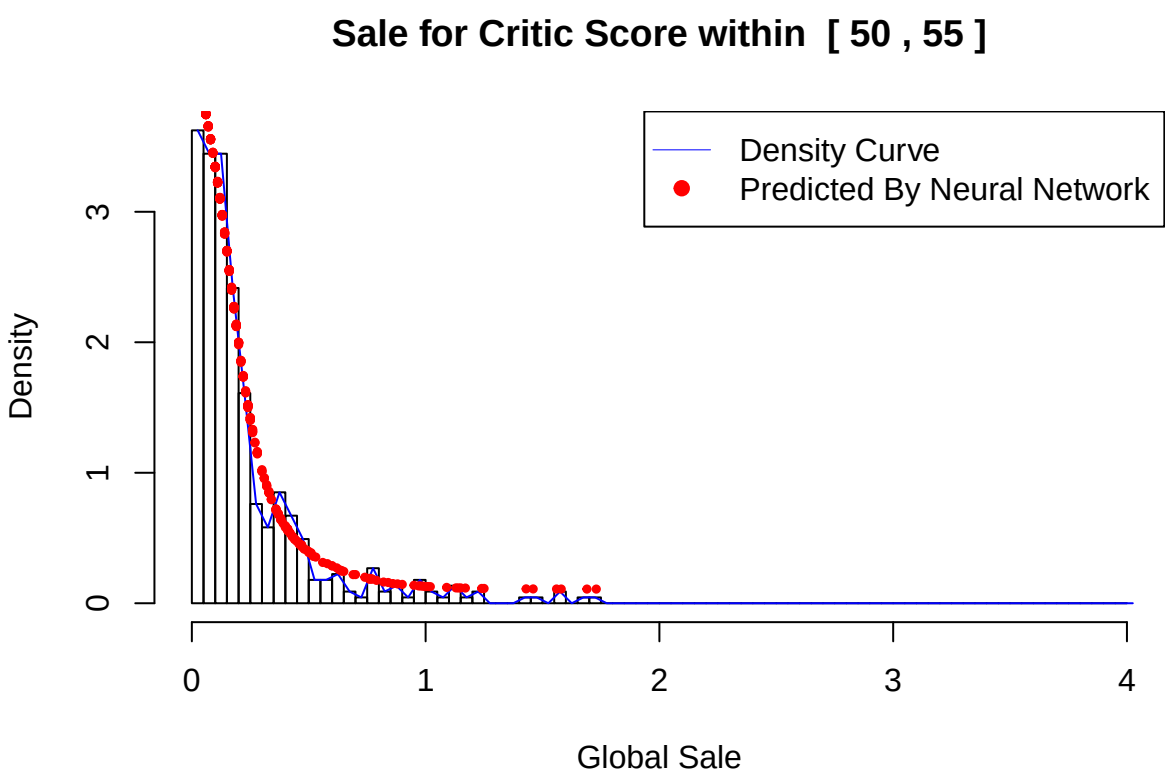


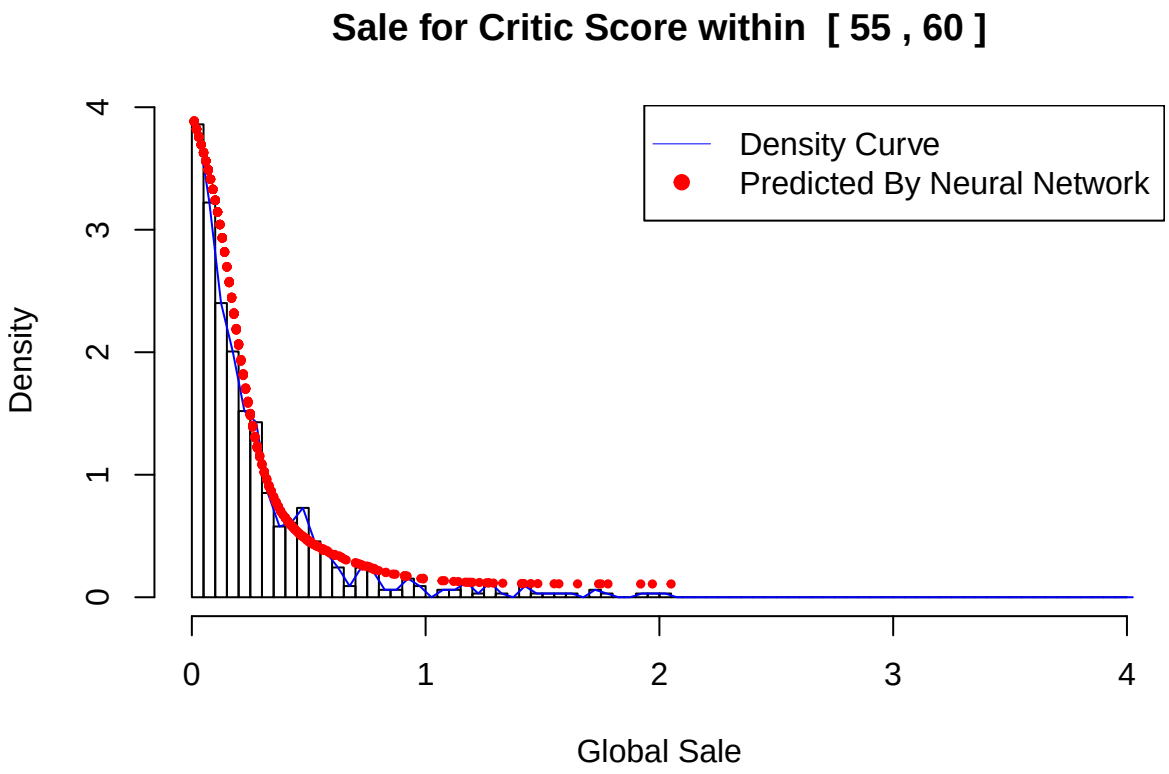

```
Pdata <- read.csv('predictedData2.csv')
for (i in seq(11)) {
  hist(Vdata$Global_Sales[ran == (32.5+5*i)], breaks=seq(0,4,by=0.05), main =
         paste("Sale for Critic Score within ", "[",as.character(32.5+5*i-2.5), ",",
        as.character(32.5+5*i+2.5),"]"), xlab="Global Sale", probability = TRUE)
  lines(seq(0,4,by=0.05)+0.025, get_density(Vdata$Global_Sales[ran == (32.5+5*i)], seq(0,4,by=0.05), 0.
  gs <- Pdata$gs[Pdata$cs_range == (32.5+5*i)]
  y <- Pdata[[6]][Pdata$cs_range == (32.5+5*i)]
  points(gs, y, col='red', pch=19,cex=0.5)
  legend("topright", c("Density Curve", "Predicted By Neural Network"), lwd = c(0.5, NA), col = c("blue
}
```

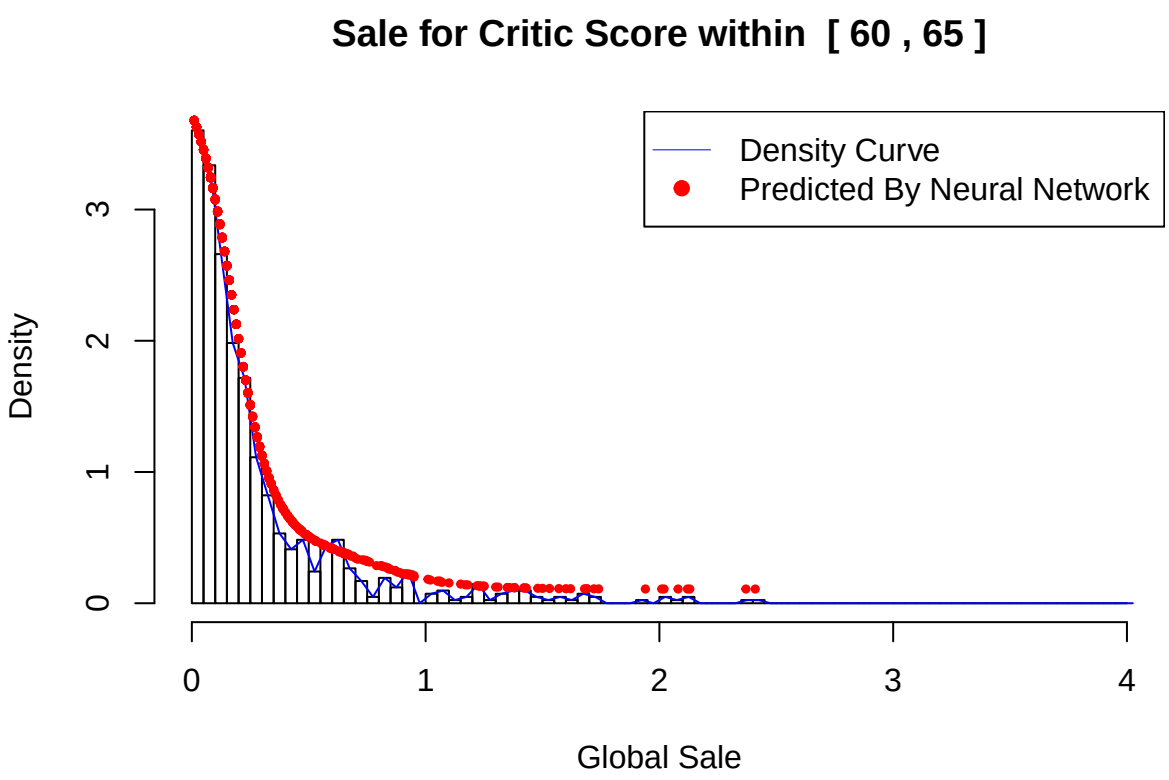

Sale for Critic Score within [35 , 40]
Density Curve
Predicted By Neural Network
Density
0
1
2
3
4
5
0
1
2
3
4
Global Sale

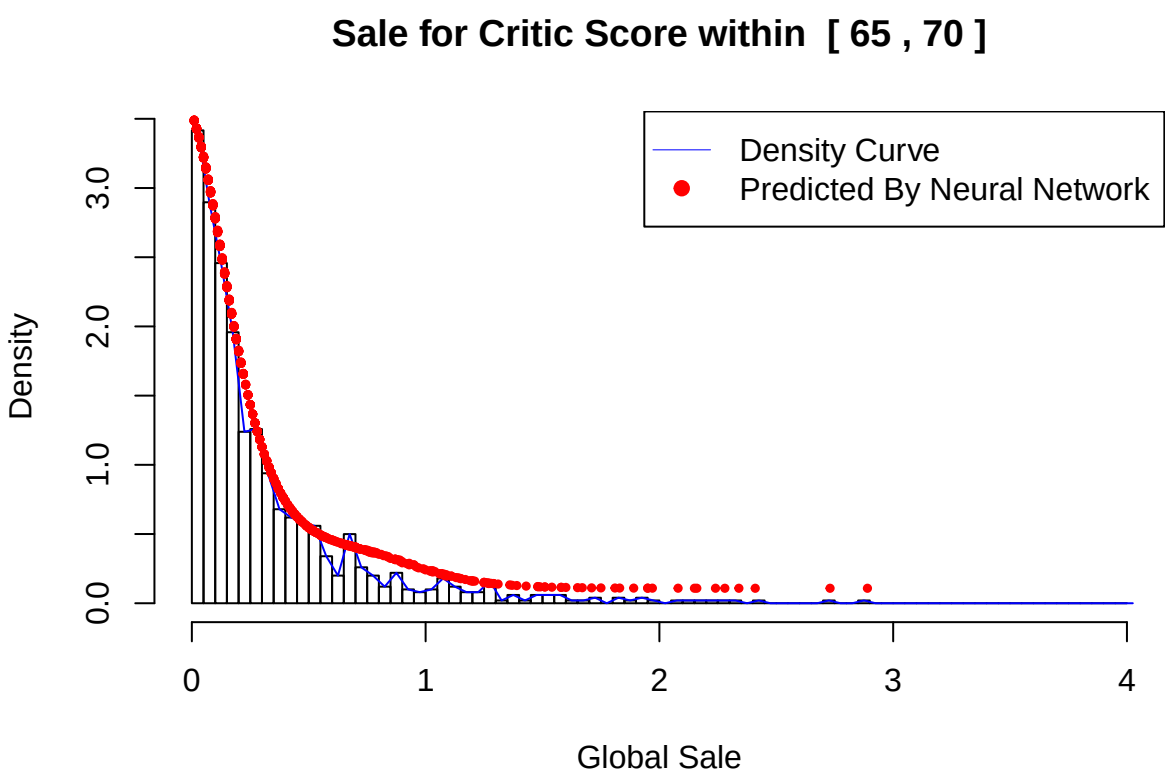

Sale for Critic Score within [40 , 45]
Density Curve
Predicted By Neural Network
Density
0
1
2
3
4
5
0
1
2
3
4
Global Sale

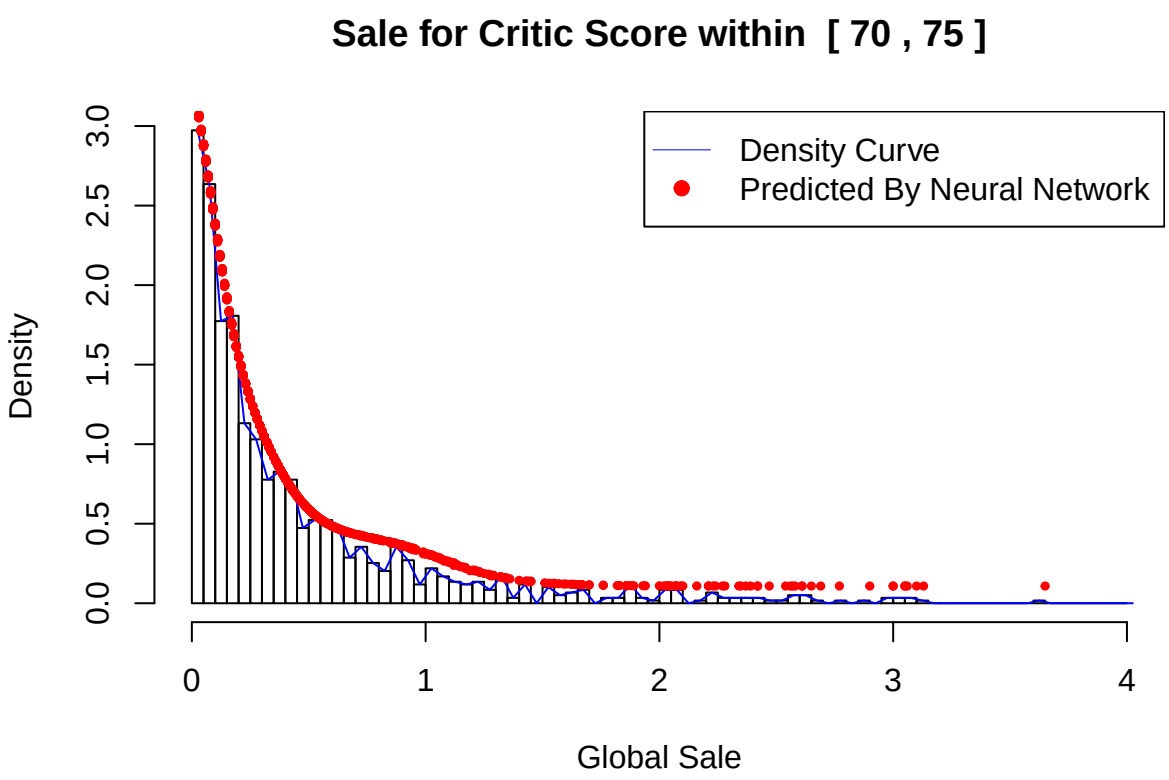

Sale for Critic Score within [45 , 50]
Density Curve
Predicted By Neural Network
Density
0
1
2
3
4
Global Sale

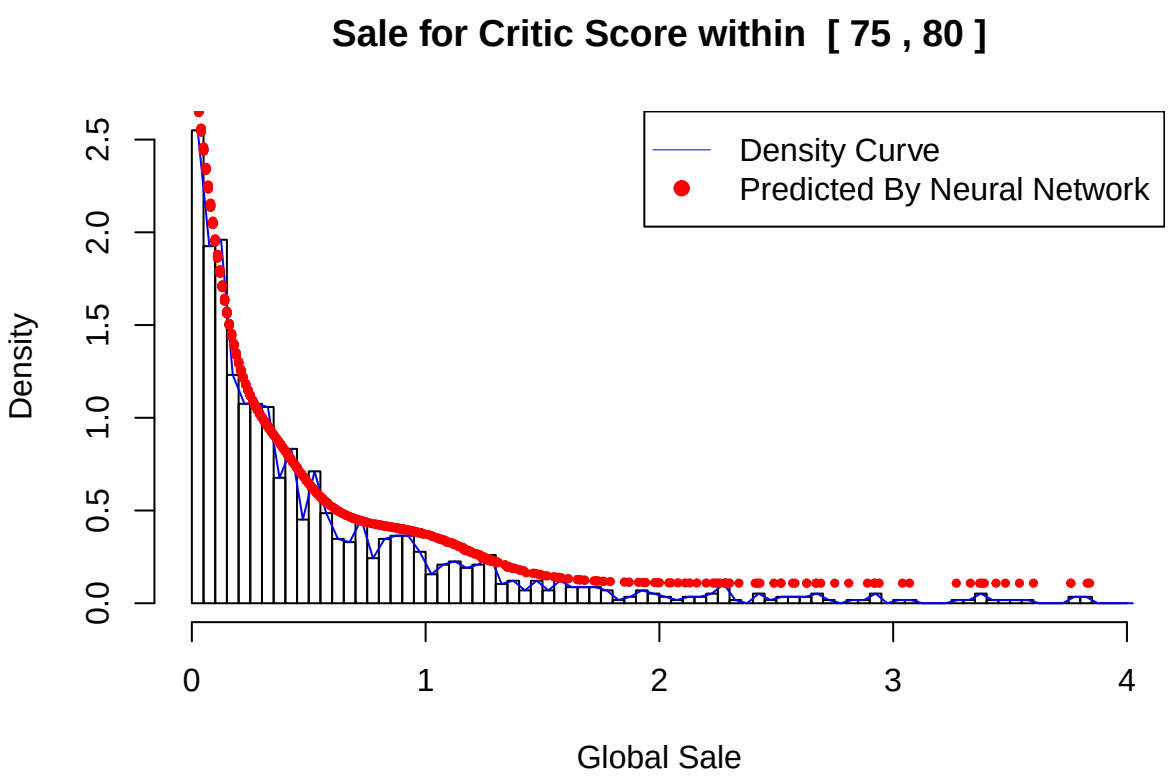

Sale for Critic Score within [50 , 55]
Density Curve
Predicted By Neural Network
Density
0
1
2
3
0
1
2
3
4
Global Sale

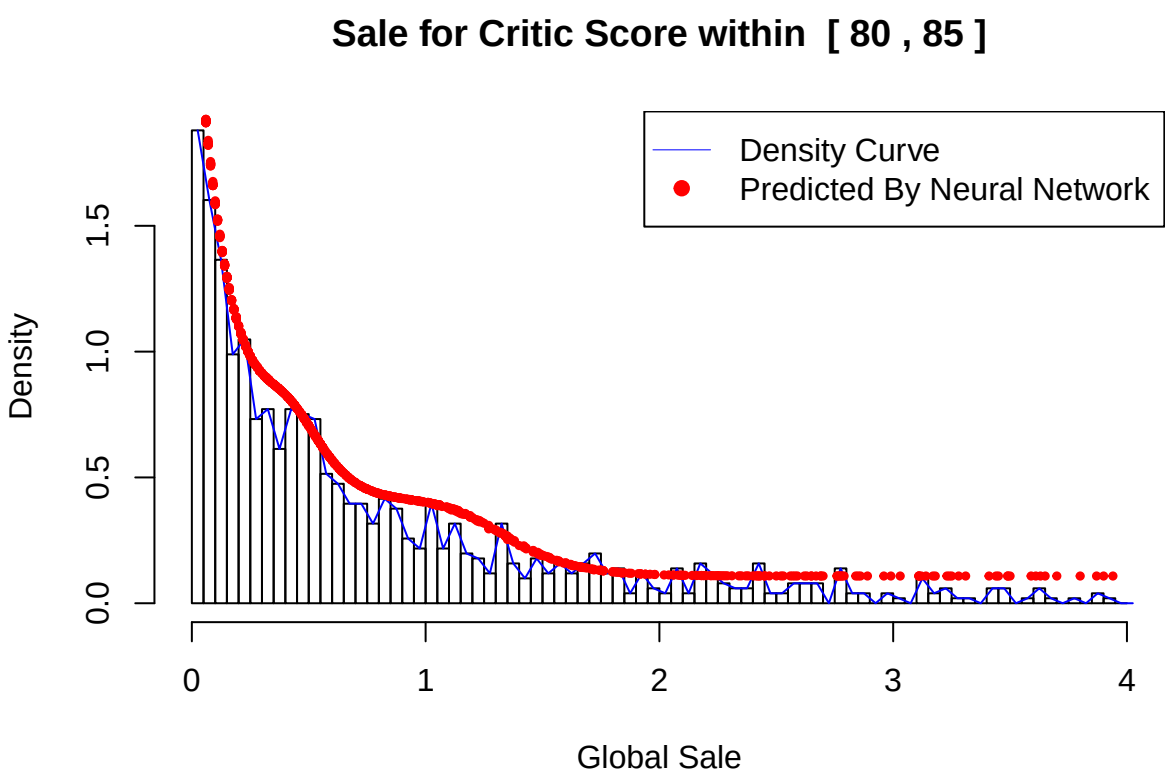

Sale for Critic Score within [55 , 60]
Density Curve
Predicted By Neural Network
Density
0
1
2
3
4
Global Sale

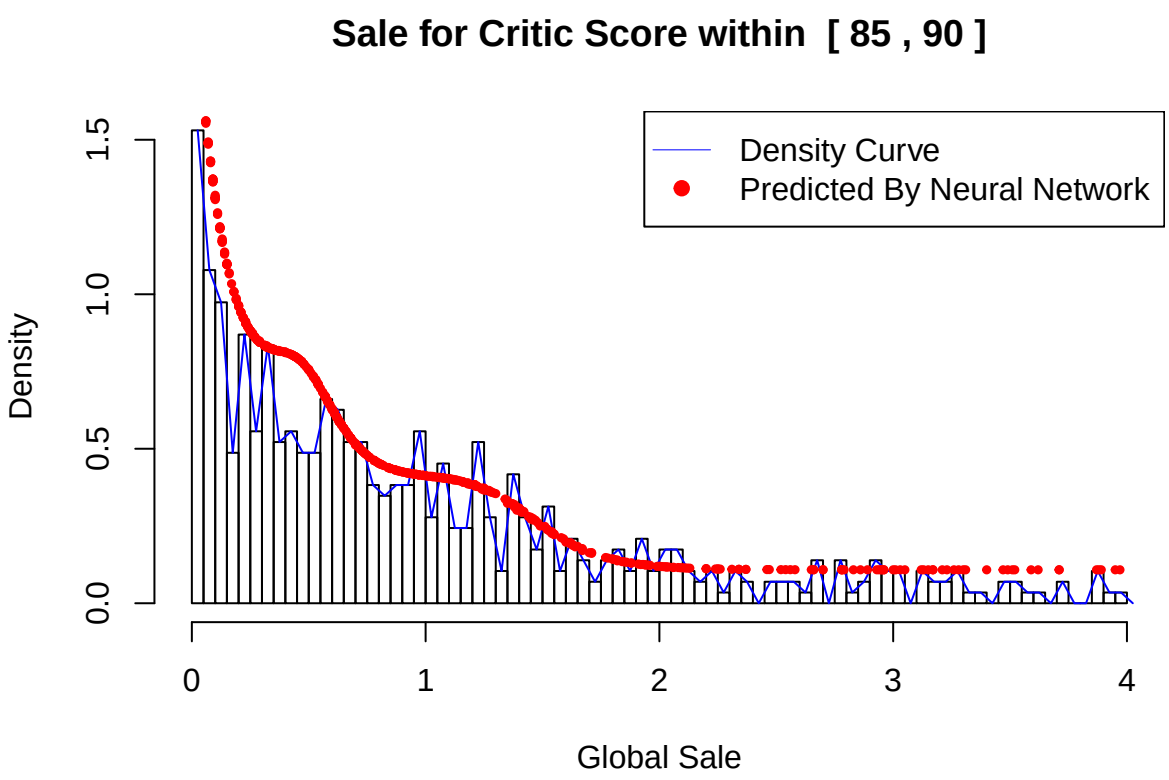

Sale for Critic Score within [60 , 65]
Density Curve
Predicted By Neural Network
Density
0
1
2
3
0
1
2
3
4
Global Sale


Sale for Critic Score within [65 , 70]
Density Curve
Predicted By Neural Network
Density
0.0
1.0
2.0
3.0
0
1
2
3
4
Global Sale


Sale for Critic Score within [70 , 75]
Density Curve
Predicted By Neural Network
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
Global Sale


Sale for Critic Score within [75 , 80]
Density Curve
Predicted By Neural Network
Density
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3
4
Global Sale


Sale for Critic Score within [80 , 85]
Density Curve
Predicted By Neural Network
Density
0.0
0.5
1.0
1.5
0
1
2
3
4
Global Sale


Sale for Critic Score within [85 , 90]
Density Curve
Predicted By Neural Network
Density
0.0
0.5
1.0
1.5
0
1
2
3
4
Global Sale