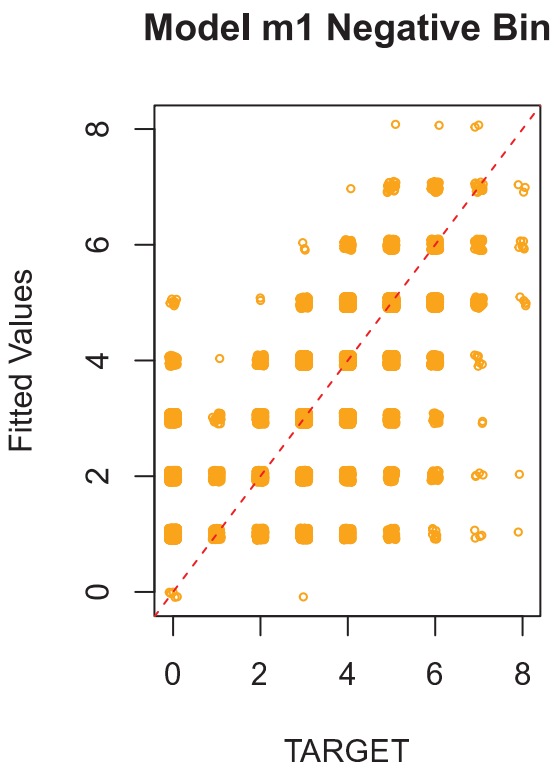
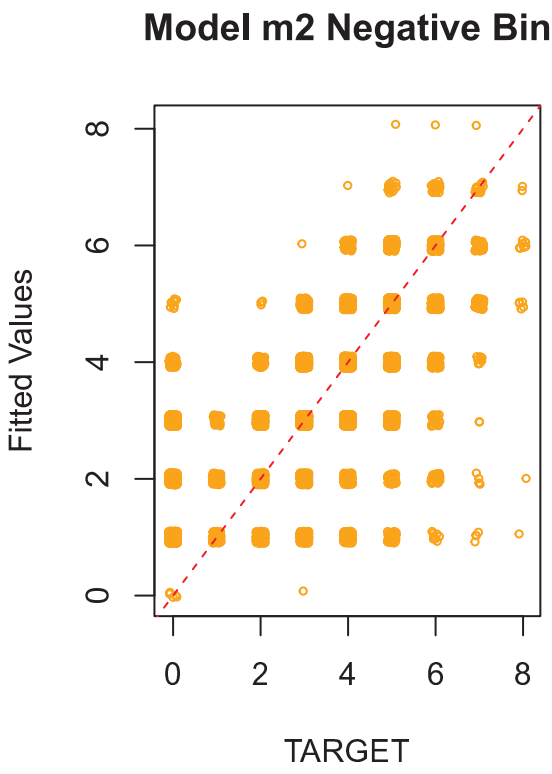


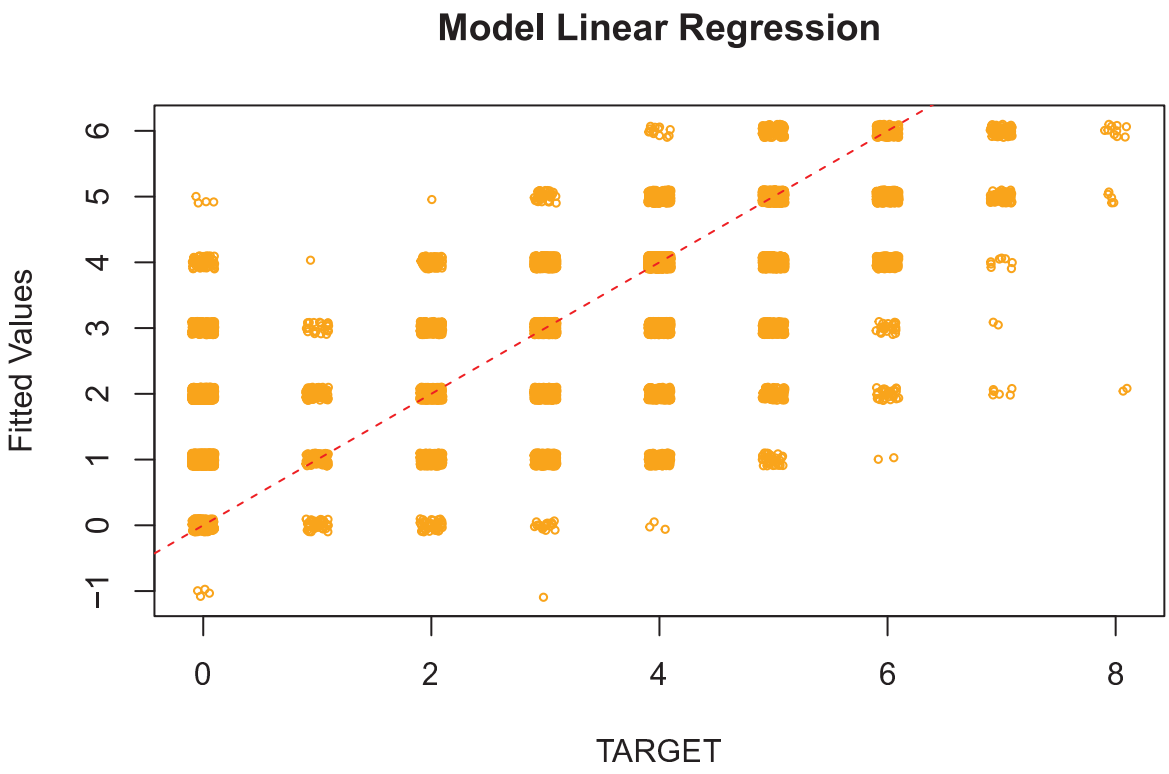




Finally, a simple linear regression model is developed as a comparison.

```
l1<-lm(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex,wine)
summary(l1)$coefficients %>% kable("latex",booktabs=T) %>% kable_styling(full_width=F)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.3309778 | 0.0983954 | 33.852971 | 0.00e+00 |
| as.factor(STARS)2 | 1.0416526 | 0.0326904 | 31.864151 | 0.00e+00 |
| as.factor(STARS)3 | 1.6111490 | 0.0377670 | 42.660175 | 0.00e+00 |
| as.factor(STARS)4 | 2.2962485 | 0.0599319 | 38.314279 | 0.00e+00 |
| as.factor(LabelAppeal)-1 | 0.3634081 | 0.0630624 | 5.762677 | 0.00e+00 |
| as.factor(LabelAppeal)0 | 0.8289754 | 0.0614908 | 13.481286 | 0.00e+00 |
| as.factor(LabelAppeal)1 | 1.2935158 | 0.0642224 | 20.141193 | 0.00e+00 |
| as.factor(LabelAppeal)2 | 1.8781253 | 0.0846105 | 22.197320 | 0.00e+00 |
| stars_I_FLAG | -1.3771624 | 0.0329978 | -41.735016 | 0.00e+00 |
| Alcohol | 0.0129347 | 0.0032863 | 3.935903 | 8.33e-05 |
| AcidIndex | -0.2043639 | 0.0089217 | -22.906338 | 0.00e+00 |

Jitter plot of the regression model shows that there are some negative value forecasts as well which is problematic given that we know the output variable to zero or strictly greater than zero.



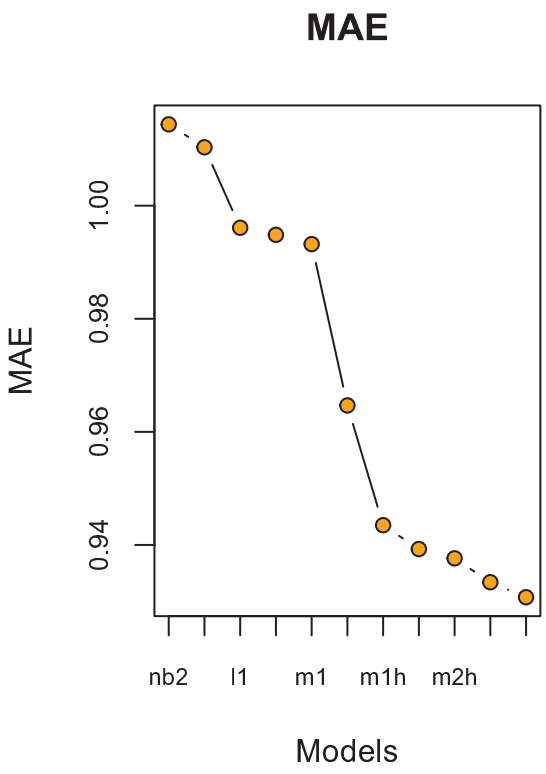

**4. Select Models (25 Points)**

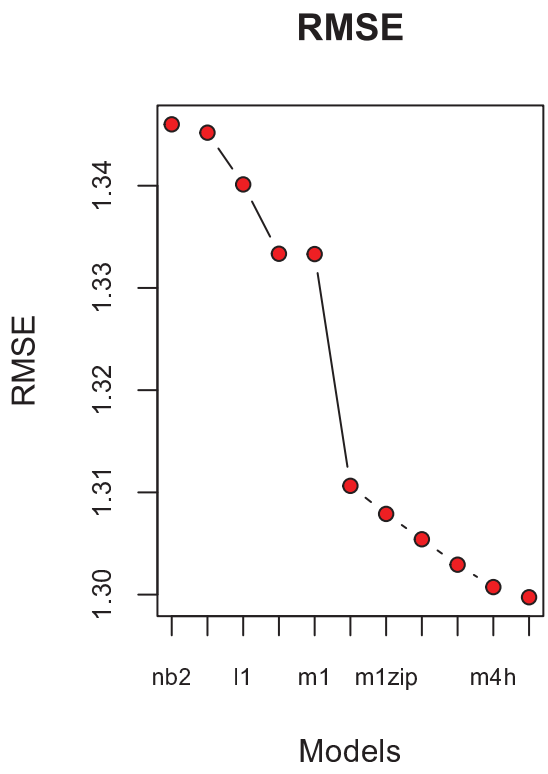
Decide on the criteria for selecting the best count regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models. For the count regression model, will you use a metric such as AIC, average squared error, etc.? Be sure to explain how you can make inferences from the model, and discuss other relevant model output. If you like the multiple linear regression model the best, please say why. However, you must select a count regression model for model deployment. Using the training data set, evaluate the performance of the count regression model. Make predictions using the evaluation data set.

**Solution**

Shown below are plots showing the Root Mean Squared Error (RMSE) and Mean Average Error (MAE) for all the models estimated in the analysis. The measures are calculated using functions in the **DESCTOOLS** package and are evaluated comparing the fitted values to the actual **TARGET** variable.

The analysis does indicate that model m2zip, m1zip and m4h are the most promising models. If we had to select a single model then it does seem that model m2zip performs the best.




**Appendix**

```
library(kableExtra)
library(knitr)
library(ggplot2)
library(gridExtra)
library(tidyr)
library(dplyr)
library(robustbase)
library(leaps)
library(DescTools)
library(nlme)
library(tinytex)
library(Hmisc)
library(reshape2)
library(pscl)
library(MASS)
```

**Question 1**

```
p1<-ggplot(wine,aes(TARGET))+geom_bar(color="orange", fill="orange", alpha=0.5)
p1
```

```
#a=data.frame()
round(summarise_all(wine[,2:16],mean,na.rm=TRUE),2) %>% kable("latex", booktabs=T) %>% kable_styling(ful
```

```
round(summarise_all(wine[,2:16],median,na.rm=TRUE),2) %>% kable("latex", booktabs=T) %>% kable_styling(
```

```
round(summarise_all(wine[,2:16],mean,na.rm=TRUE),2) %>% kable("latex", booktabs=T) %>% kable_styling(ful
round(summarise_all(wine[,2:16],median,na.rm=TRUE),2) %>% kable("latex", booktabs=T) %>% kable_styling(
```

```
#negative values
negative_vals<-as.data.frame(table(negative=wine$FixedAcidity<0, useNA="no"))
b<-as.data.frame(table(negative=wine$VolatileAcidity<0, useNA="no"))
c<-as.data.frame(table(negative=wine$CitricAcid<0, useNA="no"))
d<-as.data.frame(table(negative=wine$ResidualSugar<0, useNA="no"))
e<-as.data.frame(table(negative=wine$Chlorides<0, useNA="no"))
f<-as.data.frame(table(negative=wine$FreeSulfurDioxide<0, useNA="no"))
g<-as.data.frame(table(negative=wine$TotalSulfurDioxide<0, useNA="no"))
h<-as.data.frame(table(negative=wine$Sulphates<0, useNA="no"))
i<-as.data.frame(table(negative=wine$Alcohol<0, useNA="no"))

#merge data frames
negative_vals<-merge(negative_vals,b,by="negative")
negative_vals<-merge(negative_vals,c,by="negative")
negative_vals<-merge(negative_vals,d,by="negative")
negative_vals<-merge(negative_vals,e,by="negative")
negative_vals<-merge(negative_vals,f,by="negative")
negative_vals<-merge(negative_vals,g,by="negative")
negative_vals<-merge(negative_vals,h,by="negative")
negative_vals<-merge(negative_vals,i,by="negative")

names(negative_vals)<-c("Negative","FixedAcidity","VolatileAcidity","CitricAcid","ResidualSugar","Chlor

negative_vals %>% kable("latex", booktabs=T) %>% kable_styling(full_width=F, latex_options=c("scale_dow
round(prop.table(as.matrix(negative_vals[,2:10]),2),3) %>% kable("latex",booktabs=T) %>% kable_styling(
#delete temp tables
rm(b,c,d,e,f,g,h,i)
```

```
## put histograms on the diagonal
panel.hist <- function(x, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, ...)
}

pairs(TARGET~abs(FixedAcidity)+abs(VolatileAcidity)+pH+abs(CitricAcid)+abs(ResidualSugar)+abs(Chlorides)
```

```
pairs(jitter(TARGET,amount=0.1)~abs(FreeSulfurDioxide)+abs(TotalSulfurDioxide)+Density+abs(Sulphates)+al
```

```
pairs(jitter(TARGET,amount=0.1)~jitter(LabelAppeal,amount=0.2)+jitter(AcidIndex, amount=0.2)+jitter(STA
```

```
p2<-ggplot(wine,aes(LabelAppeal))+geom_bar(color="orange", fill="orange", alpha=0.5)
p3<-ggplot(wine,aes(AcidIndex))+geom_bar(color="orange", fill="orange", alpha=0.5)
p4<-ggplot(wine,aes(STARS))+geom_bar(color="orange", fill="orange", alpha=0.5)
grid.arrange(p2,p3,p4,nrow=1)
```

**Question 2**

```
#deal with negative values
#Fixed Acidity
wine$F_ACID_N_FLAG<-0
wine$F_ACID_N_FLAG[wine$FixedAcidity<0]<-1
wine<-wine %>% mutate(FixedAcidity=abs(FixedAcidity))
#Volatile Acidity
wine$V_ACID_N_FLAG<-0
wine$V_ACID_N_FLAG[wine$VolatileAcidity<0]<-1
wine<-wine %>% mutate(VolatileAcidity=abs(VolatileAcidity))
#Citric Acid
wine$C_ACID_N_FLAG<-0
wine$C_ACID_N_FLAG[wine$CitricAcid<0]<-1
wine<-wine %>% mutate(CitricAcid=abs(CitricAcid))
#Residual Sugar
wine$Rsugar_N_FLAG<-0
wine$Rsugar_N_FLAG[wine$ResidualSugar<0]<-1
wine<-wine %>% mutate(ResidualSugar=abs(ResidualSugar))
#Chlorides
wine$Chl_N_FLAG<-0
wine$Chl_N_FLAG[wine$Chlorides<0]<-1
wine<-wine %>% mutate(Chlorides=abs(Chlorides))
#Freesulphur
wine$free_sox_N_FLAG<-0
wine$free_sox_N_FLAG[wine$FreeSulfurDioxide<0]<-1
wine<-wine %>% mutate(FreeSulfurDioxide=abs(FreeSulfurDioxide))
#Totalsulphur
wine$tot_sox_N_FLAG<-0
wine$tot_sox_N_FLAG[wine$TotalSulfurDioxide<0]<-1
wine<-wine %>% mutate(TotalSulfurDioxide=abs(TotalSulfurDioxide))
#Sulphates
wine$sulphates_N_FLAG<-0
wine$sulphates_N_FLAG[wine$Sulphates<0]<-1
wine<-wine %>% mutate(Sulphates=abs(Sulphates))
#Alcohol
wine$alcohol_N_FLAG<-0
wine$alcohol_N_FLAG[wine$Alcohol<0]<-1
wine<-wine %>% mutate(Alcohol=abs(Alcohol))
```

```
#impute missing variables
#impute ResdiualSugar
wine$ResidualSugar_I_FLAG<-0
wine$ResidualSugar_I_FLAG[is.na(wine$ResidualSugar)]<-1
wine$ResidualSugar[is.na(wine$ResidualSugar)]<-median(wine$ResidualSugar,na.omit(TRUE))
#Chlorides
wine$Chlorides_I_FLAG<-0
wine$Chlorides_I_FLAG[is.na(wine$Chlorides)]<-1
wine$Chlorides[is.na(wine$Chlorides)]<-median(wine$Chlorides,na.omit(TRUE))
#Free SulfurDioxide
wine$free_sox_I_FLAG<-0
wine$free_sox_I_FLAG[is.na(wine$FreeSulfurDioxide)]<-1
wine$FreeSulfurDioxide[is.na(wine$FreeSulfurDioxide)]<-median(wine$FreeSulfurDioxide,na.omit(TRUE))
#Total SulfurDioxide
wine$tot_sox_I_FLAG<-0
```

```
wine$tot_sox_I_FLAG[is.na(wine$TotalSulfurDioxide)]<-1
wine$TotalSulfurDioxide[is.na(wine$TotalSulfurDioxide)]<-median(wine$TotalSulfurDioxide,na.omit(TRUE))
#pH
wine$ph_I_FLAG<-0
wine$ph_I_FLAG[is.na(wine$pH)]<-1
wine$pH[is.na(wine$pH)]<-median(wine$pH,na.omit(TRUE))
#Sulphates
wine$sulphates_I_FLAG<-0
wine$sulphates_I_FLAG[is.na(wine$Sulphates)]<-1
wine$Sulphates[is.na(wine$Sulphates)]<-median(wine$Sulphates,na.omit(TRUE))
#Alcohol
wine$alcohol_I_FLAG<-0
wine$alcohol_I_FLAG[is.na(wine$Alcohol)]<-1
wine$Alcohol[is.na(wine$Alcohol)]<-median(wine$Alcohol,na.omit(TRUE))
```

```
#plot of Stars
par(mfrow=c(1,2))
hist(wine$TARGET[!is.na(wine$STARS)],breaks=20,freq=TRUE, col="blue",xlab="TARGET", main="STARS Not Miss
hist(wine$TARGET[is.na(wine$STARS)],breaks=20,freq=TRUE, col="blue",xlab="TARGET", main="STARS Missing"
par(mfrow=c(1,1))
```

```
#STARS
wine$stars_I_FLAG<-0
wine$stars_I_FLAG[is.na(wine$STARS)]<-1
#median TARGET for NA / Missing stars
median(wine$TARGET[is.na(wine$STARS)])
#Impute median for missing stars
wine$STARS[is.na(wine$STARS)]<-median(wine$STARS[wine$TARGET==0],na.omit(TRUE))
```

**Question 3**

```
# Get upper triangle of the correlation matrix
get_upper_tri <- function(cormat){
  cormat[lower.tri(cormat)]<- NA
  return(cormat)
}

#get correlations
res2<-rcorr(as.matrix(wine[,2:33]))
#extract upper triangle of correlations
res2_corr<-get_upper_tri(res2$r)
melted_res2 <- melt(res2_corr, na.rm = TRUE)


#correlation plot
ggplot(data = melted_res2, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation")+theme(axis.text.x = element_text(angle = 90,vjust =
```

```
sort(round(res2$r[,1],4))%>% kable("latex",booktabs=T) %>% kable_styling(full_width=F) %>% add_header_al
```

```
#Develop a simple Poisson Model
m1<-glm(TARGET~STARS+LabelAppeal+stars_I_FLAG+Alcohol+AcidIndex,family=poisson,wine)
```

```
m2<-glm(TARGET~STARS+LabelAppeal+stars_I_FLAG+Alcohol+AcidIndex+VolatileAcidity+TotalSulfurDioxide,famil
#summary of results
summary(m1)$coefficients%>% kable("latex",booktabs=T) %>% kable_styling(full_width=F)
summary(m2)$coefficients%>% kable("latex",booktabs=T) %>% kable_styling(full_width=F)
```

```
#comparison of fitted plots
par(mfrow=c(1,2))
plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(m1)),amount=0.1),cex=0.5,col="blue",xlab="TARGET
abline(0,1,lty=2,col="red")
plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(m2),0),amount=0.1),cex=0.5,col="blue",xlab="TARG
abline(0,1,lty=2,col="red")
#Histograms of low values
hist(fitted(m1)[fitted(m1)<2],breaks=20,col="orange",xlab="Fitted Values",main="Model m1")
hist(fitted(m2)[fitted(m2)<2],breaks=20,col="orange",xlab="Fitted Values",main="Model m2")
```

```
#develop hurdle rate models
m1h<-hurdle(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex,dist = c("pois

m2h<-hurdle(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex+VolatileAcidi

m3h<-hurdle(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex+VolatileAcidi

m4h<-hurdle(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex+VolatileAcidi

m5h<-hurdle(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex+VolatileAcidi
```

```
hist(fitted(m1h)[fitted(m1h)<2],breaks=20,col="orange",xlab="Fitted Values",main="Model m1 Hurdle")
hist(fitted(m2h)[fitted(m2h)<2],breaks=20,col="orange",xlab="Fitted Values",main="Model m2 Hurdle")

plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(m1h),0),amount=0.1),cex=0.5,col="blue",xlab="TAR
abline(0,1,lty=2,col="red")
plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(m2h),0),amount=0.1),cex=0.5,col="blue",xlab="TAR
abline(0,1,lty=2,col="red")
plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(m3h),0),amount=0.1),cex=0.5,col="blue",xlab="TAR
abline(0,1,lty=2,col="red")
plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(m4h),0),amount=0.1),cex=0.5,col="blue",xlab="TAR
abline(0,1,lty=2,col="red")
```

```
#develop ZIP models
m1zip<-zeroinfl(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex,dist = c(
m2zip<-zeroinfl(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex+VolatileA

hist(fitted(m1zip)[fitted(m1zip)<2],breaks=20,col="orange",xlab="Fitted Values",main="Model m1 Zip")
hist(fitted(m2zip)[fitted(m2zip)<2],breaks=20,col="orange",xlab="Fitted Values",main="Model m2 Hurdle")

plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(m1zip),0),amount=0.1),cex=0.5,col="orange",xlab=
abline(0,1,lty=2,col="red")
plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(m2zip),0),amount=0.1),cex=0.5,col="orange",xlab=
abline(0,1,lty=2,col="red")
```

```
#negative binomial models
nb1<-glm(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex, negative.binomia
nb2<-glm(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex+VolatileAcidity+
par(mfrow=c(1,2))
plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(nb1),0),amount=0.1),cex=0.5,col="orange",xlab="
```

```
abline(0,1,lty=2,col="red")
plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(nb2),0),amount=0.1),cex=0.5,col="orange",xlab="
abline(0,1,lty=2,col="red")
```

```
l1<-lm(TARGET~as.factor(STARS)+as.factor(LabelAppeal)+stars_I_FLAG+Alcohol+AcidIndex,wine)
summary(l1)$coefficients %>% kable("latex",booktabs=T) %>% kable_styling(full_width=F)
plot(jitter(wine$TARGET,amount=0.1),jitter(round(fitted(l1),0),amount=0.1),cex=0.5,col="orange",xlab="T
abline(0,1,lty=2,col="red")
```

### Question 4

```
par(mfrow=c(2,2))
plot(m1h$fitted.values,m1h$residuals,cex=0.5, xlab="Fitted Values", ylab="Residuals", col="blue", main=
plot(m2h$fitted.values,m2h$residuals,cex=0.5, xlab="Fitted Values", ylab="Residuals", col="blue", main=
plot(m3h$fitted.values,m3h$residuals,cex=0.5, xlab="Fitted Values", ylab="Residuals", col="blue", main=
plot(m4h$fitted.values,m4h$residuals,cex=0.5, xlab="Fitted Values", ylab="Residuals", col="blue", main=
```

```
par(mfrow=c(2,2))
plot(nb1$fitted.values,nb1$residuals,cex=0.5, xlab="Fitted Values", ylab="Residuals", col="blue", main=
plot(nb2$fitted.values,nb2$residuals,cex=0.5, xlab="Fitted Values", ylab="Residuals", col="blue", main=
plot(m1$fitted.values,m1$residuals,cex=0.5, xlab="Fitted Values", ylab="Residuals", col="blue", main="M
plot(m2$fitted.values,m2$residuals,cex=0.5, xlab="Fitted Values", ylab="Residuals", col="blue", main="M
```

```
par(mfrow=c(2,2))
plot(l1$fitted.values,l1$residuals,cex=0.5, xlab="Fitted Values", ylab="Residuals", col="blue", main="M
plot(m1zip$fitted.values,m1zip$residuals,cex=0.5, xlab="Fitted Values", ylab="Residuals", col="blue", ma
plot(m2zip$fitted.values,m2zip$residuals,cex=0.5, xlab="Fitted Values", ylab="Residuals", col="blue", ma
```

```
#Mean Average Error
mae_table<-data.frame("m1h",MAE(round(fitted(m1h)),wine$TARGET))
names(mae_table)<-c("Model","MAE")
b<-data.frame("m2h",MAE(round(fitted(m2h)),wine$TARGET))
names(b)<-c("Model","MAE")
c<-data.frame("m3h",MAE(round(fitted(m3h)),wine$TARGET))
names(c)<-c("Model","MAE")
d<-data.frame("m4h",MAE(round(fitted(m4h)),wine$TARGET))
names(d)<-c("Model","MAE")
f<-data.frame("m1",MAE(round(fitted(m1)),wine$TARGET))
names(f)<-c("Model","MAE")
g<-data.frame("m2",MAE(round(fitted(m2)),wine$TARGET))
names(g)<-c("Model","MAE")
h<-data.frame("m1zip",MAE(round(fitted(m1zip)),wine$TARGET))
names(h)<-c("Model","MAE")
i<-data.frame("m2zip",MAE(round(fitted(m2zip)),wine$TARGET))
names(i)<-c("Model","MAE")
j<-data.frame("l1",MAE(round(fitted(l1)),wine$TARGET))
names(j)<-c("Model","MAE")
k<-data.frame("nb1",MAE(round(fitted(nb1)),wine$TARGET))
names(k)<-c("Model","MAE")
l<-data.frame("nb2",MAE(round(fitted(nb2)),wine$TARGET))
names(l)<-c("Model","MAE")


mae_table<-rbind(mae_table,b,c,d,f,g,h,i,j,k,l)
mae_table<-arrange(mae_table,desc(MAE))
```

```
#RMSE table
rmse_table<-data.frame("m1h",RMSE(round(fitted(m1h)),wine$TARGET))
names(rmse_table)<-c("Model","RMSE")
b<-data.frame("m2h",RMSE(round(fitted(m2h)),wine$TARGET))
names(b)<-c("Model","RMSE")
c<-data.frame("m3h",RMSE(round(fitted(m3h)),wine$TARGET))
names(c)<-c("Model","RMSE")
d<-data.frame("m4h",RMSE(round(fitted(m4h)),wine$TARGET))
names(d)<-c("Model","RMSE")
f<-data.frame("m1",RMSE(round(fitted(m1)),wine$TARGET))
names(f)<-c("Model","RMSE")
g<-data.frame("m2",RMSE(round(fitted(m2)),wine$TARGET))
names(g)<-c("Model","RMSE")
h<-data.frame("m1zip",RMSE(round(fitted(m1zip)),wine$TARGET))
names(h)<-c("Model","RMSE")
i<-data.frame("m2zip",RMSE(round(fitted(m2zip)),wine$TARGET))
names(i)<-c("Model","RMSE")
j<-data.frame("l1",RMSE(round(fitted(l1)),wine$TARGET))
names(j)<-c("Model","RMSE")
k<-data.frame("nb1",RMSE(round(fitted(nb1)),wine$TARGET))
names(k)<-c("Model","RMSE")
l<-data.frame("nb2",RMSE(round(fitted(nb2)),wine$TARGET))
names(l)<-c("Model","RMSE")

rmse_table<-rbind(rmse_table,b,c,d,f,g,h,i,j,k,l)
rmse_table<-arrange(rmse_table,desc(RMSE))
```

```
par(mfrow=c(1,2))
plot(mae_table$MAE,xaxt="n",xlab="Models",type='b',ylab="MAE",pch=21,bg="orange",main="MAE",cex.axis=0.8
axis(1,at=seq(1,11,1),labels=mae_table$Model,cex.axis=0.75)

plot(rmse_table$RMSE,xaxt="n",xlab="Models",type='b',ylab="RMSE",pch=21,bg="red",main="RMSE",cex.axis=0
axis(1,at=seq(1,11,1),labels=rmse_table$Model,cex.axis=0.75)
```