CUNY DATA 621 HW 5

```
library( ggplot2)
library(reshape2)
library(corrplot)
library (MASS)
library(Deducer)
library(ROCR)
library(dplyr)
```

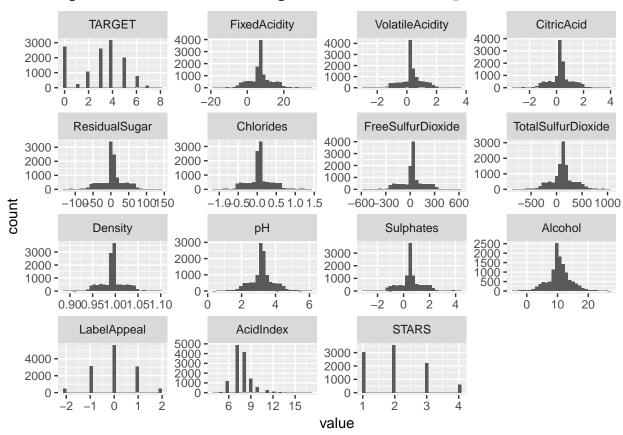
Basic Data Exploration

```
wine_df <- read.csv('wine-training-data.csv')</pre>
wine_df <- wine_df[,-c(1)]</pre>
summary(wine_df)
##
        TARGET
                     FixedAcidity
                                      VolatileAcidity
                                                           CitricAcid
##
   Min.
           :0.000
                    Min.
                         :-18.100
                                      Min.
                                              :-2.7900
                                                         Min.
                                                                :-3.2400
   1st Qu.:2.000
                    1st Qu.: 5.200
                                      1st Qu.: 0.1300
                                                         1st Qu.: 0.0300
   Median :3.000
                    Median: 6.900
                                      Median : 0.2800
                                                         Median: 0.3100
           :3.029
   Mean
                    Mean
                          : 7.076
                                      Mean
                                             : 0.3241
                                                         Mean
                                                               : 0.3084
   3rd Qu.:4.000
                    3rd Qu.: 9.500
##
                                      3rd Qu.: 0.6400
                                                         3rd Qu.: 0.5800
##
   Max.
           :8.000
                          : 34.400
                                              : 3.6800
                                                         Max.
                                                              : 3.8600
                    Max.
##
                                         FreeSulfurDioxide TotalSulfurDioxide
##
   ResidualSugar
                         Chlorides
##
   Min.
          :-127.800
                       Min.
                              :-1.1710
                                         Min.
                                                 :-555.00
                                                            Min.
                                                                   :-823.0
##
   1st Qu.: -2.000
                       1st Qu.:-0.0310
                                         1st Qu.:
                                                    0.00
                                                            1st Qu.: 27.0
   Median :
              3.900
                       Median : 0.0460
                                         Median :
                                                   30.00
                                                            Median: 123.0
                              : 0.0548
##
   Mean
         : 5.419
                       Mean
                                         Mean
                                                :
                                                   30.85
                                                            Mean : 120.7
   3rd Qu.: 15.900
                       3rd Qu.: 0.1530
                                         3rd Qu.: 70.00
                                                            3rd Qu.: 208.0
##
          : 141.150
                                                 : 623.00
##
                              : 1.3510
                                                                   :1057.0
   Max.
                       Max.
                                         Max.
                                                            Max.
                                         NA's
   NA's
           :616
                       NA's
                              :638
                                                 :647
                                                            NA's
                                                                   :682
##
                                       Sulphates
                                                           Alcohol
       Density
                           Нq
##
   Min.
           :0.8881
                     Min.
                            :0.480
                                     Min.
                                           :-3.1300
                                                        Min.
                                                               :-4.70
   1st Qu.:0.9877
                     1st Qu.:2.960
                                     1st Qu.: 0.2800
                                                        1st Qu.: 9.00
   Median :0.9945
                     Median :3.200
                                     Median : 0.5000
                                                        Median :10.40
   Mean
          :0.9942
                           :3.208
                                     Mean
                                           : 0.5271
                                                               :10.49
##
                     Mean
                                                        Mean
##
   3rd Qu.:1.0005
                     3rd Qu.:3.470
                                     3rd Qu.: 0.8600
                                                        3rd Qu.:12.40
          :1.0992
##
   Max.
                     Max.
                            :6.130
                                     Max.
                                            : 4.2400
                                                        Max.
                                                               :26.50
##
                     NA's
                            :395
                                     NA's
                                             :1210
                                                        NA's
                                                               :653
##
    LabelAppeal
                          AcidIndex
                                             STARS
                        Min.
                               : 4.000
##
   Min.
           :-2.000000
                                         Min.
                                                 :1.000
   1st Qu.:-1.000000
                        1st Qu.: 7.000
                                         1st Qu.:1.000
##
   Median : 0.000000
                        Median : 8.000
                                         Median :2.000
   Mean :-0.009066
                        Mean
                               : 7.773
                                         Mean
                                                 :2.042
##
   3rd Qu.: 1.000000
                        3rd Qu.: 8.000
                                         3rd Qu.:3.000
   Max. : 2.000000
                               :17.000
                                                 :4.000
                        Max.
                                         Max.
##
                                         NA's
                                                 :3359
```

```
wine_melt_df <- melt(wine_df[,sapply(wine_df, is.numeric)])
## No id variables; using all as measure variables
ggplot(wine_melt_df,aes(x = value)) +
   facet_wrap(~variable,scales = "free") +
   geom_histogram()</pre>
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

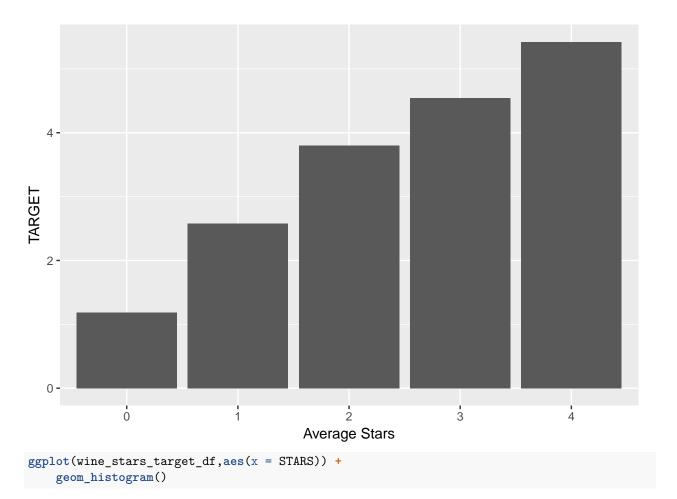
Warning: Removed 8200 rows containing non-finite values (stat_bin).



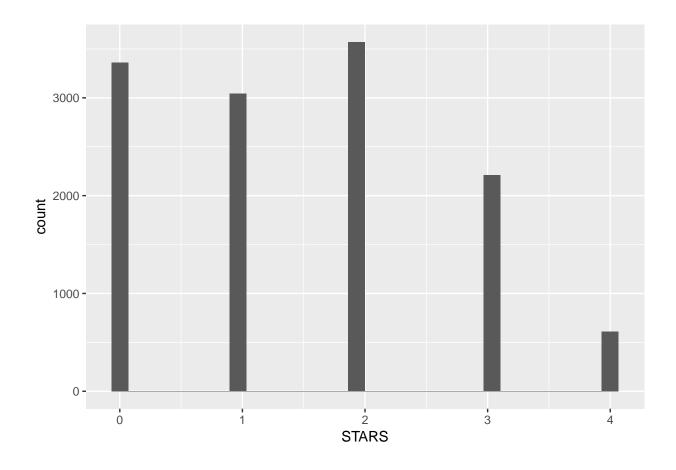
Explore relationship of stars

The min value of stars is 1 therefore it seems logical that missing data in stars means zero stars.

```
wine_stars_target_df <- wine_df[,c("TARGET", "STARS")]
wine_stars_target_df[is.na(wine_stars_target_df[,"STARS"]), "STARS"] <-0
ggplot(wine_stars_target_df, aes(x=factor(STARS), y=TARGET)) + stat_summary(fun.y="mean", geom="bar") +</pre>
```



`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



Data imputations

For this dataset we will want to treat stars as if it is a categorical variable, with all missing values treated as 0. The rest of the missing data will be replaced with the median of the value.

```
str(wine_df)
  'data.frame':
                    12795 obs. of 15 variables:
##
   $ TARGET
                        : int
                               3 3 5 3 4 0 0 4 3 6 ...
##
   $ FixedAcidity
                              3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
                        : num
   $ VolatileAcidity
                               1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
                        : num
                               -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
   $ CitricAcid
##
                        : num
##
   $ ResidualSugar
                        : num 54.2 26.1 14.8 18.8 9.4 ...
   $ Chlorides
                              -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##
                        : num
   $ FreeSulfurDioxide : num NA 15 214 22 -167 -37 287 523 -213 62 ...
##
   $ TotalSulfurDioxide: num
                               268 -327 142 115 108 15 156 551 NA 180 ...
##
   $ Density
                               0.993 1.028 0.995 0.996 0.995 ...
                        : num
   $ pH
##
                        : num
                               3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##
   $ Sulphates
                               -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
                        : num
                               9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##
   $ Alcohol
                        : num
##
   $ LabelAppeal
                        : int
                               0 -1 -1 -1 0 0 0 1 0 0 ...
                               8 7 8 6 9 11 8 7 6 8 ...
##
   $ AcidIndex
                        : int
                               2 3 3 1 2 NA NA 3 NA 4 ...
   $ STARS
##
                        : int
wine_df[is.na(wine_df[,"STARS"]), "STARS"] <-0</pre>
```

```
wine_df <- wine_df %>%
    mutate_all(~ifelse(is.na(.), median(., na.rm = TRUE), .))
wine_df$STARS<- as.factor(wine_df$STARS)</pre>
```

Transforms

All data is relatively normal. No transforms needed.

Build Models

Possion

Build a possion model with all variables

```
model_posion1 <- glm(TARGET ~ . , data=wine_df, poisson )
summary(model_posion1)</pre>
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = wine_df)
## Deviance Residuals:
      Min
                10
                    Median
                                  30
                                          Max
## -3.2777 -0.6623 -0.0025
                              0.4490
                                       3.7623
##
## Coefficients:
##
                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)
                      1.105e+00 1.959e-01
                                             5.642 1.68e-08 ***
## FixedAcidity
                      6.117e-05 8.197e-04
                                           0.075 0.940515
## VolatileAcidity
                     -3.045e-02 6.527e-03 -4.666 3.08e-06 ***
## CitricAcid
                      4.963e-03 5.897e-03
                                           0.842 0.399964
## ResidualSugar
                      5.321e-05
                                1.547e-04
                                            0.344 0.730836
## Chlorides
                     -3.853e-02 1.647e-02 -2.340 0.019304 *
## FreeSulfurDioxide
                    9.253e-05 3.504e-05
                                             2.640 0.008280 **
## TotalSulfurDioxide 7.909e-05 2.273e-05
                                             3.479 0.000503 ***
## Density
                     -2.699e-01 1.918e-01
                                           -1.407 0.159315
## pH
                     -1.203e-02 7.651e-03 -1.573 0.115718
                     -1.256e-02 5.752e-03 -2.184 0.028981 *
## Sulphates
                      3.758e-03 1.409e-03
                                             2.668 0.007637 **
## Alcohol
## LabelAppeal
                      1.594e-01 6.126e-03 26.012 < 2e-16 ***
## AcidIndex
                     -7.982e-02 4.574e-03 -17.451 < 2e-16 ***
## STARS1
                      7.668e-01 1.954e-02 39.242 < 2e-16 ***
                      1.089e+00 1.822e-02
                                           59.789 < 2e-16 ***
## STARS2
## STARS3
                      1.208e+00 1.920e-02 62.909 < 2e-16 ***
## STARS4
                      1.323e+00 2.431e-02 54.403 < 2e-16 ***
## Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
## Null deviance: 22861 on 12794 degrees of freedom
## Residual deviance: 13652 on 12777 degrees of freedom
## AIC: 45630
##
## Number of Fisher Scoring iterations: 6
```

Build Negaive bionomial Model

```
model_neg_binomial <- glm(TARGET ~ . , data=wine_df, quasipoisson )</pre>
summary(model neg binomial)
##
## glm(formula = TARGET ~ ., family = quasipoisson, data = wine_df)
##
## Deviance Residuals:
      Min
                10
                     Median
                                  3Q
                                          Max
## -3.2777 -0.6623 -0.0025
                              0.4490
                                       3.7623
## Coefficients:
##
                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)
                      1.105e+00 1.842e-01
                                          6.001 2.01e-09 ***
## FixedAcidity
                     6.117e-05 7.707e-04
                                           0.079 0.936735
## VolatileAcidity
                     -3.045e-02 6.136e-03 -4.963 7.04e-07 ***
## CitricAcid
                     4.963e-03 5.544e-03
                                           0.895 0.370645
## ResidualSugar
                      5.321e-05 1.454e-04
                                           0.366 0.714427
## Chlorides
                     -3.853e-02 1.548e-02 -2.489 0.012836 *
## FreeSulfurDioxide 9.253e-05 3.294e-05 2.809 0.004983 **
## TotalSulfurDioxide 7.909e-05 2.137e-05 3.701 0.000216 ***
                     -2.699e-01 1.803e-01 -1.497 0.134412
## Density
## pH
                     -1.203e-02 7.193e-03 -1.673 0.094313 .
## Sulphates
                     -1.256e-02 5.407e-03 -2.323 0.020204 *
## Alcohol
                      3.758e-03 1.324e-03
                                            2.838 0.004552 **
## LabelAppeal
                      1.594e-01 5.760e-03 27.669 < 2e-16 ***
                     -7.982e-02 4.300e-03 -18.563 < 2e-16 ***
## AcidIndex
## STARS1
                      7.668e-01 1.837e-02 41.742 < 2e-16 ***
## STARS2
                      1.089e+00 1.712e-02
                                           63.597 < 2e-16 ***
## STARS3
                      1.208e+00 1.805e-02 66.916 < 2e-16 ***
## STARS4
                      1.323e+00 2.286e-02 57.868 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for quasipoisson family taken to be 0.8838223)
##
      Null deviance: 22861 on 12794 degrees of freedom
## Residual deviance: 13652 on 12777 degrees of freedom
## Number of Fisher Scoring iterations: 6
```

compare models

Becasue of the fact that I transformed the STARS variable into a categorical variable the model is not overdispersed, therefore the possion and negative binomials are the same.

Generate predications

```
wine_eval_df <- read.csv("wine-evaluation-data.csv")</pre>
wine_eval_df[is.na(wine_eval_df[,"STARS"]), "STARS"] <-0</pre>
wine_df <- wine_df %>%
   mutate_all(~ifelse(is.na(.), median(., na.rm = TRUE), .))
wine_eval_df$STARS <- as.factor(wine_eval_df$STARS)</pre>
str(wine_df)
  'data.frame':
                     12795 obs. of 15 variables:
                                 3 3 5 3 4 0 0 4 3 6 ...
##
    $ TARGET
                          : int
##
    $ FixedAcidity
                                3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
    $ VolatileAcidity
                                 1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
                          : num
    $ CitricAcid
                                 -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
                          : num
    $ ResidualSugar
##
                          : num 54.2 26.1 14.8 18.8 9.4 ...
                                 -0.567 -0.425 0.037 -0.425 0.046 0.556 0.06 0.04 -0.007 -0.277 ...
##
   $ Chlorides
                          : num
   $ FreeSulfurDioxide : num
                                 30 15 214 22 -167 -37 287 523 -213 62 ...
##
    $ TotalSulfurDioxide: num
                                 268 -327 142 115 108 15 156 551 123 180 ...
##
    $ Density
                          : num 0.993 1.028 0.995 0.996 0.995 ...
                                 3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##
    $ pH
                          : num
                                 -0.59 0.7 0.48 1.83 1.77 1.29 1.21 0.5 0.26 0.75 ...
##
    $ Sulphates
                          : num
                                 9.9 10.4 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##
    $ Alcohol
                          : num
##
    $ LabelAppeal
                          : int
                                 0 -1 -1 -1 0 0 0 1 0 0 ...
    $ AcidIndex
                          : int
                                 8 7 8 6 9 11 8 7 6 8 ...
    $ STARS
                                 3 4 4 2 3 1 1 4 1 5 ...
##
                          : int
round(predict(model_posion1, wine_eval_df))
##
      1
           2
                 3
                      4
                            5
                                 6
                                       7
                                            8
                                                  9
                                                      10
                                                                 12
                                                                       13
                                                                            14
                                                                                  15
                                                            11
##
      0
           1
                      1
                            0
                                 2
                                       1
                                            0
                                                  0
                                                       0
                                                            NA
                                                                  0
                                                                        1
                                                                            NA
                                                                                  0
##
     16
          17
                           20
                                21
                                      22
                                           23
                                                 24
                                                      25
                                                            26
                                                                 27
                                                                       28
                                                                            29
                                                                                 30
                18
                     19
##
     NA
           1
                 0
                     NA
                            2
                                 1
                                       0
                                                  1
                                                       1
                                                             2
                                                                  1
                                                                        2
                                                                             2
                                                                                 NA
     31
          32
                     34
                                36
                                      37
                                                                 42
                                                                                 45
##
                33
                           35
                                           38
                                                 39
                                                      40
                                                            41
                                                                       43
                                                                            44
##
     NA
           0
                 2
                      1
                            0
                                 1
                                       1
                                            1
                                                  0
                                                       0
                                                             1
                                                                 NA
                                                                        0
                                                                             0
                                                                                 NA
          47
##
     46
                48
                     49
                           50
                                51
                                      52
                                           53
                                                 54
                                                      55
                                                            56
                                                                 57
                                                                       58
                                                                            59
                                                                                  60
##
      1
           1
                 2
                      1
                            1
                                 1
                                      NA
                                            1
                                                  0
                                                      NA
                                                            NA
                                                                 NA
                                                                        1
                                                                             0
                                                                                  0
                                                            71
                                                                                  75
##
          62
                63
                     64
                                66
                                      67
                                                 69
                                                      70
                                                                 72
                                                                       73
                                                                            74
     61
                           65
                                           68
##
     NA
          NA
                NA
                      1
                           NA
                                 1
                                      NA
                                            2
                                                  0
                                                      NA
                                                            NA
                                                                 NA
                                                                        1
                                                                             2
                                                                                  1
     76
          77
                78
                     79
                           80
                                      82
                                                      85
                                                            86
                                                                 87
                                                                            89
                                                                                 90
##
                                81
                                           83
                                                 84
                                                                       88
##
      1
           0
                NA
                      2
                           NA
                                 1
                                      NA
                                            1
                                                  1
                                                       1
                                                            NA
                                                                  1
                                                                        2
                                                                            NA
                                                                                   2
##
     91
          92
                93
                     94
                           95
                                96
                                      97
                                           98
                                                 99
                                                     100
                                                           101
                                                                102
                                                                      103
                                                                           104
                                                                                 105
                 0
                                                       0
##
      1
           0
                      1
                            1
                                NA
                                       1
                                            1
                                                  1
                                                            NA
                                                                 NA
                                                                        1
                                                                             1
                                                                                 NA
##
    106
         107
               108
                    109
                          110
                               111
                                     112
                                          113
                                                114
                                                     115
                                                           116
                                                                117
                                                                      118
                                                                           119
                                                                                 120
##
      0
          NA
                NA
                       2
                            0
                                NA
                                      NA
                                                 NA
                                                            NA
                                                                            NA
                                                                                   0
                                            1
                                                       0
                                                                  1
                                                                      NΑ
```