Homework 5 - DATA 621

Business Analytics & Data Mining

Ohannes (Hovig) Ohannessian
ohannes.ohannessian16@spsmail.cuny.edu

7/12/2018

Contents

Introduction

In business world, who doesn't want to predict its sales? Or have competitive ways to expand its sales?

This analysis will develop models to predict the number of cases of wine samples to provide understanding to wine manufacturer how to maximize wine sales.

If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

1. Overview

In this homework assignment, you will explore, analyze and model a dataset containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

Your objective is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. You can only use the variables given to you (or variables that you derive from the variables provided).

Below is a short description of the variables of interest in the data set:

VARIABLE	NAME DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET	Number of Cases Purchased	None
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average	
Alcohol	Alcohol Content	
Chlorides	Chloride content of wine	
CitricAcid	Citric Acid Content	
Density	Density of Wine	
FixedAcidity	Fixed Acidity of Wine	
FreeSulfurDioxide	Sulfur Dioxide content of wine	
LabelAppeal	Marketing Score indicating the appeal of label design for consumers. Higher numbers suggest customers like the label design. Negative numbers suggest customes don't like the design.	Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales.
ResidualSugar	Residual Sugar of wine	
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor	A high number of stars suggests high sales
Sulphates	Sulfate conten of wine	
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	
VolatileAcidity	Volatile Acid content of wine	
pH	pH of wine	

2. Data Exploration

Describe the size and the variables in the wine training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

The data set contains 12,795 cases (no pun intended), with an indentification variable (`INDEX`), 14 predictors, and one response variable. Each case is a commerically available wine, with the response variable being the number of cases purchased by restaurants and wine shops after sampling the wine. Of the 14 predictor variables, 12 are related to chemical preoperties of the wine, while the other two have to do with a rating and label design.

A summary of each variable is presented below:

Table 1: Table continues below

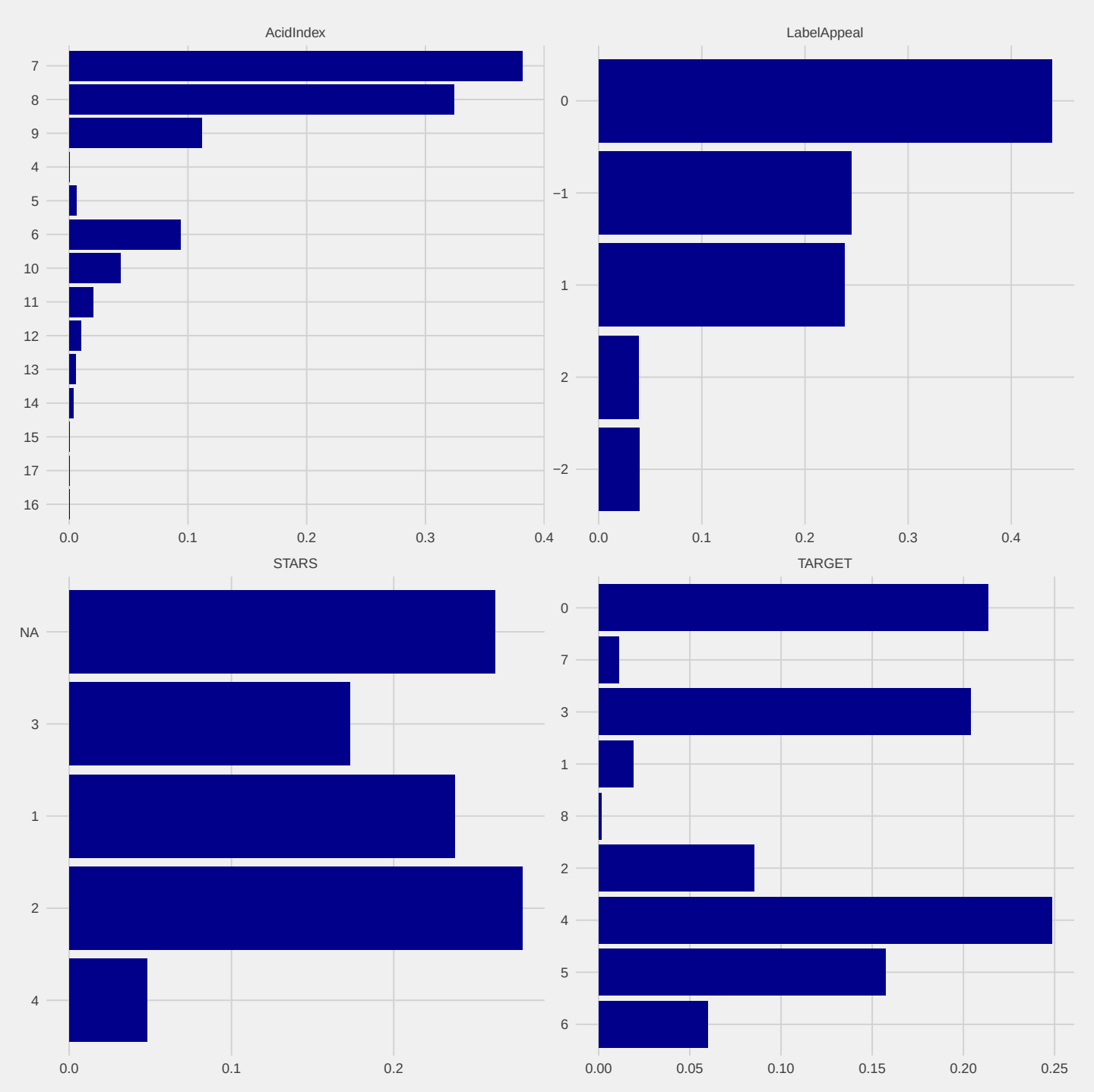
	MEAN	MIN	MEDIAN	MAX	IQR	STD. DEV
TARGET	3.03	0	3	8	2	1.93
FixedAcidity	7.08	-18.1	6.9	34.4	4.3	6.32
VolatileAcidity	0.32	-2.79	0.28	3.68	0.51	0.78
CitricAcid	0.31	-3.24	0.31	3.86	0.55	0.86
ResidualSugar	5.42	-127.8	3.9	141.2	17.9	33.75
Chlorides	0.05	-1.17	0.05	1.35	0.18	0.32
FreeSulfurDioxide	30.85	-555	30	623	70	148.7
TotalSulfurDioxide	120.7	-823	123	1057	181	231.9
Density	0.99	0.89	0.99	1.1	0.01	0.03
pH	3.21	0.48	3.2	6.13	0.51	0.68
Sulphates	0.53	-3.13	0.5	4.24	0.58	0.93
Alcohol	10.49	-4.7	10.4	26.5	3.4	3.73
LabelAppeal	-0.01	-2	0	2	2	0.89
AcidIndex	7.77	4	8	17	1	1.32
STARS	2.04	1	2	4	2	0.9

	SKEW	r_{TARGET}	NAs
TARGET	-0.33	1	0
FixedAcidity	-0.02	-0.01	0
VolatileAcidity	0.02	-0.08	0
CitricAcid	-0.05	0	0
ResidualSugar	-0.05	0	616
Chlorides	0.03	-0.03	638
FreeSulfurDioxide	0.01	0.02	647
TotalSulfurDioxide	-0.01	0.02	682
Density	-0.02	-0.05	0
pH	0.04	0	395
Sulphates	0.01	-0.02	1210
Alcohol	-0.03	0.07	653
LabelAppeal	0.01	0.5	0
AcidIndex	1.65	-0.17	0
STARS	0.45	0.55	3359

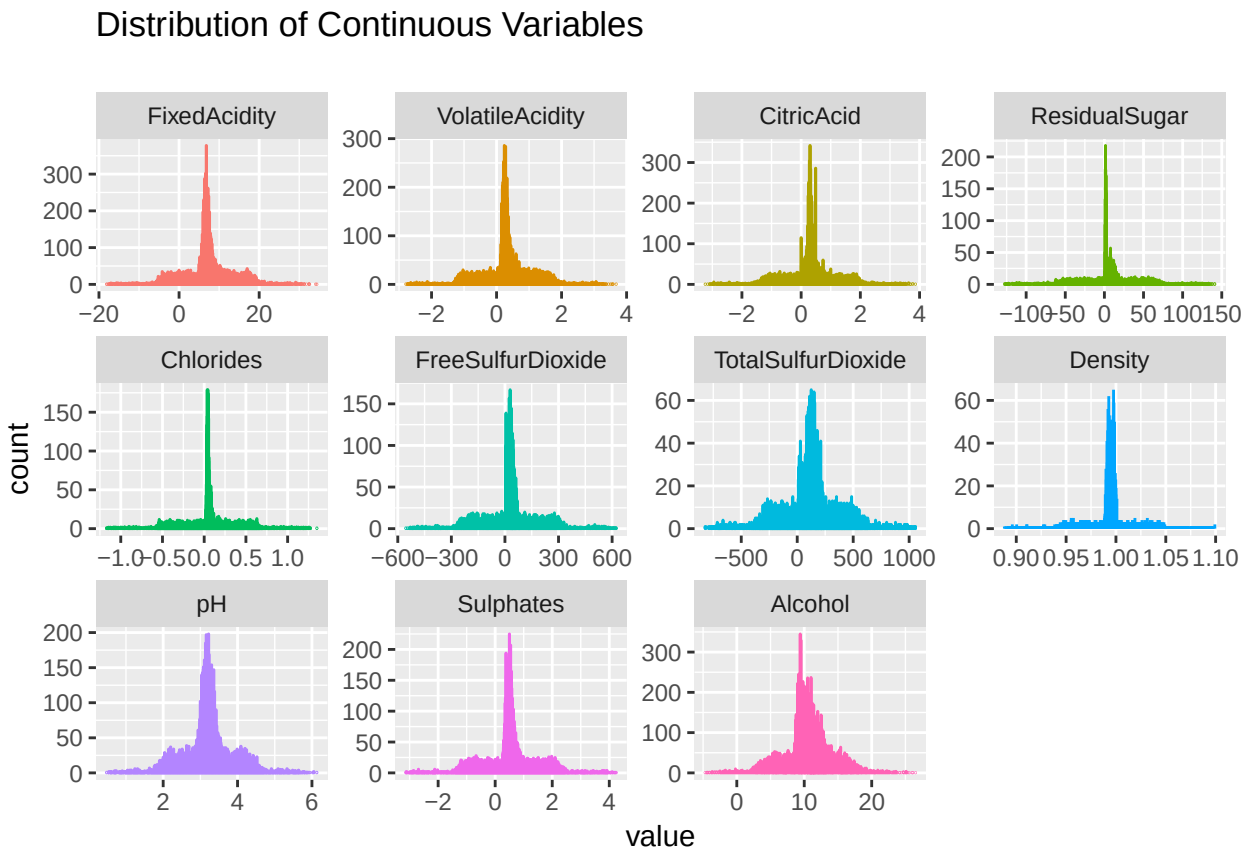
There are eight variables with missing values, with the proportion of values missing ranging from slightly over 3% to roughly 9.5%; these missing values will need to either be imputed or excluded from the dataset before modeling.

Discrete Variable Frequencies

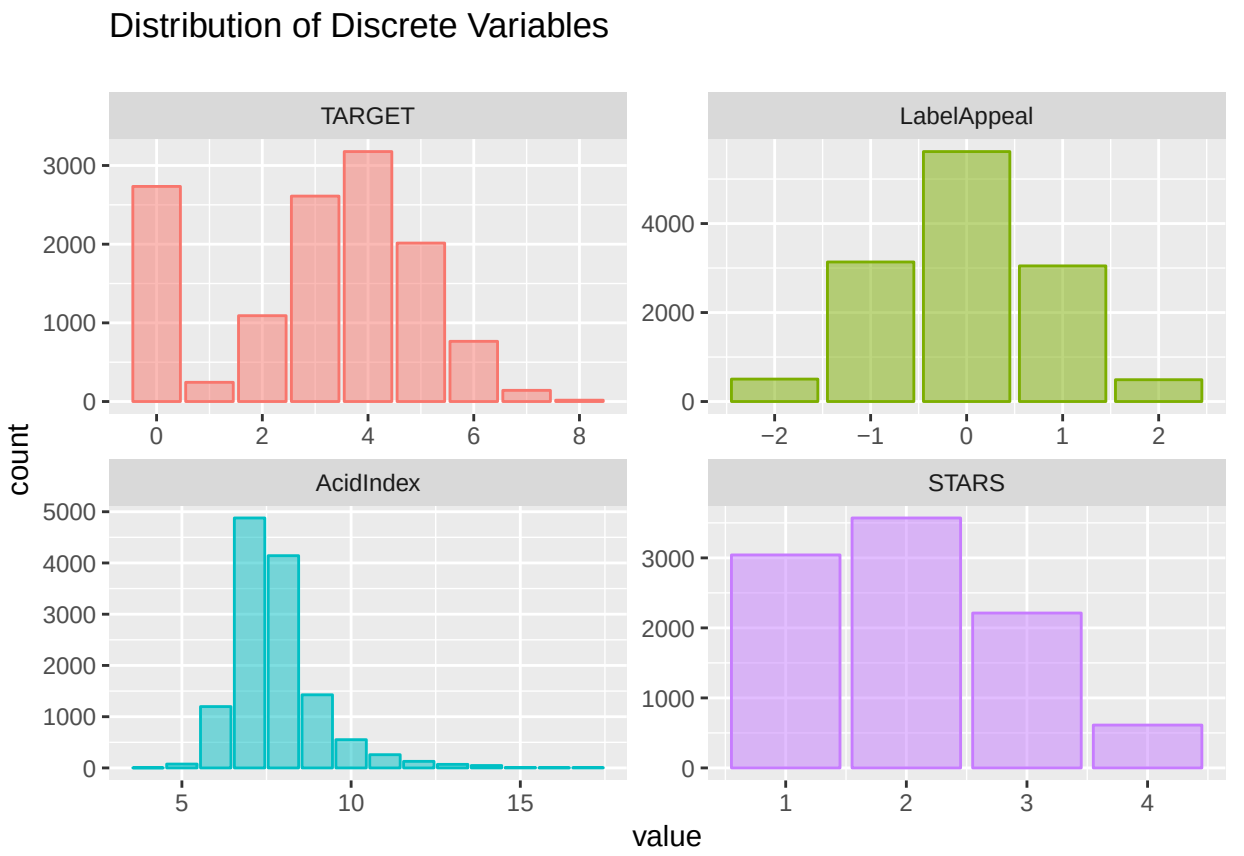
In the barplots below, we notice that more than 20% of the `TARGET` values are zero, which indicates that the data may be a good candidate for a zero-inflated Poisson regression model.



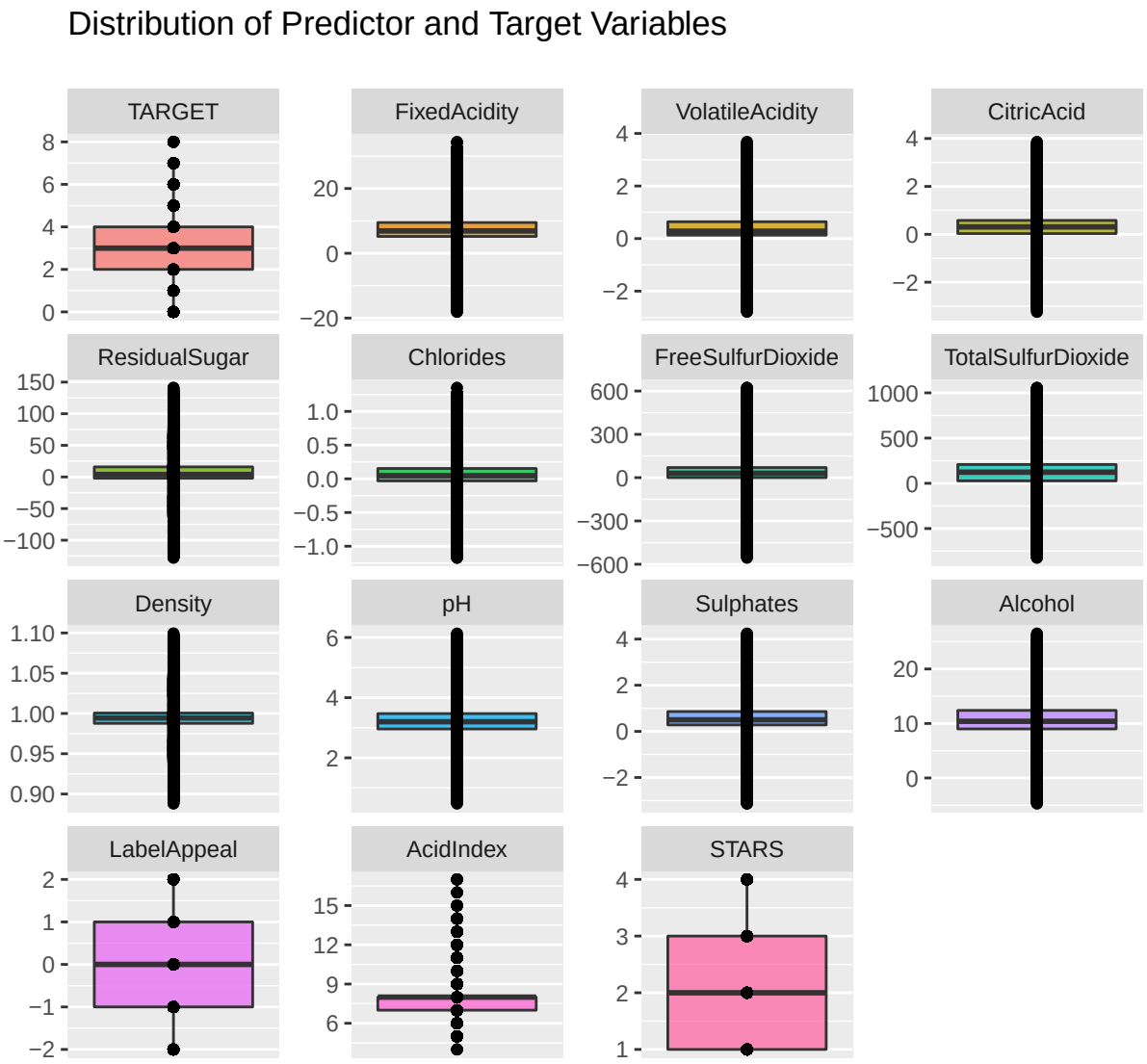
With the exception of the predictor variables `LabelAppeal`, `AcidIndex`, `STARS`, and our response variable, the remainder of the variables are continuous, and appear to have a fairly normal distribution with a small spread; many of the values are centered around the mean. Due to the size of the dataset, observations outside of 3 sd from the mean do exist. Comparing the means and median, there is very little skew in all of these predictors.



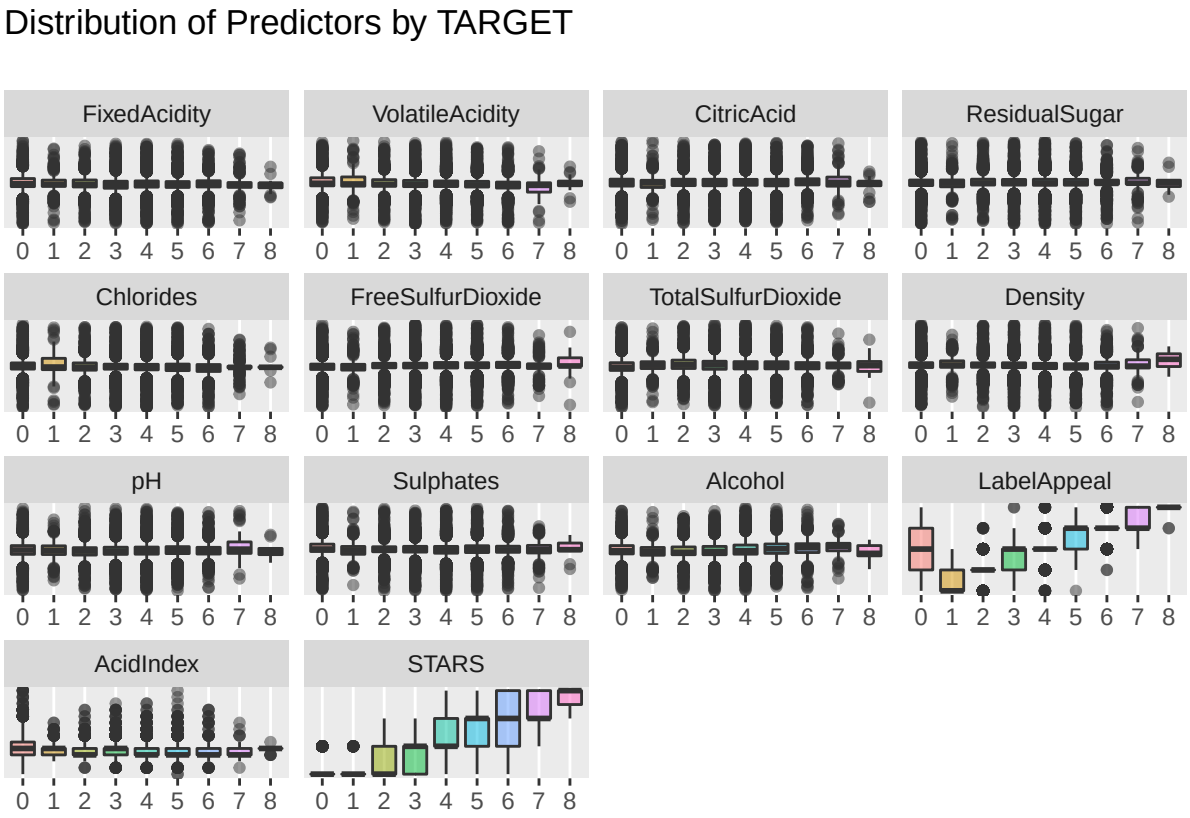
The other variables are discrete, taking only whole number values, and are therefore binomial distributions.



We will revisit the `STARS` variable in the next section, as over 1/4 of the cases contain NAs. It is suspected that these values could be equal to a 0 rating, rather than a missing value.



Generating boxplots and dividing them up by the response variable, the three variables that show some type of correlation to `TARGET`, `LabelAppeal`, `AcidIndex` and `STARS`, become more apparent. Clearly the effect of bottle aesthetics and ratings of the wine by experts seems to have a greater effect on the decision to purchase the wine or not than any of the chemical properties. The larget majority of wine purchased in higher case numbers (> 4 cases) have higher label likability (`LabelAppeal` is not negative).

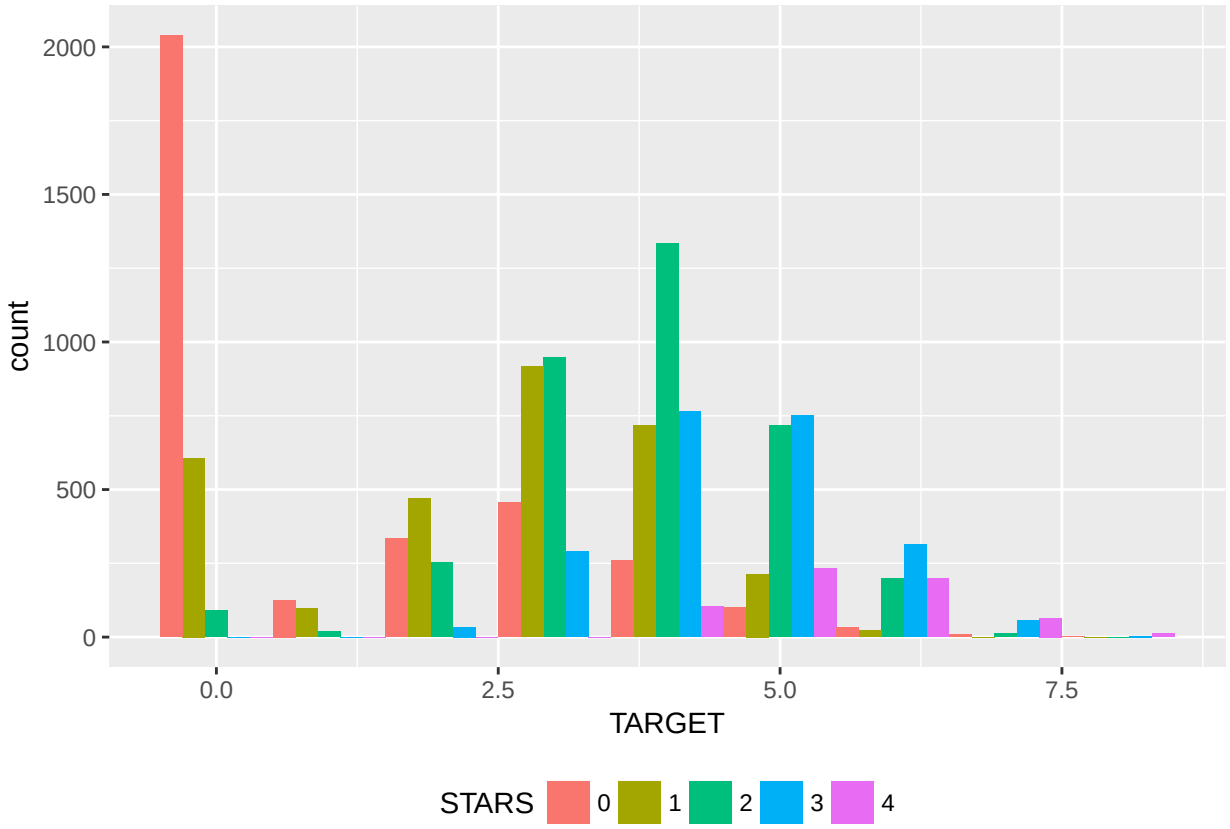


3. Data Preparation

```
Describe how you have transformed the data by changing the original variables or creating new variables. If yo
transform the data or create new variables, discuss why you did this. Here are some possible transformations.
a. Fix missing values (maybe with a Mean or Median value)
b. Create flags to suggest if a variable was missing
c. Transform data by putting it into buckets
d. Mathematical transforms such as log or square root (or use Box-Cox)
e. Combine variables (such as ratios or adding or multiplying) to create new variables
```

Before attempting to combine or transform any of our variables, the predictor variables with missing values must be addressed. Eight of the predictor variables have missing values, with almost a third of our cases containing an NA. As concluded from our examination of the data in the previous section, the `STARS` variable has over 3000 NA values, which are most likely associated with a 0 rating. Using this reasoning, we will replace the NA values in `STARS` with zeros.

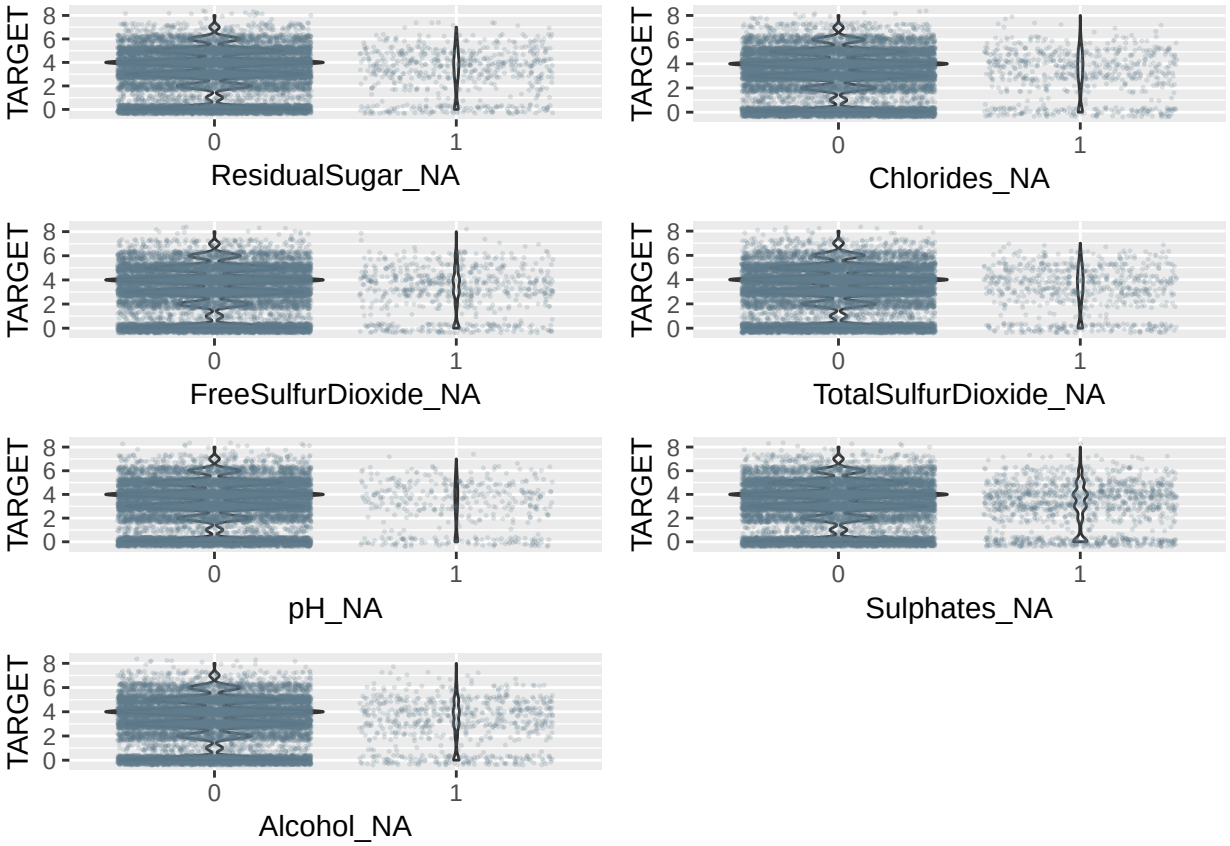
The majority of the zero-star ratings we have used to replace the NAs have been placed into cases where our `TARGET` variable is equal to zero. This will increase the correlation between the predictor and the response, and `STARS` will remain the variable with the highest correlation with `TARGET`.



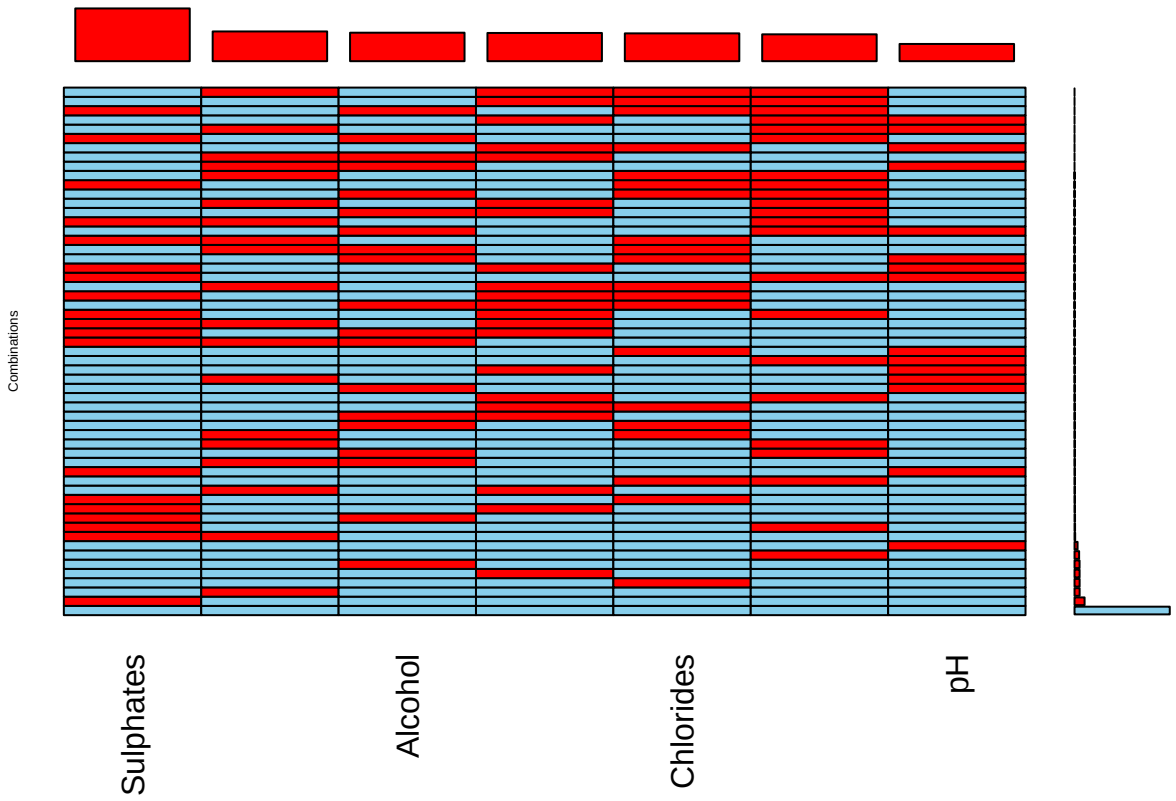
The remaining variables with NAs, along with the count and proportion of the total cases is found in the table below:

Var Name	No. NAs	% of Cases
ResidualSugar	616	0.048
Chlorides	638	0.05
FreeSulfurDioxide	647	0.051
TotalSulfurDioxide	682	0.053
pH	395	0.031
Sulphates	1210	0.095
Alcohol	653	0.051

With the exception of `Sulphates`, the missing values are approximately 5% or less of the each of the above variables. After replacing NAs in the `STARS` variable, nearly one-third of our dataset contains a row with an NA. The plots below show the distribution of the NA values for each of the remaining predictors, which appear to be missing at random.

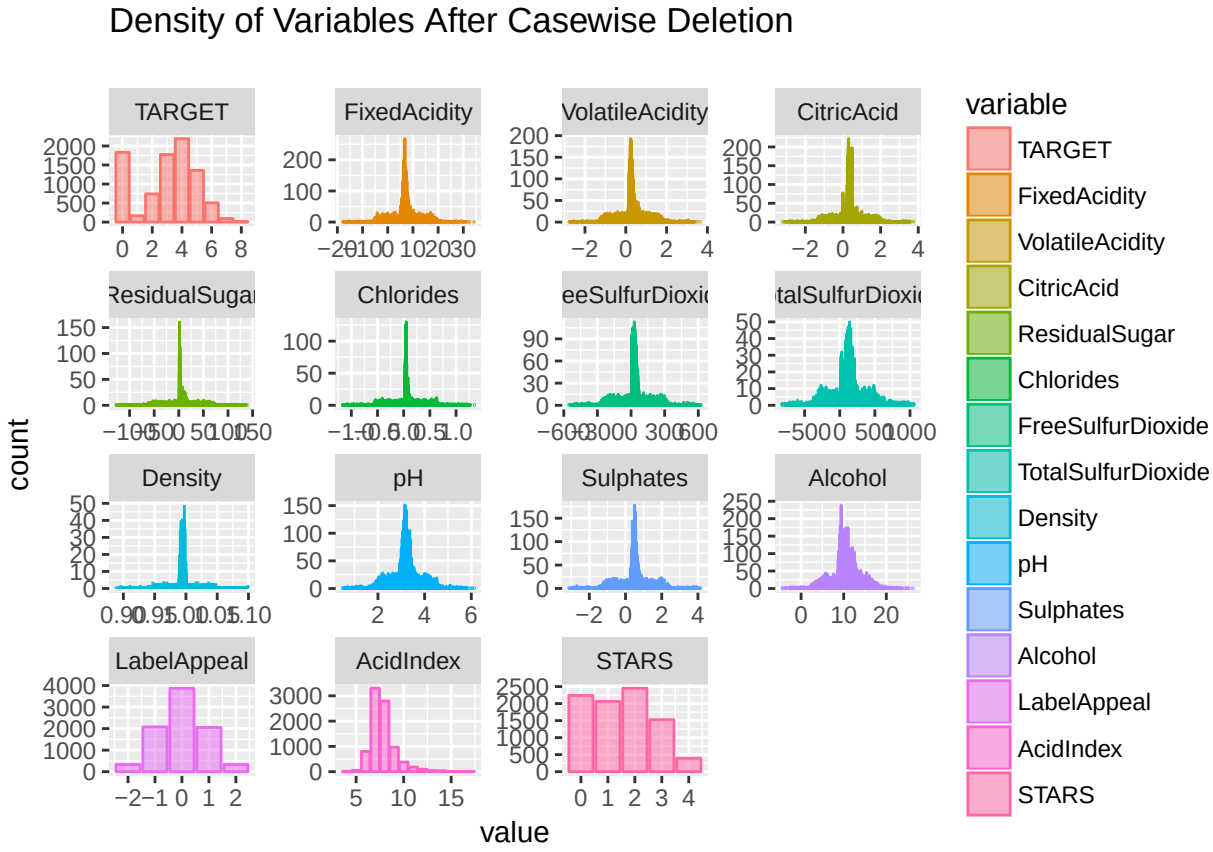


Even though we have such a large number of observations, deleting such a large quantity of cases may affect the prediction ability of our models. Deleting all of the cases with missing values will result in the reduction of our data set by about 1/3, but due to the size of the data set, the predictors have very little skew, and the majority of values being clustered around the mean, the cases with NA values will be removed from the data set. Imputing the mean or median will simply add to the clustering of values around the center of each distribution. A visual comparison of the predictor variables and the quantitiy of NAs is below:




```
Variables sorted by number of missings:
          Variable Count
         Sulphates  1210
TotalSulfurDioxide   682
           Alcohol   653
 FreeSulfurDioxide   647
         Chlorides   638
     ResidualSugar   616
                pH   395
```

Histograms of each variable in the data set after casewise deleteion is below:



Examinging all of the variables, there do not appear to be any values far outside of the ranges, values that are nonsensical (negative cases purchased, for example), or need to be removed before moving forward. The combination of variables, or ratios of existing predictors did not seem to yield any effective results.

Log or square root transformations may help to improve models using some of the slightly skewed predictors. In particular, `AcidIndex` appears the most skewed of the predictors. Log and square root transformations are performed to observe the change in correlation of this variable with `TARGET` — both improve the correlation by a very small amount, but the log-transform yields a greater improvement.

Correlations

Very few of the predictors are correlated with the response variable. As we would expect from the previous boxplots, `STARS` and `LabelAppeal` have moderate positive correlations with `TARGET`, and `AcidIndex` has a slight negative correlation. Additionally, very few of the predictor variables are correlated with each other, suggesting that our models will not have much multicollinearity. `STARS` is only slightly correlated with `LabelAppeal` and `AcidIndex`.

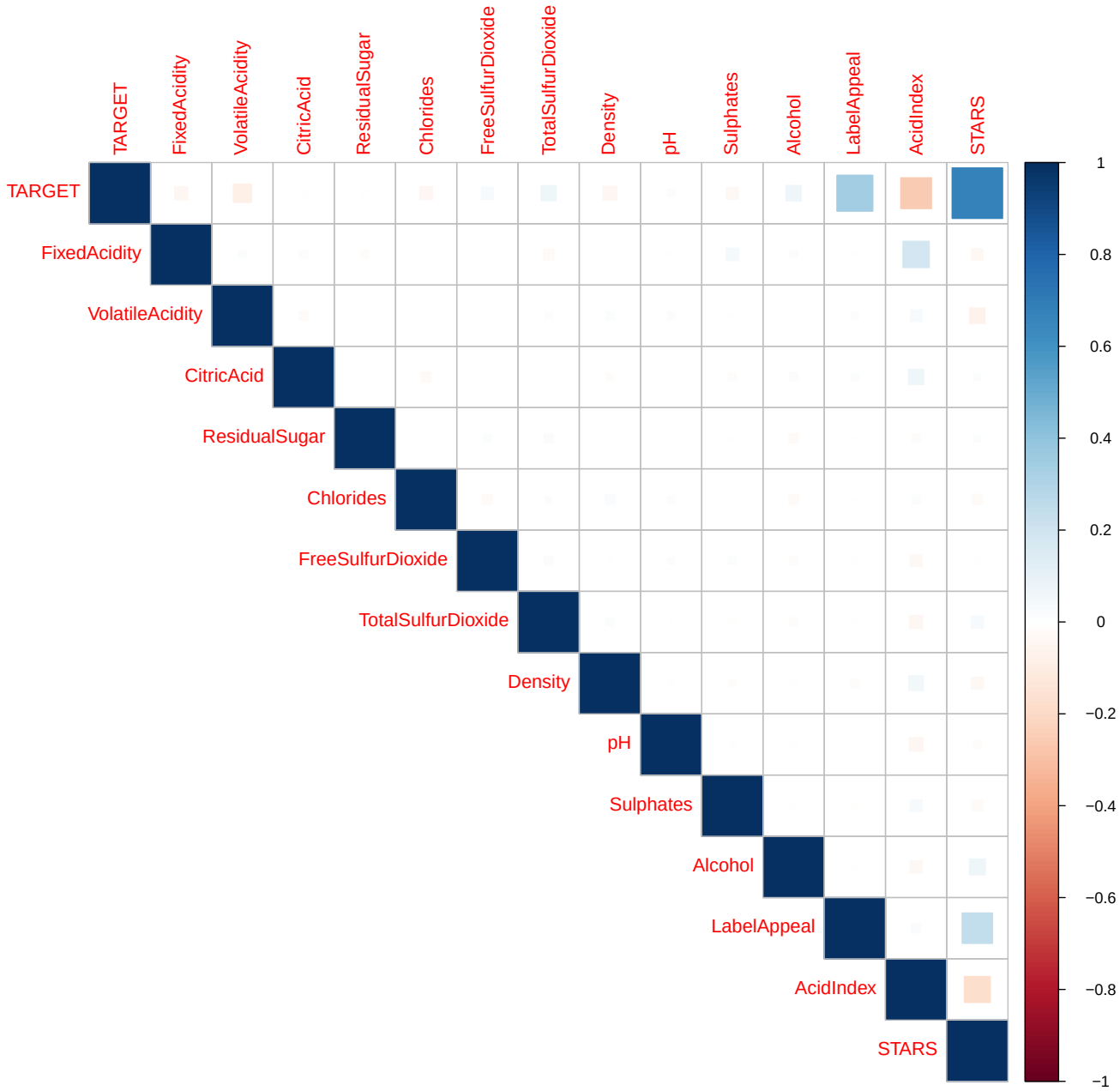


Table 4: Top Correlated Variable Pairs

	Var1	Var2	Correlation	Rsquared
1	TARGET	STARS	0.68	0.46
2	TARGET	LabelAppeal	0.34	0.12
3	TARGET	AcidIndex	-0.25	0.06
4	LabelAppeal	STARS	0.25	0.06
5	FixedAcidity	AcidIndex	0.18	0.03
6	AcidIndex	STARS	-0.18	0.03
7	TARGET	VolatileAcidity	-0.09	0.01
8	Alcohol	STARS	0.07	0.00
9	VolatileAcidity	STARS	-0.06	0.00
10	TARGET	Alcohol	0.06	0.00

4. Model Creation

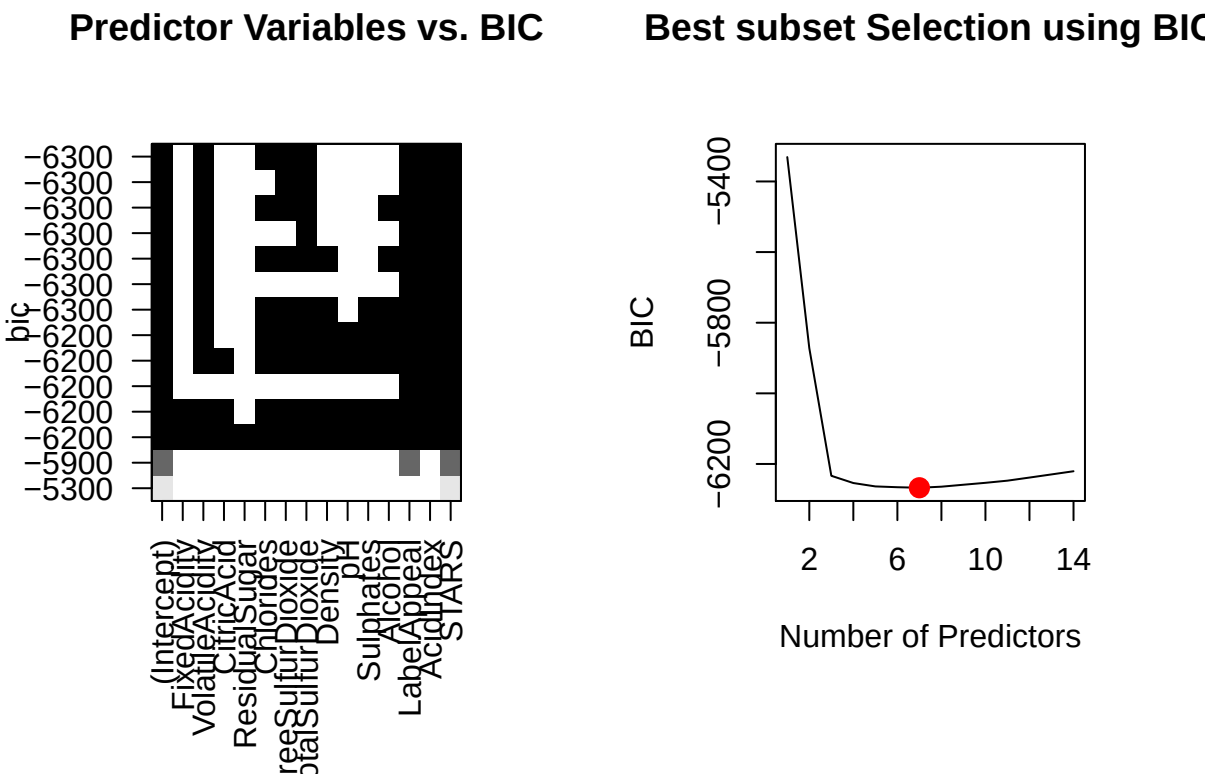
```
Using the training data set, build at least two different poisson regression models, at least two
different negative binomial regression models, and at least two multiple linear regression
models, using different variables (or the same variables with different transformations). Sometimes
poisson and negative binomial regression models give the same results. If that is the case, comment
on that. Consider changing the input variables if that occurs so that you get different models.
Although not covered in class, you may also want to consider building zero-inflated poisson
and negative binomial regression models. You may select the variables manually, use an approach such
as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques.
Describe the techniques you used. If you manually selected a variable for inclusion into the model
or exclusion into the model, indicate why this was done.

Discuss the coefficients in the models, do they make sense? In this case, about the only thing you can
comment on is the number of stars and the wine label appeal. However, you might comment on the coefficient
and magnitude of variables and how they are similar or different from model to model. For example, you
might say "pH seems to have a major positive impact in my poisson regression model, but a negative
effect in my multiple linear regression model". Are you keeping the model even though it is counter
intuitive? Why? The boss needs to know.
```

Only one response variable exists, and we will use at least two versions of three different types of models. Because the `TARGET` variable is a poisson distribution, we will create two poisson regression models, followed by two models using the negative binomial regression model, and lastly at least two using our familiar multiple linear regression model.

Before creating the different models, we will investigate using Bayesian Information Criteria (BIC) and Mallow's C_p to determine the quantity of predictors, and which ones specifically to use in our models.

BIC Predictor Selection

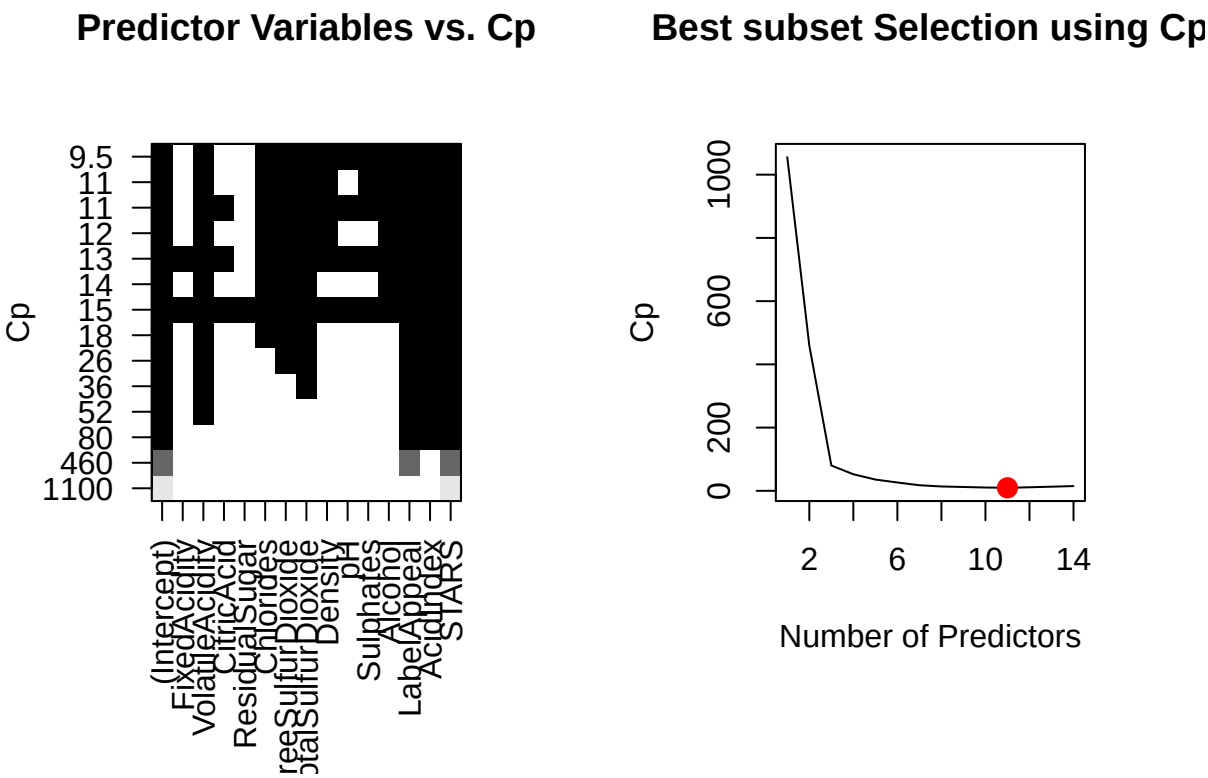
First, we will look at predictor selection using Bayesian Information Criteria:



The first plot shows that our most significant predictor, `STARS` would appear in every model, but more effective models would

contain the `AcidIndex` and `LabelAppeal` predictors as well. The plot on the right shows the lowest BIC values for models using 7 predictors. We will investigate if adding additional predictors into the model is worth giving up the simplicity of the model. The difference in BIC values for 3 vs. 7 predictors is not drastic.

Mallow's C_p Predictor Selection



Using Mallow's C_p, the smallest Cp values are associated with models with 11 predictors. The more parsimonious models would contain the higher correlated variables (`STARS`, `LabelAppeal`, and `AcidIndex`), but the lowest Cp value model also contains `VolatileAcidity`, `Chlorides`, `FreeSulfurDioxide`, `TotalSulfurDioxide`, `Alcohol`, `Density`, `Sulphates`, and `pH`.

4.1 - Poisson Regression

Model 0: Full Model

First, a full model using Poisson regression and all 14 of the predictors is built:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.618	0.2368	6.83	8.491e-12
FixedAcidity	-0.0001785	0.001001	-0.1783	0.8585
VolatileAcidity	-0.03296	0.007888	-4.178	2.936e-05
CitricAcid	0.004358	0.007178	0.6071	0.5438
ResidualSugar	-5.403e-05	0.0001831	-0.2951	0.7679
Chlorides	-0.04827	0.01939	-2.489	0.01281
FreeSulfurDioxide	0.0001275	4.173e-05	3.057	0.002239
TotalSulfurDioxide	9.401e-05	2.698e-05	3.484	0.0004932
Density	-0.3618	0.2332	-1.552	0.1208
pH	-0.01708	0.009073	-1.883	0.05976
Sulphates	-0.01092	0.006657	-1.64	0.101
Alcohol	0.001492	0.001677	0.89	0.3735
LabelAppeal	0.1324	0.007369	17.96	3.764e-72
AcidIndex	-0.08671	0.005479	-15.82	2.112e-56
STARS	0.3094	0.005532	55.94	0

(Dispersion parameter for poisson family taken to be 1)

Null deviance:	15334 on 8674 degrees of freedom
Residual deviance:	9962 on 8660 degrees of freedom

Seven of the predictor coeffcients are significant under a reasonable α of 0.05, and one more, `pH` being just over the threshold. The coefficients themselves are fairly small, with most of the more significant predictors having slopes of greater magnitude than those which are not. Eight of the coefficients have negative slopes, while the other six are positive. Those with positive coefficients, such as `STARS` and `LabelAppeal`, make sense as greater ratings and more appealing label aesthetics would entice buyers to purchase more wine.

Model 1: Reduced Model

A reduced model is created, using the predictors recommended by using Mallow's C_p to determine the best fit. Here, most of the predictors are the highly significant ones from the full poisson regression model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.619	0.2368	6.837	8.104e-12
VolatileAcidity	-0.03305	0.007885	-4.191	2.774e-05
Chlorides	-0.04856	0.01939	-2.504	0.01226
FreeSulfurDioxide	0.0001275	4.172e-05	3.056	0.002241
TotalSulfurDioxide	9.381e-05	2.697e-05	3.478	0.0005043
LabelAppeal	0.1324	0.007368	17.97	3.381e-72
AcidIndex	-0.08664	0.005411	-16.01	1.026e-57
STARS	0.3094	0.00553	55.96	0
Alcohol	0.001523	0.001676	0.9084	0.3637
Density	-0.3641	0.2331	-1.562	0.1184
Sulphates	-0.01101	0.006653	-1.655	0.09793
pH	-0.01706	0.009072	-1.88	0.06007

(Dispersion parameter for poisson family taken to be 1)

Null deviance:	15334 on 8674 degrees of freedom
Residual deviance:	9963 on 8663 degrees of freedom

Of the 11 predictors, five have positive coefficients, six have negative coefficients. Increases in acidity and chlorides seem to have a negative effect on the number of cases purchased; this may have to do with any wines with more extreme values in these predictors to demand a specific pallette. Four variables are not significate at the $\alpha = 0.5$ level: `Alcohol`, `Density`, `Sulphates`, and `pH`. The predictors most correlated to `TARGET` have the most significant statistical significance. The reduced model using the recommended predictors from Mallow's C_p only returns a very slight reduction in AIC.

Model 2: Significant Reduced Model

For our last poisson regression model, the predictors which were the most significant and highly correlated to our response variable are used:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.223	0.04415	27.69	8.413e-169
LabelAppeal	0.1324	0.007361	17.98	2.665e-72
AcidIndex	-0.08813	0.005374	-16.4	1.889e-60
STARS	0.3123	0.005493	56.86	0

(Dispersion parameter for poisson family taken to be 1)

Null deviance:	15334 on 8674 degrees of freedom
Residual deviance:	10018 on 8671 degrees of freedom

Here the AIC value increases again slightly, but only three predictors out of 14 are used. The difference in goodness of fit may not be enough to justify using a simpler model. This will be investigated further in the model selection section. There is also evidence of overdispersion, given that the residual deviance divided by the df is > 1. The overdispersion may be the result of outliers, possibly in the response variable, given that the largest number of cases ordered happens less frequently than for the other amounts.

4.2 - Negative Binomial Regression

One way of possibly dealing with the overdispersion is using a negative binomial (NB) regression model. Negative binomial regression has the k parameter, which is the dispersion parameter. As this parameter gets larger, the variance will converge to the mean.

Model 3: BIC Predictors

For the first negative binomial regression model, the predictors selected using BIC are used.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.173	0.09221	34.41	2.696e-243
VolatileAcidity	-0.09812	0.01822	-5.384	7.464e-08
Chlorides	-0.1453	0.04462	-3.257	0.001129
FreeSulfurDioxide	0.0003159	9.621e-05	3.283	0.00103
TotalSulfurDioxide	0.0002625	6.213e-05	4.226	2.405e-05
AcidIndex	-0.2069	0.01087	-19.03	3.623e-79
LabelAppeal	0.4307	0.01669	25.8	1.713e-141
STARS	0.9728	0.01276	76.26	0

(Dispersion parameter for gaussian family taken to be 1.768483)

Null deviance:	31832 on 8674 degrees of freedom
Residual deviance:	15327 on 8667 degrees of freedom

The AIC for this model is 31704, similar to what was achieved using the predictors selected by Mallow's C_p for the Poisson regression. Also similar are the coefficients and directions of the slopes, which is a result of the overdispersion.

Model 4: Significant Predictors

For the second negative binomial model, only the predictors with highly significant coefficients from the previous model are used:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.186	0.09212	34.58	1.263e-245
VolatileAcidity	-0.09854	0.01824	-5.401	6.792e-08
TotalSulfurDioxide	0.0002682	6.219e-05	4.313	1.627e-05
LabelAppeal	0.4306	0.01671	25.77	3.475e-141
AcidIndex	-0.2087	0.01088	-19.18	2.334e-80
STARS	0.9738	0.01277	76.27	0

(Dispersion parameter for gaussian family taken to be 1.772552)

Null deviance:	31832 on 8674 degrees of freedom
Residual deviance:	15366 on 8669 degrees of freedom

The AIC for this model increased slightly to 31715. The coefficients for the reduced model are quite similar to the full BIC criteria model, but their significance has increased.

4.3 - Multiple Linear Regression

In response to the high level of dispersion exhibited, two multiple linear regression models are created on the data. A full model, including the log-transformed acid index variable, is built (but not presented here) to aid in variable selection.

Model 5: Reduced Model

The 7 most significant predictors from the full model are used to create a modified linear model. Most of these predictors match the BIC-selected predictors (with the exception of the log transformation on `AcidIndex`).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.997	0.1928	25.92	1.013e-142
VolatileAcidity	-0.0999	0.01825	-5.475	4.504e-08
Chlorides	-0.1452	0.04468	-3.249	0.001161
FreeSulfurDioxide	0.0003199	9.634e-05	3.321	0.0009007
TotalSulfurDioxide	0.0002704	6.22e-05	4.347	1.396e-05
LabelAppeal	0.4283	0.01671	25.63	9.401e-140
log(AcidIndex)	-1.687	0.09186	-18.37	6.251e-74
STARS	0.9757	0.01276	76.46	0

(Dispersion parameter for gaussian family taken to be 1.773397)

Null deviance:	31832 on 8674 degrees of freedom
Residual deviance:	15370 on 8667 degrees of freedom

This model yields an AIC of 29584. The signs of these coefficients match those of the BIC-criteria negative binomial model. Many of the coefficients have similar magnitudes, but in this model, the effects of `LabelAppeal` and `STARS` are greater by roughly three-fold. Additionally, the coefficient for the `AcidIndex` variable has increased roughly 20-fold, but this can be largely attributed to the fact that it refers to the transformed variable. All predictors show very high statistical significance.

Model 6: Significant Predictors

A second linear model is created, including any non-transformed predictors that held statistical significance at the $\alpha = 0.10$ level.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.327	0.5428	7.971	1.77e-15
VolatileAcidity	-0.09739	0.01822	-5.345	9.281e-08
FreeSulfurDioxide	0.0003192	9.619e-05	3.318	0.0009089
TotalSulfurDioxide	0.0002638	6.212e-05	4.247	2.187e-05
Chlorides	-0.1448	0.04462	-3.246	0.001173
Density	-1.032	0.538	-1.919	0.05508
pH	-0.03663	0.02101	-1.743	0.08131
Sulphates	-0.02866	0.01534	-1.868	0.06174
LabelAppeal	0.4303	0.01669	25.79	2.387e-141
AcidIndex	-0.2062	0.0109	-18.91	3.616e-78
STARS	0.9712	0.01276	76.1	0

(Dispersion parameter for gaussian family taken to be 1.76701)

Null deviance:	31832 on 8674 degrees of freedom
Residual deviance:	15309 on 8664 degrees of freedom

This multiple linear model returns an AIC of 29570, the lowest of any models created. The coefficients have the same sign and very similar magnitudes as those of the reduced model above. The three predictors not included in model 5 have similar coefficients to those of model 3 above. These three predictors – `Density`, `pH`, and `Sulphates` – are not significant at the $\alpha = 0.5$ level, but two of them are marginally not significant, and all three would still be significant at the $\alpha = 0.10$ level.

5. Model Selection and Prediction

```
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.
```

5.1 - Model Comparsion

The characteristics and performance of the seven multiple linear regression models from the previous section are compared below:

Model #	Type	# of Predictors	AIC
0	Poisson	14	31705
1	Poisson	11	31700
2	Poisson	3	31739
3	Negative Binomial	7	31704
4	Negative Binomial	5	31750
5	Mult. Linear Reg.	7	29598
6	Mult. Linear Reg.	10	29570

5.2 - 10-fold Cross Validation

Mean CV Error

Poisson Model 0	6.739
Poisson Model 1	6.738
Poisson Model 2	6.743
NB Model 3	6.738
NB Model 4	6.74
MLR Model 5	1.776
MLR Model 6	1.771

Multiple Linear Model 6 exhibits both the lowest AIC and the lowest mean cross-validation error. This may seem surprising given the nature of the `TARGET` data, but is a sensible outcome given the overdispersion apparent in the data. The linear nature of this model provides the benefit of being easily understood by a wide audience as compared to Poisson or negative binomial models.

The linear model is applied to a test dataset containing response variables for 3335 cases.

The predicted ratings are converted to even case values by rounding to the nearest integer. A table of the proportion of ratings at each number of cases is presented for both the test and training datasets.

	0	1	2	3	4	5	6	7	8
Test	0	0.002	0.1	0.177	0.161	0.099	0.035	0.001	0
Train	0.211	0.019	0.085	0.205	0.253	0.157	0.058	0.011	0.001

The full sets of predictions – both raw predicted values and even case values – for the evaluation dataset are available in Appendix B.

Appendix A - Results from Predictive Model

Index	Value	Cases	Index	Value	Cases
3	NA	NA	7908	NA	NA
9	3.648	4	7917	1.691	2
10	2.222	2	7924	2.817	3
18	2.098	2	7948	4.046	4
21	NA	NA	7950	6.265	6
30	5.775	6	7955	2.257	2
31	3.399	3	7957	NA	NA
37	NA	NA	7959	NA	NA
39	NA	NA	7967	5.91	6
47	NA	NA	7969	NA	NA
60	NA	NA	7971	4.035	4
62	NA	NA	7974	NA	NA
63	3.449	3	7976	NA	NA
64	NA	NA	7986	5.86	6
68	NA	NA	7987	NA	NA
75	NA	NA	7993	3.862	4
76	2.354	2	7996	3.642	4
83	NA	NA	7998	3.207	3
87	3.682	4	8018	3.166	3
92	5.608	6	8019	NA	NA
98	2.6	3	8027	NA	NA
106	NA	NA	8036	NA	NA
107	NA	NA	8040	4.332	4
113	2.266	2	8044	3.678	4
120	3.845	4	8050	3.162	3
123	4.717	5	8052	NA	NA
125	2.917	3	8054	NA	NA
126	6.168	6	8057	3.652	4
128	5.513	6	8058	4.85	5
129	2.122	2	8059	NA	NA
131	NA	NA	8066	3.842	4
135	NA	NA	8070	5.594	6
141	4.709	5	8072	5.323	5
147	2.776	3	8078	2.562	3
148	NA	NA	8079	3.334	3
151	3.647	4	8080	4.325	4
156	2.98	3	8081	4.498	4
157	4.373	4	8088	NA	NA
174	NA	NA	8091	2.316	2
186	NA	NA	8094	2.319	2
193	2.136	2	8095	NA	NA
195	NA	NA	8099	4.933	5
212	NA	NA	8101	5.02	5
213	NA	NA	8102	5.252	5
217	NA	NA	8116	4.852	5
223	3.924	4	8125	4.796	5
226	2.737	3	8134	NA	NA
228	4.284	4	8139	NA	NA
230	4.479	4	8141	4.039	4
241	2.194	2	8147	2.704	3
243	3.617	4	8158	NA	NA
249	NA	NA	8160	2.394	2
281	4.458	4	8165	NA	NA
288	NA	NA	8187	2.12	2
294	NA	NA	8205	NA	NA
295	1.842	2	8209	2.787	3

Index	Value	Cases	Index	Value	Cases
300	NA	NA	8211	NA	NA
302	3.845	4	8232	3.609	4
303	NA	NA	8236	4.019	4
308	NA	NA	8237	NA	NA
319	NA	NA	8238	6.594	7
320	NA	NA	8245	4.464	4
324	NA	NA	8256	NA	NA
331	2.343	2	8268	NA	NA
343	NA	NA	8269	2.414	2
347	2.161	2	8270	NA	NA
348	3.778	4	8286	2.386	2
350	4.095	4	8289	NA	NA
357	NA	NA	8301	3.676	4
358	3.505	4	8305	NA	NA
360	NA	NA	8310	2.778	3
366	3.141	3	8312	NA	NA
367	2.276	2	8318	5.077	5
368	5.098	5	8321	4.43	4
376	1.89	2	8328	NA	NA
380	2.949	3	8331	NA	NA
388	NA	NA	8334	NA	NA
396	NA	NA	8344	3.879	4
398	5.635	6	8345	1.901	2
403	NA	NA	8352	4.74	5
410	1.959	2	8358	NA	NA
412	NA	NA	8359	NA	NA
420	2.242	2	8360	2.354	2
434	2.089	2	8365	3.185	3
440	2.667	3	8366	3.286	3
450	3.33	3	8369	4.824	5
453	2.364	2	8373	NA	NA
464	5.603	6	8378	3.746	4
465	NA	NA	8392	3.256	3
466	5.433	5	8397	3.092	3
473	2.038	2	8399	1.981	2
476	NA	NA	8400	2.664	3
478	NA	NA	8405	2.5	3
479	3.303	3	8406	NA	NA
493	2.442	2	8410	NA	NA
497	NA	NA	8413	4.477	4
503	3.715	4	8414	NA	NA
504	3.135	3	8416	NA	NA
505	2.954	3	8426	2.156	2
507	NA	NA	8434	4.127	4
513	2.439	2	8439	NA	NA
519	NA	NA	8440	NA	NA
521	3.71	4	8475	4.784	5
522	3.828	4	8480	4.979	5
545	NA	NA	8497	1.854	2
549	NA	NA	8499	NA	NA
551	NA	NA	8500	3.126	3
556	NA	NA	8501	3.161	3
557	6.217	6	8502	NA	NA
559	NA	NA	8518	4.583	5
560	NA	NA	8520	3.851	4
566	3.503	4	8523	4.012	4
569	3.873	4	8525	NA	NA
573	NA	NA	8532	NA	NA

Index	Value	Cases	Index	Value	Cases
578	NA	NA	8535	2.015	2
579	NA	NA	8543	NA	NA
582	4.661	5	8554	NA	NA
596	NA	NA	8560	5.78	6
598	NA	NA	8561	5.312	5
599	NA	NA	8563	NA	NA
602	2.658	3	8566	NA	NA
605	NA	NA	8570	3.78	4
617	NA	NA	8572	4.17	4
619	5.548	6	8582	1.166	1
630	3.068	3	8583	NA	NA
634	3.288	3	8587	NA	NA
643	NA	NA	8592	1.653	2
645	NA	NA	8593	NA	NA
647	4.041	4	8607	NA	NA
649	2.835	3	8609	4.129	4
656	3.506	4	8610	NA	NA
657	4.593	5	8614	NA	NA
658	NA	NA	8616	5.332	5
667	4.181	4	8622	5.053	5
692	NA	NA	8623	3.981	4
693	4.626	5	8624	NA	NA
698	NA	NA	8633	6.055	6
699	3.458	3	8641	NA	NA
700	NA	NA	8644	NA	NA
704	3.309	3	8649	3.844	4
707	NA	NA	8653	2.732	3
708	3.948	4	8657	5.544	6
709	2.651	3	8658	2.476	2
713	NA	NA	8663	3.246	3
714	2.9	3	8672	1.608	2
716	1.748	2	8680	NA	NA
718	3.643	4	8684	2.457	2
722	4.162	4	8687	NA	NA
729	NA	NA	8688	3.931	4
731	2.261	2	8690	3.687	4
733	3.218	3	8712	NA	NA
746	2.669	3	8717	3.032	3
747	3.262	3	8730	4.831	5
748	NA	NA	8739	3.342	3
753	NA	NA	8744	3.92	4
757	NA	NA	8747	NA	NA
763	3.396	3	8748	4.62	5
767	5.065	5	8751	3.293	3
774	2.507	3	8758	3.239	3
776	NA	NA	8761	NA	NA
788	NA	NA	8763	NA	NA
794	3.61	4	8764	NA	NA
799	NA	NA	8765	5.465	5
803	4.562	5	8773	NA	NA
806	4.445	4	8780	2.48	2
807	3.4	3	8781	3.016	3
811	4.764	5	8782	NA	NA
816	NA	NA	8785	NA	NA
818	3.203	3	8786	NA	NA
819	NA	NA	8797	NA	NA
831	4.961	5	8799	4.651	5
835	4.821	5	8807	NA	NA

Index	Value	Cases	Index	Value	Cases
837	NA	NA	8816	NA	NA
841	NA	NA	8817	4.065	4
846	NA	NA	8826	NA	NA
856	NA	NA	8833	2.82	3
861	3.671	4	8834	NA	NA
862	NA	NA	8835	NA	NA
863	3.069	3	8840	5.072	5
865	NA	NA	8843	3.016	3
871	2.435	2	8849	3.215	3
879	NA	NA	8855	2.998	3
880	2.288	2	8861	2.031	2
881	2.972	3	8862	4.057	4
885	3.732	4	8865	NA	NA
887	NA	NA	8868	5.565	6
892	NA	NA	8870	3.966	4
898	3.646	4	8880	NA	NA
900	NA	NA	8885	NA	NA
904	NA	NA	8894	NA	NA
906	5.076	5	8895	4.914	5
910	4.161	4	8899	4.894	5
912	4.089	4	8912	NA	NA
913	2.446	2	8922	2.339	2
919	5.551	6	8924	NA	NA
924	NA	NA	8928	3.945	4
925	2.903	3	8932	NA	NA
930	3.26	3	8943	5.283	5
940	NA	NA	8945	NA	NA
941	3.328	3	8946	4.1	4
946	NA	NA	8954	1.784	2
949	4.967	5	8958	NA	NA
951	NA	NA	8960	NA	NA
962	3.933	4	8965	1.448	1
966	2.096	2	8966	4.084	4
967	6.008	6	8967	1.946	2
971	NA	NA	8969	3.277	3
981	3.767	4	8980	NA	NA
982	NA	NA	8984	NA	NA
983	NA	NA	8985	NA	NA
984	NA	NA	8988	NA	NA
989	2.09	2	8989	4.133	4
990	4.15	4	8995	NA	NA
992	2.609	3	9004	2.77	3
995	5.093	5	9010	NA	NA
996	NA	NA	9012	NA	NA
998	NA	NA	9018	NA	NA
1001	5.169	5	9036	NA	NA
1007	NA	NA	9037	1.842	2
1008	NA	NA	9040	NA	NA
1016	2.38	2	9041	4.769	5
1022	NA	NA	9044	5.81	6
1027	4.931	5	9045	NA	NA
1032	NA	NA	9047	2.162	2
1033	3.36	3	9049	NA	NA
1041	4.519	5	9061	NA	NA
1065	NA	NA	9062	2.394	2
1074	NA	NA	9076	3.518	4
1075	NA	NA	9079	2.352	2
1081	NA	NA	9081	3.808	4

Index	Value	Cases	Index	Value	Cases
1094	4.683	5	9082	3.847	4
1099	3.279	3	9089	NA	NA
1105	2.478	2	9092	3.048	3
1123	NA	NA	9094	NA	NA
1135	NA	NA	9115	NA	NA
1142	2.154	2	9117	4.254	4
1155	2.037	2	9118	2.884	3
1169	NA	NA	9120	NA	NA
1176	NA	NA	9124	NA	NA
1178	3.816	4	9128	NA	NA
1180	3.36	3	9135	NA	NA
1184	NA	NA	9136	2.41	2
1185	NA	NA	9138	4.421	4
1193	NA	NA	9157	NA	NA
1196	NA	NA	9176	NA	NA
1199	NA	NA	9183	NA	NA
1203	2.243	2	9187	2.858	3
1205	2.324	2	9188	NA	NA
1207	2.087	2	9190	4.525	5
1208	NA	NA	9197	NA	NA
1212	NA	NA	9200	4.081	4
1213	NA	NA	9201	NA	NA
1222	NA	NA	9203	NA	NA
1223	NA	NA	9212	3.534	4
1226	NA	NA	9213	NA	NA
1227	5.412	5	9214	NA	NA
1229	NA	NA	9217	2.743	3
1230	NA	NA	9219	3.406	3
1231	2.642	3	9220	4.985	5
1241	NA	NA	9221	5.898	6
1243	4.664	5	9237	NA	NA
1244	6.135	6	9240	3.8	4
1246	4.829	5	9241	2.105	2
1248	2.442	2	9248	3.407	3
1249	3.794	4	9253	6.38	6
1252	3.366	3	9259	4.019	4
1261	3.221	3	9267	NA	NA
1275	4.206	4	9271	2.381	2
1281	NA	NA	9273	NA	NA
1285	NA	NA	9285	5.995	6
1288	NA	NA	9290	NA	NA
1290	3.48	3	9291	2.953	3
1291	NA	NA	9293	NA	NA
1304	3.791	4	9294	3.761	4
1305	3.54	4	9301	3.44	3
1323	3.888	4	9302	NA	NA
1342	NA	NA	9312	2.134	2
1348	NA	NA	9316	3.514	4
1353	3.523	4	9319	NA	NA
1363	3.049	3	9328	4.627	5
1371	3.262	3	9331	3.366	3
1372	NA	NA	9338	NA	NA
1378	NA	NA	9350	2.655	3
1381	3.823	4	9356	3.555	4
1382	4.497	4	9359	1.798	2
1393	NA	NA	9362	3.246	3
1394	5.183	5	9364	2.631	3
1398	5.446	5	9370	3.127	3

Index	Value	Cases	Index	Value	Cases
1404	NA	NA	9380	NA	NA
1405	3.4	3	9386	NA	NA
1419	NA	NA	9394	NA	NA
1421	2.822	3	9407	3.648	4
1426	NA	NA	9411	2.855	3
1431	NA	NA	9422	3.551	4
1435	NA	NA	9423	NA	NA
1437	NA	NA	9429	NA	NA
1438	NA	NA	9433	NA	NA
1442	NA	NA	9439	NA	NA
1464	NA	NA	9451	4.479	4
1471	NA	NA	9452	2.993	3
1473	4.043	4	9453	NA	NA
1476	3.377	3	9460	NA	NA
1478	1.98	2	9465	4.582	5
1479	3.943	4	9470	NA	NA
1487	5.452	5	9476	NA	NA
1492	4.111	4	9485	3.483	3
1496	3.326	3	9486	NA	NA
1497	NA	NA	9488	2.722	3
1515	NA	NA	9507	5.848	6
1519	NA	NA	9508	NA	NA
1522	3.096	3	9517	6.554	7
1526	3.297	3	9521	4.777	5
1537	2.661	3	9528	3.289	3
1538	4.777	5	9532	NA	NA
1540	1.64	2	9536	2.917	3
1543	NA	NA	9540	5.201	5
1548	NA	NA	9542	3.986	4
1549	NA	NA	9546	4.177	4
1556	2.909	3	9548	4.585	5
1564	NA	NA	9549	NA	NA
1570	4.306	4	9554	5.661	6
1577	1.949	2	9555	3.972	4
1585	4.711	5	9558	NA	NA
1590	3.796	4	9573	NA	NA
1592	NA	NA	9575	5.279	5
1594	NA	NA	9584	3.041	3
1596	6.423	6	9586	NA	NA
1598	5.815	6	9588	5.149	5
1603	NA	NA	9591	NA	NA
1607	NA	NA	9592	NA	NA
1612	5.912	6	9597	NA	NA
1627	4.167	4	9600	4.23	4
1629	3.265	3	9603	NA	NA
1630	NA	NA	9605	2.268	2
1640	5.75	6	9614	NA	NA
1641	4.44	4	9616	NA	NA
1646	3.778	4	9622	3.612	4
1662	NA	NA	9624	4.15	4
1668	NA	NA	9629	NA	NA
1671	NA	NA	9633	NA	NA
1672	4.747	5	9640	4.382	4
1673	NA	NA	9644	1.355	1
1686	4.008	4	9645	NA	NA
1688	4.253	4	9646	2.69	3
1696	4.138	4	9648	2.782	3
1701	5.472	5	9649	NA	NA

Index	Value	Cases	Index	Value	Cases
1707	NA	NA	9660	3.651	4
1708	3.248	3	9664	NA	NA
1713	3.622	4	9675	NA	NA
1715	NA	NA	9679	2.911	3
1717	2.65	3	9680	3.035	3
1721	3.39	3	9682	NA	NA
1724	NA	NA	9697	2.721	3
1725	3.098	3	9701	4.349	4
1730	3.685	4	9704	2.878	3
1731	4.461	4	9705	NA	NA
1734	3.167	3	9707	4.225	4
1740	2.987	3	9714	NA	NA
1748	2.628	3	9718	2.34	2
1749	3.648	4	9722	5.225	5
1750	5.787	6	9739	2.535	3
1763	2.837	3	9747	5.821	6
1768	4.862	5	9751	NA	NA
1773	NA	NA	9757	1.877	2
1777	NA	NA	9759	4.428	4
1778	NA	NA	9760	NA	NA
1780	3.041	3	9764	2.597	3
1782	NA	NA	9776	NA	NA
1784	4.129	4	9778	2.531	3
1786	3.595	4	9786	NA	NA
1787	NA	NA	9803	3.563	4
1792	NA	NA	9804	NA	NA
1800	NA	NA	9815	4.897	5
1801	3.119	3	9824	NA	NA
1803	2.297	2	9825	NA	NA
1804	4.334	4	9826	2.832	3
1807	2.392	2	9827	3.81	4
1818	5.336	5	9833	3.543	4
1821	3.192	3	9835	NA	NA
1822	4.757	5	9860	3.718	4
1828	2.615	3	9865	2.877	3
1833	3.748	4	9871	4.465	4
1844	4.346	4	9874	NA	NA
1847	2.823	3	9880	2.564	3
1850	2.422	2	9882	NA	NA
1854	3.63	4	9885	2.88	3
1858	4.606	5	9888	4.012	4
1864	4.149	4	9892	1.898	2
1867	NA	NA	9893	4.881	5
1876	NA	NA	9896	NA	NA
1880	NA	NA	9902	3.051	3
1881	NA	NA	9906	3.854	4
1891	2.843	3	9910	4.792	5
1894	NA	NA	9914	NA	NA
1895	3.936	4	9918	NA	NA
1901	NA	NA	9920	NA	NA
1905	3.841	4	9926	3.524	4
1912	5.41	5	9931	5.136	5
1918	2.995	3	9935	3.575	4
1921	3.815	4	9945	3.849	4
1923	2.941	3	9953	NA	NA
1924	NA	NA	9957	NA	NA
1931	NA	NA	9963	NA	NA
1941	4.729	5	9972	3.836	4

Index	Value	Cases	Index	Value	Cases
1950	NA	NA	9976	NA	NA
1951	4.772	5	9979	NA	NA
1954	4.279	4	9980	NA	NA
1961	3.478	3	9982	NA	NA
1966	NA	NA	9991	NA	NA
1979	4.099	4	10000	4.353	4
1982	NA	NA	10003	3.462	3
1987	2.711	3	10005	2.001	2
1997	NA	NA	10014	3.367	3
2004	4.49	4	10032	3.113	3
2011	5.276	5	10034	NA	NA
2015	3.187	3	10041	NA	NA
2025	6.329	6	10042	3.902	4
2033	NA	NA	10044	5.13	5
2034	NA	NA	10045	NA	NA
2035	NA	NA	10054	3.622	4
2036	NA	NA	10061	NA	NA
2053	2.525	3	10062	NA	NA
2059	NA	NA	10073	NA	NA
2060	NA	NA	10081	NA	NA
2073	NA	NA	10084	NA	NA
2084	2.775	3	10086	NA	NA
2089	4.545	5	10093	NA	NA
2092	NA	NA	10101	NA	NA
2109	5.694	6	10105	4.474	4
2129	4.229	4	10110	NA	NA
2134	4.911	5	10113	2.787	3
2135	4.663	5	10115	2.636	3
2148	NA	NA	10119	3.325	3
2149	NA	NA	10121	4.294	4
2150	2.122	2	10124	NA	NA
2165	2.569	3	10126	5.772	6
2166	NA	NA	10127	3.224	3
2168	NA	NA	10145	3.045	3
2170	NA	NA	10147	NA	NA
2171	2.171	2	10148	2.193	2
2172	2.591	3	10162	NA	NA
2176	5.112	5	10163	2.36	2
2182	3.13	3	10166	1.888	2
2189	1.741	2	10172	3.365	3
2191	NA	NA	10173	2.042	2
2197	2.647	3	10175	1.909	2
2202	NA	NA	10180	NA	NA
2203	3.292	3	10186	NA	NA
2204	NA	NA	10192	4.441	4
2206	4.165	4	10199	3.09	3
2218	1.795	2	10209	NA	NA
2219	3.416	3	10210	6.422	6
2221	NA	NA	10214	NA	NA
2226	NA	NA	10215	NA	NA
2228	NA	NA	10216	NA	NA
2232	NA	NA	10232	2.726	3
2236	NA	NA	10239	4.66	5
2241	2.394	2	10249	4.285	4
2245	5.115	5	10253	5.055	5
2251	NA	NA	10255	NA	NA
2255	5.202	5	10262	NA	NA
2256	3.619	4	10264	2.212	2

Index	Value	Cases	Index	Value	Cases
2259	NA	NA	10266	NA	NA
2263	3.491	3	10268	NA	NA
2264	NA	NA	10271	3.143	3
2267	NA	NA	10272	5.111	5
2273	1.696	2	10276	NA	NA
2277	4.058	4	10277	2.118	2
2287	3.962	4	10279	2.575	3
2289	3.749	4	10281	1.597	2
2291	NA	NA	10285	NA	NA
2296	1.517	2	10294	NA	NA
2299	NA	NA	10300	2.882	3
2306	3.217	3	10304	2.913	3
2314	NA	NA	10307	NA	NA
2317	2.632	3	10309	5.086	5
2318	4.342	4	10310	NA	NA
2321	4.599	5	10312	NA	NA
2324	NA	NA	10321	3.335	3
2340	3.315	3	10332	NA	NA
2343	NA	NA	10336	3.435	3
2349	NA	NA	10368	2.699	3
2352	4.921	5	10369	4.547	5
2353	NA	NA	10375	NA	NA
2365	2.14	2	10376	NA	NA
2370	NA	NA	10379	NA	NA
2378	NA	NA	10380	2.581	3
2390	NA	NA	10383	NA	NA
2399	NA	NA	10385	5.079	5
2402	NA	NA	10387	NA	NA
2403	NA	NA	10397	2.521	3
2404	NA	NA	10412	NA	NA
2414	5.026	5	10413	NA	NA
2422	3.449	3	10418	NA	NA
2424	NA	NA	10420	3.325	3
2430	3.827	4	10426	3.862	4
2435	3.236	3	10427	2.358	2
2439	NA	NA	10428	3.18	3
2442	3.826	4	10430	NA	NA
2445	3.615	4	10435	NA	NA
2449	NA	NA	10436	NA	NA
2451	NA	NA	10446	6.27	6
2461	2.877	3	10448	4.046	4
2464	NA	NA	10449	4.145	4
2465	3.754	4	10463	2.098	2
2472	NA	NA	10469	1.906	2
2476	2.228	2	10470	NA	NA
2482	2.853	3	10471	NA	NA
2487	5.087	5	10473	NA	NA
2498	3.93	4	10476	2.869	3
2501	NA	NA	10482	2.737	3
2504	2.327	2	10500	3.301	3
2511	NA	NA	10511	5.046	5
2518	4.972	5	10512	3.647	4
2521	NA	NA	10514	3.266	3
2530	2.959	3	10515	4.929	5
2543	4.167	4	10526	NA	NA
2545	3.618	4	10546	NA	NA
2561	NA	NA	10549	NA	NA
2566	NA	NA	10553	NA	NA

Index	Value	Cases	Index	Value	Cases
2572	2.567	3	10558	1.981	2
2577	2.979	3	10575	NA	NA
2578	NA	NA	10581	2.294	2
2580	2.382	2	10583	NA	NA
2581	4.907	5	10584	NA	NA
2582	3.834	4	10585	NA	NA
2584	NA	NA	10610	NA	NA
2590	4.909	5	10611	NA	NA
2598	3.016	3	10616	3.682	4
2602	NA	NA	10618	2.935	3
2605	5.019	5	10628	NA	NA
2616	3.766	4	10632	NA	NA
2618	3.164	3	10642	3.468	3
2619	NA	NA	10648	3.103	3
2624	NA	NA	10649	3.441	3
2632	NA	NA	10650	NA	NA
2640	3.77	4	10654	1.756	2
2646	4.66	5	10656	5.015	5
2651	NA	NA	10661	4.756	5
2660	NA	NA	10663	NA	NA
2661	4.006	4	10672	2.765	3
2668	NA	NA	10678	4.878	5
2670	NA	NA	10685	NA	NA
2680	2.648	3	10690	4.752	5
2681	3.803	4	10702	3.74	4
2689	NA	NA	10706	3.158	3
2694	4.553	5	10708	2.282	2
2695	NA	NA	10716	3.229	3
2696	NA	NA	10717	NA	NA
2702	2.429	2	10720	4.93	5
2704	3.239	3	10729	NA	NA
2708	5.335	5	10730	5.135	5
2709	2.425	2	10745	NA	NA
2714	NA	NA	10753	NA	NA
2716	NA	NA	10754	1.865	2
2723	NA	NA	10762	2.015	2
2725	NA	NA	10766	NA	NA
2738	NA	NA	10776	NA	NA
2750	3.692	4	10783	2.349	2
2756	NA	NA	10789	2.751	3
2758	NA	NA	10790	4.672	5
2766	NA	NA	10797	NA	NA
2767	NA	NA	10807	2.755	3
2771	2.676	3	10810	NA	NA
2775	3.417	3	10817	2.879	3
2776	NA	NA	10820	2.445	2
2779	NA	NA	10822	3.308	3
2780	2.798	3	10828	3.962	4
2781	NA	NA	10829	NA	NA
2782	4.85	5	10830	3.06	3
2783	3.522	4	10831	6.34	6
2796	3.316	3	10841	5.233	5
2798	3.711	4	10847	NA	NA
2800	NA	NA	10856	NA	NA
2803	6.456	6	10860	NA	NA
2806	NA	NA	10861	NA	NA
2813	3.252	3	10863	2.674	3
2818	NA	NA	10875	NA	NA

Index	Value	Cases	Index	Value	Cases
2821	5.179	5	10884	4.52	5
2825	4.681	5	10895	NA	NA
2829	1.695	2	10897	2.64	3
2830	2.788	3	10898	NA	NA
2833	4.547	5	10903	NA	NA
2839	NA	NA	10908	NA	NA
2843	5.325	5	10924	2.237	2
2846	3.981	4	10926	1.996	2
2847	2.934	3	10927	2.681	3
2848	NA	NA	10928	2.458	2
2856	NA	NA	10933	NA	NA
2863	NA	NA	10939	5.963	6
2867	4.812	5	10942	3.215	3
2869	3.549	4	10945	3.33	3
2873	4.868	5	10949	3.738	4
2874	3.445	3	10950	2.737	3
2875	3.223	3	10958	5.282	5
2880	NA	NA	10963	3.741	4
2886	4.194	4	10967	3.695	4
2887	NA	NA	10971	NA	NA
2888	2.811	3	10972	NA	NA
2889	2.186	2	10974	3.453	3
2890	NA	NA	10976	5.528	6
2892	NA	NA	10980	2.589	3
2901	NA	NA	10991	NA	NA
2902	1.648	2	10995	NA	NA
2905	1.763	2	11014	4.992	5
2917	NA	NA	11017	NA	NA
2922	2.407	2	11019	4.07	4
2924	3.355	3	11022	NA	NA
2930	NA	NA	11030	NA	NA
2931	4.782	5	11031	3.308	3
2946	NA	NA	11041	NA	NA
2955	3.833	4	11042	3.539	4
2962	NA	NA	11044	NA	NA
2964	NA	NA	11047	NA	NA
2965	NA	NA	11048	NA	NA
2967	4.218	4	11049	2.866	3
2970	5.644	6	11052	2.884	3
2973	5.592	6	11058	NA	NA
2974	NA	NA	11069	3.215	3
2976	NA	NA	11070	NA	NA
2977	NA	NA	11073	NA	NA
2978	NA	NA	11074	NA	NA
2986	4.792	5	11078	NA	NA
2988	4.897	5	11079	NA	NA
2989	NA	NA	11085	NA	NA
2995	4.069	4	11088	3.332	3
3005	4.695	5	11106	NA	NA
3011	4.643	5	11110	NA	NA
3013	2.819	3	11114	3.958	4
3019	NA	NA	11118	2.258	2
3021	2.107	2	11129	4.294	4
3022	5.112	5	11130	3.5	4
3029	NA	NA	11131	3.302	3
3037	3.454	3	11133	NA	NA
3042	2.232	2	11138	4.448	4
3043	3.83	4	11143	3.622	4

Index	Value	Cases	Index	Value	Cases
3049	3.547	4	11146	5.5	5
3050	6.204	6	11153	3.789	4
3053	NA	NA	11162	3.32	3
3058	NA	NA	11170	6.209	6
3062	NA	NA	11171	NA	NA
3063	NA	NA	11201	NA	NA
3065	NA	NA	11216	NA	NA
3080	2.549	3	11219	NA	NA
3088	NA	NA	11222	4.231	4
3093	NA	NA	11234	NA	NA
3096	NA	NA	11238	3.428	3
3101	5.928	6	11244	3.322	3
3103	NA	NA	11246	NA	NA
3107	5.313	5	11248	NA	NA
3109	4.848	5	11250	NA	NA
3111	5.905	6	11256	3.877	4
3113	4.149	4	11259	1.892	2
3116	4.798	5	11263	NA	NA
3132	NA	NA	11264	NA	NA
3141	5.141	5	11270	NA	NA
3153	NA	NA	11274	NA	NA
3154	NA	NA	11281	NA	NA
3160	NA	NA	11285	NA	NA
3167	NA	NA	11300	1.682	2
3170	3.62	4	11305	2.73	3
3173	2.228	2	11317	4.212	4
3174	4.007	4	11319	2.193	2
3177	4.971	5	11330	NA	NA
3179	NA	NA	11334	4.695	5
3184	NA	NA	11335	NA	NA
3190	NA	NA	11336	2.641	3
3193	NA	NA	11356	4.561	5
3199	4.552	5	11358	2.726	3
3201	NA	NA	11360	NA	NA
3202	2.676	3	11364	NA	NA
3203	2.892	3	11373	6.215	6
3206	NA	NA	11379	4.479	4
3209	3.154	3	11382	NA	NA
3210	2.087	2	11383	NA	NA
3217	4.77	5	11385	4.41	4
3220	NA	NA	11387	NA	NA
3228	NA	NA	11391	2.679	3
3232	2.686	3	11397	2.199	2
3239	NA	NA	11404	1.643	2
3243	3.342	3	11405	NA	NA
3245	2.709	3	11409	3.679	4
3246	NA	NA	11419	3.056	3
3251	5.474	5	11430	4.78	5
3253	5.401	5	11434	5.937	6
3257	NA	NA	11436	2.101	2
3260	NA	NA	11440	NA	NA
3261	3.227	3	11443	NA	NA
3263	5.049	5	11449	2.988	3
3278	3.816	4	11452	NA	NA
3281	3.105	3	11453	NA	NA
3283	4.884	5	11456	3.79	4
3290	1.902	2	11457	NA	NA
3297	3.281	3	11459	3.823	4

Index	Value	Cases	Index	Value	Cases
3304	3.633	4	11471	NA	NA
3305	3.419	3	11476	3.072	3
3307	4.16	4	11479	NA	NA
3308	2.751	3	11481	3.471	3
3313	2.873	3	11485	2.797	3
3314	2.813	3	11486	NA	NA
3317	NA	NA	11487	2.329	2
3348	NA	NA	11488	1.793	2
3350	1.888	2	11498	NA	NA
3359	4.078	4	11506	2.56	3
3367	3.909	4	11511	NA	NA
3376	2.803	3	11515	NA	NA
3378	3.717	4	11518	NA	NA
3384	NA	NA	11521	NA	NA
3386	2.759	3	11523	NA	NA
3387	NA	NA	11524	NA	NA
3388	2.95	3	11525	3.44	3
3390	3.926	4	11528	NA	NA
3391	NA	NA	11530	2.428	2
3396	NA	NA	11531	3.115	3
3398	3.912	4	11533	2.524	3
3404	5.899	6	11535	3.583	4
3406	NA	NA	11537	3.563	4
3407	3.289	3	11538	2.935	3
3414	4.607	5	11541	NA	NA
3419	3.327	3	11548	4.196	4
3423	NA	NA	11552	2.539	3
3427	2.796	3	11558	NA	NA
3432	NA	NA	11560	NA	NA
3434	2.657	3	11566	NA	NA
3438	NA	NA	11572	NA	NA
3442	NA	NA	11573	5.118	5
3443	NA	NA	11582	4.292	4
3448	NA	NA	11586	2.256	2
3456	3.307	3	11590	4.02	4
3464	5.777	6	11591	2.296	2
3470	NA	NA	11601	3.885	4
3475	3.544	4	11611	NA	NA
3477	NA	NA	11617	NA	NA
3490	NA	NA	11619	3.747	4
3493	3.667	4	11624	3.546	4
3502	NA	NA	11626	6.09	6
3508	NA	NA	11644	1.998	2
3516	NA	NA	11652	NA	NA
3517	NA	NA	11656	NA	NA
3525	NA	NA	11658	3.551	4
3532	3.207	3	11659	5.152	5
3535	NA	NA	11663	NA	NA
3536	4.288	4	11665	NA	NA
3540	3.223	3	11683	4.656	5
3547	2.663	3	11685	NA	NA
3550	4.039	4	11691	2.66	3
3557	NA	NA	11694	3.171	3
3562	NA	NA	11698	NA	NA
3563	3.23	3	11700	3.439	3
3564	NA	NA	11703	2.13	2
3570	3.21	3	11705	2.217	2
3573	NA	NA	11710	NA	NA

Index	Value	Cases	Index	Value	Cases
3577	2.78	3	11711	3.166	3
3579	NA	NA	11714	NA	NA
3581	NA	NA	11731	NA	NA
3587	3.235	3	11732	NA	NA
3602	4.291	4	11742	NA	NA
3609	3.223	3	11744	3.667	4
3612	3.647	4	11745	2.881	3
3621	2.942	3	11749	NA	NA
3642	NA	NA	11756	3.024	3
3647	NA	NA	11761	NA	NA
3649	NA	NA	11762	4.684	5
3654	NA	NA	11766	4.851	5
3660	4.175	4	11767	6.886	7
3665	NA	NA	11769	2.408	2
3669	3.359	3	11770	3.764	4
3673	3.747	4	11771	4.222	4
3675	NA	NA	11777	3.288	3
3678	3.572	4	11778	4.932	5
3680	3.21	3	11779	NA	NA
3686	5.074	5	11788	NA	NA
3693	4.027	4	11790	3.936	4
3710	2.403	2	11794	3.566	4
3713	4.485	4	11801	NA	NA
3718	NA	NA	11807	NA	NA
3725	3.523	4	11812	3.988	4
3726	2.997	3	11817	NA	NA
3747	2.1	2	11818	NA	NA
3753	NA	NA	11825	NA	NA
3754	5.775	6	11828	NA	NA
3760	5.921	6	11833	4.702	5
3763	2.156	2	11837	NA	NA
3765	4.351	4	11838	NA	NA
3769	NA	NA	11842	NA	NA
3771	3.321	3	11853	4.455	4
3784	2.366	2	11857	NA	NA
3787	NA	NA	11858	1.7	2
3794	NA	NA	11860	4.929	5
3796	3.449	3	11867	NA	NA
3798	4.021	4	11868	5.617	6
3809	3.595	4	11871	NA	NA
3812	5.007	5	11875	NA	NA
3819	NA	NA	11881	5.405	5
3828	NA	NA	11890	NA	NA
3831	5.096	5	11892	NA	NA
3833	3.074	3	11894	2.91	3
3837	4.992	5	11896	NA	NA
3839	NA	NA	11903	3.694	4
3843	NA	NA	11905	4.31	4
3846	NA	NA	11907	2.491	2
3854	4.851	5	11909	6.291	6
3861	NA	NA	11911	2.754	3
3864	3.428	3	11915	NA	NA
3868	2.704	3	11918	NA	NA
3869	4.968	5	11920	5.688	6
3870	2.624	3	11923	4.493	4
3883	2.422	2	11924	2.831	3
3886	2.521	3	11926	1.718	2
3889	NA	NA	11931	NA	NA

Index	Value	Cases	Index	Value	Cases
3894	NA	NA	11933	4.528	5
3907	2.337	2	11940	4.385	4
3910	3.741	4	11951	2.997	3
3913	NA	NA	11953	2.326	2
3914	2.853	3	11973	NA	NA
3921	5.016	5	11984	NA	NA
3923	NA	NA	11985	NA	NA
3929	2.741	3	11991	2.51	3
3931	NA	NA	12002	2.782	3
3932	5.003	5	12006	NA	NA
3937	NA	NA	12008	4.878	5
3943	2.33	2	12013	NA	NA
3956	4.206	4	12015	4.436	4
3957	2.694	3	12016	4.743	5
3961	6.149	6	12023	NA	NA
3971	3.037	3	12029	NA	NA
4004	NA	NA	12036	NA	NA
4005	2.778	3	12038	3.247	3
4006	4.579	5	12041	NA	NA
4011	2.491	2	12049	NA	NA
4013	4.696	5	12050	NA	NA
4014	5.489	5	12054	3.184	3
4016	NA	NA	12060	2.935	3
4017	5.371	5	12062	5.288	5
4020	NA	NA	12065	2.349	2
4022	3.628	4	12079	NA	NA
4026	NA	NA	12083	NA	NA
4032	NA	NA	12090	4.442	4
4043	1.96	2	12091	3.973	4
4045	2.981	3	12094	4.594	5
4048	5.269	5	12099	3.762	4
4051	NA	NA	12101	2.797	3
4052	3.584	4	12110	2.831	3
4056	2.861	3	12116	NA	NA
4059	1.456	1	12122	4.468	4
4069	NA	NA	12127	6.468	6
4074	3.697	4	12133	3.465	3
4076	2.524	3	12142	NA	NA
4077	NA	NA	12147	NA	NA
4079	NA	NA	12156	2.316	2
4081	NA	NA	12157	4.881	5
4088	NA	NA	12158	NA	NA
4105	NA	NA	12161	4.403	4
4125	3.616	4	12163	4.312	4
4134	3.05	3	12166	2.253	2
4139	1.915	2	12170	NA	NA
4146	1.945	2	12174	NA	NA
4149	NA	NA	12183	NA	NA
4151	NA	NA	12188	NA	NA
4155	NA	NA	12189	4.776	5
4157	2.581	3	12192	NA	NA
4168	NA	NA	12201	3.643	4
4170	2.08	2	12204	NA	NA
4174	2.361	2	12207	NA	NA
4179	5.081	5	12208	1.775	2
4185	4.528	5	12209	NA	NA
4199	NA	NA	12210	6.273	6
4205	NA	NA	12217	3.144	3

Index	Value	Cases	Index	Value	Cases
4208	NA	NA	12227	NA	NA
4211	2.664	3	12231	NA	NA
4212	NA	NA	12232	3.665	4
4215	2.918	3	12239	5.289	5
4217	NA	NA	12240	2.03	2
4219	NA	NA	12251	3.144	3
4226	4.255	4	12256	4.051	4
4227	NA	NA	12261	NA	NA
4229	NA	NA	12263	2.496	2
4231	1.893	2	12266	NA	NA
4233	NA	NA	12267	2.782	3
4237	NA	NA	12268	3.496	3
4243	3.91	4	12279	NA	NA
4248	4.749	5	12280	NA	NA
4255	5.106	5	12283	2.346	2
4262	2.415	2	12284	NA	NA
4266	NA	NA	12285	NA	NA
4268	NA	NA	12286	4.978	5
4270	NA	NA	12292	4.266	4
4273	NA	NA	12295	NA	NA
4276	3.787	4	12301	NA	NA
4277	3.545	4	12314	2.167	2
4279	NA	NA	12315	NA	NA
4299	3.575	4	12318	NA	NA
4313	NA	NA	12332	5.095	5
4322	NA	NA	12334	NA	NA
4324	NA	NA	12337	3.376	3
4328	2.872	3	12338	4.789	5
4331	2.654	3	12349	4.857	5
4335	1.847	2	12350	4.24	4
4337	3.409	3	12359	5.836	6
4338	NA	NA	12360	5.264	5
4343	1.457	1	12373	NA	NA
4347	1.526	2	12374	NA	NA
4355	NA	NA	12380	3.621	4
4357	NA	NA	12382	3.308	3
4359	6.027	6	12383	3.253	3
4362	NA	NA	12390	3.867	4
4368	2.432	2	12398	4.005	4
4374	NA	NA	12405	3.766	4
4375	5.116	5	12407	2.557	3
4378	3.376	3	12410	6.258	6
4381	NA	NA	12418	5.358	5
4387	3.906	4	12421	NA	NA
4400	1.882	2	12422	2.99	3
4423	3.599	4	12439	NA	NA
4424	NA	NA	12444	3.787	4
4428	4.578	5	12463	5.903	6
4433	5.917	6	12465	NA	NA
4436	NA	NA	12470	NA	NA
4437	NA	NA	12471	3.686	4
4439	6.099	6	12480	4.19	4
4449	3.552	4	12482	4.542	5
4456	3.319	3	12484	2.997	3
4463	5.693	6	12487	4.504	5
4467	NA	NA	12491	NA	NA
4468	NA	NA	12503	NA	NA
4469	2.418	2	12507	NA	NA

Index	Value	Cases	Index	Value	Cases
4472	3.981	4	12526	2.227	2
4473	4.7	5	12533	2.769	3
4476	2.503	3	12540	NA	NA
4500	2.549	3	12543	2.291	2
4509	4.179	4	12552	NA	NA
4513	NA	NA	12555	5.521	6
4521	NA	NA	12556	4.302	4
4527	2.374	2	12570	4.581	5
4530	1.48	1	12579	3.208	3
4532	2.851	3	12588	2.194	2
4533	2.971	3	12600	NA	NA
4535	NA	NA	12615	5.308	5
4536	NA	NA	12624	3.686	4
4542	3.385	3	12629	NA	NA
4551	2.112	2	12634	2.403	2
4554	3.638	4	12638	NA	NA
4555	1.733	2	12646	2.616	3
4564	NA	NA	12650	NA	NA
4572	NA	NA	12665	NA	NA
4573	5.986	6	12674	NA	NA
4577	NA	NA	12676	NA	NA
4579	NA	NA	12678	NA	NA
4583	NA	NA	12685	2.717	3
4584	5.926	6	12690	3.011	3
4596	3.282	3	12698	4.277	4
4599	3.945	4	12702	3.833	4
4607	3.672	4	12704	2.73	3
4609	NA	NA	12705	2.255	2
4610	NA	NA	12710	4.219	4
4616	1.968	2	12715	4.996	5
4617	2.656	3	12720	3.253	3
4633	4.706	5	12734	2.115	2
4638	NA	NA	12744	3.737	4
4641	NA	NA	12747	NA	NA
4653	6.367	6	12757	4.81	5
4655	4.558	5	12758	NA	NA
4659	NA	NA	12766	NA	NA
4669	NA	NA	12782	3.162	3
4678	NA	NA	12787	NA	NA
4685	3.838	4	12799	2.12	2
4686	NA	NA	12804	4.351	4
4691	NA	NA	12809	2.782	3
4695	4.144	4	12813	3.007	3
4698	3.726	4	12816	3.318	3
4700	5.607	6	12821	NA	NA
4711	3.5	3	12826	2.798	3
4722	NA	NA	12831	3.467	3
4727	4.24	4	12832	3.333	3
4756	6.042	6	12833	3.304	3
4762	NA	NA	12835	3.564	4
4763	NA	NA	12842	NA	NA
4766	5.112	5	12844	3.681	4
4770	NA	NA	12847	NA	NA
4784	4.238	4	12852	NA	NA
4791	2.899	3	12856	3.517	4
4795	5.063	5	12857	2.99	3
4799	NA	NA	12858	NA	NA
4802	4.477	4	12861	2.49	2

Index	Value	Cases	Index	Value	Cases
4805	3.441	3	12869	3.617	4
4814	4.044	4	12876	3.912	4
4816	NA	NA	12877	NA	NA
4817	3.766	4	12879	3.058	3
4822	2.775	3	12882	3.866	4
4827	2.313	2	12883	NA	NA
4833	NA	NA	12887	2.264	2
4836	NA	NA	12889	NA	NA
4842	NA	NA	12891	5.361	5
4844	NA	NA	12894	4.159	4
4845	2.489	2	12895	NA	NA
4849	4.606	5	12899	NA	NA
4850	NA	NA	12905	6.349	6
4860	4.654	5	12913	NA	NA
4863	2.293	2	12916	NA	NA
4871	3.773	4	12917	2.69	3
4878	3.422	3	12925	3.96	4
4881	2.531	3	12934	5.296	5
4888	NA	NA	12939	4.489	4
4900	5.915	6	12943	NA	NA
4906	3.316	3	12950	NA	NA
4909	NA	NA	12961	NA	NA
4916	3.75	4	12963	2.441	2
4918	5.537	6	12973	NA	NA
4926	NA	NA	12979	3.655	4
4928	NA	NA	12980	NA	NA
4941	2.221	2	12981	NA	NA
4946	5.043	5	12982	2.298	2
4949	NA	NA	12992	2.654	3
4956	2.532	3	12994	NA	NA
4966	4.445	4	12999	2.644	3
4969	2.268	2	13002	3.714	4
4973	3.796	4	13004	NA	NA
4978	4.831	5	13010	NA	NA
4982	2.664	3	13013	3.226	3
4985	4.11	4	13015	4.038	4
4991	3.05	3	13019	4.494	4
4998	3.198	3	13030	NA	NA
5000	2.325	2	13031	NA	NA
5004	NA	NA	13036	3.1	3
5005	2.047	2	13037	4.776	5
5011	3.927	4	13042	NA	NA
5016	NA	NA	13054	2.411	2
5018	3.826	4	13060	NA	NA
5034	3.761	4	13072	NA	NA
5038	NA	NA	13073	NA	NA
5042	NA	NA	13079	5.126	5
5046	2.682	3	13081	NA	NA
5051	NA	NA	13086	NA	NA
5054	NA	NA	13087	3.59	4
5057	4.673	5	13090	NA	NA
5062	3.608	4	13098	2.704	3
5063	NA	NA	13100	2.603	3
5065	NA	NA	13105	NA	NA
5066	NA	NA	13106	2.815	3
5076	2.394	2	13107	3.535	4
5089	3.256	3	13113	6.413	6
5092	2.755	3	13115	5.048	5

Index	Value	Cases	Index	Value	Cases
5093	4.646	5	13117	NA	NA
5094	4.323	4	13118	NA	NA
5098	3.589	4	13121	NA	NA
5102	3.68	4	13137	1.688	2
5112	4.703	5	13146	NA	NA
5117	2.379	2	13150	NA	NA
5127	NA	NA	13151	3.625	4
5130	2.451	2	13152	NA	NA
5131	1.827	2	13156	NA	NA
5132	NA	NA	13165	5.862	6
5135	NA	NA	13169	3.633	4
5136	NA	NA	13178	3.776	4
5147	4.337	4	13180	NA	NA
5157	3.661	4	13183	4.496	4
5160	1.983	2	13184	NA	NA
5165	NA	NA	13188	3.508	4
5166	NA	NA	13191	3.621	4
5172	2.259	2	13196	NA	NA
5173	2.486	2	13203	NA	NA
5179	NA	NA	13206	2.35	2
5184	5.06	5	13211	NA	NA
5187	NA	NA	13219	NA	NA
5191	2.98	3	13223	5.193	5
5193	NA	NA	13226	3.598	4
5194	NA	NA	13228	2.293	2
5199	NA	NA	13230	4.222	4
5212	NA	NA	13240	3.887	4
5213	2.724	3	13249	5.256	5
5224	3.359	3	13250	NA	NA
5226	3.7	4	13256	3.546	4
5239	4.724	5	13261	NA	NA
5252	NA	NA	13263	2.239	2
5264	NA	NA	13268	4.139	4
5266	NA	NA	13275	NA	NA
5271	4.853	5	13277	NA	NA
5273	3.75	4	13283	3.661	4
5276	3.182	3	13284	NA	NA
5278	4.7	5	13285	2.314	2
5281	1.794	2	13286	NA	NA
5283	4.047	4	13287	NA	NA
5291	NA	NA	13290	2.719	3
5294	5.629	6	13291	NA	NA
5296	3.611	4	13294	4.889	5
5297	NA	NA	13295	2.896	3
5313	3.64	4	13303	3.272	3
5314	2.604	3	13306	NA	NA
5321	NA	NA	13311	4.321	4
5325	3.136	3	13322	NA	NA
5326	NA	NA	13331	NA	NA
5328	NA	NA	13337	4.453	4
5334	2.665	3	13344	NA	NA
5338	3.943	4	13362	4.443	4
5344	1.274	1	13364	1.984	2
5348	NA	NA	13366	NA	NA
5352	NA	NA	13368	3.319	3
5353	3.847	4	13370	1.879	2
5354	NA	NA	13377	4.927	5
5361	NA	NA	13378	2.039	2

Index	Value	Cases	Index	Value	Cases
5364	2.809	3	13388	NA	NA
5365	3.734	4	13392	3.914	4
5367	NA	NA	13398	4.04	4
5379	4.163	4	13403	NA	NA
5382	2.984	3	13404	3.975	4
5386	NA	NA	13409	3.628	4
5395	3.875	4	13416	NA	NA
5410	NA	NA	13422	NA	NA
5411	2.909	3	13427	3.483	3
5416	4.601	5	13433	5.197	5
5424	3.585	4	13438	2.55	3
5426	2.392	2	13441	NA	NA
5428	NA	NA	13449	5.125	5
5430	NA	NA	13450	3.081	3
5433	NA	NA	13453	2.075	2
5437	3.136	3	13460	2.77	3
5440	NA	NA	13461	4.294	4
5442	4.873	5	13465	3.375	3
5445	2.817	3	13468	NA	NA
5449	NA	NA	13481	3.438	3
5452	3.624	4	13485	1.965	2
5460	NA	NA	13487	NA	NA
5461	1.914	2	13490	NA	NA
5465	NA	NA	13493	2.767	3
5467	NA	NA	13497	2.288	2
5471	4.029	4	13508	NA	NA
5474	NA	NA	13516	NA	NA
5475	NA	NA	13525	3.063	3
5480	NA	NA	13533	NA	NA
5481	3.469	3	13535	NA	NA
5484	NA	NA	13538	3.355	3
5494	4.723	5	13545	NA	NA
5495	NA	NA	13566	4.051	4
5497	NA	NA	13581	2.344	2
5499	3.141	3	13584	NA	NA
5507	NA	NA	13588	NA	NA
5510	3.129	3	13596	2.483	2
5515	NA	NA	13600	5.606	6
5516	1.963	2	13604	3.419	3
5517	NA	NA	13608	2.859	3
5524	4.335	4	13611	2.642	3
5530	NA	NA	13612	NA	NA
5534	2.718	3	13615	3.527	4
5543	NA	NA	13616	4.241	4
5545	2.672	3	13618	3.443	3
5558	3.641	4	13625	NA	NA
5562	NA	NA	13628	3.925	4
5573	6.491	6	13629	NA	NA
5581	4.476	4	13630	5.147	5
5583	5.306	5	13633	3.363	3
5587	NA	NA	13637	3.812	4
5589	1.811	2	13640	NA	NA
5591	5.014	5	13641	2.176	2
5596	2.571	3	13651	NA	NA
5606	4.381	4	13674	3.532	4
5608	4.235	4	13684	3.674	4
5611	3.195	3	13690	NA	NA
5612	4.162	4	13707	3.449	3

Index	Value	Cases	Index	Value	Cases
5614	NA	NA	13709	5.146	5
5620	3.82	4	13710	NA	NA
5623	4.958	5	13713	4.346	4
5624	2.822	3	13724	NA	NA
5626	5.015	5	13725	NA	NA
5633	NA	NA	13731	2.051	2
5635	2.775	3	13736	1.752	2
5640	NA	NA	13740	NA	NA
5643	NA	NA	13745	NA	NA
5644	6.032	6	13748	NA	NA
5653	5.076	5	13751	NA	NA
5663	NA	NA	13758	NA	NA
5664	4.777	5	13762	NA	NA
5667	2.12	2	13764	4.912	5
5671	NA	NA	13765	NA	NA
5673	3.549	4	13769	NA	NA
5676	3.074	3	13770	2.144	2
5678	1.933	2	13774	5.169	5
5698	2.038	2	13787	2.428	2
5700	6.09	6	13791	1.871	2
5705	3.425	3	13802	1.67	2
5706	NA	NA	13807	2.4	2
5711	NA	NA	13808	1.678	2
5712	4.966	5	13809	NA	NA
5716	2.721	3	13810	NA	NA
5719	2.791	3	13822	NA	NA
5725	NA	NA	13823	3.368	3
5728	6.47	6	13825	3.475	3
5734	NA	NA	13826	NA	NA
5735	4.64	5	13833	2.959	3
5743	2.53	3	13837	2.823	3
5754	NA	NA	13842	3.214	3
5755	NA	NA	13846	2.257	2
5756	NA	NA	13852	4.347	4
5766	2.994	3	13853	NA	NA
5770	3.642	4	13858	NA	NA
5774	NA	NA	13860	NA	NA
5775	2.118	2	13866	NA	NA
5776	3.616	4	13886	NA	NA
5778	6.05	6	13887	NA	NA
5786	2.969	3	13890	3.452	3
5787	3.49	3	13891	3.448	3
5791	4.796	5	13893	2.024	2
5794	NA	NA	13902	5.34	5
5803	2.99	3	13903	NA	NA
5804	NA	NA	13908	3.276	3
5808	NA	NA	13912	1.94	2
5810	3.506	4	13924	NA	NA
5813	3.545	4	13928	2.431	2
5828	3.501	4	13929	3.246	3
5839	6.058	6	13938	4.139	4
5842	4.343	4	13939	NA	NA
5843	1.868	2	13941	4.668	5
5844	3.437	3	13951	2.173	2
5847	2.305	2	13962	NA	NA
5851	2.112	2	13964	2.111	2
5854	NA	NA	13967	2.928	3
5857	NA	NA	13971	4.158	4

Index	Value	Cases	Index	Value	Cases
5866	NA	NA	13972	4.63	5
5874	NA	NA	13975	NA	NA
5886	NA	NA	13977	2.428	2
5895	1.559	2	13979	4.687	5
5897	2.53	3	13983	NA	NA
5898	2.529	3	13984	2.948	3
5900	NA	NA	13987	NA	NA
5902	NA	NA	13994	NA	NA
5908	NA	NA	13999	2.683	3
5909	3.095	3	14003	4.855	5
5912	6.169	6	14008	4.257	4
5913	NA	NA	14011	NA	NA
5917	5.512	6	14012	2.326	2
5918	2.736	3	14016	3.212	3
5921	NA	NA	14017	NA	NA
5931	NA	NA	14020	NA	NA
5942	3.907	4	14027	NA	NA
5943	NA	NA	14038	3.23	3
5950	NA	NA	14040	5.073	5
5954	NA	NA	14042	2.427	2
5983	NA	NA	14055	NA	NA
5995	NA	NA	14057	4	4
6002	2.263	2	14060	NA	NA
6005	3.706	4	14081	NA	NA
6009	5.975	6	14091	3.758	4
6011	NA	NA	14111	3.369	3
6012	5.325	5	14117	5.082	5
6019	NA	NA	14121	NA	NA
6021	2.897	3	14122	NA	NA
6029	NA	NA	14125	NA	NA
6036	NA	NA	14129	NA	NA
6037	2.758	3	14135	NA	NA
6038	NA	NA	14148	NA	NA
6043	1.976	2	14157	3.684	4
6045	3.979	4	14161	NA	NA
6047	NA	NA	14163	NA	NA
6048	2.75	3	14172	NA	NA
6061	4.051	4	14180	NA	NA
6063	2.79	3	14182	NA	NA
6064	4.849	5	14188	NA	NA
6068	4.642	5	14191	2.815	3
6069	NA	NA	14201	6.205	6
6070	3.641	4	14202	3.556	4
6071	5.252	5	14213	NA	NA
6074	NA	NA	14220	NA	NA
6079	3.382	3	14224	1.898	2
6082	NA	NA	14231	NA	NA
6088	4.708	5	14241	4.914	5
6094	2.399	2	14243	NA	NA
6095	NA	NA	14245	2.47	2
6098	2.485	2	14247	NA	NA
6102	NA	NA	14248	2.621	3
6105	NA	NA	14252	2.684	3
6113	NA	NA	14254	4.592	5
6116	2.53	3	14260	4.109	4
6120	NA	NA	14269	3.802	4
6121	2.232	2	14272	NA	NA
6126	2.811	3	14274	3.306	3

Index	Value	Cases	Index	Value	Cases
6144	2.953	3	14279	2.966	3
6145	3.733	4	14280	5.256	5
6153	2.147	2	14290	3.476	3
6156	NA	NA	14298	NA	NA
6159	4.167	4	14308	4.633	5
6162	NA	NA	14313	3.833	4
6184	2.399	2	14316	3.779	4
6188	2.513	3	14319	NA	NA
6189	3.107	3	14322	NA	NA
6191	NA	NA	14323	NA	NA
6211	NA	NA	14325	4.99	5
6216	3.406	3	14337	4.291	4
6218	NA	NA	14339	4.351	4
6222	NA	NA	14341	2.727	3
6235	3.251	3	14342	4.621	5
6245	NA	NA	14346	6.041	6
6248	3.412	3	14351	3.325	3
6253	3.229	3	14354	NA	NA
6256	NA	NA	14355	NA	NA
6257	3.598	4	14358	5.341	5
6259	NA	NA	14359	NA	NA
6266	4.174	4	14364	3.525	4
6268	NA	NA	14374	NA	NA
6275	NA	NA	14376	NA	NA
6280	2.844	3	14382	2.467	2
6283	NA	NA	14384	2.763	3
6288	NA	NA	14393	NA	NA
6289	2.069	2	14398	2.832	3
6301	4.193	4	14403	4.586	5
6308	3.56	4	14406	NA	NA
6314	NA	NA	14408	NA	NA
6315	NA	NA	14411	NA	NA
6316	3.656	4	14414	4.187	4
6317	NA	NA	14418	NA	NA
6318	3.659	4	14423	NA	NA
6323	4.22	4	14442	4.415	4
6329	NA	NA	14443	3.602	4
6336	NA	NA	14444	3.097	3
6341	4.308	4	14446	NA	NA
6348	4.056	4	14455	2.806	3
6349	NA	NA	14456	5.178	5
6365	NA	NA	14458	3.433	3
6372	3.081	3	14464	NA	NA
6376	NA	NA	14466	4.603	5
6378	NA	NA	14467	2.578	3
6379	NA	NA	14469	3.901	4
6382	NA	NA	14483	NA	NA
6383	NA	NA	14484	2.731	3
6389	4.751	5	14490	2.544	3
6390	NA	NA	14491	2.505	3
6392	5.127	5	14494	2.9	3
6394	NA	NA	14496	3.795	4
6402	NA	NA	14503	NA	NA
6404	3.578	4	14504	2.48	2
6405	2.407	2	14505	3.162	3
6406	NA	NA	14506	NA	NA
6409	NA	NA	14507	NA	NA
6410	4.821	5	14512	NA	NA

Index	Value	Cases	Index	Value	Cases
6411	3.703	4	14520	4.164	4
6421	4.397	4	14527	1.812	2
6428	6.117	6	14531	4.708	5
6429	2.196	2	14532	NA	NA
6432	4.356	4	14535	2.057	2
6436	NA	NA	14543	NA	NA
6437	3.797	4	14554	NA	NA
6438	NA	NA	14556	5.6	6
6445	2.701	3	14557	3.697	4
6447	NA	NA	14561	1.982	2
6450	3.677	4	14562	NA	NA
6462	3.901	4	14567	NA	NA
6467	3.679	4	14568	NA	NA
6478	3.546	4	14574	4.286	4
6484	3.959	4	14575	2.583	3
6492	3.814	4	14579	3.826	4
6497	NA	NA	14581	3.446	3
6504	2.948	3	14582	NA	NA
6505	NA	NA	14586	NA	NA
6513	3.074	3	14591	2.33	2
6525	5.651	6	14598	2.772	3
6526	NA	NA	14599	NA	NA
6528	2.614	3	14600	NA	NA
6540	NA	NA	14612	5.519	6
6542	NA	NA	14613	NA	NA
6544	5.068	5	14624	4.549	5
6548	3.687	4	14626	2.511	3
6552	NA	NA	14630	4.777	5
6558	3.426	3	14633	2.425	2
6567	3.186	3	14639	NA	NA
6569	4.693	5	14642	NA	NA
6572	5.765	6	14643	5.268	5
6577	5.216	5	14649	2.493	2
6581	3.397	3	14650	NA	NA
6588	3.021	3	14653	2.254	2
6591	NA	NA	14655	1.972	2
6594	2.335	2	14656	2.178	2
6600	5.098	5	14662	NA	NA
6602	NA	NA	14663	NA	NA
6604	2.438	2	14673	2.519	3
6605	3.423	3	14674	2.221	2
6614	NA	NA	14676	3.586	4
6616	NA	NA	14682	NA	NA
6621	NA	NA	14685	5.636	6
6640	5.235	5	14689	2.439	2
6641	NA	NA	14693	NA	NA
6643	2.214	2	14697	2.51	3
6644	NA	NA	14700	2.937	3
6649	2.234	2	14704	NA	NA
6650	NA	NA	14710	3.502	4
6655	6.031	6	14719	2.867	3
6661	1.519	2	14724	4.341	4
6672	5.674	6	14728	4.572	5
6677	NA	NA	14735	NA	NA
6688	2.079	2	14736	NA	NA
6689	NA	NA	14741	NA	NA
6691	NA	NA	14744	1.691	2
6692	4.3	4	14753	NA	NA

Index	Value	Cases	Index	Value	Cases
6694	3.963	4	14756	4.574	5
6702	NA	NA	14762	4.92	5
6714	NA	NA	14765	5.365	5
6716	4.894	5	14783	5.793	6
6724	3.347	3	14784	2.155	2
6725	NA	NA	14786	NA	NA
6730	3.327	3	14790	1.156	1
6735	2.671	3	14793	3.112	3
6738	2.767	3	14796	6.237	6
6739	2.923	3	14801	NA	NA
6743	3.605	4	14807	NA	NA
6747	2.461	2	14812	4.001	4
6750	6.53	7	14815	5.086	5
6751	NA	NA	14831	4.611	5
6753	4.597	5	14833	5.93	6
6754	3.081	3	14836	6.192	6
6755	NA	NA	14856	4.699	5
6762	3.385	3	14859	NA	NA
6764	3.42	3	14861	3.06	3
6772	NA	NA	14863	NA	NA
6774	NA	NA	14865	NA	NA
6787	2.931	3	14880	NA	NA
6789	2.953	3	14881	3.859	4
6793	4.609	5	14883	NA	NA
6798	NA	NA	14884	NA	NA
6799	NA	NA	14894	3.73	4
6800	4.021	4	14896	4.424	4
6802	NA	NA	14899	3.382	3
6808	NA	NA	14900	3.511	4
6809	4.007	4	14901	NA	NA
6812	NA	NA	14906	3.721	4
6814	NA	NA	14907	NA	NA
6816	2.35	2	14915	4.861	5
6822	NA	NA	14919	NA	NA
6829	NA	NA	14926	5.784	6
6834	2.372	2	14927	2.124	2
6836	5.067	5	14933	3.216	3
6839	NA	NA	14937	3.083	3
6840	NA	NA	14939	NA	NA
6843	NA	NA	14940	NA	NA
6846	1.741	2	14943	NA	NA
6848	NA	NA	14953	3.331	3
6852	NA	NA	14954	2.297	2
6856	5.766	6	14969	3.312	3
6860	4.078	4	14999	NA	NA
6866	4.887	5	15008	3.62	4
6870	4.414	4	15009	NA	NA
6878	NA	NA	15018	3.951	4
6880	4.941	5	15023	3.635	4
6885	NA	NA	15025	NA	NA
6897	4.657	5	15034	NA	NA
6902	4.389	4	15036	4.546	5
6904	4.623	5	15051	NA	NA
6907	4.438	4	15052	3.293	3
6909	NA	NA	15064	3.275	3
6914	1.868	2	15070	3.208	3
6915	4.277	4	15074	3.843	4
6922	NA	NA	15077	4.189	4

Index	Value	Cases	Index	Value	Cases
6924	1.973	2	15081	3.586	4
6933	2.916	3	15086	5.982	6
6934	4.096	4	15093	NA	NA
6941	3.891	4	15094	NA	NA
6957	NA	NA	15103	2.061	2
6960	NA	NA	15104	NA	NA
6969	NA	NA	15110	NA	NA
6975	2.33	2	15112	1.844	2
6980	NA	NA	15115	NA	NA
6983	NA	NA	15131	NA	NA
6987	2.835	3	15139	3.624	4
6994	2.552	3	15141	3	3
6997	4.313	4	15148	NA	NA
7002	3.088	3	15154	NA	NA
7010	2.123	2	15156	3.26	3
7015	NA	NA	15161	1.787	2
7019	NA	NA	15167	2.834	3
7022	NA	NA	15178	NA	NA
7025	2.225	2	15205	4.87	5
7029	3.811	4	15207	NA	NA
7031	2.364	2	15222	3.316	3
7037	NA	NA	15223	4.789	5
7038	3.391	3	15225	4.337	4
7043	NA	NA	15228	NA	NA
7049	3.134	3	15239	NA	NA
7052	3.004	3	15241	NA	NA
7053	4.756	5	15246	NA	NA
7056	3.286	3	15247	NA	NA
7057	4.732	5	15249	2.878	3
7080	NA	NA	15255	NA	NA
7086	NA	NA	15257	NA	NA
7087	3.22	3	15267	NA	NA
7105	NA	NA	15277	3.053	3
7108	NA	NA	15280	4.88	5
7121	NA	NA	15289	NA	NA
7122	1.985	2	15297	NA	NA
7125	NA	NA	15302	NA	NA
7132	3.112	3	15304	NA	NA
7134	3.195	3	15312	NA	NA
7151	NA	NA	15321	NA	NA
7152	4.135	4	15325	NA	NA
7157	2.814	3	15326	4.116	4
7159	NA	NA	15333	NA	NA
7166	NA	NA	15337	NA	NA
7167	NA	NA	15338	3.458	3
7177	2.82	3	15340	4.223	4
7179	NA	NA	15342	NA	NA
7181	NA	NA	15344	3.427	3
7183	4.227	4	15347	NA	NA
7186	3.913	4	15349	NA	NA
7193	5.089	5	15355	2.855	3
7205	NA	NA	15359	NA	NA
7207	NA	NA	15366	NA	NA
7209	3.976	4	15367	NA	NA
7216	NA	NA	15368	NA	NA
7232	NA	NA	15369	NA	NA
7235	NA	NA	15380	3.65	4
7238	2.27	2	15381	3.255	3

Index	Value	Cases	Index	Value	Cases
7240	2.17	2	15387	NA	NA
7243	NA	NA	15388	2.704	3
7252	2.232	2	15389	NA	NA
7269	2.253	2	15392	NA	NA
7275	NA	NA	15400	2.973	3
7281	NA	NA	15405	2.91	3
7283	NA	NA	15407	3.343	3
7287	5.541	6	15408	6.255	6
7289	5.194	5	15411	2.668	3
7291	NA	NA	15413	4.501	5
7294	NA	NA	15418	5.599	6
7304	5.325	5	15419	4.856	5
7308	3.547	4	15421	2.568	3
7313	3.058	3	15425	2.598	3
7319	NA	NA	15436	3.668	4
7325	NA	NA	15438	4.223	4
7326	2.381	2	15440	4.268	4
7330	5.075	5	15443	3.767	4
7332	3.85	4	15460	2.311	2
7337	NA	NA	15464	NA	NA
7341	NA	NA	15465	2.94	3
7346	4.116	4	15473	NA	NA
7353	NA	NA	15475	2.697	3
7354	4.468	4	15483	NA	NA
7361	NA	NA	15494	5.562	6
7366	NA	NA	15495	NA	NA
7368	5.131	5	15498	4.245	4
7372	3.978	4	15499	2.567	3
7375	3.027	3	15500	NA	NA
7377	5.353	5	15501	NA	NA
7380	NA	NA	15510	1.822	2
7382	NA	NA	15512	2.149	2
7385	2.494	2	15516	2.434	2
7392	2.22	2	15518	NA	NA
7395	3.9	4	15519	3.501	4
7397	3.792	4	15524	2.989	3
7403	3.971	4	15527	NA	NA
7406	NA	NA	15529	NA	NA
7409	NA	NA	15530	NA	NA
7410	3.669	4	15538	NA	NA
7412	3.758	4	15539	2.637	3
7419	3.107	3	15541	NA	NA
7425	3.403	3	15546	NA	NA
7435	6.473	6	15547	NA	NA
7438	3.722	4	15548	NA	NA
7440	3.765	4	15552	1.638	2
7447	NA	NA	15556	4.819	5
7449	NA	NA	15567	NA	NA
7456	4.464	4	15572	2.133	2
7464	NA	NA	15573	3.801	4
7478	NA	NA	15574	3.273	3
7480	NA	NA	15577	NA	NA
7481	5.255	5	15579	3.432	3
7483	3.134	3	15581	2.309	2
7484	NA	NA	15589	NA	NA
7491	4.258	4	15596	2.042	2
7494	1.88	2	15598	4.637	5
7501	NA	NA	15599	4.47	4

Index	Value	Cases	Index	Value	Cases
7503	NA	NA	15605	5.252	5
7509	2.034	2	15606	4.895	5
7517	NA	NA	15608	3.876	4
7518	6.091	6	15616	4.1	4
7519	1.75	2	15618	NA	NA
7521	NA	NA	15621	NA	NA
7522	NA	NA	15626	NA	NA
7536	3.725	4	15638	2.535	3
7539	NA	NA	15639	NA	NA
7547	5.376	5	15642	1.971	2
7549	2.913	3	15644	NA	NA
7552	NA	NA	15646	NA	NA
7554	NA	NA	15649	3.781	4
7556	3.149	3	15656	NA	NA
7564	NA	NA	15659	NA	NA
7566	NA	NA	15680	NA	NA
7570	4.736	5	15686	4.613	5
7571	2.913	3	15693	NA	NA
7572	3.563	4	15697	4.822	5
7575	1.858	2	15699	5.928	6
7586	2.787	3	15701	4.199	4
7589	4.927	5	15705	NA	NA
7590	2.544	3	15714	2.296	2
7597	3.622	4	15722	4.669	5
7602	3.817	4	15728	NA	NA
7604	4.504	5	15734	3.359	3
7605	2.643	3	15752	1.621	2
7612	1.894	2	15756	3.773	4
7615	4.443	4	15760	3.835	4
7617	NA	NA	15762	4.448	4
7624	3.406	3	15767	4.432	4
7632	4.66	5	15768	NA	NA
7639	3.488	3	15773	NA	NA
7642	2.889	3	15774	NA	NA
7643	3.23	3	15781	NA	NA
7649	3.265	3	15782	2.605	3
7650	NA	NA	15784	6.786	7
7653	3.893	4	15791	NA	NA
7654	NA	NA	15796	4.123	4
7657	5.138	5	15798	4.95	5
7662	NA	NA	15806	NA	NA
7669	3.767	4	15814	NA	NA
7671	2.399	2	15819	2.8	3
7675	NA	NA	15825	3.608	4
7678	4.691	5	15826	3.97	4
7682	NA	NA	15831	NA	NA
7688	NA	NA	15835	6.02	6
7689	NA	NA	15836	NA	NA
7690	4.747	5	15839	5.61	6
7692	2.234	2	15845	2.896	3
7699	NA	NA	15858	NA	NA
7705	3.421	3	15859	3.662	4
7712	NA	NA	15876	4.059	4
7726	4.02	4	15878	NA	NA
7728	NA	NA	15880	NA	NA
7735	3.127	3	15886	NA	NA
7737	2.248	2	15888	NA	NA
7739	3.346	3	15891	3.131	3

Index	Value	Cases	Index	Value	Cases
7743	NA	NA	15900	NA	NA
7744	NA	NA	15902	4.354	4
7746	2.108	2	15904	2.467	2
7749	NA	NA	15908	NA	NA
7750	3.999	4	15910	NA	NA
7752	3.785	4	15917	3.459	3
7755	4.918	5	15919	4.883	5
7756	NA	NA	15924	2.717	3
7762	NA	NA	15927	2.65	3
7764	NA	NA	15937	NA	NA
7769	2.227	2	15946	3.506	4
7770	2.046	2	15949	3.104	3
7776	5.055	5	15957	4.061	4
7778	5.205	5	15961	4.402	4
7784	2.283	2	15964	NA	NA
7786	NA	NA	15965	3.541	4
7789	3.545	4	15966	NA	NA
7793	3.639	4	15978	NA	NA
7794	NA	NA	15983	NA	NA
7804	NA	NA	15987	NA	NA
7811	NA	NA	15988	2.465	2
7813	2.62	3	15998	2.471	2
7815	2.574	3	16004	2.976	3
7817	NA	NA	16008	4.228	4
7818	NA	NA	16011	NA	NA
7821	6.062	6	16023	2.675	3
7825	4.538	5	16024	NA	NA
7830	5.152	5	16025	2.417	2
7832	NA	NA	16048	NA	NA
7835	NA	NA	16050	NA	NA
7839	5.572	6	16051	NA	NA
7842	5.141	5	16057	NA	NA
7849	4.628	5	16059	4.913	5
7856	NA	NA	16060	3.235	3
7857	2.432	2	16075	4.858	5
7863	3.929	4	16094	5.149	5
7866	4.839	5	16096	4.788	5
7871	3.419	3	16116	NA	NA
7875	NA	NA	16118	NA	NA
7882	1.862	2	16121	2.91	3
7887	NA	NA	16122	3.535	4
7888	4.515	5	16124	5.793	6
7891	1.687	2	16125	3.655	4
7895	NA	NA	16126	NA	NA
7901	NA	NA	16130	4.58	5
7906	2.257	NA	NA	NA	NA

Appendix B - Code

```
library(knitr)
library(dplyr)
library(tidyr)
library(VIM)
library(MASS)
library(caret)
library(pROC)
library(grid)
library(leaps)
library(vcd)
library(glmnet)
library(e1071)
library(car)
library(pander)
library(ggplot2)
library(reshape2)
library(corrplot)
library(gridExtra)
opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE,
    comment = NA, fig.align = "center")

train_url = "https://raw.githubusercontent.com/hovig/MSDS_CUNY/master/DATA621/hw5/wine-training-data.csv"
train = read.csv(train_url)
train <- train[-1]
eval_url = "https://raw.githubusercontent.com/hovig/MSDS_CUNY/master/DATA621/hw5/wine-evaluation-data.csv"
test = read.csv(eval_url)

# summary
train_means <- sapply(train, function(y) mean(y, na.rm = TRUE))
train_mins <- sapply(train, function(y) min(y, na.rm = TRUE))
train_medians <- sapply(train, function(y) median(y, na.rm = TRUE))
train_maxs <- sapply(train, function(y) max(y, na.rm = TRUE))
train_IQRs <- sapply(train, function(y) IQR(y, na.rm = TRUE))
train_SDs <- sapply(train, function(y) sd(y, na.rm = T))
train_skews <- sapply(train, function(y) skewness(y, na.rm = TRUE))
train_cors <- as.vector(cor(train$TARGET, train[, 1:ncol(train)],
    use = "complete.obs"))
train_NAs <- sapply(train, function(y) sum(length(which(is.na(y)))))

d_sum <- data.frame(train_means, train_mins, train_medians, train_maxs,
    train_IQRs, train_SDs, train_skews, train_cors, train_NAs)
colnames(d_sum) <- c("MEAN", "MIN", "MEDIAN", "MAX", "IQR", "STD. DEV",
    "SKEW", "$r_{TARGET}$", "NAs")
d_sum <- round(d_sum, 2)
pander(d_sum)

summary_metrics <- function(df) {
    metrics_only <- df[, sapply(df, is.numeric)]
    df_metrics <- psych::describe(metrics_only, quant = c(0.25,
        0.75))
    df_metrics$unique_values = rapply(metrics_only, function(x) length(unique(x)))
    df_metrics <- dplyr::select(df_metrics, n, unique_values,
        min, Q.1st = Q0.25, median, mean, Q.3rd = Q0.75, max,
        range, sd, skew, kurtosis)
    return(df_metrics)
}
```

```
metrics_df <- summary_metrics(train)
discrete_vars_freq <- train %>% dplyr::select(rownames(metrics_df)[metrics_df$unique_values <
    15]) %>% gather("var", "value") %>% group_by(var) %>% count(var,
    value) %>% mutate(prop = prop.table(n))

ggplot(data = discrete_vars_freq, aes(x = reorder(value, prop),
    y = prop)) + geom_bar(stat = "identity", fill = "darkblue") +
    facet_wrap(~var, scales = "free") + coord_flip() + ggthemes::theme_fivethirtyeight()

# variable plots
cont_vars <- train %>% dplyr::select(-c(TARGET, LabelAppeal,
    AcidIndex, STARS))
melted <- melt(cont_vars)
ggplot(melted, aes(value)) + geom_bar(aes(fill = variable, col = variable),
    alpha = 0.5, show.legend = FALSE) + facet_wrap(~variable,
    scale = "free") + ggtitle("Distribution of Continuous Variables \n")
poisson_vars <- train %>% dplyr::select(c(TARGET, LabelAppeal,
    AcidIndex, STARS))
melted <- melt(poisson_vars)
ggplot(melted, aes(value)) + geom_bar(aes(fill = variable, col = variable),
    alpha = 0.5, show.legend = FALSE) + facet_wrap(~variable,
    scales = "free") + ggtitle("Distribution of Discrete Variables \n")
density_df2 <- train

melted2 <- melt(density_df2, id = 1)
melted2$TARGET <- factor(melted2$TARGET)

ggplot(melted2, aes(TARGET, value)) + geom_boxplot(aes(fill = TARGET),
    alpha = 0.5) + facet_wrap(~variable, scales = "free") + scale_fill_discrete(guide = FALSE) +
    scale_y_continuous("", labels = NULL, breaks = NULL) + scale_x_discrete("") +
    ggtitle("Distribution of Predictors by TARGET\n")

# treat NAs
train$STARS[is.na(train$STARS)] <- 0

ggplot(train, aes(TARGET, fill = factor(STARS))) + geom_histogram(binwidth = 1,
    position = "dodge") + scale_fill_discrete(name = "STARS") +
    theme(legend.position = "bottom")

row_has_NA <- apply(train, 1, function(x) {
    any(is.na(x))
})
sum(row_has_NA)
wine_new <- train
wine_new$ResidualSugar_NA <- factor(ifelse(is.na(train$ResidualSugar),
    1, 0))
wine_new$Chlorides_NA <- factor(ifelse(is.na(train$Chlorides),
    1, 0))
wine_new$FreeSulfurDioxide_NA <- factor(ifelse(is.na(train$FreeSulfurDioxide),
    1, 0))
wine_new$TotalSulfurDioxide_NA <- factor(ifelse(is.na(train$TotalSulfurDioxide),
    1, 0))
wine_new$pH_NA <- factor(ifelse(is.na(train$pH), 1, 0))
wine_new$Sulphates_NA <- factor(ifelse(is.na(train$Sulphates),
    1, 0))
wine_new$Alcohol_NA <- factor(ifelse(is.na(train$Alcohol), 1,
    0))

p1 <- ggplot(wine_new, aes(x = ResidualSugar_NA, y = TARGET)) +
```

```
    geom_violin(scale = "count") + geom_jitter(alpha = 0.2, size = 0.2,
    col = "lightskyblue4") + xlab(colnames(wine_new)[16])
p2 <- ggplot(wine_new, aes(x = Chlorides_NA, y = TARGET)) + geom_violin(scale = "count") +
    geom_jitter(alpha = 0.2, size = 0.2, col = "lightskyblue4") +
    xlab(colnames(wine_new)[17])
p3 <- ggplot(wine_new, aes(x = FreeSulfurDioxide_NA, y = TARGET)) +
    geom_violin(scale = "count") + geom_jitter(alpha = 0.2, size = 0.2,
    col = "lightskyblue4") + xlab(colnames(wine_new)[18])
p4 <- ggplot(wine_new, aes(x = TotalSulfurDioxide_NA, y = TARGET)) +
    geom_violin(scale = "count") + geom_jitter(alpha = 0.2, size = 0.2,
    col = "lightskyblue4") + xlab(colnames(wine_new)[19])
p5 <- ggplot(wine_new, aes(x = pH_NA, y = TARGET)) + geom_violin(scale = "count") +
    geom_jitter(alpha = 0.2, size = 0.2, col = "lightskyblue4") +
    xlab(colnames(wine_new)[20])
p6 <- ggplot(wine_new, aes(x = Sulphates_NA, y = TARGET)) + geom_violin(scale = "count") +
    geom_jitter(alpha = 0.2, size = 0.2, col = "lightskyblue4") +
    xlab(colnames(wine_new)[21])
p7 <- ggplot(wine_new, aes(x = Alcohol_NA, y = TARGET)) + geom_violin(scale = "count") +
    geom_jitter(alpha = 0.2, size = 0.2, col = "lightskyblue4") +
    xlab(colnames(wine_new)[22])

grid.arrange(p1, p2, p3, p4, p5, p6, p7, ncol = 2)

df_varHasNA <- train %>% dplyr::select(-c(TARGET, FixedAcidity,
    VolatileAcidity, CitricAcid, Density, LabelAppeal, AcidIndex,
    STARS))

aggr(df_varHasNA, prop = TRUE, numbers = TRUE, sortVars = TRUE,
    cex.lab = 0.4, cex.axis = par("cex"), cex.numbers = par("cex"),
    delimiter = "_imp", combined = TRUE)
train <- data.frame(train[complete.cases(train), ])
melted3 <- melt(train)
ggplot(melted3, aes(value)) + geom_bar(aes(fill = variable, col = variable),
    alpha = 0.5, show.legend = FALSE) + facet_wrap(~variable,
    scale = "free") + ggtitle("Density of Variables After Casewise Deletion\n")

cm <- cor(train, use = "pairwise.complete.obs")
corrplot(cm, method = "square", type = "upper")

cor_df <- correlation_df(cm)
kable(head(cor_df, 10), digits = 2, row.names = T, caption = "Top Correlated Variable Pairs")

# predictor selection
regfit.full = regsubsets(TARGET ~ ., data = train, nvmax = 14)
reg.summary <- summary(regfit.full)

par(mfrow = c(1, 2))
plot(regfit.full, scale = "bic", main = "Predictor Variables vs. BIC")
plot(reg.summary$bic, xlab = "Number of Predictors", ylab = "BIC",
    type = "l", main = "Best subset Selection using BIC")
points(7, reg.summary$bic[7], col = "red", cex = 2, pch = 20)

par(mfrow = c(1, 2))
plot(regfit.full, scale = "Cp", main = "Predictor Variables vs. Cp")
plot(reg.summary$cp, xlab = "Number of Predictors", ylab = "Cp",
    type = "l", main = "Best subset Selection using Cp")
points(11, reg.summary$cp[11], col = "red", cex = 2, pch = 20)
par(mfrow = c(1, 1))
```

```
# Poisson models
model0 <- (glm(TARGET ~ ., family = "poisson", data = train))
pander(summary(model0))

model1 <- glm(TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
    TotalSulfurDioxide + LabelAppeal + AcidIndex + STARS + Alcohol +
    Density + Sulphates + pH, family = "poisson", data = train)
pander(summary(model1))

model2 <- (glm(TARGET ~ LabelAppeal + AcidIndex + STARS, family = poisson,
    data = train))
pander(summary(model2))

# NB models
model3 <- glm(TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
    TotalSulfurDioxide + AcidIndex + LabelAppeal + STARS, data = train)
pander(summary(model3))

model4 <- glm(TARGET ~ VolatileAcidity + TotalSulfurDioxide +
    LabelAppeal + AcidIndex + STARS, data = train)
pander(summary(model4))

# linear models
model5 <- glm(TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
    TotalSulfurDioxide + LabelAppeal + log(AcidIndex) + STARS,
    family = "gaussian", data = train)
pander(summary(model5))

model6 <- glm(TARGET ~ VolatileAcidity + FreeSulfurDioxide +
    TotalSulfurDioxide + Chlorides + Density + pH + Sulphates +
    LabelAppeal + AcidIndex + STARS, family = gaussian, data = train)
pander(summary(model6))

# cross validation
k = 10
set.seed(1306)
folds = sample(1:k, nrow(train), replace = TRUE)

cv.errors0 = matrix(NA, k, 10, dimnames = list(NULL, paste(1:10)))
cv.errors1 = matrix(NA, k, 10, dimnames = list(NULL, paste(1:10)))
cv.errors2 = matrix(NA, k, 10, dimnames = list(NULL, paste(1:10)))
cv.errors3 = matrix(NA, k, 10, dimnames = list(NULL, paste(1:10)))
cv.errors4 = matrix(NA, k, 10, dimnames = list(NULL, paste(1:10)))
cv.errors5 = matrix(NA, k, 10, dimnames = list(NULL, paste(1:10)))
cv.errors6 = matrix(NA, k, 10, dimnames = list(NULL, paste(1:10)))

for (j in 1:k) {
    model0 <- glm(TARGET ~ ., family = "poisson", data = train[folds !=
        j, ])
    model1 <- glm(TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
        TotalSulfurDioxide + LabelAppeal + AcidIndex + STARS +
        Alcohol + Density + Sulphates + pH, family = "poisson",
        data = train[folds != j, ])
    model2 <- glm(TARGET ~ LabelAppeal + AcidIndex + STARS, family = "poisson",
        data = train[folds != j, ])
    model3 <- glm.nb(TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
        TotalSulfurDioxide + Alcohol + AcidIndex + LabelAppeal +
        STARS, data = train[folds != j, ])
    model4 <- glm.nb(TARGET ~ VolatileAcidity + TotalSulfurDioxide +
```

```
        LabelAppeal + AcidIndex + STARS, data = train[folds !=
        j, ])
    model5 <- glm(TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
        TotalSulfurDioxide + LabelAppeal + log(AcidIndex) + STARS,
        family = "gaussian", data = train[folds != j, ])
    model6 <- glm(TARGET ~ VolatileAcidity + FreeSulfurDioxide +
        TotalSulfurDioxide + Chlorides + Density + pH + Sulphates +
        LabelAppeal + AcidIndex + STARS, family = "gaussian",
        data = train[folds != j, ])

    # best.fit = regsubsets(y ~ ., data = train_df[folds != j, ],
    # nvmax = 10)
    for (i in 1:10) {
        f = train[folds == j, ]
        # f = f[complete.cases(f),]

        pred0 = predict(model0, f, id = i)
        cv.errors0[j, i] = mean((train$TARGET[folds == j] - pred0)^2,
            na.rm = TRUE)

        pred1 = predict(model1, f, id = i)
        cv.errors1[j, i] = mean((train$TARGET[folds == j] - pred1)^2,
            na.rm = TRUE)

        pred2 = predict(model2, f, id = i)
        cv.errors2[j, i] = mean((train$TARGET[folds == j] - pred2)^2,
            na.rm = TRUE)

        pred3 = predict(model3, f, id = i)
        cv.errors3[j, i] = mean((train$TARGET[folds == j] - pred3)^2,
            na.rm = TRUE)

        pred4 = predict(model4, f, id = i)
        cv.errors4[j, i] = mean((train$TARGET[folds == j] - pred4)^2,
            na.rm = TRUE)

        pred5 = predict(model5, f, id = i)
        cv.errors5[j, i] = mean((train$TARGET[folds == j] - pred5)^2,
            na.rm = TRUE)

        pred6 = predict(model6, f, id = i)
        cv.errors6[j, i] = mean((train$TARGET[folds == j] - pred6)^2,
            na.rm = TRUE)
    }
}

mean.cv.errors0 <- apply(cv.errors0, 2, mean)
mean.cv.errors1 <- apply(cv.errors1, 2, mean)
mean.cv.errors2 <- apply(cv.errors2, 2, mean)
mean.cv.errors3 <- apply(cv.errors3, 2, mean)
mean.cv.errors4 <- apply(cv.errors4, 2, mean)
mean.cv.errors5 <- apply(cv.errors5, 2, mean)
mean.cv.errors6 <- apply(cv.errors6, 2, mean)

all.cv.error = data.frame(mean(mean.cv.errors0), mean(mean.cv.errors1),
    mean(mean.cv.errors2), mean(mean.cv.errors3), mean(mean.cv.errors4),
    mean(mean.cv.errors5), mean(mean.cv.errors6))

names(all.cv.error) = c("Poisson Model 0", "Poisson Model 1",
```

```
    "Poisson Model 2", "NB Model 3", "NB Model 4", "MLR Model 5",
    "MLR Model 6")
all.cv.error = t(all.cv.error)
names(all.cv.error) = c("Model", "Mean CV Error")
pander(all.cv.error)

# prediction
predicted_cases <- predict(model6, test, type = "response")
predicted_cases_int = round(predicted_cases, 0)

table_test <- table(predicted_cases_int)/length(predicted_cases_int)
table_test <- c(0, table_test, 0)
names(table_test)[1] <- "0"
names(table_test)[9] <- "8"
table_train <- table(train$TARGET)/length(train$TARGET)
table_ratings <- rbind(table_test, table_train)
row.names(table_ratings) <- c("Test", "Train")
pander(table_ratings, round = 3)

# prediction results
prediction = data.frame(matrix(NA, nrow = 1667, ncol = 6))
prediction[, 1] = test$IN[1:1667]
prediction[, 2] = predicted_cases[1:1667]
prediction[, 3] = predicted_cases_int[1:1667]
prediction[, 4] = test$IN[1669:3335]
prediction[, 5] = predicted_cases[1669:3335]
prediction[, 6] = predicted_cases_int[1669:3335]
prediction <- rbind(prediction, c(test$IN[1668], predicted_cases[1668],
    predicted_cases_int[1688], rep(NA, 3)))

names(prediction) = rep(c("Index", "Value", "Cases"), 2)
pander(prediction)
```