DS621-Homework 5

George Cruz Deschamps[1], Karim Hammoud[1], Maliat Islam[1], Matthew Lucich[1], Gabriella Martinez[1], Ken Popkin[1]

Abstract

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. Our 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.

Email addresses: georg4re@gmail.com (George Cruz Deschamps), cunykarim@gmail.com (Karim Hammoud), maliat.islam21@gmail.com (Maliat Islam), matt.lucich@gmail.com (Matthew Lucich), gpmmrtzz@gmail.com (Gabriella Martinez), krpopkin@gmail.com (Ken Popkin)

Data Exploration

Let's take an initial look at our data. The summary looks like this:

Characteristic	**N = 12,795**
TARGET	
0	2,734 / 12,795 (21%)
1	244 / 12,795 (1.9%)
2	1,091 / 12,795 (8.5%)
3	2,611 / 12,795 (20%)
4	3,177 / 12,795 (25%)
5	2,014 / 12,795 (16%)
6	765 / 12,795 (6.0%)
7	142 / 12,795 (1.1%)
8	17 / 12,795 (0.1%)
FixedAcidity	7.08 (6.32)
VolatileAcidity	0.32 (0.78)
CitricAcid	0.31 (0.86)
ResidualSugar	5.42 (33.75)
(Missing)	616
Chlorides	0.05 (0.32)
(Missing)	638
FreeSulfurDioxide	30.85 (148.71)
(Missing)	647
TotalSulfurDioxide	120.71 (231.91)
(Missing)	682
Density	0.99 (0.03)
pH	3.21 (0.68)
(Missing)	395
Sulphates	0.53 (0.93)
(Missing)	1,210
Alcohol	10.49 (3.73)
(Missing)	653
LabelAppeal	
-2	504 / 12,795 (3.9%)
-1	3,136 / 12,795 (25%)
0	5,617 / 12,795 (44%)
1	3,048 / 12,795 (24%)
2	490 / 12,795 (3.8%)
AcidIndex	7.77 (1.32)
STARS	
1	3,042 / 9,436 (32%)
2	3,570 / 9,436 (38%)
3	2,212 / 9,436 (23%)
4	612 / 9,436 (6.5%)

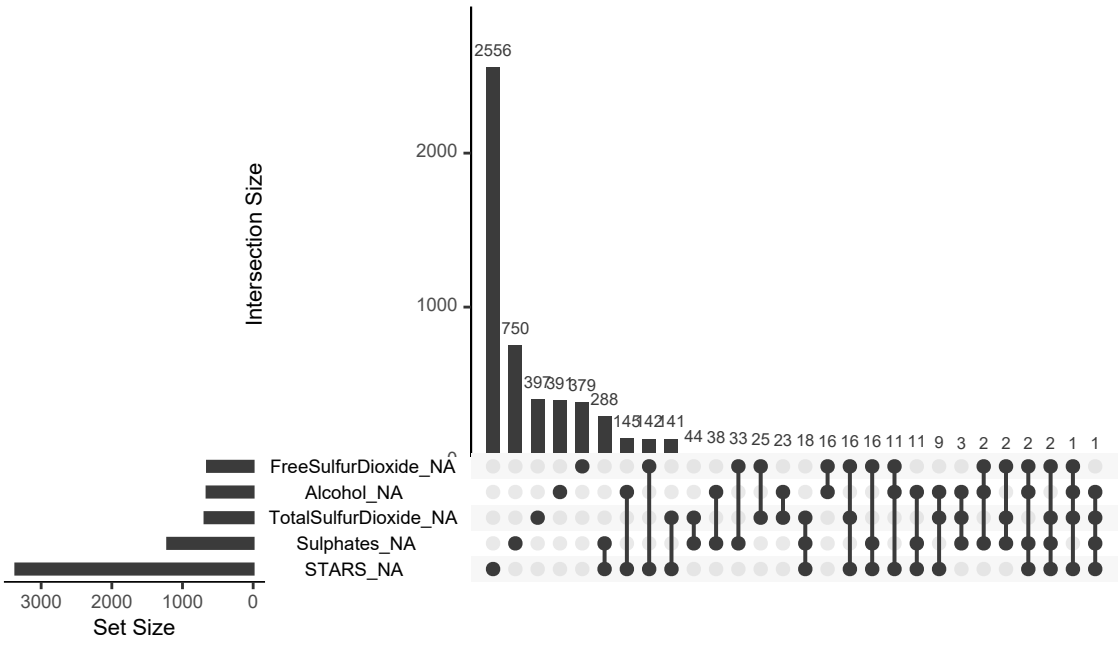
Characteristic	N = 12,795
(Missing)	3,359

Our dataset consists of 15 variables and 12,795 observations. There are some missing values on ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, and STARS variables. TARGET is our response variable. LabelAppeal, AcidIndex, and STARS are discrete variables and the rest are continuous.

```
## Dataset for this assignment is 12795 instances and 16 features
```

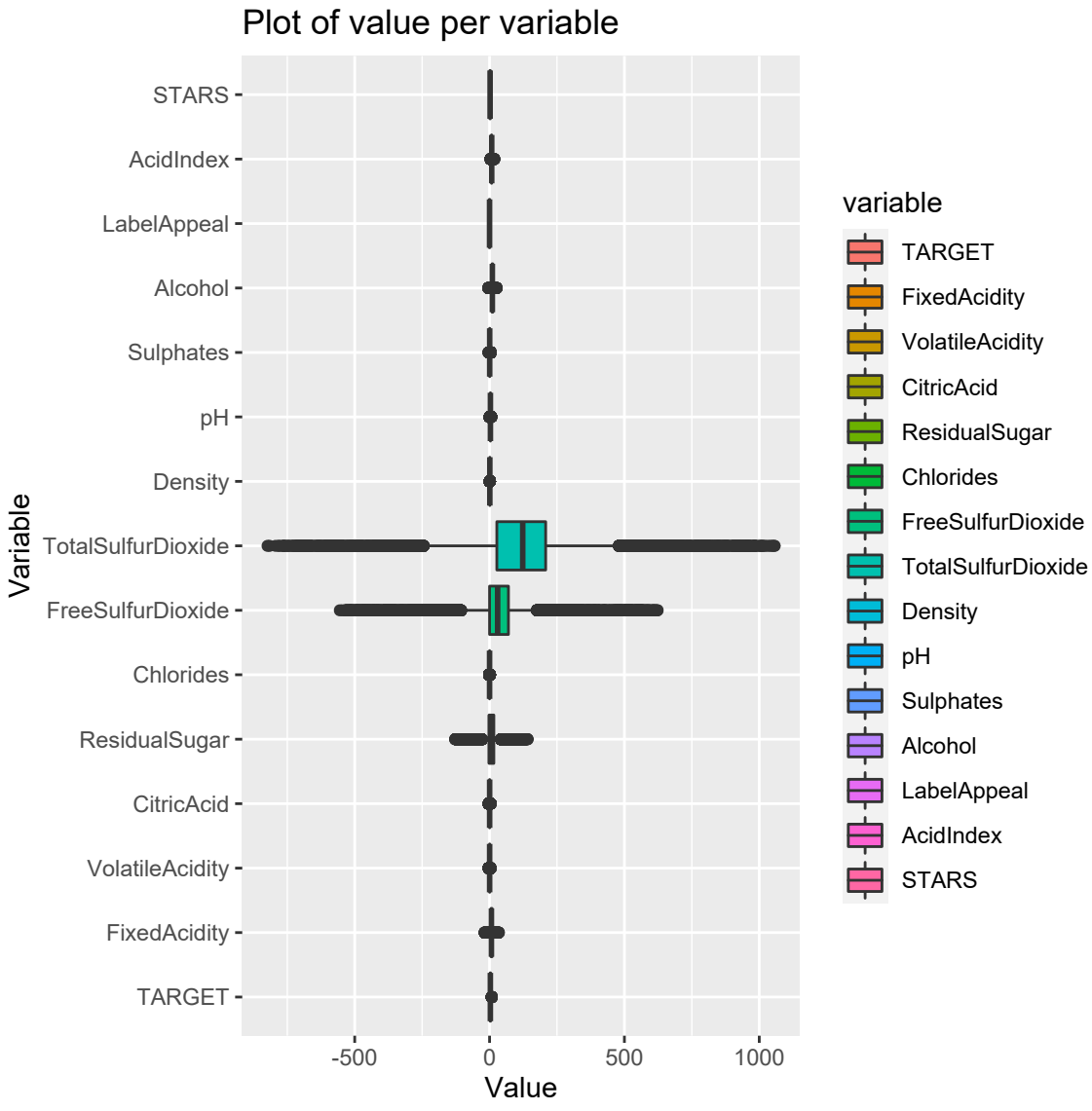
Missing Values Count:

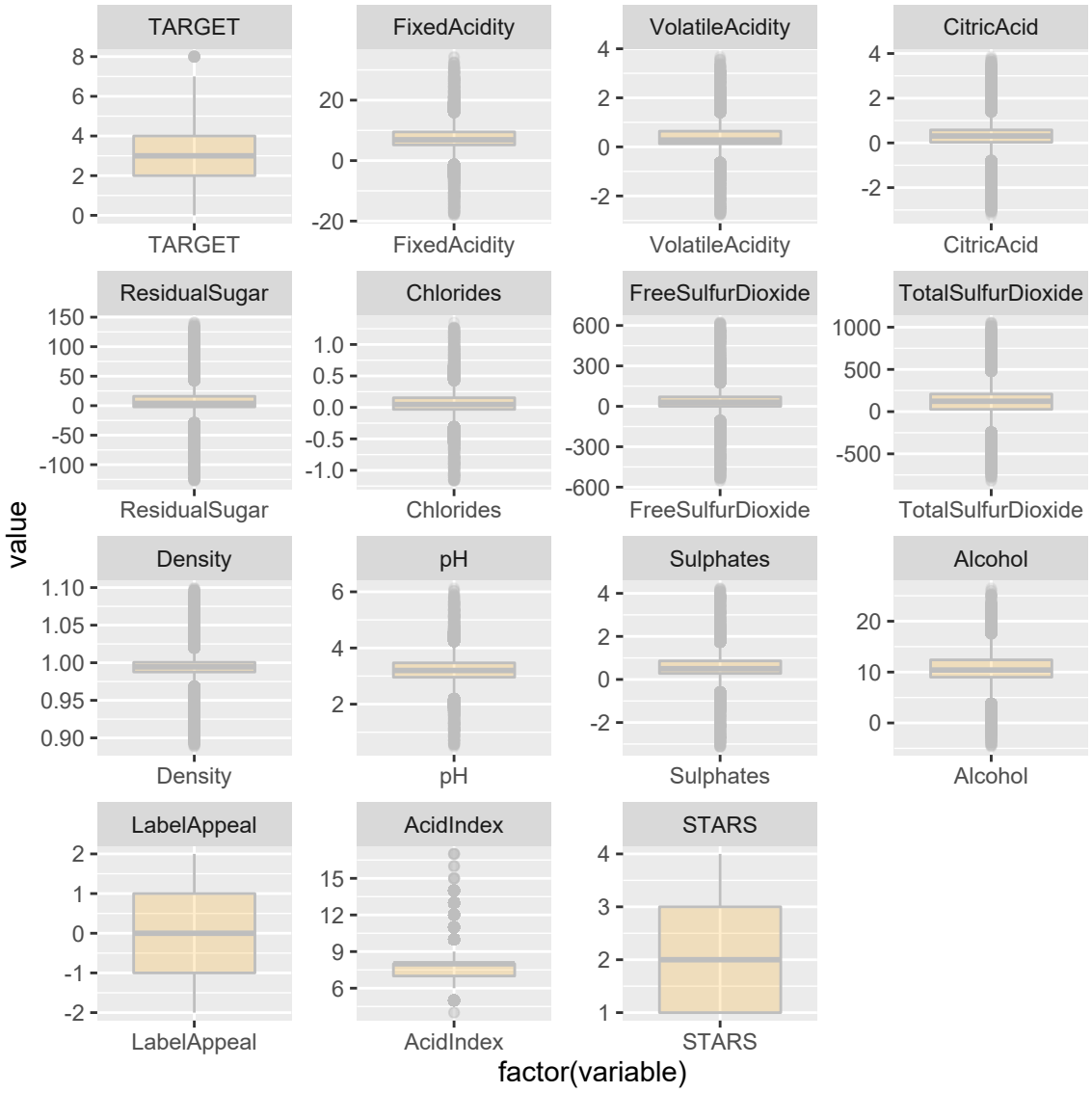
Variable	Missing
ResidualSugar	616
Chlorides	638
FreeSulfurDioxide	647
TotalSulfurDioxide	682
pH	395
Sulphates	1210
Alcohol	653
STARS	3359



Data Visualization

In the box plots below, we can see `TotalSulfurDioxide`, `FreeSulfurDioxide`, and `ResidualSugar` variables have large ranges compared to the other variables. We can tell a high number of variables have outliers. Almost all of the variables are centered around zero and at least four of the variables have negative values.

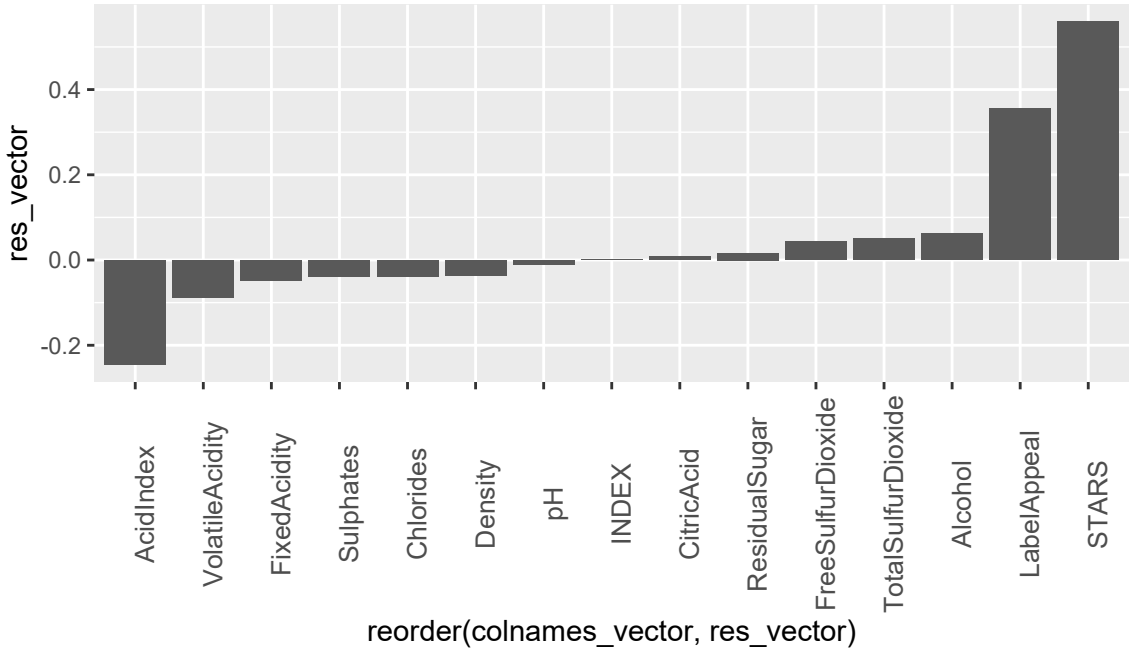



TARGET
FixedAcidity
VolatileAcidity
CitricAcid
ResidualSugar
Chlorides
FreeSulfurDioxide
TotalSulfurDioxide
Density
pH
Sulphates
Alcohol
LabelAppeal
AcidIndex
STARS
value
factor(variable)

Data correlations

We also take a look at the correlation between the variables.

	colnames_vector	res_vector
2	STARS	0.5588
3	LabelAppeal	0.3565
4	Alcohol	0.0621
5	TotalSulfurDioxide	0.0515
6	FreeSulfurDioxide	0.0438
7	ResidualSugar	0.0165
8	CitricAcid	0.0087
9	INDEX	0.0013
10	pH	-0.0094
11	Density	-0.0355
12	Chlorides	-0.0383
13	Sulphates	-0.0388
14	FixedAcidity	-0.0490
15	VolatileAcidity	-0.0888
16	AcidIndex	-0.2460



In the above correlation table and plot `STARS` and `LabelAppeal` appear to be most positively correlated variables with the response variable. We can also see some mild negative correlation between the response variable and `AcidIndex`.

Data Manipulation

Recommendations

1. **Address Missing Values** The features below have missing values. Recommendation to address this for each feature is:

a. **Residual Sugar (616):** replace missing values with mean because normal dist, corr w/target is 0, & most target values have a similar mean.
b. **Chlorides (638):** same as Residual Sugar (replace with mean)
c. **Free Sulfur Dioxide (647):** same as Residual Sugar
d. **Total Sulfur Dioxide (682):** same as Residual Sugar (considered using median, but mean is 120 and median is 123, so not much difference)
e. **PH (395):** same as Residual Sugar
f. **Sulphates (1210):** same as Residual Sugar

g. **Alcohol (653):** replace missing values with mean. Correlation is higher on this feature at around 12%
h. **STARS (3359):** values are 1 to 4. replace missing values with a 0 to indicate no ranking provided.

2. **Apply Scalar** Max values are wide ranging across the features. For example, `Volatile Acidity`'s max value is 4 and `SulfurDioxide's` max value is 1038. Recommend applying StandardScalar prior to modeling.

Data Modeling

Poisson model (with inputted missing values)

Characteristic	IRR	95% CI	p-value
FixedAcidity	1.00	1.00, 1.00	0.7
VolatileAcidity	0.97	0.95, 0.98	<0.001
CitricAcid	1.01	1.00, 1.02	0.2
ResidualSugar	1.00	1.00, 1.00	0.7
Chlorides	0.96	0.93, 0.99	0.012
FreeSulfurDioxide	1.00	1.00, 1.00	<0.001
TotalSulfurDioxide	1.00	1.00, 1.00	<0.001
Density	0.75	0.52, 1.10	0.14
pH	0.98	0.97, 1.00	0.040
Sulphates	0.99	0.98, 1.00	0.027
Alcohol	1.00	1.00, 1.00	0.12
LabelAppeal	1.14	1.13, 1.16	<0.001
AcidIndex	0.92	0.91, 0.92	<0.001
STARS	1.37	1.35, 1.38	<0.001

Poisson model with significant variables.

Characteristic	IRR	95% CI	p-value
VolatileAcidity	0.96	0.95, 0.98	<0.001
AcidIndex	0.92	0.91, 0.93	<0.001
STARS	1.41	1.40, 1.42	<0.001

Negative Binomial (with inputted missing values)

Characteristic	IRR	95% CI	p-value
INDEX	1.00	1.00, 1.00	0.7
FixedAcidity	1.00	1.00, 1.00	0.7
VolatileAcidity	0.97	0.95, 0.98	<0.001
CitricAcid	1.01	1.00, 1.02	0.2
ResidualSugar	1.00	1.00, 1.00	0.7
Chlorides	0.96	0.93, 0.99	0.012
FreeSulfurDioxide	1.00	1.00, 1.00	<0.001
TotalSulfurDioxide	1.00	1.00, 1.00	<0.001
Density	0.75	0.52, 1.10	0.14
pH	0.98	0.97, 1.00	0.040
Sulphates	0.99	0.98, 1.00	0.028
Alcohol	1.00	1.00, 1.00	0.12
LabelAppeal	1.14	1.13, 1.16	<0.001

Characteristic	IRR	95% CI	p-value
AcidIndex	0.92	0.91, 0.92	<0.001
STARS	1.37	1.35, 1.38	<0.001

Negative binomial model with significant values.

Characteristic	IRR	95% CI	p-value
VolatileAcidity	0.97	0.96, 0.98	<0.001
FreeSulfurDioxide	1.00	1.00, 1.00	0.011
TotalSulfurDioxide	1.00	1.00, 1.00	0.001
Alcohol	1.00	1.00, 1.01	0.002
as.factor(LabelAppeal)			
-2			
-1	1.27	1.18, 1.37	<0.001
0	1.54	1.43, 1.65	<0.001
1	1.76	1.63, 1.89	<0.001
2	2.01	1.85, 2.19	<0.001
as.factor(AcidIndex)			
4			
5	0.86	0.48, 1.73	0.6
6	0.90	0.51, 1.79	0.7
7	0.87	0.49, 1.73	0.7
8	0.84	0.48, 1.68	0.6
9	0.75	0.43, 1.50	0.4
10	0.64	0.36, 1.29	0.2
11	0.45	0.25, 0.90	0.012
12	0.44	0.24, 0.89	0.012
13	0.52	0.29, 1.06	0.047
14	0.47	0.25, 0.98	0.028
15	0.74	0.34, 1.69	0.5
16	0.39	0.12, 1.09	0.082
17	0.30	0.09, 0.84	0.028
as.factor(STARS)			
0			
1	2.13	2.05, 2.21	<0.001
2	2.93	2.83, 3.04	<0.001
3	3.30	3.18, 3.43	<0.001
4	3.72	3.55, 3.90	<0.001

Linear Model with scaled significant data.

Characteristic	Beta	95% CI	p-value
VolatileAcidity	-0.04	-0.05, -0.03	<0.001
FreeSulfurDioxide	0.02	0.01, 0.03	<0.001
TotalSulfurDioxide	0.03	0.01, 0.04	<0.001
Alcohol	0.03	0.01, 0.04	<0.001
as.factor(LabelAppeal)			
-2.23426998624856			
-1.11204793733397	0.19	0.13, 0.25	<0.001
0.0101741115806247	0.43	0.37, 0.50	<0.001
1.13239616049522	0.68	0.61, 0.74	<0.001
2.25461820940981	1.0	0.89, 1.1	<0.001
as.factor(AcidIndex)			
-2.84964768611181			
-2.09431867252628	-0.16	-0.94, 0.62	0.7
-1.33898965894075	-0.10	-0.87, 0.66	0.8
-0.583660645355226	-0.15	-0.92, 0.61	0.7
0.171668368230302	-0.21	-1.0, 0.56	0.6
0.926997381815829	-0.37	-1.1, 0.40	0.3
1.68232639540136	-0.53	-1.3, 0.24	0.2
2.43765540898688	-0.78	-1.5, -0.01	0.049
3.19298442257241	-0.78	-1.6, -0.01	0.047
3.94831343615794	-0.79	-1.6, -0.01	0.048
4.70364244974347	-0.71	-1.5, 0.08	0.077
5.45897146332899	-0.32	-1.2, 0.58	0.5
6.21430047691452	-0.89	-1.9, 0.08	0.073
6.96962949050005	-1.0	-1.9, -0.07	0.035
as.factor(STARS)			
-1.26902307229896			
-0.42623524866846	0.70	0.67, 0.74	<0.001
0.416552574962037	1.2	1.2, 1.3	<0.001
1.25934039859254	1.5	1.5, 1.6	<0.001
2.10212822222303	1.9	1.8, 1.9	<0.001

Linear model without the scaled value.

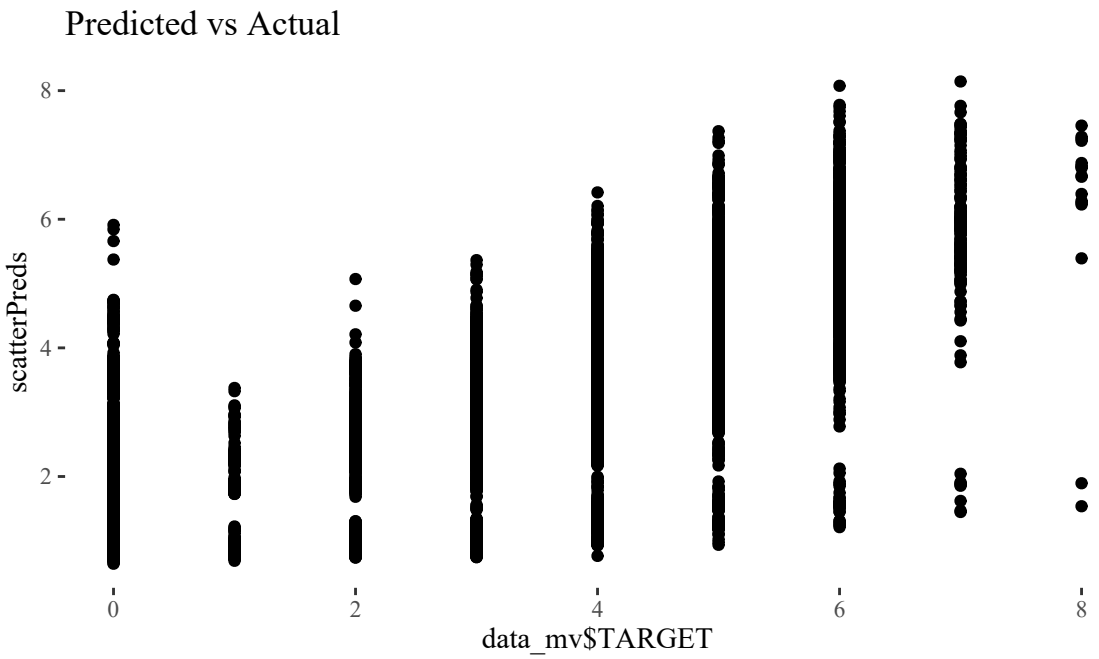
Characteristic	Beta	95% CI	p-value
INDEX	0.00	0.00, 0.00	0.7
FixedAcidity	0.00	0.00, 0.00	>0.9
VolatileAcidity	-0.10	-0.13, -0.07	<0.001
CitricAcid	0.02	-0.01, 0.05	0.13
ResidualSugar	0.00	0.00, 0.00	0.6
Chlorides	-0.12	-0.20, -0.05	<0.001
FreeSulfurDioxide	0.00	0.00, 0.00	<0.001
TotalSulfurDioxide	0.00	0.00, 0.00	<0.001

Characteristic	Beta	95% CI	p-value
Density	-0.80	-1.7, 0.07	0.071
pH	-0.03	-0.07, 0.00	0.049
Sulphates	-0.03	-0.06, -0.01	0.013
Alcohol	0.01	0.00, 0.02	<0.001
LabelAppeal	0.43	0.41, 0.46	<0.001
AcidIndex	-0.21	-0.23, -0.19	<0.001
STARS	1.0	1.0, 1.0	<0.001

Zero inflation with inputted Missing Values.

```
##
## Call:
## zeroinfl(formula = TARGET ~ . | STARS, data = data_mv, dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.31701 -0.53688  0.01918  0.40829  2.89146
##
## Count model coefficients (negbin with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.543e+00         NA      NA       NA
## INDEX                1.809e-07         NA      NA       NA
## FixedAcidity         3.013e-04         NA      NA       NA
## VolatileAcidity     -1.499e-02         NA      NA       NA
## CitricAcid           1.243e-03         NA      NA       NA
## ResidualSugar       -4.969e-05         NA      NA       NA
## Chlorides           -2.335e-02         NA      NA       NA
## FreeSulfurDioxide    3.718e-05         NA      NA       NA
## TotalSulfurDioxide -2.958e-06         NA      NA       NA
## Density             -2.732e-01         NA      NA       NA
## pH                   2.922e-03         NA      NA       NA
## Sulphates           -1.947e-03         NA      NA       NA
## Alcohol              6.542e-03         NA      NA       NA
## LabelAppeal          2.240e-01         NA      NA       NA
## AcidIndex           -3.130e-02         NA      NA       NA
## STARS                1.007e-01         NA      NA       NA
## Log(theta)           1.797e+01         NA      NA       NA
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3804         NA      NA       NA
## STARS        -2.2238         NA      NA       NA
##
## Theta = 63997386.9979
## Number of iterations in BFGS optimization: 23
```

```
## Log-likelihood: -2.083e+04 on 19 Df
```



The scaled dataset was transformed into absolute value dataset as Poisson and Negative Binomial were unable to work with negative data.

Absolute Value Poisson

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = absdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8550  -0.5038  -0.0667   0.4644   1.5768
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.920316   0.041520 -22.166   <2e-16 ***
## INDEX               0.012386   0.019653   0.630    0.529
## FixedAcidity        0.001218   0.013445   0.091    0.928
## VolatileAcidity     0.017113   0.013199   1.296    0.195
## CitricAcid          0.002713   0.013125   0.207    0.836
## ResidualSugar      -0.004542   0.013001  -0.349    0.727
## Chlorides           0.001761   0.012668   0.139    0.889
## FreeSulfurDioxide  -0.002295   0.012879  -0.178    0.859
## TotalSulfurDioxide  0.003254   0.013308   0.245    0.807
```

```
## Density             0.008868   0.012986   0.683    0.495
## pH                  0.004886   0.013442   0.363    0.716
## Sulphates           0.013197   0.012680   1.041    0.298
## Alcohol            -0.007084   0.013916  -0.509    0.611
## LabelAppeal         0.162795   0.013839  11.764   <2e-16 ***
## AcidIndex           0.120342   0.012365   9.732   <2e-16 ***
## STARS               0.488151   0.018842  25.907   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 6665.5  on 12794  degrees of freedom
## Residual deviance: 5715.0  on 12779  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 5
```

Model Selection

Let's select the models:

	MSE	AIC
Poisson	6.749981e+00	6.793446
Reduced Poisson	6.749968e+00	6.704008
Neg Binomial	4.568016e-01	1.751305
Neg Binomial Reduced	1.717355e+00	1.417136
Linear Model	4.670023e+04	47205.691870
LM Unscaled	4.670437e+04	45540.340259
Zero Inflation	Inf	6.749981
Poisson Abs	6.793446e+00	6.749968

Lower AIC and MSE Values indicate a better model.

Several models have a lower AIC and MSE value. 1. The negative binomial model with data_mv dataset. 2. The negative binomial model with only significant value. 3. The linear model with significant values from the scaled dataset. 4. The linear model with data_mv dataset. 5. The Scaled Poisson model with absolute values.

Test Data

Now lets see the output of the Models using test data.

Poisson	6.7499809
Reduced Poisson	6.7934458
Neg Binomial	6.7499676
Neg Binomial Reduced	6.7040075
Linear Model	0.4568016
LM Unscaled	1.7513053
Zero Inflation	1.7173548
Abs Poisson	1.4171356

Conclusions

We have used squared loss to validate the model. We will use the squared difference to select a model (MSE) from predictions on the training sets. As a lower number indicates a better fit model, (Poisson model with scaled absolute value dataset) will be selected.

References

Kieran Healy. "Data Visualization | A practical introduction". Duke University https://socviz.co/modeling.html#modeling. Accessed 22 Nov. 2021.

"Best subset model selection with R." http://jadianes.me/best-subset-model-selection-with-R. Accessed 22 Nov. 2021.

"ZERO-INFLATED POISSON REGRESSION | R DATA ANALYSIS EXAMPLES." https://stats.idre.ucla.edu/r/dae/zip/. Accessed 22 Nov. 2021.

"Introduction." R-Packages cran.r-project.org/web/packages/gtsummary/vignettes/tbl_summary.html. Accessed 22 Nov. 2021.

Kristalli, Anna. "Reproduce a Paper in Rmd." Annakrystalli.me, annakrystalli.me/rrtools-repro-research/paper.html#render_final_document_to_pdf. Accessed 22 Nov. 2021.