Assignment #3

Nitin Gaonkar PREDICT 411 section 56

INTRODUCTION

The purpose of this project is to analyze the data of around 12000 commercially available wines and build a model to predict the number of cases of wine that will be sold given certain properties of wine. Each record in the given data represents wine characteristics, here the target variable is the number of sample cases of wine that were purchased by wine distributing company after sampling a wine. We will build a model, which can predict the number of cases so that the manufacturing company will be able to adjust their wine offering to maximum sales. We will build 5 models with poison distribution, negative binomial, Zero Inflated Poisson distribution, Zero Inflated Negative Binomial distribution and regression using standard reg proc. Once we are done with model building, we will pick the best model using the metrics such as AIC or Average squared error.

1. Data Exploration:

The auto insurance dataset has around 12000 records and each record in the given data represents the wine characteristics, the variables are mostly related to the chemical properties of the wine being sold. The target value is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine.

We begin our data exploration by examining the data dictionary and the definitions given in the dictionary, after observing the data dictionary, we can consider both stars and label appeal as the categorical variables.

VARIABLE NAME	Defination
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. High numbers suggest customers like the label design. Negative numbers suggest customes don't like the design.
ResidualSugar	Residual Sugar of wine
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor
Sulphates	Sulfate content of wine
TotalSulfurDioxide	Total Sulfur Dioxide of Wine
VolatileAcidity	Volatile Acid content of wine
PH	PH of wine

a. Just to give a bit insight on the data, I have calculated and listed the mean and the standard deviation along with the min and max value of each variable in the below table.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
INDEX	12795	8069.98	4656.91	1.0000000	16129.00
TARGET	12795	3.0290739	1.9263682	0	8.0000000
FixedAcidity	12795	7.0757171	6.3176435	-18.1000000	34.4000000
VolatileAcidity	12795	0.3241039	0.7840142	-2.7900000	3.6800000
CitricAcid	12795	0.3084127	0.8620798	-3.2400000	3.8600000
ResidualSugar	12179	5.4187331	33.7493790	-127.8000000	141.1500000
Chlorides	12157	0.0548225	0.3184673	-1.1710000	1.3510000
FreeSulfurDioxide	12148	30.8455713	148.7145577	-555.0000000	623.0000000
TotalSulfurDioxide	12113	120.7142326	231.9132105	-823.0000000	1057.00
Density	12795	0.9942027	0.0265376	0.8880900	1.0992400
pH	12400	3.2076282	0.6796871	0.4800000	6.1300000
Sulphates	11585	0.5271118	0.9321293	-3.1300000	4.2400000
Alcohol	12142	10.4892363	3.7278190	-4.7000000	26.5000000
LabelAppeal	12795	-0.0090660	0.8910892	-2.0000000	2.0000000
AcidIndex	12795	7.7727237	1.3239264	4.0000000	17.0000000
STARS	9436	2.0417550	0.9025400	1.0000000	4.0000000

Table 1

So let's focus on the continuous variables and try to explore those variables and see get the means and the missing values from the data:

b. Missing values:

Based on the below table we can see that the there are few values which are missing for few variables, we will use imputation where we use the mean for the missing values. We will further analyze the variables by creating the histogram as well tests for normality. From the table 2 we can see that we have missing values for the variables like residual sugar, chlorides, freesulphurdioxide, totalsulfurdioxide, ph, sulphate,alcohol and stars.

The MEANS Procedure

Variable	N Miss
INDEX	0
TARGET	0
FixedAcidity	0
VolatileAcidity	0
CitricAcid	0
ResidualSugar	616
Chlorides	638
FreeSulfurDioxide	647
TotalSulfurDioxide	682
Density	0
pH	395
Sulphates	1210
Alcohol	653
LabelAppeal	0
AcidIndex	0
STARS	3359

Table 2

c. Below table gives us the correlation of the variables with the target flag: From the below table we can see that none of the variable are highly correlated with the target, other than label appeal and stars. They have a correlation of 0.35650 and 0.55879 respectively.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations															
	INDEX	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
TARGET	0.00126 0.8871 12795	-0.04901 <.0001 12795	-0.08879 <.0001 12795	0.00868 0.3260 12795	0.01649 0.0688 12179	-0.03826 <.0001 12157	0.04382 <.0001 12148	0.05148 <.0001 12113	-0.03552 <.0001 12795	-0.00944 0.2930 12400	-0.03885 <.0001 11585	0.06206 <.0001 12142	0.35650 <.0001 12795	-0.24605 <.0001 12795	0.55879 <.0001 9436

d. In this section we will explore the distribution of variables, after going through all the distribution of the variables we could see that there are few variables, which have extreme values and the high skewness. Below we will discuss few of the variables with the extreme values and skewness.

TARGET:

As we can observe from the below graph that the mean is around 3.0290 the data is normally distributed as the skewness is about -0.3263 which is good and by looking at the graph we can say that this variable passes the test of normality.

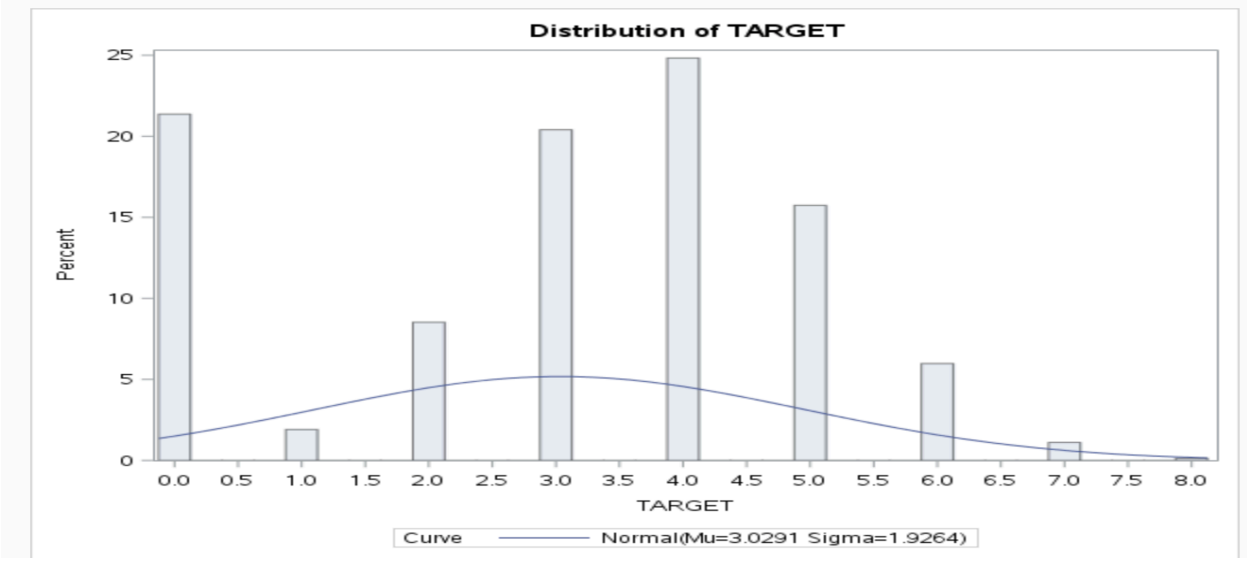

Fig: 1 distributions of TARGET

ACIDINDEX:

As we can observe from the below graph that the mean is around 7.7727 but there is long tail for this graph and the skewness is about 1.6488 which is bit high, thus we would definitely would require to have a look at this variable.

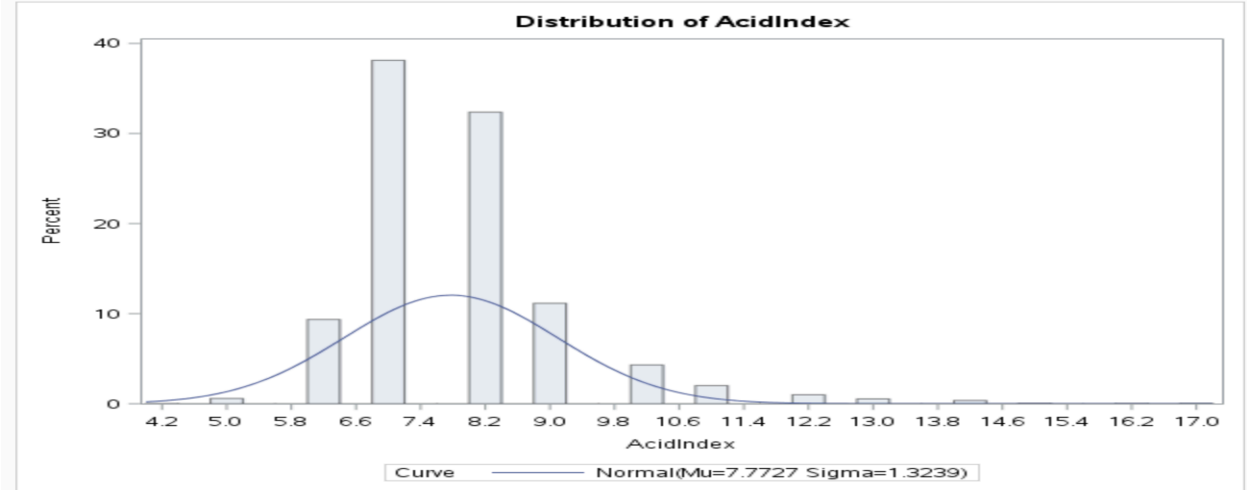

Fig :2 distributions of acidindex

ALCOHOL:

As we can observe from the below graph that the mean is around 10.48924 but there is long tail for this graph and the skewness is about -0.030 which is good, thus this variable passes the normality test.

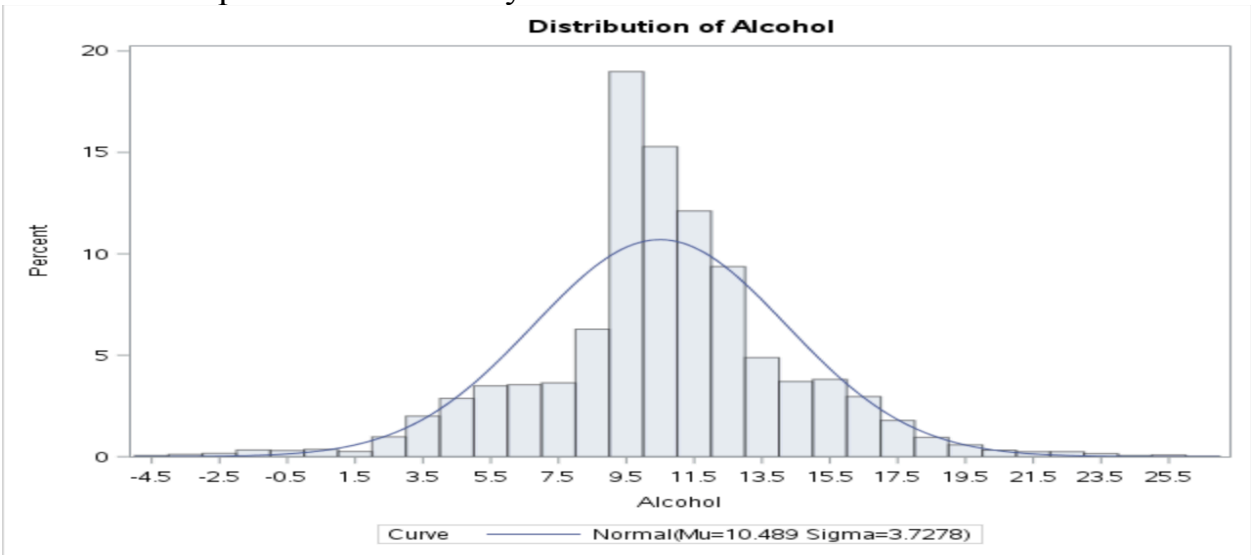

Fig :3 distribution of alcohol

DENSITY:

As we can observe from the below graph that the mean is around 0.9942 but there is long tail for this graph and the skewness is about -0.01 which is good, looks like the data is normally distributed.

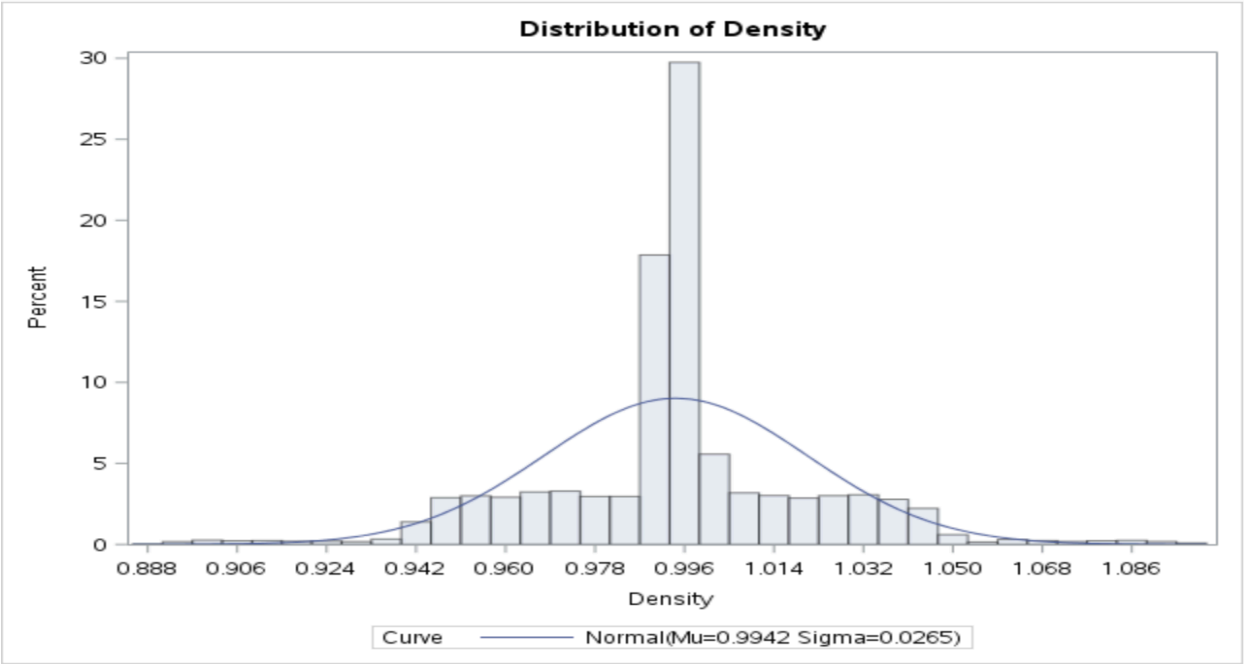

Fig: 4 distributions of Density

RESIDUAL SUGAR:

As we can observe from the below graph that the mean is around 5.41 and the skewness -0.053136 which is good

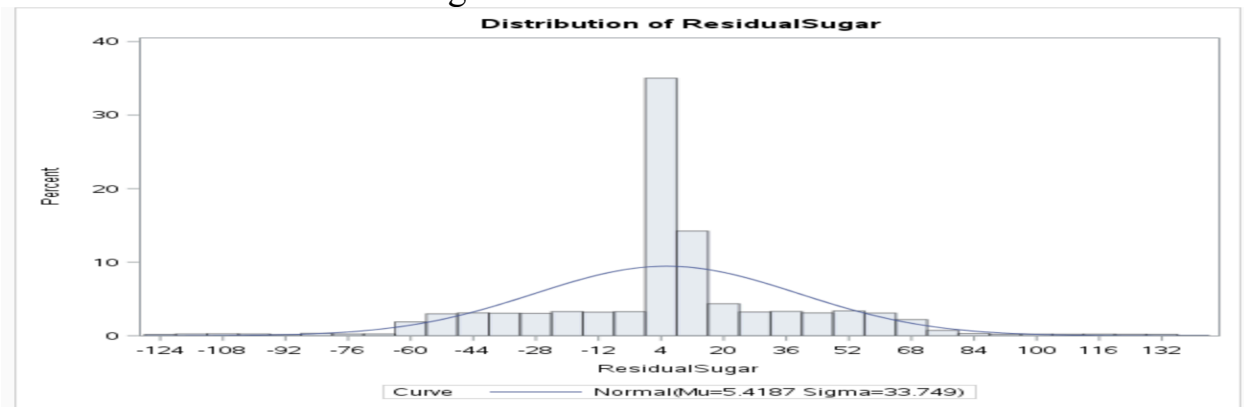

Fig: 5 distributions of RESIDUALSUGAR

By observing all the distributions via a histogram as well as test statistics for normality that includes a series of goodness of fit test based on the empirical distribution function, we found that despite the way the variables appear on the histogram, our good-ness fit test indicated that we should not reject the null hypothesis, which means that our variables are normally distributed, but there are few very extreme values and asymmetrical distributions in few of the variables, these issues can be addressed by various techniques like deleting the extreme values, use bucketing, transformation.

2. Data Preparation:

After the EDA in the above section we could see that two of the variables are strongly correlated to the dependent variable label appeal and stars, but as we can see that the missing table that for the variable STAR 25 % of the data is missing and there are few other variables which have missing data, we need to impute before we proceed with the modeling.

Missing values:

On reviewing the missing value chart, we can see that below variables have missing values:

The MEANS Procedure

Variable	N Miss
INDEX	0
TARGET	0
FixedAcidity	0
VolatileAcidity	0
CitricAcid	0
ResidualSugar	616
Chlorides	638
FreeSulfurDioxide	647
TotalSulfurDioxide	682
Density	0
pH	395
Sulphates	1210
Alcohol	653
LabelAppeal	0
AcidIndex	0
STARS	3359

Just to be on the safer side and we may use the other variables in our modeling later so we have imputed all the variables which have missing values with their mean.

Below new variables have been created:
IMP_ALCOHOL
I_IMP_ALCOHOL
IMP_CHLORIDES
I_IMP_CHLORIDES
IMP_FREESULPHURDIOXIDE
I_ IMP_FREESULPHURDIOXIDE
IMP_RESIDUALSUGAR
I_ IMP_RESIDUALSUGAR
I_IMP_STARS
I_STARS
IMP_SULPHATE
I_ IMP_SULPHATE
IMP_TOTALSULFURDIOXIDE
I_ IMP_TOTALSULFURDIOXIDE
IMP_PH
I_IMP_PH

Let's see the distribution of the imputed variables now:

IMP_HOME_VAL
As we can observe from the below graph that the mean is around 2.03,But there is long tail for this graph and the skewness is about 0. 0.5628 which is good.

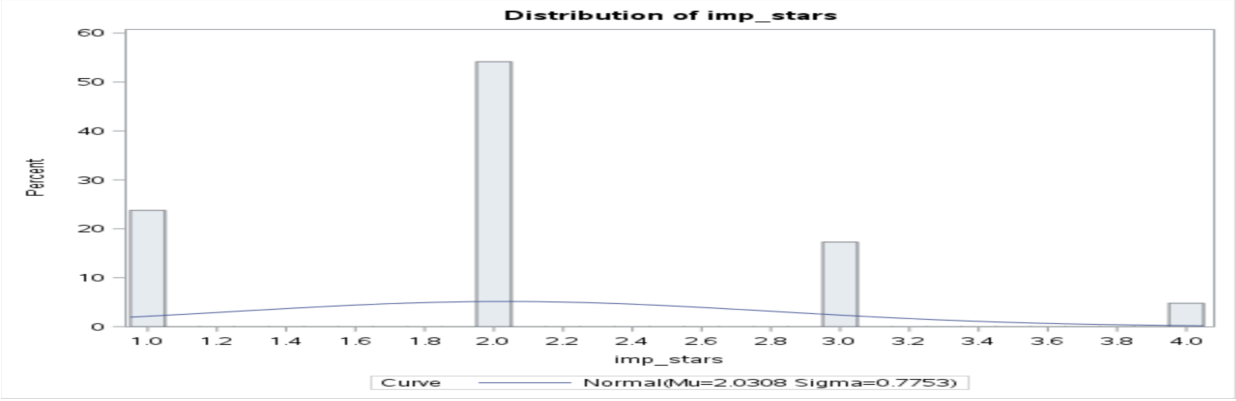

Fig:6 distributions of IMP_STARS

IMP_PH

As we can observe from the below graph that the mean is around 3.20 but there is long tail for this graph and the skewness is about 0.04 which is good as the data looks normally distributed.

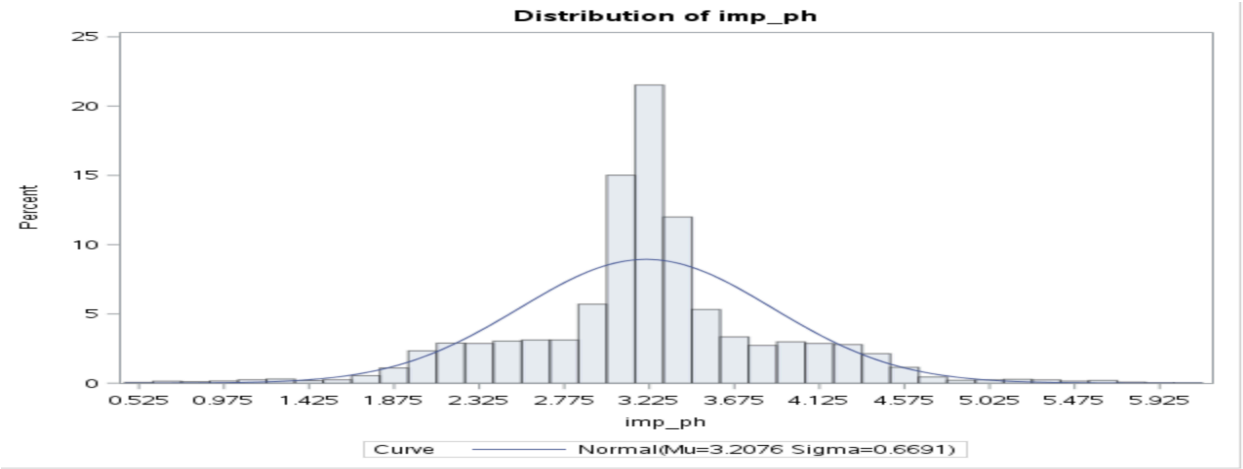

Fig:7 distributions of IMP_PH

After looking at the correlation tables in the EDA section we have decided to leave out the variables that doesn't have at least 0.1 correlation with the dependent variable, we do see high correlation between the qualitative review variables like label appeal and stars with the dependent variable, after leaving out the variables with below 0.1 we are left with acidindex,labelappeal, imp_stars and i_imp_stars, interestingly indicator variable for stars has the highest correlation with the dependent variable.

BUILD MODELS:

MODEL1:
Poisson model:

We will be using Genmod for Poisson distribution; below variables have been used for modeling:
acidindex
labelappeal
imp_stars
i_imp_stars

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.3752	0.0476	1.2820	1.4684	836.24	<.0001
AcidIndex		1	-0.0814	0.0045	-0.0902	-0.0726	328.69	<.0001
LabelAppeal	-2	1	-0.6958	0.0424	-0.7789	-0.6126	269.03	<.0001
LabelAppeal	-1	1	-0.4597	0.0250	-0.5086	-0.4107	338.98	<.0001
LabelAppeal	0	1	-0.2702	0.0228	-0.3149	-0.2254	139.87	<.0001
LabelAppeal	1	1	-0.1377	0.0232	-0.1831	-0.0923	35.38	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
imp_stars	1	1	-0.5647	0.0216	-0.6071	-0.5224	682.89	<.0001
imp_stars	2	1	-0.2431	0.0199	-0.2820	-0.2041	149.78	<.0001
imp_stars	3	1	-0.1207	0.0202	-0.1602	-0.0811	35.77	<.0001
imp_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
i_imp_stars	0	1	1.0926	0.0182	1.0569	1.1283	3599.71	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Table: Poisson Model Analysis of Maximum Likelihood Parameter Estimates

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	13E3	13700.3624	1.0716
Scaled Deviance	13E3	13700.3624	1.0716
Pearson Chi-Square	13E3	11331.6014	0.8863
Scaled Pearson X2	13E3	11331.6014	0.8863
Log Likelihood		8775.9792	
Full Log Likelihood		-22821.1920	
AIC (smaller is better)		45662.3841	
AICC (smaller is better)		45662.4013	
BIC (smaller is better)		45736.9522	

Table: Poisson Model Criteria For Assessing Goodness Of Fit

The model equation is:
$Y=B0+B1X+ B2X2+B3X3+B4X4 +E$

Interpretation of the above model is that the one-unit increase in acid index is a 8 % decrease in the expected number of cases purchased.
Now looking at the label appeal we can see that we have highest rating of 2.
For negative 2 rating 50 % decrease in the expected number of cases purchased.
For 1 rating 36 % decrease in the expected number of cases purchased.
Zero and 1 rating its 23 % and 12 % decrease in the expected number of cases purchased respectively.

Now for stars we can see that it decreases from obtaining a 4 rating, the model indicates a 98% increase in the expected number of cases to be purchased if the wine was given a star rating rather than being imputed.

Model 2:
Negative binomial distribution:

We will be using Genmod for negative binomial distribution; below variables have been used for modeling:
acidindex
labelappeal
imp_stars
i_imp_stars

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.3752	0.0476	1.2820	1.4684	836.24	<.0001
AcidIndex		1	-0.0814	0.0045	-0.0902	-0.0726	328.69	<.0001
LabelAppeal	-2	1	-0.6958	0.0424	-0.7789	-0.6126	269.03	<.0001
LabelAppeal	-1	1	-0.4597	0.0250	-0.5086	-0.4107	338.98	<.0001
LabelAppeal	0	1	-0.2702	0.0228	-0.3149	-0.2254	139.87	<.0001
LabelAppeal	1	1	-0.1377	0.0232	-0.1831	-0.0923	35.38	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
imp_stars	1	1	-0.5647	0.0216	-0.6071	-0.5224	682.89	<.0001
imp_stars	2	1	-0.2431	0.0199	-0.2820	-0.2041	149.78	<.0001
imp_stars	3	1	-0.1207	0.0202	-0.1602	-0.0811	35.77	<.0001
imp_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
i_imp_stars	0	1	1.0926	0.0182	1.0569	1.1283	3599.71	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		0	0.0000	0.0000	0.0000	0.0000		

Table: Negative Binomial Analysis of Maximum Likelihood Parameter Estimates

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13700.3624	1.0716
Scaled Deviance	13E3	13700.3624	1.0716
Pearson Chi-Square	13E3	11331.5923	0.8863
Scaled Pearson X2	13E3	11331.5923	0.8863
Log Likelihood		8775.9792	
Full Log Likelihood		-22821.1920	
AIC (smaller is better)		45664.3841	
AICC (smaller is better)		45664.4047	
BIC (smaller is better)		45746.4090	

Table: Negative binomial Model Criteria For Assessing Goodness Of Fit

The interpretation is similar to the model, we tried with the different varied inputs, the parameters are very close to the above model due to the mean and variance are so close.
Interpretation of the above model is that the one-unit increase in acid index is a 8 % decrease in the expected number of cases purchased.
Now looking at the label appeal we can see that we have highest rating of 2.
For negative 2 rating 50 % decrease in the expected number of cases purchased.
For 1 rating 36 % decrease in the expected number of cases purchased.
Zero and 1 rating its 23 % and 12 % decrease in the expected number of cases purchased respectively. Now for stars we can see that it decreases from obtaining a 4 rating, the model indicates a 98% increase in the expected number of cases to be purchased if the wine was given a star rating rather than being imputed.

Model 3:
We will be using Genmod for zero inflated Poisson, below variables have been used for modeling:
acidindex
labelappeal
imp_stars
i_imp_stars

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.8750	0.0499	1.7773	1.9727	1413.66	<.0001
AcidIndex		1	-0.0223	0.0049	-0.0320	-0.0126	20.34	<.0001
LabelAppeal	-2	1	-0.9652	0.0439	-1.0512	-0.8793	484.26	<.0001
LabelAppeal	-1	1	-0.5995	0.0260	-0.6504	-0.5486	533.13	<.0001
LabelAppeal	0	1	-0.3390	0.0236	-0.3852	-0.2928	206.89	<.0001
LabelAppeal	1	1	-0.1567	0.0238	-0.2032	-0.1101	43.46	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
imp_stars	1	1	-0.4170	0.0230	-0.4622	-0.3718	327.39	<.0001
imp_stars	2	1	-0.2012	0.0199	-0.2403	-0.1621	101.76	<.0001
imp_stars	3	1	-0.1049	0.0202	-0.1445	-0.0653	26.98	<.0001
imp_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
i_imp_stars	0	1	0.1868	0.0196	0.1483	0.2253	90.62	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Table: Zero Inflated Poisson Model Analysis Of Maximum Likelihood Parameter Estimates

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		41927.6145	
Scaled Deviance		41927.6145	
Pearson Chi-Square	13E3	6122.3756	0.4790
Scaled Pearson X2	13E3	6122.3756	0.4790
Log Likelihood		10633.3640	
Full Log Likelihood		-20963.8072	
AIC (smaller is better)		41953.6145	
AICC (smaller is better)		41953.6429	
BIC (smaller is better)		42050.5530	

Table: Zero Inflated Poisson Model Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.4731	0.1989	-3.8628	-3.0833	305.03	<.0001
AcidIndex		1	0.4773	0.0247	0.4288	0.5258	372.21	<.0001
i_imp_stars	0	1	-3.6189	0.0919	-3.7990	-3.4388	1550.75	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.

Table: Zero Inflated Poisson Model Criteria For Assessing Goodness Of Fit

Interpretation of the above model is that the one-unit increase in acid index is a 2 % decrease in the expected number of cases purchased.
Now looking at the label appeal we can see that we have highest rating of 2.
For negative 2 rating 61% decrease in the expected number of cases purchased.
For 1 rating 45% decrease in the expected number of cases purchased.
Zero and 1 rating its 28 % and 14 % decrease in the expected number of cases purchased respectively.

Now for stars we can see that it decreases from obtaining a 4 rating, for 1 rating 34 % decrease in the expected number of cases purchased, 2 rating 18% decrease in the expected number of cases purchased and 3 rating 10% decrease in the expected number of cases purchased.

Model 4:

We will be using Genmod for zero inflated negative binomial, below are the variables that will be used for modelling.

acidindex
labelappeal
imp_stars
i_imp_stars

The model equation is:
Y=B0+B1X+ B2X2+B3X3+B4X4+E

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.8705	0.0499	1.7726	1.9684	1403.01	<.0001
AcidIndex		1	-0.0214	0.0049	-0.0310	-0.0117	18.71	<.0001
LabelAppeal	-2	1	-0.9704	0.0440	-1.0566	-0.8842	487.27	<.0001
LabelAppeal	-1	1	-0.6029	0.0260	-0.6539	-0.5519	536.36	<.0001
LabelAppeal	0	1	-0.3409	0.0236	-0.3872	-0.2945	207.84	<.0001
LabelAppeal	1	1	-0.1574	0.0238	-0.2041	-0.1106	43.55	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
imp_stars	1	1	-0.4068	0.0230	-0.4519	-0.3618	312.88	<.0001
imp_stars	2	1	-0.1999	0.0200	-0.2391	-0.1606	99.53	<.0001
imp_stars	3	1	-0.1046	0.0203	-0.1444	-0.0648	26.56	<.0001
imp_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
i_imp_stars	0	1	0.1854	0.0197	0.1469	0.2239	88.92	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		0	0.0019	0.0000	0.0019	0.0019		

Table: Zero Inflated Negative Binomial Model Analysis Of Maximum Likelihood Parameter Estimates

results of this model are very similar to the model 3
Interpretation of the above model is that the one-unit increase in acid index is a 2 % decrease in the expected number of cases purchased.
Now looking at the label appeal we can see that we have highest rating of 2.
For negative 2 rating 61% decrease in the expected number of cases purchased.

For 1 rating 45% decrease in the expected number of cases purchased.
Zero and 1 rating its 28 % and 14 % decrease in the expected number of cases purchased respectively.

Now for stars we can see that it decreases from obtaining a 4 rating, for 1 rating 34 % decrease in the expected number of cases purchased, 2 rating 18% decrease in the expected number of cases purchased and 3 rating 10% decrease in the expected number of cases purchased.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		41984.4131	
Scaled Deviance		41984.4131	
Pearson Chi-Square	13E3	6016.3408	0.4707
Scaled Pearson X2	13E3	6016.3408	0.4707
Log Likelihood		-20992.2065	
Full Log Likelihood		-20992.2065	
AIC (smaller is better)		42012.4131	
AICC (smaller is better)		42012.4459	
BIC (smaller is better)		42116.8084	

Table: Zero Inflated Negative Binomial Model Criteria For Assessing Goodness Of Fit

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.3657	0.1930	-3.7439	-2.9875	304.21	<.0001
AcidIndex		1	0.4637	0.0240	0.4168	0.5107	374.47	<.0001
i_imp_stars	0	1	-3.4689	0.0828	-3.6311	-3.3067	1757.03	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.

Table: Zero Inflated Negative Binomial Model Analysis of Maximum Likelihood Zero Inflation Parameter Estimates

Model 5:

We will be using prog reg and run a regression model, below are the variables that will be used for modelling.

acidindex
labelappeal
imp_stars
i_imp_stars

The model equation is:
Y=B0+B1X+ B2X2+B3X3+B4X4+E

Root MSE	1.31520	**R-Square**	0.5340
Dependent Mean	3.02907	**Adj R-Sq**	0.5339
Coeff Var	43.41913		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.63526	0.07894	46.05	<.0001
AcidIndex	1	-0.20642	0.00895	-23.06	<.0001
LabelAppeal	1	0.46504	0.01372	33.90	<.0001
imp_stars	1	0.78650	0.01570	50.08	<.0001
i_imp_stars	1	-2.26538	0.02698	-83.96	<.0001

as we can see from the above results that the ADJ r square is good about 0.5339 which assures us that the model is good, after going through the ODS out put we could see that the QQ plots, the residual plots, cook's d and the histogram in which the SAS curve covered all the points. Overall a good model.

MODEL COMPARISON

We have five models which we have built, Model 1, Model 2, Model 3, Model 4 and Model 5.

Model1- **Poisson model:**
Model2- **Negative binomial distribution:**
Model3- **Zero inflated Poisson**
Model4- **Zero inflated negative binomial**
Model5- **Regression using proc reg.**

AIC values:

	Model1	Model2	Model3	Model4
AIC value	45562	45664	41953	42012

We see a little difference between Poisson and negative binomial due to the mean and variance being very close in value. When comparing all the models, I was more impressed by the Regression model as the Adj r square value was pretty high. Since the project requires us to chose one among the first four models
And since that we observed the target value being zero inflated and also comparing the AIC values we chose to continue with the Zero Inflated negative binomial model. We compared the methods using test procedures and found that the zero inflated model fits the best among all other models.

Below is the equation of the model4

The equation of the model selected is below:

P_TARGET = 1.8705 +
- 0.9704 * ACIDINDEX
+ (label Appeal in (-2)) * -0.9704
+(label Appeal in (-2)) * -0.6029
+(label Appeal in (-2)) * -0.3409
+(label Appeal in (-2)) * -0.1574
+(imp_stars in (1) * -0.4068
+ (imp_stars in (1) * -0.1999

+ (imp_stars in (0)) * 0.1854;

MODEL DEPLOYMENT CODE:

```
*///DEPLOYMENT///*;

libname mydata "/sscc/home/n/ngg135/assignment3/" access=readonly;

proc datasets library=mydata;
run;
quit;

data testing;
set mydata.wine_test;


data testing_fixed;
    set testing;

    imp_stars = stars;
    i_imp_stars = 0;
    if missing(imp_stars) then do;
       imp_stars = 2.0;
       i_imp_stars = 1;
    end;



data testing_score;
    set testing_fixed;

    TEMP = -3.3657
```

```
    + AcidIndex * 0.4637
    + (i_imp_stars in (0)) * -3.4689;

   P_SCORE_ZERO = exp(TEMP) / (1 + exp(TEMP));

    temp = 1.8705
    + AcidIndex * -0.0214
    + (LabelAppeal in (-2)) * -0.9704
    + (LabelAppeal in (-1)) * -0.6029
    + (LabelAppeal in (0)) * -0.3409
    + (LabelAppeal in (1)) * -0.1574
    + (imp_stars in (1)) * -0.4068
   + (imp_stars in (2)) * -0.1999
    + (imp_stars in (3)) * -0.1046
    + (i_imp_stars in (0)) * 0.1854;

P_SCORE_ZIP_ALL = exp(TEMP);

P_TARGET = P_SCORE_ZIP_ALL * (1 - P_SCORE_ZERO);

keep index P_TARGET;

data home.Wine_final_prediction_score;
 set testing_score;
 run;
```

SCORED DATA FILE:

Scored data file is attached with the name Wine_final_prediction_score sas7bdat.
This file will have two columns one is the index and p_target

Conclusion:

We developed several models for this project using the data of the wine, we built 5 models and we chose to go ahead with Zero inflated negative binomial, we had a good exposure of different models in this project. We also built the regression model and the model fitted data pretty good with a good adj r square value.
But overall this was a good project, where we were able to build models/deploy and make the model re-usable so that people can re-use our model. Overall we can say that the logistic model and poison models have initial complexity when building out interpretation, however once locked in the interpretation of these feels more natural than other techniques that we used in our course.

SAS CODE:

```
libname mydata "/sscc/home/n/ngg135/assignment3/" access=readonly;

proc datasets library=mydata;
run;
quit;

data training;
set mydata.wine;

proc contents data=training;
run;

proc means data=training ;
run;

proc print data=training (obs=10);
run;

proc means nmiss data=training;
run;

proc means data=training NMISS N;
run;

proc corr data=training;
with target;
run;


proc univariate data=training;
histogram TARGET /normal;
```

```
run;

proc univariate data=training;
histogram acidindex /normal;
run;

proc univariate data=training;
histogram Alcohol /normal;
run;

proc univariate data=training;
histogram Density /normal;
run;


proc univariate data=training;
histogram ResidualSugar /normal;
run;



proc univariate data=training;
   var acidindex alcohol chlorides citricacid density fixedacidity freesulfurdioxide
labelappeal residualsugar stars sulphates totalsulfurdioxide volatileacidity ph;
   histogram;

data imp_training;
   set training;

   imp_alcohol = alcohol;
   i_imp_alcohol = 0;
   if missing(imp_alcohol) then do;
      imp_alcohol = 10.4892363;
      i_imp_alcohol = 1;
   end;

   imp_chlorides = chlorides;
   i_imp_chlorides = 0;
   if missing(imp_chlorides) then do;
      imp_chlorides = 0.0548225;
```

```
    i_imp_chlorides = 1;
end;

imp_freesulfurdioxide = freesulfurdioxide;
i_imp_freesulfurdioxide = 0;
if missing(imp_freesulfurdioxide) then do;
    imp_freesulfurdioxide = 30.8455713;
    i_imp_freesulfurdioxide = 1;
end;

imp_residualsugar = residualsugar;
i_imp_residualsugar = 0;
if missing(imp_residualsugar) then do;
    imp_residualsugar = 5.4187331;
    i_imp_residualsugar = 1;
end;

imp_stars = stars;
i_imp_stars = 0;
if missing(imp_stars) then do;
    imp_stars = 2.0;
    i_imp_stars = 1;
end;

imp_sulphates = sulphates;
i_imp_sulphates = 0;
if missing(imp_sulphates) then do;
    imp_sulphates = 0.5271118;
    i_imp_sulphates = 1;
end;

imp_totalsulfurdioxide = totalsulfurdioxide;
i_imp_totalsulfurdioxide = 0;
if missing(imp_totalsulfurdioxide) then do;
    imp_totalsulfurdioxide = 120.7142326;
   i_imp_totalsulfurdioxide = 1;
end;

imp_ph = ph;
i_imp_ph = 0;
```

```
    if missing(imp_ph) then do;
       imp_ph = 3.2076282;
       i_imp_ph = 1;
    end;


proc univariate data=IMP_training;
histogram IMP_STARS /normal;
run;

proc univariate data=IMP_training;
histogram IMP_PH /normal;
run;

proc corr data=imp_training;
with target;
run;

*///Model Poisson///;

proc genmod data=imp_training;
    class labelappeal imp_stars i_imp_stars;
    model target = acidindex labelappeal imp_stars i_imp_stars / link=log dist=poi;
    output out=imp_training p=pr1;

    *///Negative Binomial///;

proc genmod data=imp_training;
   class labelappeal imp_stars i_imp_stars;
   model target = acidindex labelappeal imp_stars i_imp_stars / link=log dist=nb;
   output out=imp_training p=nbr1;

   *///zero inflared Binomial///;
proc genmod data=imp_training;
    class labelappeal imp_stars i_imp_stars;
    model target = acidindex labelappeal imp_stars i_imp_stars / link=log dist=ZIP;
    zeromodel acidindex i_imp_stars / link=logit;
    output out=imp_training p=zip1;
```

```
    *///zero inflated negative Binomial///;

proc genmod data=imp_training;
 class labelappeal imp_stars i_imp_stars;
   model target = acidindex labelappeal imp_stars i_imp_stars / link=log
dist=ZINB;
    zeromodel acidindex i_imp_stars / link=logit;
    output out=imp_training p=zinb1 pzero=zzinb1;


   ods graphics on;
proc reg data=imp_training;
    model  target = acidindex labelappeal imp_stars i_imp_stars;
    output out=imp_training p=yhat;

    *///Negative Binomial///;

libname mydata "/sscc/home/n/ngg135/assignment3/" access=readonly;

proc datasets library=mydata;
run;
quit;

data testing;
set mydata.wine_test;


data testing_fixed;
    set testing;

    imp_stars = stars;
    i_imp_stars = 0;
    if missing(imp_stars) then do;
       imp_stars = 2.0;
       i_imp_stars = 1;
    end;



data testing_score;
```

```
    set testing_fixed;

    TEMP = -3.3657
    + AcidIndex * 0.4637
    + (i_imp_stars in (0)) * -3.4689;

   P_SCORE_ZERO = exp(TEMP) / (1 + exp(TEMP));

    temp = 1.8705
    + AcidIndex * -0.0214
    + (LabelAppeal in (-2)) * -0.9704
    + (LabelAppeal in (-1)) * -0.6029
    + (LabelAppeal in (0)) * -0.3409
    + (LabelAppeal in (1)) * -0.1574
    + (imp_stars in (1)) * -0.4068
   + (imp_stars in (2)) * -0.1999
    + (imp_stars in (3)) * -0.1046
    + (i_imp_stars in (0)) * 0.1854;

P_SCORE_ZIP_ALL = exp(TEMP);

P_TARGET = P_SCORE_ZIP_ALL * (1 - P_SCORE_ZERO);

keep index P_TARGET ;

data home.Wine_final_prediction_score;
 set testing_score;
 run;
```