Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Wine Sales Project

BINGO BONUS:

(20 Pts) Develop a LOGISTIC/POISSON model (this model is our champion). Explanation of the model and its predictive capabilities are included in this write-up.

(10 Pts) Used SAS Macros for model scoring routine.

Executive Summary & Introduction

Our analysis produced a Hurdle (Logistic+ Poisson) model that represents the number of cases of wine purchased by wine distribution companies after sampling a wine.

The purpose of this assignment is to use Poisson/Negative Binomial regression to develop a model that predicts the number of sample cases of wine purchased by distribution companies after tasting a wine. The sample cases purchased represent the earning potential of the wine. The logic is that the number of sample cases purchased translates to the number of tasting samples provided to restaurants and wine stores around the country. As the number of tasting samples increases, so does the probability of the wine being sold at a high-end restaurant. Theoretically, there is an assumed correlation between sample cases sold and the earning potential of a wine. Our analysis will seek to predict the number sample cases sold based on chemical properties of the wine, an expert rating system, and label marketability to gain insight on the earning potential of a specific wine. With this information, a manufacturer will be able to adjust their wine offerings to maximize sales to wine distribution companies. We will initiate our efforts with exploratory data analysis to understand our data, progress to data manipulation (imputation and transformations) to prepare our data for logistic regression, then finally initiate model development and selection based on model validation criteria (AIC, ROC Curve, Mean Squared Error) to develop the model that best predicts the number of sample cases sold. Once we've selected the best model for our analysis, we will develop a scoring routine that will produce the predicted sample cases sold to wine distribution companies (TARGET).

Data Exploration

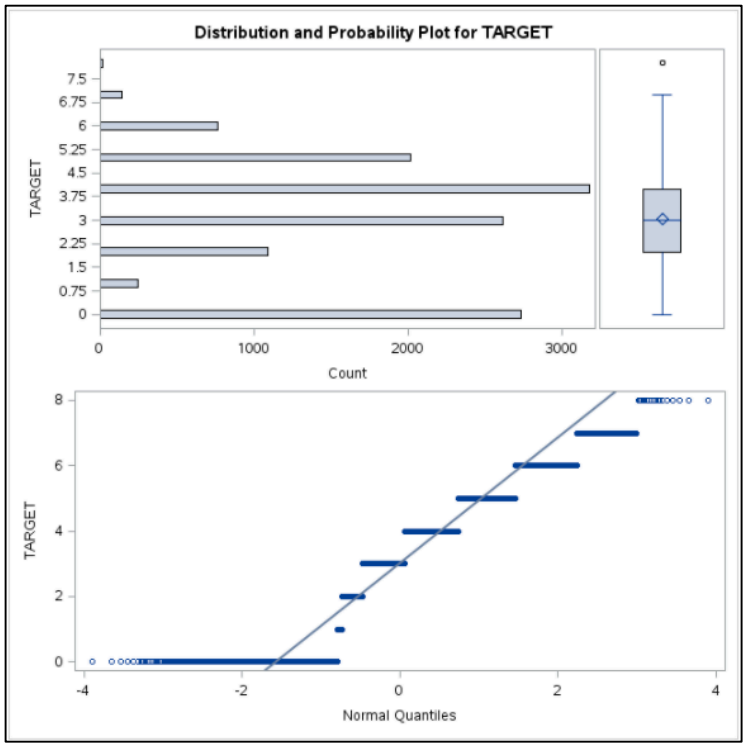
Our data set contains approximately 12,795 records that represent the chemical properties, expert ratings, and label marketability of commercially available wines. The target variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. There are 14 other variables; 12 of the variables (AcidIndex, Alcohol, Chorides, Citric Acid, Density, FixedAcidity, FreeSulfurDioxide, ResidualSugar, Sulphates, TotalSulfurDioxide, VolatileAcidity, and pH) represent chemical properties of the wine while the remaining two variables (STARS and LabelAppeal) represent marketing efforts related to the wine. Of the 14 regressor variables, eight (ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alchohol, and STARS) are missing values and will require imputation.

Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Further analysis of the regressor variables using the PROC CORR, PROC MEANS and PROC FREQ procedures in SAS will help us gain additional details about the variables in our dataset and the possible relationships that exist. Information gained from the data analysis phase will direct our actions for imputation and transformations as we seek to prepare the data for Poisson/Negative Binomial regression. For each variable in the data that we plan on using in the model, we will examine its measures of central tendency (mean, median, mode, standard deviation), and relationships (via Pearson Correlation Coefficients (R)) with the TARGET variable.

TARGET

The target variable TARGET represents the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. TARGET is discrete and has a zero-inflated Poisson/Negative Binomial distribution with mean 3.85, mode 4.00, and variance 1.55 after removing all wines where TARGET is not greater than zero (TARGET > 0). Due to the minimal amount of difference between the mean and variance I believe that this data can be modeled either using a Poisson or Negative Binomial distribution. If the difference between the mean and variance were significant it would indicate extradispersion and we would rely only on a Negative Binomial distribution to model our data. It appears that roughly 22% of the wines have zero sample cases sold. A box plot of TARGET shows minimal outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Using the PROC CORR procedure on the continuous regressor variables and the PROC FREQ procedure with the categorical/ordinal variables allows us to assess the proposed relationships between the predictor variable TARGET and its regressor variables. The PROC CORR procedure produces the Pearson Correlation Coefficient (R) that either confirms or denies our hypothesis regarding the relationship between the regressor variable and target variable. Utilizing the PROC FREQ procedure to assess relationship between the target and categorical/ordinal variables is a little more difficult; this method compares the breakdown between the variable's categories against the options present in the target variable.

TARGET Relationships with Regressor Variables

Variable	Description	Proposed Relationship	Actual Relationship
FixedAcidity	Fixed Acidity of Wine	Unknown	R = -0.049; minimal negative corollary relationship
VolatileAcidity	Volatile Acid Content of Wine	Unknown	R = -0.089; minimal negative corollary relationship
CitricAcid	Citric Acid Concentration of Wine	Unknown	R = 0.009; minimal positive corollary relationship
ResidualSugar	Residual Sugar Content of Wine	Unknown	R = 0.016; minimal positive corollary relationship
Chlorides	Chloride Content of Wine	Unknown	R = -0.038; minimal negative corollary relationship
FreeSulfurDioxide	Free Sulfur Dioxide Content	Negative, more sulphates means less sales. Some people have allergic reactions to sulphates.	0.044; minimal positive corollary relationship
TotalSulfurDioxide	Total Sulfur Dioxide Content	Negative, more sulphates means less sales. Some people have allergic reactions to sulphates.	R = 0.051; minimal positive corollary relationship
Density	Density of Wine	Unknown	R = -0.036; ; minimal negative corollary relationship
pH	pH of Wine	Unknown	R = -0.009; minimal negative corollary relationship
Sulphates	Sulfate content of Wine	Unknown	R = -0.039; ; minimal negative corollary relationship
Alcohol	Alcohol Content of Wine	Unknown	R = 0.062; minimal positive corollary relationship
LabelAppeal	Marketing score indicating the appeal of label design for consumers.	Many consumers purchase based on on the visual appeal of the wine label. Higher numbers suggest better sales.	R = 0.357; strong positive corollary relationship
AcidIndex	Proprietary method of testing told acidity of wine by using a weighted average.	Unknown	R = -0.246, strong negative corollary relationship
STARS	Wine rating by a team of experts. 4 stars = Excellent, 1 Star = Poor.	A higher number of stars suggest better sales.	R = 0.559; strong positive corollary relationship

Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Due to the zero-inflated distribution of the predictor variable TARGET, we have separate the target variable into two separate derivative variables, TARGET_FLAG and TARGET_AMT, for modeling and assessment purposes. TARGET_FLAG represents the probability of any sample cases of wine being sold with TARGET_FLAG = 1 meaning sample cases were sold and TARGET_FLAG = 0 meaning no sample cases were sold. TARGET_AMT represents the number of sample cases of wine sold if any sample cases of wine are sold (TARGET_FLAG = 1). The relationship between the target composite variables TARGET_FLAG and TARGET_AMT is displayed to show each regressor probability of affecting the sale of sample cases and affect upon the number of cases sold.

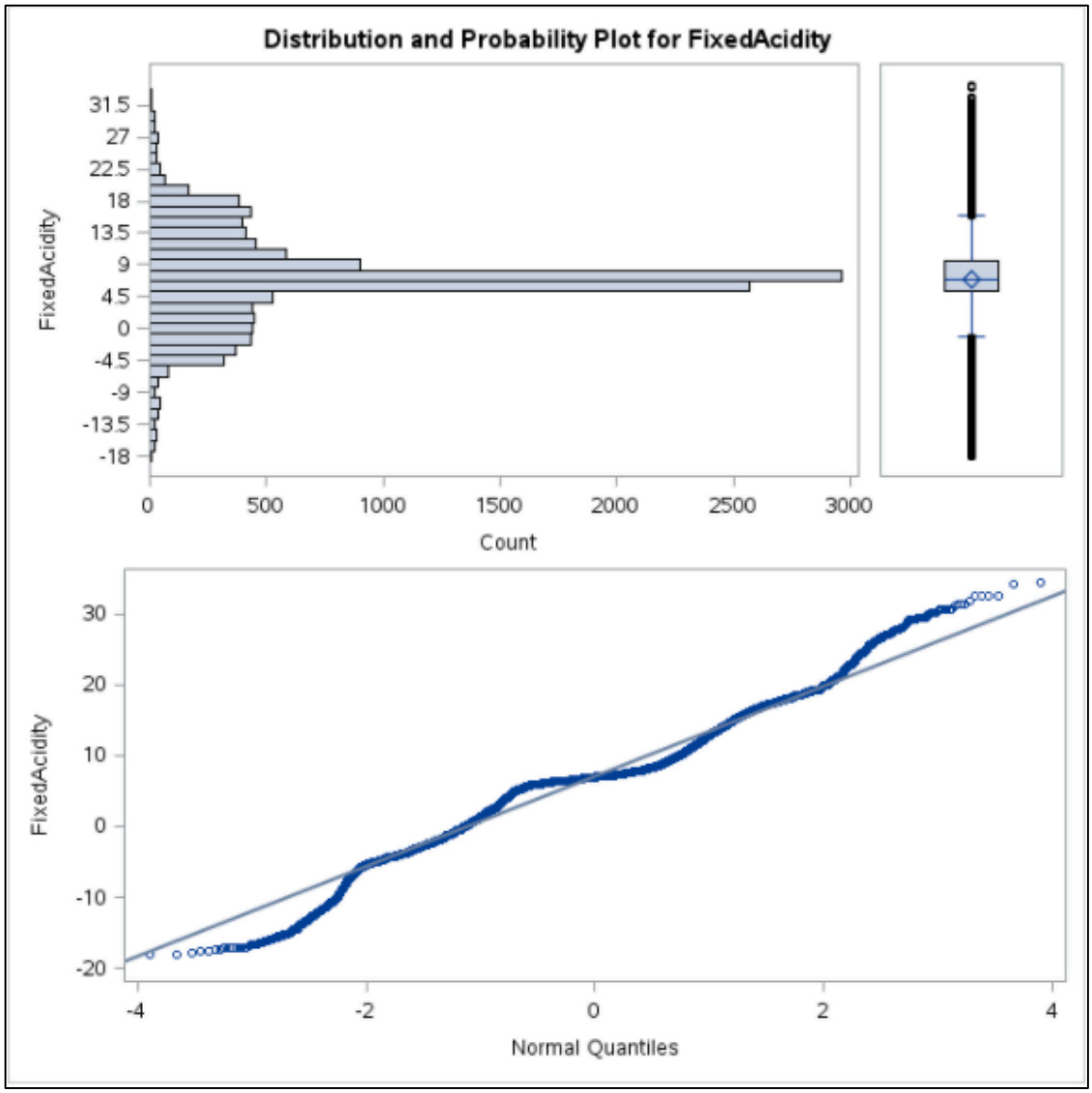
TARGET_FLAG & TARGET_AMT Relationships with Regressor Variables

Variable	TARGET_FLAG	TARGET_AMT
FixedAcidity	R = -0.054; minimal negative corollary relationship	R = -0.09; minimal negative corollary relationship
VolatileAcidity	R = -0.081; minimal negative corollary relationship	R = -0.044; minimal negative corollary relationship
CitricAcid	R = 0.006; minimal positive corollary relationship	R = 0.008; ; minimal positive corollary relationship
ResidualSugar	R = 0.022; ; minimal positive corollary relationship	R = -0.003; ; minimal negative corollary relationship
Chlorides	R = -0.035; minimal negative corollary relationship	R = -0.02; minimal negative corollary relationship
FreeSulfurDioxide	R = 0.045; ; minimal positive corollary relationship	R = 0.013; ; minimal positive corollary relationship
TotalSulfurDioxide	R = 0.08; positive corollary relationship	R = -0.028; minimal negative corollary relationship
Density	R = -0.021; minimal negative corollary relationship	R = -0.036; minimal negative corollary relationship
pH	R = -0.03; minimal negative corollary relationship	R = 0.029; positive corollary relationship
Sulphates	R = -0.047; minimal negative corollary relationship	R = -0.001; minimal negative corollary relationship
Alcohol	R = 0.008; ; minimal positive corollary relationship	R = 0.108; positive corollary relationship
LabelAppeal	R = -0.005; minimal negative corollary relationship	R = 0.711; strong positive corollary relationship
AcidIndex	R = -0.268; negative corollary relationship	R = -0.063; minimal negative corollary relationship
STARS	R = 0.286; positive corollary relationship	R = 0.52; strong positive corollary relationship

Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

FixedAcidity

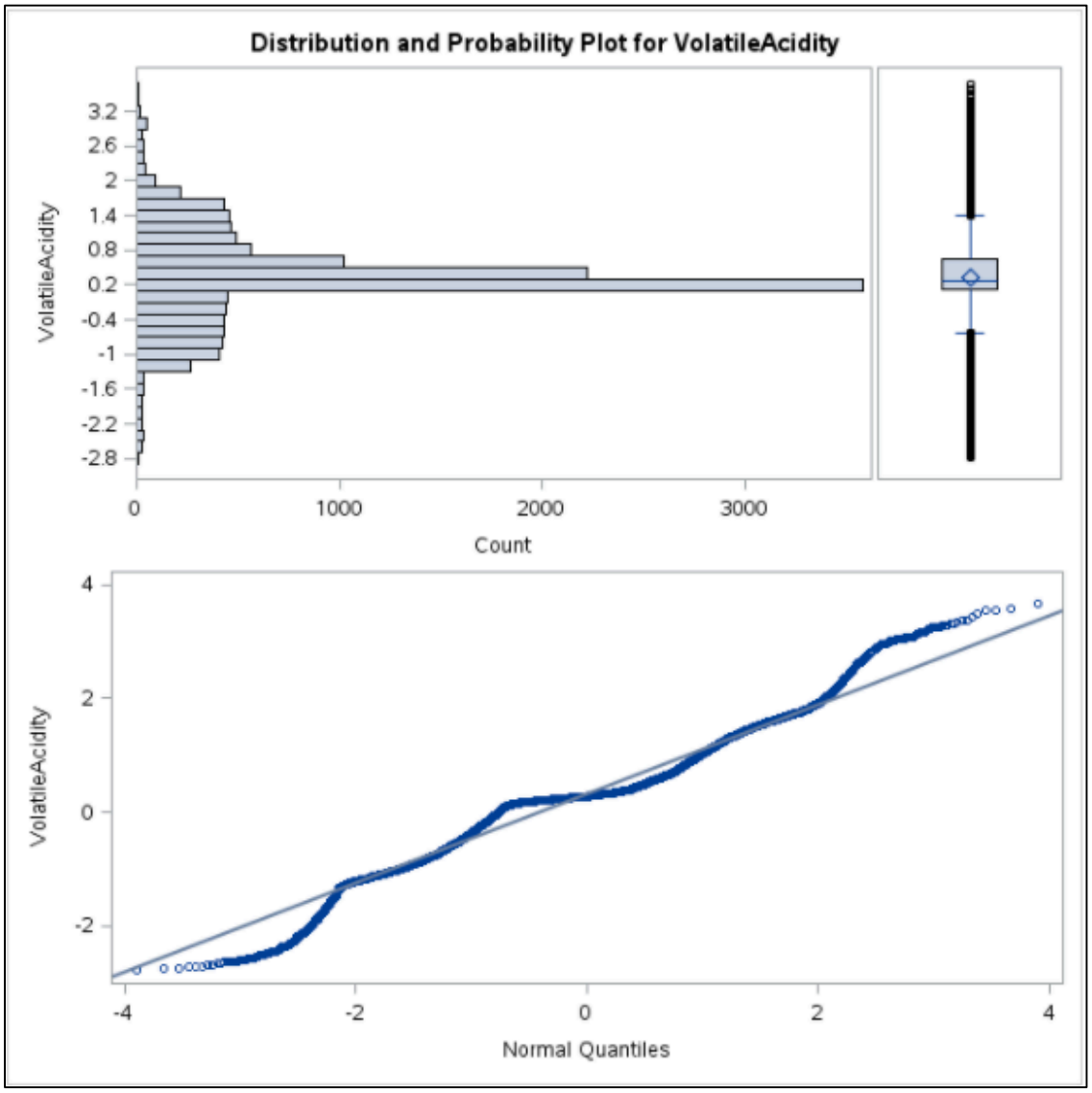
The regressor variable FixedAcidity is a continuous variable with mean 7.1, median 6.9, mode 6.8, and standard deviation 6.32. FixedAcidity represents the amount of fixed acidity in a wine. A frequency plot of the variable represents a normal distribution with the highest count of values occurring near the mean. A review of the quantiles plot shows only a slight departure from normality at the extreme lower and higher values. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. FixedAcidity appears to have minimal negative corollary effect on the TARGET which can be interpreted as wines with higher levels of fixed acidity don't sell as well.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

VolatileAcidity

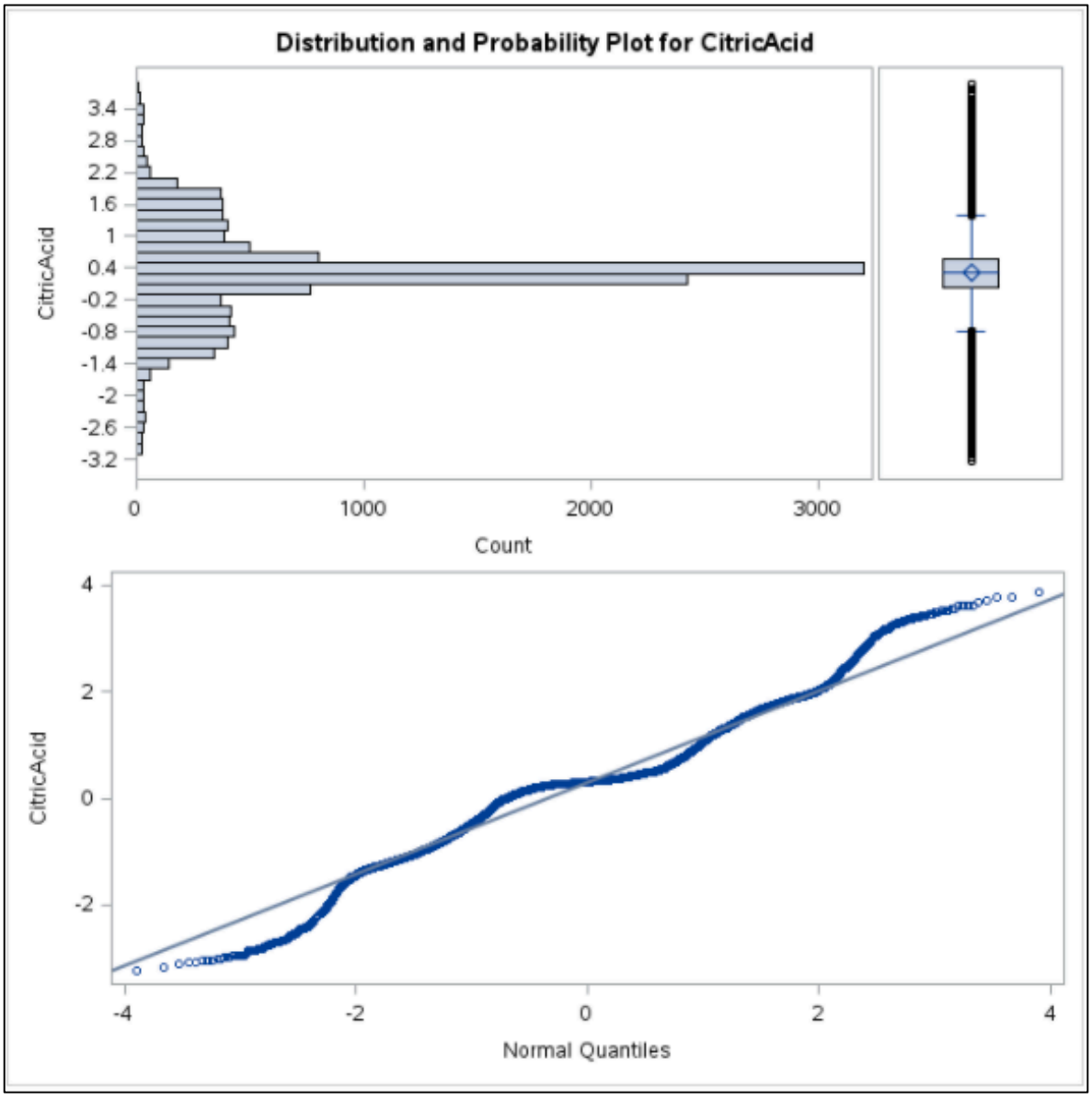
The regressor variable VolatileAcidity is a continuous variable with mean 0.32, median 0.28, mode 0.28, and standard deviation 0.78. VolatileAcidity represents the amount of volatile acidity in a wine. A frequency plot of the variable represents a normal distribution with the highest count of values occurring near the mean. A review of the quantiles plot shows only a slight departure from normality at the extreme lower and higher values. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. VolatileAcidity appears to have minimal negative corollary effect on the TARGET which can be interpreted as wines with higher levels of volatile acidity don't sell as well.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

CitricAcid

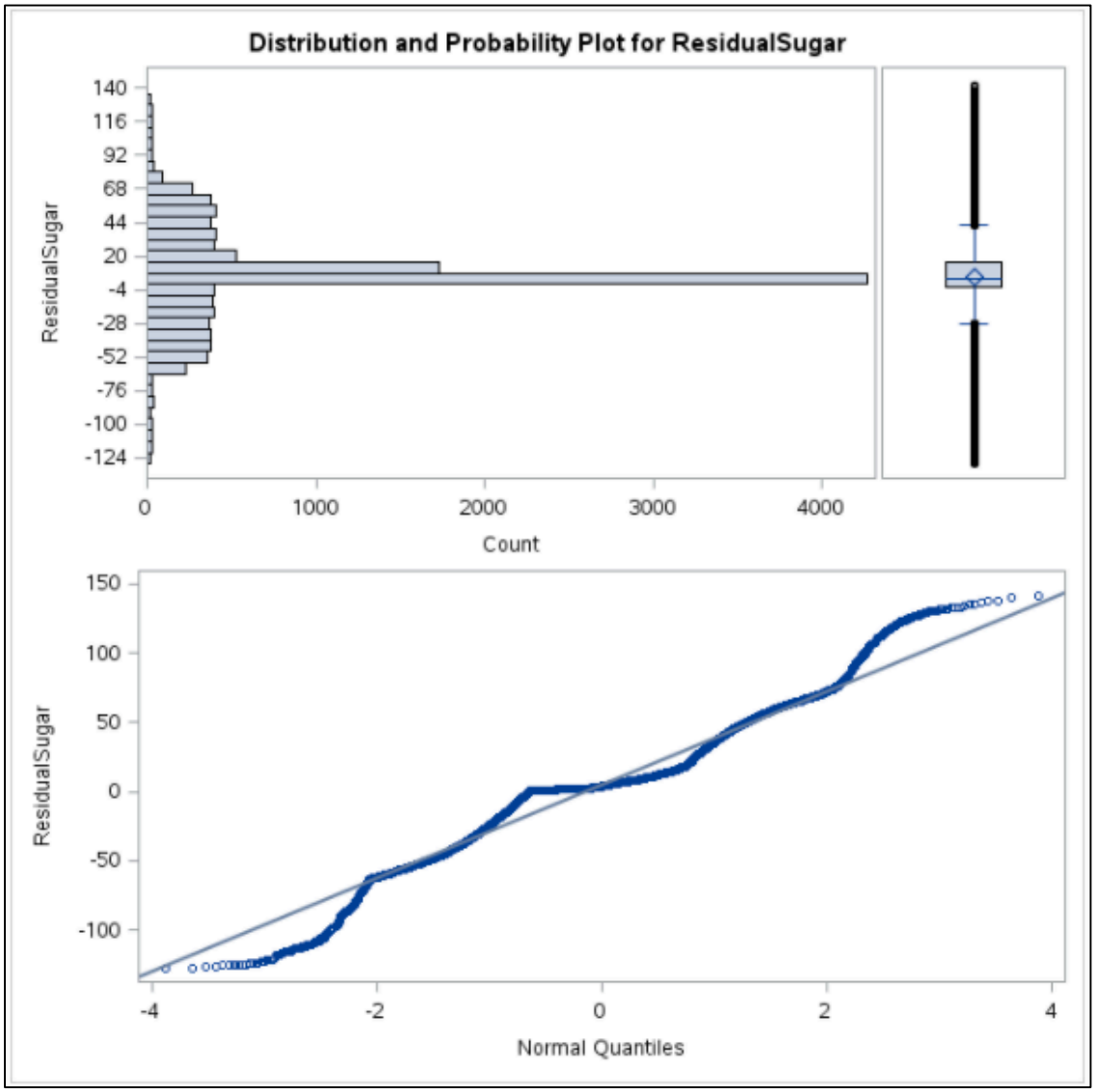
The regressor variable CitricAcid is a continuous variable with mean 0.31, median 0.31, mode 0.30, and standard deviation 0.86. CitricAcid represents the concentration of citric acid in a wine. A frequency plot of the variable represents a normal distribution with the highest count of values occurring near the mean. A review of the quantiles plot shows only a slight departure from normality at the extreme lower and higher values. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. Citric Acid appears to have a minimal positive corollary effect on the predictor variable which can be interpreted as wines with higher levels of citric acid sell better. This contrasts with the minimal negative corollary effect we observed with *FixedAcidity* and *VolatileAcidity*.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

ResidualSugar

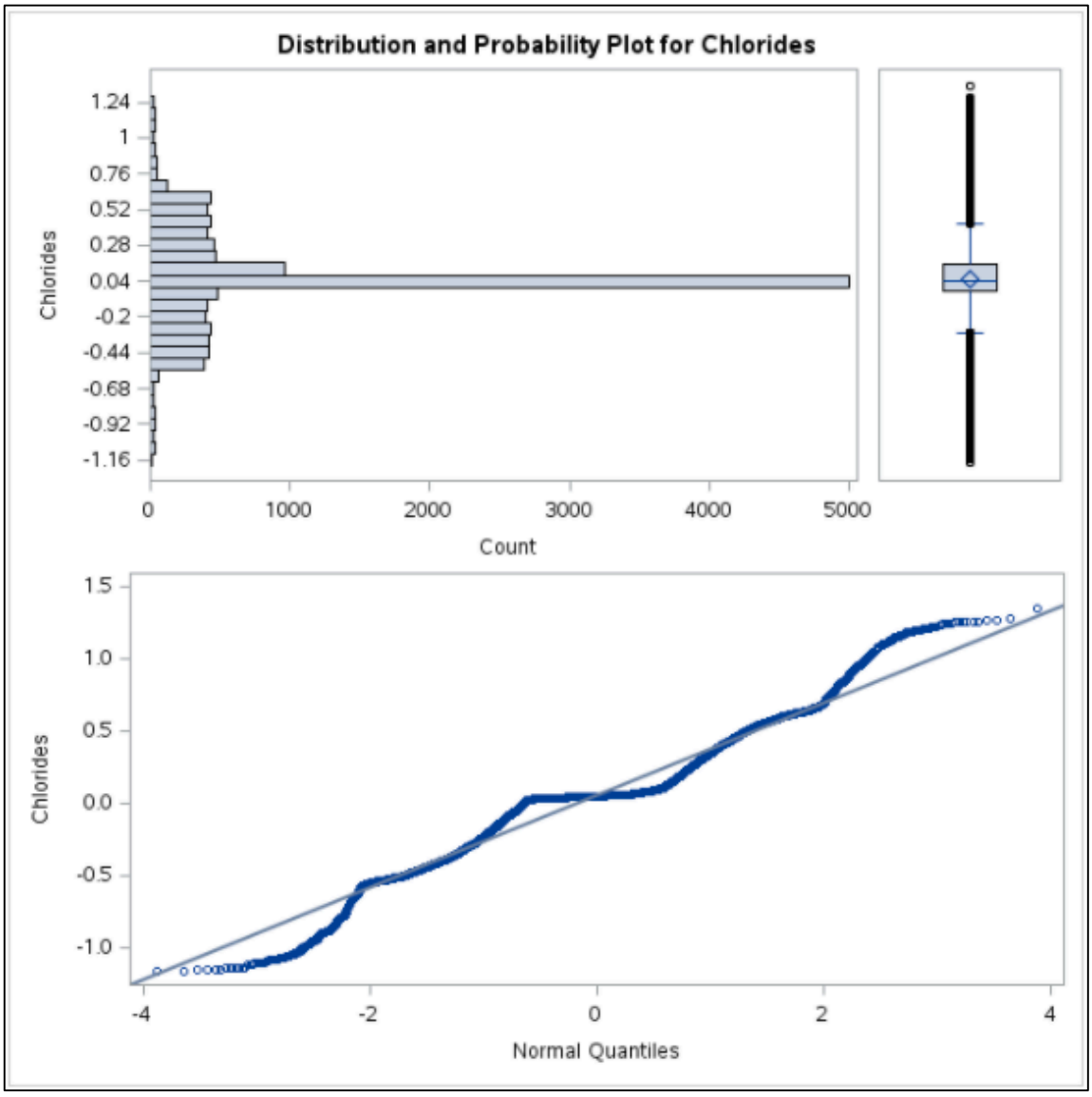
The regressor variable ResidualSugar is a continuous variable with mean 5.42, median 3.90, mode 1.40, and standard deviation 33.75. ResidualSugar represents the concentration of residual sugars in a wine. A frequency plot of the variable represents a normal distribution with the highest count of values occurring near the mean. A review of the quantiles plot shows only a slight departure from normality at the extreme lower and higher values. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. ResidualSugar appears to have a minimal positive corollary effect on the predictor variable which can be interpreted as wines with higher levels of residual sugar sell better. ResidualSugar is missing 616 out of 12795 records and will require imputation for our analysis.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Chlorides

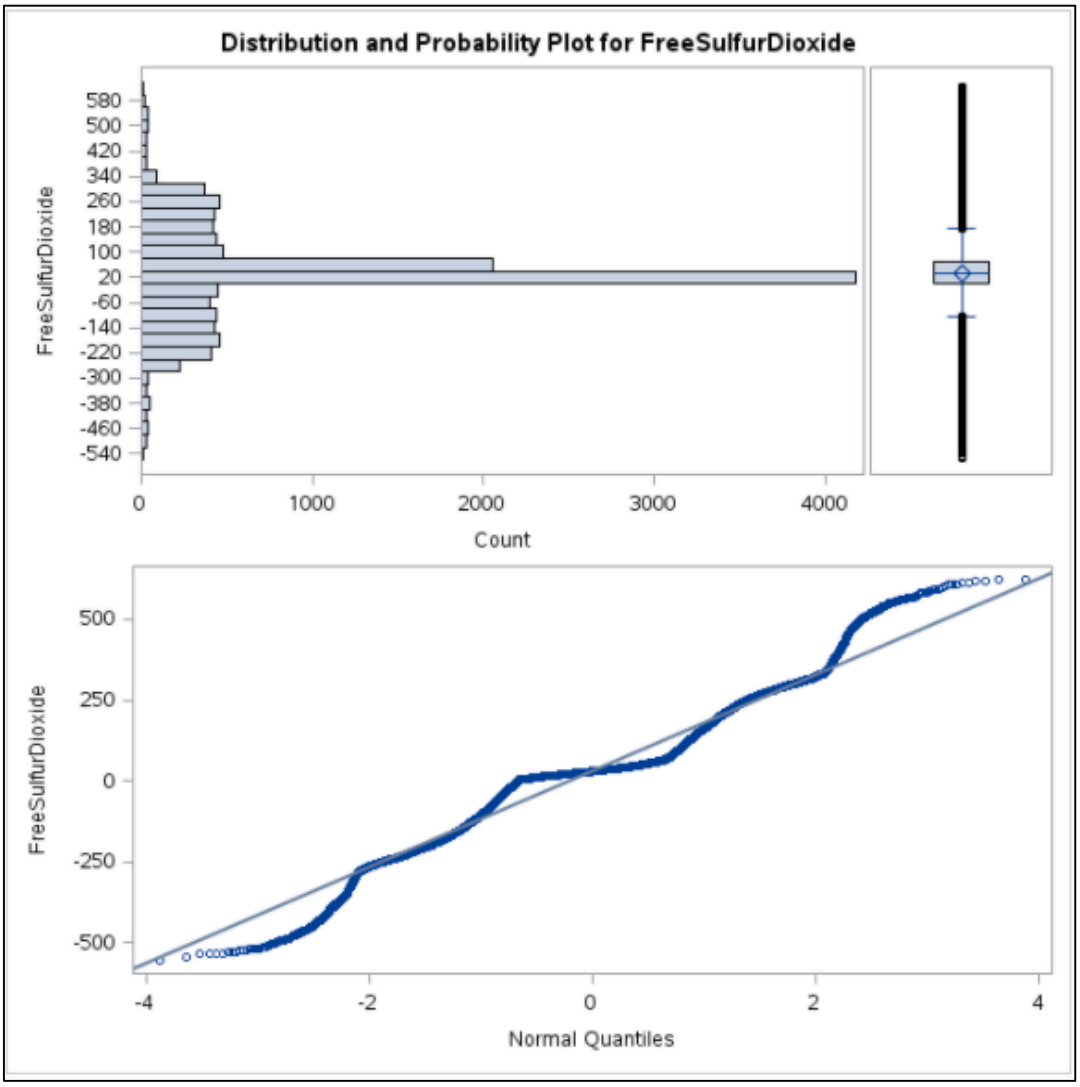
The regressor variable Chlorides is a continuous variable with mean 0.05, median 0.05, mode 0.04, and standard deviation 0.32. Chlorides represents the concentration of chlorides in a wine. A frequency plot of the variable represents a normal distribution with the highest count of values occurring near the mean. A review of the quantiles plot shows only a slight departure from normality at the extreme lower and higher values. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. Chlorides appears to have a minimal negative corollary effect on the predictor variable which can be interpreted as wines with higher levels of chlorides don't sell as well. Chlorides is missing 638 out of 12795 records and will require imputation for our analysis.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

FreeSulfurdioxide

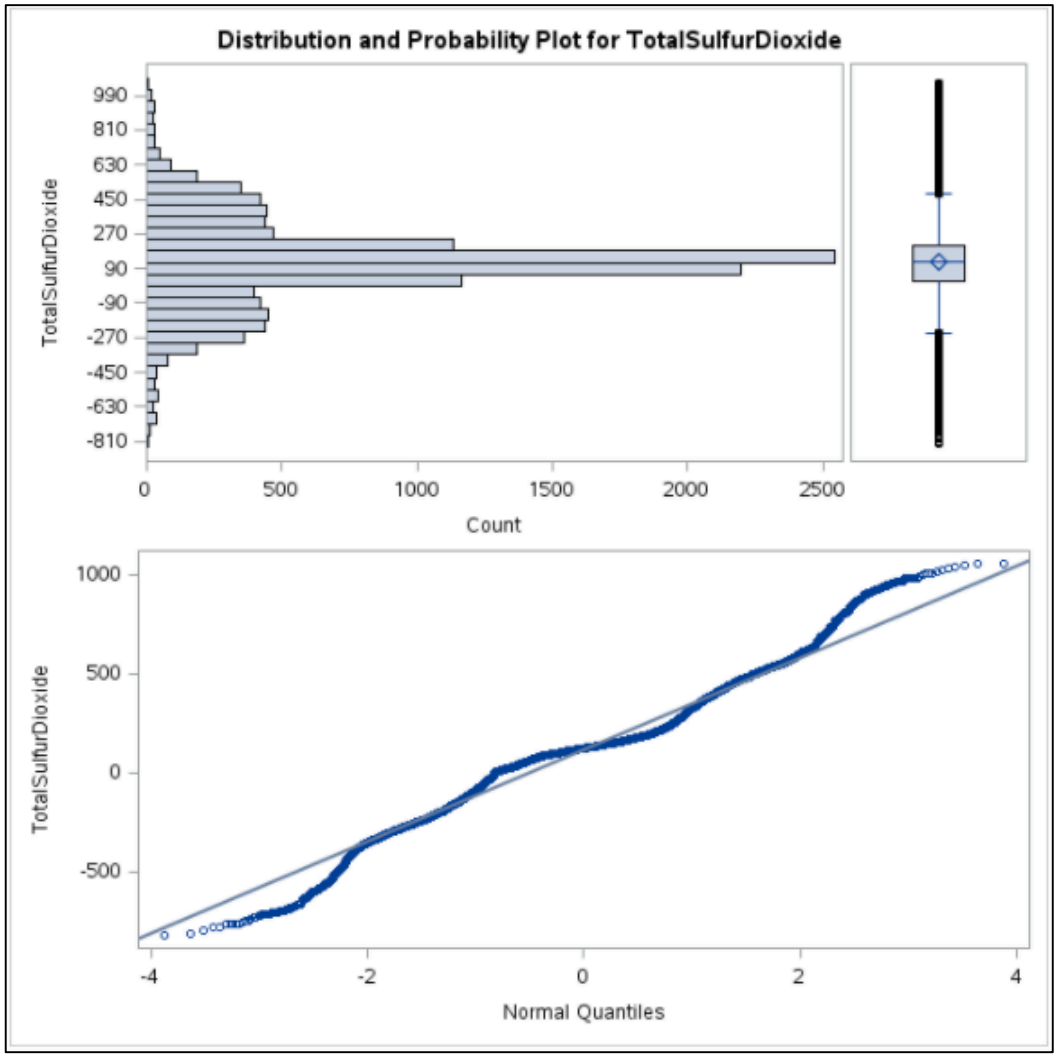
The regressor variable FreeSulfurDioxide is a continuous variable with mean 30.85, median 30.00, mode 29.00, and standard deviation 148.71. FreeSulfurDioxide represents the concentration of sulfur dioxide in a wine. A frequency plot of the variable represents a normal distribution with the highest count of values occurring near the mean. A review of the quantiles plot shows only a slight departure from normality at the extreme lower and higher values. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. FreeSulfurDioxide appears to have a minimal positive corollary effect on the predictor variable which can be interpreted as wines with higher levels of sulfur dioxide sell better. Chlorides is missing 647 out of 12795 records and will require imputation for our analysis. I believe this variable may be related to the regressor variable TotalSulfurDioxide.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

TotalSulfurDioxide

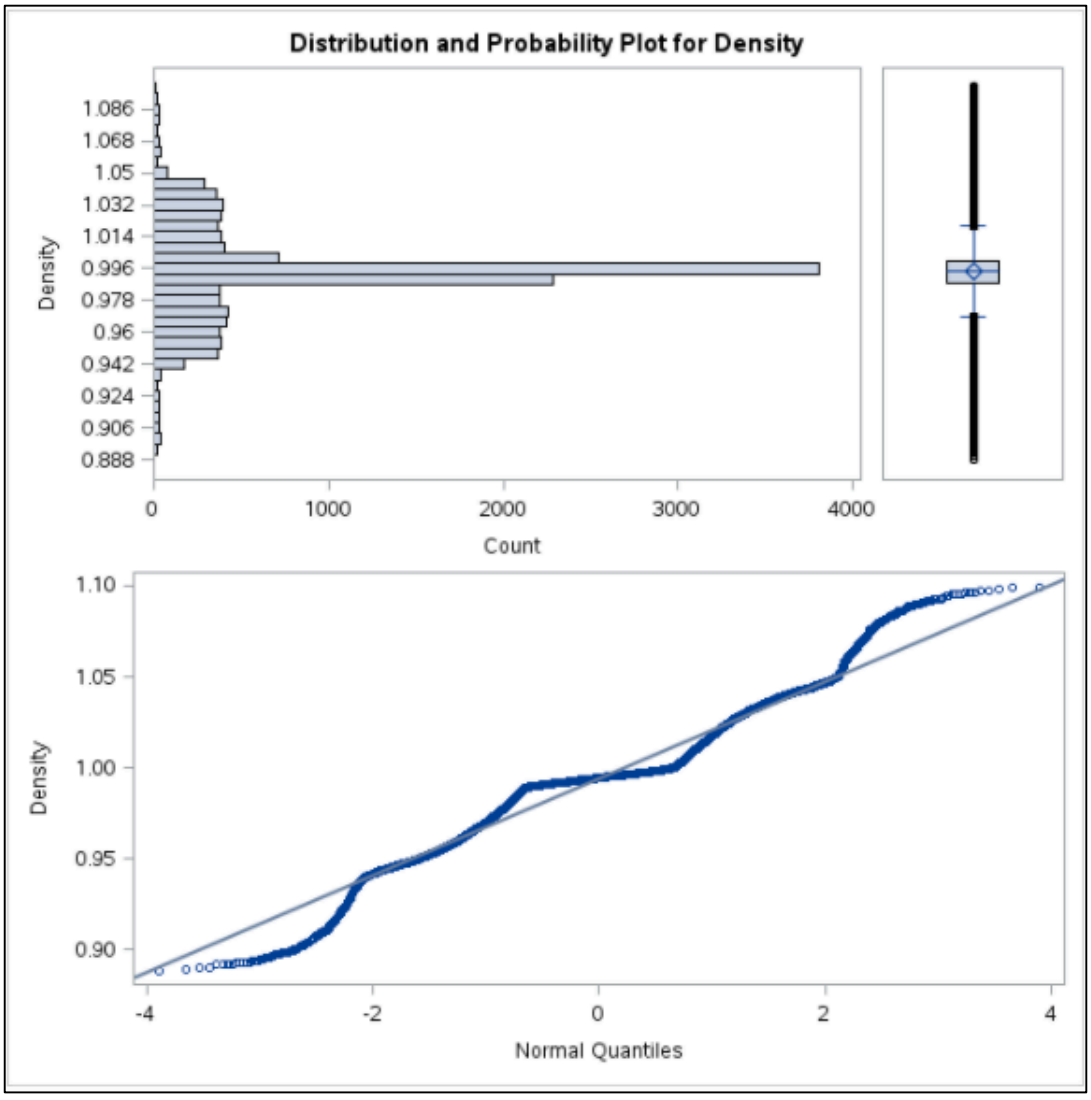
The regressor variable TotalSulfurDioxide is a continuous variable with mean 120.71, median 123.00, mode 125.00, and standard deviation 231.91. TotalSulfurDioxide represents the total concentration of sulfur dioxide in a wine. A frequency plot of the variable represents a normal distribution with the highest count of values occurring near the mean. A review of the quantiles plot shows only a slight departure from normality at the extreme lower and higher values. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. TotalSulfurDioxide appears to have a minimal positive corollary effect on the predictor variable which can be interpreted as wines with higher levels of sulfur dioxide sell better. Chlorides is missing 647 out of 12795 records and will require imputation for our analysis. I believe this variable may be related to the regressor variable FreeSulfurDioxide.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Density

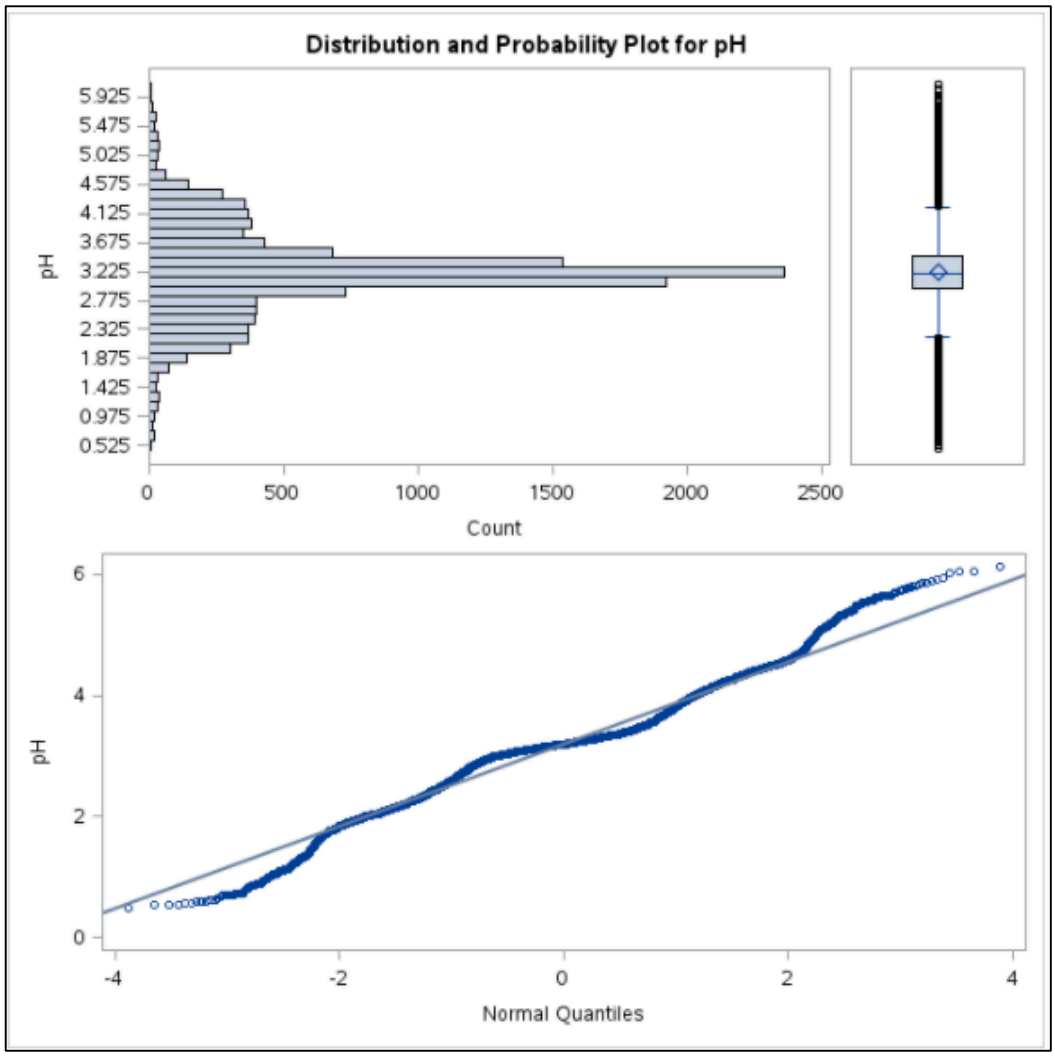
The regressor variable Density is a continuous variable with mean 0.99, median 0.99, mode 1.00, and standard deviation 0.03. Density represents the thickness a wine (total dissolved solids). A frequency plot of the variable represents a normal distribution with the highest count of values occurring near the mean. A review of the quantiles plot shows only a slight departure from normality at the extreme lower and higher values. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. Density appears to have a minimal negative corollary effect on the predictor variable which can be interpreted as wines with higher densities don't sell as well. I'm noticing that the frequency distribution plots of many of the regressors are the same, which suggest that wine makers are targeting specific chemical properties that are known to be popular.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

pH

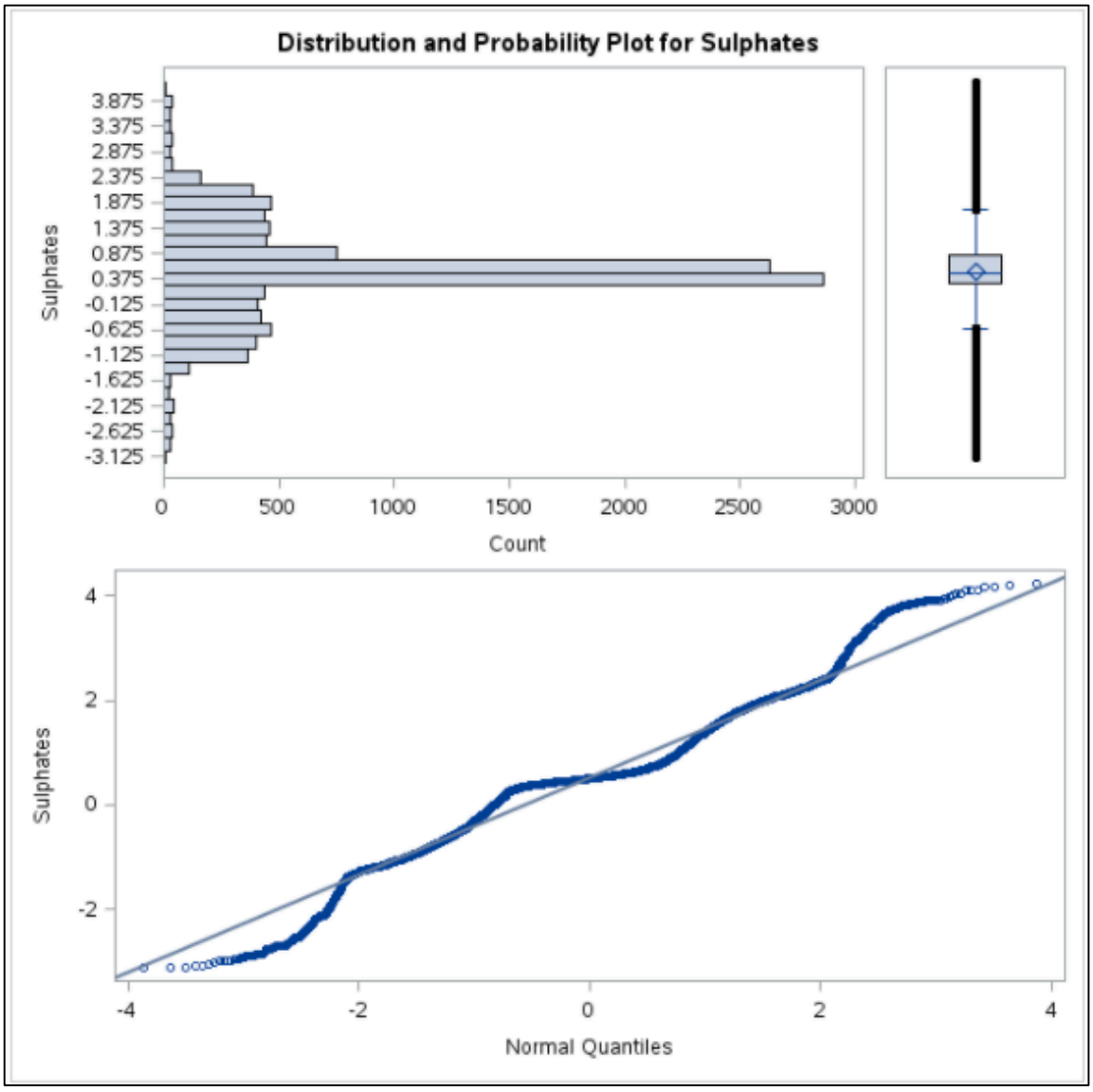
The regressor variable pH is a continuous variable with a range of 0 - 14, mean 3.21, median 3.20, mode 3.16, and standard deviation 0.68. pH represents the concentration of hydrogen ions in solution with anything between 0 – 7 being acidic and anything from 7 – 14 being basic. A frequency plot of the variable represents a normal distribution with the highest count of values occurring near the mean. It appears that all wines that are sold for distribution are highly acidic as their pH values are between 0 and 7. A review of the quantiles plot shows only a slight departure from normality at the extreme lower and higher values. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. pH appears to have a minimal negative corollary effect on the predictor variable which can be interpreted as wines with higher acidities don't sell as well. I believe that this variable may be related to acidindex, citricacid, fixedacidity, and volatile acidity. pH is missing 395 of its records and will require imputation for analysis.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Sulphates

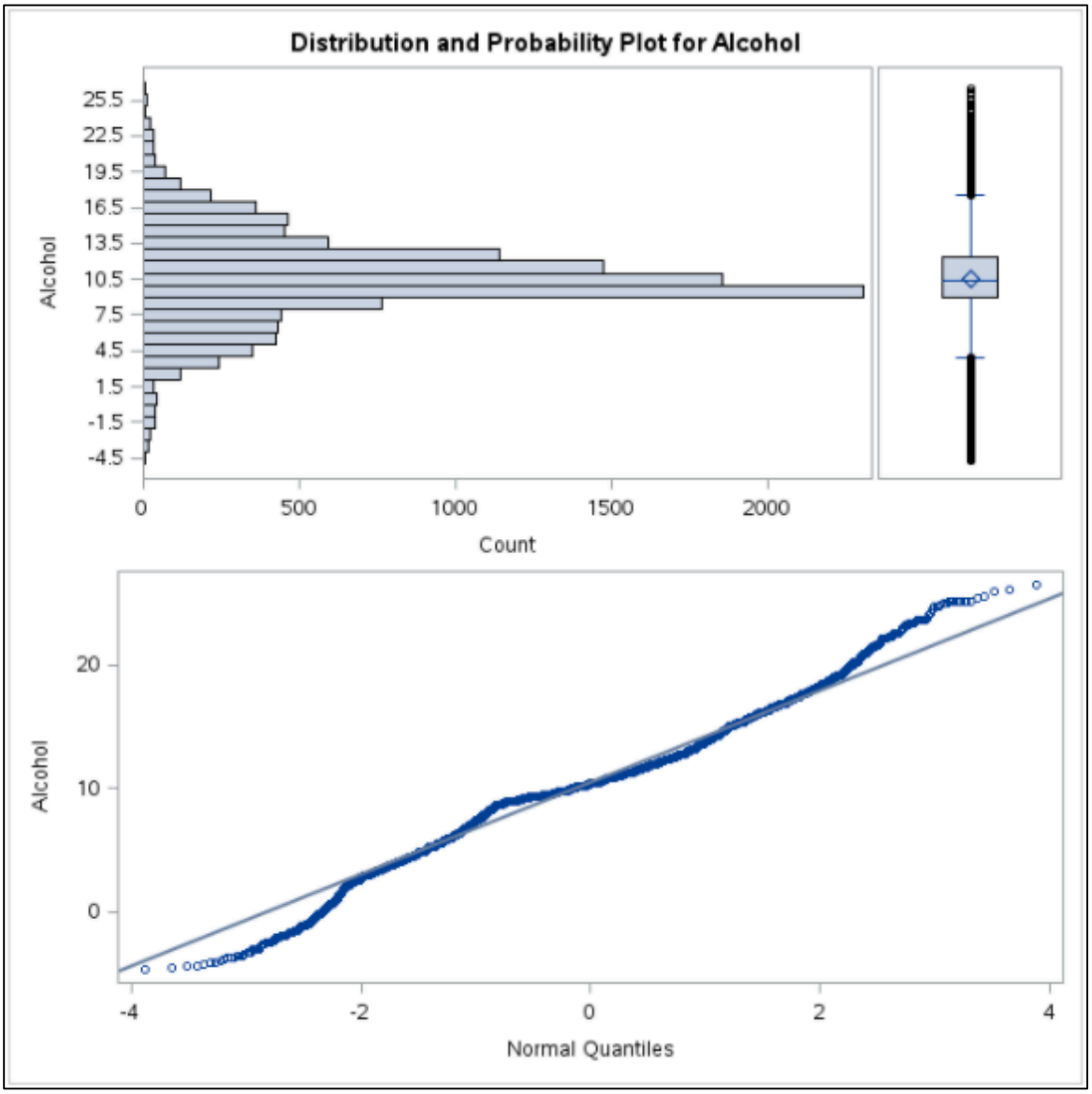
The regressor variable Sulphates is a continuous variable with mean 0.53, median 0.50, mode 0.50, and standard deviation 0.93. Sulphates represents the sulfate concentration of the sample wine. A frequency plot of the variable represents a normal distribution with the highest count of values occurring near the mean. A review of the quantiles plot shows only a slight departure from normality at the extreme lower and higher values. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. Sulphates appears to have a minimal negative corollary effect on the predictor variable which can be interpreted as wines with higher concentrations of sulfates don't sell as well. I believe that this variable may be related to freesulfurdioxide and totalsulfurdioxide. Sulphates is missing 1210 of its 12795 records and will require imputation for analysis.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Alcohol

The regressor variable Alcohol is a continuous variable with mean 10.49, median 10.40, mode 9.4, and standard deviation 3.73. Alcohol represents the alcohol by volume concentration of the sample wine. A frequency plot of the variable represents a normal distribution with the highest count of values occurring near the mean. A review of the quantiles plot shows only a slight departure from normality at the extreme lower and higher values. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. Alcohol appears to have a minimal positive corollary effect on the predictor variable which can be interpreted as wines with higher alcohol concentrations sell better. This variable is missing 653 of its records and will require imputation.



Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

LabelAppeal

The regressor variable LabelAppeal is a discrete ordinal variable with mean -0.01, median 0.00 mode 0.00, and standard deviation 0.89. LabelAppeal represents the appeal of a wine labels design to consumers. A positive value represents label appeal while negative numbers suggest that customers don't like the label design. LabelAppeal appears to have a strong positive corollary effect on the predictor variable which can be interpreted as wines with marketable labels sell better.

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of LabelAppeal by TARGET

LabelAppeal	TARGET 0	1	2	3	4	5	6	7	8	Total
-2	102 0.80 20.24 3.73	136 1.06 26.98 55.74	177 1.38 35.12 16.22	74 0.58 14.68 2.83	14 0.11 2.78 0.44	1 0.01 0.20 0.05	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	504 3.94
-1	671 5.24 21.40 24.54	89 0.70 2.84 36.48	755 5.90 24.08 69.20	1118 8.74 35.65 42.82	413 3.23 13.17 13.00	88 0.69 2.81 4.37	2 0.02 0.06 0.26	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3136 24.51
0	1193 9.32 21.24 43.64	19 0.15 0.34 7.79	152 1.19 2.71 13.93	1347 10.53 23.98 51.59	1972 15.41 35.11 62.07	775 6.06 13.80 38.48	155 1.21 2.76 20.26	4 0.03 0.07 2.82	0 0.00 0.00 0.00	5617 43.90
1	660 5.16 21.65 24.14	0 0.00 0.00 0.00	7 0.05 0.23 0.64	70 0.55 2.30 2.68	765 5.98 25.10 24.08	1040 8.13 34.12 51.64	425 3.32 13.94 55.56	79 0.62 2.59 55.63	2 0.02 0.07 11.76	3048 23.82
2	108 0.84 22.04 3.95	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 0.02 0.41 0.08	13 0.10 2.65 0.41	110 0.86 22.45 5.46	183 1.43 37.35 23.92	59 0.46 12.04 41.55	15 0.12 3.06 88.24	490 3.83
Total	2734 21.37	244 1.91	1091 8.53	2611 20.41	3177 24.83	2014 15.74	765 5.98	142 1.11	17 0.13	12795 100.00

Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

AcidIndex

The regressor variable AcidIndex is a discrete ordinal variable with mean 2.04, median 8.00, mode 2.00, and standard deviation 1.32. AcidIndex is proprietary method of testing the total acidity of a wine by using a weighted average. AcidIndex appears to have a strong positive corollary effect on the predictor variable which can be interpreted as adding validity to the worth of this regressor.

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of AcidIndex by TARGET

AcidIndex	TARGET 0	1	2	3	4	5	6	7	8	Total
4	1	0	0	0	0	2	0	0	0	3
	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02
	33.33	0.00	0.00	0.00	0.00	66.67	0.00	0.00	0.00	
	0.04	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	
5	11	0	3	16	21	15	8	1	0	75
	0.09	0.00	0.02	0.13	0.16	0.12	0.06	0.01	0.00	0.59
	14.67	0.00	4.00	21.33	28.00	20.00	10.67	1.33	0.00	
	0.40	0.00	0.27	0.61	0.66	0.74	1.05	0.70	0.00	
6	170	18	91	259	318	234	86	21	0	1197
	1.33	0.14	0.71	2.02	2.49	1.83	0.67	0.16	0.00	9.36
	14.20	1.50	7.60	21.64	26.57	19.55	7.18	1.75	0.00	
	6.22	7.38	8.34	9.92	10.01	11.62	11.24	14.79	0.00	
7	727	88	461	996	1324	867	354	57	4	4878
	5.68	0.69	3.60	7.78	10.35	6.78	2.77	0.45	0.03	38.12
	14.90	1.80	9.45	20.42	27.14	17.77	7.26	1.17	0.08	
	26.59	36.07	42.25	38.15	41.67	43.05	46.27	40.14	23.53	
8	782	89	351	904	1090	632	238	44	12	4142
	6.11	0.70	2.74	7.07	8.52	4.94	1.86	0.34	0.09	32.37
	18.88	2.15	8.47	21.83	26.32	15.26	5.75	1.06	0.29	
	28.60	36.48	32.17	34.62	34.31	31.38	31.11	30.99	70.59	
9	437	28	126	285	305	174	57	14	1	1427
	3.42	0.22	0.98	2.23	2.38	1.36	0.45	0.11	0.01	11.15
	30.62	1.96	8.83	19.97	21.37	12.19	3.99	0.98	0.07	
	15.98	11.48	11.55	10.92	9.60	8.64	7.45	9.86	5.88	
10	257	13	40	102	73	54	9	3	0	551
	2.01	0.10	0.31	0.80	0.57	0.42	0.07	0.02	0.00	4.31
	46.64	2.36	7.26	18.51	13.25	9.80	1.63	0.54	0.00	
	9.40	5.33	3.67	3.91	2.30	2.68	1.18	2.11	0.00	
11	169	6	12	29	23	16	2	1	0	258
	1.32	0.05	0.09	0.23	0.18	0.13	0.02	0.01	0.00	2.02
	65.50	2.33	4.65	11.24	8.91	6.20	0.78	0.39	0.00	
	6.18	2.46	1.10	1.11	0.72	0.79	0.26	0.70	0.00	
12	92	2	3	9	8	8	5	1	0	128
	0.72	0.02	0.02	0.07	0.06	0.06	0.04	0.01	0.00	1.00
	71.88	1.56	2.34	7.03	6.25	6.25	3.91	0.78	0.00	
	3.37	0.82	0.27	0.34	0.25	0.40	0.65	0.70	0.00	
13	42	0	2	4	11	7	3	0	0	69
	0.33	0.00	0.02	0.03	0.09	0.05	0.02	0.00	0.00	0.54
	60.87	0.00	2.90	5.80	15.94	10.14	4.35	0.00	0.00	
	1.54	0.00	0.18	0.15	0.35	0.35	0.39	0.00	0.00	
14	32	0	2	6	2	2	3	0	0	47
	0.25	0.00	0.02	0.05	0.02	0.02	0.02	0.00	0.00	0.37
	68.09	0.00	4.26	12.77	4.26	4.26	6.38	0.00	0.00	
	1.17	0.00	0.18	0.23	0.06	0.10	0.39	0.00	0.00	
15	4	0	0	1	2	1	0	0	0	8
	0.03	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.06
	50.00	0.00	0.00	12.50	25.00	12.50	0.00	0.00	0.00	
	0.15	0.00	0.00	0.04	0.06	0.05	0.00	0.00	0.00	
16	4	0	0	0	0	1	0	0	0	5
	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.04
	80.00	0.00	0.00	0.00	0.00	20.00	0.00	0.00	0.00	
	0.15	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	
17	6	0	0	0	0	1	0	0	0	7
	0.05	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.05
	85.71	0.00	0.00	0.00	0.00	14.29	0.00	0.00	0.00	
	0.22	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	
Total	2734	244	1091	2611	3177	2014	765	142	17	12795
	21.37	1.91	8.53	20.41	24.83	15.74	5.98	1.11	0.13	100.00

Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

STARS

The regressor variable STARS is a discrete ordinal variable with mean 2.04, median 2.00, mode 2.00, and standard deviation 0.90. STARS is the number of stars the wine was rated by a panel of experts with more stars indicating a better review, thus a higher quality wine. STARS appears to have a strongest positive corollary effect of all the variables in the data set on the TARGET which can be interpreted as adding validity to the worth of this regressor. STARS is missing 3359 of 12795 possible requires and will require imputation for this analysis.

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of STARS by TARGET

STARS	TARGET 0	1	2	3	4	5	6	7	8	Total
.	2038 15.93 60.67 74.54	126 0.98 3.75 51.64	335 2.62 9.97 30.71	457 3.57 13.61 17.50	260 2.03 7.74 8.18	101 0.79 3.01 5.01	32 0.25 0.95 4.18	8 0.06 0.24 5.63	2 0.02 0.06 11.76	3359 26.25
1	607 4.74 19.95 22.20	98 0.77 3.22 40.16	469 3.67 15.42 42.99	916 7.16 30.11 35.08	716 5.60 23.54 22.54	214 1.67 7.03 10.63	22 0.17 0.72 2.88	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3042 23.77
2	89 0.70 2.49 3.26	20 0.16 0.56 8.20	253 1.98 7.09 23.19	948 7.41 26.55 36.31	1333 10.42 37.34 41.96	716 5.60 20.06 35.55	199 1.56 5.57 26.01	12 0.09 0.34 8.45	0 0.00 0.00 0.00	3570 27.90
3	0 0.00 0.00 0.00	0 0.00 0.00 0.00	34 0.27 1.54 3.12	290 2.27 13.11 11.11	764 5.97 34.54 24.05	750 5.86 33.91 37.24	313 2.45 14.15 40.92	57 0.45 2.58 40.14	4 0.03 0.18 23.53	2212 17.29
4	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	104 0.81 16.99 3.27	233 1.82 38.07 11.57	199 1.56 32.52 26.01	65 0.51 10.62 45.77	11 0.09 1.80 64.71	612 4.78
Total	2734 21.37	244 1.91	1091 8.53	2611 20.41	3177 24.83	2014 15.74	765 5.98	142 1.11	17 0.13	12795 100.00

Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Data Preparation

Data preparation efforts consisted of imputation to replace missing values with measures of central tendency and variable transformation for Poisson, Negative Binomial, and Logistic regression. No outliers were removed from the data to their effects upon the selected Logistic and Poisson regression model. Utilizing a combination of SAS procedures, we identified eight variables within the data that are missing records. Specifically, out of the 14 regressor variables used in this analysis, eight (ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, and STARS) were missing values and required imputation. Each variable that had missing values is designated in the data as IMP_XXXX and has a flag value F_XXXX that indicates which records were replaced.

A summary of the performed imputations are below:

- ResidualSugar: The regressor ResidualSugar had 616 missing records replaced with its mean value of 5.42. The mean value was used because it best represented the central tendency of the variable's distributions and was thought to have the highest probability of reflecting the missing value.
- Chlorides: The regressor Chlorides had 638 missing records replaced with its mean & mode value of 0.05. The mean value was used because it best represented the central tendency of the variable's distributions and was thought to have the highest probability of reflecting the missing value.
- FreeSulfurDioxide: The regressor FreeSulfurDioxide had 647 missing records replaced with its mode value of 30.85. FreeSulfurDioxide's mode value was used for imputation because it was thought have the highest probability of reflecting the missing value in the skewed distribution.
- TotalSulfurDioxide: The regressor TotalSulfurDioxide had 647 missing records replaced with its mode value of 125. TotalSulfurDioxide's mode value was used for imputation because it was thought have the highest probability of reflecting the missing value in the skewed distribution. Additionally, I wanted to follow the same process of imputation for TotalSulfurDioxide as FreeSulfurDioxide.
- pH: The regressor pH had 395 missing records replaced with its mean value of 3.21. The mean value was used because it best represented the central tendency of the variable's distributions and was thought to have the highest probability of reflecting the missing value.
- Sulphates: The regressor Sulphates had 1210 missing records replaced with its mean value of 0.53. The mean value was used because it best represented the central tendency of the variable's distributions and was thought to have the highest probability of reflecting the missing value.
- Alcohol: The regressor Alcohol has 653 missing records replaced with its mode value of 9.4. The regressor's mode value was used for imputation because it was thought have the highest probability of reflecting the missing value in the skewed distribution.

- STARS: The regressor STARS had 3359 missing records replaced with the value of 1. A one-star wine represents an unfavorable rating and was used as the imputed value of all non-rated wines (missing STAR values).

Data transformations were focused on altering variables so they would best conform to the normality assumptions normally used for OLS regression. We altered the variables VolatileAcidity, CitricAcid, ResidulSugar, Chlorides, FreeSulfurDioxide, and TotalSulfurDioxide through systematic trial and error until we had a distribution/frequency plot that minimized the AIC value for our Poisson & Negative Binomial models and maximized the area under the ROC curve for our Logistic models.

A summary of the performed transformations are below:

- FixedAcidity = sqrt(abs(FixedAcidity) + 1)
- VolatileAcidity = log(abs(VolatileAcidity))
- CitricAcid = sqrt(abs(CitricAcid));
- ResidualSugar = log(abs(ResidualSugar)+ 1)
- Chlorides = sqrt(abs(Chlorides))
- FreeSulfurDioxide = log(abs(IMP_FreeSulfurDioxide) + 1)
- TotalSulfurDioxide = log(abs(IMP_TotalSulfurDioxide) + 1)

Build Models

A process of trial & error was used to create the Poisson, Negative Binomial, and Logistic models that were developed for this analysis. We began by utilizing the PROC CORR procedure to see if the regressor's Pearson Correlation Coefficient values changed due to data imputations and transformations. We then progressed to using stepwise automated variable selection (AVS) in OLS regression to select our variables for Poisson, Negative Binomial, and eventually Logistic (for our champion hurdle model) regression. Stepwise AVS was used simply because of preference as all the of the variable selection procedures (stepwise, forward, backward) yielded the same adjusted R^2 value of 0.5984. Additionally, we utilized AVS to select the regressors for our derived TARGET values TARGET_FLAG and TARGET_AMT. As a reminder, the derived variable TARGET_FLAG represents the probability of any cases of sample wine being sold while TARGET_AMT represents the number of sample cases sold if a sale did take place i.e. if TARGET_FLAG = 1 then TARGET_AMT > 0.

The primary metrics used for model validation were the ROC curve, AIC values, and mean squared error where we attempted to maximize the area under the ROC curve and minimize the AIC value and mean square error of our models. In total, variations of five different models were evaluated for this analysis:

- GENMOD with Poisson Distribution
- GENMOD with Negative Binomial Distribution
- GENMOD with Zero Inflated Poisson (ZIP) Distribution

- GENMOD with Zero Inflated Negative Binomial Distribution
- Hurdle Logistic/Poisson Distribution

One item to note is that the mean and variance of the target variable's distribution were assessed to not be significantly different. As a result, our initial GENMOD Poisson distribution model and GENMOD Negative Binomial distribution model yielded the same results. Since both procedures were effectively doing the same thing, the decision was made to evaluate a different set of regressor's for the negative binomial model to open the aperture to include more models.

A summary of the evaluated models follows:

- GENMOD with Poisson Distribution

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13729.3886	1.0757
Scaled Deviance	13E3	13729.3886	1.0757
Pearson Chi-Square	13E3	11259.8092	0.8822
Scaled Pearson X2	13E3	11259.8092	0.8822
Log Likelihood		8732.8956	
Full Log Likelihood		-22809.7616	
AIC (smaller is better)		45647.5232	
AICC (smaller is better)		45647.5562	
BIC (smaller is better)		45751.8989	

- GENMOD with Negative Binomial Distribution

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13728.0462	1.0758
Scaled Deviance	13E3	13728.0462	1.0758
Pearson Chi-Square	13E3	11259.8328	0.8824
Scaled Pearson X2	13E3	11259.8328	0.8824
Log Likelihood		8733.5668	
Full Log Likelihood		-22809.0904	
AIC (smaller is better)		45652.1808	
AICC (smaller is better)		45652.2288	
BIC (smaller is better)		45778.9227	

- GENMOD with Zero Inflated Poisson (ZIP) Distribution

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		40694.7078	
Scaled Deviance		40694.7078	
Pearson Chi-Square	13E3	5759.7005	0.4517
Scaled Pearson X2	13E3	5759.7005	0.4517
Log Likelihood		11195.3034	
Full Log Likelihood		-20347.3539	
AIC (smaller is better)		40746.7078	
AICC (smaller is better)		40746.8179	
BIC (smaller is better)		40940.5482	

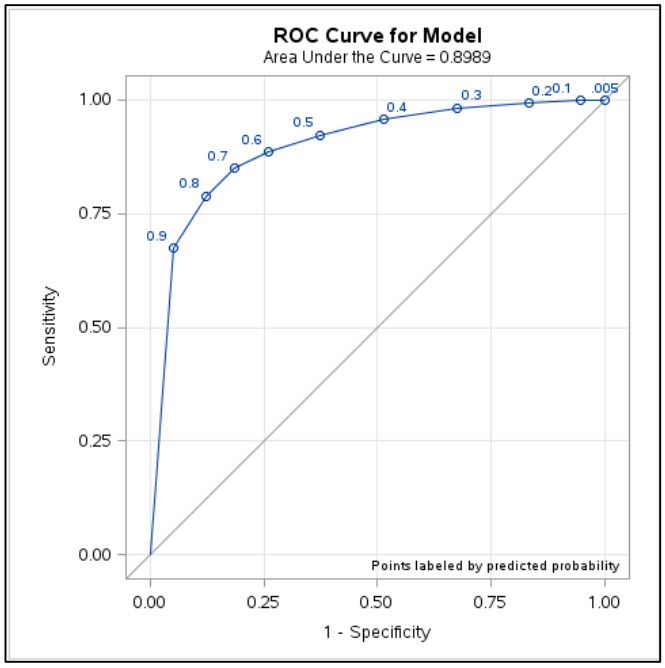
- GENMOD with Zero Inflated Negative Binomial Distribution

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		40859.8610	
Scaled Deviance		40859.8610	
Pearson Chi-Square	13E3	5557.7498	0.4359
Scaled Pearson X2	13E3	5557.7498	0.4359
Log Likelihood		-20429.9305	
Full Log Likelihood		-20429.9305	
AIC (smaller is better)		40913.8610	
AICC (smaller is better)		40913.9796	
BIC (smaller is better)		41115.1568	

Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

- Hurdle Logistic/Poisson Distribution

Logistic Model ROC Curve for TARGET_FLAG



Poisson Goodness of Fit criteria for TARGET_AMT

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	1E4	2941.7973	0.2932
Scaled Deviance	1E4	2941.7973	0.2932
Pearson Chi-Square	1E4	2612.1751	0.2604
Scaled Pearson X2	1E4	2612.1751	0.2604
Log Likelihood		2996.5858	
Full Log Likelihood		-15615.1042	
AIC (smaller is better)		31256.2084	
AICC (smaller is better)		31256.2447	
BIC (smaller is better)		31350.0012	

- Scoring Routine Output

Model Error Comparison

Obs	P	_FREQ_	ERROR_MEAN	ERROR_POI	ERROR_NB	ERROR_ZIP	ERROR_ZINB	ERROR_HURDLE_NB	ERROR_HURDLE_POI
1	1	12795	1.59558	1.03454	1.03468	1.83057	1.85167	1.34901	1.34901

Model Error Comparison

Obs	P	_FREQ_	ERROR_MEAN	ERROR_POI	ERROR_NB	ERROR_ZIP	ERROR_ZINB	ERROR_HURDLE_NB	ERROR_HURDLE_POI
1	1.5	12795	1.86745	1.17915	1.17919	1.99156	2.01250	1.48407	1.30108

Model Error Comparison

Obs	P	_FREQ_	ERROR_MEAN	ERROR_POI	ERROR_NB	ERROR_ZIP	ERROR_ZINB	ERROR_HURDLE_NB	ERROR_HURDLE_POI
1	2	12795	2.09479	1.31552	1.31547	2.12764	2.14843	1.61000	1.26886

Where P =1 assesses the average error of the models, P = 1.5 assess the exponent 1.5 error of the models, and P = 2 assess the mean squared error of the models.

Select Models

Our selected Logistic/Poisson Hurdle Regression model to predict the number of cases of sample wine sold to distributors.

Logistic Model for predicting derivative target variable TARGET_FLAG takes form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \varepsilon$$

With values:

In Model	In Data	Beta	Value
Y	Target_Flag	B0	**0.7802**
X1	AcidIndex	B1	**-0.3819**
X2	IMP_Alcohol	B2	**-0.0184**
X3	CitricAcid	B3	**0.2617**
X4	IMP_FreeSulfurDioxide	B4	**0.0783**
X5	LabelAppeal	B5	**-0.4649**
X6	IMP_STARS	B6	**2.5383**
X7	F_STARS	B7	**-1.8202**
X8	IMP_ResidualSugar	B8	**0.0547**
X9	IMP_Sulphates	B9	**-0.1056**
X10	IMP_TotalSulfurDioxide	B10	**0.2201**
X11	VolatileAcidity	B11	**-0.1497**
X12	IMP_pH	B12	**-0.1799**

Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

The chosen probability model has the following notes:

- LabelAppeal has a negative coefficient which runs contrary to my assumptions regarding the ability of well-designed wine labels to help wines sell.
- F_Stars, the flag for imputed STAR ratings, has a negative coefficient which matches my assumptions that wines with lower ratings < 3 don't sell as well.

Poisson Model for predicting derivative target variable TARGET_AMT takes form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \varepsilon$$

With Values:

In Model	In Data	Beta	Value
Y	Target_AMT	B0	0.7874
X1	AcidIndex	B1	-0.0205
X2	IMP_Alcohol	B2	0.009
X3	CitricAcid	B3	0.0083
X4	IMP_FreeSulfurDioxid	B4	0.0052
X5	LabelAppeal	B5	0.2952
X6	IMP_ResidualSugar	B6	-0.0021
X7	IMP_STARS	B7	0.1211
X8	F_STARS	B8	-0.0866
X9	IMP_Sulphates	B9	0.0003
X10	IMP_TotalSulfurDioxi	B10	-0.0049
X11	VolatileAcidity	B11	-0.0132
X12	IMP_pH	B12	0.0103

The chosen probability model has the following notes:

- F_Stars, the flag for imputed STAR ratings, has a negative coefficient which matches my assumptions that wines with lower ratings < 3 don't sell as well.

Assignment #3
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Conclusion

The purpose of this assignment was to use Poisson/Negative Binomial regression to develop a model that predicts the number of sample cases of wine purchased by distribution companies after tasting a wine. Over the course of our analysis we decided to break the predictor variable TARGET into the two derivative target variables TARGET_FLAG and TARGET_AMT. We then proceeded to model the probability of sample wine case sale using TARGET_FLAG and Logistic regression and used Poisson regression to model TARGET_AMT to predict the number of sample cases of wine sold, if a sale took place. The separate models where then combined into a Logistic/Poisson "hurdle" regression model that combined the objectives of the two separate models. We selected our Logistic/Poisson probability model based off model validation criteria such as the ROC curve and AIC. Our final analysis emulates the process performed by wine companies in determining the "popularity" of a wine as projection of potential sales.