# Predicting wine quality based on wine characteristics for wine distributors

Group 13: Joby George, Charles Klein, David Smith and David Zhang

Professor V
STAT 3200, Section 2
Due: October 5, 2017

**Pledge:**________________________________________

_____________________________________________

## Introduction: Predicting Wine Quality through Key Drivers

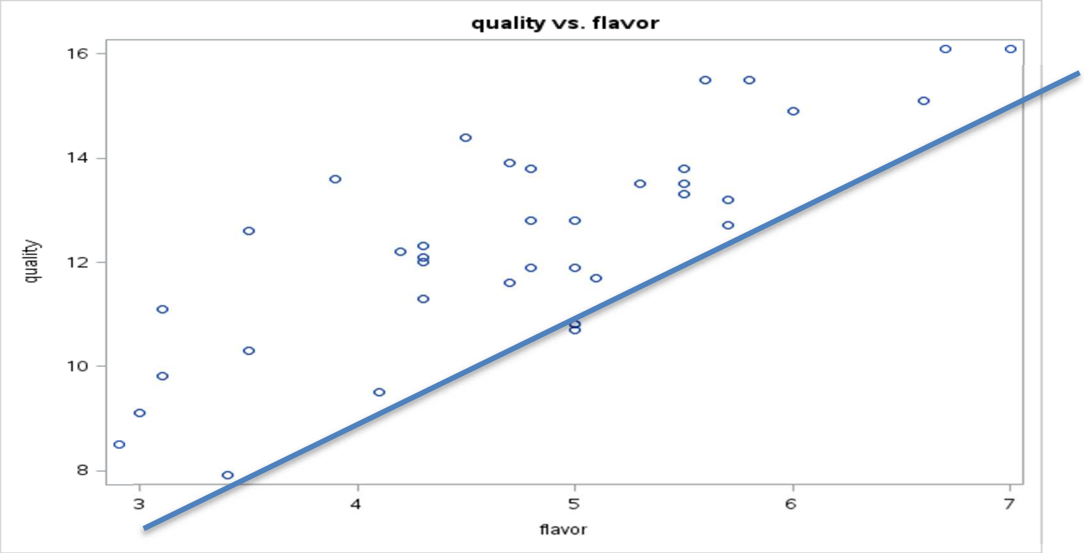
Pinot Noir quality is extremely important in determining its price. Even marginal improvements in its quality can drive large price premiums. To help wine distributors better understand wine quality given key drivers, we can use linear regression, a technique that looks at how variables interact and comes up with a linear prediction equation of wine quality. With this information, wine distributors will purchase better inventory and better satisfy their customers.

## Data Summary: Understanding our variables and the relationship between wine quality

There are six explanatory variables which affect overall quality that the wine distributor provided data on: clarity, aroma, body, flavor, oakiness, and the region where the wine was produced.

| Variable Name | Variable description |
|---|---|
| Aroma | Aroma refers to the smell of the wine, with higher scores implying stronger scents. |
| Body | Body refers to the weight of the wine, or how it feels within your mouth; higher scores imply a fuller body (or heavier wine). |
| Clarity | Clarity refers to the brightness and transparency of the wine, it was measured in a 0-1 scale, with 1 implying perfect clarity |
| Flavor | Flavor of the wine is either dry or sweet, a higher rating implies dryer wine and a lower rating implies sweeter wine. |
| Oakiness | Oak indicates the presence of oak during the fermentation process; higher ratings imply a greater presence of oak. |
| Region | Region refers to the region the wine was produced in, there are three regions in this dataset. |

To better understand the relationship between these variables and wine quality, we produced scatterplot, plotting each variable on the x axis, and quality on the y axis. Additionally, we created interaction and quadratic variables (creating a variable called flavor squared for example, by multiplying its value by itself) to test whether interaction variables better modeled quality, or the varaibles had a quadratic relationsihp with quality. Below is an example scatter plot that we found to show a relationship with quality (with additional scatterplots in **Appendix 1**).



Clarity did not appear to have any significant relationship to quality, so that scatter plot is not pictured, and we have not included that variable in our model. The plot of quality vs. aroma (**see Appendix 1, Figure 1)**, suggests a moderately strong positive relationship. The correlation coefficient is about 0.707, confirming this. The relationship could be linear, but it might also be quadratic, as the points appear to curve. We will evaluate both possibilities in our analysis.

The plot of quality vs. body (**see Appendix 1, Figure 2)** shows a moderately strong positive relationship, the correlation coefficient in this case being 0.549. Again, it appears to curve a bit downwards, so we will test for a quadratic relationship in our analysis.

The quality vs. flavor graph above shows a fairly strong positive relationship, with a correlation coefficeint equaling 0.79, also appearing possibly quadratic.

The plot of quality vs. oakiness did not show any relationship, so we have excluded it from our model.

The region vs quality plot above shows that region has an effect on the quality of the wine, wines coming from region 3 are more often higher quality, while wines coming from region 2 are likelier to be lower quality. Since region is a categorical variable, we needed to use two indicator variables to represent it in our analysis.

Lastly, we checked for interactions. It seems intuitive that aroma and flavor might interact, so we plotted quality vs. aroma grouping by flavor. The graph (**see Appendix 1, Figure 3)** shows of an interaction, with data points having the same flavor appearing to be grouped together. Thus, the variables that might be significant are aroma, body, and flavor, all of which might be quadratic based on the graphs; region, which is represented by two indicator variables; and an interaction term of aroma times flavor.

**Analysis: Determining which variables truly drive wine quality**

We have as our preliminary model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 (x_1)^2 + \beta_5 (x_2)^2 + \beta_6 (x_3)^2 + \beta_7 x_1 x_3 + \beta_8 D_3 + \beta_9 D_2 + \varepsilon$, where y is quality, $x_1$ is aroma, $x_2$ is body, $x_3$ is flavor, $D_3$ is an indicator variable which is 1 if the wine is from region 3 and 0 otherwise, and similarly for $D_2$ with region 2. We analyzed this model in SAS, and found the adjusted r squared value, which measures how much of the variance in quality is explained by the variance in the model, to be 0.7843, and the standard error, which measures the accuracy of our predictions, to be 0.95003. Looking at the overall F-statistic, we see from the p-value of less than 0.0001 that there is extremely strong evidence that at least one variable in our regression model is significant. Thus, we proceed to determine which variables in our model are significant. We ran the regression for our preliminary model, and removed the variable with the highest p-value, and repeated this process until every variable remaining was significant at an alpha level of 0.05. We removed the variables aroma, body, aroma squared, body squared, flavor squared, and the interaction variable aroma/flavor. Additionally, we can run a partial F-test to determine the significance of the region variable, represented by the two indicator variables $D_3$ and $D_2$. The partial F-statistic is calculated as F(partial) = [(SSE$_R$ - SSE$_F$)/(k - g)] / [SSE$_F$/(n - (k + 1))] = [(49.98597 - 25.2718)/2] / [25.2718/(38 - 10)] = 13.691. Comparing this to F$_{[0.01]}$ = 5.45, we see that F(partial) is greater, and can conclude that there is very strong evidence that at least two regions have different effects on mean quality.

We now remove the insignificant variables from the model, and reevaluated the model in SAS. Our new model is $Wine\ quality = 7.09431(flavor\ score) + 1.11555DD_3 - 1.53348D_2 + \varepsilon$ . It has an adjusted r squared value of 0.8087, which is higher than the adjusted r squared value of 0.7843 in the previous model. This is good, since it means more of the variance in quality is explained by the variance in the model. Furthermore, the standard error in the new model is 0.89464, which is smaller than the standard error of the previous model, which was 0.95003. This means that our current model makes more accurate predictions. Additionally, the current model has far fewer variables, and is therefore more parsimonious, which is always desirable. Looking at the overall F-statistic, we see that the p-value is less than 0.0001, confirming that there is extremely strong evidence that our model is significant. Looking at the individual variables, they are now all highly significant. Furthermore, running the partial F-test for the region variable again, we find that F(partial) = [(58.17344 - 27.21309)/2] / [27.21309/(38 - 4)] = 19.341. Since this is greater than F$_{[0.01]}$ ≈ 5.3, we conclude that there is very strong evidence that at least two regions have different effects on mean quality. Putting all of this together, we choose to use the $Wine\ quality = 7.09431(flavor\ score) + 1.11555DD_3 - 1.53348D_2 + \varepsilon$, and its associated prediction equation $\widehat{wine\ quality} = 7.09431(flavor\ score) + 1.11555DD_3 - 1.53348D_2$ to make predictions and answer questions about Pinot Noir wine quality. This can be divided into three equations:

**Equation 1: Wine from Region 1**

Wine Quality = 7.09431 + 1.11555*(flavor score)

**Equation 2: Wine from Region 2**

$$\text{Wine Quality} = 5.56083 + 1.11555*(\text{flavor score})$$

**Equation 3: Wine from Region 3**

$$\text{Wine Quality} = 8.31768 + 1.11555*(\text{flavor score})$$

The intercept is interpreted as such: for a wine from region 1 and a flavor rating of 0, the predicted quality of the wine is 7.09431; for a wine from region 2 and a flavor rating of 0, the predicted quality of the wine is 5.56083; for a wine from region 3 and a flavor rating of 0, the predicted quality of the wine is 8.31768. This interpretation makes little sense, as flavor ratings of 0 should yield a very low quality; the intercept is an inaccurate extrapolation from the data. The slope of the flavor score variable is interpreted as such: for every one increase in flavor score, the predicted quality increases by 1.11555. We reach this prediction equation under the key assumption that the error term is independently, identically, normally distributed with a mean of 0 and some constant standard deviation.

**Conclusion: Wine distributors should priortize wines from Region 3**

In conclusion, we used multiple regression methods to conclude a final wine quality prediction equation of predicted wine quality score = 7.09431 + 1.11555(flavor score) + 1.22337(If the wine is from region 3) - 1.53348(If the wine is from region 2). For an example wine from region 1 with a flavor score of 5, we would predict the wine's quality score to be 12.67206.

Our findings imply that this wine distributor should focus more resources on wines from region 3, as they tend to be of higher quality, and fewer resources on wines from region 2, which tend to be of lower quality. In the future for further analysis, this wine distributor could potentially utilize more data in their analysis. This data could perhaps pertain to the grapes the wines are made from, or additional regions where wines are made.

## Appendix 1: Scatterplots of the independent variables and quality

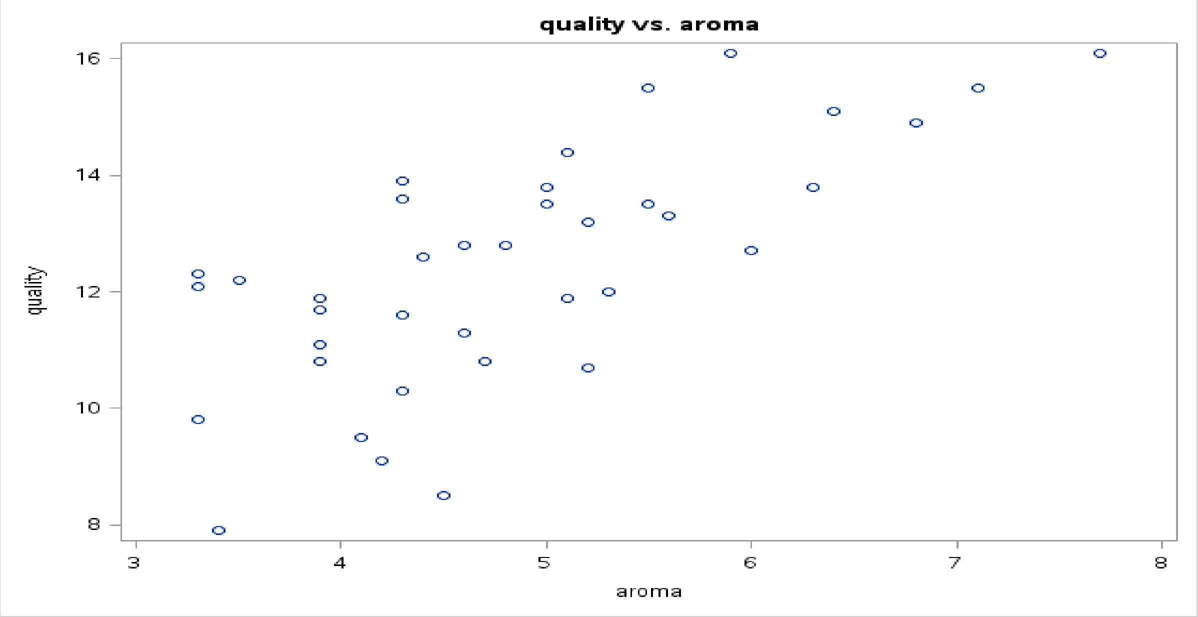


**Figure 1**: This scatterplot shows a relatively strong linear relationship between aroma and quality.

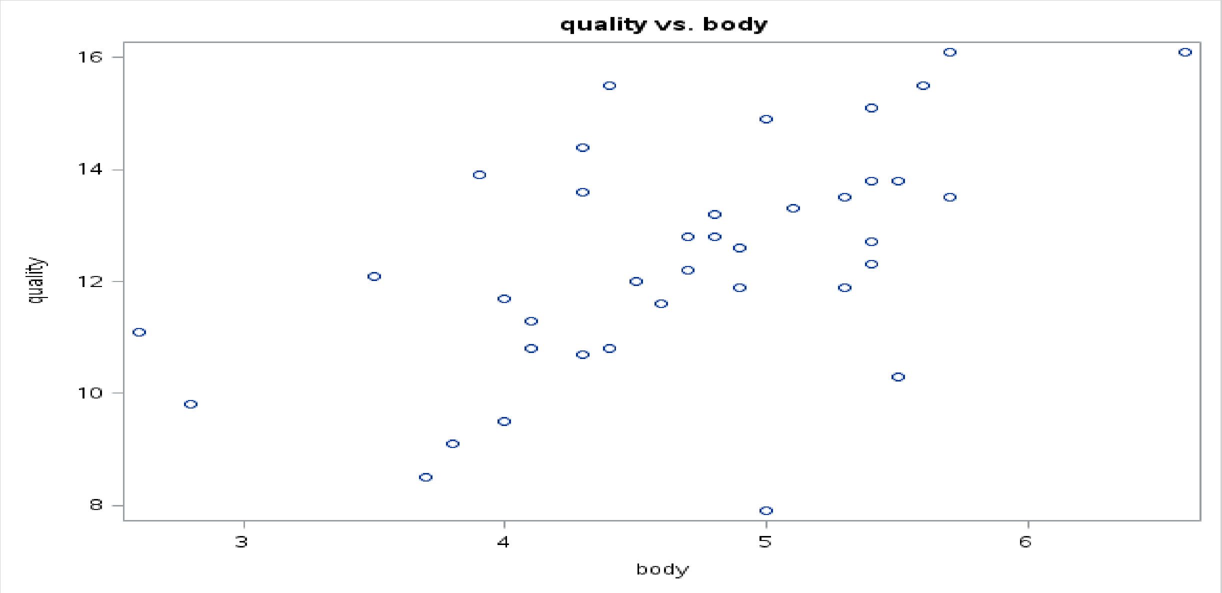


**Figure 2**: This scatterplot shows a slight linear relationship between body and quality

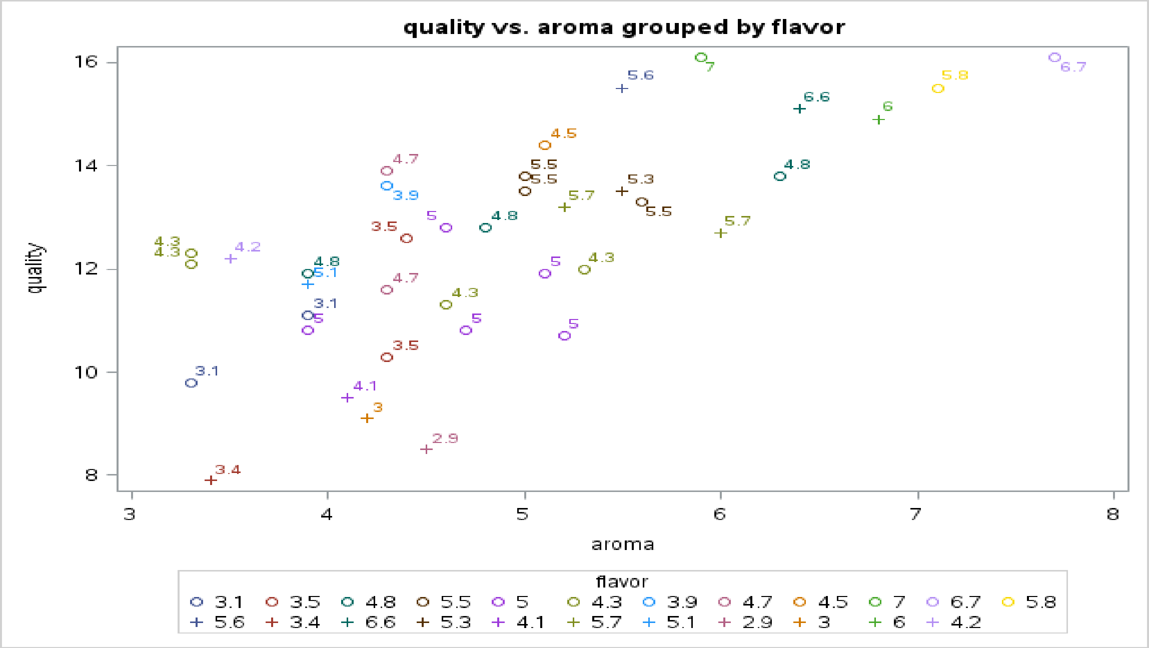


**Figure 3:** This graph shows how the interaction of aroma and flavor impact quality. The relationship is unclear if not relevant.

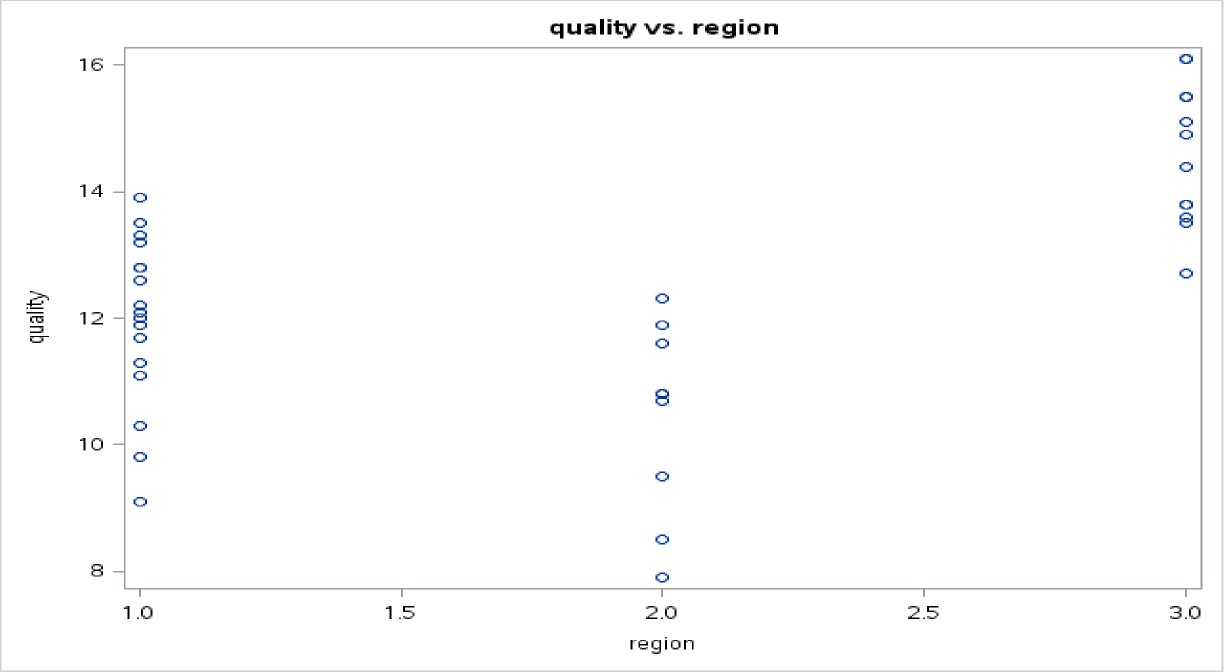


**Figure 4:**

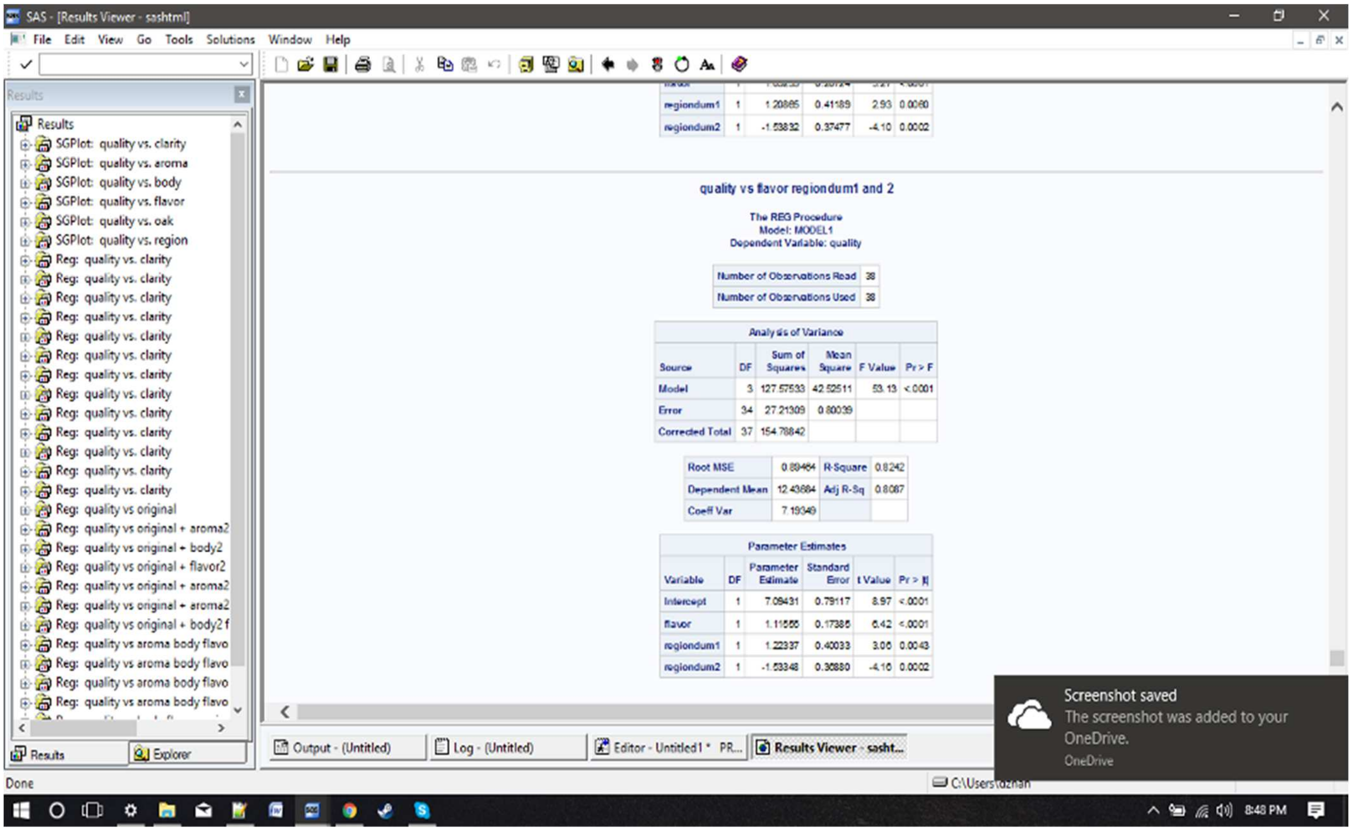
This graph shows that the qualitative variable region has a clear impact upon quality.

## Appendix 2: SAS Code

```
data wine;
input quality clarity aroma body flavor oak region;
cards;
9.8 1 3.3 2.8 3.1 4.1 1
12.6 1 4.4 4.9 3.5 3.9 1
11.9 1 3.9 5.3 4.8 4.7 1
11.1 1 3.9 2.6 3.1 3.6 1
13.3 1 5.6 5.1 5.5 5.1 1
12.8 1 4.6 4.7 5 4.1 1
12.8 1 4.8 4.8 4.8 3.3 1
12 1 5.3 4.5 4.3 5.2 1
13.6 1 4.3 4.3 3.9 2.9 3
13.9 1 4.3 3.9 4.7 3.9 1
14.4 1 5.1 4.3 4.5 3.6 3
12.3 0.5 3.3 5.4 4.3 3.6 2
16.1 0.8 5.9 5.7 7 4.1 3
16.1 0.7 7.7 6.6 6.7 3.7 3
15.5 1 7.1 4.4 5.8 4.1 3
15.5 0.9 5.5 5.6 5.6 4.4 3
13.8 1 6.3 5.4 4.8 4.6 3
13.8 1 5 5.5 5.5 4.1 3
11.3 1 4.6 4.1 4.3 3.1 1
7.9 0.9 3.4 5 3.4 3.4 2
15.1 0.9 6.4 5.4 6.6 4.8 3
13.5 1 5.5 5.3 5.3 3.8 3
10.8 0.7 4.7 4.1 5 3.7 2
9.5 0.7 4.1 4 4.1 4 2
12.7 1 6 5.4 5.7 4.7 3
11.6 1 4.3 4.6 4.7 4.9 2
11.7 1 3.9 4 5.1 5.1 1
11.9 1 5.1 4.9 5 5.1 2
10.8 1 3.9 4.4 5 4.4 2
8.5 1 4.5 3.7 2.9 3.9 2
10.7 1 5.2 4.3 5 6 2
9.1 0.8 4.2 3.8 3 4.7 1
12.1 1 3.3 3.5 4.3 4.5 1
14.9 1 6.8 5 6 5.2 3
13.5 0.8 5 5.7 5.5 4.8 1
```

```
12.2 0.8 3.5 4.7 4.2 3.3 1
10.3 0.8 4.3 5.5 3.5 5.8 1
13.2 0.8 5.2 4.8 5.7 3.5 1
;
run;

proc sgplot data=wine;
scatter y=quality x=clarity;
title "quality vs. clarity";
run;

proc sgplot data=wine;
scatter y=quality x=aroma;
title "quality vs. aroma";
run;

proc reg data=wine plots=none;
model quality = aroma;
run;

proc sgplot data=wine;
scatter y=quality x=body;
title "quality vs. body";
run;

proc reg data=wine plots=none;
model quality = body;
run;

proc sgplot data=wine;
scatter y=quality x=flavor;
title "quality vs. flavor";
run;

proc reg data=wine plots=none;
model quality = flavor;
run;

proc sgplot data=wine;
scatter y=quality x=oak;
```

```
title "quality vs. oak";
run;
proc sgplot data=wine;
scatter y=quality x=region;
title "quality vs. region";
run;

data wine1;
set wine;
regiondum1 = 0;
regiondum2 = 0;
if region = '3' then regiondum1 = 1;
if region = '2' then regiondum2 = 1;
aflavor = aroma*flavor;
aroma2 = aroma**2;
body2 = body**2;
flavor2 = flavor**2;
run;

proc reg data=wine1 plots=none;
model quality = aroma body flavor aroma2 body2 flavor2 regiondum1 regiondum2 aflavor;
title 'quality vs aroma body flavor aroma2 body2 flavor2 regiondum1 regiondum2 aflavor';
run;

proc reg data=wine1 plots=none;
model quality = aroma body flavor aroma2 body2 flavor2 aflavor;
title "Partial F-test for Region";
run;

proc reg data=wine1 plots=none;
model quality = flavor regiondum1 regiondum2;
title "quality vs. flavor regiondum1 regiondum2";
run;

proc reg data=wine1 plots=none;
model quality = flavor;
title "Partial F-test for Region";
run;
```


SAS - [Results Viewer - sashtml]
File Edit View Go Tools Solutions Window Help
Results
SGPlot: quality vs. clarity
SGPlot: quality vs. aroma
SGPlot: quality vs. body
SGPlot: quality vs. flavor
SGPlot: quality vs. oak
SGPlot: quality vs. region
Reg: quality vs. clarity
Reg: quality vs original
Reg: quality vs original + aroma2
Reg: quality vs original + body2
Reg: quality vs original + flavor2
Reg: quality vs original + body2 f
Reg: quality vs aroma body flavo
Results
Explorer
Output - (Untitled)
Log - (Untitled)
Editor - Untitled1 * PR...
Results Viewer - sasht...
Done
C:\Users\dznan
8:48 PM

| | | | | | |
|---|---|---|---|---|---|
| regiondum1 | 1 | 1.20865 | 0.41189 | 2.93 | 0.0060 |
| regiondum2 | 1 | -1.53832 | 0.37477 | -4.10 | 0.0002 |

quality vs flavor regiondum1 and 2

The REG Procedure
Model: MODEL1
Dependent Variable: quality

| | |
|---|---|
| Number of Observations Read | 38 |
| Number of Observations Used | 38 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 127.57533 | 42.52511 | 53.13 | <.0001 |
| Error | 34 | 27.21309 | 0.80039 | | |
| Corrected Total | 37 | 154.78842 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.89464 | R-Square | 0.8242 |
| Dependent Mean | 12.43684 | Adj R-Sq | 0.8087 |
| Coeff Var | 7.19349 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 7.09431 | 0.79117 | 8.97 | <.0001 |
| flavor | 1 | 1.11566 | 0.17386 | 6.42 | <.0001 |
| regiondum1 | 1 | 1.22337 | 0.40033 | 3.06 | 0.0043 |
| regiondum2 | 1 | -1.53348 | 0.36880 | -4.16 | 0.0002 |

Screenshot saved
The screenshot was added to your OneDrive.
OneDrive

**Figure 5:** Screenshot of SAS output for the finalized regression model