MWSUG 2019 - Paper PO-097

Exploring Wine Reviews: How Language and Word Use Varies in Wine Reviews

Abigail Zysk, Grand Valley State University, Allendale, MI
Kylie Springer, Grand Valley State University, Allendale, MI

ABSTRACT

With thousands of varieties of wine, wine descriptions are diverse and unique to the individual describing the wine. A dataset including the wine variety, reviewer, and wine descriptor/flavor words was used to explore the frequency of each word used within certain varieties of wine, for individual reviewers, and for the combination of variety and reviewer. By examining the word usage for different varieties of wine for the top reviewers, we saw that the most used wine descriptors were not exclusive to varieties of wine, but dependent on the wine reviewer. Roger Voss, who had the largest amount of reviews, used the word 'rich' when using a wine descriptor 17.79% of the time. This was reflected across different varieties of wines. When reviewing Bordeaux-Style Red Blends he used the word 'rich' 21.71% of the time when using any wine descriptor, 17.39% for Chardonnay, and 18.14% for Pinot Noir. When looking at word usage for a single variety of wine, we concluded that the words favored by reviewers could influence the results. When exploring the most used wine descriptors for Chardonnay and Pinot Noir wines, the word 'rich' was one of the top words for both wines, which could be due to Voss frequently using the word. The results from this study suggest it would be helpful to look at reviews from a multitude of different wine reviewers to get an accurate description of the wine.

INTRODUCTION

Wine is a popular beverage that is enjoyed and produced in nearly every country. Wine has existed and flourished throughout history, and the concept of toasting was originally established using wine in Ancient Rome. Today, the top wine-producing countries are Italy, China, France, Spain, USA, and Australia. Around 267 million hectoliters of wine was produced in the world in 2016 alone (OIV, 2016). Within different countries, the production of wine is often specialized to certain varieties due to different wine varieties being able to flourish under certain climates. The most produced wine in China is called Kyoho, which accounts for 44% of China's total vineyard area (OIV, 2017). With thousands of different varieties of wine grapes, there is a large amount of diversity in the varieties of wine. Due to the diversity of wines, they are often easiest to classify by the way they taste. For example, a Moscato is usually sweet and slightly fizzy. Wine reviews are a description of how a wine tastes, provided by a wine expert. An individual selecting a bottle of wine may reference a wine review to determine if the bottle of wine matches their own taste preferences.

While performing the analysis, learning objectives that focused around SAS were utilizing PROC SQL, features of the mosaic plot within PROC FREQ, and searching character strings. We wished to utilize PROC SQL as a learning experience because it is a useful tool for data cleaning and manipulation. This is an introductory paper, but some PROC SQL will be utilized and illustrated. While exploring the mosaic plot we specifically wanted to learn more about changing colors and adding labels and titles. To maximize the consistency of our analysis we wanted to control for punctuation marks at the end of the word. To do this we explored different ways to search and alter strings.

DATA

We obtained the wine review dataset from Kaggle Inc, a website that allows users to explore and publish data sets. The dataset contained 130,000 observations, which each observation representing one wine review. The variables included in the dataset were the country the wine came from, description of the wine by the taster, the vineyard within the winery where the grapes which made the wine are from (designation), points, price, province, region, taster's name, taster's twitter handle, the title of the wine review, the wine variety, and the winery the wine came from. In our analysis, we used the description of the wine, taster's name, and the wine variety. In the dataset there were 707 different varieties of wine, reviewed by 19 wine tasters. Nothing alarming came from initial exploration of the data. We created the

variable wine word (*word*), by constructing a list of wine descriptors that we obtained from the California Wine Club. In the list, we also included fruit flavor words, and wine descriptors that contained punctuation marks at the end of the word to ensure the consistency of our results. For the purpose of this study the variables used were wine variety (*variety*), wine reviewer(*taster*), wine description (*description)*, and a list of words pertaining to wine (*word*).

DATA CLEANING AND VALIDATION

We performed all data cleaning and validation in SAS 9.4 (SAS Institute, Cary NC). The first step in our data cleaning process was delimiting the wine reviews by spaces. The delimited dataset was transposed using an array so there would be one word per observation, wine variety, and wine reviewer. Next, we merged in the list of wine words from the California wine club that contained a column of ones, with the variable name *num*. Lastly, all non-wine descriptors were deleted from the dataset. This was done by deleting all words where *num* was not equal to "1". The final dataset contained the variables representing variety, wine reviewer, wine descriptors, and a count of "1" for each word. The count of "1" was the result of the variable *num*. For analysis, the count of each word was summed across different levels using PROC SQL. An example of the PROC SQL code where word count was at the wine variety-taster level can be seen here:

```
proc sql;
create table word_counts_variety_taster;
select variety, taster_name, word, count(word) as word_count
    from wine
    where taster_name ne ""
    group by 1,2,3;
quit;
```

We discovered multiple ways to control for punctuation marks at the end of the word. One of these ways was by using the INDEX function. The INDEX function can be used to identify words with a punctuation mark at the end, such as ",", ".", "!", etc. Then the variable "word" was trimmed by 1 character at the end of the word to remove the ending punctuation, which can be seen here:

```
data data2;
set data1;
  If index(word,'.') then substr(word, length(a.word)-1);
  else if index(word,',') then substr(word, length(a.word)-1);
Run;
```

Another way to control for punctuation marks at the end of a word is to use PROC SQL and the 'like %_%' operation, which can be seen here:

```
proc sql;
create table data2 as method init();
select a.*, (case
             when a.word like '%.%' then substr(a.word,1,length(a.word)-1)
             when a.word like '%,%' then substr(a.word,1,length(a.word)-1)
             else word end) as word1
    from data1 as a;
quit;
```

Both techniques provided the same results. We thought using the INDEX function was better because it used more concise and efficient code, opposed to having to create a new PROC. Additionally, the complexity involved with 'case when' in PROC SQL was reduced.

ANALYSIS

DESCRIPTIVE STATISTICS

For certain varieties of wine, we explored the distribution of wine reviews by variety, wine reviews by reviewer, and the frequency of wine reviews by an individual reviewer for certain varieties of wine using

PROC FREQ. Out of the 707 varieties of wine, the top 3 varieties of wine reviewed the most were Pinot Noir with 13,272 reviews, Chardonnay with 11,753 reviews and Cabernet Sauvignon with 9,472 reviews. There were 140 varieties of wine that were reviewed only once. Out of the 19 reviewers, the top 3 wine reviewers were Roger Voss with 25,514 reviews, Michael Schachner with 15,134 reviews, and Kerin O'Keefe with 10,776 reviews. Of the 25,514 reviews that Roger Voss made, the top 3 wines he reviewed were Bordeaux-style Red Blend which he reviewed 4,710 times, Chardonnay (2,786 times), and Portuguese Red (2,462 times). Of Michael Schachner's 15,134 wine reviews, Malbec, Red Blend, and Tempranillo were among his top 3 wines reviewed with Schachner reviewing Malbec 1,652 times, Red Blend 1,496 times, and Tempranillo 1,439 times. Kerin O'Keefe's top 3 wines reviewed were Red Blend, Nebbiolo, and Sangiovese with O'Keefe rating Red Blend 2,507 times, Nebbiolo 1,930 times, and Sangiovese 1,584 times. The bottom 3 reviewers who reviewed more than 1 variety of wine were Carrie Dykes with 139 reviews, Fiona Adams with 27 reviews, and Christina Pickard with 6 reviews. There were many reviewers who reviewed only 1 variety of wine, so for the consistency of our analysis we focused on the top 3 individuals who had the most wine reviews in the dataset.

PROCEDURE APPLICATIONS

When first examining this dataset, we wanted to look at popular words for each variety-taster combination. To determine word usage across different wine reviewers and wine varieties, we summed the word counts for each wine type and wine taster creating the variable *word_count*. Next, a new dataset was created using the output statement in PROC FREQ, which obtained the frequency of each word used for each variety-taster combination. To display the data, we created a PROC PRINT from a restricted data set only containing the top 3 wine reviewers to make the data manageable to view. The results can be seen below in Table 1.

Wine Type	Name of Wine Reviewer and Taster	Word	Word Count	Percentage of Word Use
Bordeaux-style Red Blend	Roger Voss	fruity	521	13.86
Bordeaux-style Red Blend	Roger Voss	rich	816	21.71
Bordeaux-style Red Blend	Roger Voss	structure	562	14.95
Cabernet Sauvignon	Michael Schachner	green	169	12.37
Chardonnay	Michael Schachner	nose	84	13.50
Chardonnay	Roger Voss	rich	482	17.39
Pinot Noir	Roger Voss	rich	334	18.14
Pinot Noir	Roger Voss	soft	236	12.82
Pinot Noir	Roger Voss	structure	320	17.38
Red Blend	Kerin O'Keefe	fresh	262	14.59
Red Blend	Kerin O'Keefe	nose	256	14.25
Red Blend	Roger Voss	rich	56	18.92
Riesling	Roger Voss	perfumed	93	18.13
Riesling	Roger Voss	rich	64	12.48
Sauvignon Blanc	Michael Schachner	green	405	42.01
Sauvignon Blanc	Roger Voss	green	164	14.40
Sauvignon Blanc	Roger Voss	rich	176	15.45

Table 1. Most Used Words for Different Wine Variety and Taster Combinations

Looking at Table 1, on the previous page, we can see many interesting attributes about the word usage based on variety and taster. First, Roger Voss used the word 'rich' for every variety of wine on the table. Looking at the percentages we can see that on average, he used the word 'rich' when using a wine descriptor about 13% of the time for all varieties shown. Similarly, Michael Schachner used the word 'green' quite often when using wine descriptors. Michael used the word 'Finish' when using a wine descriptor pertaining to the variety Sauvignon Blanc around 21.09% of the time.

Since Table 1 gives the impression that wine reviewers use the same word for describing any variety of wine, our next question was "what words the wine reviewers used the most regardless of variety". To answer this question, we took the original cleaned dataset and summed a word that was used for each taster using PROC SQL creating the variable *word_count*. Then we created a mosaic plot from a limited dataset using PROC FREQ of wine descriptors by taster, weighted by word count.

Two features that were utilized to produce our mosaic plot were specifying "(colorstat=stdres)" within the tables statement of PROC FREQ and specifying "sge=on" within the "ods listing" statement. Using "(colorstat=stdres)" colored the tiles based on the standardized residuals of the corresponding table cells. So, in terms of our mosaic plot, tiles that are a darker red were used more than colors that are lighter terms of red. A complication that we ran into while creating our mosaic plot was changing the title, using the "title" statement did not work. So, specifying "sge=on" within the "ods listing" statement creates an editable SGE file. The SGE file can be edited by selecting the graph file in the explorer window, where a new window appears that allows you to select and edit certain features of the graph. The following code was used to create the mosaic plot:

```
ods listing sge=on;
proc freq data=data2;
  tables word*taster_name/plots=mosaic (colorstat=stdres);
  weight word_count;
  label taster_name="Taster Name"
        word="Word";
run;
```

For the mosaic plot, a restricted data set was used to make the figure readable and easy to understand. The mosaic plot can be seen on the following page in Figure 1.

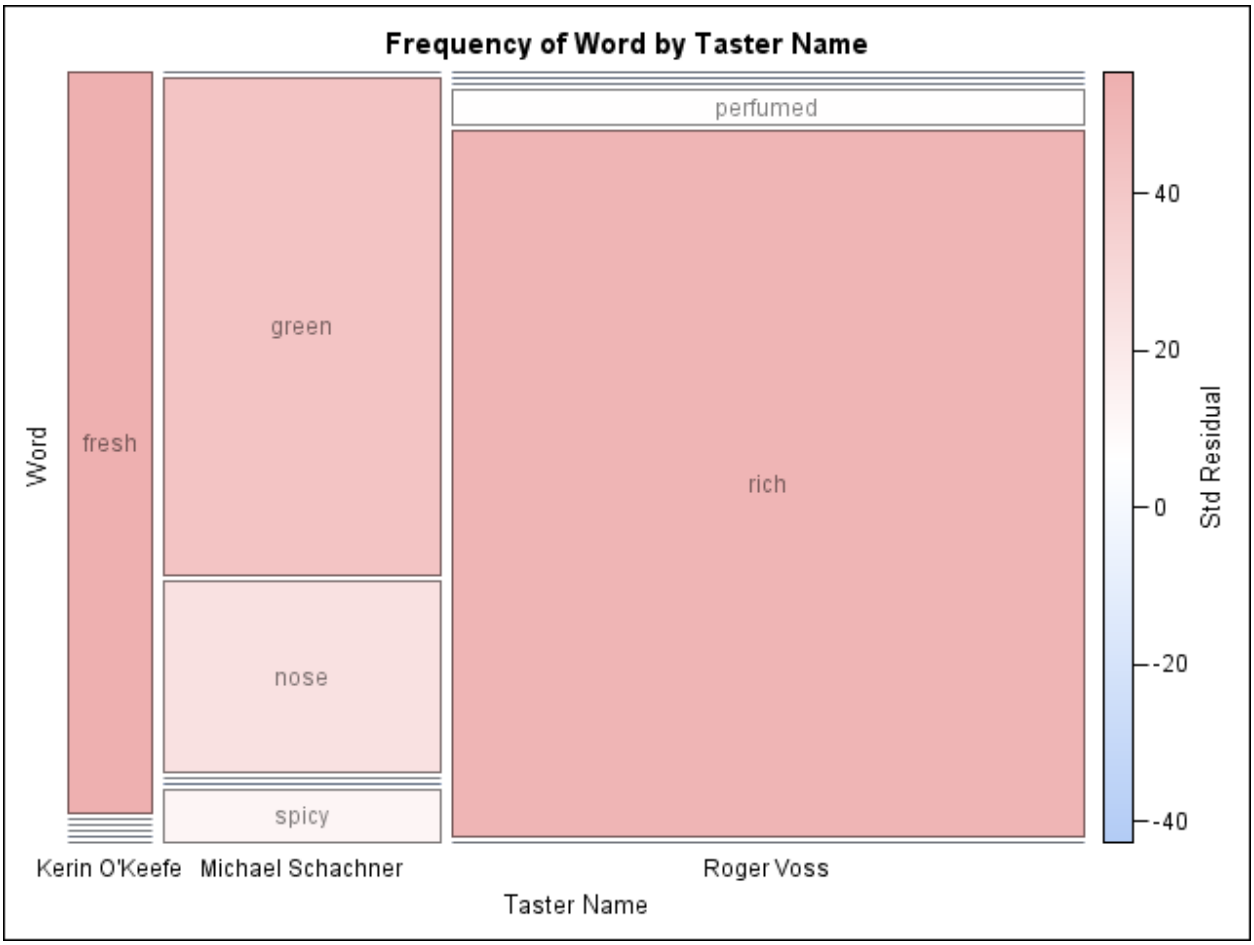

Figure 1. Frequency of Word Usage by Taster

The mosaic plot above in Figure 1 helps give us an idea of each individual word usage. Looking at the plot we can see that for the top 3 wine reviewers, there was not a large diversity in their word usage. For Roger, we can see that the word 'rich' dominates the other wine descriptors that he uses. Similarly, the wine descriptor 'fresh' overruled all other wine descriptors for the reviewer Kerin O'Keefe. Michael Schachner had a larger diversity in wine descriptors that he used most often. The three wine descriptors that Schachner used most often were 'green', 'nose', and 'spicy'. Even though Schachner did have a wider diversity of words he most often used, the word 'green' was still used more often than the words 'nose' and 'spicy'. The frequency in word usage being shown in the mosaic plot for each reviewer is consistent with what we saw in Table 1. In Table 1, we saw that Roger used the word 'rich' quite often for every variety of wine shown in the table, which is consistent with the results in the Mosaic plot.

One issue that could exist with examining word usage this way, is that the words a wine reviewer uses could be dependent on the wine variety. So, if a wine taster reviews a variety of wine more often than others, words pertaining to that variety would appear more often. Table 1 shows that across multiple varieties of wine Voss employed the word 'rich' and while exploring descriptive statistics we saw that there was a good amount of diversity in the varieties of wine reviewed by each taster in the Kaggle dataset.

After seeing how certain wine descriptors dominated the word usage for individual reviewers, we were curious to see if a reviewer's most used words were present in a varieties most popular wine descriptors, while only looking at variety. To further examine this thought we created a new dataset from the "out=" statement within PROC FREQ, to obtain the percentages of the times a word was used per variety. Next, we created two horizontal bar graphs to look at the most used wine descriptors for each variety. We chose the wine varieties Chardonnay and Pinot Noir for the horizontal bar graphs because the number of times these varieties were reviewed by each of the top wine reviewers was large and balanced between

the three. Also, included in these bar graphs are fruit flavor words. Fruit flavor words are included in these visualizations, but excluded from Table 1 and Figure 1, because fruit flavors are heavily dependent on the variety of wine, and the variable wine taster was in Table 1 and Figure 1, whereas Figure 2 and 3 are solely based on variety. Below, we can see the most used words pertaining to the wine varieties Chardonnay (Figure 2) and Pinot Noir (Figure 3).

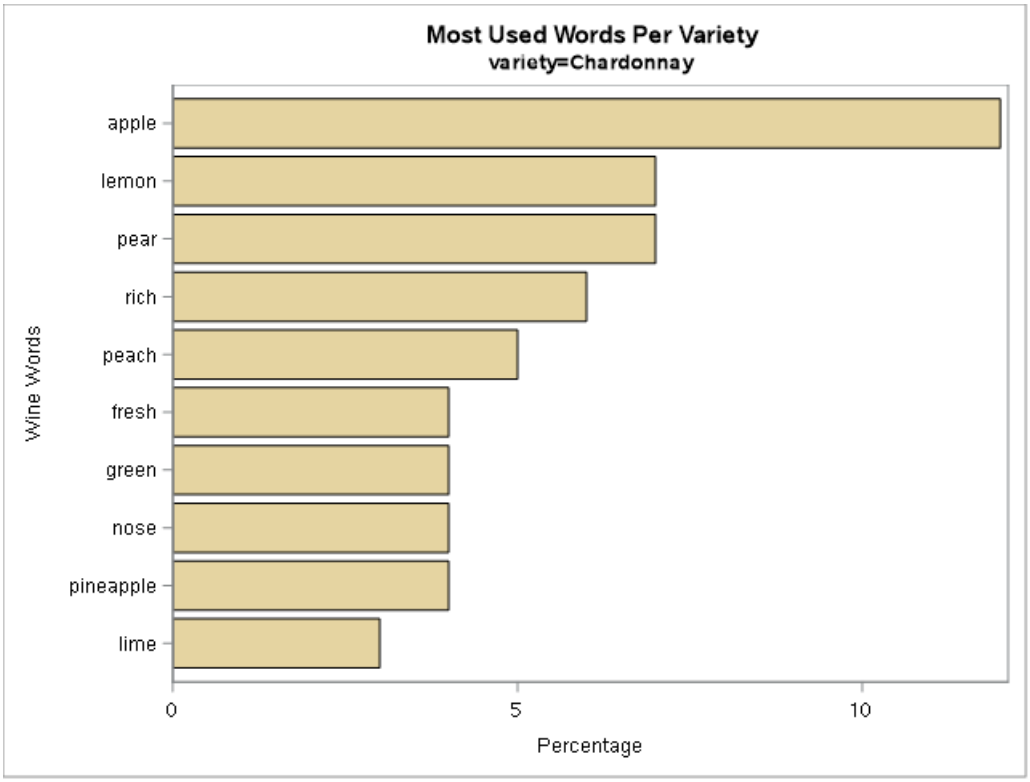

Figure 2. Most Used Words for the Wine Variety Chardonnay

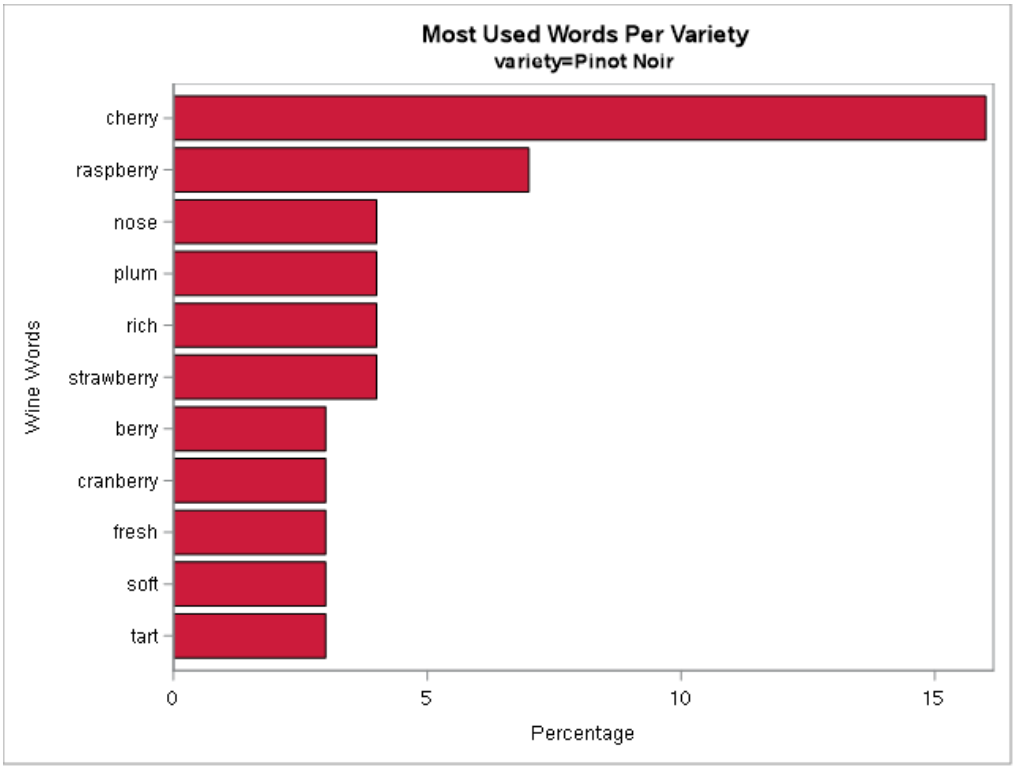

Figure 3. Most Used Words for the Wine Variety Pinot Noir

When examining Figure 2 and Figure 3 we can see that the wine words fit the ideal description of the wines. When looking at the bar graph for chardonnay the flavors of apple, pear, lemon, etc. are what one would expect from a white wine. Similarly, for Pinot Noir flavors of cherry, raspberry, plums, etc. accurately represent the variety. But while looking at the descriptor words we can see that rich, fresh, and green are the most popular descriptor word for both Chardonnay and Pinot Noir. Previously, we discussed how Voss's most used wine descriptor was the word 'rich', for Schachner the words were 'nose' and 'green', and for O'Keefe the most used wine descriptor was 'fresh'. Describing these wines as such could easily be based on the wine reviewer favoring these words and not the wine itself. If wine reviewers do not distinguish their wine description by the wine type, a wine review can be similar to an ingredient list.

SUGGESTIONS

There are many possibilities of exploration within wine reviews. With this Kaggle dataset, it would be interesting to try and predict the wine type based on wine review. Other prediction models would be interesting as well. For example, predicting the wine score based on wine descriptors and wine variety. On the topic of word usage within wine reviews, further analysis would require a dataset where the frequency of wine reviews per variety and taster are more consistent in number. This would allow us to see if the wine descriptors which a taster uses really affect the description of a wine variety.

One possible limitation of our analysis is that some of the wine descriptions were translated from another language and contained special characters. Words with special characters would have been excluded from our analysis due to the vast number of words in the wine descriptions, and there was no way to control for all possible special characters. Another limitation is that the list of wine words from the California Wine Club may be incomplete. Multiple lists of wine words from various data sources could be compiled and compared to ensure that the list of wine words used is complete.

CONCLUSION

In our exploration of the dataset involving wine reviews, we utilized various SAS techniques such as PROC SQL and mosaic plot options in PROC FREQ to examine various aspects of the dataset. One new

thing that we learned was the INDEX function. Prior to this project, we were unaware that this function existed. The INDEX function is a very useful tool that can be utilized frequently for analyzing or identifying information within strings.

Although our dataset had a few imperfections, many interesting insights came to light about word usage among varieties of wine and wine reviewers. We saw that different wine reviewers favor certain wine descriptors more than others. For example, we saw that across many different varieties of wine, Roger Voss's most frequently used wine descriptor was often 'rich'. Furthermore, when examining Voss's most frequently used wine descriptors, regardless of variety, 'rich' was still the most used word. When looking at the most popular wine descriptors for individual varieties of wine, the word 'rich' appeared across many different varieties as one of the most popular words. This shows how Voss's word usage affects the description of a wine variety. Based on these facts, there is evidence to believe that wine reviews could be influenced on the wine reviewers preferred vocabulary instead of the wine itself. So, when picking out a bottle of wine, an individual should reference multiple reviews to make an informed decision on the best wine for themselves.

REFERENCES

Distribution of the world's grapevine varieties(Rep.). (n.d.). Retrieved http://www.oiv.int/public/medias/5888/en-distribution-of-the-worlds-grapevine-varieties.pdf

OIV Statistical Report on World Vitiviniculture(Rep.). (n.d.). Retrieved http://www.oiv.int/public/medias/5479/oiv-en-bilan-2017.pdf

ACKNOWLEDGMENTS

The authors would like to thank Dr. Robert Downer for his guidance and expertise, Dr. Laura Kapitula for sharing her passion and knowledge of SAS, and the GVSU Department of Statistics for guiding us in our statistical careers.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Abbie Zysk
Grand Valley State University
zyska@mail.gvsu.edu

Kylie Springer
Grand Valley State University
springek@mail.gvsu.edu