## Vienna 2019 Abstract Submission

**Title**
Mid-Atlantic Wine Industry Target Market Analysis Using Decision Trees

**I want to submit an abstract for:**
Conference Presentation

**Corresponding Author**
Chen Yue

**E-Mail**
jane.chenyue@rutgers.edu

**Affiliation**
Rutgers University

**Co-Author/s**

| Name | E-Mail | Affiliation |
|---|---|---|
| Ramu Govindasamy | govind@sebs.rutgers.edu | Rutgers University |
| Kathy Kelly | kmk17@psu.edu | Pennsylvania State University |

**Keywords**
Survey Research, Decision Tree Models, Market Segmentation, Machine Learning, Data Mining

**Research Question**
Who is the target market of Mid-Atlantic wines? What characteristics do they have? How do they think of Mid-Atlantic wines so far?

**Methods**
An Internet Survey (n=977) was used as the data source. The decision tree model is used to identify target markets and evaluate the most important factors that segment the market.

**Results**
The target market of wines that are produced in each state of the Mid-Atlantic region is identified. The most important feature that identifies the market is compared between the states.

**Abstract**
Introduction
Recently, as the development of computers and machine learning algorithms advance, it has become possible to apply this technology to data collected from wine consumers to learn about their attitudes and behaviors. Though readily available, only a few studies have applied algorithms in this manner to discover wine consumer insights. For this reason, it is necessary to start studying their application.
In this study, we specifically focus on the Mid-Atlantic wine production region (New York, New Jersey, and Pennsylvania), as there have been few wine-consumer studies that have highlighted any potential unique characteristics specific to wine drinkers in the region. Although Kelley et al. (2015) segmented the Mid-Atlantic market based on the frequency (super core, core, and marginal) with which they consumed the beverage (the summation of both domestic and imported wine), little is known specifically about these consumers consumption of

local wines. Considering its market share and reputation, the Mid-Atlantic wine industry needs to identify its uniqueness and differentiate itself from the general wine market to stay competitive. Because the Mid-Atlantic Region does not have very high market share, only targeting the frequent wine drinkers can be misleading. Thus, this study fills the gap with analysis performed based solely on wines produced in each of the three Mid-Atlantic states. Using such an approach, we were able to identify points of differentiation for wines from each of the three states and the respective target markets.

## Methodology
### Source of Data
A consumer survey approach was taken. A 15-minute Internet survey was administered on 22-24 October 2014 to 977 of Survey Sampling International (Shelton, CT) panelists residing in New Jersey, New York, and Pennsylvania. Panelists were screened for not being a member of the wine industry, being at least 21 years old, and having purchased and drunk wine at least once within the previous year.

### Decision Tree Models
In addition to descriptive statistics, we employed a decision tree approach to characterize the target market for each of the three states. Decision trees use learning algorithms that define the splits at each internal node. The tree starts from a root node that then splits into two or more children nodes (subgroups) that visually looks like an "upside down" or "sideways" tree. Each child node, if a further separation occurs, becomes a parent node. If no further separation occurs, the child node becomes a leaf node with no possibility of splitting any further. Except for the root node and the leaf nodes, each node will have exactly one parent and two or more children.

The decision tree model is specialized in handling categorical data. Compared to a logistic regression model, a decision tree offers many advantages. One advantage that implementation is fast and variable selection is automatic during model fitting (Deshpande, 2011). Additionally, missing data are well handled and normalization or scaling is not required prior to fitting the tree. The model is also insensitive to outliers as the splits are made based on the proportion of samples instead of on the absolute values (Deshpande, 2011). When working with survey data, the tree can also eliminate the need to create excessive of excessive dummy variables for multinomial data as each category are assigned to the subgroup that optimizes the model. Moreover, the model does not assume linear relationships between the variables, unlike logistic models. Unlike other machine learning algorithm (i.e. neuron networks), decision trees are highly interpretable, as its graphical representation enables the non-technical audience to understand the output (SPSS Decision Trees, 2016). As we learned its advantages, we chose the decision tree model to analyze the characteristics of the wine consumers in each of the three states who consume Mid-Atlantic wine.

Despite the number of advantages, a decision model has a risk of being overfitting. Overfitting happens when the model has overfit the current training set that fails to generalize. Several pruning methods are designed to avoid the appearance of overfitting. In this work, we chose Chi-pruning and cross-validation as ways to avoid overfitting. In this research, we used a decision tree model in SPSS with Exhaustive CHAID as the growing method. We selected dependent variable, Targetstate, which indicates the characteristics that describe a wine consumer living in New Jersey, New York, and Pennsylvania who consumes wine produced in the region. Integrating the independent and dependent variables together, we proposed the following initial hypothesis:

## RESULTS
### Survey Profile
Of the 977 participants who met the criteria and completed the survey, responses to demographic questions include: 62% were female, with a nearly equal distribution among age ranges. Most respondents resided in New York (48%), and, of those who provided a response to the question regarding marital status, 60% were married or in a partnership. When answering questions regarding their level of education and household income, 39% (n=972) had a bachelor's degree and 18% had a master's degree or above, while more than half of the participants had an annual household income of less than $76,000 (n=971).

Among all participants who indicated they consumed wines from New Jersey, New York, and Pennsylvania, about half of the respondents claimed they drank wines from the three states less than once a month. Thus, this variable was used to separate the target market. Pertaining to differences in consumption frequency of wine produced in their state of residence, New York participants reported drinking the state wine more frequently than consumers

residing in New Jersey and Pennsylvania.

Survey participants were asked their opinions about New York, New Jersey, and Pennsylvania wines. In terms of consumer retention rate (willingness to continue purchasing wine from the state), New Jersey has the lowest at 47% compared to New York (70%). Pennsylvania has the rate at 53% among the target market. In terms of recommendation rate (willingness to recommend wine from the state to others), New York has the highest at 58.4%. Pennsylvania is acceptable at 46.5%, and New Jersey is much lower at 36.1%.

In terms of perceived wine quality, New York wine target market have the highest perceived quality with minimal differences between varieties (about 70% on average). Pennsylvania wine target market also have comparably high perceived quality (about 65%) than New Jersey (about 45%).

Pertaining to sources of knowledge, most survey respondents in the target market claimed they primarily learn from their family and friends (NJ 68%, NY 74%, PA 72%), followed by wine and liquor store staff (NJ 52%, NY 54%, PA 51%) and winery tasting staff (NJ 45%, NY 51%, PA 49%). Among indirect marketing, food and cooking magazines (NJ 38%, NY 36%, PA 37%) is the best method. For this reason, wineries can recommend wine pairings for recipes in food and cooking magazines.

Mid-Atlantic Wine Target Market Decision Tree Model by State
1. New Jersey Decision Tree
The model has an overall 71.5% correct prediction. The results, in conclusion, suggests that New Jersey wine producers should target Mid-Atlantic wine consumers who are young (45 years old) non-Pennsylvania residents; 2) Younger (<45 years old) non-super core wine consumers who live in Pennsylvania.

Figure 3: Simplified Decision Tree for Mid-Atlantic Wine Consumers who Drink Pennsylvania Wine

Discussion
Similarities and Differences between New Jersey, New York, and Pennsylvania Wine Target Market
Similarities exist pertaining to characteristics that describe participants likely to consume wine from each of the three targeted states: a) being a super-core wine consumer, 2) age less than 45 years old, and 3) resident of the state where the wine is produced. Compared to participants' attitudes about New Jersey and Pennsylvania wines, New York wines have a bigger target market since they can not only target New York super core consumers but also New York non-super core wine consumers and young (age<45) super core wine consumers from New Jersey and Pennsylvania. Target market of New Jersey and Pennsylvania wines, however, can only attract super cores living locally. Hence, the most important factor that identifies the target market of New York wines is residency, different from New Jersey and Pennsylvania that are super cores.

Marketing channel effectiveness was similar for wine consumers who resided in each of the three states, with learning about wine from family and friends being the most effective way to reach target consumers. Food and cooking magazines are the most effective indirect marketing channel.

Several differences are revealed. New York wines are more widely accepted, and its perceived quality is higher than the two other states. Perceived quality of Pennsylvania wine is slightly lower than New York wine but much higher than New Jersey wines. None of the states has very high perceived quality.

Relationship Between Super Core Consumers and Mid-Atlantic Target Markets
In this research, we observed that being a super-core wine consumer is one of the decisive factors when reaching consumers in the Mid-Atlantic region who drink Mid-Atlantic wine. However, we also observed that target market consumers also reside locally. For this reason, Mid-Atlantic wine target consumers also highlighted the word "local" when they try to reach their target consumers. In addition, Mid-Atlantic wine target consumers are oftentimes younger (age 45 years old) (Chen et. al AAWE Presentation, 2017). Considering the reputation and retention rate of the Mid-Atlantic wines, we conclude that Mid-Atlantic wine consumers should not only be a super core wine consumers who drink wine frequently but also a wine enthusiasts at the same time who actively look for and try out new wines.

Figure 4: Generalized Target Market for Mid-Atlantic Region

Conclusion

This research introduced a decision tree algorithm to the field of wine economics. Compared to some previous studies that used decision trees, this research emphasized how to interpret and generate insights from the mode. Our research also highlighted the uniqueness of the Mid-Atlantic wine production region as well as the different characteristics between the target consumers from the Mid-Atlantic states that were not studied before.

The typical target market of the Mid-Atlantic region should be both super core consumers who drink wine frequently and wine enthusiasts who consume wine produced in their state of residence at the same time. However, the Mid-Atlantic region does not have a high perceived quality. This low perceived quality may result in lower retention rate and recommendation rate. Since word of mouth (i.e. learning from family and friends) is most popular among wine consumers, wineries in the Mid-Atlantic region should keep working on improving their perceived qualities.

**File Upload (PDF only)**

- Decision-Tree-AAWE-Abstract-Review-RG-2.pdf

**Privacy**

- Mid-Atlantic Wine Industry Target Market Analysis Using Decision Trees
- By using this form you agree with the storage and handling of your data by this website.

**Mid-Atlantic Wine Industry Target Market Analysis Using Decision Trees**

Chen Yue [1], Ramu Govindasamy [1], Kathleen Kelley [2]

1. Department of Agricultural, Food, and Resource Economics, Rutgers University. New Brunswick, NJ, USA.

2. Department of Plant Science, The Pennsylvania State University, University Park, PA, USA.

**Introduction**

Recently, as the development of computers and machine learning algorithms advance, it has become possible to apply this technology to data collected from wine consumers to learn about their attitudes and behaviors. Though readily available, only a few studies have applied algorithms in this manner to discover wine consumer insights. For this reason, it is necessary to start studying their application.

In this study, we specifically focus on the Mid-Atlantic wine production region (New York, New Jersey, and Pennsylvania), as there have been few wine-consumer studies that have highlighted any potential unique characteristics specific to wine drinkers in the region. Although Kelley *et al.* (2015) segmented the Mid-Atlantic market based on the frequency (super core, core, and marginal) with which they consumed the beverage (the summation of both domestic and imported wine), little is known specifically about these consumers consumption of local wines. Considering its market share and reputation, the Mid-Atlantic wine industry needs to identify its uniqueness and differentiate itself from the general wine market to stay competitive. Because the Mid-Atlantic Region does not have very high market share, only targeting the frequent wine drinkers can be misleading. Thus, this study fills the gap with analysis performed based solely on wines produced in each of the three Mid-Atlantic states. Using such an approach, we were able to identify points of differentiation for wines from each of the three states and the respective target markets.

**Methodology**

*Source of Data*

A consumer survey approach was taken. A 15-minute Internet survey was administered on 22-24 October 2014 to 977 of Survey Sampling International (Shelton, CT) panelists residing in New Jersey, New York, and Pennsylvania. Panelists were screened for not being a member of the wine industry, being at least 21 years old, and having purchased and drunk wine at least once within the previous year.

*Decision Tree Models*

In addition to descriptive statistics, we employed a decision tree approach to characterize the target market for each of the three states. Decision trees use learning algorithms that define the splits at each internal node. The tree starts from a root node that then splits into two or more children nodes (subgroups) that visually looks like an "upside down" or "sideways" tree. Each child node, if a further separation occurs, becomes a parent node. If no further separation occurs, the child node becomes a leaf node with no possibility of splitting any further. Except for the root node and the leaf nodes, each node will have exactly one parent and two or more children.

The decision tree model is specialized in handling categorical data. Compared to a logistic regression model,

a decision tree offers many advantages. One advantage that implementation is fast and variable selection is automatic during model fitting (Deshpande, 2011). Additionally, missing data are well handled and normalization or scaling is not required prior to fitting the tree. The model is also insensitive to outliers as the splits are made based on the proportion of samples instead of on the absolute values (Deshpande, 2011). When working with survey data, the tree can also eliminate the need to create excessive of excessive dummy variables for multinomial data as each category are assigned to the subgroup that optimizes the model. Moreover, the model does not assume linear relationships between the variables, unlike logistic models. Unlike other machine learning algorithm (i.e. neuron networks), decision trees are highly interpretable, as its graphical representation enables the non-technical audience to understand the output (SPSS Decision Trees, 2016). As we learned its advantages, we chose the decision tree model to analyze the characteristics of the wine consumers in each of the three states who consume Mid-Atlantic wine.

Despite the number of advantages, a decision model has a risk of being overfitting. Overfitting happens when the model has overfit the current training set that fails to generalize. Several pruning methods are designed to avoid the appearance of overfitting. In this work, we chose Chi-pruning and cross-validation as ways to avoid overfitting.

In this research, we used a decision tree model in SPSS with Exhaustive CHAID as the growing method. We selected dependent variable, $Target_{state}$, which indicates the characteristics that describe a wine consumer living in New Jersey, New York, and Pennsylvania who consumes wine produced in the region. Integrating the independent and dependent variables together, we proposed the following initial hypothesis:

$$Target_{State} = f(SuperCore, Bottle, Type, Knowledge, Occasion, Gender, Age, Residency, MaritalStatus, Education, Income, Limitations, Family_{DrinkWine}, Donation, SuperCore)$$

## RESULTS

*Survey Profile*

Of the 977 participants who met the criteria and completed the survey, responses to demographic questions include: 62% were female, with a nearly equal distribution among age ranges. Most respondents resided in New York (48%), and, of those who provided a response to the question regarding marital status, 60% were married or in a partnership. When answering questions regarding their level of education and household income, 39% (n=972) had a bachelor's degree and 18% had a master's degree or above, while more than half of the participants had an annual household income of less than $76,000 (n=971).

Among all participants who indicated they consumed wines from New Jersey, New York, and Pennsylvania, about half of the respondents claimed they drank wines from the three states less than once a month. Thus, this variable was used to separate the target market. Pertaining to differences in consumption frequency of wine produced in their state of residence, New York participants reported drinking the state wine more frequently than consumers residing in New Jersey and Pennsylvania.

Survey participants were asked their opinions about New York, New Jersey, and Pennsylvania wines. In terms

of consumer retention rate (willingness to continue purchasing wine from the state), New Jersey has the lowest at 47% compared to New York (70%). Pennsylvania has the rate at 53% among the target market. In terms of recommendation rate (willingness to recommend wine from the state to others), New York has the highest at 58.4%. Pennsylvania is acceptable at 46.5%, and New Jersey is much lower at 36.1%.

In terms of perceived wine quality, New York wine target market have the highest perceived quality with minimal differences between varieties (about 70% on average). Pennsylvania wine target market also have comparably high perceived quality (about 65%) than New Jersey (about 45%).

Pertaining to sources of knowledge, most survey respondents in the target market claimed they primarily learn from their family and friends (NJ 68%, NY 74%, PA 72%), followed by wine and liquor store staff (NJ 52%, NY 54%, PA 51%) and winery tasting staff (NJ 45%, NY 51%, PA 49%). Among indirect marketing, food and cooking magazines (NJ 38%, NY 36%, PA 37%) is the best method. For this reason, wineries can recommend wine pairings for recipes in food and cooking magazines.

*Mid-Atlantic Wine Target Market Decision Tree Model by State*

1. New Jersey Decision Tree

The model has an overall 71.5% correct prediction. The results, in conclusion, suggests that New Jersey wine producers should target Mid-Atlantic wine consumers who are young (< 45 years old), who are super-core wine drinkers (those who drink any wine more than once a week) and either: a) purchased a bottle of wine in the past because profits were donated to a non-profit organization or b) they did not graduate from a four-year college.  Mid-Atlantic wine consumers who: a) are not super-core wine consumers, b) are over the age of 45 years, and/or c) are age 45 years or younger, have purchased wine because profits were donated to a non-profit, have purchased wine, and have a four-year college degree are less likely to purchase New Jersey wine.

Figure 1: Simplified Decision Tree for Mid-Atlantic Wine Consumers who Drink New Jersey Wine
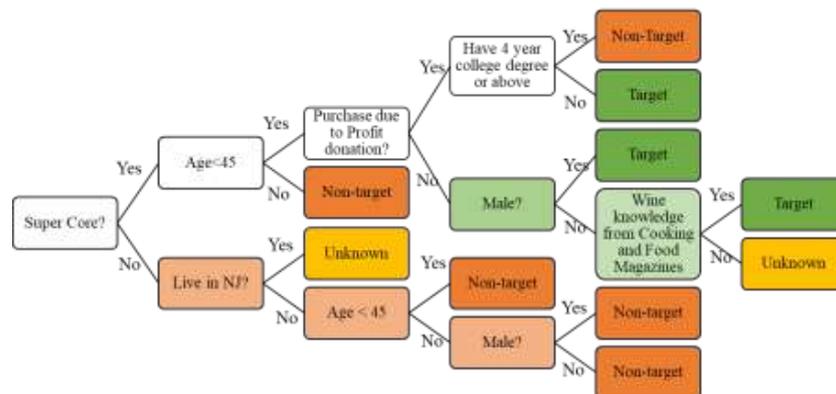


2. New York Decision Tree

Statistically, the New York model has a 71.1% overall correct prediction rate. The target markets of New York wines are 1) New York residents; and 2) New Jersey and Pennsylvania young super core wine consumers.

New York wines were more widely accepted since New Jersey and Pennsylvania young consumers are included in the target market in addition to New York consumers. For those living in New York, the difference between super cores and non-super cores is relatively small. New York wine generally has a bigger market than New Jersey.
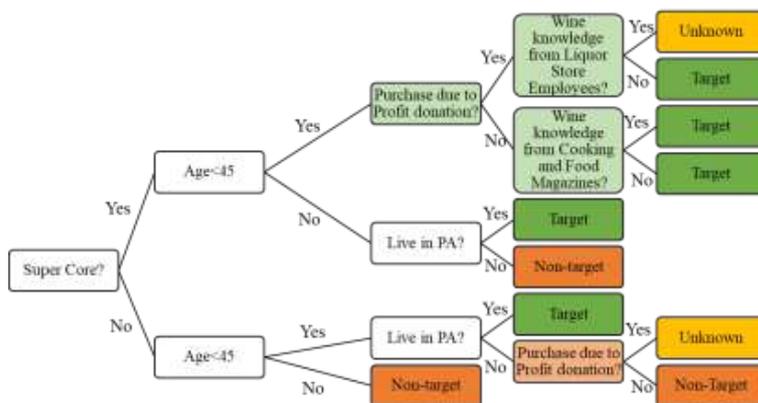
Figure 2: Simplified Decision Tree for Mid-Atlantic Wine Consumers who Drink New York Wine
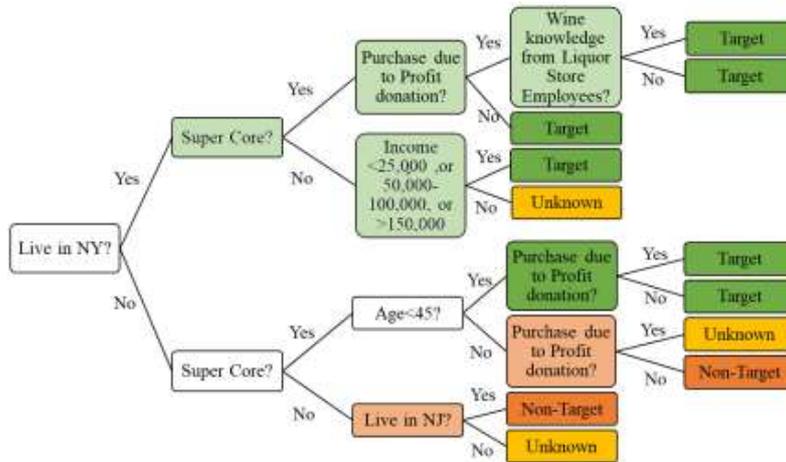


3. Pennsylvania Decision Tree

The Pennsylvania decision tree has an overall 70.9% correct prediction rate. The target market for Pennsylvania wine is 1) all super core wine consumers except those elder (>45 years old) non-Pennsylvania residents; 2) Younger (<45 years old) non-super core wine consumers who live in Pennsylvania.

Figure 3: Simplified Decision Tree for Mid-Atlantic Wine Consumers who Drink Pennsylvania Wine

**Discussion**

*Similarities and Differences between New Jersey, New York, and Pennsylvania Wine Target Market*

Similarities exist pertaining to characteristics that describe participants likely to consume wine from each of the three targeted states: a) being a super-core wine consumer, 2) age less than 45 years old, and 3) resident of the state where the wine is produced. Compared to participants' attitudes about New Jersey and Pennsylvania wines, New York wines have a bigger target market since they can not only target New York super core consumers but also New York non-super core wine consumers and young (age<45) super core wine consumers from New Jersey and Pennsylvania. Target market of New Jersey and Pennsylvania wines, however, can only attract super cores living locally. Hence, the most important factor that identifies the target market of New York wines is residency, different from New Jersey and Pennsylvania that are super cores.
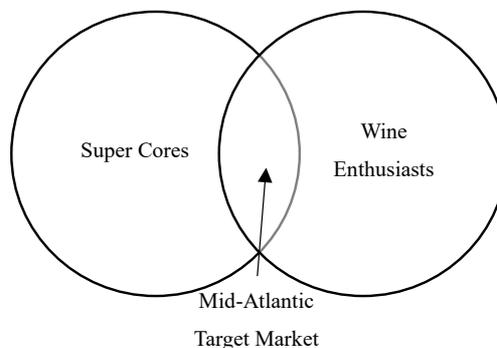
Marketing channel effectiveness was similar for wine consumers who resided in each of the three states, with learning about wine from family and friends being the most effective way to reach target consumers. Food and cooking magazines are the most effective indirect marketing channel.

Several differences are revealed. New York wines are more widely accepted, and its perceived quality is higher than the two other states. Perceived quality of Pennsylvania wine is slightly lower than New York wine but much higher than New Jersey wines. None of the states has very high perceived quality.

*Relationship Between Super Core Consumers and Mid-Atlantic Target Markets*

In this research, we observed that being a super-core wine consumer is one of the decisive factors when reaching consumers in the Mid-Atlantic region who drink Mid-Atlantic wine. However, we also observed that target market consumers also reside locally. For this reason, Mid-Atlantic wine target consumers also highlighted the word "local" when they try to reach their target consumers. In addition, Mid-Atlantic wine target consumers are oftentimes younger (age < 45 years old), unlike super core wine consumers being older (>45 years old) (Chen *et. al* AAWE Presentation, 2017). Considering the reputation and retention rate of the Mid-Atlantic wines, we conclude that Mid-Atlantic wine consumers should not only be a super core wine consumers who drink wine frequently but also a wine enthusiasts at the same time who actively look for and try out new wines.

Figure 4: Generalized Target Market for Mid-Atlantic Region



Super Cores

Wine Enthusiasts

Mid-Atlantic Target Market

**Conclusion**

This research introduced a decision tree algorithm to the field of wine economics. Compared to some previous studies that used decision trees, this research emphasized how to interpret and generate insights from the mode. Our research also highlighted the uniqueness of the Mid-Atlantic wine production region as well as the different characteristics between the target consumers from the Mid-Atlantic states that were not studied before.

The typical target market of the Mid-Atlantic region should be both super core consumers who drink wine frequently and wine enthusiasts who consume wine produced in their state of residence at the same time. However, the Mid-Atlantic region does not have a high perceived quality. This low perceived quality may result in lower retention rate and recommendation rate. Since word of mouth (i.e. learning from family and friends) is most popular among wine consumers, wineries in the Mid-Atlantic region should keep working on improving their perceived qualities.