Choice based Conjoint Analysis on Wine Preference of consumers in Italy

Kumar Swaviman

swaviman.kumar@studenti.unitn.it

Abstract **- This project aims to conduct a random survey design for collecting responses regarding wine preferences of Italian consumers. Furthermore, it attempts to understand how preference share gets affected as we vary different attributes associated with wine.**

Keywords - Conjoint Analysis, part worth estimate, mlogit, customer heterogeneity, random design

1. Introduction

Wine Industry in Italy takes one of the leading places in the local economy. Furthermore, Italy is a world-leading wine-producing country both in terms of volume and exports. There are more than 100 wine brands from more than 20 regions across Italy. In the previous studies (The Casini et al, 2009 paper) dedicated to wine consumer behavior, the main criteria for customers to evaluate and purchase wine are based on price, awards, region, the taste of wine or even personal recommendations [7]. Researchers also found (Lockshin et al, 2006 paper) some differences in respondents' preferences according to age, involvement level, and the geographical part they were from [8]. The research method we used, called Conjoint Analysis, is designed on the view that consumers' values are based on the utility offered by products' attributes.

Our study involves a series of interrelated stages which can be classified into 4 main steps. The first step in conducting the analysis is to identify suitable attributes and levels as motivators for consumer choice & conducting the survey based on the design chosen. The second is to perform data pre-processing and exploratory analysis on the collected data. Thirdly, we have to implement the Multinomial Choice model with Conditional Logistic Regression. The fourth & the last step is to implement the Multinomial Choice model but taking into account the customer heterogeneity as well. With these, further preference shares can be derived.

2. Problem definition: purpose of study

This project aims to investigate the choice-making process and preferences for wine among customers in Trento. In particular, the main goal is to determine how customers value different features of the wine and find the optimal levels of product attributes that maximize sales. Moreover, this project helps to answer questions about more popular types of wine among youth, providing information about the segmentation of customers and brands positioning. At the core of the project is conjoint analysis – a survey-based statistical technique, which allows to perform market research and estimate market share for products or services. The chosen method of the project is a choice-based conjoint survey, subsequently analyzed by using generalized linear models. Methods of conjoint analysis yield estimates of attribute trade-offs using a formal model for analysis.

3. Design of Conjoint Survey

The basic principles of designing a conjoint survey involve four different steps, such as determination of the type of study, identifying the relevant attributes, specifying the attributes' levels

and designing a questionnaire.As regards the former step, it has been conducted a branded choice-based conjoint survey, where customers are asked to choose from several alternatives of the product profiles.

Considering the study design, it has selected the salient attributes and levels for constructing the hypothetical product profiles of wine. Since attributes of a product can be divided broadly into categorical (nominal, ordinal, binary) and quantitative classes, in our study it has been used 5 attributes in total of both types. Particularly, nominal quantitative attributes of the survey include price, aging time of wine and percentage of alcohol, while categorical attributes involve brand and type of wine. Each attribute occurs at some level, therefore, it has been chosen from 4 to 5 levels for the attributes, based on the understanding of the consumer's choice process.

3.1 Attributes and their levels:

1. **Price per bottle**: Value (3-10€), Popular (10-15€), Premium (15-30€), Luxury (50-100€)
2. **Brands**: Cavit, Mezzacorona, Ferrari, Cantina Toblino
3. **Type of wine**: red wine, white wine, sparkling wine, rose wine.
4. **Alcohol Percentage**: 5.5%, 7%, 12%, 18%
5. **Wine aging time**: 1 year, 2 years, 3 years, 4 years, 5 years

In order to collect data for the conjoint analysis, the choice-based conjoint survey was conducted. The designed survey consists of 7 questions with 4 alternatives for each question and the questionnaire deals with 4-5 attributes with each level. As a result, each respondent had to evaluate a total number of **7 × 4 = 28** product profiles. To cover more combinations, random design of the conjoint survey is chosen. Thus, from **4× 4 × 4 × 4 × 5 = 1280** combinations in total, the questionnaires cover 840 combinations of product profiles. Considering time and resource limitations in conducting the survey, the determined sample size of the survey is 30 respondents in total. The respondents were asked to choose one out of 4 product profiles of wine, which they are more likely to buy. For example, the choice between four bottles of wine with different characteristics:

	☐	☐	☐	☐
Options	Option 1	Option 2	Option 3	Option 4
Price	Popular(10€-15€)	Value(3€-10€)	Popular(10€-15€)	Value(3€-10€)
Brand	Cantina Toblino	Ferrari	Ferrari	Mezzacorona
Type of Wine	Sparkling wine	White wine	Sparkling wine	Rose wine
Percentage of Alcohol	12 %	7 %	7 %	12 %
Aging time of Wine	3 years	2 years	2 years	5 years

Fig.1 Example of a question in the conjoint survey

Moreover, in addition to the choice-based conjoint questions, social-demographic data of the respondents, such as age, gender, level of education, employment status, ethnicity, usage of social media platforms and level of daily physical activity was collected.

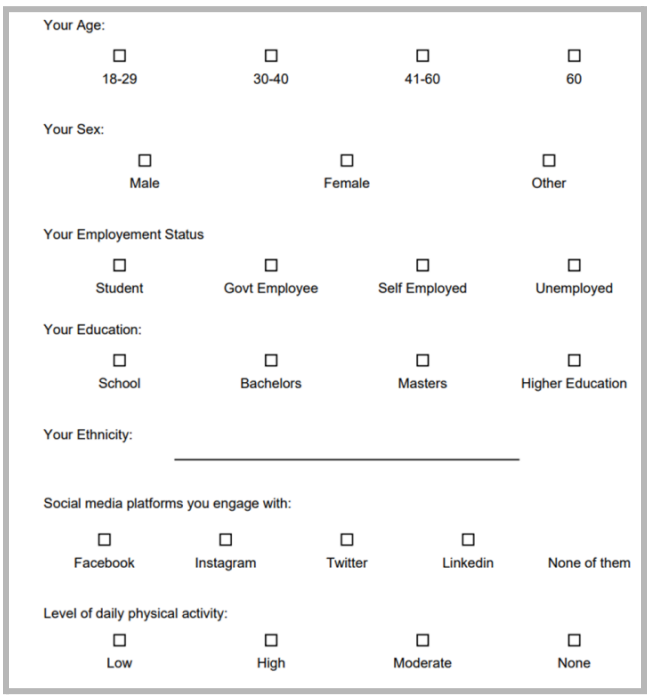


Fig.2 Example of social-demographic questions in the conjoint survey

4. Exploratory Data Analysis

It is important to analyze basic descriptive statistics of the collected data set to understand its initial features and summarize the main characteristics. Descriptive statistics also help to understand if there are any imbalances as well as missing values in the data. Firstly, we analyzed the descriptive social-demographic statistics of the respondents.

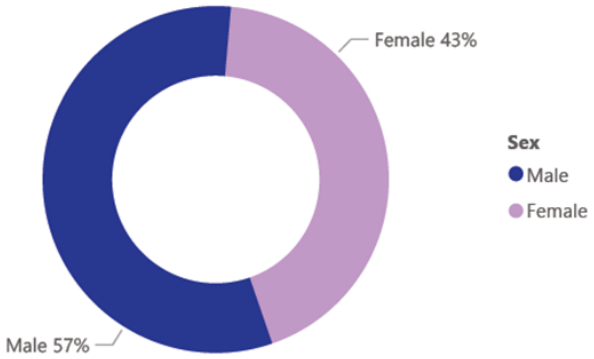


Fig.3 Respondents distribution by gender

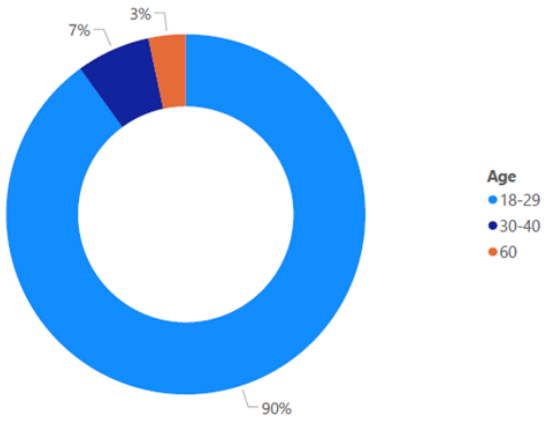


Fig.4 Respondents distribution by age

From the data we can see that females are 43% while male represent around 57% of the population. The vast majority of the respondents belong to the age group of 18 to 29 years. The remaining participants are included in the following groups: 30-40 years old-7% of the respondents; 60+ years old-3% of the respondents.

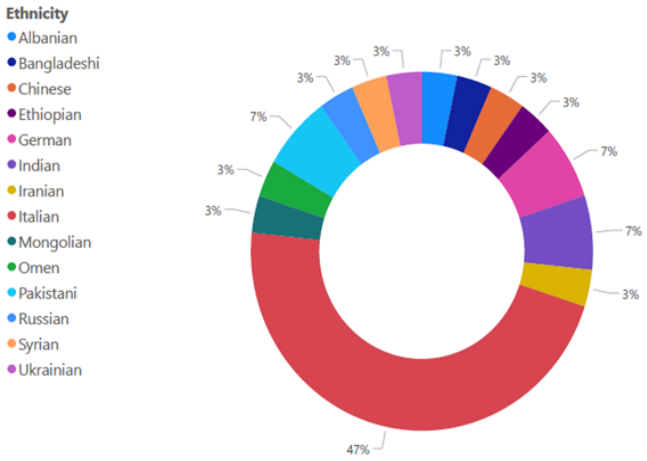


Fig. 5 Respondents distribution by ethnicity

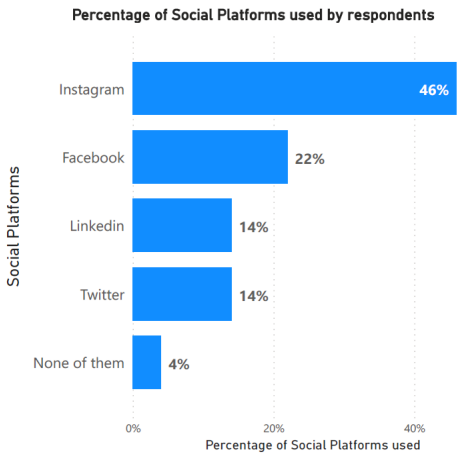


Fig.6 Percentage of Social Platforms used by respondents

The graph indicates that more than 45% of respondents are Italians, while other nationalities spread in the proportion of 3% and 7%. The most popular Social Platform used by 46% of respondents is Instagram, followed by Facebook, Linkedin, and Twitter with 22% and 14% of respondents respectively**.** Summarizing the collected data on choices of respondents, we estimated that the most preferred attributes among price segments and brands are Ferrari and Popular (10-15€) price segments.

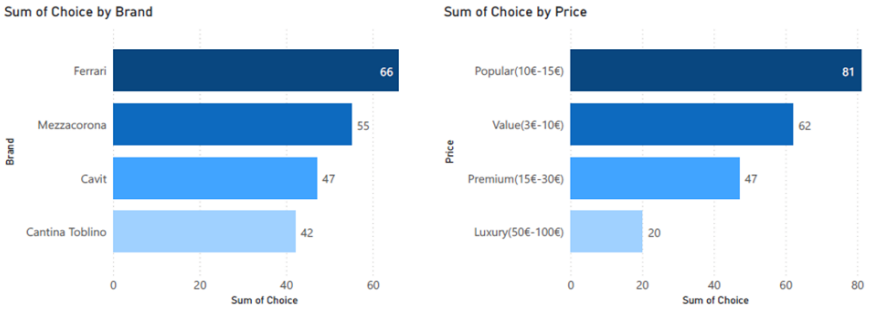


Fig.7 Choice distribution by brand and by price

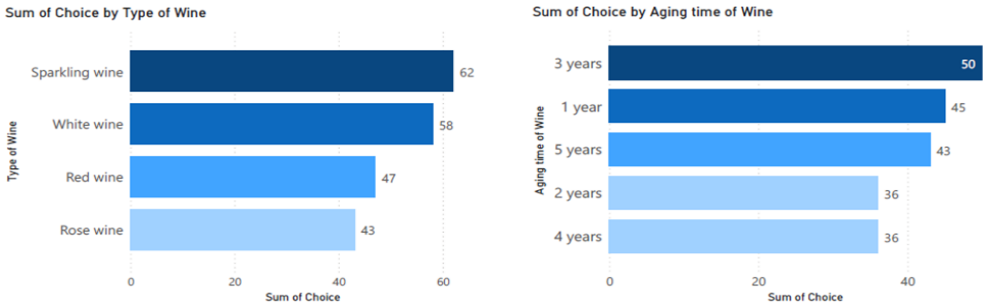


Fig. 8 Choice distribution by type of wine and by aging time

In addition, it is noticeable that when evaluating white wine options, respondents prefer Mezzacorona and Cavit brands the most, whereas for other types of wine the most widely chosen brand is Ferrari.

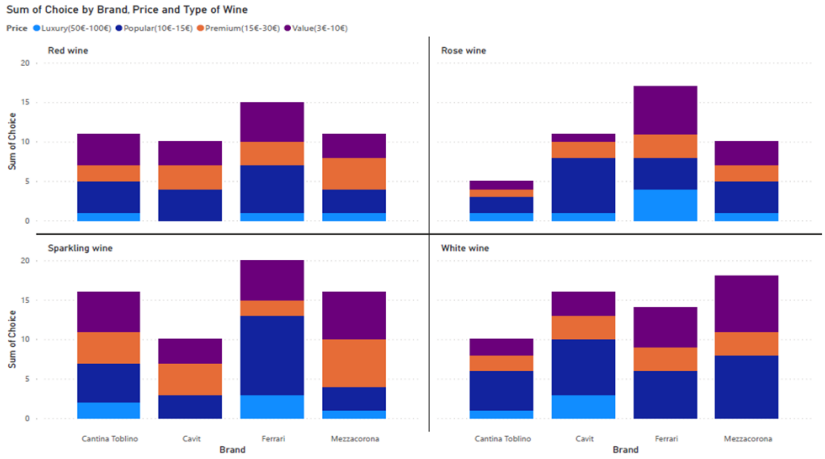
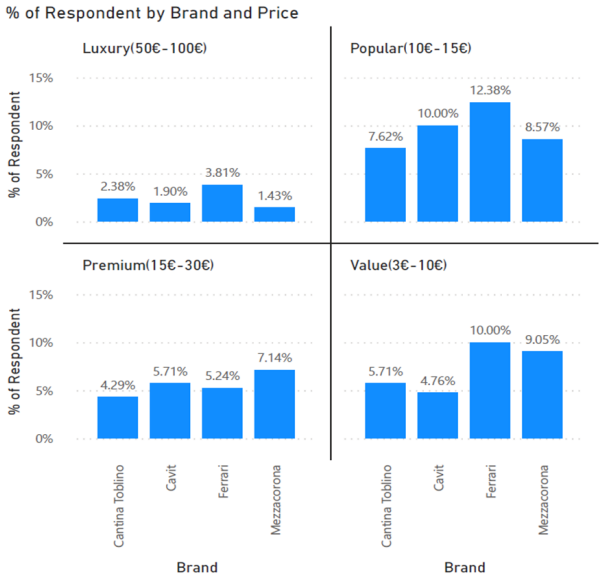


Fig.9 Choice distribution by type of wine, brand, and price



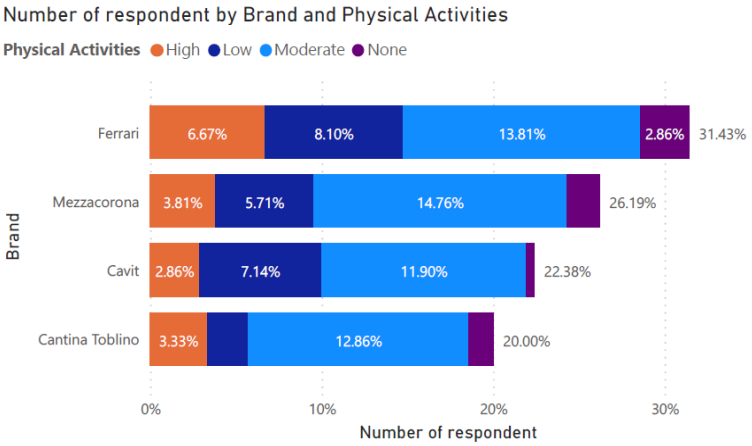
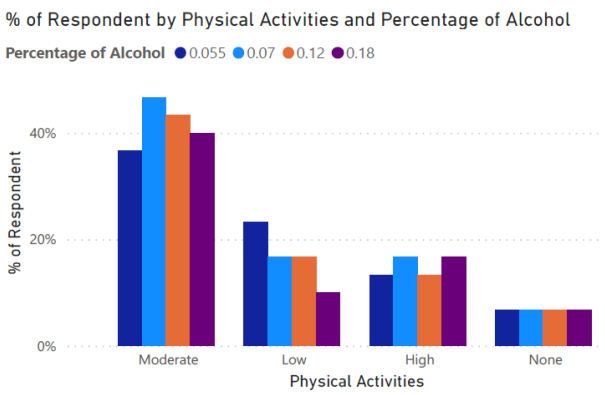


Fig.10 Percentage of respondents distribution by brand and price

Fig.11 No. of respondents by brand and physical activities

Moreover, if we analyze Fig.10, in the Value (3-10€) price segment the 2 popular brands are Ferrari and Mezzacorona & in the Popular (10-15€) price segment the popular brands are Ferrari and Cavit. Nevertheless, in the Premium(15-30€) segment Mezzacorona is taking the leading place, followed by Cavit. The following Fig. 11 represents the chosen brands and the level of physical activities of the respondents. Overall, the majority of people with moderately active daily life chose Mezzacorona.



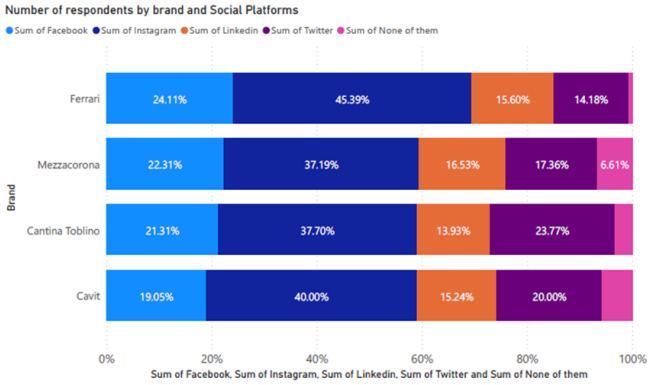


Fig. 12 Percentage of respondents distributed by physical activities and percentage of Alcohol

Fig. 13 Number of respondents by brand and Social Platforms

Analyzing the graph Fig. 12, it is noticeable that respondents with a moderate and high level of physical activity prefer 7% of alcohol, while respondents with a low level of activity prefer 5.5% of alcohol.

The Fig. 13 indicates that respondents, who chose the Ferrari brand in the questionnaire, use Twitter less than the other platforms. In contrast, respondents who chose Cantina Toblino are engaged with Twitter more. Moreover, among Ferrari's choices, only a small percentage of respondents don't use any social platforms. Such analysis shows certain platforms s.a. Instagram & Facebook can be leveraged by brands in running effective marketing campaigns.

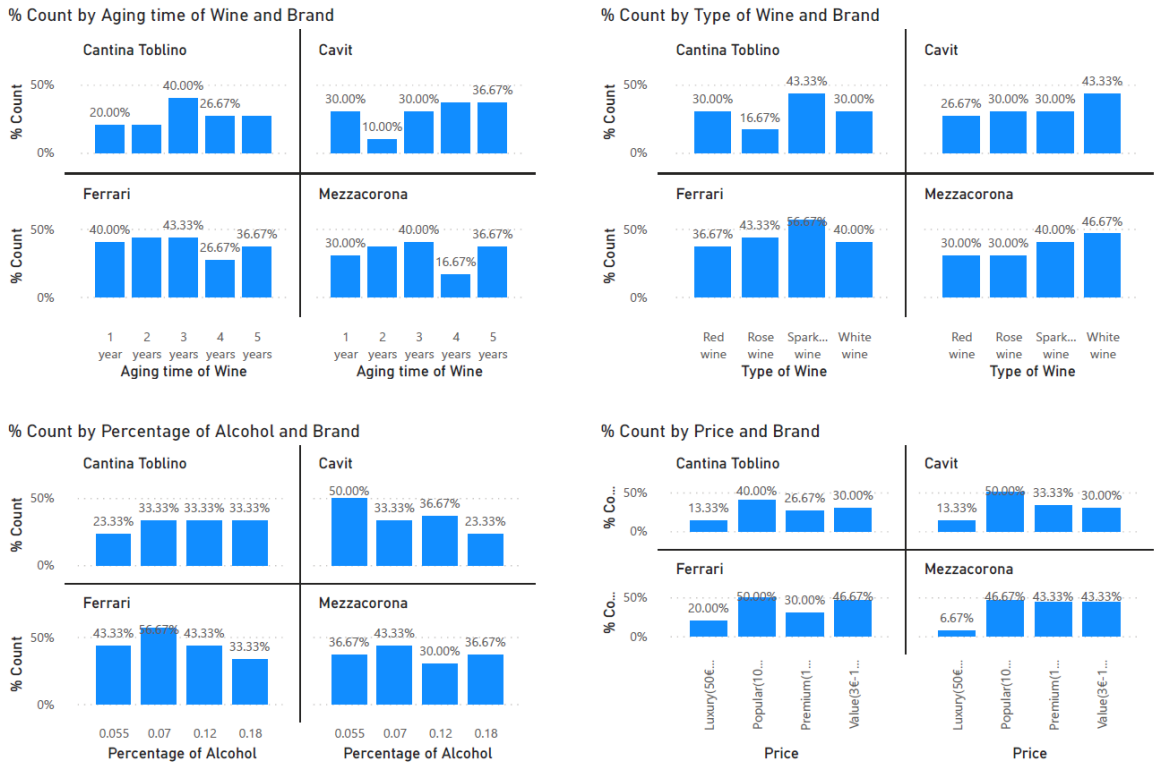


Fig. 14 Distribution by brand and other 3 attributes

5. Choice Model with Conditional Logistic Regression

5.1 Utility function

Chosen forms of component utility functions are part worth function and vector models:

$$Y_i = U_1(x_{i1}) + U_2(x_{i2}) + U_3(x_{i3}) + U_4(x_{i4}) + U_5(x_{i5}) + Error$$

Y_i-preference rating given to the ith product

x_{i1}-level of the price for the ith product

x_{i2}- level of the brand for the ith product

x_{i3}- level of the type of wine for the ith product

x_{i4}- level of the alcohol percentage for the ith product

x_{i5}- level of the wine aging time for the ith product

U_1 - partworth function for price

U_2- partworth function for brand

U_3 - partworth function for type of wine

U_4 - partworth function for alcohol percentage

U_5 - partworth function for wine aging time

In this case, the probability of choosing the alternative i is increasing with the biggest value Y_1. Therefore, probability of choice of any alternative for the individual is equal to:

$$P_1 = \frac{e^{Y_1}}{e^{Y_1}+e^{Y_2}+e^{Y_3}+e^{Y_4}}, P_2 = \frac{e^{Y_2}}{e^{Y_1}+e^{Y_2}+e^{Y_3}+e^{Y_4}}, P_3 = \frac{e^{Y_2}}{e^{Y_1}+e^{Y_2}+e^{Y_3}+e^{Y_4}}, P_4 = \frac{e^{Y_4}}{e^{Y_1}+e^{Y_2}+e^{Y_3}+e^{Y_4}},$$

$$\sum_{j=1}^{3} P_j = 1 \ , \ 0 \le P_j \le 1, \ \forall_i = 1, 2, 3, 4$$

5.2 Multinomial Logit Model

Model description

Multinomial logit model is used for analyzing choice-based conjoint data. In the usual multinomial logit model, the expected utilities are modeled in terms of the characteristics of the individuals. Utilities derived by the individual for four alternatives are:

$$U(x_{ij}) = \alpha_j + \beta x_{ij} + \delta_j z_{ij} + \gamma_j \omega_i,$$

where three kinds of covariates are considered:

- alternative-specific attributes x_{ij} with common coefficients β
- alternative-specific attributes z_{ij} with alternative-specific coefficients δ_j
- individual-level variables ω_i with alternative-specific coefficients γ_j

Thus, probability of choice:

$$p_{ij} = \frac{exp(x_i \beta_j)}{1 + \sum_{h=2}^{J} exp(x_i \beta_h)}, j = 2, \ldots, J, \ p_{i1} = 1 - \sum_{j=2}^{J} p_{ij}$$

In order to use the collected dataset in the multinomial logit model, it has been organized into the following data format:

```
       X              resp.id          ques          alt
 Min.   :  0.0   Min.   : 1.0   Min.   :1   Length:840
 1st Qu.:209.8   1st Qu.: 8.0   1st Qu.:2   Class :character
 Median :419.5   Median :15.5   Median :4   Mode  :character
 Mean   :419.5   Mean   :15.5   Mean   :4
 3rd Qu.:629.2   3rd Qu.:23.0   3rd Qu.:6
 Max.   :839.0   Max.   :30.0   Max.   :7
            Price                  Brand           Type.of.Wine
 Luxury(50€-100€):198   Cantina Toblino:231   Red wine      :199
 Popular(10€-15€):247   Cavit           :203   Rose wine     :194
 Premium(15€-30€):200   Ferrari         :213   Sparkling wine:234
 Value(3€-10€)   :195   Mezzacorona     :193   White wine    :213
```

```
 Percentage.of.Alcohol Aging.time.of.Wine      Age            Sex
 12 % :218             1 year :162        Min.   :1.000   Length:840
 18 % :192             2 years:163        1st Qu.:1.000   Class :character
 5.5 %:202             3 years:185        Median :1.000   Mode  :character
 7 %  :228             4 years:159        Mean   :1.167
                       5 years:171        3rd Qu.:1.000
                                          Max.   :4.000
 Employment.status   Education          Ethnicity         Social.Platforms
 Length:840          Length:840         Length:840        Length:840
 Class :character    Class :character   Class :character  Class :character
 Mode  :character    Mode  :character   Mode  :character  Mode  :character


 Physical.Activities     Choice
 Length:840          Min.   :0.00
 Class :character    1st Qu.:0.00
 Mode  :character    Median :0.00
                     Mean   :0.25
                     3rd Qu.:0.25
                     Max.   :1.00
```

5.3 Fitting the Multinomial Logit Model without intercepts

A model with only alternative specific variables is called a conditional logit model. Considering that this model does not have intercepts α_j=0, the equation for the utilities can be simplified into:

$$U(x_{ij}) = \beta x_{ij}\ , \quad Y_i = \beta x_{ij} + e_i\ ,$$

To calculate the data, it was used the mlogit package:

```
Call:
mlogit(formula = Choice ~ 0 + Price + Brand + Type.of.Wine +
    Percentage.of.Alcohol + Aging.time.of.Wine, data = cbc.mlogit,
    method = "nr")

Frequencies of alternatives:choice
     1      2      3      4
0.20476 0.26667 0.22381 0.30476

nr method
4 iterations, 0h:0m:0s
g'(-H)^-1g = 7.07E-07
gradient close to zero

Coefficients :
                          Estimate Std. Error z-value  Pr(>|z|)
PricePopular(10€-15€)     1.503812   0.274514  5.4781 4.299e-08 ***
PricePremium(15€-30€)     1.034239   0.294471  3.5122 0.0004444 ***
PriceValue(3€-10€)        1.423789   0.285615  4.9850 6.197e-07 ***
```

```
BrandCavit                  0.353767   0.250166  1.4141 0.1573237
BrandFerrari                0.875427   0.243341  3.5975 0.0003213 ***
BrandMezzacorona            0.663960   0.250648  2.6490 0.0080736 **
Type.of.WineRose wine      -0.083027   0.254559 -0.3262 0.7443021
Type.of.WineSparkling wine  0.217729   0.236474  0.9207 0.3571916
Type.of.WineWhite wine      0.238447   0.241399  0.9878 0.3232645
Percentage.of.Alcohol18 %   0.224076   0.246609  0.9086 0.3635468
Percentage.of.Alcohol5.5 %  0.388347   0.241419  1.6086 0.1077041
Percentage.of.Alcohol7 %    0.311218   0.239271  1.3007 0.1933633
Aging.time.of.Wine2 years  -0.437539   0.277900 -1.5744 0.1153839
Aging.time.of.Wine3 years   0.048895   0.257447  0.1899 0.8493700
Aging.time.of.Wine4 years  -0.229575   0.275379 -0.8337 0.4044683
Aging.time.of.Wine5 years  -0.024921   0.267110 -0.0933 0.9256665
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -260.09
```

The Estimate shows the mean values for each level, which must be understood in relation to each attribute's base levels with a magnitude of preferences from -2 to 2 on the logit scale. Considering this case study, the Estimate for price is measured relative to Luxury (50-100€) price segment base level. This leads to the fact that our simulated customers strongly preferred Popular (10-15€) and then Value (3-10€) price segments to the Luxury segment. Considering The Estimate values of brands, it is noticeable that the most preferred brand is Ferrari, followed by Mezzacorona and Cavit relative to Cantina Toblino brand. Likewise, it can be estimated that the simulated customers slightly dislike Rose wine in relation to Red wine and they tend to prefer 1 and 3 years of aging time to others.

The Std. Error indicates the level of precision of the estimates, which is directly proportional to the size of the conjoint survey data. These coefficients mean that the value of the estimate will fall

within the *estimate ±1.96× the std.error*. For example, assuming the data are representative, there is a 0.95 probability that the coefficient for price Popular (10-15€) varies between 2.02 and 0.97.

5.4 Fitting the Multinomial Logit Model with intercepts

The second estimated model in the study case is a multinomial logit with intercepts. The equation for the utility functions is the following:

$$U(x_{ij}) = \alpha_j + \beta x_{ij} ,$$

$$Y_i = \alpha_j + \beta x_{ij} + e_i ,$$

To estimate the second Multinomial Logit model, we used mlogit with a formula describing the model:

```
Call:
mlogit(formula = Choice ~ Price + Brand + Type.of.Wine + Percentage.of.Alcohol +
    Aging.time.of.Wine, data = cbc.mlogit, method = "nr")

Frequencies of alternatives:choice
     1      2      3      4
0.20476 0.26667 0.22381 0.30476

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 1.07E-06
successive function values within tolerance limits

Coefficients :
                          Estimate Std. Error z-value  Pr(>|z|)
(Intercept):2             0.270761   0.214650  1.2614 0.2071636
(Intercept):3             0.141093   0.223449  0.6314 0.5277566
(Intercept):4             0.389553   0.210497  1.8506 0.0642213 .
PricePopular(10€-15€)     1.496454   0.275723  5.4274 5.719e-08 ***
PricePremium(15€-30€)     1.011992   0.298370  3.3917 0.0006945 ***
PriceValue(3€-10€)        1.392920   0.287908  4.8381 1.311e-06 ***
```

```
BrandCavit                0.380192   0.253446  1.5001 0.1335906
BrandFerrari              0.873869   0.244195  3.5786 0.0003455 ***
BrandMezzacorona          0.696337   0.253610  2.7457 0.0060382 **
Type.of.WineRose wine    -0.079270   0.256657 -0.3089 0.7574297
Type.of.WineSparkling wine 0.246208   0.237565  1.0364 0.3000234
Type.of.WineWhite wine    0.252377   0.243251  1.0375 0.2994950
Percentage.of.Alcohol18 % 0.218244   0.248314  0.8789 0.3794536
Percentage.of.Alcohol5.5 % 0.418252   0.243099  1.7205 0.0853419 .
Percentage.of.Alcohol7 %  0.316370   0.240103  1.3176 0.1876230
Aging.time.of.Wine2 years -0.422761   0.279153 -1.5144 0.1299138
Aging.time.of.Wine3 years  0.045125   0.258801  0.1744 0.8615808
Aging.time.of.Wine4 years -0.231177   0.276693 -0.8355 0.4034370
Aging.time.of.Wine5 years -0.022488   0.267952 -0.0839 0.9331148
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -258.16
McFadden R^2:  0.10552
Likelihood ratio test : chisq = 60.91 (p.value = 3.6749e-07)
```

The alternative specific constants indicate a preference for the different positions in the question (1,2,3,4) in relation to the first position.

```
                          Estimate Std. Error z-value  Pr(>|z|)
(Intercept):2             0.270761   0.214650  1.2614 0.2071636
(Intercept):3             0.141093   0.223449  0.6314 0.5277566
(Intercept):4             0.389553   0.210497  1.8506 0.0642213
```

(Intercept):4 illustrates the relative preference of the fourth position in the question (versus the first), as well as (Intercept):2 and (Intercept):3 indicate the preference for the second and third positions to varying degrees relative to the first one. The estimated alternative specific constants are not significantly different from zero.

Considering the estimated values of part-worth coefficients of the multinomial logit model with intercepts, there is no significant difference from the multinomial logit model without intercepts.

5.5 Model Comparison

We built logistic regression models with/without intercepts and checked impact on results using a likelihood ratio test.

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
	<dbl>	<dbl>	<dbl>	< dbl>	<dbl>
1	16	-260.0942	NA	NA	NA
2	19	-258.1610	3	3.86641	0.2762554

Likelihood ratio test gives a p-value of 0.27, which is way higher than the conventional threshold of 0.05. Hence, we can conclude that the two models m1 and m2 fit the data equally well. We cannot reject the null hypothesis that there is no significant difference between m1 and m2 and we don't need the alternative specific constants to fit the present data. This also means, excluding intercepts from our model doesn't affect the result significantly. Hence for further analysis, we will only consider the m1 model which excludes intercepts.

5.6 Preference Shares Simulator

With the help of a share simulator, we can specify a variety of alternatives and then use the model to anticipate which options customers will select. In our case, we analyzed preference shares for different types of wine in different price segments among different brands.

We computed choice shares using a multinomial logit model without intercepts. Based on our exploratory analysis we observe that within Brand Cavit, customers prefer the Popular(10€-15€) price segment, white wine, 5.5% alcohol with 4 years of aging the most. Hence we design a combination for that which does not exist in our initial survey design (to be used as base design). We will compare this new proposed design against various combinations of other brands, based on the most chosen attributes within each brand.

	share	Price	Brand	Type.of.Wine	Alcohol	Aging.time
822	0.08791050	Popular(10€-15€)	Cavit	White wine	5.5 %	4 years
610	0.09992091	Popular(10€-15€)	Cantina Toblino	Sparkling wine	7 %	3 years
674	0.11776021	Popular(10€-15€)	Cantina Toblino	Sparkling wine	12 %	3 years
738	0.10901894	Popular(10€-15€)	Cantina Toblino	Sparkling wine	18 %	3 years
362	0.14743278	Popular(10€-15€)	Ferrari	Sparkling wine	7 %	2 years
618	0.23980020	Popular(10€-15€)	Ferrari	Sparkling wine	7 %	3 years
638	0.19815646	Popular(10€-15€)	Mezzacorona	White wine	7 %	3 years

As a result from the above table, we found that, among this set of the most popular wine profiles, we would expect 24% of consumers to choose Sparkling wine from Ferrari brand in the Popular(10€-15€) price segment with 7% of alcohol and 3 years of aging time.

5.7 Trade-off Table & Sensitivity Plot

A trade-off table and a sensitivity plot show the relative importance of different attributes and levels in a choice model, and how changing the levels of one attribute affects the likelihood of choosing one option over another. The trade-off table shows the part-worth or utility value of each attribute level, while the sensitivity plot shows how these values change as one attribute level is changed while keeping all others constant. These tools help to understand the trade-offs that consumers make between different attributes and levels and the degree to which they are willing to compromise on one attribute to get another. They also provide insights into how consumers value different attributes and how changes to them may affect demand.

A data.frame: 21 × 3

	level	share	increase
	<chr>	<dbl>	<dbl>
Price1	Value(3€-10€)	0.02097490	-0.066935605
Price2	Popular(10€-15€)	0.08791050	0.000000000
Price3	Premium(15€-30€)	0.05684019	-0.031070312
Price4	Luxury(50€-100€)	0.08170214	-0.006208358
Brand1	Cantina Toblino	0.06337662	-0.024533879
Brand2	Cavit	0.08791050	0.000000000
Brand3	Ferrari	0.13970298	0.051792480
Brand4	Mezzacorona	0.11616831	0.028257810
Type.of.Wine1	Red wine	0.07057659	-0.017333907
Type.of.Wine2	Rose wine	0.06532078	-0.022589726
Type.of.Wine3	Sparkling wine	0.08626338	-0.001647117
Type.of.Wine4	White wine	0.08791050	0.000000000
Percentage.of.Alcohol1	5.5 %	0.08791050	0.000000000
Percentage.of.Alcohol2	7 %	0.10761445	0.019703951
Percentage.of.Alcohol3	12 %	0.12443653	0.036526033
Percentage.of.Alcohol4	18 %	0.11627368	0.028363182
Aging.time.of.Wine1	1 year	0.10814370	0.020233202
Aging.time.of.Wine2	2 years	0.07260251	-0.015307992
Aging.time.of.Wine3	3 years	0.11295071	0.025040205
Aging.time.of.Wine4	4 years	0.08791050	0.000000000
Aging.time.of.Wine5	5 years	0.10576348	0.017852976

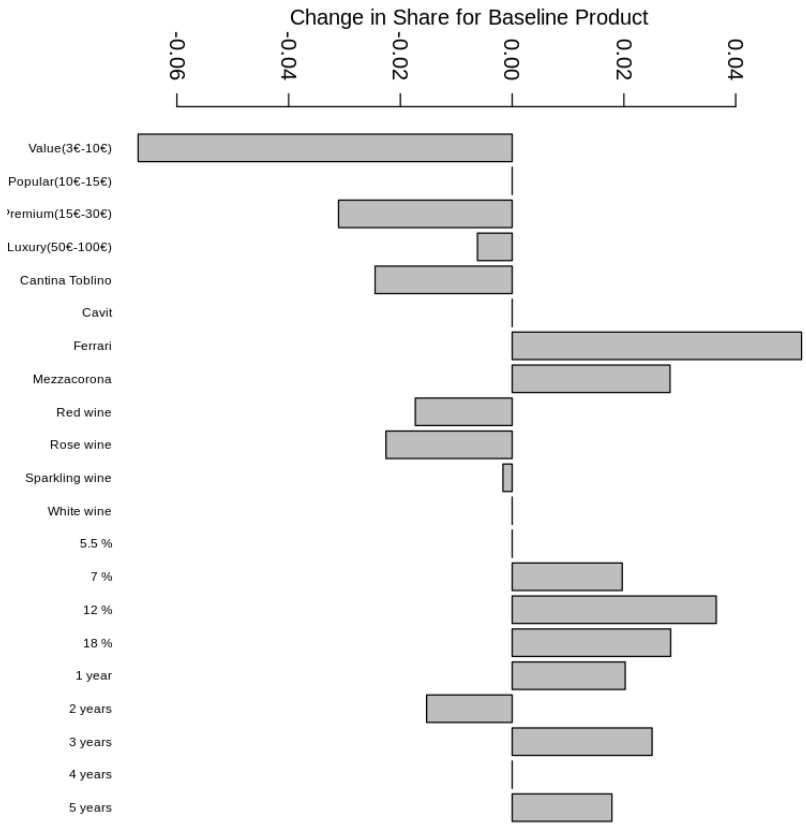


Fig. 15 Sensitivity plot for the MNL model m1

Based on the estimated preference shares, we can build a plot, representing how shares would change if we changed each of the attributes of the product profiles. This information on consumer preferences can be used to examine which product characteristics appeal to customers the most and how they trade off desirable qualities and pricing.Fig. 15 Sensitivity plot for the MNL model m1

The trade-off table gives a clearer picture of how the choice shares would get affected, when attributes are varied from the baseline design. For attributes “Price” & "Type.of.Wine" the base design offers the highest market share. For the rest of the attributes changing attribute levels to other alternatives increases the share positively. For example, in case of "Aging.time.of.Wine" changing from base design i.e. 4 years to 3 years shows 2.5% increase in market share.

6. Choice model considering Customer Heterogeneity

6.1 Mixed Multinomial logit model

Mixed Multinomial logit models allow to estimate individual-level coefficients for each respondent, which leads to possibilities to better fit data and produce more accurate predictions than sample-level models, since different persons have varied preferences. The statistical term for coefficients that vary across respondents is random coefficients, hence, the utility function is equal to:

$$U(x_{ij}) = \alpha_j + \beta x_{ij} \ , \ where \ \alpha_j = 0$$

$$Y_i = \beta x_{ij} + e_i \ , \ \beta = \sum_{r=1}^{R} \beta_r$$

In choice modeling, Correlated Random Parameters means that the parameters (part worths) that represent the preferences of individuals for different attributes (e.g. price, brand, etc) are not independent and have some correlation between them. On the other hand, Independent Random Parameters means that the parameters are completely uncorrelated and have no relationship between them.

The difference between the two methods is in the estimation of the parameters and the assumptions made about the underlying structure of the preferences. In the Correlated Random Parameters approach, the parameters are estimated in such a way that the correlation between them is taken into account. This leads to more accurate and robust results, as it accounts for the fact that individual preferences for different attributes are often not completely independent. In the Independent Random Parameters approach, the parameters are estimated without taking into account the correlations, leading to less accurate results and a higher risk of overfitting. For our project we have tried implementing both the options, first with independent parameters and later with correlation taken into account.

6.2 Mixed Model with Independent parameters

We can either allow the random parameters to be correlated or independent. If correlation is true, the correlation between the random parameters is taken into consideration by estimating the components of the Cholesky decomposition of the covariance matrix. However we first analyze a simpler case, with an assumption of independence among all parameters. All of the coefficients in our first model are assumed to follow a normal distribution across the population, and they make up the following vector m1.rpar.

```
m1.rpar <- rep("n", length=length(m1$coef))
names(m1.rpar) <- names(m1$coef)
```

```
Call:
mlogit(formula = Choice ~ 0 + Price + Brand + Type.of.Wine +
    Percentage.of.Alcohol + Aging.time.of.Wine, data = cbc.mlogit,
    rpar = m1.rpar, correlation = FALSE, panel = TRUE)

Frequencies of alternatives:choice
      1       2       3       4
0.20476 0.26667 0.22381 0.30476

bfgs method
34 iterations, 0h:0m:3s
g'(-H)^-1g = 4.87E-07
gradient close to zero

Coefficients :
                           Estimate Std. Error z-value  Pr(>|z|)
PricePopular(10€-15€)      2.066206   0.507813  4.0688 4.725e-05 ***
PricePremium(15€-30€)      1.383331   0.474098  2.9178 0.0035250 **
PriceValue(3€-10€)         1.598766   0.533355  2.9976 0.0027215 **
BrandCavit                 0.565028   0.423147  1.3353 0.1817780
BrandFerrari               1.422795   0.435506  3.2670 0.0010870 **
BrandMezzacorona           0.925398   0.429523  2.1545 0.0312026 *
Type.of.WineRose wine      0.027485   0.434740  0.0632 0.9495898
Type.of.WineSparkling wine 0.294434   0.411843  0.7149 0.4746597
Type.of.WineWhite wine     0.389964   0.350598  1.1123 0.2660169
Percentage.of.Alcohol18 %  0.174192   0.442555  0.3936 0.6938722
Percentage.of.Alcohol5.5 % 0.402435   0.387824  1.0377 0.2994221
Percentage.of.Alcohol7 %   0.280964   0.430537  0.6526 0.5140202
Aging.time.of.Wine2 years -0.258031   0.482987 -0.5342 0.5931762
Aging.time.of.Wine3 years  0.108704   0.397334  0.2736 0.7844053
Aging.time.of.Wine4 years -0.370008   0.405894 -0.9116 0.3619864
Aging.time.of.Wine5 years -0.208951   0.449284 -0.4651 0.6418770
```

```
sd.PricePopular(10€-15€)       1.822012   0.436281  4.1762 2.964e-05 ***
sd.PricePremium(15€-30€)       1.286740   0.325470  3.9535 7.702e-05 ***
sd.PriceValue(3€-10€)          1.625442   0.562105  2.8917 0.0038315 **
sd.BrandCavit                 -0.274731   0.412973 -0.6653 0.5058903
sd.BrandFerrari                1.175362   0.466189  2.5212 0.0116950 *
sd.BrandMezzacorona            0.760893   0.545387  1.3951 0.1629726
sd.Type.of.WineRose wine       0.568787   0.548433  1.0371 0.2996837
sd.Type.of.WineSparkling wine  1.497187   0.504948  2.9650 0.0030265 **
sd.Type.of.WineWhite wine     -0.023321   0.316193 -0.0738 0.9412041
sd.Percentage.of.Alcohol18 %   1.475199   0.383312  3.8486 0.0001188 ***
sd.Percentage.of.Alcohol5.5 %  1.838691   0.525430  3.4994 0.0004663 ***
sd.Percentage.of.Alcohol7 %    1.354252   0.568486  2.3822 0.0172092 *
sd.Aging.time.of.Wine2 years   0.504445   0.369534  1.3651 0.1722261
sd.Aging.time.of.Wine3 years  -0.376999   0.368555 -1.0229 0.3063508
sd.Aging.time.of.Wine4 years   0.269475   0.464019  0.5807 0.5614152
sd.Aging.time.of.Wine5 years  -0.257163   0.421165 -0.6106 0.5414659
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
random coefficients
                                Min.    1st Qu.     Median       Mean    3rd Qu.
PricePopular(10€-15€)         -Inf  0.8372783  2.06620645  2.06620645  3.29513459
PricePremium(15€-30€)         -Inf  0.5154378  1.38333062  1.38333062  2.25122347
PriceValue(3€-10€)            -Inf  0.5024217  1.59876595  1.59876595  2.69511017
BrandCavit                    -Inf  0.3797253  0.56502830  0.56502830  0.75033129
BrandFerrari                  -Inf  0.6300255  1.42279496  1.42279496  2.21556443
BrandMezzacorona              -Inf  0.4121833  0.92539804  0.92539804  1.43861281
Type.of.WineRose wine         -Inf -0.3561559  0.02748508  0.02748508  0.41112605
Type.of.WineSparkling wine    -Inf -0.7154031  0.29443404  0.29443404  1.30427122
Type.of.WineWhite wine        -Inf  0.3742341  0.38996412  0.38996412  0.40569412
Percentage.of.Alcohol18 %     -Inf -0.8208142  0.17419230  0.17419230  1.16919877
Percentage.of.Alcohol5.5 %    -Inf -0.8377436  0.40243490  0.40243490  1.64261338
Percentage.of.Alcohol7 %      -Inf -0.6324646  0.28096436  0.28096436  1.19439330
Aging.time.of.Wine2 years     -Inf -0.5982737 -0.25803055 -0.25803055  0.08221259
Aging.time.of.Wine3 years     -Inf -0.1455778  0.10870380  0.10870380  0.36298543
Aging.time.of.Wine4 years     -Inf -0.5517659 -0.37000777 -0.37000777 -0.18824966
Aging.time.of.Wine5 years     -Inf -0.3824045 -0.20895105 -0.20895105 -0.03549757
```

To estimate a multinomial logit model with random coefficients using mlogit, we build a vector specifying which coefficients vary across customers. For the first run, we assume that random parameters are not correlated, therefore we used correlation = FALSE.

Comparison can be made between the coefficients and the standard deviation coefficients. Thus, for price attributes, there is no significant difference in the values. Likewise, for Cavit brand the estimate of sd.BrandCavit is about 0.274 lower than the mean estimate of 0.565, which suggests that in general people prefer Cavit over Cantina Toblino. We can also analyze the range of respondent-level coefficients in the random coefficients section. Thus, for brand Cavit the first quartile is 0.379 (showing a preference for Cavit over Cantina Toblino), the third quartile is 0.750 and the mean is 0.565 (again indicating a preference for Cavit over Cantina Toblino).

Considering the distribution of respondent-level coefficients for rose wine, the majority of respondents slightly prefer rose wine to red wine, as the mean value is 0.027, very close to that of red wine. Similarly, in the random coefficients section, a number of people tend to prefer red wine over sparkling wine (the first quartile is equal to -0.715), while the majority prefer sparkling wine (the mean is about 0. 294).

Unlike, for 2 years of aging time attribute, the mean estimate is around -0.258, showing the preference of 1 year of aging time over 2 years. Moreover, the negative mean values and the random coefficients for Aging.time.of.Wine 5 years show that there is no fraction of people who prefer 5 years of aging over 1 year aging of wine.

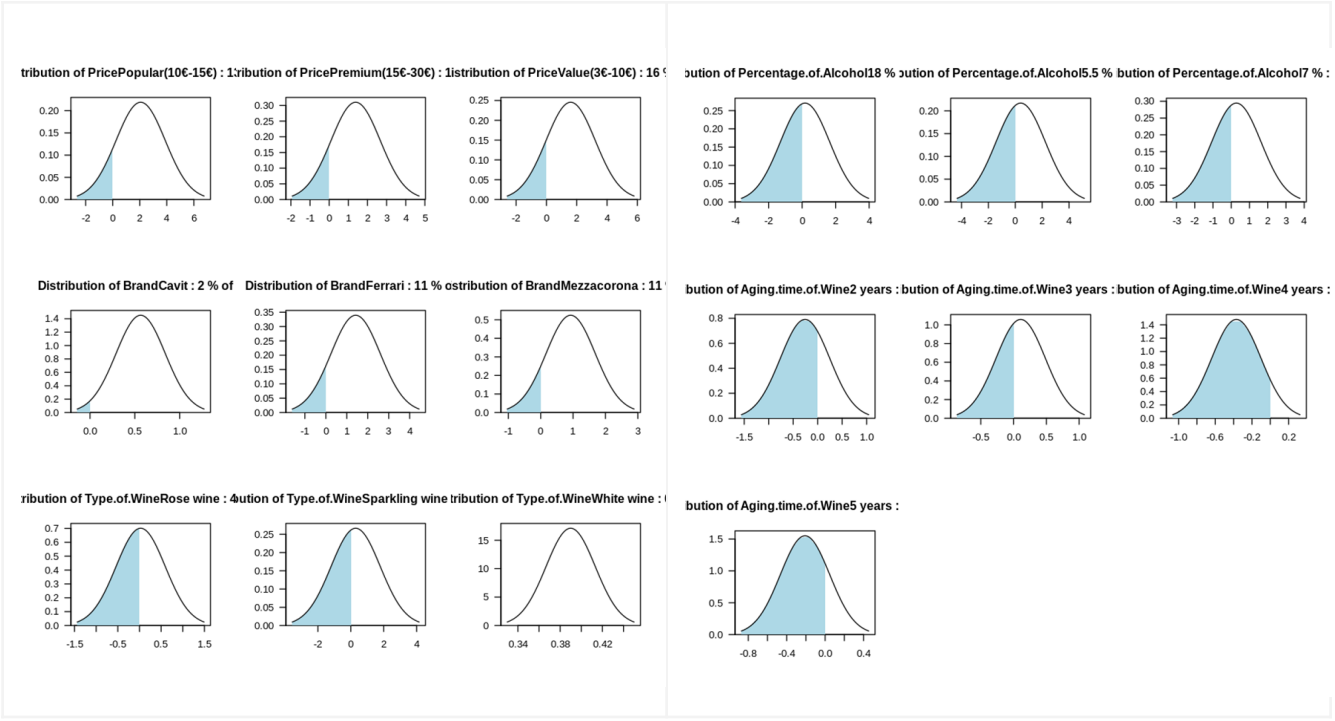


Fig. 16 Distributions of the attributes

6.3 Mixed model with correlated parameters

In this model by allowing the parameters to be correlated, the mixed multinomial logit model can account for more complex relationships between the predictors and the outcome variable and can provide more accurate predictions of the probabilities of each category for each individual.

This type of model is useful where the outcome of interest has multiple categories and the relationship between the predictors and the outcome is complex and may vary across individuals and categories. Including correlations in the random coefficients, it allows us to estimate based on the data if people who like one attribute also like another attribute.

Since our design matrix is not invertible, we practically can not build the same mixed model with correlation. Our data contains some strongly correlated variables. Hence we examined the pairwise covariance (or correlation) of our variables to investigate if there are any variables that can potentially be removed. For this we ran chi-square tests between variables. We found that a lower p-value and a higher X-squared indicates stronger evidence against the null hypothesis and therefore stronger evidence of dependence between the two variables. Based on this, we could say that the two variables "Brand" and "Aging time of Wine" are more dependent, with a p-value of 0.02779. So we decided to drop the feature, "Aging time of Wine".

```
chisq.test(cbc.mlogit$Brand, cbc.mlogit$Aging.time.of.Wine,correct=FALSE)

        Pearson's Chi-squared test

data:  cbc.mlogit$Brand and cbc.mlogit$Aging.time.of.Wine
X-squared = 22.992, df = 12, p-value = 0.02779
```

Our model with correlation set to TRUE gives a log-likelihood score. The log-likelihood is a measure of how well the specified model fits the observed data. A higher log-likelihood value indicates a better fit, and it is often used as an optimization criterion in statistical modeling. In our case, the log-likelihood value is -233.3 which is the same as the previous case when correlation was set to FALSE. It means that both the models fit the observed data reasonably well. However, the Correlated Random Parameters approach is more practical as it has no underlying assumptions of independence.

6.4 Preference Share Simulator

The preference share table summarizes the results of the Conjoint Analysis and provides insights into the factors that drive customer preferences and choice behavior. It helps to identify the most preferred and least preferred attribute levels, which can be used to inform product design and positioning decisions. As we build the preference share simulator again for this model, we get results as follows. The first highlighted entry refers to the base design and the rest are competitor profiles.

	colMeans(shares)	Price	Brand	Type.of.Wine	Alcohol	Aging.time.
822	0.22511020	Popular(10€-15€)	Cavit	White wine	5.5 %	4 years
610	0.02773092	Popular(10€-15€)	Cantina Toblino	Sparkling wine	7 %	3 years
674	**0.22545170**	**Popular(10€-15€)**	**Cantina Toblino**	**Sparkling wine**	**12 %**	**3 years**
738	0.15106438	Popular(10€-15€)	Cantina Toblino	Sparkling wine	18 %	3 years
362	0.10127301	Popular(10€-15€)	Ferrari	Sparkling wine	7 %	2 years
618	0.10127301	Popular(10€-15€)	Ferrari	Sparkling wine	7 %	3 years
638	0.16809678	Popular(10€-15€)	Mezzacorona	White wine	7 %	3 years

Cantina Toblino Sparkling wine with 12% alcohol content has the highest preference share of 22.54%, followed by Cavit white wine with 5.5% of alcohol and 4 years of aging time. Since the model controlling for heterogeneity usually predicts a bit larger preference share to niche products, we can assume that there is a non-negligible fraction of consumers who would prefer the products profiles listed above.

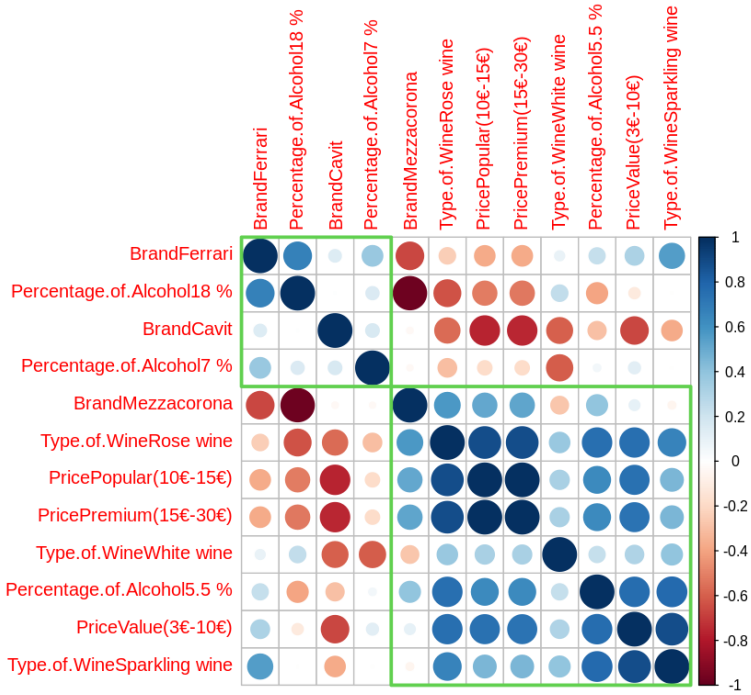


Fig. 17 Correlation plot of the parameters

For further analysis, we did a correlation plot among all the attribute levels as shown in figure 17. The correlation matrix of part worth estimates provides insights into the relationship between different levels of attributes and helps us understand how these levels are valued by consumers. For instance, a high positive correlation among **rose wine** and **price popular** suggests that consumers tend to prefer wines that are both of type red and priced popular. Similarly, positive correlation between the part-worth estimates of **price premium** and **price popular** suggests that people who prefer popular are also likely to prefer **premium**. On the other hand, a low correlation among **Mezzacorona** and **18%** alcohol content tells us the respondents are less likely to prefer **Mezzacorona** brand wine with **18%** alcohol in it.

6.5 Trade off table & Sensitivity Plot

The trade-off table created from the share simulator shows the part-worth or utility value of each attribute level, while the sensitivity plot shows how these values change as one attribute level is changed while keeping all others constant. For baseline and competitor design we reuse the same choices we made previously. The tradeoff table we got is as follows.

	Level	share	increase	Level	share	increase
	<chr>	<dbl>	<dbl>			
Price1	Value(3€-10€)	0.1409499	-0.06179818	Percentage.of.Alcohol1 5.5 %	0.2146006	0.01185251
Price2	Popular(10€-15€)	0.2313434	0.02859531	Percentage.of.Alcohol2 7 %	0.1503057	-0.05244239
Price3	Premium(15€-30€)	0.1851513	-0.01759674	Percentage.of.Alcohol3 12 %	0.2670324	0.06428435
Price4	Luxury(50€-100€)	0.1919294	-0.01081869	Percentage.of.Alcohol4 18 %	0.1507745	-0.05197357
Brand1	Cantina Toblino	0.2166581	0.01391005			
Brand2	Cavit	0.2422183	0.03947019	Aging.time.of.Wine1 1 year	0.2105376	0.00778950
Brand3	Ferrari	0.1880909	-0.01465716	Aging.time.of.Wine2 2 years	0.2241645	0.02141639
Brand4	Mezzacorona	0.2331560	0.03040796	Aging.time.of.Wine3 3 years	0.2156525	0.01290442
Type.of.Wine1	Red wine	0.1763217	-0.02642635	Aging.time.of.Wine4 4 years	0.2213280	0.01857996
Type.of.Wine2	Rose wine	0.1805134	-0.02223466	Aging.time.of.Wine5 5 years	0.2148196	0.01207149
Type.of.Wine3	Sparkling wine	0.1429285	-0.05981955			
Type.of.Wine4	White wine	0.2256721	0.02292402			

The sensitivity plot looks as below.

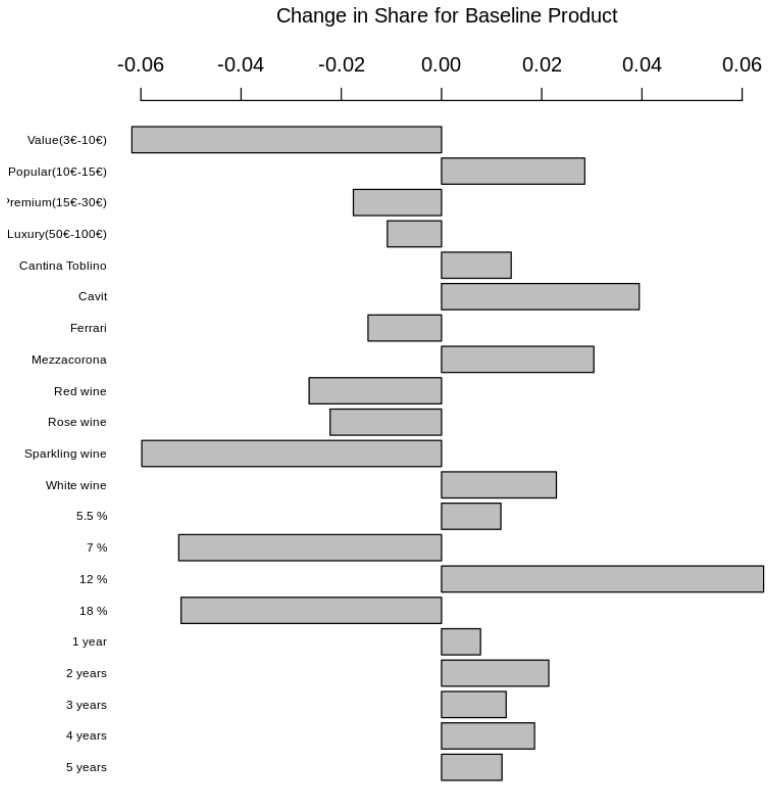


Fig. 18 Sensitivity plot for the mixed MNL with correlated parameters

From the trade-off table and the subsequent sensitivity plot, we observed that the base design constitutes dominant market share for attributes such as "**Price**", "**Brand**" & "**Type.of.Wine**". However, much like our previous analysis for the attribute, "**Percentage.of.Alcohol**", the preference share increases 6.4% if the value is changed from 5.5% to 12%.

7. Future works

While we have taken consumer heterogeneity into account in our analysis, we have not yet taken our uncertainty in the parameter estimations into account. This makes it challenging to ascertain what would result in a (statistically significant) share difference between the two alternatives. Although it is feasible to estimate the prediction intervals for these models in a frequentist framework, a Bayesian framework makes this task simpler. So as an enhancement to the current work, we could implement Hierarchical Bayesian Choice Model.

Even if we considered brand of wine as an attribute in our survey, due to practical limitations we had to limit our questionnaire to choice based design only. User's preference may vary purely based on how they associate a brand with a certain type of perception. If we collected responses regarding the perceptual adjectives people associate every brand of wine with, that could help us come up with positioning maps / perceptual maps. They allow us to address relevant questions such as, what constitutes the class or category of a brand or what are the characteristics of consumers' perception etc.

In real life scenarios, a survey is generally conducted with way more respondents than what we could consider in our survey which in our case resulted in large errors. As an improvement to our current work we could broaden our scope of survey in order to accommodate a larger respondent pool. In that case the respondent sample could be representative of the actual consumer population.

8. Conclusion

The study analyzed the use of conjoint analysis in understanding customer value in the wine market among customers in Italy. This information can be used to create optimal wine products for consumers. The results of the conjoint analysis can also predict how the market will respond to changes in product attributes or price before a production decision is made.

On average, **Brand** and **Price** were found to be the most important factors in wine purchasing decisions, while **Wine Aging Time** was not significant. Using the mixed MNL model, brand and wine aging time were found to be highly significant dependent variables. The highest market share was observed for the price range of **Popular (10€-15€)** and **White wine** with **alcohol content of 12%**.

The preference share simulation showed that the product combination of **Cantina Toblino** with a **popular price** range, **sparkling wine with 12%** alcohol, and **3 years of aging** would be the most popular. The most recommended attributes would be the brand **Cavit**, priced as **Popular (10€-15€)**, **white wine with 12% alcohol**, and aged for **2 years**. Additionally, **Mezzacorona** appeared to be the second most popular brand attribute, showing a little difference in the preference share changing from Cavit.

9. Appendix

https://drive.google.com/drive/folders/1lHDRYqaT2l-X294QYyNd-0vSMMsGnTNQ?usp=share_link

10. References

[1] Barr, Dale J. "Learning Statistical Models through Simulation in R." *Chapter 5 Introducing Linear Mixed-Effects Models*, https://psyteachr.github.io/stat-models-v1/introducing-linear-mixed-effects-models.html.

[2] Chapman, Chris, and Elea McDonnell Feit. *R For Marketing Research and Analytics*. Springer International Publishing, 2019.

[3] Clark, Michael. "Mixed Models with R." *Mixed Models*, https://m-clark.github.io/mixed-models-with-R/random_intercepts.html.

[4] Orme, Bryan K. *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research*. Research Publishers LLC, 2020.

[5] Schwarz, Jason S., et al. *Python for Marketing Research and Analytics*. Springer Nature, 2021.

[6] Walton, Noah, and Haitao Du. "How Do I Avoid Computationally Singular Matrices in R?" *Cross Validated*, 1 Jan. 1964, https://stats.stackexchange.com/questions/250350/how-do-i-avoid-computationally-singular-matrices-in-r.

[7] Seghieri, Chiara & Casini, Leonardo & Torrisi, Francesco. (2007). The wine consumer's behaviour in selected stores of Italian major retailing chains. International Journal of Wine Business Research. 19. 139-151. 10.1108/17511060710758696.

[8] Lockshin, L., & Corsi, A. M. (2012). Consumer behaviour for wine 2.0: A review since 2003 and Future Directions. *Wine Economics and Policy*, *1*(1), 2–23. https://doi.org/10.1016/j.wep.2012.11.003