Choice-based Conjoint Analysis for services on Virtual Servers for CS students.

Laboratory of customer and business analytics

Hamid Omidi, Md Ashraful Alam Hazari | Data science 2022-2023

1. Introduction

A company's leadership must have a clear grasp of the value that its goods or services provide to customers for the business to operate successfully. A better-educated approach can be implemented across the board, from long-term planning to pricing and sales, thanks to this insight. For this reason, we can see the *conjoint analysis* which is so common in the marketing field. To have a more precise application of this technique, this can be used in different fields like pricing, sales and marketing, research, and development, and so on.

Companies use conjoint analysis, a type of statistical analysis, in market research to comprehend how customers value various aspects or qualities of their goods or services. It is founded on the idea that any product can be reduced to a set of characteristics that ultimately affect consumers' perceptions of the worth of a good or service. **Choice-Based Conjoin**t (CBC) Analysis is one of the most common forms of conjoint analysis and reveals how respondents evaluate feature combinations in specific products.

During a choice-based conjoint analysis procedure, each person is given a number of choice sets from which to choose. Each choice alternative is often represented by a collection of attributes. In order to create a function that connects attribute levels to the likelihood of choice, these choice data are then examined using a choice model (often a multinomial logit model. This method's use of choices rather than ratings to gauge preferences is a significant benefit.

In this paper, we will explain the steps for implementing choice-based conjoint analysis on a data set provided by a survey. Therefore, we made a survey questionnaire and after analysing that interpret the results based on this methodology.

This analysis is a part of the course name Laboratory of customer and business analytics as a final project.

2. Problem description

Virtual computing service business is one of the most popular and competitive businesses in the tech world. And under this Virtual computing service business, VPS(Virtual private server) service is widely known for customer demand and popular choice in the market. Under this segment, every virtual computing service company provides competitive and close alternatives for their customers. Here product features are very crucial based on the customer segment. Our objective of this study is to understand the relative importance of different product features and the trade-offs consumers are willing to make between these features. In this study, our target audience is university students, more specifically Computer Science students who would like to use Virtual private servers for their academic purposes, small-scale businesses, and also for start-ups. In general most of the virtual computing service provider companies introduce or segment their product based on the business and they do not have any specific product on this specific segment so in our study we will focus on this specific customer group which is students and we will try to design some of products and analysis those products features based on the customer preferences we collected through the survey. So in our data, we tried to find out which features customers prefer most, how much they are willing to spend, and also within the price range what features they are looking for.
As our study is on a technology-based product which is VPS (Virtual private server), at first we tried to define the product attributes and the levels that we want to study. So, in this study, we took five attributes into our consideration for the analysis and conducting the survey. Then we designed the survey with CBC tools where we created a design matrix that contains all possible combinations of product attributes and levels. This matrix will be used to generate the hypothetical product profiles that participants will see during

the survey. A representative sample of the target population is selected for the study. Participants are presented with a set of product profiles and asked to choose the most preferred product profile. And the data is collected through the choice survey questionnaire. The data collected from the choice survey is analysed using multinomial logit to determine the trade-offs that participants are willing to make between attributes. The results of the analysis are used to determine the relative importance of each attribute and the trade-offs that participants are willing to make between attributes. The results of the CBC analysis are used to determine the relative importance of each attribute and the trade-offs that participants are willing to make between attributes. The results can be used to inform product design, pricing strategies, and marketing campaigns.

3. Survey design

Our research objective for conducting this survey was to know the preferred attributes of VPS(Virtual private server) for Computer Science students or university students who are interested in having VPS service. So, we developed the product profiles by creating a set of product profiles by combining the attributes and levels. We designed the survey questionnaire in a clear and concise way to present the product profiles to participants and collect their choices. We surveyed 50 university students from different dormitories in Trento and among those students, most of them are from computer science backgrounds or related to information technology. We chose these sample participants from a specific background because these participants have some idea about the products and survey we are studying. Each of the participants answered 10 questions with different product profiles where each question has 3 alternatives and from these 3 alternatives participants can choose one alternative which s/he prefers most.

We use the cbc-tools package for making the questions. This CBC tool package provides a set of tools for designing surveys and conducting power analyses for choice-based conjoint survey experiments in R. Each function in this cbcTools package begins with cbc_ and supports a step in the following process for designing and analysing surveys.

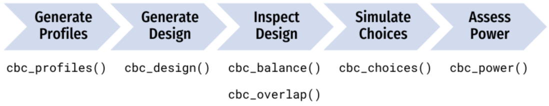
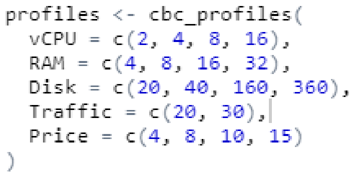


Fig: Process for designing and analysing surveys

We used the remote version of the package from Github using the {remotes} library. For designing the survey, we first defined the attributes and levels for our experiment and then generated all the profiles with each possible combination of those attributes and levels. We took 5 attributes into our consideration which are vCPU, RAM, Disk, Traffic, and Price to generate the profiles.

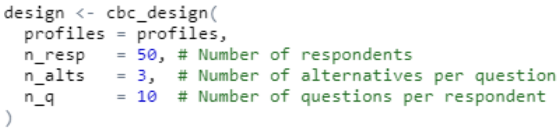


```
profiles <- cbc_profiles(
  vCPU = c(2, 4, 8, 16),
  RAM = c(4, 8, 16, 32),
  Disk = c(20, 40, 160, 360),
  Traffic = c(20, 30),
  Price = c(4, 8, 10, 15)
)
```

Fig: Generating profiles using cbcTool

Here for the vCPU, possible labels are 2,4,8, and 16 cores. For RAM labels are 4,8,16,32 gb, the disk has 20,40,160 360 gb storage, and for Traffic possible labels are 20 and 30.

Once a set of profiles is obtained, a randomised conjoint survey can then be generated using the cbc_design() function.The below figure shows the code for making the ques.



```
design <- cbc_design(
  profiles = profiles,
  n_resp   = 50, # Number of respondents
  n_alts   = 3,  # Number of alternatives per question
  n_q      = 10  # Number of questions per respondent
)
```

Fig: Generating random designs cbcTool

Here for the number of participants, we are generating 10 questions for each participant where the participants will choose 1 alternative from each question.

We tried to conduct our research with randomised machine-generated data but we were not satisfied with the data as it had some patterns which might mislead our study. So, we conducted the survey by using the google form where we used the above-mentioned questions to create our survey dataset.

4. Exploratory Data Analysis

After collecting the answers we check the frequency based on some demographic data. Also, we will use this information as the respondent-level variables for checking the relations between them and the references. Below you can see the frequency of the respondents based on their education degree and the sex.

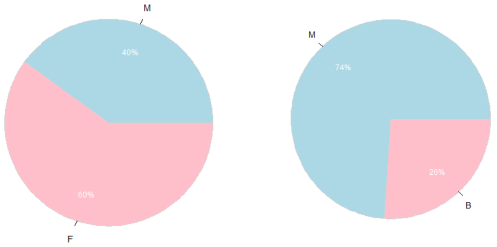
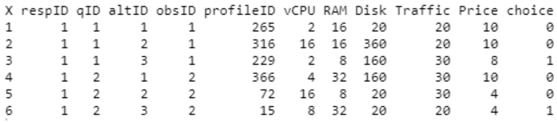


Fig: Pie chart for participant's demographic data

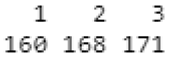
As we mentioned before fifty Computer science students have participated in this survey. For the first step, some exploratory analysis has been done to make clear insights into the data set that we collect. In the following figure, the first six rows of the data set can be seen.



```
X respID qID altID obsID profileID vCPU RAM Disk Traffic Price choice
1      1   1     1     1       265    2  16   20      20    10      0
2      1   1     2     1       316   16  16  360      20    10      0
3      1   1     3     1       229    2   8  160      30     8      1
4      1   2     1     2       366    4  32  160      30    10      0
5      1   2     2     2        72   16   8   20      30     4      0
6      1   2     3     2        15    8  32   20      20     4      1
```

Fig: Dataset

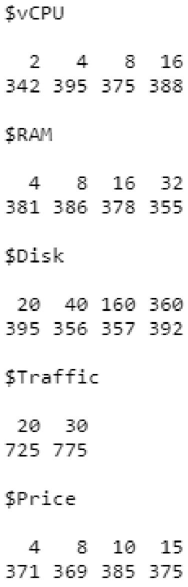
As it can be guessed the columns are based on their name, and the last column shows the choices of the participants. At first, we checked the frequency of the three different alternatives to be sure that the choices are not based on the positions of the alternatives.



```
  1   2   3 
160 168 171 
```

Fig: Frequency of the three different alternatives

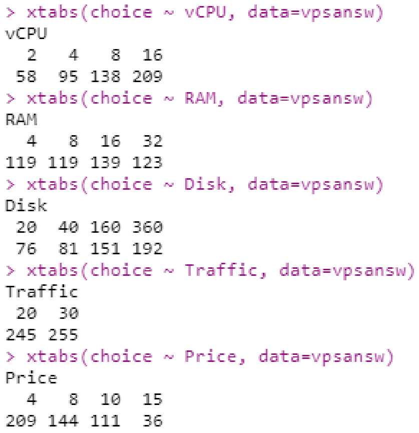
As it can be seen the numbers are somehow balanced and there are no intentional preferences in the choosing process between different alternatives. For checking if we have a sampling bias we checked the frequency distribution of the attributes.



```
$vCPU

  2   4   8  16 
342 395 375 388 

$RAM

  4   8  16  32 
381 386 378 355 

$Disk

 20  40 160 360 
395 356 357 392 

$Traffic

 20  30 
725 775 

$Price

  4   8  10  15 
371 369 385 375 
```

Fig: Frequency distribution of the attributes

So it seems that the attribute levels are distributed kind of balanced among the questions.
In the following figures, we can see some summary of the individual and pairwise counts of each level of each attribute across all choice questions. These figures help us to spot some intuition about the tendency of the participants and their choices.



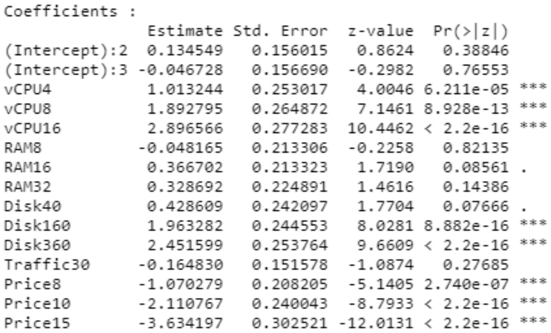
```
> xtabs(choice ~ vCPU, data=vpsansw)
vCPU
  2   4   8  16 
 58  95 138 209 
> xtabs(choice ~ RAM, data=vpsansw)
RAM
  4   8  16  32 
119 119 139 123 
> xtabs(choice ~ Disk, data=vpsansw)
Disk
 20  40 160 360 
 76  81 151 192 
> xtabs(choice ~ Traffic, data=vpsansw)
Traffic
 20  30 
245 255 
> xtabs(choice ~ Price, data=vpsansw)
Price
  4   8  10  15 
209 144 111  36 
```

Fig: one-way frequency relation with participant's preference

For example, we can see that there is no big difference between the RAM size for participants as the choices are somehow distributed and balanced between them, while, about more than 41 percent of the participants prefer the low price. Moreover, we can see some preferences on the number of virtual CPUs and the disk space. Therefore, as a quick result maybe the company would prefer to offer low RAM in the student packages as it reduces the costs of the services because we can say that the RAM does not have that much impact on the choices and also RAM is always among the pricy hardware.

5. Models

After changing the data format with **dfidx** we tried to make the models. For the first try multinomial logistic regression model was while considering the alternative variables. In the following figure, you can see the outcome of the first model.



```
Coefficients :
                 Estimate Std. Error  z-value  Pr(>|z|)
(Intercept):2  0.134549   0.156015   0.8624   0.38846
(Intercept):3 -0.046728   0.156690  -0.2982   0.76553
vCPU4          1.013244   0.253017   4.0046 6.211e-05 ***
vCPU8          1.892795   0.264872   7.1461 8.928e-13 ***
vCPU16         2.896566   0.277283  10.4462 < 2.2e-16 ***
RAM8          -0.048165   0.213306  -0.2258   0.82135
RAM16          0.366702   0.213323   1.7190   0.08561 .
RAM32          0.328692   0.224891   1.4616   0.14386
Disk40         0.428609   0.242097   1.7704   0.07666 .
Disk160        1.963282   0.244553   8.0281 8.882e-16 ***
Disk360        2.451599   0.253764   9.6609 < 2.2e-16 ***
Traffic30     -0.164830   0.151578  -1.0874   0.27685
Price8        -1.070279   0.208205  -5.1405 2.740e-07 ***
Price10       -2.110767   0.240043  -8.7933 < 2.2e-16 ***
Price15       -3.634197   0.302521 -12.0131 < 2.2e-16 ***
```

Fig: Model Summary

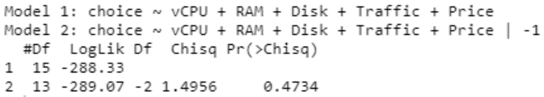
The two first rows of the table (estimated intercepts) represent the alternative-specific constants that show preferences for the positions of the alternatives in each question. As we expect that the estimations are not significantly different from zero so we can ignore them. That is, as we mentioned before there were no preferences on the position of the alternatives on the questions.

we can see the estimated average part worths for each level in the above table. The point here is that they have to be interpreted with respect to the reference level of each attribute [3].

At the first glance, we can see that the p values of the RAM levels and the Traffic levels are not significantly different from zero. For the interpretation of the vCPU, we can say that on average respondents are more attracted to 16-core virtual CPUs in comparison with CPUs with 2, and as the estimation score is so high it means that they really prefer that. It is true for CPUs with 8 and 4 cores but the strengths of them are different.

For the price variable, we can see that the estimated numbers are negative, which means that the reference level (price4) on average is more attractive for the participants. The other important thing here is that among the Disk levels, the 40 level does not have a significant difference from zero with respect to its p-value. As we know, we do not have a huge number of participants so we can not expect great precision with respect to the standard error (The more participants given the attributes the smaller the standard error).

For testing, if removing the intercepts in the previous model would have any effect on the model fitting we run a model without engaging the alternatives (equal intercepts to zero) and then we will test them with a **likelihood ratio test**. The Below figure shows the test.



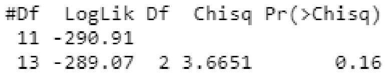
```
Model 1: choice ~ vCPU + RAM + Disk + Traffic + Price
Model 2: choice ~ vCPU + RAM + Disk + Traffic + Price | -1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  15 -288.33
2  13 -289.07 -2 1.4956     0.4734
```

Fig: Likelihood ratio test for fitting choice model

From the output, we can see that the Chi-Squared test statistic is 1.4956 and the corresponding p-value is 0.4734. Since this p-value is not less than .05, we will fail to reject the null hypothesis.

This means the complete model and the nested model fit the data equally well. Thus, we should use the nested model because the additional predictor variables in the full model don't offer a significant improvement in fit.

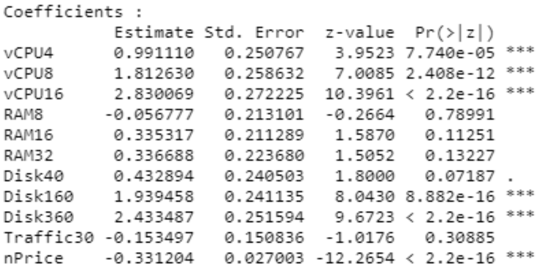
for having more interpretable concepts we can use the willingness to pay. For this reason, we should have the price attribute as a quantitative variable. At first, we will change the model with the quantitative price and see the impact of that on the result.



```
#Df  LogLik Df  Chisq Pr(>Chisq)
 11 -290.91
 13 -289.07  2 3.6651       0.16
```

Fig: Likelihood ratio test for new model

As we can see, again this change did not have any big changes to the model with respect to improvement and both models are equally well.



```
Coefficients :
            Estimate Std. Error  z-value  Pr(>|z|)
vCPU4       0.991110   0.250767   3.9523 7.740e-05 ***
vCPU8       1.812630   0.258632   7.0085 2.408e-12 ***
vCPU16      2.830069   0.272225  10.3961 < 2.2e-16 ***
RAM8       -0.056777   0.213101  -0.2664   0.78991
RAM16       0.335317   0.211289   1.5870   0.11251
RAM32       0.336688   0.223680   1.5052   0.13227
Disk40      0.432894   0.240503   1.8000   0.07187 .
Disk160     1.939458   0.241135   8.0430 8.882e-16 ***
Disk360     2.433487   0.251594   9.6723 < 2.2e-16 ***
Traffic30 -0.153497   0.150836  -1.0176   0.30885
nPrice     -0.331204   0.027003 -12.2654 < 2.2e-16 ***
```

Fig: Summary of New Model

For calculating the Willingness to pay we can use this formula:

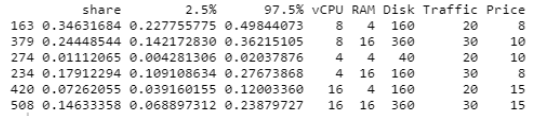
$$WTP = \frac{\beta}{\beta_{\text{price}}}$$

As we saw that maybe the RAM and Traffic levels do not have that much impact on the model, we will check the WTP of some other alternatives. WTP for the vCPU with 16 cores will be -8.544783 which means on average, the individual is willing to pay at most 8.5 euros to have 16 core virtual instead of 2 core virtual CPU. As the maximum price is 15 this preference can cover more than half of the income on the maximum price. WTP for the CPU with 8 cores and 4 cores in comparison with 2 cores would be around 5.47 and 3 euros respectively. In the below table WTP for all variables calculated:

vCPU4 : -2.992443 vCPU8 : -5.472845 vCPU16 : -8.544783 RAM8 : 0.1714251 RAM16 : -1.012419	RAM32 : -1.016558 Disk40 : -1.307029 Disk160 : -5.855775 Disk360 : -7.347389 Traffic30 : 0.4634513

Fig: WTP for all variables

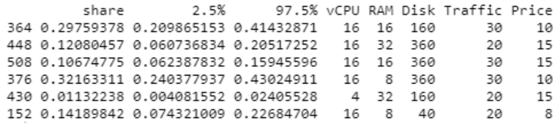
As we know another useful approach to assess the role of product attributes consists of using the model to obtain preference share predictions. Based on the mentioned things and based on our knowledge of this industry and assuming that the students are our customers we try to keep the situation kind of economical by simultaneously considering the information we have gained so far. based on that we made 3 imaginary different services and checked their simulated preference share in comparison with real options that we have in the real world. Our candidate configuration would be 8 cores of CPU, 4 gigs of ram, 160 of disk space, 20 gigs of traffic, and 8 euro price per month (profile number 163). In the following figure, we can see the preference share with the bootstrap confidence interval.



```
         share        2.5%      97.5% vCPU RAM Disk Traffic Price
163 0.34631684 0.227755775 0.49844073    8   4  160      20     8
379 0.24448544 0.142172830 0.36215105    8  16  360      30    10
274 0.01112065 0.004281306 0.02037876    4   4   40      20    10
234 0.17912294 0.109108634 0.27673868    4  16  160      30     8
420 0.07262055 0.039160155 0.12003360   16   4  160      20    15
508 0.14633358 0.068897312 0.23879727   16  16  360      30    15
```

Fig: Preference of Share

Based on the table we can expect that about 35% of the time students will prefer our configuration in comparison with the other configuration in the table. The second service would be more focused on being a high-performance one (profile number 364). The below figure shows the result.



```
         share        2.5%      97.5% vCPU RAM Disk Traffic Price
364 0.29759378 0.209865153 0.41432871   16  16  160      30    10
448 0.12080457 0.060736834 0.20517252   16  32  360      20    15
508 0.10674775 0.062387832 0.15945596   16  16  360      30    15
376 0.32163311 0.240377937 0.43024911   16   8  360      30    10
430 0.01132238 0.004081552 0.02405528    4  32  160      20    15
152 0.14189842 0.074321009 0.22684704   16   8   40      20     8
```

Fig: Preference of Share

This time the preference share would be 29% which sounds good enough for the company to sell this service. To make a clearer decision process we also make the sensitivity chart as you can see in the below figure.

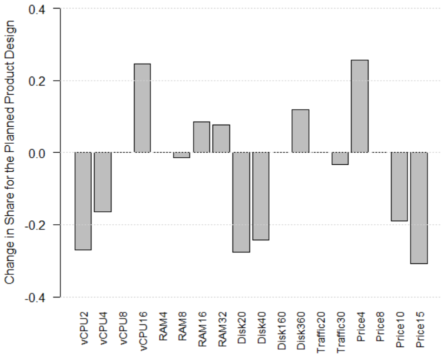


Fig: Sensitivity Chart for VPS design with 8gb vCPU, 4gb RAM, 160 Disk space, 20 Traffic for a price of 8 Euro

So for this pic, we can say changing the number of the virtual CPU from 8 to 16 can increase the preference share by more than 0.2. Based on that, if the provider uses the dynamic virtual CPU allocation system, they can shift to the 16 virtual CPUs by a little bit increasing the price as usually students do not use all the cores in a time and the provider can allocate them exactly when they need it. That is, in some sense, we can share some vCPU. But the important thing here is that we should also inform the customers of this system and see if this change can have an impact on the preference share as using shared resources always has a negative impact when someone wants to pay. Also, changing the disk volume from 160 to 360 will increase the preference share by roughly 0.17. Therefore, the provider can make its decision based on the different costs between 160 to 360.

The below figure shows the sensitivity chart for the second preference share table associated with the high-performance VPSs.

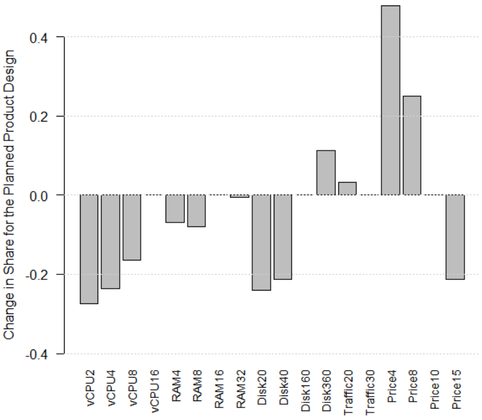
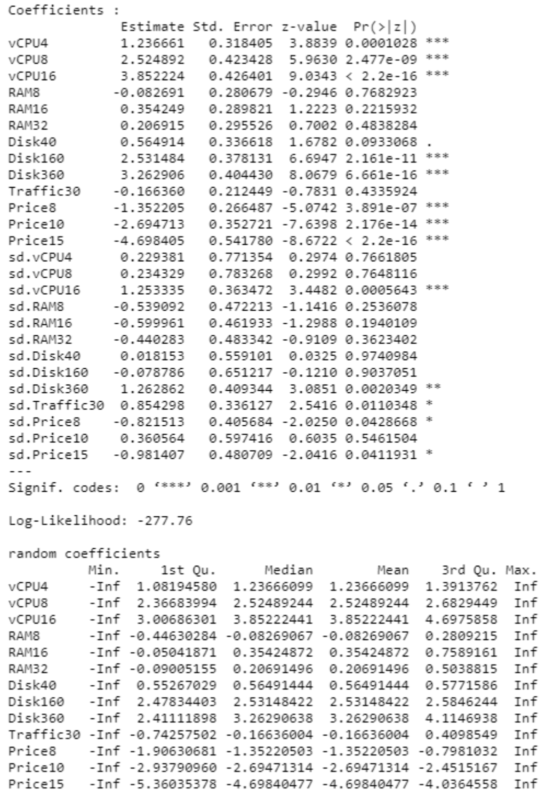


Fig: Sensitivity Chart for VPS design with 16gb vCPU, 16gb RAM, 160 Disk space, 30 Traffic for a price of 10 Euro

For the second proposal configuration as you can see most of the changes will decrease the average preference share for students. While changing the price to 4 euro will overshoot the preference share, it is impossible to keep the price on that level as the resources that have been used in this configuration is so pricey.

For considering the customers (students) effect and checking the respondent coefficient we should use the mixed multinomial logistic regression. Usually, we do that because estimated consumer-level coefficients can have higher goodness of fit and provide more accurate preference share predictions than models with only fixed effects [3]. As we should have a pr assumption for the coefficient distribution across the population we assume that all the coefficients are normally distributed for this analysis. The following table shows the outcome of the modelling. We did not consider the correlation between the random parameters [correlation=False].



```
Coefficients :
               Estimate Std. Error z-value  Pr(>|z|)    
vCPU4          1.236661   0.318405  3.8839 0.0001028 ***
vCPU8          2.524892   0.423428  5.9630 2.477e-09 ***
vCPU16         3.852224   0.426401  9.0343 < 2.2e-16 ***
RAM8          -0.082691   0.280679 -0.2946 0.7682923    
RAM16          0.354249   0.289821  1.2223 0.2215932    
RAM32          0.206915   0.295526  0.7002 0.4838284    
Disk40         0.564914   0.336618  1.6782 0.0933068 .  
Disk160        2.531484   0.378131  6.6947 2.161e-11 ***
Disk360        3.262906   0.404430  8.0679 6.661e-16 ***
Traffic30     -0.166360   0.212449 -0.7831 0.4335924    
Price8        -1.352205   0.266487 -5.0742 3.891e-07 ***
Price10       -2.694713   0.352721 -7.6398 2.176e-14 ***
Price15       -4.698405   0.541780 -8.6722 < 2.2e-16 ***
sd.vCPU4       0.229381   0.771354  0.2974 0.7661805    
sd.vCPU8       0.234329   0.783268  0.2992 0.7648116    
sd.vCPU16      1.253335   0.363472  3.4482 0.0005643 ***
sd.RAM8       -0.539092   0.472213 -1.1416 0.2536078    
sd.RAM16      -0.599961   0.461933 -1.2988 0.1940109    
sd.RAM32      -0.440283   0.483342 -0.9109 0.3623402    
sd.Disk40      0.018153   0.559101  0.0325 0.9740984    
sd.Disk160    -0.078786   0.651217 -0.1210 0.9037051    
sd.Disk360     1.262862   0.409344  3.0851 0.0020349 ** 
sd.Traffic30  0.854298   0.336127  2.5416 0.0110348 *  
sd.Price8     -0.821513   0.405684 -2.0250 0.0428668 *  
sd.Price10     0.360564   0.597416  0.6035 0.5461504    
sd.Price15    -0.981407   0.480709 -2.0416 0.0411931 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -277.76

random coefficients
          Min.     1st Qu.      Median        Mean    3rd Qu. Max.
vCPU4     -Inf  1.08194580  1.23666099  1.23666099  1.3913762  Inf
vCPU8     -Inf  2.36683994  2.52489244  2.52489244  2.6829449  Inf
vCPU16    -Inf  3.00686301  3.85222441  3.85222441  4.6975858  Inf
RAM8      -Inf -0.44630284 -0.08269067 -0.08269067  0.2809215  Inf
RAM16     -Inf -0.05041871  0.35424872  0.35424872  0.7589161  Inf
RAM32     -Inf -0.09005155  0.20691496  0.20691496  0.5038815  Inf
Disk40    -Inf  0.55267029  0.56491444  0.56491444  0.5771586  Inf
Disk160   -Inf  2.47834403  2.53148422  2.53148422  2.5846244  Inf
Disk360   -Inf  2.41111898  3.26290638  3.26290638  4.1146938  Inf
Traffic30 -Inf -0.74257502 -0.16636004 -0.16636004  0.4098549  Inf
Price8    -Inf -1.90630681 -1.35220503 -1.35220503 -0.7981032  Inf
Price10   -Inf -2.93790960 -2.69471314 -2.69471314 -2.4515167  Inf
Price15   -Inf -5.36035378 -4.69840477 -4.69840477 -4.0364558  Inf
```

Fig:

With a comparison between the estimated standard deviation for the part-worth of RAM8 (-0.539) against the average estimate of that parameter (-0.082). We can see this situation for other RAM options too. So it seems that we cannot turn a blind eye to the fact that although students prefer the RAM8 but there are some that prefer RAM4. Therefore, if we check the random coefficient table we can see that there is a sign change in the 3rd quartile of the distribution. Moreover, we can see this charge on the RAM16 and RAM32.

If we draw the RAM8 distribution we can see this situation more clearly. The below figure shows the mentioned explanation.

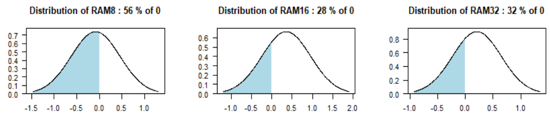


Fig: Visual summary of the distribution of random effects of different RAM

As it can be seen we should also consider RAM4 despite the willingness of the other alternatives. In fact, we should consider it to keep the price lower as the RAM is one of the most expensive hardware so we can say with consideration that as part of consumers, we can at the same time keep the expenses low and also keep that type of students that prefer the RAM4. For the price we can say that there is no significant variety in the students' choice so for example the student that prefers 4 euro fee payment with comparison to euro, almost all of them prefer 4 euro.

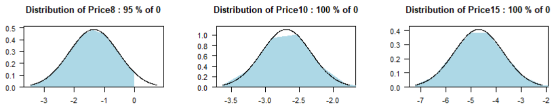


Fig: Visual summary of the distribution of random effects of different prices

As the above figure shows we can see the same situation for two other options: the price. As we can remember the Traffic variable was not that important as the coefficient was not that much and the p-value was big but we can also check the variability of choice between the two options.

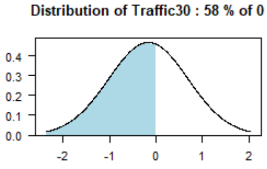
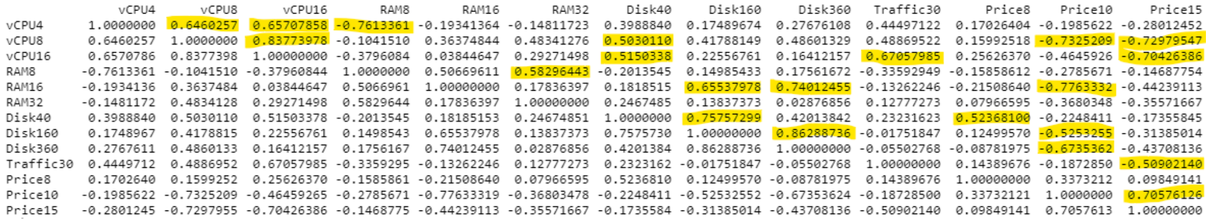


Fig: distribution of specific random effects for Trafic 30

Now we also check the correlation between the random effects and see if we consider that situation and what will happen to the results. In the following figure the association between the random coefficients as we see the covariance matrix.

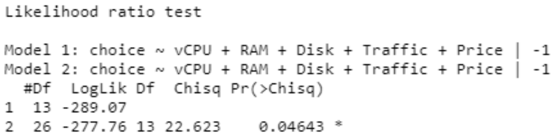
The most powerful associations are indicated by the yellow highlight. We will restrict the correlation to only random parameters with significant associations. we can see some association in each level of the attributes with other ones.



```
                vCPU4      vCPU8      vCPU16       RAM8       RAM16       RAM32      Disk40     Disk160     Disk360   Traffic30      Price8     Price10     Price15
vCPU4      1.0000000  0.6460257  0.65707858 -0.7613361 -0.19341364 -0.14811723  0.3988840  0.17489674  0.27676108  0.44497122  0.17026404 -0.1985622 -0.28012452
vCPU8      0.6460257  1.0000000  0.83773978 -0.1041510  0.36374844  0.48341276  0.5030110  0.41788149  0.48601329  0.48869522  0.15992518 -0.7325209 -0.72979547
vCPU16     0.6570786  0.8377398  1.00000000 -0.3796084  0.03844647  0.29271498  0.5150338  0.22556761  0.16412157  0.67057985  0.25626370 -0.4645926 -0.70426386
RAM8      -0.7613361 -0.1041510 -0.37960844  1.0000000  0.50669611  0.58296443 -0.2013545  0.14985433  0.17561672 -0.33592949 -0.15858612 -0.2785671 -0.14687754
RAM16     -0.1934136  0.3637484  0.03844647  0.5066961  1.00000000  0.17836397  0.1818515  0.65537978  0.74012455 -0.13262246 -0.21508640 -0.7763332 -0.44239113
RAM32     -0.1481172  0.4834128  0.29271498  0.5829644  0.17836397  1.00000000  0.2467485  0.13837373  0.02876856  0.12777273  0.07966595 -0.3680348 -0.35571667
Disk40     0.3988840  0.5030110  0.51503378 -0.2013545  0.18185153  0.24674851  1.0000000  0.75757299  0.42013842  0.23231623  0.52368100 -0.2248411 -0.17355845
Disk160    0.1748967  0.4178815  0.22556761  0.1498543  0.65537978  0.13837373  0.7575730  1.00000000  0.86288736 -0.01751847  0.12499570 -0.5253255 -0.31385014
Disk360    0.2767611  0.4860133  0.16412157  0.1756167  0.74012455  0.02876856  0.4201384  0.86288736  1.00000000 -0.05502768 -0.08781975 -0.6735362 -0.43708136
Traffic30  0.4449712  0.4886952  0.67057985 -0.3359295 -0.13262246  0.12777273  0.2323162 -0.01751847 -0.05502768  1.00000000  0.14389676 -0.1872850 -0.50902140
Price8     0.1702640  0.1599252  0.25626370 -0.1585861 -0.21508640  0.07966595  0.5236810  0.12499570 -0.08781975  0.14389676  1.00000000  0.3373212  0.09849141
Price10   -0.1985622 -0.7325209 -0.46459265 -0.2785671 -0.77633319 -0.36803478 -0.2248411 -0.52532552 -0.67353624 -0.18728500  0.33732121  1.0000000  0.70576126
Price15   -0.2801245 -0.7297955 -0.70426386 -0.1468775 -0.44239113 -0.35571667 -0.1735584 -0.31385014 -0.43708136 -0.50902140  0.09849141  0.7057613  1.00000000
```

Fig: Correlation matrix of the random effects

We will make a mixed model with the most associated ones and will compare them with the previous ones to see the result. The below figure shows the result of the ML ratio test. The first result is the outcome of a comparison between fixed effects models with uncorrelated random model effects.



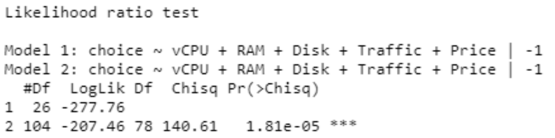
```
Likelihood ratio test

Model 1: choice ~ vCPU + RAM + Disk + Traffic + Price | -1
Model 2: choice ~ vCPU + RAM + Disk + Traffic + Price | -1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  13 -289.07
2  26 -277.76 13 22.623    0.04643 *
```

Fig: Likelihood ratio test

As we can see from the output the Chi-Squared test statistic is 22.623 and the corresponding p-value is 0.04643. Since this p-value is less than .05, we won't fail to reject the null hypothesis.

This means that the complete model and the random effect model do not fit the data equally well. And we can say that the uncorrelated random effect fits the model better.

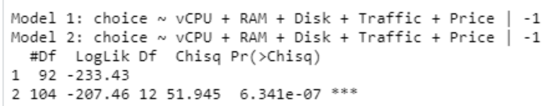
The next figure shows the maximum likelihood ratio test for Uncorrelated random effects and all correlated random effects (without selecting the most associated ones).



```
Likelihood ratio test

Model 1: choice ~ vCPU + RAM + Disk + Traffic + Price | -1
Model 2: choice ~ vCPU + RAM + Disk + Traffic + Price | -1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  26 -277.76
2 104 -207.46 78 140.61   1.81e-05 ***
```

Fig: Likelihood ratio test

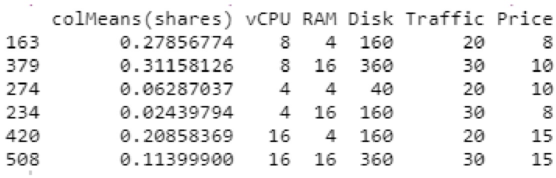
As we also expect that the new model fits the data better as we have a very small p. And the last figure shows the same test between partially correlated random effects and all correlated random effects.



```
Model 1: choice ~ vCPU + RAM + Disk + Traffic + Price | -1
Model 2: choice ~ vCPU + RAM + Disk + Traffic + Price | -1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  92 -233.43
2 104 -207.46 12 51.945  6.341e-07 ***
```

Fig: Likelihood ratio test

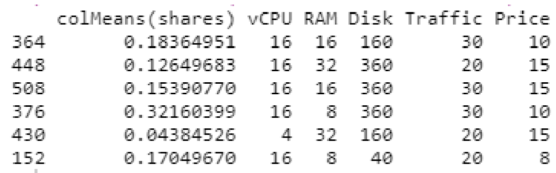
And as we expected again we can see that the smaller model can fit the data better. So now we use the new model with all correlations between the random effects to calculate the preference share. The below figure shows the result for the first configurations (profile number 163).



```
    colMeans(shares) vCPU RAM Disk Traffic Price
163       0.27856774    8   4  160      20     8
379       0.31158126    8  16  360      30    10
274       0.06287037    4   4   40      20    10
234       0.02439794    4  16  160      30     8
420       0.20858369   16   4  160      20    15
508       0.11399900   16  16  360      30    15
```

Fig: Preference share prediction

We can see about a 0.07 change in the preference share of our first configuration and the preference share shifted a little bit. Like in the second row, we can see more preference shares in comparison to the previous preference share table with the same configurations. The next figure shows the second configuration related to the performance.



```
    colMeans(shares) vCPU RAM Disk Traffic Price
364       0.18364951   16  16  160      30    10
448       0.12649683   16  32  360      20    15
508       0.15390770   16  16  360      30    15
376       0.32160399   16   8  360      30    10
430       0.04384526    4  32  160      20    15
152       0.17049670   16   8   40      20     8
```

Fig: Preference share prediction

Again we have changed as you can see in the figure. Technically we will lose 0.1 in the new model and as it seems it is more realistic we should keep going with the new one.

At the end we also check the respondent level variables sex and education on the important level. For assessing if consumer behaviour heterogeneity can be explained by their individual characteristics we can study the relationship between the individual part worth and the individual-level variables. So, we chose the most important variables as mentioned before. Below chart shows the relationship between sex and price.

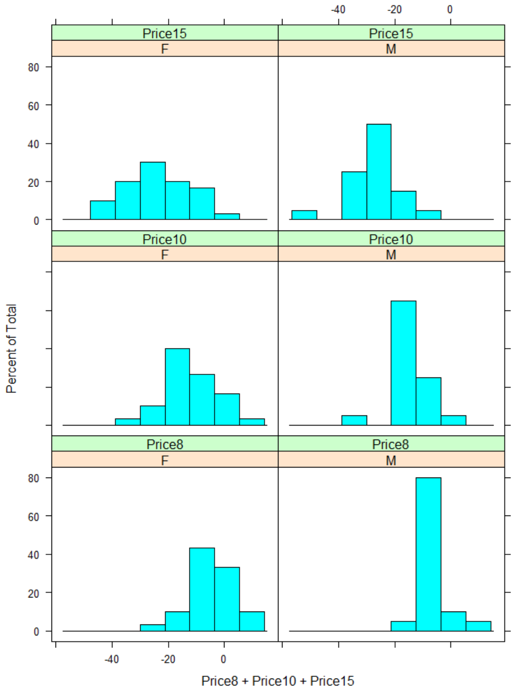
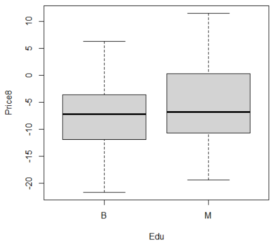
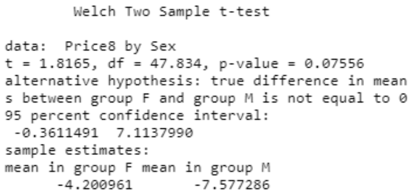


Fig: Relationship between sex and price of the participants

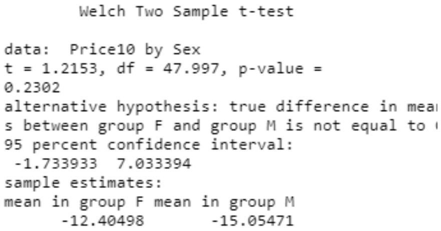
We can see a little bit of difference between them. So we tried to draw it by boxplot and do the t-test to see if it is clear. For 8 euro price:





```
          Welch Two Sample t-test

data:  Price8 by Sex
t = 1.8165, df = 47.834, p-value = 0.07556
alternative hypothesis: true difference in mean
s between group F and group M is not equal to 0
95 percent confidence interval:
 -0.3611491  7.1137990
sample estimates:
mean in group F mean in group M
      -4.200961       -7.577286
```

With drawing the box plot and t-test we can see there is no significant difference between them. The t-test for two other prices can be seen.



```
          Welch Two Sample t-test

data:  Price10 by Sex
t = 1.2153, df = 47.997, p-value =
0.2302
alternative hypothesis: true difference in mea
s between group F and group M is not equal to
95 percent confidence interval:
 -1.733933  7.033394
sample estimates:
mean in group F mean in group M
      -12.40498       -15.05471
```

Same as the previous one there is no significant difference. So we check the price and education.

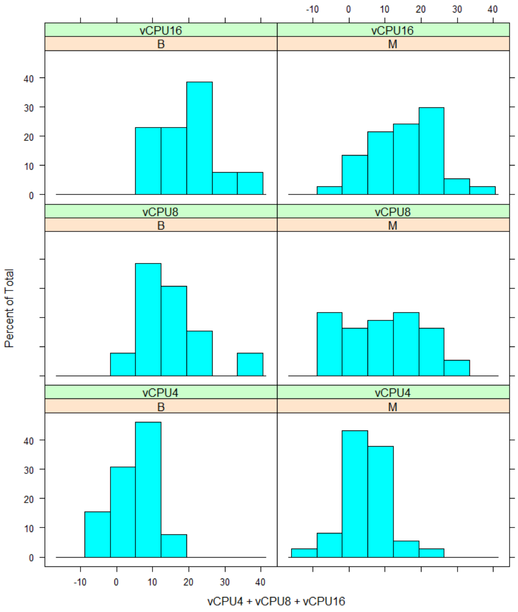
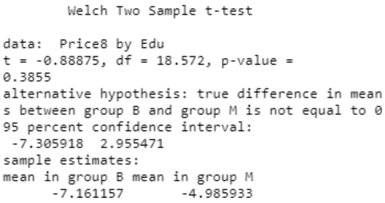


Fig: Relationship between sex and price of the participants

We will check the price 8 as it seems they may have a significant difference.



```
          Welch Two Sample t-test

data:  Price8 by Edu
t = -0.88875, df = 18.572, p-value =
0.3855
alternative hypothesis: true difference in mean
s between group B and group M is not equal to 0
95 percent confidence interval:
 -7.305918  2.955471
sample estimates:
mean in group B mean in group M
      -7.161157       -4.985933
```

Fig:

The test shows that it is not right and we do not have a significant difference as the p-value is not equal to zero. By doing these tests on all the important variables we saw that there is not that much relation between these variables at the respondent level.

7. Discussion and Conclusion

In the different steps of the analysis we had different results and based on that different results and conclusions.

The first thing we should mention is that we can not turn a blind eye to the sample size here. As we did not have sufficient resources (time, community and …) of course we can not expect the high precision as the sample size can have an impact on that. But we try to consider the representativeness of the analysis by doing the survey in a really representative community of the research field.

Of course, the results can be analysed more precisely by having more information about the cost of the services and the resources in the industry, despite that we tried to consider the general situations about that to make our interpretation close to the real world. for example, we do not know how much one unit of RAM would cost for the business owner as it really depends on the brand and also the architecture of the RAM and so on, but we consider general facts that the CPUs are the most expensive, or the RAMs are always more expensive than Disks with the same capacity.

We saw that the alternative positions did not have the impact on the choice process of the individuals as we expected. So we reduce the model to the nested one to make it more optimised.

One thing to mention is that, as we saw that the disk space is the important variable between the students, maybe we can offer some alternatives with a lower gap between the volumes to control the sensitivity of the changes. For example, we can put something like 80 gig volume options to observe the preference in the choices (It seems that distance between 40 and 160 is huge with respect to the disk volume).

Based on the analysis the RAM and the Traffic levels are the less important part of the models. So businesses should stay focused on the other variables and mostly on the virtual CPUs as they are flexible for them.

Despite some good results of the two first sensitivity charts for the business's newly selected services, we saw that after engaging the respondent level effect, we have a big change in the preference share. In this case, the company should start to change the configuration a little bit to improve the real preference share. Again because in the technical context, we have more flexibility on vCPUs we can consider that as a good variable to change. Another good suggestion for changing this configuration is that we can work on the Disk variable as it has a big impact and at the same time the cost of the development of that is really lower than the CPUs and RAMs.

We saw that we should also consider RAM4 despite the willingness of the other options. In fact, we should consider it to keep the price lower as the RAM is one of the most expensive hardware so we can say with consideration that (for some services) part of consumers we can at the same time keep the expenses low and also keep that type of students that prefer the RAM4. At the same time, because the RAM was not that impactful, we can put lower attention to this part.

After extracting the mixed-level model, we try to find the relationship between the respondent variables that we collected (Sex, and education), and the tests and the figures showed that at least these variables are not the related one. So as we sat at the mixed model to model the data better we can guess there are some parameters in the respondent level that can impact the voice of them but they are not the ones that we collect. So heterogeneity for preference of the important variables can not be explained by the respondent level

8. References:

[1]Stobierski, T.(2020). *What is Conjoint Analysis, And how can it be used?*

https://online.hbs.edu/blog/post/what-is-conjoint-analysis

[2]Helveston, J.(2022). *cbcTools. https://github.com/jhelvy/cbcTools*

[3] Rao, V. R. (2014). Applied Conjoint Analysis. Springer Publishing.

[4] professor Diego Giuliani’s presentation, University of Trento, LCBA course, 2022

[5] McFadden, Daniel, and Kenneth Train. 2000. “Mixed MNL Models for Discrete Response.” Journal of Applied Econometrics 15 (5): 447–70.

Train, Kenneth E. 2009. Discrete Choice Methods with Simulation. 2nd ed. Cambridge University Press.