Phase 6 Report: Data Analysis

1. **Data Cleaning & Preparation:**

We needed to do additional modifications to the data to make it even more suitable for the analytical methods we planned to use.

* We dropped 2 things from the table, careless-ness responses and the columns we wouldn’t use or don’t have a use for in the analysis itself

Such as: ['ID', 'point of view','graphics card model','how powerful from 1\_to\_5.1', 'main device', 'is it intel ', 'sum of the time spent', 'own DeskTop', 'own Laptop', 'Data scientist?', 'would he chage his device']

* Cleaned text strings
* Created 2 columns that measure the percentage or the ratio of laptop/pc usage
* Replaced null values in the price column with the average price of all devices
* Replaced null values in the “how powerful is your device” column with average scale of power
* Anyone having +5 devices will be considered as only having 5, for easier scalability.
* Anyone above the age of 21 will be considered as 21, for easier scalability.
* Created 2 columns for visualization’s sake, the 2 columns being a ratio for which device you want to rebuy/replace.

1. **Descriptive Statistics:**

We first took a descriptive look at the data. Visualizing the data that we have using charts and histograms we were able to obtain some insights and knowledge about the data we have. First we made histograms of all important columns that we have:

**General Information about demographic:**

**A picture containing screenshot, plot, diagram, line

Description automatically generatedA picture containing screenshot, text, plot, diagram

Description automatically generated**

**A picture containing screenshot, text, diagram, square

Description automatically generatedA picture containing screenshot, text, diagram, plot

Description automatically generated**

**A picture containing screenshot, diagram, text, plot

Description automatically generatedA picture containing screenshot, text, plot, diagram

Description automatically generated**

**Importance :**

**A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, number, plot

Description automatically generated**

**A picture containing text, screenshot, number, plot

Description automatically generatedA picture containing text, screenshot, plot, number

Description automatically generated**

**A picture containing text, screenshot, plot, diagram

Description automatically generated**

**A picture containing text, screenshot, number, plot

Description automatically generated**

**A picture containing text, screenshot, diagram, plot

Description automatically generated**

**A picture containing text, screenshot, diagram, plot

Description automatically generated**

**A picture containing text, screenshot, rectangle, number

Description automatically generated**

**A picture containing screenshot, text, plot, diagram

Description automatically generatedUsage hours:**

**A picture containing screenshot, diagram, plot, line

Description automatically generated**

**A picture containing text, screenshot, plot, diagram

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generated**

**A picture containing screenshot, diagram, plot, text

Description automatically generatedA picture containing text, screenshot, plot, diagram

Description automatically generated**

**A picture containing screenshot, plot, line, rectangle

Description automatically generatedDevice related information:**

**A picture containing text, screenshot, diagram, plot

Description automatically generated**

**A picture containing text, screenshot, plot, line

Description automatically generated**

**A picture containing screenshot, text, diagram, plot

Description automatically generatedA picture containing screenshot, text, diagram, plot

Description automatically generated**

**A picture containing text, screenshot, plot, number

Description automatically generatedA picture containing text, screenshot, plot, number

Description automatically generated**

We tried to test the correlation between a few attributes and these are examples of what we found:

* Correlation coefficient between people who would rebuy a laptop and hours spent gaming was -0.4129, which means the more hours people spend on gaming, the more they are likely to get a PC instead.
* **A picture containing screenshot, diagram, circle, text

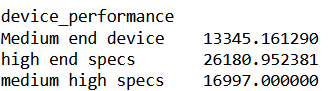
  Description automatically generated**Looking for more information about device performance:

We see that the sampling is not representative enough for low end devices, which is expected given that the sample is drawn from Computer science majors and professionals, we focus on analyzing the differences between the other 3 segments:

**A pie chart with different colors

Description automatically generated with low confidence**

The target audience seems to be in favor of medium high devices.

And when we look at the average prices of each device specs : 

We see that device price is consistent with device performance.

A picture containing screenshot, diagram, rectangle, line

Description automatically generatedAlso:

The people who would prefer keeping or rebuying laptops have a higher average income than those who would prefer to buy PCs, yet the price range is shorter which stands out well with general laptop prices.

**3.Inferential statistics:**

* A picture containing screenshot, diagram, plot, line

  Description automatically generatedWe tried to find out whether people spend more on laptops or on PCs. We would use the alternative hypothesis, but because the sample size is very small and the distributions are hard to identify, we had to resort to empirical statistical methods. The most suitable here would be **Bootstrap Algorithm**.

These are the results from running the algorithm:

* A picture containing screenshot, diagram, text, plot

  Description automatically generatedBootstrap mean = 110.28260869565234
* Bootstrap confidence interval is constructed by taking the value at
  + (alpha\*iterations) and(iteration-alpha\*iteration)
* 95% confidence interval for mu1 - mu2: [ -5252.17391304348 , 5250.0 ]
* The bootstrap hypothesis test is performed by checking whether the value for the variable lies within the acceptance region or not, since the null hypothesis states that *𝑋*=*0* where *𝑋*=*𝜇1*−*𝜇2*, and 0 lies inside the acceptance region, therefore we do not have enough evidence to reject the null hypothesis, that people tend to spend more on one device than another.
* We also tried to find out if the two variables, gender and purchase choice, are dependent. So we used **Chi-Square Test (95% confidence level)**.

𝐻*0*: The gender has no relationship with one's purchase choice.

𝐻*1*: The two variables, gender and purchase choice, are dependent.

* The p-value we extracted was 0.12639457846098398
* Since p-value = 0.12639 > *𝛼*, we cannot reliably reject the null hypothesis. **There is not enough evidence to say that the two variables, gender and purchase decision are dependent.**
* We conducted a similar test for age and purchase decision. **(95% conf)**
  + P-value = 0.4463577738771113
  + Since p-value = 0.44635 >*𝛼*, we cannot reliably reject the null hypothesis. **There is not enough evidence to say that the two variables, age and purchase decision are dependent.**
* When we did the same thing with the factors matrix, this is what we gathered. Since all the variables are ordinal, we use the chi-square test to test all of them for dependency.
  + With alpha = 0.05, the only two factors that show dependency is how much a user cares about portability and working without cable, which are highly correlated because the latter is part of the first.
  + Using alpha = 0.1, we can be at least 90% confident that the screen size also shows dependency.
  + For other tests, there is not enough evidence for dependency.
  + Other factors will be checked through logistic regression analysis.

**4.Model building and interpretation:**

* Dropped ['working\_without\_cable\_importance', 'affordability\_importance', 'processor\_generation'] columns due to multicollinearity
* The main statistical algorithm used is **Logistic Regression**, to obtain odds ratio in the presence of more than one explanatory variable based on prior observation of a dataset.
* Used **One-hot encoding** from the **sklearn** library, which can be used to transform one or more categorical features into numerical dummy features useful for training machine learning model.
* We split data for training and testing.
* Upon testing the model we found:
  + The accuracy is better than the baseline.
  + We evaluated the model, and we can see that it's valuable. Now we can use the whole data for training.
* Using the model for the entire dataset the results when it comes to importance for Laptops were:
  + The 'portability importance' is the most effective feature for increasing the odds of buying a laptop as the next device. An increase of the 'portability importance' feature by one unit increases the odds of buying a laptop by a factor of 3.5 when all other features remain the same.
  + An increase of the 'income level' feature by one unit increases the odds of buying a laptop by a factor of 2.42 when all other features remain the same.
  + An increase of the 'laptop usage percentage' feature by one unit increases the odds of buying a laptop by a factor of 2.21 when all other features remain the same.
  + An increase of the 'CD existence importance ' feature by one unit increases the odds of buying a laptop by a factor of 1.87 when all other features remain the same.
  + Using the device for studying mainly increases the odds of buying a laptop by a factor of 1.74 when all other features remain the same.
* Using the model for the entire dataset the results when it comes to importance for PCs were:
  + The 'desktop usage percentage' is the most effective feature for increasing the odds of buying a desktop as the next device. An increase of the 'desktop usage percentage' feature by one unit increases the odds of buying a desktop by a factor of 2.48 when all other features remain the same.
  + Having a Core i5 processor increases the odds of buying a desktop by a factor of 1.98 when all other features remain the same.
  + An increase of the 'performance importance' feature by one unit increases the odds of buying a desktop as the next by a factor of 1.84 when all other features remain the same.
  + An increase of the 'screen size importance' feature by one unit increases the odds of buying a desktop by a factor of 1.83 when all other features remain the same.
  + Being a Data scientist increases the odds of buying a desktop by a factor of 1.6 when all other features remain the same.

**5.****Reliability and** **validity**

Starting with reliability , we went for **Cronbach's alpha** approach which Is a way of assessing reliability by comparing the amount of shared variance, or covariance, among the items making up an instrument to the amount of overall variance

* Started by some preprocessing ,

- dropping the columns that will not be used as ["main\_usage\_device",

"income\_level",  "own\_tablet"]

- changing the gender “string” values to numerical values

- same goes for the “device\_performance” changed it to a ranged

numerical values from 1 to 4

* Then we used Cronbach alpha from pingouin lib to get general Reliabilityvalue



* Then we normalized the “device price” dividing it by it’s max value

And redid Cronbach



For the validity stage we resent the survey for some of the respondents asking them to refill it if possible , getting a 10% of the original respondents to refill it half of which didn’t write their id’s properly the first time , making the validity data 5%

* For those 5% data , we calculated how identical there answer was to theirs first response ,result was (90%, 81%, 87%, 0.87%, 48%)
* So we have an average validity value of 79%