Statistical Modelling and Design of Experiments

Assignment I

Stefano Petrilli
stefano.petrilli@upc.edu

Pol Segura
pol.segura.retana@estudiantat.upc.edu

Jakob Eberhardt
jakob.eberhardt@estudiantat.upc.edu

April 28, 2024

Contents

List of Figures

Introduction

In this assignment, we employ different statistical methods to data sets and evaluate the results to draw the respective conclusions. In the first part of the report, we work on a laptop data set. Section 1 includes descriptive and inferential statistics about this data set. In section 2, we study relationships between the data points by using the ANOVA test. In section 3, we develop and test a linear regression model. In section 4, we carry out a multivariate analysis of sports competition data.

1 Visualisation, Chi-Square and t-Test

In this section, we will start to statistically describe the Laptop Price Prediction cleaned Dataset [6]. The following table shows the included data and the types we assigned to them. In total, the set includes data about more than 1270 unique laptop models. In the following analysis, we will focus on `Weight`, `Price` and the categorical variables of `Ultrabook`, `Notebook` and `2 in 1 Convertible`. This subset still consists of 1016 rows, which is more than sufficient for a meaningful analysis.

Column Name	Type	Unit
Company	factor	-
TypeName	factor	-
Ram	factor	-
Weight	numeric	Kilogram
Price	numeric	Thousands of Rupees
TouchScreen	factor	-
Ips	factor	-
Ppi	numeric	Pixels per inch
Cpu_brand	factor	-
HDD	factor	-
SSD	numeric	Gigabyte
Gpu_brand	factor	-
Os	factor	-

Table 1: Data Frame Column Classes (1.a)

1.1 Initial data summary (1.b and 1.c)

A small sample of the data once filtered is the following:

Company	Type Name	Ram	Weight	Price	Touch Screen	Ips	Ppi	Cpu Brand	HDD	SSD	Gpu Brand	Os
Apple	Ultrabook	8	1.37	11.176	0	1	226.983	Intel Core i5	0	128	Intel	Mac
Apple	Ultrabook	8	1.34	10.777	0	0	127.678	Intel Core i5	0	0	Intel	Mac
HP	Notebook	8	1.86	10.330	0	0	141.212	Intel Core i5	0	256	Intel	Others
Apple	Ultrabook	16	1.83	11.814	0	1	220.535	Intel Core i7	0	512	AMD	Mac
Apple	Ultrabook	8	1.37	11.473	0	1	226.983	Intel Core i5	0	256	Intel	Mac
Acer	Notebook	4	2.10	9.967	0	0	100.455	AMD Processor	500	0	AMD	Windows

Table 2: Small sample of the filtered dataset as outputted from the command `head(filtered_data)`

Min	1st Quartile	Median	Mean	3rd Quartile	Max.
0.690	1.440	1.920	1.862	2.200	4.420

Table 3: Summary Statistics for Weight as outputted from the command `summary(filtered_data$Weight)`

Min	1st Quartile	Median	Mean	3rd Quartile	Max.
9.254	10.302	10.732	10.716	11.145	12.472

Table 4: Summary Statistics for Price as outputted from the command `summary(filtered_data$Price)`

The rest of the summary executed for questions 1.b and 1.c can be found in the code, in file `first_question.R`

1.2 Cross Classification (1.d)

The conditional probability of having a touchscreen given the laptop type, denoted $P(\text{TouchScreen} \mid \text{TypeName})$, is presented in the table below. We can see that a *2 in 1 Convertibles* is very likely to have a touch screen while only some *Ultrabooks* and almost no *Notebook* have such a screen.

Laptop Type	**No Touch Screen**	**Touch Screen**
2 in 1 Convertible	0.0172	0.9828
Notebook	0.9731	0.0269
Ultrabook	0.7680	0.2320

Table 5: Conditional Probabilities of Touch Screen Presence by Laptop Type

The next table shows the obtained conditional probabilities for the types of laptops given that a laptop has a touch screen or not which is denoted by $P(\text{TypeName} \mid \text{TouchScreen})$. If a laptop has a touch screen, the likelihood of it being a *2 in 1 Convertibles* is 64% while the chance of it being a *Notebook* or *Ultrabook* is only 10% and 25%, respectively.

Laptop Type	**No Touch Screen**	**Touch Screen**
2 in 1 Convertible	0.0024	0.6404
Notebook	0.8198	0.1067
Ultrabook	0.1778	0.2528

Table 6: Conditional Probabilities of Touchscreen Presence by Laptop Type

Laptop Type	**No Touch Screen**	**Touch Screen**
2 in 1 Convertible	2	114
Notebook	687	19
Ultrabook	149	45

Table 7: Contingency Table of a subset of Laptop types

1.3 Association between Type and Touch Screen (1.e)

As we have an adequate sample size and two categorical variables, we can test the association between the type of computer and the presence of a touch screen by applying a Chi-Square [9] test. The H_0 of this test states that both variables are *independent*. The test yields a p-value of 2.2e-16 which gives us enough evidence to reject the null hypothesis. Hence, `Type` and `TouchScreen` are statistically dependent in this sample.

1.4 Analysing the Distribution of Price (1.f)

We will now analyze the distribution of the `Price` in the data set and in the subgroups `2 in 1 Convertible`, `Notebook`, and `Ultrabook`. We will first analyze the distributions visually and then continue to use the Shapiro-Wilk test [12]. The H_0 of this test assumes a normal distribution of a numeric variable, hence, it allows us to significantly and formally validate our visual impression.

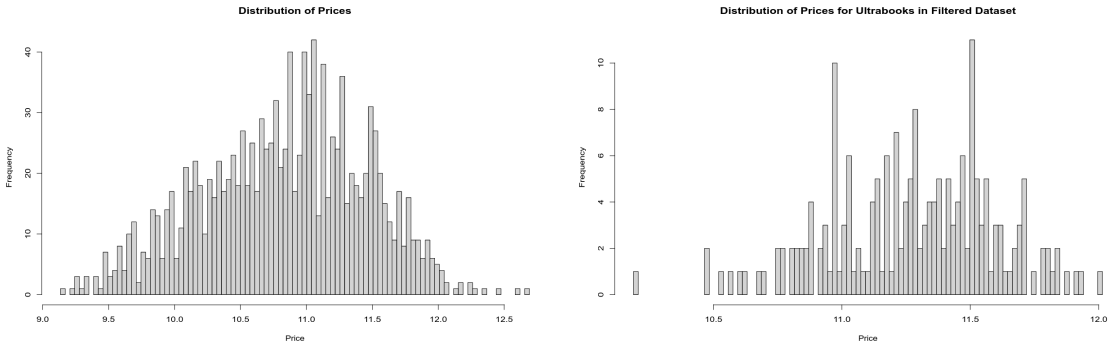

Figure 1: The left figure shows the distribution of the `Price` variable across the whole data set. The right-hand side plot shows the distribution of prices among `Ultrabook` models

Regarding the distribution of `Price` in the full data set, we can see from the left plot of figure 1 that it does not follow a bell shape because we have some extremely expensive models in the data set. Yielding a p-value of 4.074e-06, the respective Shapiro-Wilk test confirms this as we have to reject the null hypothesis. If we look at the right-hand side plot which shows the histogram of `Price` among ultrabooks, we can see that even though it also includes some high values, it still generally follows a normal distribution. We can confirm this with another Shapiro-Wilk test which returns a p-value of 0.1248. Therefore, we have enough evidence to accept the H_0 and state that the price among the ultrabook subset follows a normal distribution.

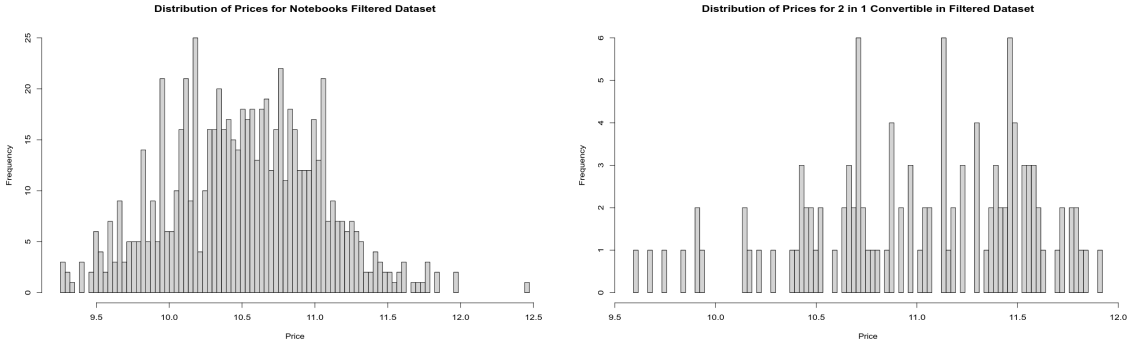

Figure 2: The normal distribution of the prices among `Notebook` laptops can be seen in the left plot. Yet, the distribution of prices in `2 in 1 Convertible` models is clearly not normal

Similarly, we can see on the left histogram of figure 2 that the prices of laptops belonging to the `Notebook` category also appear normally distributed. The subsequent Shapiro-Wilk test confirms this, as we accept the H_0 with a p-value of 0.1131. In opposite to that, the distribution of prices of `2 in 1 Convertible` laptops does not exactly suggest a bell shape. With a p-value of 0.001445, we also must reject the null hypothesis of the Shapiro-Wilk test, therefore there is significant evidence that the data do not follow a normal distribution.

1.5 Distribution of Price cross categories of Type (1.h)

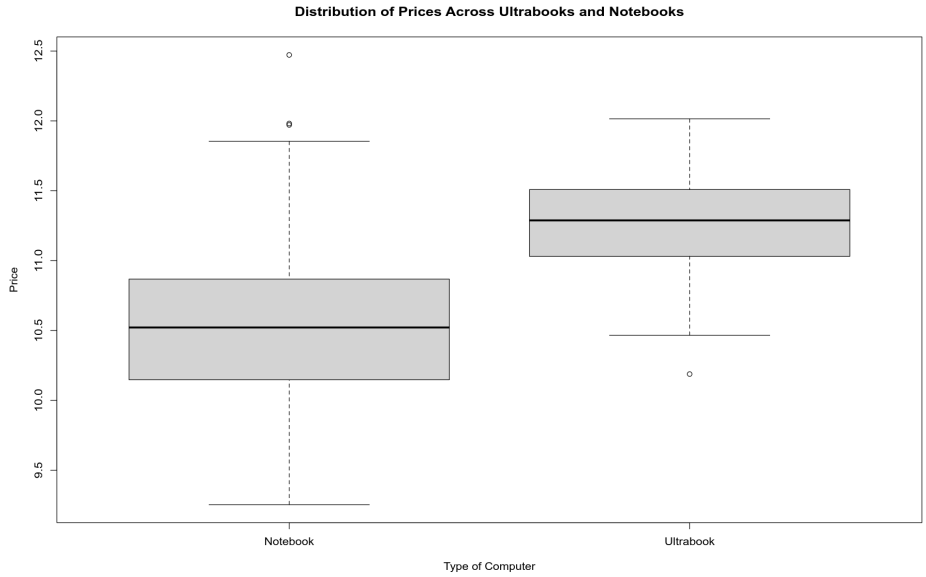

Figure 3: A Boxplot of the distribution of `Price` across the categories

We can tell by the Boxplot shown in figure 3 that `Ultrabook` laptops are on average more expensive than laptops considered a `Notebook`. Also, both categories include outliers in this data set. In the case of `Notebook`, the outliers are more expensive. One could assume that these might be Apple models, but after we looked up the outliers, we found that all three are Lenovo. Among the category of `Ultrabook`, the only outlier seems unusually cheaper.

1.6 Comparison of Average Price of Ultrabooks and Notebooks (1.i)

In this section, we compare the average `Price` of ultrabooks and notebooks. To this end, we want to apply a Student's t-test [13]. Before, we check the respective assumptions which are

1. The data are continuous
2. The distribution is approximately normal.
3. There is homogeneity of variance (i.e., the data variability in each group is similar).

We already know that the `Price` data is continuous and, at least for the subgroups `Ultrabooks` and `Notebooks`, normally distributed (see section 1.4). To check the homogeneity of variance, we use an F-test. The F-test null hypothesis states that the variances between groups are equal. The F-test yields a p-value of 3.784e-13, providing substantial evidence that the variances between our groups are significantly different. To accommodate the variance heterogeneity, we apply the Welch transformation [7] to adjust the t-test analysis. The null hypothesis assumes equal means between the groups. However, the Welch t-test returns a p-value of 2.2e-16, leading us to reject the H_0 . This indicates a statistically significant difference in the average prices between ultrabooks and notebooks.

2 ANOVA

The goal of this section is to examine the analyze the relationship between the brand of the computer and its respective price and its weight by using the ANOVA test [3].

2.1 Data (2.a)

We are using a subset of the data which includes the models of the companies *Dell*, *Acer* and *HP*. From summarizing the variable `Company`, we can see that *Dell* and *HP* have a similar number of observations (291

and 268 respectively), whereas *Acer* has less (101). However, we do not expect this to lead to a lack of statistical power.

To analyze the distributions of the `Weight` and `Price` in this subset, we can visualize the data using a histogram. In the second step, we apply a Shapiro-Wilk test to validate that the distribution of both variables is normal.

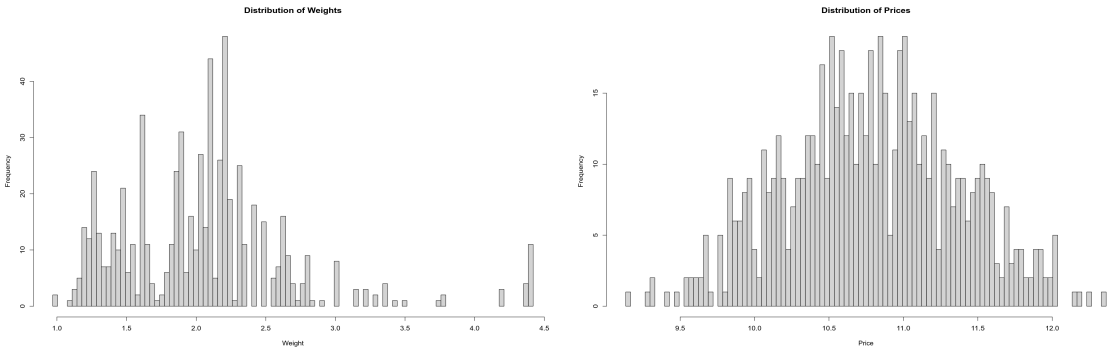

Figure 4: Weight and Price distribution of computers of the brands *Dell*, *Acer* and *HP*

From the executed tests we can state that the `Price` follows a normal distribution (p-value = 0.28), yet, the `Weight` is right-skewed and therefore fails the Shapiro test with a p-value = 2.2e-16.

2.2 Checking the Assumptions (2.b)

Now, aiming to analyze the relationship between the brand of the computer and its price and its weight, we will perform an ANOVA test. To this end, we first have to check the respective assumptions:

- The distribution of the population must be normal.
- Homogeneity of variances: Equal variance among the groups.
- Observations are independent. As stated in the disclaimer, generally it is impossible to verify this, but, as this is a regression analysis, the Durbin-Watson test can be used on the residuals to detect the presence of autocorrelation.

Normality has to be verified on the input data while the assumptions of homogeneity of variances and independence of observations have to be verified on the residuals.

2.3 Attemt to normalize Weight

In section 2.1, we have seen that the weight variable is not normally distributed. To be able to apply ANOVA, we will try to remove outliers and do a log transformation [11]. We remove outliers by using the inter-quartile range (IQR) method along with the 1.5 IQR rule: First, we compute the first and third quartiles, then the IQR, define lower and upper bounds for outliers and remove the values outside the boundaries [10]. Checking with Shapiro-Wilk we see that removing the outliers is not enough to obtain a normal distribution. Therefore, we continue by applying a logarithmic transformation. But it is again not enough for normalizing the data (Shapiro-Wilk p-value = 3.621e-11), thus, we cannot proceed with the ANOVA test for this variable, as the results will not be reliable.

2.4 Applying the ANOVA Test (2.c)

Now we want to know if the brand of the computer has a significant effect on its price, so we proceed by fitting the ANOVA model with `Price` as the response and `Company` as the independent variable. With a p-value of 2e-16, we have enough evidence to reject the null hypothesis. This means at least one mean is significantly different. In this case, as we only use one independent variable, we can be sure that the means of `Price` among `Company` are different.

As a post-hoc test, we can apply Turkey's [5] test, from which we can interpret that *Dell* has a significantly higher price compared to *Acer*, with a mean difference of approximately 0.623 (95% CI: [0.476, 0.769]) and an adjusted p-value of 0.000. *HP* also has a significantly higher price compared to *Acer*, with a mean difference of approximately 0.503 (95% CI: [0.356, 0.651]) and an adjusted p-value of 0.000. However, there is no

significant difference in price between *HP* and *Dell*, as the mean difference is approximately -0.119 (95% CI: [-0.227, -0.012]) with an adjusted p-value of 0.025. We also perform post-hoc tests that adjust the p-value to reduce type I errors, Bonferroni is more conservative than FDR (False Discovery Rate). We can interpret the results of the tests as follows:

- *Acer* vs. *Dell*: The p-value for comparing prices between *Acer* and *Dell* is extremely small (¡ 2e-16), which indicates a highly significant difference in prices between these two companies.
- *Acer* vs. *HP*: The p-value for comparing prices between *Acer* and *HP* is also very small, indicating a highly significant difference in prices between these two companies.
- *Dell* vs. *HP*: The p-value for comparing prices between *Dell* and *HP* is 0.027, which is smaller than alpha (0.05), indicating a statistically significant difference in prices between these two companies, but not as extreme as the differences observed between *Acer* and each of *Dell* and *HP*.

Lastly, we have to verify the assumptions that are left. To check the homogeneity of variances we can use a Q-Q plot. As can be seen in figure 5, we can state that the distribution of the residuals is normal. However, we can further check this with Levene's test. The p-value obtained is 0.067. As it is greater than alpha, we fail to reject the null hypothesis of homoscedasticity, therefore the assumption is fulfilled.

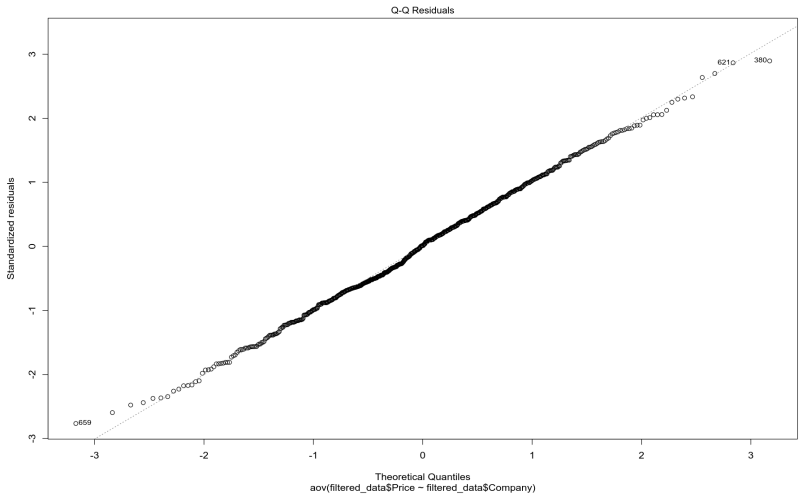

Figure 5: A Q-Q Plot of `Price` and `Company`

We checked the independence of observations, which we tested using the Durbin-Watson test. The test yields a p-value of 0.1393, hence, we reject the null hypothesis which means that the observations are independent. All the assumptions for ANOVA are hence verified for `Price`.

2.5 Brand, Touch Screen and Price (2.d)

For a last analysis, we check the effect of brand and touch screen characteristics together on the price. As demonstrated previously in section 2.2, the distribution of `Price` is normal. Hence, we proceed to use ANOVA to verify the interaction of the two terms.

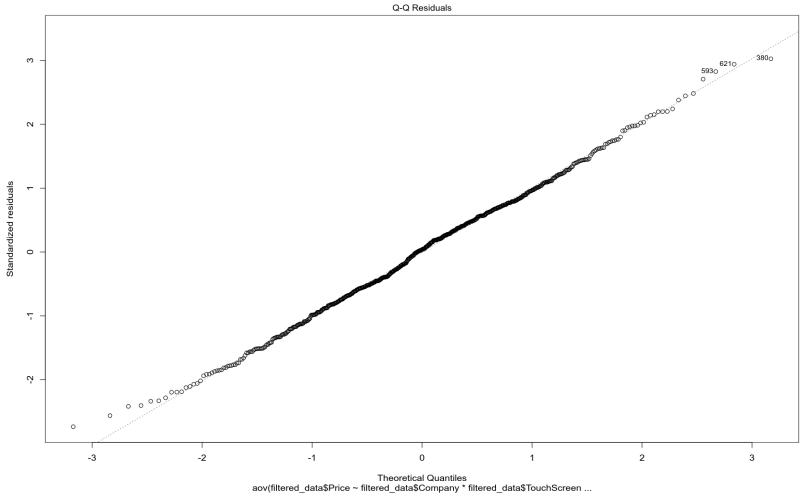

Figure 6: A Q-Q Plot of `Price`, `Company`, and `TouchScreen`

In the Q-Q plot shown in figure 6, we can see that the residuals follow a normal distribution. Furthermore, if we check Levene's test on the interaction model, we see that we yield a p-value of 0.47, which confirms that homoscedasticity is given. Lastly, we check the assumption regarding the independence of observation with the Durbin-Watson test. It yields a Durbin-Watson value of 1.92, which is sufficiently close to 2 and hence confirms the independence of errors.

As the assumptions are fulfilled, we can proceed with the analysis of the ANOVA test for the interaction model. The test is significant for this model in both variables (***), hence, we have enough evidence to reject the null hypothesis. This means at least one mean is significantly different. We now apply Tukey's as a post-hoc test. The results presented in table 8 show that all of the differences are significant, which agrees with the results obtained previously.

	Dell	HP	Acer
Dell	-	-0.119	0.623
HP	0.119	-	0.503
Acer	-0.623	-0.503	-

Table 8: Mean differences between groups

3 Regression Analysis

In this section, we take into account the numerical variables in our dataset to develop regression models. We consider `Weight`, `Ppi` and `SSD`. The `SSD` variable distribution fits into a relatively small number of categories. Despite this, we still believe there are sufficient distinct categories in the variable to consider it numerical.

3.1 Simple Linear Regression (3.a)

In the following, we fit each simple linear regression that can be done with the variables mentioned. We always use `Price` as the dependent variable. For every model, we check the assumptions that have to be addressed when performing linear models, which are:

1. **Homoscedasticity**: The variance of the residuals should be constant across all levels of the independent variables. We the property using the Breusch-Pagan test.

2. **Normality of residuals**: The residual (the difference between observed and predicted values) should be normally distributed. We check this visually with a Q-Qplot and with a Shapiro-Wilk test on the residuals of the model as in previous sections.

3. **No multicolinearity**: There can not exist a correlation between the independent variables. We check this with the `vif()` function, in which values have to be lower than 9 to state that there is no colinearity.

4. **Independence of errors**: The residual (the difference between observed and predicted values) should be independent. We check this with the Durbin-Watson test, which has the hypothesis that the auto-correlation of the errors is 0.

In the case of simple linear regression, the colinearity between independent variables does not have to be checked, as we only use one variable as a predictor. To check which model performs better, we can use the Adjusted R-squared value by using the R command `summary()`. The R-squared value is the amount of variance explained by the model. It is a measure of how well the model fits the data. The higher it is, the better the model. Moreover adjusted R-squared will decrease if additional variables do not contribute to the model's explanatory power.

Factor	Adjusted R-squared	Coefficient
Weight	0.02215	-
Ppi	0.2305	0.006838
SSD	0.4336	0.002188

Table 9: Simple regression R-squared per factor and coefficients

With `Weight` as the explanatory variable, we obtain an adjusted R-squared of 0.0222, yet, the homoscedasticity assumption is not given, as the Breusch-Pagan test yields a p-value of 6.509e-5. However, `Ppi` and `SSD` fulfil the assumptions and yield a p-value of 0.347 and 0.05876 in the Breusch-Pagan. To check the normality of residuals, we apply a Shapiro test on `Ppi` (0.2226) and `SSD` (0.07079) to confirm that both fulfil the assumption. Regarding the performance, with `Ppi` we find an adjusted R-squared of 0.2305. By using `SSD` we achieve an R-squared of 0.4336. Lastly, we check the independence of errors in the remaining models using a Durbin-Watson. In the `Ppi` model, we find a Durbin-Watson value of 2.05 which indicates that the errors are independent. The same applies to `SSD`, where we obtain a value of 2.04.

Regarding the simple linear regression, we can state that `SSD` is a better predictor for `Price` compared to `Ppi`. We do not take `Weight` into account, as it does not fulfill the homoscedasticity assumption and therefore might produce unreliable results.

3.2 Multivariate Linear Regression (3.b)

We will proceed with multiple linear regressions in the same manner.

Factor	Adjusted R-squared
Weight + Ppi	0.3336
Weight + SSD	0.409
Ppi + SSD	0.4637

Table 10: Multivariate regression R-squared per factor

With these results, we can state that the best model for predicting `Price` is `Ppi+SSD`, as it has the higher adjusted R-squared. We can do further checking with partial F-tests. For the partial F-tests, we will use ANOVA. The ANOVA test suggests that the model `Ppi+SSD` provides a significantly better fit to the data compared to the model only considering `SSD`. This conclusion is based on the significantly lower residual sum of squares (RSS) and the associated F-statistic with a very low p-value of 2.2e-16, indicating strong evidence against H_0 .

3.3 Add a factor to multivariate regression (1.c)

The model could still be further improved by adding a factor. To check that we do a for loop adding individually each one of the factors and checking the adjusted R-squared values.

Next, we start check the assumptions of the model starting from high to low values. In that order, the models with `Cpu_brand`, `Ram` and `TypeName` fail the assumption of homoscedasticity. Finally, the model including `Gpu_brand` fulfills all assumptions. As in the previous two model comparisons, we now check which model (with or without the factor) is a better predictor for `Price` by using ANOVA.

The ANOVA test suggests that the model with the factor provides a significantly better fit to the data compared to the model without the factor. This means that adding `Gpu_brand` is better for explaining the

Factor	Adjusted R-squared
Company	0.5471
TypeName	0.6320
Ram	0.6842
TouchScreen	0.4651
Ips	0.4745
Cpu_brand	0.6946
HDD	0.5146
Gpu_brand	0.5494
OS	0.5056

Table 11: Adjusted R-squared per factor

variation in the response variable `Price`. This conclusion is based on the significantly lower residual sum of squares (RSS) and the associated F-statistic with a very low p-value (***), indicating strong evidence against the null hypothesis.

Lastly, we check the validity of the chosen model (`PPI+SSD+Gpu_Brand` model). We check visually how good is the model by plotting 'Actual vs Predicted values' and the 'Residuals vs Fitted Values'.

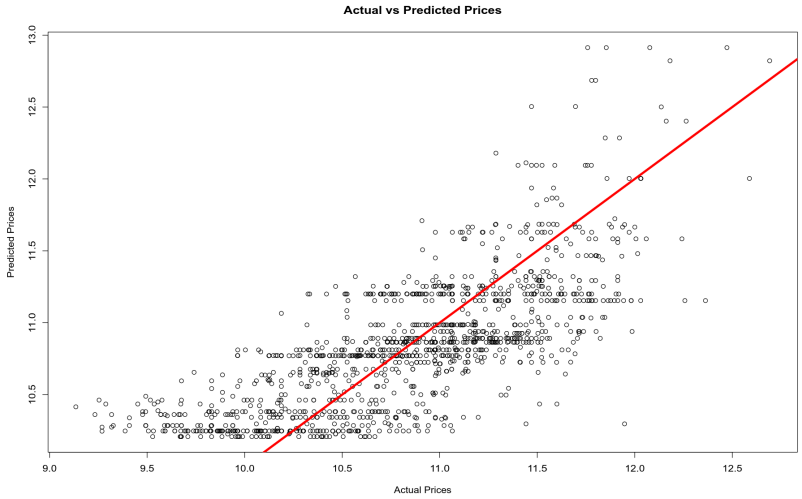

Figure 7: Redisual vs Fitted of the `PPI+SSD+Gpu_brand` model

3.4 Benchmarking the Models

To verify the model, we take a subset of 10 random laptops from the dataset and verify how accurate the model is.

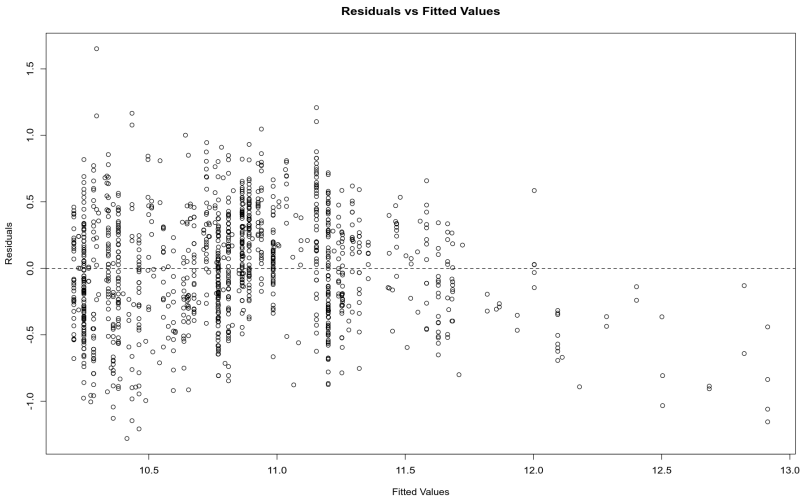

Figure 8: Redisual vs Fitted of the `PPI+SSD+Gpu_brand model`

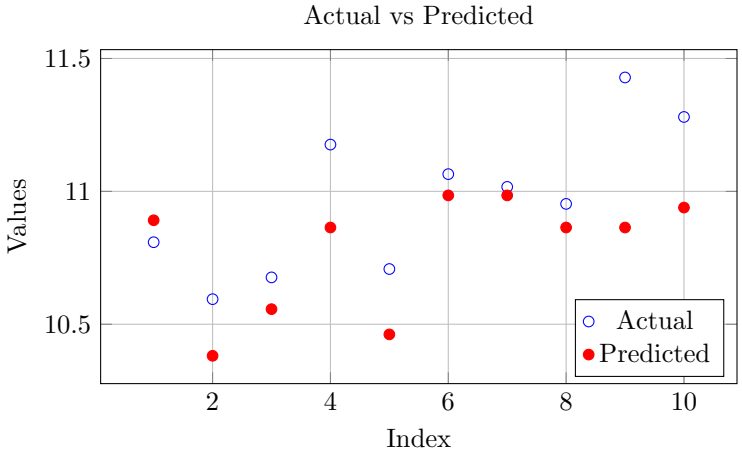

Figure 9: Actual vs Predicted

Predicted	Actual	Difference	Percentage (%)
10.89124	10.80859	0.08265075	0.7646763
10.38124	10.59430	-0.21305816	-2.0110640
10.55674	10.67629	-0.11955347	-1.1198033
10.86393	11.17599	-0.31205279	-2.7921724
10.46177	10.70777	-0.24599927	-2.2973900
10.98480	11.06480	-0.08000619	-0.7230692
10.98480	11.01680	-0.03199970	-0.2904629
10.86393	10.95284	-0.08890924	-0.8117458
10.86393	11.42854	-0.56461023	-4.9403516
10.93904	11.27992	-0.34088446	-3.0220462

Table 12: Comparison of Predicted and Actual Values on a random sample of the dataset using `PPI+SSD+Gpu_Brand`

3.5 Analysis of the result

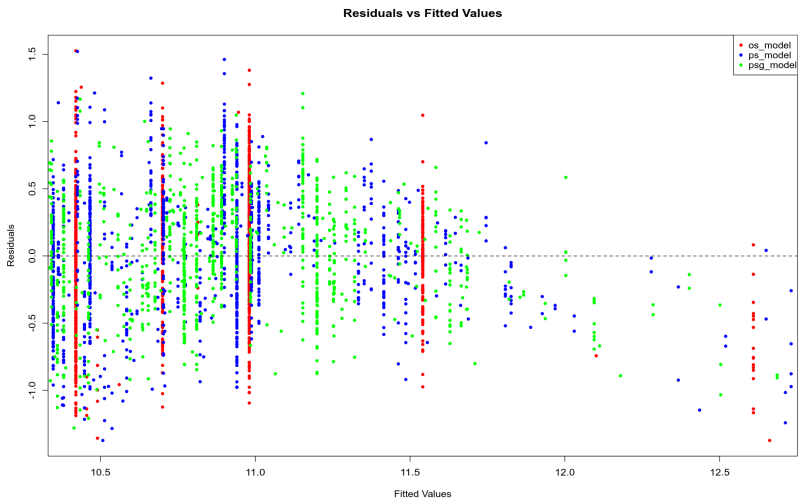

Figure 10: A comparison of the three discussed models. The model fitting `PPI+SSD+Gpu_Brand` (green) shows the least residual. The `SSD` (red) and `SSD+PPI` (blue) include skewed residuals

As can be seen in table 12, the values are accurate, so we can conclude that the `PPI+SSD+Gpu_Brand` model is indeed predicting `Price` well. From graph 10, we can also say that the residuals of the `PPI+SSD+Gpu_Brand` model are not just less skewed concerning the other models but are also less cluttered. This leads us to the conclusion that the `PPI+SSD+Gpu_Brand` model is not just accurate, but also more precise than the other models taken into consideration. As the display, the SSD and the GPU are among the most expensive components in a laptop it does not come as a surprise that those are also good predictors of the final prices.

4 Principal Component Analysis

The following parts include a Principal Component Analysis (PCA) [8] applied to the Decathlon dataset [2].

4.1 Principal Component Analysis

Not all the variables are suitable for the Principal Component Analysis (PCA). We removed from the dataset the variables *competition*, *points* and *rank*. We removed these variables respectively for the following reasons: they are not numerical; we only wish to include values relevant to a discipline; and since we aim to predict the variable *rank*, including them in the prediction model would be nonsensical. We first normalize and center the data of the remaining columns to standardize the dataset. To continue with the PCA we made sure that the data fulfills the assumption of sphericity, we do that with Barlett's test, which needs the correlation matrix and the number of rows of the dataset. As is possible to notice from 11, not all the variables are following the same direction, which could be affecting the KMO value. This is related to the fact that in certain disciplines a higher score is better (*long jump*) while in other disciplines a lower score is better (*100m*). We redirected the variables regarding *100m*, *400m*, *110m hurdles* and *1500m* disciplines to solve this problem.

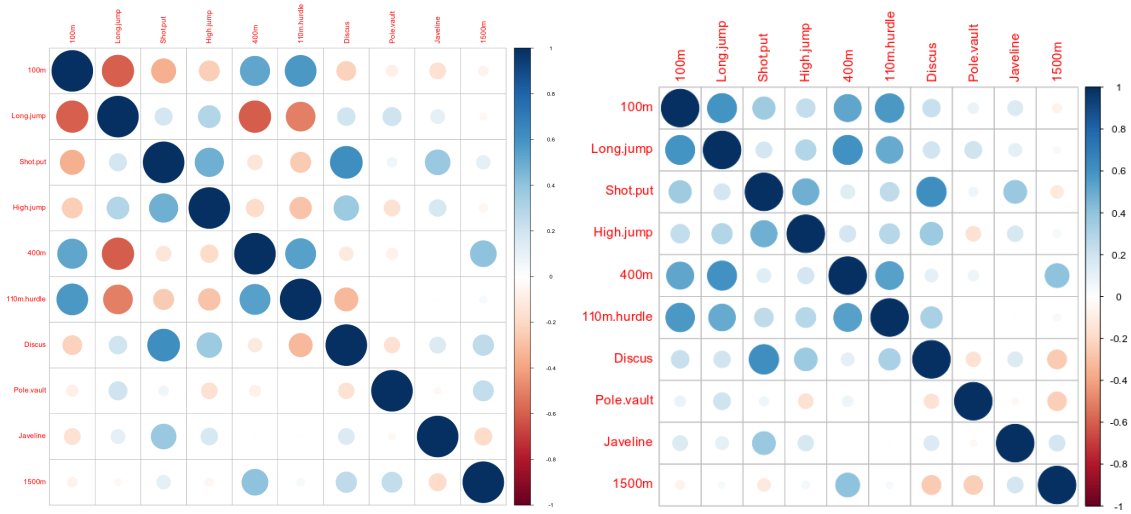

Figure 11: The left figure shows the correlation matrix before changing the directions of some disciplines. The right image shows the correlation matrix after directing all the variables in the same direction.

Using the correct correlation matrix we perform Bartlett's test [1]. As the p-value of1.15685e-10 is < 0.05 we have enough evidence to reject the null hypothesis, so the assumption of correlation between features is fulfilled. Before performing the PCA we check which Principal Components (PCs) we should use with the KMO test. The Kaiser-Meyer-Olkin (KMO) [14] [4] represents the degree to which each observed variable is predicted by the other variables in the dataset, and this indicates the suitability for factor analysis. The first KMO showed a value of 0.596, as it was very low, we decided to remove the variable that is contributing the least, in this case, *pole vault*. The second KMO value was still low so we proceeded, in the same way, removing *1500m*. Finally, we obtained an acceptable KMO value of 0.74. As we see in table 13 reproduced from [14] a value of 0.74 is classified as *Middling*. The paper confirms that values above 0.6 are sufficient.

KMO Value	**Degree of Common Variance**
0.90 to 1.00	Marvelous
0.80 to 0.89	Meritorious
0.70 to 0.79	Middling
0.60 to 0.69	Mediocre
0.50 to 0.59	Miserable
0.00 to 0.49	Unacceptable (No factor)

Table 13: Interpretation Guidelines for the Kaiser-Meyer-Olkin Test [14]

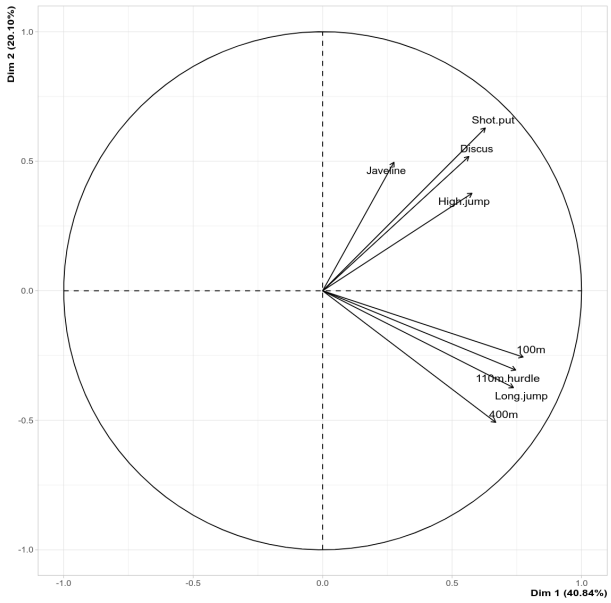

Figure 12: PCA Plot of Variables

Looking at the PCA we can state that:

- The variables have a strong relation with the first dimension which explains over 51% of the variance of the data.
- It seems that the first two components are enough to explain the variance of the data overall.

Only the first and second PC (*100 meters* and *long jump* respectively) have eigenvalues greater than 1, so we can conclude that these two PCs are enough to explain the variance of the data. We want to use a Scree plot to have another result to support the conclusion of only using the first two PCs.

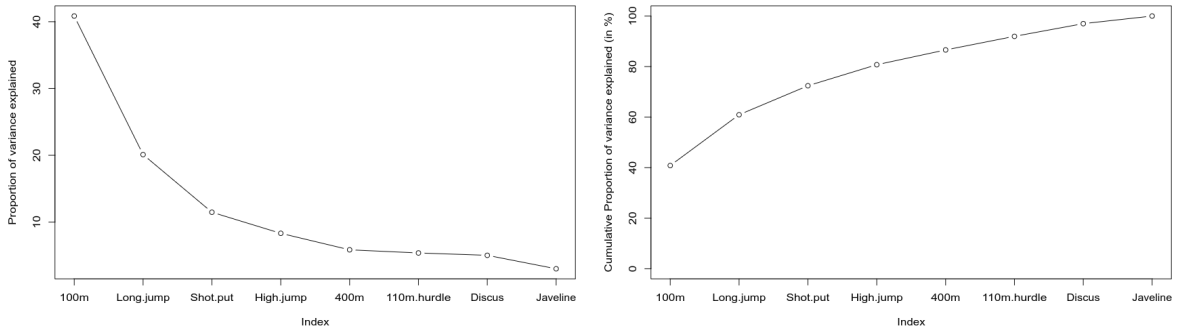

Figure 13: On the left, the proportion of variance explained by each discipline.
On the right, the cumulative variance explanation.

From this plot 13, we can see that indeed the first two PCs explain more than 60% of the variance, so they are enough for the model. Moreover, we want to know in which component the variables are most explained. We can see it through the cos of the models (`pca$var$cos2`). In this case, the first PC explains the vast majority (6 out of 7) of the variables. Except *shot put* and *discus* which are better explained by the second dimension. This could be related to the fact that the disciplines are similar (both involve thrusting an object) and measure the score in meters.

4.2 Principal Component Regression

As in the previous section, we have concluded that the first two PCs (*100m* and *long jump*) are enough to explain the variance of the data, we will perform the linear model with these two PCs as independent

variables. After fitting the model we have to check the assumptions. Firstly we check homoscedasticity with the Breusch-Pagan test. As the p-value (0.727) > 0.05 we have enough evidence for accepting the null hypothesis, so we can state that the homoscedasticity assumption is fulfilled.

Secondly, we check the normality of the residuals with a Q-Q plot (14) and with the Shapiro-Wilk test (14). In this case, the p-value (0.4211) > 0.05 so we have enough evidence for accepting the null hypothesis, and the assumption of normality in the residuals is fulfilled. Both from the Q-Q plot and Shapiro-Wilk we can conclude that the residuals are normally distributed.

Test Statistic (W)	***p*-value**
0.97272	0.4211

Table 14: Shapiro-Wilk Normality Test Results

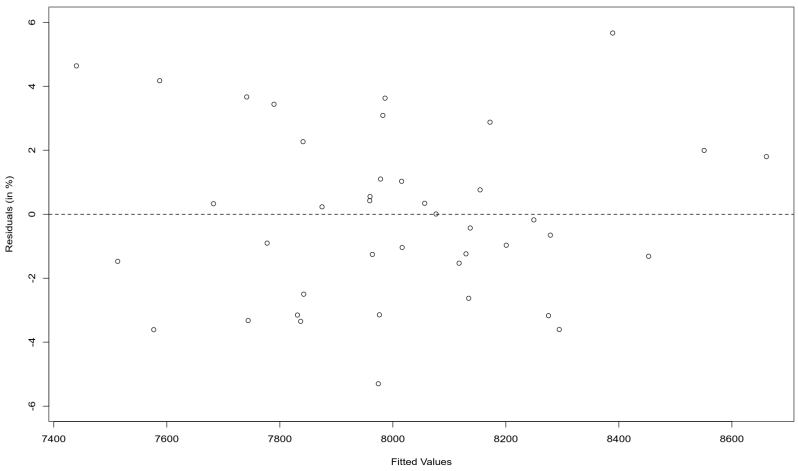

Figure 14: Redisual vs Fitted values

Thirdly we check that there is no multicollinearity between the independent variables, for this specific model the variables *100m* and *long jump*. We will use the function `vif()` for this purpose. As the correlation between these variables is 1.55864, we can state that the assumption of no multicollinearity between the independent variables is fulfilled. Lastly, we have to check the independence of errors, the Durbin-Watson test can be used to verify this. The test yields a `DW` value equal to 1.70, which confirms the independence of errors. To check how well the model performs we will plot *Actual vs Predicted Prices* (Figure 15) and a *Residuals vs Fitted Values* (Figure 16). From the plot of *Residual vs Fitted Values* (Figure 16), we can see that all the values are within a 6% of error which means that the model is accurate.

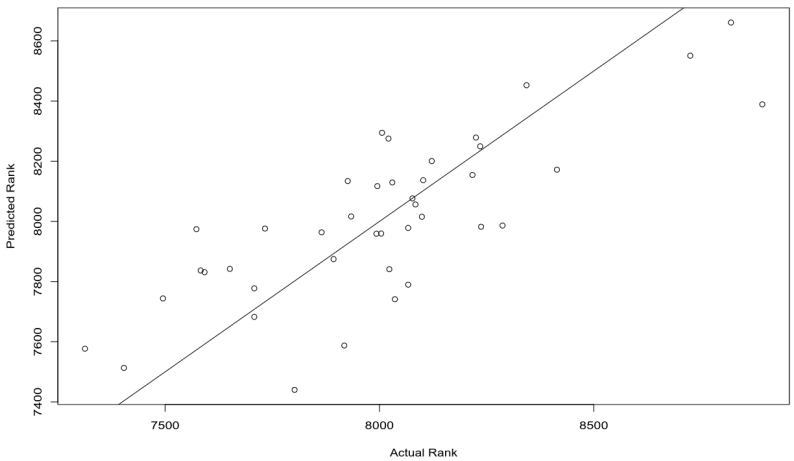

Figure 15: PCA Regression actual vs predicted.

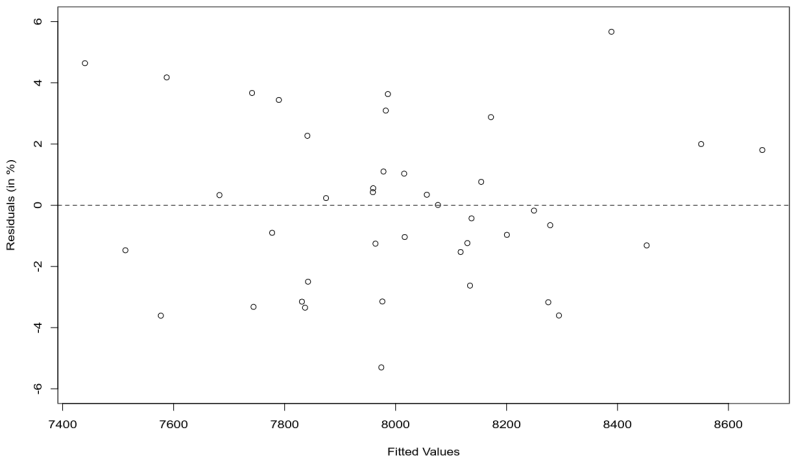

Figure 16: PCA Regression Residual vs Fitted.

5 Summary

In this assignment, we statistically described and summarized the given laptop dataset. We obtained and interpreted probabilities regarding touch screens in laptops. We found significant evidence for a dependence between the type of laptops and the presence of a touch screen. Further, we could conclude that ultrabooks are on average the most expensive type in the sample. We continued the analysis by comparing the prices of different companies. To this end, we applied the ANOVA test and found significant differences among the means of price. We developed a simple and multivariate linear regression model to predict the price of a laptop based on different input factors. We benchmarked our models and found that the multivariate model delivers accurate results. Lastly, we worked with the Decathlon dataset. Using Principal Component Analysis, we determined the variables that explain the majority of the variance, which turned out to be the 100 meters and long jump disciplines. Yet, shot put and discus are better explained by the second dimension, as they are similar disciplines related to throwing. We applied a Principal Component Regression and found an accurate model that delivers values within 6% of error.

Figure 17: A picture of the group at the seaside after finishing the assignment.

References

[1] Hossein Arsham and Miodrag Lovric. Bartlett's test. *International encyclopedia of statistical science*, 2:20–23, 2011.

[2] CRAN. *Decathlon dataset*, 2004. Data from the 2004 Olympic Games and the Decastar.

[3] Ellen R Girden. *ANOVA: Repeated measures*. Sage, 1992.

[4] Brent Dale Hill. *The sequential Kaiser-Meyer-Olkin procedure as an alternative for determining the number of factors in common-factor analysis: A Monte Carlo simulation*. Oklahoma State University, 2011.

[5] HJ Keselman and Joanne C Rogan. The tukey multiple comparison test: 1953–1976. *Psychological Bulletin*, 84(5):1050, 1977.

[6] Gyanprakash Kushwaha. Laptop price prediction cleaned dataset. `https://www.kaggle.com/datasets/gyanprakashkushwaha/laptop-price-prediction-cleaned-dataset`, 2023. Accessed: 2023-04-20.

[7] Zhenqiu Lu and Ke-Hai Yuan. Welch's t test, 01 2010.

[8] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

[9] Mary L McHugh. The chi-square test of independence. *Biochemia medica*, 23(2):143–149, 2013.

[10] Neil C Schwertman, Margaret Ann Owens, and Robiah Adnan. A simple more general boxplot method for identifying outliers. *Computational statistics & data analysis*, 47(1):165–174, 2004.

[11] Philip Sedgwick. Log transformation of data. *BMJ*, 345, 2012.

[12] S. S. SHAPIRO and M. B. WILK. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, dec 1965.

[13] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[14] Lütfi Sürücü, İbrahim Yikilmaz, and Ahmet Maslakci. Exploratory factor analysis (efa) in quantitative researches and practical considerations, 12 2022.