Analysis of Laptop Sale Data

Junwei Deng(junweid@umich.edu)

University of Michigan

1 Motivation

Back to my first year in the college, my parents asked me to choose a type of laptop as their present for me to celebrate that I entered a good college. When I opened the shopping website, I was shocked by the great variety of laptops and I could not find the one that has the highest price-performance ratio. Which brand has the most choices that meet my(an engineering student) requirement? How much RAM is the popular configuration nowadays? So many problems need to be answered by a proper exploratory data analysis.
My main motivation is to dig some insights from a laptop sale data to find out how the attributes of a laptop affect its price. I will focus on three main problems and some auxiliary problems based on them.

1. How do the price varies from laptop's attributes, such as Ram, Brand, Screen resolution, Type and Operating system?
2. Can we perform a linear regression model based on the laptop's attributes to predict the price?
   Auxiliary/Follow-up:
   (a) Does our model predict the price well?
   (b) Can some variables be dropped or combined in our model?
3. Can we find some meaningful clusters among all the laptops?
   Auxiliary/Follow-up problems:
   (a) Are Apple laptops outliers among all the laptops?

2 Data Source

I will use one dataset for this project.

2.1 Laptop Prices(https://www.kaggle.com/ionaskel/laptop-prices)

This dataset contains 1303 laptops with their attributes listed as following(Format, Example):

1. Company Name(Str, Apple)
2. Product Name(Str, MacBook Pro)
3. Laptop Type(Str, Ultrabook)
4. Screen Inches(Num, 13.3)
5. Screen Resolution(Str, IPS Panel Retina Display 2560x1600)
6. CPU Model(Str, Intel Core i5 2.3GHz)

7. RAM Characteristics(Str, 8GB)
8. Memory(Str, 128GB SSD)
9. GPU Characteristics(Str, Intel Iris Plus Graphics 640)
10. Operating System(Str, macOS)
11. Laptop's Weight(Str, 1.37kg)
12. Laptop's Price(Num, 1339.69)

3 Methods

3.1 Question 1

How did you manipulate the data to prepare it for analysis? I load the csv file to a R dataframe and then transform the dataframe to a datatable because of the speed problems. The R datatable will be used in the next two problems either. In this problem, we will find the distribution of price for different Ram, Brand, Screen resolution, Type and Operating system. The main method is to visualize the variables. For the Ram, Brand, Type and Operating system variables, I ensure them to be factor variables. For the Screen resolution, I find there are too many levels of factor, so I extract the resolution(e.g. "Full HD 1920x1080" to "1920x1080") only from the 'Screen Resolution' variable. I use R regex and the 'stringr' package to extract the string.
Furthermore, I will draw one graph for each attribute vs the price, so there should not be too many factor levels in an attribute, otherwise the plot will be ugly and crowd. I examine the number of factors by explore the summary of the laptop datatable and decide to draw a facet bar chart for price distribution on different brands and boxplots for the other attributes.

How did you handle missing, incomplete, or noisy data? The dataset is really clean and there are no data missing or incomplete. Still I find there are two factor levels in "Operating System" which is "macOS" and "Mac OS X". I want to clarify them. Apple changed the name from "Mac OS X" to "macOS" recently, which means there the laptop running "macOS" is kind of the old version of Apple laptops. I will not combine them to one factor level, so that we can compare the new Apple laptops and the old Apple laptops.

What challenges did you encounter and how did you solve them? The string extraction is really a big problem. The original data is a messy in many columns. They have too many redundant information. We have used regex in python but we did not use the regex on R. I tried to learn the 'stringr' package and solved this problem.

3.2 Question 2

How did you manipulate the data to prepare it for analysis? I extract more data from the string using the R regex, including the Ram, screen resolution and the cpu frequency. I apply the same way just like in Question 1. For the memory, I tried to split the data from one column to three columns according to the type of the memory. For example a "256 GB SSD + 512 GB HDD + 2 GB Flash" can be translate to [256,512,2]. After the translating, the dataset is done for both question2 and question3. The summary is as following.

```
       X               Company                Product                        TypeName        Inches
 Min.   :   1.0   Dell   :297   XPS 13               :  30   2 in 1 Convertible:121   Min.   :10.10
 1st Qu.: 331.5   Lenovo :297   Inspiron 3567        :  29   Gaming            :205   1st Qu.:14.00
 Median : 659.0   HP     :274   250 G6               :  21   Netbook           : 25   Median :15.60
 Mean   : 660.2   Asus   :158   Legion Y520-15IKBN:  19   Notebook          :727   Mean   :15.02
 3rd Qu.: 990.5   Acer   :103   Vostro 3568          :  19   Ultrabook         :196   3rd Qu.:15.60
 Max.   :1320.0   MSI    : 54   Inspiron 5570        :  18   Workstation       : 29   Max.   :18.40
                  (Other):120   (Other)              :1167
                                    ScreenResolution                               Cpu               Ram
 Full HD 1920x1080                          :507     Intel Core i5 7200U 2.5GHz :190    8GB    :619
 1366x768                                   :281     Intel Core i7 7700HQ 2.8GHz:146    4GB    :375
 IPS Panel Full HD 1920x1080                :230     Intel Core i7 7500U 2.7GHz :134    16GB   :200
 IPS Panel Full HD / Touchscreen 1920x1080: 53     Intel Core i7 8550U 1.8GHz : 73    6GB    : 41
 Full HD / Touchscreen 1920x1080            : 47     Intel Core i5 8250U 1.6GHz : 72    12GB   : 25
 1600x900                                   : 23     Intel Core i5 6200U 2.3GHz : 68    2GB    : 22
 (Other)                                    :162     (Other)                    :620    (Other): 21
                 Memory                         Gpu                 OpSys            Weight       Price_euros
 256GB SSD           :412   Intel HD Graphics 620  :281   Windows 10:1072   2.2kg  :121   Min.   : 174
 1TB HDD             :223   Intel HD Graphics 520  :185   No OS     :  66   2.1kg  : 58   1st Qu.: 599
 500GB HDD           :132   Intel UHD Graphics 620 : 68   Linux     :  62   2.4kg  : 44   Median : 977
 512GB SSD           :118   Nvidia GeForce GTX 1050: 66   Windows 7 :  45   2.3kg  : 41   Mean   :1124
 128GB SSD +  1TB HDD: 94   Nvidia GeForce GTX 1060: 48   Chrome OS :  27   2.5kg  : 38   3rd Qu.:1488
 128GB SSD           : 76   Nvidia GeForce 940MX   : 43   macOS     :  13   2kg    : 35   Max.   :6099
 (Other)             :248   (Other)                :612   (Other)   :  18   (Other):966
 ScreenResolution_alt   Cpu_alt            Memory_alt_SSD     Memory_alt_HDD     Memory_alt_FS
 Length:1303          Length:1303        Length:1303        Length:1303        Length:1303
 Class :character     Class :character   Class :character   Class :character   Class :character
 Mode  :character     Mode  :character   Mode  :character   Mode  :character   Mode  :character



 Memory_alt_SSD_num Memory_alt_HDD_num Memory_alt_FS_num  Cpu_alt_num      Weight_num       Ram_num
 Min.   :   0       Min.   :   0.0     Min.   :  0.000   Min.   :0.900   Min.   :0.690   Min.   : 2.000
 1st Qu.:   0       1st Qu.:   0.0     1st Qu.:  0.000   1st Qu.:2.000   1st Qu.:1.500   1st Qu.: 4.000
 Median : 256       Median :   0.0     Median :  0.000   Median :2.500   Median :2.040   Median : 8.000
 Mean   : 183       Mean   : 421.7     Mean   :  4.556   Mean   :2.299   Mean   :2.039   Mean   : 8.382
 3rd Qu.: 256       3rd Qu.:1024.0     3rd Qu.:  0.000   3rd Qu.:2.700   3rd Qu.:2.300   3rd Qu.: 8.000
 Max.   :1024       Max.   :2048.0     Max.   :512.000   Max.   :3.600   Max.   :4.700   Max.   :64.000
```

The PCA processing will be written in the 'Analysis and Result" section.

How did you handle missing, incomplete, or noisy data? The dataset is really clean and there are no data missing or incomplete.

What challenges did you encounter and how did you solve them? Some of the data in memory may write "1TB" rather than "1024 GB", if I don't pay attention to this problem, the data may be really inaccurate. I tried to detect this problem by setting a threshold, if I have a number less than 5, I will multiple them by 1024.

3.3 Question 3

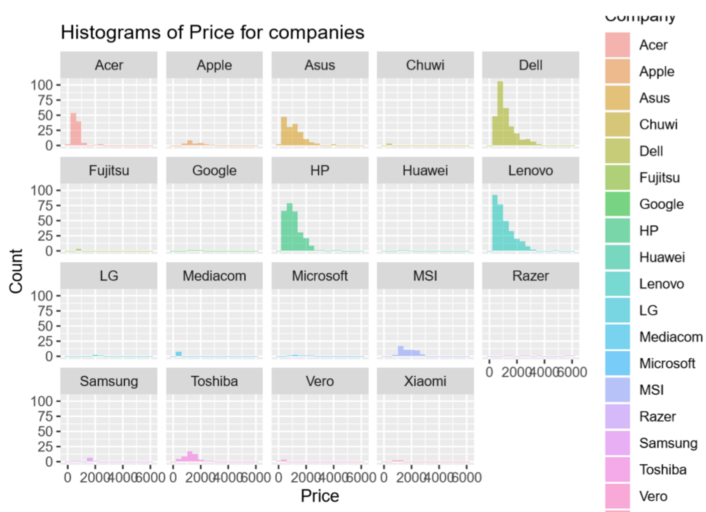
Nothing new, the same as Question 2.

4 Analysis and Results

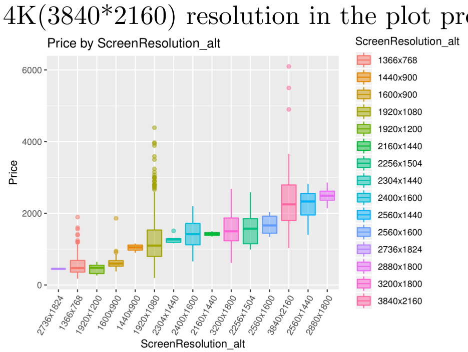
4.1 Question 1

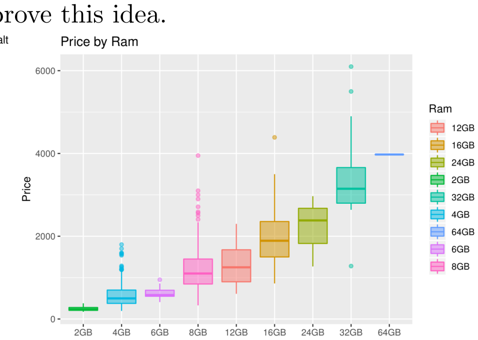
Workflow I construct a datatable and extract the useful part for some columns with string format. The I use ggplot and graph grammar to draw the plots we need.

Analysis From the Brand vs Price plot, we can find out that Acer, Asus, Dell, HP and Lenovo are the largest brands in the market. Lenovo tends to sell more low-price laptops and other brands tend to sell laptops with price around 1000 euro dollars.

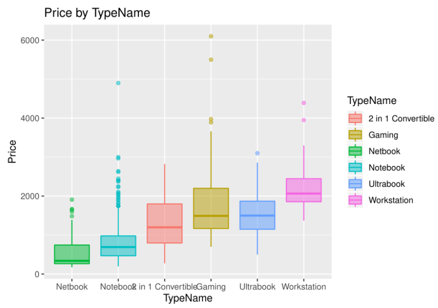


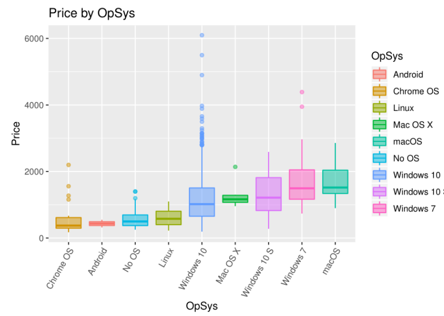
From the ScreenResolution vs Price and Ram vs Price plot, We can find out that The larger the Ram, the higher the price. And the trend is generally proper for Screen Resolution. The most popular HD(1366*768), FHD(1920*1080) and 4K(3840*2160) resolution in the plot prove this idea.





For the Type vs Price and Operating System vs Price plot, we can find that the distribution basically fit our expectation. Gaming laptop usually has a high configuration than normal notebook. MacOS is no doubt the operating system on the laptop with the highest price. and the laptop with new macOS is more expensive than the old Mac OS X.





It is clear that the quantitative and catalogic variable can be good predictor for the price of a laptop, which is what I will do in next section.

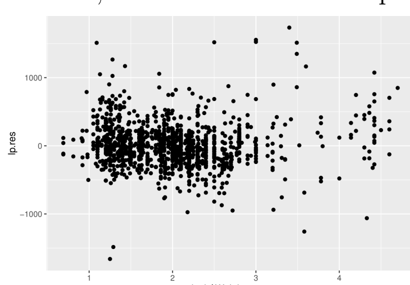
4.2 Question 2

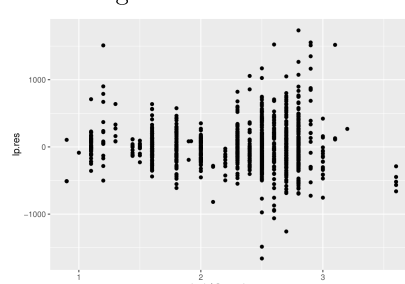
Workflow I first tried to involve all the data I can use to construct a linear model to predict the price of the laptop. Then I will try to diagnose this linear model by examine some residual plot. After that I check the pair relationship between each quantitative variable and try to apply PCA to drop/combine some of the variables in the original linear model.

Analysis The naive linear model performs properly and it really has some explanatory power(because all the variable is linearly combined). Furthermore, the predictive power is good and we have a R^2 near 0.8.

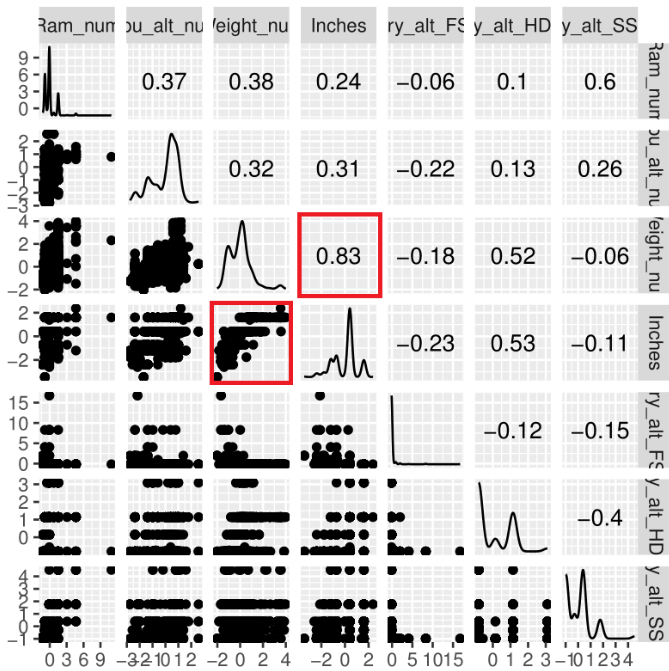
```
Call:
lm(formula = Price_euros ~ Ram_num + Cpu_alt_num + Weight_num +
    Company + TypeName + Inches + ScreenResolution_alt + Memory_alt_FS_num +
    Memory_alt_HDD_num + Memory_alt_SSD_num, data = lp.tb)

Residual standard error: 332.2 on 1260 degrees of freedom
Multiple R-squared:  0.7814,    Adjusted R-squared:  0.7741
F-statistic: 107.2 on 42 and 1260 DF,  p-value: < 2.2e-16
```

I examine some residual plots on different variable, and I find the residual is uniform, which means one level polynomial is good for this model.

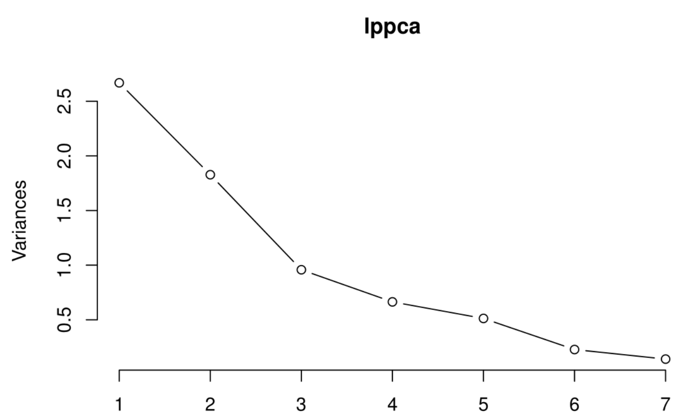




To drop some of the variable to make the model easier and more predictive. I examine the pair relationship and find some of the variable are highly correlated. Such as the Inches and the weight(no doubt a larger screen results in a larger weight)(I have performed scaling already)



Then I perform PCA and find out the first three components have significantly larger variance than the other four, so I choose the first three components as my second model variable.



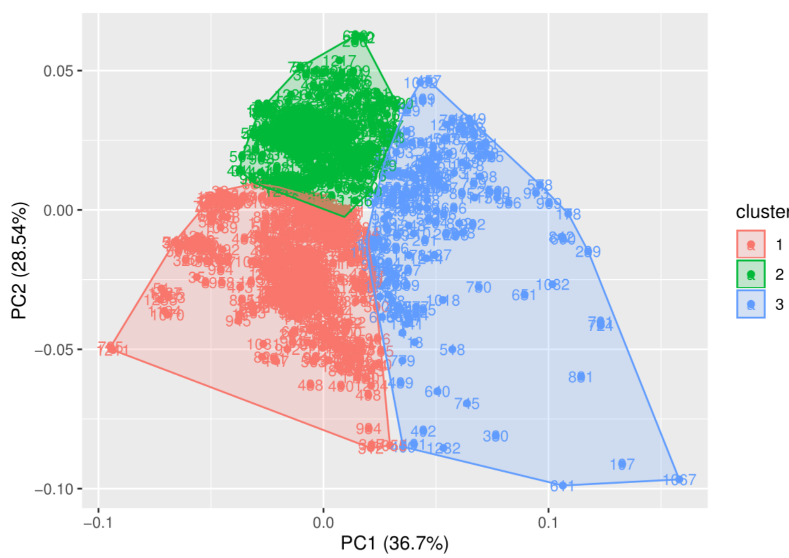
The result is really good with only these three components. The R^2 raises to around 0.84.

```
Call:
lm(formula = Price_euros ~ lppca$x[, 1] + lppca$x[, 2] + lppca$x[,
    3], data = lp.tb)

Residuals:
     Min       1Q   Median       3Q      Max
-1660.20  -162.22    -2.79   140.93  1664.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1123.687      7.747  145.04   <2e-16 ***
lppca$x[, 1]  289.487      4.523   64.00   <2e-16 ***
lppca$x[, 2] -257.480      5.129  -50.20   <2e-16 ***
lppca$x[, 3]  115.115      7.773   14.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 279.7 on 1299 degrees of freedom
Multiple R-squared:  0.8403,    Adjusted R-squared:  0.8399
F-statistic:  2279 on 3 and 1299 DF,  p-value: < 2.2e-16
```

4.3 Question 3

Workflow I carried out scaling before I cluster the laptops through a unsupervised learning method K-means. I visualized the cluster and then perform the Anomaly Detection using the clustering method. Then I visualized to check the outliers.

Analysis From the visualization, we can find that the cluster is not that clear. It seems that the majority of the laptop are in one cluster and the k means algorithm has a great randomness on the cluster result.



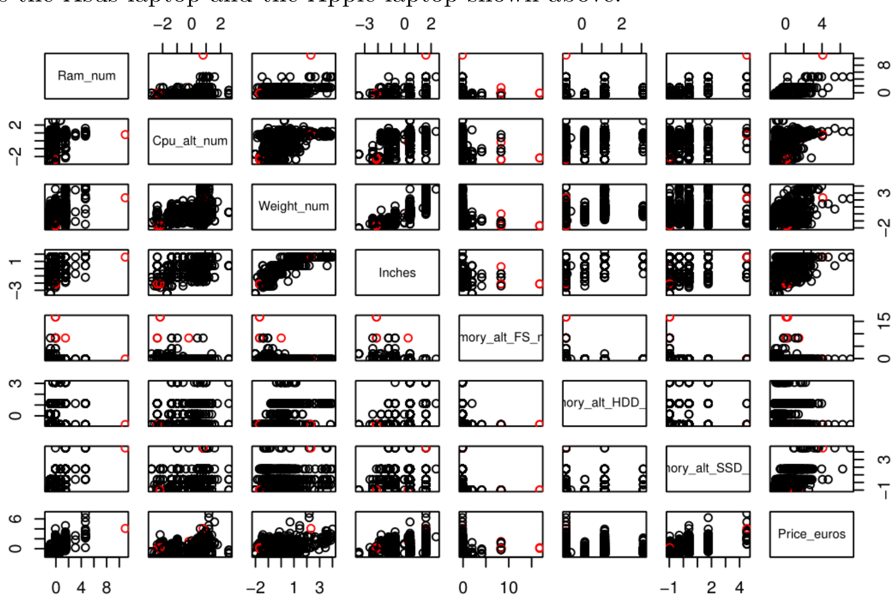
Still, we can still find the outlier of the dataset. I set the distance metric to be

$$Distance = (cluster\_center - scaled\_data)^2$$

The we can find the top-5 outliers, they are:

```
##        X Company     Product  TypeName Inches
## 1:   803   Apple MacBook 12" Ultrabook   12.0
## 2:  1228   Apple MacBook 12" Ultrabook   12.0
## 3:  1081    Asus  ROG G701VO    Gaming   17.3
## 4:     7   Apple MacBook Pro Ultrabook   15.4
## 5:  1211   Apple MacBook 12" Ultrabook   12.0
```

It doesn't surprise us because Apple laptop use different and unique screen resolution, screen inches and tends to use flash storage rather than SSD or HDD. As for the Asus laptop, it has a 64 GB Ram which makes it unique. We can find that the red outliers of Ram and the FS memory from the visualization, which is the Asus laptop and the Apple laptop shown above.



5 Conclusion

We have done a fruitful exploratory data analysis on the laptop dataset and find out the distribution of price for different laptops, an advanced linear model and the outliers in the dataset.