KOLOKIUM PENYELIDIKAN 2021

INSTITUT TADBIRAN AWAM NEGARA (INTAN)

ANALYSIS OF THE MOBILE PHONES PRICES IN MALAYSIA USING WEB SCRAPED DATA

NUR HURRIYATUL HUDA ABDULLAH SANI

DEPARTMENT OF STATISTICS MALAYSIA

PRESENTATION OUTLINE

01 Introduction

Background of the study | Project Focus (Objectives)

02 Methodology

Data Acquisition | Data Preparation | Item Selected | Method of Analysis

03 Results and Discussion

Clustering | Price Dispersion | Regression Analysis

04 Conclusion and Recommendation

Conclusion | Recommendation

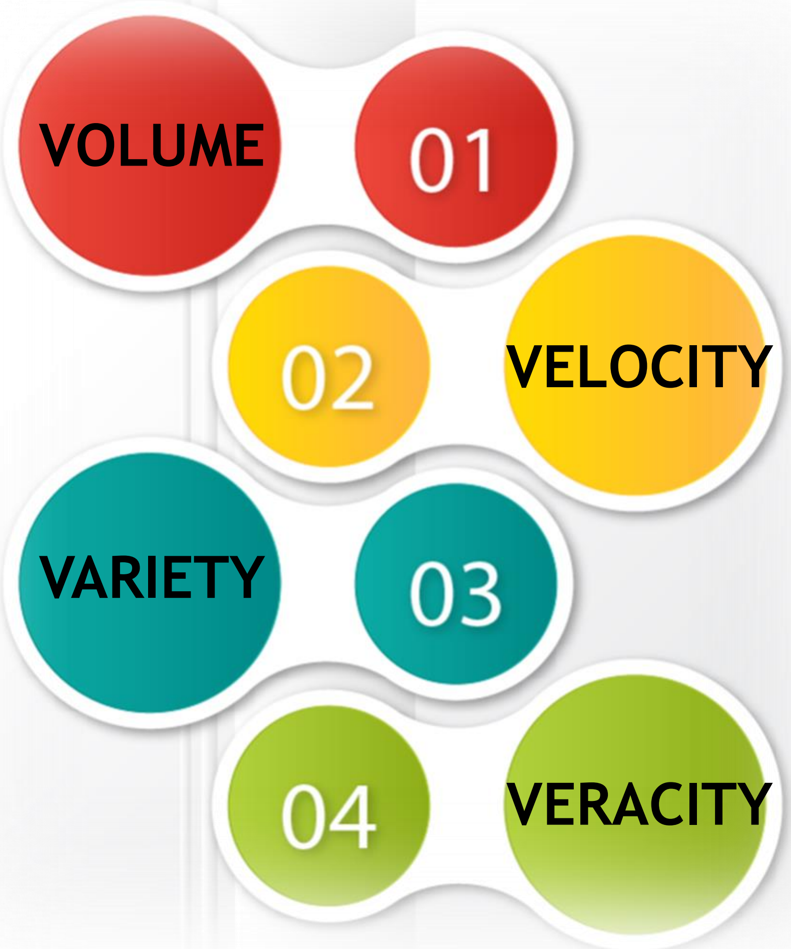
INTRODUCTION

- UN Working Group on Big Data, 2014 explore the utilization of new information sources and technology advancement for the official statistics.

- Concern on the data collections of online price.

- NSOs (UK, US, Korea, Italy, Netherland, Japan, MALAYSIA…) have started to consider the use of online data in official Consumer Price Index (CPI), (Cavallo, 2017).

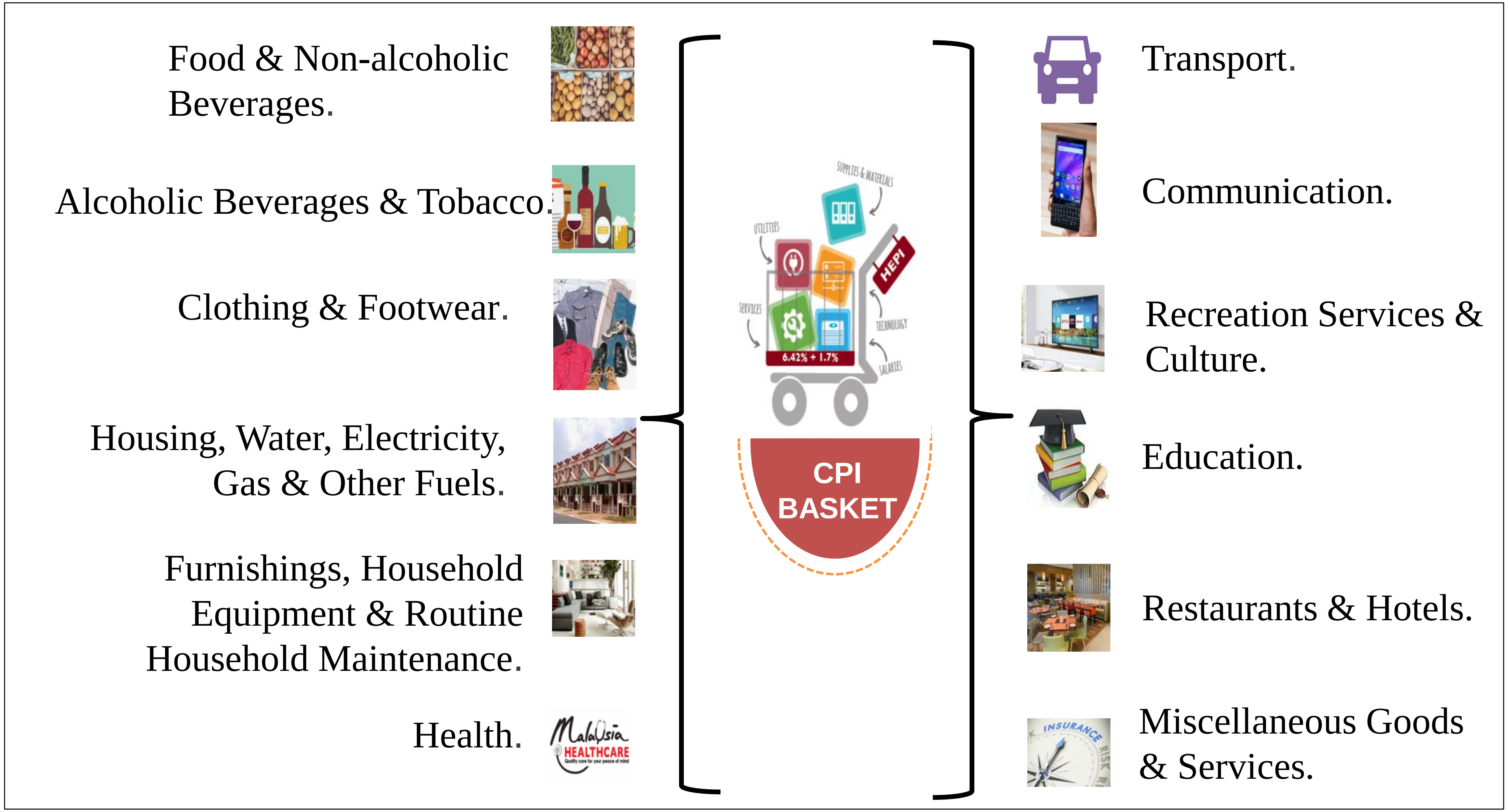
- DOSM (StatsBDA), 2017 has developed Price Intelligence (PI) as an alternative and compliment approach for the data collection method.



- 2.5 quintillion bytes of data produced everyday. 90% data is **unstructured,**(Dobre and Xhafa, 2014)**.**

- Largest source of data is **online data**

INTRODUCTION



Food & Non-alcoholic Beverages.

Alcoholic Beverages & Tobacco.

Clothing & Footwear.

Housing, Water, Electricity, Gas & Other Fuels.

Furnishings, Household Equipment & Routine Household Maintenance.

Health.

Transport.

Communication.

Recreation Services & Culture.

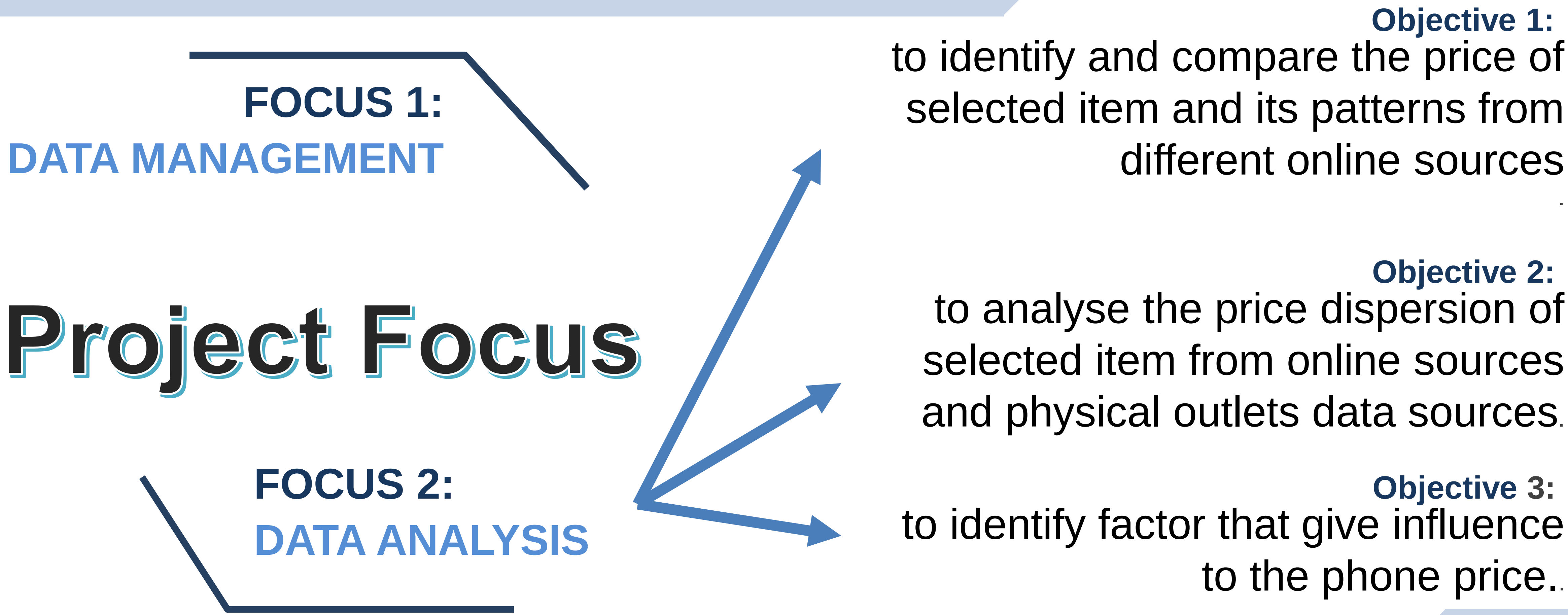
Education.

Restaurants & Hotels.

Miscellaneous Goods & Services.

- The CPI measures the percentage change in price through time in a constant basket of goods and services.
- CPI represents the average pattern of purchases made by a particular population group in a specified time period.
- The price basket is a consumer goods to define the CPI using sample of goods and services available at the consumer market place.
- The goods and services covered in the price "basket" are broadly classified using 12 groups in COICOP

INTRODUCTION



FOCUS 1:
DATA MANAGEMENT

Project Focus

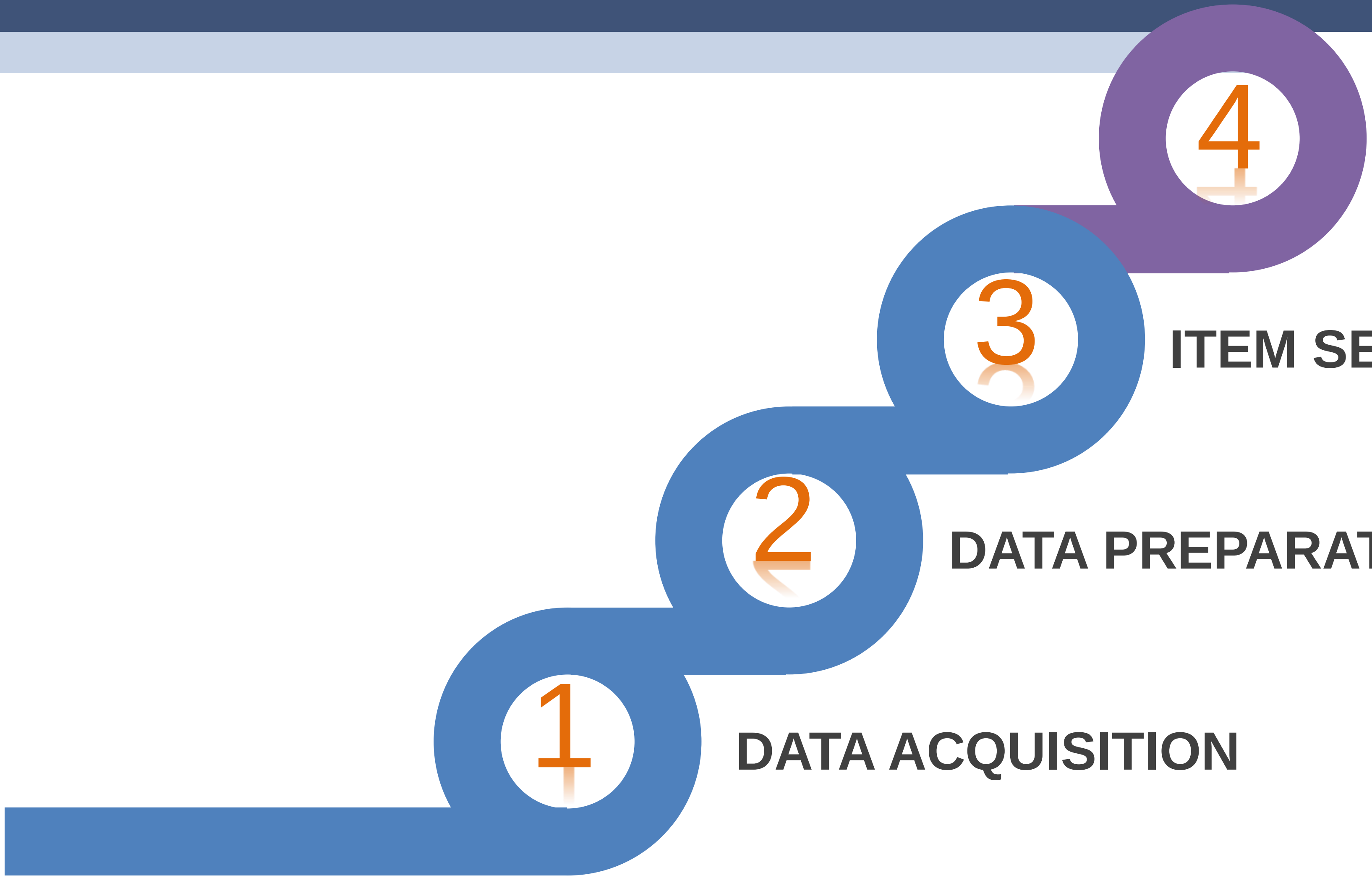
FOCUS 2:
DATA ANALYSIS

Objective 1:
to identify and compare the price of selected item and its patterns from different online sources
.

Objective 2:
to analyse the price dispersion of selected item from online sources and physical outlets data sources.

Objective 3:
to identify factor that give influence to the phone price..

METHODOLOGY



1 DATA ACQUISITION

2 DATA PREPARATION

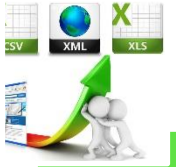
3 ITEM SELECTION

4 METHOD OF ANALYSIS

METHODOLOGY : **Data Acquisition**

DATA SOURCE

- DOSM STATSBDA
- 4 WEBSITES
- 176K price quotation
- 2,800 unique items
- 5.2M data
- Manual Price Data Collection (physical outlet)
- Jan – Feb 2018



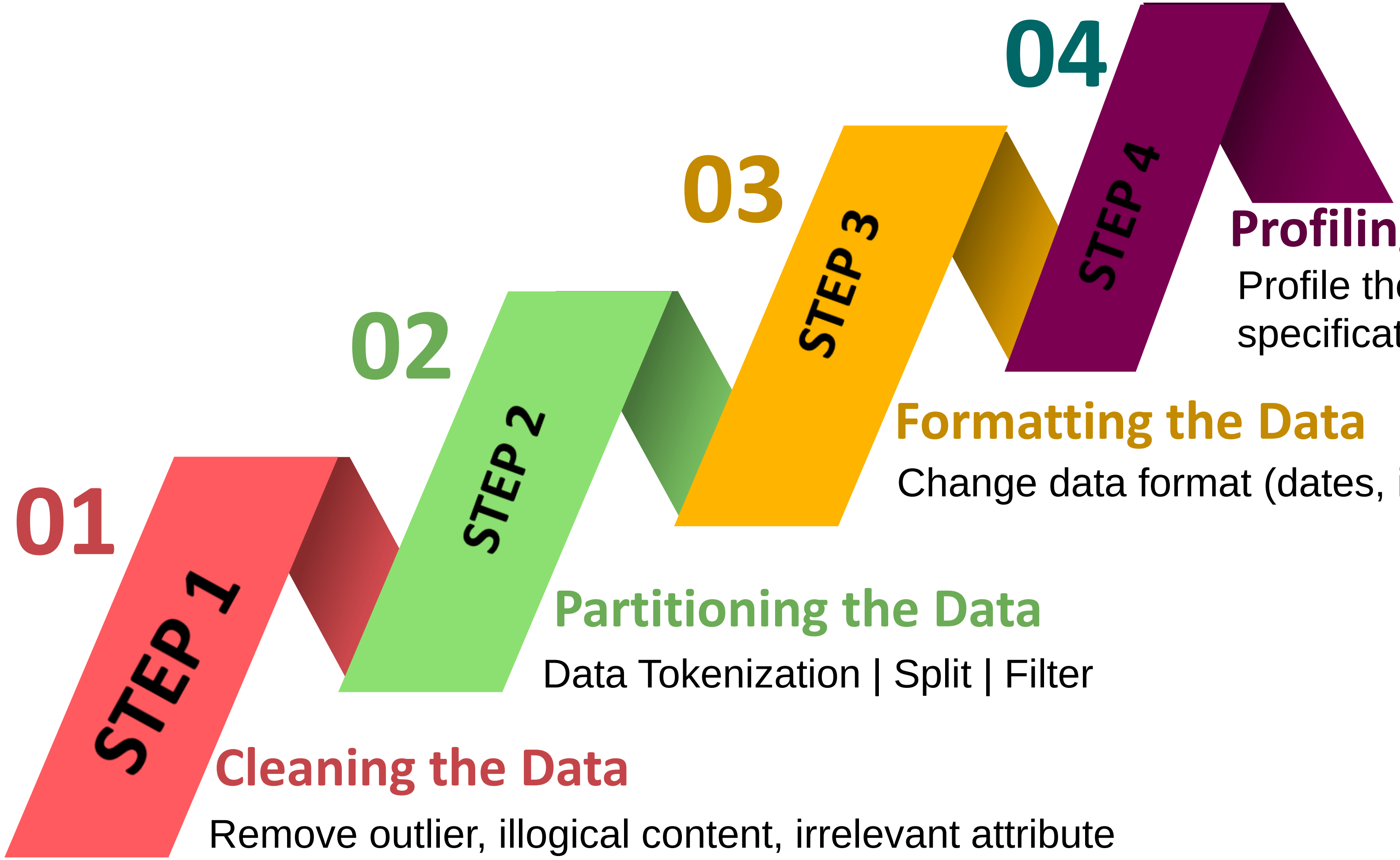
TYPE OF DATA

- SEMI STRUCTURED

TOOL

- Web scrap using **python selenium**
- Data Preparation using python
- Data Analysis using R

METHODOLOGY : **Data Preparation**



01

STEP 1

Cleaning the Data

Remove outlier, illogical content, irrelevant attribute

02

STEP 2

Partitioning the Data

Data Tokenization | Split | Filter

03

STEP 3

Formatting the Data

Change data format (dates, integer, string)

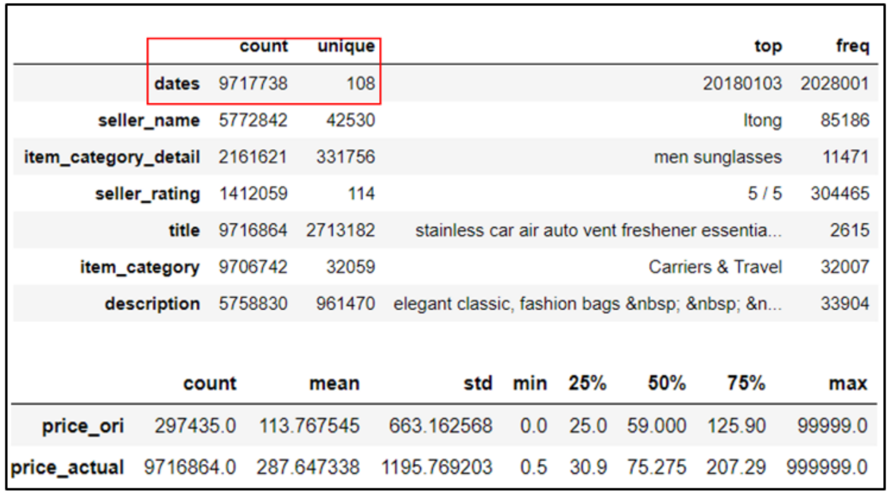
04

STEP 4

Profiling the Data

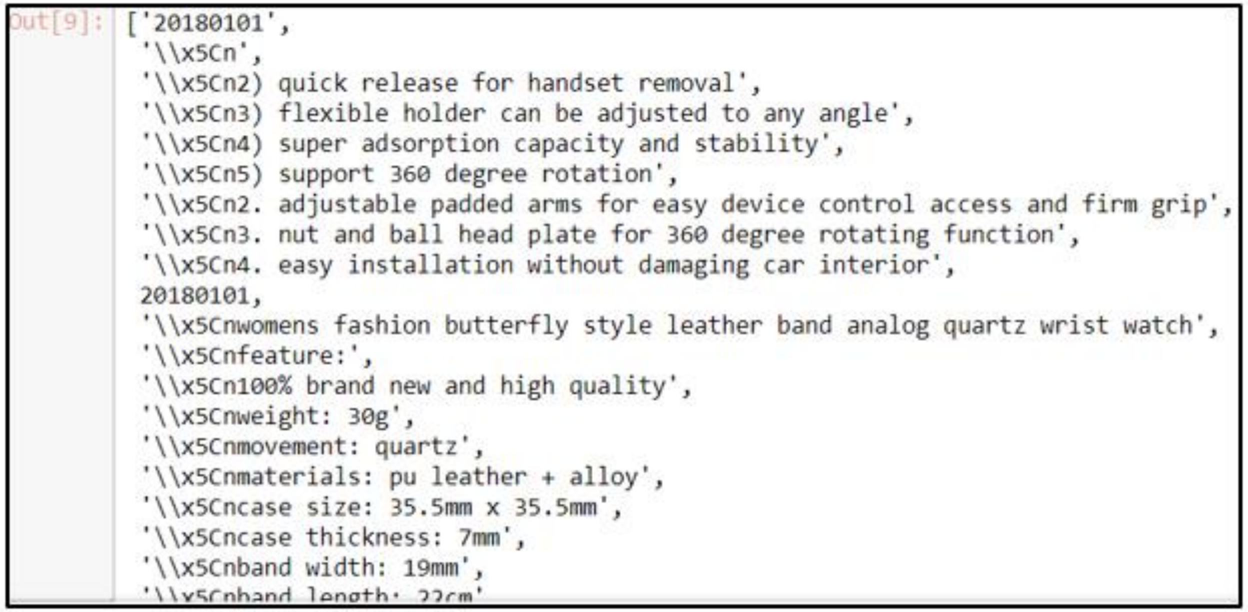
Profile the data into category, brand, other item specification

STEP 1: Cleaning the Data



	count	unique	top	freq
dates	9717738	108	20180103	2028001
seller_name	5772842	42530	ltong	85186
item_category_detail	2161621	331756	men sunglasses	11471
seller_rating	1412059	114	5 / 5	304465
title	9716864	2713182	stainless car air auto vent freshener essentia...	2615
item_category	9706742	32059	Carriers & Travel	32007
description	5758830	961470	elegant classic, fashion bags &n...	33904

	count	mean	std	min	25%	50%	75%	max
price_ori	297435.0	113.767545	663.162568	0.0	25.0	59.000	125.90	99999.0
price_actual	9716864.0	287.647338	1195.769203	0.5	30.9	75.275	207.29	999999.0

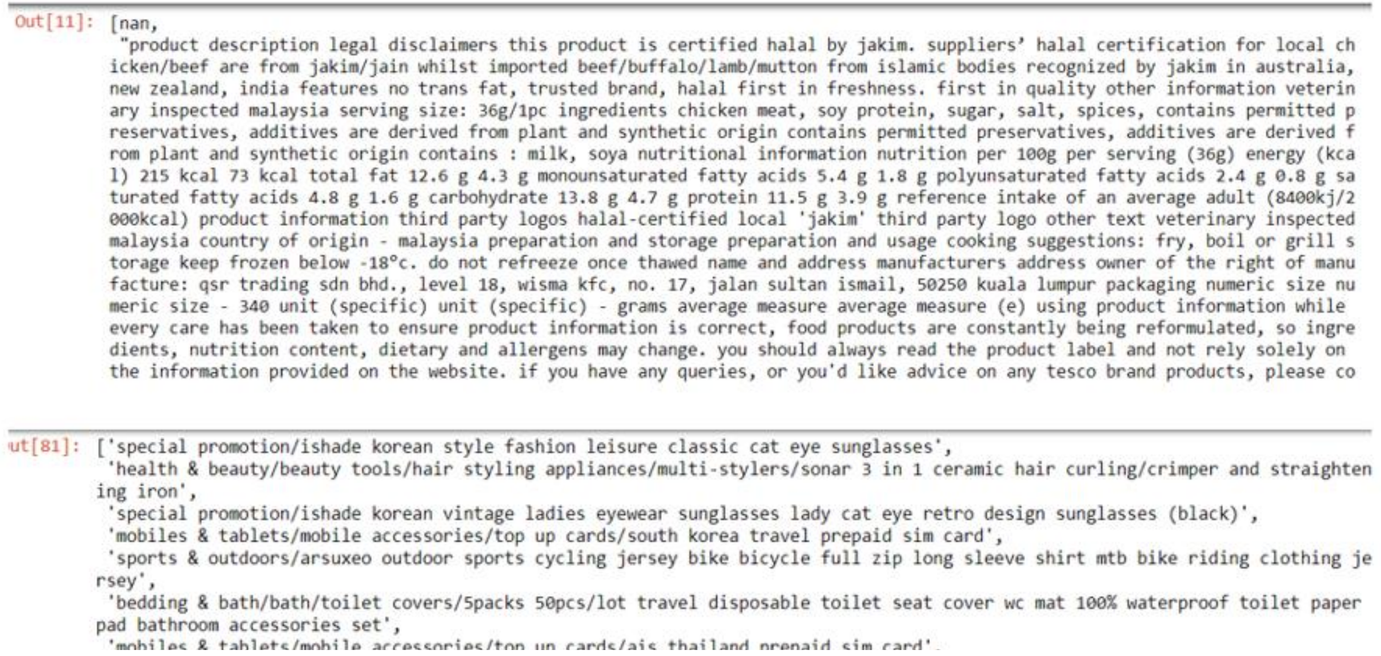


```
Out[9]: ['20180101',
          '\\x5Cn',
          '\\x5Cn2) quick release for handset removal',
          '\\x5Cn3) flexible holder can be adjusted to any angle',
          '\\x5Cn4) super adsorption capacity and stability',
          '\\x5Cn5) support 360 degree rotation',
          '\\x5Cn2. adjustable padded arms for easy device control access and firm grip',
          '\\x5Cn3. nut and ball head plate for 360 degree rotating function',
          '\\x5Cn4. easy installation without damaging car interior',
          20180101,
          '\\x5Cnwomens fashion butterfly style leather band analog quartz wrist watch',
          '\\x5Cnfeature:',
          '\\x5Cn100% brand new and high quality',
          '\\x5Cnweight: 30g',
          '\\x5Cnmovement: quartz',
          '\\x5Cnmaterials: pu leather + alloy',
          '\\x5Cncase size: 35.5mm x 35.5mm',
          '\\x5Cncase thickness: 7mm',
          '\\x5Cnband width: 19mm',
          '\\x5Cnband length: 22cm'
```

Remove outlier, illogical content and irrelevant attribute:

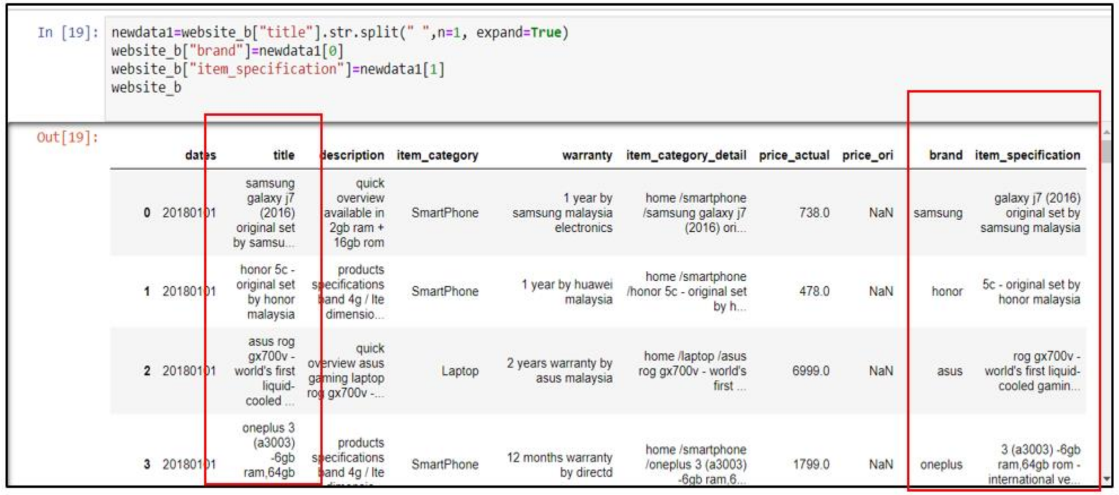
- Using `describe ()`, `unique ()` function to identify the data summary/ content
- Using `isin()` function to delete the unwanted content.

STEP 2: Partitioning the Data



```
Out[11]: [nan,
 "product description legal disclaimers this product is certified halal by jakim. suppliers' halal certification for local ch
icken/beef are from jakim/jain whilst imported beef/buffalo/lamb/mutton from islamic bodies recognized by jakim in australia,
new zealand, india features no trans fat, trusted brand, halal first in freshness. first in quality other information veterin
ary inspected malaysia serving size: 36g/1pc ingredients chicken meat, soy protein, sugar, salt, spices, contains permitted p
reservatives, additives are derived from plant and synthetic origin contains permitted preservatives, additives are derived f
rom plant and synthetic origin contains : milk, soya nutritional information nutrition per 100g per serving (36g) energy (kca
l) 215 kcal 73 kcal total fat 12.6 g 4.3 g monounsaturated fatty acids 5.4 g 1.8 g polyunsaturated fatty acids 2.4 g 0.8 g sa
turated fatty acids 4.8 g 1.6 g carbohydrate 13.8 g 4.7 g protein 11.5 g 3.9 g reference intake of an average adult (8400kj/2
000kcal) product information third party logos halal-certified local 'jakim' third party logo other text veterinary inspected
malaysia country of origin - malaysia preparation and storage preparation and usage cooking suggestions: fry, boil or grill s
torage keep frozen below -18°c. do not refreeze once thawed name and address manufacturers address owner of the right of manu
facture: qsr trading sdn bhd., level 18, wisma kfc, no. 17, jalan sultan ismail, 50250 kuala lumpur packaging numeric size nu
meric size - 340 unit (specific) unit (specific) - grams average measure average measure (e) using product information while
every care has been taken to ensure product information is correct, food products are constantly being reformulated, so ingre
dients, nutrition content, dietary and allergens may change. you should always read the product label and not rely solely on
the information provided on the website. if you have any queries, or you'd like advice on any tesco brand products, please co
```

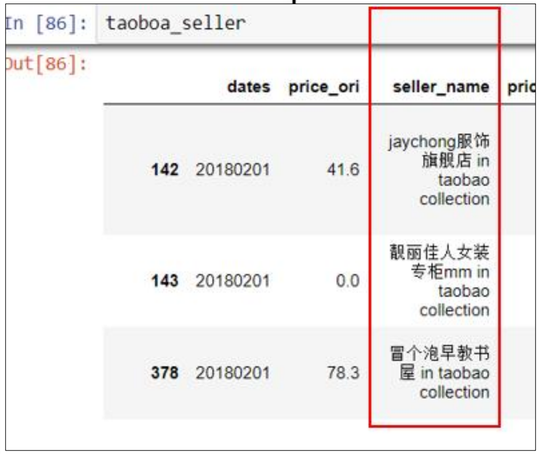
```
ut[81]: ['special promotion/ishade korean style fashion leisure classic cat eye sunglasses',
 'health & beauty/beauty tools/hair styling appliances/multi-stylers/sonar 3 in 1 ceramic hair curling/crimper and straighten
ing iron',
 'special promotion/ishade korean vintage ladies eyewear sunglasses lady cat eye retro design sunglasses (black)',
 'mobiles & tablets/mobile accessories/top up cards/south korea travel prepaid sim card',
 'sports & outdoors/arsuxeo outdoor sports cycling jersey bike bicycle full zip long sleeve shirt mtb bike riding clothing je
rsey',
 'bedding & bath/bath/toilet covers/5packs 50pcs/lot travel disposable toilet seat cover wc mat 100% waterproof toilet paper
pad bathroom accessories set',
 'mobiles & tablets/mobile accessories/top up cards/ais thailand prepaid sim card',
```



```
In [19]: newdata1=website_b["title"].str.split(" ",n=1, expand=True)
         website_b["brand"]=newdata1[0]
         website_b["item_specification"]=newdata1[1]
         website_b
```

Out[19]:

	dates	title	description	item_category	warranty	item_category_detail	price_actual	price_ori	brand	item_specification
0	20180101	samsung galaxy j7 (2016) original set by samsu...	quick overview available in 2gb ram + 16gb rom	SmartPhone	1 year by samsung malaysia electronics	home /smartphone /samsung galaxy j7 (2016) ori...	738.0	NaN	samsung	galaxy j7 (2016) original set by samsung malaysia
1	20180101	honor 5c - original set by honor malaysia	products specifications band 4g / lte dimensio...	SmartPhone	1 year by huawei malaysia	home /smartphone /honor 5c - original set by h...	478.0	NaN	honor	5c - original set by honor malaysia
2	20180101	asus rog gx700v - world's first liquid-cooled ...	quick overview asus gaming laptop rog gx700v -...	Laptop	2 years warranty by asus malaysia	home /laptop /asus rog gx700v - world's first ...	6999.0	NaN	asus	rog gx700v - world's first liquid-cooled gamin...
3	20180101	oneplus 3 (a3003) -6gb ram,64gb	products specifications band 4g / lte	SmartPhone	12 months warranty by directd	home /smartphone /oneplus 3 (a3003) -6gb ram,6...	1799.0	NaN	oneplus	3 (a3003) -6gb ram,64gb rom - international ve...



```
In [86]: taoboa_seller
```

Out[86]:

	dates	price_ori	seller_name	pric
142	20180201	41.6	jaychong服饰旗舰店 in taobao collection	
143	20180201	0.0	靓丽佳人女装专柜mm in taobao collection	
378	20180201	78.3	冒个泡早教书屋 in taobao collection	

The long text description need to be partitioned/ split before the product can be categorized into proper category:

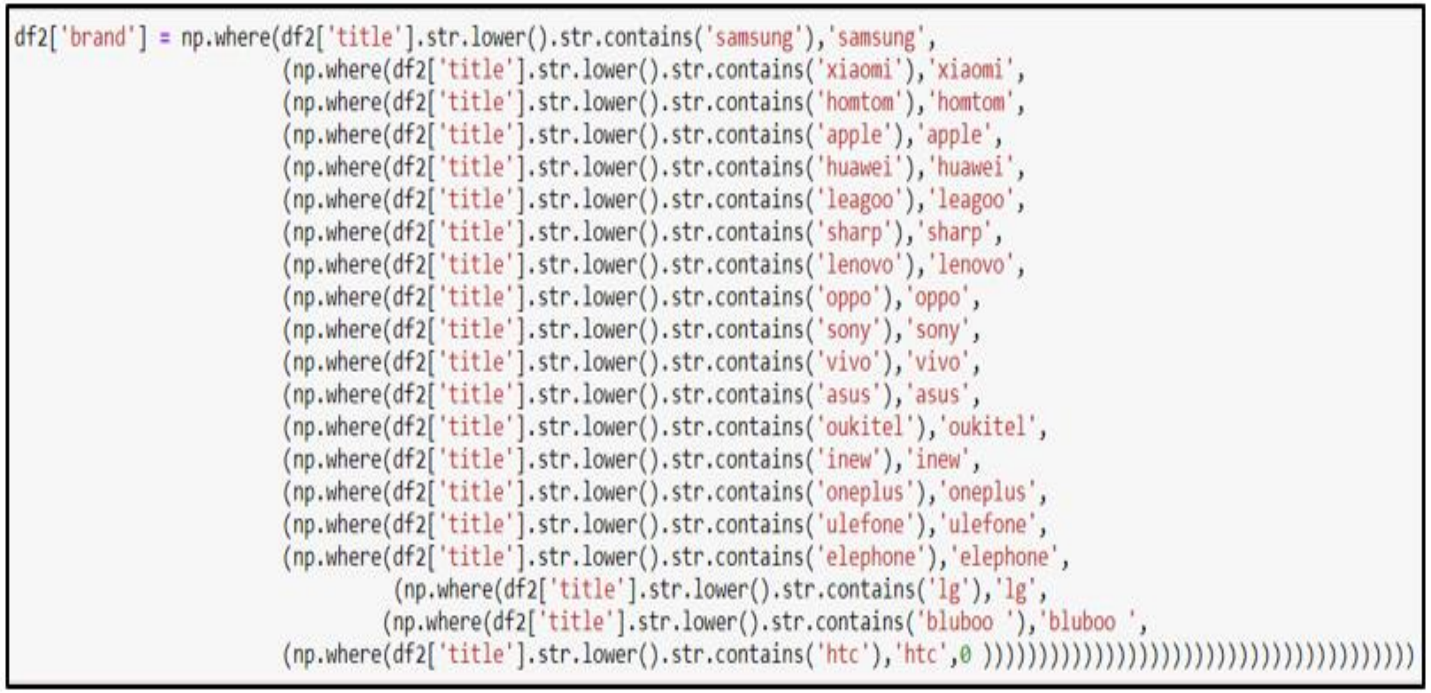
- Using `split()` function to breaks-up the string at the specific separator and then returns a list of strings.
- Using `str.cont ()` function to filter the specific word to specify the data
- Using `np.where()` function to partitioning the data according to the item specification (e.g. brand, model, shipping, warranty, etc.)

STEP 3: Formatting the Data

The data type function is used to identify the data format for each attribute, and the format can be changed if necessary.

In this data set, attribute ‘dates’ been assign as ‘object’, therefore it has been changed to date and time format using `datetime()` function.

STEP 4: Profiling and Finding Misclassification Data



```
df2['brand'] = np.where(df2['title'].str.lower().str.contains('samsung'),'samsung',
                        (np.where(df2['title'].str.lower().str.contains('xiaomi'),'xiaomi',
                        (np.where(df2['title'].str.lower().str.contains('homtom'),'homtom',
                        (np.where(df2['title'].str.lower().str.contains('apple'),'apple',
                        (np.where(df2['title'].str.lower().str.contains('huawei'),'huawei',
                        (np.where(df2['title'].str.lower().str.contains('leagoo'),'leagoo',
                        (np.where(df2['title'].str.lower().str.contains('sharp'),'sharp',
                        (np.where(df2['title'].str.lower().str.contains('lenovo'),'lenovo',
                        (np.where(df2['title'].str.lower().str.contains('oppo'),'oppo',
                        (np.where(df2['title'].str.lower().str.contains('sony'),'sony',
                        (np.where(df2['title'].str.lower().str.contains('vivo'),'vivo',
                        (np.where(df2['title'].str.lower().str.contains('asus'),'asus',
                        (np.where(df2['title'].str.lower().str.contains('oukitel'),'oukitel',
                        (np.where(df2['title'].str.lower().str.contains('inew'),'inew',
                        (np.where(df2['title'].str.lower().str.contains('oneplus'),'oneplus',
                        (np.where(df2['title'].str.lower().str.contains('ulefone'),'ulefone',
                        (np.where(df2['title'].str.lower().str.contains('elephone'),'elephone',
                                  (np.where(df2['title'].str.lower().str.contains('lg'),'lg',
                                 (np.where(df2['title'].str.lower().str.contains('bluboo '),'bluboo ',
                        (np.where(df2['title'].str.lower().str.contains('htc'),'htc',0 )))))))))))))))))))))))))))))))))))))))
```

- Using `np.where()` function to profile the data according to the item specification (e.g. brand, model, shipping, warranty, etc.).

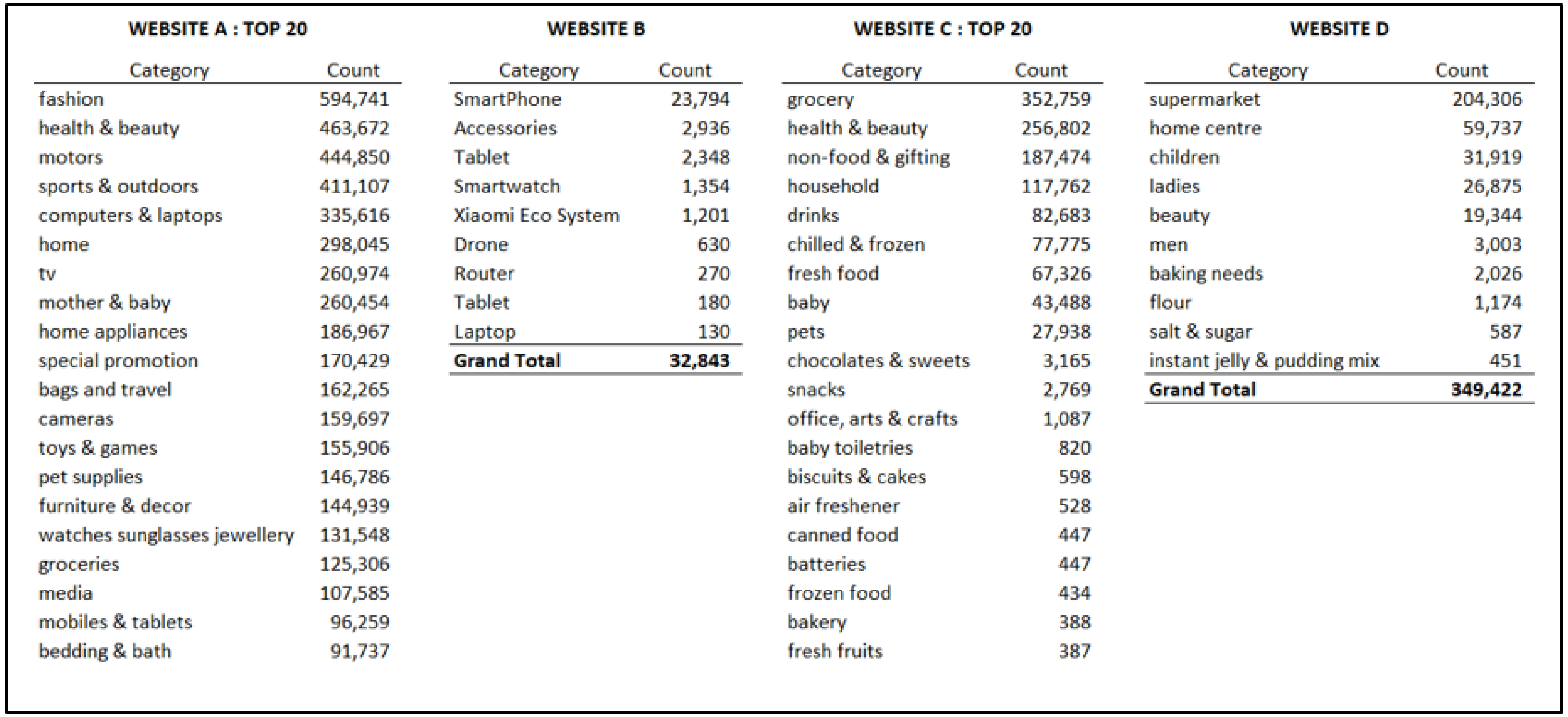
- Using `unique ()` function and plot the data to identify the potential misclassification of the data.
- For example, ‘apple’ can be misclassified under electronics category which are either iPhone or mac book which also carry the same name of ‘apple’. It also can be classified into fruit, beverages and also fashion category as there exist woman jeans with brand of ‘apple mint’.
- Finalise the item category after data profiling because could be happened between different categories. For example, Samsung accessories (phone case, keypad, earphone, etc) can still falls into mobiles category instead of accessories

STEP 4: Profiling the Data

After initial cleaning process, the following results are obtained.

Table 2.2: Number of unique item category and seller of four selected websites

Website	Number of unique item category	Number of unique seller
Website A	2,890	17,366
Website B	9	1
Website C	49	1
Website D	10	1



WEBSITE A : TOP 20

Category	Count
fashion	594,741
health & beauty	463,672
motors	444,850
sports & outdoors	411,107
computers & laptops	335,616
home	298,045
tv	260,974
mother & baby	260,454
home appliances	186,967
special promotion	170,429
bags and travel	162,265
cameras	159,697
toys & games	155,906
pet supplies	146,786
furniture & decor	144,939
watches sunglasses jewellery	131,548
groceries	125,306
media	107,585
mobiles & tablets	96,259
bedding & bath	91,737

WEBSITE B

Category	Count
SmartPhone	23,794
Accessories	2,936
Tablet	2,348
Smartwatch	1,354
Xiaomi Eco System	1,201
Drone	630
Router	270
Tablet	180
Laptop	130
Grand Total	**32,843**

WEBSITE C : TOP 20

Category	Count
grocery	352,759
health & beauty	256,802
non-food & gifting	187,474
household	117,762
drinks	82,683
chilled & frozen	77,775
fresh food	67,326
baby	43,488
pets	27,938
chocolates & sweets	3,165
snacks	2,769
office, arts & crafts	1,087
baby toiletries	820
biscuits & cakes	598
air freshener	528
canned food	447
batteries	447
frozen food	434
bakery	388
fresh fruits	387

WEBSITE D

Category	Count
supermarket	204,306
home centre	59,737
children	31,919
ladies	26,875
beauty	19,344
men	3,003
baking needs	2,026
flour	1,174
salt & sugar	587
instant jelly & pudding mix	451
Grand Total	**349,422**

Figure 2.14: Item category of four selected websites

METHODOLOGY : Item Selection

Five Selection Criteria :

Selected item: Mobile Phones from Websites A and B

1. Item contained in CPI basket
2. The item found in at least from two different website
3. The availability of the item by date and days (less missing days)
4. Popularity of the most purchased item online and relevant to issue in the country
5. Item that can replicate the data acquisition and preparation process against other item categories

METHODOLOGY : **Method of Analysis**

3 Analysis were conducted in this project to identify the price pattern :

- ***k*-mean clustering**

Clustering is a grouping of data that share similar features together in the same group. *K*-mean is one of the most commonly used

- **t-test**

The *t*-test compares two averages (means) and tells if the observations are different from each other. The *t*-test also tells how significant the differences are.

- **Regression analysis**

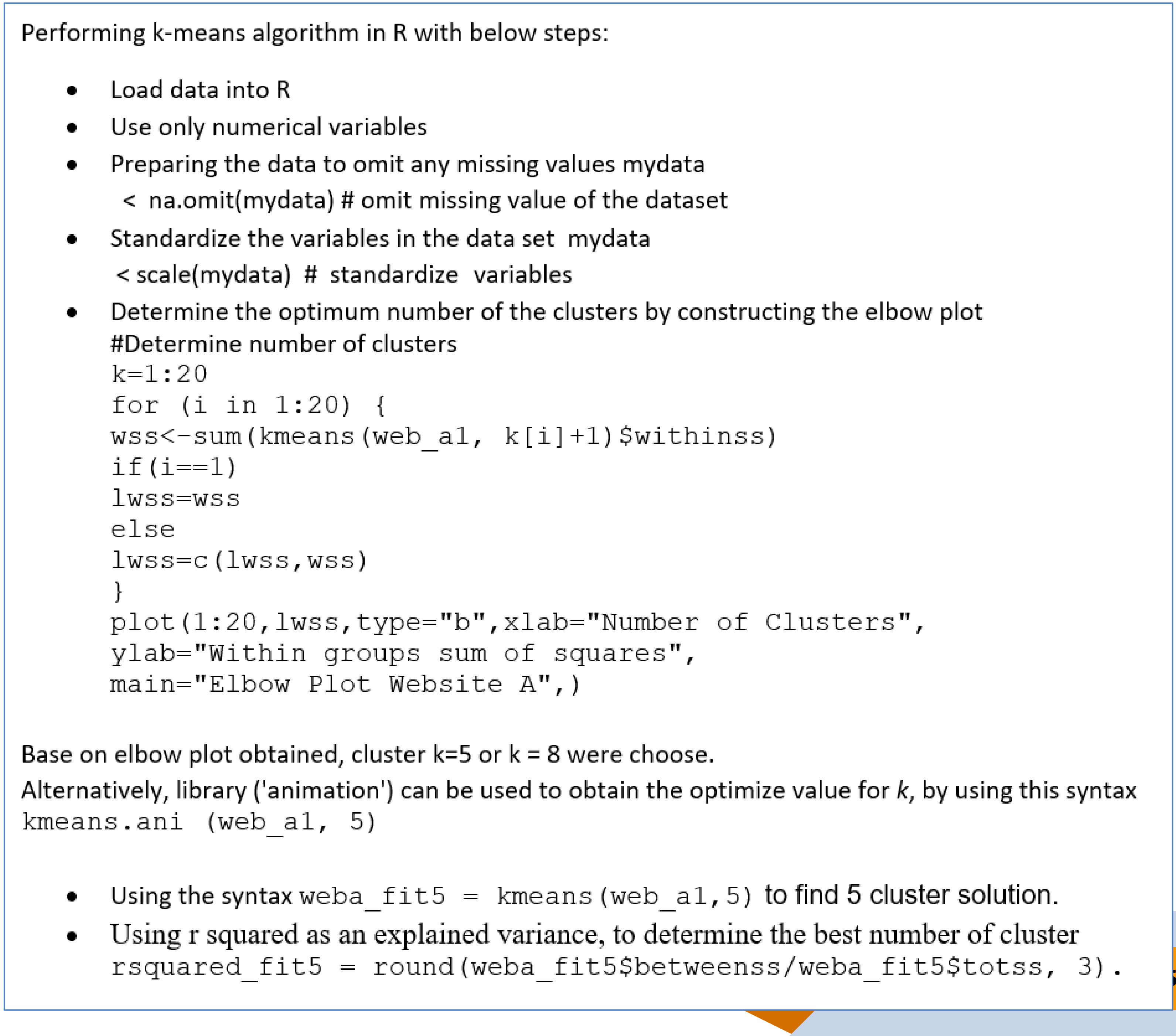
Regression analysis is used to examine the relationship between two or more variables of interest. It is used to examine the influence of one or more independent variables (predictor variable) on a dependent variable (response variable)

k-mean clustering

- The *K*-means clustering algorithm is used to find groups which have not been explicitly labelled in the data.

- *K*-mean clustering was carried out to get a rough idea of the mobile phone group that is in the market based on price, specification and brand.

- Why *K*-mean ?
  - ✓ most popular algorithms used for clustering practice because of its simplicity and speed;
  - ✓ Can be applied to large size of dataset that has a small number of dimensions, numeric, and continuous.
  - ✓ Commonly use for market segmentation (customer and products)



Performing k-means algorithm in R with below steps:

- Load data into R
- Use only numerical variables
- Preparing the data to omit any missing values mydata
  < na.omit(mydata) # omit missing value of the dataset
- Standardize the variables in the data set mydata
  < scale(mydata) # standardize variables
- Determine the optimum number of the clusters by constructing the elbow plot

```
#Determine number of clusters
k=1:20
for (i in 1:20) {
wss<-sum(kmeans(web_a1, k[i]+1)$withinss)
if(i==1)
lwss=wss
else
lwss=c(lwss,wss)
}
plot(1:20,lwss,type="b",xlab="Number of Clusters",
ylab="Within groups sum of squares",
main="Elbow Plot Website A",)
```

Base on elbow plot obtained, cluster k=5 or k = 8 were choose.

Alternatively, library ('animation') can be used to obtain the optimize value for *k*, by using this syntax `kmeans.ani (web_a1, 5)`

- Using the syntax `weba_fit5 = kmeans(web_a1,5)` to find 5 cluster solution.
- Using r squared as an explained variance, to determine the best number of cluster `rsquared_fit5 = round(weba_fit5$betweenss/weba_fit5$totss, 3).`

t-test

- Using independent t-test to understand if there is price difference between website A and B and also from both websites A and B compare to average price collected via physical outlet.

- Since the number of samples in each group is different, and the variance of the two data sets is also different, the unequal variance t-test (Welch's t-test) is used.

- The null hypothesis for the independent t-test: H_0: $u_1 = u_2$
- The alternative hypothesis: H_A: $u_1 \neq u_2$

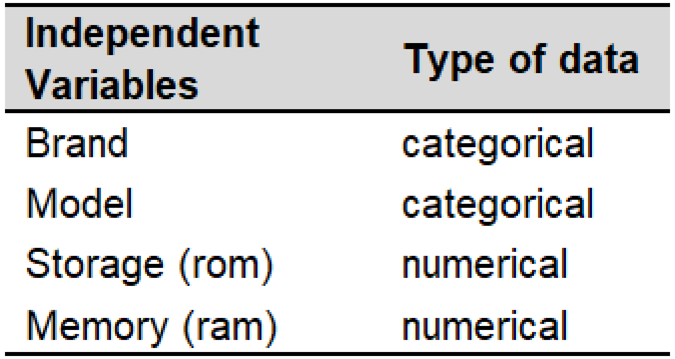
- The following syntax is used for t-test analysis using R.

```
>t.test(apple_a,apple_b)
>Var(apple_a$price_a)
```

Regression analysis

- Regression analysis is used to understand the factor that give influence to the phone price.

- Dependent variable is **Prices**
- Four independent used as below:



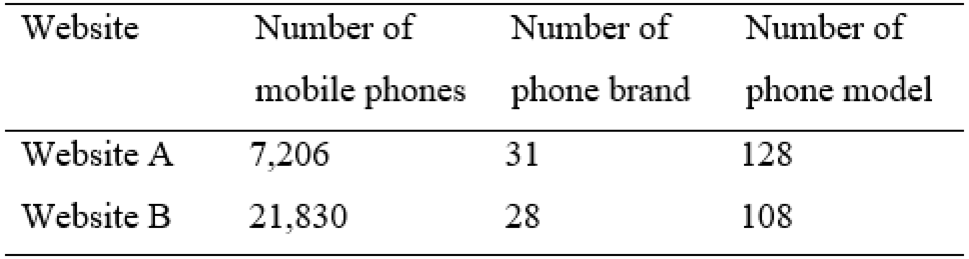
Independent Variables	Type of data
Brand	categorical
Model	categorical
Storage (rom)	numerical
Memory (ram)	numerical

- Below syntax is used to perform regression analysis:

```
>j1 = lm(data=web_b,price_actual~-1+factor(model)+storage + memory)
>summary(j1)
>anova(j1)
```

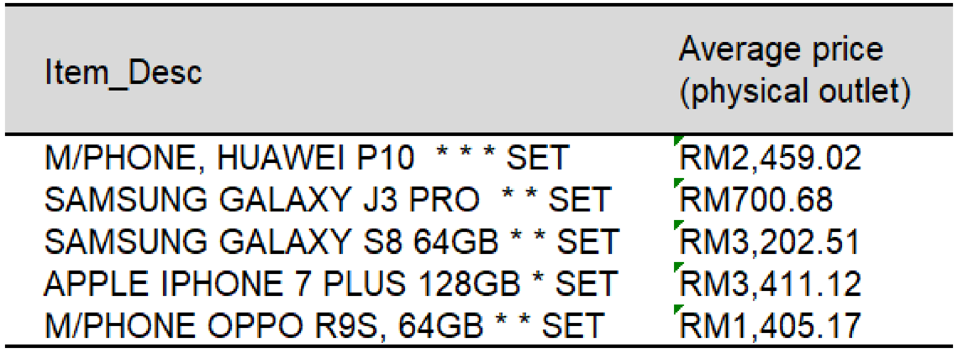
RESULTS & DISCUSSION

Selected item for this project is mobile phones from websites A and B.
Below results were obtained from both websites.



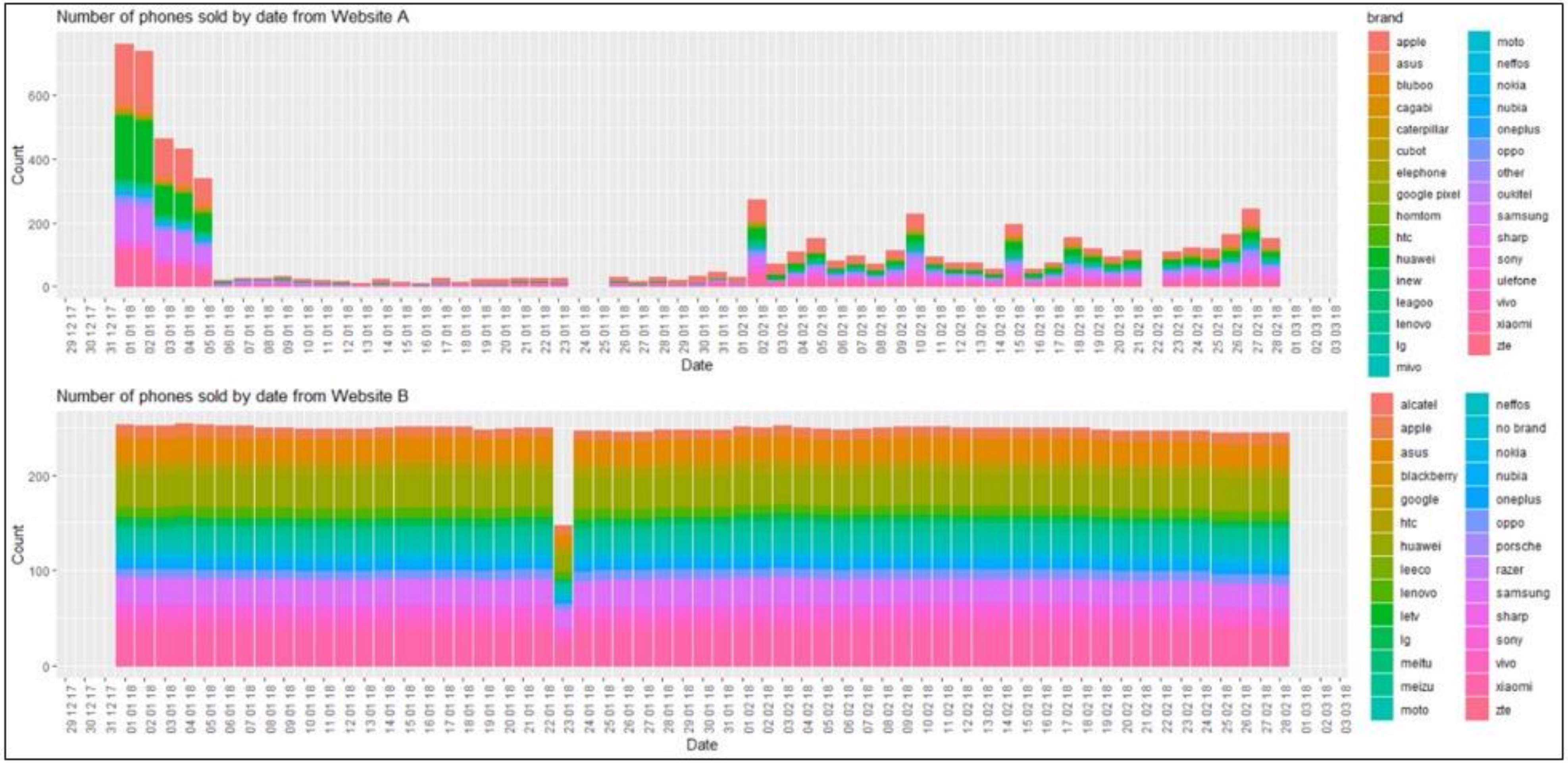
Website	Number of mobile phones	Number of phone brand	Number of phone model
Website A	7,206	31	128
Website B	21,830	28	108

Average phone price from physical outlet data collection for month of January and February 2018 is as below:



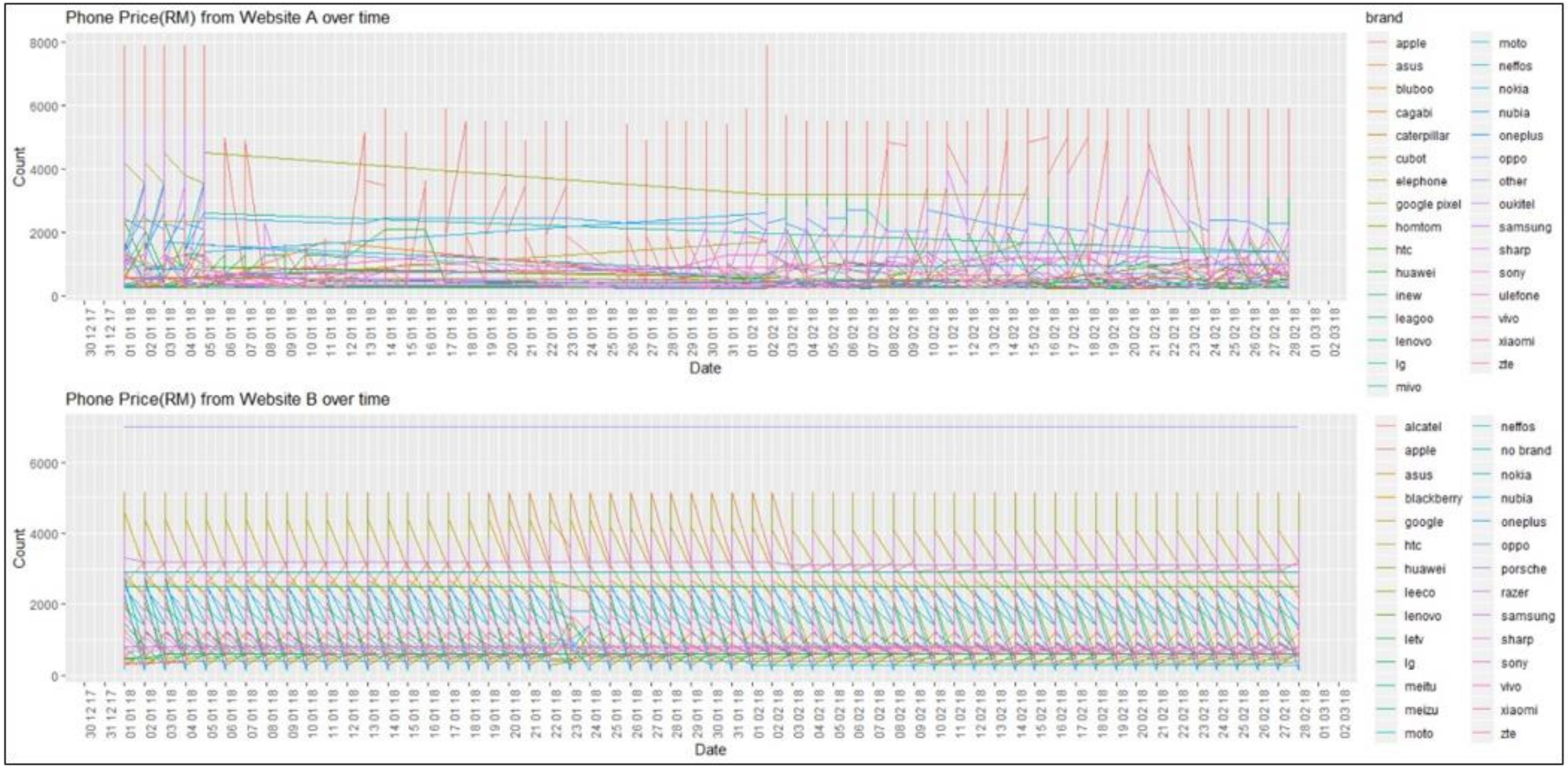
Item_Desc	Average price (physical outlet)
M/PHONE, HUAWEI P10 * * * SET	RM2,459.02
SAMSUNG GALAXY J3 PRO * * SET	RM700.68
SAMSUNG GALAXY S8 64GB * * SET	RM3,202.51
APPLE IPHONE 7 PLUS 128GB * SET	RM3,411.12
M/PHONE OPPO R9S, 64GB * * SET	RM1,405.17

RESULTS & DISCUSSION



Number of phones offer by date from website A and website B

RESULTS & DISCUSSION

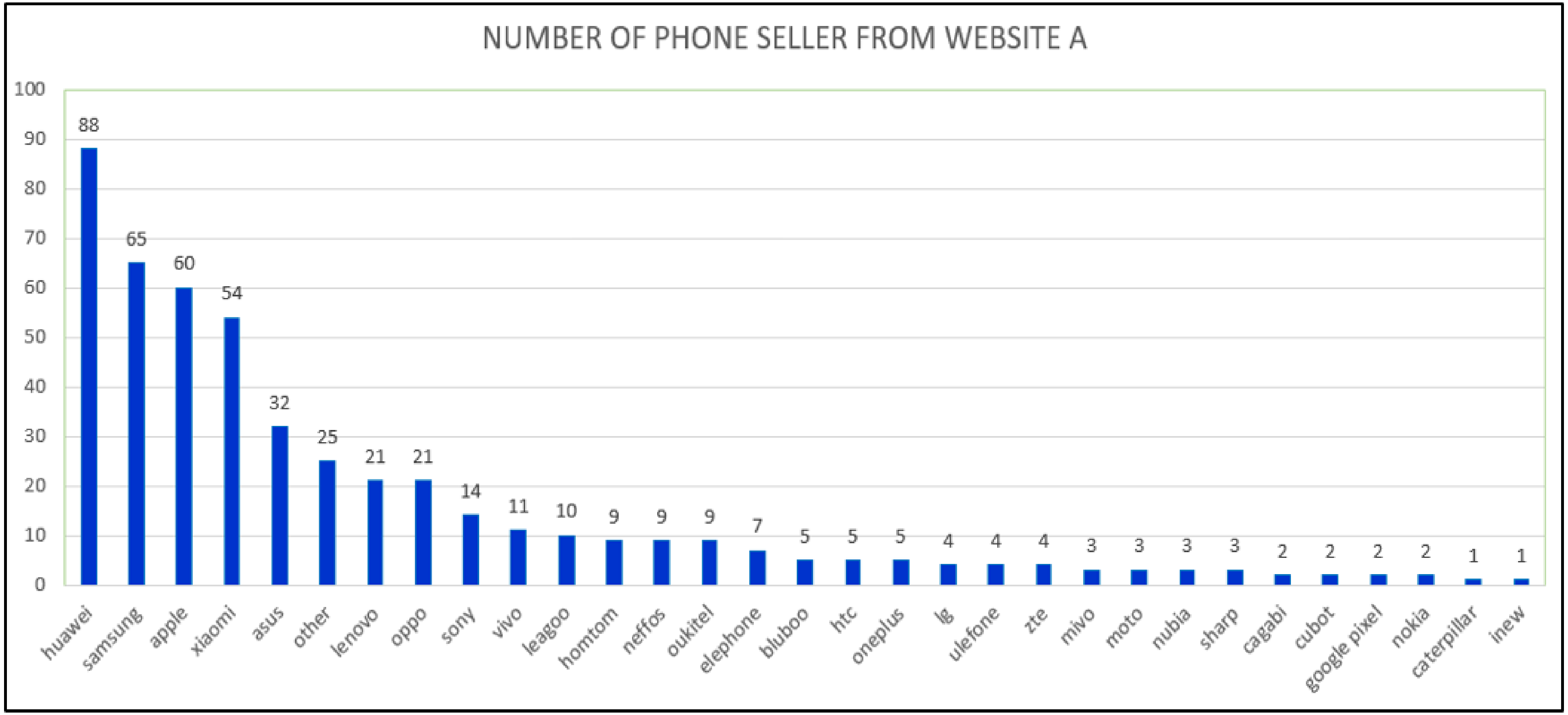


Phone Price(RM) from Website A over time

Phone Price(RM) from Website B over time

Daily phones price offer by website A and website B

RESULTS & DISCUSSION

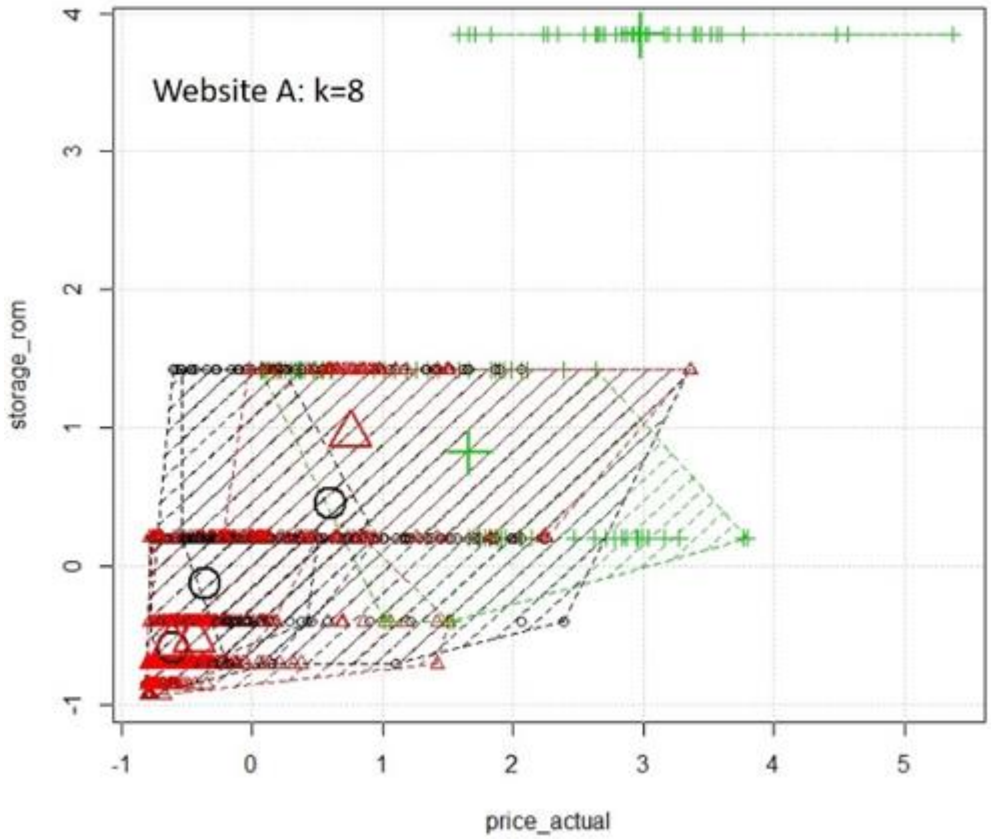


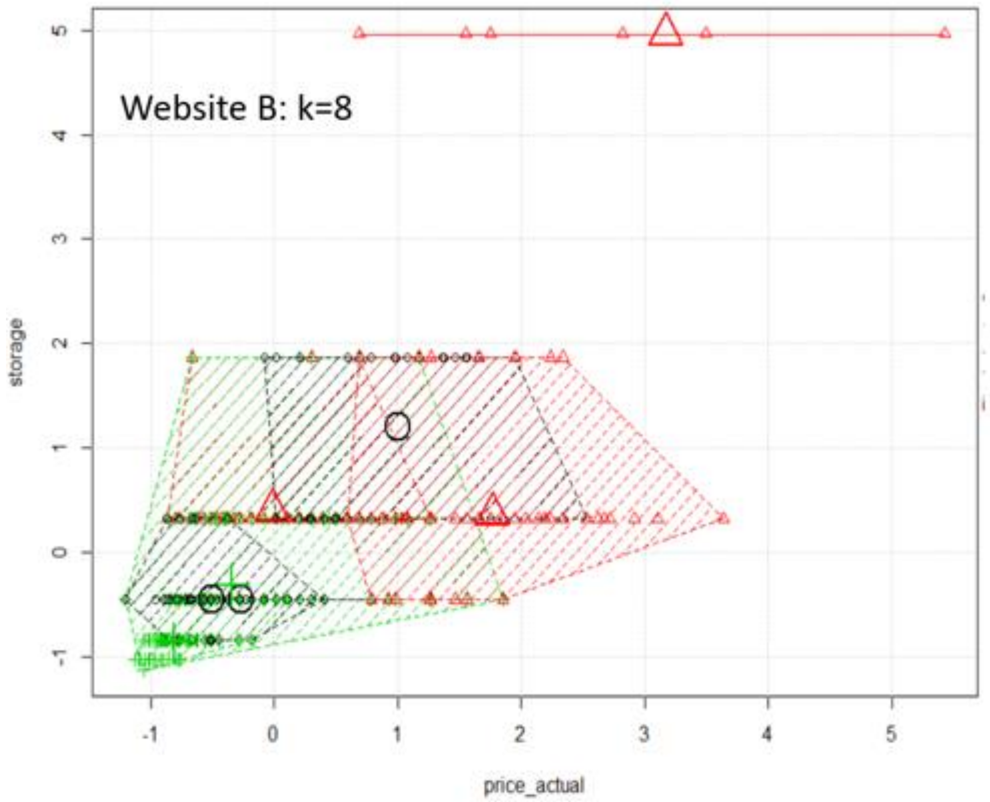
Number of phone seller base on brand from website A

There are 240 unique sellers for phone from website A, while website B is the single seller

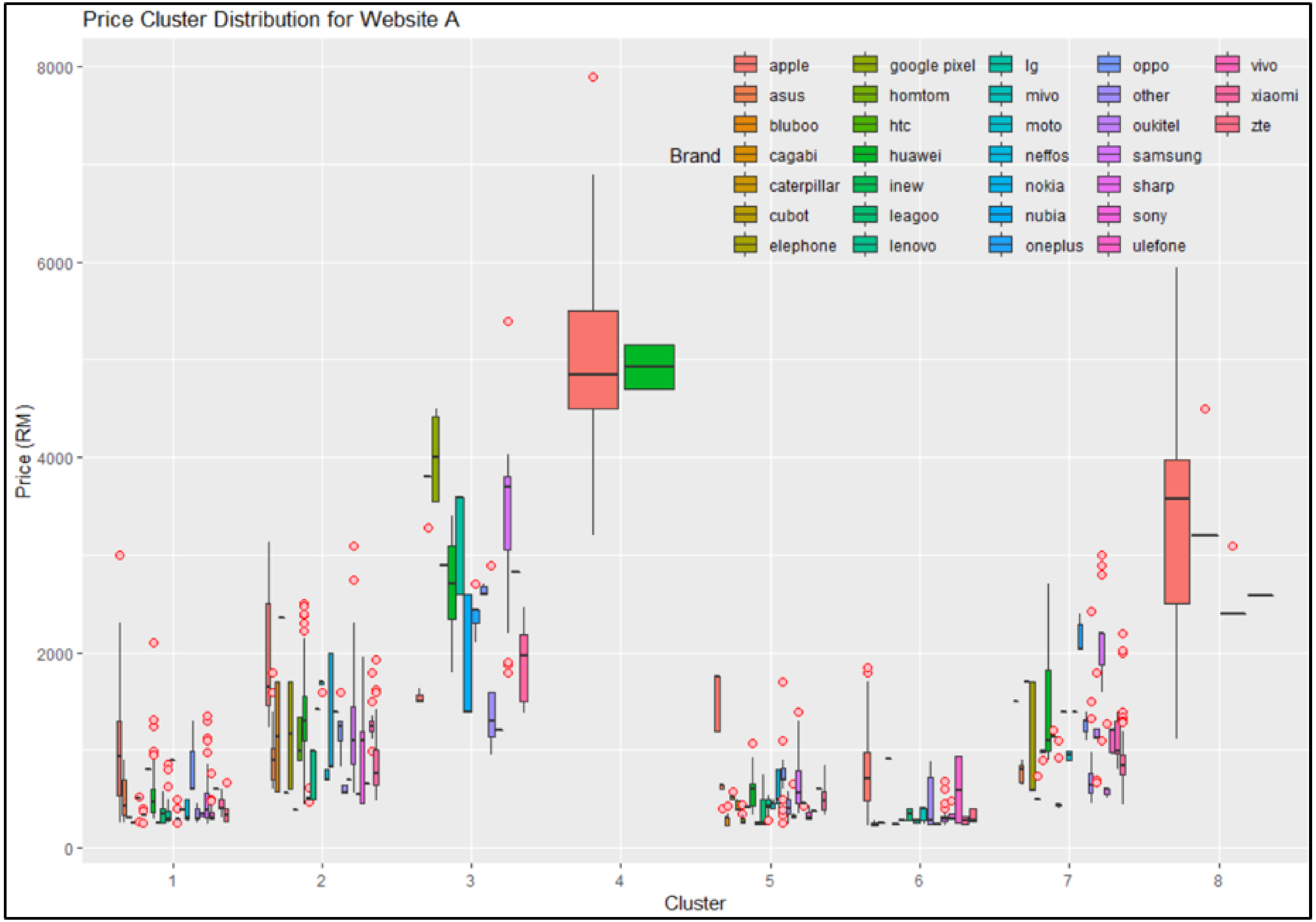
RESULTS & DISCUSSION : *K*-mean clustering

Phone price can be grouped using *k*-mean clustering.





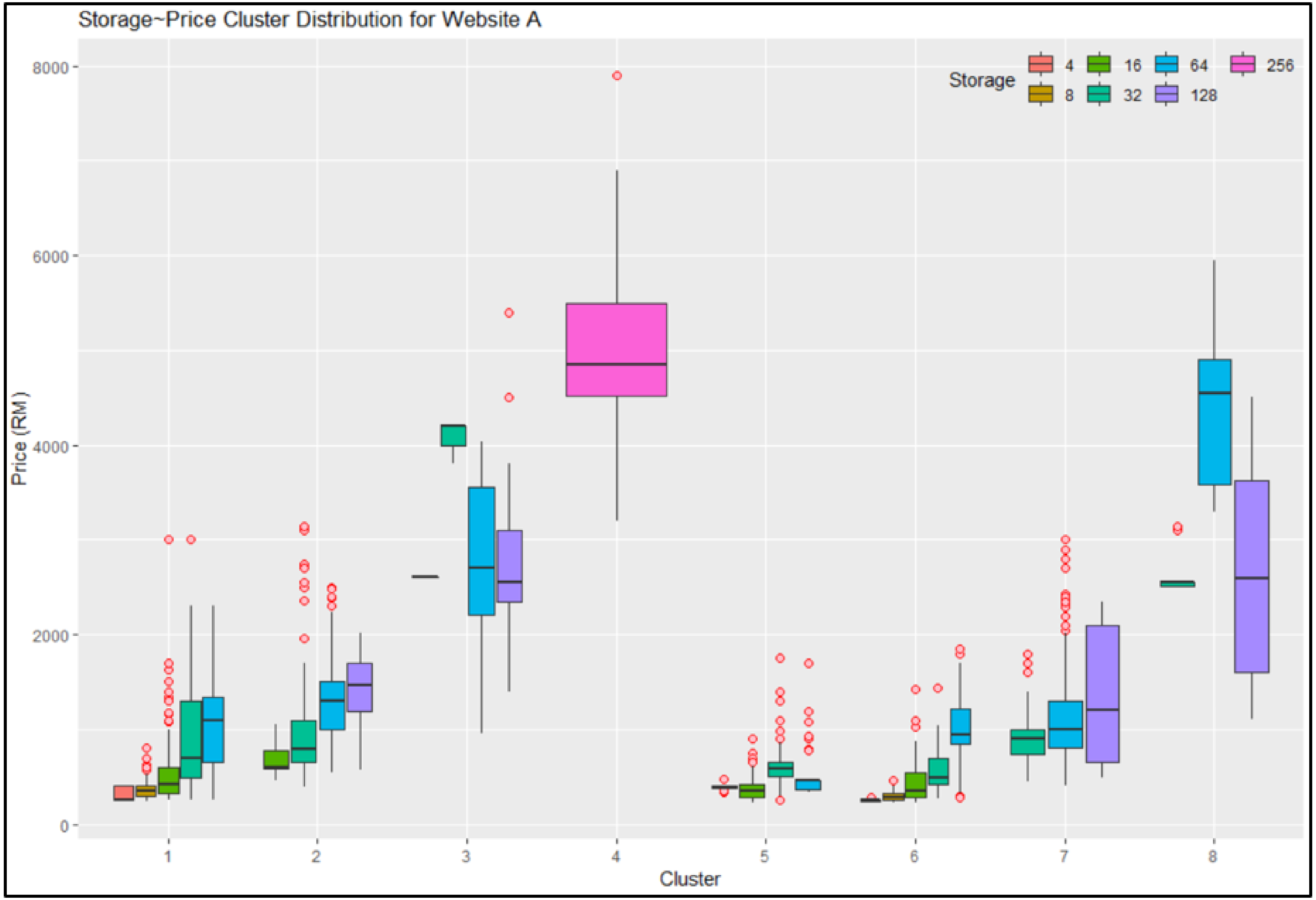
Website A



The 8 cluster group for brand-price distribution of website A

K-mean clustering

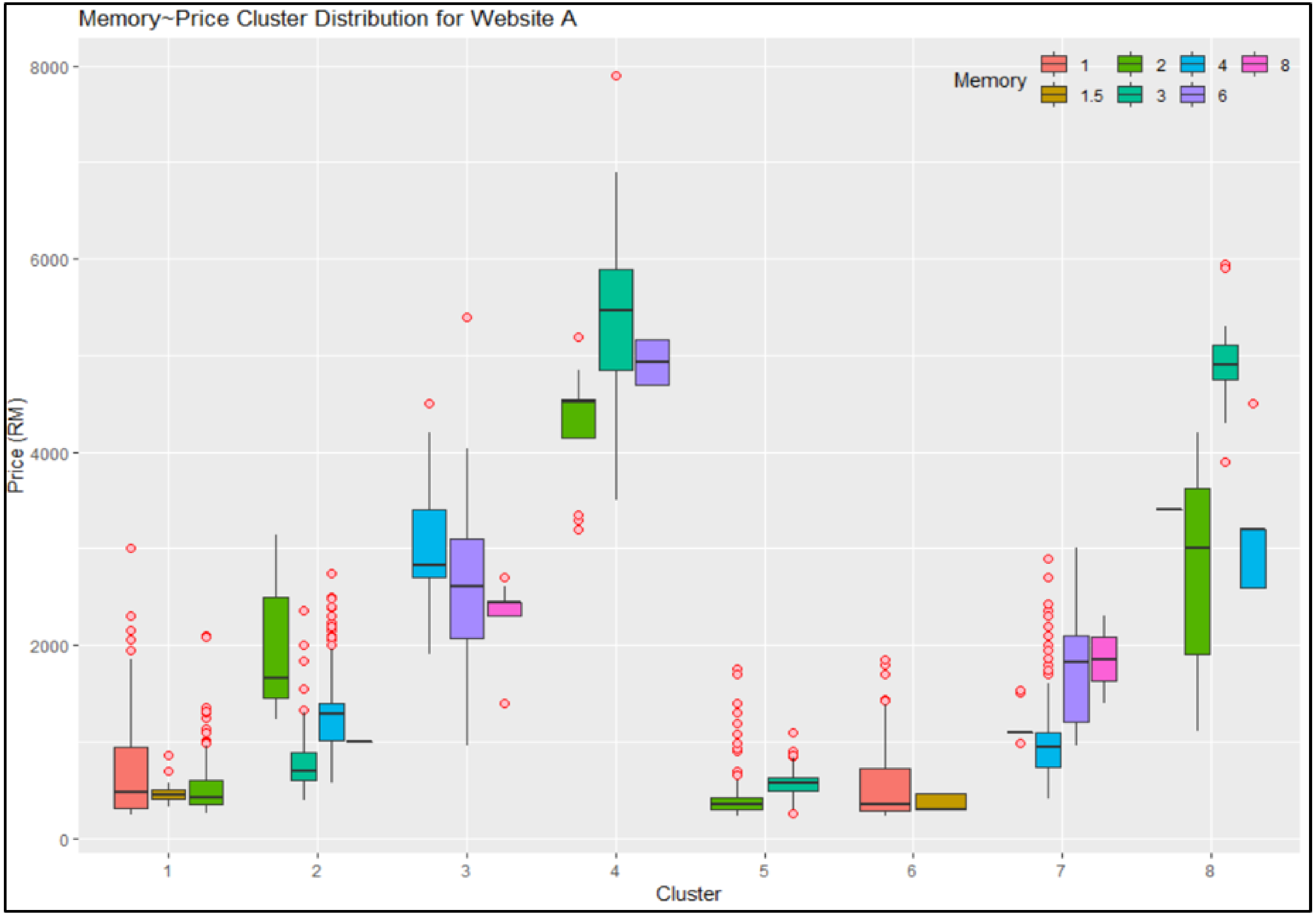
Website A



The 8 cluster group for storage-price distribution of website A

K-mean clustering

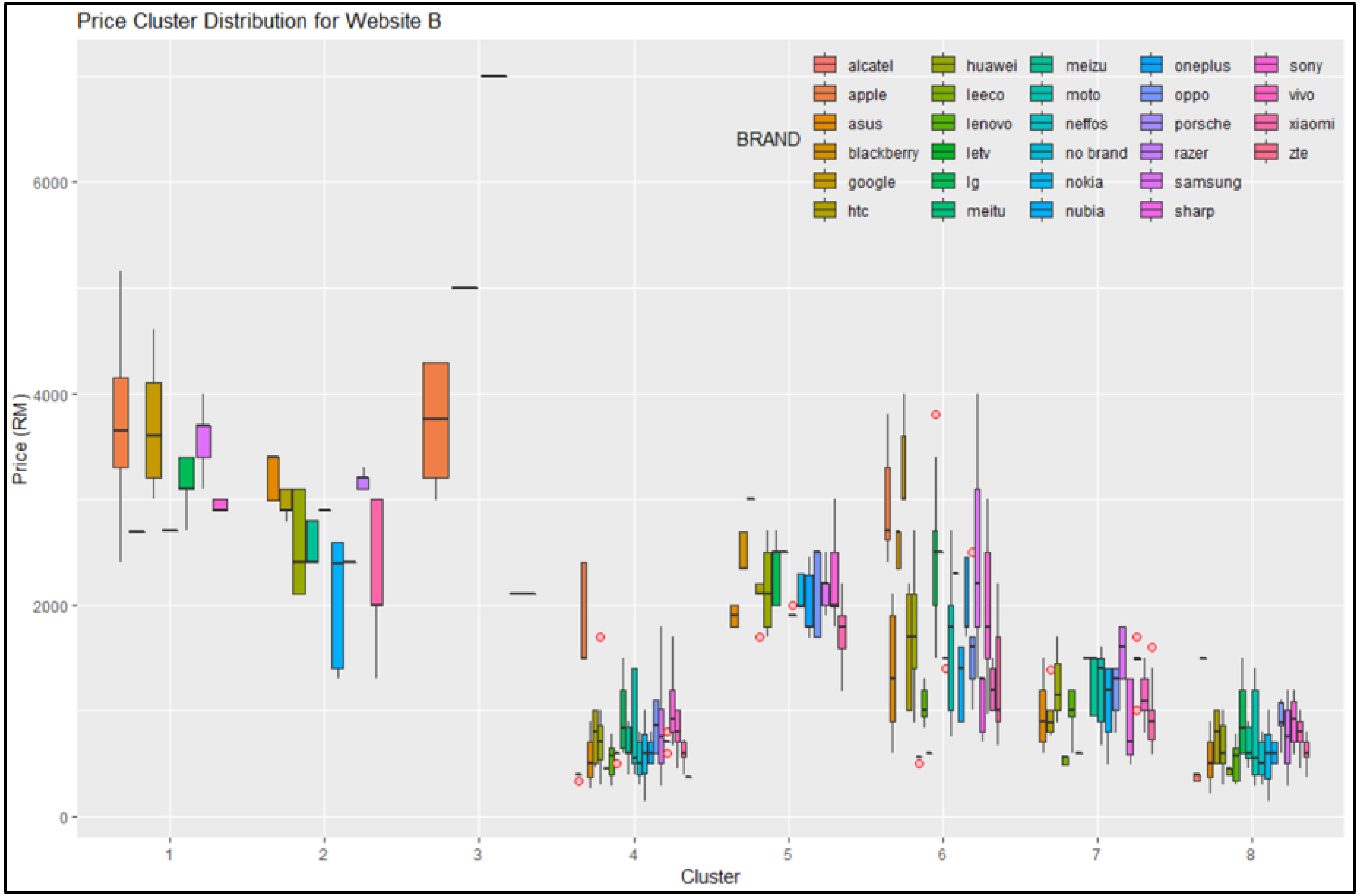
Website A



The 8 cluster group for memory-price distribution of website A

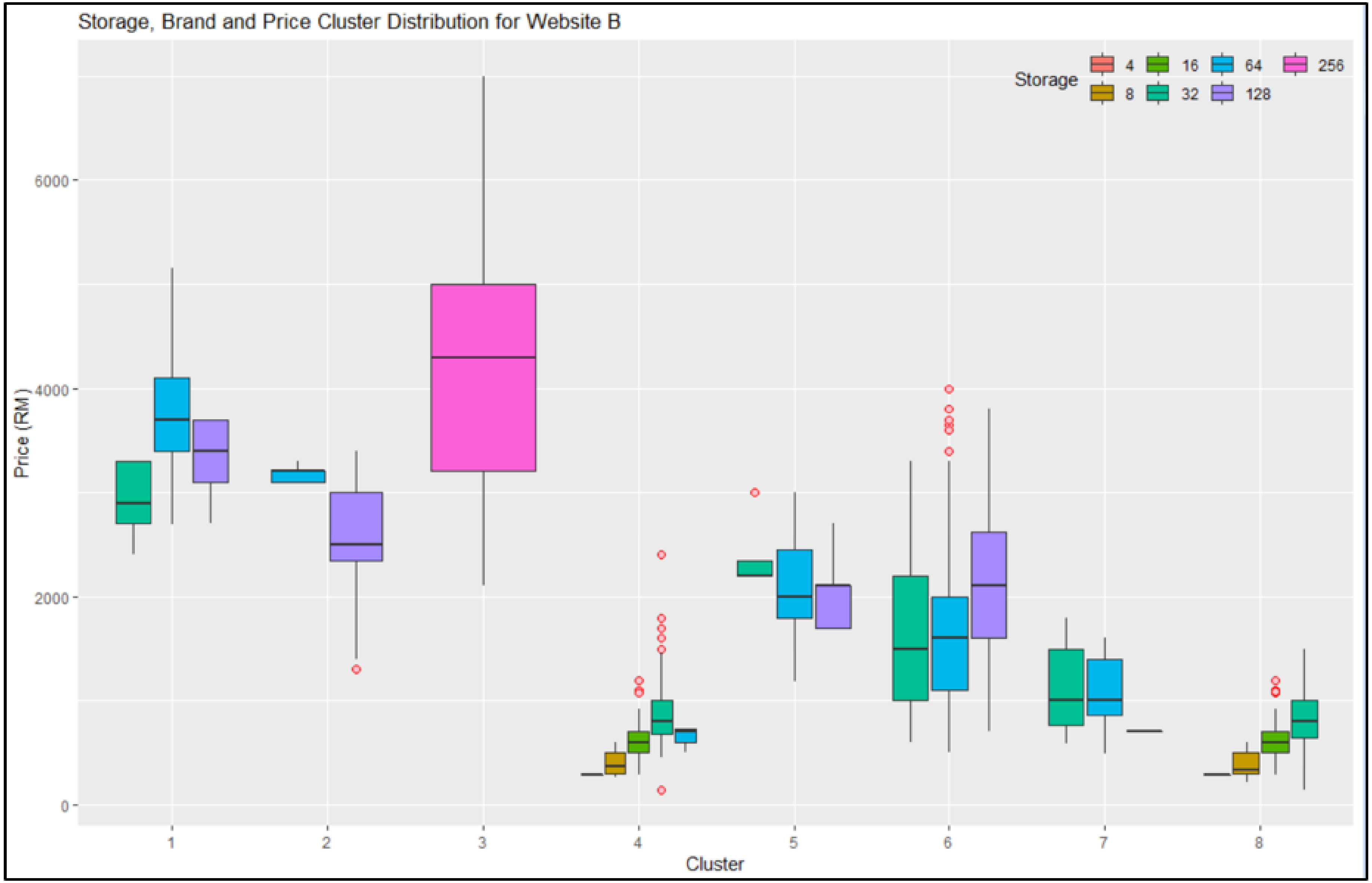
K-mean clustering

Website B



The 8 cluster group for brand-price distribution of website B

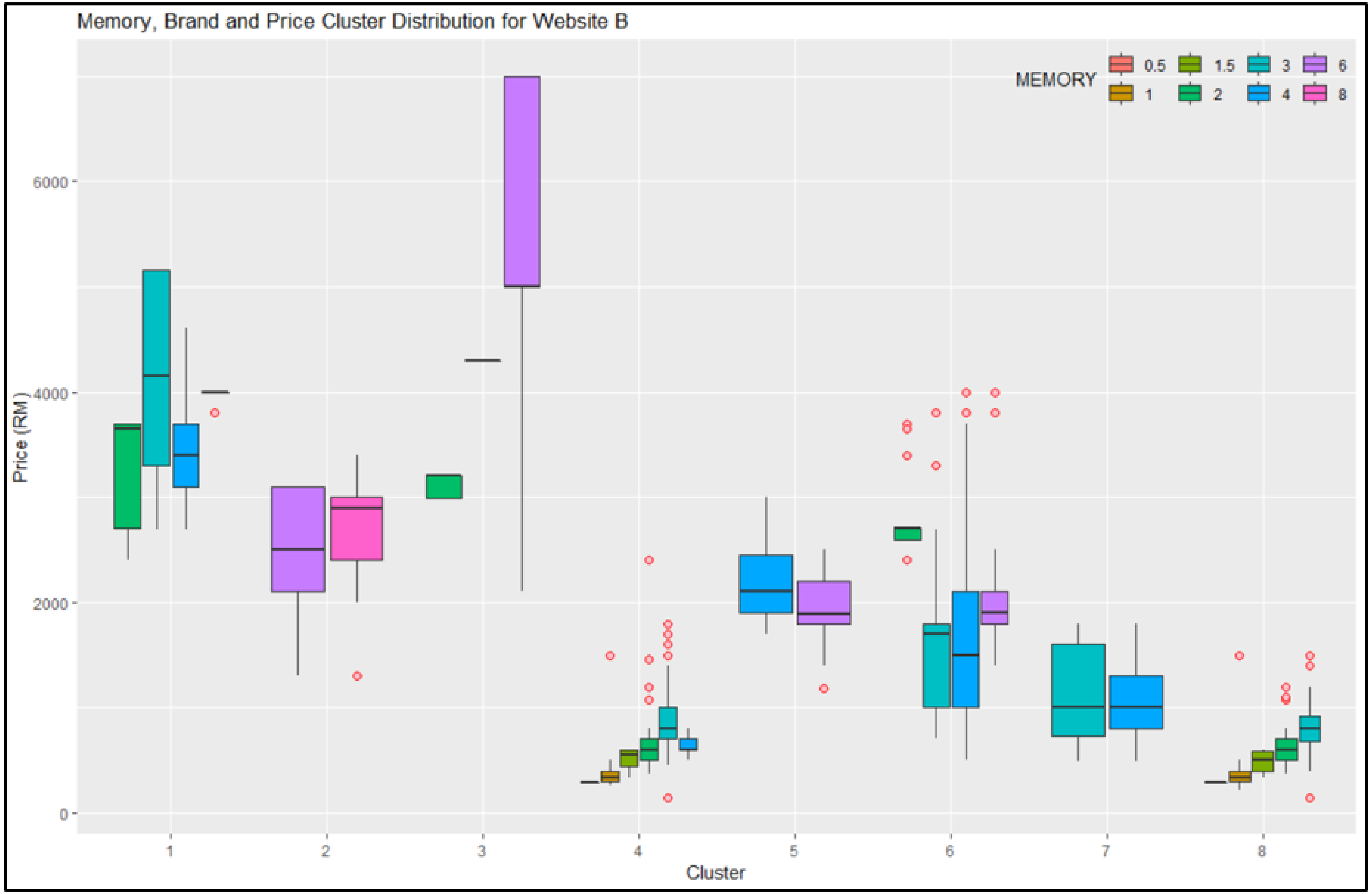
Website B



The 8 cluster group for storage-price distribution of website B

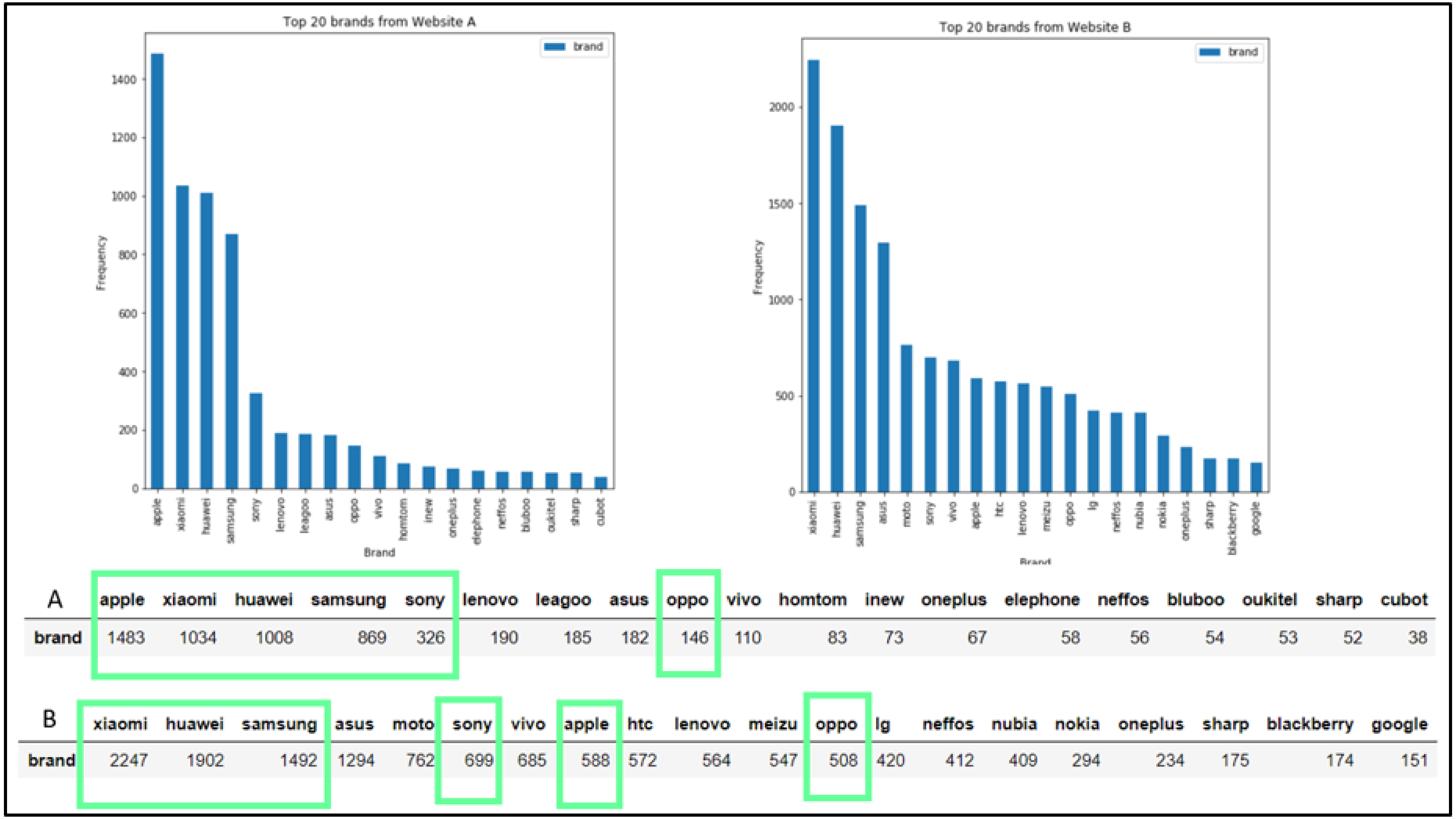
K-mean clustering

Website B



The 8 cluster group for memory (RAM)-price distribution of website B

RESULTS & DISCUSSION : *Price Dispersion*



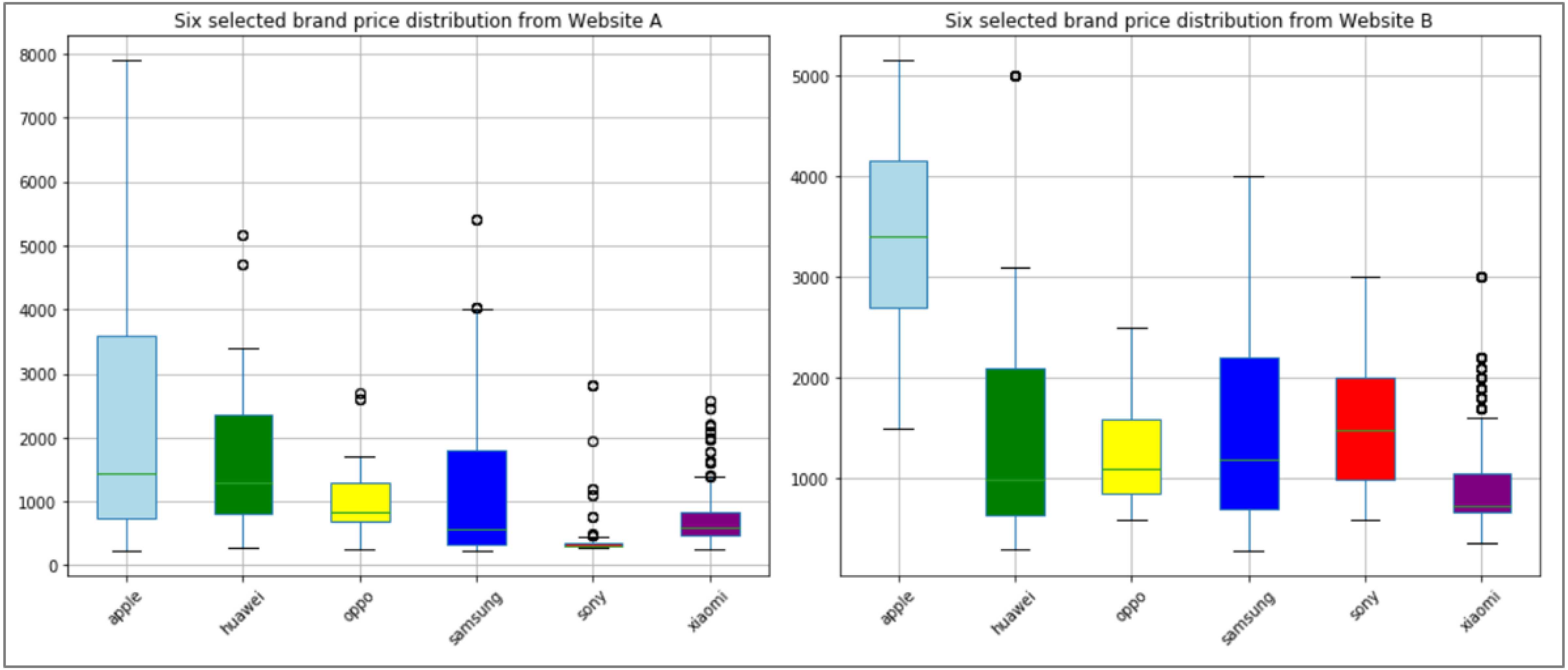
A	apple	xiaomi	huawei	samsung	sony	lenovo	leagoo	asus	oppo	vivo	homtom	inew	oneplus	elephone	neffos	bluboo	oukitel	sharp	cubot
brand	1483	1034	1008	869	326	190	185	182	146	110	83	73	67	58	56	54	53	52	38

B	xiaomi	huawei	samsung	asus	moto	sony	vivo	apple	htc	lenovo	meizu	oppo	lg	neffos	nubia	nokia	oneplus	sharp	blackberry	google
brand	2247	1902	1492	1294	762	699	685	588	572	564	547	508	420	412	409	294	234	175	174	151

Top 20 mobile phones from website A and B

Top ranking of mobile phones has been selected and matched with the physical outlets price data collection

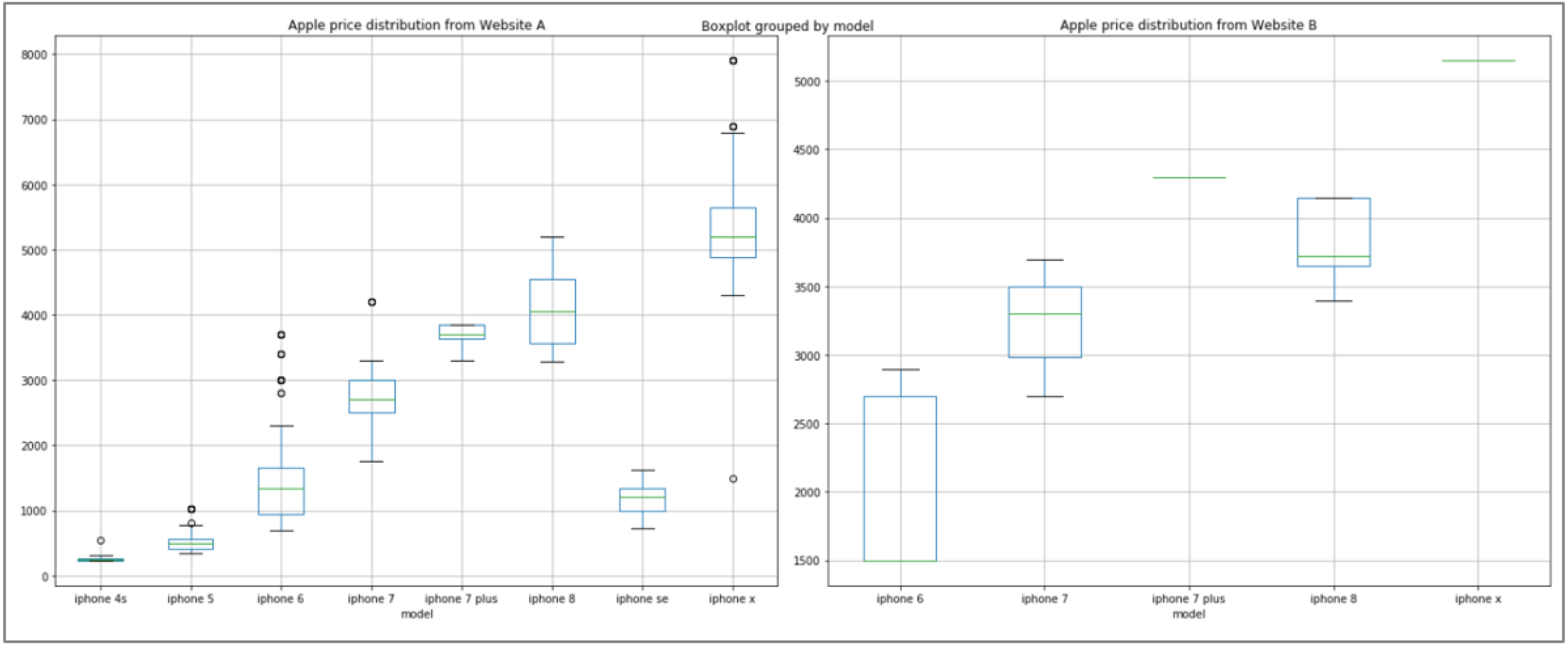
Price Dispersion



Price distribution for six mobile phones brand from website A and B

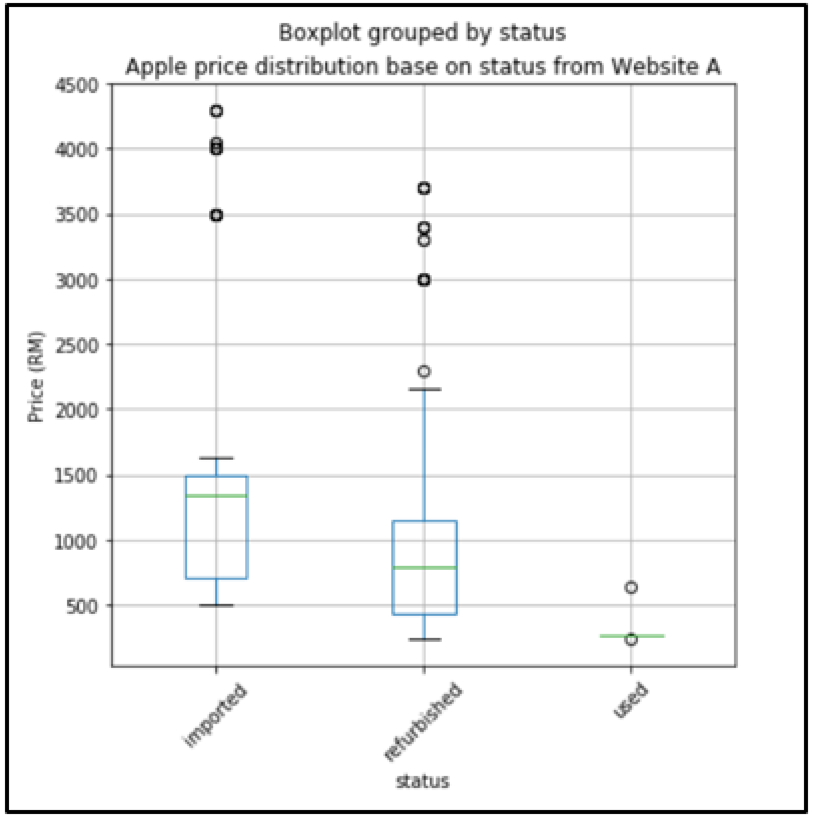
Price Dispersion

APPLE



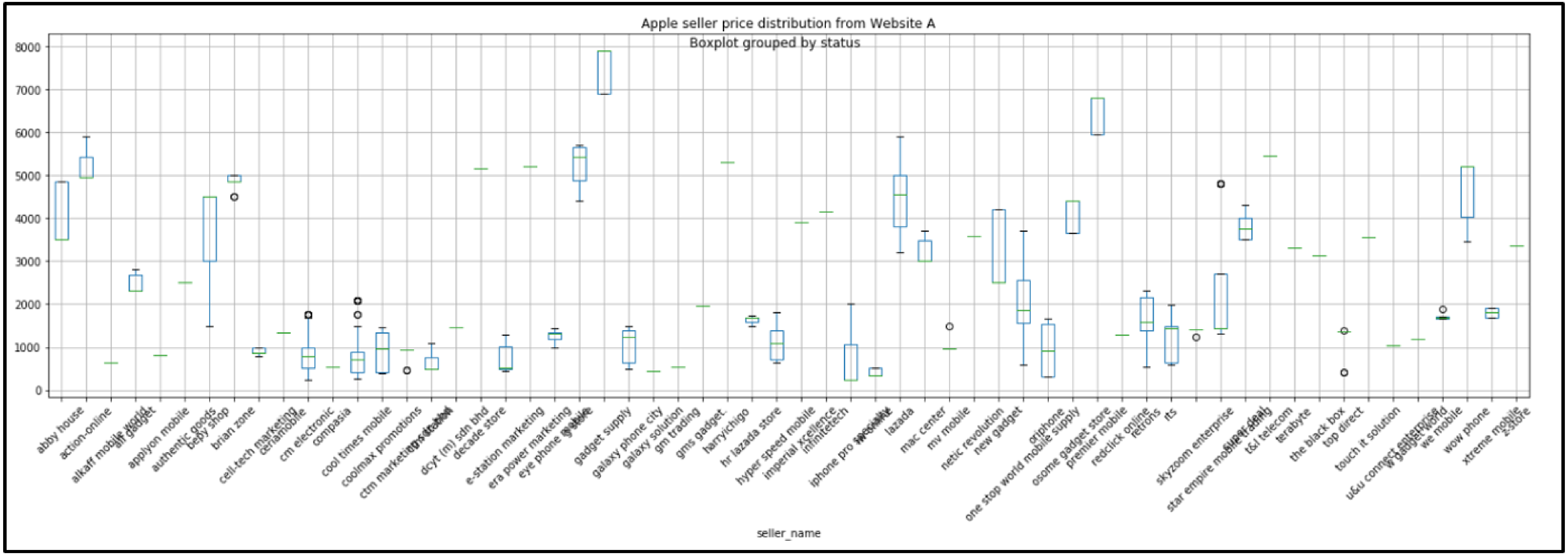
Price distribution for Apple models from website A and B

Website A: Apple



- 8 Apple model from Website A and 5 Apple model from Website B.
- The price range are different between website A and B.
- Lower price below RM 1,000 in website A occurred because used and refurbish set of iPhone is offered from this website

Website A: Apple



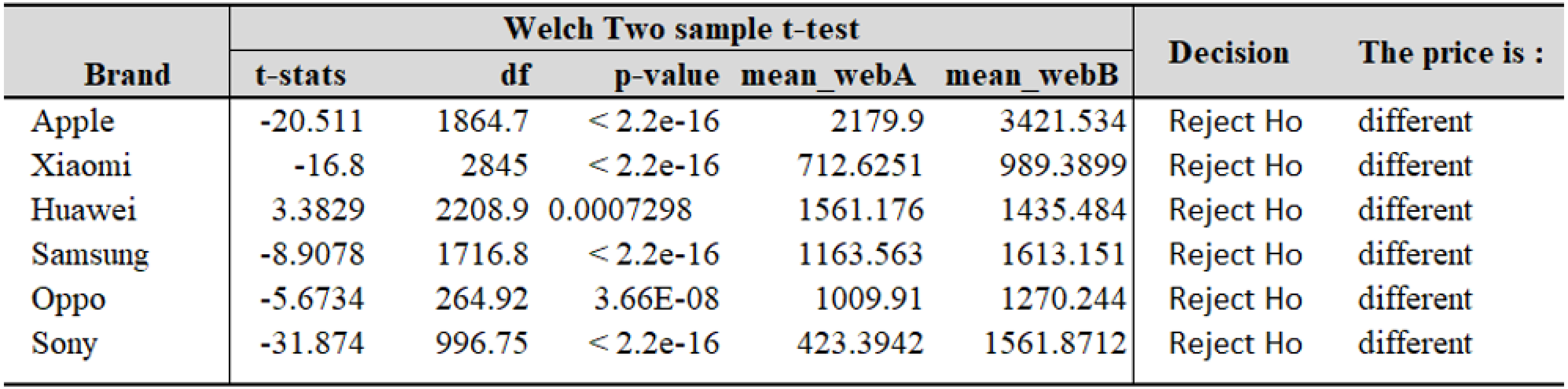
Apple models price distribution base on seller name from website A

RESULTS & DISCUSSION : *T*-test

Hypothesis 1:

Null Hypothesis: Average price of mobile phone brand from website A and B are the same,

Alternative Hypothesis: Average Price of mobile phone brand from websites A and B are different.



Brand	Welch Two sample t-test					Decision	The price is :
	t-stats	df	p-value	mean_webA	mean_webB		
Apple	-20.511	1864.7	< 2.2e-16	2179.9	3421.534	Reject Ho	different
Xiaomi	-16.8	2845	< 2.2e-16	712.6251	989.3899	Reject Ho	different
Huawei	3.3829	2208.9	0.0007298	1561.176	1435.484	Reject Ho	different
Samsung	-8.9078	1716.8	< 2.2e-16	1163.563	1613.151	Reject Ho	different
Oppo	-5.6734	264.92	3.66E-08	1009.91	1270.244	Reject Ho	different
Sony	-31.874	996.75	< 2.2e-16	423.3942	1561.8712	Reject Ho	different

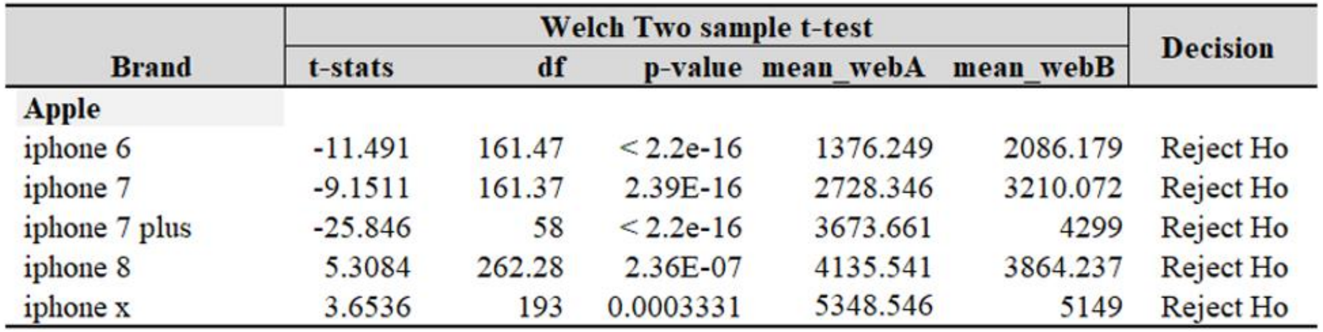
All p-value in the table are lower than significant level of 5%. Therefore, the Null Hypothesis is rejected and concluded that in overall, **this six brands have different prices between website A and B**

RESULTS & DISCUSSION : *T*-test

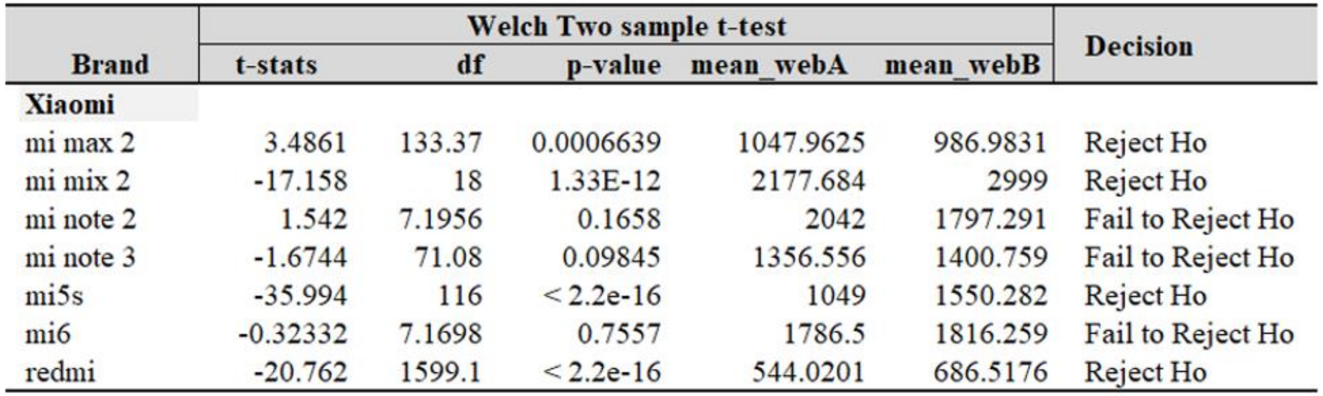
Hypothesis 2:

Null Hypothesis: Average price of **mobile phone model** from website A and B are the same.

Alternative Hypothesis: Average price of **mobile phone model** from website A and B are different.



Brand	Welch Two sample t-test					Decision
	t-stats	df	p-value	mean_webA	mean_webB	
Apple						
iphone 6	-11.491	161.47	< 2.2e-16	1376.249	2086.179	Reject Ho
iphone 7	-9.1511	161.37	2.39E-16	2728.346	3210.072	Reject Ho
iphone 7 plus	-25.846	58	< 2.2e-16	3673.661	4299	Reject Ho
iphone 8	5.3084	262.28	2.36E-07	4135.541	3864.237	Reject Ho
iphone x	3.6536	193	0.0003331	5348.546	5149	Reject Ho

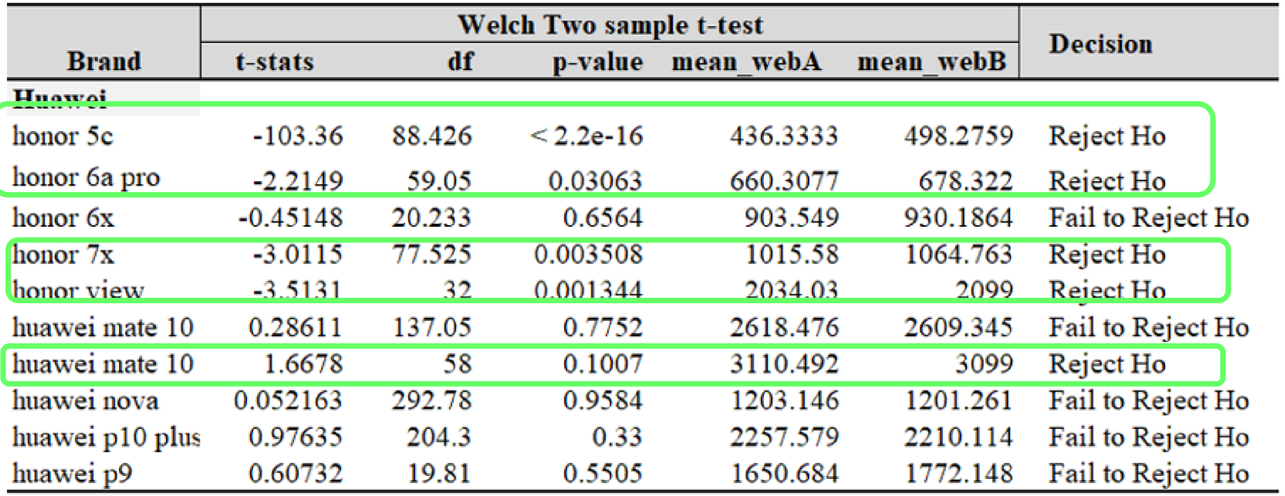


Brand	Welch Two sample t-test					Decision
	t-stats	df	p-value	mean_webA	mean_webB	
Xiaomi						
mi max 2	3.4861	133.37	0.0006639	1047.9625	986.9831	Reject Ho
mi mix 2	-17.158	18	1.33E-12	2177.684	2999	Reject Ho
mi note 2	1.542	7.1956	0.1658	2042	1797.291	Fail to Reject Ho
mi note 3	-1.6744	71.08	0.09845	1356.556	1400.759	Fail to Reject Ho
mi5s	-35.994	116	< 2.2e-16	1049	1550.282	Reject Ho
mi6	-0.32332	7.1698	0.7557	1786.5	1816.259	Fail to Reject Ho
redmi	-20.762	1599.1	< 2.2e-16	544.0201	686.5176	Reject Ho

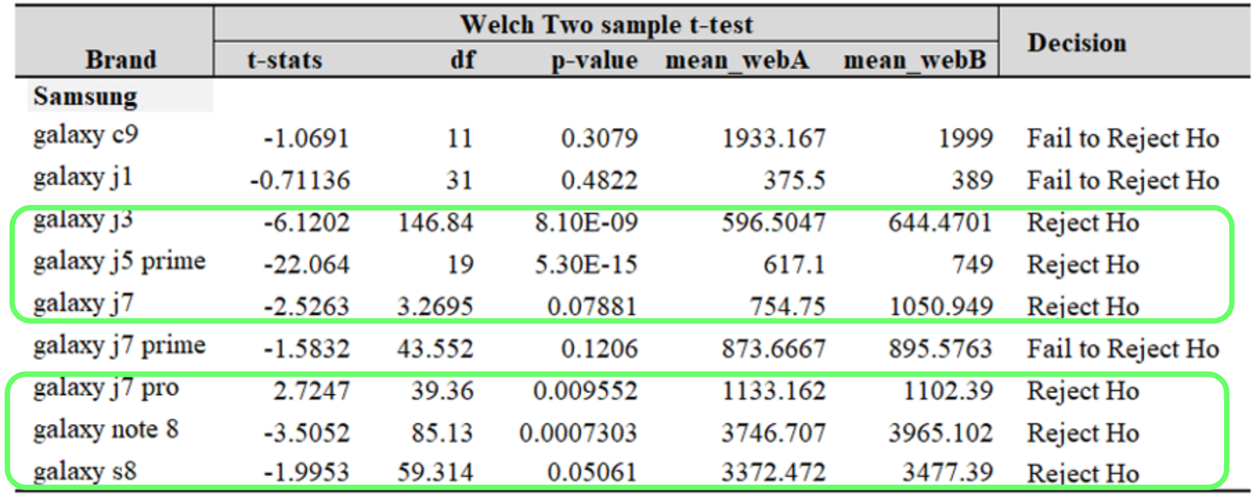
- Apple brand is significantly different between website A and B

- 4 Xiaomi model is significantly different between website A and B

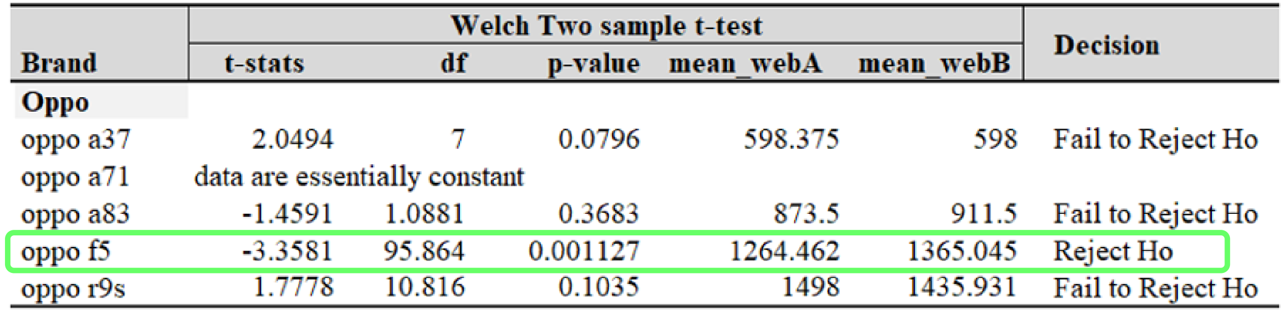
HYPOTHESIS 2



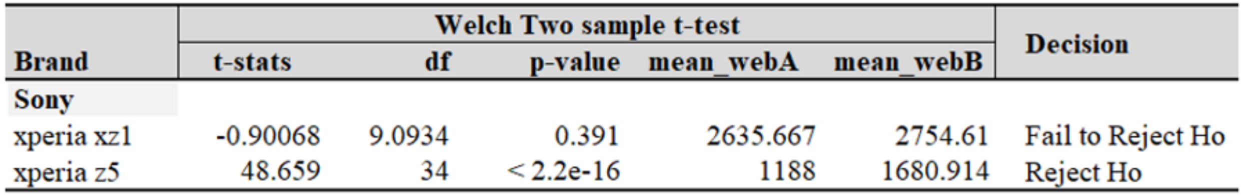
	Welch Two sample t-test					Decision
Brand	t-stats	df	p-value	mean_webA	mean_webB	
Huawei						
honor 5c	-103.36	88.426	< 2.2e-16	436.3333	498.2759	Reject Ho
honor 6a pro	-2.2149	59.05	0.03063	660.3077	678.322	Reject Ho
honor 6x	-0.45148	20.233	0.6564	903.549	930.1864	Fail to Reject Ho
honor 7x	-3.0115	77.525	0.003508	1015.58	1064.763	Reject Ho
honor view	-3.5131	32	0.001344	2034.03	2099	Reject Ho
huawei mate 10	0.28611	137.05	0.7752	2618.476	2609.345	Fail to Reject Ho
huawei mate 10	1.6678	58	0.1007	3110.492	3099	Reject Ho
huawei nova	0.052163	292.78	0.9584	1203.146	1201.261	Fail to Reject Ho
huawei p10 plus	0.97635	204.3	0.33	2257.579	2210.114	Fail to Reject Ho
huawei p9	0.60732	19.81	0.5505	1650.684	1772.148	Fail to Reject Ho



	Welch Two sample t-test					Decision
Brand	t-stats	df	p-value	mean_webA	mean_webB	
Samsung						
galaxy c9	-1.0691	11	0.3079	1933.167	1999	Fail to Reject Ho
galaxy j1	-0.71136	31	0.4822	375.5	389	Fail to Reject Ho
galaxy j3	-6.1202	146.84	8.10E-09	596.5047	644.4701	Reject Ho
galaxy j5 prime	-22.064	19	5.30E-15	617.1	749	Reject Ho
galaxy j7	-2.5263	3.2695	0.07881	754.75	1050.949	Reject Ho
galaxy j7 prime	-1.5832	43.552	0.1206	873.6667	895.5763	Fail to Reject Ho
galaxy j7 pro	2.7247	39.36	0.009552	1133.162	1102.39	Reject Ho
galaxy note 8	-3.5052	85.13	0.0007303	3746.707	3965.102	Reject Ho
galaxy s8	-1.9953	59.314	0.05061	3372.472	3477.39	Reject Ho



	Welch Two sample t-test					Decision
Brand	t-stats	df	p-value	mean_webA	mean_webB	
Oppo						
oppo a37	2.0494	7	0.0796	598.375	598	Fail to Reject Ho
oppo a71	data are essentially constant					
oppo a83	-1.4591	1.0881	0.3683	873.5	911.5	Fail to Reject Ho
oppo f5	-3.3581	95.864	0.001127	1264.462	1365.045	Reject Ho
oppo r9s	1.7778	10.816	0.1035	1498	1435.931	Fail to Reject Ho



	Welch Two sample t-test					Decision
Brand	t-stats	df	p-value	mean_webA	mean_webB	
Sony						
xperia xz1	-0.90068	9.0934	0.391	2635.667	2754.61	Fail to Reject Ho
xperia z5	48.659	34	< 2.2e-16	1188	1680.914	Reject Ho

RESULTS & DISCUSSION : *T*-test

Hypothesis 3:

Null Hypothesis:

i) Average prices of mobile phones models from website A are the same with Average prices of mobile phones from physical outlets data collection.

ii) Average prices of mobile phone model from website B are the same with Average prices of mobile phone from physical outlets data collection.

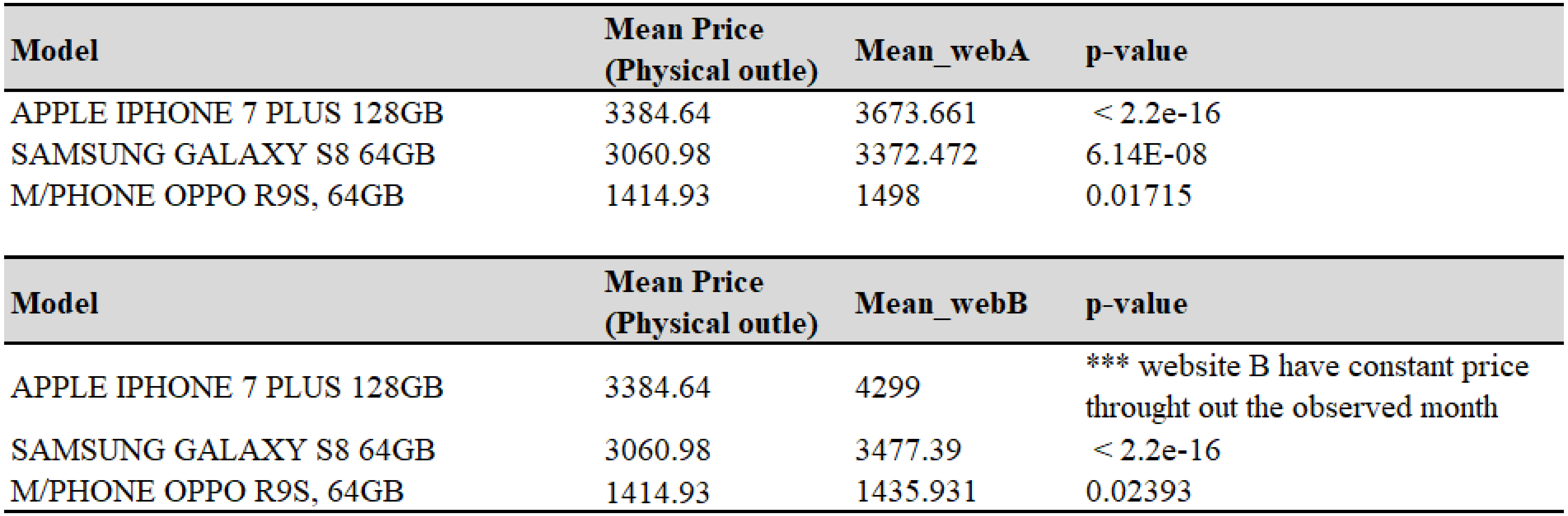
Alternative Hypothesis:

i) Average prices of mobile phones models from website A are higher compare with Average prices of mobile phones from physical outlets data collection.

ii) Average prices of mobile phones models from website B are higher as compared to Average prices of mobile phones from physical outlets data collection.

T-test

HYPOTHESIS 3

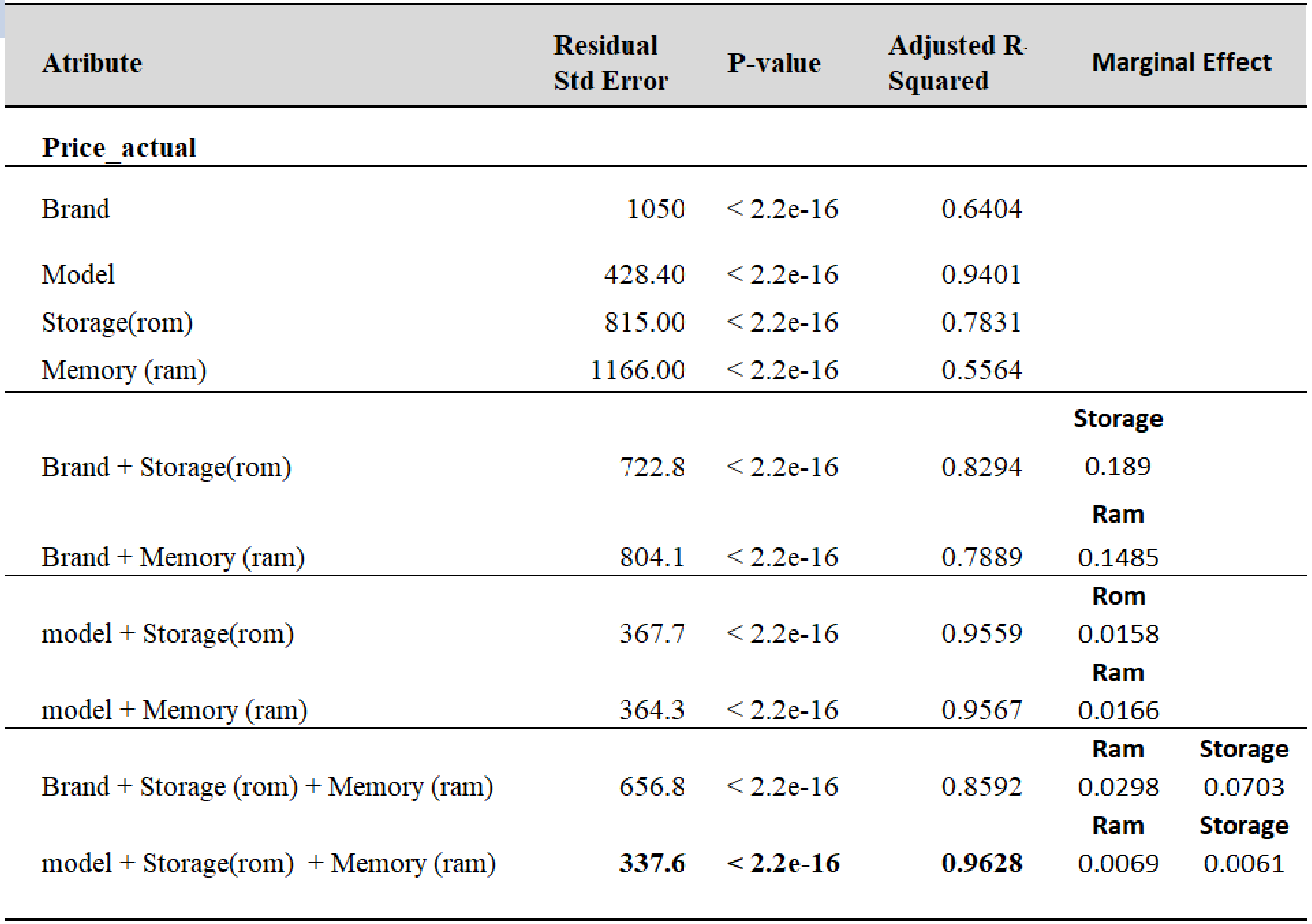


Model	Mean Price (Physical outle)	Mean_webA	p-value
APPLE IPHONE 7 PLUS 128GB	3384.64	3673.661	< 2.2e-16
SAMSUNG GALAXY S8 64GB	3060.98	3372.472	6.14E-08
M/PHONE OPPO R9S, 64GB	1414.93	1498	0.01715

Model	Mean Price (Physical outle)	Mean_webB	p-value
APPLE IPHONE 7 PLUS 128GB	3384.64	4299	*** website B have constant price throught out the observed month
SAMSUNG GALAXY S8 64GB	3060.98	3477.39	< 2.2e-16
M/PHONE OPPO R9S, 64GB	1414.93	1435.931	0.02393

- Base on the findings, the price from physical outlet is different compare to both website.
- The online prices are higher than the physical outlets average prices.
- This is only for this three specific model at the mentioned time. (Jan~Feb2018)

RESULTS & DISCUSSION : *Regression Analysis*



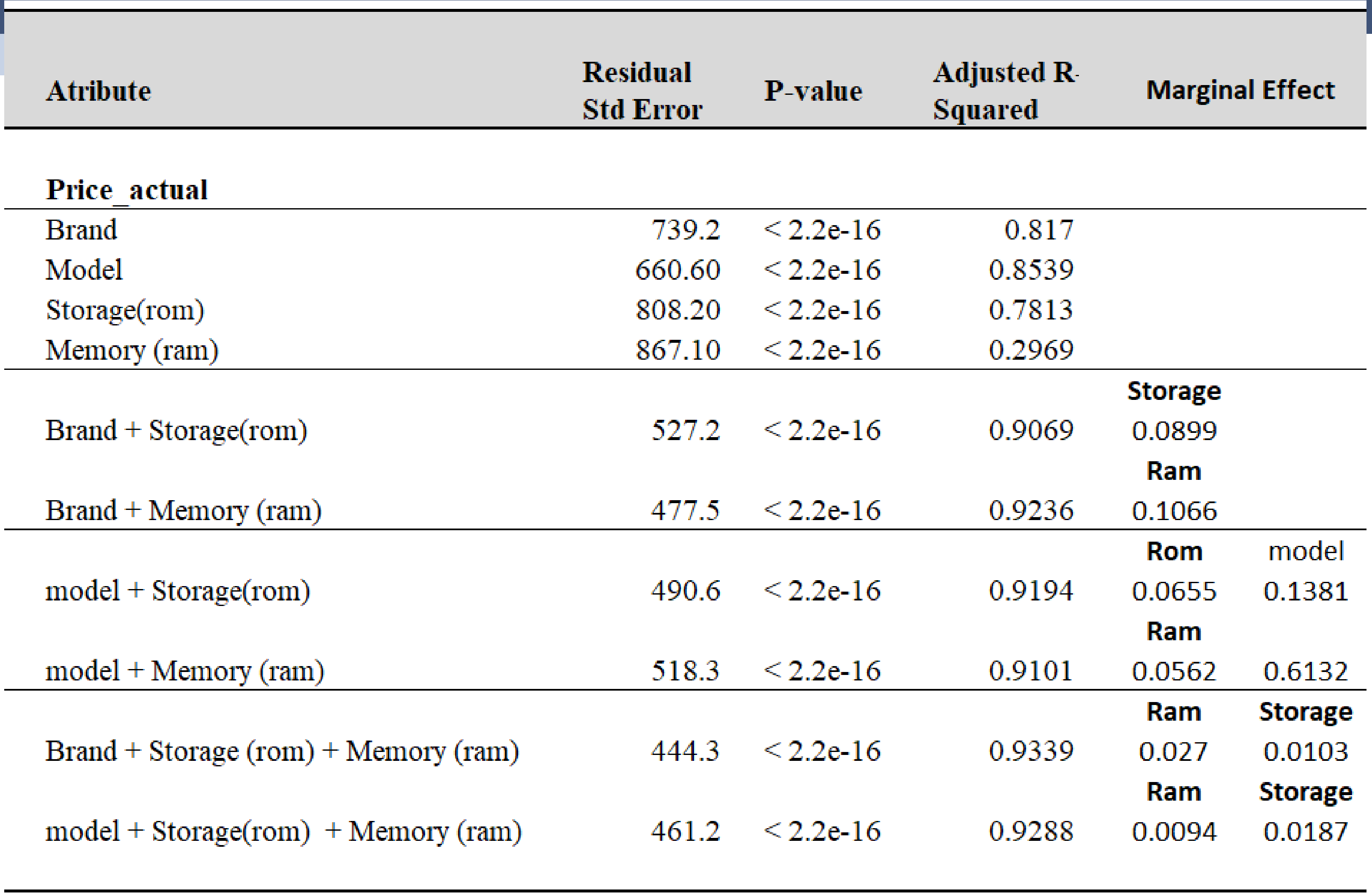
Atribute	Residual Std Error	P-value	Adjusted R-Squared	Marginal Effect	
Price_actual					
Brand	1050	< 2.2e-16	0.6404		
Model	428.40	< 2.2e-16	0.9401		
Storage(rom)	815.00	< 2.2e-16	0.7831		
Memory (ram)	1166.00	< 2.2e-16	0.5564		
Brand + Storage(rom)	722.8	< 2.2e-16	0.8294	**Storage** 0.189	
Brand + Memory (ram)	804.1	< 2.2e-16	0.7889	**Ram** 0.1485	
model + Storage(rom)	367.7	< 2.2e-16	0.9559	**Rom** 0.0158	
model + Memory (ram)	364.3	< 2.2e-16	0.9567	**Ram** 0.0166	
Brand + Storage (rom) + Memory (ram)	656.8	< 2.2e-16	0.8592	**Ram** 0.0298	**Storage** 0.0703
model + Storage(rom) + Memory (ram)	**337.6**	**< 2.2e-16**	**0.9628**	**Ram** 0.0069	**Storage** 0.0061

Table 3.16: Summary Result for Regression Analysis of Website A

Table 3.16 shows that regression model with **variables phone model, storage (rom) and memory (ram)** give the best model with high adjusted R-squared and the lowest residual standard error.

Phone model and storage size shows high variation in price for website A

RESULTS & DISCUSSION : *Regression Analysis*



Atribute	Residual Std Error	P-value	Adjusted R-Squared	Marginal Effect	
Price_actual					
Brand	739.2	< 2.2e-16	0.817		
Model	660.60	< 2.2e-16	0.8539		
Storage(rom)	808.20	< 2.2e-16	0.7813		
Memory (ram)	867.10	< 2.2e-16	0.2969		
Brand + Storage(rom)	527.2	< 2.2e-16	0.9069	**Storage** 0.0899	
Brand + Memory (ram)	477.5	< 2.2e-16	0.9236	**Ram** 0.1066	
model + Storage(rom)	490.6	< 2.2e-16	0.9194	**Rom** 0.0655	model 0.1381
model + Memory (ram)	518.3	< 2.2e-16	0.9101	**Ram** 0.0562	0.6132
Brand + Storage (rom) + Memory (ram)	444.3	< 2.2e-16	0.9339	**Ram** 0.027	**Storage** 0.0103
model + Storage(rom) + Memory (ram)	461.2	< 2.2e-16	0.9288	**Ram** 0.0094	**Storage** 0.0187

Table 3.17: Summary Result for Regression Analysis of Website B

Table 3.17 shows that the last 2 regression models give the best model with high adjusted R-squared and the lowest residual standard error.

Brand, model and storage give high impact on the phone prices sold in website B.

RESULTS & DISCUSSION : *Regression Analysis*

Regression analysis is to identify which variables have impact on the prices.

Base on regression analysis, **storage sizes and models**, significantly give influence for the phone prices.

However this is not the best regression model to be used because other factors are not fully taken into account such as multicollinearity, autocorrelation, heteroscedasticity and outliers.

Further analysis can be carried out in future studies to find the best model to fit phone pricing in Malaysia. Details product specification and information can also be enriched for the better results.

CONCLUSION

01 There exist price **differences** between website A and website B

02 There exist price **differences** between average online price and the average price data collection through physical outlets

03 The online prices are **higher** than average price data collection through physical outlets

04 The price differences are basically **contributed** by the brands and the specifications of a model

STUDY LIMITATION

- New area / initial project in DOSM.
- Limited past research study on the initial topics.
- Limited access to data and data collection technique.
- Dashboard.
- Time.

RECOMMENDATION

- Further analysis can be conducted with better product specification through details data enrichment.
- Analysis on price differences due to **competition between merchants or sellers** are proposed to be done in the future.
- Better understanding of the **influenced factors** of the price (e.g. **shipping cost, seller ratings, warranty, loyalty/membership**, etc) can also be obtained through more enrichment data process.
- Extended the study by comparing the consumer price index for goods online and existing field methods.

BENEFITS OF RESEARCH

- This study as a paving way for further consumer pricing studies/ task in DOSM.
- Big data analytics team is currently developing the dashboard related to the consumer price index.
- The government will have the signal when there is an increase of the price, hence the appropriate solutions can be draw in order to control the inflation in the country.

THANK YOU …

NUR HURRIYATUL HUDA ABDULLAH SANI
DEPARTMENT OF STATISTICS MALAYSIA

Master of Science (Data Science and Analytics)
School of Mathematical Sciences
Universiti Kebangsaan Malaysia

nurhurriyatulhuda@gmail.com