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

By:

NUR HURRIYATUL HUDA BINTI ABDULLAH SANI

CORE TEAM MALAYSIAN BUREAU OF LABOUR STATISTICS

DEPARTMENT OF STATISTICS MALAYSIA

PRESENTATION OUTLINE

01 Introduction

Background of the study | Project Focus (Objectives)

02 Literature Review

Web Scrapping| Price Pattern Analysis

03 Methodology

Handling Unstructured Data| Method of Analysis

04 Results and Discussion

Price Dispersion | T-test

05 Conclusion and Recommendation

Conclusion | Recommendation

1. Introduction

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 for 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,**(Do bre and Xhafa, 2014)**.**
- Largest source of data is **online data**

Introduction

Food & Non-alcoholic Beverages.

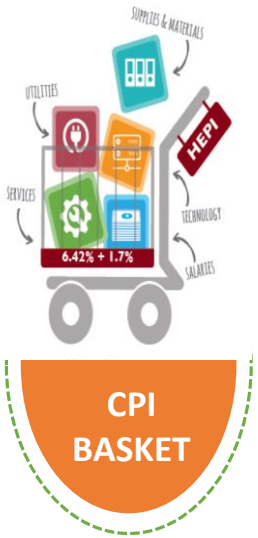
Alcoholic Beverages & Tobacco.

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

FOCUS 1:

DATA MANAGEMENT

→ Handling Unstructured Data

Project Focus

FOCUS 2:

DATA ANALYSIS

→ Price Pattern & Dispersion Analysis

2. Literature Review

WEB SCRAPPING

Author	Title	Findings
Cavallo, 2017	Are online and offline prices similar? evidence from large multi-channel retailers.	Online price information is increasingly being used as economy indicator measurement and research application. The data collections of online price are one of the main initiative that is concerned by majority of the official statistical agencies around the world**. National Statistical Offices (NSOs) have recently started to consider the use of online data in official Consumer Price Index (CPI)**

WEB SCRAPPING

Author	Title	Findings
• Polidoro, F. et. al., 2015 • Nygaard, R. 2015 • Josef Auer and Ingolf Boettcher, 2016	• Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. • The use of online prices in the Norwegian Consumer Price Index. Statistics Norway.	The purpose of modernization in this consumer price data collection is to understand and evaluate the suitability, quality of data and potential of cost reduction if other data sources can be used as an alternative or complementary for CPI measurements and inflation. Among the countries involved are **The Italian National Statistical Institute (Istat)** who did the research for online data collection using open source tools i.e. **iMacros** , while **Statistics Austria, and Norway Statistics** using **Import.io software as web scraper tools.**

Literature Review

Author	Title	Findings
Breton R, et. al., 2015	Research indices using web scraped data: May 2016 update. Newport: Office for National Statistics.	Office for National Statistics United Kingdom (ONS) using Python scrapy module to scrap data from three online supermarket chains in UK which are Tesco, Sainsbury and Waitrose. The web scrapers managed to collect estimates of 6,500 daily price quotation i.e. approximately 200 thousand price quotation monthly. This is a huge amount of price quotes compare to price collected in the traditional approach
Chuanyang and Joseph, 2016	Experiences with the use of Online Prices in Consumer Price Index	Statistics Singapore also uses open source software i.e**. Import.io** as a web scraper to scraped prices for airline tickets fares. There are two types of web crawlers that have been used i.e. customized web crawlers and the 'point-and-click' .
Boeing, G., & Waddell, P. (2017)	New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings	The study of rental housing market across the United States, used web scraped tool to collect housing rental data in the US to see the trend of home-rental rates in US urban areas. The python software has been used as a web scraper tools in the study

Literature Review

Author	Title	Findings
Polidoro, F. et all, 2015	Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. Statistical Journal of the IAOS, 31(2), 165-176.	Istat uses web scrap techniques to collect prices for electronic goods and airline tickets fares. Open source software is used for scraped data purposes. Istat found that there was an **improvement in terms of 30% time saving in carrying out price quotations for electronic goods.**
Nygaard, R, (2015)	The use of online prices in the Norwegian Consumer Price Index. Statistics Norway.	Statistics Norway has started a pilot study for price data collection using web scraped as early as in year 2014. Despite the increase in online sales, the reality is unsure whether price changes are the same or not. **Thus the researcher suggest that it is important in the preparation of the CPI, both methods of collecting prices i.e. web scraped for online price data and traditional approach of price collections are used as complementary to the CPI and inflation measurement.**

PRICE PATTERN ANALYSIS

Author	Title	Findings
Tang and Xing (2001)	Will the growth of multi-channel retailing diminish the pricing efficiency of the web?.	The spread of prices for pure e-commerce seller can be considered lower than the prices of multi-channel retailers.
Pan et al., 2002, p. 433	Pricing on the Internet. Journal of Product & Brand Management	Price dispersion is the variation of the price for the same product. It is defined as “the distribution of prices of an item with the same measured characteristics across sellers, as indicated by measures such as range and standard deviation of prices.
Smith and Brynjolfsson (2001)	Consumer decision-making at an Internet shopbot: Brand still matters.	The price sensitive consumers on internet respond very strongly to well-known and heavily branded retailers which show that branding can be important in determine the price. They also found that consumer who care about shipping time are likely to prefer well-known brands, because promised shipping times are difficult to enforce

Literature Review

Author	Title	Findings
Petrescu, M., (2011)	Online price dispersion– more than imperfect information	**…** aims to analyse the price dispersion for online shopping mall and to identify factors that affected the price dispersion. Regression analysis for price dispersion on five factors was carried out i.e. the average price of goods, types of goods, number of sellers, shipping costs and reviews and ratings from users. The results show that price dispersion is influenced by the average price of goods, types of goods, shipping costs and reviews and ratings of users. The number of sellers does not affect the price dispersion because users can easily make price comparisons across sellers
Cavallo (2017)	Are online and offline prices similar? evidence from large multi-channel retailers	… through the BPP shows that there is a price difference between price collection through physical outlet and online price collection using web scraping technique. But this difference is small where eight countries show online prices are lower than prices at physical outlets, while Australia and Argentina show otherwise. However, this research only involves multi-channel retailer, a retailer with both business platforms, physical outlet and online website

3. METHODOLOGY

METHODOLOGY

4. HANDLING UNSTRUCTURED DATA

Data Acquisition

DATA SOURCE

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

TYPE OF DATA

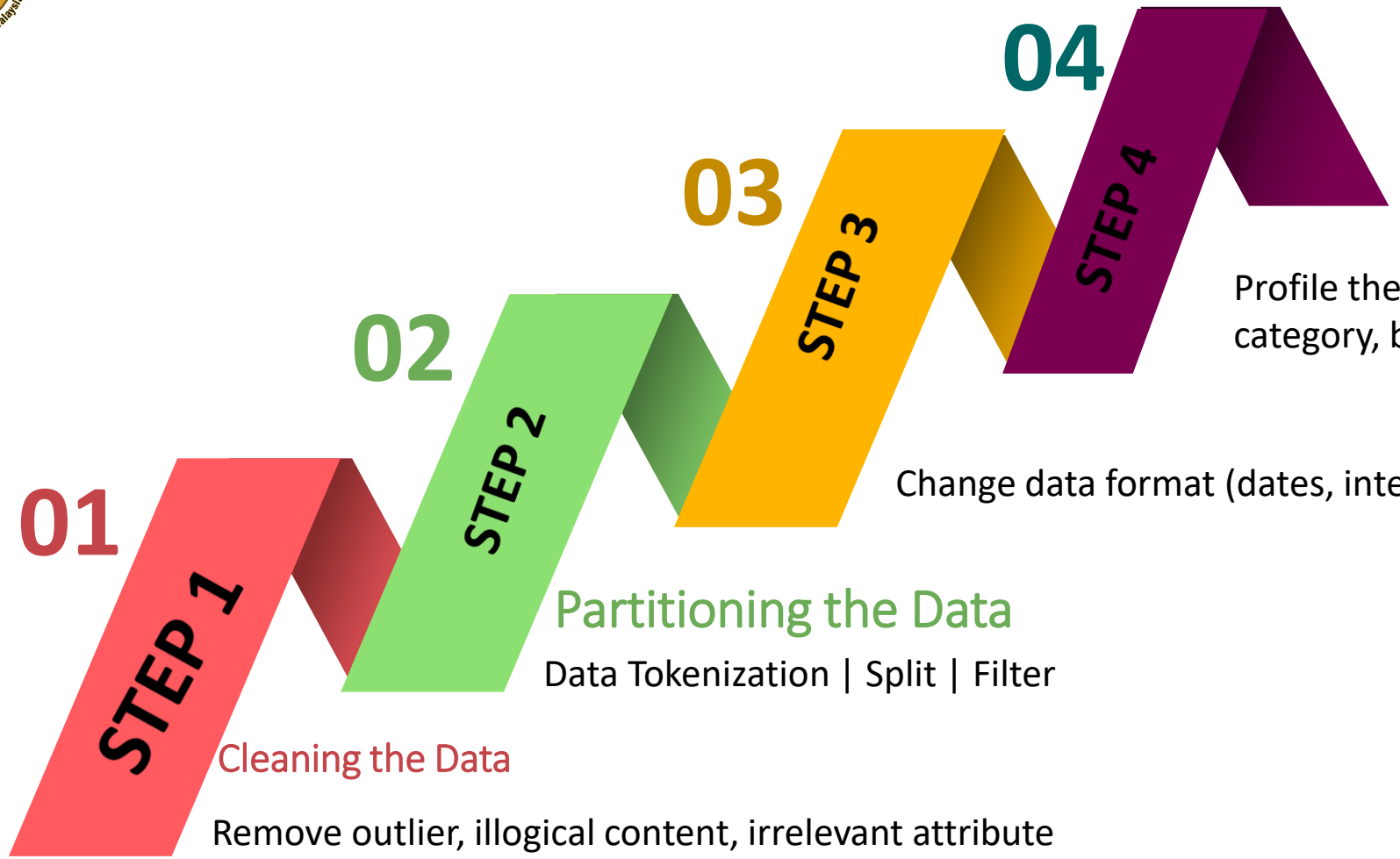
- SEMI STRUCTURED

* Semi structured dataset is when the data has field attribute form but the elements under the attribute are in unstructured form i.e. in this case is long text form. The items are organized differently according to the website preferences

TOOLS

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

Data Preparation



Cleaning the Data

Remove outlier, illogical content, irrelevant attribute

Partitioning the Data

Data Tokenization | Split | Filter

Change data format (dates, integer, string)

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

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.

```
[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'
```

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',
 nan,
 'bedding & bath/bath/towel rails & warmers/pjy towel rack 3038(cappuccino)',
 'special promotion/pjy towel rack 3065 (cappuccino)',
 "fashion/women/clothing/women's hoodies & sweatshirts/hequ autumn hooeded sweatshirt women embroidery flower long sleeve pul
lover streetwear fleece hoodies burgundy",
 "fashion/boys/accessories/boys' hats & caps/tactical chief adjustable baseball cap shooting cap hat black",
 "fashion/women/clothing/women's hoodies & sweatshirts/hequ autumn hooeded sweatshirt women long sleeve pullover streetwear f
leece hoodies grey",
 'fashion/women/clothing/pants & leggings/leggings/fashion women stretch skinny leggings pencil pants slim casual trousers -
```

Step 2: Partitioning the Data

```
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	

Step 2: Partitioning the Data

```
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...

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

TOTAL 14,066,121

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

TOTAL 1,251,376

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**

Item Selection

Five Selection Criteria :

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

Selected item: Mobile Phones from Websites A and B

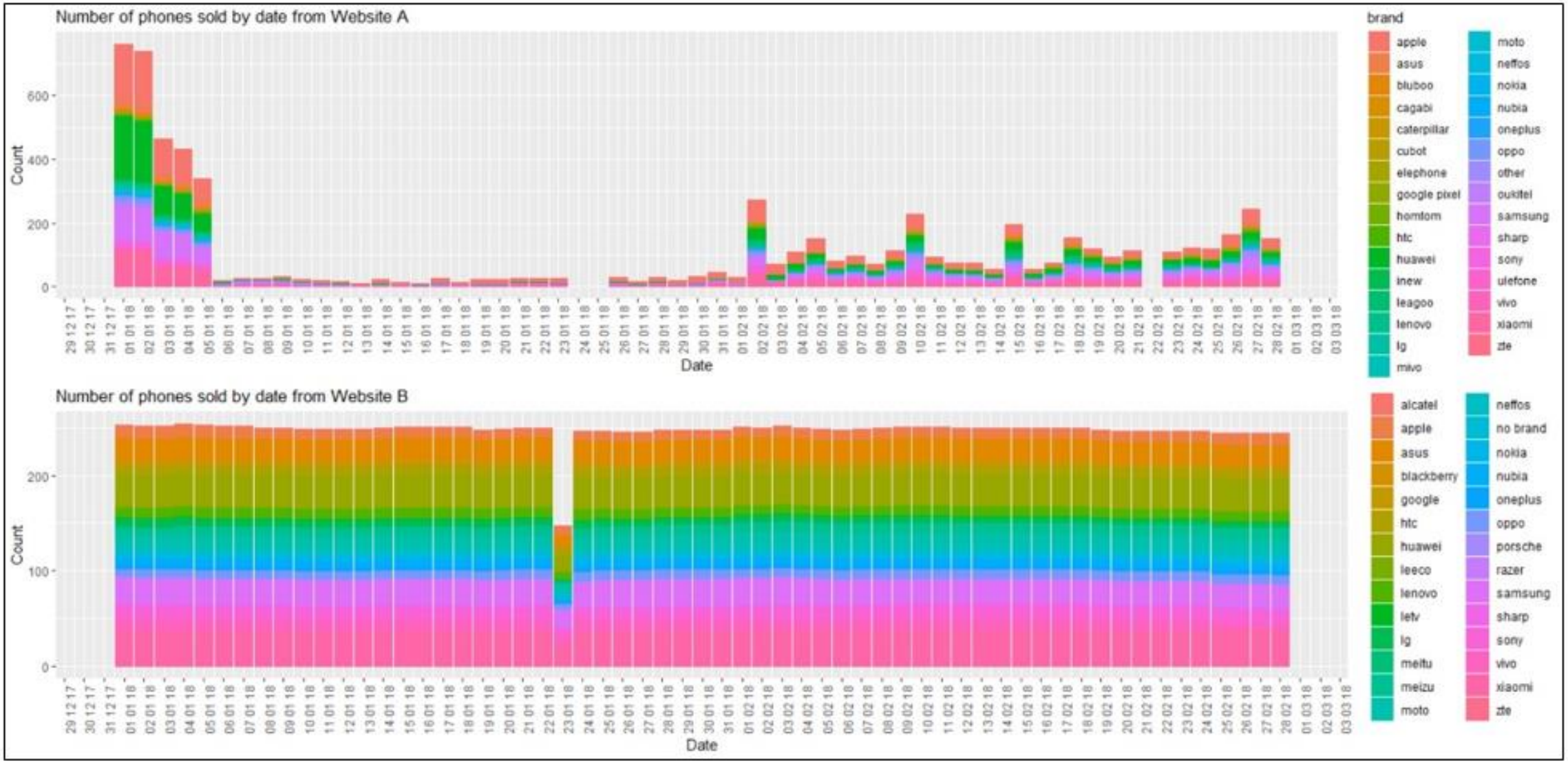
5. PRICE PATTERN ANALYSIS

Findings

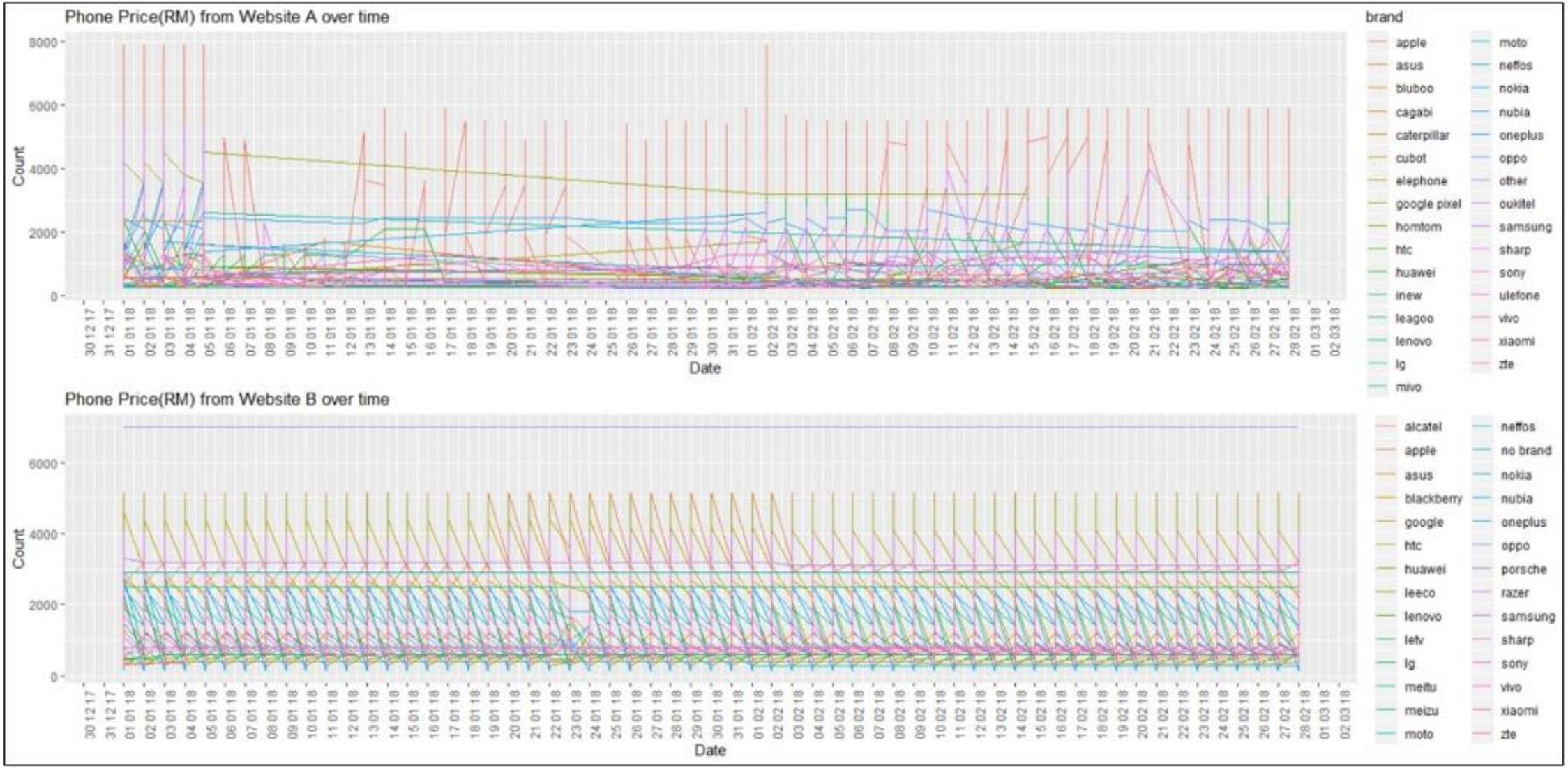
Selected item for this project is mobile phones from websites A and B. Below results were obtained from both websites.

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

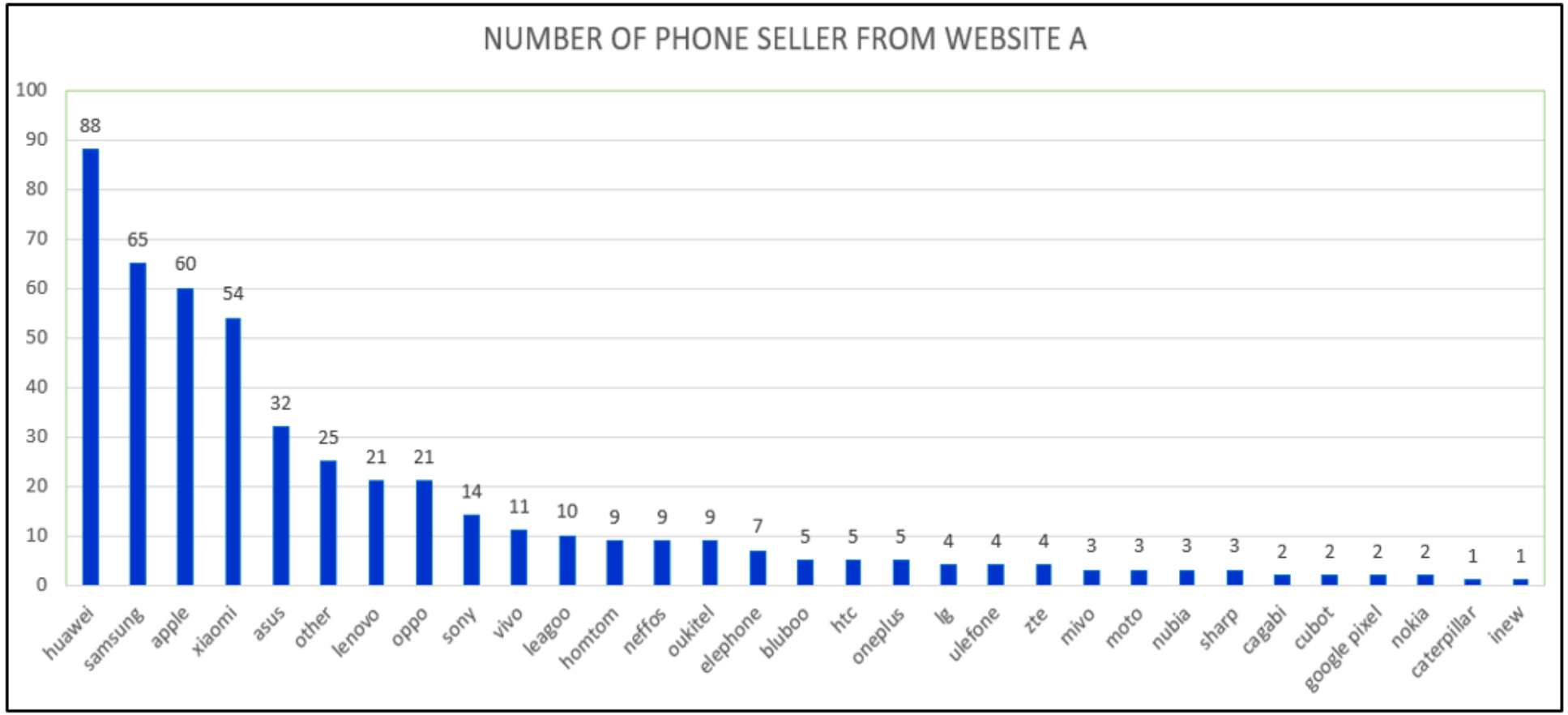
Findings : Daily quantity offered



Findings: Daily price offered



Findings



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

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

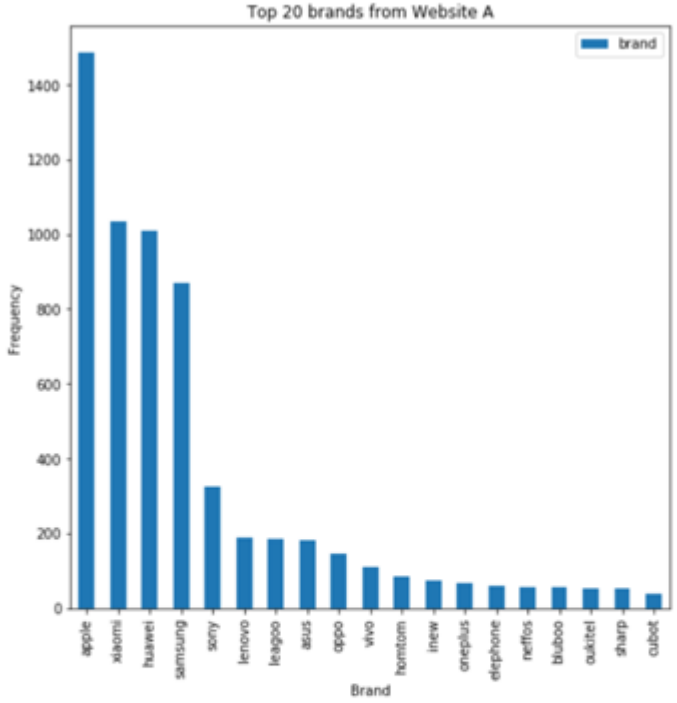
- **t-test**

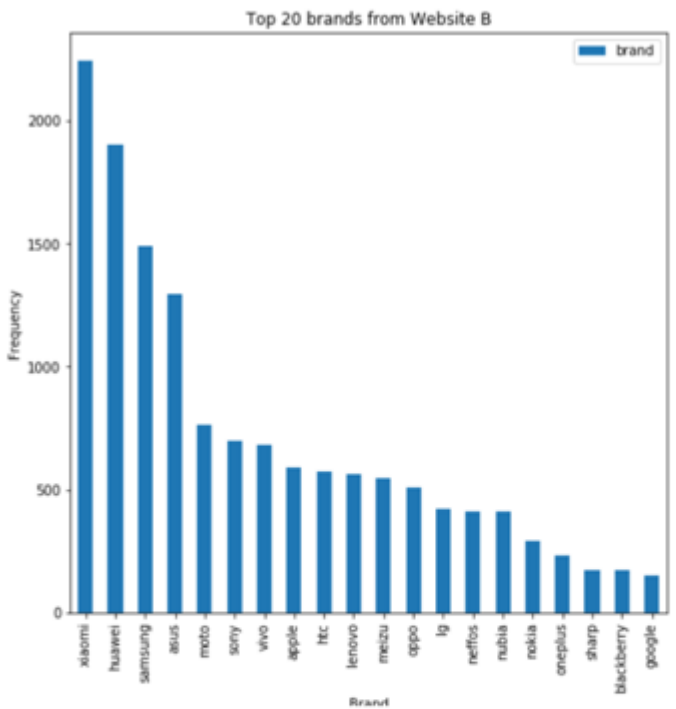
The *t*-test compares two averages (means) and tells if the observations are different from each other. The *t*-test also tells how significant the differences are.

- **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)

PRICE DISPERSION

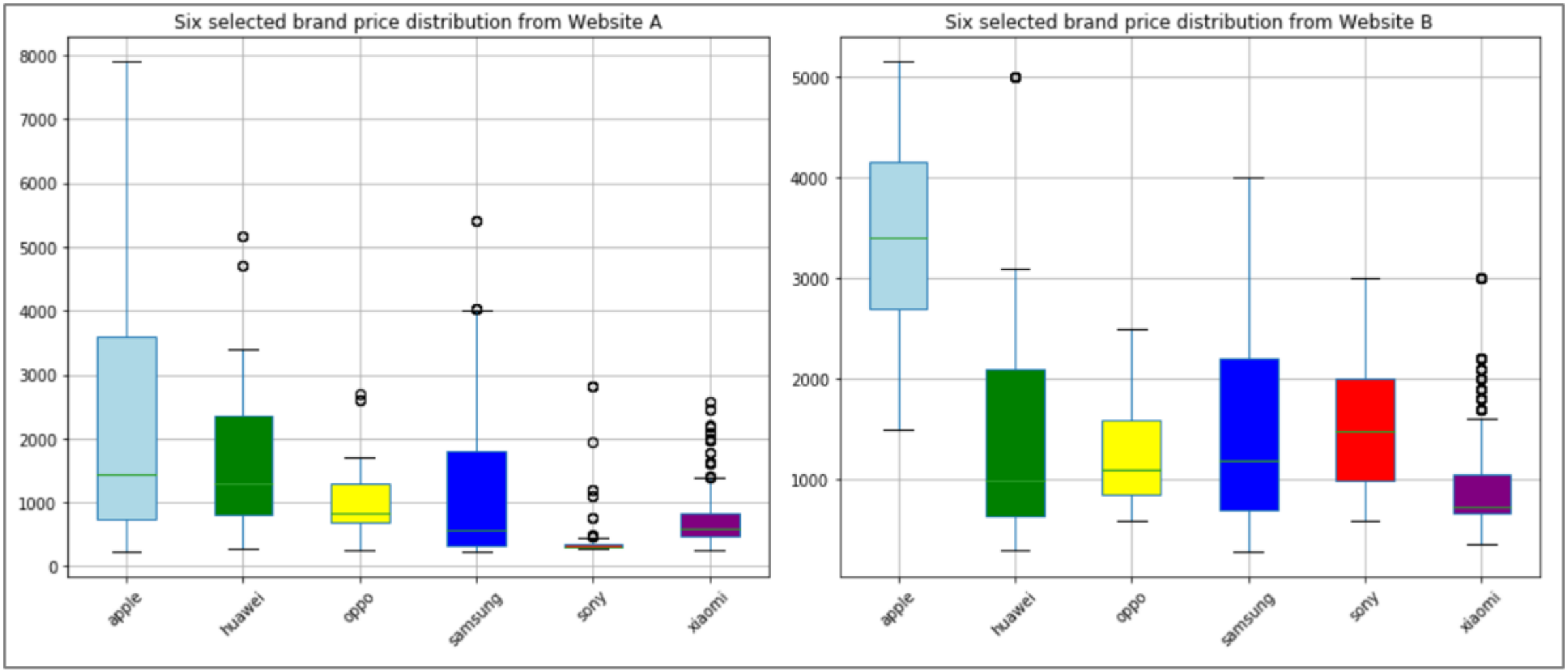




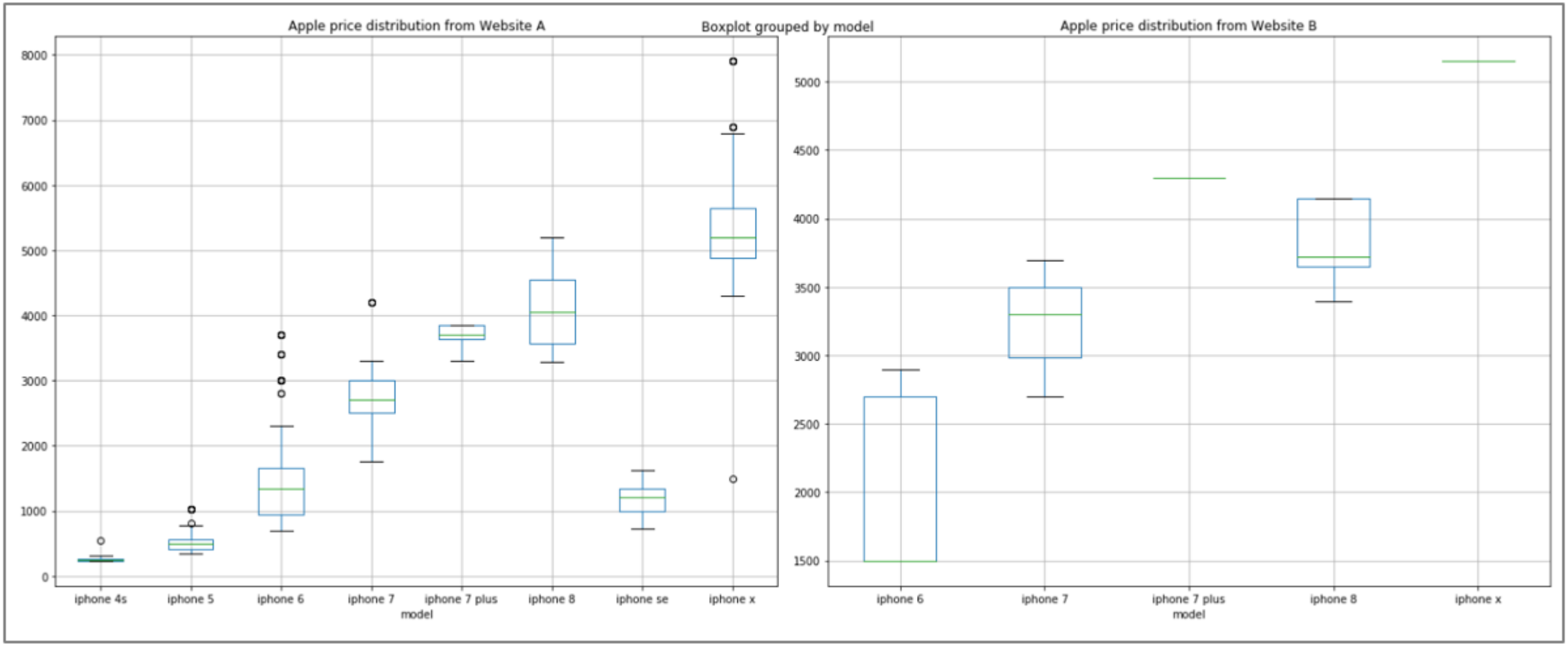
A	apple	xiaomi	huawei	samsung	sony	lenovo	leagoo	asus	oppo	vivo	homtom	inew	oneplus	elephone	neffos	bluboo	oukitel	sharp	cubot
brand	1483	1034	1008	869	326	190	185	182	146	110	83	73	67	58	56	54	53	52	38

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

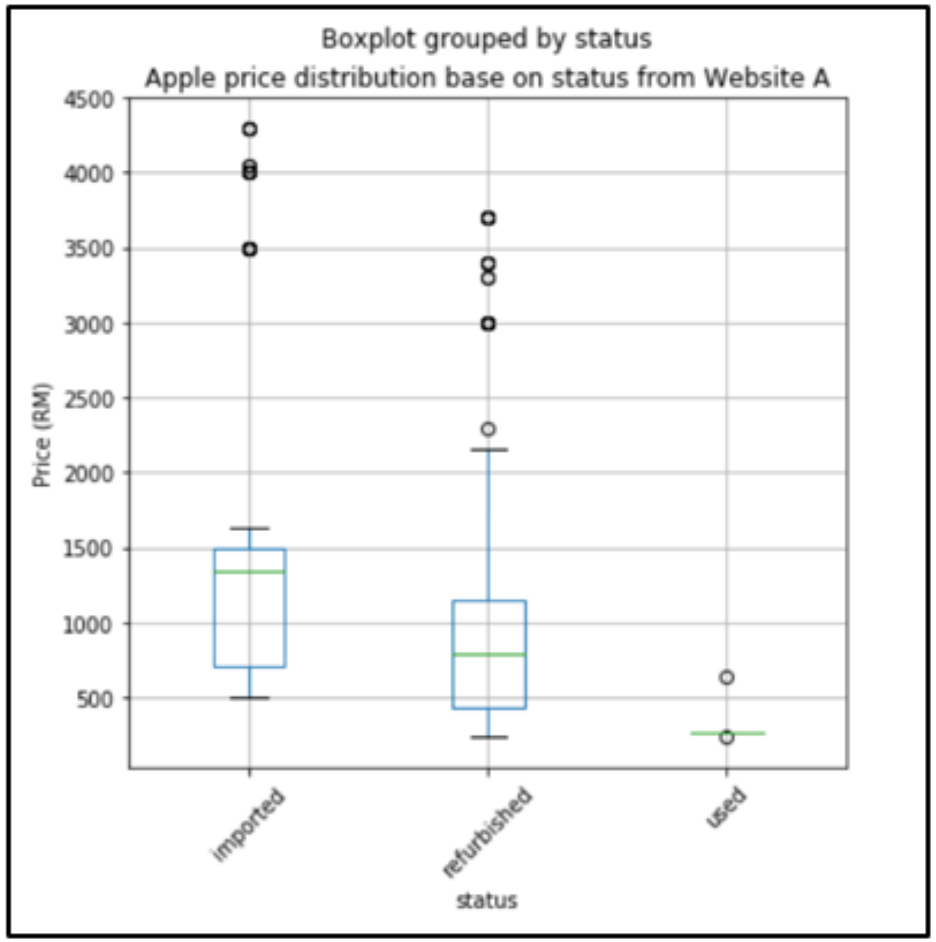


Price distribution for six mobile phones brand from website A and B



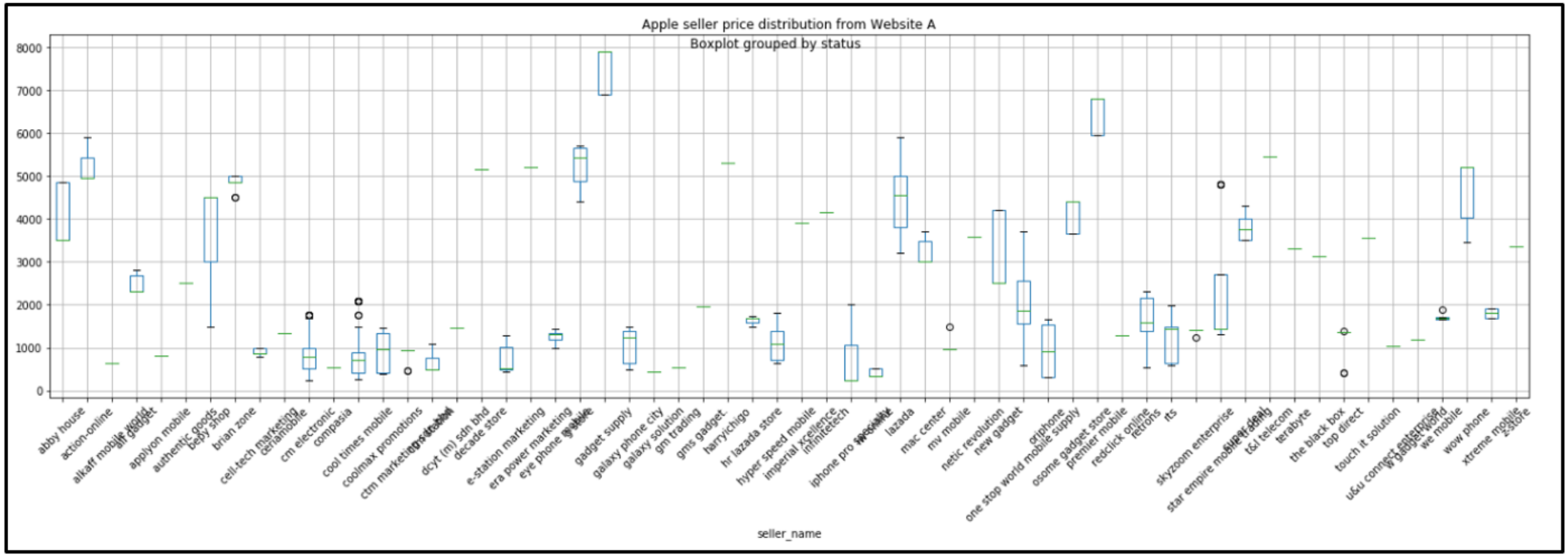
Price distribution for Apple models from website A and B

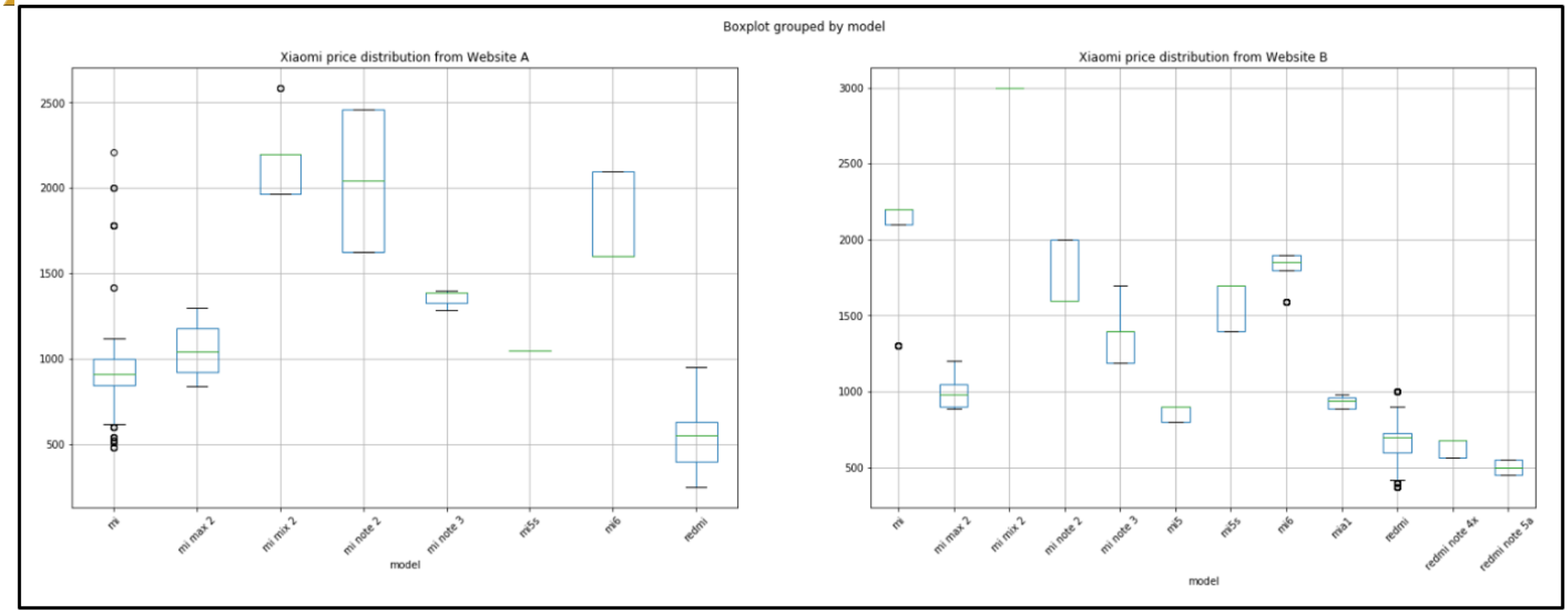
Website A: Apple



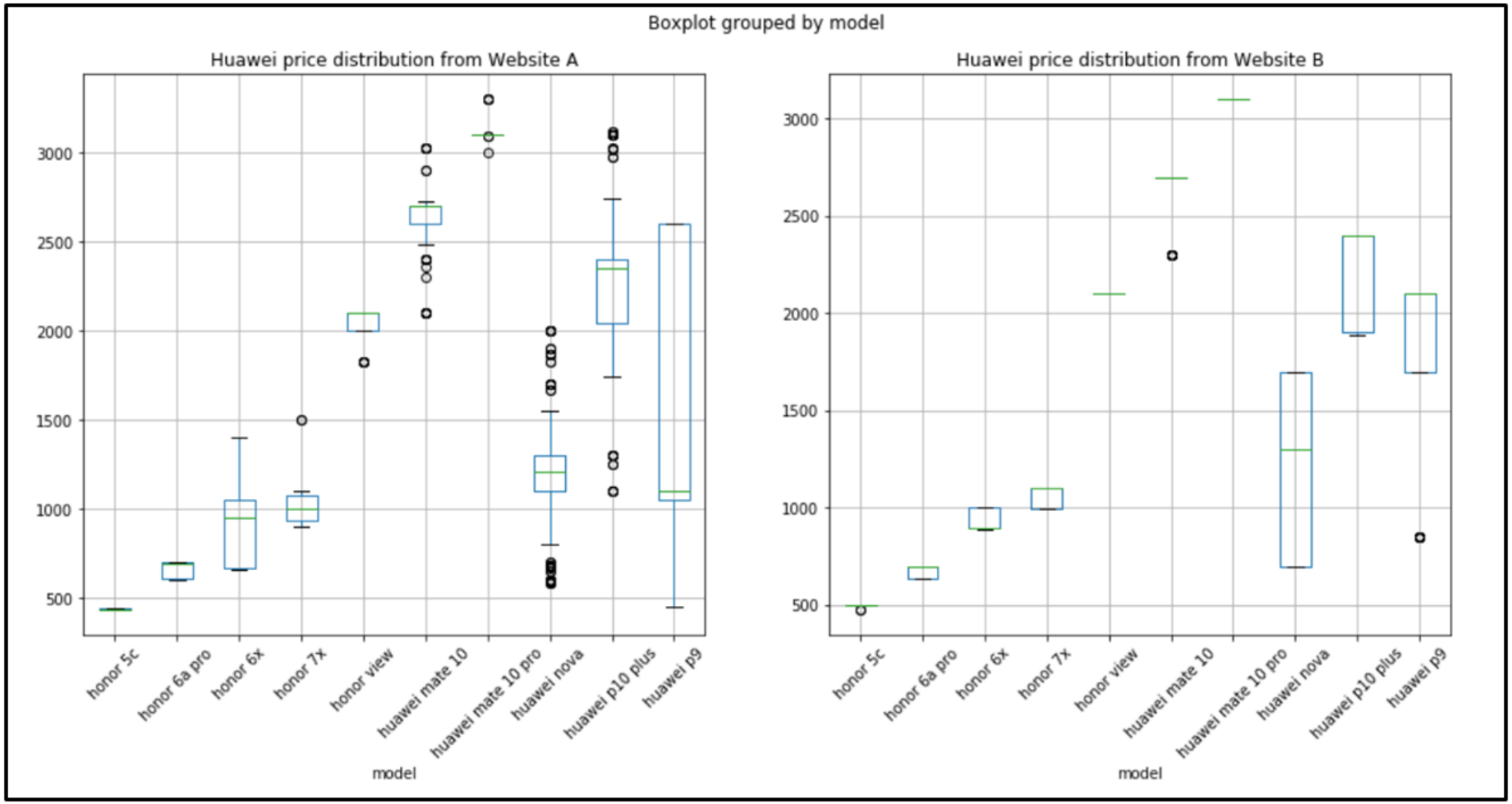
- 8 Apple model from Website A and 5 Apple model from Website B.
- The price range are different between website A and B.
- Lower price below RM 1,000 in website A occurred because used and refurbish set of iPhone is offered from this website

Website A: Apple



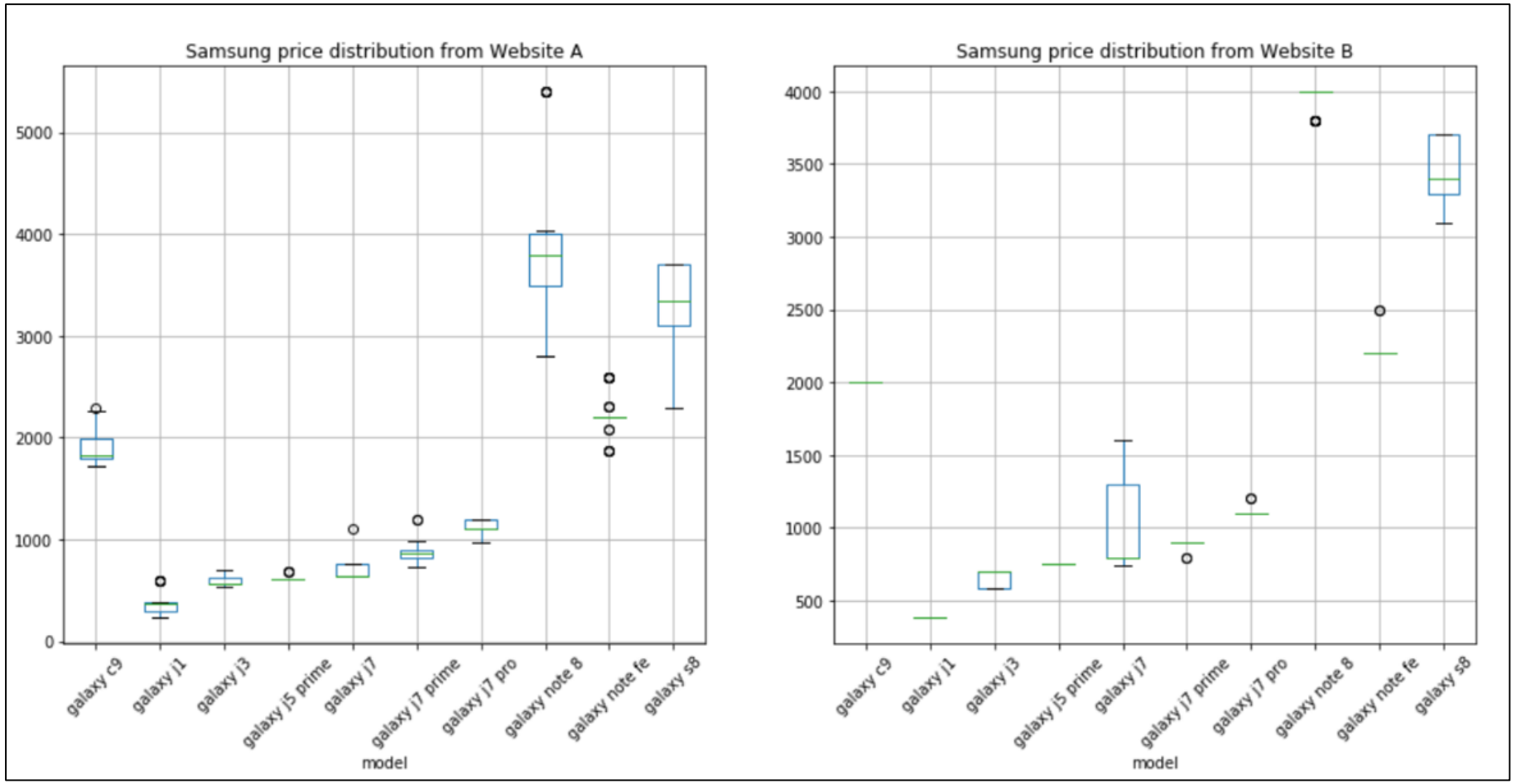


There are 8 Xiaomi model from Website A and 12 Xiaomi model from Website B.
The price range from both websites is almost the same



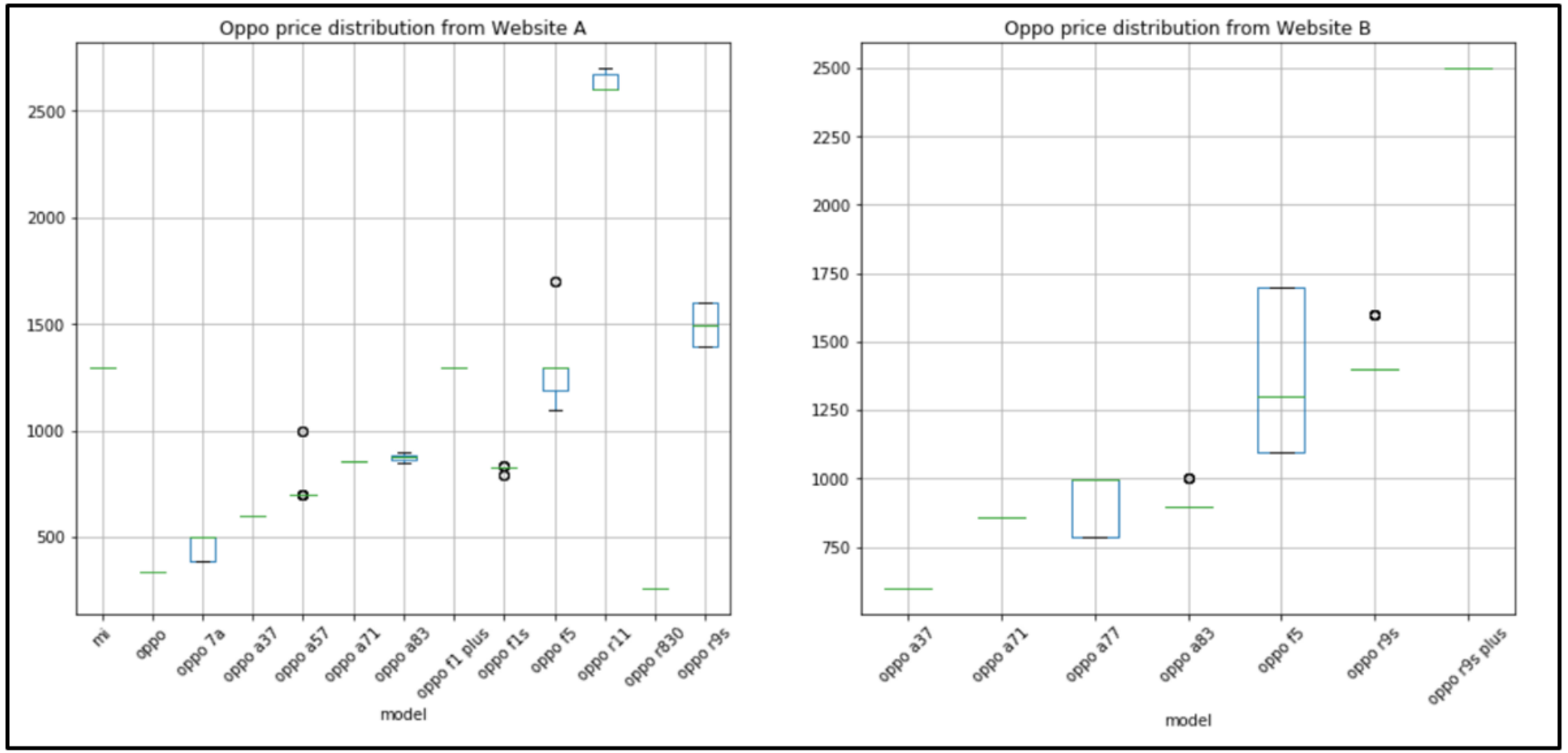
There are 37 Huawei models from Website A and 25 Huawei models from Website B.

Only 10 match models were selected from both websites but must include Huawei P10 because it is one of mobile phone item in physical price data collection



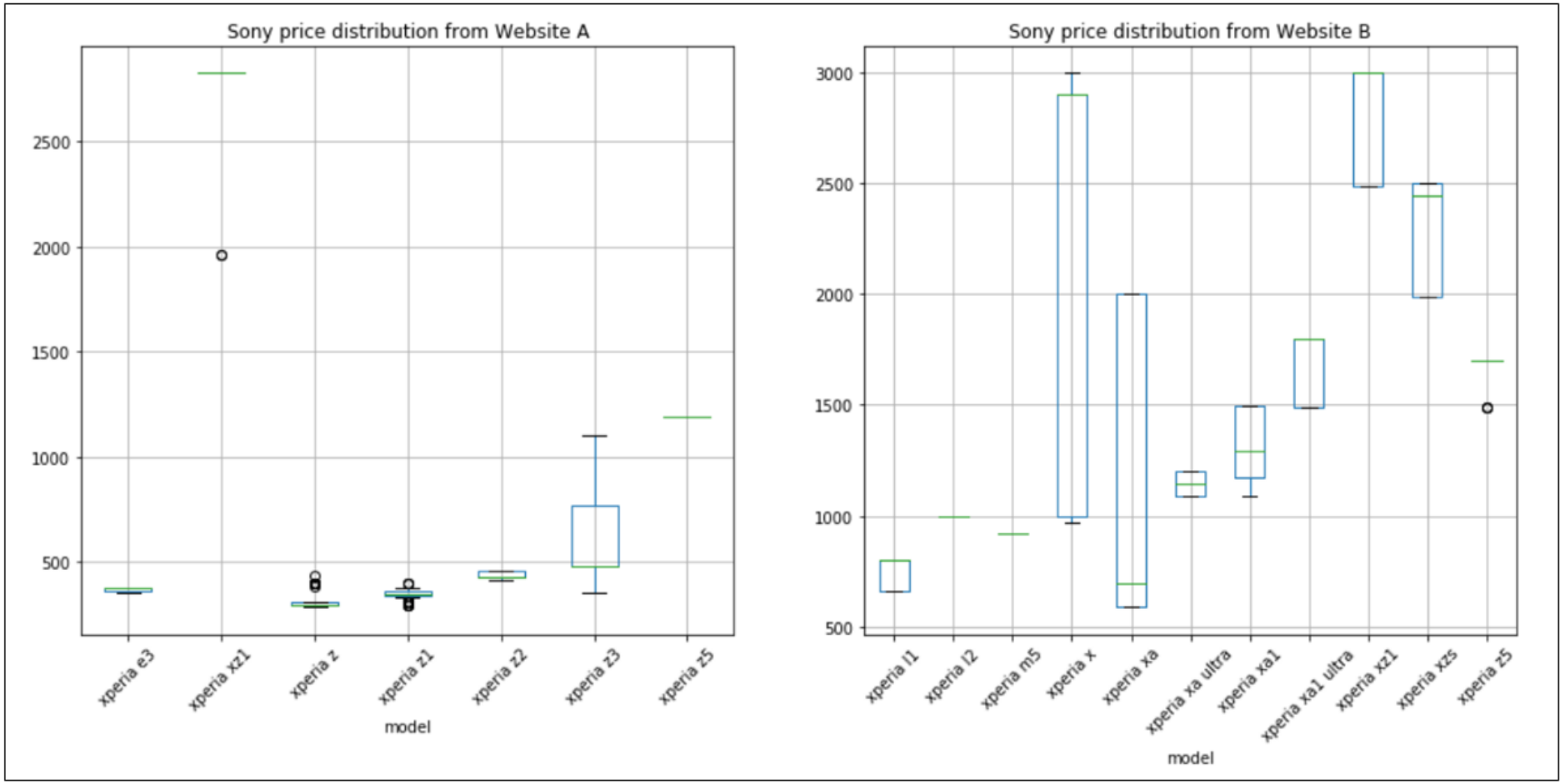
There are 34 Samsung models from Website A and 21 Samsung models from Website B.

Ten match models were selected from both websites but **must include Samsung J3 Pro and Samsung S8** as it is one of mobile phone item in physical price data collection



13 Oppo models from Website A and

7 Oppo models from Website B



7 Sony models from Website A and

11 Sony models from Website B

T-TEST

T-test : Hypothesis 1

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**

T-test : Hypothesis 2

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:

Brand	Welch Two sample t-test					Decision
	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

- 5 Huawei model is significantly different between website A and B

Brand	Welch Two sample t-test					Decision
	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

- 1 Oppo model is significantly different between website A and B

Hypothesis 2:

Brand	Welch Two sample t-test					Decision
	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

- 6 Samsung model is significantly different between website A and B

Brand	Welch Two sample t-test					Decision
	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

- 1 Sony model is significantly different between website A and B

T-test : Hypothesis 3

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.

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)

CONCLUSION & RECOMMENDATION

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

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.
- Conducting regression analysis to have better understanding of the **influenced factors** of the price (e.g. **shipping cost, seller ratings, warranty, loyalty/membership**, etc) which can also be obtained through more enrichment data process.

TERIMA KASIH & THANK YOU

HARI STATISTIK NEGARA	62nd ISI World Statistics Congress ISI 2019 Kuala Lumpur 18 - 23 August 2019	POPULATION & HOUSING CENSUS MALAYSIA 2020 YOUR DATA IS OUR FUTURE	HIES/BA 2019 Data Shaping People's Wellbeing	WRC 2019 Bancian Perdagangan Borong & Runcit Wholesale & Retail Trade Census	SUSTAINABLE DEVELOPMENT GOALS MALAYSIA
20 OCT	18 - 23 AUG 2019	JULY 2020 (ACTUAL MYCENSUS)	JAN – DEC 2019	MAR – SEPT 2019	2015 - 2030

#StatsMalaysia | #MyStatsDay | #ISIWSC2019 | #MyCensus2020 | #HIES2019 | #MyRetailCensus2019 | #LeaveNoOneBehind