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

Nur Hurriyatul Huda Abdullah Sani[1]

ABSTRAK

Maklumat berkaitan harga barangan dan perkhidmatan dalam talian semakin mendapat tempat sebagai salah satu pengukuran indikator ekonomi dan aplikasi penyelidikan. Pengumpulan data harga dalam talian adalah salah satu daripada program utama yang terlibat oleh majoriti agensi statistik rasmi di seluruh dunia. Organisasi Statistik Kebangsaan (NSOs) telah mula meneroka kaedah dan faedah penggunaan data harga menerusi platform dalam talian bagi penentuan Indeks Harga Pengguna rasmi (IHP). Jabatan Perangkaan Malaysia (DOSM) melalui inisiatif Analitik Data Raya (StatsBDA) pada tahun 2017 telah membangunkan satu portal harga yang dikenali sebagai Price Intelligence (PI) dengan matlamatnya sebagai pendekatan alternatif dan pelengkap untuk kaedah pengumpulan data harga. Melalui inisiatif ini, sejumlah besar data dari laman web e-dagang telah dikikis setiap hari. Projek mercu ini mempunyai dua tumpuan utama, pertama untuk membincangkan kaedah yang digunakan dalam penyediaan data dan proses perbersihan data dan kedua untuk menganalisis corak harga serta penyebaran harga dalam talian bagi item terpilih, iaitu telefon bimbit. Telefon bimbit dipilih dalam projek kerana ia adalah salah satu item dalam bakul harga. Selain itu, kaedah penyediaan data boleh direplikasi untuk item lain yang mempunyai ciri yang sama seperti item dari kategori elektronik. Hasil utama membuktikan kewujudan perbezaan harga bagi telefon bimbit yang dijual secara dalam talian berbanding harga purata melalui kaedah pungutan secara tradisional menerusi lawatan ke kedai atau outlet fizikal. Purata harga dalam talian adalah lebih tinggi berbanding harga purata dari kedai fizikal untuk model telefon bimbit tertentu.

Kata kunci: Pengikisan Web, Perangkak Web, Harga Dalam Talian, Analisis Kluster

ABSTRACT

Online price information is increasingly being used as an economy indicator measurement and research application. The data collections of online prices are one of the main programs 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). The Department of Statistics Malaysia (DOSM) through the Big Data Analytics initiative (StatsBDA) in 2017 has developed an intelligence price portal known as Price Intelligence (PI) with its aim as an alternative and compliment approach for the price data collection method. Through this initiative, large amount of data from e-commerce website has been

[1] Nur Hurriyatul Huda Abdullah Sani is currently Senior Assistant Director of Malaysian Bureau of Labour Statistics (MBLS), Department of Statistics Malaysia.

scraped on daily basis. This capstone project has two main focuses, first to discuss the methods used in the data preparation and wrangling process and secondly to discuss the online price pattern and dispersion analysis on selected item i.e. mobile phones. Mobile phones are chosen in the project because they are one of the items in the price basket. Besides, the method of data preparation can be replicated for other items that have similar features such as electronics category. The main finding validated the existence of price differences inter websites and the average price data collection through physical outlets. It is also found that the average online prices are higher than average prices from physical outlets for certain mobile phone models.

Keywords: Web Scraping, Web Crawling, Online Price, Clustering Analysis

1. INTRODUCTION

Today’s world revolves around data and information. The amount of data that has been generating is so big and requires good data management to utilise the value of the data (Agarwal & Dhar, 2014). Data can be in various form of which mainly are structured and unstructured data. Dobre and Xhafa (2014) reported that every day, the world produces around 2.5 quintillion bytes of data (i.e. 2.5 Exabyte or 2.5 billion gigabytes), of which 90.0% of the data is unstructured data. Technology advancement has created a huge diversity of new data sources. The largest source of data today is online data. Almost all transactions, are carried out online and it is challenged to analysed them. Previously the data are stored and analysed in a structured form as it is easy to conduct descriptive, predictive or even prescriptive analyses.

The UN Statistical Commission has started exploring and identifying new information sources through technological advances to provide better data for official statistics. One of the new information sources is online price data. Online price information is increasingly being used as economy indicator measurement and research application. The data collections of online prices are one of the main initiatives 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) (Cavallo, 2017).

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 in Malaysia consists of 13 main groups i.e. 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 and Miscellaneous Goods & Services and Health & Medical Care. The prices used in the calculation of the CPI are retail prices or transacted prices including all taxes imposed on those goods and services.

The aim of this study is to utilise web-scraped online price data as an alternative source for analysing mobile phone prices. The study will first identify and compare the pricing patterns of mobile phones across various online platforms. In the second part, the analysis will explore the price dispersion of mobile phones between online retailers and physical outlets. Lastly, the study will examine the factors influencing mobile phone prices.

2. LITERATURE REVIEW

Department of Statistics, Malaysia (DOSM) is currently using method in price data collection for the preparation of CPI. DOSM conducts monthly price data collection by visiting the selected outlets in 14 states in Malaysia which include both urban and rural areas. The monthly price data collection is conducted by DOSM's field enumerators. However, in line with the global initiative, DOSM through the Big Data Analytics initiative (StatsBDA) in 2017 has developed an intelligence price portal known as Price Intelligence (PI) which is an alternative and complimentary approach of the data collection method. Through this initiative, large amount of data from e-commerce website can be scraped on daily basis.

Since online shopping is growing rapidly, e-commerce websites are also growing due to the trends of consumers purchasing behaviour via online. According to Wee Huay Neo, Director of e-commerce Enablement at MDEC, as reported in *The Sun Daily* on 6 February 2017, Malaysia has approximately 19 million online users. Among these, 7.0% of online shoppers make purchases almost daily, 26.0% shop weekly, 54.0% once a month, and 13.0% shop only once a year (*The Sun Daily*, 2017). Survey conducted by openmindsresources.com in 2012 showed that 65.0% Malaysians shopped online while 35.0% stated they did not do online shopping. There are 8 categories of items that are often bought online with the clothing & accessories category being on the top of the list with 69.0% followed by Electronics, 43.0%. The other six categories are computers (31.0%), DVD, music & books (29.0%), Sports & outdoors (22.0%), Jewellery & Watches (21.0%), Gifts & Flowers (16.0%) and Toys & Baby products at 13.0%.

2.1 Web Scraping

Through the pilot study and research conducted, NSOs agree that the use of web scraping can serve as a complement method for consumer price data collection but should not replace the traditional approach (Office for National Statistics, 2017; European Commission, 2017). Web scraping is the process of extracting information from the websites using web scraper tools. This is an automated process of retrieving information from websites using its content and structure, to identify and get the relevant information. This method has potential to become new sources of compiling the CPI. On top of physical data collection through outlet visit, using web scraping tools is another way to complement the data collection method. The European Commission has conducted a Multi-Purpose Price Statistics (MPS) project with the purpose of modernising data collection methods. 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 to collect prices for electronic goods and airline tickets fares. Istat found that there was an improvement in terms of 30.0% time saving in carrying out price quotations for electronic goods (Polidoro et al., 2015). Meanwhile Statistics Austria (Josef Auer and Ingolf Boettcher, 2016) and Norway Statistics (Nygaard, 2015) using Import.io software as web scraper tools. Statistics Norway has started a pilot study for price data collection using web scraped as early as 2014. According to Nygaard (2015), 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. 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 successfully collected an estimated 6,500 daily price quotations, totalling approximately 200,000 price quotation per month. This volume of price quotations is significantly higher compared to those collected using the traditional approach (Breton et. al., 2015). Statistics Singapore uses open source software, such as Import.io, as a web scraper to gather prices for airline tickets fares. Two types of web crawlers that have been used, such as customized web crawlers and the 'point-and-click' method (Chuanyang & Joseph, 2016).

2.2 Price Pattern Analysis

The growth of internet users globally is undeniable. The Hand Phone Users Survey 2017 (HPUS 2017), conducted by the Malaysian Communications and Multimedia Commission (MCMC) since 2012, reports a 26.0% increase in smartphone users in Malaysia who access to the internet via their mobile devices. This widespread use of the internet has permeated nearly every aspect of daily life, presenting a significant opportunity for retailers to capitalize on the internet as a business platform.

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” (Pan et al., 2002). Petrescu (2011) aims to analyse the price dispersion for online shopping mall and to identify factors that affected the price dispersion. The study was conducted on 100 watches models from seven brands, and 100 camera models from nine brands offer on Amazon.com. 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 indicate that price dispersion is influenced by the average price of goods, types of goods, shipping costs and user reviews and ratings. The number of sellers does not affect the price dispersion because users can easily make price comparisons across sellers. Cavallo (2017), through the Billion Prices Project shows that there is a price difference between price collection through physical outlet and online price collection using web scraping technique. Nevertheless, 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 focused on multi-channel retailers, which operate both physical outlets and online platforms. Cavallo recommends that NSO consider using online price data as a complementary tool for calculating the Consumer Price Index (CPI), particularly for items in the clothing and electronics categories.

Based on studies by Boeing and Waddell (2017), using descriptive statistics, it can be demonstrated the market price pattern of home-rental rates and property values for urban areas throughout the US. However, if there are longer time series data, time series analysis can be conducted to predict more precise home market prices.

There are a wide range of topics that can be analysed using online price data collection such as price dispersion, price comparison, price stickiness, real exchange rate, market segmentation and international relative prices (Chevalier et al., 2003). Even though one of the aims for PI initiative is to explore new method in calculating CPI, however before we can proceed with that, it is best to understand the challenges and pattern of online data that has been scraped from the websites. Therefore, the aim of this study is to utilise the web scraped of online data price as an alternative data source by performing analyses to identify and compare the price of selected item and its patterns from different online sources, to analyse the price dispersion of selected item from online sources and physical outlets data sources, and finally to identify factor that give impact to the phone price.

3. METHODOLOGY

The online price data collection used web scraping technique to collect data on daily basis from January 2018 to December 2018. The data source used in this project is from DOSM. Online price data that has been scraped from four websites and the average price data through traditional field work data collection. DOSM developed web scrapers to scrap data from more than 20 e-commerce websites. The scrapers are programmed using python selenium package to automate the web browser interaction from python. The initial output obtained from the online data that was scraped is in a semi structured form. Therefore, the dataset needs to go through data wrangling process before it can be used for further analysis.

The web scraper collected approximately 176,000 price quotations per day for average 2,800 of unique items, which approximately 5.2 million price collection per month for these four websites. Meanwhile, in traditional price collections, on average collects 167 thousand of price quotation monthly from 551 groups of goods. There are days that have no data recorded and it is called ‘missing days. Missing days was due to websites structure change, internet and other IT system failure. It is also possible that the item is not available on that day as it became out of stock.

There are wide range of topics can be analysed using online price data collection. In this capstone project, analysis is done on price pattern and dispersion inter websites A and B for item category mobile phones. Analysis was conducted using descriptive statistics, k-mean clustering and t-test to understand the price pattern distribution and dispersion (if any) for websites A and B for month of January and February 2018.

3.1 Clustering Analysis

The 11th Malaysia Plan (RMK-11), under the 6th pillar on the digital nation, aimed to reduce the digital gap in urban and rural area, old and young people, and various income groups. At this moment, the focus is given on the affordable "internet line" for the people to subscribe. However good internet line needs to be supported with good device to facilitate access to information. According to Household Expenditure Survey (HES) report 2016, Malaysians allocate 5.0% for expenses under communication groups. The monthly mean household consumption expenditure for Malaysia in 2016 is RM 4,033, i.e. RM 4,402 in urban area and RM 2,725 in rural area. Therefore, having clustering group of mobile phones, it provides an overview, phone and specification that can be found in Malaysia. In this project, clustering analysis using K-means was carried out to get a rough idea of the mobile phone group that is in the market based on price, specification and brand.

Clustering is a grouping of data that share similar features together in the same group. K-mean is one of the most commonly used. The K-mean clustering algorithm is used to find groups which have not been explicitly labelled in the data. The K-mean clustering is used because of a few reasons. K-mean is the most popular algorithms used for clustering practice because of its simplicity and speed. It can be applied to large size of dataset that has a small number of dimensions, numeric, and continuous. This clustering technique also commonly used for market segmentation analysis.

3.2 T-test Analysis

The t-test is a statistical method used to compare two means and determine whether they differ significantly from each other. The t-test also tells how significant the differences are; in other words, it lets the readers know if those differences could have happened by chance. In this project, Welch’s t-test is used to understand if there is price difference between website A and website 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, that is why the unequal variance t-test (Welch’s t-test) is used.

3.3 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). In this study, both simple and multiple regression analysis will be used to identify which variables have impact on the mobile phones prices. There are four independent variables will be used in the price dataset i.e. brand, model, storage size and memory size to identify which factors that may influence the price of mobile phones.

4. RESULT

Total number of data falls under category mobile phones are 7,206 for websites A and 21,830 for websites B. However, there are at least five reasons found to remove 6.0% of the total data from both websites due to wrong categorization which the data refer to phone accessories, data with no price, phone name not clear (mix up more than one brand within same observation) and finally wrong and not enough information to categorise the data. The total number of unique phone brand in websites A is 31 with 128 models, while there are 28 unique phone brands with 108 models from website B.

Through the statistical test conducted, price differences exist between website A and B. It is also found that 22 out of 38 phones models tested using t-test, are statistically difference between both websites. Price differences also exist for website A, website B as compared to the average price data collection through physical outlets.

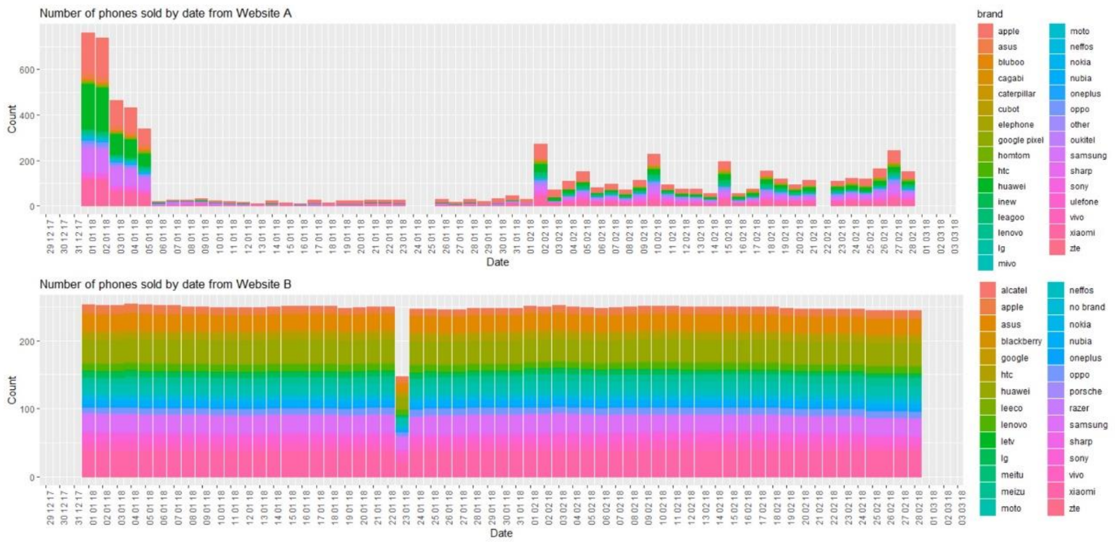

Figure 4.1: Number of Phones Offer by Date from Website A and Website B

From Figure 4.1, it was found that website A is quite fibrous in terms of quantity and price of phones offered daily as compared to website B. This is due to website A nature of business as it is a global marketplace with many sellers selling phones from the same brand. There are 240 unique sellers for phone from website A, while website B is the single seller.

4.2 Clustering

The purpose of the clustering analysis is to identify and visualise patterns in the mobile phone data, grouping similar phones together based on their characteristics such as brand, model, storage size, and memory size. By applying the K-means clustering algorithm, the analysis aims to reveal distinct groups or clusters of phones, helping to understand how different phones relate to one another in terms of pricing and specifications. This can provide insights into price segmentation and help in identifying trends or anomalies within the dataset.

The phone pattern can be visualised using clustering analysis. K-mean is one of the most commonly used. From the elbow plot, the number of cluster k that can be chosen is either k=5 or k=8. Using r squared as an explained variance, to validate the best number of clusters, cluster 5 obtained 70.0% and cluster 8 have 80.0% of variance explained. However, the clustering for both k=5 and k=8 indicates duplicates point for each group. When cluster sum of square of the distance from each data point (‘withinss’) is checked, the value is high for both but higher at k=5. The between-cluster distance (‘betweenss’) for k=8 give slightly higher sum of the squared distance compare to k=5. Ideally the cluster centres should be far apart from each other. Hence, k=8 is selected for both website A and B. The clustering result for both websites A and B does not show the best result as there is an overlap in the phone model and the average price in each cluster. The cluster analysis can be improved by adding other phone specifications as value added in making comparisons such as screen sizes, camera, touch pens, and other security features. The detailed data enrichment can be done by data matching base on the specification from the official website through another web scraping exercise.

4.2 Price Dispersion

The analysis is conducted to identify if there is price difference between websites A and B, and also to identify the price difference for mobile phones sold online compare to physical outlets average prices. Top ranking mobile phones has been selected and matched with the physical outlets price data collection.

The box plot in Figure 4.2 below shows an obvious huge price distribution obviously for Apple, Samsung and Huawei brands. Outlier price exist for Xiaomi and Sony and some in Huawei, Oppo and Samsung. This is due to influenced by the models in each of the brands.

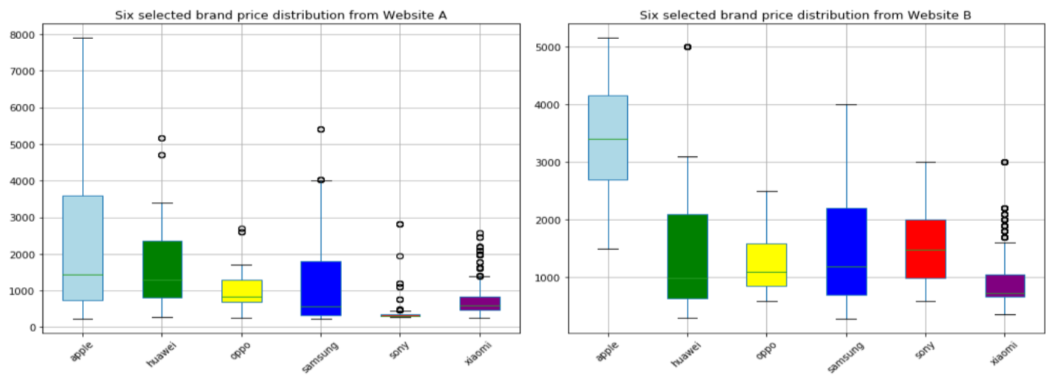

Figure 4.2: Price Distribution for Six Mobile Phones Brand from Website A and B

4.3 T-test

T-test is conducted for three hypotheses at 5% significance level.

4.3.1 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.

All p-value in Table 4.1 are lower than significant level of 5%. Therefore, the Null Hypothesis is rejected and concluded that in overall, these six brands have different prices between website A and B.

Table 4.1: Result for Mobile Phone Brands

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

4.3.2 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.

Based on the results presented in Table 4.2, it can be concluded that the price for Apple brands products on website A are different with price in website B. However, as shown in Table 4.3 to Table 4.7, certain models on website A have a p-value smaller than the 5% significance level, therefore the models are statistically different from the average price in website B, while others are not significantly different.

Table 4.2: Result for Mobile Phone Apple

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

Table 4.3: Result for Mobile Phone Xiomi

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

Table 4.4: Result for Mobile Phone Huawei

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 mate10	0.28611	137.05	0.7752	2618.476	2609.345	Fail to Reject Ho
huawei mate10 pro	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

Table 4.5: Result for Mobile Phone Samsung

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

Table 4.6: Result for Mobile Phone Oppo

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

Table 4.7: Result for Mobile Phone Sony

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

Prices from physical outlets were collected for five models which were Huawei P10, Samsung Galaxy J3 Pro, Samsung Galaxy S8 (64GB), Apple Iphone 7 Plus (32GB), and Oppo RS9. However, for comparison with between online prices and physical data collection average price, only Samsung Galaxy S8, Apple IPhone 7 Plus (32GB), and Oppo RS9 were used as both websites A and B do not have Samsung Galaxy J3 Pro and Huawei P10 models for the same month.

4.3.3 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.

When comparing the average price from online and offline outlet, it is concluded that online prices are higher than the physical outlets average prices based on statistical test in Table 4.8 below, the p-value is lower than 5% significance level. Thus, null hypothesis is rejected.

Table 4.8: Result for 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

4.4 Regression Analysis

Table 4.9 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.

Table 4.9: Summary Result for Regression Analysis of Website A

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	**Memory** 0.1485	
model + Storage(rom)	367.7	< 2.2e-16	0.9559	**Storage** 0.0158	
model + Memory (ram)	364.3	< 2.2e-16	0.9567	**Memory** 0.0166	
Brand + Storage (rom) + Memory (ram)	656.8	< 2.2e-16	0.8592	**Memory** 0.0298	**Storage** 0.0703
model + Storage(rom) + Memory (ram)	**337.6**	**< 2.2e-16**	**0.9628**	**Memory** 0.0069	**Storage** 0.0061

However, referring to Table 4.10, it is shows that brand, model and storage have high impact on the phone prices sold in website B. Table 4.10 shows that regression model with variables phone brand, storage (ROM) and memory (RAM) give the best model with high adjusted R-squared and the lowest residual standard error. Brands and models cannot be in the same model as multicollinearity exists.

Table 4.10: Summary Result for Regression Analysis of Website B

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	**Memory** 0.1066	
model + Storage(rom)	490.6	< 2.2e-16	0.9194	**Storage** 0.0655	
model + Memory (ram)	518.3	< 2.2e-16	0.9101	**Memory** 0.0562	
Brand + Storage (rom) + Memory (ram)	444.3	< 2.2e-16	0.9339	**Memory** 0.027	**Storage** 0.0103
model + Storage(rom) + Memory (ram)	461.2	< 2.2e-16	0.9288	**Memory** 0.0094	**Storage** 0.0187

This is not the best regression model to be used because other factors are not fully considered such as multicollinearity, autocorrelatiom, heteroscedasticity and outliers. Further analysis can be cariied 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.

5. CONCLUSION AND RECOMMENDATION

This study concludes that base on the data source, there exist price differences between website A and website B, and there also exist price differences between average online price and the average price data collection through physical outlets. The price differences are basically contributed by the brands and the specifications of a model. Surprisingly the study shows that the online prices are higher than average price data collection through physical outlets. 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 can also be obtained through more enrichment data process. The study can be extended by comparing the consumer price index for goods online and existing field methods.

This study paves the 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 to control the inflation in the country.

REFERENCES

Agarwal, R., & Dhar, V. (2014). *Big data, data science, and analytics: The opportunity and challenge for IS research*.

Boeing, G., & Waddell, P. (2017). New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. *Journal of Planning Education and Research*, *37*(4), 457–476.

Breton, R., Flower, T., Mayhew, M., Metcalfe, E., Milliken, N., Payne, C., ... & Woods, A. (2016). Research indices using web scraped data: May 2016 update. *Newport: Office for National Statistics*. Available from www.ons.gov.uk/file.

Cavallo, A. (2017). Are online and offline prices similar? evidence from large multi-channel retailers. *American Economic Review, 107*(1), 283-303.

Cavallo, A., & Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives, 30*(2), 151-78.

Chawla, M., Khan, M. N., & Pandey, A. (2015). Online Buying Behaviour: A Brief Review and Update. *Journal of Management and Research, 9*(2/4).

Data Preparation, Preprocessing and Wrangling in Deep Learnin. https://www.xenonstack.com/blog/data-science/preparation-wrangling-machine-learning-deep/

Dobre, C., & Xhafa, F. (2014). Intelligent services for big data science. *Future Generation Computer Systems, 37*, 267–281.

Malaysians are online shopaholics. (n.d.). http://www.thesundaily.my/news/2151542.

Malaysias Online Shopping Behaviour. (2018). http://www.talented.com.my/malaysias-online-shopping-behaviour/. [18 September 2018]

Mudinas, A., Zhang, D., & Levene, M. (2012, August). Combining lexicon and learning based approaches for concept-level sentiment analysis. *In Proceedings of the first international workshop on issues of sentiment discovery and opinion mining* 5. ACM.

Nygaard, R. (2015). The use of online prices in the Norwegian Consumer Price Index. *Statistics Norway*.

Petrescu, M. (2011). Online price dispersion–more than imperfect information. *Journal of Product & Brand Management, 20*(7), 541-548.

Polidoro, F., Giannini, R., Conte, R. L., Mosca, S., & Rossetti, F. (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.

Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59*(1), 1-34.

Study: Digital interactions dictate majority of in-store spending. (2018). https://www.retaildive.com/news/study-digital-interactions-dictate-majority-of-in-store-spending/426356/. [18 September 2018]

Tang, F. F., & Xing, X. (2001). Will the growth of multi-channel retailing diminish the pricing efficiency of the web? *Journal of Retailing, 77*(3), 319-333.

Xing, X., Tang, F. F., & Yang, Z. (2004). Pricing dynamics in the online consumer electronics market. *Journal of Product & Brand Management, , 13*(6), 429-441.