IS428 Visual Analytics for Business Intelligence

Visualizing Private Car Resale Market in Singapore

Data Tsunami: Dong Ruiyan, Zhang Qian, Lee Ting Kok Jeremy

Abstract—Singapore is lauded to be one of the most expensive cities to own a car, with many costs involved with car ownership that go beyond its purchase price. Hence, the second hand car market has seen a huge growth in recent years as used cars are becoming more popular amongst car dealers and online resale. To understand this growing market and provide better insights for resale car buyers and sellers we designed and developed a dynamic and interactive visual analytics dashboard. We explore and analyse car resale data to gain valuable insights through the usage of treemaps, bubble charts, scatter plot matrix and radar charts. Ultimately, our visual application reflects the existing climate of the car resale market and assists with resale car purchasing and sale decisions.

Key Words – Seconhand car, Car resale market, Treemap, Bubble chart, Scatter plot matrix, Radar Chart

1 Introduction

According to Forbes, there has been a huge increase in demand for used cars, as a result, the used car market has seen a stellar growth of up to 68% since 2009. In fact, latest statistics on used cars published by the Land Transport Authority also shows an increasing trend. January saw 9,281 used cars sold, compared to 7,306 in the same period last year (Junn, 2017). In addition, with increasing COE prices many Singaporeans are turning towards used cars. Many motorists are opting to purchase six- to nine-year-old cars to tide them over, as they hope for COE prices to fall, in anticipation of higher de-registrations (Junn, 2017).

This has led to huge changes in car buying behavior, marketplaces like sgCarMart are one of the key platforms paving way the growth of the used car industry. sgCarMart is one of Singapore's biggest online car resale marketplace. Specifically, it facilitates the resale of cars between a buyer and seller.

As a result, we tried to understand this market and its dynamics by crawling data from the sgCarMart's used car listings website. Data analysed reflects the existing cars available in the market currently.

This paper outlines our research and development effort to design and implement a web-enabled client-based visual analytics tool for supporting the analysis and visualisation of the second-hand car resale market in Singapore. It consists of five sections. Section 1 provides a general introduction of the paper. Section 2 provides an overview of the motivation and objectives of our research efforts. It follows by a detailed outline of our approaches taken. In section 4 we discuss our visualisation designs in greater details and respective insights and findings from reach visualisation. Lastly, the paper concludes by highlighting the future direction of the research and challenges faced.

2 Motivations & Objectives

Our project aims to understand the growing used car market in Singapore to enable better decision making for the different stakeholders (Buyers & Sellers) involved. This is especially relevant in Singapore due to high COE prices and changing consumer car buying behavior - many are turning to the used car market to afford a car.

Prices of new cars can be too expensive for price sensitive individuals to afford, especially in Singapore due to high COE prices. However, through the used car market one will be able to afford the convenience of owning a car. For budget conscious individuals, buying a used can be a great way to save money. On the other hand, owners of existing cars interested to make a sale can enjoy savings from its successful sale. Hence, understanding the used car market can prove to be useful for individuals looking to sell / buy a existing car.

We have created a visualization application that helps users perform the following:

1. Uncover the top 5 most common brands for the Singapore car resale market
2. Visualise age profiles of second hand cars based on: COE Time Remaining, Depreciation, Engine power, Mileage, Price
3. Identify relationships and correlations across different factors affecting resale prices

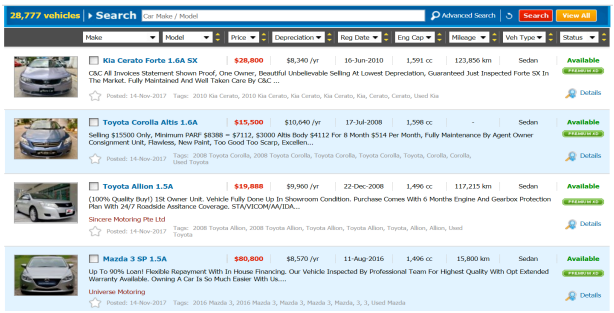
3 Approach

Before we could effectively analyse our data, we have to undergo the data preparation phase. Since there were no publicly available datasets regarding the resale car market in Singapore we had to begin by crawling our data from online sites such as sgcarmart. Thereafter we had to clean the data and preprocess it for it to be ready for analysis.

3.1 Data Crawling

With the lack of an available dataset about the secondhand car resale market in Singapore we had to write a web crawler to attain such data from second hand car resale websites such as www.sgcarmart.com.

Fig. 1. Sgcarmart used car listings.



We then crawled 28,028 records into a csv file with the following attributes: Car brand, Model, Price, Depreciation, Registration Date, Engine Power (cc), Mileage (km), Vehicle Type, Availability Status, Post Date as seen in figure 2 below.

Fig. 2. Crawled dataset into csv format

Car Model	Engine Po	Mileage (k	Depreciati	Price	Brand	Country	Continent	PowerBy	COE Rema	Car Age	Post date	Registration Date	Status	Vehicle Type
Audi A6 2.0A TFSI MU	1984	43000	15940	121999	Audi	Germany	Europe	Petrol	7	3	25/10/2017	25/3/2014	Available	Luxury
Audi Q3 2.0A TFSI Quattro	1984	37000	14960	117999	Audi	Germany	Europe	Petrol	7	3	25/10/2017	30/5/2014	Available	SUV
Audi A3 Sportback 1.4A TFSI S-tro	1395				Audi	Germany	Europe	Petrol	6	4	25/10/2017	27/11/2013	SOLD	Hatchback
Audi A4 1.8A TFSI MU	1798	82000	13380	106999	Audi	Germany	Europe	Petrol	7	3	25/10/2017	31/12/2014	Available	Luxury
Audi A6 2.0A TFSI MU S-Line	1984	117389	12430	47288	Audi	Germany	Europe	Petrol	2	8	28/10/2017	29/10/2009	Available	Luxury
Audi A4 1.8A TFSI MU	1798	95000	13300	72800	Audi	Germany	Europe	Petrol	5	5	28/10/2017	15/2/2012	Available	Luxury
Audi A4 1.8A TFSI MU	1798	83000	13260	93800	Audi	Germany	Europe	Petrol	6	4	28/10/2017	28/8/2013	Available	Luxury
Audi RS5 Coupe 4.2A FSI Quattro	4163	27800	26380	138900	Audi	Germany	Europe	Petrol	3	7	28/10/2017	20/12/2010	Available	Sports
Audi R8 5.2A FSI Quattro R-tronic	5204	75000	54280	228888	Audi	Germany	Europe	Petrol	3	7	28/10/2017	12/4/2010	Available	Sports
Audi A5 Sportback 2.0A TFSI Quat	1984	23300	15780	149000	Audi	Germany	Europe	Petrol	8	2	30/10/2017	29/6/2015	Available	Luxury
Audi A4 1.8A TFSI MU	1798				Audi	Germany	Europe	Petrol	6	4	25/10/2017	22/4/2013	SOLD	Luxury

3.2 Data Cleaning

As the data is crawled from an online source, there were a huge amount of data cleaning and pre-processing to be done before we can effectively analyse the data.

Missing Values: some values are missing, such as mileage, depreciation/ yr, engine power, and prices especially from sold cars. In addition, there we were multiple missing values and inconsistent values to indicate unavailable data such as "nil" / "null" / "-". Hence, we had to standardise and change such values to an empty value so it can be recognised by the JavaScript programme.

Furthermore, there were different naming convention for the same data- due to input errors from the users. This is especially common as data was from a crowdsource database. Hence, this has resulted in names of the same car model to be inconsistent. We had to standardise and aggregate the different naming conventions Eg "MercedesBenz", "Mercedes-Benz", "Mercedes Benz" to a single usable format.

3.3 Data Preprocessing

To assist in giving more context and value from the raw data we had to undergo a certain degree of data preprocessing.

These are the additional calculated fields and preprocessing we have added into the raw data:

Car Age

The age of the car was not available on the website. Hence, we had to calculate car age based on its registration date. Car Age = Existing timestamp – registration data.

COE Remaining

As the remaining COE of a car is not available from the crawled data. We had to also calculate it's remaining COE from its existing car age.

If (Car Age < 10 years) COE Remaining = 10 years – Car Age.

Else COE Remaining = 10years - (Car Age - 10years).

*Assuming COE renewal is 10 years per renewal.

Geographical data

There were no geographical data indicating which country & continent the car brand originated from. Hence, we had to create our own database and a lookup table to append the country and continent information to our existing crawled dataset.

Fig. 3. Car Brand Geographical database

Brand	Country	Continent
Alfa	Italy	Europe
Ankai	China	Asia
Aston	UK	Europe
Audi	Germany	Europe
Austin	UK	Europe
Bentley	UK	Europe
Bertone	Italy	Europe
BMW	Germany	Europe
Chana	China	Asia
Chery	China	Asia
Chevrolet	America	North America
Chrysler	America	North America

Separate car brand & model

As data on the website appends the car brand and model together in one single field. We have to split it into 2 separate fields (Car brand and car model) in order for further analysis.

Fig. 4. Splitting Car brand & model

Car Model
Mercedes Benz E300 AMG Line C
Mercedes-Benz E-Class E230
Mercedes-Benz C-Class C200 Ava

Car Brand	Car Model
Mercedes-Benz	E300 AMG Line Coupe
Mercedes-Benz	E-Class E230
Mercedes-Benz	C-Class C200 Avantgarde

Standardizing data

As the data range for different attribute varies greatly. Eg: Price and car age ranges vastly. Hence, we use x_new = (x-

min)/(max-min) to make sure that all data points are within 0 and 1.

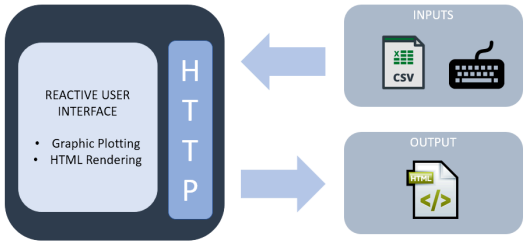
3.4 Tools & Platform

We utilized Scrapy a free and open source web crawling framework, written in Python for our web crawling needs to crawl information from http://www.sgcarmart.com/used_cars/listing.php.

We also utilized excel to help with our data preprocessing and cleaning. In addition, we have also utilized tableau to build our low-fidelity prototypes and perform explorative visual analysis on the data set. Lastly, we utilized D3.js for our front end interactive visualizations. The system architecture is a simple 2-tiered architecture using D3 a JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of the widely implemented SVG, HTML5, and CSS standards.

Fig. 5. Tools utilised

Fig. 6. System architecture



4 VISUALISATION DESIGNS

For our visualisation designs we have made use of a Treemap to provide an overall market analysis on resale cars to show the most popular car brands available for sale.

In addition, we also utilised a Bubble Chart to provide a deeper understanding of different vehicle types available in the resale market based on price and quantity.

Furthermore, through a Scatter Plot Matrix we could identify relationships and correlations across different key factors affecting the price of a resale car.

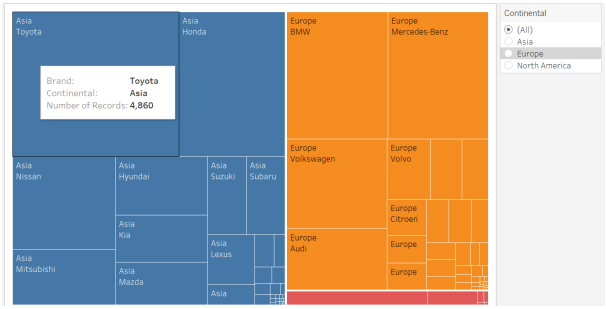
Lastly, with a Radar Chart we were able to help buyers understand different car age profiles based on critical factors such as COE remaining, depreciation, engine power, mileage and price.

4.1 Treemap

Through the treemap we were able to uncover the most popular car brands available for sale in the market right now. From our visualization, we can see that about 60% of the Singapore car resale market is dominated by Asian cars, while 35% of the market is dominated by continental (European) cars and only 5% dominated by North- American cars. The color of the treemap represents the continent where the car is originally created from, we have blue demarking Asia cars, orange representing continental (European) cars and lastly red showing North American cars. The size of the area represents the number of cars available in the market for sale.

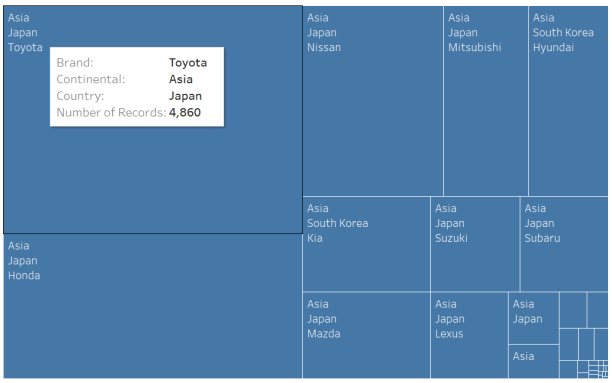
Hence, we can see that the top 5 most popular car brands (Refer to Fig 7) are Toyota, Honda, BMW, Mercedes-Benz and lastly Nissan (popularity in descending order). This shows that 3/5 of the popular car brands are all from Asia.

Fig. 7. Treemap



Upon further inspection through a drill down focusing on Asian cars (Refer to Fig 8). We can see that most of the Asian cars available for sale are from Japan with brands such as Toyota, Honda and Nissan dominating the market. While only a small part of the Asian cars available for sales are from South Korea with brands such as Hyundai and Kia.

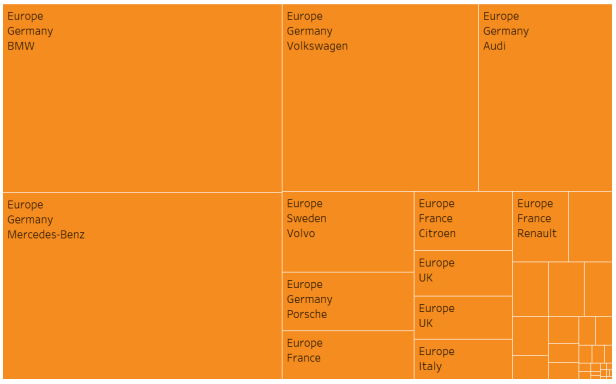
Fig. 8. Treemap- Asian cars drill down



Taking a further look at the Continental (European) cars market (Refer to Fig 9), we can see that the 2 largest brands dominating the market are BMW and Mercedes-Benz with Volkswagen and Audi trailing behind.

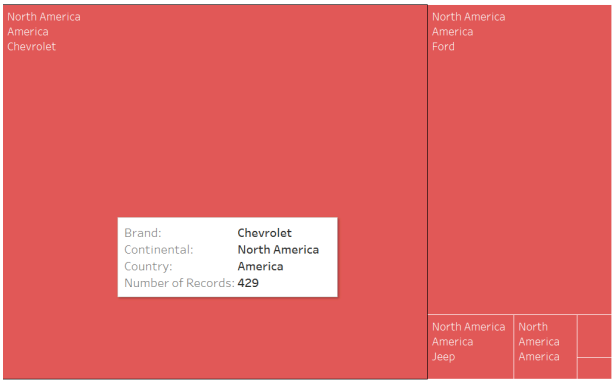
We can also see that the above mentioned 4 brands originated from Germany indicating that most of our resale continental cars are from Germany. While smaller brands such as Volvo originates from Sweden.

Fig. 9. Treemap- Continental cars drill down



Lastly looking at the North American resale cars avaialble in the local market (Refer to Fig 10), we can see that there are lesser brands originating from North America. It is also clear that the North American car market locally is largely dominated by chevrolet cars, while Ford holds the remaining significance.

Fig. 10. Treemap- North American cars drill down



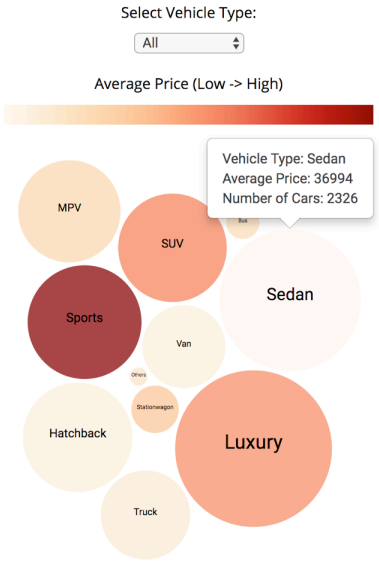
Such information will prove useful for potential buyers surveying the market to understand available brands for sale in the market. In addition, this will also translate to brands that will be more easily serviceable due to its larger presence in the local market.

4.2 Bubble chart

The force bubble chart using D3.js is to provide a deeper understanding of different vehicle types available in the resale market based on price and quantity. The color of the bubble represents the average price level. The darker the color, the higher the average price. The size of the bubble represents the number of available cars in the market. By default, the bubble chart provides an overview of the market by showing the distribution of vehicle types in the resale market.

Figure 11 shows an overall view of all the vehicle types available in the car resale market. It shows that the available cars in the market is largely dominated by luxury cars, followed by sedan cars as the second largest dominant vehicle type. We can also see that Sports car has the highest average price, comparing with other vehicle types and it is the 3rd most popular vehicle types available in the market. This suggest that the top 3 most popular vehicle types in the car resale market are Luxury, Sedan and Sports cars.

Fig. 11. Bubble Chart- All Vehicle Type



An interactive filter is applied in the bubble chart to allow for a further drill down for each vehicle type by the car brands. As the sedan vehicle type holds the second largest market share (Refer to Figure 11) we decided to investigate this market segment further. Figure 12 shows the sedan car market by car brands, we can see that Toyota has the largest number of available sedan cars. If we mouse over the bubble a tooltip with details of the brand will be shown, as such we can see that the average price of this brand of vehicle is $35,586 with 504 cars available in the car resale market.

Fig. 12. Sedan Vehicle Type

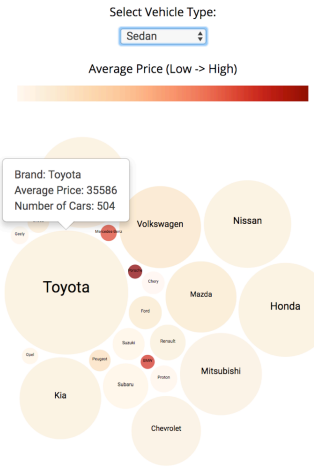
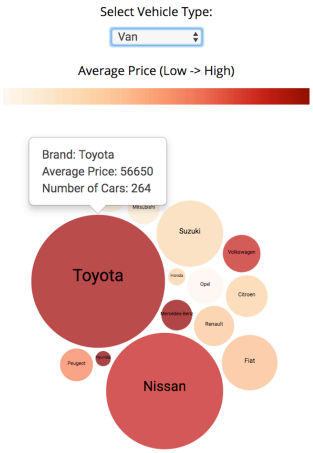


Fig 13. Van Vehicle Type



Looking at a drill down by van vehicle types in Figure 13. We uncovered an interesting insight showing that although the average prices of Toyota and Nissan are much higher as compared to the other brands, these 2 brands are still the most dominant car brands in the van car resale market. This is an unlikely trend as compared to the other vehicle type

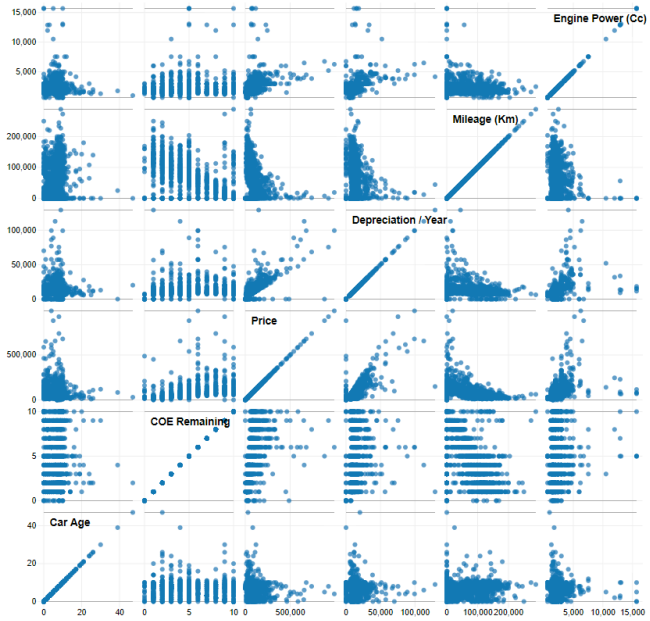
markets suggesting that more expensive vans are more popular within this market segment.

4.3 Scatter plot Matrix

We have also implemented a D3.js brushing enabled Scatter Plot Matrix. This will allow us to determine any correlations between multiple variables such as Car age, COE remaining, Price, Depreciation/year, Mileage, Engine Power (cc). This is particularly helpful in pinpointing specific variables that might have similar correlations across multiple variables (S.wak, 2013).

The variables are written in a diagonal line from top right to bottom left grid. Then each variable is plotted against each other, with each grid representing a correlation between each variable with a corresponding x & y axis (S.wak, 2013). The same plots are replicated across the diagonal line reflecting a mirror image.

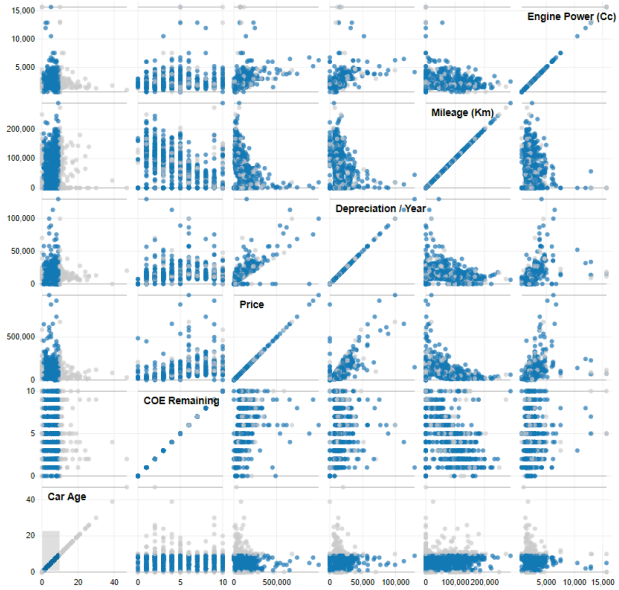
Fig. 14. Scatter Plot Matrix



We can identify variables that are correlated (Refer to Figure 14). As we are more interested to see how each variable affect the price of the resale vehicle we can focus on 4th row of the Scatter Plot Matrix. From the 4th row we notice that there is a positive correlation between the price of the car and its depreciation/ year. Hence as prices of a car in the resale market increases the depreciation/ year of the car increases as well. As such we can infer that more expensive cars generally tend to experience a higher depreciation rate, indicating that an individual loses more of its car value every year if he/she purchase a more expensive vehicle.

In addition, from Figure 14 we can also identify that there is an inverse relationship between price and mileage of a vehicle. This suggest that as the mileage of vehicle increases the price it fetches on the resale market decreases. This suggest that the greater the utility of the vehicle the less value it fetches in the resale market.

Fig. 15. Scatter Plot Matrix - Brushing

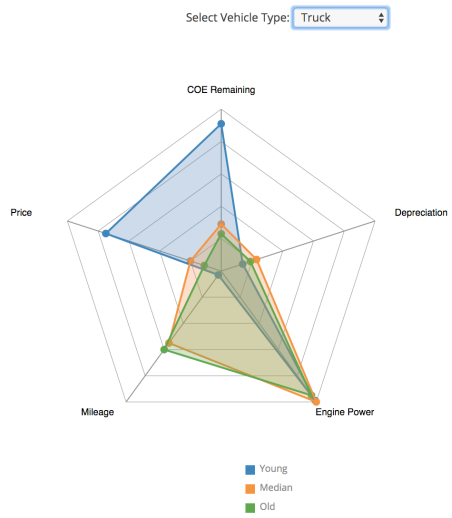


The scatter plot matrix we implemented is dynamically linked (Refer to Figure 15), so by clicking and dragging on any parts of one scatterplot it highlights the individual points in all the other scatterplots. This enables the brushing feature which allows us to highlight a specific part of the scatter plot to examine other relationships across different variables. For example, we can highlight car ages between 0 -10 years old to see how it correlates across other variables. With the brushing feature, we can see that car ages ranging from 0-10 years fetches the highest price value in the resale car market and prices start to decline for cars age more than 10 years.

4.4 Radar chart

The radar chart implemented with D3.js provides an overview of used cars from different age categories. Radar chart was selected as the choice of visualization because it is good for displaying multivariate data and users could compare cars with different ages easily.

Fig. 16. Radar Chart - Truck



We have divided the car age range into three bins specifically. If the car age is less than 5 years, it is defined as "young". If the car age is greater than 5 years but smaller than 10 years, it is categorized as a "median" age profile. If the car age is above 10 years, it is considered "old". The radar chart is then plotted against five attributes namely "COE remaining", "depreciation", "Engine Power", "Price" and "Mileage".

After conducting some preliminary analysis, we realize that the car profiles vary drastically across different vehicle types. Hence, we provided a drop-down list to allow users to dynamically explore and drill down based on vehicle types.

Here are some interesting findings:

For truck vehicle types (Refer to Figure 16). , the profile for median aged cars and old cars are very similar. Choosing either of them will make not a great difference However, for young trucks it shows a great difference in characteristics, especially in price. We can see that trucks age <5 years old cost drastically higher on an average as compared to the 2 other age profiles.

For luxury cars (Refer to Figure 17), we observe similar patterns as identified above, the mileage, engine power and price of a median aged car does not vary significantly from an old age car as well. In addition, we also noticed that old cars tend to have more COE remaining and cost lower in price as compared to median aged cars. Hence, we will recommend that if a consumer has tight budget, buying an old car seems to be an optimal choice as it maximize utility (in terms of COE available) for the best possible price.

Fig. 17. Radar Chart – Luxury cars

5 Conclusion

The visualization application created was largely useful as a descriptive analysis on the existing resale cars available in the market. However, due to the limitation of the dataset our analysis was only restricted to current data of available cars for sale and not past data of sold cars. It will be more useful in the future if we will be able to attain past transactions of cars sold. This will provide us with the ability to provide a more accurate analysis and possible predictions / forecasting for the resale car market. In addition, to add more layers and effectiveness of possible predictions we hope to aggregate the data with the existing economic climate and car COE quota to provide a better context and insights on the resale car market.

Acknowledgments

Our team will like to thank Professor Kam Tin Seong for his time and guidance throughout the project. Prof Kam has been very kind and helpful along the way as he pointed us towards the right direction of the project. He also guided us when we are unsure about certain areas of analysis and visualisations.

References

[1] Costs of Car Ownership in Singapore 2017. (2017, July 11). Retrieved November 25, 2017, from https://www.valuepenguin.sg/costs-car-ownership-singapore

[2] Junn, L. C. (2017, March 16). Used cars becoming more popular: Car dealers. Retrieved November 25, 2017, from http://www.channelnewsasia.com/news/singapore/used-cars-becoming-more-popular-car-dealers-8158106

[3] Bloomberg - Scientific Proof that Americans are Completely Addicted to Trucks: https://www.bloomberg.com/graphics/2015-auto-sales/

[4] Predict second hand car price using artificial neural network: http://csidsocialmedia.github.io/2014/05/02/Predict-second-hand-car-price-using-artificial-neural-network.html

[5] The Perfect Storm Hits Used-Car Values: The Foundation Of The Auto Industry Is Faltering: http://www.zerohedge.com/news/2017-05-21/perfect-storm-hits-used-car-values-foundation-auto-industry-faltering

[6] Kaggle - Used car database: https://www.kaggle.com/orgesleka/used-cars-database

[7] Kaggle -Data Crunchers: https://www.kaggle.com/timucinanuslu/data-crunchers

[8] S.wak (2013, January 31). Scatterplot Matrices. Retrieved November 25, 2017, from https://www.r-bloggers.com/scatterplot-matrices/

[9] D3.js: https://d3js.org/

[10] Chart.js: http://www.chartjs.org/