DAO2702: Programming for Business Analytics
Group Project Report
Tutorial: A21
Team: BIG6

Group members	Matric Number
Anthea Yeo Chyi Yin	A0189992B
Benedict Liew Weng Chee	A0189836H
Brandon Koh Wai Loong	A0183500R
Brandon Yeap Yan Ting	A0182624E
Liu Jia Rui	A0200045A
Wong Soon, Mark	A0183072J

Table of Contents

1. Introduction

1.1 Problem Statement

BIG6 Associates is a US-based strategy consulting firm specialising in the automotive industry. Our firm was approached by Python Inc., a global luxury car dealership, who wants to tap on growing car sales in the US by expanding into the US market. Our client requires our expertise to formulate a profit-maximising market entry strategy comprising recommendations of **(1)** a US city to launch its first dealership, **(2)** a selection of luxury car brands to sell, and **(3)** a price-setting guideline based on cars' attributes.

1.2 Terminologies

- **Invoice Price** - Price manufacturer charges dealer for car (i.e., the cost of dealer's inventory)[1]
- **True Dealer Cos**t – Actual cost dealer pays manufacturer for car (Includes hidden mark-up costs such as holdback, and other dealer fees/incentives, built into price)[2]
- **Manufacturer's Suggested Retail Price (MSRP)** - Price the car manufacturer suggests car dealers should sell at
- **Actual Retail Price/Selling Price** – The actual price that Python Inc. will sell its cars at

2. Methodology

Our firm has decided to adopt a **4-step approach** to assist Python Inc. in solving its business problem.
Step 1: We determine the optimal US city for Python Inc. to set up its car dealership
Step 2: Next, we identify the top 3 luxury car brands that Python Inc. should sell
Step 3: Through a correlation and multiple linear regression analysis, we will identify the car attributes which can command higher prices in luxury cars
Step 4: After variable selection, we will visualise the variables identified for each chosen car brand to curate a pricing guideline that allows Python Inc. to increase its price-setting ability by upselling car models based on car attributes

3. Data Source

Our income dataset (A) is a combination of 3 datasets: i) "States Lat Long.csv" acquired from Latlong.net[3], ii) "states.csv" acquired from Kaggle, and (iii) "Income Dataset.csv" acquired from DQYDJ[4]. It provides insights into the median household income of US states, which acts as a proxy to the purchasing power of its consumers. Our second dataset, "Maryland Cities Population.csv" (B), is obtained from Maryland Demographics by Cubit[5]. It provides insights into the population sizes of cities in the state of Maryland. We also obtained Maryland's land area data from the United States Census Bureau in our calculations for population density. Our third dataset, "Car_sales (Cleaned).csv" (C) was acquired from Kaggle[6]. It contains information for 12,000 car models sold in the US between 1990 and 2018, which provides insights into how car attributes and brands affect MSRP.

The variables in our datasets are summarised below:

A. Income Dataset	B. Population Dataset	C. Car Features Dataset		
1. Abbreviation	1. City	1. Market Category	6. Year	11. Engine Fuel Type
2. State	2. Population	2. Engine HP	7. Engine Cylinders	12. Transmission Type
3. Median Household Income	3. Land Area	3. Driven Wheels	8. Number of doors	13. Vehicle Size
4. Latitude	4. Population Density	4. Vehicle Style	9. Highway MPG	14. City MPG
5. Longitude		5. Popularity	10. Make	15. Model

4. Assumptions

1. Data from the car features dataset (C) is representative of the general US population and is relatively similar throughout all US states and cities in terms of consumers' tastes and preferences
2. Non-market factors such as economic policies, political stability, and legal regulation, will not influence our selection of dealership location, car brands, and car types
3. Invoice price is calculated as a 30% markdown from MSRP

The third assumption is necessary as our analysis of car dealership data[7] and related articles[8] found that the profit margin of luxury cars increases with MSRP and popularity[9]. However, if Python Inc. wants to increase profit margins, it must also consider the Cost of Goods Sold (i.e., inventory purchased from car manufacturers). Since data on the luxury car market is limited, we analysed MSRP and invoice price data of cars with MSRP > $70,000 instead and calculated the average percentage markdown from MSRP to be 29.2% (*Refer to Annex A, Appendix).* Given the potential of higher profit margins and mark-up costs for luxury cars (which we have defined to be those with MSRP > $100,000), our team used a conservative **30% reduction of MSRP as our invoice price.** This gives Python Inc. a more reliable estimate of the true dealer cost in its inventory cost calculations. Overall, Python Inc. should set pricing decisions based on both revenue and cost factors, to effectively increase profit margins.

5. Market Entry Strategy

5.1 Selection of State for Market Entry

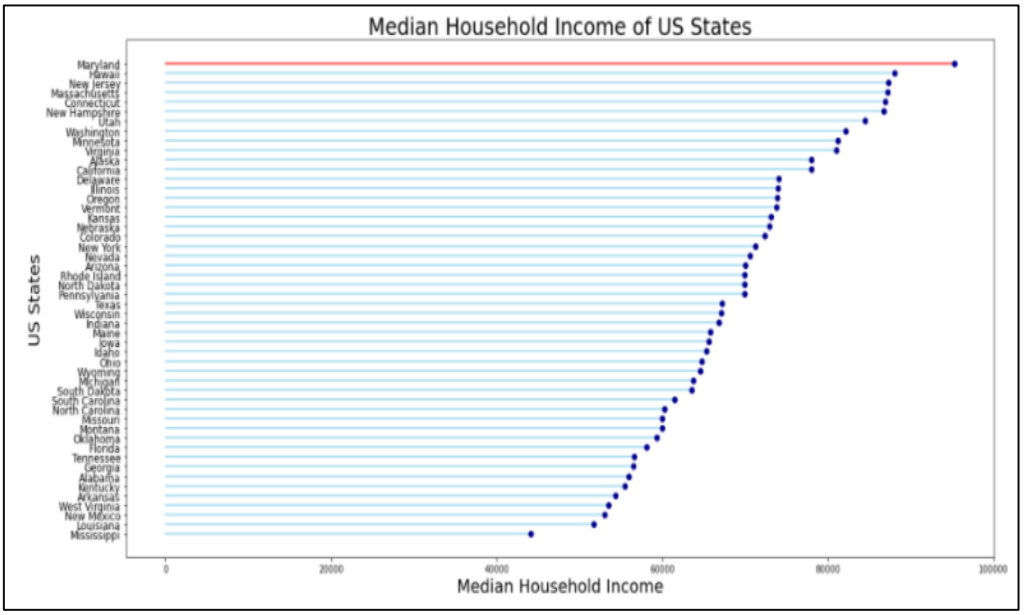


Figure 1: Lollipop Chart of Household Income across US States

Since Python Inc. is a luxury car dealer, we will identify its target group of high-income consumers who are more likely to purchase luxury cars.

Thus, we compared median household income across US states (refer to Fig. 1) and identified Maryland as the state with the highest median household income ($95,310), which signifies the highest average purchasing power and demand for luxury cars.

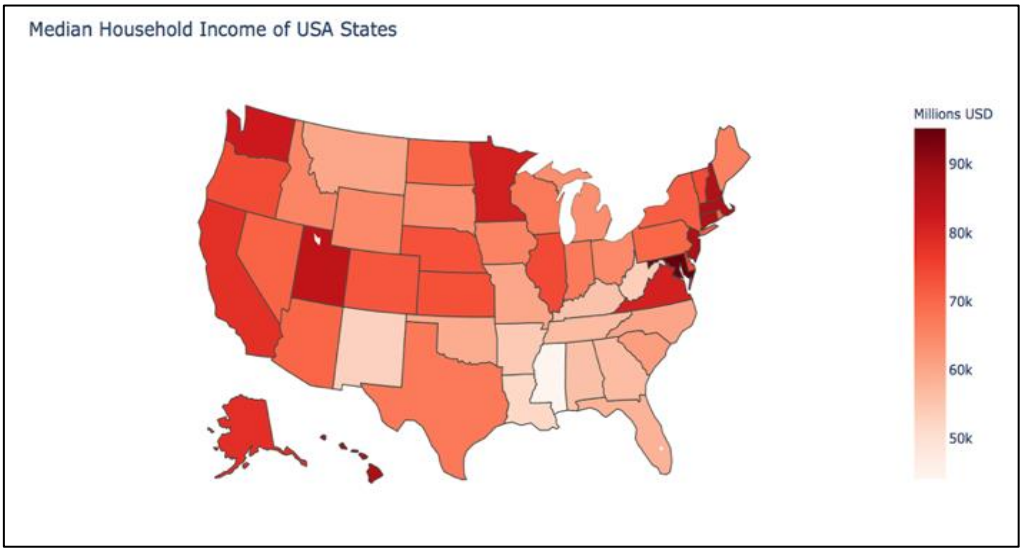


Figure 2: Choropleth Map of US States by Household Income

However, we acknowledge that our client might have other considerations such as the accessibility to ports and other non-market factors mentioned in **Section 4.**

Thus, we used a choropleth map to illustrate the median household income across all US states, where the colour intensity of the shaded regions increases with greater income levels (refer to Fig. 2). This will help our client pinpoint potential alternative locations for its dealership.

5.2 Selection of City for Market Entry

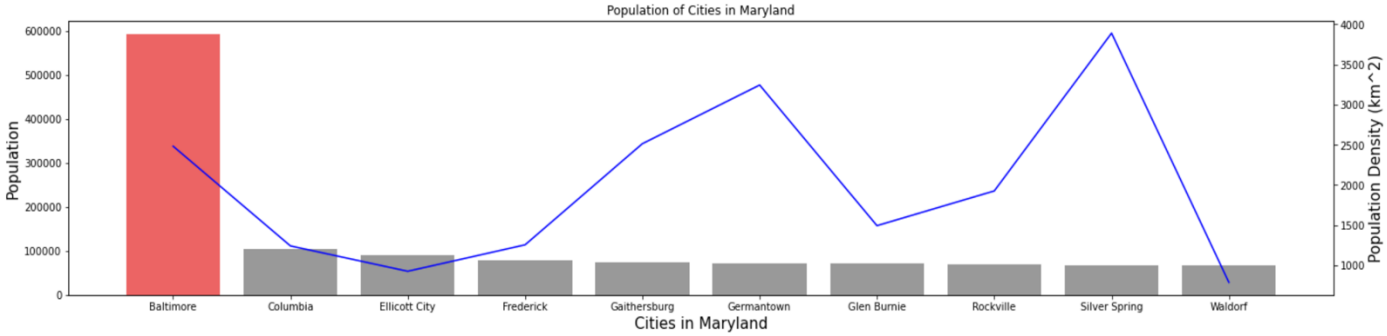


Figure 3: Population Size and Population Density of Cities in Maryland

Next, assuming median household income and vehicle ownership per capita (31.8%)[10] hold relatively constant across cities in Maryland, we shortlisted the top 10 largest cities based on population size. From our analysis, we identified Baltimore as the city with the largest population with a relatively high population density as well. As luxury cars are in a niche market, consumers are more likely to travel a longer distance to purchase cars from a luxury car dealership. This suggests that our client will still be able to capture the consumer base farther away from its dealership location, which diminishes the impact of its slightly lower population density. Furthermore, given that Baltimore has the largest land area, its residents have a greater need for a car as a means of transport around the city. Thus, we overweight the larger population size compared to the slightly lower population density as it signifies a more substantial luxury car consumer base. Overall, the highest median household income in Maryland, coupled with the largest population size in Baltimore, suggests the greatest potential effective demand in terms of the willingness and ability of its consumers to purchase luxury cars. Therefore, Python Inc. should launch its new dealership in Baltimore, Maryland.

6. Selection of Car Brands

As Python Inc. is still in the market-entry stage, we decided to provide a more conservative selection of recommended car brands for its dealership. Thus, we compared the popularity and quantity sold across all luxury brands to identify the top 3 brands with the highest projected demand for luxury cars, which are likely to maximise potential profits for our client.

6.1 Data Cleaning

Before analysing our car features dataset (C), we first cleaned the dataset to account for missing values and removed data with "NaN values" or inputs equivalent to a null entry for the data column such as "0" or "Unknown".

6.2 Segmentation of Car Brands into Distinct Categories

Next, we sorted the car brands into 3 distinct car categories which we defined according to their MSRP (refer to Fig. 4):

Category	Budget	Middle	Luxury
MSRP (P)	P < $40,000	$40,000 < P < $100,000	P > $100,000

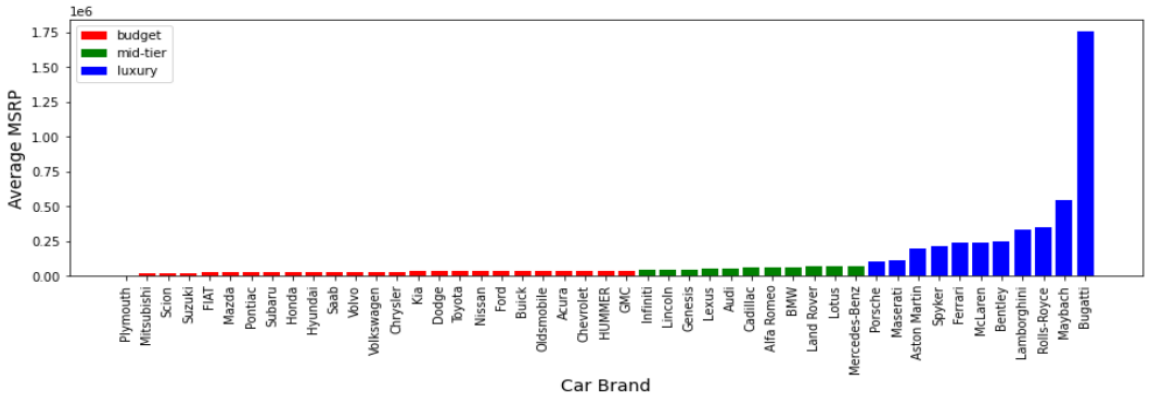


Figure 4: Categorisation of Car Brands

6.3 Luxury Car Brands vs Popularity

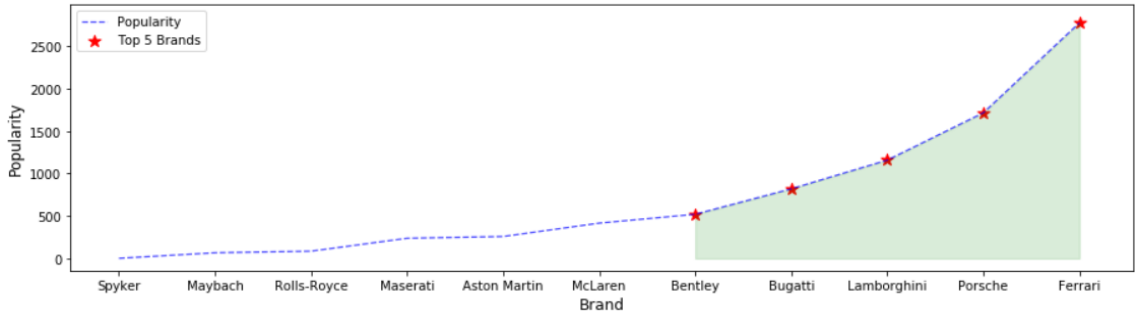


Figure 5: Popularity of Luxury Car Brands

'Popularity' in this case refers to the number of times the brand was mentioned on social media throughout the time recorded in the dataset (C). Given that we previously found a positive relationship between profit margin and popularity as stated in **Section 4**, we zoomed in on the popularity of each luxury car brand (refer to Fig. 5). Thus, we narrowed our selection down to the 5 most popular luxury brands: **Bentley, Bugatti, Lamborghini, Porsche, and Ferrari.**

6.4 Luxury Car Brands vs Quantity Sold

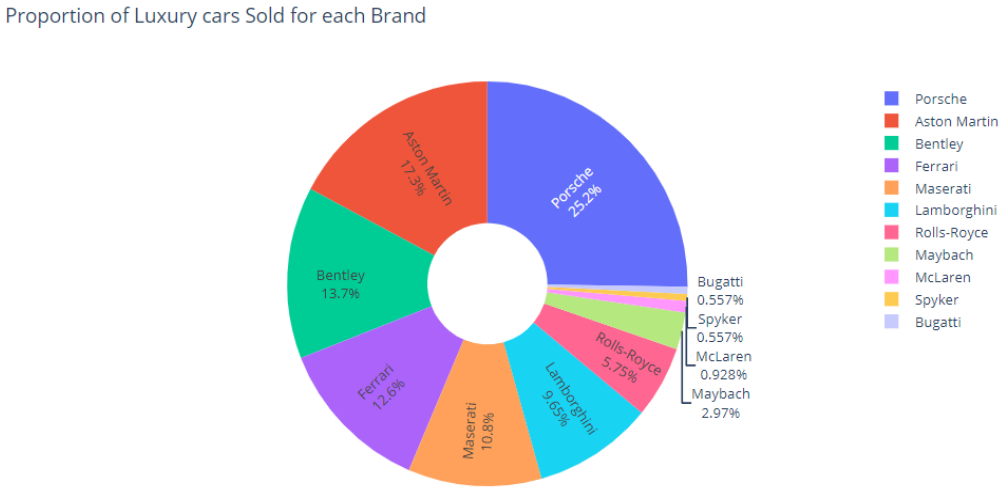


Figure 6: Proportion of Total Quantity Sold of Luxury Cars for each Brand

Next, assuming that the demand of luxury cars is relatively price-inelastic, and the supply of each car brand is relatively similar, comparing the quantity of cars sold across the luxury car brands allows us to determine the brands with the highest demand. Thus, the top 5 brands with the largest quantity of cars sold are: **Maserati, Ferrari, Bentley, Aston Martin and Porsche** (refer to Fig. 6). In addition, these 5 brands make up 75% of market share in the luxury cars segment, illustrating their significance.

6.5 Luxury Car Brands vs Quantity Sold & Popularity

By corroborating our findings for the top 5 brands in both popularity and quantity sold, we determine the top 3 car brands in the luxury car market to be: **Porsche, Ferrari and Bentley** (refer to Annex B, Appendix). Thus, our recommendation to Python Inc. would be to engage in dealerships for these 3 car brands.

7. Analysis of Luxury Car Brand Attributes

7.1 Correlation Matrix for Luxury Car Brand

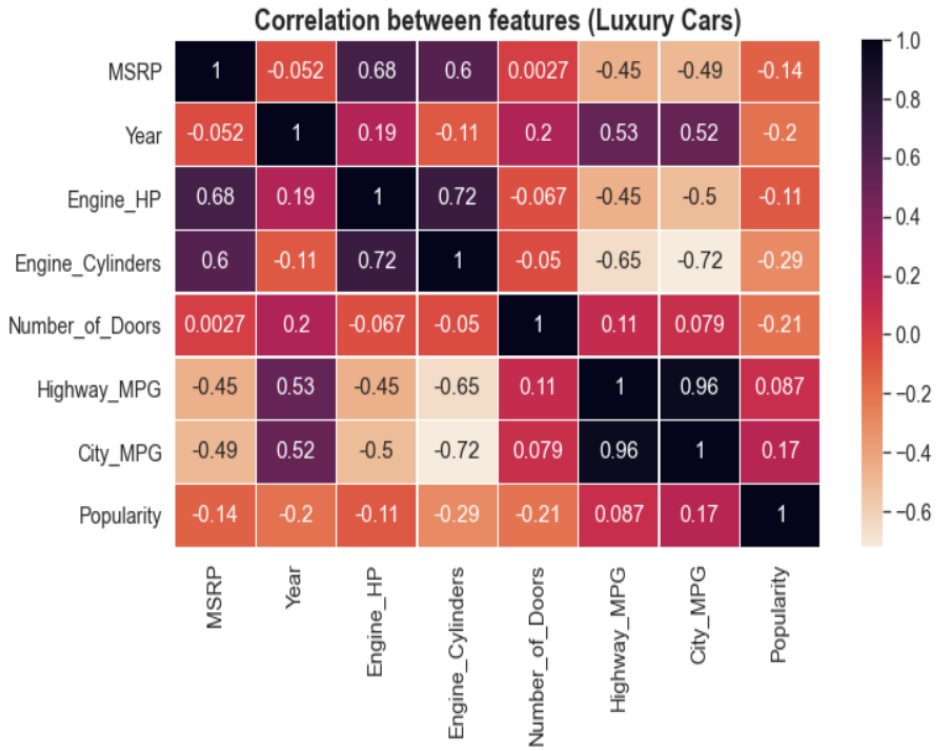


Figure 7: Correlation Matrix between Car Attributes

Next, using a correlation matrix to determine the correlation between independent variables (refer to Fig. 7), we identified the luxury car attributes which would fetch a higher MSRP. Assuming that the variation of MSRP with car attributes is applicable to all luxury car brands, Python Inc. will be able to use this information to achieve a greater price setting ability for its selected car brands.

With a Pearson's correlation value of 0.5 set as a cut-off to indicate a moderately significant correlation, we found that **'Engine_HP'** and **'Engine_Cylinders'** have the strongest relationship with MSRP. However, the correlation coefficient alone is insufficient to conclude whether these variables directly affect the MSRP of luxury cars. Thus, we will run a multiple regression analysis to confirm these relationships.

7.2 Multiple Linear Regression of Attributes

For our regression analysis, we conducted repeated variable selection to remove variables with p-values above 0.05 (i.e., variables did not affect MSRP at a 5% level of significance). Thus, we obtained the following regression results as shown in Fig. 8:

OLS Regression Results

Dep. Variable:	MSRP	R-squared:	0.645
Model:	OLS	Adj. R-squared:	0.639
Method:	Least Squares	F-statistic:	120.2
Date:	Tue, 03 Nov 2020	Prob (F-statistic):	7.14e-114
Time:	14:38:09	Log-Likelihood:	-7019.8
No. Observations:	539	AIC:	1.406e+04
Df Residuals:	530	BIC:	1.410e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.815e+06	2.05e+06	4.304	0.000	4.79e+06	1.28e+07
Engine_HP	999.6145	67.917	14.718	0.000	866.195	1133.034
Year	-5030.1398	1019.264	-4.935	0.000	-7032.434	-3027.846
d_Large	6.819e+04	1.73e+04	3.952	0.000	3.43e+04	1.02e+05
d_4dr_SUV	-9.287e+05	8.36e+04	-11.108	0.000	-1.09e+06	-7.64e+05
d_Sedan	-9.674e+05	7.92e+04	-12.210	0.000	-1.12e+06	-8.12e+05
d_flex_fuel_premium_unleaded_required_E85	-1.306e+05	2.4e+04	-5.442	0.000	-1.78e+05	-8.34e+04
Engine_Cylinders	1.077e+04	2953.681	3.645	0.000	4962.861	1.66e+04
Number_of_Doors	4.669e+05	4e+04	11.662	0.000	3.88e+05	5.46e+05

Omnibus:	664.486	Durbin-Watson:	1.140
Prob(Omnibus):	0.000	Jarque-Bera (JB):	74428.597
Skew:	5.945	Prob(JB):	0.00
Kurtosis:	59.327	Cond. No.	8.90e+05

Figure 8: Multi-linear Regression Results

Hence, our model will be: 'MSRP ~ Engine_HP + Year + d_Large + d_4dr_SUV + d_Sedan + d_flex_fuel_premium_unleaded_required_E85 + Engine_Cylinders + Number_of_Doors'

Overall, based on our multi-variable linear regression model (refer to Fig. 8), our team has identified the car attributes which significantly affect the MSRP of luxury cars as follows:

1. Engine HP	**3. Vehicle Size - Large**	**5. Vehicle Style - Sedan**	**7. No. of Engine Cylinders**
2. Year	**4. Vehicle Style - SUV**	**6. Engine Fuel Type - Premium unleaded required/E85**	**8. No. of Doors**

Figure 9: Visualising Coefficients of our Regression Model

To visualise the magnitude and nature of relationships between variables identified from our model, we plotted a diverging bars graph of the coefficients (refer to Fig. 9).

The negative coefficient of vehicle types **'4-drive SUV'** and '**Sedan'** showed a negative relationship with MSRP, while the positive coefficient of other variables showed a positive relationship with MSRP.

Thus, incorporating this into our pricing guideline will allow our client to vary its prices based on the car attributes.

7.3 Train-Test Split

The train-test split procedure is used to estimate our model's performance when used to make predictions on data not used to train the model. With a low Normalised Root Mean Squared Error of **0.035**, it shows our model is accurate in showing the variation of MSRP with our selected variables. Moreover, the model will have an adjusted R^2 value of **0.639**, implying strong goodness-of-fit.

8. Unique Selling Point for Chosen Car brands

After determining the variables which significantly affect MSRP of luxury cars, we proceeded to analyse the impact of these variables for each of our chosen luxury brands.

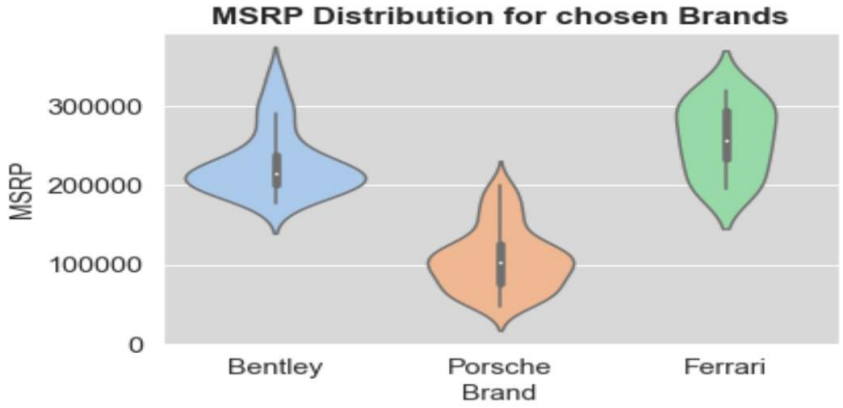


Figure 10: Violin plot of MSRP for Chosen Brands

From Figure 10, we can conclude that Ferrari has the highest average **'MSRP'**, followed by Bentley, then Porsche.

However, Bentley seems to have the largest concentration of cars around $200,000, whereas Porsche's largest concentration of prices is at around $100,000. The concentration of MSRP for Ferrari is more evenly spread out, indicating a wider price range.

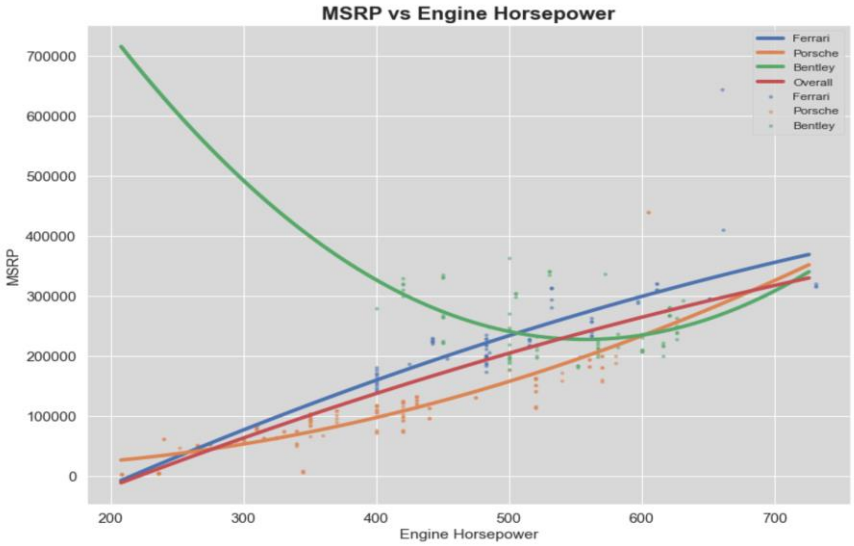


Figure 11: Engine HP against MSRP

From Figure 11, where engine horsepower is plotted against MSRP, we conclude that generally, the higher the **'engine horsepower'**, the higher the MSRP.

However, the MSRP for Bentley cars is seen to decrease until approximately 550HP, before increasing. This implies a quadratic relationship for Bentley's MSRP and Engine HP. Thus, for Bentley, Python Inc. should only increase the price of Bentley cars if that model's 'Engine HP' is above 550HP.

The distribution of 'Engine HP' for all 3 car brands is illustrated in Annex C, Appendix.

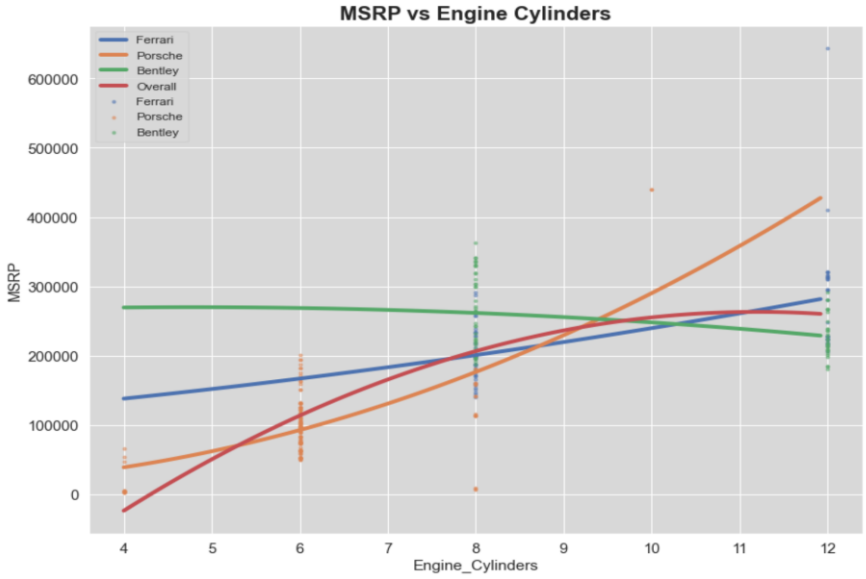
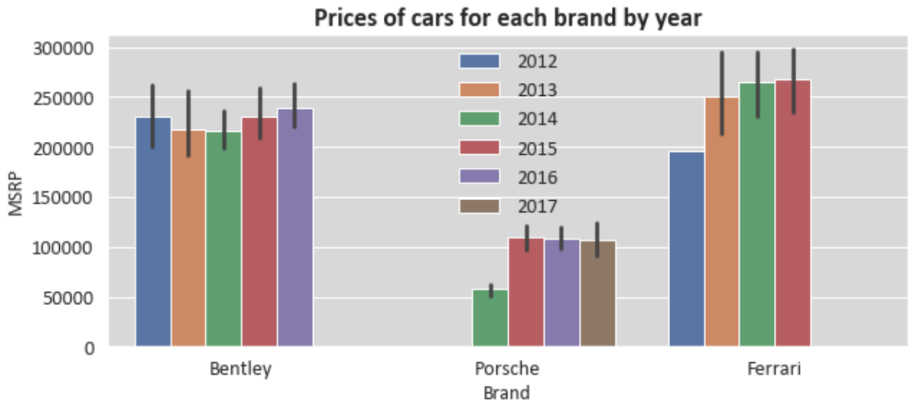


Figure 12: Engine Cylinders against MSRP

From Figure 12, where number of **'engine cylinders'** is plotted against MSRP, we conclude that generally, the greater the number of **'engine cylinders'**, the higher the MSRP.

However, for Bentley cars, MSRP decreases with increasing number of **'engine cylinders'**. This means that Bentley cars with 4 **'engine cylinders'** should be priced the highest.

For Porsche and Ferrari, a higher number of **'engine cylinders'** would allow Python Inc. to set higher prices.



From Figure 13, where the **'year'** of car models from 2012 to 2017 is plotted against MSRP, we observe that, generally, the more recent the car models have higher MSRP.

Hence, for these 3 car brands, Python Inc. should sell newer car models at a higher price.

Figure 13: Years against MSRP

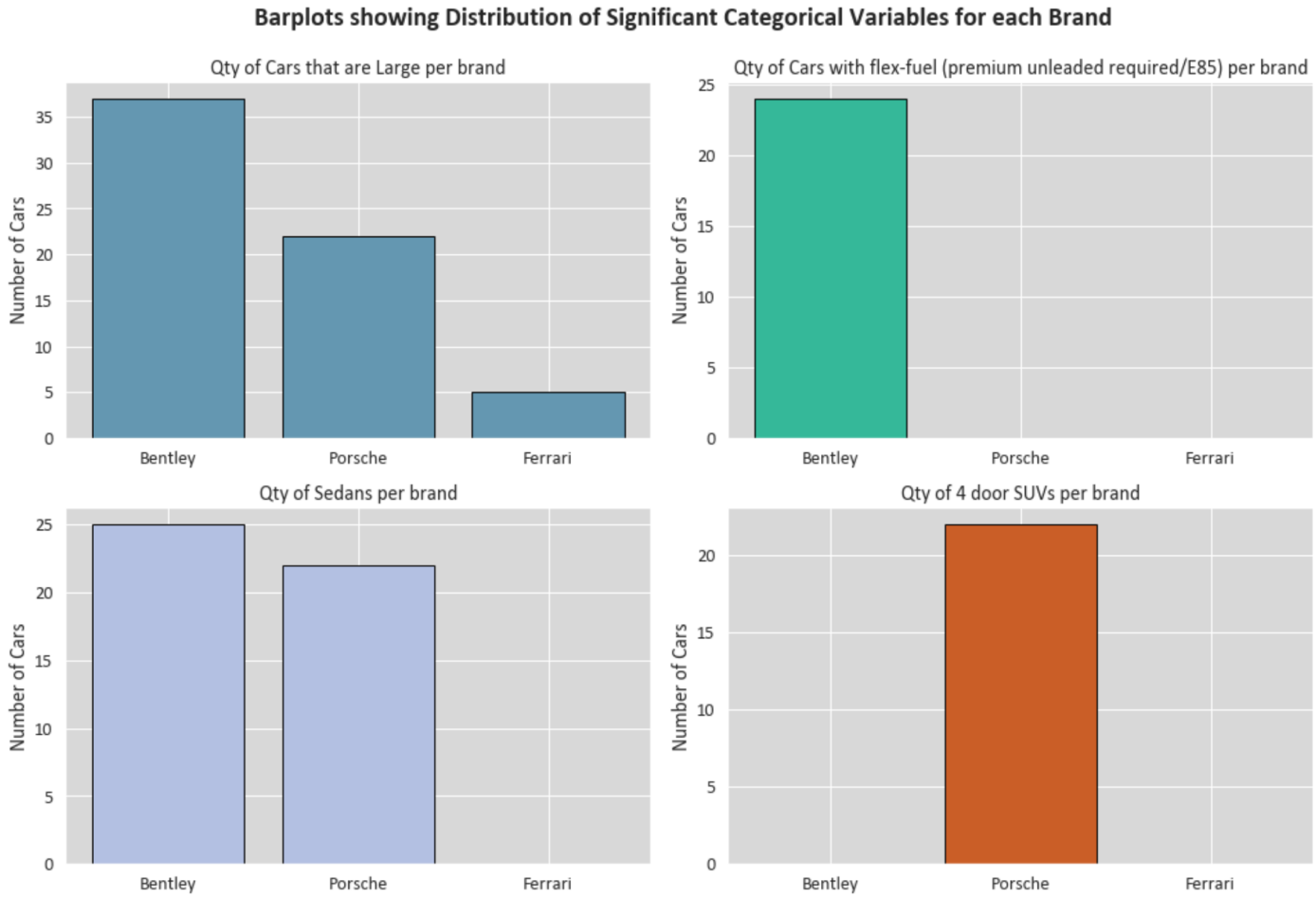


Figure 14: Significant Categorical Variables present in each car brand

For the remaining categorical variables from our regression model, we decided to look for their significance in each car brand (i.e., how many cars of each brand contains the categorical attributes that affect price). As shown previously in **Section 7.2,** luxury cars with a '**large'** vehicle size and a fuel type of '**flex-fuel premium unleaded required E85**' command a higher MSRP. On the other hand, the '**sedan'** and **'4-drive SUV'** vehicle types are generally negatively related to MSRP as the correlation coefficient is negative.

Looking at Figure 14, some of our target brands do not contain the categorical variables that affect MSRP. There are also more '**large'** Bentleys than both '**large'** Porsches and Ferraris combined. Ultimately, there are different categorical considerations for each of the 3 selected brands which affects our client's price setting ability and what types of car designs they might want to buy from the manufacturer.

9. Limitations of Analysis

Firstly, our car features dataset (C) contains information that was dated from 2012 to 2017. Since our analysis was based on information collected within this timeframe, the slight difference in timing could limit the accuracy of our recommendations provided. However, we chose to rely on this dataset as it was the most recent and complete data source available.

On the same note, using a historical dataset may not yield the same results in future as it does not take into account changes in market forces, especially amidst the current global uncertainty surrounding the US political elections, US-China trade war, and the global recession from the COVID-19 pandemic.

Next, when identifying the optimal location for Python Inc. to set up operations in **Section 5**, we assumed that the median household income and vehicle ownership per capita hold relatively constant across Maryland. In reality, these variables may not be uniform and might affect our recommendations. However, the median household income of Maryland is significantly higher than other states and signifies a general greater purchasing power of the state as a whole. With regards to the vehicle ownership per capita, there is a lack of vehicle ownership data segregated into the respective Maryland cities. Thus, we selected the city with the highest population for reasons outlined in **Section 5.2**, as it would likely translate into the highest vehicle owner density there.

Moreover, elimination of 'NaN' and 'UNKNOWN' values from our car dataset resulted in the removal of **33.44%** of observations. As the amount of data removed was significant, this may have restricted our data analysis and influenced our recommendations proposed. Despite this restriction, most of the data removed due to "no categorization of market category" were car brands in the "Middle" segment and would not have significantly affected our analysis within the "Luxury" car segment.

In addition, while our recommendations are made based on correlations to price factors, we acknowledge that correlation does not imply causation. Thus, it might not be entirely possible to predict the certainty of higher prices with a positively correlated variable in our model.

Lastly, we also acknowledge that we are limited by the variables provided in our car features dataset (C), and there could be other car features not included in the model which could be good predictors in our models such as interior design and technological functions.

10. Recommendations & Conclusion

Recommendation 1 – Set up dealership in Baltimore, Maryland

Based on our analysis in **Section 5**, we have found Baltimore, Maryland to be the most ideal city for Python Inc to set up its dealership as it has a sizable market for luxury cars based on median household income and vehicle ownership.

Recommendation 2 - Sell Porsche, Ferrari, and Bentley

Since this is a market entry strategy, our analysis in **Section 6 suggests** that Python Inc. should focus its resources on the car brands we suggest in order to optimise sales and maintain its branding as the world's leading luxury car dealership.

Recommendation 3 - Optimise selling price above True Dealer Cost

Based on our analysis in **Sections 7.2 and 8**, after considering the continuous and categorical variables we identified, and the relationship they have with MSRP, Python Inc. should have different pricing guidelines for their 3 car brands.

- Generally, Python Inc. should set higher prices for car models of all 3 recommended car brands which are either **'new'**, have **'large'** vehicle types, or both.

- For Bentley, Python Inc. should only increase the price of Bentley cars if that model's **'engine horsepower'** is above 550HP. Since Bentley is the only car brand out of the 3 selected recommendations that offers car models with the fuel type **'flex-fuel premium unleaded required/E85'**, those should be set to higher prices.

- Python Inc. should put a higher price tag on Porsche and Ferrari car models with increasing number of **'engine cylinders'** and **'engine horsepower'**.

- Python Inc. should avoid selling the vehicle types of **'sedan'** and **'4-door SUV'** due to the lower prices associated with these car models, assuming that there are no business synergies in selling these products at a lower profit margin.

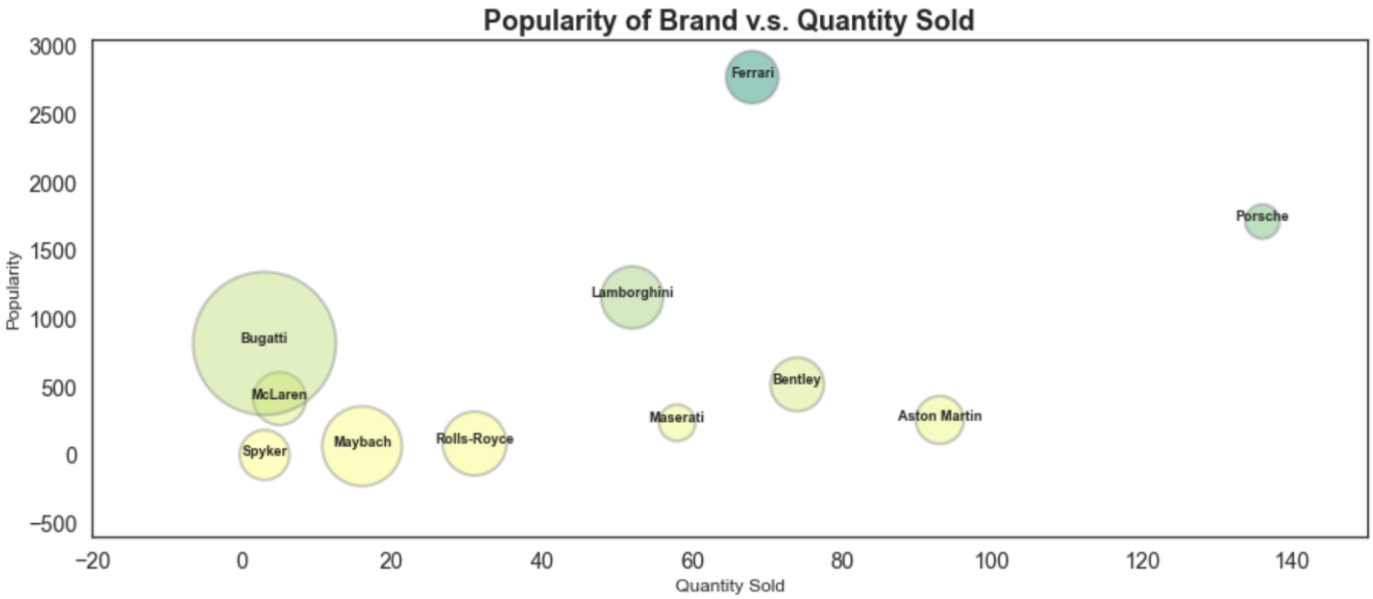
11. References

1. Car and Driver Research. (2020, April 22). Car msrp vs. Invoice: Everything you need to know. Car and Driver. Retrieved November 9, 2020, from https://www.caranddriver.com/research/a31874008/car-msrp-vs-invoice/
2. CarBuyingStrategies. (2020). 2020 mercedes-benz prices: Msrp, invoice price & dealer cost. Retrieved November 9, 2020, from https://www.car-buying-strategies.com/mercedes-benz-invoice-prices.html
3. LatLong.net. (2020). States in united states with lat long. Retrieved November 9, 2020, from https://www.latlong.net/category/states-236-14.html
4. DQYDJ. (2020, October 2). Average income by state, median, top & percentiles [2020]—Dqydj. Retrieved November 9, 2020, from https://dqydj.com/average-income-by-state-median-top-percentiles/
5. Cubit. (2019). Maryland cities by population. Retrieved November 9, 2020, from https://www.maryland-demographics.com/cities_by_population
6. Kaggle. (2018). Car features and msrp. Retrieved November 9, 2020, from https://kaggle.com/CooperUnion/cardataset
7. iSeeCars. (2020). 2020 bmw 3 series price | 2020 bmw 3 series invoice | 2020 bmw 3 series msrp—Iseecars. Com. Retrieved November 9, 2020, from https://www.iseecars.com/car/2020-bmw-3_series-price#:~:text=2020%20BMW%203%20Series%20base,goes%20from%20%2438%2C400%20to%20%2452%2C580
8. iSeeCars. (2020). 2020 jaguar f-type price | 2020 jaguar f-type invoice | 2020 jaguar f-type msrp—Iseecars. Com. Retrieved November 9, 2020, from https://www.iseecars.com/car/2020-jaguar-f__type-price#:~:text=2020%20Jaguar%20F%2DTYPE%20base,goes%20from%20%2457%2C904%20to%20%24119%2C098
9. Autotrader. (n.d.). Buying a Car: How Much Do Dealers Mark Up a Car Over the Invoice Price? Retrieved November 9, 2020, from https://www.autotrader.com/car-reviews/buying-car-how-much-do-dealers-mark-car-over-invoice-price-228247
10. Peterson, B. (2020, October 20). Car Ownership Statistics (2020 Report). Retrieved November 09, 2020, from https://www.valuepenguin.com/auto-insurance/car-ownership-statistics
11. J. Rumsey, D. (n.d.). How to interpret a correlation coefficient r. How to Interpret a Correlation Coefficient r; Dummies. Retrieved November 9, 2020, from https://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/

12. Appendix

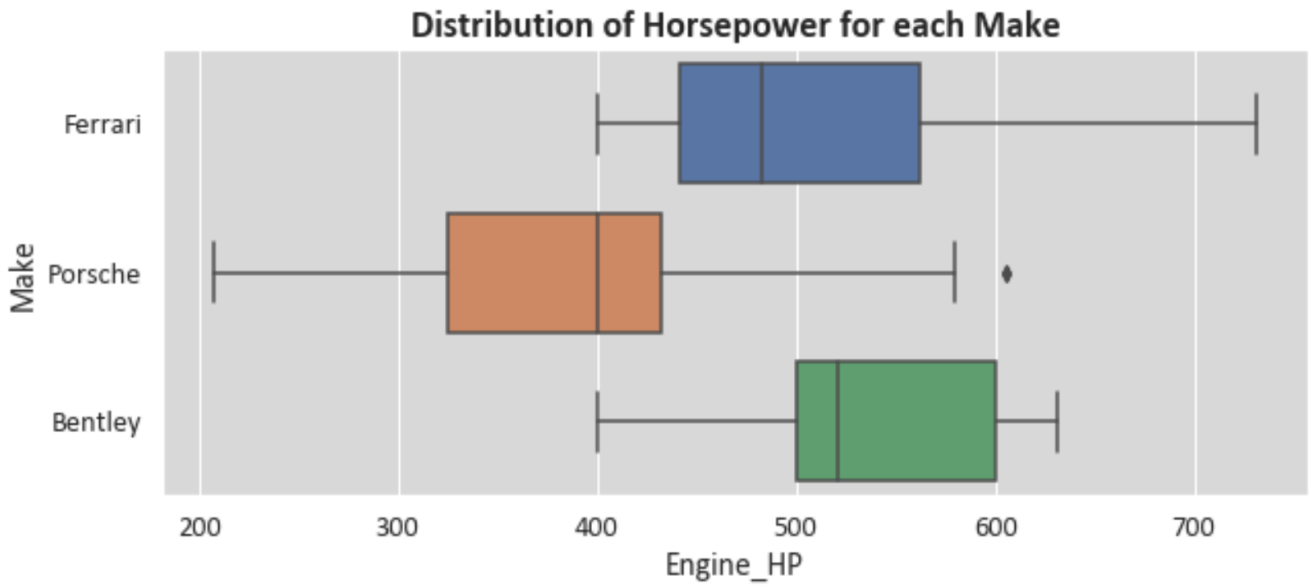
Category	Avg MSRP	Average Invoice	Markdown %
RWD Coupe	$71,475	$54,924	23.2%
Small Convertible	$72,080	$56,677	21.4%
Gasoline Convertible	$73,346	$48,483	33.9%
Convertible	$76,834	$48,193	37.3%
2-Passenger Convertible	$81,528	$58,154	28.7%
RWD Sedan	$85,482	$56,073	34.404%
RWD Convertible	$86,596	$51,020	41.1%
AWD Coupe	$94,058	$76,167	19.0%
2-Passenger Coupe	$103,807	$78,231	24.6%
AWD Convertible	$119,478	$85,856	28.1%
	$86,468	**$61,378**	**29.2%**

Annex A: Deriving Average Mark-up



Annex B: Comparing Popularity of Brands vs Quantity of Cars Sold (%)

Based on Annex B, we selected the 3 brands that were in both top 5 luxury car brands in terms of popularity and quantity sold. These 3 car brands were Ferrari, Porsche, and Bentley. The size of the bubble represents the average MSRP for the car brand. While it seems that either Lamborghini or Aston Martin may be a good choice, Bentley was ultimately chosen to be one of the 3 luxury car brands recommended for our client to engage in a dealership due to it being in both top 5 luxury car categories.



Annex C: Distribution of Engine Horsepower for Selected Car Brands