



SCHOOL OF INFORMATION AND
COMMUNICATION TECHNOLOGY

DA3304 – APPLIED PROGRAMMING FOR DATA

Data Visualisation

Lecturer's Name

Amal Hazimah binti Mohamed Yusoff

Submitted by

Awangku Muhammad Yamin Bin Pengiran Ibnu 22FTT1344

Semester 3

Academic Session 2023/2024

August 2023

Table of Content

Introduction.....	3
Business Understanding.....	4
Business Understanding :	4
Data Understanding.....	5
Attribute Description.....	5
Attributes contribute to data analysis.....	6
Data Preparation.....	7
Feature Engineering:.....	7
Detecting and Treating Outlier.....	8
Treating Missing Value:.....	9
Data Visualisation.....	11
6 Visualization.....	11
4 Important Visualization.....	15
Recommendations.....	17
Conclusions.....	18
References.....	19

Introduction

To create meaningful data visualizations for the selected dataset. Individuals are required to go through data analysis processes starting from business understanding to data preparation. Individuals are also required to document and to produce detailed reports on steps performed to achieve the result.

In the early phases of the report, a step-by-step CRISP-DM analysis of the Hyundai Used Car Listing dataset was performed. During the Business Understanding phase, the dataset's focus was on Hyundai used vehicle listings, with the goal of uncovering critical insights influencing price and market dynamics. The dataset's properties were rigorously reviewed during Data Understanding, revealing light on critical variables like model, registration year, transmission type, mileage, fuel type, road tax, mpg, and engine size. Following that, feature engineering was undertaken in the Data Preparation phase to improve data quality, efficiently detect and eliminate outliers, and carefully handle missing values to maintain the dataset's integrity.

Moving on to Data Visualization, 10 unique graphics were created, each exposing a different aspect of the information. Notably, four crucial visualizations were emphasized, offering substantial insights into market demand, price drivers, the influence of fuel efficiency, and the impact of road tax. These visuals enable companies to make more informed decisions. Actionable suggestions were produced in the latter sections, emphasizing the necessity of exploiting these insights for strategic positioning and operational excellence in the used Hyundai automobile market.

Business Understanding

Business Understanding :

Gaining a thorough grasp of the variables that affect pricing and attractiveness is essential for success in the dynamic world of used automobile sales. This dataset, which collected from Hyundai Used Car Listings, provides a thorough look at the complex web of factors influencing the resale value of Hyundai cars. The dataset provided information from the model and year of registration to the gearbox type, mileage, fuel economy, road tax, and engine capacity. With the knowledge gained from this research, firms should be able to confidently and accurately negotiate the competitive environment of the used Hyundai automobile market when making decisions about pricing strategies, inventory management, and market positioning.

The variables impacting the pricing and performance of Hyundai cars are the focus of the business information that can be gleaned from the dataset. The following conclusions are conceivable:

1. **Market Demand:** Companies may determine which Hyundai models are more in demand in the used automobile market by looking at the distribution of Hyundai models and their level of popularity over time. This knowledge might assist dealerships in concentrating on supplying and advertising the most popular models.
2. **Tax Implications:** Understanding how road tax influences used Hyundai car prices may assist businesses comprehend how government laws and tax rates affect the resale market. This knowledge may be beneficial in forecasting changes in consumer behavior as a result of tax changes.
3. **Transmission Preferences:** Companies can determine if automatic or manual transmission vehicles are more popular in the used Hyundai market, which can help them make inventory selections.
4. **Year of Registration:** Examining how the car's age (year of registration) influences price can give insight into the depreciation rate of Hyundai vehicles as well as their perceived worth over time.
5. **Data Quality:** It is critical to ensure data quality and consistency in aspects such as mileage and engine size. Inaccuracies in these factors may result in inaccurate pricing or misleading insights.

Data Understanding

The dataset that was chosen is based on Hyundai Used Car Listing from kaggle. It consists of 9 attributes. Model, Year, Price, Transmission, Mileage, FuelType, Tax, Mpg, EngineSize.

Below are Description for each attributes :-

Hyundai Used Car Listing

Attribute Description

Attribute Name	Description	Dependent / Independent	Type Data
Model	Hyundai Model	Independent	String
Year	Year of Registration	Dependent	Float
Price	Price in Euro	Independent	Float
Transmission	Type of Gear Box	InDepended	String
Mileage	Mileage Distance Used	InDependent	Float
FuelType	Engine Fuel	InDependent	String
Tax(£)	Road Tax	InDependent	Float
Mpg	Mile Per Gallon	InDependent	Float
EngineSize	Size In Liters	InDependent	Float

Attributes contribute to data analysis

1. **Model:** The Hyundai automobile's model may be an important aspect for data analysis since it can assist in discovering certain car models that are in great demand or command premium prices in the used car market. This data is useful for inventory management and marketing tactics.
2. **Year of Registration:** Understanding how a car's age affects its price and appeal requires knowing the year of registration. Newer models are often more expensive and may be in more demand.
3. **Transmission Type:** Consumer preferences can be shown by analyzing the distribution of automated and manual transmissions. This data can have an impact on inventory selections and pricing tactics.
4. **Mileage** is an important aspect in judging the condition and value of a used vehicle. Lower mileage frequently corresponds with higher prices, making it an important factor in pricing analyses.
5. **Fuel Type:** Because of variances in fuel costs and environmental concerns, fuel type, such as gasoline or diesel, might influence customer preferences. Understanding how fuel type affects price might assist firms in tailoring their products.
6. **Tax:** A road tax is a governmental expense that varies by location and can affect the entire cost of ownership. Analyzing road tax data can reveal how regulatory issues influence pricing.
7. **Miles per Gallon (mpg):** Fuel efficiency, as measured by mpg, is a desirable feature, particularly in an environmentally concerned market. Automobiles with greater mpg ratings may cost a higher premium.
8. **Engine Size:** The size of an engine in liters can affect the power and performance of a vehicle. Larger engine sizes may result in higher pricing, and this feature might aid in market segmentation based on performance preferences.

Data Preparation

Data preparation is an important step in the CRISP-DM technique because it cleans, transforms, and preprocesses the data so that it is suitable for modeling. This step is essential because the quality of the data utilized for modeling has a direct impact on the model's accuracy and efficacy. Overall, data preparation is an important step in the CRISP-DM process, and it necessitates meticulous attention to detail to guarantee that the data is of good quality and acceptable for modeling.

In this report, there are several technique that been applied for data preparation, that will be explained Below:-

Feature Engineering:

The process of developing new features (variables) from existing ones in order to improve the performance of machine learning models or extract more relevant information from data is known as feature engineering.

Below are some feature engineering:-



```
df['model'] = df['model'].str.replace(' ', '')
df
✓ 0.0s
```

[Fig 1.0]

In fig 1.0, in the dataset column 'model', which has a unique value there is an empty space in the beginning for each value, by using `str.replace(' ', '')`. It will remove the empty space in the beginning of value for each row. For example: ' Tucson' to 'Tucson'.

```
bins = [-1,10000,20000,30000,40000,50000,60000,140000]
groupname = ['0<10000', '10000>20000', '20000>30000', '30000>40000', '40000>50000', '50000>60000', '>60000']
df['category_mileage'] = pd.cut(df['mileage'],bins,labels=groupname)
```

✓ 0.0s

Python

[Fig 1.1]

```
bins = [1999,2005,2010,2015,2020]
groupname = ['2000>2005', '2005>2010', '2010>2015', '2015>2020']
df['category_year'] = pd.cut(df['year'], bins, labels= groupname)
```

✓ 0.0s

Python

[Fig 1.2]

```
bins = [-1,100,150,200,600]
groupname = ['<100', '100>150', '150>200', '>200']
df['category_tax'] = pd.cut(df['tax(£)'],bins,labels=groupname)
```

✓ 0.0s

Python

[Fig 1.3]

In (fig 1.1, fig1.2, fig 1.3), In order to generate new columns, it uses pd.cut() method where used to bin (or categorize) continuous data into discrete intervals, it requires both the bins and labels parameters. as bins, The bins parameter indicates how the continuous data should be divided into intervals. It establishes the limits of these intervals. Labels: The labels option specifies which labels or names should be assigned to the generated bins.

Detecting and Treating Outlier

```
print ("Before")
print(df.loc[(df['mileage'] < 1)])
print(df.loc[(df['year'] > 2023) | (df['year'] < 2000)])
df.drop(df[(df['year'] < 2000) | (df['year'] > 2023)].index, inplace=True)
df.drop(df[(df['mileage'] < 1)].index, inplace=True)
print ("After")
print(df.loc[(df['mileage'] < 1)])
print(df.loc[(df['year'] > 2023) | (df['year'] < 2000)])
```

✓ 0.0s

[Fig 1.4]

In fig 1.4. In this dataset, it collected the model year from 2000 to the current year, but in this dataset, there's a present outlier where the model is below 2000 and above current year. There's also an outlier present on column mileage where the mileage is below 0. With code on Fig 1.4, firstly will detect then remove it by using df.drop() with inplace='True', function inplace is a code where it will update the original dataset instead of putting it into another variable.

Treating Missing Value:

```
df.drop(df[df.isna().sum(axis=1) >= 2].index, inplace=True)
print (df)
```

✓ 0.0s

[Fig 1.5]

In fig 1.5, Using code 'df.drop(df[df.isna().sum(axis=1)>=2].index, inplace = True)' drops any row that contains 2 or more missing values. The reason behind dropping 2 or more missing values for each row, is that it cannot be fixed due to lack of characteristic and even the possibility of outlier is high.

```
df.dropna(subset=['model'], inplace=True)
modeyear_model = df.groupby('model')['year'].apply(lambda x: x.mode().iloc[0])
df['year'].fillna(df['model'].map(modeyear_model), inplace=True)

modeprice_model = df.groupby('model')['price'].apply(lambda x: x.mode().iloc[0])
df['price'].fillna(df['model'].map(modeprice_model), inplace=True)

modetransmission_modelyear = df.groupby(['model', 'year'])['transmission'].apply(lambda x: x.mode().iloc[0]).reset_index()
df = df.merge(modetransmission_modelyear, on=['model', 'year'], how='left', suffixes=('', '_mode'))
df['transmission'].fillna(df['transmission_mode'], inplace=True)
df.drop(columns='transmission_mode', inplace=True)

modeengine_model = df.groupby('model')['engineSize'].apply(lambda x: x.mode().iloc[0])
df['engineSize'].fillna(df['model'].map(modeengine_model), inplace=True)

modempg_modelyeartransmission = df.groupby(['model', 'year', 'transmission'])['mpg'].transform(lambda x: x.mode().iloc[0])
df['mpg'].fillna(modempg_modelyeartransmission, inplace=True)

modetax_mpgmodelyeartransmission = df.groupby(['model', 'year', 'transmission', 'fuelType'])['tax(£)'].transform(lambda x: x.mode().iloc[0])
df['tax(£)'].fillna(modetax_mpgmodelyeartransmission, inplace=True)
```

[Fig 1.6]

In Fig 1.6, After remove 2 or more missing value that present in each row, there will be missing value still exist in dataset, with this code above will fill those missing value as more explanation below :-

df.drop (subset=['model'],inplace = True):

Dropping missing value where the model is missing, the reason for removing it from the dataset, it cannot be filled in because the code cannot find the same characteristic of the model that has missing value, and also if using replace , the chance of the wrong model is high.

modeyear_model:

Using function lambda and mode, it will try to find the characteristic that has on each model of the car by using car model, then after getting the result , it will take the mode of the year from the result and fill it to the missing value that has on year columns. For eg. tucson 2015 , then cars that have missing value of year will be filled by the mode of year which have the same name model.

modetransmission_modelyear:

Using function lambda and mode, it will try to find the characteristic that has on each model of the car by using car model and car year, then after getting the result , it will take the mode of the transmission from the result and fill it to the missing value that has on transmission columns.

modeengine_model

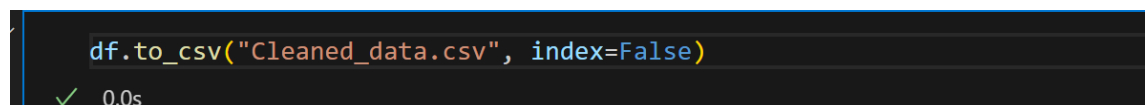
Using function lambda and mode, it will try to find the characteristic that has on each model of the car by using car model, then after getting the result , it will take the mode of the engine on specific characteristic from the result and fill it to the missing value that has on engine columns.

Modempg_modelyeartransmission:

Using function lambda and mode, it will try to find the characteristic that has on each model of the car by using car mode,year and transmission, then after getting the result , it will take the mode of the mpg on specific characteristic from the result and fill it to the missing value that has on mpg columns.

Modetax_mpgmodelyeartransmission:

Using function lambda and mode, it will try to find the characteristic that has on each model of the car by using car mode,year, transmission and mpg, then after getting the result , it will take the mode of the tax on specific characteristic from the result and fill it to the missing value that has on tax columns.



```
df.to_csv("Cleaned_data.csv", index=False)
```

✓ 0.0s

[fig .16]

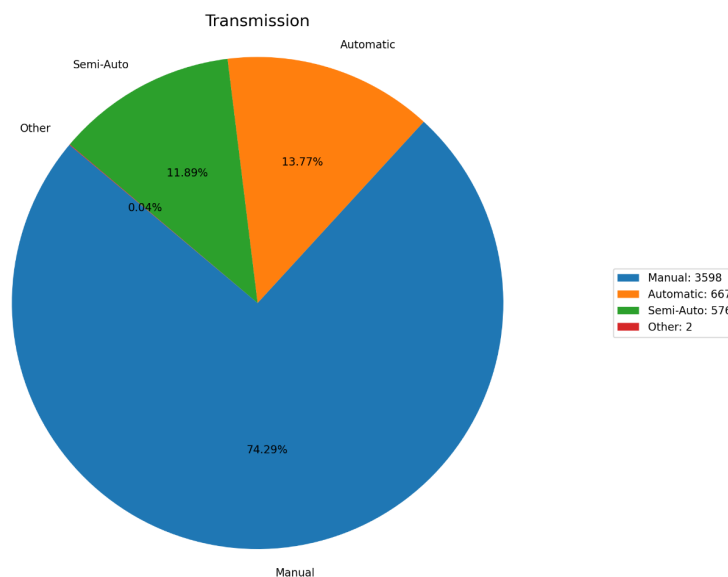
After cleaning the dataset, with removing and filling missing values and outliers, code on fig 1.6 will save the cleaned data into the computer path.

Data Visualisation

Data visualization is an important stage in the data interpretation phase of CRISP-DM (Cross-Industry Standard Process for Data Mining). It entails visually exploring and analyzing data via graphical representations such as charts, graphs, and plots. Data visualization assists data analysts and data scientists in gaining insights into the dataset's properties, trends, and interactions between variables. It helps them to detect trends, outliers, and anomalies, making informed decisions about data pretreatment, feature selection, and modeling techniques simpler. Finally, data visualization in CRISP-DM helps with data interpretation, which is necessary for constructing efficient data mining models and making data-driven business choices.

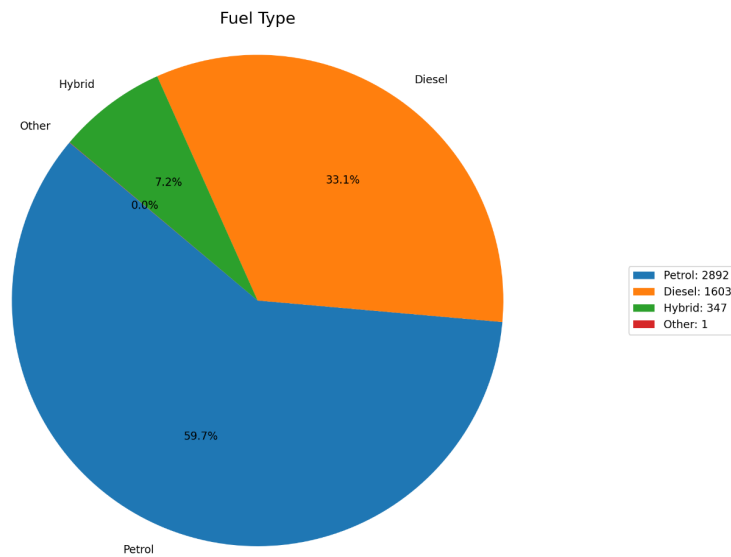
Below are 10 visualization based on dataset :-

6 Visualization



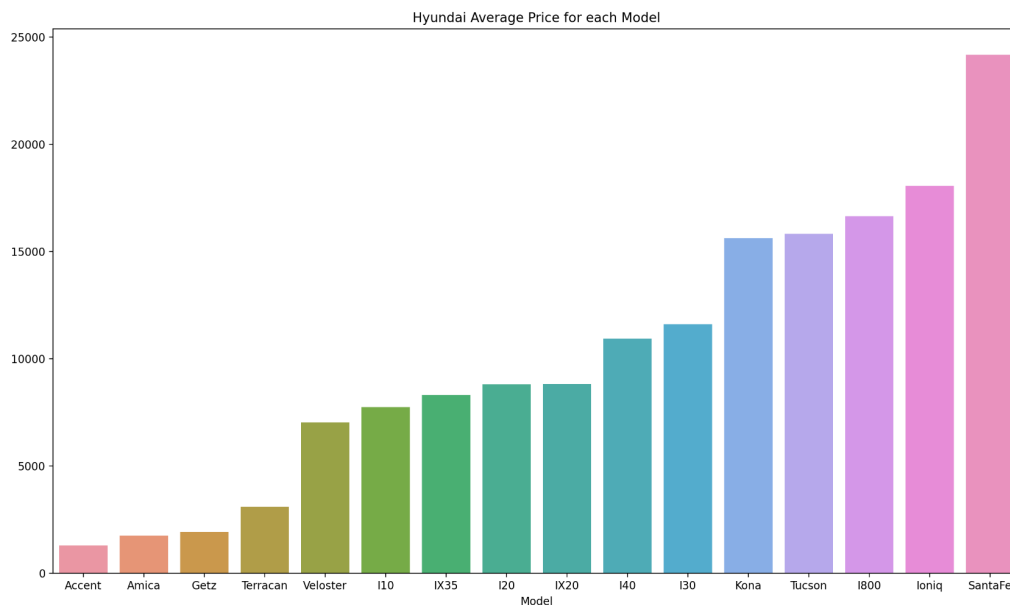
[fig 2.1]

Figure 2.1, show that the highest value is Manual with percentage of 74.29% which the majority of the car that been sell is manual car, however the lowest is 0.04% belong to other type of Transmissions



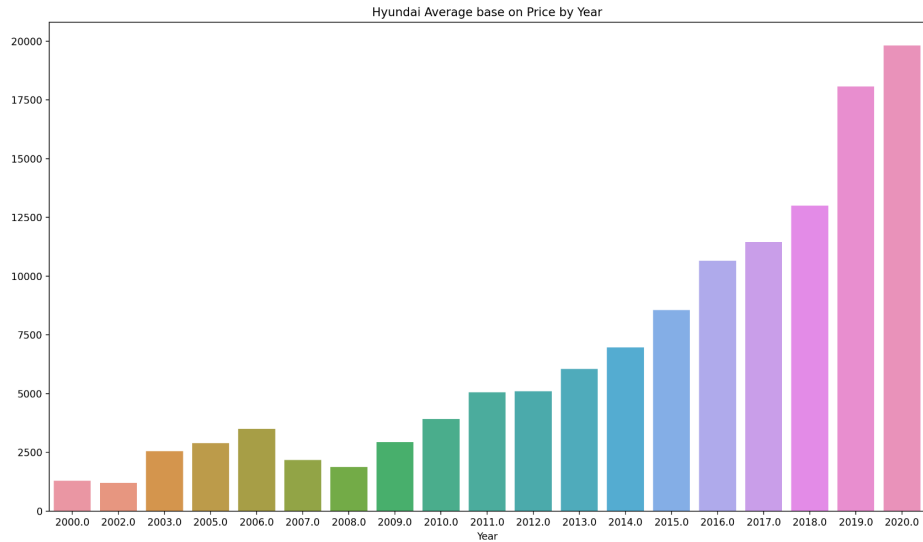
[fig 2.2]

In fig 2.2, show a pie chart where the highest value of fuel type is Petrol with 59.7% while the lowest value is other with 0.02%.



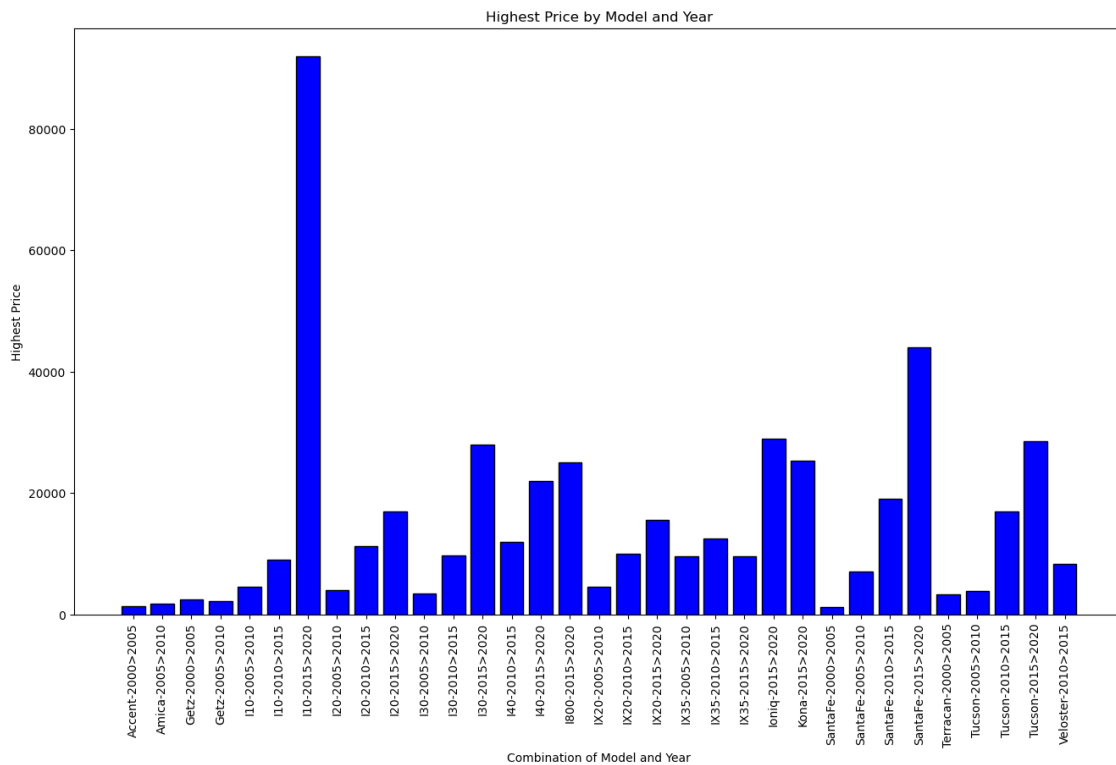
[Fig 2.4]

In fig 2.4, Bar graphs show Car models based on average price, the highest car model is SanteFe while the lowest average price is model Accent.



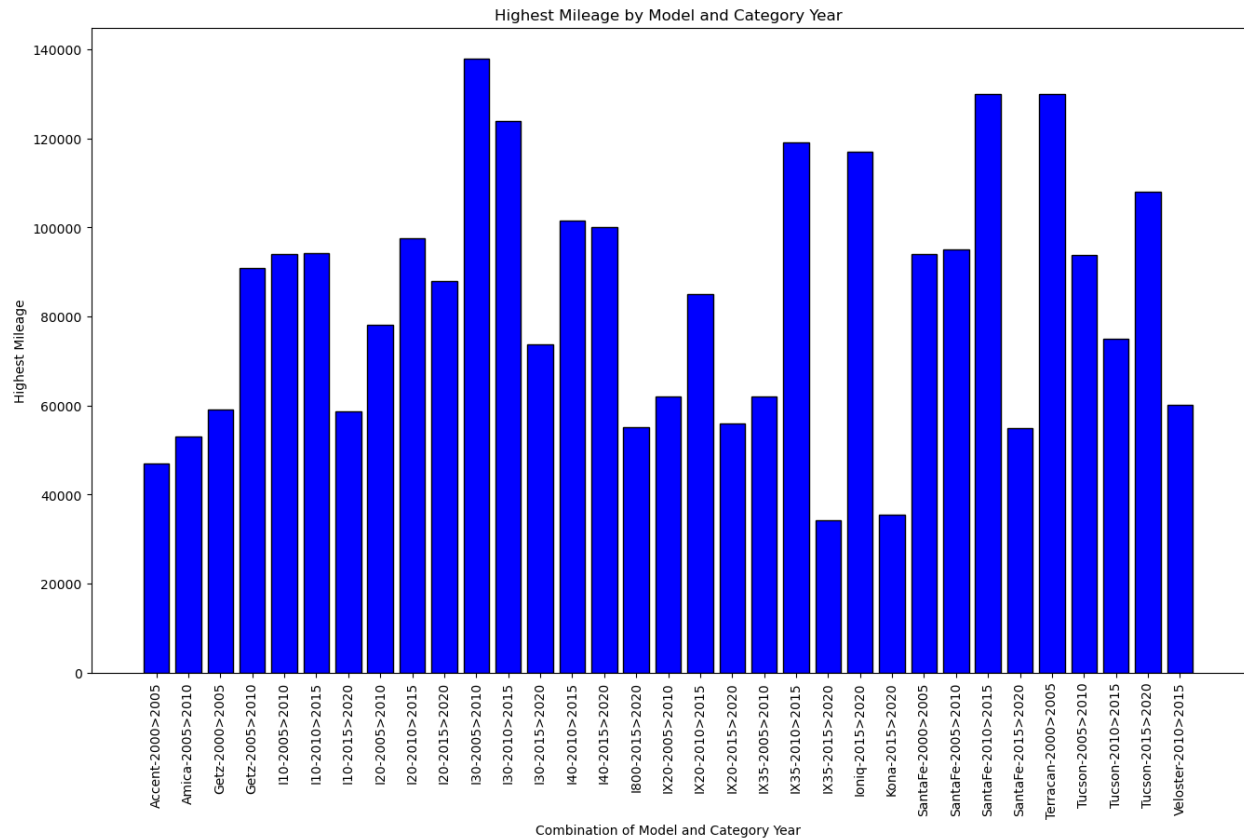
[Fig 2.5]

In fig 2.5, Bar graphs show car model by year based on average price, it tells that 2020 model car is the highest, while the lowest year price is 2002.



[Fig 2.5]

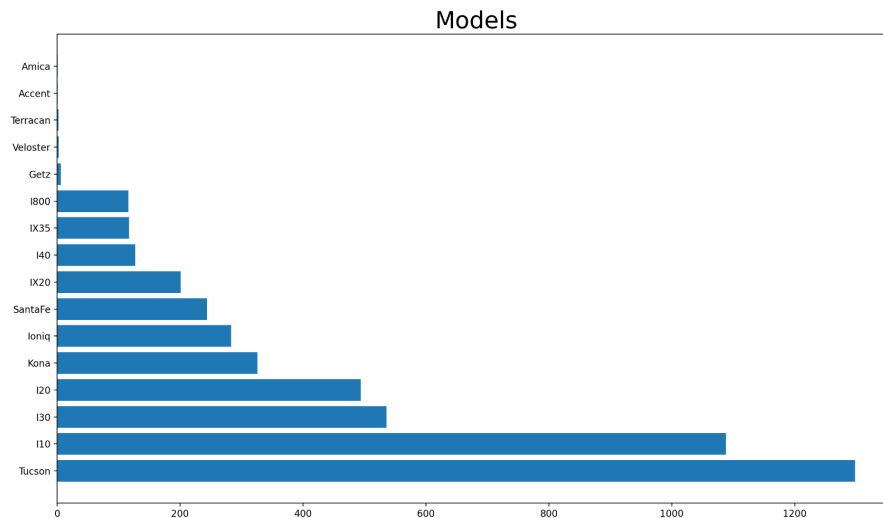
In Figure 2.5, bar graphs show that the highest price based on Model and year of the models, the highest price is car model I20 from 2015>2020. While the lowest is Accent from 2000>2005 model.



[Fig 2.6]

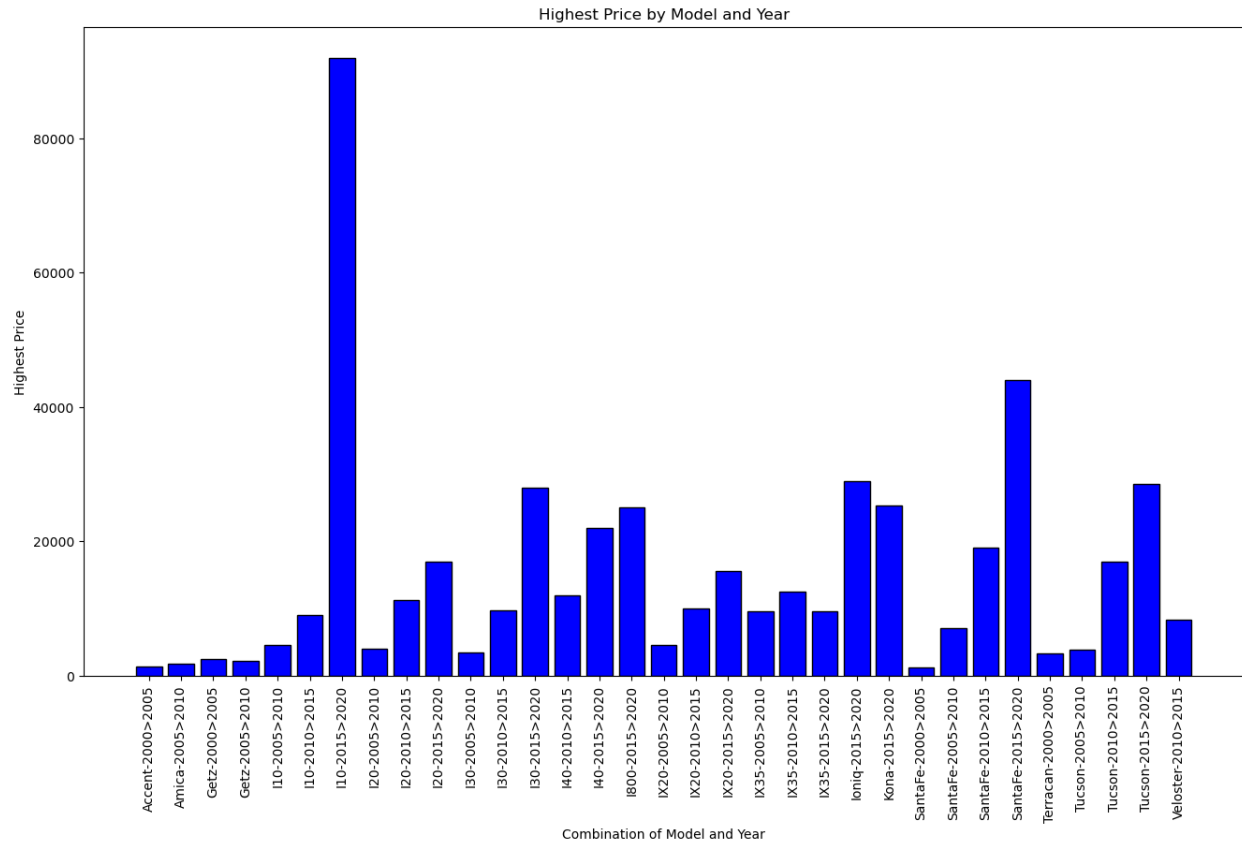
In Fig 2.6, Bar graphs show the highest mileage based on model and year of the car is I30 from 2005>2010. However the lowest mileage is IX35 from 2005>2010.

4 Important Visualization



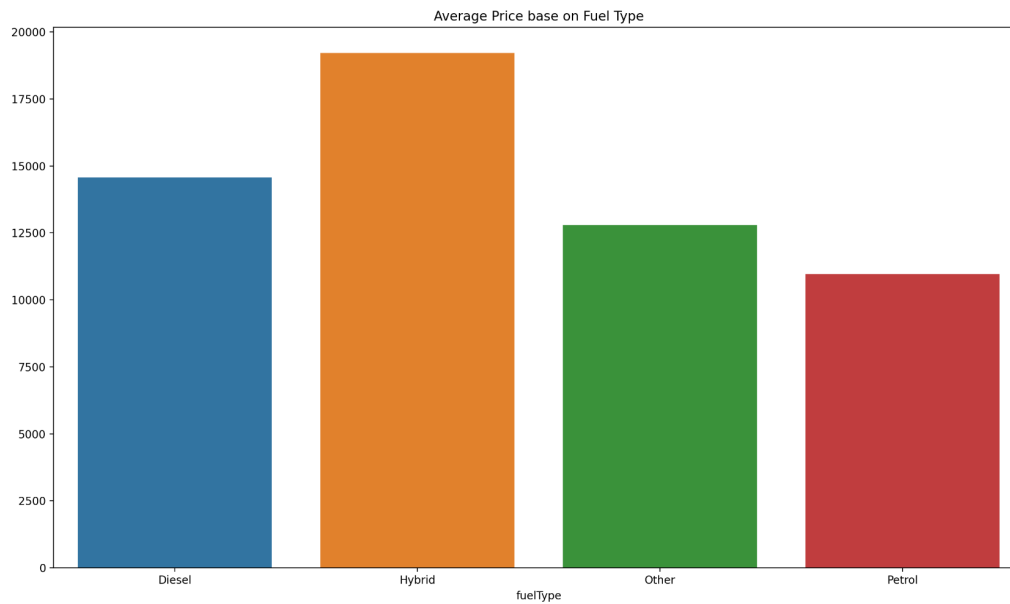
[fig 3.1]

In Figure 3.1, it show a bar in horizontal that the highest car that been sell in the market which is Tucson, while the lowest Amica and accent, Using a case study Hyundai's Tucson model is one of the company's core products in Europe, and it is the company's best-selling SUV internationally, with more than 7 million units sold worldwide since its inception in 2004, including 1.4 million in Europe. (Capparella, 2021) While Amica and Accent, Consumer preferences have shifted: Sales of the B-car sector, which includes models such as the Accent and Amica, have declined in recent years. This might be attributed to a trend in customer tastes toward bigger cars like SUVs and crossovers. (*New Accent Launching Amid B-Car Slump*, 2017)



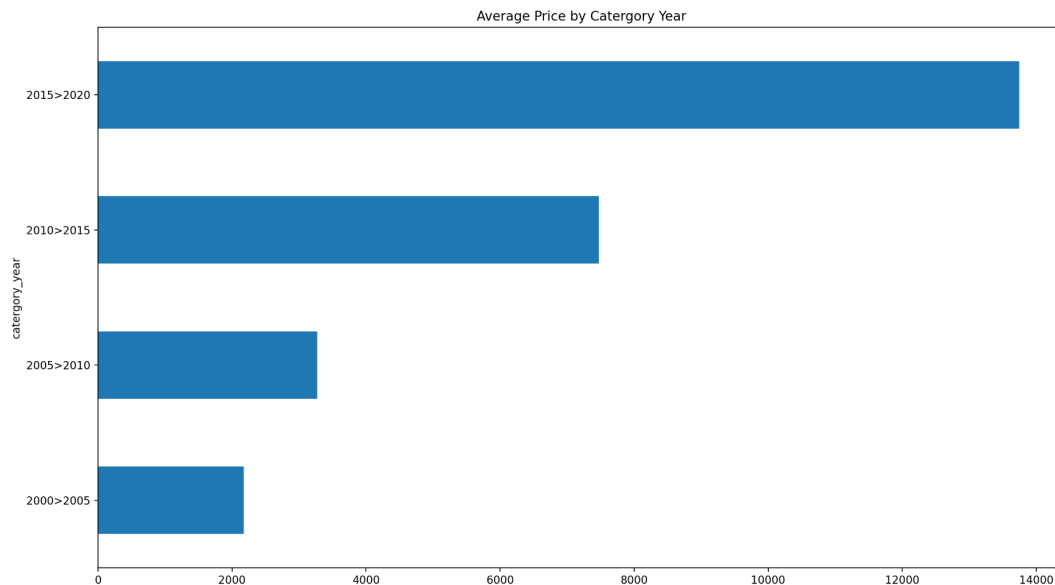
[Fig 3.2]

In fig 3.2 , show that I10 from 2015 to 2020 has the Highest price with above 80000, while the lowest is santafe from 2000 to 2005. According to Autocar, the second-generation Hyundai i20 is a highly large, well-equipped, and reasonably priced entrant to the competitive supermini sector, (Hyundai I20 2015-2020, 2023). This could tell that the reason behind the I10 having the highest price is its newest model that comes into the business , while santafe is an old model from 2 decades ago, it could be the reason why the price of it is low.



[Fig 3.3]

In Fig 3.3, tell that Hybrid fuel type is the highest value. The potential reason behind it could be with hybrid type it could be easy for the customer or user to fill their fuel with different type of fuel and not stuck with 1 fuel only.



[Fig 3.4]

In fig 3.4 , show a horizontal bar graphs show that model from 2015-2020 has the highest price among other model year, it tell that most of people looking for the latest model due to new model could much last longer than old one. The EU's emission standards are growing increasingly stringent, putting pressure on auto makers to develop more fuel-efficient and ecologically friendly vehicles. This might be another reason why Europeans are more interested in Hyundai's newer models, since they are more likely to fulfill these pollution standards.

Recommendations

These visualizations may assist organizations in understanding how elements impact model and year, as well as other attributes that may be the reason why their firm is trending, particularly personal aspects. The properties in the dataset are straightforward, making it easy to view the data. However, some visualizations cannot be further described since the data provided is insufficient to establish the meaning of the visualization generated. It is advised that a specific dataset be provided in order to produce a more accurate result and examine the depth of created visuals. Aside from that, building visualizations is limited because the data provided is primarily classified and dependent. Due to the constraint, businesses may not get the desired objectives. As a result, a greater variety of data in the dataset is advised for practice and obtaining the desired results.

Conclusions

Data visualization has been shown to have an impact since it provides information and insight through various types of representations. The goal of this study is to develop useful data visualizations after going through data analysis techniques. Due to missing variables, the hyundai used vehicle listing dataset from Kaggle must go through Data Preparation before it can be visualized. A total of ten visualizations are developed with Visual Studio Code, and four of them are judged to be essential in connection to the business objectives.

These visualizations can provide a concise overview of how company objectives might be met, as well as methods for improving the firm and competing with other competitors.

References

Hyundai used car listing. (2021, September 24). Kaggle.

<https://www.kaggle.com/datasets/mysarahmadbhat/hyundai-used-car-listing?datasetId=1609240&sortBy=voteCount>

User Guide — pandas 2.1.0 documentation. (n.d.).

https://pandas.pydata.org/docs/user_guide/index.html#user-guide

Examples — Matplotlib 3.7.2 documentation. (n.d.).

<https://matplotlib.org/stable/gallery/index.html>

Kimseg. (2023). [eng, kor]Hyundai_Car(EDA, XGB, GridSearchCV). *Kaggle*.

<https://www.kaggle.com/code/kimseg/eng-kor-hyundai-car-eda-xgb-gridsearchcv#Feature-engineering>

Yutotsubaki. (2021). EDA, Visualization, and Linear Reg ($R^2 > 0.93$)! *www.kaggle.com*.

<https://www.kaggle.com/code/yutotsubaki/eda-visualization-and-linear-reg-r-2-0-93>

Capparella, J. (2021, January 27). 2022 Hyundai Tucson N Line previewed by Europe-Spec Model. *Car And Driver*.

<https://www.caranddriver.com/news/a35337651/2022-hyundai-tucson-n-line-europe/>

New accent launching amid B-Car slump. (2017, September 28). WardsAuto.

<https://www.wardsauto.com/industry/new-accent-launching-amid-b-car-slump>

Hyundai i20 2015-2020. (2023, February 21). Autocar.

<https://www.autocar.co.uk/car-review/hyundai/i20-2015-2020>

Boston, W. (2023, January 16). EVs Made Up 10% of All New Cars Sold Last Year. *WSJ*.

<https://www.wsj.com/articles/evs-made-up-10-of-all-new-cars-sold-last-year-1167381838>