Assignment 3: Data Analytics, Group 60

188.429 Business Intelligence (VU 4,0) – WS 2022/23

William Amminger
e12229199@student.tuwien.ac.at
12229199
Person B

Zsombor Iszak
e11709501@student.tuwien.ac.at
11709501
Person A

ACM Reference Format:
William Amminger and Zsombor Iszak. 2023. Assignment 3: Data Analytics, Group 60: 188.429 Business Intelligence (VU 4,0) – WS 2022/23. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Business Understanding

(a) **Scenario:** This dataset contains the latest information on car prices in Australia for the year 2023. It covers various brands, models, types, and features of cars sold in the Australian market. It provides useful insights into the trends and factors influencing the car prices in Australia. The dataset has over 16k records of car listings from various online platforms in Australia.
The business scenario is that a car dealership located in Germany wants to expand and get into the Australian car market. However, before penetrating the market some research on the market needs to be done in order to set up a business model and a pricing strategy.

(b) **Business Objectives:** The main business objective of the company is to get an idea about the current pricing practice in the Australian car market.

(c) **Business Success Criteria:** This overall picture about the pricing strategies is crucial in order to be able to determine, for what price the cars can be sold. This prediction impacts the profitability of the enterprise as the selling price directly impacting the profitability of the company. It helps the dealer to better approximate what is the realistic selling price for given car is.

(d) **Data mining goal:** Extract patterns and dependencies that influence the price of the cars. Finally a regression model needs to be built on top of those patterns, that predict the future selling price of the cars.

(e) **Data Mining Success criteria:** Come up with a model, that delivers reasonable results based on the input features of a car.

(f) **AI Risk Aspects:** A concern of ours is that both selling price and local parameters appear in the data. Therefore one can run into conclusions and categorizations about the financial state of inhabitants of specific towns. To be more precise the data shows, that in 450 cases out of 16734 the location is not given. This might be just due to poor data maintenance, however it should be considered, that for some cases it was a sensitive information and was not published intentionally.

2 Data Understanding

(a) **Attribute Types:** For the first view on the data, one can immediately recognise, that most of the columns contain strings, however in those string in many cases numeric information is stored with inclusive units, such as '5 Doors' or '6.7 L / 100 km'. Regarding the missing value it turns out, that beside the missing cells, some cells contain the string '-'. The number of missing values can be seen in the figure 1. Additionally the columns FuelType and BodyType contain a limited number of "Other" string value. In those cases it is also not quite clear what was meant by that value and as the factors covered in our opinion all the possibilities, we assumed, that "Other" means some unique features and is not a homogeneous class by its own, therefore those values were also labeled as NaN.
On a per column basis:

- Brand: `string` - the manufacturer (make) of the vehicle
- Year: `integer` - the model year
- Model: `String` - the model name itself
- Car/Suv: `String` - this column is not quite so clear, as sometimes it contains a custom string about the car type, while usually it mirrors the "BodyType" column
- Title: `String` - full title of the car, including the year, make, model, & sometimes the trim type
- UsedOrNew: `String` - indicates USED, NEW, or DEMO
- Transmission: `String` - Automatic or Manual

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

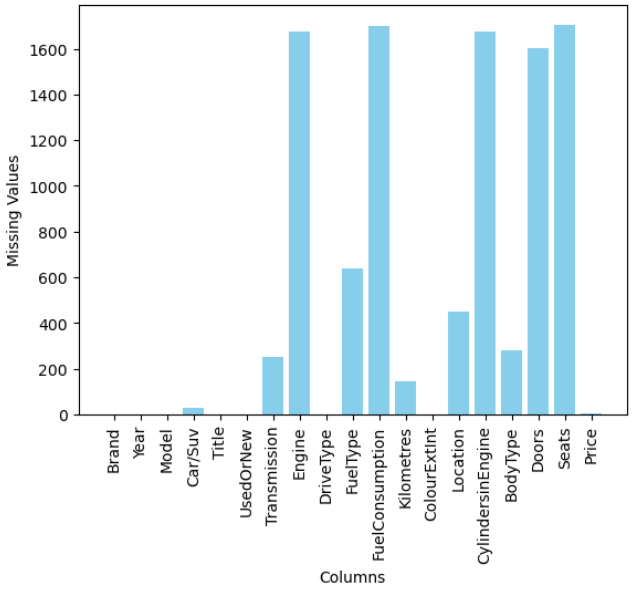

Figure 1. The missing values for each column in the dataset

- Engine: `String` - specifications in the format # cyl, # L
- DriveType: `String` - indicating which axles have power (AWD, Front, Rear)
- FuelType: `String` - fuel type required by car, such as Diesel, Premium, Unleaded, ...
- FuelConsumption: `String` - efficiency in the format # L / 100 km
- Kilometres: `integer` - current odometer reading
- ColourExtInt: `String` - color in the format (color) / (color), though this data is very heterogenous & sometimes is missing the interior color or contains the trim color
- Location: `String` - location in the format (City), (Territory)
- CylindersinEngine: `String` - number of cylinders in the format # cyl
- BodyType: `String` - body style of the car, such as SUV or Coupe. Most of the information this implies is encoded in other columns
- Doors: `String` - in the format # Doors
- Seats: `String` - in the format # Seats
- Price: `integer` - the target column, being the selling price of the vehicle

(b) **Statistical Properties:** The first step was to determine the number of unique values for each column:

- `Brand.....76`
- `Year.....45`
- `Model.....781`
- `Car/Suv.....618`
- `Title.....8804`
- `UsedOrNew.....3`
- `Transmission.....3`
- `Engine.....106`
- `DriveType.....5`
- `FuelType.....9`
- `FuelConsumption.....157`
- `Kilometres.....14262`
- `ColourExtInt.....834`
- `Location.....618`
- `CylindersinEngine.....11`
- `BodyType.....10`
- `Doors.....13`
- `Seats.....13`
- `Price.....3794`

Another interesting aspect is the correlation matrix, 2, which does not show any immediately promising correlations besides the year & Kilometres. As for the the target variable of "price", no variable correlates above |.35|

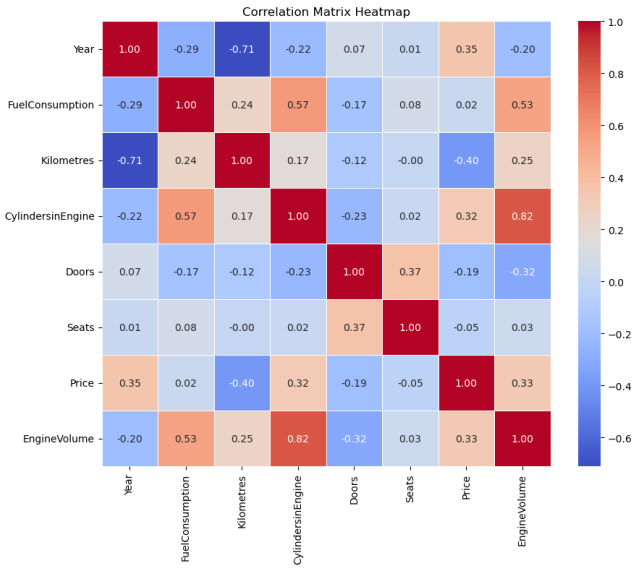

Figure 2. Correlation matrix between numerical columns

(c) **Data Quality:** To explain we took a closer look on examples with missing values but we couldn't identify any specific patterns. The Price column contains in 50 cases the string "POA" and we assume it stands for "Price on Asking", thus the prices is negotiable. However, we converted to column to numbers and "POA" was changed to NaN. After extracting the numeric values, the distribution of the values of the numeric columns is shown in figure 3.
As one can see the distribution plot for the target column Price is hard to read. Our assumption was, that the skewness can be explained by an outlier and by plotting the boxplot (figure 4) one can also show, that

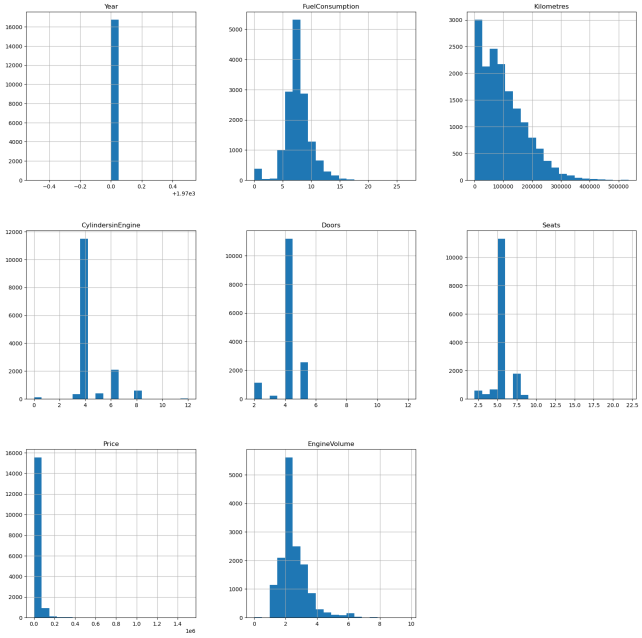

Figure 3. The distribution of the numeric columns

truly there is a significant outlier that is responsible for the skewness of the distribution plot. Another helpful visualization is a logarithmic histogram of the prices (figure 5)

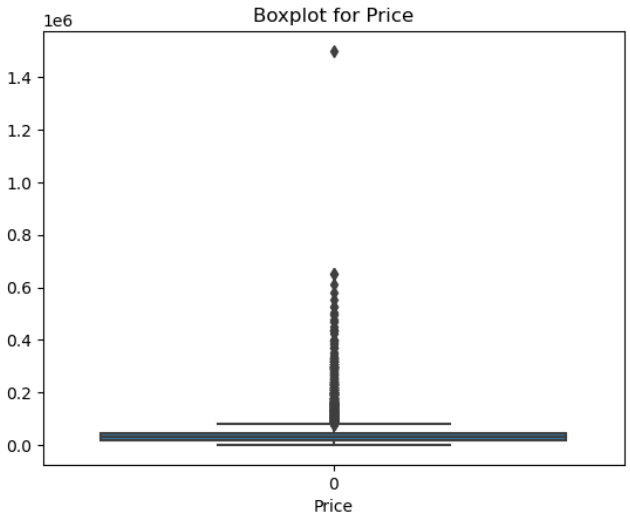

Figure 4. The statistical analysis of the target value

(d) **Visual Exploration:** Here it is useful to plot the distributions of some numeric variables. (Brand names 6) (Fuel Types 7) (Transmissions 8) (Used/New 9)

(e) **Ethically Sensitive:** In terms of data representation, there are many classes that are unbalanced, as shown by the previous graphs. There are many more cheap cars than expensive cars for example, which may cause the model to assume that everything is cheaper in order to maintain the distribution at the expense of true accuracy.

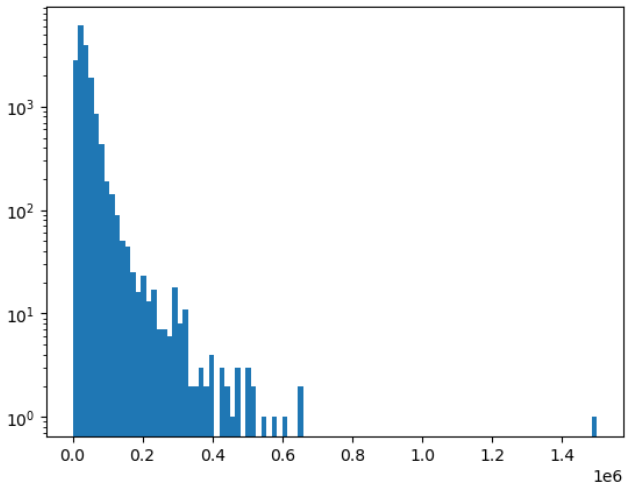

Figure 5. Histogram of the target value

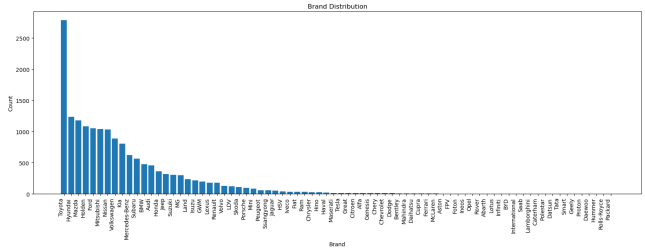

Figure 6. Distribution of brand names

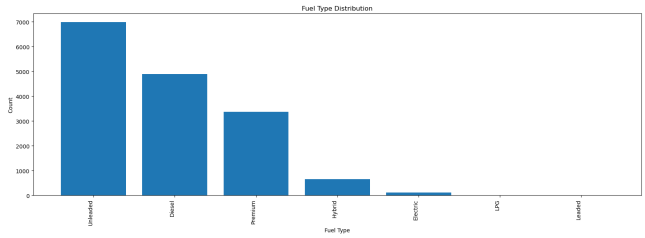

Figure 7. Fuel Type distributions

(f) **Risks & Bias:** A possible risk is that since a single entry has many values, it is possible to locate the owner of the vehicle. This is why the location was removed as mentioned previously. Since the title also does not contain information usable by machine learning, it will also be dropped. Furthermore, if this was used by people trying to sell their own car, the model could underprice it since the training data has a distribution skewed towards lower prices.

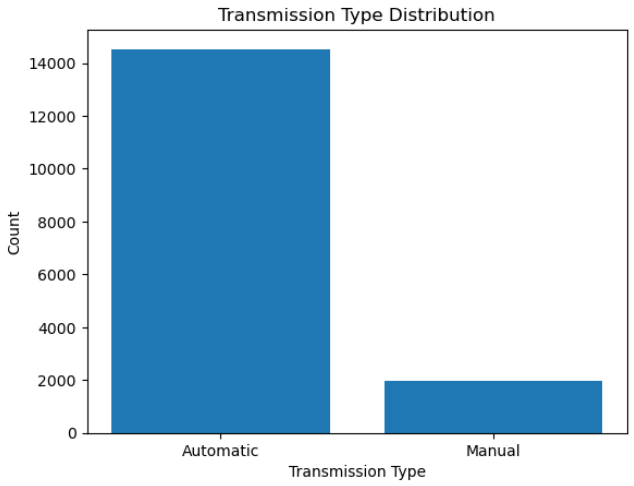

Figure 8. Transmission class balance

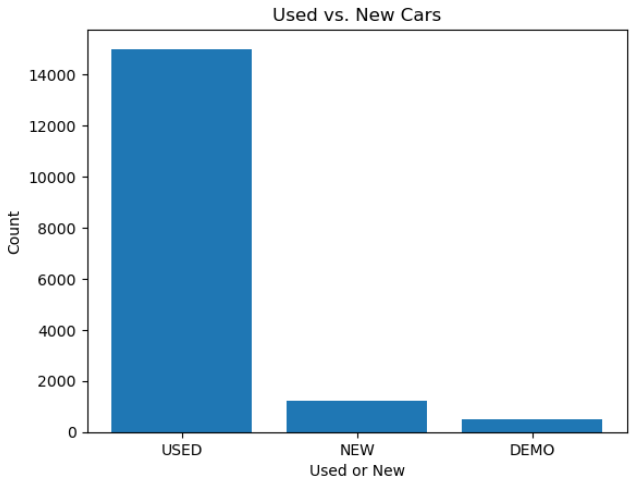

Figure 9. Used/New class balance

(g) **Required Actions:** One possible solution is normalizing the price information, so that the model receives & predicts a normal distribution. The result could then be reskewed back to the correct prices.

3 Data Preparation

(a) **Necessary Actions:**

- A number of operations involved converting string columns into integers with little complexity, such as FuelConsumption, Doors, & Seats, which involved using regex to extract the numerical values.
- The UsedOrNew column contained the values "DEMO", "USED", & "NEW". After some research, we determined that while demo cars are mostly new, they are slightly used by the dealership. Thus, some ordinality can be applied to these values. More specifically, converted Used, Demo, New to 1, 2, 3
- The transmission type was changed to the boolean IsAutomatic column. The "-" values were set to np.nan.
- The BodyType was also a subject of discussion. It was eventually concluded that to retain the most information, the categories would be ordered by their relative size to imply some ordinality. Although these body types contain other information (such as Ground Clearance, Number of Doors), this information was either found in other columns or of minimal importance. In the end the categories were sorted in increasing order as follows: Convertible, Coupe, Hatchback, Sedan, Wagon, SUV, Ute Tray, People Mover, Commercial and converted to numbers of the same order.

(b) **Derived Attributes:**

- The Engine column was split up into EngineCyl & EngineVolume. As the data about the number of cylinders was already included the column of CylindersinEngine, the EngineCyl was dropped.
- The ColourExtInt column was turned into simply the ExtCol column as most of the interior columns were missing.
- The Territory was extracted from the Location column to keep the data more anonymized while still retaining useful information
- The DriveType was split up into 2 columns: DriveFront & DriveRear, 2 boolean columns indicating if the vehicle was powered on that axle set.

(c) **External Data Sources:** There are numerous possibilities for expanding the business potential with more data.

- Supply chain information can show the trends of the new car market, which the used car market mimics to some extend. Disruptions in the supply chain of new vehicles can cause people to seek out used vehicles as alternatives, shifting demand. This information shows the relative abundance of vehicles with their pricing, allowing to make more informed decisions.
- Technological trends are playing an increasingly important role in the Auto industry. Datasets including more technical features of cars such as entertainment systems impact the price of modern cars. Having a better understanding of the technical offerings of cars can lead to more accurate pricing & exploitation of market trends in this respect.
- Economic indicators like unemployment rates, consumer confidence indices, & historical demand could allow the company to better forecast the potential selling price of a vehicle. For example, an increasing

job market in an area would also potentially increase vehicle demand & thus their prices.

-

(d) **Pre-Processing Steps:** Another issue encountered were several categorical variables that could not be encoded with some ordinality. These columns (Brand, ExtCol, State, FuelType) were 1-hot encoded, which means that each category is split into a seperate boolean columns indicating whether the observation had this feature or not.