Practical Data Science

Student Name: Kelvin Young
Student ID: s3899733

Data Preparation

The dataset used in this report has been provided by RMIT University; it contains information about the ownership of different car models. The features of this dataset, and their possible values, are listed below (extracted from the supplied data_descriptions.txt file):

Manufacturer ∈ ['Renault', 'BMW', 'Volkswagen', 'Peugeot', 'Ford', 'Nissan', 'Honda', 'Toyota', 'Mercedes', 'Audi', 'Citroen', 'Skoda', 'Land-Rover', 'Seat', 'Fiat', 'Mini', 'Saab', 'Hyundai', 'Kia', 'Jaguar', 'Mazda', 'Suzuki', 'Volvo', 'Rover', 'Mitsubishi', 'Smart', 'Porsche', 'Subaru', 'MG', 'Chrysler', 'Chevrolet', 'Alfa-Romeo', 'Daihatsu', 'Bentley', 'Daewoo', 'Dacia', 'Dodge', 'Lotus', 'Aston-Martin', 'Abarth', 'Ssangyong', 'Lexus', 'Maserati', 'Opel', 'Ferrari', 'TVR', 'Triumph', 'Lada', 'Daimler', 'Lancia', 'Datsun', 'Morris']

Model ∈ ['Clio', '320i', 'Polo', '206', 'Mondeo', 'Micra', 'Civic', 'Ka+', 'Megane', 'Yaris', 'CLA', 'A4', 'Passat', 'A3', '307', '207', 'Xsara', 'Fabia', '118i', 'Freelander', 'C3', 'Corolla', '535i', 'Avensis', 'E', 'Scenic', 'Ibiza', 'Range', 'Qashqai+2', 'Octavia', '500L', 'A', 'Almera', 'C4', 'C5', 'C6', 'Two', 'One', 'RAV4', 'RAV5', '9-3X', '107', 'Aygo', 'I10', 'I11', 'A6', 'Picanto', 'Note', 'Leon', 'XE', 'Fusion', 'Accord', 'MX-5', '6', 'C1', 'Swift', 'Galaxy', '308', '3', 'TT', 'Laguna', '2', 'Panda', 'Auris', '306', '106', 'V70', 'Saxo', 'Touran', 'Beetle', 'Rio', 'C-MAX', 'X5', 'Juke', 'Vitara', 'Ceed', '406', 'Getz', 'X-Trail', 'CLK', 'S40', '75', 'Berlingo', 'Alto', 'Primera', 'I20', 'SLK', 'S-MAX', 'ML', 'Sportage', '407', 'C2', 'I30', 'XF', 'A1', 'Colt', 'A5', 'S-Type', 'V50', 'Escort', 'XC90', 'V40', 'X3', 'Tiguan', 'B', 'Fortwo', 'Bora', '45', '911', 'XJ', 'Impreza', '09-May', 'Celica', 'Sharan', 'Santa', '208', 'S60', 'DS3', 'Modus', 'ZR', 'Voyager', 'Sorento', 'Z4', '2000', 'Up', 'Jimny', 'Matiz', 'Boxster', 'Wagon', '3008', '323', 'Puma', 'Stilo', 'Scirocco', 'XK', 'Altea', 'C30', 'Accent', 'Lupo', 'Sprinter', 'Ix35', 'Kangoo', '5', 'Jetta', 'Alhambra', 'Espace', 'Ignis', 'SX4', 'Carens', 'Q5', 'Partner', 'Yeti', 'Touareg', 'Seicento', 'Superb', 'Bravo', 'X1', 'Q7', '147', 'MR2', '740i', 'TF', 'S80', 'Fox', 'Twingo', 'Lancer', 'Doblo', 'Eos', 'M3', 'PT', 'Matrix', 'IQ', 'SLS', 'A2', 'Z3', 'C70', 'Roomster', 'Splash', 'Previa', 'Starlet', 'Aveo', 'XC70', 'Tucson', 'Legacy', 'Cayenne', '156', '640i', 'Spark', 'Kalos', 'Outlander', 'Pixo', 'FR-V', 'Sirion', 'Mini', 'Venga', 'Streetka', 'Felicia', 'Terrano', 'Space', 'Carisma', 'Lacetti', 'ZT', 'CLC', 'Defender', 'Arosa', 'HR-V', 'City-coupe', 'Terios', 'Carina', 'Multipla', 'A8', 'Citigo', 'Liana', 'Continental', 'Captiva', 'Pathfinder', '205', '900', 'B-MAX', '940', 'Toledo', 'Ix20', '850', 'Verso-S', 'S3', '159', 'Lanos', 'C8', '626', 'Sandero', '807', '5008', 'Cayman', '300C', '300', 'Duster', '1007', '508', 'Xantia', 'RCZ', '350Z', 'Demio', 'S2000', '190', 'Trajet', 'Tacuma', 'GT', 'Premacy', 'ZX', 'V60', 'Brava', 'S4', 'X6', 'Caliber', 'Maverick', 'Streetwise', 'CityRover', 'Cerato', 'Elise', 'Cougar', 'Roadster-coupe', 'Vantage', 'I40', 'Cruze', 'Exeo', 'Urban', 'Mii', 'A7', '405', '500C', 'Transporter', 'Spider', 'DS4', 'Picnic', 'Viano', 'Ulysse', 'Sierra', 'DB9', '80', 'Neon', 'Terracan', 'M550i', 'Magentis', 'AX', 'RS4', 'Captur', 'Grandis', 'Crossfire', 'CR-Z', 'R', 'Stream', 'Idea', 'Prelude', 'Elantra', '607', 'Rexton', 'CC', 'Roadster', '806', '9000', 'S5', 'Nemo', 'Sonata', 'Capri', 'Sedici', 'Brera', '944', 'Cordoba', 'C-Crosser', 'Cabriolet', 'Cinquecento', 'Galant', 'Qubo', 'Lantra', 'Baleno', '440', 'GS', '2008', 'S70', 'GT86', 'Caddy', '4007', 'GLK', 'Camry', 'Nitro', 'Sebring', 'Shuma', 'Veloster', 'Patrol',

'Expert', 'R8', 'Wind', 'Rapid', 'DB7', 'Corrado', 'Supra', 'Marea', 'Serena', 'Vaneo', 'Panamera', 'Copen', 'Logo', 'Paceman', '960', 'Tribute', 'M135i', 'Shuttle', 'Trafic', 'Journey', 'Phaeton', 'Justy', 'Orlando', 'Vito', 'Granada', 'Bipper', '370Z', 'RS6', 'John', 'Hilux', 'Legend', 'MX-3', 'GranCabrio', 'Scorpio', 'Rodius', 'Mustang', '121', 'Coupe', 'Integra', 'Carrera', 'Croma', 'Zafira', 'GT-R', 'DS5', 'Korando', '309', '100', 'Quattro', 'Probe', 'Focus', 'Fiesta', 'Golf', 'Discovery', 'XKR', 'Pride', '19', 'Concerto', 'Leganza', 'Bluebird', 'MPV', 'Xedos', 'Pajero', 'Orion', 'Vento', 'Nexia', 'Mentor', '460', '166', 'Maxima', 'Paseo', 'F430', '850i', 'Uno', 'Favorit', '146', 'MX-6', '145', 'Safrane', 'Move', '928', 'Musso', 'Cooper', 'Tipo', 'Espero', 'XM', 'CRX', 'Tuscan', 'Vel', '21', 'Prairie', 'SJ', 'Esprit', 'Samurai', '360', 'Tempra', '155', 'Barchetta', '90', 'TR7', 'Pony', '605', '25', '164', '11', '505', 'Marbella', 'Acclaim', 'X-90', 'Samara', '305', 'Griffith', '9', '33', 'Manta', 'Applause', '4', '3.6', 'Delta', 'Silvia', 'Sunny', '126', 'Tercel', 'Niva', 'Stellar', '99', '120', 'Kadett', '18', 'Dedra', 'Cherry', 'Marina', 'Laurel', 'Regata', '130', 'Thema', 'Stanza', 'Ital', 'Strada', 'Santana', 'Double', '105', 'Malaga', 'Ascona', 'Monza', 'Fuego', 'Derby', '104', 'Transit', 'Prisma', 'Senator', '127', '200', 'Rekord', '929', 'Cressida', 'Quintet', '14', '20', '1200', '504']

Price ∈ (0.0, 650.0]
(unspecified unit)
Transmission ∈ (0.0; 10.0]
(unspecified unit)
Power ∈ (0.0, 500.0]
(BHP)
Engine size ∈ (0.0, 6500.0]
(CC)
Fuel ∈ [‘petrol’, ‘diesel’, ‘automatic’]
Male ∈ (0, +∞)
(number of owners)
Female ∈ (0, +∞)
(number of owners)
Unknown ∈ (0, +∞)
(will be referred to as unclassified/unclassified gender from this point on)
Total ∈ (0, +∞)
(defined as Male + Female + Unknown)

This information is provided here as it will be frequently referenced further in this report, particularly within the data cleaning process.

The steps taken in order to clean the data are as follows (within the accompanying .ipynb file these steps have been labelled as comments on their respective code cells):

1. Drop rows with null values
2. Typecast numerical discrete features from string into int
3. Eliminate numeric data beyond constraints
4. Eliminate categorical data beyond constraints
5. Fix gender count summation
6. Remove outliers

Error 1: Null values

We check for null values first to avoid raising errors when trying to clean other parts of the data. First we check how many null values exist in the dataset, revealing only 44 null values spread across all features. We consider substituting null values with interpolated values, however according to our subject matter expert “it is impossible to predict for example which type of fuel a car model would use, given no other information about the specs themselves”.

This dilemma extends to interpolating numeric features such as price; conventionally substituting column averages would be unreasonable as price would no longer be proportional to a specific car’s other preexisting attributes. Our subject matter expert also credits the existence of other factors such as brand value, nostalgia value, and scarcity value (of which are beyond the scope of this dataset), as legitimate considerations of a car to have a statistically questionable value.

Therefore it has been decided the simplest and least destructive approach is to simply remove the rows containing null values. Only 5 rows have been lost from this approach.

Error 2: Mismatching data types

We check the column data types to ensure the data is properly usable for graphing. Upon manual inspection of the data the Male, Female, Unknown, and Total columns are found to be stored as strings (defined as object in ipython3; an object constitutes any non-numeric data type) instead of a numerical data type. In addition, commas have been used for digit grouping.

The Male, Female, Unknown, and Total features will be type casted into integers. The digit separation commas must be removed in order to prevent a type cast error; the following functions are used:

pandas.DataFrame.str.replace

pandas.DataFrame.astype

Error 3: Illegal numerical data entries

To eliminate numerical data beyond the given constraints, we split this task to separately tackle negatives, positives, and zeros.

First we check for negative numeric feature values, revealing that the price, transmission, and power features contain illegal negative values.

It would be theoretically possible for data entry to accidentally input a negative symbol; compounded by the fact that the absolute value of a negative value is often still usable (for example a value of -94 could have possibly been a misinput of +94, well within the supplied constraints).

Therefore instead of culling negative values, an absolute function is mapped across the numerical continuous features to preserve as much usable data as possible; this solution is also completely indestructive. The following function is used:

pandas.DataFrame.abs()

Now we check for positive numeric feature values beyond their given constraints. The solution to tackle very large positive values is to limit them to their feature maximums. This only works because theoretically it is almost impossible that a car of price 0.1 exists, however a car of price 649.9 may exist because as our subject matter expert has said earlier, the price of a car may be arbitrarily influenced by outside factors.

Case in point, there exists a lower limit where no cars are sold under a certain price, or no cars are produced under a minimum power rating or engine size (in the name of practicality). The limit process is done conditionally, cell by cell, using:

pandas.DataFrame.iloc[]

Now we remove any rows which have values of 0, for the same reason mentioned directly above. 127 rows are lost with this method; this is done using:

pandas.DataFrame.loc[]

Error 4: Illegal categorical data entries

Now we check for values within the categorical features that are not within the list of allowed values. Output reveals model and manufacturer values with extra whitespaces, and misspellings of petrol/diesel as peatrol/diasel.

The extra whitespaces need to removed and the typos need to be fixed; these are done using:

pandas.DataFrame.str.replace()

Error 5: Incorrect values

Now we check for invalid summations in the total feature, more specifically we seek to identify the row indexes where Male + Female + Unknown != Total. There are four rows with incorrect summations for the gender features, these are corrected using:

pandas.DataFrame.iloc[]

Error 6: Handle outliers

Even after curbing numerical features to fit within their constraints, we still check for outliers; a value can be considered both legal and an outlier simultaneously. One method was to use an interquartile range to determine which rows to cull, but we found that this method removed too many legitimate values. Despite not using this method, the code has been left in for completion and comparison.

The method we ended up using was a 1-99 quartile based method that specifically targeted values closer to the extreme minimums and maximums, while preserving as many legitimate values as possible.

In order to define the term ‘legitimate value’, the following example is given: using the iqr method the resulting range of ‘cleaned’ price values is [3.3322, 78.74625]. As a subset of the defined legal values for price (0, 650.0], it is clear that there are too many ‘legitimate’ values within the 75 and above range that are being culled. In comparison to the 1-99 quartile method the resulting range is [6.895, 158.823], preserving a wider range of ‘legitimate’ values.

Interesting to note is that both methods are equally destructive in an objective sense, with 745 or 747 rows being destroyed depending on the method. Therefore, in order to decide which method to use, we must instead consider which method is better at preserving legitimate values, and distorts the range of values the least.

The output contrasts the efficacy of the two methods, specifically in the overall wider range of preserved values achieved by the 1-99 quartile method compared to interquartile ranges.

Data Exploration

Superscript numbers [123456789] correspond to their respective Figure values in the .ipynb.

Task 2.1

Of the top ten car brands, six of them (Escort, Golf, Focus, Mondeo, 320i, and 2000) are predominantly owned by men[1], while the other four (Fiestra, Clio, Micro, and Polo) are predominantly owned by women[1]. The unclassified genders have no majority ownership of any of the top ten car models[1].

Notably, for the car models where the ownership is women dominated, the number of women do not outnumber the men nearly as much, as compared to the car models predominantly owned by men[2]. That is to say, the objective difference is larger for majority male owned cars.

The Escort model has the largest discrepancy between the number of Male and Female owners, while the Fiesta model has the smallest discrepancy[2]. The 320i model has the largest discrepancy ratio between male and female owners, at about 3:1 visually[2]; there is a large enough statistical difference to consider that certain car models are more popular with one gender than another. Now we seek to understand which car attributes influence this observation.

The average power rating and engine size of the car model 320i is higher than the micra, while the 320i's ownership is male dominated as compared to the micra's female dominated ownership[3]. This shows that male drivers tend to value larger engine sizes and power ratings as compared to female drivers. The outstanding average price of the 320i also suggests that males tend to purchase more expensive cars than females[3].

Task 2.2

We can see a mostly linear relationship between the price and power rating of a car[4]. There are no particularly significant stand out values seen.

It is clear that any significant errors in the price or power features have been eliminated in the data cleaning process, therefore there is no further work to be done here in this regard.

There is a point to be made about the definition of an 'error', specifically in the context of a car's price or power rating. There are many variables which can contribute to the price of a car, as previously mentioned by our subject matter expert these can include, but are not limited to: brand value, nostalgia value, and scarcity value. Consideration of these factors beyond the scope of the data set then means that, if any error price were to show as a rogue data point in Figure 4, it may not be an error at all. One example our subject matter

expert uses is the Porsche Carerra, where the main contributing factor to its very high price is in its scarcity. The prevalence of its power, in terms of its price, is therefore weaker.

Task 2.3

The price distribution of male car owners appears to be normally distributed, and heavily right-skewed, with a large concentration of owners with prices between (0, 100)[5]. There are many large spikes in ownership count at certain prices, such as (visually) 15-20, 30-35, 45-50, 55-60, and 75-85[5], suggesting that males tend to purchase cars at certain price brackets.

The transmission distribution of male car owners appears to be normally distributed, and is mostly symmetrical around a value of 5, if not slightly left-skewed[6]. The symmetric distribution suggests that transmission has no meaningful correlation in terms of buying decisions.

The power distribution of male car owners appears to be normally distributed, and right skewed with a large concentration of owners owning cars with power measures between (50, 150)[7]. The shape of the power distribution is notably similar to the price distribution as seen in Figure 5; suggesting a measurable linear correlation between the power and price, in relation to the distribution of male owners.

The Engine CC distribution of male car owners appears to be normally distributed, and right skewed with a large concentration of owners owning cars between (1000, 2500) CC[8]. The shape of the engine size distribution is notably similar to both the price and power distributions as seen in Figures 5 and 7. This suggests a measurable correlation between the power, price, and in addition the engine size of a car model, in relation to the distribution of male owners.

The fuel type distribution of male car owners is heavily biassed towards the petrol fuel type[9]. According to our subject matter expert, the physical cost of cars that use the petrol fuel type are cheaper than the diesel fuel type; more male owners are purchasing relatively cheaper cars, as measured by fuel type. This is further corroborated by Figure 5, where there is an intense concentration of males purchasing petrol type cars with a price value between (10, 30), compared to the price range of diesel type cars between (30, 50)[5].

References

Subject matter expert: Spencer Buckley Morrison - s3906025