**Data Processing:**

### Dataset Summary:

* The dataset was imported from a csv file, containing 5330 values, with 26 columns and 205 records.The columns include:

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| Symboling | Integer | Rating that corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk. A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe. |
| Normalized-losses | String | Represents the average loss per car per year. |
| Make | String | The make of the vehicle |
| Fuel-type | String | The type of the fuel that the vehicle takes. |
| Aspiration | String | The more air that enters into the combustion chambers of an engine, the more fuel can be added - creating bigger explosions and generating more power. A vehicle can be either naturally aspirated, turbocharged, or supercharged.(Mat Foundry Group Ltd, N.D.) |
| Num-of-doors | String | The number of doors that a vehicle has. |
| Body-style | String | The body type that the vehicle shape is |
| Drive-wheels | String | Describes the driving system of the vehicle. A driving system is defined by where the power from the vehicle transmits this power according to the vehicle wheels.(Wuling, 2023) |
| Engine-location | String | Where the engine of the vehicle is located |
| Wheel-base | Float | The size of the wheel-base |
| Length | Float | The length of the vehicle |
| Width | Float | The width of the vehicle |
| Height | Float | The height of the vehicle |
| Curb-weight | Integer | Curb weight refers to the total weight of a vehicle, inclusive of standard equipment and necessary operating fluids such as motor oil, transmission oil and brake fluid, but without passengers or cargo. It’s essentially the weight of a ready-to-drive vehicle without any load. (Deana Beltsis, 2024) |
| Engine-type | String | The type of engine of the vehicle |
| Num-of-cylinders | String | The number of cylinders in the vehicle engine |
| Fuel-system | String | The fuel system of the vehicle |
| Bore | String | The depth of the cylinder that a piston travels up and down within the engine of the vehicle (Automative Training Centre, N.D.) |
| Stroke | String | The diameter of the cylinder that a piston travels up and down within the engine of the vehicle (Automative Training Centre, N.D.) |
| Compression-ratio | Float | The compression ratio is defined as the ratio between the volume of the cylinder with the piston in the bottom position, Vbottom (largest volume), and in the top position, Vtop (smallest volume). The higher this ratio, the greater will be the [power output](https://www.sciencedirect.com/topics/engineering/power-output) from a given engine.  (Gudmundsson, 2014) |
| Horsepower | String | The power that the engine produces |
| Peak-rpm | String | The power that is being transferred by the engine of the vehicle |
| City-mpg | Integer | The total number of miles a vehicle will travel on a gallon of fuel in the city |
| Highway-mpg | Integer | The total number of miles a vehicle will travel on a gallon of fuel on a highway |
| Price | String | The total value of the vehicle. |

### Data Cleaning:

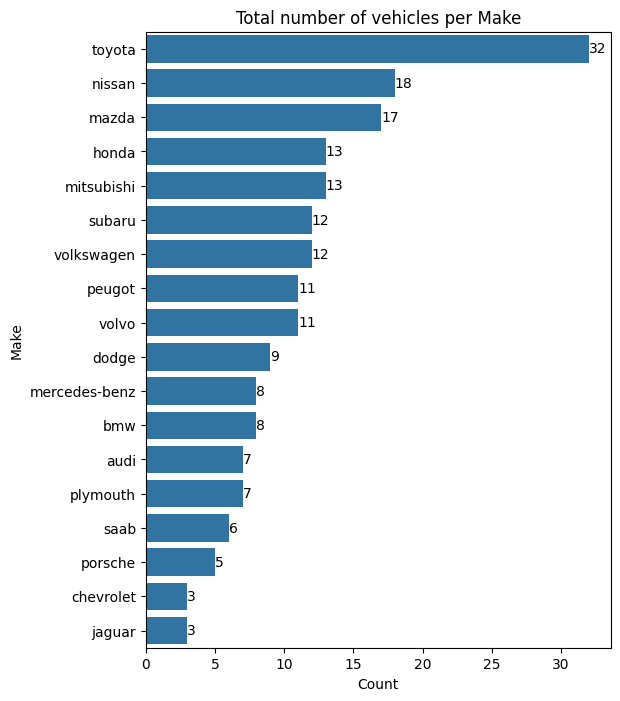
1. **Duplicated Values**:
   1. Identified the total number of duplicated values by chaining ***.duplicate()*** and ***.sum()*** to return the total number of duplicate values within the dataset
   2. There were no duplicate values within the dataset
2. **Null Values**:
   1. Identified the total number of null values by applying the numpy function ***np.count\_zero()*** and passed the dataframe with the method ***.isnull()*** chained with the ***.sum()*** method into the function.
   2. There were no null values within the dataset.
3. **Missing values**:
   1. It was found that some features with numeric values were being read as features with string values. This was achieved by applying the method ***.select\_dtypes()*** and passing the argument ‘number’ in the parameter *‘exclude’*.
   2. A user defined function, ***special\_characters()*** was used to identify the missing values being represented as ‘?’ in the following columns:  ***['normalized-losses', 'bore', 'stroke', 'horsepower', 'peak-rpm', 'price'].***
   3. Another user defined function, ***mean\_blank\_imputation()*** was used to impute the missing values with the mean values by the mean value of that column by that vehicle-make.
   4. There were values that could not be imputed because there were no values for that particular vehicle-make to be able to impute the mean values. In total, there were ten of these values.
   5. It was decided that the missing values can be dropped because the values only account for 5% of the total records within the dataset. The records were dropped using the index values of the records, using the ***.drop()***  method on the dataframe.
4. **Cast Features**:
   1. The method ***.astype()*** was applied on the dataframe, and a dictionary was passed as the argument, containing the name of the columns as keys, and the values as the data types which we intend on casting the columns to.
5. **Outliers**:
   1. The outliers were visualized using the seaborn library’s function, ***sns.boxplot()***
   2. The user defined function, ***outlier\_imputation().*** This function implemented the IQR to find the total number of outliers in each numeric column. This would then be outputted to find
   3. The same function was then responsible for imputing the outlier values, by either the mean, median, or upper and lower limits that were defined within the function.
   4. The distribution of the features were then visualized using the ***sns.histplot()*** to compare the distribution of the feature before the imputation, and after the imputation.

Data Transformation:

1. **Feature Engineering:**
   1. Use one hot encoding on categorical features with less than 5 values in the column. The user defined function ***add\_dummies()*** was used.

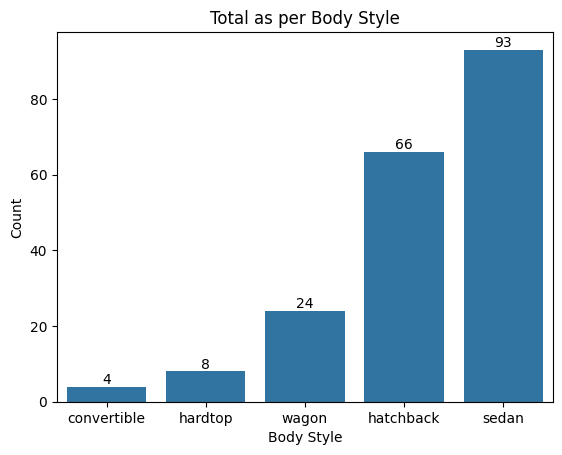
Data Visualization:

*Total Number of vehicles per Make:*



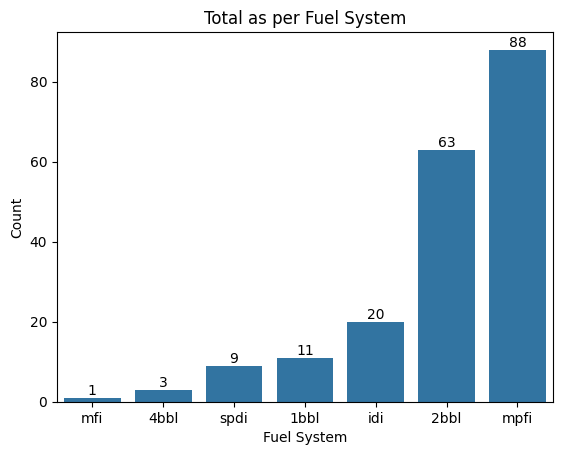
Based on the bar chart, we can see that within the dataset, we have Toyota having the most vehicles within the dataset. The top 5 of the total makes is completed by: Nissan, Mazda, Honda, and Mitsubishi in descending order.

*Total Number of vehicles per Make:*



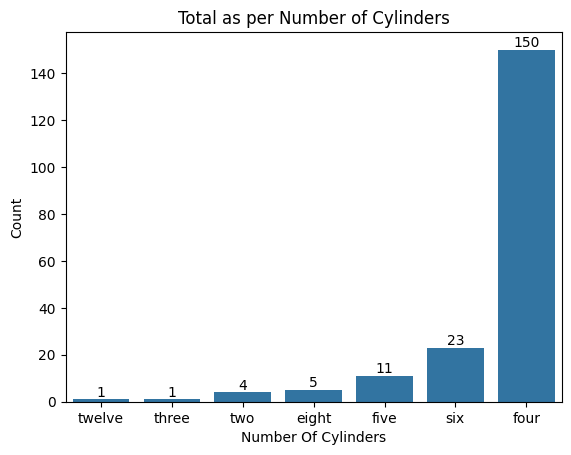
Based on the bar plot, we can see that the most popular body style is a sedan, followed by a hatchback, wagon body style, hardtop and a convertible in descending order.

*Total vehicles per fuel System:*



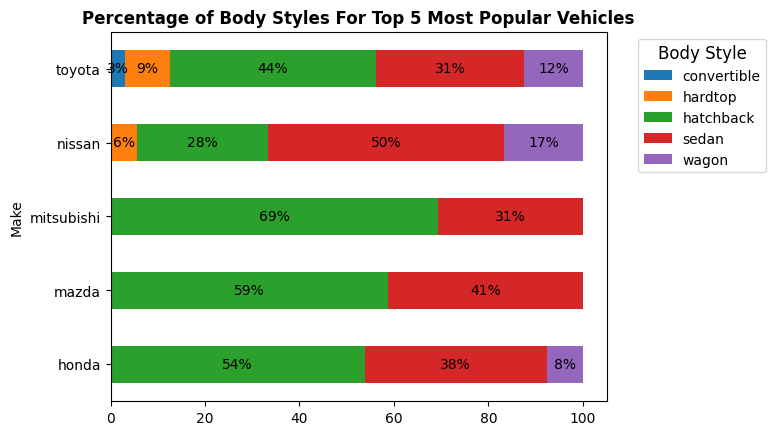
In descending order, we can see that the most frequent fuel system within the dataset is that of the mpfi fuel system. This is followed in descending order with the 2bbl, idi, 1bbl, spdi, 4bbl and finally the mfi fuel system.

*Total Number of cylinders within the dataset:*

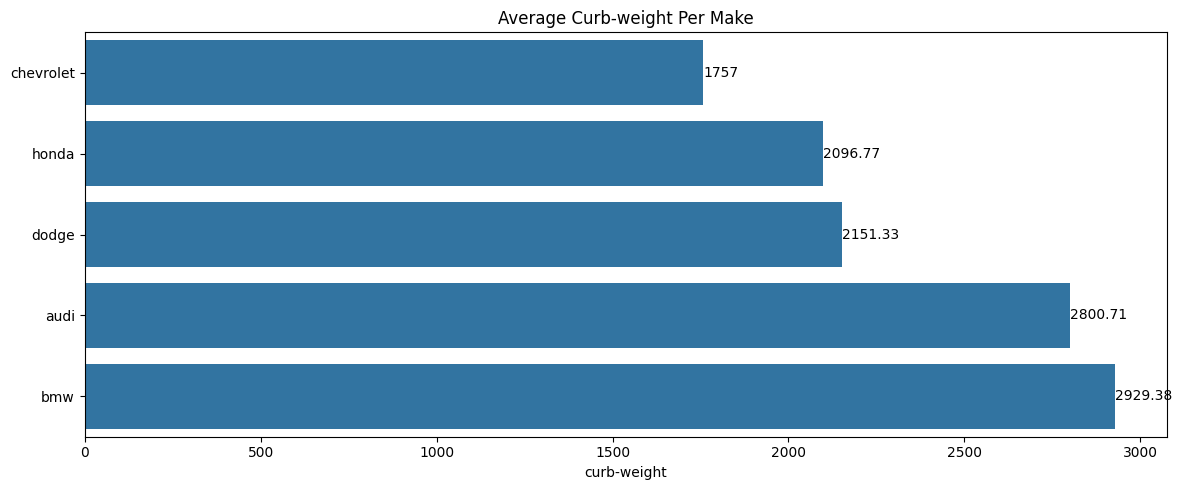
**

In ascending order, within the dataset we find that we have the 12 cylinder, then the three cylinder, the two cylinder, the eight cylinder, the five cylinder, the six cylinder, and most frequently, the four cylinder.

*Total Number of cylinders within the dataset:*

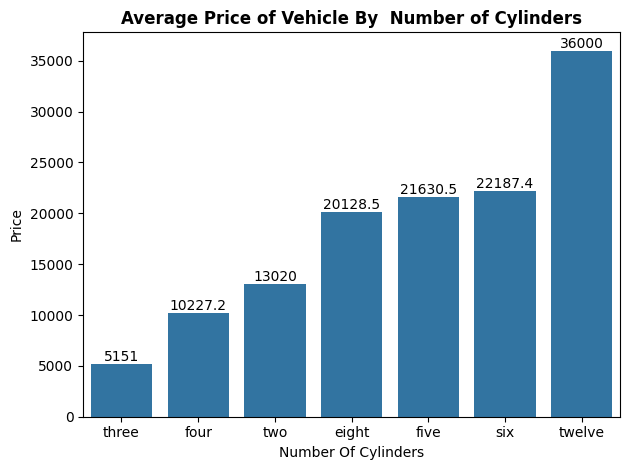
From the stacked bar chart, we can see that the most popular body style type for the top 5 vehicles-make within the dataset is the hatchback and the sedan, and the least popular being the convertible. 

*Average curb-weight per Make:*



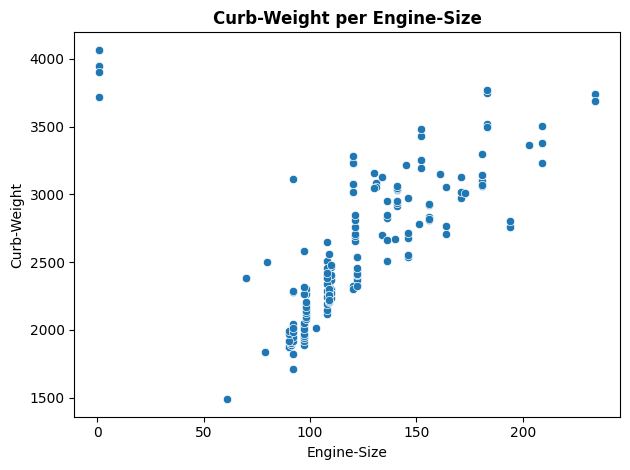
The largest curb-weight, based on the dataset includes, in ascending order: Chevrolet, Honda, Dogge, Audi and BMW.

*Average price per number of cylinders:*



In ascending average price of a vehicle based on the number of cylinders, we have the following in ascending order: three pistons, four pistons, two pistons, eight pistons, five pistons, six pistons and twelve pistons.

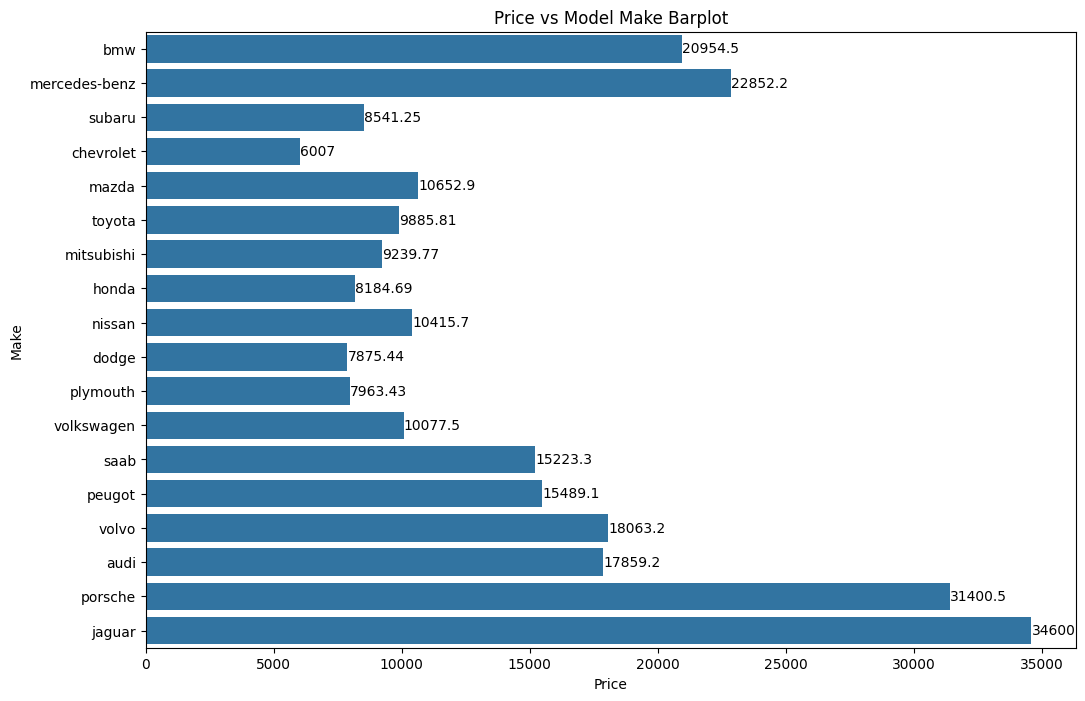
*Curb-weight per engine-size relationship:*



Through the scatterplot, we can note a linear relationship between the curb weight and the engine size of a vehicle within the dataset.

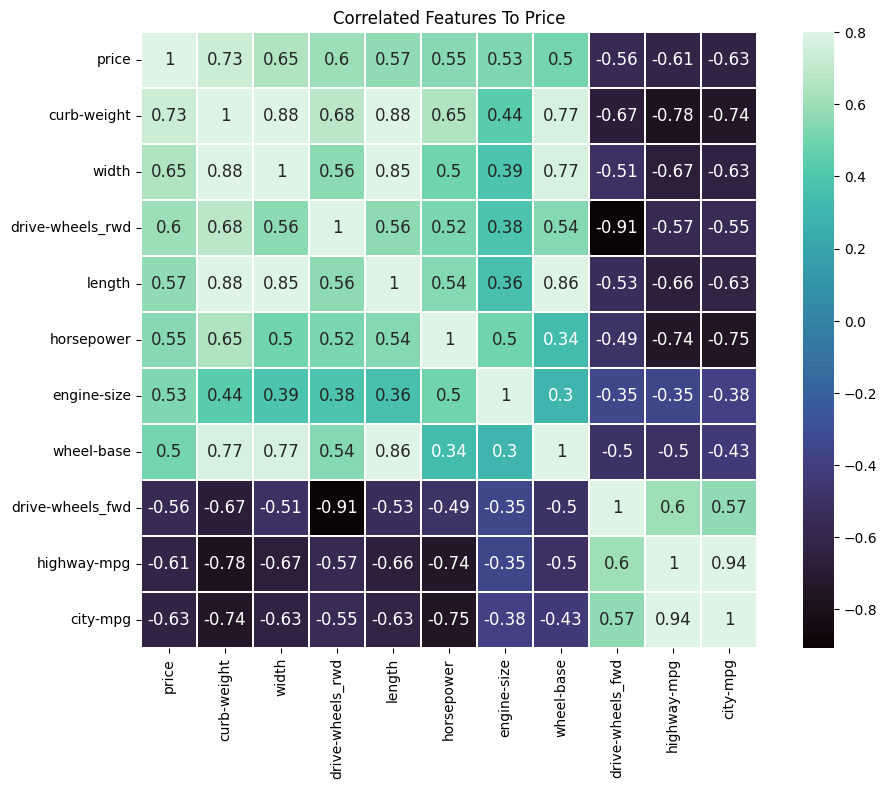
The strength of the correlation can be seen with how densely (or not) a gradient line. In this instance, the weak correlation str

*Average price per vehicle make:*



From the bar plot we can see that the jaguar has the highest value on average per model make.

*Average price per vehicle make:*

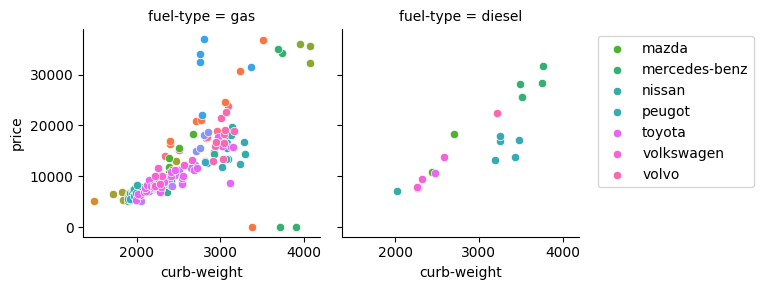
Based on the correlation matrix above, we can find that the most correlated feature to price is curb-weight with a pearson correlation coefficient of 0.73. 

*Distribution and relationships per feature:*

**

The paiplot from the above shows the relationship and the distribution per feature.

*Distribution and relationships per feature:*



Based on the correlation coefficient, it is evident to see that there is a moderately strong positive correlation between the price of a vehicle, and the curb-weight of the vehicle. In an article by Deana Beltsis, curb weight is referred to as the total weight of a vehicle, inclusive of standard equipment and necessary operating fluids such as motor oil, transmission oil and brake fluid, but without passengers or cargo. It’s essentially the weight of a ready-to-drive vehicle without any load. Deana Beltsis (2024).

We can also see that from the scatterplot looking at the relationship between price and curb-weight, that fuel-type brings variation in the data. This can be seen with the number of outliers where the points are the furthest from any linear gradient. These variations are less compared to when looking at the relationship between the two variables when the vehicle is that of a diesel vehicle.

*Hypothesis testing:*

Based on the null hypothesis, we incur that there is no linear relationship between the price of a vehicle and the vehicle's curb-weight.

The alternative hypothesis is that there is a linear relationship between the price of a vehicle and the vehicle curb-weight

The significance level we have decided is 5%

Based on the statistical test, using the t\_test, where we used the ***scipy.stats()*** library and the ***pearsonr()*** coefficient to determine the coefficient and compare it to the significance level. If the p-value of the t\_test results is less than 5%, then the null hypothesis can be rejected.

However, in the case between the price and the curb-weight features, we can see that the p-value is greater than 5% and therefore, we can reject the null hypothesis, and can accept the alternative hypothesis, which states that there is a linear relationship between the price of a vehicle and the vehicle curb-weight.

Bibliography:

1. Harika Bonthu (2024), \_Analytics Vidhya\_, retrieved 24/07/2024,

<https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/>

2. Eric Kleppen (2023), 'How To Find Outliers in Data Using Python (and How To Handle Them)', \_Career Foundry Blog\_, 11 May,

<https://careerfoundry.com/en/blog/data-analytics/how-to-find-outliers/>

3. Jules J. Berman (2016), \_Data Simplification Taming Information with Open Source Tools\_, [\_Online\_], Morgan Kaufmann, Available at:

<https://www.sciencedirect.com/topics/computer-science/pearson-correlation#:~:text=The%20P%20value%20is%20typically,coefficient%20is%20not%20statistically%20significant.>

4. Deana Beltsis (2024), 'Curb Weight vs Gross Weight : A Comprehensive Guide', \_Geotab Blog\_, 31 May, Available at:

<https://www.geotab.com/blog/curb-weight-vs-gross-weight/>

5. Jim Frost (N.D.), \_Statistics By Jim Making statistics intuitive\_, Available at:

<https://statisticsbyjim.com/glossary/significance-level/>

6. Avijeet Biswal (2024), 'What is Hypothesis Testing in Statistics? Types and Examples', \_Simplilearn Tutorials\_, 9 July, Available at:

<https://www.simplilearn.com/tutorials/statistics-tutorial/hypothesis-testing-in-statistics#null\_hypothesis\_and\_alternate\_hypothesis>

7. Rebecca Bevans (2020), 'An Introduction to t Tests | Definitions, Formula and Examples', \_Scribbr\_, Available at:

<https://www.scribbr.com/statistics/t-test/>

(Mat Foundry Group Ltd, N.D.)

(Wuling, 2023)

<<https://wuling.id/en/blog/autotips/fwd-vs-rwd-drive-system-whats-the-difference#:~:text=FWD%20vehicles%20are%20driven%20through,are%20more%20balanced%20than%20FWD>.>

(Automative Training Centre, N.D.)

<[https://www.autotrainingcentre.com/blog/strokes-bores-explained-auto-parts-specialist-t](https://www.autotrainingcentre.com/blog/strokes-bores-explained-auto-parts-specialist-training/)

[<](https://www.autotrainingcentre.com/blog/strokes-bores-explained-auto-parts-specialist-training/)<https://www.matfoundrygroup.com/blog/Turbos_Superchargers_and_Naturally_Aspirated_Engines#:~:text=Naturally%20Aspirated%20or%20naturally%20'breathing,explosions%20and%20generating%20more%20power>[.>raining/](https://www.autotrainingcentre.com/blog/strokes-bores-explained-auto-parts-specialist-training/)>

(Snorri Gudmundsson BScAE, MScAE, FAA DER(ret.), 2014)

<<https://www.sciencedirect.com/topics/engineering/compression-ratio#:~:text=The%20compression%20ratio%20is%20defined,output%20from%20a%20given%20engine>.>