Task 1 Data Acquisition and Preparation

The datafiles are acquired from the Swinburne canvas and all the files are adapted form Automobile data set. The data contains all the records of the cars with specification of attributes and observed that it has got numerical data and objects, where I can easily analyze the data by going through the attributes, that it shows the different data and statistics for each car. It gave an idea that each car got it positives and negatives with different ranges. In this document, Analysis is done by doing comparison between the attributes with different methods.

1.1 Merging the data

First the path is created for the three data files and later it was loaded into the panda's data frames. Next the data1 and data2 are merged by keeping id as unique. later the merged file is done concat with data 3 file. As data 1 file has few attributes and data2 has remaining attributes of the same set, first the merging is done between them and data 3 has continuation of merged data file with all 27 attributes, so it just done concat.

1.2 Data cleaning

Methods and techniques used

1. **Removing duplicates:** Firstly, encountered the mistake by finding the duplicates all at once it shows only one duplicate. Later it takes 2 duplicates if "id" Is taken as a preference, it shows there are two ids with same number and got a problem that 16th column as two distinct names for same value. To encounter this problem firstly identifies the unique values by unique () in that column and replaced with their original names. After replacing the names, the duplicated rows are dropped by drop_duplicates () and data is shaped from 199 rows to 197 rows.

2. **Finding Null values:** To find the null values, it is needed to sum up all the null values in the dataset by using a method Isnull. After summing up the count it shows that missing elements are less than 50%. so, the null values are not possible to remove.

3. **Replacing Null values:** As the normalised-losses and price got few null values. The mean value is calculated for both the columns and it is replaced in the Null places.

4. **Convert number stored as text into numbers:** In the data num of cylinders and num of doors are integers but the numbers are stored into text format. So firstly, replaced the text with numbers and encountered a problem that few of the text were not changing and the problem is resolved by converting the data type from object to integer. By converting it is easy to explore the data and visualize.

5. **Changing text to proper case:** This technique is very important as it used to check the casing in the dataset. So that it will not encounter any problems in the future. In the present data set while checking the unique values by unique () came up with a wrong casing in the 18th column. So, words are replaced by replace () with the correct wordings. Finally, upper case is changed into lower case.

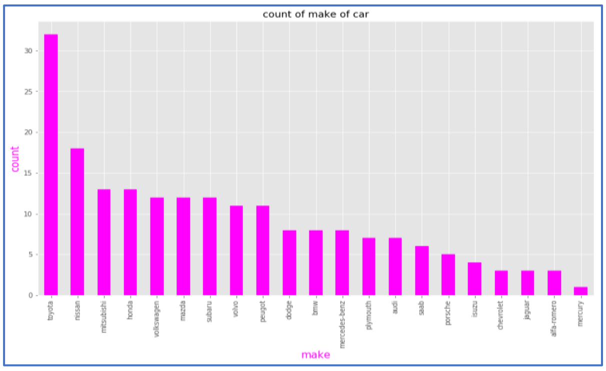
6. **Getting rid of extra spaces**: checking Scaping is very important in data cleaning it might encounter problem while visualizing. In this dataset in the 4th column fount an extra spacing in front of the text and the space is removed and replaced with correct wording.

7. **Converting the decimal points:** It is noticed that price got a datatype of float and the decimal values are too big to make it short and flexible to calculate. All the values of the price are rounded to 2 decimal points.

Task 2 Data exploration

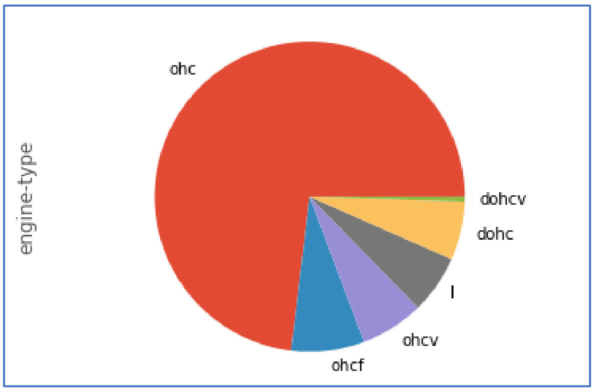
2.1 Univariant analysis

Highest car manufacturer



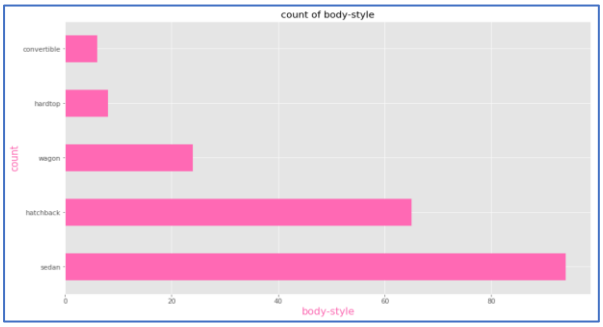
The above graph illustrates the make of the car, this graph is taken because it easy to compare and show which group is the highest and more common and can know which make of car produces the highest number of cars. It is known that Toyota has large amount of car making more then other brands. From this graph we can analyze that how Toyota is more car and risks involved behind.

Highest frequency of engine types used in the car



The above figure is pie chart as the engine-types are less it is easy to show in pie chart and know that which engine type is mostly used in the car. By taking this column it can be predicted that which engine-type has good curb-weight and price. It is good to analyze the data and make predictions based on the engine-type. It is shows that ohc has frequency of engine type used in the car.

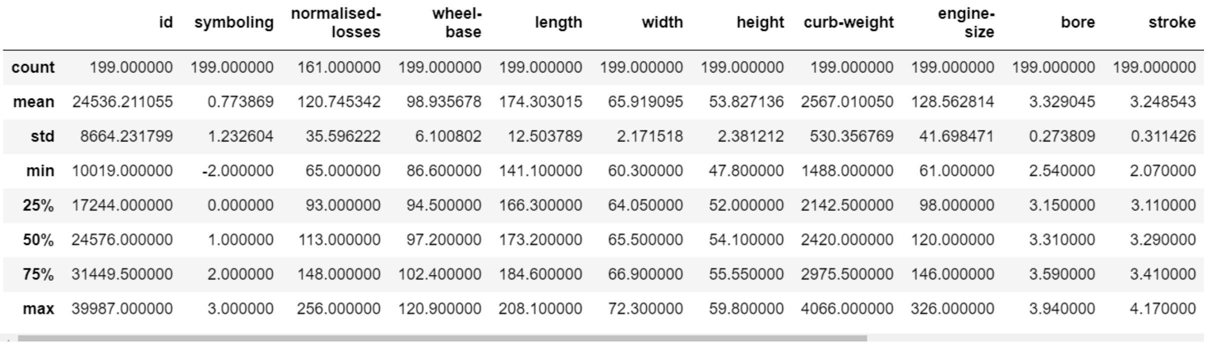
Types of body styles used for the cars



The above figure is the horizontal bar graph, it is taken as it is used to know which group has highest or lowest usage. By taking this, it can be predicted that which body-style has more usage and why it is preferred more that other styles. It is shown that sedan has highest amount of usage.

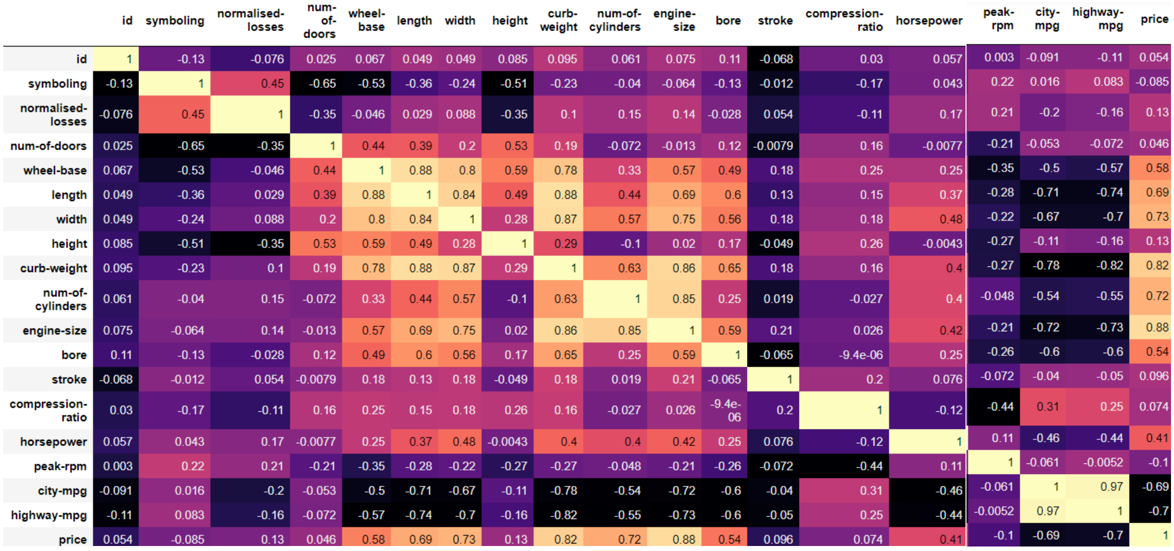
2.2 Bivariant analysis

Descriptive statistics (can be referred full in python code)



	id	symboling	normalised-losses	wheel-base	length	width	height	curb-weight	engine-size	bore	stroke
count	199.000000	199.000000	161.000000	199.000000	199.000000	199.000000	199.000000	199.000000	199.000000	199.000000	199.000000
mean	24536.211055	0.773869	120.745342	98.935678	174.303015	65.919095	53.827136	2567.010050	128.562814	3.329045	3.248543
std	8664.231799	1.232604	35.596222	6.100802	12.503789	2.171518	2.381212	530.356769	41.698471	0.273809	0.311426
min	10019.000000	-2.000000	65.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000	2.540000	2.070000
25%	17244.000000	0.000000	93.000000	94.500000	166.300000	64.050000	52.000000	2142.500000	98.000000	3.150000	3.110000
50%	24576.000000	1.000000	113.000000	97.200000	173.200000	65.500000	54.100000	2420.000000	120.000000	3.310000	3.290000
75%	31449.500000	2.000000	148.000000	102.400000	184.600000	66.900000	55.550000	2975.500000	146.000000	3.590000	3.410000
max	39987.000000	3.000000	256.000000	120.900000	208.100000	72.300000	59.800000	4066.000000	326.000000	3.940000	4.170000

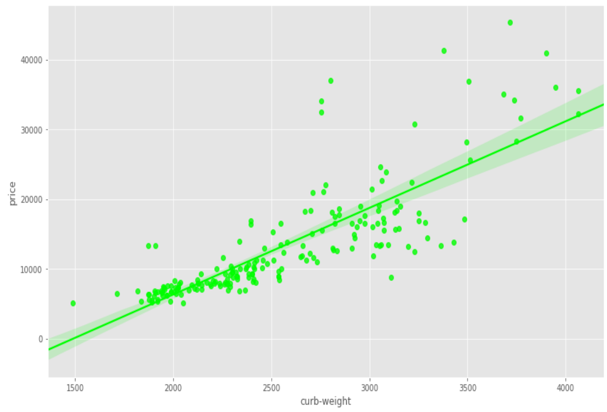
Co-relation matrix

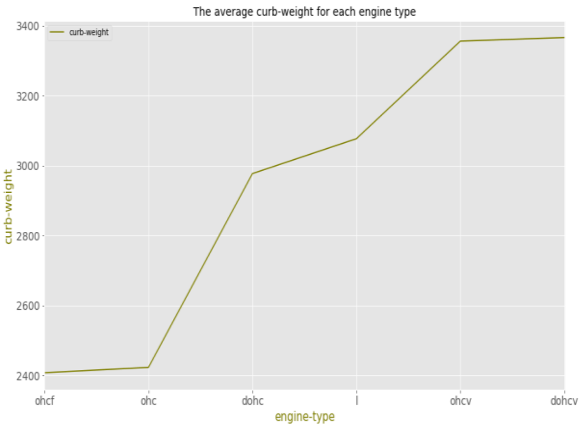


	id	symboling	normalised-losses	num-of-doors	wheel-base	length	width	height	curb-weight	num-of-cylinders	engine-size	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price
id	1	-0.13	-0.076	0.025	0.067	0.049	0.049	0.085	0.095	0.061	0.075	0.11	-0.068	0.03	0.057	0.003	-0.091	-0.11	0.054
symboling	-0.13	1	0.45	-0.65	-0.53	-0.36	-0.24	-0.51	-0.23	-0.04	-0.064	-0.13	-0.012	-0.17	0.043	0.22	0.016	0.083	-0.085
normalised-losses	-0.076	0.45	1	-0.35	-0.046	0.029	0.088	-0.35	0.1	0.15	0.14	-0.028	0.054	-0.11	0.17	0.21	-0.2	-0.16	0.13
num-of-doors	0.025	-0.65	-0.35	1	0.44	0.39	0.2	0.53	0.19	-0.072	-0.013	0.12	-0.0079	0.16	-0.0077	-0.21	-0.053	-0.072	0.046
wheel-base	0.067	-0.53	-0.046	0.44	1	0.88	0.8	0.59	0.78	0.33	0.57	0.49	0.18	0.25	0.25	-0.35	-0.5	-0.57	0.58
length	0.049	-0.36	0.029	0.39	0.88	1	0.84	0.49	0.88	0.44	0.69	0.6	0.13	0.15	0.37	-0.28	-0.71	-0.74	0.69
width	0.049	-0.24	0.088	0.2	0.8	0.84	1	0.28	0.87	0.57	0.75	0.56	0.18	0.18	0.48	-0.22	-0.67	-0.7	0.73
height	0.085	-0.51	-0.35	0.53	0.59	0.49	0.28	1	0.29	-0.1	0.02	0.17	-0.049	0.26	-0.0043	-0.27	-0.11	-0.16	0.13
curb-weight	0.095	-0.23	0.1	0.19	0.78	0.88	0.87	0.29	1	0.63	0.86	0.65	0.18	0.16	0.4	-0.27	-0.78	-0.82	0.82
num-of-cylinders	0.061	-0.04	0.15	-0.072	0.33	0.44	0.57	-0.1	0.63	1	0.85	0.25	0.019	-0.027	0.4	-0.048	-0.54	-0.55	0.72
engine-size	0.075	-0.064	0.14	-0.013	0.57	0.69	0.75	0.02	0.86	0.85	1	0.59	0.21	0.026	0.42	-0.21	-0.72	-0.73	0.88
bore	0.11	-0.13	-0.028	0.12	0.49	0.6	0.56	0.17	0.65	0.25	0.59	1	-0.065	-9.4e-06	0.25	-0.26	-0.6	-0.6	0.54
stroke	-0.068	-0.012	0.054	-0.0079	0.18	0.13	0.18	-0.049	0.18	0.019	0.21	-0.065	1	0.2	0.076	-0.072	-0.04	-0.05	0.096
compression-ratio	0.03	-0.17	-0.11	0.16	0.25	0.15	0.18	0.26	0.16	-0.027	0.026	-9.4e-06	0.2	1	-0.12	-0.44	0.31	0.25	0.074
horsepower	0.057	0.043	0.17	-0.0077	0.25	0.37	0.48	-0.0043	0.4	0.4	0.42	0.25	0.076	-0.12	1	0.11	-0.46	-0.44	0.41
peak-rpm	0.003	0.22	0.21	-0.21	-0.35	-0.28	-0.22	-0.27	-0.27	-0.048	-0.21	-0.26	-0.072	-0.44	0.11	1	-0.061	-0.0052	-0.1
city-mpg	-0.091	0.016	-0.2	-0.053	-0.5	-0.71	-0.67	-0.11	-0.78	-0.54	-0.72	-0.6	-0.04	0.31	-0.46	-0.061	1	0.97	-0.69
highway-mpg	-0.11	0.083	-0.16	-0.072	-0.57	-0.74	-0.7	-0.16	-0.82	-0.55	-0.73	-0.6	-0.05	0.25	-0.44	-0.0052	0.97	1	-0.7
price	0.054	-0.085	0.13	0.046	0.58	0.69	0.73	0.13	0.82	0.72	0.88	0.54	0.096	0.074	0.41	-0.1	-0.69	-0.7	1

This is done to check the co-relation between each column

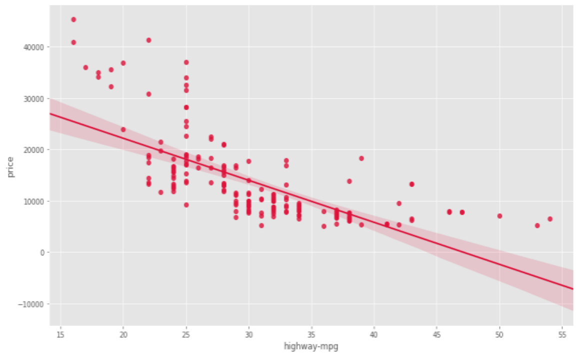
Finding 1:

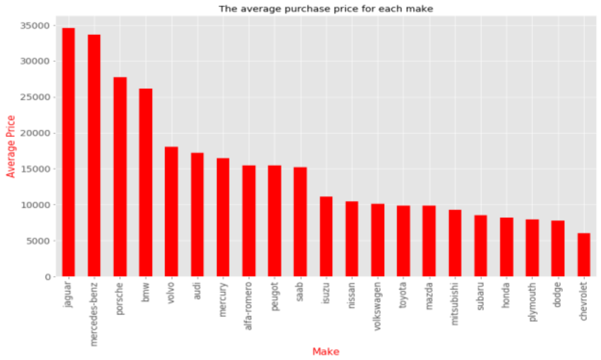




From the co-relation matrix it can be known that which columns are positively co-related and regression plot is made because the co-relation between them is near to 1 and it is positively co-related. It is shown that as curb-weight increases the price also increases. As the dohcv has more curb-weight it is also can be said that it has got more price. As it is known that ohc has more usage of engine-type because it has predicted that it got less curb-weight and less price.

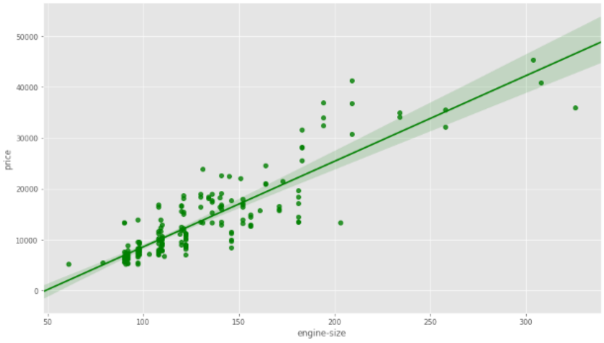
Finding 2:

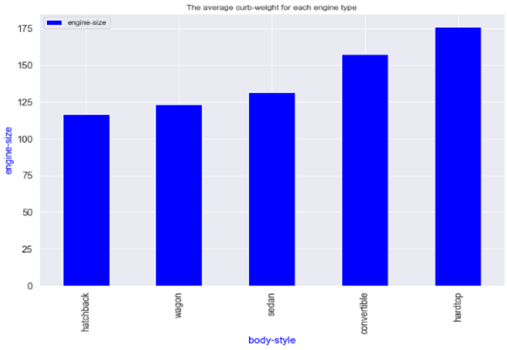




The regression graph shows that mileage and price are negatively co-related as the co-relation between them is near to -1. It is shown that if the price is more it gives less mileage and if it compared with the make of the car as jaguar has more price it can be said as it has got less mileage. As Chevrolet has less pricing it can be said that it has more mileage.

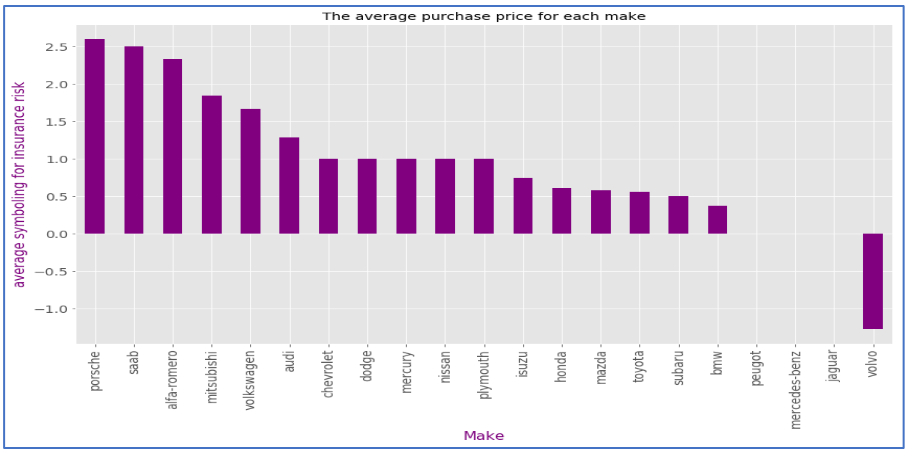
Finding 3:





The regression graph that price and engine-size are positively co-related. It says that if the engine-size increases the price also increases. After finding the average engine size for the body it is known that hardtop has more engine size. By comparing both the graphs it can be said as the body style with more engine-size has also got more pricing. As a greater number of cars use Sweden because it has got an average size of engine and price is reasonable compared to others.

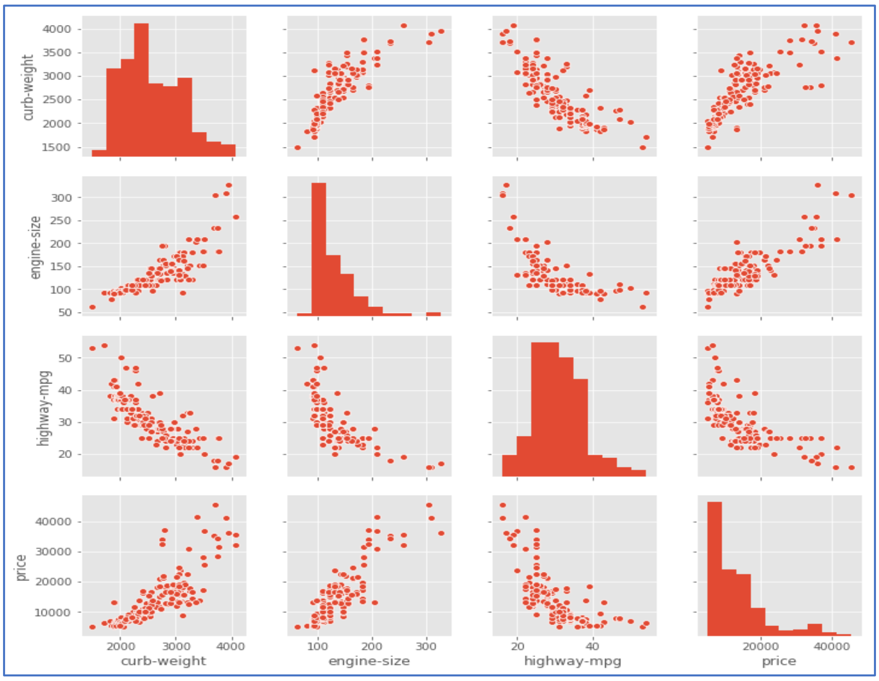
Finding 4:



The above figure depicts the bar graph, which shows the insurance risk for each make it shows that values near to -3 is less risky and value near to +3 is risky. Volvo is less risky make of car compared to other cars, but make of cars are more for Toyota as it has average risk and got reasonable pricing .

2.3 scatter matrix

The scatter matrix with all the numerical columns as the data is huge my findings are done in a pair plot for certain number of columns which are important for my findings. The bigger matrix can be referred in the python code



My findings are based mainly on mileage, engine-size, curb-weight and price. From the above plot some finding are made by comparing with each other.

- It is known that as the curb-weight and engine-size increases price increase, if the highway-mpg decreases the price will be decreasing.

- If the curb-weight and engine-size Increases the mileage is decreasing and the price increases the mileage is decreasing.
- Curb weight and price increases the engine size also increases, but if the mileage increases engine size will decrease
- Engine size and price increases the curb-weight will raise up, but if the mileage more the curb-weight will be less

The fuel-type and make of car can be selected based on these findings, for suppose if the car has good pricing it will getting good engine size and curb weight, but the mileage might be less.

References

Anon., 2015. *Datascience made simple.* [Online]
Available at: http://www.datasciencemadesimple.com/append-concatenate-rows-python-pandas-row-bind/
[Accessed 30 april 2020].

Anon., 2019. *Medium.* [Online]
Available at: https://towardsdatascience.com/visualizing-data-with-pair-plots-in-python-f228cf529166
[Accessed 28th April 2020].

Anon., 2019. *W3schools.* [Online]
Available at: https://www.w3schools.com/python/python_ml_multiple_regression.asp
[Accessed 28th April 2020].

Anon., 2020. *ROZF.* [Online]
Available at: http://queirozf.com/entries/pandas-dataframe-plot-examples-with-matplotlib-pyplot
[Accessed 29th April 2020].

Begg, T. C. a., 2005. Database Systems: A Practical Approach to Design, Implementation, and Management, 6th Edition. In: *pearson education.* s.l.:s.n.