**DATA ANALYSIS AND VISUALIZATION ON PRE-OWNED CARS**

Aadil Mohammed  
 Computer Science  
 University of Missouri - Kansas City  
 Kansas City Missouri USA  
 amfwc@umsystem.edu

Jahnavi Kolla  
 Computer Science  
 University of Missouri - Kansas City  
 Kansas City Missouri USA  
 jk55h@umsystem.edu

Niharika Guntupalli  
 Computer Science  
 University of Missouri - Kansas City  
 Kansas City Missouri USA  
 ngtqb@umsystem.edu

Pavan Kumar Reddy Palanki Priyanka Munnangi Sashi Kanth Pepalla

Computer Science Computer Science Computer Science

University of Missouri – Kansas City University of Missouri – Kansas City University of Missouri – Kansas City

Kansas City Missouri USA Kansas City Missouri USA Kansas City Missouri USA

[ppcq4@umsystem.edu](mailto:ppcq4@umsystem.edu) [pmd9r@umsystem.edu](mailto:pmd9r@umsystem.edu) spdtc@umsystem.edu

ABSTRACT

Now-a-days most of us prefer to buy used cars as they are usually low cost. Everyone wants good quality cars and to have a good lifespan, we all know that there are some factors that help a car be judged and estimated of how it is. Like for example the mileage of a car (number of miles driven), production year, engine capacity, fuel consumption, number of cylinders, price the owner setup are all some key factors to decide if the car is a good fit to buy or not. If we could approximately estimate the overall status of the car based on these factors and generate graphs on which type of car is best in terms of which feature based on the historical data, then those graphs must compare the differences between the features accordingly to decide from the customer perspective.

To achieve this, we took a dataset from ‘Kaggle.com’ which has 4 lakh of records that has data from past years about the entire story of each car that was sold. This data set is used to compare the features and bring out concepts based on the features and their comparisons in between them.

**INTRODUCTION**

Anything in this world can be predicted or at least assessed based on the data available and the way we draw concepts based on the data. In this paper, we tried to conceptualize and visualize the previous data of used cars and we took a data set of size 1.5 Gigabytes from the very popular website called Kaggle.com. This data help in giving the overall picture of how the condition of the car is. With the python notebook we can generate the visualized graphs to make it easy for the users to see, differentiate and compare among the different models. What if the data is not accurate? There are chances that the data is an influence or has the hand of someone to manipulate it. So, the data always must be confidential but because this is a basic paper about how the visualization trends go, we followed a website to have a dataset. Usually, the data from any firm is confidential and it must be. Because anything can happen if the data is open, for example let us consider an example of our dataset and one brand of cars have very high pricing even if the condition of the car is not up to the mark or let us say the engine capacity is lesser comparatively. If the data is in the hands of this brand and there is always a chance to manipulate the data. Because after all it is just a set of numbers and rows which will define the overall brand value and the car performance with efficiency in multiple ways. So, the underlying point is to have the data secured, real time data is always gold, and it is helpful forever. Based on the concepts we draw out from the data we can predict and come to conclusions, so it is always a catch to handle it securely. All the data we got is not complete, there is always duplication, redundancy and many factors that make the data incomplete and it must be cleaned. This cleaning is a process to follow as we have huge number of rows to handle. We fill up the empty rows with some zeros or unknown values, we removed duplicates and made sure the data is cleaned. This process took long as there are 25 columns, and each column is supposed to be done separately. The pandas library helps in visualizing the content and brining everything at one place to make the entire data visually clear is the task and the representation is worth the whole procedure as we could see the trends that are running all over. We tried to generate concepts and visualize them in a timely manner and thus the procedure is discussed in the rest of the paper.

**DATASET DETAILS**

The dataset has an enormous number of records which is collected over the past years. There are a few websites which provide data for public use like Kaggle.com. We used one dataset for our project named vehicles.csv. In this dataset All the details about the data are in 26 columns. There are around four lakh and twenty thousand rows in this dataset. The columns in the dataset are URL, id, region, region URL, price, year, manufacturer, model, condition, Cylinder, fuel, Odometer, title status, transmission, VIN, drive, size, type, paint color, image URL, description, County, State, latitude, longitude, posting date.

Each column in the dataset adds a value to the dataset and provides a lot of information about the vehicle. We have constrained using only 18 columns for the project. We are taking consideration of below columns: Region, it tells us from which region the listing is made. E.g., Hudson Valley, Elpaso etc. Price, it tells us the price of the product. The value in this column will be dollars as its unit. Year, in which year was the vehicle manufactured. It just displays the year in which is manufactured. Manufacturer, it tells us the company which manufactured this product. Values in this column can be Chevrolet, Toyota, ford etc. Model columns tells us the model of the vehicle. Condition columns tells us the condition of the vehicle. As it is a used car values for condition of the vehicle can be good, Excellent etc. Cylinder. The values in this column can be a number like 6,4 etc. Odometer column represents the miles the vehicle has traveled. The values of this column will be the miles usage. Title status is one of the important columns where it displays the title status of the vehicle. Vehicle titles are certified for legal ownership of a vehicle. Values in this can be clean it has not met with any accidents. Salvaged/rebuild value represents it met with an accident and it has few legal complications. Transmission column tells about the gear box if it is an automatic, manual or any other kind. It contains values like manual, automatic, others if not any. The drive column contains Eg: AWD, FWD, RWD, etc. Size is one of the important columns where it tells us whether the vehicle is full-size, mid-size, compact. Based on this, a person can categorize which vehicle they needs. Type column tells us more about the vehicle if it is truck, sedan, SUV, pickup, convertible, hatchback. Latitude and Longitude columns tells us the location details when the listing is made. Posting date shows the date on which the listing is made. State column shows the details of the state. For example, if the state is Kansas, we display only the short form like KS.

Here is the data set link -

<https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data?resource=download>

Table

Description automatically generated

**USAGE**

This project will be helpful for those who are interested in buying used cars. It will be useful for analyzing different cars based on the factors like price, model, company etc. We have different representations based on the different parameters. Each representation depicts all usage and high-level visualization. For example, I am new to a place where I don’t know what vehicles to buy and which is efficient, cost-effective and easy to commute. I refer different websites of each car company to get to know more about it. Now what I can do is just use these representations analyze based on our preferences. For say, a person budget is 6k for buying a car he would get to see what he can afford under 6k and he chooses the best among different car companies. This is the advantage of this project. There is no necessity for him to go to each company website and filter under the limit and select among them. This is not only used for who are interested in buying preowned cars. It is useful for those who wants to buy a car as well. They will know the market trend as well for the cars they are interested in buying. Prices are fluctuating now a days based on some factors. In these times, to get a better understanding of market is also important. For some customers brand is important for buying a car and for few people miles covered is important. They want the vehicle which has covered less miles. We are making multiple representations of them. Mostly this will be benefit for the trades who trade the cars. They get a better picture of the vehicles based on the features. Using this project, both buyers and sellers gets a lot of benefits. Easy to analyze and understand the vehicle conditions and details is easily represented. If needed, we can add few more columns from the actual dataset and represent the data in data visualization.

**USED TOOLS**

We used python for coding, Open Datasets to download datasets. Open Datasets is a curated collection of datasets for data analysis & machine learning. Then we installed NumPy library to perform a wide variety of mathematical operations on arrays. NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. We installed pandas library to analyze data. Pandas is a software library written for the Python programming language for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series. In order to create static, animated and interactive visualizations we used Matplotlib library. Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tintern, wxPython, Qt, or GTK.

**DATA CLEANING**

For any data in a dataset that is huge, must be introspected to verify and make sure the data is cleaned. Data Cleaning is a procedure that needs to be handled to ensure clean data from the dataset. that the redundant data is handled, inconsistent data is taken care of and missing values, unnecessary data is removed from the dataset results in clean data and complete data with which we can generate patterns and visualize them for a better understanding and experience. For example, we have incomplete data in some columns like mileage, price, odometer, and many other columns. So, we filled the incomplete columns with values like ‘unknown’, ‘na’.

A clear picture of the data processed and handled using the complete dataset based on the column values are compared and conceptualized. This is an example of a column that has NaN values for Model, Condition, Cylinders, Fuel, Title Status, Size, Drive.

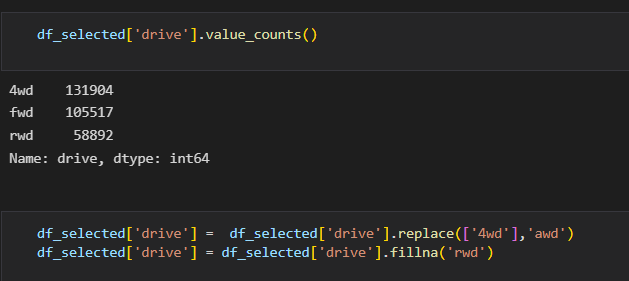
Below are a few figures of examples on Data Cleaning.

Text

Description automatically generated

**Fig 1. Data cleaning of different columns.**





A screenshot of a computer screen

Description automatically generated with medium confidence

**DATA PREPROCESSING**

In Data Preprocessing, the open datasets are downloaded from the library. Once the dataset is obtained, we read it using the pandas library and study the data frame. We also chose the number of columns that need to be included in this project and dropped the ones that are not going to be used.

All the out-of-range values or any compatible issues are clarified here in this process. They handle the dataset as a whole and do the initial procedures to clean the dataset.

**Graphical user interface, text, calendar

Description automatically generated** **Chart

Description automatically generated**

**Fig 2**.

**DATA VISUALIZATION**

Words don't always convey an idea clearly. Raw data may not necessarily provide the most interesting narrative. Visual information is highly well-received by the human mind that’s why Data visualization is a potent tool for communication. But don't worry if "data visualization" seems difficult and complex, it needn't be. This manual will provide an accessible explanation of the principles of data visualization. Numerous examples of various data visualizations are provided, along with information on when to use each one for reports, presentations, marketing, and other purposes.

**RESULTS**

**DATA VISUALIZATION ON USED CARS**

This paper gives all the information and representation of comparisons between different features in a car and all the required graphical representation is clearly depicted in the below figures. Each graph has a concept that is delivered to show case all the aspects of a car are mentioned, and some graphs could also be included for further more analysis and future works. Follow the rest of the paper with all the graphs we generated.

**AVERAGE CAR PRICE BY BRAND-DATA VISUALIZATION**

The two factors proposed in this graph are Price and Brand. For price we took the mean of each price of a car for a particular brand, and the brand remains the same. The Price shows from 0-500000 dollars and shows nearly 42 brands in the graph. This graph illustrates the Mercedes Benz and Volvo having high price and Mercury having low price. Most of the car brands are between 10k to 50k.

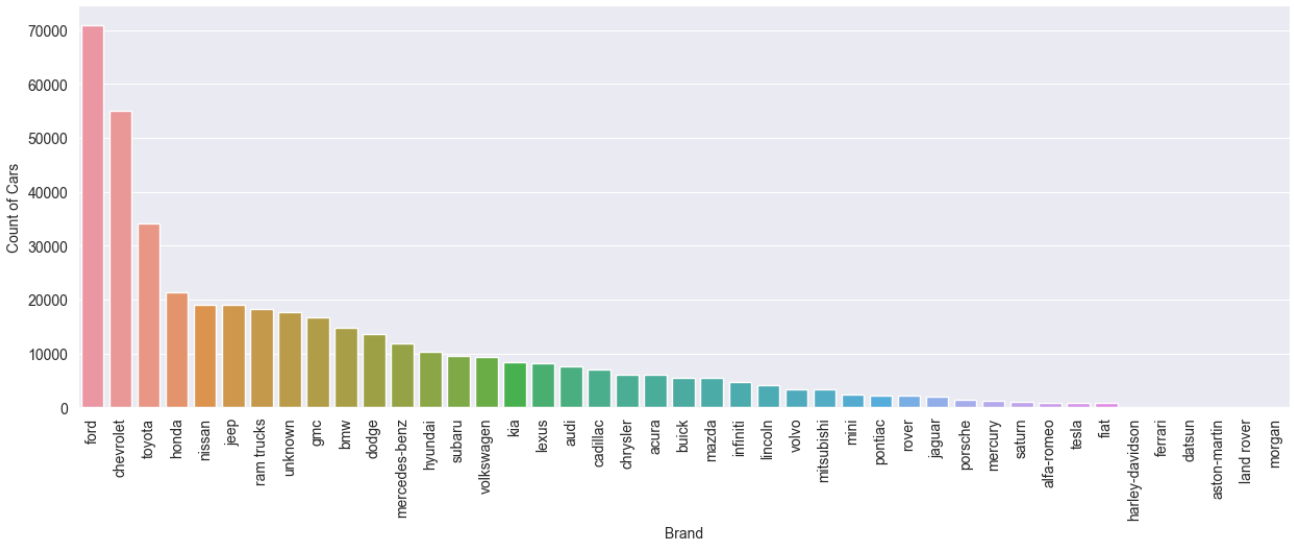
Chart, bar chart, histogram

Description automatically generated

**Fig 3. Price ($) Vs Brand**

**ANALYZE THE TOPMOST MANUFACTURER OF THE CAR BRANDS SOLD**

For any company, there should be a track of how many cars are sold overall by their customers. Also, the one who wants to buy a car will have an idea about the range of people buying a particular brand. To analyze this sort of concept we came up with comparing the Number of Cars Sold to the Brand of the Car. The below graph depicts this idea, and it clearly says that ‘Ford’ is the brand of car that is most bought by customers. The reasons could be the price availability, condition, life span and many others. Whereas ‘Chevrolet’ is the other brand people prefer and it is competing with Ford.



**Fig 4. Car Count Vs Brand**

**FORD CARS CHANGE IN PRICE OVER THE YEARS FROM 1990 TO 2022**

As the most sold car is FORD since so many years, we analyzed the data of the average price it got sold in each year from 1990 to 2022. The graph depicts the situation about how the hike in price changed in the pandemic time and ho wit got decreased after 2020.

**Chart, bar chart, histogram

Description automatically generated**

**Fig 5. Price ($) Vs Year Manufactured (Ford)**

**TESLA CARS CHANGE IN PRICE OVER THE YEARS FROM 2000 TO 2021**

Tesla is a completely electric car that was established in 2010. Its brand value is high, and its price is always a standard and how its price changed since it got released in the market is depicted in this graph. The price decreased drastically after the pandemic based on the dataset. Since it is electric as we saw in the other graph, electric cars are usually costly compared to the gas type of Fuel cars.

**Chart

Description automatically generated**

**Fig 6. Price ($) Vs Year (Tesla)**

**SALES TREND IN THE USA**

For any product that is available has a certain value in a particular region and with time, region, area, and weather conditions the number of people buying cars changes. Mostly the area of the state and the population living effects a lot. So here is a graph that gives us a picture of which state has the highest car sales and using this data car manufacturers can concentrate on the advertising and manufacturing of cars in those states.

Based on the dataset, California is the state that has the highest car sales. Florida has the next highest sales and is not even close to the first one. Next is close to Florida and it is ‘Texas’ which is the biggest state in the United States. So if the area is big does not mean that the car sales will be more, but they change with many other factors as well is one conclusion with this graph.

**Chart, bar chart, histogram

Description automatically generated**

**Fig 7. Car Sales Vs States**

**A picture containing map

Description automatically generated**

**Fig 7. Geo Location Graph**

**AVERAGE PRICE CHANGE WITH THE TYPE OF FUEL USED IN THE CAR**

This is something that comes under the maintenance of the vehicle. If the customer is ok with the price of the vehicle electric costs less in maintenance compared to gas but Gas type of vehicles are less in terms of price. The performance is somewhere similar between the two.

**Chart, bar chart

Description automatically generated**

**Fig 8. Price ($) Vs Fuel Type**

**COST COMPARISION OF CAR BODY WITH AVERAGE PRICE**

Everyone has their own choice and based on the work the customer does, the vehicle is chosen. For a businessman who owns a store probably needs a Truck or Minivan to meet their daily needs and it is costly compared to a good Sedan type. SUV type cars are costly as their luxury and comfort is different. So, this graph basically shows their difference in prices clearly on an average basis.

**Chart, bar chart

Description automatically generated**

**Fig 9. Price ($) Vs Car Body**

**COST COMAPRISION WITH TYPE OF DRIVE WHEEL**

Any car that drives in a Midwest region needs to be AWD type at least during the Winter to adhere to the Snowfall. It is safe to have AWD in such a region. So, if it is a city like ‘Dallas’ the AWD is not a compulsory feature as most of the time the weather is ‘Hot’ and does not snow there. But at a place like ‘Denver’ an AWD type car is a must to avoid accidents and drive safe.

**Chart, bar chart

Description automatically generated**

**Fig 10. Average Price ($) Vs Drive Wheel Type**

**AVERAGE PRICE VS CAR SIZE TYPE**

For a Full-Size car there are many options that are available and it helps in enabling so many features and luxuries available in a car. So the average price of a Full-size is obviously high but what we should observe is that there is not much difference in price than a mid-size car.

**Chart, bar chart

Description automatically generated**

**Fig 11. Average Price ($) Vs Car Size Type**

**AVERAGE PRICE VS NUMBER OF CYLINDERS**

Number of Cylinders of a car affects the performance of the car. It also affects the gas or electricity it consumes. The more cylinders the performance is more and consumption of fuel is more.

**Chart, bar chart

Description automatically generated**

**Fig 12. Price ($) Vs Cylinders**

**TREND OF NUMBER OF CARS FOR SALE VS THE MANUFACTURED YEAR**

**Chart, histogram

Description automatically generated**

**Fig 13. Density (No of Cars for Sale) Vs Year**

**CAR PRICE VS MANUFACTURED YEAR**

Car price by manufacture year visualization graph shows the price of a different brand of cars according to its manufacture year. It shows the difference between the price of different cars manufactured in a particular year, like this it shows the difference in price of car of different brands in every year. This will help the users to analyze the value of a car without any difficulty and save time by not going through huge data.

Chart, bar chart

Description automatically generated

**Fig 13. Average Car Price ($) Vs Year Manufactured**

**COMAPRING ODOMETER DISTANCE TRAVELLED WITH CAR BRANDS**

For any car, the number of miles driven is the most important and the life span of the car is clearly shown in these numbers. So the odometer gives this data from the car and this data compared with each car brand gives an idea about which brand of car is the most driven and how often people are using these brands is show in the Fig 14.

**Chart, bar chart, histogram

Description automatically generated**

**Fig 14. Odometer (Miles Driven(mph)) Vs Brands**

**PRICE AND MILES DRIVEN BY MOST DRIVEN CAR DATA VISUALIZATION**

Price and Miles driven analysis is very important for preowned cars. This visualization illustrates Price the pre-owned owner sold it for, and the Miles driven of that car by the seller. Ford has highest miles used for old cars and Morgan is the lowest miles used car. The scatter plot graph provides the data about the ten brands, which are Ford, Chevrolet, Mercedes-Benz, Toyota, BMW, Jeep Volkswagen, Audi, Lexus, and Tesla. The price range is between 0-40k, and distance driven range is 0-400k. This graph shows lot of data on these ten car brands with the car details clearly mentioned in the mouse hover. It gives the details about the Manufacturer, Price, Distance Driven, Year it was manufactured, Model of the car and the Condition of the car.

A screenshot of a computer

Description automatically generated with medium confidence

**Fig 15. Miles Driven(mph) Vs Price ($)**

Graphical user interface

Description automatically generated with low confidence

**RELATED WORK**

Initially it was purely our thoughts on doing something with the used cars dataset. With time we came across this work done by a Data scientist professional ‘Harshit Gupta’ in India. He used ‘Python’, ‘Pandas’, ‘Seaborn’ and ‘Plotly’ and generated some graphs from a dataset. The initial steps on cleaning data and preprocessing were done clearly in his work. The visualization concepts he used are not satisfying and thus we came up with improving the concepts used to analyze the data. We developed so many graphs to conclude about all the features that affect a car and what are all the aspects a buyer looks at before buying one. Luckily, we found a good dataset from ‘Kaggle’ that has so many columns on features about different brands of cars. ‘Seaborn’ and ‘Matplotlib’ libraries helped in enabling options to visualize the graph dynamically. Finally resulted in this project with comparison between features and bringing out so many conclusions about which is the best or better or chosen option and which are not, and which comes under the average category to buy a car based on the features involved in the dataset. The data in the dataset is not completely true to the best of our knowledge so that is something that needs to be improved.

CONCLUSION

Now a days cars usage has increased a lot and people are interested in buying a used car. This project will be useful when we need to analyze the market and which cars are highly used lately while considering all the features. Our analysis clearly says we are supposed to compare the features available and depending on the prefer3enec of the customer the patterns are changed, and the conclusions are altered. So here are a few conclusions based on the graphs generated and the data in the dataset.

1.Most cars are ‘Gas’ type of Fuel and if you are someone interested in ‘electric’ then the price to pay for the car is going to be high but the other important aspect is, it has low maintenance during the long run.

2.Our wait time is not at all an option to buy a car as the price changes are unpredictable.

3.Four Cylinder cars are cheaper in price and help a budget friendly customer.

4.There is no big difference between a Compact or Mid-Size car, so it is better to choose a Mid-size car and if the customer is fine with some couple grands, then full size is an option to choose.

5.Regions that have snowfall prefer the All-Wheel Drive cars to the FWD or RWD.

6.Sedan model cars’ average price is less compared to other SUVs or Coupe models.

7. There are very few cars which are driven the most and have the highest price. So that’s because of the brand and people don’t consider much of a brand if the miles driven are more.

Let us conclude by taking the example of a customer who is so concerned about the price and wants a budget friendly car. The opinion could be a Ford car with ‘$7500’ price and 100,000 miles driven with ‘Gas’ Fuel Type, Cylinders as ‘4’ could be a considerably good option. Likewise based on the customer preferences we could come to conclusions about each feature and decide based on the filtered options available at the end of the comparisons.

FUTURE WORK

To improve this project there are a few ways. One is to build a Web or Mobile application and arrange different filters in each section. All the filters of the car features have to be selected and based on the customer's requirement, they will give the input of their choice and with considering all those possibilities, the result is a scatter plot that has all the available cars and their details like ‘Fig 15.’ for the range of filters given by the customer. This result is something that is not single, and the customer can decide among the generated results. If the choices of each filter are more precise then there could be very few after all the selections are made and ultimately it will be easy for the customer to choose among a few of them. Also, the customer can alter their choices and see what are the after results if those choices are made.

REFERENCES

[1] H. A. Ayad, L. A. Al-Kilani, R. Arshad, M. A. Al-Obadi, H. T. Hussein and M. Kucukvar, "Developing an Interactive Data Visualization Platform to Present the Adaption of Electrical Vehicles in Washington, California and New York," 2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA), 2020, pp. 731-737, doi: 10.1109/ICIEA49774.2020.9101928.

.

[2] Sharan Kumar Paratala Rajagopal, 2020, Data Visualization for Vehicle Selection Process, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 07 (July 2020).

[3] *Matplotlib Documentation*[*https://matplotlib.org*](https://jovian.ai/outlink?url=https%3A%2F%2Fmatplotlib.org)

[4] Shaun Bangay, “Visview: A system for the visualization of multi-dimensional data”, in “Visual Data Exploration and Analysis V”. (TA 1505 Pse 3298)

[5] Harshit Gupta, 2021, Exploratory Data Analysis of User Cars in the United States using Python, Seaborn, Pandas and Plotly

[*https://blog.jovian.ai/understanding-used-cars-market-in-usa-52489e10d551*](https://blog.jovian.ai/understanding-used-cars-market-in-usa-52489e10d551)

[6] Moosavi, Shoban, M.H.S.S.P., Ramnath., R.: “a country wide traffic accident dataset” (2019)

AUTHOR CONTRIBUTIONS

Sashi Kanth Pepalla – Data Preprocessing, Data Cleaning, Presentation, Documentation, Dataset Study, PPT preparation.

Priyanka Munnangi – Mid Project Presentation, Documentation, Data Analysis, Data Set Study.

Jahnavi Kolla – Data Analysis, Documentation, Data Set Study and active discussion.

Aadil Mohammed - Data Analysis, Documentation, discussions about the car concepts.

Pavan Kumar Reddy Palanki – Data Analysis, documentation, discussions about the car concepts.

Niharika Guntupalli – Data Analysis, Matplotlib and other libraries their usage explanation to the team, documentation.