|  |
| --- |
|  |
| Challenge Proposal |
|  |

|  |
| --- |
| Oerlemans,Timo T.  13-10-2022 |

# Introduction

My first thought of the project was something to do with sport but when I asked for feedback and looked at it closer it was not a good idea. There were too many features that are not really predictable and human aspects. After that I went to my group and asked if they could help me brainstorm some ideas.

That is when this idea came into my head. My idea for the challenge is that I want to predict how much cars will cost. I found a dataset that has a lot of brands of cars and models. I am going to use this to predict how much a car is going to cost. The few last years there are more electric cars on the market and in the future maybe other fuel types.

Content

[Introduction 1](#_Toc120002393)

[Phase 1 Proposal 3](#_Toc120002394)

[Research questions 3](#_Toc120002395)

[Main question 3](#_Toc120002396)

[Sub questions 3](#_Toc120002397)

[Why do I do this? 3](#_Toc120002398)

[What I am going to do? 3](#_Toc120002399)

[Who? 3](#_Toc120002400)

[When am I going to do what? 3](#_Toc120002401)

[How am I going to do it? 4](#_Toc120002402)

[Domain understanding 4](#_Toc120002403)

[Interview 5](#_Toc120002404)

[Data sourcing 5](#_Toc120002405)

[Analytical approach 6](#_Toc120002406)

[Delivery 7](#_Toc120002407)

[Demonstration 7](#_Toc120002408)

[Feedback 7](#_Toc120002409)

[Appendix 8](#_Toc120002410)

# Phase 1 Proposal

## Research questions

### Main question

What is going to be the price of cars?

### Sub questions

* What are the best features to predict the price?

## Why do I do this?

I want to do this because each year there are many cars sold but the new generation has not that much money to spend. Even some of the older generations have not enough money for a car. According to the CBS 1 on 538 people who are above 18 years old have a car. A car yearly cost is around the 3500 and 7500 euros but the net income is around the 30.000 euros. So almost 1/4 of the net income goes to the yearly cost for the car. That is why most people want to have a cheap car that can just bring them from one place to another. For the consumer it is very handy to know how much a car is going to cost because they need to spare money to buy one or loan money from a bank. Almost 15% of all people in the Netherlands have a loan to have enough money for a car.

This selling price I want to predict is a helping tool for the individual sellers. In the other document(the notebook) you can see that 83,2% of the sold cars are sold by these type of sellers. I spoke to one director of a car dealer and he worked in this business for almost 25 years. He said to me that most of the times he could just see from a few things what the price roughly could be. But sometimes when there is something different that he had to do some research on how much this car could sell for. What if an unexperienced person wants to sell his own car how much research does he need to do. So that is why I make this model to help the individual seller.

## What I am going to do?

I am going to predict what the price will be for used cars. I will take a look at different variables and compare them with each other to see if there is a connection with the price.

## Who?

This challenge will only be made by myself. There are no external stakeholders so I can choose my own idea and requirements. There are however 4 teachers who are going to grade my work(Sabina, Qin, Nick and Frank).

As I said in the “why do I do this” I am going to make this model for the individual seller who have not as much experience as the most car dealers.

## When am I going to do what?

The first few weeks of this semester I am going to do some exercises and things for the group project. After week 5 I start to do more for my individual challenge. In week 5 I am going to deliver my iteration 0. In week 7/8 I want to deliver the second iteration and every 2/3 weeks after that an other iteration.

I want to ask every week for feedback. So sometimes it will be before my new iteration and sometimes the week after I delivered it. The feedback I ask from the teachers will be from all teachers so not just one. Every week I am going to ask 1 or 2 teachers for feedback or questions if there are some. Beside the feedback for my actual work I want to ask feedback in general from my semester coaches(Sabina and Frank).

After 12 weeks of working on this challenge I need to deliver my final product on canvas. This will be on the 20th of November.

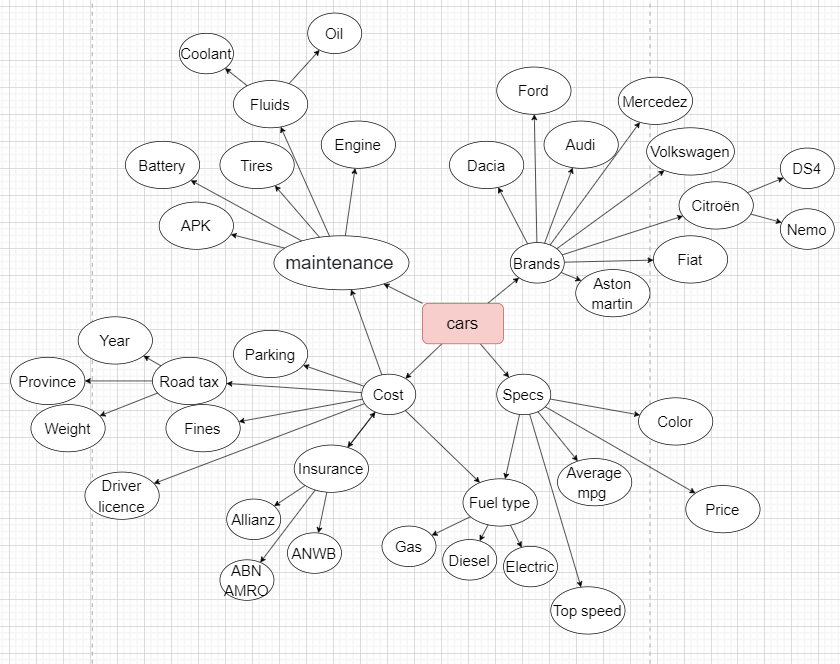
## How am I going to do it?

I am going to do this by working 4 days every on this challenge. Every Thursday afternoon and Friday morning I work on the group project but every other hour on this. Like I said before it is till week 12 and after that I have to work fulltime on the group project.

I am going to make this project by using a dataset that I found on Kaggle. It has all the requirements that I needed to make this model. The model it self will be an regression model because I want to predict an quantity not a category. In phase 3 in my notebook I have explained more how I came up with this conclusion.

## Domain understanding

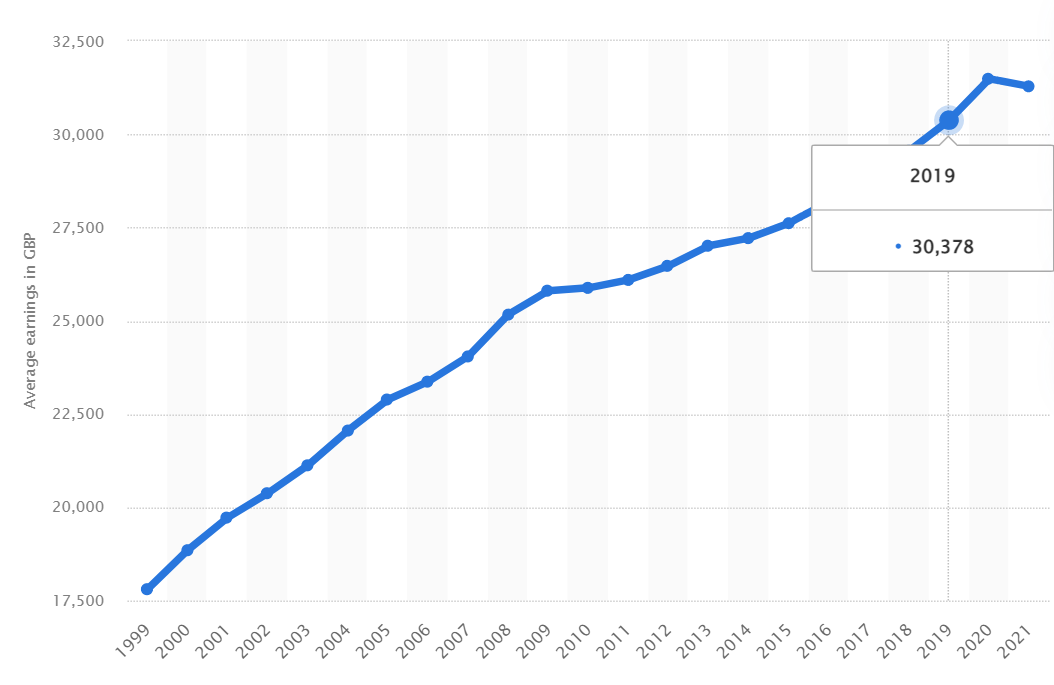
For the domain understanding I first made a cognitive scheme.



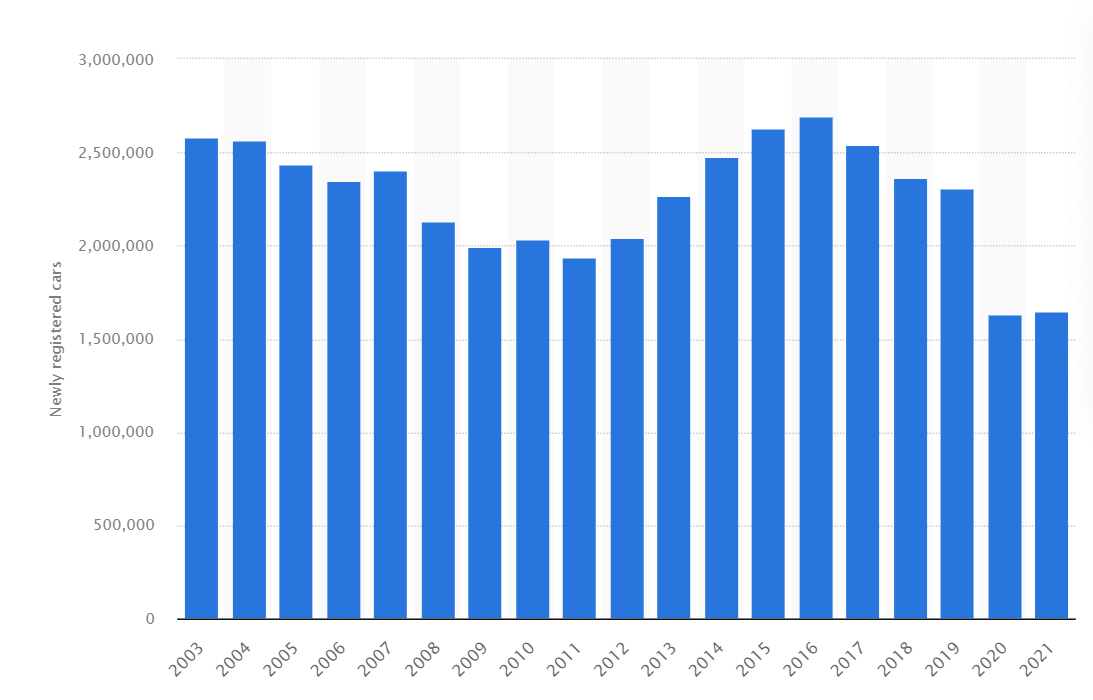
After the cognitive scheme I went on the internet and searched for the information I wanted to know more of.

**For an example I used the car price in the UK!**

In the UK a small car is around the 12.000 and 17.000 pounds, a medium car cost around 22.000 and 36.000 pounds. That is a lot if you compare it to the income of these countries. In 2020 the average annual income is around the 31.000 pounds. If you want to by a car you can spend almost nothing of your income that you have left after your monthly cost.



In 2020 there were considerable less cars sold if you look at 2019 or the years before that. It was because of covid-19 because most people had to work from home and didn’t need a car at the moment. And an other reason of the dip in cars that were sold is that many people got fired because here was not a lot of work in that period.



## Interview

I have held an expert interview with an car dealer. I first called some dealers that are established in the town I live in but they all said on the phone that they were to busy. So the next day I went to one dealer in the same street I live in. All the dealers were to busy with customers so I asked if I could speak to the director. The next day I could have an conversation with him.

In the we spoke about the impact of AI in the car sector, my challenge and what he thought. The impact of AI in the car sector in the sales direction is not there because they don’t need it. The dealers can come up with a price so quick that AI is not necessary for this direction. He said that the best thing AI does for the car sector is the self riding cars because many people want such car. For the entire interview you can see it on canvas.

## Data sourcing

I went on the internet to search for a dataset that has the information I think there has to be. I found a site named Kaggle and they had a dataset(Vehicle dataset. (2020, 24 oktober). Kaggle. <https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=Car+details+v3.csv>). There are 13 columns: name, year, selling\_price, km\_driven, fuel, seller\_type, transmission, owner, mileage, engine, max\_power, torque and seats.

After I took a little peak into the data I personally found the torque data unnecessary so I dropped this column. The other columns had different types of data in them. As you can see in the image below there were 3 columns as type int64, 1 as type float and 9 of them were objects. I wanted some of the columns to be int64 so I changed the data as you can see in the right image. I also changed the year into a age column so you didn’t had to calculate how old a car is.

Afbeelding met tekst, naamplaatje

Automatisch gegenereerde beschrijvingAfbeelding met tekst, naamplaatje

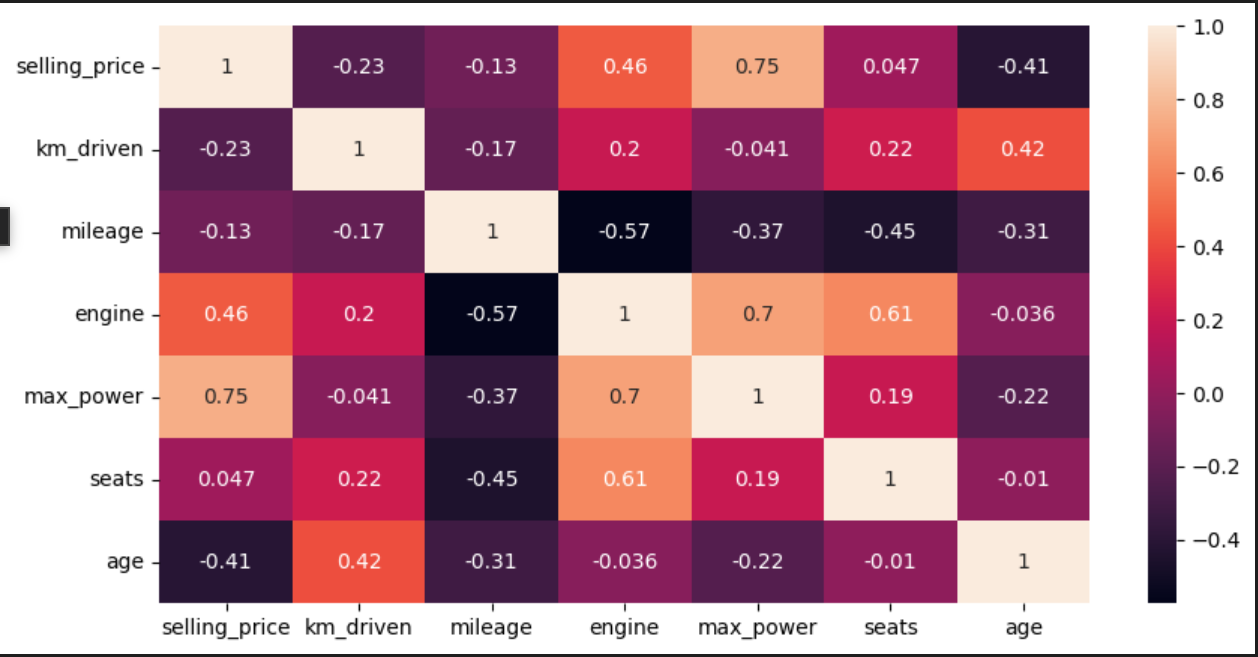
Automatisch gegenereerde beschrijving

After that I looked at how many data there was and there was missing data in some columns(mileage, engine, max\_power and seats). So I wanted to fill in those missing data values. I thought about what should be in those missing data and eventually came with the answer. The mean of every column should fill in those missing values.

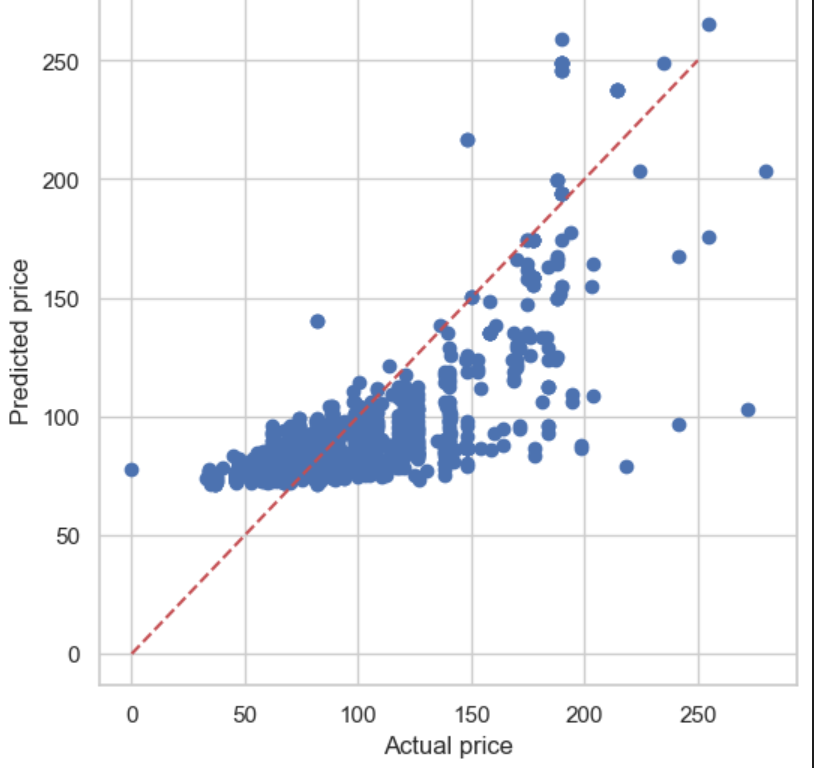
## Analytical approach

I have used this diagram from scikit-learn to see what kind of problem it is I am dealing with. It came out to be a regression problem because I want to predict a quantity and not a category that already is defined. It is also no dimensionality so it is a regression problem. I also have less then 100k samples so SDG regressor is not recommended. The Lasso and Elastic Net are recommended for my problem.

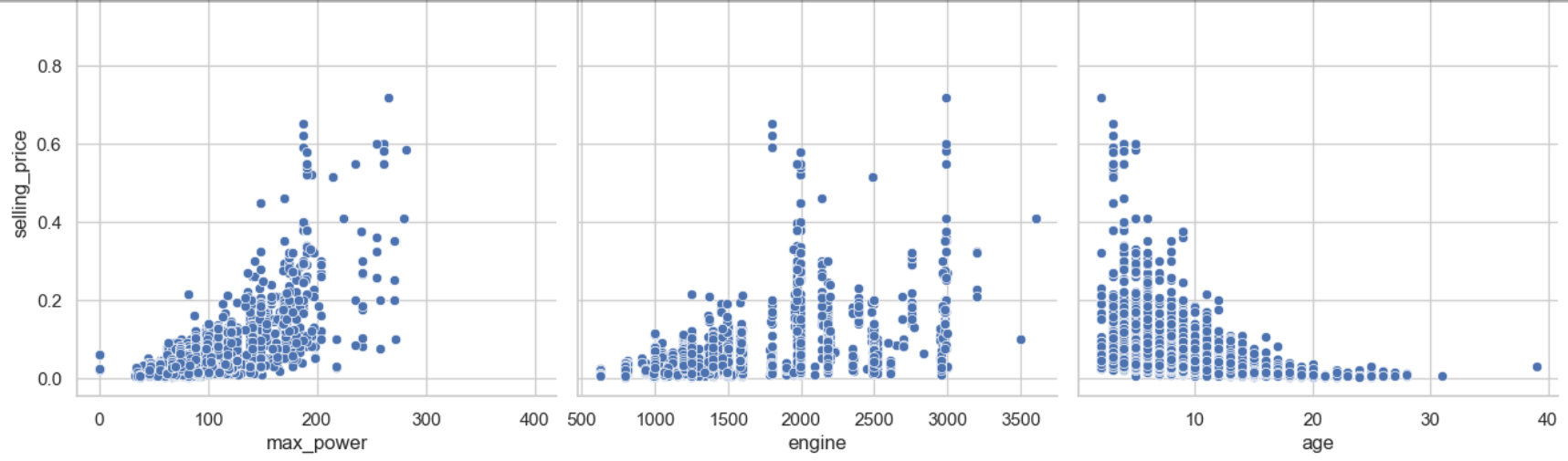
My target variable will be selling\_price because my goal is to predict the price for a second hand used car. The nearest neighbour can be used for classification as well as for regression so I am going to use that. But there are more kind of models like linear or vector machines. First I want to look at the results when I use linear regression and after that I am going to look at what I think is better.



As you can see in the heatmap there are a few good correlations. Like between selling\_price with max\_power or selling\_price with age. So these variables I look closer to then correlation that are not that good.



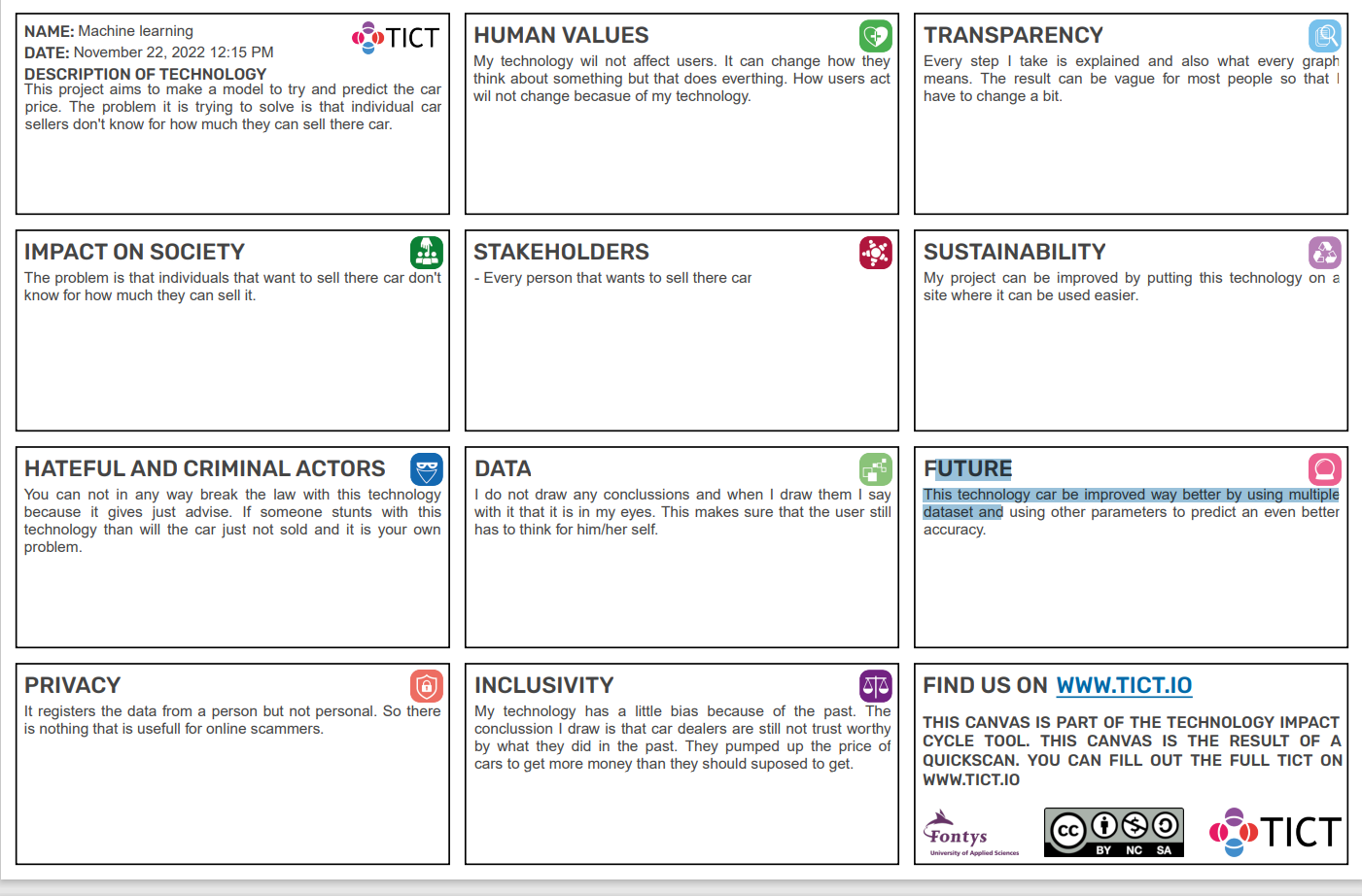
I have used linear regression but there is no linear connection between those variables as you can see in the image below. So in the next iteration I am going to use other models to predict the price better.



# Delivery

## Demonstration

Before I went to a stakeholder I wanted to do a impact scan. In canvas there is a link to TIC-tool that I used for my scan. The results are in the image below:



For my delivery I let my mother have a look through my documentation. She speaks for the individual seller because everyone who doesn’t have any car knowledge could be a stakeholder. I chose for my mother because a few months ago she wanted to maybe sell her car. Before this demonstration I haven’t talked about my project so she didn’t have any information about my project. I am going to do a face to face demonstration and let her scroll through the document. I have only told her that if something isn’t clear or if she has questions she should tell me.

## Feedback

During the demonstration I got some feedback from my stakeholder. She thought at first that is was a good idea. She had a question about my “who” question because it wasn’t that clear for her. This feedback was also given by Kaan an student who had to look at my project. Further she had some questions about the images but I told her that these images are explained in my other document. The other parts in the proposal where pretty clear to her. In the EDA she questioned why I wrote under the heading “About this document” about the dataset. I told her that some steps overlap each other. This way everybody knows where everything is. She didn’t understand everything under every heading but some information is not useful for her. Like the heading “How often do I want to retrieve the data” this is something for myself. The visualization was all clear except for 1 graph: boxplot. She didn’t know this graph so I had to explain how the boxplot works. After I explained it she said that it didn’t really bring more insight. She said that she understand why you should use certain values to predict the price.

For the prediction is self she didn’t know what the values meant. And Looking back, I overlooked the fact that I didn't explain the numbers properly. So for demonstration I explained the values quick so that she could give me more feedback. She said afterwards that she wouldn’t really use this because she has some connections and knows a good garage. The months that she doubted if she wanted to sell the car she went to a this garage. The garage just helped her to estimated a price she could still ask for it. Many people have connection these days and that is why I think this isn’t going to be used that much. Only people who don’t have a connection with someone who is an expert with cars.

# Appendix

Yurday, E. (2021, 11 februari). Average Cost of Cars UK 2021. NimbleFins. <https://www.nimblefins.co.uk/cheap-car-insurance/average-cost-cars-uk>

Statista. (2022, 26 oktober). Average annual earnings for full-time employees in the UK 1999-2022. <https://www.statista.com/statistics/1002964/average-full-time-annual-earnings-in-the-uk/>

Statista. (2022a, april 25). Volume of new passenger cars registered in the UK 2003-2021. <https://www.statista.com/statistics/299240/volume-of-new-passenger-cars-registered-in-the-united-kingdom/>

Choosing the right estimator. (z.d.-b). scikit-learn. <https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html>