# 

STAT 112 - Introduction to Data Processing and Visualization Project

The Most Sustainable and Equipped Tractor to Buy

# by

Name

January 20, 2023

# 

# 

# 

# Abstract

This research is about tractors on sale in different provinces having different features and prices. The data required precise data tidying and cleaning in order to be used and being interpreted throughout the research questions examining mostly the diff erences between tractor brands and the tractors’ safety levels. As a result of our research questions on the clean data, we interpreted relationships between several numerical and categorical data.

# Introduction

In this research we have prepared and analyzed a dataset that contains data from 115 unique tractors with different features. Main goal of this project is to decide on which tractors are the most ideal ones to buy in terms of safety, reliability and horsepower. Original dataset contained 10 variables:

| *variable name* | *description* | *scale* |
| --- | --- | --- |
| price | price of the tractor | interval, continuous |
| horsepower | horsepower of the tractor | ratio, continuous |
| brand | brand of the tractor | nominal |
| number\_of\_cylinder | number of cylinder of tractor | ratio, discrete |
| number\_of\_clicking\_rate | number of clicking rate in tractor’s website | interval, continuous |
| number\_of\_failure | number of failures since production | ratio, discrete |
| province | province that tractor has been sold | nominal |
| warranty\_period | warranty period of the tractor | ratio, continuous |
| warranty\_period\_unit | unit of warranty\_period | nominal |
| age | age of the tractor | ratio, discrete |

Categorical columns contained errors in string formatting and all of the columns had missing values. Also the warranty period had two units, disturbing the uniformity. We cleaned those errors methodically. Details of the cleaning process can be found in following section.

Exploratory data analysis gave us several insights about the data. Then we tried to answer our research questions by diving deeper to the data.

# Data Tidying and Cleaning

In the data pre-processing stage, our goal was to find and fix the inconsistencies in the data set. These include the formatting of strings, uniformity of the units, detecting the outliers and eliminating them in cases where necessary, removing duplicate rows, and imputing the missing values where possible. We methodically cleaned the data set by following these steps.

## String formatting

We formatted the strings mainly by using two functions: str.strip() and str.title(). Used together, these two functions fixed the case formatting errors in the categorical columns. The number of unique values was reduced as expected:

* 5 unique values in **brand** column reduced to 4,
* 4 unique values in **warranty\_period\_unit** column reduced to 2,
* **province** column did not have any misspellings or trailing zeros, thus remained the same.

Upon further examination of the unique values in the brand column revealed that one of the brands, namely “New Netherland”, does not actually exist. We think that this might have been due to someone mixing the word “Holland” with “Netherland”, so we replaced it accordingly. At the end we were left with following unique values:

* **brand**: New Holland, Massey Ferguson, Solis
* **warranty\_period\_unit**: Month, Year
* **province**: Konya, Aydın, Ankara

## Verifying the uniformity of units

The units of two variables are important in this dataset. One of them is the price of the tractors and the other is the warranty period of the tractors. When the outliers are removed, the price values fall within close range. Also the coefficient of variation is 0.07, indicating low variance. Thus, we concluded that the price values are all of the same unit.

In the **warranty\_period\_unit** column we had two different units: years and months. Dividing the warranty\_period value by 12 where the unit was months allowed us to convert all values into years. Afterwards, we checked that the new values are acceptable and that there are no obvious outliers to make sure everything worked perfectly.

## Outliers

Speaking of outliers, in the next step we detected the outliers and removed them. One of the ways of detecting the outliers include calculating the interquartile range and using the value obtained by multiplying the IQR by 1.5. We created a special function that checked to see if there are any values outside the and range for a given column. We then iterated over numerical columns and printed the outliers. The function detected outliers in three of the columns:

* The price column had 6 outliers. All of them were significantly more than the rest of the values, so we nullified these values altogether.
* Same could be said for the age column. There were 6 outliers that were promptly nullified as well.
* As for the number of failure column, we decided to keep the outliers since they were still within acceptable range and might still contribute to the analysis.

## Missing values

We had missing values in all of the columns and the number had only increased after removing the outliers. At first, we wanted to impute the values by using the mean and the mode. However, we decided against it since we didn’t know the mechanism by which the missingness had occurred originally. Some of the numerical columns had strong linear correlation between each other and imputing the missing values with the mean would decrease the value of the correlation coefficient. Since the value of the correlation coefficient was high between the variables: horsepower - number of clicking rate and price - horsepower, 1.0 and 0.98 respectively, we decided to create a function that would calculate the necessary coefficients for a linear line that fits the scatter plot. We used the *polyfit()* function from the *numpy* library. This function returned coefficients for the first order linear regression in the form of . Our function then imputed the missing values based on the value on the other column. Imputing the missing values using this approach has its drawbacks as well. In fact, it has the opposite effect of enforcing the linear correlation and actually increasing the correlation coefficient. However, since the amount of missing values is relatively low, we figured that the difference would be negligible. Using this method, we were able to fill in all the missing values in the horsepower, the number of clicking rate and the price columns. This approach works when there is linear correlation. However, not all the variables are correlated in this way.

The number of cylinders of a tractor is correlated with the horsepower. Based on the scatterplot, we can see that the number of cylinders takes specific values when the horsepower is between a certain range. In order to figure out the intervals of horsepower for the number of cylinders we grouped the data set by the number of cylinders and extracted the minimum and the maximum of the horsepower. Based on the results, we imputed the number of cylinders. Since the horsepower had all been filled in in the previous step, there should no longer exist any missing values in the number of cylinders column. As for the rest of the numeric columns, since there was no apparent correlation, we used simple mean imputation in order to fill them.

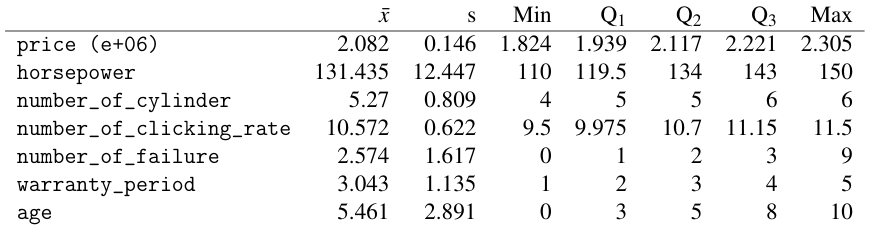
When it comes to the categorical columns, upon further analysis of the data set, a pattern had emerged. The brand and the province columns had values that were sorted in order. We figured that the sampling technique used to gather this data did so in a way that prioritized these two columns. We extracted the index ranges that contained the brand names and the provinces and imputed the missing values based on the index. With that, the data set was completely free of missing values.

## Duplicates

We checked the duplicates using *pd.DataFrame.duplicated()* both before and after the cleaning process. It did not contain any duplicated rows.

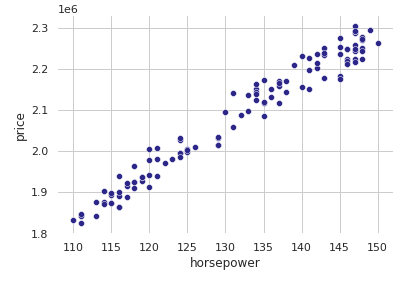
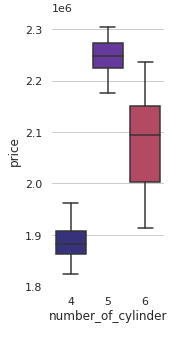
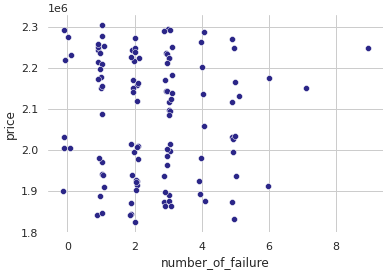
# Exploratory Data Analysis

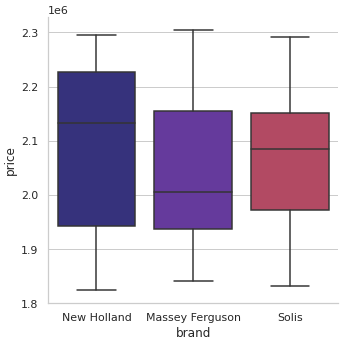
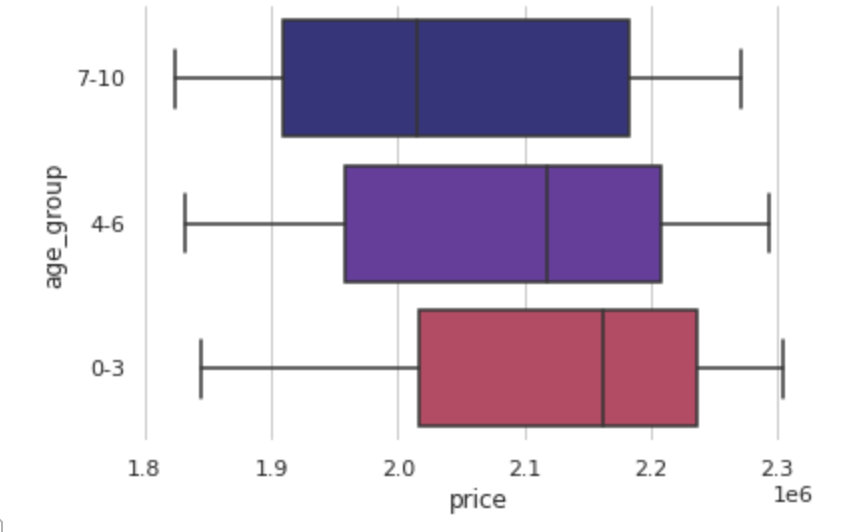
After the data cleaning process, we started working on analyzing the data set which contains properties of 115 tractors produced by three different brands. There are 9 columns, 7 of which are quantitative and the other 2 are qualitative. We created the following table that contains the descriptive statistics for numerical columns, so that we can investigate and understand the data properly. In this way, we were able to observe the mean, the standard deviation, the quartiles, the minimum and the maximum values of each quantitative data.



In the following chapters, we explain the meanings of these findings.

### Tractor prices and what they're affected by

Price is a quantitative, continuous and a ratio scaled variable. All of the values are between 1.824M and 2.305M. However, the interquartile range includes prices between 1.939M and 2.221M. The mean is around 2.082M. The median value is 2.117M, which is close to the mean, however it is a little more than it. Which shows that the distribution of the price is possibly left-skewed. The standard deviation is high at around 0.146M. We created appropriate plots of price with each variable in order to understand what and how the price is affected by them.

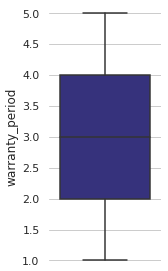
We can directly observe that there is a positive correlation between price and horsepower, which also means that it has the same correlation with number of clicking rate as well. This shows us that as the power of the engine and popularity of the tractor increase, its price also increases. Apart from these, another thing that caught our attention was the effect of the number of cylinders on the price. While the most expensive tractors have 5 cylinders, they are followed by tractors with 6 cylinders. We can say that the cheapest tractors have 4 cylinders. Another important finding was the relationship between the price and the number of failures. We were expecting the tractors that have high number of failures, say more than 5, to have low price points, which was not the case. We have 4 data points that fit this condition, and 3 of them are for sale at prices above the mean price. Another important case is the relationship between brands and price. To see this we made a boxplot for each brand. The brand with the widest price range is New Holland. In general, we can say that New Holland has the highest median vehicle price, followed by Solis. According to the data we have, any brand has an outlier. And of course, the most important part is that we can say that the most expensive tractor brand in general is New Holland. The other intriguing question is **“Does the age of the tractor affect its price?”** We divided age into three groups: 0-3, 4-6, and 7-10. Then we created a boxplot to see how it changes based on price. The medians of the prices increase while the age decreases -the tractors get newer- which is an expected outcome. However, the price ranges are approximately equal, so there are different types of tractors having weaker/stronger features or shorter/longer warranties for the brands to sell in all age groups.

## 

## Research Questions

### What can be the causes of the tractors with the highest numbers of failures?

There are two clear outliers in the number of failure distribution, which are called high failure risks.

First quartile of warranty period is labeled as short term warranty, last quartile as long term warranty and the interquartile range as standard warranty. According to the boxplot, warranty period seems to have a symmetrical distribution, which shows that there is a variety of tractors to be sold between 1 and 5 years of warranty periods.

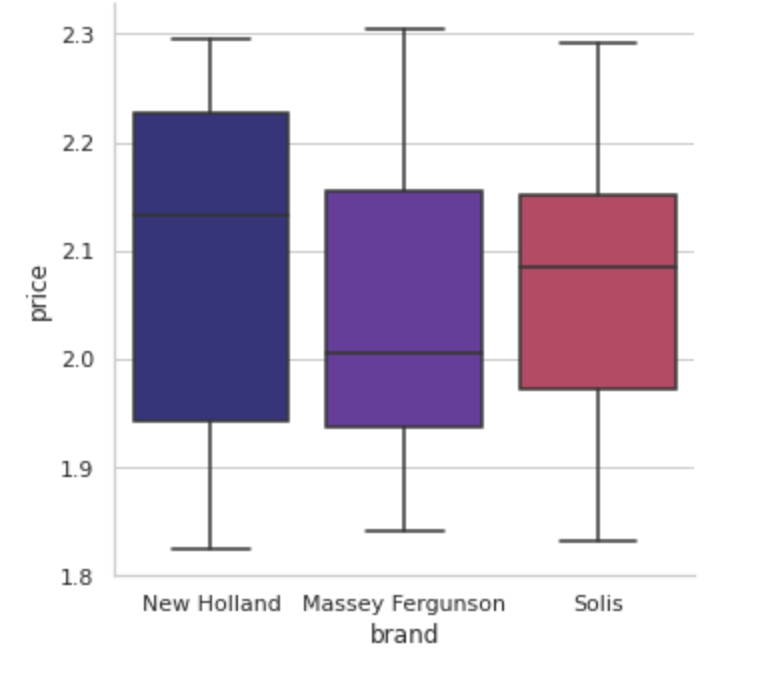
The number of failures 7 and 9 are the outliers and considered as the tractors with the highest failure risk. Since their safeties due to success are very low, their sales will be more successful if they have longer terms of warranties.

### Which features make the farmers decide on the safest tractors to use?

Warranty periods and average failure risks are the most important elements of safety which the farmers check before selecting the tractor most suitable for them. Failure risk here calculated by dividing number of failures with tractors’ age and clustering in three categories. The examination of the stacked bar chart including these two elements illustrates that tractors with normal failure risks are the most to find in all lengths of warranty periods, which are suitable for all budgets.

### How do the features of the tractors change in the variance of brands?

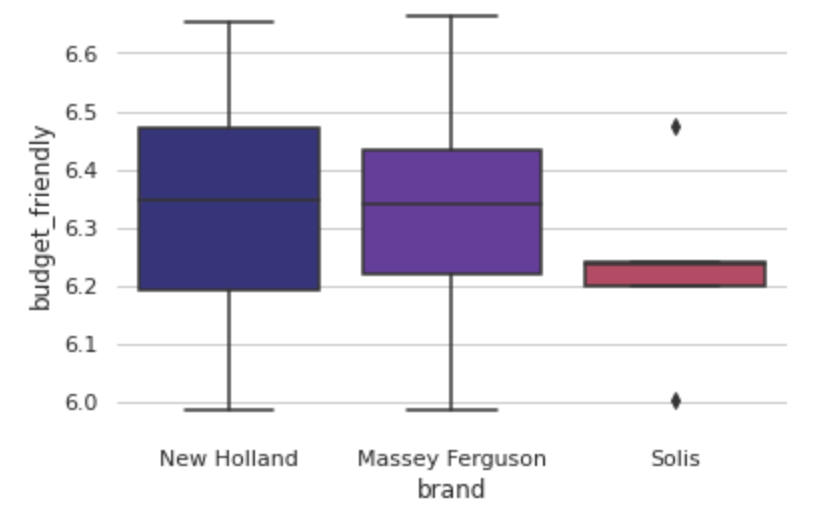
Here, the New Holland has the largest range and interquartile range in the number of clicking rate categories. Moreover, New Holland has the highest median value in terms of number of clicking rates. While the medians of New Holland and Solis are close to each other, Massey Ferguson has a significantly lower median. Additionally, New Holland and Solis have left skewness while Massey Ferguson has a right skewness.



How do the average tractor prices of different brands differ from each other?

According to average prices, the ranges of the 3 brands are close to each other. However, their median values are significantly different from each other. New Holland has the highest median value, while the other two brands have lower median value. Additionally, Massey Ferguson and Solis have close interquartile range while New Holland has highest interquartile range value. Solis and New Holland have left skewnesses while Massey F. has a right skewness.

### Which brand of tractors offer more horsepower per price paid?

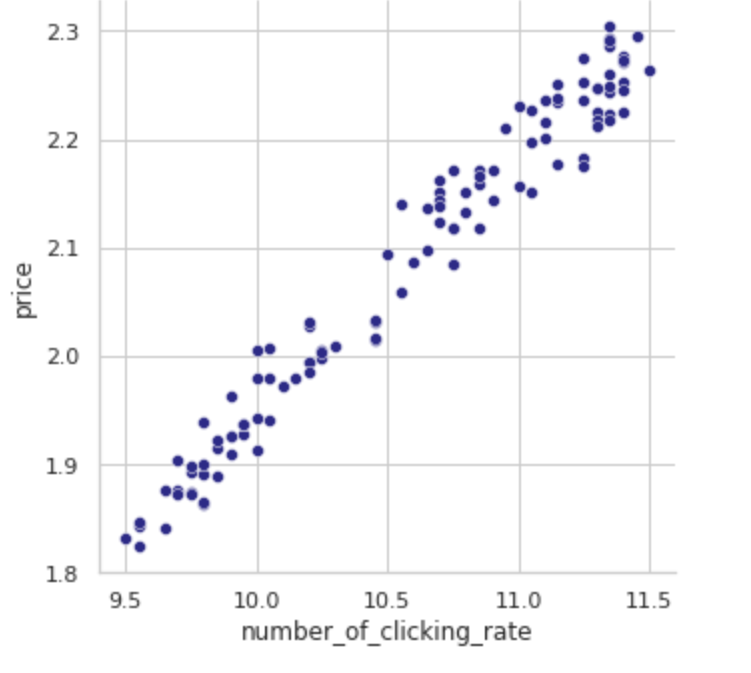


The horsepower of a tractor per the paid price is named as the level of the tractors being budget friendly. Solis is the last brand to be chosen since it has the shortest range without any whiskers and two opposite outliers. On the other hand, the distributions of New Holland and Massey Ferguson are approximately equal, having a median of a budget-friendliness above average. Both brands can be considered since they offer the most horsepower per price paid.

### What brand of tractors do farmers in different provinces prefer to sell?

Firstly, Massey Furnginson and Solis branded tractors are only on sale in Aydın, while New Holland branded tractors are on sale in all of the provided provinces. Also, the stacked bar illustrates that the lowest number of tractors are on sale in Ankara while there are approximately equal numbers of tractors on sale in other provinces.

### Are overpriced tractors more popular than the others?

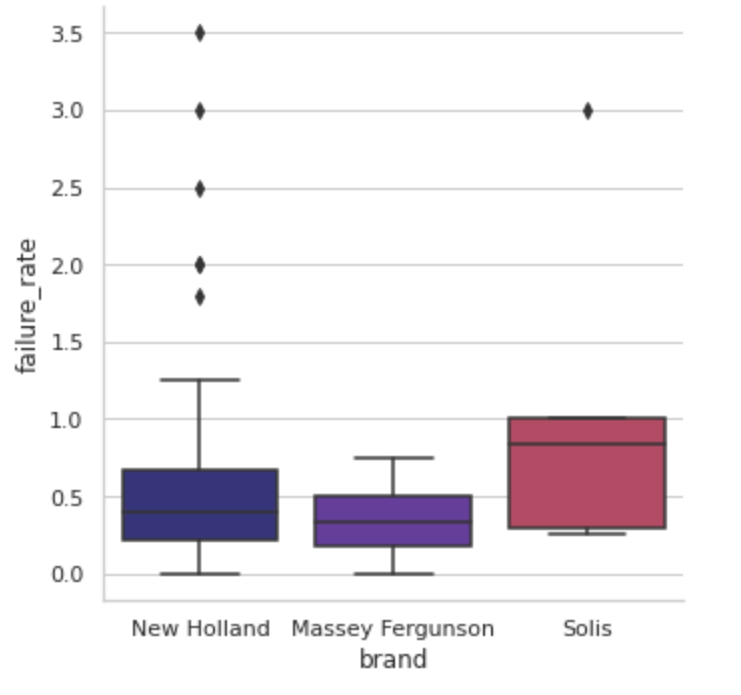


The scatter plot between the two numerical data illustrates that there is a strong positive relation between price and popularity. There is also an intensity of tractors being both overpriced and popular. Thus, farmers tend to sell tractors over about 2.1 million and have the most clicking rates on the site.

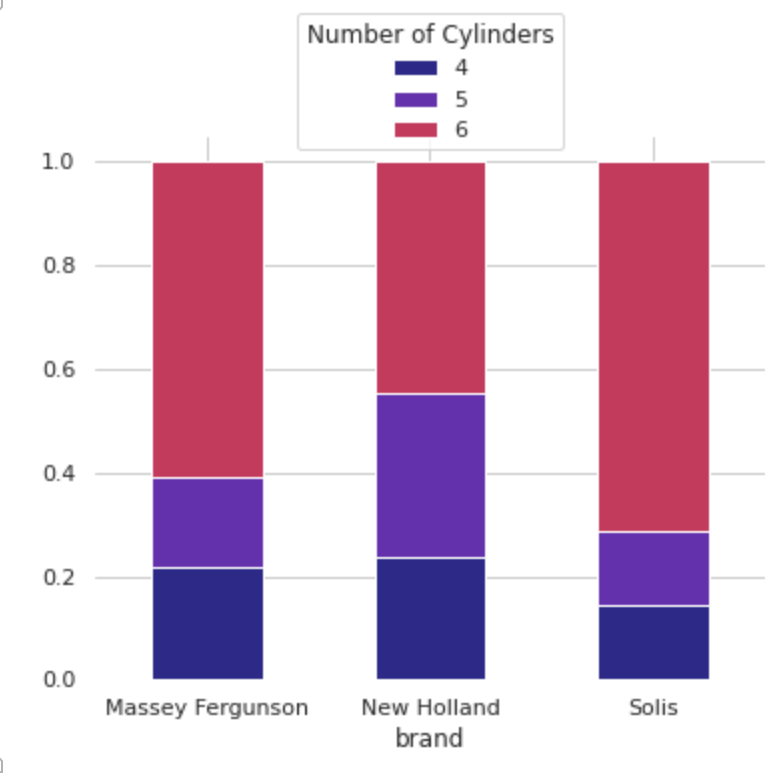
### 

### 

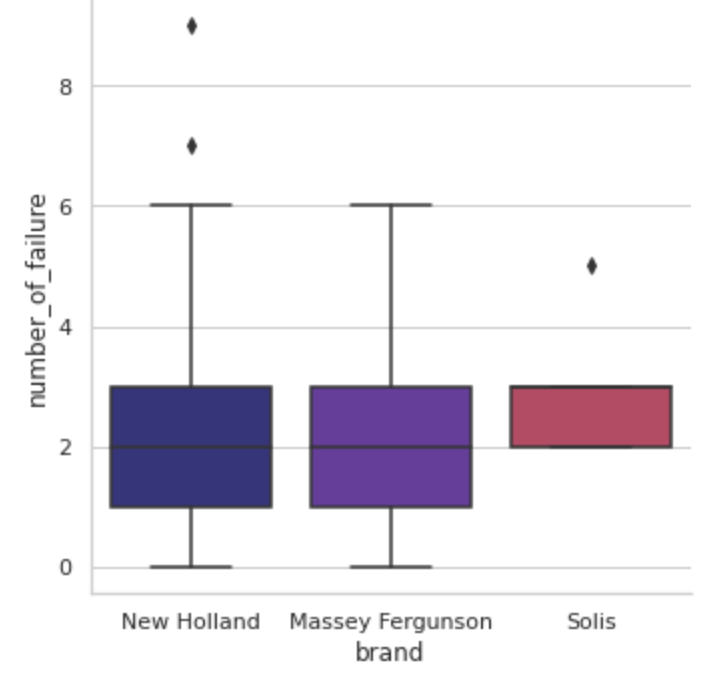
### How successful are the tractor brands in terms of their failure rates?



By the interpretation of the box-plots given, New Holland branded tractors have the most outliers as being the least successful brand. Followingly, Solis also cannot be considered as a highly successful brand since it has an outlier with a really high failure rate and its interquartile range reaches the most failure rates in the given brand group. However, Massey Ferguson has the most success rate, since it doesn’t have any outliers and has the most squeezed range among other brands.



All of the brands sell tractors with 6 cylinders the most. It is most likely for a farmer to select a tractor with 6 cylinders and least likely to select one with 4 cylinders overall. Even though the tractors with 4 cylinders are the minority, Massey Ferguson branded tractors with 4 cylinders are on sale more than ones with 5. Compared to the other two brands, the sales rate of 5 cylinders tractors is higher in the New Holland brand. However, if we look at the general, we can say that 6 cylinders tractors are preferred more in all brands.



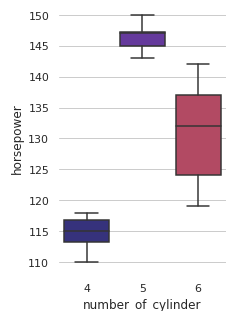
New Holland branded tractors have an approximately same distribution as Massey Fergusons’ but due to the fact that they have two outliers with higher number of failures, they are less safe than Massey F. Moreover, Solis has a compact range without having any whiskers but only an outlier which lies below the other brands’ ranges. Thus, we can say that Solis is the most successful brand to sell tractors since its failure rates are in a more predictable and narrow range -2 or 3 failures-.

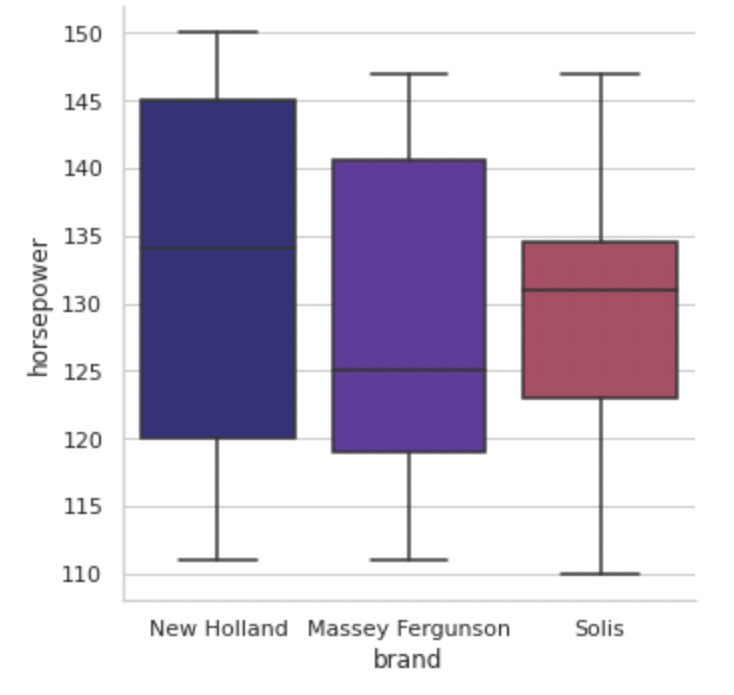
## Interpretations

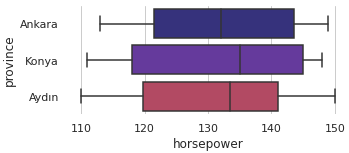
### What does tractors’ horsepower have a relationship with? ?

When we say a vehicle, one of the first and most important things that comes to mind is the power of the vehicle’s engine, so we wondered whether the horsepower of the tractor is really that much important, and we started to examine it.

Horsepower is a quantitative, continuous and a ratio scaled variable. All the data points are between 110 and 150. However, the interquartile range includes horsepower between 119 and 143 and the mean is around 131. The median value is 134 , close to the mean, however it is a little more than it. Which shows that the distribution of the price is possibly left-skewed. The standard deviation is low at around 12.5, which means that the data has a low dispersion rate.



We have already examined the relation of tractors’ horsepower and their prices. Since the correlation coefficient between the horsepower and the number of clicks the tractor gets is exactly 1 we can say that as the power of the tractor increases its number of clicks also increases, so does its popularity. This was actually a result we expected, since both variables have a positive relationship with price, therefore they have a positive relationship with each other, too. Another important point that caught our attention was its relationship with the number of cylinders. As you can see, the tractors with the highest horsepower have 5 cylinders, while the weakest tractors have 4 cylinders.

**“How does the horsepower differ from brand to brand?”** As with the price, the widest range in horsepower is New Holland's on the other hand Solis has the narrowest. New Holland has the highest median and then Solis follows it. The interesting finding is that Massey Ferguson's horsepower range is wider, nevertheless Solis's median is higher. **What is the distribution of horsepower of the tractors sold in different provinces?** To examine this question, we created a separate boxplot for the horsepower of the tractors sold in each province. Although the medians are very close to each other, we see that Konya has a higher median. Again, although the ranges of the 3 cities are very close to each other, Aydın has the widest range. 

### The number of cylinders

The number of cylinders is a quantitative, discrete and a ratio scaled variable. There are only 4, 5, and 6 cylinder engined tractors in the data set.

We have examined the relationship between the number of cylinders and the price and horsepower above. We were wondering if there is any connection between the number of cylinders and the number of clicks the tractor gets? What we saw on the table was exactly what we were expecting. As we observe, the most popular tractors have 5 cylinders. Opposite to that the least ones have 4 cylinders.

# Conclusion

In conclusion, throughout our project on the interpretations and questionings on the tidied and cleaned data, we tried to show a path to the farmers who will buy a tractor for their farms. The area we most cared about was the tractors being safe and how their features varied between different brands. We focused on the features which increased the efficiency, the warranty periods which would make the buyers relieved and the prices which have a great effect on the choosing process. …