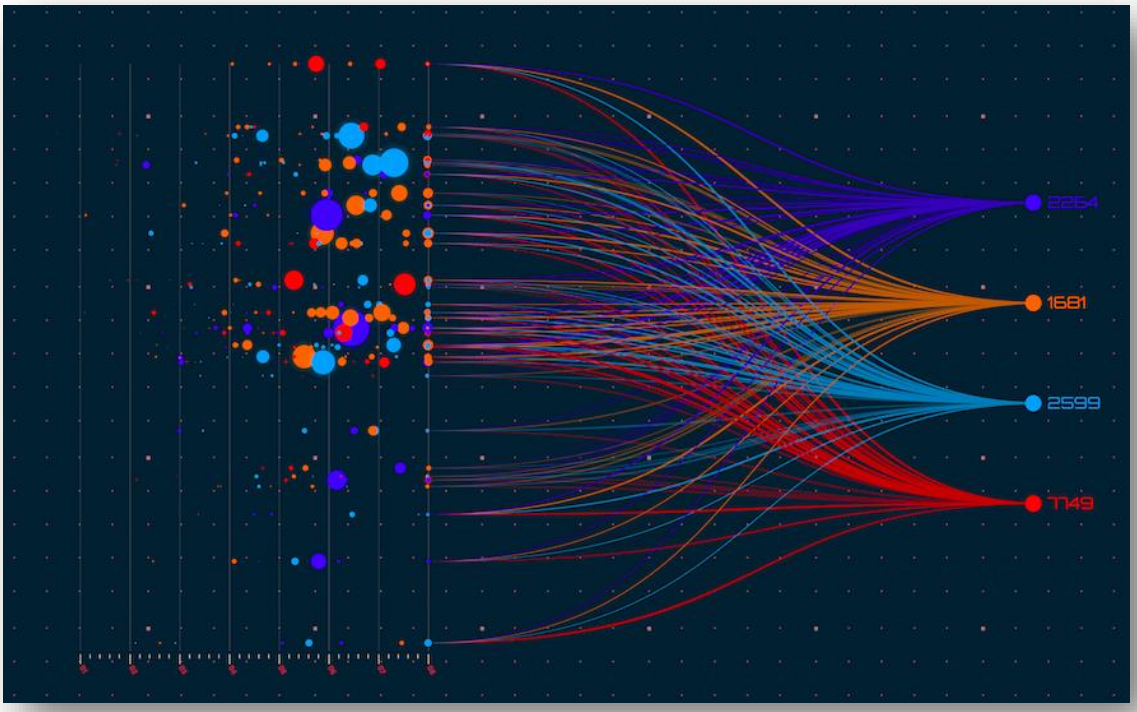
DTU
DTU Engineering Technology
Department of Engineering Technology and Didactics

# Machine Learning: Project 1




*Source: Treehouse Technology Group,*
*https://treehousetechgroup.com/wp-content/uploads/2020/01/Screen-Shot-2020-01-28-at-10.13.28-AM.png*

**Course**
02450 Introduction to Machine Learning and Data Mining

**Professor**
Bjørn Sand Jensen

**Group Number**
105

**Group Members**
Jonas Rose Lund Pedersen (s225269),
Søren Rønnekær Holgreen Graae (s225266)

**Hand-In Date**
20-03-2024

# I. Responsibilities

| Chapter | Student | Responsibility |
|---|---|---|
| 1 | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 20%<br>80% |
| 2 | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 60%<br>40% |
| 3 | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 10%<br>90% |
| 4 | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 70%<br>30% |
| 5 | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 90%<br>10% |
| 6 | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 50%<br>50% |

# II. Updates

| Chapter | Change | Whom | Responsibility |
|---|---|---|---|
| II. | Chapter added. | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 0%<br>100% |
| 2. | A better description of our future goals for the dataset regarding the machine learning aim has been given. | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 40%<br>60% |
| 2.1. | Chapter added. | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 50%<br>50% |
| 3.1. | Title added.<br>Description of different feature types has been given. | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 80%<br>20% |
| 3.2. | Chapter added. | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 10%<br>90% |
| 4.1. | Chapter added. | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 60%<br>40% |
| 4.2. | A better description of outlier detection and it's use case has been given. | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 85%<br>15% |
| 4.3. | Chapter added. | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 30%<br>70% |
| 4.4. | Chapter added.<br>A better description of PCA, it's use case, and plots has been given. | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 15%<br>85% |
| 5. | A better fitting discussion has been given. | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 80%<br>20% |
| 6. | Chapter added. | Jonas Rose Lund Pedersen<br>Søren Rønnekær Holgreen Graae | 50%<br>50% |

# Table of Contents

# 1. Introduction

The purpose of this report is to show the students understanding of the subject "Data: Feature extraction, and visualization" for the course **02450 Introduction to Machine Learning and Data Mining**.

This is to be done by finding an appropriate dataset, where feature extraction and visualization will be applied; basic descriptions and plots.

# 2. Dataset: Overview

The Automobile data for this project has been obtained from archive.ics.uci.edu.[1] We've chosen this dataset on the following guideline criteria:

1. At least 60 observations
   These observations should not have any missing or erroneous values.
2. 5 attributes
   At least three of the attributes should be continuous.

This dataset has information on cars from 1985, covering three big ideas. First, it talks about what each car is like, including things like its size, how it's made, and how it works. Next, it discusses how risky a car is for insurance companies, based on its price. Every car gets a starting risk level based on how much it costs. If a car is thought to be riskier or less risky than others, its risk level can go up or down. A car with a +3 rating is seen as risky, while a -3 means it's likely safer.

The third part of the dataset looks at how much money insurance companies typically pay out for damages involving each car in a year, making sure the comparison is fair among different kinds of cars, like small ones, wagons, or sports cars. This shows the average money lost on a car each year.

As a summary, the dataset gathers details about cars, focusing on what makes each car unique, how likely it is to cost more for insurance than expected, and how much money insurance companies typically pay out for it yearly.

Looking at previous uses and results from the data we can refer to a paper done by Haofan Zhang, **Spectral Ranking and Unsupervised Feature Selection for Point, Collective and Contextual Anomaly Detection**[2]. In this paper, H. Zhang researched the opportunity for improving the detection of outliers. By using the dataset seen in this project, H. Zhang uses it to demonstrate how feature selection using Hilbert-Schmidt Independence Criteria (HSIC) feature selection helps optimizing the interpretability of the outlier ranking result - this is out of our skill scope, but relevant for the topic.

While we could not find any specific transformations or calculations done on the data in this paper, there are multiple references to other papers utilizing it as well - some done by H. Zhang. This makes us believe that the data is not just relevant for data manipulation, but also relevant to this course, which should be understood as the opportunity to not just work with but also learn from the dataset regarding the subject 'machine learning'.

[1] Schlimmer, Jeffrey. (1987). Automobile. UCI Machine Learning Repository. https://doi.org/10.24432/C5B01C.
[2] Zhang, Haofan et al. "Spectral ranking and unsupervised feature selection for point, collective, and contextual anomaly detection." International Journal of Data Science and Analytics 9 (2018): 57 - 75.

Given that the dataset consists of insurance prices and vehicle characteristics, it seems natural to use the characteristics to label cars into risk classes based on variables as engine size, horsepower (hp), price, and perhaps the make, to gather insight into whether a certain manufacturers costumers have increased insurance risks (*symboling target*). This would be done using a classifier model, e.g. a Decision Tree- or Logistic Regression (the name being very misleading) model, predicting the class label symboling.

Taking an interest in cars, as both the students do, learning more about how the different characteristics impact each other is a great motivation for a Linear Regression model: In the next project it would be interesting to train a model on features such as fuel, fuels systems, and engine types to predict the fuel economy of the given car, both city- and highway driving - even a characteristic as horsepower would likely vary depending these features.

## 2.1. Manipulation

It's common practice to standardize data before using it for analysis and machine learning due to the fact, that the scale of features can vary heavily, thus confusing the learning model.

Standardizing the data involves finding the empirical mean ($\hat{\mu}$) and empirical standard deviation ($\hat{\sigma}$) of a feature ($X$).[3] This is done using the following formulas:

$$(1) \quad \hat{\mu}_j = \frac{1}{N} \cdot \sum_{i=1}^{N} X_{ij}$$

$$(2) \quad \hat{\sigma}_j = \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^{N} \left(X_{ij} - \hat{\mu}\right)^2}$$

Having found the empirical mean and empirical standard deviation, the transformed feature column can be calculated using the following formula:

$$(3) \quad \tilde{X}_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

To explain it a bit further: The empirical mean, $\hat{\mu}$, is used to normalize the data, meaning, that we take away the scaling problem - the data is centered around zero. Afterwards dividing by the empirical standard deviation, we ensure that the standard deviation of the feature is 1; the data should follow the 68-95-99.7 rule.

Dealing with datasets that contains features which have a large variance in values, it can be beneficial to transform the data logistically.

$$(4) \quad \tilde{X}_{ij} = \begin{bmatrix} \log\left(X_{11}\right) & \cdots & \log\left(X_{1j}\right) \\ \vdots & \ddots & \vdots \\ \log\left(X_{i1}\right) & \cdots & \log\left(X_{ij}\right) \end{bmatrix}$$

[3] Note: We find the empirical statistics due to the fact, that unbiased estimates are considered superior for datasets with a small sample count.

In the section *Dataset: Pre-processing & Visualization*, the effect of transforming the continuous part of our dataset can be seen.

Furthermore, our data contains categorical features. These will need to be encoded - we have chosen one-of-K as our method - for the features to be useful.

# 3. Dataset Features: Detailed explanation

The automobile dataset contains 25 features of different types. Most of the features are continuous, but most of the earlier mentioned types are discreet.

## 3.1. Feature types

Explaining the features further, we will expand on their type, give a description, and note if it contains missing values.

| **Feature** | **Type** | **Reasoning** | **Description** |
|---|---|---|---|
| Engine Size | Continuous, Ratio | An engine size of $0.00\ L$ would mean a non-existent engine. | The engine size can be a value in the range of $es \in [61; 326], es \in \mathbb{Z}$. |
| Horsepower | Continuous, Ratio | A horsepower value of 0 would mean that there is no power. | The horsepower can be a value in the range of $hp \in [48; 288], hp \in \mathbb{R}$.<br><br>Missing values. |
| Make | Discrete, Nominal | The make is a category, it can only be categorized. | The make can be categorized into the following:<br><br>{Alfa Romeo, Audio, BMW, Chevrolet, Dodge, Honda, Isuzu, Jaguar, Mazda, Mercedes Benz, Mercury, Mitsubishi, Nissan, Peugot, Plymouth, Porsche, Renault, Saab, Subaru, Toyota, Volkswagen, Volvo} |
| Price | Continuous, Ratio | Having a price of currency 0, would mean an absence of any price. | The price can be a value in the range of $p \in [5118; 45400], p \in \mathbb{R}$.<br><br>Missing values. |
| Symboling* | Discrete, Ordinal | The symboling factor can be larger than or less than, let us say, 0. | Symboling describes the risk factor that is associated with a car, and therefore the insurance price. It can be in the range:<br>$rf \in [-3; 3], rf \in \mathbb{Z}$ |

| | | | |
|---|---|---|---|
| City-/highway MPG | Continuous, Ratio | An MPG value of 0, would mean that you get 0 miles pr. Gallon, therefore nothing. | MPG describes the miles pr. gallon a given car gets. This can be in the range $mpg \in [13; 49], mpg \in \mathbb{Z}$ |
| Engine Type | Discrete, Nominal | The engine type can be categorized, just like the make. | Engine type can be put into the following categories: {DOHC, DOHCV, L, OHC, OHCF, OHCV, Rotor} |
| Fuel Type | Discrete, Nominal | Fuel type is categorized as well. | Fuel can be either of the following: {diesel, gas}, making it a binary feature. |
| Fuel System | Discrete, Nominal | Fuel system is also categorized. | The fuel system can be either of the following: {1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi} |

*Table 1: Used dataset features. *Symboling is a target, not a feature.*

To keep the section focused on our chosen goal, we have described our features (and target) with somewhat deeper detail. In table 1 we have gone over different types of features:

- Continuous, Ratio
  - Integer
  - Float
- Discreet, Nominal
- Discreet, Ordinal

The dataset consists of the three (five) types above.

As mentioned earlier, categorical features should be (one-of-K) encoded, and looking at table 1, most of our features are categorical.

## 3.2. Summary statistics

When looking at a dataset, the summary statistics can be useful to get a sense of the bigger picture that the individual feature is contributing towards. In this context, it can provide an overview over which cars are most present. Is it Jaguar that dominates the insurance lists, or is it simply an everyday Honda with its four cylinders?[4]

Summary statistics for aspiration
Most occurring: std

| | std | turbo | Total |
|---|---|---|---|
| Count | 168.0 | 37.0 | 205.0 |
| Proportion | 0.82 | 0.18 | 1.0 |

*Table 2: Summary statistics for aspiration.*

Summary statistics for body-style
Most occurring: sedan

| | sedan | hatchback | wagon | hardtop | convertible | Total |
|---|---|---|---|---|---|---|
| Count | 96.0 | 70.0 | 25.0 | 8.0 | 6.0 | 205.0 |
| Proportion | 0.47 | 0.34 | 0.12 | 0.04 | 0.03 | 1.0 |

*Table 3: Summary statistics for body-style.*

[4] To answer the question, it was in 1985 Toyota that ruled the insurance lists; table 9.

Summary statistics for drive-wheels
Most occurring: fwd

| | fwd | rwd | 4wd | Total |
|---|---|---|---|---|
| Count | 120.0 | 76.0 | 9.0 | 205.0 |
| Proportion | 0.59 | 0.37 | 0.04 | 1.0 |

*Table 4: Summary statistics for drive-wheels.*

Summary statistics for engine-location
Most occurring: front

| | front | rear | Total |
|---|---|---|---|
| Count | 202.0 | 3.0 | 205.0 |
| Proportion | 0.99 | 0.01 | 1.0 |

*Table 5: Summary statistics for engine-location.*

Summary statistics for engine-type
Most occurring: ohc

| | ohc | ohcf | ohcv | dohc | l | rotor | dohcv | Total |
|---|---|---|---|---|---|---|---|---|
| Count | 148.0 | 15.0 | 13.0 | 12.0 | 12.0 | 4.0 | 1.0 | 205.0 |
| Proportion | 0.72 | 0.07 | 0.06 | 0.06 | 0.06 | 0.02 | 0.0 | 1.0 |

*Table 6: Summary statistics for engine-type.*

Summary statistics for fuel-system
Most occurring: mpfi

| | mpfi | 2bbl | idi | 1bbl | spdi | 4bbl | mfi | spfi | Total |
|---|---|---|---|---|---|---|---|---|---|
| Count | 94.0 | 66.0 | 20.0 | 11.0 | 9.0 | 3.0 | 1.0 | 1.0 | 205.0 |
| Proportion | 0.46 | 0.32 | 0.1 | 0.05 | 0.04 | 0.01 | 0.0 | 0.0 | 1.0 |

*Table 7: Summary statistics for fuel-system.*

Summary statistics for fuel-type
Most occurring: gas

| | gas | diesel | Total |
|---|---|---|---|
| Count | 185.0 | 20.0 | 205.0 |
| Proportion | 0.9 | 0.1 | 1.0 |

*Table 8: Summary statistics for fuel-type.*

Summary statistics for make
Most occurring: toyota

| | toyota | nissan | mazda | mitsubishi | honda | volkswagen | subaru | peugot | volvo | dodge | mercedes-be | bmw | audi | plymouth | saab | porsche | isuzu | jaguar | chevrolet | alfa-romero | renault | mercury | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Count | 32.0 | 18.0 | 17.0 | 13.0 | 13.0 | 12.0 | 12.0 | 11.0 | 11.0 | 9.0 | 8.0 | 8.0 | 7.0 | 7.0 | 6.0 | 5.0 | 4.0 | 3.0 | 3.0 | 3.0 | 2.0 | 1.0 | 205.0 |
| Proportion | 0.16 | 0.09 | 0.08 | 0.06 | 0.06 | 0.06 | 0.06 | 0.05 | 0.05 | 0.04 | 0.04 | 0.04 | 0.03 | 0.03 | 0.03 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.0 | 1.0 |

*Table 9: Summary statistics for make.*

Summary statistics for num-of-cylinders
Most occurring: 4

| | 4 | 6 | 5 | 8 | 2 | 3 | 12 | Total |
|---|---|---|---|---|---|---|---|---|
| Count | 159.0 | 24.0 | 11.0 | 5.0 | 4.0 | 1.0 | 1.0 | 205.0 |
| Proportion | 0.78 | 0.12 | 0.05 | 0.02 | 0.02 | 0.0 | 0.0 | 1.0 |

*Table 10: Summary statistics for num-of-cylinders.*

Summary statistics for num-of-doors
Most occurring: 4.0

| | 4.0 | 2.0 | Total |
|---|---|---|---|
| Count | 114.0 | 89.0 | 203.0 |
| Proportion | 0.56 | 0.44 | 1.0 |

*Table 11: Summary statistics for num-of-doors.*

Summary statistics for numerics features
(1/2)

| | normalized-losses | wheel-base | length | width | height | curb-weight | engine-size |
|---|---|---|---|---|---|---|---|
| count | 164.0 | 205.0 | 205.0 | 205.0 | 205.0 | 205.0 | 205.0 |
| mean | 122.0 | 98.76 | 174.05 | 65.91 | 53.72 | 2555.57 | 126.91 |
| std | 35.44 | 6.02 | 12.34 | 2.15 | 2.44 | 520.68 | 41.64 |
| min | 65.0 | 86.6 | 141.1 | 60.3 | 47.8 | 1488.0 | 61.0 |
| 25% | 94.0 | 94.5 | 166.3 | 64.1 | 52.0 | 2145.0 | 97.0 |
| 50% | 115.0 | 97.0 | 173.2 | 65.5 | 54.1 | 2414.0 | 120.0 |
| 75% | 150.0 | 102.4 | 183.1 | 66.9 | 55.5 | 2935.0 | 141.0 |
| max | 256.0 | 120.9 | 208.1 | 72.3 | 59.8 | 4066.0 | 326.0 |

*Table 12: Summary statistics for numeric features, first half.*

Summary statistics for numerics features
(2/2)

| | bore | stroke | compression-ratio | horsepower | peak-rpm | city-mpg | highway-mpg | price |
|---|---|---|---|---|---|---|---|---|
| count | 201.0 | 201.0 | 205.0 | 203.0 | 203.0 | 205.0 | 205.0 | 201.0 |
| mean | 3.33 | 3.26 | 10.14 | 104.26 | 5125.37 | 25.22 | 30.75 | 13207.13 |
| std | 0.27 | 0.32 | 3.97 | 39.71 | 479.33 | 6.54 | 6.89 | 7947.07 |
| min | 2.54 | 2.07 | 7.0 | 48.0 | 4150.0 | 13.0 | 16.0 | 5118.0 |
| 25% | 3.15 | 3.11 | 8.6 | 70.0 | 4800.0 | 19.0 | 25.0 | 7775.0 |
| 50% | 3.31 | 3.29 | 9.0 | 95.0 | 5200.0 | 24.0 | 30.0 | 10295.0 |
| 75% | 3.59 | 3.41 | 9.4 | 116.0 | 5500.0 | 30.0 | 34.0 | 16500.0 |
| max | 3.94 | 4.17 | 23.0 | 288.0 | 6600.0 | 49.0 | 54.0 | 45400.0 |

*Table 13: Summary statistics for numeric features, second half.*

On all the tables above, table 2 to table 13, summary statistics for all the features can be seen. For the numeric features, that is table 12 and table 13, we have included[5] the following statistics:

- Count of feature
- Mean of feature
- Standard deviation of feature
- Minimum value of feature
- $25^{th}$ percentile of feature
- $50^{th}$ percentile of feature
- $75^{th}$ percentile of feature
- Maximum value of feature

Looking at the count statistic, we can effectively pinpoint which features have missing values, and, for good reason, all the noticeable features are also present in the missing values bar graph mentioned earlier. Later, the observations with missing values, in a relevant feature, will be dropped. This is a choice, that is different from filling in the missing values - we do not see this providing any value to the dataset, as it might just end up as oddly placed values.

Including the mean and median ($50^{th}$ percentile) in the statistics, can also make it possible to spot outliers early on. In our case, there is no apparent reason to suspect outliers in this stage of the data analysis, given that there is no feature in which the mean and median are far enough from each other.

For the categorical features, that is table 2 to table 11 we have included two statistics:

- Count of category
- Proportion of category

The count shows have many times a given category is present in a given feature. E.g., how many cars are equipped with four doors. If the count does not add up to 205, there is missing values. The proportion statistic shows the proportion of a given category in a given feature. E.g., the percentage of cars with four doors.

[5] Using the pandas 'df.describe()' method, the mentioned statistics have been found. We have not made a conscious choice to include all of them; some of them are redundant for the purpose of this report.

# 4. Dataset: Pre-processing & Visualization

## 4.1. Distribution

To get insight into the distribution of our dataset, we plotted histograms of all the numerical features, without any manipulation.

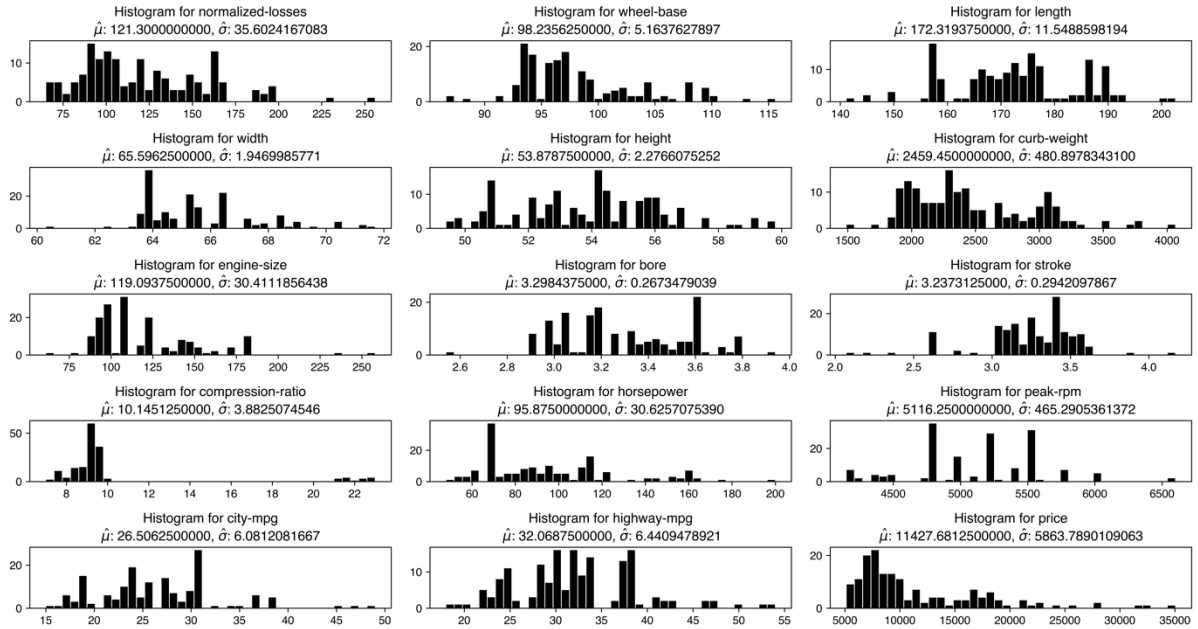



*Figure 1: Histogram of each original numerical feature.*

In the above figure, figure 1, a histogram for each numerical feature has been plotted. At this stage, the missing values have been removed.

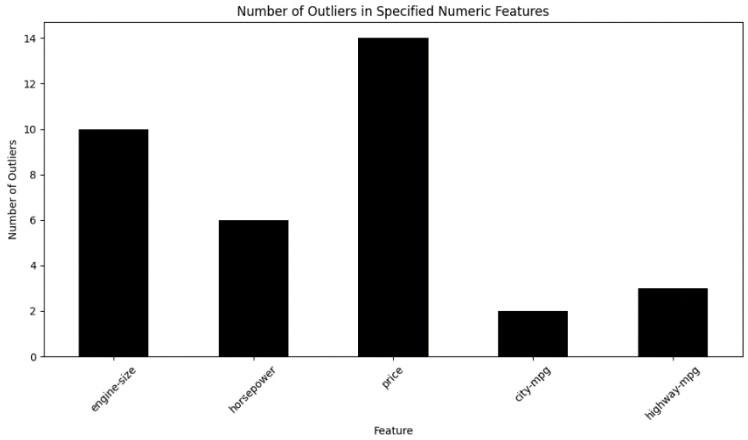
None of the discreet features are plotted. This is due to the fact, that the plotting of their distribution won't be meaningful nor provide any useful insight.

As it is obvious to see, the data is not sorted in any way, shape, or form; the scales are not standardized and the data is not normal distributed, as can be expected from a real-world dataset. It is also noticeable, that there are a few data points that could be outlier-like. Unfortunately, we cannot make any conclusion based on these plots, yet.

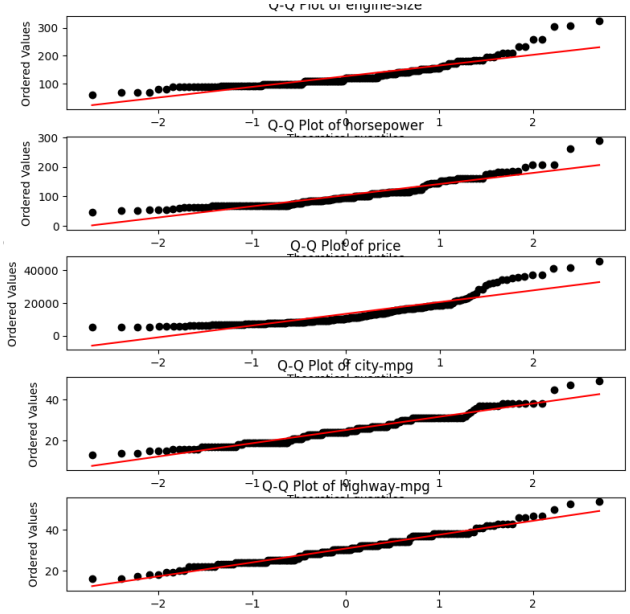
## 4.2. Outliers

For finding outliers we use Interquartile Range (IQR). IQR is calculated by specified numeric features. IQR is the difference between third quartile and first quartile. So, the outliers are detected if any of the data points are beyond 1.5 times below the first quartile or above the third quartile then it is considered an outlier. Another method is looking at the Histograms and Q-Q plots.

The IQR method provides a quantitative way to find data points as outliers based on their numeric value where the histograms and Q-Q plots offers a visual approach to identifying outliers.




We will examine the outliers to determine whether they are the result of data entry errors or rare events. Upon reviewing our dataset, the outliers did not appear to be data entry errors but rather rare events. This is because the dataset includes luxury cars, which typically have larger engines, explaining the high horsepower and elevated prices. We know this because the values are not unreasonably high.




As observed in the Q-Q plots, there are no values that are unreasonably off, confirming our theory that the outliers represent rare events caused by luxury or sports cars.

## 4.3. Transformation: Logistical- and standardized

Before doing a principal component analysis, it is common practice to transform the data.

In our case, we have logistically transformed it following equation (4), as well as standardized it following equation (1), (2), and (3).[6] This ensures that we rid the dataset of any heavy scaling problems, and we make sure that the data has a standard variation of 1 and is zeroed; the mean of all features is zero.

[6] The standardization calculations have been done using the Pythons built-in methods ‘mean()’ and ‘std()’.

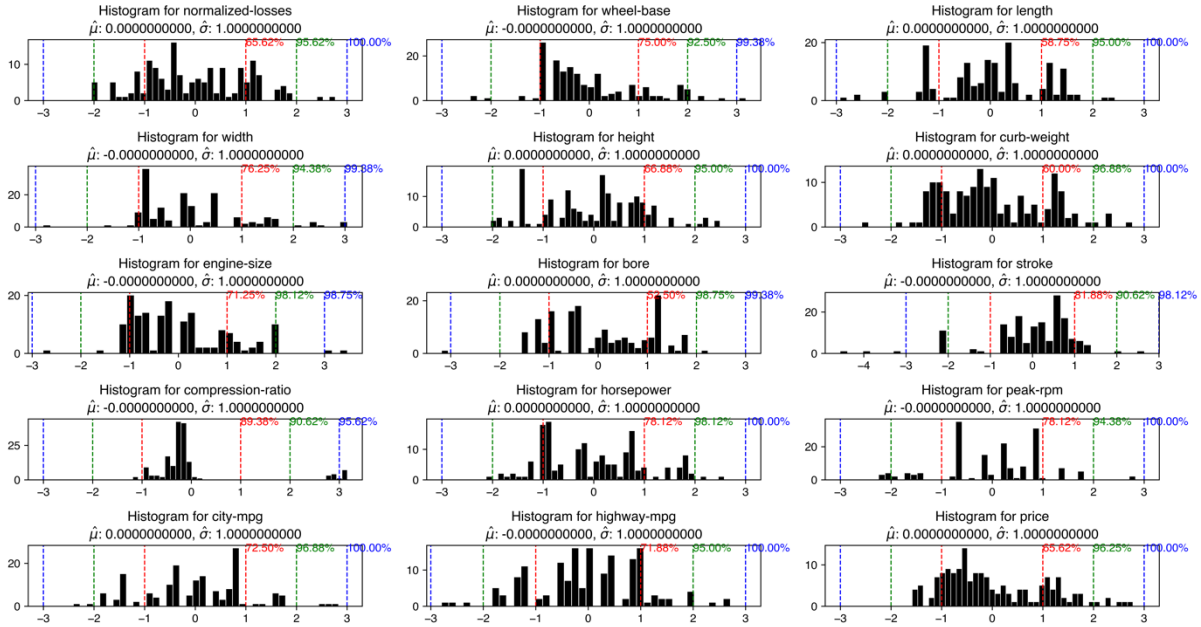



*Figure 2: Histogram of each transformed numerical feature.*

As can be seen above, in figure 3, the data has been standardized and scaled, showing a few data points which value is greater than three times the standard deviation. This could indicate outliers - but not necessarily! Some cars are known for their huge engines, and thus horsepower that comes along, their great fuel efficiency, and their expensive price tags. This has already been clarified in the section *Outliers* though.

As mentioned in an earlier section, the standardization of the data should ensure that the data follows the 68-95-99.7 rule.

Out of curiosity we have plotted lines and added annotations for each standard deviation in the histograms in figure 3, where we can see that it somewhat follows the rule for all features. Some of the features do show a significant detachment from the rule. This can be expected on a dataset like our own, due to the natural differences that occurs when different manufactures produce different cars. This will develop outlier-like data points.

Besides the numerical values, we also have the categorical. For these to be useful for a machine learning model, they have been encoded, as earlier mentioned as a necessary pre-processing step.

## 4.4. Principal Component Analysis

With the data transformed, it is possible to perform a principal component analysis (PCA). Performing a PCA will help us reduce the dimensionality of our dataset, while keeping a certain variety. Before optimizing the dataset for machine learning, a PCA allows us to view the explained variance of each principal component, making it possible to get an idea about how much we can reduce the dimensionality of the dataset.

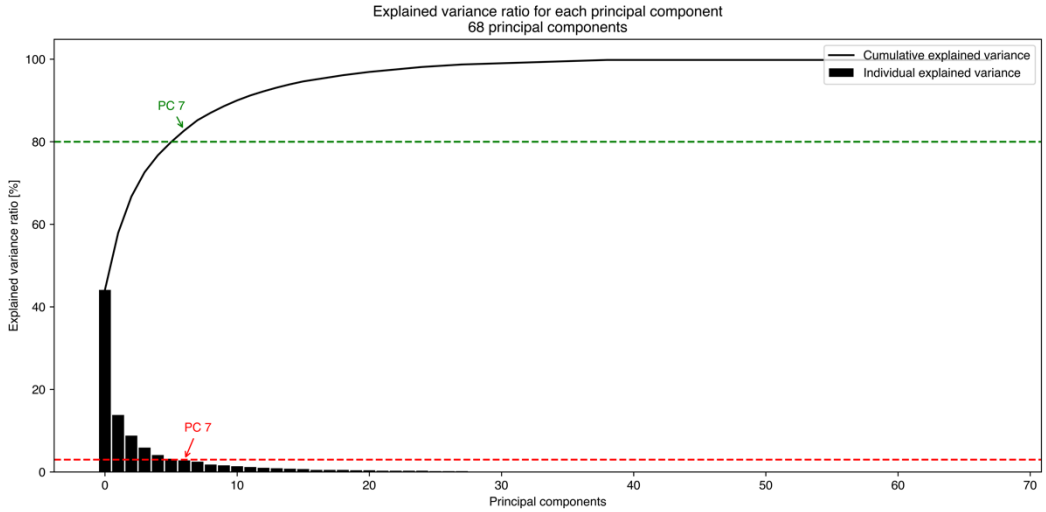



*Figure 3: Visualized PCA of dataset.*

Figure 4 above shows the explained variance gathered from our PCA. Looking at the figure, it is noticeable that it takes quite a few principal components (PC) before the total explained variance reaches an acceptable point of 80%. This is marked with the *green* arrow-annotation, 'PC 7'. We have decided on a threshold of 80%, given the context of our dataset; assuming that an insurance company wouldn't want to lose a valuable chunk of variance. We do also have to keep in mind, that we do not want to keep any principal component that cannot explain a justifiable proportion of the variation, which has a cut-off index of 7, the *red* arrow-annotation, where the variance explained drops below 3%, meaning that PC 1 to (and including) PC 6, each has an explained variance of $\sigma^2 > 0.03$. We thereby declare that PC 1 through PC 6 will be our choice of components, given that the total explained variance reaches a value reasonably close to our 80% threshold.

Although six PCs are still too many for a proper visualization, the reduction from 68 to 6 PCs (~91% reduction) is perfectly reasonable, and also very useful, for training models on the dataset in the future.

Besides looking at the explained variance, we have plotted the direction of the first seven PCs as well.

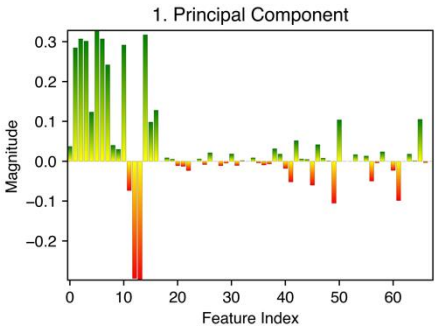



*Figure 4: PC 1 with its feature magnitudes and directions plotted as bars.*

Looking at figure 5, we can see PC 1 plotted as bar graphs with gradients. The gradient indicates the direction as well as the magnitude of the relationship between feature and PC.

Plotting the PCs contributes towards the overview of the features and their contribution to the variance. We can see in appendix 1, that generally feature 20 through 30 have little to no influence on the final PC value. This essentially means that these features could at a future point of development, be removed

from the dataset, considering that they do not contribute very much, compared to some other features. Obviously, this depends on what the model is being trained to predict, and which features have been specified as training features. It can also be seen that the features around 10, generally have quite an influence, so these would generally not be considered for removal.

As a final analysis of this section, we have also plotted the two first PCs against each other.

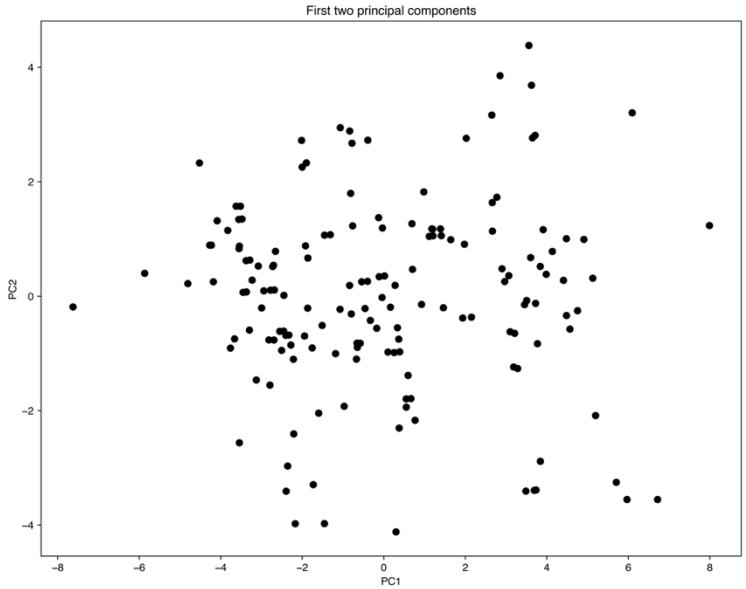



*Figure 5: Scatter plot of PC1 and PC2.*

Figure 6 shows a scatter plot of our two first principal components. Noticeably, there's a lack of clear clusters in the plot. This can be explained by the total explained variance of the two components. They reach just about 50% of the explained variance when used together, as seen in figure 4, meaning that the other components still contain 50% of the variance, which could contain information that would reveal clusters.

What is visible from figure 6 though, is that a standardization of the data depicted has been made, as the data has an origin of zero. In our case, we already knew that the data had been transformed, but if in some cases it can be valuable information. Given our dataset, it might be reasonable that there aren't any apparent clusters, although this could be up for discussion, dependent on the reader.

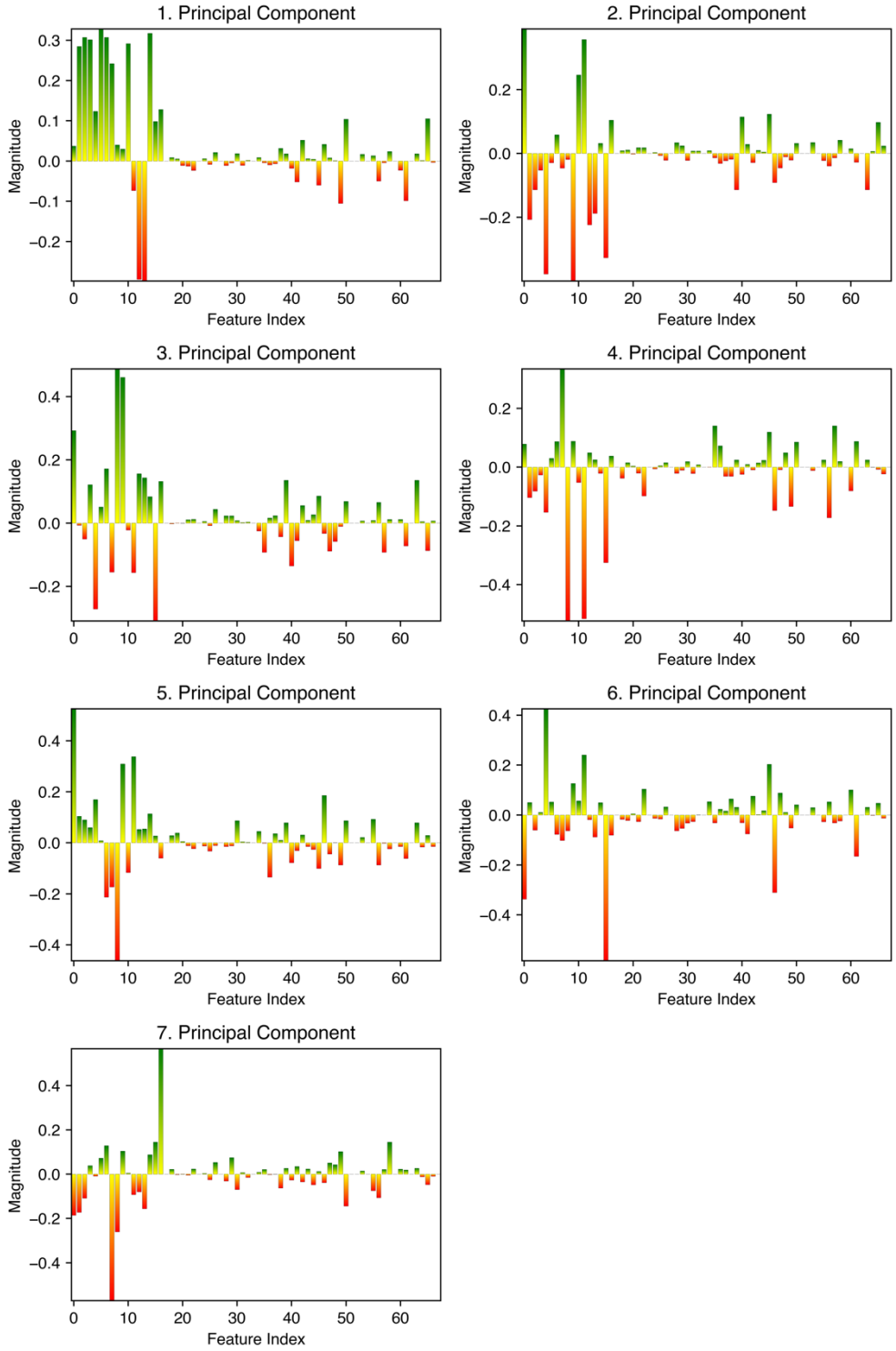
# 5. Discussion

It is unclear whether our primary machine aim is feasible when considering our scatter plot in figure 6 but given the fact that the explained variance increases quite a bit until PC 7, we conclude that it's reasonable to say that the training of machine learning models is definitely possible on this dataset.

From the analysis of our dataset, we have learned a multitude of things, enabling us to describe the dataset further in a future report - furthermore we can confidently already consider which methods could be used in training models on the dataset, and what sort of pre-processing is absolutely necessary for the training to be useful, e.g., standardization.

# 6. Appendix

## 6.1. Appendix 1: Principal Components



# 7. Exam Problems

## 7.1. Question 1, Spring 2019, question 1

Given that the variable Time of Day is a different type in each answer, we can answer by looking at this one variable. Time of Day is a 30-minute interval between to timeslots (i.e. 17.00 to 17.30), and based on this information, we would say that Time of Day is nominal; it belongs to a category ($x_1 = 1$: 07.00 to 07.30, $x_1 = 2$: 07.30 to 08.00, etc.).

Therefore, the **answer A is correct**.

## 7.2. Question 2, Spring 2019, question 2

We start off by checking whether the answer A is correct, as that is the simplest calculation:

$$d_{p=\infty}(x_{14}, x_{18}) = \max\{x_{14}\} - \max\{x_{18}\} = 26 - 19 = 7$$

Given the above calculation, **answer A is correct**.

## 7.3. Question 3, Spring 2019, question 3

The data from question 1 is being used in this question as well but transformed with a singular value decomposition. We've been given the matrix **S** containing the singular values.
We convert these to eigenvalues:

$$\sigma_1^2 = 13{,}9^2 = 193{,}21$$
$$\sigma_2^2 = 12{,}47^2 = 155{,}50$$
$$\sigma_3^2 = 11{,}48^2 = 131{,}79$$
$$\sigma_4^2 = 10{,}03^2 = 100{,}60$$
$$\sigma_5^2 = 9{,}45^2 = 89{,}30$$

This gives the total variance $\sigma^2 = 670{,}40$. Checking whether answer A is correct we calculate the variance for the first four principal components:

$$\sigma^2_{[1;4]} \frac{193{,}21 + 155{,}50 + 131{,}79 + 100{,}60}{670{,}40} = 0{,}82$$

Since answer A states that the result should be greater than 0,8, we can say that **answer A is correct**.

## 7.4. Question 4, Spring 2019, question 4

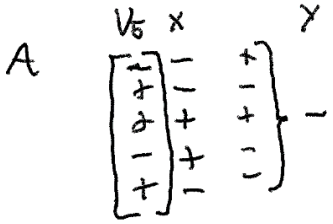



*Figure 6: Illustration of solving answer A*

Finding the answer for this question, we setup a table like the one in figure 1, for each of the answers. $v_5$ is the principal component analysis for $x_5$. The column contains the numeral sign of each value in the vector. The column x shows the numeral sign of the given statements in each answer. As in answer A, it is said that a low value of Time of Day, making the numeral sign negative. Then we multiply the signs, at the end revealing the overall positive or negative effect, y.
After doing this we conclude that **answer D is correct.**