Data Scientist

Tim Hellemans

May 28, 2023

Contents

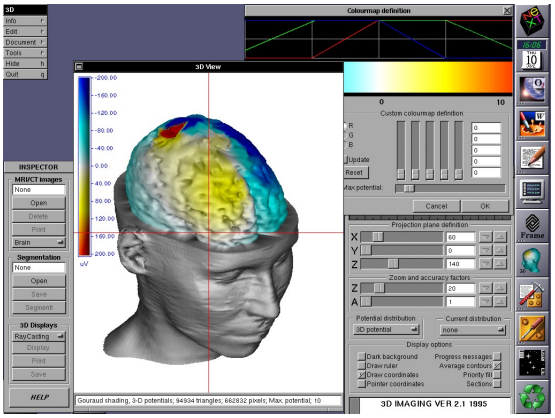

Figure 1: Example of data visualization in the medical sector.

1 Visualizing your results

In this chapter, we take a deeper look into data visualization. We start our discussion by motivating why data visualizations are important and a general remark on which type of visualizations you should use for which audience (something we regularly refer back to throughout the whole chapter). We then go through a list of the most common categories of data visualizations, these categories are:

- Statistical charts.
- Geospatial visualization.
- Time series visualization
- Network visualization.
- Hierarchical visualization.
- Interactive visualization.
- Flow visualization.
- Multidimensional visualization.
- Financial & business visualization.

Remark. These are *categories* of data visualizations, each category is further subdivided into many more types. Moreover, there are still many (often topic specific) visualizations that we will not touch on (e.g. medical visualization, see Figure 1). However, in most cases, you will be able to create an appropriate visual using one presented in this work, but you should always remain on the lookout for new & alternative ways to visualize your data!

1.1 Motivation

In the realm of data science, information is abundant, but knowledge is priceless. Data visualization acts as the brushstroke that brings life to raw data, transforming it into a vivid and intuitive masterpiece. By leveraging the power of visual representation, data visualization enables us to unlock the true potential of our data, making complex concepts more accessible, patterns more discernible, and insights more compelling. In this section, we explore the options one has in visualizing data. As a data scientist/data analyst/machine learning engineer or whichever role in data, we spent days and days exploring data, developing models, testing hypotheses,... We are passionate about this work and appreciate the beauty in being able to predict the total number of passengers on a Boeing 747 with up to 98% accuracy. However, in the end, all our hard work must generate *business value*. In order for this to happen, people must be made aware of our work, its value and applicability. This mostly boils down to having many meetings and presentations where you want to create a mutual understanding of what the data tells us and how this story can help us create business value.

Example. We often need to create schedules in order to process a certain number of people/things in a day (e.g. number of passengers at the airport, number of visitors arriving to a theme park, number of orders coming in at Amazon, ...). Let's look at the example of looking at the expected number of passengers at the different locations of an airport. We want to generate a quick overview of these numbers while still displaying all information a decision maker needs. The easiest way to do this is simply by putting all numbers for a specific day into a table, a quick and simple way to make this table easier to read without losing information is to round all numbers to the closest integer and using a heatmap to make high values stand out more.
We clearly observe that there are 2 peaks in number of passengers. Given the data, one could supplement this heatmap with historic information on these peaks and the amount of personnel that was required to handle those peaks. This can then be weighed against the peaks we see in this heatmap.

There are many reasons why data visualization is important, some of them are:

- **Clarity in Complexity**: Data, in its raw form, often appears as an intricate web of numbers, symbols, and patterns.

	Parking	Check in	Screening	Gate
00:00	210.5263	212.766	217.3913	0
01:00	210.5263	212.766	217.3913	227.2727
02:00	210.5263	212.766	217.3913	227.2727
03:00	210.5263	212.766	217.3913	227.2727
04:00	210.5263	212.766	217.3913	227.2727
05:00	421.0526	212.766	217.3913	227.2727
06:00	631.5789	425.5319	217.3913	227.2727
07:00	1684.211	638.2979	434.7826	227.2727
08:00	1684.211	1702.128	652.1739	454.5455
09:00	1684.211	1702.128	1739.13	681.8182
10:00	1052.632	1702.128	1739.13	1818.182
11:00	842.1053	1063.83	1739.13	1818.182
12:00	631.5789	851.0638	1086.957	1818.182
13:00	421.0526	638.2979	869.5652	1136.364
14:00	421.0526	425.5319	652.1739	909.0909
15:00	421.0526	425.5319	434.7826	681.8182
16:00	1052.632	425.5319	434.7826	454.5455
17:00	1263.158	1063.83	434.7826	454.5455
18:00	1473.684	1276.596	1086.957	454.5455
19:00	1684.211	1489.362	1304.348	1136.364
20:00	1684.211	1702.128	1521.739	1363.636
21:00	842.1053	1702.128	1739.13	1590.909
22:00	631.5789	851.0638	1739.13	1818.182
23:00	421.0526	638.2979	869.5652	1818.182

	Parking	Check in	Screening	Gate
00:00	211	213	217	0
01:00	211	213	217	227
02:00	211	213	217	227
03:00	211	213	217	227
04:00	211	213	217	227
05:00	421	213	217	227
06:00	632	426	217	227
07:00	1684	638	435	227
08:00	1684	1702	652	455
09:00	1684	1702	1739	682
10:00	1053	1702	1739	1818
11:00	842	1064	1739	1818
12:00	632	851	1087	1818
13:00	421	638	870	1136
14:00	421	426	652	909
15:00	421	426	435	682
16:00	1053	426	435	455
17:00	1263	1064	435	455
18:00	1474	1277	1087	455
19:00	1684	1489	1304	1136
20:00	1684	1702	1522	1364
21:00	842	1702	1739	1591
22:00	632	851	1739	1818
23:00	421	638	870	1818

Figure 2: We show the number of passengers passing by a number of locations at the airport per 1 hour slot. On the left numbers are shown without and on the right with heatmap.

Unveiling its inherent complexity can be a daunting task. However, data visualization rises to the challenge, offering a visual narrative that simplifies the complexity and enhances our understanding. By transforming abstract numbers into intuitive charts, graphs, and infographics, data visualization provides clarity and reveals the underlying story within the data. It enables us to grasp intricate relationships, spot trends, and identify anomalies at a glance, empowering us to make more informed interpretations.

- **Insightful Communication**: Data holds tremendous insights, but these insights are of little value if they remain hidden within spreadsheets and databases. Data visualization acts as a bridge between the data and its audience, facilitating effective communication of information, ideas, and discoveries. By presenting data in visually appealing and interactive formats, it engages the viewer's attention and facilitates a deeper understanding of the underlying message. Whether it's presenting research findings, conveying business metrics, or explaining scientific concepts, data visualization empowers us to communicate complex ideas with clarity, brevity, and impact. There are many meetings in life that could be replaced by a couple of good visualizations.
- **Decision-Making Catalyst**: In a world driven by data, making informed decisions is critical. Data visualization plays a pivotal role in this process by transforming data into actionable insights. By presenting data in visual formats that highlight key trends, comparisons, and patterns, it enables decision-makers to discern relevant information swiftly. Interactive dashboards and real-time visualizations further enhance the decision-making process by providing dynamic and up-to-date views of critical metrics. Whether it's optimizing business strategies, improving operational efficiency, or identifying opportunities, data visualization acts as a catalyst, empowering decision-makers to act with confidence and precision.
- **Driving Innovation**: Innovation thrives on exploration and the ability to perceive connections where others see chaos. Data visualization fuels this spirit of innovation by enabling us to spot patterns, identify gaps, and make creative leaps. By visualizing data from diverse sources and perspectives, it fosters interdisciplinary collaboration, encouraging experts from different domains to contribute their unique insights. Visualization tools, such as network graphs and geographic mapping, unravel hidden relationships and expose untapped opportunities. With data visualization as our guide, we navigate uncharted territories and push the boundaries of what is possible.

1.2 Some general (unwritten) rules in data visualization

Some golden rules which you should always keep in mind when creating data visualizations.

- Understand your audience and objectives: Clearly define your target audience and the purpose of the visualization. Understand what insights or messages you want to convey and tailor your visualization accordingly.
- Choose the appropriate visualization type: Select the visualization type that best suits your data and goals. Consider factors such as the data structure, relationships, and the type of analysis you want to perform.
- Keep it simple and uncluttered: Avoid visual clutter and unnecessary complexity. Simplify your visualizations by removing irrelevant elements and focusing on the key information. Strive for a clean and minimal design that enhances clarity and readability.
- Use color strategically: Choose colors purposefully to highlight important information, create visual contrast, and guide the viewer's attention. Maintain consistency and ensure that color choices are accessible and meaningful. Consider color-blindness and use color palettes that accommodate different types of color vision. In general: only

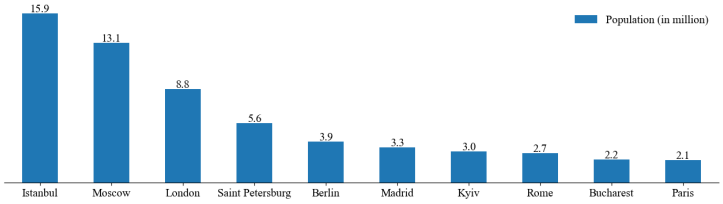

Figure 3: Example of data visualization in the medical sector.

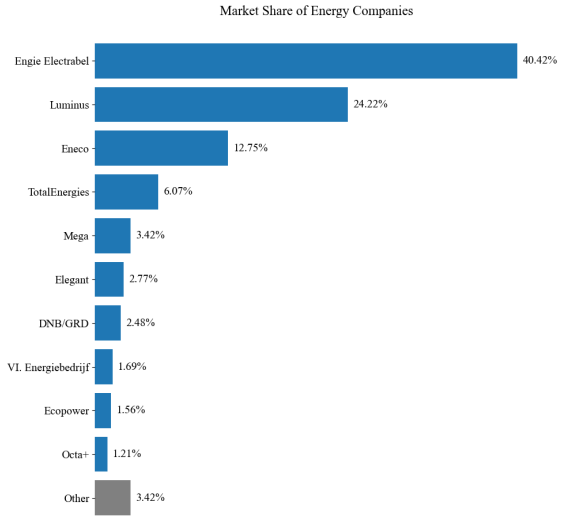

Figure 4: Example of some data visualization in the energy sector.

add additional colors when you feel like this provides additional information for your audience (otherwise the color is a form of clutter).

- Ensure appropriate scaling: Set the scale of your visualization appropriately to accurately represent the data. Avoid distorting or misleading interpretations by using appropriate axis scales and starting axes from zero, unless there is a specific reason to do otherwise.
- Label clearly and provide context: Labels are crucial for conveying information accurately. Clearly label data points, axes, and any relevant components of the visualization. Provide additional context through captions, titles, or annotations to help viewers interpret the data correctly.
- Use appropriate data order and grouping: Arrange the data in a logical order to facilitate understanding. Group related data together to highlight patterns or comparisons. Consider sorting data based on magnitude, time, or any other relevant factor to support meaningful analysis.
- Incorporate interactive elements (if applicable): If your visualization is interactive, provide intuitive controls or tooltips that allow users to explore the data and gain deeper insights. Interactivity can enhance engagement and enable users to interact with the visualization according to their specific interests or queries.
- Validate and iterate: Test your visualization with a sample audience or colleagues to gather feedback and identify areas for improvement. Iterate and refine your visualization based on the feedback received to enhance its effectiveness.

1.3 Statistical charts

When it comes to displaying statistical charts, one crucial decision is whether to use total numbers or distributional numbers. Both approaches have their merits and can provide valuable insights depending on the context and purpose of the chart. Let's explore each option further.

Using total numbers in statistical charts involves presenting the absolute values or counts of a particular variable. This approach offers a straightforward representation of the data and allows viewers to understand the actual quantities involved. Total numbers are particularly useful when the goal is to convey the magnitude or scale of a phenomenon. For example, if you're comparing the population sizes of different cities, displaying the total numbers can help in easily identifying which cities have larger or smaller populations (see Figure **??**).

On the other hand, distributional numbers focus on the proportions or percentages within different categories or groups. Instead of presenting raw counts, this approach highlights the relative distribution of values across the data set. Distributional numbers are valuable when the emphasis is on understanding patterns, trends, or relationships between different groups. For instance, if you're analyzing the market share of various companies within a specific industry, using

distributional numbers can reveal the relative dominance or competitiveness of each player. An example of this is show in Figure **??**.

The choice between total numbers and distributional numbers depends on the specific objectives and audience of the statistical chart. If you aim to provide a comprehensive view of the data and prioritize a clear understanding of the overall size or quantity, total numbers may be more appropriate. On the other hand, if your intention is to highlight the relative proportions or share of different components within a dataset, distributional numbers can offer a more insightful representation.

It's worth noting that in many cases, a combination of both approaches can provide a well-rounded depiction of the data. For instance, you can use a chart that presents the total numbers as well as accompanying percentages or proportions within each category. This approach enables viewers to grasp the absolute values while also understanding the relative significance of different groups.

Ultimately, the choice between total numbers and distributional numbers in statistical charts should align with the goals of your analysis and the message you want to convey. By carefully considering the context, audience, and objectives, you can select the most suitable approach to effectively communicate your data-driven insights.

When displaying the values in a distributional sense, you should first be aware of the question whether the distribution your data is coming from is discrete or continuous. Let us first dive into the details on what each type is and how these type of distributions are defined.

1.3.1 Random variables

A random variable is a mathematical concept used in probability theory and statistics to represent an uncertain quantity or outcome. It is a variable whose value is determined by the outcome of a random event, such as the roll of a die, the flip of a coin, or the measurement of a physical quantity subject to random variation.

Formally, a random variable is defined as a function that assigns a numerical value to each possible outcome of a random experiment. The set of all possible values that a random variable can take is called its sample space, and the probability distribution of the random variable describes the likelihood of each value occurring.

There are two types of random variables: discrete and continuous.

Discrete Random Variables

A discrete distribution is a probability distribution that describes the probabilities of a random variable taking on specific values. In other words, it represents the likelihood of each possible outcome in a discrete set of values. Unlike continuous distributions, which can take on any value within a range, discrete distributions have a countable set of possible outcomes. These values can either be categorical (such as companies, gender, ...) or discrete numbers (such as the number of eyes on a die, daily production of smartphones, ...). These distributions are defined by assigning a probability to each possible outcome.

Remark. In the example of Figure **??**, one could argue that the random variable we are looking at is *total population* and it takes on the given values for the different cities. Alternatively, if we were only interested in the population of Madrid, we could define our random variable as the population of Madrid and the value of 3.3 million is one observation from this value at a specific point in time. We could then argue that it might be of interest to visualize this total value as a function of time and look at that distribution.

Continuous Random Variables

A continuous random variable can take any value within a certain range or interval. For example, the height of a person, the time it takes for a car to travel a certain distance and the market share of Engie at different points in time are continuous random variables.

1.3.2 Bar charts

Bar charts are commonly used for visualizing categorical or discrete data and displaying comparisons between different categories or groups. They are particularly effective in representing data that can be organized into distinct categories or groups, such as survey responses, sales figures, population distribution, and frequency counts.

Here are some common uses of bar charts:

- **Data Comparison**: Bar charts are ideal for comparing the values or frequencies of different categories. The length of each bar represents the quantity or magnitude of the variable being measured, allowing for easy visual comparison. This makes bar charts useful for analyzing data trends, identifying patterns, and drawing insights.
- **Data Distribution**: Bar charts can show the distribution of data across categories or groups. Each category is represented by a separate bar, and the height of the bar indicates the frequency, count, or proportion associated with that category. This helps in understanding the distributional characteristics of the data.
- **Data Representation**: Bar charts provide a clear and concise way to present data to an audience. They are easy to understand and interpret, making them suitable for conveying information in presentations, reports, and publications.
- **Data Exploration**: Bar charts can be used as an exploratory tool to identify patterns, outliers, or anomalies in the data. By visualizing the data in a bar chart, you can quickly spot variations or discrepancies between different categories, which may lead to further investigation or analysis.
- **Data Comparison over Time**: Bar charts can also be used to compare data across different time periods. By

arranging bars chronologically, you can observe changes or trends in the data over time. This is particularly useful for tracking progress, analyzing historical data, or identifying seasonal patterns.

Overall, bar charts provide a simple and effective way to display and analyze categorical data, making them a popular choice for data visualization and communication purposes. As such, a number of different types of bar charts have been developed. Some general tips to keep in mind when creating a bar chart:

- Choose appropriate colors: It's recommended to use the same colors for similar categories while using contrasting colors between sufficiently different categories. Avoid using colors that may convey unintended meaning or bias. In the example of Figure 4, we used a single color for all actual companies and we *greyed out* the "Other"category.
- Maintain a logical data order: Arrange the bars in a logical and meaningful order to facilitate easy comprehension. Typically, it is best to sort the categories in ascending or descending order based on the variable being measured. This helps in identifying patterns, trends, or outliers more effectively.
- Avoid excessive clutter: Keep the design clean and uncluttered by removing unnecessary elements that may distract or confuse viewers. Minimize gridlines, borders, or background patterns if they do not add value to the chart. Focus on presenting the data in a clear and straightforward manner.
- Use appropriate scales: Ensure that the scale of the axes is appropriate for the data being represented. Avoid distorting the perception of data by using uneven or misleading scales. Start the y-axis (vertical axis) from zero unless there is a specific reason to use a different baseline.

Vertical bar charts

This is the most common type of bar chart, where the bars are displayed vertically along the y-axis. Each bar represents a category or group, and the length of the bar corresponds to the value or frequency associated with that category. Vertical bar charts are useful for comparing data across different categories or groups. We already had an example of this type of bar charts in Figure 3. Some things to keep in mind when creating this type of bar chart

1.3.3 Horizontal bar charts

In a horizontal bar chart, the bars are displayed horizontally along the x-axis. Like vertical bar charts, each bar represents a category or group, and the length of the bar indicates the value or frequency. Horizontal bar charts are often used when the category labels are long or when there are many categories to display.

1.3.4 Grouped bar charts

A grouped bar chart displays multiple sets of bars side by side, allowing for easy comparison between different groups. Each group represents a category, and within each group, there can be multiple bars representing subcategories or different variables. Grouped bar charts are useful for comparing values across multiple categories simultaneously. These can of course be either vertical or horizontal.

1.3.5 Stacked bar charts

In a stacked bar chart, the bars are stacked on top of each other, with each segment of the bar representing a subcategory or component. The total height of the bar represents the cumulative value or frequency of all the subcategories. Stacked bar charts are suitable for showing the composition or breakdown of a variable across different categories.

1.3.6 100% stacked bar charts

Similar to a stacked bar chart, a 100% stacked bar chart also displays bars stacked on top of each other. However, the height of each bar is normalized to represent the relative proportion or percentage of each subcategory within the total. This type of chart is useful for comparing the relative contribution of different subcategories across categories.

1.3.7 Clustered bar charts

A clustered bar chart displays bars grouped together in clusters, with each cluster representing a category. Within each cluster, there can be multiple bars representing different variables or subcategories. Clustered bar charts are suitable for comparing values between different categories and within each category.

2 Statistics for Data Science

Insert drawings to explain math in data science and the whole pipeline pointing at where you use which math/skills.
Thus far, we have mainly focused on the technical skills that allow you to work with data, this includes:

- Basic python to think programmatically (if/for/while loops, variables, errors, ...).
- We zoomed in on the python you need to work with data specifically; this includes working with files/folders on your computer but also to connect to an external database, scrape information from a website etc..
- We then focused on tabular data, this is mainly done with Pandas, a python package which allows you to quickly combine data from multiple tables and process data in a table.
- We also looked at numpy which contains functionality specifically to work with tensor data, these include 1 dimensional *vectors*, but also 2 dimensional *matrices* and we also looked at three dimensional tensors in the form of RGB images.

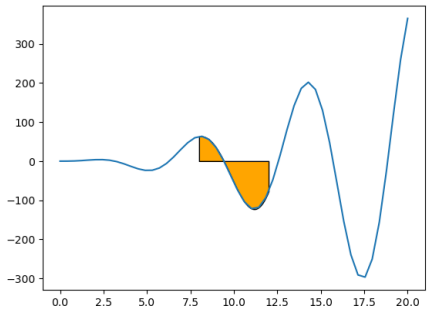

Figure 5: Integral $\int_8^1 2\sin(x) \cdot x^2 \, dx$ visualized.

This chapter concludes the more technical part of the course and from here on out, in this second part of the course, we are going to focus more on the conceptual side and less on the technical. We will still have many exercises in Python and you are still expected to be able to "do"everything we discuss in Python (or any other programming language) but the technical implementation will no longer be the focal point. We start this chapter by providing an overview of the position of math in the data scientist's workflow.

In the subsequent sections, we introduce basic statistics that lie at the heart of data science. These contain fundamentals of statistics (such as the definition of a probability density function, how to visualize probability distributions etc.), but also some classical statistics results that are often (mis)used in the real world. One of the most important results in statistics is the central limit theorem. This theorem is often interpreted as *you can model any reasonable random variable as a normal distribution*, but this is not at all the case! Simply being able to spot when the normal distribution is being misused can already add a lot of value to a company (see also section ref naar sectie CLT).

2.1 Required math for data science

For many aspiring data scientists, a good foundation in math may seem daunting. However, when starting to learn data science/machine learning, it is not necessary to immediately delve into learning all aspects of mathematics. While mathematics plays a fundamental role in understanding the underlying concepts of machine learning, it can be overwhelming to tackle all math topics at once. A few reasons why you shouldn't start with learning all math for data science are:

1. **Steep learning curve**: Mathematics can be complex, and diving into advanced topics without a solid foundation can make it difficult to grasp the concepts. It's better to approach mathematics progressively, starting with the basics and gradually building up your knowledge as you gain more experience with machine learning.
2. **Practicality**: Machine learning frameworks and libraries provide high-level abstractions that allow practitioners to apply machine learning techniques without extensive mathematical knowledge. These tools offer pre-implemented algorithms and functions, making it easier to get started and achieve meaningful results without delving into the underlying mathematics.
3. **Focus on application**: Initially, it's more important to understand the practical aspects of machine learning, such as data preprocessing, model selection, and evaluation. By focusing on these areas, you can gain hands-on experience and develop intuition about how machine learning works in real-world scenarios.
4. **Learning by doing**: Machine learning is a highly practical field. Engaging in projects and implementing machine learning algorithms will help you understand the core concepts and their mathematical underpinnings more effectively. This hands-on approach will allow you to see the direct application of mathematics in machine learning.

Remark. The approach we will be taking is the learning by doing, that is, we provide the math required as we come across it. Moreover, this explanation will be on the intuition rather than arithmetics. For example:
When someone asks you to compute $\int_8^1 2\sin(x) \cdot x^2 \, dx$ and *explain* how you did it, I do not expect you to be able to take out pen and paper and be able to exactly compute this integral. You should be able to (with the help of google) use Python to compute this value through the function `np.trapz`. Ideally, you should also be able to plot the function $f(x) = \sin(x) \cdot x^2$ and explain that the area under the curve on the interval $[8, 12]$ corresponds to the integral that this person is inquiring about (see also Figure 5). Ideally, you would also be able to explain how this integral comes into existence as the limit of the surface area of histograms with an increasing number of bins.

The part of Data Science for which one requires the most math is typically when developing a machine learning model. Typical projects consist of (a selection from) the following steps:

1. Collect the data from different sources (no math).
2. Merge and clean the data from all sources into a compatible format (almost no math).
3. Slice, dice and preprocess the data to be ready to be fed into some statistical tool (e.g. a machine learning algorithm). This mostly requires some linear algebra which is the mathematical field of doing computations on tensors.
4. Explore the data using standard data analysis tools (requires some probability/statistics knowledge).

5. Select and execute an appropriate method to tackle the given problem. Here the underlying probability theory can be regarded as the beating heart of the method. In order to understand what you are doing and what results you are getting, you will always need a basic level of probability/statistical knowledge. In order to truly understand what is happening under the hood, you will need a lot of deep probability theory knowledge. This should not be your goal at this point.
6. Calculus and optimization come into play when we have selected a model and we now want to train our model, that is we want to tweak our conceptual model to our specific case. This means that we want to find the *optimal* parameters for our problem setting and this in turn requires some basic understanding of calculus and optimization theory in order to *kind of* know what you are doing and deep knowledge of these subjects to truly understand it (which, again, is not the goal at this point in time).

Remark. In Sections 2.6 and 2.7, we will look at our first machine learning algorithm and give a high level overview of how the above process plays out for linear regression.

We now start this chapter by first looking at the basics of (applied) statistics and look at some of the most important results in statistics. The reason why we start our quest here is three-fold:

1. Nowadays, it is a popular belief that Machine Learning is the only method that can be used to solve problems using data. This belief is however completely bogus, for many problems, simple statistics are sufficient.
2. An algorithm is often used to create predictions, these predictions are by definition uncertain. Therefore having a sense of how to talk about uncertainty is required in order to talk about predicting.
3. A basis in statistics is quite standard in many school curricula, therefore this content will be somewhat familiar to many people. However, due to this *somewhat familiarity*, the results are often used in an incorrect way leading to illogical decisions. Using a combination of data visualizations and statistical knowledge these mistakes can often be laid bare and the search for a more sensible methodology can be sought.

Remark. The field of solving problems using data is of course much broader than statistics or machine learning. There's also Operations Research, graph algorithms and simulations, just to name a few alternative methods. Which methodology should be used is highly dependent on the problem at hand and it is always an important decision to select the right framework for a given problem.

Without further ado, let us dive into an overview of the types of data one may encounter.

2.2 Types of data

Section describing all types of structured data

2.3 Estimating center locations

Section describing types of center locations and when to use which.

2.4 Variability

2.5 Exploring distributions & Bayes Theorem

In this section we talk

2.6 Linear Regression & the standard ML flow

In previous section, we learned about conditional probabilities and Bayes theorem. Conditional probabilities lie at the heart of (almost) all machine learning problems, we constantly ask ourselves questions like:

- What should the prize of a house be given its surface area, location, ...
- What is the probability of being a survivor on the titanic given your age, gender, ...
- Etc.

Before we dive deeper into the *statistical* approach of solving these type of questions, let us look into the most taught machine learning method in existence: *Linear Regression*.

We are given a dataset consisting of the medical costs for a group of 1339 Americans. For these people, we are given the age, sex, bmi (body mass index), number of childen, whether they smoke, the region they live in and the total amount of medical charges they have been charged in their life.

Let us start by focusing on linear regression with only one feature, we expect a person's age to be the most telling, therefore we will try to predict the a person's medical cost solely based on his age. Therefore age is the input feature and the medical cost is the target or dependent variable. We can visualize the relation between the two features using a scatterplot, see Figure 6.

Remark. From this scatterplot, we can immediately see that there clearly is some connection between age and medical cost, but that we still require some additional information in order to split the dataset further. For ease of notation we will nonetheless first focus on the case with only one input feature (age), but afterwards we will investigate how we can effectively use multiple features in order to more accurately predict the medical charges.

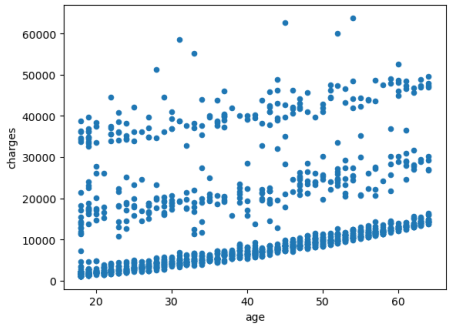

Figure 6: Scatterplot comparing the total medical cost of a person with his age.

We have chosen for linear regression with only one feature, therefore our model looks like:

$$h(x) = \theta_0 + \theta_1 \cdot x, \tag{1}$$

that is a straight line that passes through the $y - axis$ at θ_0, we call θ_0 the intercept. It then linearly increases linearly with a slope equal to θ_1. We are now looking for a *optimal* values for θ_0 and θ_1 such that the line defined by $y = h(x)$ fits the distribution of points in Figure 6 as closely as possible.

To this end, we first need to quantify the error made by a prediction, this is done by defining a loss function $\ell(x)$. The most typical loss function used for linear regression is the Root Mean Squared Error (or RMSE in short), it is defined by:

$$\ell_\theta(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(h(x_i) - y_i)^2}{n}}. \tag{2}$$

Here, we denote by n the total number of samples, by x_i the age of sample i and by y_i the actual medical costs charged to the sample.

Remark. This loss function in (2) (referred to as the RMSE) can alternatively be written as:

$$\ell(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}, \tag{3}$$

with $\hat{y}_i = h(x_i)$, the prediction for sample i. Using this notation one can clearly see that this is a general formula which can be used to measure the performance of any model.

Remark. There is a strong resemblance between the RMSE and the standard deviation! If we would use the average of all y_i as prediction, we find that (2.6) is exactly the variance of the sample data $(y_i)_i$. This *mean estimate* is often used as a base case and any model that you use can be seen as improving this base case, that is: predicting (a part of) the variability in the distribution $(y_i)_i$.

In the loss function in we have x_i, y_i and n which are fixed given the dataset. We can however still determine the values of θ_0 and θ_1 in order to find a good fit. We can therefore visualize $\ell_\theta(x)$ as a function of θ_0 and θ_1. We

2.7 Gradient descent

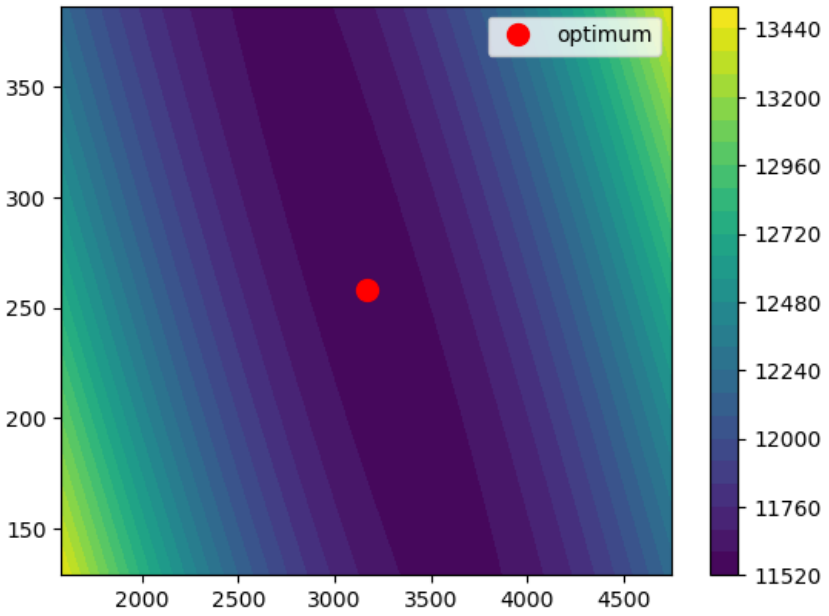

Figure 7: The loss function in (2) visualized for the medical charges example visualized.