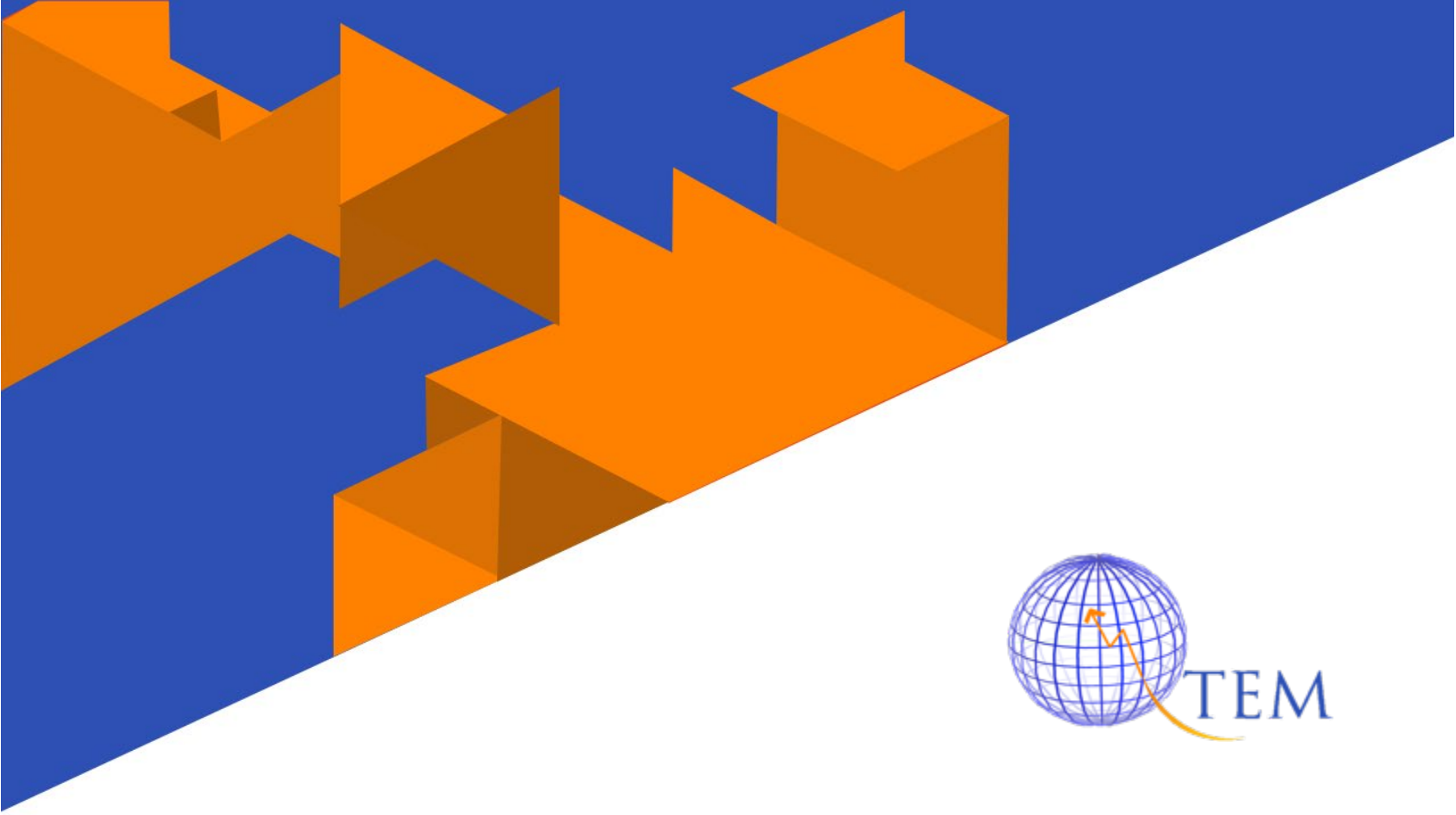

Name of the module	Reacfin Academy Introduction to Data Visualization

Powered by Reacfin

Contents

Scene 1: Introduction

Slide 1.1: Title

Slide 1.2: Introduction

In this module we will discuss data visualization and why it is so crucial today to master this kind of techniques. We will highlight some common techniques of data visualization, illustrate good and bad practices and provide use cases for the insurance sector.

At the end of this module, you will find a short 10-question quiz. This quiz will help you to test your knowledge and will allow you to proceed to the next modules.

Scene 2: Overview of data visualization + 17min30sec

Slide 2.1: Overview of data visualization

As stated in Module 1.1, data visualization is defined as the creation and analysis of the visual representation of data. It is an art based on interdisciplinary scientific facts that aims to exploit human eye and brain abilities to understand a content more easily.

In this first section, we will present the history of data visualization, its purposes and its characteristics. We will also discuss common issues through the illustration of good and bad practices. Finally, we will list the main players, tools and packages available in the data visualization market.

Slide 2.2: History of data visualization

Contrary to general belief, the concept of using pictures to understand data is a very old science.

Did you know for instance that the first maps visualizing star locations were drawn by cavemen on the walls of the Lascaux Cave? This takes us back to the Pleistocene Epoch which means more than 11,700 years ago! Another example of old data visualization techniques can be found in the Inca culture. 4,000 years ago, Incas used to create quipu, which were coloured, spun, and plied thread or strings made of cotton or camelid fibre, in order to collect data and keep records, such as tax and census records or calendrical information. The principle of these tools was to encode the represented values by knots on the strings. After prehistoric times, the most common forms of data visualization included various thematic maps from different cultures and ideograms providing and interpreting information. The first map depicting a spherical earth with latitude and longitude markings was created by Claudius Ptolemy in Alexandria in the second century.

Of course, throughout history, data visualization has improved with technological progress, from the invention of paper to the development of precise instruments and methods for measurements, to the creation and development of statistics and computer science. Progresses in data collection, observation and recording are also directly correlated with the progress of data visualization. Data visualization evolved thus at an accelerated pace, and spread throughout a larger and larger range of disciplines. What basically started with handmade and hand drawn visualizations, such as the first bar and pie charts published by William Playfair in the 18th century, evolved into technical applications which eventually led to software visualization as we will discuss further at the end of this section.

Slide 2.3: Reasons for data visualization

Data is often difficult to interpret, whether at the beginning of the process or in analysis and results. This is particularly true for large dimension datasets which are increasing in number with the emergence and development of internet and data-driven business models.

Thus, it is necessary to find the most accurate and intuitive representation of your data taking the general purpose and the person receiving the results into consideration. This is where data visualization comes in.

There are many uses for this combination of art and science:

- Data visualization helps us understand information quickly: analysing information in graphical form as opposed to spreadsheets reveals a large quantity of clearly presented data which answers questions faster. Visual perception is far more immediate than the sequential scan of numbers and letters;
- It identifies relationships and patterns: some large quantity of data only begin to make sense when they are represented graphically. Identifying correlations, which is an important concern in the insurance field, is not easy when only dealing with numbers, especially with large quantities of them. But a simple graphical representation can make this information simple to understand.
- It also quickly detects emerging trends: this immediate detection can give companies an edge over the competition as well as the opportunity to cope with problems soon after their emergence, such as an important number of clients terminating their contracts.
- Another huge advantage of data visualization is that it helps you to communicate your story to others: your audience will understand the main results and insights of your analysis. As visual depictions of data are universally understood, you can propose your idea in a timely and attractive manner, without needing a more technical background. When this new business language is used throughout the company, a much broader audience has the opportunity to analyse the data. In this way, it can provide a fresh view on the results and lead to other interesting insights.

Slide 2.4: Characteristics of accurate graphic representation

In 1983, Edward Tufte defined the principles for effective graphical display in the following sentence: “Excellence in statistical graphics consists of complex ideas communicated with clarity, precision and efficiency”.

In order to obtain this excellence, some steps should always be adhered to.

The process of data visualization begins with asking yourself the three following questions:

- What do you want to show? As your graphic representation should tell a story, it is important that you keep in mind the key information that you want to convey. To do so, an implicit process is needed in order to really understand the dataset, find the key messages and thus truly achieve your objective.

- Who do you want to show that information to? It is important to have a good understanding of who your audience is and how they perceive the information. Trying to communicate to too many different people with disparate needs and backgrounds prevents you from communicating as efficiently as you could by restricting yourself to a more specific audience.

- How to show that information? Once you have clearly understood and identified all the interesting information that should be conveyed to the chosen audience, the next step is to find the best way to represent it so the audience can thoroughly explore it. This could be achieved, for instance, through interactive data visualizations that are kept simple but allow the user to see relevant additional information through different buttons, effects or added filters. The precise chart type that you should use depends on the patterns you want to identify, the number of variables and the amount of data, as we will discuss in the next part of this module.

Slide 2.5: Common issues

In this slide, we will explain and illustrate the most common issues in data visualization, while giving an example of a possible solution to the corresponding problem.

A common example of poor data visualization is **a chart containing too much information** for the audience to stay focused, to understand and to digest the key message. Being given too much information at once is known as excessive cognitive load. The brain is overloaded with information and doesn't try to understand what the key message worth extracting from the data visualization is. Adding too much information to a single chart eliminates the advantages of visual perception as the viewer has to read every element individually in order to understand the key message. Instead of overloading one single data visualization, like the chord diagram on the left which is a very complex graph, you should try changing chart types, removing or splitting up data points, simplifying colours or positions. In this case, using a Sankey diagram which is easier to read than this chord diagram is a possible solution. Making the graph interactive in order to hide some overloading features is another solution.

Another issue in the matter of data visualization is **the temptation to use too many unnecessary designs** such as too many colours, too many sizes or other visual effects. Visual

elements that take up space but don't increase understanding are called clutter. They make your visuals appear more complicated than the key message actually is, which also leads to an excessive cognitive load. In his book from 1983, Edward Tufte stated that the data-ink-ratio should be maximised. Most graphics tend to contain a stunning amount of excess ink in comparison to the information they convey.

Here, the cutting of the pie chart in addition to the use of a third dimension makes the percentages far less clear. A first solution could be coming back to a simple pie chart. A second solution consists in applying some good practices in terms of use of colours in order to obtain this last, more attractive chart. However, considering pie chart areas can be sometimes confusing, a third solution could be using a bar chart as lengths are easier to compare for the human eye.

In the opposite direction of overcomplicated designs, we also find **data visualizations that are not attractive enough** to generate the interest of the audience. Details that require attention include rotating labels on a graph in order to make it simple for the viewer to read, avoiding too bright or contrasting colours, as well as more general settings such as choosing a theme of colours while taking into account the desired tone of your visualization as colour conveys emotion, or choosing to make your graphs interactive to stimulate curiosity and encourage exploration of the data visualization.

In this chart, the unit could have been changed, as well as the over-use of gridlines and saturated colours.

Another major issue that can negatively impact conclusions drawn from the dataset is a representation which is **badly scaled**. For example, if you have one or two very tall bars in your bar charts compared to the rest of your data, you might consider using multiple charts to show both the full scale and a "zoomed in" view. Thus, no piece of information is lost in the process. Other solutions could be finding a non-distorting transformation of the data you want to represent or using a scale break.

This last point can lead to unwanted **distortion of data**, in the same way as truncation, omission and bad visual effects can. Cropping off the bottom parts of your bar charts leads to misjudgements of the data. The conventional way to set up the y-axis is to start at 0 and then go up to the highest value in your data set. By not setting the origin of the y-axis at zero, small differences can become huge as you can see on the bar chart below. Omitting certain data points is another way to distort the information and the context is lost. Trends that don't actually exist can easily be created whereas some existing highlights can be hidden. That's what was done for this graph for propaganda purposes.

This kind of misleading graphs is often used deliberately for propaganda. For example, companies can take advantage of this by omitting years with significant changes in sales to make their earnings look constant and predictable, masking the true volatility of the market.

Another type of data charts that tells a distorted story is **3D** pie graphs. As you can see here, the importance of the yellow and pink pieces of this pie cannot be easily compared to the green one on the distorted graphs but it is clear on a simple 2-dimensional pie chart.

Sometimes, the **chosen chart type could be inappropriate** to convey the key message. As we will see in the third part of this module, each chart type has a certain purpose. For example, pie charts should be used to illustrate composition when there aren't too many components. If you have several items and the key message you want to convey is the comparison between the importance of each component to the whole, you should use a comparison bar chart instead of a pie. This will be clearer for your audience and you won't use too many colours and lose clarity when illustrating the less represented components. Another inappropriate choice would have been visualizing a trend in time using a bar chart, instead of a line chart.

Slide 2.6: Other good practices

If you want to avoid making the above mentioned mistakes, keep these principles in mind when creating data visualization:

- The three minute principle: According to this concept, your visual representation should be simple enough for the audience to understand the key message in less than three minutes.
- How do visual perception and memory work in our brain? Consider preattentive processing. Those preattentive attributes (such as form, colour, motion, and spatial positioning) are processed in our sensory memory without our conscious thought. You can click on this process to discover those attributes in more details. Designers can take advantage of these pattern recognitions in their designs to help users better understand the information. For example, quantitative perception is very precise with length and spatial position, whereas it is less precise with sizes and shades. Gestalt Theory of Visual Perception provides a clear description of many basic perceptual phenomena of pattern recognition. The laws explain how individual elements may be visually organized into groups and how the brain distinguishes relevant information from noise. You can click on this theory to discover the influence of six of these principles: proximity, similarity, enclosure, closure, continuity and connection on your own visual perception.

- Other important and specific good practices concern the use of colours. A good practice consists in using maximum six colours while avoiding using combinations of warm and cold colours, or colours coming from opposite sides of the colour wheel. Keep in mind that warm and cool colours do not render the same way to the human eye. Also, colours should be used to distinguish or emphasize the key data, and not decorate them. Sometimes, using only shades and saturation is enough to get your message across

while maximizing the data-ink ratio previously mentioned. Don't forget to consider colour-blindness when creating your representation. Several colour-blindness simulators can be found on the internet so you can check what your visual representation looks like to those who suffer from this condition.

- Unnecessary multidimensional representation should always be avoided.

- Finally, keep it simple! Bad visualization confuses rather than clarifies! Indeed, there are many ways to visualize data. New tools and chart types are constantly appearing and each aims to create more attractive and complex charts than before. However, one should focus on the principle that visualization should clarify and summarize the key messages rather than confuse and overload the reader with superfluous information. Designs should be a tool used to only draw attention to key messages and the more information you convey through them, the harder it is to read.

Slide 2.7: Data visualization in the market

The data visualization market was valued at 4.51 billion dollars in 2017, and is expected to reach a value of 7.76 billion dollars by 2023. The most active player in this market is North America followed by Asia-Pacific and then, Western Europe.

The increasing interest in measuring the performance of every operation across an organization to get a daily overview of their situation is driving the demand for these solutions in the market. Technological advancements along with fast growth in big data and the growing need for faster decision-making are some of the other factors accelerating the growth of the market. However, the implementation of data visualization software is complex and lack of skills might challenge the deployment of data visualization applications in the market worldwide.

To give you an example, banks today are growing in size and are expanding geographically. Hence, the volume of transactions is growing exponentially, and manual operations and analyses have become time-consuming. As a result, a near to real-time data visualization has become common in the market and teams need automated dynamic dashboard solutions rather than complex manual reporting processes. Banks have thus started using advanced visualization tools to track their operations.

There are numerous key players in this market: Tableau Software, SAP SE, SAS Institute Inc., Microsoft Corporation, Oracle Corporation, TIBCO Software Inc., IBM Corporation, Information Builders, Dundas Data Visualization Inc., Pentaho Corporation, InetSoft Technology Corporation, MicroStrategy Inc., and many others.

As a result of innovations, statistical and quantitative data visualization, which was initially performed with the Microsoft Office Suite now relies on dashboard, software and programs such as SAS, SOFA, R, Minitab, Cornerstone, D3, Python and Javascript. There are a lot of tools, meaning software packages, developed for creating visualization using these programming languages. These are the most popular ones in R programming depending on the type of visualization:

- Ggplot2 for statistical visualization

- Plotly,
- Ggigraph,
- Ggvis,
- RCharts for classic interactive visualization

- Echarts,
- D3heatmap,
- DiagrammeR,
- NetworkD3,
- VisNetwork,
- Three for more complex interactive visualization

Scene 3: Data visualization families

Slide 3.1: Data visualization families

There are many types of data visualization depending on the story you want to tell.

In this section, we will explain the four basic data visualization families in details through their definition and different examples drawn from a same dataset per family. We will then conclude our study by giving some examples of what other less common types of data visualization can help represent.

Slide 3.2: Choosing a presentation type

There are four basic presentations of data visualization depending on what you want to show: Comparison, composition, distribution and relationship. The two most commonly used types of data analysis are comparison and composition.

Besides, many other representations exist, but we won't discuss them here for simplicity reasons.

To determine which precise type of visualization is best suited for your purpose, you must ask yourself a few questions:

- How many variables do you want to show in a single chart?
- How much data will you display for each variable?
- Will you display values over a period of time, or among items?

Dr. Andrew Abela created a high-level chart selection diagram that can help you pick the right chart to convey your key message based on these questions. You can consult this diagram by clicking on the following button.

We will now apply it to our examples, visualizing a dataset in three different ways for each family in order to observe different style effects and give some benchmarks.

Slide 3.3: Comparison

Comparison charts are data visualization methods that show the differences or similarities between values. They can be used to easily notice highest or lowest values in a dataset, as well as to compare current values to old ones in order to detect a trend.

There are many types of comparison charts Depending on whether you are showing an evolution over time or a simple comparison between items. It also depends on the amount and type of data you want to represent.

Hereunder you can see comparison charts on the number of traffic accidents occurred in Belgium, between regions and over time. Click on the buttons in order to zoom in on the different charts.

Line and bar charts are used to represent the evolution over time, whereas this map of Belgium, called a chloropleth map is well-suited to compare the number of accidents per regions in a single year. We used only cold colours here in order to respect the good practices described earlier.

Slide 3.4: Composition

Composition or parts-to-a-whole charts are data visualization methods that show the different components of a whole, as well as the importance of parts of a variable compared to its total in relative value.

Below, you can see the representation of car accidents with injuries occurred in Belgium in 2016 over the four seasons. Pie and doughnut charts are suitable for representing such information. If you want to add the precise number of car accidents per month, tree maps and sunburst charts can be used. Click on the buttons in order to zoom in on the different charts.

Slide 3.5: Distribution

Distribution charts are data visualization methods that display how data is spread over an interval, or, more simply put: the frequency of occurrence.

Below, you can see the distribution of Belgian population depending on age classes, as it was in 2016. Here, an age pyramid and an area charts were used. Click on the buttons in order to zoom in on the different charts and get some explanations.

For other types of datasets, histograms or bubble charts could also be used.

Slide 3.6: Relationship

Relationship charts are data visualization methods that show connections or correlations between data.

Below, you can see three ways of representing the correlation between fuel consumption and three characteristics of 32 car models: the number of carburettors, the number of cylinders and the horsepower. Click on the buttons in order to zoom in on the different charts.

Many other types of charts could have been used here, depending on what relationship you are trying to show or find.

Slide 3.7: And much more…

There are many other families of visualization methods to represent different types of data:

- proportions, to use size or area to show differences or similarities between values,
- hierarchy, to show how data or objects are ranked and ordered together in an organisation or system (an example of dendrogram was presented in module 2.2)
- location, to show data over geographical regions as we did with the chloropleth map exposed earlier,
- movement, for showing the flow and the transformation of data (called lineage),
- text analysis, to reveal insights from a body of text using for example a word cloud, as already mentioned in module 2.4
- and many more…

Scene 4: Examples in insurance + 4min

Slide 4.1: Examples in insurance

Data visualization is and continues to be a fundamental tool for any company that deals with a huge amount of data, including insurance companies of course. In this final section, we will explain why and how such data visualization could be used in the insurance sector.

Slide 4.2: Asset management

In asset management, we usually use Economic Scenario Generators (ESG) which are models that project the value of economic indicators (e.g. stock returns, interest rates, corporate bond spreads, property values). An ESG helps to create numerous possible scenarios for the evolution of macro-economic and market indicators. The main problem is that such models are quite sensitive, and also that output data usually comes in huge amount and not so easy to analyse. In such situations data visualization can be very helpful:

- To observe historical data to set up the ESG;
- To observe correlation between indicators and thus verify our hypothesis ;
- To observe the distribution of our results with an easy and readable view.

You can click on the different buttons to zoom in on the data visualizations.

Slide 4.3: Life insurance

In life insurance, mortality rates are key drivers that must be followed in an insurance portfolio. In recent decades, significant improvements on life expectancy have been observed. It has become critical for insurers to study such statistics. Prospective mortality tables were created in order to assist in this effort, incorporating mortality rates that take both age and generation into account. Nowadays, these tables help actuaries to model life insurance risks. But one problem remains: these tables or the results are not easy to understand when considering multiple dimensions (age, generation, etc.). Here again data visualization has proven useful:

- To observe prospective tables content in a 3D view
- To observe prospective tables content using a heatmap
- To visualize backtestings which test predictive models, using historical data, on life expectancy

You can click on the different buttons to zoom in on the data visualizations.

Slide 4.4: Non-life insurance tariff

Another example is about the development of a commercial tariff. The starting point of such an exercise involves the calculation of the technical tariff followed by additional enhancement layers that must take multiple factors into account. Competitor price analysis is not a trivial exercise and several factors must be taken into consideration so as to have access to the most complete picture. In order to respond to such needs, data visualization that enables the user to perform many comparative analyses to assess and benchmark warranty prices relative to those of competitors would be a valuable tool. Several examples are presented below:

- To visualize the distribution of the insured age for different competitors and different profiles;
- To compare and pinpoint the global position of the tariff offer compared to other insurers.

You can click on the different buttons to zoom in on the data visualizations.

Slide 4.5: Insurance premium segmentation

Last topic will deal with the premium segmentation based on location in insurance. In recent years, we observed that claims experience was not evenly distributed in some countries. Indeed, for Belgian car insurance for instance, it can be noticed that the claim frequency is more significant in big cities, for small claim amounts, whereas the costs are more important in more remote regions even though accidents are less frequent.

Insurers may therefore be interested in pricing their products while considering the place of residence of the policyholder. To respond to such needs, it can be relevant to develop data visualizations:

- To visualize the distribution of the factors by which they should multiply their tariff throughout the country of interest,
- To allow easily measuring the gap importance between the highest and the lowest factors.

You can click on the different buttons to zoom in on the data visualizations.

Scene 5: Conclusion

Slide 5.1: Conclusion title

Slide 5.2: Conclusion

In conclusion, although data visualization is not a new concept, it is essential to master this practice in the present day. We presented different examples of bad and good uses of visual representation, as well as the key elements to keep in mind and the main questions to ask ourselves when we want to represent results. We also described the four main families of data visualization which are: comparison, composition, distribution and relationship. Finally, we applied data visualization methodologies in an insurance context to analyse results related to asset management, life insurance, non-life pricing competition and insurance premium segmentation.

If you feel comfortable with all the concepts you have learned in this module, you can begin the certification questions by clicking on the right button.

If not, you can rewind the module and listen to it again by clicking on the left button.