Exam Questions
Visual Data Science VU
TU Wien, 186.868, WS 2019/20
v0.2

By a group of motivated students

January 22, 2020

Contents

1 Example Questions

Disclaimer: no guarantee for questions and answers being complete or correct.

1.1 Chapter 1

Question 1.
What is the Anscombe's Quartet?

Answer 1.

- Four groups of numbers [2 dimensional data points] that have identical statistical parameters
- Visual representations of the data are quite different
- The dataset shows that statistical analysis is abstracting over possibly important details of the dataset, exemplifying the importance of visual analysis of the whole data.

[comment: plotting the data reveals they have quite different layouts, however the statistical properties are virtually identical. Especially, the linear regression model shows the same R^2, and the correlation coefficient is also the same for all four cases. from only these measures one might conclude that the data are very similar, which they are not, as is revealed when plotting.]

Question 2.
What are the three visualisation use cases?

Answer 2.
(VDS_05, 19/20, p.25)

- Exploration
  - Searching and analysis
  - No or only minor prior knowledge about the data
  - Find potentially useful information
- Confirmation
  - Goal-oriented
  - Examination of a prior defined hypothesis
  - More prior knowledge of the data
- Presentation
  - Efficient communication of data features and findings

- Clear definition of what to show
- Often targeted towards external people

Question 3.
Name a visualisation technique for exploration/confirmation/presentation and describe it.

Answer 3.
(VDS_01, 19/20, p.30f)

Scatterplot

- Bivariate data (2 dimensions)
- Visual channel: position
- Intuitive, easy to spot outliers, clusters, correlations, and to identify distributions
- More dimensions may be added by using additional visual channels (such as color, size, shape of symbols → Bubble chart)
- For Exploration, Confirmation, and Presentation

Bubble Chart Like scatterplot, but additional dimensions can be shown

- Uses size for additional attributes → 3 dimensions
- Color used for categorical attributes [use categorical color map]
- For Exploration, Confirmation, and Presentation

Question 4.
What are parallel coordinates?

Answer 4.
(VDS_01, 19/20, p.34) Analysis of data based on more than 2 variables. Usually, variables have different domains (numbers, categories, etc.).

- Invented probably 1885, but got popular in 70s (Alfred Inselberg)
- Align data dimensions as vertical axes
- Axes need to be scaled accordingly
- Can be used to identify statistical parameters:
  - Parallel lines: positive correlation
  - X-shaped lines: negative correlation
  - Random: no correlation

- Axes order very important to spot correlations
- Alternative: Splines instead of straight lines
- Visualisation technique for multivariate data
- For Exploration and Confirmation

Drawbacks:

- Axes ordering
- Overplotting [cluttering by many lines crossing, etc.]

Overplotting can be solved by:

- Filtering [remove uninteresting items; select value range of variables?]
- Clustering [grouping together instances of the data having similar paths in the parallel coordinates plot?]
- Brushing [Selecting only a subset of data instances / a subset of lines in the parallel coordinates plot, with rectangle, etc.]

Question 5.
What are radar charts?

Answer 5.
(VDS_01, 19/20, p.51)

- Circular alignment of axes
- Axes need to be scaled
- Harder to spot correlations
- Rather used for comparisons [between different instances having same attributes]
- Interpretation may be difficult, because of radial distortion
- Axes ordering important
- Should not be used for linear data, axes should be independent
- For Exploration, Confirmation, and Presentation

Question 6.
What are the main other scientific disciplines that influence visualisation research?

Answer 6.
One of these:

- Perception
  Usage of colors and shapes, Human abilities to recognize patterns
- Human-Computer-Interaction (HCI)
  Interaction for Analysis, Input devices,
- Computer Graphics
  Rendering, Efficient handling of large data

Or:

- Cartography
- Graphic Design
- Statistics
- Psychology
- Computer Science

1.2 Chapter 2

Question 7.
What are glyphs?

Answer 7.
Small visual objects that can be used independently and constructively to depict attributes of a data record.

- Glyphs can be placed independently from others
- Glyphs may be connected to convey topological relationships
- Glyphs are to be differentiated from other types of signs such as icons, indices, and symbols

Examples:

- Chernoff faces (?)
- Star chart / plot (?)
- Vector field arrows (?)

Question 8.
How can graphs and networks be visualised?

Answer 8.
(VDS_02, 19/29, p.37)

They can be visualized with nodes (circles) and edges (lines) in the following ways to avoid visual clutters in the following layouts:

- Radial layout [nodes arranged in circle, connected by edges]
- Clustering & Box layout [Aggregating node that are very close, drawing boxes around clusters to get some visual segregation?]
- Matrix visualisations [names of nodes naming rows and columns, cell is colored if connection is present. This can show directed graphs?]

Hierarchical networks can also be represented as:

- Trees [root node(s), edges]
- Sunburst Charts [concentric rings / ring sections, innermost are root node(s) of hierarchy, outer ring sections get deeper and deeper in hierarchy]
- Treemaps [(VDS_02, 19/29, p.45, also VDS_06, 19/29, p.37) Split large rectangle into smaller rectangles, each is a certain attribute / variable, used to show proportions. Attributes mapped to rectangle size, color, position in hierarchy.]

Question 9.
Which points have to be considered when designing choropleth maps?

Answer 9.
(VDS_02, 19/29, p.52ff) General:

- Usage of relative (rather than absolute) values [?]
- Choosing the right visual mapping (more on this below)

One needs to choose the right visual mapping: e.g. the right color map (which depends on the type of data:

- is it categorical $\rightarrow$ use of different colors,
- is it diverging with a zero point $\rightarrow$ use of a diverging color map
- is it on a ratio scale $\rightarrow$ one color with different brightnesses = sequential colormaps).

The number of colors shouldn't get too large, as it gets harder and harder to distinguish different groups as there is only a limited set of colors available that are easily distinguishable (for categorical data).

Also, one should aim for a high spatial resolution, as if it is too coarse-grained, it might abstract to much and show a misleading depiction.

Question 10.
What are alternatives for choropleth maps?

(VDS_02, 19/29, p.63ff)

Answer 10.

- Circle Maps [Map with e.g. outlines of countries/states shown; each country or state has a circle plotted into it (somewhat centered), with attributes mapped to area, color, ...]
- Tile Grid Maps [Here, the countries/states/regions retain their position as on the map, but their outline/shape is replaced with square or hexagonal cells. This way, regions (etc.) can still be identified by their position (besides names, etc.). Attributes mapped to color, but possibly also size. In the latter case, the shape of the original map may get more distorted (see example on p.69)]
- Cartograms [Contiguous or non-contiguous versions available. Attributes mapped to area of country/region, distorting the original shape, but regions still "stick" together (contiguous version). Non-contiguous version: can retain shape of each region, but scales area of region. The regions are no longer "sticking" together.]

Question 11.
What is edge bundling?

Answer 11.
Edges close to each other are bundled together to get a better overview over the general direction of edges.

Used for Origin-Destination-relations (OD-relations), as a better demonstration of various flows [can reduce clutter].

Flow direction can be visualized using color gradient.

Question 12.
What are trajectories?

Answer 12.
(VDS_02, 19/20, p.79ff)

- Movers are characterized by their trajectories [Mover: object that changes spatial position over time]
- Trajectories [trace (curve) depicting a state change; here: change of position with time]:
  - Points sampled in space (usually GPS)

  - Interpolated curve
  - Optional additional attributes (e.g. speed)

Additional: Challenges

- Noise: from sensors; Solution is using (advanced) filtering, like Kalman or Particle filters, running mean or median, etc.
- Map matching: Align trajectory with road (that was likely taken) on map. Can be done geometrically (shape), topologically (connectivity of road network, probabilistic (for dealing with low sampling rates).

Additional: Visualisation

- Aggregation of trajectories (like edge-bundling?)
- Space-Time cubes: 3D visualization, time mapped to axis perpendicular to map. Shows temporal and spatial extent.

1.3 Chapter 3

Question 13.
Which two strategies can be used in visualisation to make better understand AI models?

Answer 13.

- Instance based: How do instances in the data contribute to a model‘s accuracy?
- Subset based (= input features): How do features effect the model‘s output?

Question 14.
Name an example for an “understandable AI” approach.

Answer 14.
(VDS_03, 19/20, p.16ff)
Preliminary: visualization for AI models:

- Instance-based: how do instances in the data contribute to a model's accuracy?
- Subset-based (subsets = input features): how do features effect the model's output?

Goal of explainable AI is to make neural networks understandable by humans (vs. being a black box).
Explain one of those:

- Neural Network Playground (subset based) [play around with different number of layers, layers per neuron, etc.; can show activation functions / feature maps?]

- Neuron Optimisation (instance based)
- Activation Atlas (instance and subset based)
- ActiVis (instance and subset based)

E.g. **Neuron Optimisation**:

- Idea: Uncover what neural network "sees"
- Understand how neural networks build understanding of images
- Adapt input images to maximize neuron activity
- Optimise input data [→ Allows to see what parts of an image a neuron reacts to. This way it is possible to check whether a neuron reacts to what we think it does in an image, or if it is something different that is probably not relevant for the actual class.]
- Drawbacks:
  - Optimisation only for individual neurons
  - Neurons never work in isolation [so, we can't then say if it will behave identical or at least similar when working together with other neurons?]

Question 15.
What is data wrangler?

Answer 15.
(VDS_03, p.53ff)

An approach to using AI for improving visualizations.

Software for data wrangling (pre-processing, formatting, ...). Basically machine learning system that can assist in data wrangling, can suggest operations on data based on user actions, history, data attributes. Has shared representations of the possible processing steps, i.e. representation language that is human- as well as machine-readable. Gives preview of the suggested operations (how would the outcome look like, how would the data change?).

Question 16.
What is CNN-based brushing?

Answer 16.
(VDS_03, p.60ff)

An approach to using AI for improving visualizations.

Assumption: Brushing goal [that we want to achieve through brushing] be derived from:

- Brushing interaction
- Data distribution near interaction [select data points that have similar data distribution to the ones selected manually?]
- Brushing is an interactive method for selecting data points in a visualization by drawing simple geometries onto it
- key functionality in coordinated multiple views
- consistent highlighting of the selected data in all linked views
- Paper referenced in lecture ("Fast and accurate CNN-based brushing in scatterplots"): `https://vis.uib.no/wp-content/papercite-data/pdfs/eurovis18.pdf`

Procedure: For sketching, the user clicks into the middle of the data subset to be selected and drags the pointer to the border of the subset; The CNN then "sees" the data distribution near the interaction as a 2D histogram [i.e., the distribution of data close to where interaction happens is used as input to the CNN]. The CNN delivers a degree-of-selection value per histogram bin [degree of importance of data points represented by the selected histogram bin?], from which we can compute, which data subset is selected. **Results**: reduced error rate when brushing, fast enough for interactive workflow.

1.4 Chapter 4

Question 17.
What are personas?

Answer 17.
(VDS_04, p.17)

Concept introduced around 1999 (Alan Cooper). A persona is a fictive person. A persona is meant to represent a larger number of users (a specific user group?).

(Goal: understanding of target users. There are different user groups with different requirements and backgrounds).
Persona defines:

- Characteristics [...]
- Expectation [of / towards?]
- Motivation
- Background

Usage:

- Requirement analysis is targeted towards defined personas (so, have some personas representing a larger user base in a few, manageable cases?)

- Persona description should be as detailed as possible
- Personas are shared across teams (Working on the same project / product?) for common understanding (common "design target"?)
- Used to describe typical use-case scenarios

Persona description consists of:

- Age
- Gender
- (Family status)
- Education
- Job title
- Work experience
- Use-case scenarios

Personas are basis for defining tasks (to be solved using a system / tool / visualization).

Additional: Tasks

- Users think in tasks
- Task: smaller part of user activities with meaningful outcome
- Task: use-case, scenario
- Task analysis: Understand the tasks users need to perform; understand how often tasks need to be performed.
- Tasks $\neq$ Functions

Design process (of tool/software) should be triggered by tasks that should be supported. Task frequency should influence design decisions: more frequently used tasks should be easy to access.

Question 18.

Name and describe three usability heuristics.

(VDS_04, p.34ff)

In general: Usability is *"The effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments."* (ISO 9241)

There is a list of 10 usability heuristics in the slides, 4 of them emphasized (details below):

Answer 18.

2. Match between system and the real world
4. Consistency and standards
5. Error prevention

6. Recognition rather than recall

[The following is copy-past from slides]
Match between system and real world
The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
Consistency and standards
Users should not have to wonder whether different words, situations, or actions mean the same thing.
Error prevention
Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
Recognition rather than recall
Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate [e.g. like a tooltip, i.e. name of function when doing a mouseover?].

Question 19.
What has to be considered when using shapes in visualisation?

Answer 19.
(VDS_04, p.46ff)
Example in slides: Amounts of money raised (for research) vs. number of deaths from different causes. Correct mapping is of numbers to area of circles, not diameter. Categories are differentiated in the visualization using color.

- Mapping data attributes to geometric elements (shapes / symbols?). Color, position and size can be used for additional information.
- When mapping quantitative attributes to shapes, use the **area** and not the diameter [strictly for round shapes, and anything remotely round?]
- More complex geometric objects allow to map multivariate attributes [to properties like size, angle, aspect ratio, convex or concave outlines (see p.53)]

Rules for usage:

- Fewer (different) shapes is better than many
- Shape parameters (size, angle, aspect ratio) can be mapped to attributes for multivariate data
- Always adjust **area**, not diameter / size when visualizing (mapping) quantitative attributes

Question 20.
Name and describe three Gestalt laws and their relation to visualisation.

Answer 20.
(VDS_05, p.55)
Relation to visualization:
They define laws [rather: rules, principles] of low-level pattern recognition. From visual perception theory (1920). Were established in psychology, theory of perception [comment: the visualization people often don't mention that these are actually from psychology and are a wider concept than only for visual perception]. Define [I think, rather describe] how humans perceive shapes based on size and location.
Gestalt laws are important for UI design.

- Law of proximity: Spatially close elements appear as grouped together and "as one"
- Law of similarity: When things appear similar to each other [such as common shape, common color], we group them together. We also tend to think they have the same function.
- Law of continuity: Elements that are arranged on a line (curve) are perceived to be more related than elements not on the line or curve
- Law of closure: When we look at a complex arrangement of visual elements, we tend to look for a single, recognizable pattern.
- Law of simplicity: When we see convoluted shapes in a design, the eye simplifies these by transforming them into a single, unified shape.

TO DO: put some explanatory figures in here?

Question 21.
What has to be considered when using colours?

Answer 21.
(VDS_04, p.72ff)
Color is a strong visual channel for transporting information.
Things to consider:

- Similar/dissimilar colors for similar/dissimilar attributes [categorical color maps for categorical variables; there are sequential, diverging color maps, etc.]
- Psychological effects (e.g., red vs. blue for alerts)
- Consider color blindness

Rules:

- Do not use too many colors in one chart

- Consider relation and size of elements
- Do not use gradient colors for categorical data [suggests "closeness" that may not be there between categories]
- Use intuitive colors [red → bad, blue → good, and not vice versa, etc.]
- Carefully design color maps [e.g., two hues instead of one]
- Don't use rainbow color maps, there are better ones, based on perceptual principles (perceptually uniform)

1.5 Chapter 5

Question 22.
How can data science and business intelligence be distinguished?

Answer 22.
(VDS_05, p. 8ff)

Business intelligence (BI): Monitoring the current state of business data to understand the historical performance of a business.

Slide p.9 gives an overview of differences: (KPI ... Key Performance Indicators)

	Business Intelligence	Data Science
Focus	Reports, trends, KPI	Patterns, correlations, models
Process	Static, comparative	Exploratory, experimentation, visual
Data Sources	Pre-planned, added slowly	on the fly, as needed
Transform	up front, carefully planned	in-database, on-demand, enrichment
Data Quality	Single version of truth	"Good enough", probabilities
Data Model	Schema on load	Schema on query
Analytics	Retrospective, descriptive	predictive, prescriptive, preventative

In summary, it seems BI is more concerned with question "what happened?". From existing, past data, a model is formed that is describing what happened, so I assume this should be a more explanatory model: what happened, and why, which factors / variables contributed to that? There is an existing model of the data (data warehouse) where the data is stored. Reports are created to assess the performance of e.g. a past year.

Data science on the other hand is more concerned with coming up with models from data, forming hypotheses first, building models, evaluating them and eventually using them for prediction and prescribing actions. DS methods seem to be more for looking into the future, trying to build models of what will happen, whereas BI is looking into the past and analyzes what has happened.

Question 23.
What are the main visualisation challenges in visual data science?

Answer 23.
(VDS_05, p. 29ff)
We have:

- Large data sets [many instances, variables/attributes?]
- Complex data [many attributes/variables, complex relationships between them?]
- Web-based vs. desktop applications [how to present what to whom?]
- Exploration vs. presentation [Get insights, narrow them down, present to different audiences?]

[More needed here?]

Question 24.
Name two dashboards categories and describe them.

Answer 24.
(VDS_05, p. 39ff)
There are several different categories, the slides present the 7 following:

1. Strategic decision making
2. Quantified self
3. Static Operational
4. Static Organizational
5. Operational decision-making
6. Communication
7. Dashboards evolved

In general, dashboards can have different purposes and goals (Decision making; Awareness, Motivation and Learning), are targeted at different audiences (with differing e.g. visual literacy), need to have different visual features and different levels of need for interaction (may need no interaction at all) (see Table in VDS_05, p. 50).
Examples:
Strategic decision-making (same for Operational decision making):
(To assist in decision making; how would y change if I change x?)

- Audience: at organizational level [manager dudes / dudettes?]
- Daily business
- Benchmarks
- Interaction [i.e. graphs must support interaction]

Static operational:
(So, this is mainly for observing some process, checking that everything runs as it should?)

- Support for operators
- Real-time data from sensors
- Domain knowledge needed
- No interaction

Communication:

- Audience: general public
- Communication and learning [present some results to general public]
- Interaction [i.e. graphs must support interaction, "playing around" with data / visualization?]

1.6 Chapter 6

Question 25.
Is it better to use bar charts or pie charts and why?

Answer 25.
(VDS_06, 19/20, p.12)
Better to use bar charts for comparison than pie charts, since it is

- Hard to read exact values in pie charts
- Differences in angles are harder to spot

[Both these tasks should be better solvable using bar charts]

Question 26.
What is the main perceptual trigger to interpret pie charts?

Answer 26.
(VDS_06, 19/20, p.28ff)

- Main visual clue: Area (!) [p.34]
- Also: Arc length, Angle

Donut charts (like pie chart, but ring instead of circle as basic element) also can work well. But: do not use stacked donut charts.

Question 27.
How should temporal attributes be visualised on a map?

Answer 27.
(VDS_06, 19/29, p.46ff) shows examples, tasks to be solved with such visualizations. Slide p. 50 shows different suggestions, depending on how many locations on the map are needed (single location, locations in a region, all locations), and what time instances (single point in time, time interval, or all times).

Some possibilities (see p. 50):

- Dorling Cartograms: circles that are centered on their respective location on the map (the map itself is not shown), and several Cartograms next to each other are shown, where each one represents a single year. Best for: single time, independent of location and time interval with all locations. Still good for time interval and locations in a region.
- Bar charts plotted at each location, where bars represent values at consecutive times [in the visualization shown, color of the bars was used as additional channel for a different variable]. Best for: time interval and one location; all times independent of location (i.e. for all cases single, region, all locations). Still good for time interval and locations in a region.

[I think this doesn't really is what is asked for here, but I think it also is a way to visualize temporal evolution in different places: Circular visualization, radial slots, each e.g. an hour; clock metaphor (VDS_06, 19/29, p.18f)]

1.7 Chapter 7

Question 28.
Why is it important to connect visualisation with statistical tests?

Answer 28.
(VDS_07, 19/20, p.12ff; p.18ff)

- Data with the same statistical properties can still be quite different (see Q1)
- Statistical tests: Multiple Comparison Problem → The more inferences are made, the more likely erroneous inferences are to occur; problem also evident in exploratory visualization.
- Visualisation connected with tests also avoids
  - False discoveries: discovery of interestingly looking, but random events
  - False omissions: ignoring real patterns because they looks uninteresting
- (More detailed examples are the whole content of Chapter 7)

Question 29.
Name two perceptual biases and describe them.

Answer 29.
(VDS_07, 19/20, p.27ff)

- Clustering Illusion: Clusters and groups perceived in random distributions (related to Gestalt Laws) → cognitive bias
- Multiple Comparison Problem [described in question above].
- Texas Sharpshooter Fallacy: Occurs when differences in data are ignored, but similarities are overemphasized
- Curse of Knowledge: Having different backgrounds, charts can be interpreted quite differently (difficulty for presentations). Researchers (showing results) assume audience reads graphs in same way as him/her (which might not be the case).

Additional: Data analysis pitfalls
- Gambler's (Monte Carlo) Fallacy (a form of cognitive bias)
- Birthday paradoxon

1.8 Chapter 9+10

Question 30.
Name three charting libraries and describe their differences.

Answer 30.
Python has for instance Seaborn and Bokeh:

- Seaborn
  Seaborn can produce very complex visualizations, but requires matplotlib knowledge.
  Seaborn is strongly targeted towards the Analysis task.
  Seaborn is less flexible than Bokeh, but has a similarly steep learning curve.

- Bokeh
  Is very robust, but an overkill for simple visualizations.
  Bokeh is targeted somewhere in-between the analysis and presentation task.
  Bokeh is very flexible, but has a steep learning curve.

Seaborn and Bokeh are used in a static environment [are they? Bokeh can be used in the browser, supports interaction], while for instance D3 and Processing are made for web-interactive environments.

Javascript has also many charting libraries, for instance D3. Processing is a visualization language built on Java (that can use the JavaScript interpreter for web-based usage → processing.js).

Comparison of D3 and processing:

- Processing easier to learn and better for making quick prototypes.
- D3 is not suitable for quick prototyping.
- D3 has a steep learning curve, but a large community for getting help/ideas
- More tools available for D3
- Both libraries allow to publish results online
- Both libraries are heavily targeted towards presentation.
- Both libraries are highly flexible (more flexible than Seaborn and Bokeh).
- However D3 has an even steeper learning curve than Seaborn and Bokeh, while Processing has a less steep learning curve than the rest,

---

Question 31.
Name three applications and describe their differences.

Answer 31.
We will compare Tableau, MS Power BI, Excel, JasperSoft and QlikView followingly:
Tableau has a huge set of features in every area (except automatic analysis, where it's just has some features). It's very innovative towards perception workflow, data handling/management and infrastructure and has as it's target audience the whole set of upper management, reporting manager and data analyst. It is also highly flexible.

MS PowerBI on the other hand has also a huge set of features in every area. It is very innovative in most areas (except automatic analysis) and it's target audience is also the whole set of upper management, reporting management and data analysts.

Then there is Excel, which is targeted a little more towards Analysis. It is in comparison to Tableau and MS PowerBI way less flexible.

JasperSoft has a medium set of features for Perception workflow, infrastructure and data management, while the rest is very low. It has high innovations in Data management and some innovations in infrastructure and is primarily targeting the upper management.

QlikView has many features for complex datatypes, perception workflow and data management, and some innovations for perception workflow, infrastructure and data management. It is targeted mainly upper management and reporting managers.

Tableau and PowerBI have the same capabilities to execute, while QlikView and Jaspersoft are below. Jaspersoft is in the middle of completing its vision, next comes Tableau, then Qlik and MS BI has almost completed its vision.

---

2 Additional Material

Pixel-oriented techniques

(VDS_02, 19/20, p.30)

Pixel, smallest entity on screen, represents one data point. Color is used to identify the attribute? Different ways of arranging available.