Data Visualization Concepts

BINF4234

Prof. Dr. Renato Pajarola

Exercise and Homework Completion Requirements

1. Exercises and reading assignments are **mandatory** and they must be completed successfully to finish the class and get a sufficient passing final grade.
2. Exercises are graded coarsely into categories **pass** or **fail**.
   - *A **fail** is given to failed submissions and incomplete solutions, and no points are awarded.*
   - *A **pass** indicates that the exercise is sufficiently good to receive the corresponding points.*
   - *Late submissions (up to one day) will result in "-1" point.*
3. The five exercises give rise to the following point distribution: 2 – 3 – 5 – 5.
   - *A **minimum of 7 points** from all four exercises must be achieved to pass the module. Failure to achieve this minimum will result in a failing grade for the entire module.*
   - *Thus at least two exercises have to be correctly solved, and one has to be from the more advanced ones.*
4. We give **bonus points** for students who have completed more than 8 points from all the exercises.
   - *Thus **7 points** from the exercises is required, **8 points** is still normal passing, and **9 and above** would give 1 or more extra bonus points.*
   - *Only the bonus points can and will be added directly to the final grade.*
5. Do not copy assignments, tools to detect copying and plagiarism will be used.
   - *The exercise results are an integral part of the final course grade and therefore the handed in attempts and solutions to the exercises **must be your personal work**.*

Submission Rules

- Submitted code must compile and run without errors using the indicated Python environment, using the included libraries, packages and frameworks. If additional libraries/packages are needed, please specify in a 'readme.txt' file together with your submission.
- The whole project source code must be zipped and submitted before the given deadline, including the output results (saved in .html file or as a screenshot picture).
- Submit your .zip archive named *dvc_ex1_MATRIKELNUMBER.zip* (e.g. dvc_ex1_01234567.zip) through the OLAT course page.
- **Deadline is Sunday, 29 May 2022 at 23:59h**

Exercise 4

In this exercise, you have **two options** to choose from, one is to create an animated map visualization using Python and Bokeh, and another one is to participate in the vis guideline forum and submit a short report (This will be explained by Dr. Alexandra Diehl on Wednesday May 11th). You can receive the bonus points by finishing either one. You may find **option 1 on pages 2-4** and **option 2 on pages 5-8**.

Option 1: Map visualization for New York City taxi trip data. To be specific, you are expected to create two map plots plus an animation slider. One map presents the pickup trip information and the other map presents the drop off trip information, given the fact that one trip normally starts from one specific location on the map and end at another different location. On each map, the districts in NYC will be color coded and the color corresponds to the average number of passengers from the trips in a certain district.

Task1: Data Preprocessing.

T1.1: Basic processing. Follow the instruction in the skeleton code and process the raw data.

T1.2: Create two new data frames for pickup and drop off. The reason to do this step is to prepare two data sources separately for later visualization purpose.

T1.3: Create two new data frames for the number of trips per day for pickup and drop off. Carefully read the hint and be aware that these two data frames has a different shape (row, column).

T1.4: Merge the taxi and shape data, and build a GeoJSONDataSource for pickup data and drop off data separately.

Task2: Map Visualization.

T2.1: Build the GeoJSONDataSource for plotting, and define linear color mappers for each map.

T2.2: Map and circle plotting for drop off data and pickup data. We use the bokeh patches for map plot.

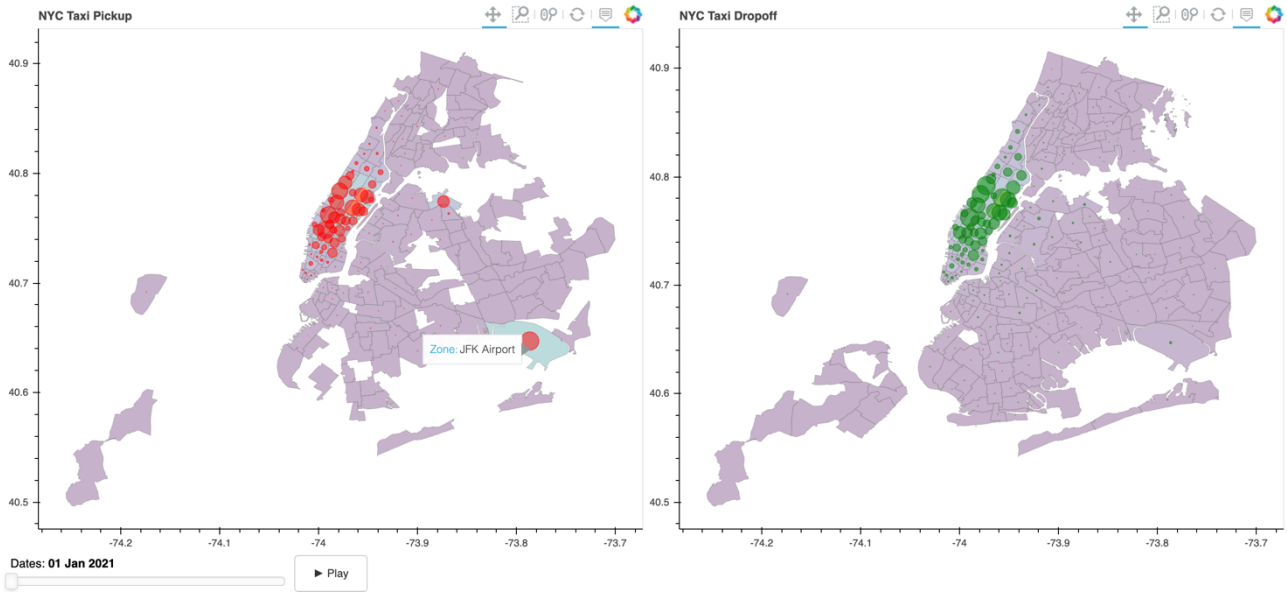
Task3: Add Interaction.

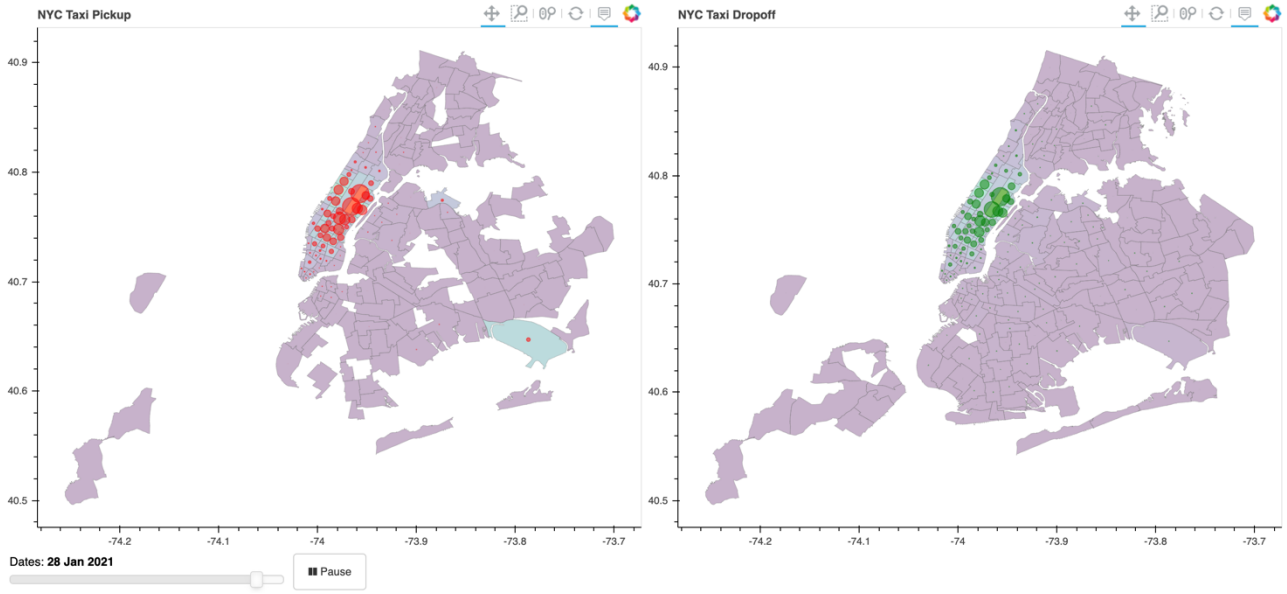
T3.1: Complete the callback function. In the callback function, you should update the circle size and then update the geosource.

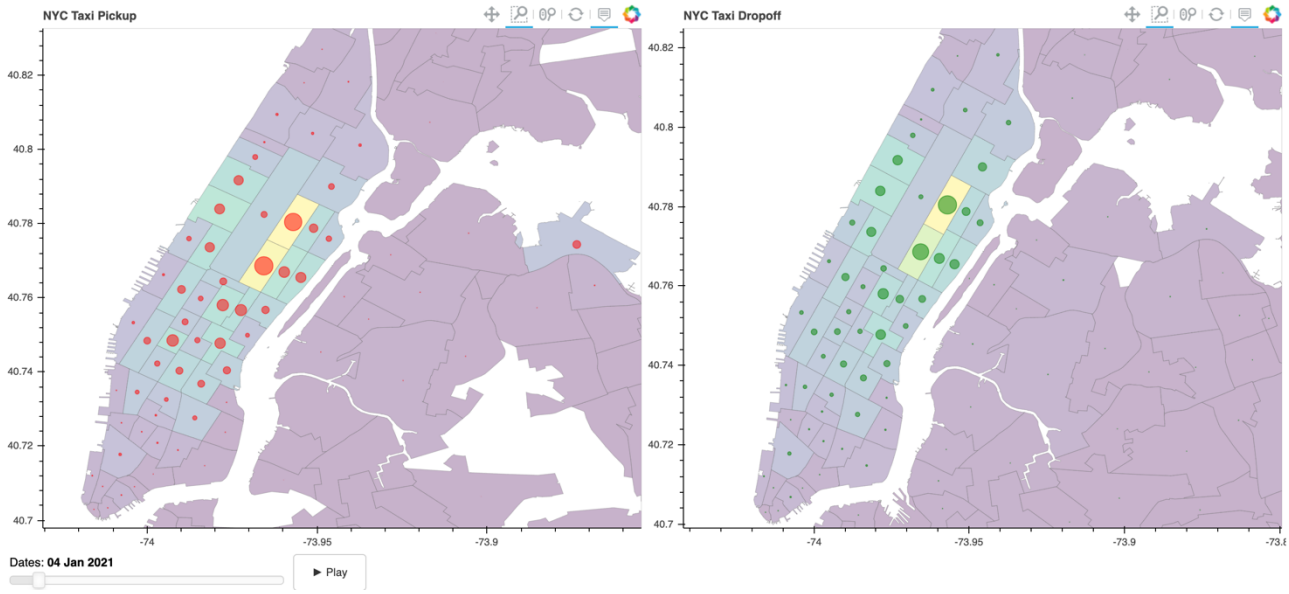
T3.2: Define the function for slider animation. Update the slider value with one day after current date, so the slider moves from left to right after clicking "Play".

T3.3: Complete the layout of the maps and animation slider. Save the result into a .html file.

The following pictures are an example for the desired (but not necessarily the same) visualization result:


NYC Taxi Pickup
NYC Taxi Dropoff
Zone: JFK Airport
Dates: 01 Jan 2021
▶ Play


NYC Taxi Pickup
NYC Taxi Dropoff
Dates: 28 Jan 2021
❚❚ Pause


NYC Taxi Pickup
NYC Taxi Dropoff
Dates: 04 Jan 2021
▶ Play

Remarks:

- In general, the code skeleton is well structured and divided into groups based on the tasks. However, you may want to change the structure of the skeleton for readability reasons of your own code.
- We recommend to use Jupyter Notebook for your implementation as it can visualize the intermediate output which helps for debugging. However, the final delivery of your code should be .py file rather than .ipynb.
- Try to make good use of the hints and references provided in the skeleton code. (**very important**)
- Try to google first for any Python related issues/bugs.
- Due to the special situation, we don't arrange in person meeting in this semester. Please contact the TA **Emine Didem Durukan (eminedidem.durukan@uzh.ch)** for technical questions regarding the exercise only if needed.
- More than one day late submission will not be accepted and graded.
- The deliverables of this exercise will be a clean version of your code with proper comments, any additional files necessary for executing it (for example, the data file), a "readme.txt" file for your comments or remarks (if necessary), dataset, as well as an export of the final output result in .html or .jpg/.png format. The absence of any required deliverable files will automatically lead to a **FAIL**.

Option 2: An Assignment on Visualization Guidelines

In the discipline of visualization and visual analytics, it is estimated that there are several hundreds of guidelines recorded in different books, research papers, blogs, and other communications. Some of such guidelines can be found at **VisGuides**. The application of many guidelines may be limited to some conditions that are yet to be specified explicitly, while some guidelines may be in conflict with others. In order to build visualization techniques on a solid theoretical foundation, it was suggested that the grounded theory methodology in social sciences can be used to study guidelines in visualization [Chen et al. 2017, DOI: 10.1109/MCG.2017.3271463] [Diehl et al. 2019, https://arxiv.org/abs/2010.09040]. "*It involves observing practical phenomena in the wild (to ground the theory in real-world data), identifying categories of the instances (events, processes, occurrences, participants, and so on), making links between categories, and establishing relationships between them. The method utilizes descriptive labelling (referred to as coding) to conceptualize discrete instances of phenomena systematically. It advocates continuous comparative analysis and negative case analysis to ensure the coding is comprehensive, meticulous, and up to date. It encourages researchers to interact with data by asking questions, broadening the sampling space by exploring related phenomena, and writing memos.*" The online discourse forum **VisGuides** offers a useful platform to enable open, democratic, evidence-based discourses in supporting the grounded theory methodology.

Assignment Specification.

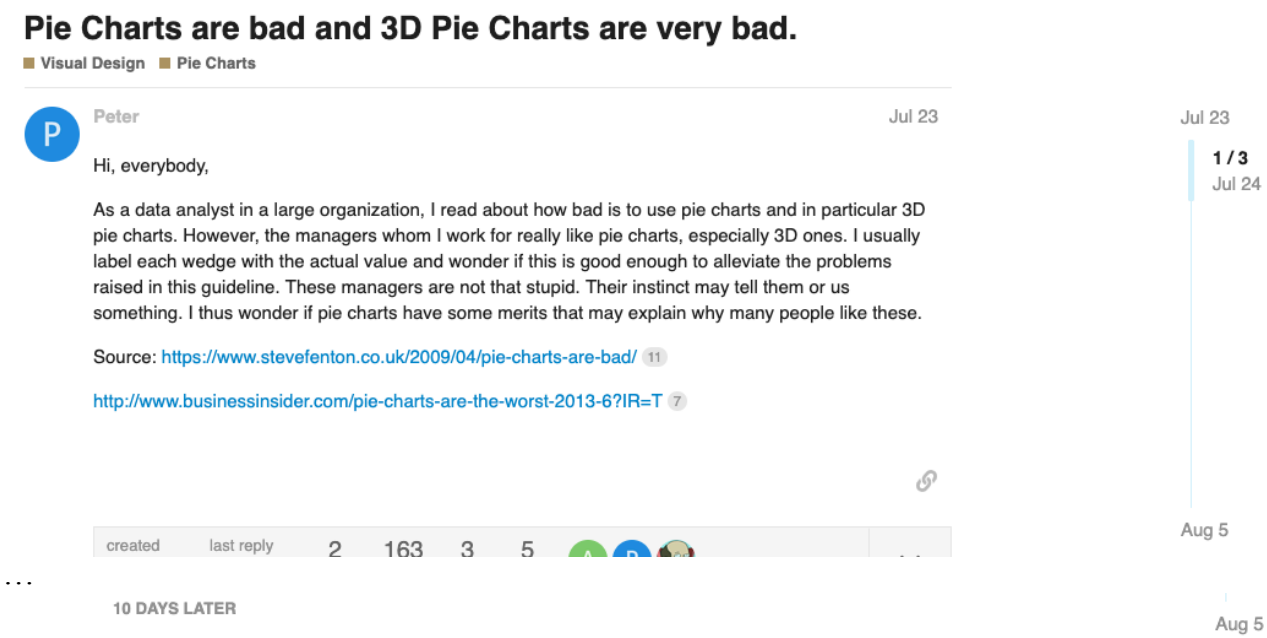
(1) You are required to participate in at least one of the ongoing discussion threads in the **VisGuides** forum. As many questions posed in the forum do not necessarily have a simple answer, you are encouraged to consider your views from different perspectives and offer your opinions confirmatively or critically. You may also start a new thread by posing a question or an assertion. Examples of the questions currently posed in the forum are attached at the end of this assignment.

(2) You are required to briefly reflect on how the discussion on the specific guidelines and to provide two real case scenarios of application? Justify why do you think that the guideline could be useful or not useful (in the case of counter-evidence). This point does not require participation in the VisGuides Forum.

Assessment. This assignment will be marked with three grades Excellent (5 points) when the two assignments are complete and very satisfactory, Good (3 points) when the two assignments are complete and just satisfactory, or Pass (2 points) when at least one assignment is complete and satisfactory, respectively.
When posting a question consider:

- The breadth and depth of the knowledge shown in your discourse.
- The level of reasoning, critique, and supporting evidence.
- The conciseness and effectiveness of the written communication, and the maturity of using academic language with appropriate vocabulary, grammatical structures, and technical terminology.
- Delivery: A document should be submitted to OLAT including: (1) a screenshot of the VisGuides and (2) a written reflection (100 to 500 characters paragraph). You don't need to use your real credentials at VisGuides. E-mail and personal data are confidential. Posts are public and available on Internet. Please check the Terms and Conditions and Data Privacy statement for further information at VisGuides.org.

Examples of Questions Posed in VisGuides.



Pie Charts are bad and 3D Pie Charts are very bad.

Visual Design Pie Charts

Peter Jul 23

Hi, everybody,

As a data analyst in a large organization, I read about how bad is to use pie charts and in particular 3D pie charts. However, the managers whom I work for really like pie charts, especially 3D ones. I usually label each wedge with the actual value and wonder if this is good enough to alleviate the problems raised in this guideline. These managers are not that stupid. Their instinct may tell them or us something. I thus wonder if pie charts have some merits that may explain why many people like these.

Source: https://www.stevefenton.co.uk/2009/04/pie-charts-are-bad/ 11

http://www.businessinsider.com/pie-charts-are-the-worst-2013-6?IR=T 7

Jul 23

1 / 3

Jul 24

Aug 5

created last reply 2 163 3 5

...

10 DAYS LATER

Aug 5

eagereyes Aug 5

Hi Peter. I've done research on pie charts and have written about them on my blog. First, we need to separate pie charts from 3D pie charts here. With a small number of slices, and for data that has a meaningful part-to-whole relationship, they work just fine. There's a bit more on this on my website 2 and I also gave a talk on some of my research a few years ago (search for "Unloved" on my website to find it, I can only post two links here as a new user).

3D pie charts are a different matter. 3D charts generally introduce lots of distortions that aren't related to the data at all, and produce large errors in reading data. I have a forthcoming short paper at VIS this year where I used the distortion created by 3D pie charts to investigate the perceptual mechanism used to read the charts. Figures 1 and 2 in the paper 2 should give you a sense of the distortion, and this is using parallel projection – perspective projection is a lot worse.

As to why people like them, my personal theory is that we inherently like round, complete shapes better than the jagged bar charts that usually work much better.

So in short, I'd recommend dropping the 3D and just going with 2D pies – assuming you don't have a lot of data points in each chart and the data fits the chart. Adding the values is a good idea, but it's really a cop-out. What purpose are the charts serving if people are just reading the numbers?

(Don’t / Do) Replicate the Real World in VR?

Visual Design

May 2018

1 / 9
May 2018

Apr 1

matthias.kraus May '18

Guideline: Don’t Replicate the Real World
Source: https://sites.umiacs.umd.edu/elm/2017/10/03/3d-visualization-for-nonspatial-data-guidelines-and-challenges/ 29

Question:
The overall message of the guideline given in the above blog is clear (Rule Nr. 7): For 3D visualizations for nonspatial data, it is unfavorable to replicate real-world controls. However, this seems to only cover 3D visualizations on the screen and not fully immersive visualization environments in VR. I think that in such environments, at least for some application cases, the exact opposite would be helpful: try to replica the real world as far as possible and extend/improve/adapt it to smooth away difficulties given by real environments (e.g., limitations given by screen sizes: the entire wall of the virtual environment could be used for display).
This would probably help the analyst to quickly orient in and interact naturally with the virtual environment without a steep learning curve. Of course, the new technology enables new ways to interact with the environment and we should exploit those possibilities, but I also think that it would increase the overall usability to create a familiar environment to the known real world and therefore use a “normal” room as the “base” of the virtual environment and implement visual metaphors.
What is your opinion on that?

• • •

elm May '18

The intention **was** to cover 3D visualization in general, regardless of whether it is shown on a flat screen or in an immersive environment. Perhaps there is a difference in understanding of what “replicating the real world” means, however. As I state in the blog post, I have seen far too many naive 3D user interface environments where people create a 3D depiction of an office in VR to somehow represent a 3D version of a desktop: to print something, just send it to the 3D printer in the corner of your office! It’s a bad idea, because it just adds back all of the physical limitations of the real world that are not necessary and do not contribute anything other than familiarity.

Your argument seems mainly to be about decreasing the learning curve while smoothing away the difficulties of the real world. I would argue that replicating the real world “as far as possible” is going to yield far too many of these difficulties to be useful in the first place. Familiarity will only get you so far. After all, we can’t rest our feet on our virtual Windows or OSX desktops, yet we are perfectly capable of understanding them as virtual workspaces after only a little instruction and training. The only exception is really when the data is truly representational of the real world, like for an architectural previsualization, in which case it **does** make sense to do this.

1 Reply

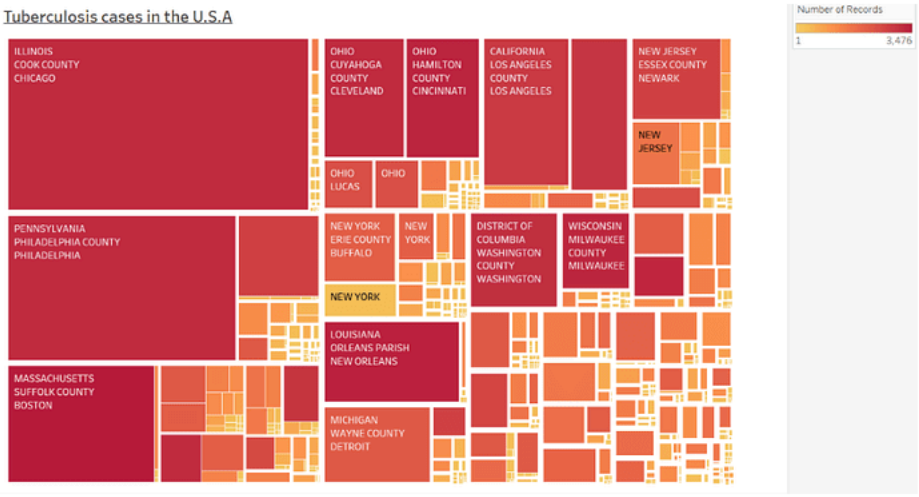
Visualizing TB cases from the USA using a hierarchical treemap

Visual Design

Anthony Mar 2

Hello, I am a Swansea University student (I am sure you've seen many of us around), and this my hierarchical treemap.

Tool: Tableau
Country: U.S.A
DOI: 10.25337/T7/ptycho.v2.0/US.56717001
Disease: Tuberculosis
Visual Mappings:
-Colour: Number of records
-Shape: Count Value
-Position: From the upper left to the lower right, the count value decreases.
-Hierarchy: State -> County-> City
Data Preparation: I have filtered all the results that would give the value "NA".



I was wondering if I could get any feedback regarding the visualisation of the data. I have noticed that the labeling is inconsistent and that there are some inequalities, especially looking at Illinois. However it was the best I could think of with the dataset given. I've tried reducing the number of states but the inconsistent labeling issue still occurs.

Thank you in advance.

created Mar 2 | last reply Mar 3 | 3 replies | 233 views | 3 users

min.chen Mar 2

What is "Count Value"? Is the 3rd level "City" encoded in the treemap? I guess that this treemap allows users to see (i) a large number of records vs a small number of records, (ii) a large count value vs a small count value, (iii) the name of the state with a large count value. Would the users want to see also, e.g., (iv) the name of the state with a large number of records, or (v) the ratio of number of records over count value? Perhaps these requirements are beyond what your class lecturer asked you to do.