Data Story Group Project Description

(Last updated: May 23, 2025)

This group project focuses on writing a data story for complex datasets. Our definition of a data story is a series of interlinked visualizations (either static or interactive) with the support of explanatory text or annotations. We urge all students to focus on creating a highly visual and/or interactive experience but not using visualization to support text.

We strongly recommend reading the "Narrative Visualization: Telling Stories with Data" paper by Edward Segel and Jeffrey Heer before doing this project to understand the process of telling stories with data visualizations. The distinction they make between author-driven and reader-driven stories is important in their analysis, as shown in the table below (taken from the paper mentioned above). It is up to you to find a good balance between them so that your stories are not completely reader-driven and also not fully author-driven.

Author-Driven	Reader-Driven
Linear ordering of scenes Heavy messaging No interactivity	No prescribed ordering No messaging Free interactivity

Language

You can choose to use English or Dutch. Do not interchange languages in the data story, as this will just cause confusion. It is OK to quote or use technical terms in another language (but do not write some paragraphs in English and others in Dutch).

Team size and forming

****You must form a team of 3 or 4 people to write ONE data story.**** You can team up with anyone in the course. We have a project registration milestone to facilitate the team forming process (see details in the "Milestones" section in this document).

****You should always aim for having a 4-person team since the data story project of 4 people will be evaluated in the same way as the team with 3 people.**** Having 4 people allows your team to be more flexible in distributing the workload. Also, if you have a 3-person team, we may assign a random person (who has no team) to your team.

****You are not allowed to do the project alone without team members.**** If you are alone after the project registration milestone, we will randomly assign you to a team. On one side, this project wants students to learn how to work in a team. Doing the project alone means losing this learning opportunity. On the other hand, this course has a high workload in management and grading due to the large number of students. Using group projects for assessment can reduce the workload for TAs and the course management team and thus make the course manageable.

If the collaboration goes very wrong to the point that you (or a part of your team) need to be separated to form a new team, your new and the original teams' workload will remain the same. This means your new smaller team will have the same requirements in the milestones as the team with 3-4 people. So, it is better to keep the collaboration going in some way (even though it is not working well) rather than splitting the team.

In principle, you are not allowed to have a team with only two people. Under some unusual situations, we may allow a different team size with an adjusted workload. For example, if there are only 2 people left while all other students have a team with 4 people, we may put these two people together to form a team with a reduced workload. If your team members somehow do not follow the course anymore in the middle of the project, and your team now has only two active people, we may reduce your team's workload. But in this case, you need to inform us ahead of time (at least 7 days before the final deliverable deadline), not at the late stage of the project.

Students are responsible for forming a team and communicating with team members. The course management team will only intervene if something goes very wrong. You can come to the seminars and use TicketVise on Canvas to find people. If you really cannot find a team, you can indicate that you want to be assigned to a team in the project registration milestone. To be clear, we cannot just randomly group people on Canvas, as some students may not be active in the course. On Canvas, there is no way to know who is not active, and putting non-active people in a team will lead to management difficulties.

By default, all team members will receive the same grade. In the case that your team's collaboration goes very wrong, you need to inform us ahead of time (at least 7 days before the final deliverable deadline) so that we can intervene, such as splitting your team into different parts. If you inform us late (less than 7 days before the deadline of the final deliverable), we will not be able to respond on time. We will definitely be unable to intervene if you let us know the problem after the deadline.

Tools and materials that you can use

You need to write the code in Python, and the visualizations need to be created through Python programming on Jupyter Notebook. You can freely choose the library that you want to use to create static or interactive visualizations. You can use Python Plotly (as shown in the second workshop) or others, such as seaborn (for static visualizations), Matplotlib, etc.

You can use graphics software (e.g., Adobe Photoshop/Illustrator, CorelDRAW, Microsoft PowerPoint), dashboarding tools (e.g., Microsoft Power BI, Google Looker Studio), or image generation models (e.g., Stable Diffusion) to create diagrams/photos to support your story. However, these will not count as data visualization components in the requirement of the data story project. **Any data visualizations that are not created using Python code will be ignored in grading.**

You can use photos to enhance your story. ****The photos that you use do not count as visualizations in the requirement of the data story project.**** If you use photos, you need to cite the sources and give the creators of the photos credit (e.g., by mentioning them in the caption). We will not deduct points for photos without credits or citations in this course. But you need to be careful about photo credits, as you may run into copyright problems in the future (e.g., when you work on other projects, or if you publish your work online publically).

You are allowed to use generative AI tools (such as ChatGPT) in the data story project. You do not need to report the usage of generative AI tools. Please check the "Policy for Using Generative AI Tools" section in the syllabus for details.

Possible topics

There are no restrictions on the type of topics. Below we curated a list of examples of data stories for inspiration, categorized by topics. Many of these examples use more complex techniques and software libraries to create the data stories, which is outside this course's scope. However, they can still serve as good inspirations and may point you to some interesting datasets that you can use in your project. You can also use these topics in your data story. Keep in mind that you can use multiple datasets and combine them into a story.

Global and local societal issues:

- The refugee crisis in five charts (by NOS news)
- More strikes in 2022, but fewer strikers (by CBS NL)
- Ask the question, visualize the answer (by FlowingData)
- The dark side of Guardian comments (by The Guardian)
- Trust and Social Connections in Times of Crisis (by World Happiness Report)

Climate change and environmental health:

- Deadly weather: the human cost of 2018's climate disasters (by The Guardian)
- Analysis of citizen-contributed smell reports in Pittsburgh (by CMU CREATE Lab)
- Trees are moving north from global warming (by The Washington Post)
- Anatomy of the Lismore disaster (by The Sydney Morning Herald)

Scientific and statistical insights:

- The Netherlands in numbers (by CBS NL)
- Hidden bias (by Google PAIR)

Humanity and culture:

- What is the most successful Hollywood movie of all time (by Information is Beautiful)
- Why do cats (by Nadieh Bremer)
- Figures in the Sky (by Nadieh Bremer)

If you choose a high-stakes topic (e.g., race, gender, ethnicity, human rights, social justice, immigration, war/conflict, hate/discrimination, etc.), please keep common/general good in mind and follow the UvA Code of Conduct. Please also be extra careful in the wording that you use and be prepared for difficult conversations using the framework in the link below. It is designed for teachers, but the framework can also be applied by anyone.

- Hard Questions: Learning to Teach Controversial Issues (by Judy Pace)

Multiple perspectives and arguments

****A key point of the project is that you think of it as a discussion, and therefore your data story must have multiple (at least two) different perspectives.**** These perspectives do not have to be opposing or contradictory. Please note that having only one perspective (or multiple very similar ones) will negatively affect your grade.

Perspectives are flexible for you to interpret. The reason that we have this requirement is that we want to encourage students to not just think about an issue from only a narrow view. Perspectives should be different and can offer diverse points of view on the same topic/issue, but they do not have to be contradictory or opposing. One way to think about this is that you can first come up with a list of possible stakeholders for your topic. Then, think about these stakeholders' perspectives by stepping into their shoes and empathizing with their situation. In this way, you can form perspectives based on stakeholders.

****Each perspective must have one or multiple arguments.**** Our definition of perspective is “an individual or a community’s point of view of a topic or issue.” Perspectives can differ among different groups of people with various backgrounds and beliefs. Our definition of argument is “a logical reasoning to support or justify a specific perspective.” Typically, arguments are used to persuade and convince people about the point that you are making, and thus, some evidence is required. In this course, we ask you to use visualizations as a primary way to communicate and support arguments.

You can choose to use either the exploratory or confirmatory approach. The exploratory approach focuses on using visualizations to discover patterns and relationships in the data so that you can form arguments. The confirmatory approach focuses on using visualizations to validate existing arguments or make them concrete. In the exploratory approach, you can use multiple perspectives to come up with multiple possible arguments that can enable follow-up research. In the confirmatory approach, you can use existing arguments (e.g., from scientific papers or from debates about social issues) and use visualizations to strengthen these arguments. You can also combine both approaches, for example, by firstly using the exploratory approach to identify several arguments and secondly using the confirmatory approach to strengthen the arguments.

Notice that the goal of having multiple (or even opposing) perspectives/arguments is not to lie or to manipulate the data. It is about summarizing the data differently, visualizing it differently, or substantiating an alternative point of view by placing the focus differently. You can also use multiple datasets to form different perspectives. But keep in mind that these different perspectives/arguments need to connect firmly back to the main topic of your story. Below, we provide some examples of these multiple perspectives/arguments:

Air pollution in the Netherlands:

- Perspective 1: Air pollution is a big problem in the Netherlands and requires immediate attention from the regulators.
    - Argument 1.1: Air pollution is highly related to chronic respiratory diseases.

- Perspective 2: Air quality is generally not a big deal in the Netherlands and does not require immediate action.
    - Argument 2.1: Air pollution measurements in the Netherlands remain low through many years and are typically below the EU regulation standards.

Climate change:

- Perspective 1: Climate change is real and it is a substantial problem that requires quick intervention before it is too late.
    - Argument 1.1: Climate change is related to the increasing frequency of natural disasters, which impact humanity negatively in the long term.
- Perspective 2: The hysteria about climate change is overblown, and there is no need to be so afraid about it.
    - Argument 2.1: Temperature changes on Earth follow a typical upward-downward pattern throughout history.

Expansion of the Schiphol Airport:

- Perspective 1: Schiphol Airport needs to be expanded as soon as possible.
    - Argument 1.1: More passengers and tourists are related to the boost of economic growth.
- Perspective 2: Schiphol Airport should reduce its number of flights.
    - Argument 2.1: Many people complain about the noise pollution in the neighborhoods, and a lot of the complaints are about flights.

Datasets

This section talks about common questions and expectations for getting datasets.

Where to find datasets?

In this project, your team needs to define the topic, form multiple perspectives, frame arguments for each perspective, and find the datasets. You may also need to merge datasets from multiple sources into a final “clean” dataset. One important learning goal is that you need to know how to search for datasets to fit your needs. Thus, we will not provide you with datasets. We suggest checking the following links to search for data. You can also find many other datasets on search engines, such as Google.

- Kaggle (search for code and datasets) – https://www.kaggle.com/datasets/
- Google dataset search – https://datasetsearch.research.google.com/
- Gemeente Amsterdam (e.g., for urban mobility data) – https://data.amsterdam.nl/
- CBS Netherlands (e.g., for census data) – https://opendata.cbs.nl/statline
- KNMI Netherlands (e.g., for weather data) – https://dataplatform.knmi.nl/
- Zenodo (search for datasets) – https://zenodo.org/
- GitHub (search for code and datasets) – https://github.com/
- Information is Beautiful (a list of datasets) – https://informationisbeautiful.net/data/
- The world bank – https://datacatalog.worldbank.org
- UC Irvine Machine Learning Repository – https://archive.ics.uci.edu/

How large should the dataset be?

There are no restrictions on dataset size. However, we recommend that you use the dataset with a sufficient number of data points. For example, you may need data for several months or years to argue that air pollution is an important issue. If there is one sensor reading per hour (e.g., the concentration of very small particles in the air) and we want one year of data, this means 24*365=8760 data points for each sensor station. You may need multiple sensor stations, such as 5, which will be 5*8760=43800 data points. The definition of “sufficient” depends on the application domains. For example, if your data story uses medical data, which is known to be hard to collect, your dataset size will probably be in the scale of hundreds but not thousands.

Also, keep in mind that for this project, avoid using datasets with millions of data records unless your team is very fluent in programming and has experience in dealing with large amounts of data. Processing a large number of data points can take a lot of time, and thus, you may run into the risk of not being able to complete the data story on time. If you really want to use large datasets in your data story, we recommend you filter the data by picking only a part of it (e.g., only the data for certain years) or aggregate the data in some way (e.g., aggregate to only one number per day) to reduce the number of data points.

One problem with using a very large dataset is that you will not be able to put it on GitHub when building the Jupyter Book (if you choose this way to submit the final deliverable). GitHub has limitations in file sizes, as documented on this webpage. If you have a very large dataset, try to remove the columns that you do not need or aggregate the rows (e.g., average sensor data within 4 hours into only one number) to reduce the dataset size. If you really need all the raw data, you can perform the analysis offline and save the analysis result only in CSV files. Then, upload the CSV files that contain your analysis results to GitHub (or in your zip file submission) to plot the visualizations.

Can we create new data for the data story project?

In the real world, often you may need to create and/or collect data to perform analysis. However, in this project, we strongly recommend that you use existing datasets for two reasons: one about the methodology and the other about the short timeline. First, this course is about visualization and not data collection. Creating and collecting data is complicated and requires advanced knowledge of the methodology. Without a proper understanding of methodology, you are highly likely to run into biases and data quality issues. Second, creating datasets takes a lot of time, typically way more than you expected. You may spend too much time in data generation/collection and not be able to finish the data story on time (or having a poorly written data story). Unless your team is very tech-savvy and has deep knowledge of data creation/collection methodology, we recommend using existing datasets.

Submission format

****Your submissions must meet the requirements described in this section. Failing to do so will have a significant negative impact on your grade (see the grading rubric).**

We are very strict about this. You also need to submit your work using Canvas, and we do not accept submissions via email.**

Your team's data story must be written in a Jupyter Notebook. You can choose to submit the story notebook in the following two methods:

- The first one is to submit all notebooks (including the main story and the data processing part) with the necessary files (e.g., datasets, photos) in a single zip file.
- The second one is to build a Jupyter Book and publish it online using GitHub Pages. Submit the URL as the final deliverable. You get bonus points for this method.

****For the first submission method, the main data story notebook must finish running within 1 minute.**** Always restart the kernel and run all the cells before submission to check the runtime and fix errors (if any). Sometimes, the results are kept in the cache and will cause errors on other people's computers.

****For the second submission method, the Jupyter Book URL must take less than 1 minute to load in a normal wireless network environment (e.g., the ground floor in the LAB42 building).**** Always put the URL in your browser and check if the website works as expected and fix errors (if any).

If the Jupyter Book URL for the final deliverable works without errors (i.e., everything can show without problems), we will give you an extra 0.6 point bonus (in the 10-point grading scale) for your final grade of the course. For example, if your final grade before the bonus is 7 (calculated based on the group project and also other assignments), you will get a 7.6 grade at the end. The first workshop of the course is designed to help you achieve this.

We also recommend submitting the zip file, even if you have a Jupyter Book URL. In this way, if your Jupyter Book URL does not work for unexpected reasons, we can still grade your original notebook in your zip file submission.

Any parts in the submissions with errors will be ignored during grading and not considered to satisfy the requirements or conditions in the grading rubric. The course management team and TAs will not notify or debug errors in your notebooks. Examples of errors may involve (but are not limited to) broken URL links, broken visualizations, missing files, code errors, infinite loops in the code, typos in variable names, etc.

You are expected to show that you have used existing data, modified/transformed the raw data, and then continue working with the modified/transformed data in the project. ****It is not allowed to hard-code any made-up data points, raw data, or transformed data in one or more variables in your notebooks.****

Suppose your original dataset is too large to be included in the zip file submission (or your GitHub repository). In that case, you can put the cleaned dataset file in the zip file submission (or your GitHub repository) and include a link to the place to download the original dataset (e.g., in the data story or the README file of your GitHub repository).

Flaws to avoid

Your data story will be assessed for argument insight, visualization design, readability coherence, and structure. ****Below are examples of the flaws that you should avoid, which will negatively impact your grade (see the grading rubric).****

Argument flaws

- ****AF1:**** Clearly nonsense arguments (e.g., making arguments using spurious correlations).
- ****AF2:**** Poorly supported arguments (e.g., making scientific arguments without the support from scientific evidence, explaining an argument without using visualizations or data).
- ****AF3:**** One-sided arguments without variety (e.g., mainly making the argument based on the finding that climate change is related to the change of global temperature).
- ****AF4:**** Shallow arguments (e.g., centering the argument around the finding that happiness is related to GDP and concluding that boosting GDP is the way to increase happiness).

Visualization flaws

- ****VF1:**** Clearly ineffective visual encodings (e.g., bad colormap design that does not make sense, such as using many different color hues to represent ordinal data).
- ****VF2:**** Inappropriate visualization figure size (e.g., a scatterplot with a small height that makes the Y-axis hard to interpret visually).
    - You may want to use large plots to make the visualizations clear. But do not make the plots too large in a way that they cannot fit into a typical computer screen. For example, making the plot height 5000 pixels is unreasonable, as most screens cannot fit a plot with 5000 pixels height. Here is a link about how to set figure size in Plotly.
- ****VF3:**** Confusing visualization design (e.g., plotting too many lines on a chart that makes the information hard to interpret, using too many colors that makes the graph looks very complicated, using colors without having a legend to explain what each color means, using colors that is not good for color-blind people, using a pie chart with too many pieces that look complicated).
- ****VF4:**** Non-meaningful interactive visualization (e.g., simply turning a static figure into an interactive one by using the default Plotly setting, non-effective exploration where readers do not really get insights from interacting with the data).
    - ****Meaningful interactive visualization**** means that the interaction has to make sense and is necessary in the storytelling. This means that the reader needs to get some insights from exploring the data by themselves when compared to only using a static visualization, such as using drop-down menus, timeline slider, animations, etc.
    - Simply making the plot interactive does not always make it meaningful. For example, without guidance, a map-based plot can be confusing to zoom and explore, as it is hard to know where to start. To make it meaningful, you can

add a timeline slider to hint that there can be differences in the data across years. Another example is a parallel coordinate plot, which also does not count as a meaningful interaction, even though users can drag the bars. You can add a drop-down menu that guides the users in reconfiguring the order of variables with various settings, which allows users to think more, and thus, this will be considered meaningful.

- **VF5:** Too many (more than 8) or too few (less than 6) visualization components.
    - **Visualization component** means anything that the plotting library (e.g, Plotly and seaborn) can do in a figure (e.g., plotly.graph_objects.Figure or matplotlib.figure). You can use multiple traces or subplots (e.g., plotly.subplots.make_subplots or matplotlib.pyplot.subplot) in one visualization component, but they count as one component (not multiple).
    - A visualization component can have multiple charts (and even with different types). You can use different types of charts in subplots to explain the same argument, but they are counted as one visualization (but not multiple).
    - Charts are defined as basic graphical representations of data, such as bar chart, line chart, pie chart, scatter plot, geographic map, heatmap, box plot, contour plot, donut chart, area chart, bubble chart, trees, radar chart, waterfall chart, violin plot, word cloud, histogram, etc.
    - Vertical bar chart, horizontal bar chart, stacked bar chart, grouped bar chart all count as the same type: bar chart.
- **VF6:** Lack of variety/richness in the visualization components (e.g., mainly using simple charts that visualize single variables, use similar charts such as violin chart and box chart, no multivariate visualizations, having too few types of charts).
    - Each visualization component should make a different point in representing and communicating information.
    - **Multivariate visualization** means charts that visualize at least two different variables in the dataset. A time series variable only counts as one variable, but not multiple ones.
    - The definition of similar charts include (but are not limited to) below:
        - Pie charts and donut charts are similar.
        - Bar charts, waterfall charts, and histograms are similar.
        - Bubble charts and scatter plots are similar.
        - Violin charts and box charts are similar.

Readability flaws

- **RF1:** Erroneous write-up to explain the arguments (e.g., grammar errors, incomplete sentences, very long and hard-to-read sentences).
- **RF2:** Bad caption and title for the visualization figures (e.g., no caption/title for figures, or the caption is not self-contained, or the title/caption is not informative).
    - **Self-contained figure caption** means that readers can get the necessary information (e.g., data insights) without going to the text body to find the part that describes the visualization.
- **RF3:** Bad flow between sentences (e.g., sentences in paragraphs are not connected well, or sentences are not connected to the argument, or some technical terms are mentioned without explanation).

- **RF4:** Bad structure (e.g., paragraphs do not connect well, or paragraphs were clearly written by separate people and later glued together in a poor way).
- **RF5:** Bad connections between the visualizations and the text (e.g., visualizations and the text are disconnected, or paragraphs do not support the visualizations).
    - Keep in mind that in this project, you should use text to support the visualization but not use the visualization to support the text.
- **RF6:** Mix multiple languages in the story in a way that is confusing (e.g., some paragraphs in English and some paragraphs in Dutch).
- **RF7:** Too many (typically more than 2500) or too few (typically less than 1500 words) words in the data story.
    - This word count includes all the text in the deliverable (e.g., including figure captions and titles), excluding the references (i.e., the section for putting the sources when you cite them).

Data story structure

Your data story for the final deliverable should have an ****overall title**** that describes your topic (typically at the top position of the notebook) and the sections below.

- A section with title: ****Introduction****
    - Describe the topic, perspectives, and arguments.
- A section with title: ****Dataset and Preprocessing****
    - Provide information about the datasets (e.g., where to download them, what are the variables in there) and also explain how you preprocess the data.
- One or more sections that describe your perspectives and arguments using visualization components and text. We also recommend adding a “Summary” section (optional) to conclude the perspectives and arguments.
- A section with title: ****Reflection****
    - Describe how the feedback from your TA and/or your peers helps you improve your data story. Also, describe self-reflections (if any) using the design guidelines about how to evaluate visualizations that we taught in the lectures.
- A section with title: ****Work Distribution****
    - Explain the distribution of work among the team members (i.e., describing who is responsible for what).

To give you a more clear idea about what the final deliverable may look like, we have created the following template to show the structure. You can also start your project using the template. Notice that the content is mostly dummy text, and the structure is just an example. Besides the required sections that are mentioned above, you are free to decide how many other sections you want to include in the data story final deliverable.

- https://multix.io/data-story-template/

We strongly recommend you wrangle and transform the original dataset(s) into a “cleaned” dataset and then work on the data story from that step. About the code that you used for pre-processing and cleaning the original dataset(s), you can put them in a separate file to demonstrate that you wrote the code. For some teams that have large datasets, you can process your raw data further and only store the necessary results for plotting visualizations.

Then, you can load the results (e.g., in CSV) in your notebook. In this way, you can avoid having the code take a lot of time to preprocess the dataset in your main story notebook.

At the very least, you should clearly describe the variables that you used to create the visualizations in the "Dataset and Preprocessing" section. Ideally, you should describe the variables in your dataset clearly, as you did in the proposal milestone when creating dataset descriptions. However, we understand that your dataset may have a large number of variables, and it would not make sense to report all of them. So, in this case, describe variables in a general and high-level way so that your readers can have a good understanding of what is inside your dataset. If you aggregate the variables in some way to create the visualizations, you should mention how you aggregate them and what the final aggregated variables mean.

Data story example

Below is a data story example from the previous year's course that gets a full score (based on the grading rubric in 2023). We have obtained the authors' consent to share the story. We strongly recommend you check the example to understand our expectations.

- Peter Adema, Aize van Basten Batenburg, Wim Berkelmans, and Kim Koomen. (2023). The salary gap in software development. University of Amsterdam, data story project of the information visualization course. https://p-adema.github.io/info-vis/ (fork — https://multix.io/info-vis-course-2023-example/)

The perspectives and arguments that the story made can help you understand their definitions. ****However, please keep in mind that the story structure in the example above may be different from the required structure for this year's course**.** Follow the instructions in the "Data story structure" section in this document for this year's expectations.

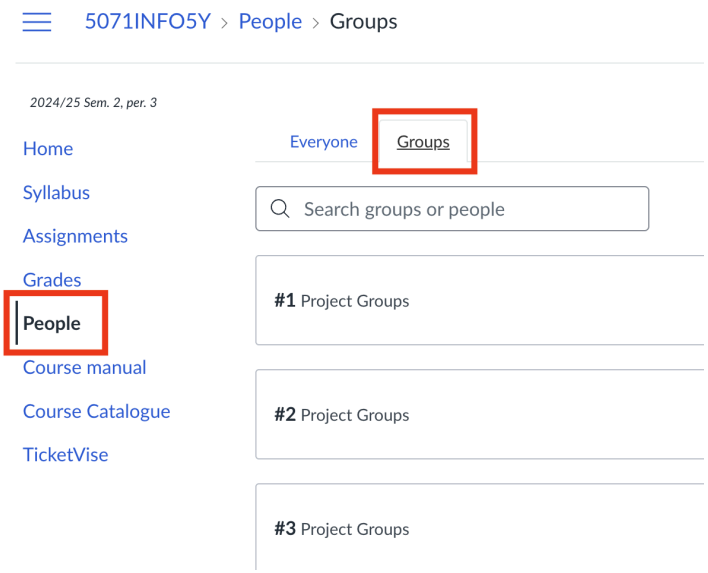
Milestones

The data story project consists of the following milestones. Check the syllabus for their deadlines and the times/dates that activities will happen.

M0: Registration

Register your project team members on Canvas. This step is to facilitate administration. Initially, we will not assign students to laptopcollege sessions. After registration, we will put members in the same team in the same laptopcollege session. ****You need to submit the registration to Canvas using the given template in the Canvas Assignments.****

We strongly recommend that you form a team of 4 students (see the "Team size and forming" section in this document). You need to self sign-up your team on Canvas, under the People -> Groups, as shown in the figure below. ****You only need to self sign-up if you have a 3-people or 4-people team.**** In other situations, indicate that you want to be randomly assigned to a team. Notice that a 2-people team is not allowed.



If you want to be assigned to a random team, you can indicate this in the project registration form template. We will later place you randomly into another team or put you together with other students who have no team. We will assume that you want to be assigned to a random team under the following situations:

- If the course management team cannot understand whether you have a project team or want to be assigned to a team based on the information in the registration form
- If you did not submit the registration form
- If you submitted the form later than the deadline (in this case, you may get a partial score due to the late submission policy, but we will assume that you have no team)
- If you form a two-people team, which is not allowed

M1: Proposal

Provide information about your team, the topic that you want to work on, the dataset that you will choose, and an initial analysis of the dataset. ****You need to submit the proposal to Canvas using the given template in the Canvas Assignments.**** The proposal is meant to ensure that you have everything in place to work on the data story project.

****For the project proposal, you need to do the following:****

- Discuss with your team members to come up with your project title, topic, and at least two different perspectives. For more information, check the “Multiple perspectives and arguments” subsection in this document.
- Find at least 2 datasets that your team wants to explore and describe these datasets by following the instructions in the template.
- Work with your team members to download, load, and clean these raw datasets using Python. Then, save the cleaned dataset to a file so that it can be later loaded using Python pandas. Include a screenshot of the result of running pandas.DataFrame.head(n=5) on your dataset in the submission.

M2: Draft

Submit your draft to Canvas. The draft submission is the version that you will use for the next milestone during the discussion and peer feedback moment. The more complete your draft submission is, the more in-depth discussion and feedback you can have.

****The minimum requirements for this draft submission are as follows:****

- Meet the requirements in the “Submission format” section.
- ****Argument insight requirements:****
  - You need to have at least two perspectives, and each perspective needs to have one or multiple arguments. Include a brief description for each perspective and argument (e.g., a few short sentences or bullet points). Indicate which perspective and argument will relate to which visualization.
- ****Visualization design requirements:****
  - You need to have at least 6 visualization components. At least 4 of them need to be nearly finished. Other visualizations need to have sketches or draft versions. They may change after the feedback, but they should be ready for your final deliverable.
- ****Structure requirements:****
  - The “Introduction” section should be complete with the content that describes the topic and your perspectives. Notice that you are writing the data story for a general audience (not domain experts), and therefore you need to explain the story in a way that lay people can understand.
  - The “Dataset and Preprocessing” section should be complete with the content that describes the dataset(s) and also how you preprocess them.

These are minimum requirements. It helps to have the visualizations, perspectives, and arguments already ordered in the manner in which they will be presented in your final data story. The further along you are with your data story the more detailed feedback you can receive. Also, if your argumentation and data is ready for the data story then you can focus more on the aesthetics and the functionality in the last week.

M3: Discussion and peer feedback

Discuss your work and give feedback on other teams’ work during the seminars. You need to document the feedback, as you need to write a section in the final deliverable to reflect on how the feedback from your TA and your peers helps you improve your data story.

Check the syllabus for the date of the seminar that the discussion and peer feedback moments will happen. If there are time conflicts, you need to discuss an alternative plan with your TA and check if your TA can arrange a separate time. But TAs do not have the obligation to provide you extra time.

During the discussion and peer feedback, use the “Communication Principles” as mentioned in the course syllabus. We aim to create a safe environment for discussions. When you give feedback to others, think about if you also want to receive that kind of feedback from others.

Discuss your draft during seminars

You need to discuss your draft during a seminar to get feedback from your TA and peers. ****During the discussion, at least one team member needs to attend the seminar that we mentioned in the syllabus (i.e., the two scheduled time slots for the "Discussion and peer feedback" milestone) to show the work to the TA and receive feedback.****

However, we strongly encourage all team members to attend the discussion. The discussion is not for the entire seminar room but only for your TA and the students (in other words, it is not a formal presentation). The discussion time should be 10 minutes for presenting your draft, which should be followed by a 10-minute Q&A. Your TA will assess your discussion and submit the grade to Canvas.

Give oral peer feedback during the seminars

You need to give oral peer feedback to another project team. ****At least one member from your team must attend at least one discussion of another data story project and give them oral feedback during the seminar that we mentioned in the syllabus (i.e., the two scheduled time slots for the "Discussion and peer feedback" milestone).****

Your TA will keep track if your team gives oral feedback to at least one other team or not, which will be reflected in your grade for this milestone. We strongly encourage all team members to attend the discussion to give feedback to the other team(s).

M4: Final deliverable

Submit your final deliverable to Canvas. ****Your submission must meet the requirement in the "Submission format" section.****

Below are recommendations, and they are not strict. However, failing to do so can have a negative impact on your grade (see the "Flaws to avoid" section in this document).

- Have a minimum 6 and maximum 8 visualization components.
- Have a minimum 1500 and maximum 2500 words with clear writing.
- Have a caption and title for each visualization that clearly explains how the visualization works and the take-away messages. For example, for a line chart, describe what the x-axis and the y-axis mean and their scale. You also need to describe the insights that you expect the readers to get from reading the line chart (i.e., the take-away message). We recommend 50-100 words for each caption.
- Have at least 1 multivariate visualization (see the definition in the "Visualization flaws" section).
- Have at least 3 different types of charts among all the visualization components. See the definition of "visualization component" and "chat" in the "Visualization flaws" section. One visualization component can have multiple charts.
- Have multiple perspectives and arguments. See the "Multiple perspectives and arguments" section for details. Each perspective should have one or multiple arguments that are supported by visualizations and text.
- Follow the structure instructions in the "Data story structure" section.

Having no interactive visualizations (or a poorly designed one) can still get you a sufficient grade. However, we strongly encourage you to craft at least one meaningful interactive visualization, which will be rewarded with higher grades (see the grading rubric). Refer to subsection “Visualization flaws” for the definition of meaningful interaction.

Although we are not grading the code, you still need to demonstrate that you use programming to generate the visualizations. In principle, the final deliverable should contain code blocks. But, if you are using the Jupyter Book URL as the submission format, you can use the feature to hide the code blocks (by making it a dropdown button) to make the story look nice and compact. Please make sure that we can still find your code. For example, you can include your GitHub repository link in the introduction section of the final deliverable.

Grading rubric

The maximum point for the entire data story project is 100, which is a sum of the points for all milestones (M0, M1, M2, M3, and M4).

M0: Registration (2 points maximum)

The maximum point for this milestone is 2, and the minimum point is 0.

Pass	Fail
The information is provided correctly so that the course management team can understand clearly whether you are already on a team or want to be assigned to a random team. You either registered a team with 3 or 4 people, or you indicated that you want to be assigned to a random team. If you have a team, all team members are registered correctly on Canvas, under the People -> Groups, as shown in the data story project description document. You submitted the form to the correct registration assignment. If you have a team, the submission should go to the “Type 1 Registration” assignment. If you do not have a team, the submission should go to the “Type 2 Registration” assignment. If you have a team, your team only makes one submission that represents all team members. +2 points	You did not submit a project registration assignment. The information is unclear in a way that the course management team is not sure if you have a team or want to be randomly assigned. If you have a team, your team did not register (self sign-up) the team members on Canvas, or your team registered the team members incorrectly (such as missing team members). You submitted this form for the wrong registration assignment on Canvas. For example, you have a team, but you submitted the form to the “Type 2 Registration” category, or vice versa. Or, you submitted the form for both “Type 1 Registration” and “Type 2 Registration” categories. If you have a team, you did this in a way that led to multiple submissions (which is likely that you forgot to self sign-up team members before submitting the assignment to Canvas). You registered a team with 2 team members, which is not allowed. 0 points

M1: Proposal (8 points maximum)

We will grade your proposal based on completeness, readability, coherence, and correctness. The total grade for this milestone consists of the sum of the points obtained for all the criteria. The maximum point for this milestone is 8, and the minimum point is 0.

Criteria	Pass	Fail
Completeness	All boxes in the “Project information” and “Dataset information” sections in the proposal template are completed. +2 points	One or more boxes in the “Project information” and “Dataset information” sections in the proposal template are missing, have no information, have unmeaningful information, or duplicate the answers to other questions. 0 points

Readability	The proposal looks neat and professional, is written clearly, and does not contain any (or more than a few) writing errors. +2 points	The proposal looks untidy, is not written clearly, or contains more than a few writing errors. 0 points
Coherence	Most chosen datasets, variables, and the planned visualization components are in general related to the topic, perspectives, and/or arguments. +2 points	Most chosen datasets, variables, and the planned visualization components connect poorly to the topic, perspectives, and/or arguments. 0 points
Correctness	All descriptive statistics (given the chosen variables) are appropriate to use. The proposed visualization components sound appropriate and suitable for the chosen variables. +2 points	Any of the following happens: ● One or more descriptive statistics (given the chosen variables) are missing, incorrect, not appropriate, or not suitable to use. ● One or more proposed visualization components are missing, incorrect, not appropriate, or not suitable for the chosen variables. 0 points

M2: Draft (10 points maximum)

We will grade your draft based on argument insight, visualization design, and structure. The total grade for this milestone consists of the sum of the points obtained for all the criteria. The maximum point for this milestone is 10, and the minimum point is 0.

Criteria	Pass	Fail
Argument insight	Satisfy all the argument insight requirements in the “M2: Draft” section Must satisfy the requirement in the “Submission format” section in this document. +3 points	Any of the following happens: ● Does not satisfy all the argument insight requirements in the “M2: Draft” section ● Fail to meet the requirements in the “Submission format” section in this document. ● Hard-code any made-up data points, raw data, or transformed data in one or more variables in your notebooks. 0 points
Visualization design	Satisfy all the visualization design requirements in the “M2: Draft” section Must satisfy the requirement in the “Submission format” section in this document. +4 points	Any of the following happens: ● Does not satisfy all the visualization design requirements in the “M2: Draft” section ● Fail to meet the requirements in the “Submission format” section in this document. ● Hard-code any made-up data points, raw data, or transformed data in one or more variables in your notebooks. 0 points
Structure	Satisfy all the structure requirements in the “M2: Draft” section	Any of the following happens: ● Does not satisfy all the structure requirements in the “M2: Draft” section ● Fail to meet the requirements in the

	Must satisfy the requirement in the “Submission format” section in this document. +3 points	“Submission format” section in this document. ● Hard-code any made-up data points, raw data, or transformed data in one or more variables in your notebooks. 0 points

M3: Discussion and peer feedback (10 points maximum)

The total grade for this milestone consists of the sum of the points obtained for all the criteria. The maximum point for this milestone is 10, and the minimum point is 0.

Criteria	Pass	Fail
Discussion	At least one team member attends the seminar that we mentioned in the syllabus (i.e., the two scheduled time slots for the “Discussion and peer feedback” milestone) to discuss the data story with the TA. +5 points	No team members attend the seminar that we mentioned in the syllabus (i.e., the two scheduled time slots for the “Discussion and peer feedback” milestone) to discuss the data story with the TA. 0 points
Peer feedback	At least one team member attends the seminar that we mentioned in the syllabus (i.e., the two scheduled time slots for the “Discussion and peer feedback” milestone) to give at least one other project team oral feedback. +5 points	No team members attend the seminar that we mentioned in the syllabus (i.e., the two scheduled time slots for the “Discussion and peer feedback” milestone) to give at least one other project team oral feedback. 0 points

M4: Final deliverable (70 points maximum)

We will grade your final deliverable based on argument insight, visualization design, readability coherence, and structure. The total grade for this milestone consists of the sum of the points obtained for all the criteria. Grading details are in the following subsections. The maximum point for this milestone is 70, and the minimum point is 0.

The design of the rubric is that students should get a “Sufficient” grade if they spend a reasonable amount of time and effort. To reach a “Good” grade, students need to spend extra effort to make sure that the data story does not suffer from the flaws described in the “Flaws to avoid” section in this document. To reach an “Excellent” grade, students need to spend substantial effort and time creating advanced visualizations, polishing the writing, and researching their data story topic to make strong perspectives/arguments.

Argument insight

Excellent +21 points	Have very convincing arguments based on the strong interplay between the text and the visualizations. Have great depth due to strong original contributions and critical views. Explain multiple different and diverse perspectives very well, and each with multiple convincing arguments. The arguments are supported by scientific published literature with citations. Arguments have no flaws (or the flaw is extremely minor). Must satisfy the requirements in the “Submission format” section in this document.
Good +15 points	Have strong arguments based on text or visualizations. Text and visualizations are reasonably connected. Have good depth through some original contributions and good distinctive views. Explain multiple perspectives well, and each with at least one good argument that may be supported by scientific evidence. Perspectives are reasonably different. Arguments have no flaws (or the flaw is extremely minor). Must satisfy the requirements in the “Submission format” section in this document.
Sufficient +12 points	Have reasonable arguments based on visualizations. Text and visualizations may not be connected well. Have some original contributions with distinctive views, but less depth due to less critical views. The views may remain at the abstract and surface level. Explain multiple perspectives reasonably, and each with at least one argument. Perspectives may not be distinctive. Arguments may have some minor flaws but no major flaws. Must satisfy the requirements in the “Submission format” section in this document.
Insufficient +9 points	Have missing arguments several times. Visualizations are not very convincing and often do not support the text. Have little depth due to weak personal contributions, a lack of critical eye, and insufficiently distinctive views. Explain only one perspective, or explain multiple ones without the support of reasonable argument. Arguments have many minor flaws and may have major flaws. Must satisfy the requirements in the “Submission format” section in this document.
Fail 0 points	Any of the following happens: ● Have factual errors in the story. Visualizations and text are not connected at all. Use primarily personal opinion without support of evidence. A critical view and a distinctive point of view are missing. ● Perspectives are missing or not clear. Arguments are poor, missing, or have factual errors. ● Arguments have major flaws. ● Fail to meet the requirements in the “Submission format” section in this document. ● Hard-code any made-up data points, raw data, or transformed data in one or more variables in your notebooks.

See the “Argument flaws” subsection in this document for the definition of flaws in this grading criterion.

Visualization design

Excellent +28 points	Use excellent visualizations with high richness and variety. There are at least 4 different types of charts, at least 2 multivariate visualizations, and at least 1 meaningful interactive visualization. Visualizations are thoughtfully and elegantly designed, including carefully chosen colors and layouts. The visualizations themselves, and the data story as a whole, look very attractive visually. The visual design is excellently used as the main component in directing the reader and providing the arguments. Visualizations have no flaws (or the flaw is extremely minor). Must satisfy the requirements in the “Submission format” section in this document.
Good +20 points	Use good visualizations and design with good richness and variety. There are at least 3 different types of charts, at least 1 multivariate visualization, and at least 1 interactive visualization. The interactive visualization may not be very meaningful. Visualizations are reasonably designed. The visualizations themselves, and the data story as a whole, look visually appealing. The visual design is reasonably used as the main component in directing the reader and providing the arguments. Visualizations have no flaws (or the flaw is extremely minor). Must satisfy the requirements in the “Submission format” section in this document.
Sufficient +16 points	There is a clear role for the visualizations and design. There are at least 3 different types of charts, and at least 1 multivariate visualization. There may be no interactive visualization, or the interactive visualization is not meaningful. The visualizations themselves, and the data story as a whole, look sufficiently well cared for visually. The visual design is often used in directing the reader and providing the arguments, but may not work well for some parts of the story. Visualizations may have some minor flaws but no major flaws. Must satisfy the requirements in the “Submission format” section in this document.
Insufficient +12 points	Visualizations and design are not used properly. There are only 1 to 3 different types of charts. And there may even be no multivariate visualization. There may be no interactive visualization, or the interactive visualization is not meaningful. The visualizations themselves, and the data story as a whole, do not look visually well cared for and have major flaws. The visual design works poorly in explaining the arguments. Visualizations have many minor flaws and may have major flaws. Must satisfy the requirements in the “Submission format” section in this document.
Fail 0 points	Any of the following happens: ● The visualizations and design are poorly used. There are only 1 to 3 different types of charts. And there may even be no multivariate visualization. There may be no interactive visualization, or the interactive visualization is not meaningful. ● The visualizations themselves are incorrect or uninterpretable, and the data story does not look neat. The visual design is disconnected from the arguments. ● Visualizations have major flaws. ● Fail to meet the requirements in the “Submission format” section in this document. ● Hard-code any made-up data points, raw data, or transformed data in one or more variables in your notebooks.

See the “Visualization flaws” subsection in this document for the definition of flaws and meaningful interactive visualization in this grading criterion.

Readability coherence

Excellent +14 points	The story is easy to understand without knowing the dataset. A well-running and excellent story with a clear structure is written in your own words. The perspectives are used as a common thread throughout the entire story. The story has a very high quality and is close to the level that can be published in major news articles. Readability has no flaws (or the flaw is extremely minor). Must satisfy the requirements in the “Submission format” section in this document.
Good +10 points	The story can be understood without knowing the dataset. In your own words, a clear story has been written with a clear structure. The perspectives are used as a common thread throughout the entire story. The story has a good quality and requires some tweaking to be able to be published on personal blogs/websites. Readability has no flaws (or the flaw is extremely minor). Must satisfy the requirements in the “Submission format” section in this document.
Sufficient +8 points	The story can be understood without knowing the dataset. The story is understandable and written in your own words that, except for a few parts, runs well and has a clear structure. The perspectives are used as the common thread for most of the story. The story has an okay quality and requires much editing before it can be published online. Readability may have some minor flaws but no major flaws. Must satisfy the requirements in the “Submission format” section in this document.
Insufficient +6 points	The story cannot be fully understood without some prior knowledge about the dataset. A story has been written that does not flow well, with a structure that is difficult to follow. The perspectives are not used as a common thread throughout the story. The story has a poor quality and requires a significant amount of editing or re-writing from experts to reach a publishable level. Readability has many minor flaws and may have major flaws. Must satisfy the requirements in the “Submission format” section in this document.
Fail 0 points	Any of the following happens: ● The story cannot be understood without studying the dataset. A story has been written that does not read well and is unclear in content and structure. The perspectives are not used as a common thread throughout the story. The story has a bad quality and will not even be considered publishable. ● Readability has major flaws. ● Fail to meet the requirements in the “Submission format” section in this document. ● Hard-code any made-up data points, raw data, or transformed data in one or more variables in your notebooks.

See the “Readability flaws” subsection in this document for the definition of flaws in this grading criterion.

Structure

Pass +7 points	Meet all structure expectations in the “Data story structure” section. Must satisfy the requirements in the “Submission format” section in this document.
Fail 0 points	Any of the following happens: ● Fail to meet all structure expectations in the “Data story structure” section (e.g., some sections are missing) ● Some sections are very shallow and not described in detail (e.g., unclear work distribution, poor reflection of the work, unclear descriptions of dataset and preprocessing methods) ● Fail to meet the requirements in the “Submission format” section in this document. ● Hard-code any made-up data points, raw data, or transformed data in one or more variables in your notebooks.