DS 544/444

Data Visualization - Fall 2024

Instructor: Dr. Dumindu Samaraweera

Team Project Instructions and Guidelines

Competence and Skill Assessment for Data Visualization Team Project

* This guide provides comprehensive information from project initiation to completion. Please read the entire document thoroughly.

1. TEAM FORMATION

Each team can have up to four students. Please find your teammates and inform me as soon as possible. Pairs of teammates with complementary skills from different majors will be combined to form a team. Depending on the requirements, with prior approval, a team may consist of 3 or 5 members.

2. PROBLEM SELECTION

There are three options to select your data visualization project topic:

a) All team members agree on a project of their own choice using your own dataset. If the problem has been previously studied (even if it is as research in progress), you must reference the prior work and clearly identify your new contribution; failure to do so will be considered plagiarism.

b) You may propose a problem that interests all team members, and I will assist you in finding an appropriate dataset.

c) Alternatively, you can choose a project from the list provided in the shared folder, and we will discuss how you can make a unique contribution based on previous research.

3. DATASETS

a) Most Popular Publicly Available Datasets:

i. Scikit-learn's built-in datasets: https://scikit-learn.org/stable/datasets

ii. Kaggle datasets: https://www.kaggle.com/datasets/

iii. UCI (University of California, Irvine) Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.php

iv. O*NET Occupational Database: https://www.onetonline.org/find/descriptor/browse

v. Finance Datasets: https://finance.yahoo.com/

vi. Bureau Of Transportation Statistics: https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp https://www.bts.gov/topics/airlines-and-airports/airline-time-performance-and-causes-flight-delays

vii. ISIC 2020 Challenge Dataset: https://challenge2020.isic-archive.com/

viii. Covid19 Datasets:

Kaggle Dataset for Traffic Analysis of Traffic Volume Post-COVIDic Volume Post-COVID: https://www.kaggle.com/terenceshin/covid19s-impact-on-airport-traffic?select=covid_impact_on_airport_traffic.csv
CDC: https://covid.cdc.gov/covid-data-tracker/#trends_dailytrendscases
WHO: https://covid19.who.int/
Harvard 1.3 acres: https://coronavirus.1point3acres.com/
Johns Hopkins University: https://coronavirus.jhu.edu/

b) Instructor Stored Datasets in Google Drive:
Datasets can be accessed using this link.

c) Datasets provided by the Molin's book: https://github.com/stefmolin/Hands-On-Data-Analysis-with-Pandas-2nd-edition

4. TEAM PROJECT PROCESS AND TASKS EACH WEEK

The project process shown in figure 1 is adapted and customized for our Data Mining and Data Driven Modeling Courses, including data visualization, from the Cross Industrial System Process for Data Mining (CRISP-DM) Reference Model.

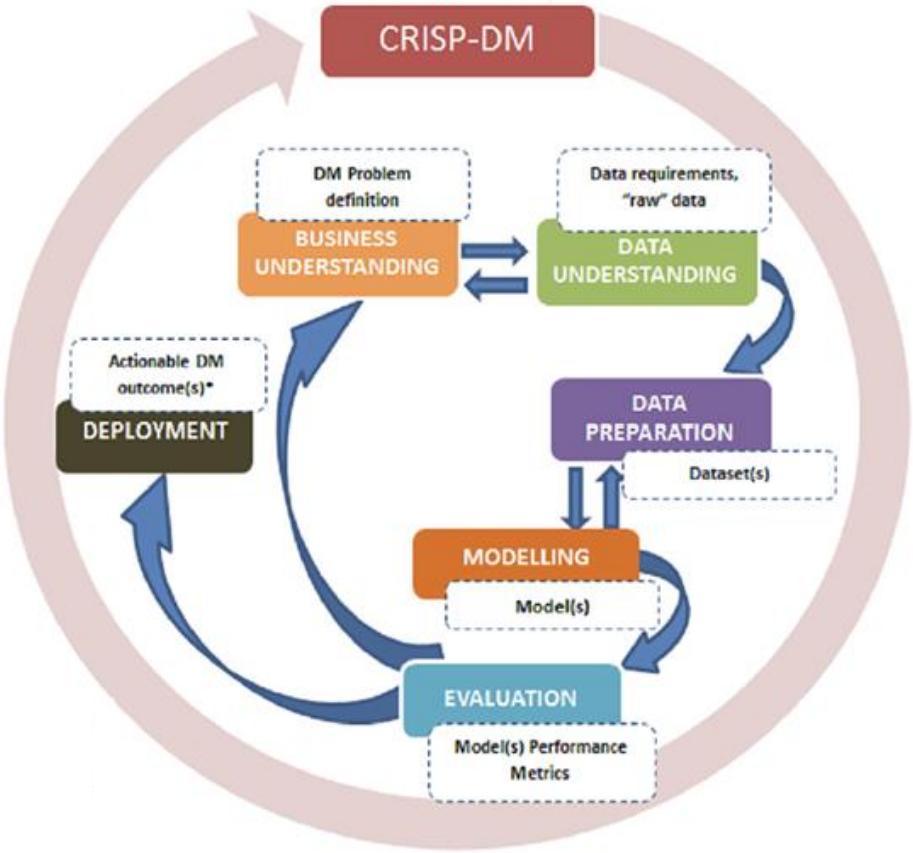

Figure 1: Modified CRISP-DM Reference Model for Data Visualization Project.

The following table describes the tasks and timeline for each week based on the aforementioned reference model.

CRISP-DM Stages & Objectives	Roles and Task Description	Deliverables (to be Shared in the Team Shared Folder)
Week 1: Business Understanding	*Domain Researchers*: One to two members with expertise in a field outside of Data Science, Math, or Computer Science. Begin a literature review and conduct internet search. Collaborate with other team members and the instructor to identify a research problem and potential data sources. *Data Retrievers*: Two members with database and programming skills. Their role is to identify, extract, and consolidate relevant heterogeneous data, including modern sources like social media, open data, and government datasets.	• *Team members and weekly meeting times* are set up and posted in the shared folder. • Who serves as domain researchers? • Who serves as data retrievers?
Week 2: Data Preparation	*One domain researcher* can continue proposing business questions that data visualization can help answer. *One Data Retriever* may focus on gathering additional datasets and merging them with existing data. *Data Wranglers:* Two (or more, as needed) members with strong programming skills. Use tools like Numpy, Pandas, and MS Excel to clean and transform the data into a tidy, structured format.	• Submit a *one-page proposal* outlining the project topic and data sources your team has agreed to work on. • Specify who will serve as the data wranglers and clearly define the roles of all other team members.
Week 3: Data Visual Exploration	*Data Wranglers:* Continue cleaning and transforming the data, incorporating feedback from the *exploratory analyzers*. *Exploratory Analyzers*: Two members use tools like Tableau, Pandas, Matplotlib, and Seaborn to perform visual analytics, identifying and ranking the most informative features for modeling. They develop a set of questions to explore data distributions, relationships, and their potential implications for business decisions.	• A *tidy data frame (tible)* is ready for data mining. • Who serve *Exploratory Analyzers*, and what are the roles of other members? • What questions to be explored? • What features are selected to answer the questions?
Week 4: Develop Data Visual Analytics	*Visual Developers:* Each member is responsible for independently creating visual analytics to address the team's key questions. The team will collectively decide which visualizations to select and integrate into the project report and final presentation.	• Develop initial visualizations (Preliminary Visual Analytics) to explain and address research questions of interest. • Evaluate whether the visualizations effectively answer the research questions?

Week 5: Discovery Communication Final presentation and report due on Dec 04, 5:00pm Eastern Time	*Dashboard developer:* Two members will be designated to create the dashboard and integrate the selected visualizations and graphs into a cohesive display. *Writers*: Each member is responsible for detailing the technical aspects of the visual analytics they developed. The primary writer, chosen by the team, will compile, refine, and polish the final report. Future work related to the data mining course is encouraged. *Presenters:* All members are expected to participate in-person for the final team presentation.	• Prepare a white paper summarizing the key findings and create presentation slides. • Additionally, ensure that seven meta questions are addressed. • Refer to the evaluation rubrics below for guidance.

5. PROJECT EVALUATION

The final project score consists of two components: 60% based on the quality of the team's project outcomes and 40% based on individual contributions during the collaborative learning process. While the team learning outcomes score will generally be distributed equally among team members (with some exceptions), individual merit scores will vary. These differences will depend on factors such as weekly input from the team leader, peer evaluations, and metadata from team meetings regarding task division, roles, and progress tracking.

Seven key questions for the Projects:

1) Who is your target audience? Identify the anticipated viewers who would be interested in your project (e.g., their job titles and career backgrounds).
2) What are the input and output attributes? Define the input attributes and the output (class) relevant to your problem.
3) How are these factors related? Identify the three most important attributes (features) and explain their significance.
4) What data visualization principles did you apply? Describe the principles used and how they shaped your design decisions.
5) What value does your project provide? Highlight the actionable insights your target audience can gain from your visual analytics.
6) Can additional data strengthen your analysis? If you had more time, what additional data sources would you explore to better address your research questions?
7) What future data features would you like to include? Identify additional features you wish you had for more quantitative answers in future work.

6. STRUCTURE OF THE PROJECT REPORT

For the report, use 12-point Times New Roman font, with 1-inch margins, single-column, and 1.2 line spacing. Ensure that all references follow APA style formatting.

* You may use LaTeX to produce your final report; however, ensure that you adhere to the prescribed document structure and styling guidelines as outlined.

1) **Title page**:
   Title of the project
   Names of the teammates in alpha-beta order
   Name of the instructor, course, etc.
   ** The report is regarded as a white paper (though the content is more like a proposal for a data analytic project) that is required to use APA Format.*
2) **Abstract** (Less than 50 words):
   Provide a concise statement outlining the goal of your project, including the datasets used, types of visualizations created, key observations made, and the value your project offers to potential viewers.
   Keywords: Key concepts related to the project (limited to 5 words or fewer)
3) **Introduction** (1 to 2 pages):
   Includes the following three sub-sections.
   Problem Statement:
   - What data-centric problem are you investigating? Describe the core issue you aim to explore.
   - What datasets will you use to address this problem? Outline the datasets that will assist in your investigation.
   - Who is the target audience for your data visualization project? Identify the primary viewers or stakeholders interested in your findings.

   Background of the problem:
   - Why is this problem significant? Explain the relevance and importance of the issue.
   - What are the challenges associated with the problem? Discuss the difficulties involved and provide at least one example of similar or relevant work, either from books or other references.

   Layout of the rest of the report:
4) **Data Preparation** (2 to 4 pages, include your source code (Python) as appendix):
   Outline the sources of the datasets, their relevance to your problem, and the steps taken to clean and transform them into tidy, usable datasets. Address the following questions to ensure originality and clarity:
   Three key questions on dataset origin:
   - How many datasets did you collect, and where did you source them?
   - What are the size and dimensions of each dataset?
   - Who are the original users of each dataset, and for what purposes were they intended?

   Three key questions on data transformation:

- What transformations were applied to convert each dataset into tidy data frames? (Include full R scripts in the project appendix).
- What data reduction techniques did you use to extract essential variables and records?
- How did you join smaller datasets into a cohesive larger dataset, and when necessary, how did you break down large datasets into smaller, manageable subsets?

5) **Problem Exploration** (3 to 6 pages):
   Break down the main problem into several sub-questions and use one or more plots to investigate each. Your analysis will be evaluated based on the Three C Principles:
   - Correctness: Ensure the data and insights are accurate.
   - Clarity: Present your findings in a clear, understandable manner.
   - Conciseness: Keep your explanations brief but informative.

   For each sub-question, use the tidy dataset(s) prepared earlier to explore and visualize potential answers. Steps 3 and 4 are iterative, meaning you can flexibly adjust the order of exploration as long as your reasoning remains logically sound. For reference, consult #1 from Homework3 and #6 and #7 from Homework4.
6) **Conclusion and Explanation** (Max 500 words, with as many pages of supporting plots as needed):
   Summarize your key observations and insights from the exploration. Create a dashboard, formatted in standard poster size, to present your findings to potential reviewers. The dashboard content should also be included in your report. Observations should be self-evident, data-driven, and intuitively understandable. As this course emphasizes qualitative insights, you are encouraged to outline potential quantitative problems for future research. The evaluation is based on the Three C Principles: Convincing, Clarity, Conciseness.
   Key Questions for Visualization and Explanation:
   - What are the key factors involved in the problem, and how are they interrelated?
   - What conclusions can be drawn, and how do these insights aid your target audience in decision-making?
   - What questions remain unanswered, particularly those requiring further quantitative analysis?
7) **References**:
   If you use any sources, cite them in APA format, including articles, books, or any other materials referenced.
8) **Appendix**:
   Include Python and Tableau scripts, along with links to the datasets used, either from your shared folders or external sources.

7. GRADING CRITERIA

The grading (max 100 points) will be based on the rubrics of the following five aspects.

Grade	Oral Presentation & Dashboard (20%)	Overall Quality of Project Report (20%)	Data Retrieval & Preparation (20%)	Visual Exploration (20%)	Visual Explanation (20%)
F	**Clueless**: Used neither appropriate terminology, nor appropriate form of representations.	**Confusing:** Used neither appropriate process, nor model assumption is identified.	**No effort:** Shown limited effort to collect data, nor effort to cleanse them, but download a tidy dataset.	**No effort:** Shown limited effort to explore the datasets, but almost no questions are asked.	**No effort:** Provided neither observations, nor justifications for the answers.
D	**Some Clue, but inconsistent:** Attempted to use appropriate form of representations (e.g. table, graphs, diagram).	**Procedure followed:** Attempted to follow a process and identify limitations.	**Some effort:** Retrieved a few available relevant datasets and made some of them tidy.	**Some effort:** Shown some effort, but the questions are unrelated, nor clearly related to the plots.	**Some effort:** Provided limited observation and attempt to justify the answers.
C	**Understandable:** Provided adequate explanations, and appropriate form of representations (e.g. table, graphs, diagram).	**Make-sense:** Followed an incomplete process and explain their methodology and identifies most of the observations & limitations.	**Good Effort with marginal outcomes:** The student retrieved a few available relevant datasets and made these dataset tidy.	**Good Effort with marginal outcomes:** Many questions are asked, and clear, but most relationships of questions, and question vs plot are unclear.	**Good Effort with marginal outcomes:** Provided some observations, identified some limitations, & justified some of the answers.
B	**Logically clear but lack smooth transitions:** Provided complete explanations, arguments, and appropriate forms of representations.	**Well-organized and articulated:** Followed a complete process and identify some insightful observations and limitations of the observations.	**Best effort but incomplete:** Retrieved most available relevant datasets, and made the original ones, and some joint datasets tidy.	**Best effort but incomplete:** A complete list of questions and the relationship to plots are clear, but the hierarchy/logic of the questions are unclear.	**Best effort but incomplete:** Provided adequate observations, limitations, and sound justification to most of answers.
A	**Context, logic, and transitions all crystal clear:** Provided a complete, concise and coherent explanations and arguments, as well as intuitive representations.	**Polished work:** Provided insightful observations to the problems and a sound trade off analysis between accuracy, certainty, and simplicity.	**Complete:** Retrieved all available relevant datasets, and made not only the original ones, but also all meaningful joint datasets tidy.	**Complete:** A complete list of questions are asked. The relationship between them and plots, and the hierarchy/ logic of the questions are all clear.	**Complete:** Provided insightful observations, with visual and verbal analytics to intuitively justify the answers, and limitations identified.

8. TEAM POLICIES

Your team will have several key responsibilities in completing problem and project assignments.

a) **Team Structure & Leadership**: Each team will consist of four members, including a team leader who is either elected by the team or appointed by the instructor. The team leader's role includes preparing the meeting agenda (after consulting with team members), documenting minutes from weekly meetings, organizing metadata, and providing reports to the instructor.

b) **Meeting Organization**: Agree on a regular meeting time, method of communication, and how the meeting outcomes and metadata will be recorded. The team leader is responsible for sharing the agenda before the meeting through email or other collaborative platforms.

c) **Task Division & Assignment**: The project should be broken down into manageable tasks, each requiring about 4-8 hours to complete and an additional 2 hours for validation and testing. Every task should have one leader responsible for implementation and one checker to assist and verify the work.

d) **Meeting Focus**: Each meeting will have three main objectives:
   a. Identify upcoming tasks and assign roles for the following week.
   b. Review completed tasks and determine who completed them.
   c. Discuss any tasks requiring assistance from the instructor or TA.

e) **Collaborative Workflow**: The task leader and checker will exchange project materials (such as documents, code, or data) and store all communication logs in a shared folder. This ensures that all project artifacts are organized and accessible.

f) **Tracking Progress**: The captain will monitor the progress of tasks, documenting the names of task leaders and checkers. If any issues arise that could delay the project, the captain must remind the responsible members and inform the instructor if the problem persists.

g) **Conflict Resolution**: If conflicts arise that cannot be resolved internally within the team, the instructor should be consulted to mediate the situation.

<u>Handling Non-Cooperative Team Members:</u>

If a team member consistently refuses to cooperate on assignments, their name should not be included in the completed work. The team should first attempt to resolve the issue by discussing it directly with the instructor. If the issue persists, the cooperating members may send a written warning to the uncooperative member, with a copy sent to the instructor, indicating that they are in danger of being removed from the team.

If there is no improvement, the team should send a formal notification (again, copied to the instructor) informing the uncooperative member that they are no longer part of the team. The dismissed member should then meet with the instructor to discuss alternative options.

Similarly, if a team member finds themselves doing the majority of the work, they are encouraged to issue a warning memo to the team, with an indication to leave if cooperation does not improve. If the situation remains unchanged, they may formally quit the team with a second memo, notifying both the team and the instructor.

Students who are either dismissed from or voluntarily leave the team must either join another group or receive zero credit for the remaining assignments.

As you will soon discover, working in a group can be challenging—team members may have conflicting schedules or other commitments that prevent them from fully participating, and differences in skill levels and work habits can lead to tension. However, when teams communicate effectively and collaborate well, the rewards far outweigh the challenges. One way to increase the likelihood of success is by clearly defining and agreeing on team expectations at the outset. This mutual understanding is the focus of the "Team Expectations Agreement", which is designed to align everyone's responsibilities and goals from the start.

Team Expectations Agreement

On a single sheet of paper, put your names and list the rules and expectations you agree as a team to adopt. You can deal with any or all aspects of the responsibilities outlined above—preparation for and attendance at group meetings, making sure everyone understands all the solutions, communicating frankly but with respect when conflicts arise, etc. Each team member should sign the sheet, indicating acceptance of these expectations and intention to fulfill them. Turn one copy into the professor and keep the remaining copy or copies for yourselves.

These expectations are for your use and benefit—they won't be graded or commented on unless you specifically ask for comments. Note, however, that if you make the list fairly thorough without being unrealistic, you'll be giving yourselves the best chance. For example, "We will each solve every problem in every assignment completely before we get together" or "We will get 100 on every assignment" or "We will never miss a meeting" are probably unrealistic, but "We will try to set up the problems individually before meeting" and "We will make sure that anyone who misses a meeting for good cause gets caught up on the work" are realistic.

Last Name: _____ First Name: _____

Signature: _____

Date: _____

9. TEAM AND PEER EVALUATION

* The following table to be filled out and evaluated by the instructor.

Name of the Team: _____

Item	Usually	Sometimes	Hardly Ever
Team meetings frequently start 5-15 minutes late, causing delays in productivity.			
Some members arrive late, leave early, or fail to attend, disrupting the flow of the meeting.			
Meetings lack a clear agenda, with members having only a general idea of the goals.			
Discussions are dominated by one or two individuals, leaving little room for balanced input.			
Members come unprepared, having not reviewed the assignment/task, researched necessary information, or completed their tasks, leading to inefficient discussions.			
Body language or verbal cues from some participants indicate disengagement or a lack of interest in the meeting.			
Conversations are frequently interrupted, and side discussions take place, preventing focused, meaningful dialogue.			
Critical issues are repeatedly postponed rather than being resolved, leading to ongoing confusion.			
No follow-up action plan is established, leaving team members unclear on next steps or responsibilities.			
A disproportionate amount of the work falls on the same few individuals, resulting in frustration and burnout.			
Meetings drag on without clear outcomes, leading to wasted time and minimal progress.			
Tasks are either submitted late or completed below standard, compromising the quality of the group's work.			

* The following form to be filled out and evaluated by the team members and submitted through the Canvas.

Name of the Team: _____

Your Name: _____

Please write the names of all of your team members, INCLUDING YOURSELF, and rate the degree to which each member fulfilled his/her responsibilities in completing the team assignments.

DO NOT LEAVE ANY COMMENTARY BLANK!

The possible ratings are as follows:

- Excellent: Consistently carried more than his/her fair share of the workload.
- Very good: Consistently did what he/she was supposed to do, very well prepared and cooperative.
- Satisfactory: Usually did what he/she was supposed to do, acceptably prepared and cooperative.
- Ordinary: Often did what he/she was supposed to do, minimally prepared and cooperative.
- Marginal: Sometimes failed to show up or complete assignments, rarely prepared.
- Deficient: Often failed to show up or complete assignments, rarely prepared.
- Unsatisfactory: Consistently failed to show up or complete assignments, unprepared.
- Superficial: Practically no participation.
- No show: No participation at all.

These ratings should reflect each individual's level of participation and effort and sense of responsibility, not his or her academic ability.

Name of the Team Member	Rating	Commentary (Do not leave blank)

Your Signature: _____

10. ADDITIONAL INFORMATION

Submission of Final Report and Presentation:

Final presentation and report due on Dec 04, 5:00 pm Eastern Time.

Please compile your final report to a single PDF document and submit it through Canvas. Include all Python and Tableau scripts in the *Appendix*, along with links to the datasets used, either from your shared folders or external sources.

Final presentation slides must also be uploaded to Canvas on or before the aforementioned date. The presentation schedule will be announced in due course.

All your document/material should be self-contained, so that the reader can understand them without additional information not contained in the document. The document should be concise, thorough, clear, well organized, and have good grammar. You may use any additional resources besides the class materials to finish the work. You can also make your own assumptions, modifications, and derivations, if required. But they need to be well documented.

If ChatGPT (or any other AI assistant) has been used to generate any part of your codes or report, you are required to share your interaction history with the AI assistant. Check this link about how to share the chat history safely.

The deadline is solid. Early submission is encouraged but will not earn extra points.