# Table of Contents

[Table of Contents 1](#_Toc16648386)

[Summary of Project Requirements 2](#_Toc16648387)

[Deliverables 2](#_Toc16648388)

[1. Project Draft – Pitch Deck 2](#_Toc16648389)

[2. Project Draft – Project Board 3](#_Toc16648390)

[3. Project Draft – Project Prototype 4](#_Toc16648391)

[4. Project Final – Presentation Deck 5](#_Toc16648392)

[5. Project Final – Project Board 5](#_Toc16648393)

[6. Project Final – Source Code 6](#_Toc16648394)

[7. Project Final – In-Class Presentation 7](#_Toc16648395)

[Group Project Ideas 7](#_Toc16648396)

[User Interfaces – A Trivial Example 8](#_Toc16648397)

# Summary of Project Requirements

The Group Project consists of designing, developing, and presenting a Python-based product of interest to some customer base. You can think of your product as a commercial product, or as a non-profit public service product.

Your product will need to scrape or otherwise obtain data from AT LEAST three different online sources, including AT LEAST one scraped web site. Your product will then need to clean, organise, and analyse the obtained data to produce interesting results for the user to look at (statistical results, tables, plots, ...) and perhaps interact with.

Since this is a Python course, all of this will need to be in Python code: you may not download data using a browser, clean, and organize the data in Excel, and then just do the analysis and display part in Python. Your entire product must be Python, beginning to end.

You MAY NOT present your product as a web site: We need to be able to install and use your product easily on our own systems, on which we have installed **Python 3.8**. We are not willing to fiddle with a web-based application and the server it needs to talk to.

# Deliverables

There are 7 deliverables for this project:

|  |  |  |
| --- | --- | --- |
| **Deliverable** | **Due Date** | **Points (Total: 1700)** |
| Project Draft – Pitch Deck | Week 3 | 100 |
| Project Draft – Project Board | Week 4 | 100 |
| Project Draft – Project Prototype | Week 4 | 100 |
| Project Final – Presentation Deck | Week 7 | 100 |
| Project Final – Project Board | Week 7 | 100 |
| Project Final – Source Code | Week 7 | 1000 |
| Project Final – In-Class Presentation | Week 7 | 200 |

The following sections discuss these deliverables in more detail. For specific breakdowns of mark allocations, see the Rubrics for each deliverable on Canvas.

## Project Draft – Pitch Deck

The first phase of the Group Project will consist of:

1. **Forming a Group.** The Group should have 4 or 5 people in it. If absolutely necessary, a group of 3 is okay and a group of 6 is okay. However, 3 people cannot usually accomplish as much as 4 or 5; and it may be hard to divide up and manage the development work of 6 people. Form your Group as soon as possible and send me an email with the Group members.
2. **Deciding on a product to develop.**
3. **Developing and submitting a "Draft Pitch Deck"**, that is, a draft PowerPoint presentation describing your group, your product, your market, and so forth. During the development of your product over the next five weeks, you will revise this Draft Pitch Deck into a Final Pitch Deck, from which your group will make its presentation during the last week of classes.

[The following video](https://www.youtube.com/watch?v=V0caTPYb6IE&index=2&list=PLPwOg45xx0bUj18qqwg1n2FdoASsr3xi0&t=0s) was originally developed for a class with only 15 students in it, so that each Project Group only had 2 or 3 people. When you watch the video, you will need to "scale it up" to the size of your Project Group: 4, 5 or possibly 6 people.

For example, the video says in various places that you need to scrape data from two web sites. In this class, you will need to obtain data from AT LEAST THREE web sites, at least one of which must be scraped: more sources of data is better. In lectures, we will look at using Beautiful Soup to scrape tabular data from a web site. If you wish to scrape data from an active JavaScript web site, you will have to learn about Selenium, as well. You can also pull data from an API at a web site. Using different scraping methods in your Group Project code will be more impressive than using just one scraping method.

The Pitch Deck template (available on Canvas) and the video are geared toward creating a commercial company with a product to sell, with the pitch delivered to some investors. You can follow that strategy, or instead think about forming an organisation that provides some societal benefit, in the form of a Python program, and aim your pitch at policy decision makers. In the end, either way, **you are making a pitch for money and support**.

You may create your own pitch deck from scratch, if it contains similar information in a similar order as shown in the template. This means that you may add additional slides to your template to spread information out and not clutter everything into a few slides (in fact, it is recommended that you increase the slide count – you are just required to stick to the template structure).

This is a DRAFT pitch deck. It does not need to be perfect, and *you can change it* as your Group Project evolves and develops, but the draft pitch deck should be complete and coherent. At the end of the Project, you will have to deliver a FINAL pitch deck, which should be professional and polished. This FINAL pitch deck will be the basis for your Group Presentation.

## Project Draft – Project Board

We will use Trello to create and manage a Kanban Board for each Group Project. Given our present circumstances, in which we should not be meeting with one another physically, appropriate use of your Trello Kanban Board for coordination among team members will be even more important than usual!

The Kanban system originated at Toyota Motors as a way of streamlining the manufacturing process. The same ideas apply very nicely to other management tasks. We will use Kanban boards to keep track of the development process for the Group Project. There are many Kanban board software packages available: we will use **Trello**, which is free, simple, and nice. Lecture time will not be spent on this, so please watch these two YouTube videos that cover the ideas very well:

[Intro to Kanban in Under 5 Minutes (What is Kanban, Learn Kanban)](https://www.youtube.com/watch?v=R8dYLbJiTUE)

[Using Trello for Online Project Management](https://www.youtube.com/watch?v=fe5ei52usQ8)

In Trello, make the board “private” and add all Group members to the board. Also, add me and my TA.

How can you know that you have broken your Group Project tasks down to the correct level for the Kanban board?

Think of each task as a discrete unit of work that an individual person could accomplish in 1, 2, 4, 8 hours or 1, 2, 3, 5 days. Shorter tasks are better, but sometimes a task can't be realistically reduced to less than 2 or 3 days. “Design and develop our website” is a goal, not a task; “use a for loop to step through the data” is a small implementation detail within a task, not a task. *All you really need are the “To Do”, “Doing”, and “Done” categories, but you can create more categories if you find that useful.*

A couple of obvious tasks would be “scrape raw data from web site A” and “reorganize the raw data from web site A into a convenient form”. Since there are AT LEAST three web sites, this is an obvious division of labour: different group members can handle the scraping and data massaging of different sites. The efficiency of parallelism!

One of the most important--and sadly overlooked--parts of organising tasks is to **agree on the interfaces between tasks**: as a group task, before the “scrape” and “reorganise” tasks, you should agree on what constitutes a “convenient form” of the data produced by each “reorganize” task, so that you can combine the three or more sets of data easily rather than awkwardly in a subsequent analysis task. (Generally, “make a group decision about X” is a task, even if it only takes a few minutes.)

A Kanban board is useful even for an individual developer to keep track of the tasks--and approximate order of tasks--necessary to complete a project. It works very well for small groups, say 5 or fewer.

For larger groups/larger projects, more machinery is necessary to keep track of everything. Scrum is very popular for this nowadays. Kanban boards can be used as a component within the scrum structure. Here is a good overview:

[Intro to Scrum in Under 10 Minutes](https://www.youtube.com/watch?v=XU0llRltyFM)

We will not need all the machinery of scrum.

## Project Draft – Project Prototype

The eventual goal of the Group Project is to use AT LEAST THREE sources of data (or AT LEAST FOUR, if you are part of a group of 6), with AT LEAST ONE (or TWO) obtained from the web: AT LEAST ONE (or TWO) by web scraping, others perhaps by direct download, by using an API, or whatever. In class we will investigate Beautiful Soup, but you might also choose to look at Selenium, Scrapy, or other web scrapers. (You may NOT use data from a "data set aggregator" site like Kaggle, where the data sets have already been cleaned for you. One purpose of the project is to gain experience dealing with dirty data. But if the original source of data provides clean data, that's fine. One more thing: ***Don't do a project about movies***. Choose finance or economics, health care (or lack of it), policing or incarceration, natural disaster risk, sports, politics, wildlife, the environment, geology, real estate, shopping or trade, transportation or travel, energy, ..., just not movies, mainly because data about movies is already too clean and easy to obtain and process.)

By the time of this deliverable’s due date, you will have learned about basic Python: int, float, str, bool, None; arithmetic operators; lists, loops and range(); equality, relational, and logical operators; if/elif/else decision; built-in data structures (list, tuple, set, dict, with named and symbolic operators); sequence slices; user and text file input/output; and Beautiful Soup for scraping HTML web sites.

But note that we will be learning more, including NumPy for N-dimensional arrays, random numbers, and statistical functions; Pandas for Series (like time series) and DataFrames (like spreadsheets); Regular Expressions for general pattern matching; matplotlib for graphics/visualization; and other data cleaning/wrangling techniques.

Hence, your Group Project code, in parallel with doing something that is at least potentially useful and valuable to some customer, should demonstrate a solid knowledge of most of the topics we will cover. This does not mean that you must use tuples, or that you must use floor division, or that you must use a binary data file. But it does mean that you must use--in a skilful way--many features of most of the Python facilities we will cover.

In this third deliverable in the Group Project, you must turn in an Excel spreadsheet that contains at least these worksheets:

* 1 cover letter, which should give a clear description of what you are trying to do and what data sources you will be using;
* 2 worksheets (raw and cleaned) PER DATA SOURCE that you plan to use;
* and 1 worksheet of the merged data for all the data sources. If including all of the data in the Excel spreadsheet would not make sense due to size, then include a *representative sample* of data from each source.

You will see in the Rubric on Canvas that we are looking for:

* The NUMBER of data sources you are using,
* The VARIETY of scraping/downloading/API technologies you are using to obtain data,
* The COMPLEXITY of the data you are using,
* The VARIETY OF KIND of data sources: don't just create an aggregated list of trucks available for purchase by downloading and combining data from several vehicle sales websites.

Please be aware that according to Heinz standard "meets expectations" is an A-, "exceeds expectations" is an A, and "superior performance" is an A+. So AT LEAST 3 (or 4) data sources means that 3 (or 4) meets expectations: more data sources than the minimum required is a good thing.

This part of the Group Project is a DRAFT: you can certainly go above and beyond what you describe in this DRAFT Project Prototype, but this deliverable should describe the minimum that you intend to do.

Please zip your spreadsheet and upload the zip archive.

## Project Final – Presentation Deck

This will be your FINAL Pitch Deck, which should be professional and polished. This FINAL Pitch Deck will be the basis for your Group Presentation, which will be delivered by you in week 7. The upload may be a PPT or a PDF file, in a zip archive.

## Project Final – Project Board

We will check your Trello Project Board to see how your project has progressed. The board should show how you divided up and completed the tasks necessary to complete your project.

## Project Final – Source Code

This deliverable requires you to submit all your source code to Canvas. We must be able to download your project code from Canvas, install it, and run it, without pounding our heads on our desks. Each time we feel the urge to pound our heads on our desks, you will lose points.

You must write exactly ONE “README” file, submitted as a PDF, MS-Word, or .txt file, that gives us instructions on how to install and run your project, including installing any additional Python modules beyond what Anaconda provides; setting environment variables *(please avoid if possible)*; obtaining an API key; or the like. (If your project requires packages/modules other than what Anaconda provides, tell us what we need to do MANUALLY to install those: do not auto-install those within your project: it is too hard for us to un-install things that were auto-installed.) Include the names and email addresses of all Group members in the “README” file. Give us screen shots to look at, when that would be helpful. DO NOT write a “README” file that refers us to other “README” files elsewhere. To repeat: ONE “README” file only, submitted as a PDF.

Additional instructions:

1. Before you submit your completed Group Project code, have one member of your team uninstall Anaconda/Python from their system, then do a fresh install of Anaconda/Python, and then install your Group Project code to check that the installation and usage instructions you have written for us in your "README" file are correct and not painful to follow.
2. The name of the zip file that you upload should include your Section and Group Number, e.g., A1\_Group5.
3. Create a short video (10 minutes or less) demonstrating your project being run. If certain things take a long time to run (like web scraping), show the beginning and ending of that in the video, and clip out the whirling dots/thumb twiddling/waiting-waiting-waiting part from your video. Post your video on YouTube and provide us with the link to your video (also, put the link in your "README" file). Make sure you record your video at high enough volume for us to hear it (we can turn the volume down, if we need to). DO NOT INCLUDE BACKGROUND MUSIC!!! The video must NOT contain information about how to INSTALL your project: that should be in the "README" file. (Likewise, the "README" file must contain written instructions about how to RUN your project: it MUST NOT just say, "watch the video.")
4. There must be ONE main program file (**group\_4\_nifty\_project.py** or whatever) that we can load into IDLE and run to get your project to work. This does NOT mean that you need to write all your source code in a single code file: if the main program file requires other modules/program files in order to work, then it can **import** those modules, etc., to make everything work.
5. NO hard-coded file or directory pathnames. Put everything in a SINGLE directory, so that we can go to that directory and run your program and have everything work. (Yes, we know this is not “realistic” but we have a lot of projects to grade in a small number of days.)
6. If web scraping takes a long time (perhaps so long that it might time out on a slower web connection), make it possible to use a previously-scraped copy of some data, rather than forcing the user to wait for a long and perhaps unsuccessful download.
7. You must have comments in each code file, including the name of the file, the names of the group members, and a description of the purpose of the code in the file, which other module(s) import it, and which other module(s) it imports. Comments do not need to appear in every single line of code, but you need to have comments before each function or large piece of code explaining what the code does and what data structures you are using. If you think a particular line of code is difficult to understand you can add comments before that specific line.
8. Jupyter notebooks are great for interactive analysis tasks, but not so great in a product that is supposed to do something useful for a user who may not be skilled at Python. We don't want to have to run Jupyter notebooks. We insist that we be able to run your project from IDLE or Spyder.
9. Do not install (and expect us to install) software that duplicates Python language or Python module functionality, even if it would be more convenient and/or more efficient and/or cooler. For example, do not install a database system: use Python lists, sets, dicts, NumPy ndarrays, and/or Pandas DataFrames and/or Series. This is a Python course, after all! We want to see how much Python you have learned, not how well you know other software like SQL or whatever.

## Project Final – In-Class Presentation

Under normal circumstances, each Project Group would give a live presentation of up to 14 minutes in length during the last lecture period in Week 7. For this course, you will need to produce a video presentation instead, of **up to 15 minutes in length**. The finished video must be available for us to watch on YouTube, or in a Google Document, or the like. You will have one extra day after your Project Final Pitch Deck, Project Board, and Source Code are due to get this video created.

Each person in the Project Group must give an approximately equal part of the presentation (within a minute or so: we won't time you, but we don't want to see presentations where one person talks for 10 minutes, and the other two talk for 2 or 3 minutes each). Under the 15-minute time constraint, do not spend time introducing yourselves: we can learn who you are from your Final Pitch Deck. Keep your presentation at a "here is how this will help our customers" level: a brief demo of your application is good, but do not dive into details of the code.

As your submission, post a brief .txt file telling us how and where we can view your video.

You do not need to do a live demo of your Group Project application during your Final Presentation, but if you would like to do a live demo then it is fine to have a copy of previously downloaded data to use for that purpose, so you don't have to do live web scraping.

The FOCUS of your live presentation should be to get potential investors (or government or NGO funding sources) excited about giving you money to get your company going. You need to show enough detail about what your application does to support this goal (like example output), but not so much detail (like source code details) that the presentation becomes boring.

# Group Project Ideas

What decisions do you make yourself that involve looking at multiple data sources, probably on the web?

For example, if you would like to take a vacation in Phoenix, you might look up:

1. Available flights and costs
2. The predicted weather in Phoenix when you plan to go
3. Restaurants and entertainment events in Phoenix during that time

One past project combined these sources of information to give a user a single application for vacation planning.

If you want to eat a nice meal, you might:

1. Select the type(s) of food you are interested in: Mexican, French, BBQ, ...
2. Look up menus, prices, and ratings of nearby restaurants offering that kind of food
3. Look up highly rated recipes that you could make at home, with cost estimates

This was another past project.

Think about what information YOU use from multiple sources to decide about where to go, what to buy, what to invest in, which player to draft in your fantasy league, which doctor or dentist to go to, where to move to, etc.

Remember that the Draft Pitch Deck is an outline of your plan for your Group Project. It does not have to be perfect, and you can change it entirely during the development of your project. All we are looking for in the Draft Pitch Deck is that you have an idea that involves data from several sources.

# User Interfaces – A Trivial Example

Your Group Project program does not need to have a fancy user interface. Here is a trivial menu-oriented program that:

1. Displays a table of data (which should have been scraped from some web site, but this trivial program doesn't show that part) of temperatures in Pittsburgh
2. Displays a plot of high and low temperatures
3. Displays simple statistics about the temperatures

[trivial\_project.py](https://canvas.cmu.edu/courses/19011/files/folder/Group%20Project%20Files?preview=5080580)

Of course, your Group Project program needs to do much more than this trivial program does, but it does not need to have a much fancier user interface than this.

If you decide to put extra effort into a nice GUI for your program, you might want to check out tkinter at <https://docs.python.org/3/library/tkinter.html>

But I am more concerned with what's going on inside your program: obtaining data, cleaning and structuring the data, analysing the data, and displaying results that are useful to the user. When you have the analytics part of your program working well, then try adding a GUI if you have more time, or if you have a Project Group member who wants something more to do.