

Phase 1 Project Description

You've made it all the way through the first phase of this course - take a minute to celebrate your awesomeness!

Now you will put your new skills to use with a large end-of-Phase project!

In this project description, we will cover:

- ***Project Overview:*** the project goal, audience, and dataset
- ***Deliverables:*** the specific items you are required to produce for this project
- ***Grading:*** how your project will be scored
- ***Getting Started:*** guidance for how to begin your first project

Project Overview

For this project, you will use data cleaning, imputation, analysis, and visualization to generate insights for a business stakeholder.

Business Problem

Your company is expanding in to new industries to diversify its portfolio. Specifically, they are interested in purchasing and operating airplanes for commercial and private enterprises, but do not know anything about the potential risks of aircraft. You are charged with determining which aircraft are the lowest risk for the company to start this new business endeavor. You must then translate your findings into actionable insights that the head of the new aviation division can use to help decide which aircraft to purchase.

The Data

In the `data` folder is a dataset from the National Transportation Safety Board that includes aviation accident data from 1962 to 2023 about civil aviation accidents and selected incidents in the United States and international waters.

It is up to you to decide what data to use, how to deal with missing values, how to aggregate the data, and how to visualize it in an interactive dashboard.

Key Points

- **Your analysis should yield three concrete business recommendations.** The key idea behind dealing with missing values, aggregating and visualizing data is to help your organization make data driven decisions. You will relate your findings to business intelligence by making recommendations for how the business should move forward with the new aviation opportunity.

- **Communicating about your work well is extremely important.** Your ability to provide value to an organization - or to land a job there - is directly reliant on your ability to communicate with them about what you have done and why it is valuable. Create a storyline your audience (the head of the aviation division) can follow by walking them through the steps of your process, highlighting the most important points and skipping over the rest.
- **Use plenty of visualizations.** Visualizations are invaluable for exploring your data and making your findings accessible to a non-technical audience. Spotlight visuals in your presentation, but only ones that relate directly to your recommendations. Simple visuals are usually best (e.g. bar charts and line graphs), and don't forget to format them well (e.g. labels, titles).

Deliverables

There are three deliverables for this project:

- A **non-technical presentation**
- A **Jupyter Notebook**
- A **GitHub repository**
- An **Interactive Dashboard**

Non-Technical Presentation

The non-technical presentation is a slide deck presenting your analysis to business stakeholders.

- ***Non-technical*** does not mean that you should avoid mentioning the technologies or techniques that you used, it means that you should explain any mentions of these technologies and avoid assuming that your audience is already familiar with them.
- ***Business stakeholders*** means that the audience for your presentation is the business, not the class or teacher. Do not assume that they are already familiar with the specific business problem.

The presentation describes the project ***goals, data, methods, and results***. It must include at least ***three visualizations*** which correspond to ***three business recommendations***.

We recommend that you follow this structure, although the slide titles should be specific to your project:

1. Beginning
 - Overview
 - Business Understanding
2. Middle
 - Data Understanding

- Data Analysis
- 3. End
 - Recommendations
 - Next Steps
 - Thank You
 - This slide should include a prompt for questions as well as your contact information (name and LinkedIn profile)

You will give a live presentation of your slides and submit them in PDF format on Canvas. The slides should also be present in the GitHub repository you submit with a file name of **presentation.pdf**.

The graded elements of the presentation are:

- Presentation Content
- Slide Style
- Presentation Delivery and Answers to Questions

See the Grading section for further explanation of these elements.

For further reading on creating professional presentations, check out:

- Presentation Content
- Slide Style

Jupyter Notebook

The Jupyter Notebook is a notebook that uses Python and Markdown to present your analysis to a data science audience.

- ***Python and Markdown*** means that you need to construct an integrated `.ipynb` file with Markdown (headings, paragraphs, links, lists, etc.) and Python code to create a well-organized, skim-able document.
 - The notebook kernel should be restarted and all cells run before submission, to ensure that all code is runnable in order.
 - Markdown should be used to frame the project with a clear introduction and conclusion, as well as introducing each of the required elements.
- ***Data science audience*** means that you can assume basic data science proficiency in the person reading your notebook. This differs from the non-technical presentation.

Along with the presentation, the notebook also describes the project ***goals, data, methods, and results***.

You will submit the notebook in PDF format on Canvas as well as in `.ipynb` format in your GitHub repository.

The graded elements for the Jupyter Notebook are:

- Business Understanding

- Data Understanding
- Data Preparation
- Data Analysis
- Code Quality

See the Grading section for further explanation of these elements.

GitHub Repository

The GitHub repository is the cloud-hosted directory containing all of your project files as well as their version history.

This repository link will be the project link that you include on your resume, LinkedIn, etc. for prospective employers to view your work. Note that we typically recommend that 3 links are highlighted (out of 5 projects) so don't stress too much about getting this one to be perfect! There will also be time after graduation for cosmetic touch-ups.

A professional GitHub repository has:

1. `README.md`
 - A file called `README.md` at the root of the repository directory, written in Markdown; this is what is rendered when someone visits the link to your repository in the browser
 - This file contains these sections:
 - Overview
 - Business Understanding
 - * Include stakeholder and key business questions
 - Data Understanding and Analysis
 - * Source of data
 - * Description of data
 - * Three visualizations (the same visualizations presented in the slides and notebook)
 - Conclusion
 - * Summary of conclusions including three relevant findings
2. Commit history
 - Progression of updates throughout the project time period, not just immediately before the deadline
 - Clear commit messages
 - Commits from all team members (if a group project)
3. Organization
 - Clear folder structure
 - Clear names of files and folders
 - Easily-located notebook and presentation linked in the README
4. Notebook(s)
 - Clearly-indicated final notebook that runs without errors
 - Exploratory/working notebooks (can contain errors, redundant code, etc.) from all team members (if a group project)

5. .gitignore

- A file called `.gitignore` at the root of the repository directory instructs Git to ignore large, unnecessary, or private files
 - Because it starts with a `.`, you will need to type `ls -a` in the terminal in order to see that it is there
- GitHub maintains a Python `.gitignore` that may be a useful starting point for your version of this file
- To tell Git to ignore more files, just add a new line to `.gitignore` for each new file name
 - Consider adding `.DS_Store` if you are using a Mac computer, as well as project-specific file names
 - If you are running into an error message because you forgot to add something to `.gitignore` and it is too large to be pushed to GitHub this blog post(friend link) should help you address this

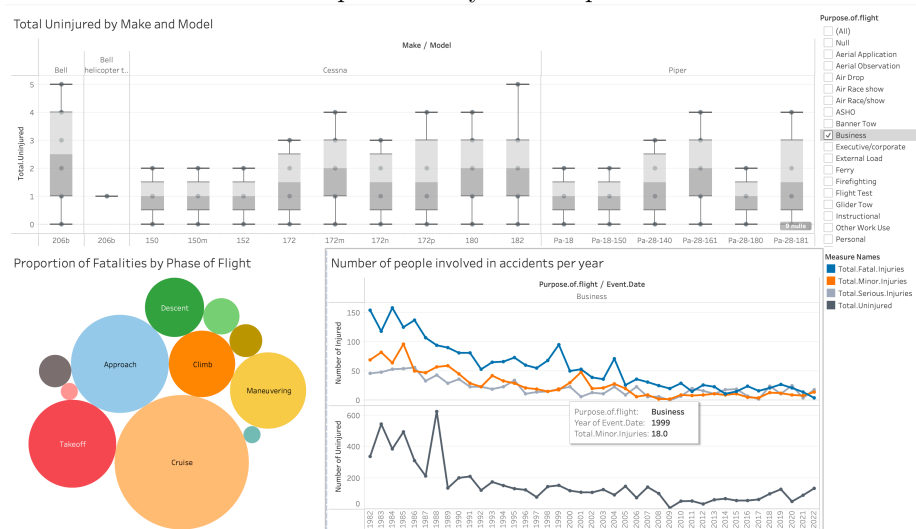
You will submit a link to the GitHub repository on Canvas.

See the Grading section for further explanation of how the GitHub repository will be graded.

For further reading on creating professional notebooks and READMEs, check out this reading.

Interactive Dashboard

The interactive dashboard is a collection of views that allows the viewer to change the views to understand different features in the data. This dashboard will be linked within your GitHub repository README.md file so that users can explore your analysis. Make sure you follow visual best practices that you have learned in this course. Below is an example of what you could produce for this as-



segment.

Grading

To pass this project, you must pass each project rubric objective. The project rubric objectives for Phase 1 are:

1. Data Communication
2. Authoring Jupyter Notebooks
3. Data Manipulation and Analysis with `pandas`
4. Interactive Data Visualization

Data Communication

Communication is a key “soft skill”. In this survey, 46% of hiring managers said that recent college grads were missing this skill.

Because “communication” can encompass such a wide range of contexts and skills, we will specifically focus our Phase 1 objective on Data Communication. We define Data Communication as:

Communicating basic data analysis results to diverse audiences via writing and live presentation

To further define some of these terms:

- By “basic data analysis” we mean that you are filtering, sorting, grouping, and/or aggregating the data in order to answer business questions. This project does not involve inferential statistics or machine learning, although descriptive statistics such as measures of central tendency are encouraged.
- By “results” we mean your *three visualizations and recommendations*.
- By “diverse audiences” we mean that your presentation and notebook are appropriately addressing a business and data science audience, respectively.

Below are the definitions of each rubric level for this objective. This information is also summarized in the rubric, which is attached to the project submission assignment.

Exceeds Objective Creates and describes appropriate visualizations for given business questions, where each visualization fulfills all elements of the checklist

This “checklist” refers to the Data Visualization checklist within the larger Phase 1 Project Checklist

Meets Objective (Passing Bar) Creates and describes appropriate visualizations for given business questions

This objective can be met even if all checklist elements are not fulfilled. For example, if there is some illegible text in one of your visualizations, you can still meet this objective

Approaching Objective Creates visualizations that are not related to the business questions, or uses an inappropriate type of visualization

Even if you create very compelling visualizations, you cannot pass this objective if the visualizations are not related to the business questions

An example of an inappropriate type of visualization would be using a line graph to show the correlation between two independent variables, when a scatter plot would be more appropriate

Does Not Meet Objective Does not submit the required number of visualizations

Authoring Jupyter Notebooks

According to Kaggle's 2020 State of Data Science and Machine Learning Survey, 74.1% of data scientists use a Jupyter development environment, which is more than twice the percentage of the next-most-popular IDE, Visual Studio Code. Jupyter Notebooks allow for reproducible, skim-able code documents for a data science audience. Comfort and skill with authoring Jupyter Notebooks will prepare you for job interviews, take-home challenges, and on-the-job tasks as a data scientist.

The key feature that distinguishes *authoring Jupyter Notebooks* from simply *writing Python code* is the fact that Markdown cells are integrated into the notebook along with the Python cells in a notebook. You have seen examples of this throughout the curriculum, but now it's time for you to practice this yourself!

Below are the definitions of each rubric level for this objective. This information is also summarized in the rubric, which is attached to the project submission assignment.

Exceeds Objective Uses Markdown and code comments to create a well-organized, skim-able document that follows all best practices

Refer to the repository readability reading for more tips on best practices

Meets Objective (Passing Bar) Uses some Markdown to create an organized notebook, with an introduction at the top and a conclusion at the bottom

Approaching Objective Uses Markdown cells to organize, but either uses only headers and does not provide any explanations or justifications, or uses only plaintext without any headers to segment out sections of the notebook

Headers in Markdown are delineated with one or more **#**s at the start of the line. You should have a mixture of headers and plaintext (text where the line does not start with **#**)

Does Not Meet Objective Does not submit a notebook, or does not use Markdown cells at all to organize the notebook

Data Manipulation and Analysis with **pandas**

pandas is a very popular data manipulation library, with over 2 million downloads on Anaconda (`conda install pandas`) and over 19 million downloads on PyPI (`pip install pandas`) at the time of this writing. In our own internal data, we see that the overwhelming majority of Flatiron School DS grads use **pandas** on the job in some capacity.

Unlike in base Python, where the Zen of Python says “There should be one—and preferably only one —obvious way to do it”, there is often more than one valid way to do something in **pandas**. However there are still more efficient and less efficient ways to use it. Specifically, the best **pandas** code is *performant* and *idiomatic*.

Performant **pandas** code utilizes methods and broadcasting rather than user-defined functions or `for` loops. For example, if you need to strip whitespace from a column containing string data, the best approach would be to use the `pandas.Series.str.strip` method rather than writing your own function or writing a loop. Or if you want to multiply everything in a column by 100, the best approach would be to use broadcasting (e.g. `df["column_name"] * 100`) instead of a function or loop. You can still write your own functions if needed, but only after checking that there isn’t a built-in way to do it.

Idiomatic **pandas** code has variable names that are meaningful words or abbreviations in English, that are related to the purpose of the variables. You can still use `df` as the name of your DataFrame if there is only one main DataFrame you are working with, but as soon as you are merging multiple DataFrames or taking a subset of a DataFrame, you should use meaningful names. For example, `df2` would not be an idiomatic name, but `movies_and_reviews` could be.

We also recommend that you rename all DataFrame columns so that their meanings are more understandable, although it is fine to have acronyms. For example, `"col1"` would not be an idiomatic name, but `"USD"` could be.

Below are the definitions of each rubric level for this objective. This information is also summarized in the rubric, which is attached to the project submission assignment.

Exceeds Objective Uses `pandas` to prepare data and answer business questions in an idiomatic, performant way

Meets Objective (Passing Bar) Successfully uses `pandas` to prepare data in order to answer business questions

This includes projects that *occasionally* use base Python when `pandas` methods would be more appropriate (such as using `enumerate()` on a `DataFrame`), or occasionally performs operations that do not appear to have any relevance to the business questions

Approaching Objective Uses `pandas` to prepare data, but makes significant errors

Examples of significant errors include: the result presented does not actually answer the stated question, the code produces errors, the code *consistently* uses base Python when `pandas` methods would be more appropriate, or the submitted notebook contains significant quantities of code that is unrelated to the presented analysis (such as copy/pasted code from the curriculum or StackOverflow)

Does Not Meet Objective Unable to prepare data using `pandas`

This includes projects that successfully answer the business questions, but do not use `pandas` (e.g. use only base Python, or use some other tool like R, Tableau, or Excel)

Interactive Data Visualization

Tableau is a powerful data analysis tool that allows data to be presented in a manner that allows it to be easily digestible with visualizations and charts to aid in the simplification of the data and its analysis. Tableau contains many customizable features and makes it easy to share in many ways. We recommend you use Tableau for your interactive data visualization now that you have experience with it.

Here are the definitions of each rubric level for this objective.

Exceeds Objective Creates an easy to use dashboard to answer business questions

Meets Objective Successfully creates a dashboard to answer business questions

Approaching Objective Creates a dashboard, but it is difficult to use

Does Not Meet Objective Unable to create a dashboard

Getting Started

Please start by reviewing the contents of this project description. If you have any questions, please ask your instructor ASAP.

Next, you will need to complete the ***Project Proposal*** which must be reviewed by your instructor before you can continue with the project.

Then, you will need to create a GitHub repository. There are three options: Interactive Data Visualization 1. Look at the Phase 1 Project Templates and Examples repo and follow the directions in the MVP branch. 2. Fork the Phase 1 Project Repository, clone it locally, and work in the `student.ipynb` file. Make sure to also add and commit a PDF of your presentation to your repository with a file name of `presentation.pdf`. 3. Create a new repository from scratch by going to github.com/new and copying the data files from one of the above resources into your new repository. This approach will result in the most professional-looking portfolio repository, but can be more complicated to use. So if you are getting stuck with this option, try one of the above options instead.

Summary

This project will give you a valuable opportunity to develop your data science skills using real-world data. The end-of-phase projects are a critical part of the program because they give you a chance to bring together all the skills you've learned, apply them to realistic projects for a business stakeholder, practice communication skills, and get feedback to help you improve. You've got this!