**Overview- Build Your Data Science Project**

In this capstone project, you will leverage what you’ve learned throughout the program to build a data science project of your choosing. Your project deliverables are:

1. A Github repository of your work.
2. A blog (or other medium for a write-up) post written for a technical audience, or a deployed web application powered by data.

You'll follow the steps of the data science process that we've discussed:

1. You will first **define** the problem you want to solve and investigate potential solutions.
2. Next, you will **analyze** the problem through visualizations and data exploration to have a better understanding of what algorithms and features are appropriate for solving it.
3. You will then **implement** the algorithms and metrics of your choice, documenting the preprocessing, refinement, and post-processing steps along the way.
4. Afterwards, you will collect **results** about your findings, visualize significant quantities, validate/justify your results, and make any concluding remarks about whether your implementation adequately solves the problem.
5. Finally, you will **construct** a blog post (or other medium for a write-up) to document all of the steps from start to finish of your project, or deploy your results into a web application.

**Setting Yourself Apart**

An important part of landing a job or advancing your career as a data scientist is setting yourself apart through impressive data science projects. By now, you've completed several guided projects, and now's your chance to show off your skills and creativity. You'll receive a review of your project with feedback from a Udacity mentor, and they will focus on how your project demonstrates your skills as a well-rounded data scientist.

This project is designed to prepare you for delivering a polished, end-to-end solution report of a real-world problem in a field of interest. When developing new technology, or deriving adaptations of previous technology, properly documenting your process is critical for both validating and replicating your results.

Things you will learn by completing this project:

* How to research and investigate a real-world problem of interest.
* How to accurately apply specific data science algorithms and techniques.
* How to properly analyze and visualize your data and results for validity.
* How to document and write a report of your work.

NEXT

## [Optional] Software Requirements

**Your project must be written in Python 3.x**. Given the free-form nature of the data scientist capstone, the software and libraries you will need to successfully complete your work will vary depending on the chosen application area and problem definition. Because of this, it is imperative that all necessary software and libraries used in your capstone project are accessible to the reviewer and clearly documented. Information regarding the software and libraries that your project makes use of should be included in your README document in your GitHub repo, along with your submission. Please note that proprietary software, software that requires private licenses, or software behind a paywall or login account should be avoided.

## Data Requirements

Every data scientist capstone project will most certainly require some form of dataset or input data structure (input text files, images, etc.). Similar to the software requirements above, the data you use must either be publicly accessible or provided by you during the submission process, and private or proprietary data should not be used without expressed permission. Please take into consideration the file size of your data — while there is no strict upper limit, input files that are excessively large may require reviewers longer than an acceptable amount of time to acquire all of your project files and/or execute the provided development code. This can take away from the reviewer's time that could be put towards evaluating your submission. If the data you are working with fits the criteria of being too large, consider whether you can work with a subset of the data instead, or provide a representative sample of the data which the reviewer may use to verify the solution explored in the project.

## Ethics

Udacity's A/B Testing course has a segment that discusses [**the sensitivity of data**](https://classroom.udacity.com/courses/ud257/lessons/3998098714/concepts/39997087540923) (free course link) and the expectation of privacy from those whose information has been collected. While most data you find has been made available to the public will not have any ethical complications, it is extremely important that you are considering where the data you are using came from, and whether that data contains any sensitive information. For example, if you worked for a bank and wanted to use customers' bank statements as part of your project, this would most likely be an unethical choice of data and should be avoided.

## Note on Spark Project

Not all of these instructions regarding software and data requirements apply if you choose the Spark project option, which comes with its own separate set of software and data requirements. You can find more details on this in the Spark course content in the Extracurriculars section of your classroom.

NEXT

### Selecting a Project

Think about a technical field or domain that you are passionate about, such as robotics, virtual reality, finance, natural language processing, or even artificial intelligence (the possibilities are endless!). Then, choose an existing problem within that domain that you are interested in which you could solve by applying data science techniques. Be sure that you have collected all of the resources needed (such as a dataset) to complete this project, and make the appropriate citations wherever necessary in GitHub (and your blog, if that is the path you decide to pursue). Below are a few suggested problem areas you could explore if you are unsure what your passion is:

* [**Robot Motion Planning**](https://docs.google.com/document/d/1ZFCH6jS3A5At7_v5IUM5OpAXJYiutFuSIjTzV_E-vdE/pub)
* [**Healthcare**](https://docs.google.com/document/d/1WzurKKa9AX2DnOH7KiB38mvozdOSemfkGpex8hdTy8c/pub)
* [**Computer Vision**](https://docs.google.com/document/d/1y-XfjkPFgUQxFIQ9bBncUSjs4HOf5E-45FrLYNBsZb4/pub)
* [**Education**](https://docs.google.com/document/d/1vjerjRQnWs1kLbZagDYT6rNqiwAG23Yj45oUY88IAxI/pub)
* [**Investment and Trading**](https://docs.google.com/document/d/1ycGeb1QYKATG6jvz74SAMqxrlek9Ed4RYrzWNhWS-0Q/pub)

In addition, you may find a technical domain (along with the problem and dataset) in a competition on platforms such as [**Kaggle**](http://kaggle.com/), or **[Devpost](http://devpost.com/" \t "_blank)**. This can be helpful for discovering a particular problem you may be interested in solving as an alternative to the suggested problem areas above. In many cases, some of the requirements for the capstone project are already defined for you when choosing from these platforms.

#### Udacity Specific Projects

* **Customer Segmentation Report for Arvato Financial Services**
* **Optimizing App Offers With Starbucks**
* **Use Convolutional Neural Networks to Identify Dog Breeds**
* **Using Spark to Predict Churn with Insight Data Science**

#### Check out Sample Projects

* Here are two projects originally created for the Machine Learning Engineer Nanodegree program, but which give you an idea of what a final blog post might look like. Each of these meets the requirements for the capstone project for the Data Scientist Nanodegree: [**sample project 1**](https://github.com/udacity/machine-learning/blob/master/projects/capstone/report-example-1.pdf) and [**sample project 2**](https://github.com/udacity/machine-learning/blob/master/projects/capstone/report-example-3.pdf).

It isn't necessary that your project be 10 or 20 pages. Rather, you should aim to solve a problem that interests you, build a project you are proud of, and explain your results to a technical audience. The details of how you decide to do this are largely up to you, but do use the rubric linked below to make sure your project meets specifications. This is intended to be an open-ended project that allows you to apply what you have learned to your own area of interest.

No matter what project you decide to complete, you will want to check your work against the [**project rubric here**](https://review.udacity.com/#!/rubrics/2345/view).

NEXT