This project is the culmination of all you’ve learned in this course!  We’ve allotted you two weeks to work on this project and you should expect to spend 10-20 hours total on the project.  **Be sure to read all of the items below before starting the project:**

* + - This project description
    - [The slides you will modify to present your work](https://docs.google.com/presentation/d/1XvqgINDstiiO_lu-Ke8CJQ7sur_7FcrHH8G_w5yr114/edit#slide=id.g1e0c2d5179_0_20)
    - The rubric for evaluating your work (next reading)

There are a number of steps outlined below, but it is **critical** that you do not view this as an entirely linear process.  Remember that the science component in data science is the creation of a hypothesis based on exploration and testing of that hypothesis is through analysis.  You may need to go through many of these steps multiple times before you arrive at meaningful hypothesis or conclusions.

**Deliverables**:  At the end of week 10, you will submit a single PDF file with your notebook and your slides.  Keep in mind, you are limited to a 5MB PDF, so be judicious about including high-resolution images in your submitted presentation and notebook.

* + - To save your Jupyter notebook as a PDF, File -> Print will open a static view of the notebook in your browser. You may then use your computer's Print or Save functionality to save the page as a PDF.

**Week 9**

**Step 1:  Find a dataset or datasets**

Based on your interest, identify a dataset which you will want to examine.  There is a reading this week on where you can find open datasets, but feel free to use other datasets you have access to and care publicly share results about.  Or, feel free to create your own dataset by getting data from the web using tools you learned in Week 8.

This step may take some time, as you’ll likely look at a number of datasets before you find one (or more) which holds promising data for the kinds of questions you want to ask.  Feel free to use a combination of datasets which you can merge in some meaningful way.

**Step 2:  Explore the dataset(s)**

In this step, you should explore what is present in the data and how the data is organized.  If using multiple datasets, you’ll need to determine what common features allow you to merge the datasets.  You’ll want to answer the following questions:

* + - Are there quality issues in the dataset (noisy, missing data, etc.)?
    - What will you need to do to clean and/or transform the raw data for analysis?

**Step 3:  Identify 1-3 research questions**

Now that you have a better understanding of the data, you will want to form a research question which is interesting to you. The research question should be broad enough to be of interest to a reader but narrow enough that the question can be answered with the data.  Some examples:

* + - **Too Narrow:**  *What is the GDP of the U.S. for 2011?*  This is just asking for a fact or a single data point.
    - **Too Broad:**  *What is the primary reason for global poverty?*  This could be a Ph.D. thesis and would still be way too broad.  What data will you use to answer this question?  Even if a single dataset offered an answer, would it be defendable given the variety of datasets out there?
    - **Good (just right):**  *Can you use sentiment analysis on comments about movies on Twitter to predict its box office earnings?*  If you have, or can obtain, tweets which refer to a variety of movies and you have their box office earnings, this is a question which you can potentially answer well.

**Step 4:  Describe your dataset(s) and your research question(s) at the end of Week 9 in the Project.**

**Checkpoint.**

**Week 10**

**Step 5:  Identify your research methods**

Based on your research question, you can now describe how you will go about answering that question.  Your methods will depend on the question being asked.  For example, if you are looking for a relationship between two items (say, CO2 emissions and GDP), you may wish to use scatter plots and statistical correlation.  If you are trying to predict an outcome based on input data, you’ll need to identify the appropriate methods from the week on Machine Learning.  Be sure to document, in your notebook, your research methods.

**Step 6:  Present your findings**

In this step, you can begin to report your findings.  What did you learn from the data and how do your findings help answer your research question?  Use visualizations to present these findings.

**Step 7:  Identify limitations to your conclusions**

Rarely will a single data analysis conclusively answer a research question.  Here, you need to identify possible limitations.  For example, are your results limited to a certain area, city, or country?  Are you making assumptions about the data which may, or may not, be valid (e.g., that students in one term are equally qualified as students in another)?  Document these limitations.

**Step 8:  (Optional) Identify future work**

You may have clear next steps at the end of your analysis.  For example, you may wish to perform the same analysis in another setting or you think your findings lead to new questions.  Feel free to document these next steps.

**Step 9: Present your work!**

[Fill in this slide template to present your project.](https://docs.google.com/presentation/d/1XvqgINDstiiO_lu-Ke8CJQ7sur_7FcrHH8G_w5yr114/edit?usp=sharing)  Feel free to add more slides, but aim to keep your entire presentation **below** **20 total slides**.  We also recommend you attach a PDF of your Jupyter notebook to this presentation in one single PDF file to be uploaded. As you are writing, please follow the guidelines below:

**General Guidelines for Project Reports and Presentations:**

* 1. This is academic writing. Keep it formal and coherent as a self-contained entity. Anyone reading your presentation should have a full understanding of the question, approach and the results.
  2. Be professional. You should be comfortable giving this presentation to the general public, your boss, or your academic advisor.
  3. Write for a diverse audience including:
     1. **General Public:** Reads only the title and abstract looking for high-level point mainly for conversational purposes.
     2. **A company CEO:** Reads introduction, research question, findings and conclusions looking for business value and details related to that.
     3. **An academic advisor (or company CTO):** Reads the full presentation AND your Jupyter Notebook paying particular attention to technical coherence, academic value, and technical data science strengths.

The final presentation should also be for the above-mentioned three audiences. Think of the diversity of the audience. The whole point is to **tell a story** - so you should motivate a reader to care based on the question you are exploring, answer that question in a clear and concise manner, provide an honest appraisal of your results, and give the reader valuable insights.  Use charts whenever possible. Avoid slides with a lot of text and bullet points - break the slide into multiple slides when this happens.  Be concise!

This is the Rubric others will use to assess your project (and what you'll use to assess other projects). It is based primarily on the slides you'll submit.  It is worth reviewing this to know how you will be assessed before starting your work and while preparing your presentation slides.

### Rubric

**<Abstract Slide> How well did the abstract describe the project?**

3 - Very well.  It described the dataset, research question, methods, and findings well.

2 - Well.  Three of the four elements (dataset, research question, methods, and findings) were explained well, one was not.

1 - Fairly well.  One or two of the four elements (dataset, research question, methods, and findings) were explained, however two or more of the above items is missing.

0 - Absent.  No abstract provided.

**<Abstract Slide> Was the abstract concise and clear?**

2 - Yes, very. It was concise, clear, and well written.

1 - Yes.  However it was longer than recommended or was hard to read.

0 - No.  No abstract provided.

**<Motivation Slide> Did the author motivate you to care about the issue/problem they are trying to address/solve well?**

2 - Well.  The motivation for the project is clear.

1 - Fairly well.  A motivation was provided but it isn’t clear why the work is valuable or for whom.

0 - Absent.  No motivation provided.

**<Dataset Slide>  Did the author describe the dataset well?**

2 - Well.  They provided information on what data is in the dataset, how much data, and where to find the dataset (if applicable).

1 - Fairly well.  They provided information about the dataset however important facets of the dataset were unclear.

0 - Absent.  No dataset provided.

**<Data Preparation Slide>  Did the author describe what data preparation and/or cleaning was required?**

1 - Yes.  They provided basic information on what data preparation was required.

0 - No.  No description of data preparation was provided.

**<Research Question(s) Slide>  How appropriate was the research question for the data?**

2 - Appropriate.  They provided a research question which was appropriate given the data in the dataset.

1 - Too Narrow.  They provided a research question, however it felt too narrow for a full data-analysis (e.g., it asks for a single datapoint, not an analysis of the data).

1 - Too Broad .  They provided a research question, however it felt too broad given this dataset (e.g., it asks to reach a sweeping conclusion which is too broad for one dataset).

0 - Absent.  No research question provided.

**<Method Slide> How well were the research methods articulated?**

2 - Well. It was clear from their methods how they were analyzing the data.

1 - Fairly well. Methods were provided but it was not clear how they were being used and for what data.

0 - Absent.  No methods provided.

**<Method Slide> How appropriate were the research methods used?**

3 - Unsure. If you are unsure or don’t feel qualified to assess the research methods, please select this option.

3 - Very appropriate.  These are reasonable research methods to answer the research question (e.g., the author is using regression to predict a numeric value).

2 - Possibly Appropriate.  The methods described may not be ideal for this approach or the methods are not described well enough to tell.

1 - Not appropriate.  Methods, as described, are inappropriate (e.g., they are using both their training and test sets when building the machine learning model).

0 - Absent.  No methods provided.

**<Findings Slide(s)>  Are their findings well presented?**  (Remember, they need not reach a definitive answer to the research question, but they need to articulate what they found clearly and how it relates to the original research question.)

3 - Very well.  The research findings are well articulated and presented.

2 - Well.  The research findings are presented, however the results were not clear.

1 - Fairly well.  The research findings were not well presented, it is unclear what they found in the study.

0 - Absent.  No findings were presented.

**<Findings Slide(s)>  Please rate the quality of their visualizations.**  Specifically, were their visualizations honest, accessible (for a general audience), and elegant?  One visualization is required, but if multiple are provided, please choose the best one and say which one you are using in the comments below.

3 - Excellent.  The visualization(s) were honest, accessible, and elegant.

2 - Very good.  However, the visualization(s) only met two of the three criteria.

1 - Fair. Visualizations were provided but the met, at most, one of the three criteria.

0 - Absent.  There were no visualizations provided.

**<Limitations Slide>  Did the author describe appropriate limitations to their results?**

2 - Yes-well.  Limitations were well described.

1 - Yes-partially.  Limitations were described but clear limitations were absent from their response (e.g., their analysis was limited to wealthy adults in a single country but they did not mention this limitation in their presentation - assuming their sampling was appropriate for all people).

0 - No.  No limitations were described.

**<Conclusions Slide>  Please rate the quality of the authors conclusions.**

3 - Excellent.  The results of the work were well summarized, accurately portrayed, and well written.

2 - Good.  The conclusion was missing applicable elements or was poorly written.

1 - Fair.  The conclusion was missing important elements or the conclusion was not warranted given the findings (e.g., too broad of claims were made, the claims were misleading, etc.)

0 - Absent.  No conclusion was provided.

**<References Slide>  Did the author provide appropriate references?**

1 - Yes. Either they did the work entirely on their own or they referenced appropriate materials.

0 - No.  They failed to reference a clear source.

**<OVERALL>  Please rate the overall analysis.**

4 - Excellent.  The analysis was very well done.  It was motivated well, the methods were appropriate, and the findings were interesting.

3 - Good.  The analysis was well done.  Most of the elements of the analysis were good, however, there were one or two elements lacking (which received lower marks earlier) which may partially limit its value.

2 - Fair.  The analysis was fairly well done.  Some of the elements of the analysis were good, however, there were a number of elements lacking (which received lower marks earlier) which limited its value.

1 - Poor.  There were serious issues throughout the analysis which severely limits its value.

0 - Absent.  No PDF of their analysis slides was provided.

Written Feedback for the project overall:

**What aspects of the project stood out to you? What did it do well? How could it be improved?**

Please be constructive and helpful with your feedback.  Be sure to praise the positive elements as well as suggesting possible improvements.

There are a surprising number of open datasets on the web.  There are also a number of excellent websites which help keep track of where to find datasets.  Rather than attempt to recreate the good work from these teams, here's where you can find these collections:

[**UCI's Machine Learning Repository**](https://archive.ics.uci.edu/ml/datasets.html)- UC Irvine maintains a fantastic collection of datasets for machine learning, tagged by machine learning task (among other things).  If you want to try out the ideas from Week 7, this is the place to start.

[**Kaggle**](https://www.kaggle.com/datasets) - Online collection of datasets.  Kaggle also has competitions for data mining and information about jobs in data science.

[**KDnuggets Dataset**](http://www.kdnuggets.com/datasets/index.html) - Another collection of datasets - most of the datasets are free. Within KDnuggets, there are links for [government data](http://www.kdnuggets.com/datasets/government-local-public.html),  [Data APIs](http://www.kdnuggets.com/datasets/api-hub-marketplace-platform.html),  and [Data Mining Competitions](http://www.kdnuggets.com/competitions/index.html).

If this seems like information overload, feel free to just get started with data at some of our favorite sites: [**US Government Data**](https://www.data.gov/), [**UK Government Data**](https://data.gov.uk/), [**Canada's Open Data Exchange**](https://codx.ca/), [**World Health Organization**](http://www.who.int/gho/en/), and the [**World Bank**](http://data.worldbank.org/).

Remember, finding a dataset of interest and exploring it is most of your job for Week 9!  So expect this to take some time - and that's both perfectly normal and completely okay.