Udacity’s Data Science Nanodegree

Sumário

[Nanodegree Summary 2](#_Toc403717476)

[Página Inicial do Curso 3](#_Toc403717477)

[Projeto 1 – Analyzing NYC Subway Dataset 4](#_Toc403717478)

[Project Overview 4](#_Toc403717479)

[How do I Complete this Project? 5](#_Toc403717480)

[Short Questions 5](#_Toc403717481)

[Evaluation and Submission 5](#_Toc403717482)

[Projeto 2 – Data Wrangle OpenStreetMaps Data 6](#_Toc403717483)

[Project Overview 6](#_Toc403717484)

[Project Details 6](#_Toc403717485)

[Project Rubric Details 6](#_Toc403717486)

[Evaluation and Submission 9](#_Toc403717487)

[Projeto 3 – Explore and Summarize Data 10](#_Toc403717488)

[Project Overview 10](#_Toc403717489)

[How do I Complete this Project? 10](#_Toc403717490)

[Project Details 10](#_Toc403717491)

[Step Five - Knit your RMD file 11](#_Toc403717492)

[Step Six - Document your Data (if you chose your own data set) 11](#_Toc403717493)

[Data Set Options 11](#_Toc403717494)

[Data Analysis with R Project Rubric 12](#_Toc403717495)

[Project Example - Diamonds 14](#_Toc403717496)

# Nanodegree Summary

**Nanodegree Summary**

This nanodegree program is the most efficient curriculum to prepare you for a job as a Data Analyst. You will learn to:

* Wrangle, extract, transform, and load data from various databases, formats, and data sources
* Use exploratory data analysis techniques to identify meaningful relationships, patterns, or trends from complex data sets
* Classify unlabeled data or predict into the future with applied statistics and machine learning algorithms
* Communicate data analysis and findings well through effective data visualizations

You will work with your peers and advisors on projects approved by leading employers as the critical indicators of job-readiness. We designed these projects with expert Data Analysts, Data Scientists, and hiring managers.

**What is a Nanodegree?**

A nanodegree is a **new type of credential**, designed **to prepare you for a job**.

It is **built with industry** for you to master skills that employers truly seek in a Data Analyst.

It is **project-based**: you'll complete several projects, supported by our community of Coaches and your fellow students, to learn and show off your skills.

It's **flexible**: take only the courses you need to ace projects!

[See how it works](https://www.udacity.com/nanodegrees#relevant-education)

**Why Take This Nanodegree?**

The Data Analyst nanodegree is designed to prepare you for a career in data science. As a Data Analyst, you will be responsible for obtaining, analyzing, and effectively reporting on data insights ranging from business metrics to user behavior and product performance.

Working with our leading industry partners, we’ve carefully designed the most efficient set of projects to train you on the fundamental skills of data science and help you become a Data Analyst.

**Prerequisites and Requirements**

Data Analyst nanodegree students...

* are interested in data science.
* have a strong grasp of descriptive and inferential statistics.
* have python programming experience.
* have a strong understanding of programming concepts such as variables, functions, loops, and basic data structures like lists and dictionaries.

Prerequisite:

The first thing you should do is take the [Readiness Assessment](https://www.udacity.com/course/viewer#!/c-none/l-2364518584/m-2782888549), which will help you determine if you have the right background to participate.

General Requirements:

* You are self-driven and motivated to learn. Participation in this program requires consistently meeting the deadlines for your cohort and devoting at least 10 hours per week to your work.
* You can communicate fluently and professionally in written and spoken English.
* You have access to a computer with a broadband connection, on which you’ll install a professional code/text editor (ie. Sublime Text or Atom) and programming languages like Python and R and associating data science libraries.
* You are willing to contribute to the success of the program, including collaborating with fellow students and giving us feedback on how we can improve.

See the [Technology Requirements](https://www.udacity.com/tech-requirements) for using Udacity.

**Nanodegree Structure**

These are the projects you'll build and the classes that will prepare you to build each of the projects. You'll have access to all these in the Nanodegree:

**Predict the New York City Subway Ridership**

In this project, you will be exposed to and learn fundamental data science skills like data wrangling, applied statistics and machine learning, effective visualization, and working with big data using MapReduce. Specifically, you will pose a question about the New York City Subway and work with the subway ridership dataset to draw an interesting conclusion about the dataset itself.

Prepare for this project with: [Intro to Data Science](https://www.udacity.com/course/ud359)

**Wrangle OpenStreetMaps Data**   
You will choose any area of the world in [Open Street Map](https://www.openstreetmap.org/)and use data munging techniques, such as assessing the quality of the data for validity, accuracy, completeness, consistency and uniformity, to clean the OpenStreetMap data for a part of the world that you care about.

Prepare for this project with: [Data Wrangling with MongoDB](https://www.udacity.com/course/ud032)

**Explore, Summarize, and Discover Interesting Insights from Datasets**   
You will use R to apply exploratory data analysis techniques that you have to the project. You will practice understanding a single variable and relationships between multiple variables, and explore a selected data set for distributions, outliers, and anomalies.

Prepare for this project with: [Data Analysis with R](https://www.udacity.com/course/ud651)

**Identify Fraud from the Enron Email Dataset**   
In this project, you will play detective, and put your machine learning skills to use by building an algorithm to identify Enron Employees who may have committed fraud based on the public Enron financial and email dataset.

Prepare for this project with: [Intro to Machine Learning](https://www.udacity.com/course/ud120)

**Tell Stories with Data Visualization**   
For this project, you will create a data visualization from a data set that tells a story or allows a reader to explore trends or patterns. Your work will be a reflection of the theory and best practices of data visualization.

Prepare for this project with: [Data Visualization](https://www.udacity.com/course/ud507)

# Página Inicial do Curso

PROJECT**P1: Analyzing the NYC Subway Dataset**

**Project Description**

In this project, you look at the NYC Subway data and figure out if more people ride the subway when it is raining versus when it is not raining.

You will wrangle the NYC subway data, use statistical methods and data visualization to draw an interesting conclusion about the subway dataset that you've analyzed.

**Learn the skills for the project**



**Intro to Data Science**

[Go to class](https://www.udacity.com/course/viewer#!/c-ud359-nd)

**Build & submit your project**   
[View project details](https://www.udacity.com/course/viewer/#!c-nd002/l-3176718735)

PROJECT**P2: Data Wrangle OpenStreetMaps Data**

**Project Description**

You will choose any area of the world in [https://www.openstreetmap.org](https://www.openstreetmap.org/) and use data munging techniques, such as assessing the quality of the data for validity, accuracy, completeness, consistency and uniformity, to clean the OpenStreetMap data for a part of the world that you care about.

**Learn the skills for the project**



**Data Wrangling with MongoDB**

[Go to class](https://www.udacity.com/course/viewer#!/c-ud032-nd)

**Build & submit your project**   
[View project details](https://www.udacity.com/course/viewer/#!c-nd002/l-3168208620)

PROJECT**P3: Explore and Summarize Data**

**Project Description**

In this project, you will use R and apply exploratory data analysis techniques to explore relationships in one variable to multiple variables and to explore a selected data set for distributions, outliers, and anomalies.

**Learn the skills for the project**



**Data Analysis with R**

[Go to class](https://www.udacity.com/course/viewer#!/c-ud651-nd)

**Build & submit your project**   
[View project details](https://www.udacity.com/course/viewer/#!c-nd002/l-3165188753)

PROJECT**P4: Identifying Fraud from Enron Email**

**Project Description**

In this project, you will play detective, and put your machine learning skills to use by building an algorithm to identify Enron Employees who may have committed fraud based on the public Enron financial and email dataset.

**Learn the skills for the project**



**Intro to Machine Learning**

[Go to class](https://www.udacity.com/course/viewer#!/c-ud120-nd)

**Build & submit your project**   
[View project details](https://www.udacity.com/course/viewer/#!c-nd002/l-3174288624)

PROJECT**P5: Make Effective Data Visualization**

**Project Description**

For this project, you will create a data visualization from a data set that tells a story or allows a reader to explore trends or patterns. You will need to use either dimple.js or d3.js to create the visualization. Your work should be a reflection of the theory and practice of data visualization, such as visual encodings, design principles, and effective communication.

**Learn the skills for the project**



**Data Visualization and D3.js**

[Go to class](https://www.udacity.com/course/viewer#!/c-ud507-nd)

**Build & submit your project**   
[View project details](https://www.udacity.com/course/viewer/#!c-nd002/l-3184238632)

# Projeto 1 – Analyzing NYC Subway Dataset

## Project Overview

In this project, you look at the NYC Subway data and figure out if more people ride the subway when it is raining versus when it is not raining.

You will wrangle the NYC subway data, use statistical methods and data visualization to draw an interesting conclusion about the subway dataset that you've analyzed.

**Note**

You do not need to complete this project if you have received a certificate from completing the Intro to Data Science course previously.

Please email **nd002subwayanalysis-project@udacity.com** with your certificate to get the completion phrase to complete this project.

**What do I Need to Install? (optional)**

If you want to complete the programming exercises on your own computer or laptop, you will need to install [Anaconda Scientific Python Distribution](https://store.continuum.io/cshop/anaconda/).

It should contain most of the libraries and packages that you need to work on the assignments.

One caveat is that Anaconda does not include pandasql (needed to complete project #2), but you can easily install pandasql through pip as below:

pip install -U pandasql

**Why this Project?**

This project will introduce you to the key concepts of data science, so you will be prepared for subsequent projects in the Data Analyst nanodegree as well as your future career as a data analyst.

In addition, you will be exposed to some of the most popular data science libraries in python, such as Pandas, Numpy, and others.

**What will I learn?**

You will be exposed to and learn fundamental data science skills like:

* data wrangling
* applied statistics and machine learning
* effective visualization
* How to work with big data using MapReduce.

**Why is this Important to my Career?**

By completing this project, you will have exhibited all of the skills needed to be a data analyst. In addition, you can add this project to your portfolio, which can help you impress recruiters and hiring managers.

Details

## How do I Complete this Project?

This project is connected to the [Intro to Data Science course](https://www.udacity.com/course/ud359-nd), but depending on your background knowledge of data science, you may not need to take the whole thing to complete this project.

Here's what you should do:

1. Complete all of the questions in Projects 2 through 5 in the [Intro to Data Science course](https://www.udacity.com/course/ud359-nd)
2. Answer these [short questions](https://docs.google.com/document/d/16T3kirC0IxvtfxlZb7n5kOz5xFF_JTwrG31J2OZj8KM/pub) and email your answers as a text file, word document, or as a pdf to nd002nysubway-project@udacity.com

Analyzing the NYC Subway Dataset

## Short Questions

**Overview**

For Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course.

For part 2 of the project, please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets.

**Section 1. Statistical Test**

1. Which statistical test did you use?
2. Why is this statistical test appropriate or applicable to the dataset?
3. What results did you get from this statistical test?
4. What is the significance of these results?

**Section 2. Linear Regression**

1. What approach did you use to compute the coefficients theta and produce prediction in your regression model:
2. Gradient descent (as implemented in exercise 3.5)
3. OLS using Statsmodels
4. Or something different?
5. What features did you use in your model?
6. Why are these features appropriate?
7. What is your model’s R2 (coefficients of determination) value?
8. What does this R2 value mean for the goodness of fit for your regression model?

**Section 3. Visualization**

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatterplots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

1. One visualization should be two histograms of ENTRIESn\_hourly for rainy days and non-rainy days
2. One visualization can be more freeform, some suggestions are:
3. Ridership by time-of-day or day-of-week
4. How ridership varies by subway station
5. Which stations have more exits or entries at different times of day

**Section 4. Conclusion**

1. From your analysis and interpretation of the data, do more people ride  
   the NYC subway when it is raining versus when it is not raining?
2. What analyses lead you to this conclusion?

**Section 5. Reflection**

1. Please discuss potential shortcomings of the dataset and the methods   
   of your analysis.

## Evaluation and Submission

Note

You do not need to complete this project if you have received a certificate from completing the Intro to Data Science course previously.

Please email **nd002subwayanalysis-project@udacity.com** with your certificate to get the pass phrase to complete this project.

Evaluation

An Udacity reviewer will review and check your completion of problem sets 2 through 5 in Intro to Data Science, as well as your answers to the questions listed in the survey.

Be sure to complete all of questions in the problem sets and the survey before sending an email to your Udacity reviewer.

Rubric

Your project will be evaluated by a Udacity reviewer according to this [project rubric](https://docs.google.com/a/knowlabs.com/document/d/1ZWdmlEgtRhreyN7AaiEfoYP70GqxOZqrajWtzIov8HM/pub). Be sure to review it thoroughly before you submit. Your "project meets specifications" if it meets specifications in all the criteria

Submission

Send an email to nd002subwayanalysis-project@udacity.com with the following:

1. Your python code to the problem sets questions.
2. A list of Web sites, books, forums, blog posts, github repositories etc that you referred to or used in this submission (Add N/A if you did not use such resources).
3. Please carefully read the following statement and include it in your email:

*“I hereby confirm that this submission is my work. I have cited above the origins of any parts of the submission that were taken from Websites, books, forums, blog posts, github repositories, etc. By including this in my email, I understand that I will be expected to explain my work in a video call with a Udacity coach before I can receive my verified certificate.”*

A Udacity reviewer will respond shortly with next steps.

# Projeto 2 – Data Wrangle OpenStreetMaps Data

## Project Overview

You will choose any area of the world in [https://www.openstreetmap.org](https://www.openstreetmap.org/) and use data munging techniques, such as assessing the quality of the data for validity, accuracy, completeness, consistency and uniformity, to clean the OpenStreetMap data for a part of the world that you care about.

**Note**

You do not need to complete this project if you have received a certificate from completing the Data Wrangling course previously.

Please email **nd002datawrangling-project@udacity.com** with your certificate to get the completion phrase to complete this project.

**Why this Project?**

What’s so hard about retrieving data from databases or various files formats? You grab some data from this file and that database, clean it up, merge it, and then feed it into your state of the art, deep learning algorithm … Right?

But the reality is this -- anyone who has worked with data extensively knows it is an absolute nightmare to get data from different data sources to play well with each other.

And this project will teach you all of the skills you need to deal with even the most nightmarish data wrangling scenarios.

**What will I learn?**

After completing the project, you will be able to:

* Assess the quality of the data for validity, accuracy, completeness, consistency and uniformity
* Parsing and gather data from popular file formats such as json, xml, csv, html
* Process data from many files and very large files that can be cleaned with spreadsheet programs
* Learn how to store, query, and aggregate data using MongoDB

**Why is this Important to my Career?**

As this [New York Times article](http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html) points out, the less heralded part of doing data science is manually collecting and cleaning data so it can be easily explored and analyzed later. Or otherwise known as “data wrangling” or “data munging” in the data science community.

Though not as glamorous as building cool machine learning models, data wrangling is a task that data scientists can spend up to 50-80% of their time doing according to many practicing data analyst and data scientists.

## Project Details

**How do I Complete this Project?**

This project is connected to the [Data Wrangling with MongoDB course](https://www.udacity.com/course/ud032), but depending on your background knowledge of data wrangling, you may not need to take the whole thing to complete this project.

Here's what you should do:

1. Solve all Lesson 6 programming exercises
2. Choose any area of the world in [https://www.openstreetmap.org](https://www.openstreetmap.org/) and apply the same techniques to this data set. The data set should be at least 50MB of size. See the project rubric for more details.
3. Read and fill in the project rubric. The project description and the the rubric can be found in [this document](https://docs.google.com/a/knowlabs.com/document/d/1gGilaago46aT_5tXnoILY4pyl7UytdB8QzXfghcZT0A/pub)

## Project Rubric Details

Final Project Overview for Data Wrangling with MongoDB

The Final Project of Data Wrangling consists of solving the Lesson 6 programming exercises, then choosing any area of the world in [https://www.openstreetmap.org](https://www.google.com/url?q=https%3A%2F%2Fwww.openstreetmap.org&sa=D&sntz=1&usg=AFQjCNGwV_1ytXhGxPJcFa9SZNl1RvAymQ) and applying the same techniques to this data set. The data set should be at least 50MB of size. After that you should be able to fill in the project rubric.

The following documents will help you understand the logistics of the final project, including grading and submission. These documents are intended for students with a Udacity Coach who have enrolled in the[full course experience](https://www.google.com/url?q=https%3A%2F%2Fwww.udacity.com%2Fsuccess&sa=D&sntz=1&usg=AFQjCNHM-Zx--lRbyXkKu51NBBvh3DydDg). If you are previewing the courseware, you are welcome to look at these documents as well (but understand that you will not submit your project).

1. [Project Summary](https://docs.google.com/document/d/1gGilaago46aT_5tXnoILY4pyl7UytdB8QzXfghcZT0A/pub#h.4e11crbdd5ze) (You must fill out this document and include it when you submit your project).
2. [Project Submission Instructions](https://docs.google.com/document/d/1gGilaago46aT_5tXnoILY4pyl7UytdB8QzXfghcZT0A/pub#h.9jd4cnblf2j9) (This will tell you how to submit your project).
3. [Project Rubric](https://docs.google.com/document/d/1gGilaago46aT_5tXnoILY4pyl7UytdB8QzXfghcZT0A/pub#h.6t25ft7ati5j) (This is the rubric that a Udacity Project Evaluator will use when he/she is evaluating your project).

**Project Summary**

What is your name?

Gundega Dekena

What E-mail address do you use to sign in to Udacity?

[example@example.com](mailto:example@example.com)

What area of the world you used for your project? Post a link to the map position and write a short description. Note that the osm file of the map should be at least 50MB.

URL: [https://www.openstreetmap.org/export#map=14/41.97316/-87.69074](https://www.google.com/url?q=https%3A%2F%2Fwww.openstreetmap.org%2Fexport%23map%3D14%2F41.97316%2F-87.69074&sa=D&sntz=1&usg=AFQjCNGWL16ZqadEAB5N1h8xIFnD9dDSow)

I chose this particular place because it is my neighbourhood, I know it well and would like its map to be improved in quality!

Is there a list  of  Web sites, books, forums, blog posts, github repositories etc that you referred to or used  in this  submission (Add N/A if you did not use  such resources)?

Use this place to list the citations.

Please carefully read the following statement and include it in your email:

*“I hereby confirm that this submission is my work. I have cited above the origins of****any****parts of the submission that were taken from Websites, books, forums, blog posts, github repositories, etc. By including this in my email, I understand that I will be expected to explain my work in a video call with a Udacity coach before I can receive my verified certificate.”*

Is there any other important information that you would want your project evaluator to know?

Use this space to communicate with your project evaluator. Is there anything you would like to communicate? Feedback or suggestions?

**Project Submission Instructions [these apply for re-submission as well]**

* Go through the [Project Rubric](https://docs.google.com/document/d/1gGilaago46aT_5tXnoILY4pyl7UytdB8QzXfghcZT0A/pub#h.6t25ft7ati5j). When you feel confident that your project will pass, continue to the next step.
* Fill out the [Project Summary section](https://docs.google.com/document/d/1gGilaago46aT_5tXnoILY4pyl7UytdB8QzXfghcZT0A/pub#h.4e11crbdd5ze).
* Send an email to: DataWrangling-project@udacity.com.
* This email should contain the following:
* 1.  Your copy of the Project Summary [from above].
* 2. Your answers to the following questions described more in the rubric below: “Problems encountered in your map”,”Overview of the data”,”Other ideas about the datasets”. This can be attached to this email as a PDF or included as text in your email if you prefer.
* **3. Please attach your code for Lesson 6 exercises to the email.** You need to send a total of five Python text files and one example map file as follows:
* mapparser.py      [Iterative Parsing Exercise]
* tags.py                [Tag Types Exercise]
* users.py              [Exploring Users Exercise]
* audit.py               [Improving Street Names Exercise]
* data.py                [Preparing for Database Exercise]
* example.osm       [Small example part from the map that you used (around 1-10 MB of size)]
* Place these six files in a folder. Add any additional code you may have (optional) inside the folder, zip it and attach the zipped folder to the email.
* You will receive a confirmation email from [support@udacity.com](mailto:support@udacity.com) that your message was received.
* Within 7 business days, you will receive an email from your Project Evaluator (who will not be the coach you’ve been working with) with a completed rubric and instructions for next steps.
* If your project was evaluated as “Passing,” you will schedule an exit interview. The purpose of this interview is to talk to you about your project and verify that you are the person who created it. Don’t worry, this interview will not be difficult or stressful.
* If you still have questions, please review the [Udacity Project Submission FAQ](https://docs.google.com/a/knowlabs.com/document/d/1BPeNySE5xpkdcTeY9bINU16UNz7aqqxy3NzqQubds_I/pub) or email your coach.

**The Rubric**

**Overview**

This rubric is here to help you understand the expectations for the project that you create. It is the same rubric that the person evaluating your project will use. We will refer to this person as the “project evaluator” in this document. We recommend you look at the rubric **before you begin working** on your project **and again before you decide to submit it**.

**How to Use: before you begin**

1. Look at the bold headings under the criteria column to understand what the project evaluator will be looking for.
2. Go through each criteria item in more detail.
3. Familiarize yourself with what is required for your project to “meet expectations”. In order to gain a certificate, you need to "meet expectations", however, to gain most benefit/learn most from the experience, we encourage you to continue working on the project and posting your results/code on GitHub, personal website, OpenStreetMap.org etc.

**How to Use: before you submit**

1. Once your project is built, go through each criteria item and do your best to honestly evaluate where you think your project falls.
2. If you think your project “does not meet expectations” for **any**criteria item, then you should make some changes to your project.
3. Once you’re confident that your project “meets expectations” go ahead and follow the [project submission instructions](https://docs.google.com/document/d/1gGilaago46aT_5tXnoILY4pyl7UytdB8QzXfghcZT0A/pub#h.9jd4cnblf2j9) to submit!

**How Grading Works**

1. Your project evaluator will use this rubric to evaluate your project.
2. Your grade will simply be “pass” or “doesn’t pass.”
3. You earn a “pass” by not having **any** criteria items in the “does not meet expectations” column.
4. If any criteria item “does not meet expectations,” you will not pass. You will be able to make changes and re-submit.

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Does not meet expectation** | **Meets expectations** |
| **Code Functionality**  All Lesson 6 problems are solved correctly. | Not all required Lesson 6 questions are solved with the submitted code. | All required Lesson 6 questions are correctly solved with the submitted code. |
| **Code Readability**  Code is well structured  Code is commented as necessary. | Code does not follow an intuitive, easy-to-follow logical structure.  Code that is not intuitively readable is not well-documented with comments. | Code follows an intuitive, easy-to-follow logical structure.  Code that is not intuitively readable is well-documented with comments. |
| **Problems encountered in your map**  Student response describes the challenges encountered while auditing, fixing and processing the dataset for the area of their choice | Student response does not show understanding of potential and actual problems with the map data | Student response shows understanding of the process of auditing, and ways to correct or standardize the data, including dealing with problems specific to the location, eg related to language or traditional ways of formatting |
| **Overview of the data**  Student provides a statistical overview about their chosen dataset, like:   * size of the file * number of unique users * number of nodes and ways * number of chosen type of nodes, like cafes, shops etc | Student response does not provide an overview of a dataset, or the dataset is smaller than 50 MB.  Student response does not includes the MongoDB queries used to obtain the statistics. | Student response provides the statistics about their chosen map area.  Student response also includes the MongoDB queries used to obtain the statistics. |
| **Other ideas about the datasets**  Student is able to analyze the dataset and recognize opportunities for using it in other projects | Student response does not show ways to process and analyze provided datasets other than the ways that were already covered. | Student proposes one or more additional ways of improving and analyzing the data and gives thoughtful discussion about the benefits and anticipated problems in implementing the improvement. |
| **Thoroughness and Succinctness of Submission**  Student submission is long enough to thoroughly answer the questions asked without giving unnecessary detail. |  | *A good general guideline is that your question responses should take about 3-6 pages.* |

## Evaluation and Submission

**Note**

You do not need to complete this project if you have received a certificate from completing the Data Wrangling with MongoDB course previously.

Please email **nd002openstreetmap-project@udacity.com** with your certificate to get the pass phrase to complete this project.

**Evaluation**

An Udacity reviewer will review and check your completion of problem set 6 in Data Wrangling with MongoDB, as well as your answers to the questions listed in the rubric.

Be sure to complete all of questions in the problem set and the rubric before sending an email to your Udacity reviewer.

**Rubric**

Your project will be evaluated by a Udacity reviewer according to this [project rubric](https://docs.google.com/a/knowlabs.com/document/d/1gGilaago46aT_5tXnoILY4pyl7UytdB8QzXfghcZT0A/pub). Be sure to review it thoroughly before you submit. Your "project meets specifications" if it meets specifications in all the criteria

**Submission**

Send an email to nd002openstreetmap-project@udacity.com with the following:

1. Your python code to the problem set questions.
2. A a link to the map position you wrangled in your project, a short description of the area and a reason for your choice.
3. A list of Web sites, books, forums, blog posts, github repositories etc that you referred to or used in this submission (Add N/A if you did not use such resources).
4. Please carefully read the following statement and include it in your email:

*“I hereby confirm that this submission is my work. I have cited above the origins of any parts of the submission that were taken from Websites, books, forums, blog posts, github repositories, etc. By including this in my email, I understand that I will be expected to explain my work in a video call with a Udacity coach before I can receive my verified certificate.”*

A Udacity reviewer will respond shortly with next steps.

# Projeto 3 – Explore and Summarize Data

## Project Overview

In this project, you will use R and apply exploratory data analysis techniques to explore relationships in one variable to multiple variables and to explore a selected data set for distributions, outliers, and anomalies.

**Note**

You do not need to complete this project if you have received a certificate from completing the Data Analysis with R course previously.

Please email **nd002eda-project@udacity.com** with your certificate to get the pass phrase to complete this project.

**What do I need to install?**

In order to complete the project, you will need to install R. You can download and [install R from the**C**omprehensive **R** **A**rchive **N**etwork (CRAN)](http://cran.r-project.org/).

After installing R, you will need to download and install [R Studio](http://www.rstudio.com/products/rstudio/download/). Choose the appropriate installation for your operating system.

Finally, you will need to install a few packages. We recommend opening R Studio and installing the following packages using the command line.

install.packages("ggplot2", dependencies = T)

install.packages("knitr", dependencies = T)

install.packages("dplyr", dependencies = T)

For more information on installing R packages, please refer to [Installing R Packages](http://www.r-bloggers.com/installing-r-packages/) on R Bloggers.

**Why this Project?**

Exploratory Data Analysis (EDA) is the numerical and graphical examination of data characteristics and relationships before formal, rigorous statistical analyses are applied.

EDA can lead to insights, which may uncover to other questions, and eventually predictive models. It also is an important “line of defense” against bad data and is an opportunity to notice that your assumptions or intuitions about a data set are violated.

**What will I learn?**

After completing the project, you will:

* Understand the distribution of a variable and to check for anomalies and outliers
* Learn how to quantify and visualize individual variables within a data set by using appropriate plots such as scatter plots, histograms, bar charts, and box plots
* Explore variables to identify the most important variables and relationships within a data set before building predictive models; calculate correlations, and investigate conditional means
* Learn powerful methods and visualizations for examining relationships among multiple variables, such as reshaping data frames and using aesthetics like color and shape to uncover more information

**Why is this Important to my Career?**

"If you are looking for a career where your services will be in high demand, you should find something where you provide a scarce, complementary service to something that is getting ubiquitous and cheap. So what’s getting ubiquitous and cheap? Data. And what is complementary to data? Analysis"

— Hal Varian, UC Berkeley, Chief Economist at Google

## How do I Complete this Project?

This project is connected to the [Data Analysis With R](https://www.udacity.com/course/ud359) course, but depending on your background knowledge of exploratory data analysis, you may not need to take the whole class to complete this project.

**Introduction**

For the final project, you will conduct your own exploratory data analysis and create an RMD file that explores the variables, structure, patterns, oddities, and underlying relationships of a data set of your choice.

The analysis should be almost like a stream-of-consciousness as you ask questions, create visualizations, and explore your data.

This project is open-ended in that we are not looking for one right answer. As John Tukey stated, "The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data." We want you to ask interesting questions about data and give you a chance to explore. We will provide some options of data sets to explore; however, you may choose to explore an entirely different data set. You should be aware that finding your own data set and cleaning that data set into a form that can be read into R can take considerable time and effort. This can add as much as a day, a week, or even months to your project so only adventure to find and clean a data set if you are truly prepared with programming and data wrangling skills.

Now, on to the details!

## Project Details

**Step One - Choose your Data Set**

First, you will choose a data set from the [**Data Set Options**](https://docs.google.com/document/d/1qEcwltBMlRYZT-l699-71TzInWfk4W9q5rTCSvDVMpc/pub) document. You should choose a data set based on your prior experiences in programming and working with data. The data set you choose will not increase or decrease your chances of passing the final project. In general, [**tidy data sets**](http://vita.had.co.nz/papers/tidy-data.pdf) are easier to work with since each variable is a column and each row is an observation; there’s no data cleaning or wrangling involved. We offer guidance below for choosing your data set. Time estimates include reading all of the project instructions and rubric, conducting the analysis, and submitting the final project.

**Step Two - Get Organized**

Eventually you’ll want to submit your project (and share it with friends, family, and employers). Get organized before you begin. We recommend creating a single folder on your desktop that will eventually contain:

1. The **RMD file** that contains the analysis, final plots and summary, and reflection (in that order)
2. The **HTML file** that will be knitted from your RMD file
3. The **data set** you used (which you will only submit if you found your own data set)

**Step Three - Explore your Data**

This is the fun part. Start exploring your data! Keep track of your thoughts as you go (in an RMD file). Please refer to the [**Example Project**](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/diamondsExample.html) that we have provided. Your report should look similar!

**Step Four - Document your Analysis**

You will want to document your exploration and analysis in an **RMD file** which you will submit. That file should be formatted in markdown and should contain (in order):

1. **A stream-of-consciousness analysis and exploration of the data.**

a. Headings and text should organize your thoughts and reflect your analysis as you explored the data.

b. Plots in this analysis do not need to be polished with labels, units, and titles; these plots are exploratory (quick and dirty).

c. You can iterate on a plot in the same R chunk, but you don’t need to show every plot iteration in your analysis.

1. **A section at the end called “Final Plots and Summary”**

You will select three plots from your analysis to polish and share in this section. The three plots should show different trends and should be polished with appropriate labels, units, and titles (see the [**Project Rubric**](https://docs.google.com/document/d/1L2Wwofs6D8Crd0QLZ1-RxBHlVoBZ3mec2xWgxrmUs5I/pub) for more information).

1. **A final section called “Reflection”**

This should contain a few sentences about your struggles, successes, and ideas for future exploration on the data set (see the [**Project Rubric**](https://docs.google.com/document/d/1L2Wwofs6D8Crd0QLZ1-RxBHlVoBZ3mec2xWgxrmUs5I/pub) for more information).

**Step Five - Knit your RMD file**

Your knitted RMD file should not be one long chunk of R code. It should contain text and plots interspersed throughout. The goal is to give the person reading the file insight into what you were thinking as you explored your data.

**Step Six - Document your Data (if you chose your own data set)**

The data set you submit (only if you chose your own) should include a text file, like those in the R documentation (e.g. ?diamonds) that describes the source of your data and an explanation of the variables in the data set (definition of any variables, units, levels of categorical variables, and the data generating process, such as how data was collected if possible).

## Data Set Options

Choose from one of the following data sets or find your own (see below if you’re finding your own).

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Set** | **Overview** | **Guiding Question** | **Time Estimate** |
| [**Red Wine Quality**](https://www.google.com/url?q=https%3A%2F%2Fs3.amazonaws.com%2Fudacity-hosted-downloads%2Fud651%2FwineQualityReds.csv&sa=D&sntz=1&usg=AFQjCNFbtvciXJWwSbQuwE2Br_pH9JquAg)**[[1]](https://docs.google.com/document/d/1qEcwltBMlRYZT-l699-71TzInWfk4W9q5rTCSvDVMpc/pub" \l "ftnt1)**  Read this [text file](https://www.google.com/url?q=https%3A%2F%2Fs3.amazonaws.com%2Fudacity-hosted-downloads%2Fud651%2FwineQualityInfo.txt&sa=D&sntz=1&usg=AFQjCNHZWanxQ_JGIKpDr2lo9rcDF9jBWQ) which describes the variables and how the data was collected. | This tidy data set contains 1,599 red wines with 11 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent). | Which chemical properties influence the quality of red wines? | 10-20 hours |
| [**White Wine Quality**](https://www.google.com/url?q=https%3A%2F%2Fs3.amazonaws.com%2Fudacity-hosted-downloads%2Fud651%2FwineQualityWhites.csv&sa=D&sntz=1&usg=AFQjCNFUg-OvPk4mpLlkpW94LVA-BGpRlA)**2**  Read this [text file](https://www.google.com/url?q=https%3A%2F%2Fs3.amazonaws.com%2Fudacity-hosted-downloads%2Fud651%2FwineQualityInfo.txt&sa=D&sntz=1&usg=AFQjCNHZWanxQ_JGIKpDr2lo9rcDF9jBWQ) which describes the variables and how the data was collected. | This tidy data set contains 4,898 white wines with 11 variables on quantifying the chemical properties of each wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent). | Which chemical properties influence the quality of white wines? | 10-20 hours |
| [**Financial Contributions to 2012 Presidential Campaigns by State**](http://www.google.com/url?q=http%3A%2F%2Ffec.gov%2Fdisclosurep%2FPDownload.do&sa=D&sntz=1&usg=AFQjCNFdHXxA8Cjy0gPRSgT_SVRgy-Kt_w) | Choose ONE state and explore financial contributions made to Presidential candidates in 2012. | Ask your own questions about this data set. You may want to add variables to this data set such as the gender or political party of the candidate. | 15-30 hours |
| [**Loan Data from Prosper**](https://www.google.com/url?q=https%3A%2F%2Fs3.amazonaws.com%2Fudacity-hosted-downloads%2Fud651%2FprosperLoanData.csv&sa=D&sntz=1&usg=AFQjCNE1BPfVPGV0q0A3xkZxq8VBcfYuoA)  Last updated 03/11/2014  This [variable dictionary](https://docs.google.com/spreadsheet/ccc?key=0AllIqIyvWZdadDd5NTlqZ1pBMHlsUjdrOTZHaVBuSlE&usp=sharing) explains the variables in the data set. | This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, borrower employment status, borrower credit history, and the latest payment information. | Ask your own questions about this data set. There are MANY variables in this data set and you are not expected to explore all of them. You should explore between 10-15 variables in your analysis. | 15-30 hours |
| **Find your own data set!** | **Remember that finding and cleaning your own data set could take significant time and effort!**See the checklist below if you want to choose your own data set. | Ask your own questions about your data set! | 30+ hours |

**If You’re finding your own data…**

Your data set should include:

* at least 1,000 observations
* contain at least one categorical variable (you may create one)
* contain at least 8 different variables
* be in a tidy format1(you may need to clean and reshape the data)
* the data set should be in a commonly used format such as .csv, .tsv, .txt, or .xls

Here are a few resources to find a data set:

* [http://www.inside-r.org/howto/finding-data-internet](http://www.google.com/url?q=http%3A%2F%2Fwww.inside-r.org%2Fhowto%2Ffinding-data-internet&sa=D&sntz=1&usg=AFQjCNGQm8WS79PZHdw0svQ_Qk56vrIVuw) (do not use the Titanic data set)
* [http://opendata.stackexchange.com/](http://www.google.com/url?q=http%3A%2F%2Fopendata.stackexchange.com%2F&sa=D&sntz=1&usg=AFQjCNFOAvr5ItZbXp_LFXHCnsFZsCytow)
* [http://www.data.gov/](http://www.google.com/url?q=http%3A%2F%2Fwww.data.gov%2F&sa=D&sntz=1&usg=AFQjCNFXOs5zqy8V6j4xTutfyzdEkG0-tg)

1 Tidy data sets are data sets that have a particular structure. Read more about tidy data in Hadley Wickham’s paper, [http://vita.had.co.nz/papers/tidy-data.pdf](http://www.google.com/url?q=http%3A%2F%2Fvita.had.co.nz%2Fpapers%2Ftidy-data.pdf&sa=D&sntz=1&usg=AFQjCNHnPBeCTuFyplFdk-OwYzm4qdPy5w)

[[1]](https://docs.google.com/document/d/1qEcwltBMlRYZT-l699-71TzInWfk4W9q5rTCSvDVMpc/pub" \l "ftnt_ref1) P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236. Available at: [@Elsevier][http://dx.doi.org/10.1016/j.dss.2009.05.016](http://www.google.com/url?q=http%3A%2F%2Fdx.doi.org%2F10.1016%2Fj.dss.2009.05.016&sa=D&sntz=1&usg=AFQjCNGt7TP-L6KyebdcUztLe_VK8XcItw) [Pre-press (pdf)] [http://www3.dsi.uminho.pt/pcortez/winequality09.pdf](http://www.google.com/url?q=http%3A%2F%2Fwww3.dsi.uminho.pt%2Fpcortez%2Fwinequality09.pdf&sa=D&sntz=1&usg=AFQjCNFQpxYsdZa2RTIOMIHjse7hiVsS6Q) [bib] [http://www3.dsi.uminho.pt/pcortez/dss09.bib](http://www.google.com/url?q=http%3A%2F%2Fwww3.dsi.uminho.pt%2Fpcortez%2Fdss09.bib&sa=D&sntz=1&usg=AFQjCNEm96TOSgOYtUoohdo43dOTuGMlLw)

## Data Analysis with R Project Rubric

**Overview**

This rubric is here to help you understand the expectations for the analysis that you create. It is the same rubric that the person evaluating your project will use. We will refer to this person as the "project evaluator" in this document. You  should look at the rubric **before you begin working** on your analysis **and before you submit it**.

**How to Use: before you begin**

1. Look at the bold headings under the criteria column to understand what the project evaluator will be looking for in your project.
2. Go through each criteria item in more detail.
3. Familiarize yourself with what is required for your project to "meet specifications" or to be "completely Udacious" (“exceeds specifications”). In order to gain a certificate, you need to "meet specifications", however, to gain the most benefit and learn most from the experience, we encourage you to aim for "completely Udacious".

**How to Use: before you submit**

1. Once your analysis is complete, go through each criteria item and do your best to honestly evaluate where you think your project falls.
2. If you think your project "does not meet specifications" for **any**criteria item, then you should make some changes to your analysis.
3. Once you’re confident that your project "meets specifications" or is "completely Udacious," go ahead and follow the Project Submission Instructions to submit!

**How Grading Works**

1. Your project evaluator will use this rubric to evaluate your analysis.
2. Your grade will simply be "pass" or "doesn’t pass."
3. You pass if your project meets or exceeds specifications in each criteria.
4. If any criteria item "does not meet specifications," you will not pass. You will be able to make changes and re-submit the project.

**The Rubric**

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria** | **Does not meet specifications** | **Meets specifications** | **Exceeds specifications**  (Completely Udacious) |
| **Code Functionality** |  |  |  |
| **Does the code work?** | Some code is not functional, producing an error or preventing the RMD document from being knit. | All code is functional. | ***Not Applicable*** |
| **Does project utilize good coding practices?** | The project sometimes uses repetitive code where a function would be more appropriate. The code uses constants or column numbers to access variables or subsets of data. | The project almost never uses repetitive code where a function would be more appropriate. The code references variables by name instead of using constants or column numbers. | The code is never repetitive and makes use of functions where appropriate and uses sound practices to access variables, subset data, or perform complex operations. |
| **Project Readability** |  |  |  |
| **Is the R code in the student’s RMD file commented in a way that is useful and not superfluous?** | Code is not commented or complex code is not adequately explained with comments. It is not always clear what the code is doing. | All complex code is adequately explained with comments. It is always clear what the code is doing. | All complex code is well explained with comments, and comments are not overused to explain obvious code. It is always clear what the code is is doing and how and why any unusual coding decisions were made. |
| **Does student’s code use formatting techniques (indents, spaces, line breaks, etc…) to improve readability?**  (Refer to [Hadley Wickham's R Style Guide](http://www.google.com/url?q=http%3A%2F%2Fadv-r.had.co.nz%2FStyle.html&sa=D&sntz=1&usg=AFQjCNGmClhIHYKrYWuiv47QKvJYLCt7AA)) | The code does not use formatting techniques or formatting techniques do not improve readability. Some lines are longer than 80 characters. | The code uses formatting techniques to improve code readability. All lines are shorter than 80 characters. | The code uses formatting techniques in a consistent and effective manner to improve code readability. All lines are shorter than 80 characters. |
| **Is Markdown used to improve the presentation of the knitted HTML file?**  (e.g. section headers, text and paragraph spacing, document styles) | There are significant areas in the knitted HTML file where use of Markdown in the RMD would greatly improve readability. | Markdown syntax is used in the code to improve readability of the knitted file. | ***Not Applicable*** |
| **Quality of Analysis** |  |  |  |
| **Is the data set explored in many ways?** | The project does not appropriately use univariate, bivariate, and multivariate plots to explore most of the expected relationships in the data set. | The project appropriately uses univariate, bivariate, and multivariate plots to explore most of the expected relationships in the data set. | The project uses many plot types to explore expected and unexpected relationships in the data. A variety of leading questions, dead-ends, and alternate approaches are presented. |
| **Are questions and observations included as text throughout the analysis?**  (i.e. Every plot or set of related plots is followed by text interpreting the plot(s).) | Questions or findings are missing in multiple places which make it unclear what the student was thinking or what the student found. | Questions and findings are placed between blocks of R code regularly so it is clear what the student was thinking throughout the analysis. | ***Not Applicable*** |
| **Is the flow of the analysis easy to follow?** | Reasons for making each plot or set of plots are not always clear from the text. | Reasoning is provided for the plots made throughout the analysis. Plots made follow a logical flow. | ***Not Applicable*** |
| **Are there a variety of relevant visualizations and statistical summaries?** | The project contains fewer than 20 visualizations. Relevant statistics, such as means, medians, quartiles, or confidence intervals, are not reported to support inferences regarding the data. | The project contains at least 20 visualizations. The visualizations are varied and show multiple comparisons and trends. Relevant statistics are computed throughout the analysis when an inference is made about the data. | The project contains a variety of visualizations that show multiple comparisons and trends. Relevant statistics are calculated throughout the analysis where they would benefit inferences and are included in visualizations. |
| **Is the data visualized using appropriate plots and parameter choices?** | There are visualizations in the project that would benefit interpretability through choice of different plot type, variables plotted, or parameter choice (e.g. bin width, color, axis breaks). | Visualizations made in the project depict the data in an appropriate manner that allows plots to be readily interpreted. | Visualizations made in the project depict the data in a way that demonstrates a deep understanding of the data’s structure. |
| **Final Plots and Summary** |  |  |  |
| **Has a Final Plots and Summary section been included in the project?** | The project does not include a dedicated Final Plots and Summary section containing three plots and commentary, or at least one plot does not follow from what was explored in the main analysis. | The project includes a Final Plots and Summary section containing three plots and commentary. Plots reflect what has been explored in the main body of the analysis. | ***Not Applicable*** |
| **Are the final three plots varied and do they meet some of the following criteria:**   * Draw comparisons. * Identify trends. * Engage a wide audience. * Explain a complicated finding. * Clarify a gap between perception and reality. * Enable the reader to digest large amounts of information. | The plots chosen for the section seem to have been selected arbitrarily, do not fulfill at least 2 ofthe criteria, or are all of the same plot type. | The plots are well chosen and the plots fulfill at least 2 of the criteria. The plots are varied and reveal interesting trends and relationships. | Each plot reveals an important and different comparison or trend in the data. The plots incorporate many of the variables from the data set in a way that allows the plots to convey a lot of information while still being interpreted easily. The plots fulfill 4 or more of the criteria. |
| **Are the plots appropriate?** | One or more plots can be significantly improved by selection of a different variable, different plot type, or different sequence of plots. | All plots have appropriately selected variables and are plotted in a way that accurately conveys the data/information. | ***Not Applicable*** |
| **Are the plots polished?**   * are axes labeled? * are units labeled on each axis? * are plots titled? * are all labels, titles, and units readable? | One or more plots are missing axes labels, plot titles, axes units, or are scaled inappropriately. | All plots are labeled appropriately and can be read and interpreted easily. | ***Not Applicable*** |
| **Are the plots explained?** | The reasoning and findings are not explained for each plot, the text about one plot is not descriptive enough to stand alone, or comments do not reflect the contents of an associated plot. | The reasoning and findings from each plot are explained and the text about each plot is descriptive enough to stand alone. Comments reflect the contents of the plots that they are associated with. | The reasoning and findings from each plot are explained concisely with appropriate variable transformations, other plot decisions, and/or statistics. The text about each figure is descriptive and adds information that the graphic itself would not easily explain. |
| **Reflection** |  |  |  |
| **Has a Reflection section been included in the project?** | The project does not include a dedicated Reflection section to reflect upon the analysis performed. | The project includes a Reflection section discussing the analysis performed. | ***Not Applicable*** |
| **Does the section provide a written reflection of the analysis?** Consider the following in your reflections:   * Where did I run into difficulties in the analysis? * Where did I find successes? * How could the analysis be enriched in future work (e.g. additional data and analyses)? | The section does not communicate struggles, successes, and ideas for improvement. | The section reflects on how the analysis was conducted and reports on the struggles and successes throughout the analysis. The section provides at least one idea or question for future work. | The section provides a rich and well-written reflection of struggles, successes, and lessons learned. The section poses ideas or questions for future work. The section explains any important decisions in the analysis and how those decisions affected the analysis. |

## Project Example - Diamonds

[Separate Document](Diamonds%20Exploration%20by%20Chris%20Saden.html)

## Project Template

[**Project Template File**](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/projectTemplate.Rmd)

Please download the [**project template file**](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/projectTemplate.Rmd) to get started on your analysis.

**Formatting Notes**

We want you to submit a readable RMD file. To help you prepare your project, please look over the following notes.

1. The knitted HTML output should be readable. Be sure to review your knitted HTML file and check that the code and plots appear correct.

1. Comments for R code in a RMD or R-Markdown file are included inside of r blocks by using a hash or pound symbol.

```{r}

library(ggplot2)

# This is an example of a comment that is not actual code.

```

1. In a RMD or R-Markdown file, use of the hash or pound symbol (#) outside of r blocks of code creates an H1 header.

**THIS IS AN H1 HEADER**

*You won't see the hash symbol in front of the text above once you knit the HTML file. See*[*Markdown Syntax*](http://daringfireball.net/projects/markdown/syntax)*for additional help with Markdown formatting.* 

1. Check that all your plots can be viewed and that they are sized appropriately for the output, which is the knitted HTML file.

## Evaluation and Submission

**Note**

You do not need to complete this project if you have received a certificate from completing the Data Analysis with R course previously.

Please email **nd002eda-project@udacity.com** with a copy of your certificate to get the pass phrase to complete this project.

**Evaluation**

Use the [**Project Rubric**](https://docs.google.com/document/d/1L2Wwofs6D8Crd0QLZ1-RxBHlVoBZ3mec2xWgxrmUs5I/pub) to review your project. If you are happy with your submission, then you are ready to submit! If you see room for improvement in **any** category in which you do not meet specifications, keep working!

Your project will be evaluated by a Udacity reviewer according to the same [**Project Rubric**](https://docs.google.com/document/d/1L2Wwofs6D8Crd0QLZ1-RxBHlVoBZ3mec2xWgxrmUs5I/pub). Your project must "meet specifications" or "exceed specifications" in each category in order for your submission to pass.

**Submission**

Send an email to **nd002exploredata-project@udacity.com** with the following:

1. the RMD file containing the analysis (final plots and summary, and reflection)
2. the HTML file knitted from the RMD file using the knitr package
3. the original data set and source if you used your own data rather than one recommended by Udacity (Note: do not submit a data set if you used one that Udacity recommended)
4. A list of Web sites, books, forums, blog posts, github repositories, etc. that you referred to or used in creating your submission (add N/A if you did not use any such resources).

Please carefully read the following statement and include it in your email:

*“I hereby confirm that this submission is my work. I have cited above the origins of any parts of the submission that were taken from Websites, books, forums, blog posts, github repositories, etc. By including this in my email, I understand that I will be expected to explain my work in a video call with a Udacity coach before I can receive my verified certificate.”*

A Udacity reviewer will respond shortly with next steps.

## Example Project

Diamonds Again

[**Example Project**](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/diamondsExample.html)

Your final project will be an analysis in which you analyze the variables and relationships within a data set. Your **final project** should look similar to this [**Example Project**](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/diamondsExample.html). It should follow the same structure with an **Analysis** section, a **Final Plots** section, and a **Reflection** section. We will provide a template for you to use with these sections already included.

Take at least 10 minutes to review the example project to get a sense of what you will need to do before starting your project.

**Common Problems with Project Submissions**

To help you succeed, we recommend comparing your project submission against the following list of common problems. Your project must **avoid** these common problems in order to pass.

* Data processing or transformations (creating a categorical variable) should be included in the RMD file and the final knitted HTML output.
* Reflection section is not included.
* Reflection section is not the last section in the RMD file.
* Final Plots section is not included.
* Final Plots section is not at the end of the RMD file before the Reflection section.
* Final Plots section does not contain three plots.
* One or more plots in the Final Plots section do not reveal a finding or pattern in the data set.
* Final Plots are not polished and are missing titles or units.
* Inappropriate plots are chosen for data in the Final Plots section.

[**Creating Effective Plots**](https://docs.google.com/document/d/1-f3wM3mJSkoWxDmPjsyRnWvNgM57YUPloucOIl07l4c/pub)

The **Final Plots** section in your RMD file should contain three polished plots that give insight into the data set that you investigated.

**Each plot should also contain a caption or description about what the plot shows.**

In determining whether or not you have three strong plots, please consult the document, [**Creating Effective Plots**](https://docs.google.com/document/d/1-f3wM3mJSkoWxDmPjsyRnWvNgM57YUPloucOIl07l4c/pub). The document covers four common problems that project evaluators have encountered in the past and how those plots can be improved.

## Creating Effective Plots

[Titles, Labels, and Legends](https://docs.google.com/document/d/1-f3wM3mJSkoWxDmPjsyRnWvNgM57YUPloucOIl07l4c/pub#h.6eh6p4aoyo1)

[Variable Types Determine Plot Types](https://docs.google.com/document/d/1-f3wM3mJSkoWxDmPjsyRnWvNgM57YUPloucOIl07l4c/pub#h.4cjvtuxjsa56)

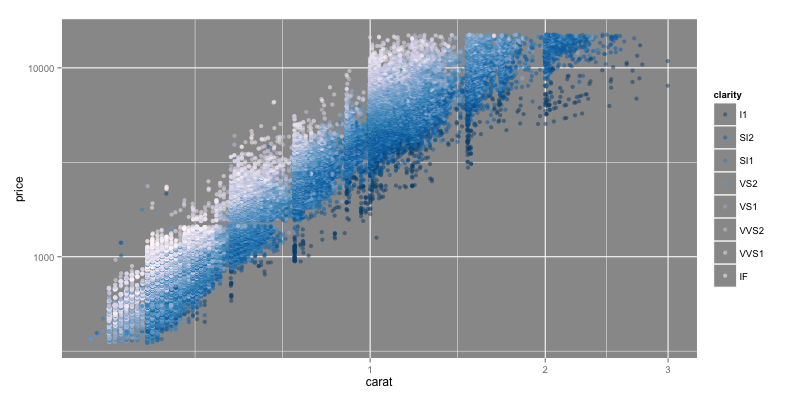
[Beware of Information Overload](https://docs.google.com/document/d/1-f3wM3mJSkoWxDmPjsyRnWvNgM57YUPloucOIl07l4c/pub#h.9eqmtuy7fgh2)

[Use of Color as a Visual Encoding](https://docs.google.com/document/d/1-f3wM3mJSkoWxDmPjsyRnWvNgM57YUPloucOIl07l4c/pub#h.ud4utugd1kn2)

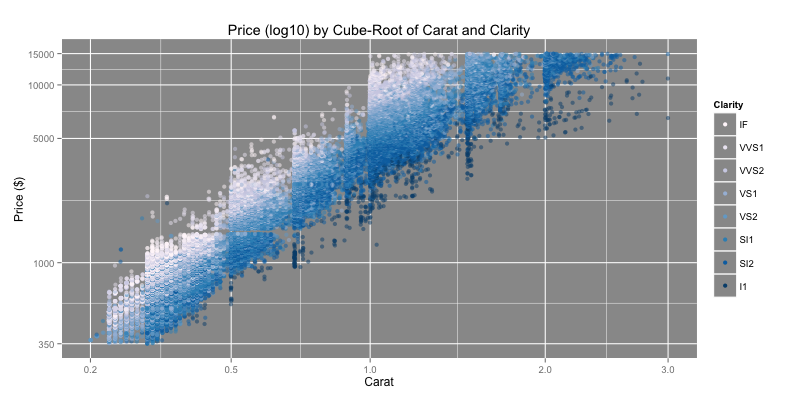
Creating plots that are efficient and effective is important for two major reasons. Firstly, when demonstrating your results to others, an effective plot will convey the results you want to present and provide an easy-to-follow story to the observer. Secondly, when you are performing the initial exploration, an appropriate plot will allow you to understand patterns or trends in the data, saving you time in following the trails that bring you to the most interesting results. This document is meant to provide some tips in the creation of effective plots to further both goals.

**Titles, Labels, and Legends**

It will help others understand your plots when they are properly annotated. Have an informative, but succinct title on your plots. Similarly, for plots meant to be read by others, have descriptive axis labels and legends when you have third variables to plot. Make sure that you have a sufficient number of axis lines for the reader to understand the scale of the data being plotted.



**Figure 1-1:**This plot is okay, but we can make it better by updating the labels and giving it a title.

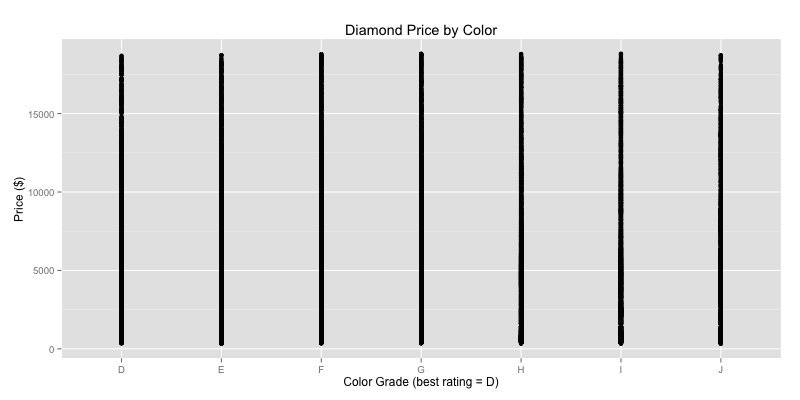


**Figure 1-2:**The added title and axis breaks make it clear that a transform has been applied to both axes. The legend has also been reversed so that better clarity grades are higher on the scale.

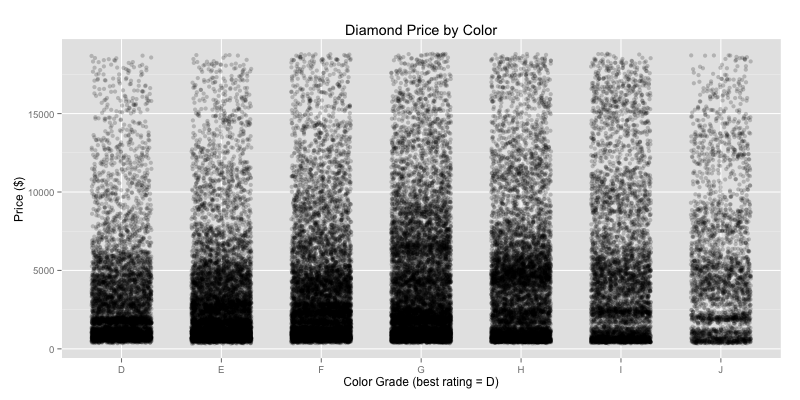
**Variable Types Determine Plot Types**

It is important to consider the values that a variable takes when electing how it should be plotted. There are a number of different variable types to consider, including nominal (categorical variable whose factor levels have no inherent ordering), ordinal (categorical variable whose factor levels have a natural inherent order), discrete quantitative (variable that takes numerical values with a fixed number of values), and continuous quantitative (variable that takes numerical values to arbitrary precision). Plots and visual elements that work well for quantitative variables may need to be modified or changed for if categorical variables are used instead.

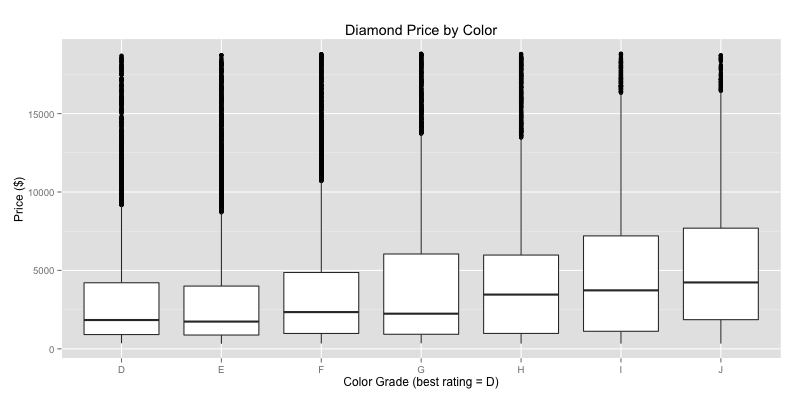
Example: When you are plotting two variables against one another, a straight scatterplot is not necessarily the best option, even if both variables are quantitative in nature. If one of the variables takes only discrete values, then it may be appropriate to introduce jitter in point positions along that axis to spread the data. The gives a better idea of the variable’s density. Alternatively, we may want to use a different plot instead of a scatterplot. If one of the variables is non-numeric, then a scatterplot doesn’t make much sense. Instead, a box plot, violin plot, density plot, or bar graph will be a more effective data representation.



**Figure 2-1:** This plot is not effective, since the density of points on categories make them impossible to distinguish.



**Figure 2-2:** We make a scatterplot more effective by adding horizontal jitter and using transparency to reduce overplotting.



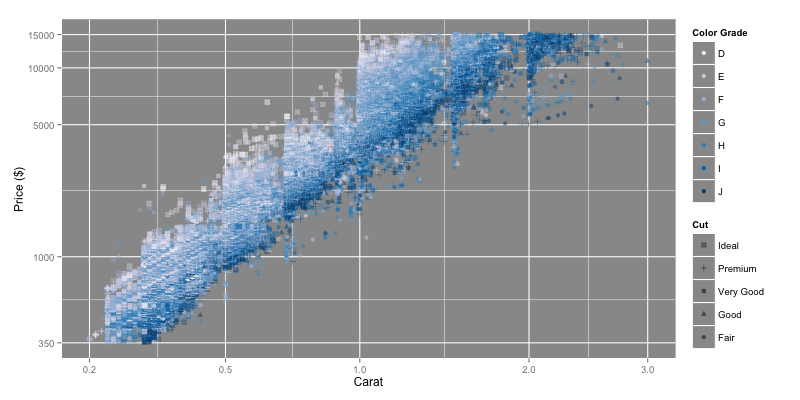
**Figure 2-3:** Perhaps even better, a box plot summarizes the distribution on each category in an easy-to-understand way. It is clearest in this plot that something unintuitive is going on with the price across color grades.

**Beware of Information Overload**

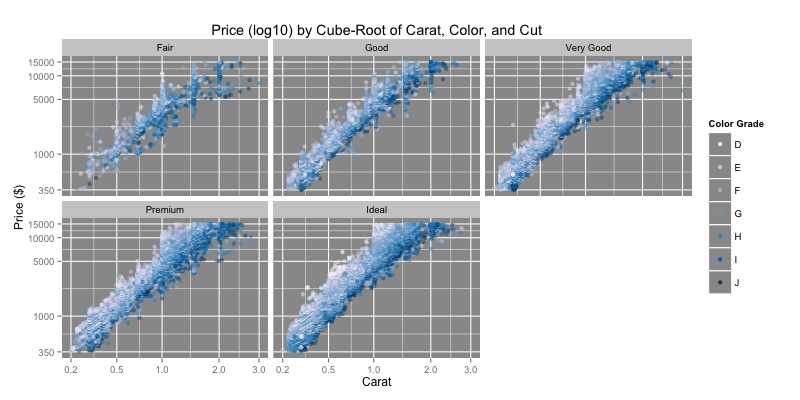
Simply because you can add additional aesthetic features to a plot does not mean that you should. A careful balance in terms of visual elements should be used to keep your plots interpretable.

Think very carefully before you let points in your plot represent more than three variables. Each point in a plot will typically contain information about two variables from its x- and y- position with additional variables contained in other visual encodings (color, shape, size, etc.). If a single point contains multiple non-positional encodings, the plot may contain too much information that patterns and trends are not easy to spot. It is better to make multiple simpler plots than it is to try and fit everything into a single plot.

Be careful about redundant information as well. For example, if a variable is being plotted on one axis, think if you need to show that information again by adding a visual encoding such as color for emphasis. Having a legend match color to axis labels or position is an example of redundancy. While “double encoding” a variable can provide emphasis to demonstrate a variable’s importance, there is potential for information overload or confusion. In some cases, simpler graphics are be better. If you don’t need a legend, then don’t use one. Readers can focus on the main data display and any relevant labels rather than having the cognitive overhead of processing the legend and having their eyes jump from the legend to the data.



**Figure 3-1:** Considering the number of data points being plotted, having even two simultaneous non-positional encodings makes things difficult to interpret.



**Figure 3-2:**We can plot all four variables by faceting across cut. Here, we can observe that the color grade trend is similar across Cut factor levels and that there are a lot more diamonds of the higher cut grades.

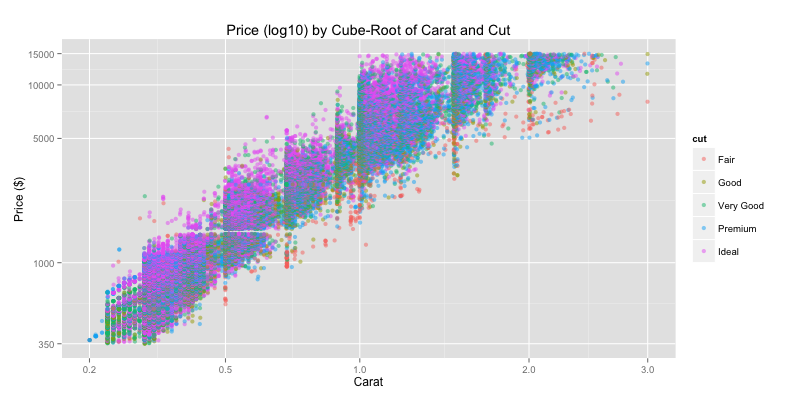
**Use of Color as a Visual Encoding**

Color is a very common visual encoding for third variables in plots. R has multiple color options, especially when using ColorBrewer. Be sure to use a color set that is appropriate for the variable type being encoded by the color scheme. There are three types of color schemes available: qualitative (use the type = “qual” option with ColorBrewer-related functions), sequential (type = “seq”), and diverging (type = “div”).

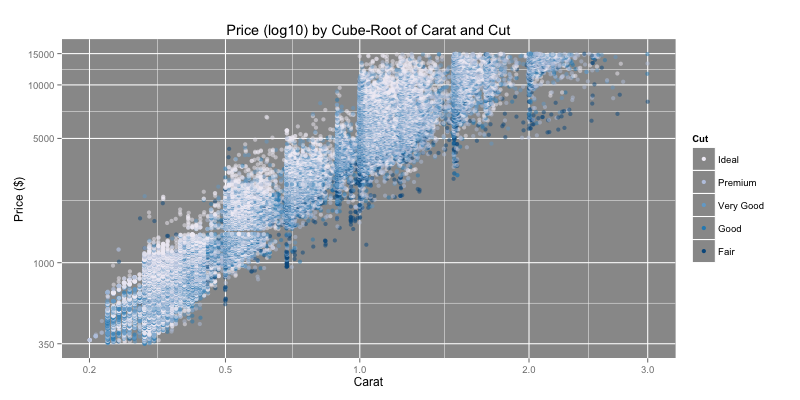
Use a qualitative color scheme when you are plotting a nominal variable. Subsequent levels in a qualitative color palette tend to vary greatly in hues to emphasize that levels are not related to one another. When we make a plot using a qualitative palette, we don’t expect to see a color trend across factor levels, but we may see that certain levels cluster in interesting ways that help us identify interesting patterns.

For variables that follow an ordering, either ordinal or quantitative, we will prefer a sequential or diverging color scheme. Each is used in different ways. Most of the time, a sequential color palette is most appropriate: these schemes vary saturation from light to dark and sometimes along a range of hues to demonstrate values increasing along a scale. Lighter colors are used to indicate small values, while darker colors indicate large values. Alternatively, the opposite ordering may be applied for the scale, in which lighter colors indicate “better” values and darker colors “worse” values.

On the other hand, some variables have a critical center point or zero where values above the center and values below the center are important features to identify. For these, we use a diverging palette, where the central value is lightest, and values farthest from the center are darkest. Values larger than the center take one hue, while values smaller than the center tend towards a second hue.



**Figure 4-1:** The default coloring separates category by hue along the full rainbow spectrum. While you can see some trend across Cut, the plot does not fully exploit the ordinal relationship between factor levels.



**Figure 4-2:** Changing the color scale to a sequential palette is arguably better, since it more tightly exploits the ordinal relationship in factor levels of the Cut variable.

# Projeto 4 – Identifying Fraud from Enron Mail

## Project Overview

In this project, you will play detective, and put your machine learning skills to use by building an algorithm to identify Enron Employees who may have committed fraud based on the public Enron financial and email dataset.

**Note**

You do not need to complete this project if you have received a certificate from completing Udacity's Intro to Machine Learning course previously.

Please email **nd002introml-project@udacity.com** with your certificate to get the completion phrase to complete this project.

**Why this Project?**

This project will teach you the end-to-end process of investigating data through a machine learning lens.

It will teach you how to extract/identify useful features that best represents your data, a few of the most commonly used machine learning algorithms today, and how to evaluate the performance of your machine learning algorithms.

**What will I learn?**

By the end of the project, you will be able to:

* Deal with an imperfect, real-world dataset
* Validate a machine learning result using test data
* Evaluate a machine learning result using quantitative metrics
* Create, select and transform features compare the performance of machine learning algorithms
* Tune machine learning algorithms for maximum performance
* Communicate your machine learning algorithm results clearly

**Why is this Important to my Career?**

Machine Learning is a first-class ticket to the most exciting careers in data analysis today.

As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions.

Machine learning brings together computer science and statistics to harness that predictive power.

## Project Details

**How do I Complete this Project?**

This project is connected to the [Intro to Machine Learning course](https://www.udacity.com/course/ud120), but depending on your background knowledge of machine learning, you may not need to take the whole thing to complete this project.

A note before you begin: the projects in the Intro to Machine Learning class were mostly designed to have lots of data points, give intuitive results, and otherwise behave nicely. This project is significantly tougher in that we're now using the real data, which can be messy and doesn't have as many data points as we usually hope for when doing machine learning. Don't get discouraged--imperfect data is something you need to be used to as a data analyst! If you encounter something you haven't seen before, take a step back and think about a smart way around. You can do it!

**Project Overview**

In 2000, Enron was one of the largest companies in the United States. By 2002, it had collapsed into bankruptcy due to widespread corporate fraud. In the resulting Federal investigation, there was a significant amount of typically confidential information entered into public record, including tens of thousands of emails and detailed financial data for top executives. In this project, you will play detective, and put your new skills to use by building a person of interest identifier based on financial and email data made public as a result of the Enron scandal. To assist you in your detective work, we've combined this data with a hand-generated list of persons of interest in the fraud case, which means individuals who were indicted, reached a settlement, or plea deal with the government, or testified in exchange for prosecution immunity.

**Resources Needed**

You should have python and sklearn running on your computer, as well as the starter code (both python scripts and the Enron dataset) that you downloaded as part of the first mini-project in the Intro to Machine Learning course. You can get the starter code on git: *git clone*[*https://github.com/udacity/ud120-projects.git*](https://github.com/udacity/ud120-projects.git)

The starter code can be found in the final\_project directory of the codebase that you downloaded for use with the mini-projects. Some relevant files:   
  
poi\_id.py : starter code for the POI identifier, you will write your analysis here   
  
final\_project\_dataset.pkl : the dataset for the project, more details below   
  
tester.py : when you turn in your analysis for evaluation by your Udacity coach, you will submit the algorithm, dataset and list of features that you use (these are created automatically in poi\_id.py). That coach will then use this code to test your result, to make sure we see performance that’s similar to what you report. You don’t need to do anything with this code, but we provide it for transparency and for your reference.   
  
emails\_by\_address : this directory contains many text files, each of which contains all the messages to or from a particular email address. It is for your reference, if you want to create more advanced features based on the details of the emails dataset.

**Steps to Success**

We will provide you with starter code, that reads in the data, takes your features of choice, then puts them into a numpy array, which is the input form that most sklearn functions assume. Your job is to engineer the features, pick and tune an algorithm, test, and evaluate your identifier. Several of the mini-projects were designed with this final project in mind, so be on the lookout for ways to use the work you’ve already done.

The features in the data fall into three major types, namely financial features, email features and POI labels. financial features: ['salary', 'deferral\_payments', 'total\_payments', 'loan\_advances', 'bonus', 'restricted\_stock\_deferred', 'deferred\_income', 'total\_stock\_value', 'expenses', 'exercised\_stock\_options', 'other', 'long\_term\_incentive', 'restricted\_stock', 'director\_fees'] (all units are in US dollars) email features: ['to\_messages', 'email\_address', 'from\_poi\_to\_this\_person', 'from\_messages', 'from\_this\_person\_to\_poi', 'poi', 'shared\_receipt\_with\_poi'] (units are generally number of emails messages; notable exception is ‘email\_address’, which is a text string) POI label: [‘poi’] (boolean, represented as integer)

You are encouraged to make, transform or rescale new features from the starter features. If you do this, you should store the new feature to my\_dataset, and if you use the new feature in the final algorithm, you should also add the feature name to my\_feature\_list, so your coach can access it during testing. For a concrete example of a new feature that you could add to the dataset, refer to the lesson on Feature Selection.

## Evaluation and Submission

**Note**

You do not need to complete this project if you have received a certificate from completing the Intro to Machine Learning course previously.

Please email **nd002introml-project@udacity.com** with your certificate to get the pass phrase to complete this project.

**General Submission and Evaluation Overview**

Your submission will have 2 parts: the code/classifier you create and some written documentation of your work. We will evaluate your project according to the rubric [here,](https://docs.google.com/a/knowlabs.com/document/d/17-JwNQH1aRxtqMkJ6zpCL_68kh5F6uSbDXcJS26vZWY/edit?usp=sharing) only projects that satisfy **all**"meets expectations" items will pass. **Please self-evaluate before you submit!** If you don't think your project meets all the criteria, your coach likely won't either.

Email your submission to **nd002introml-project@udacity.com**

**Code/Classifier**

When making your classifier, you will create three pickle files (my\_dataset.pkl, my\_classifier.pkl, my\_feature\_list.pkl) and turn those in along with your code to your Udacity coach when you submit your writeup; your coach will test them using the tester.py script. You are encouraged to use this script before submitting to gauge if your performance is good enough.

**Documentation of Your Work**

Document the work you've done by answering (in about a paragraph each) the questions found [here.](https://docs.google.com/document/d/1NDgi1PrNJP7WTbfSUuRUnz8yzs5nGVTSzpO7oeNTEWA/edit?usp=sharing)You can write your answers in a word document, text file, or similar format and submit it to your coach when all your answers are complete.

Your coach will review your answers, and ask a few more followup questions that will probe specific aspects of your project in more detail.

**Good Luck!**

## Questions for Scaled Project

**Enron Submission Free-Response Questions**

A critical part of machine learning is making sense of your analysis process, and communicating it to others.  The questions below will help us understand your decision-making process and allow us to give feedback on your project.  Please answer each question; your answers should be 1-2 paragraphs per question.  If you find yourself writing much more than that, take a step back and see if you can simplify your response!

When your coach evaluates your responses, he or she will use a specific list of rubric items to assess your answers.  Here is the link to that rubric: [Link to the rubric](https://docs.google.com/a/knowlabs.com/document/d/17-JwNQH1aRxtqMkJ6zpCL_68kh5F6uSbDXcJS26vZWY/edit)

Each question has one or more specific rubric items associated with it, so before you submit an answer, take a look at that the rubric.  If your response does not meet expectations, you will be asked to resubmit.

Once you’ve submitted your responses, your coach will take a look and ask a few more focused follow-up questions on one or more of your answers.

We can’t wait to see what you’ve put together for this project!

1. Summarize for us the goal of this project and how machine learning is useful in trying to accomplish it.  As part of your answer, give some background on the dataset and how it can be used to answer the project question.  Were there any outliers in the data when you got it, and how did you handle those?  [relevant rubric items: “data exploration”, “outlier investigation”]
2. What features did you end up using in your POI identifier, and what selection process did you use to pick them?  Did you have to do any scaling?  Why or why not?  As part of the assignment, you should attempt to engineer your own feature that doesn’t come ready-made in the dataset--explain what feature you tried to make, and the rationale behind it.  (You do not necessarily have to use it in the final analysis, only engineer and test it.)  If you used an algorithm like a decision tree, please also give the feature importances of the features that you use.  [relevant rubric items: “create new features”, “properly scale features”, “intelligently select feature”]
3. What algorithm did you end up using?  What other one(s) did you try? [relevant rubric item: “pick an algorithm”]
4. What does it mean to tune the parameters of an algorithm, and what can happen if you don’t do this well?  How did you tune the parameters of your particular algorithm?  (Some algorithms don’t have parameters that you need to tune--if this is the case for the one you picked, identify and briefly explain how you would have done it if you used, say, a decision tree classifier). [relevant rubric item: “tune the algorithm”]
5. What is validation, and what’s a classic mistake you can make if you do it wrong?  How did you validate your analysis?  [relevant rubric item: “validation strategy”]
6. Give at least 2 evaluation metrics, and your average performance for each of them.  Explain an interpretation of your metrics that says something human-understandable about your algorithm’s performance. [relevant rubric item: “usage of evaluation metrics”]

## Intro to Machine Learning Project Rubric

Intro to Machine Learning Final Project Rubric

**Overview**

This rubric is here to help you understand the expectations for how your project will be evaluated. It is the same rubric that the person evaluating your project will use. You  should look at the rubric **before you begin working** on this project **and before you submit it**.

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Does Not Meet Expectations** | **Meets expectations** |
| **Quality of Code** |  |  |
| **Functionality** | Code does not perform the functions documented in the writeup, or the writeup does not clearly specify the final analysis strategy. | Code reflects the description in the documentation. |
| **Usability** | Dataset, list of features used, or algorithm not exported properly, so that tester.py does not execute. | Dataset, list of features and algorithm are exported using code in poi\_id.py, so that it can be checked easily using tester.py |
| **Understanding the Dataset and Question** |  |  |
| **Data Exploration (related mini-project: Lesson 5)** | Student response does not address characteristics of the dataset, or ignores/gets wrong important characteristics. | Student response addresses the most important characteristics of the dataset and uses these characteristics to inform analysis. Important characteristics include:   * total no. of data points * allocation across classes (POI/non-POI) * no. of features used |
| **Outlier Investigation (related mini-project: Lesson 7)** | Outliers in financial data are not identified, and/or not removed if removal would be appropriate. | Student response identifies outlier(s) in the financial data, and explains how they are removed or otherwise handled. |
| **Optimize Feature Selection/Engineering** |  |  |
| **Create new features (related mini-project: Lesson 11)** | The only features investigated as possible algorithm inputs are the original features in the dataset provided. | At least one new feature implemented. Justification for that feature is provided in the written response, and the effect of that feature on the final algorithm performance is tested. |
| **Intelligently select features (related mini-project (related mini-project: Lesson 11)** | Feature selection is not performed or documented. | Univariate or recursive feature selection is deployed, or features are selected by hand (different combinations of features are attempted, and the performance is documented for each one).  For an algorithm that supports getting the feature importances (e.g. decision tree), those are documented as well. |
| **Properly scale features (related mini-project: Lesson 9)** | Feature scaling is not deployed when called for. | If algorithm calls for scaled features, feature scaling is deployed. |
| **Pick and Tune an Algorithm** |  |  |
| **Pick an algorithm  (related mini-project: Lessons 1-3)** | Only one algorithm attempted. | At least 2 different algorithms attempted, and their performance is compared with the more performant one used in the final analysis. |
| **Tune the algorithm (related mini-project: Lessons 2, 3, 13)** | No parameter tuning attempted for an algorithm that has tunable parameters. | At least one important parameter tuned, with at least 3 settings investigated; or any of the following are true:   * GridCV used for parameter tuning * Several parameters tuned * Parameter tuning incorporated into algorithm selection (i.e. parameters tuned for more than one algorithm, and best algorithm-tune combination selected for final analysis) |
| **Validate and Evaluate** |  |  |
| **Usage of Evaluation Metrics (related mini-project: Lesson 14)** | Less than 2 quantitative performance metrics deployed, or student cannot articulate what those metrics measure. | Precision and recall are used to evaluate performance performance, and student articulates what those metrics measure. |
| **Validation Strategy (related mini-project: Lesson 13)** | There is no data held out for testing. | The data is split into training and testing sets, with the testing data used for assessing overall analysis performance; or k-fold cross validation is deployed |
| **Analysis Performance** | Precision and recall both consistently below 0.3 | Precision and recall above 0.3 |

**Before you Submit**

1. After you’ve completed all the auto-graded questions and answered the written questions, go through each rubric item and do your best to honestly evaluate where you think your project falls.
2. If you think your project "does not meet expectations" for **any** criteria item, you should make any necessary changes.
3. Once you’re confident that your project "meets expectations" or "exceeds expectations," you can submit by emailing your written responses to introml-project@udacity.com

**How Grading Works**

1. Your project evaluator will be able to see all your code submissions. They will use this rubric to evaluate your code as well as your written responses.
2. Your grade will simply be "pass, meets expectations," “pass, exceeds expectations,” or "doesn’t pass,"
   1. You earn “pass, meets expectations” if **all** criteria “meet expectations.”
   2. You earn “pass, exceeds expectations,” if all criteria “exceed expectations” (when possible).
   3. Your project “doesn’t pass” if **any** criteria are graded as “doesn’t meet expectations.” In this case, you will have the opportunity to revise and resubmit.

# Projeto 5 – Make Effective Data Visualization

**Project Overview**

For this project, you will create a data visualization from a data set that tells a story or allows a reader to explore trends or patterns. You will need to use either [dimple.js](http://dimplejs.org/) or [d3.js](http://d3js.org/) to create the visualization. Your work should be a reflection of the theory and practice of data visualization, such as visual encodings, design principles, and effective communication.

**Note**

You do not need to complete this project if you have received a certificate from completing the Data Visualization course previously.

Please email **nd002datavis-project@udacity.com** with your certificate to get the pass phrase to complete this project.

**What do I need to install?**

To work on your data visualization, you will need to start a local server on your computer. To start a local web server, you will need to have [Python 2.7.8 or higher](https://www.python.org/downloads/) installed on your machine.

If you do not have Python installed on your machine, please watch the instructions for [Downloading Python](https://www.udacity.com/course/viewer#!/c-ud036/l-990110642/m-1007918581). These instructions come from the [Programming Foundations with Python](https://www.udacity.com/course/ud036) course.

Once you have Python installed, you can start a local web server and view your data visualization. Refer to the following [video](https://www.udacity.com/course/viewer#!/c-ud507/l-3168988586/m-3063989000) to see how to do so.

Remember, you must start your web server in the top level directory to serve all code and data files. If you do not use this folder as the root directory for the web server, be aware that you will need to change the file paths.

There are other ways to start a local web server. To learn more about why you need to start a local web server and other ways of setting up a local web server, please read [Setting Up A Local Web Server](http://chimera.labs.oreilly.com/books/1230000000345/ch04.html#_setting_up_a_web_server)from Scott Murray's book, Interactive Data Visualization for the Web.

**Why this Project?**

This project will touch on the overarching attitudes and beliefs important to effective data visualization, such as:

* visualization is a dialog
* showcasing and sharing visualization with others
* visualization is a fluid process that typically requires multiple iterations of improvements

You will have an opportunity to experience the end-to-end process of creating effective data visualizations and highlighting important information from data that may otherwise be hidden or hard to uncover.

**What will I learn?**

After completing the project, you will be able to:

* Demonstrate the ability to choose optimal visual elements to encode data and critically assess the effectiveness of the visualization
* Communicate a story or finding to the appropriate audience using interactive visualizations
* Undergo the iterative process of creating a visualization, and build interactive visualizations with dimple.js or d3.js.

**Why is this Important to my Career?**

Data analyst are storytellers that can translate data findings that other people can easily understand. They view data visualization as an important form of communication.

If you, as a data analyst, can create visualizations to explore data, articulate clear findings to drive business decisions, or use data to elicit consensus from diverse perspectives, then you will be a deeply invaluable member on your team.

## Project Details

**How do I Complete this Project?**

This project is connected to the [**Data Visualization**](https://www.udacity.com/course/ud507) course, but depending on your background knowledge of data visualization, [**dimple.js**](http://dimplejs.org/), and [**d3.js**](http://d3js.org/) you may not need to take the whole course to complete this project.

After completing Lesson 2 and Problem Set 2 of the course, you will be able to complete this project since you will have learned about dimple.js.

If you want to become more technical and expand your skill set, you can continue to Lesson 3 and Lesson 4, in which you will learn more about narrative structures and how to create graphics using d3.js. The d3.js library has a steeper learning curve, and we encourage you to take on the challenge if you desire.

The process for evaluating your project is not affected by your choice of using [**dimple.js**](http://dimplejs.org/) or [**d3.js**](http://d3js.org/).

**Note**

All of the instructions below can be found in the [**Project Description**](https://docs.google.com/document/d/1uOT-uAJSySO5jWA_1xvUGMMzn1cgLapX9lH4iWGKnvc/pub) document.

**Introduction**

For the final project, you will create a data visualization from a data set that tells a story or allows a reader to explore trends or patterns. Your work should be a reflection of the theory and practice of data visualization, and you must use either [**dimple.js**](http://dimplejs.org/) or [**d3.js**](http://d3js.org/).

We will provide some options of data sets to explore; however, you may choose to explore an entirely different data set. You should be aware that finding your own data set and cleaning it using Python, R, or some other language can take considerable time and effort. This can add as much as a day, a week, or even months to your project so embark on the adventure to find and clean a data set if you are truly prepared with programming and data wrangling skills.

There are **three difficulty** levels to this project, and you should choose an appropriate level depending on your experience with data munging and exploratory data analysis. The difficulty level you choose will not affect the evaluation of the project.

* **Beginner** - *Choose this option if you have no experience with cleaning data or exploratory data analysis.*   
    
  Select one of the beginner data sets, which already has a summary of findings, from the [**Data Set Options**](https://docs.google.com/document/d/1w7KhqotVi5eoKE3I_AZHbsxdr-NmcWsLTIiZrpxWx4w/pub) document. Then, create a visualization that communicates the findings.
* **Intermediate** - *Choose this option if you have some experience cleaning and analyzing data.*   
    
  Select one of the intermediate data sets from the [**Data Set Options**](https://docs.google.com/document/d/1w7KhqotVi5eoKE3I_AZHbsxdr-NmcWsLTIiZrpxWx4w/pub) document. These data sets are not necessarily clean or tidy data sets. You will investigate the data set to share a story or message about the data and then create a suitable visualization.
* **Advanced** - *Choose this option if you are comfortable finding, cleaning, and analyzing a data set.*  
    
  Find a data set, investigate it, and share your findings in a visualization. Your final graphic should primarily be explanatory, but it may also contain exploratory components. You can find a list of recommended websites to find data sets in the [**Data Set Options**](https://docs.google.com/document/d/1w7KhqotVi5eoKE3I_AZHbsxdr-NmcWsLTIiZrpxWx4w/pub) document. You should be aware that finding your own data set, cleaning the data set, and analyzing it (using R, iPython Notebook, or another tool) can take considerable time and effort. This can lengthen the time you spend on your project by days, weeks, or even months. Choose the option only if you feel prepared for a challenge!

Now, on to the details!

**Step One - Choose a Data Set**

First, you will choose a data set from the [**Data Set Options**](https://docs.google.com/document/d/1w7KhqotVi5eoKE3I_AZHbsxdr-NmcWsLTIiZrpxWx4w/pub) document or find a data set to explore and visualize. You should choose a data set based on your prior experiences in programming and working with data. The data set you choose will not increase or decrease your chances of passing this project.

**Step Two - Get Organized**

Eventually you’ll want to submit your project and share it. If you are familiar with [**GitHub**](https://github.com/), we encourage you to create a public repository or a public [**Gist**](https://gist.github.com/) for your project to track changes. Otherwise, you need to create the following files.

* an **index.html** file containing the code to create your visualization (you may include the JavaScript and CSS in this file or separate them in other files)
* a **README.md** file that includes four sections...
  + **Summary** - in no more than 4 sentences, briefly introduce your data visualization and add any context that can help readers understand it
  + **Design** - explain any design choices you made including changes to the visualization after collecting feedback
  + **Feedback** - include all feedback you received from others on your visualization from the first sketch to the final visualization
  + **Resources** - list any sources you consulted to create your visualization
* **data files**
  + the final data set used to create the visualization (usually .csv, .tsv, or .json file)
  + a codebook or other files related to the data set (description, readme, license)
* OPTIONAL FOLDERS IF YOU USE [**GITHUB**](https://github.com/)
  + **data** folder to include all the data related files
  + **js** folder to include .js files (not needed if javascript is in the index.html file)
  + **css** folder to include .css files (not needed if CSS is in the index.html file)

**Step Three - Find a Data Story**

Explore your data set and craft a message or story around your data! Think about the overall message you want to convey and think about the comparison(s) or relationship(s) you want your readers to see.

**Step Four - Create Your Visualization**

First, sketch ideas for your visualization. Once you settle on a sketch, explain any design choices in that sketch, such as chart type, visual encodings, and layout, in the **Design** section of the **README.md**file. Then, write code to create your visualization using either [**dimple.js**](http://dimplejs.org/) or [**d3.js**](http://d3js.org/). The visualization must include animation, interaction, or both. See the [**Project Rubric**](https://docs.google.com/document/d/1zRVs73M7P5ACKB0n3Di4k0AskId3pc6lIpMBmmydETk/pub) for more information.

**Step Five - Get Feedback**

Share your visualization with **at least 3 other people** and document their feedback. There are many ways to get feedback, and more feedback is generally better! Here are some options. - Share your visualization with others in person and have them think aloud as they read and explore the graphic so you can document what stands out to them and how they interpret the graphic. - Share a link to your repository in the discussions and ask others to share constructive criticisms. Be sure to offer advice to others who are seeking feedback too!

* Create and share a [**Gist**](https://gist.github.com/), which contains an **index.html** file, data file, and any .js or .css files). Directions for creating and sharing a Gist can be found at [**http://bl.ocks.org/**](http://bl.ocks.org/).
  + **Box Plots Gist EXAMPLE**:
    - <https://gist.github.com/mbostock/4061502>
    - <http://bl.ocks.org/mbostock/4061502>

You might need to ask specific questions to prompt the reader. Here are some questions to help you. You can, of course, ask others.

* What do you notice in the visualization?
* What questions do you have about the data?
* What relationships do you notice?
* What do you think is the main takeaway from this visualization?
* Is there something you don’t understand in the graphic?

**Step Six - Document Feedback and Improve the Visualization**

For each person that gives you feedback, add the person’s feedback to your **README.md** file in the**Feedback** section. As you improve and iterate on your visualization, update your code **AND** describe any changes in the **Design** section of the **README.md** file. You should be building evidence that you have shared your visualization, received feedback, and responded to that feedback. You will need to submit the different versions of your visualization. At the least you need to submit the initial version of the index.html file and the final index.html file.

**Step Seven - Review**

Use the [**Project Rubric**](https://docs.google.com/document/d/1zRVs73M7P5ACKB0n3Di4k0AskId3pc6lIpMBmmydETk/pub) to review your project. If you are happy with your submission, then you’re ready to submit your project. If you see room for improvement, keep working to improve your project!

**Step Seven - Submit**

Follow the instructions in the [**Project Submission Instructions**](https://docs.google.com/document/d/13nj9Q0UhsP7dcm3zK9K6Lsr6egeM31NRz55rhldQJW0/pub) document.

## Project Description for Data Visualization

**Introduction**

For the final project, you will create a data visualization from a data set that tells a story or allows a reader to explore trends or patterns. Your work should be a reflection of the theory and practice of data visualization, and you must use either [dimple.js](http://www.google.com/url?q=http%3A%2F%2Fdimplejs.org&sa=D&sntz=1&usg=AFQjCNH6kBG22shhDc8n1rCSDecWuEjxaQ) or [d3.js](http://www.google.com/url?q=http%3A%2F%2Fd3js.org&sa=D&sntz=1&usg=AFQjCNEOj7G19Lrx2IKZoyjb9FSuVi1jig).

We will provide some options of data sets to explore; however, you may choose to explore an entirely different data set. You should be aware that finding your own data set and cleaning it using Python, R, or some other language can take considerable time and effort. This can add as much as a day, a week, or even months to your project so embark on the adventure to find and clean a data set if you are truly prepared with programming and data wrangling skills.

There are**three difficulty levels** to this project, and you should choose an appropriate level depending on your experience with data munging and exploratory data analysis. The difficulty level you choose will not affect the evaluation of the project.

**Beginner**

*Choose this option if you have no experience with cleaning data or exploratory data analysis.*

Select one of the beginner data sets, which already has a summary of findings, from the [**Data Set Options**](https://docs.google.com/document/d/1w7KhqotVi5eoKE3I_AZHbsxdr-NmcWsLTIiZrpxWx4w/pub)document. Then, create a visualization that communicates the findings.

**Intermediate**

*Choose this option if you have some experience cleaning and analyzing data.*

Select one of the intermediate data sets from the [**Data Set Options**](https://docs.google.com/document/d/1w7KhqotVi5eoKE3I_AZHbsxdr-NmcWsLTIiZrpxWx4w/pub)document. These data sets are not necessarily clean or tidy[[1]](https://docs.google.com/document/d/1uOT-uAJSySO5jWA_1xvUGMMzn1cgLapX9lH4iWGKnvc/pub#ftnt1) data sets. You will investigate the data set to share a story or message about the data and then create a suitable visualization.

**Advanced**

*Choose this option if you are comfortable finding, cleaning, and analyzing a data set.*

Find a data set, investigate it, and share your findings in a visualization. Your final graphic should primarily be explanatory, but it may also contain exploratory components. You can find a list of recommended websites to find data sets in the [**Data Set Options**](https://docs.google.com/document/d/1w7KhqotVi5eoKE3I_AZHbsxdr-NmcWsLTIiZrpxWx4w/pub) document. You should be aware that finding your own data set, cleaning the data set, and analyzing it (using R, iPython Notebook, or another tool) can take considerable time and effort. This can lengthen the time you spend on your project by days, weeks, or even months. Choose the option only if you feel prepared for a challenge!

Now, on to the details!

**Step One - Choose a Data Set**

First, you will choose a data set from the [**Data Set Options**](https://docs.google.com/document/d/1w7KhqotVi5eoKE3I_AZHbsxdr-NmcWsLTIiZrpxWx4w/pub)document or find a data set to explore and visualize. You should choose a data set based on your prior experiences in programming and working with data. The data set you choose will not increase or decrease your chances of passing this project.

**Step Two - Get Organized**

Eventually you’ll want to submit your project and share it.  If you are familiar with [GitHub](https://www.google.com/url?q=https%3A%2F%2Fgithub.com%2F&sa=D&sntz=1&usg=AFQjCNHReqsuKT6C86HcgL4TbSevF24rxQ), we encourage you to create a public repository or a public [Gist](https://www.google.com/url?q=https%3A%2F%2Fgist.github.com%2F&sa=D&sntz=1&usg=AFQjCNG9eZTMJSrRIUGgGjzZU6NvSwJaPQ) for your project to track changes. Otherwise, you need to create the following files.

1. an **index.html** file containing the code to create your visualization (you may include the JavaScript and CSS in this file or separate them in other files)
2. a **README.md** file that includes **four sections**...
3. **Summary** - in no more than 4 sentences, briefly introduce your data visualization and add any context that can help readers understand it
4. **Design** - explain any design choices you made including changes to the visualization after collecting feedback
5. **Feedback** - include all feedback you received from others on your visualization from the first sketch to the final visualization
6. **Resources**- list any sources you consulted to create your visualization
7. **data files**
8. the final data set used to create the visualization (usually .csv, .tsv, or .json file)
9. a codebook or other files related to the data set (description, readme, license)
10. *OPTIONAL FOLDERS IF YOU USE*[*GITHUB*](https://www.google.com/url?q=https%3A%2F%2Fgithub.com%2F&sa=D&sntz=1&usg=AFQjCNHReqsuKT6C86HcgL4TbSevF24rxQ)
11. **data** folder to include all the data related files
12. **js** folder to include .js files (not needed if  javascript is in the index.html file)
13. **css** folder to include .css files (not needed if CSS is in the index.html file)

**Step Three - Find a Data Story**

Explore your data set and craft a message or story around your data! Think about the overall message you want to convey and think about the comparison(s) or relationship(s) you want your readers to see.

**Step Four - Create Your Visualization**

First, sketch ideas for your visualization. Once you settle on a sketch, explain any design choices in that sketch, such as chart type, visual encodings, and layout, in the **Design** section of the **README.md** file. Then, write code to create your visualization using [dimple.js](http://www.google.com/url?q=http%3A%2F%2Fdimplejs.org%2F&sa=D&sntz=1&usg=AFQjCNFzf9EI7W8pwtiWmb9ppvMbCQRLzQ) or [d3.js](http://www.google.com/url?q=http%3A%2F%2Fd3js.org%2F&sa=D&sntz=1&usg=AFQjCNGRDjOIx27VEdt3kwBvGqXUgHgu7A). The visualization must include animation, interaction, or both. See the [**Project Rubric**](https://docs.google.com/document/d/1zRVs73M7P5ACKB0n3Di4k0AskId3pc6lIpMBmmydETk/pub) for more information.

**Step Five - Get Feedback**

Share your visualization with **at least 3 other people** and document their feedback.There are many ways to get feedback, and more feedback is generally better! Here are some options.

* Share your visualization with others in person and have them think aloud as they read and explore the graphic so you can document what stands out to them and how they interpret the graphic.
* Share a link to your repository in the discussions and ask others to share constructive criticisms. Be sure to offer advice to others who are seeking feedback too!
* Create and share a [Gist](https://www.google.com/url?q=https%3A%2F%2Fgist.github.com%2F&sa=D&sntz=1&usg=AFQjCNG9eZTMJSrRIUGgGjzZU6NvSwJaPQ), which contains an **index.html** file, data file, and any .js or .css files. Directions for creating and sharing a Gist can be found at [http://bl.ocks.org/](http://www.google.com/url?q=http%3A%2F%2Fbl.ocks.org%2F&sa=D&sntz=1&usg=AFQjCNGEXjmUwrCSkVSI4e_Ax4Rj-s3HqA).
* **Box Plots Gist EXAMPLE**:
* [https://gist.github.com/mbostock/4061502](https://www.google.com/url?q=https%3A%2F%2Fgist.github.com%2Fmbostock%2F4061502&sa=D&sntz=1&usg=AFQjCNHx6RQgzz97dPjavKDqH2are1Hdfg)
* [http://bl.ocks.org/mbostock/4061502](http://www.google.com/url?q=http%3A%2F%2Fbl.ocks.org%2Fmbostock%2F4061502&sa=D&sntz=1&usg=AFQjCNEVOwaamLJRCWilzM8_0OBDOnOG-A)

You might need to ask specific questions to prompt the reader. Here are some questions to help you. You can, of course, ask others.

* What do you notice in the visualization?
* What questions do you have about the data?
* What relationships do you notice?
* What do you think is the main takeaway from this visualization?
* Is there something you don’t understand in the graphic?

**Step Six - Document Feedback and Improve the Visualization**

For each person that gives you feedback, add the person’s feedback to your **README.md** file in the **Feedback** section. As you improve and iterate on your visualization, update your code**AND** describe any changes in the **Design** section of the **README.md** file. You should be building evidence that you have shared your visualization, received feedback, and responded to that feedback. You will need to submit the different versions of your visualization.  At the least you need to submit the initial version of the index.html file and the final index.html file.

**Step Seven - Review**

Use the [**Project Rubric**](https://docs.google.com/document/d/1zRVs73M7P5ACKB0n3Di4k0AskId3pc6lIpMBmmydETk/pub) to review your project. If you are happy with your submission, then you’re ready to submit your project. If you see room for improvement, keep working to improve your project!

**Step Seven - Submit**

Follow the instructions in the [**Project Submission Instructions**](https://docs.google.com/document/d/13nj9Q0UhsP7dcm3zK9K6Lsr6egeM31NRz55rhldQJW0/pub) document.

[[1]](https://docs.google.com/document/d/1uOT-uAJSySO5jWA_1xvUGMMzn1cgLapX9lH4iWGKnvc/pub#ftnt_ref1)  Tidy data sets are data sets that have a particular structure where every row is an observation and every column contains data related to one variable. The type of data contained in a column is the same for each column. Read more about tidy data in Hadley Wickham’s paper, [http://vita.had.co.nz/papers/tidy-data.pdf](http://www.google.com/url?q=http%3A%2F%2Fvita.had.co.nz%2Fpapers%2Ftidy-data.pdf&sa=D&sntz=1&usg=AFQjCNHnPBeCTuFyplFdk-OwYzm4qdPy5w)

## Data Set Options

Choose from one of the following data sets or find your own. Additional resources for finding a data set are included at the bottom of this document.

|  |  |  |  |
| --- | --- | --- | --- |
| **Difficulty Level** | **Data Set** | **Overview** | **Notes** |
| **Beginner** | [**Titanic Data**](https://www.google.com/url?q=https%3A%2F%2Fwww.kaggle.com%2Fc%2Ftitanic-gettingStarted&sa=D&sntz=1&usg=AFQjCNFbJiKjyF-dX0enJMQe4G1LPMDAXg) | Contains demographics and passenger information of the 2224 passengers and crew onboard the Titanic. | Create a visualization that shows the demographics or passenger information between those passengers who survived and those who died |
| **Beginner** | [**Baseball Data**](https://www.google.com/url?q=https%3A%2F%2Fs3.amazonaws.com%2Fudacity-hosted-downloads%2Fud507%2Fbaseball_data.csv&sa=D&sntz=1&usg=AFQjCNEkK8NRImfPdhM7cLkivKaJ0WldFA) | A data set containing 1,157 baseball players including their handedness (right or left handed), height, weight, batting average, and home runs. | Create a visualization that shows differences among the performance of the baseball players. |
| **Intermediate** | [**Flights**](http://www.google.com/url?q=http%3A%2F%2Fstat-computing.org%2Fdataexpo%2F2009%2Fthe-data.html&sa=D&sntz=1&usg=AFQjCNEbIjTiS2ESi1oTxhG2CXsRsm4_aA) | The data set which contains information on United State flight delays and performance comes from [RITA](http://www.google.com/url?q=http%3A%2F%2Fwww.transtats.bts.gov%2FOT_Delay%2FOT_DelayCause1.asp&sa=D&sntz=1&usg=AFQjCNFd7ah4TD2BpvCZpjqAIKF1ZtZlwA). You can download the data directly from RITA or as bzipped csv files from the Flights link. The files on the Flights link are organized by year and are more compressed than the originals.  Additional details about the data can be found at here. | Investigate the performance of flights over time or simply look at data for a given year and create a graphic that showcases your finding(s). |
| **Intermediate** | [**Loan Data from Prosper**](https://www.google.com/url?q=https%3A%2F%2Fs3.amazonaws.com%2Fudacity-hosted-downloads%2Fud651%2FprosperLoanData.csv&sa=D&sntz=1&usg=AFQjCNE1BPfVPGV0q0A3xkZxq8VBcfYuoA)  Last updated 03/11/2014  This [data dictionary](https://docs.google.com/spreadsheet/ccc?key=0AllIqIyvWZdadDd5NTlqZ1pBMHlsUjdrOTZHaVBuSlE&usp=sharing) explains the variables in the data set. | This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, and many others. | Ask your own questions about this data set to find interesting trends in the data. |
| **Advanced** | [**PISA Data**](https://www.google.com/url?q=https%3A%2F%2Fs3.amazonaws.com%2Fudacity-hosted-downloads%2Fud507%2Fpisa2012.csv.zip&sa=D&sntz=1&usg=AFQjCNEyguncpUFRtJbBuRnS67mCWwix9g)  [**PISA Data Dictionary**](https://www.google.com/url?q=https%3A%2F%2Fs3.amazonaws.com%2Fudacity-hosted-downloads%2Fud507%2Fpisadict2012.csv&sa=D&sntz=1&usg=AFQjCNH8vf9t3J5YGBYuy01lp7RQw-KyhQ)  Note: The unzipped PISA Data csv file is 2.75 GB. | PISA is a survey of students' skills and knowledge as they approach the end of compulsory education. It is not a conventional school test. Rather than examining how well students have learned the school curriculum, it looks at how well prepared they are for life beyond school.  Around 510,000 students in [65 economies](http://www.google.com/url?q=http%3A%2F%2Fwww.oecd.org%2Fpisa%2Faboutpisa%2Fpisa-2012-participants.htm&sa=D&sntz=1&usg=AFQjCNGoUwGrQi-6Z0RFt_B1aAQt2EN5wg) took part in the PISA 2012 assessment of reading, mathematics and science representing about 28 million 15-year-olds globally. Of those economies, 44 took part in an assessment of creative problem solving and 18 in an assessment of financial literacy.  The data and topics of investigation come from the [PISA Data Visualization Competition](http://www.google.com/url?q=http%3A%2F%2Fwww.oecd.org%2Fpisa%2Fpisaproducts%2Fdatavisualizationcontest.htm&sa=D&sntz=1&usg=AFQjCNGn92U2dAsE-hggpGaUSeuj8sCosg). For inspiration and examples, see the winners and submissions [here](http://www.google.com/url?q=http%3A%2F%2Fbeta.icm.edu.pl%2FPISAcontest%2F&sa=D&sntz=1&usg=AFQjCNGyXaByGXwTZsQiwdPEyryZ8r6d4A). | Consider creating a graphic that explores one of the following topics.  The importance of school factors in explaining academic performance.  Differences in achievement based on gender, location, or student attitudes.  Differences in achievement based on teacher practices and attitudes.  Inequalities in academic achievement. |
| **Varies**  Depends on your experience working with data. | **Find your own data set!** | **Remember that finding and cleaning your own data set could take significant time and effort!**See the checklist below if you want to choose your own data set. | Pose your own question and find data to answer it. Alternatively, find a data set and ask questions about it until you find something interesting you want to share. |

**If you’re finding your own data set… (see next page)**

The data set that you eventually submit should:

* be in a tidy format1(you may need to clean and reshape the data)
* be in a commonly used format of loading data with dimple.js or d3.js such as .csv, .tsv, .txt, .json, .xml, or .html

Here are a few resources to find a data set:

* [http://www.pewglobal.org/category/datasets/](http://www.google.com/url?q=http%3A%2F%2Fwww.pewglobal.org%2Fcategory%2Fdatasets%2F&sa=D&sntz=1&usg=AFQjCNFyKVICv4eQIo3FzBfobc25LrHZoQ)
* [http://databank.worldbank.org/data/home.aspx](http://www.google.com/url?q=http%3A%2F%2Fdatabank.worldbank.org%2Fdata%2Fhome.aspx&sa=D&sntz=1&usg=AFQjCNGZsX6JQayrl70pgCwCT3_vibdTqA)
* [http://www.data.gov/](http://www.google.com/url?q=http%3A%2F%2Fwww.data.gov%2F&sa=D&sntz=1&usg=AFQjCNFXOs5zqy8V6j4xTutfyzdEkG0-tg)
* [http://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public](http://www.google.com/url?q=http%3A%2F%2Fwww.quora.com%2FWhere-can-I-find-large-datasets-open-to-the-public&sa=D&sntz=1&usg=AFQjCNEke8sYjCkrOJvqjJUwcnr_q87Yew)
* [http://www.inside-r.org/howto/finding-data-internet](http://www.google.com/url?q=http%3A%2F%2Fwww.inside-r.org%2Fhowto%2Ffinding-data-internet&sa=D&sntz=1&usg=AFQjCNGQm8WS79PZHdw0svQ_Qk56vrIVuw)
* [https://www.edsurge.com/n/2014-01-21-education-datapalooza](https://www.google.com/url?q=https%3A%2F%2Fwww.edsurge.com%2Fn%2F2014-01-21-education-datapalooza&sa=D&sntz=1&usg=AFQjCNG7WuPZaD98b53DqPRWjXpn8k_slQ)
* [1,001 Data Sets](https://www.google.com/url?q=https%3A%2F%2Fdreamtolearn.com%2Fdoc%2F2HDNJH3XJU6CVGKZ7SDM4MCSW&sa=D&sntz=1&usg=AFQjCNEbtDyu2fZNZdXMlGK-foYqNJo9Og)

1 Tidy data sets are data sets that have a particular structure. Read more about tidy data in Hadley Wickham’s paper, [http://vita.had.co.nz/papers/tidy-data.pdf](http://www.google.com/url?q=http%3A%2F%2Fvita.had.co.nz%2Fpapers%2Ftidy-data.pdf&sa=D&sntz=1&usg=AFQjCNHnPBeCTuFyplFdk-OwYzm4qdPy5w)

## Data Visualization Project Rubric

**Overview**

This rubric is here to help you understand the specifications for the data visualization that you create. It is the same rubric that the person evaluating your project will use. We will refer to this person as the "project evaluator" in this document. You  should look at the rubric **before you begin working** on your analysis **and before you submit it**.

**How to Use: before you begin**

1. Look at the bold headings under the criteria column to understand what the project evaluator will be looking for in your project.
2. Go through each criteria item in more detail.
3. Familiarize yourself with what is required for your project to "meet specifications" or to be "completely Udacious." In order to gain a certificate, you need to "meet specifications", however, to gain the most benefit and learn most from the experience, we encourage you to aim for "completely Udacious".

**How to Use: before you submit**

1. Once your project is complete, go through each criteria item and do your best to honestly evaluate where you think your project falls.
2. If you think your project "does not meet specifications" for **any**criteria item, then you should make some changes to your analysis.
3. Once you’re confident that your project "meets specifications" or is "completely Udacious," go ahead and follow the Project Submission Instructions to submit!

**How Grading Works**

1. Your project evaluator will use this rubric to evaluate your analysis.
2. Your grade will simply be "pass" or "doesn’t pass."
3. You earn a "pass" by not having **any** criteria items in the "does not meet specifications" column.
4. If any criteria item "does not meet specifications," you will not pass. You will be able to make changes and re-submit the project.

**The Rubric**

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria** | **Does not meet specifications** | **Meets specifications** | **Exceeds specifications**  (Completely Udacious) |
| **Code Structure and Functionality** |  |  |  |
| **Does the code work?** | The code does not render the visualization, some components of the visualization do not appear correct, or some interaction or animation breaks while a reader interacts with the visualization. | The visualization renders and any interactions or animations work as the reader interacts with the visualization. | ***Not Applicable*** |
| **Is the code in the**index.html**commented in a way that is useful and not excessive?** | Code is not commented or complex code is not adequately explained with comments. | Large code chunks are commented, and all complex code is adequately explained with comments. Comments are not overused to explain obvious code. | ***Not Applicable*** |
| **Does student’s code use formatting techniques (indents, spaces, line breaks, etc…) to improve readability?**  (Refer to [Google’s Style Guide for Javascript](https://google-styleguide.googlecode.com/svn/trunk/javascriptguide.xml)) | The code does not use formatting techniques or formatting techniques do not improve readability. Some lines are longer than 80 characters. | The code uses formatting techniques in a consistent and effective manner to improve code readability. All lines are shorter than 80 characters. | ***Not Applicable*** |
| **Design** |  |  |  |
| **Is the data visualization explanatory? Does the data visualization communicate a clear story or message about the data?** | The visualization does not have a clear story or message about the data. The visualization may confuse readers based on design choices such as chart type, visual encodings, layout, legends, or hierarchy. | The visualization communicates a clear finding or relationship in the data. Design choices foster communication between the reader and the visualization. | ***Not Applicable*** |
| **Does the written summary reflect what a reader would interpret from the graphic?** | A reader would be confused by graphic, or a reader would not able to identify the main point(s) or relationship(s) that the graphic attempts to convey. | A reader’s summary of the graphic would closely match the written summary in the README.md file, or a reader would identify at least 1 main point or relationship that the graphic attempts to convey. | ***Not Applicable*** |
| **Does the data visualization incorporate interaction or animation?** | The visualization is static and does not include any interaction or animation which could allow the reader to better understand the data. | The visualization includes interaction or animation. The interaction or animation may be simple, such as a hover, tooltip, or transition. Interaction or animation enhance understanding of the data. | The data visualization incorporates more advanced techniques beyond the scope of the class. The techniques enhance the reader’s ability to understand the data and interact with the graphic. |
| **Are initial design decisions documented?** | The student does not explain initial design decisions such as chart type, visual encodings, layout, legends, or hierarchy, or the student does not incldue a Design section in the README.md file. | The student explains initial design decisions such as chart type, visual encodings, layout, legends, or hierarchy. These are included at the beginning of the Design section in the README.md file. | ***Not Applicable*** |
| **Feedback and Iteration** |  |  |  |
| **Does the student collect feedback after sharing the initial visualization?**  *We encourage you to collect feedback from the first sketch to the final visualization.* | The student does not include a Feedback section in the README.md file, or the student does not collect feedback from at least three different people. | The student collects feedback from at least three people throughout the process of creating the data visualization. The feedback is documented in the Feedback section of the README.md file. | The student collects feedback from many people throughout the process of creating the data visualization. The student provides other evidence of feedback, such as audio files, videos, discussion forum links, or images of sketches with handwritten comments. |
| **Does the student iterate on the visualization?**  **Does the student incorporate feedback to improve the visualization? If not, does the student explain why the design of the visualization did not change?** | The student has not presented any evidence that the visualization has been improved since the first sketch or the first coded version of the visualization, or the student has not defended why no changes were not made to the visualization after gathering feedback. | The student presents evidence that the visualization has been improved since the first sketch or the first coded version of the visualization. The student has listed all of the feedback in the Feedback section of the README.md file. Most design choices and changes are accounted for in the Design section of the README.md file. | The student presents overwhelming evidence that the visualization has been improved. The student has listed all of the feedback in the Feedback section of the README.md file. All design choices and changes are accounted for in the Design section of the README.md file. |

## Data Vis Project Submission Instructions

1. Go through the following checklist:

* I am proud of the data visualization I created and would share it with others.
* I prepared my data visualization according to the instructions in the [**Project Description**](https://docs.google.com/document/d/1uOT-uAJSySO5jWA_1xvUGMMzn1cgLapX9lH4iWGKnvc/pub).
* I checked my project and all related components against the [**Project Rubric**](https://docs.google.com/document/d/1zRVs73M7P5ACKB0n3Di4k0AskId3pc6lIpMBmmydETk/pub).

1. Once you’ve completed the checklist, email [datavis-project@udacity.com](mailto:eda-project@udacity.com) with:

* the *original***index.html** file for the first version of your graphic
* the *final***index.html**file for the final version of your graphic
* the **README.md** file with the sections Summary, Design, Feedback, and Resources
* the Resources section in the README.md file should contain a list of websites, books, forums, blog posts, GitHub repositories, etc. that you referred to or used in creating your submission (add N/A if you did not use any such resources).
* the **final data set file** used for the graphic (usually .csv, .tsv, or .json)
* OPTIONAL: additional versions of your **index.html** as you iterated based on feedback (index1.html, index2.html, index3.html, … , index\_final.html)
* OPTIONAL: instead of sending all of items separately, send a link to the public GitHub repository or public Gist that contains your project and all supporting files
* Please carefully read the following statement and include it in your email:

*“I hereby confirm that this submission is my work. I have cited above the origins of****any****parts of the submission that were taken from websites, books, forums, blog posts, github repositories, etc. By including this in my email, I understand that I will be expected to explain my work in a video call with a Udacity coach before I can receive my verified certificate.”*

1. After you send your email, you will receive a confirmation email from support@udacity that your message was received.
2. Within 7 business days, you will receive an email from your project evaluator (who will not be the coach you’ve been working with) with a completed rubric and instructions for next steps.
3. If your project was evaluated as "passing," you will schedule an exit interview. The purpose of this interview is to talk to you about your analysis and verify that you are the person who created it. Don’t worry, this interview will not be difficult or stressful (assuming you are the person who completed the project).
4. If you still have questions, please review the [Udacity Project Submission FAQ](https://docs.google.com/a/knowlabs.com/document/d/1BPeNySE5xpkdcTeY9bINU16UNz7aqqxy3NzqQubds_I/pub) or email your coach.