# Data Analyst Nanodegree Syllabus

Discover Insights from Data



Thank you for your interest in the Data Analyst Nanodegree! In order to succeed in this program, we recommend having experience programing in Python. If you've never programmed before, or want a refresher, you can prepare for this Nanodegree with Lessons 1-4 of <u>Intro to Computer Science</u>.

# Project 0: Analyze Bay Area Bike Share Data

This project will introduce you to the key steps of the data analysis process. You'll do so by analyzing data from a bike share company found in the San Francisco Bay Area. You'll submit this project in your first 7 days, and by the end you'll be able to:

- → Use basic Python code to clean a dataset for analysis
- → Run code to create visualizations from the wrangled data
- → Analyze trends shown in the visualizations and report your conclusions
- → Determine if this program is a good fit for your time and talents

### Project 1: Test a Perceptual Phenomenon

In this project, you'll use descriptive statistics and a statistical test to analyze the Stroop effect, a classic result of experimental psychology. Communicate your understanding of the data and use statistical inference to draw a conclusion based on the results.

#### Supporting Lesson Content: Statistics

| Lesson Title                 | Learning Outcomes  |
|------------------------------|--|
| INTRO TO RESEARCH<br>METHODS | → Identify several statistical study methods and describe the positives and negatives of each  |
| VISUALIZING DATA             | → Create and interpret histograms, bar charts, and frequency plots                             |
| CENTRAL TENDENCY             | → Compute and interpret the 3 measures of center for distributions: the mean, median, and mode |



| VARIABILITY               | <ul> <li>→ Quantify the spread of data using the range and standard deviation</li> <li>→ Identify outliers in data sets using the interquartile range</li> </ul>                 |
|---------------------------|--|
| STANDARDIZING             | <ul> <li>→ Convert distributions into the standard normal distribution using the Z-score</li> <li>→ Compute proportions using standardized distributions</li> </ul>              |
| NORMAL DISTRIBUTION       | <ul> <li>→ Use normal distributions to compute probabilities</li> <li>→ Use the Z-table to look up the proportions of observations above, below, or in between values</li> </ul> |
| SAMPLING<br>DISTRIBUTIONS | → Apply the concepts of probability and normalization to sample data sets  |
| ESTIMATION                | → Estimate population parameters from sample statistics using confidence intervals   |
| HYPOTHESIS TESTING        | → Use critical values to make decisions on whether or not a treatment has changed the value of a population parameter  |
| T-TESTS                   | → Test the effect of a treatment or compare the difference in means for two groups when we have small sample sizes   |

### Project 2: Investigate a Dataset

In this project, you'll choose one of Udacity's curated datasets and investigate it using NumPy and pandas. You'll complete the entire data analysis process, starting by posing a question and finishing by sharing your findings.

### Supporting Lesson Content: Introduction to Data Analysis

| Lesson Title                    | Learning Outcomes  |
|---------------------------------|--|
| DATA ANALYSIS PROCESS           | <ul> <li>→ Identify the key steps in the data analysis process</li> <li>→ Complete an analysis of Udacity student data using pure<br/>Python, with minimal reliance on additional libraries</li> </ul> |
| NUMPY AND PANDAS FOR<br>1D DATA | → Use NumPy arrays, pandas series, and vectorized operations to ease the data analysis process   |
| NUMPY AND PANDAS FOR<br>2D DATA | <ul> <li>→ Use two-dimensional NumPy arrays and pandas DataFrames</li> <li>→ Understand how to group data and to combine data from multiple files</li> </ul>   |

### Project 3: Wrangle OpenStreetMap Data

In this project, you'll 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.

### Supporting Lesson Content: Data Wrangling with MongoDB or SQL

| Lesson Title                      | Learning Outcomes   |
|-----------------------------------|---|
| DATA EXTRACTION<br>FUNDAMENTALS   | <ul> <li>→ Properly assess the quality of a dataset</li> <li>→ Understand how to parse CSV files and XLS with XLRD</li> <li>→ Use JSON and Web APIs</li> </ul>  |
| DATA IN MORE COMPLEX<br>FORMATS   | <ul> <li>→ Understand XML design principles</li> <li>→ Parse XML &amp; HTML</li> <li>→ Scrape websites for relevant data</li> </ul>   |
| DATA QUALITY                      | <ul> <li>→ Understand common sources for dirty data</li> <li>→ Measure the quality of a dataset &amp; apply a blueprint for cleaning</li> <li>→ Properly audit validity, accuracy, completeness, consistency, and uniformity of a dataset</li> </ul>  |
| WORKING WITH<br>MONGODB           | <ul> <li>→ Understand how data is modeled in MongoDB</li> <li>→ Run field and projection queries</li> <li>→ Import data into MongoDB using mongoimport</li> <li>→ Utilize operators like \$gt, \$lt, \$exists, \$regex</li> <li>→ Query arrays and using \$in and \$all operators</li> <li>→ Change entries using \$update, \$set, \$unset</li> </ul> |
| ANALYZING DATA                    | <ul> <li>→ Identify common examples of the aggregation framework</li> <li>→ Use aggregation pipeline operators \$match, \$project, \$unwind,<br/>\$group</li> </ul>   |
| SQL FOR DATA ANALYSIS             | <ul> <li>→ Understand how data is structured in SQL</li> <li>→ Run queries to summarize data</li> <li>→ Use joins to combine information across tables</li> <li>→ Create tables and import data from csv</li> </ul>   |
| CASE STUDY:<br>OPENSTREETMAP DATA | <ul> <li>→ Use iterative parsing for large datafiles</li> <li>→ Understand XML elements in OpenStreetMap</li> </ul>   |



# Project 4: Explore and Summarize Data

In this project, you'll use R and apply exploratory data analysis techniques to explore a selected data set for distributions, outliers, and anomalies.

### Supporting Lesson Content: Data Analysis with R

| Lesson Title                      | Learning Outcomes  |
|-----------------------------------|--|
| WHAT IS EDA?                      | → Define and identify the importance of exploratory data analysis (EDA)  |
| R BASICS                          | <ul> <li>→ Install RStudio and packages</li> <li>→ Write basic R scripts to inspect datasets</li> </ul>  |
| EXPLORE ONE VARIABLE              | <ul> <li>→ Quantify and visualize individual variables within a dataset</li> <li>→ Create histograms and boxplots</li> <li>→ Transform variables</li> <li>→ Examine and identify tradeoffs in visualizations</li> </ul>                    |
| EXPLORE TWO VARIABLES             | <ul> <li>→ Properly apply relevant techniques for exploring the relationship between any two variables in a data set</li> <li>→ Create scatter plots</li> <li>→ Calculate correlations</li> <li>→ Investigate conditional means</li> </ul> |
| EXPLORE MANY<br>VARIABLES         | → Reshape data frames and use aesthetics like color and shape to uncover information   |
| DIAMONDS AND PRICE<br>PREDICTIONS | → Use predictive modeling to determine a good price for a diamond  |



# Project 5: Intro to Machine Learning

In this project, you'll 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.

### Supporting Lesson Content: Introduction to Machine Learning

| Lesson Title                    | Learning Outcomes  |
|---------------------------------|--|
| SUPERVISED<br>CLASSIFICATION    | <ul> <li>→ Implement the Naive Bayes algorithm to classify text</li> <li>→ Implement Support Vector Machines (SVMs) to generate new features independently on the fly</li> <li>→ Implement decision trees as a launching point for more sophisticated methods like random forests and boosting</li> </ul>  |
| DATASETS AND<br>QUESTIONS       | → Wrestle the Enron dataset into a machine-learning-ready format in preparation for detecting cases of fraud   |
| REGRESSIONS AND<br>OUTLIERS     | → Use regression algorithms to make predictions and identify<br>and clean outliers from a dataset  |
| UNSUPERVISED LEARNING           | → Use the k-means clustering algorithm for pattern-searching on unlabeled data   |
| FEATURES, FEATURES,<br>FEATURES | <ul> <li>→ Use feature creation to take your human intuition and change raw features into data a computer can use</li> <li>→ Use feature selection to identify the most important features of your data</li> <li>→ Implement principal component analysis (PCA) for a more sophisticated take on feature selection</li> <li>→ Use tools for parsing information from text-type data</li> </ul> |
| VALIDATION AND<br>EVALUATION    | <ul> <li>→ Implement the train-test split and cross-validation to validate<br/>and understand machine learning results</li> <li>→ Quantify machine learning results using precision, recall, and<br/>F1 score</li> </ul>   |

# Project 6: Make an Effective Visualization

In this project, you'll create a data visualization from a data set that tells a story or highlights trends or patterns in the data. 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, harnessing visual encodings and design principles for effective communication.

### Supporting Lesson Content: Data Visualization and D3.js

| Lesson Title                  | Learning Outcomes  |
|-------------------------------|--|
| VISUALIZATION<br>FUNDAMENTALS | → Identify the elements of great visualization in the context of data science  |
| D3 BUILDING BLOCKS            | <ul> <li>→ Use the open standards of the web to create graphical elements</li> <li>→ Select elements on a page</li> <li>→ Add and style SVG elements</li> </ul>  |
| DESIGN PRINCIPLES             | → Select the appropriate graph and color to create an effective visualization for different datasets   |
| DIMPLE.JS                     | → Create graphics using the Dimple JavaScript library  |
| NARRATIVES                    | <ul> <li>→ Incorporate different narrative structures into your visualizations</li> <li>→ Identify different types of bias in the data visualization process</li> <li>→ Add context to your data visualizations</li> </ul> |
| ANIMATION AND<br>INTERACTION  | → Incorporate animation and interaction to bring more audience insights into your visualizations using D3.js   |



# Project 7: Design an A/B Test

In this project, you'll make design decisions for an A/B test, including which metrics to measure and how long the test should be run. Analyze the results of an A/B test that was run by Udacity and recommend whether or not to launch the change.

### Supporting Lesson Content: A/B Testing

| Lesson Title                              | Learning Outcomes  |
|---|--|
| OVERVIEW OF A/B<br>TESTING                | → Identify the key concepts and considerations when designing and conducting an A/B test   |
| POLICY AND ETHICS FOR<br>EXPERIMENTS      | <ul> <li>→ Adequately protect the participants in experiments</li> <li>→ Identify the four main ethical principles to consider when designing experiments</li> </ul>   |
| CHOOSING AND<br>CHARACTERIZING<br>METRICS | <ul> <li>→ Identify techniques for brainstorming metrics</li> <li>→ List possible alternatives when unable to directly measure a desired metric</li> <li>→ Identify characteristics to consider when validating metrics</li> </ul>                   |
| DESIGNING AN<br>EXPERIMENT                | <ul> <li>→ Identify the proper users to be in control and experiment groups</li> <li>→ Calculate the number of events necessary to reach significance</li> <li>→ Define how different design decisions affect the size of your experiment</li> </ul> |
| ANALYZING RESULTS                         | <ul> <li>→ Identify the key steps for analyzing the results of an experiment</li> <li>→ Measure multiple metrics within a single experiment</li> <li>→ Understand why statistically significant results may disappear at launch</li> </ul>           |

