Data Analyst Nanodegree Syllabus

Discover Insights from Data

Before You Start

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 Intro to Computer Science.

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 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	➔ Quantify the spread of data using the range and standard deviation ➔ Identify outliers in data sets using the interquartile range
STANDARDIZING	➔ Convert distributions into the standard normal distribution using the Z-score ➔ Compute proportions using standardized distributions
NORMAL DISTRIBUTION	➔ Use normal distributions to compute probabilities ➔ Use the Z-table to look up the proportions of observations above, below, or in between values
SAMPLING 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	➔ Identify the key steps in the data analysis process ➔ Complete an analysis of Udacity student data using pure Python, with minimal reliance on additional libraries
NUMPY AND PANDAS FOR 1D DATA	➔ Use NumPy arrays, pandas series, and vectorized operations to ease the data analysis process
NUMPY AND PANDAS FOR 2D DATA	➔ Use two-dimensional NumPy arrays and pandas DataFrames ➔ Understand how to group data and to combine data from multiple files

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

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	➔ Install RStudio and packages ➔ Write basic R scripts to inspect datasets
EXPLORE ONE VARIABLE	➔ Quantify and visualize individual variables within a dataset ➔ Create histograms and boxplots ➔ Transform variables ➔ Examine and identify tradeoffs in visualizations
EXPLORE TWO VARIABLES	➔ Properly apply relevant techniques for exploring the relationship between any two variables in a data set ➔ Create scatter plots ➔ Calculate correlations ➔ Investigate conditional means
EXPLORE MANY VARIABLES	➔ Reshape data frames and use aesthetics like color and shape to uncover information
DIAMONDS AND PRICE 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 CLASSIFICATION	➔ Implement the Naive Bayes algorithm to classify text ➔ Implement Support Vector Machines (SVMs) to generate new features independently on the fly ➔ Implement decision trees as a launching point for more sophisticated methods like random forests and boosting
DATASETS AND QUESTIONS	➔ Wrestle the Enron dataset into a machine-learning-ready format in preparation for detecting cases of fraud
REGRESSIONS AND OUTLIERS	➔ Use regression algorithms to make predictions and identify and clean outliers from a dataset
UNSUPERVISED LEARNING	➔ Use the k-means clustering algorithm for pattern-searching on unlabeled data
FEATURES, FEATURES, FEATURES	➔ Use feature creation to take your human intuition and change raw features into data a computer can use ➔ Use feature selection to identify the most important features of your data ➔ Implement principal component analysis (PCA) for a more sophisticated take on feature selection ➔ Use tools for parsing information from text-type data
VALIDATION AND EVALUATION	➔ Implement the train-test split and cross-validation to validate and understand machine learning results ➔ Quantify machine learning results using precision, recall, and F1 score

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 FUNDAMENTALS	➔ Identify the elements of great visualization in the context of data science
D3 BUILDING BLOCKS	➔ Use the open standards of the web to create graphical elements ➔ Select elements on a page ➔ Add and style SVG elements
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	➔ Incorporate different narrative structures into your visualizations ➔ Identify different types of bias in the data visualization process ➔ Add context to your data visualizations
ANIMATION AND 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 TESTING	➔ Identify the key concepts and considerations when designing and conducting an A/B test
POLICY AND ETHICS FOR EXPERIMENTS	➔ Adequately protect the participants in experiments ➔ Identify the four main ethical principles to consider when designing experiments
CHOOSING AND CHARACTERIZING METRICS	➔ Identify techniques for brainstorming metrics ➔ List possible alternatives when unable to directly measure a desired metric ➔ Identify characteristics to consider when validating metrics
DESIGNING AN EXPERIMENT	➔ Identify the proper users to be in control and experiment groups ➔ Calculate the number of events necessary to reach significance ➔ Define how different design decisions affect the size of your experiment
ANALYZING RESULTS	➔ Identify the key steps for analyzing the results of an experiment ➔ Measure multiple metrics within a single experiment ➔ Understand why statistically significant results may disappear at launch