# Syllabus - Data Science for Economists - Summer 2023

**Proposed Course Number** - Economics 148

**Unit Value:** 4 units

Proposed for Summer 2023

**Course Description**

This course will give the undergraduate student the basic computational building blocks needed to be a good consumer and producer of applied economics work. The outline for the class will be to follow the chronological order of doing an applied economics data science project from start to finish. A starting point would be to add skills and experience working with data Application Programming Interface (APIs) and making datasets from census data or working with open archived datasets. Another part of the course is to give students tools and experience wrangling data, working with incomplete or unstructured data, joining and merging data, all of which would not be included in the classic econometrics class. The next part of the Data Science lifecycle, students will undertake exploratory data analysis (EDA) to examine a new dataset and build Data Visualization to describe aspects of the data. Students will then cover basics of preparing data for econometric analysis. Practices around literate code, open science tools, reproducibility practices, and data management would also be covered. Students will undertake peer review and project-based work, partly in order to have project-based examples of literate and reproducible code. Guest Lecturers will be invited throughout the semester to introduce real examples of applications of Data Science within Economics research.

**Prerequisites**

1. DATA C8\COMPSCI C8\INFO C8\STAT C8

or

1. Stat 20 + familiarity with Python recommended

**Lectures**

There will be 10 hours of lecture per week - 4x2.5-hour lectures for six weeks. When appropriate these lectures will include data explorations and live coding.

**Assignments**

Data science is about analyzing real-world data sets, and so a series of projects involving real data are a required part of the course. Students may work with a single partner on each project, and we strongly recommend that you find a partner in your data/code challenge section. Each student must submit each homework independently, but you are allowed to discuss problems with other students.

*Weekly homework:* This will be done within computational notebooks.

*Group Projects:* The first project will be a group project and will require alternating roles between the students between project design, coding and implementation, and review. The goal of the project will be to give students lessons on making their project and code well documented and reproducible.  
*Individual Project:* The final project will be to carry out a reproduction analysis of a published journal article. This may be carried out using the framework.

**Grading**

Weekly Homework (6): 30% - These will be due on the last class day of the week at 5pm a week after they are released.

Data/code Challenge (4): 30% - These will be due on the third-class day of the week at 5pm. The class will work in informal, web-based groups.

Group Project: 20% - This will be due on the last class day of Week 5 at 5pm.

Final Project: 20% - This will be due on the last class day of the last week at 5pm.

**Late Policy**

Students are allowed to submit data/code challenge and homework late for a 50% penalty until the Wednesday after they are due at 5:00 PM, after which they will receive no credit. Every student will have one data/code challenge drop. There is no drop for projects. Every day late will result in a 50 % reduction in grade.

**Disabled Students Policy**

UC Berkeley is committed to creating a learning environment that meets the needs of its diverse student body including students with disabilities. If you anticipate or experience any barriers to learning in this course, please feel welcome to discuss your concerns with me.

If you have a disability, or think you may have a disability, you can work with the Disabled Students' Program (DSP) to request an official accommodation. The Disabled Students' Program (DSP) is the campus office responsible for authorizing disability-related academic accommodations in cooperation with the students themselves and their instructors. You can find more information about DSP, including contact information and the application process here: dsp.berkeley.edu. If you have already been approved for accommodations through DSP, please meet with me so we can develop an implementation plan together.

Students who need academic accommodations or have questions about their accommodations should contact DSP, located at 260 César Chávez Student Center. Students may call 642-0518 (voice), 642-6376 (TTY), or e-mail [dsp@berkelely.edu](mailto:dsp@berkelely.edu) (link sends e-mail).

**Learning outcomes**

An understanding of how data can be used to explore economics research questions and carry out research. Appreciation for the sources of datasets that economists work with and how datasets are created. This course is based on the elements of the data science life cycle; from data exploration to formulating questions and visualization. By looking at a variety of datasets students will be exposed to a broad range of Data Science applications to economic questions. Students will get a combination of skills and tools needed to be a successful research assistant or do independent applied economics work.

**Materials & Resources**

*References will be drawn from the following according to the topics*

**Books**

Bekes, G. and G. Kezdi, 2021. *Data Analysis for Business, Economics, and Policy,* Cambridge University Press.

Cleff, T., 2014. *Exploratory Data Analysis in Business and Economics: An Introduction Using SPSS, Stata, and Excel,* Springer.

Hastie, T., R. Tibshirani, and J. Friedman, 2016. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* Second Ed., Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2021. *An Introduction to Statistical Learning: with Applications in R*, Second Ed., Springer.

Kuhn, M. and K. Johnson, 2013. *Applied Predictive Modeling*, Springer.

Vanderplas, J., 2017. *Python for Data Analysis*, O'Reilly.

**Suggested Readings**

Athey, S. 2021. ‘The Impact of Machine Learning on Economics’, Eds. A. Agrawal, J. Gans, and A. Goldfarb, *The Economics of Artificial Intelligence,* Ch. 21, University of Chicago Press, 507-552.

Breiman, L., 2001. ‘Statistical Modeling: The Two Cultures’, *Statistical Science*, 16(3), 199-231.

Ludwig, J, and S. Mullainathan, 2021. ‘Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System’, *Journal of Economic Perspectives*, 35(4), 71–96.

Mullainathan, S. and J. Spiess, 2017. ‘Machine Learning: An Applied Econometric Approach’, *Journal of Economic Perspectives*, 31(2), 87–106.

Wager, S. and S. Athey, 2019. ‘Machine Learning Methods That Economists Should Know About’, ***Annual Review of Economics,*** 11, 685-725.

**Further Readings**

Chan, F. and L. Matyas, 2022. *Econometrics with Machine Learning*, Springer.

Semenova, V. and V. Chernozhukov**, 2021. ‘Debiased Machine Learning for Conditional Average Treatment Effects and Other Causal Functions’, *The*** *Econometrics Journal*, 24(2), 264-289.

Semenova, V., 2023. ‘Debiased Machine Learning of Set-Identified Linear Models’. *Journal of Econometrics*,

Wager, S. and S Athey, 2018. ‘Estimation and Inference of Heterogeneous Treatment Effects using Random Forests’, *Journal of the American Statistical Association,* 113, 1228-1242.

**SECTION I: INTRODUCTION TO DATA SCIENCE: BASICS**

Technology for Data Science

Python

Common Libraries/Modules

SQL

GitHub

Origins of Data

What is Data?

Data Structures: Structured vs Unstructured

Data Types

Data Quality

Sources of Data and Collecting Data

Data Sampling: Selection Bias vs Random Sampling

Data Science Flow for Reading Data

Big Data, Statistical Inference, and External Validity

Data Wrangling: Preparing Raw Data for Analysis

Types of Variables

Types of Observations

Linking Relational Data with Tables

Organizing Data Tables for a Project

Exploratory Data Analysis

Frequencies and Probabilities (with application in Python)

Visualizing Distributions (with application in Python)

Extreme Values: Missing Values, Outliers, and Sparsity (with application in Python)

Some Graphs for Data Visualization (with application in Python)

EDA for Numerical and Categorical Variables (with application in Python)

Summary Statistics (with application in Python)

SECTION II: DATA ANALYSIS IN DATA SCIENCE: FUNDAMENTALS

Bias -Variance Tradeoff

Overfitting and Underfitting

Splitting Data into Training and Test Sets (with application in Python)

The Curse of Dimensionality

Dimensionality Reduction by Feature Extraction

Dimensionality Reduction by Feature Selection

Principal Component Analysis (PCA)

Feature Generation

Model Selection, Evaluation, and Validation

Regularization, Parameters, and Hyperparameters

Testing Hypotheses

The Logic of Hypothesis Testing

Null vs Alternative

The -test and the -Value

Making a Decision: False Negatives vs False Positives

Testing Hypotheses with Big Data

Comparison and Correlation

The and the

Describing Patterns of Association

Conditioning

Conditional Probabilities

Conditional Distribution and Expectation

Dependence, Covariance, and Correlation

Sources of Variation in

Introduction to Regression

Simple Regression

Regression in Prediction

Regression in Causal Inference

Non-Parametric and Parametric Regression

Assumption vs Approximation in Regression Analysis

Regression Coefficient

OLS

Predicted Values and Residuals

Multiple Regression

Case of Two Regressors

Visual Representation

Many Explanatory Variables

Non-Linearity

Using Qualitative Variables

Generalizing Regression Results

Linear Regression in ML vs Linear Regression in Econometrics (with application in Python)

Time Series Forecasting (with application in Python)

Time Series Specialties

Data Preparation

Aggregation

What is Special in Time Series?

Stationary

Serial Correlation: Order and Sign

Serial Correlation: Magnitude

Non-Stationary: Random Walk

Unit Root Tests

Seasonality in Time Series

Predicting Probabilities

Binary Events

Linear Probability Model (LPM) and its Interpretation

Predicted Values in LPM

Maximum Likelihood Estimation

Predictions for LMP, Logit, and Probit

Logistic Regression in ML vs Logistic Regression in Econometrics (with application in Python)

SECTION III: DATA ANALYSIS IN DATA SCIENCE: ML ALGORITHMS

Supervised and Unsupervised Models

Prediction in ML

Framework for Prediction

Prediction Setup

Predictive Analysis: What is New?

Regression and Prediction

Prediction Error

Decomposing Prediction Error

Interval Prediction for Quantitative Target Variable

Loss Functions

Squared Loss

Adding Up – Mean Squared Error (MSE)

MSE Decomposition: Bias and Variance

Model Selection

External Validity, Avoiding Overfitting, and Model Selection

Comparing Models: Overfit vs Underfit

Finding the Best Model using Best Fit and Penalty: The BIC

Model Fit Evaluation

Finding the Best Model using Training and Test Samples

Finding the Best Model by Cross-Validation

5-Fold Cross-Validation

External Validity and Stable Patterns

Machine Learning and the Role of Algorithms

Machine Learning Algorithms

Regression

Business Question and Defining

Steps of Prediction in ML

Sample Design: Filtering

Label Engineering:

Defining Target

Log vs Level

Feature Engineering

What Features

In What Functional Form

What to Do with Different Types of Variables

Predicting Airbnb Prices (with application in Python)

Model Building

Evaluating the Prediction using a Holdout Set

Selecting Variables in Regressions by RIDGE and LASSO

Penalized/Generalized Regression Models

Regression Trees and Forest

Post-prediction diagnostics (with application in Python)

Variable Importance (with application in Python)

ROC Curve

Classification (with application in Python)

Logistic Regression

Decision Trees

Random Forests

Boosting

Super Vector Machines

K-Nearest Neighbors

Deep Learning and Neural Networks

Clustering (with application in Python)

K-Means

K-Modes

K-Prototypes

Hierarchical Clustering

DBSCAN

Experimental Design, Reinforcement Learning, and Multi-Armed Bandits

SECTION IV: A FRAMEWORK FOR CAUSAL ANALYSIS USING MACHINE LEARNING

Data Science and Machine Learning

Data Science and Economics

Econometrics and ML

vs

Comparison vs Correlation

Association vs Causation

Model Identification vs Algorithmic Data Modeling

Causal questions

Causality, Intervention, and Variation

The Setup: Intervention, Treatment, Subjects, and Outcomes

Causal Question

Potential Outcomes Framework

Individual Treatment Effect

Heterogeneous Treatment Effects

Average Treatment Effect

ATE as Average / Expected ATE

Average Effects in Subgroups and ATET

ATE when Quantitative Causal Variables

Quantitative Causal Variables

ATE and Quantitative Causal Variables

Causal Maps (DAGs) to Uncover Causal Structure

Causal Maps: Simplest Case

DAG: Mechanisms

Random Assignment

Random Assignment and ATE

Random Assignment, ATE, and ATET

Sources of Variation in the Causal Variable

An Exogenous and an Endogenous Source of Variation in

Good and Bad Sources

Experimenting versus Conditioning: 1 Controlled Experiments

Controlled Experimental Variation in

Experimenting versus Conditioning: 2 Natural experiments

Experimenting versus Conditioning: 3 Conditioning

Confounders in Observational Data

Three Types of Confounders

Common Cause Confounder

Mechanism of Reverse Causality

Unwanted Mechanism

Confounders in Practice: Selection

From Latent Variables to Measured Variables

Omitted Variable Bias

The Three Types of Bad Conditioning Variables

**Calendar**

|  |  |  |
| --- | --- | --- |
| **Week** | Topics | Assignments |
| **Week 1: INTRODUCTION TO DATA SCIENCE: BASICS**  **- Technology for Data Science**  **- Origins of Data**  **- Data Structures: Structured vs Unstructured**  **- Collecting Data**  **- Selection Bias vs Random Sampling**  **- Reading Data** | Lecture 1- Introduction to Course,  Overview of Course Goals,  Introduction to Technology for Data Science  Python  Common Libraries/Modules  SQL  GitHub  Lecture 2 - Origins of Data  What is Data?  Data Structures: Structured vs Unstructured  Data Types  Data Quality  Sources of Data and Collecting Data  Data Sampling: Selection Bias vs Random Sampling  Data Science Flow for Reading Data  Big Data, Statistical Inference, and External Validity  Lecture 3 - Data Wrangling: Preparing Raw Data for Analysis  Types of Variables  Types of Observations  Linking Relational Data with Tables  Organizing Data Tables for a Project  Exploratory Data Analysis  Frequencies and Probabilities (with application in Python)  Visualizing Distributions (with application in Python)  Lecture 4 - Extreme Values: Missing Values, Outliers, and Sparsity (with application in Python)  Some Graphs for Data Visualization (with application in Python)  EDA for Numerical and Categorical Variables (with application in Python)  Summary Statistics (with application in Python) | Data/code challenge 1 - Introduction to Notebooks  EDA  Frequencies and Probabilities  Visualizing Distributions  Data/code challenge 2 - Extreme Values: Missing Values, Outliers, and Sparsity  EDA for Numerical and Categorical Variables Summary Statistics |
| **Week 2: SECTION II: DATA ANALYSIS IN DATA SCIENCE: FUNDAMENTALS**  **- Fundamentals of Feature Engineering**  **- Testing Hypotheses**  **- Comparison and Correlation** | Lecture 1 - Bias-Variance Tradeoff  Overfitting and Underfitting  Splitting Data into Training and Test Sets with Application in Python  The Curse of Dimensionality  Dimensionality Reduction by Feature Extraction (with Application in Python)  Lecture 2 - Dimensionality Reduction by Feature Selection (with Application in Python)  Principal Component Analysis (with Application in Python)  Feature Generation and Encoding (with Application in Python)  Lecture 3 - Model Selection, Evaluation, and Validation;  Regularization, Parameters, and Hyperparameters;  Testing Hypotheses  The Logic of Hypothesis Testing  Null vs Alternative  The t-test and the p-Value  Making a Decision: False Negatives vs False Positives  Testing Hypotheses with Big Data  Lecture 4 - Comparison and Correlation  The and the  Describing Patterns of Association  Conditioning  Conditional Probabilities  Conditional Distribution and Expectation  Dependence, Covariance, and Correlation  Sources of Variation in | Data/code challenge 3 - Dimensionality Reduction by Feature Extraction  Dimensionality Reduction by Feature Selection  Data/code challenge 4 -  Principal Component Analysis  Feature Generation and Encoding |
| **Week 3: SECTION II: DATA ANALYSIS IN DATA SCIENCE: FUNDAMENTALS**  **- Introduction to Regression**  **- Multiple Regression**  **- Time Series Forecasting**  **- Predicting Probabilities** | Lecture 1- Introduction to Regression  Simple Regression  Regression in Prediction  Regression in Causal Inference  Non-Parametric and Parametric Regression  Assumption vs Approximation in Regression Analysis  Regression Coefficient  OLS  Predicted Values and Residuals  Lecture 2 - Multiple Regression  Case of Two Regressors  Visual Representation  Many Explanatory Variables  Non-Linearity  Using Qualitative Variables  Generalizing Regression Results  Linear Regression in ML vs Linear Regression in Econometrics (with application in Python)  Lecture 3 - Time Series Forecasting  Time Series Specialties  Data Preparation  Aggregation  What is Special in Time Series?  Stationary  Serial Correlation: Order and Sign  Serial Correlation: Magnitude  Non-Stationary: Random Walk  Unit Root Tests  Seasonality in Time Series  An application in Python  Lecture 4 - Predicting Probabilities  Binary Events  Linear Probability Model (LPM) and its Interpretation  Predicted Values in LPM  Maximum Likelihood Estimation  Predictions for LMP, Logit, and Probit  Logistic Regression in ML vs Logistic Regression in Econometrics (with application in Python) | Project 1 - Selected data project - import, clean, describe  Data/code challenge 5 - Linear Regression in ML vs Linear Regression in Econometrics  Logistic Regression in ML vs Logistic Regression in Econometrics |
| **Week 4: SECTION III: DATA ANALYSIS IN DATA SCIENCE: ML ALGORITHMS - Framework for Prediction in ML**  **- Machine Learning Algorithms**  **- Regression Models in ML** | Lecture 1 - Supervised and Unsupervised Models  Framework for Prediction in ML  Prediction Setup  Predictive Analysis: What is New?  Regression and Prediction  Prediction Error  Decomposing Prediction Error  Interval Prediction for Quantitative Target Variable  Loss Functions  Squared Loss  Adding Up – Mean Squared Error (MSE)  MSE Decomposition: Bias and Variance  Lecture 2 - Model Selection  External Validity, Avoiding Overfitting, and Model Selection  Comparing Models: Overfit vs Underfit  Finding the Best Model using Best Fit and Penalty: The BIC  Model Fit Evaluation  Finding the Best Model using Training and Test Samples  Finding the best model by cross-validation  5-fold cross-validation  External validity and stable patterns  Lecture 3 - Machine Learning and the Role of Algorithms  Machine Learning Algorithms  Regression  Business question and defining  Steps of prediction in ML  Sample Design: Filtering  Label Engineering:  Defining Target  Log vs Level  Feature Engineering  What Features  In What Functional Form  What to Do with Different Types of Variables  Predicting Airbnb Prices (with application in Python)  Lecture 4 - Model Building  Evaluating the Prediction using a Holdout Set  Selecting Variables in Regressions by RIDGE and LASSO  Penalized/Generalized Regression Models  Regression Trees and Forest  Post-prediction diagnostics (with application in Python)  Variable Importance (with application in Python)  ROC Curve | Data/code challenge 6 - Predicting Airbnb Prices  Data/code challenge 7 –  Post-prediction diagnostics  Variable Importance  ROC Curve |
| **Week 5: SECTION III: DATA ANALYSIS IN DATA SCIENCE: ML ALGORITHMS - Classification**  **- Clustering**  **- Reinforcement**  **SECTION IV: A FRAMEWORK FOR CAUSAL ANALYSIS USING MACHINE LEARNING**  **- vs**  **- Potential Outcomes Framework** | Lecture 1 - Classification (with application in Python)  Logistic Regression  Decision Trees  Random Forests  Boosting  Super Vector Machines  K-Nearest Neighbors  Deep Learning and Neural Networks  Lecture 2 - Clustering (with application in Python)  K-Means  K-Modes  K-Prototypes  Hierarchical Clustering  DBSCAN  Experimental Design, Reinforcement Learning, and Multi-Armed Bandits  Lecture 3 – Data Science, and Machine Learning (ML)  Data Science and Economics  Econometrics and ML  vs  Association vs Causation  Model Identification vs Algorithmic Data Modeling  Causal Questions  Measuring Causality Requires Intervention and Variation  The Setup: Intervention, Treatment, Subjects, and Outcomes  The Causal Question Again  Lecture 4 - Potential Outcomes Framework  Individual Treatment Effect  Heterogeneous Treatment Effects  Average Treatment Effect  ATE as Average / Expected ITE  Average Effects in Subgroups and ATET  ATE When Quantitative Causal Variables  Quantitative Causal Variables  ATE and Quantitative Causal Variables  Causal Maps (DAGs) to Uncover Causal Structure  Causal Maps: Simplest Case  DAG: Mechanisms  Random Assignment  Random Assignment and ATE  Random Assignment, ATE, and ATET  Sources of Variation in the Causal Variable  An Exogenous and an Endogenous Source of Variation in  Good and Bad Sources | Data/code challenge 8 - Classification and Clustering Models  Project 3 - Code to Cleaning to Visualization to Outputs in Classification and Clustering Models |
| **Week 6: SECTION IV: A FRAMEWORK FOR CAUSAL ANALYSIS USING MACHINE LEARNING**  **- Controlled Experiments**  **- Natural Experiments**  **- Observational Data**  **- Presentations and Telling a story with Data**  **- Data Management**  **- Reproduction and Reproducibility** | Lecture 1- Experimenting versus Conditioning: 1 Controlled Experiments  Controlled Experimental Variation in  Experimenting versus Conditioning: 2 Natural Experiments  Experimenting versus Conditioning: 3 Conditioning  Confounders in Observational Data  Three Types of Confounders  Common Cause Confounder  Mechanism of Reverse Causality  Unwanted Mechanism  Confounders in Practice: Selection from Latent Variables to Measured Variables  Omitted Variable Bias  The Three Types of Bad Conditioning Variables  Lecture 2 – An Overview: Data Science, Machine Learning, and Causal Inference  Lecture 3 - Presentations and Telling a Story with Data  Lecture 4 - Final | Data/code challenge 9 - Making a Presentation using Data  Group Project -Reproducible Code and Peer Review  Project 4 - Journal Article Reproduction - Key Elements  Individual Project - [Social Science Reproduction Platform Project](https://www.socialsciencereproduction.org/) |