**The Analytics Edge – Dataset and Links**

**Week1:**

**Practice** – WHO.csv

[Global Health Observatory Data Repository](http://apps.who.int/gho/data/node.main)

**Food/Nutrition Recitation** – USDA.csv

[USDA National Nutrient Database for Standard Reference](http://ndb.nal.usda.gov/)

**Assignment – Analytical Detective**: mvtWeek1.csv

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>

* **ID**: a unique identifier for each observation
* **Date**: the date the crime occurred
* **LocationDescription**: the location where the crime occurred
* **Arrest**: whether or not an arrest was made for the crime (TRUE if an arrest was made, and FALSE if an arrest was not made)
* **Domestic**: whether or not the crime was a domestic crime, meaning that it was committed against a family member (TRUE if it was domestic, and FALSE if it was not domestic)
* **Beat**: the area, or "beat" in which the crime occurred. This is the smallest regional division defined by the Chicago police department.
* **District**: the police district in which the crime occured. Each district is composed of many beats, and are defined by the Chicago Police Department.
* **CommunityArea**: the community area in which the crime occurred. Since the 1920s, Chicago has been divided into what are called "community areas", of which there are now 77. The community areas were devised in an attempt to create socially homogeneous regions.
* **Year**: the year in which the crime occurred.
* **Latitude**: the latitude of the location at which the crime occurred.
* **Longitude**: the longitude of the location at which the crime occurred.

**Assignment – Stock Dynamics:**

IBMStock.csv, GEStock.csv, ProcterGambleStock.csv, CocaColaStock.csv, and BoeingStock.csv

<http://www.infochimps.com/>

**Assignment – Employment**: CPSData.csv , MetroAreaCodes.csv , CountryCodes.csv

<http://thedataweb.rm.census.gov/ftp/cps_ftp.html>

* **PeopleInHousehold**: The number of people in the interviewee's household.
* **Region**: The census region where the interviewee lives.
* **State**: The state where the interviewee lives.
* **MetroAreaCode**: A code that identifies the metropolitan area in which the interviewee lives (missing if the interviewee does not live in a metropolitan area). The mapping from codes to names of metropolitan areas is provided in the file [MetroAreaCodes.csv](https://courses.edx.org/assets/courseware/v1/fd88455abc1b5b69112daf70f3bb0c77/asset-v1:MITx+15.071x+1T2020+type@asset+block/MetroAreaCodes.csv).
* **Age**: The age, in years, of the interviewee. 80 represents people aged 80-84, and 85 represents people aged 85 and higher.
* **Married**: The marriage status of the interviewee.
* **Sex**: The sex of the interviewee.
* **Education**: The maximum level of education obtained by the interviewee.
* **Race**: The race of the interviewee.
* **Hispanic**: Whether the interviewee is of Hispanic ethnicity.
* **CountryOfBirthCode**: A code identifying the country of birth of the interviewee. The mapping from codes to names of countries is provided in the file [CountryCodes.csv](https://courses.edx.org/assets/courseware/v1/763710fa6703caea1cf9c708e31e99a3/asset-v1:MITx+15.071x+1T2020+type@asset+block/CountryCodes.csv).
* **Citizenship**: The United States citizenship status of the interviewee.
* **EmploymentStatus**: The status of employment of the interviewee.
* **Industry**: The industry of employment of the interviewee (only available if they are employed).

**Assignment – Internet Privacy poll:** AnonymityPoll.csv

<http://www.pewinternet.org/2013/09/05/anonymity-privacy-and-security-online/>

* **Internet.Use**: A binary variable indicating if the interviewee uses the Internet, at least occasionally (equals 1 if the interviewee uses the Internet, and equals 0 if the interviewee does not use the Internet).
* **Smartphone**: A binary variable indicating if the interviewee has a smartphone (equals 1 if they do have a smartphone, and equals 0 if they don't have a smartphone).
* **Sex**: Male or Female.
* **Age**: Age in years.
* **State**: State of residence of the interviewee.
* **Region**: Census region of the interviewee (Midwest, Northeast, South, or West).
* **Conservativeness**: Self-described level of conservativeness of interviewee, from 1 (very liberal) to 5 (very conservative).
* **Info.On.Internet**: Number of the following items this interviewee believes to be available on the Internet for others to see: (1) Their email address; (2) Their home address; (3) Their home phone number; (4) Their cell phone number; (5) The employer/company they work for; (6) Their political party or political affiliation; (7) Things they've written that have their name on it; (8) A photo of them; (9) A video of them; (10) Which groups or organizations they belong to; and (11) Their birth date.
* **Worry.About.Info**: A binary variable indicating if the interviewee worries about how much information is available about them on the Internet (equals 1 if they worry, and equals 0 if they don't worry).
* **Privacy.Importance**: A score from 0 (privacy is not too important) to 100 (privacy is very important), which combines the degree to which they find privacy important in the following: (1) The websites they browse; (2) Knowledge of the place they are located when they use the Internet; (3) The content and files they download; (4) The times of day they are online; (5) The applications or programs they use; (6) The searches they perform; (7) The content of their email; (8) The people they exchange email with; and (9) The content of their online chats or hangouts with others.
* **Anonymity.Possible**: A binary variable indicating if the interviewee thinks it's possible to use the Internet anonymously, meaning in such a way that online activities can't be traced back to them (equals 1 if he/she believes you can, and equals 0 if he/she believes you can't).
* **Tried.Masking.Identity**: A binary variable indicating if the interviewee has ever tried to mask his/her identity when using the Internet (equals 1 if he/she has tried to mask his/her identity, and equals 0 if he/she has not tried to mask his/her identity).
* **Privacy.Laws.Effective**: A binary variable indicating if the interviewee believes United States law provides reasonable privacy protection for Internet users (equals 1 if he/she believes it does, and equals 0 if he/she believes it doesn't).

**Week2: Linear Regression**

**Wine**: wine.csv , wine\_test.csv

<http://www.liquidasset.com/winedata.html>

**Baseball**: baseball.csv

<http://www.baseball-reference.com/>

**Recitation – NBA:** nba\_train.csv , nba\_test.csv

<http://www.basketball-reference.com/>

**Assignment – Climate Change:** climate\_change.csv

The file [climate\_change.csv](https://courses.edx.org/asset-v1:MITx+15.071x+2T2017+type@asset+block@climate_change.csv) contains climate data from May 1983 to December 2008. The available variables include:

* *Year*: the observation year.
* *Month*: the observation month.
* *Temp*: the difference in degrees Celsius between the average global temperature in that period and a reference value. This data comes from the [Climatic Research Unit at the University of East Anglia](http://www.cru.uea.ac.uk/cru/data/temperature/).
* *CO2*, *N2O*, *CH4*, *CFC.11*, *CFC.12*: atmospheric concentrations of carbon dioxide (CO2), nitrous oxide (N2O), methane  (CH4), trichlorofluoromethane (CCl3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl2F2; commonly referred to as CFC-12), respectively. This data comes from the [ESRL/NOAA Global Monitoring Division](http://www.esrl.noaa.gov/gmd/ccgg/data-products.html).
  + CO2, N2O and CH4 are expressed in ppmv (parts per million by volume  -- i.e., 397 ppmv of CO2 means that CO2 constitutes 397 millionths of the total volume of the atmosphere)
  + CFC.11 and CFC.12 are expressed in ppbv (parts per billion by volume).
* *Aerosols*: the mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun's energy is reflected back into space. This data is from the [Godard Institute for Space Studies at NASA](http://data.giss.nasa.gov/modelforce/strataer/).
* *TSI*: the total solar irradiance (TSI) in W/m2 (the rate at which the sun's energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. This data is from the [SOLARIS-HEPPA project website](http://solarisheppa.geomar.de/solarisheppa/cmip5).
* *MEI*: multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the [El Nino/La Nina-Southern Oscillation](http://en.wikipedia.org/wiki/El_nino) (a weather effect in the Pacific Ocean that affects global temperatures). This data comes from the [ESRL/NOAA Physical Sciences Division](http://www.esrl.noaa.gov/psd/enso/mei/table.html).

**Assignment – Reading Test scores** : pisa2009train.csv , pisa2009test.csv

<http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011038>

**grade:**The grade in school of the student (most 15-year-olds in America are in 10th grade)

**male:**Whether the student is male (1/0)

**raceeth:**The race/ethnicity composite of the student

**preschool:**Whether the student attended preschool (1/0)

**expectBachelors:**Whether the student expects to obtain a bachelor's degree (1/0)

**motherHS:**Whether the student's mother completed high school (1/0)

**motherBachelors:**Whether the student's mother obtained a bachelor's degree (1/0)

**motherWork:**Whether the student's mother has part-time or full-time work (1/0)

**fatherHS:**Whether the student's father completed high school (1/0)

**fatherBachelors:**Whether the student's father obtained a bachelor's degree (1/0)

**fatherWork:**Whether the student's father has part-time or full-time work (1/0)

**selfBornUS:**Whether the student was born in the United States of America (1/0)

**motherBornUS:**Whether the student's mother was born in the United States of America (1/0)

**fatherBornUS:**Whether the student's father was born in the United States of America (1/0)

**englishAtHome:**Whether the student speaks English at home (1/0)

**computerForSchoolwork:**Whether the student has access to a computer for schoolwork (1/0)

**read30MinsADay:**Whether the student reads for pleasure for 30 minutes/day (1/0)

**minutesPerWeekEnglish:**The number of minutes per week the student spend in English class

**studentsInEnglish:**The number of students in this student's English class at school

**schoolHasLibrary:**Whether this student's school has a library (1/0)

**publicSchool:**Whether this student attends a public school (1/0)

**urban:**Whether this student's school is in an urban area (1/0)

**schoolSize:** The number of students in this student's school

**readingScore:**The student's reading score, on a 1000-point scale

**Assignment – Flu** : FluTrain.csv , FluTest.csv

<https://www.google.org/flutrends/us/#US>

The csv file [FluTrain.csv](https://courses.edx.org/assets/courseware/v1/df331a605387ca8382972c88d2853ddf/asset-v1:MITx+15.071x+1T2020+type@asset+block/FluTrain.csv) aggregates this data from January 1, 2004 until December 31, 2011 as follows:

"Week" - The range of dates represented by this observation, in year/month/day format.

"ILI" - This column lists the percentage of ILI-related physician visits for the corresponding week.

"Queries" - This column lists the fraction of queries that are ILI-related for the corresponding week, adjusted to be between 0 and 1 (higher values correspond to more ILI-related search queries).

**Assignment – State :** statedata.csv or in built

* **Population** - the population estimate of the state in 1975
* **Income** - per capita income in 1974
* **Illiteracy** - illiteracy rates in 1970, as a percent of the population
* **Life.Exp** - the life expectancy in years of residents of the state in 1970
* **Murder** - the murder and non-negligent manslaughter rate per 100,000 population in 1976
* **HS.Grad** - percent of high-school graduates in 1970
* **Frost** - the mean number of days with minimum temperature below freezing from 1931–1960 in the capital or a large city of the state
* **Area** - the land area (in square miles) of the state
* **state.abb** - a 2-letter abreviation for each state
* **state.area** - the area of each state, in square miles
* **x** - the longitude of the center of the state
* **y** - the latitude of the center of the state
* **state.division** - the division each state belongs to (New England, Middle Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain, or Pacific)
* **state.name** - the full names of each state
* **state.region** - the region each state belong to (Northeast, South, North Central, or West)

**Assignment – Elantra**: Elantra.csv

* **Month**= the month of the year for the observation (1 = January, 2 = February, 3 = March, ...).
* **Year**= the year of the observation.
* **ElantraSales** = the number of units of the Hyundai Elantra sold in the United States in the given month.
* **Unemployment** = the estimated unemployment percentage in the United States in the given month.
* **Queries** = a (normalized) approximation of the number of Google searches for "hyundai elantra" in the given month.
* **CPI\_energy** = the monthly consumer price index (CPI) for energy for the given month.
* **CPI\_all**= the consumer price index (CPI) for all products for the given month; this is a measure of the magnitude of the prices paid by consumer households for goods and services (e.g., food, clothing, electricity, etc.).

**Week3: Logistic Regression**

**Modelling the expert:** quality.csv

* **MemberID** numbers the patients from 1 to 131, and is just an identifying number.
* **InpatientDays** is the number of inpatient visits, or number of days the person spent in the hospital.
* **ERVisits** is the number of times the patient visited the emergency room.
* **OfficeVisits** is the number of times the patient visited any doctor's office.
* **Narcotics** is the number of prescriptions the patient had for narcotics.
* **DaysSinceLastERVisit** is the number of days between the patient's last emergency room visit and the end of the study period (set to the length of the study period if they never visited the ER).
* **Pain** is the number of visits for which the patient complained about pain.
* **TotalVisits** is the total number of times the patient visited any healthcare provider.
* **ProviderCount** is the number of providers that served the patient.
* **MedicalClaims** is the number of days on which the patient had a medical claim.
* **ClaimLines** is the total number of medical claims.
* **StartedOnCombination** is whether or not the patient was started on a combination of drugs to treat their diabetes (TRUE or FALSE).
* **AcuteDrugGapSmall** is the fraction of acute drugs that were refilled quickly after the prescription ran out.
* **PoorCare** is the outcome or dependent variable, and is equal to 1 if the patient had poor care, and equal to 0 if the patient had good care

**Farmington**: Farmington.csv

https://biolincc.nhlbi.nih.gov/media/teachingstudies/framdoc.pdf

**Recitation - Election Forecast** : PollingData.csv

<http://www.realclearpolitics.com/>

**Assignment Music Records**: songs.csv

Taking an analytics approach, we aim to use information about a song's properties to predict its popularity. The dataset [songs.csv](https://courses.edx.org/assets/courseware/v1/796af11b376799bbe7769dc645564867/asset-v1:MITx+15.071x+1T2020+type@asset+block/songs.csv) consists of all songs which made it to the Top 10 of the Billboard Hot 100 Chart from 1990-2010 plus a sample of additional songs that didn't make the Top 10. This data comes from three sources: [Wikipedia](http://en.wikipedia.org/wiki/Billboard_Hot_100), [Billboard.com](http://www.billboard.com/), and [EchoNest](http://echonest.com/).

* **year** = the year the song was released
* **songtitle** = the title of the song
* **artistname** = the name of the artist of the song
* **songID** and **artistID** = identifying variables for the song and artist
* **timesignature** and **timesignature\_confidence** = a variable estimating the time signature of the song, and the confidence in the estimate
* **loudness** = a continuous variable indicating the average amplitude of the audio in decibels
* **tempo** and **tempo\_confidence** = a variable indicating the estimated beats per minute of the song, and the confidence in the estimate
* **key** and **key\_confidence** = a variable with twelve levels indicating the estimated key of the song (C, C#, . . ., B), and the confidence in the estimate
* **energy** = a variable that represents the overall acoustic energy of the song, using a mix of features such as loudness
* **pitch** = a continuous variable that indicates the pitch of the song
* **timbre\_0\_min**, **timbre\_0\_max**, **timbre\_1\_min**, **timbre\_1\_max**, . . . , **timbre\_11\_min**, and **timbre\_11\_max** = variables that indicate the minimum/maximum values over all segments for each of the twelve values in the timbre vector (resulting in 24 continuous variables)
* **Top10** = a binary variable indicating whether or not the song made it to the Top 10 of the Billboard Hot 100 Chart (1 if it was in the top 10, and 0 if it was not)

**Assignment – Parole**: parole.csv

<http://www.icpsr.umich.edu/icpsrweb/NACJD/series/38/studies/26521?archive=NACJD&sortBy=7>

* **male**: 1 if the parolee is male, 0 if female
* **race**: 1 if the parolee is white, 2 otherwise
* **age**: the parolee's age (in years) when he or she was released from prison
* **state**: a code for the parolee's state. 2 is Kentucky, 3 is Louisiana, 4 is Virginia, and 1 is any other state. The three states were selected due to having a high representation in the dataset.
* **time.served**: the number of months the parolee served in prison (limited by the inclusion criteria to not exceed 6 months).
* **max.sentence**: the maximum sentence length for all charges, in months (limited by the inclusion criteria to not exceed 18 months).
* **multiple.offenses**: 1 if the parolee was incarcerated for multiple offenses, 0 otherwise.
* **crime**: a code for the parolee's main crime leading to incarceration. 2 is larceny, 3 is drug-related crime, 4 is driving-related crime, and 1 is any other crime.
* **violator**: 1 if the parolee violated the parole, and 0 if the parolee completed the parole without violation.

**Assignment – Loan** : loans.csv

https://www.lendingclub.com/info/download-data.action

* **credit.policy**: 1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.
* **purpose**: The purpose of the loan (takes values "credit\_card", "debt\_consolidation", "educational", "major\_purchase", "small\_business", and "all\_other").
* **int.rate**: The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.
* **installment**: The monthly installments ($) owed by the borrower if the loan is funded.
* **log.annual.inc**: The natural log of the self-reported annual income of the borrower.
* **dti**: The debt-to-income ratio of the borrower (amount of debt divided by annual income).
* **fico**: The FICO credit score of the borrower.
* **days.with.cr.line**: The number of days the borrower has had a credit line.
* **revol.bal**: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
* **revol.util**: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
* **inq.last.6mths**: The borrower's number of inquiries by creditors in the last 6 months.
* **delinq.2yrs**: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
* **pub.rec**: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

**Assignment – Baseball** : baseball.csv

<http://www.baseball-reference.com/>

* **Team**: A code for the name of the team
* **League**: The Major League Baseball league the team belongs to, either AL (American League) or NL (National League)
* **Year**: The year of the corresponding record
* **RS**: The number of runs scored by the team in that year
* **RA**: The number of runs allowed by the team in that year
* **W**: The number of regular season wins by the team in that year
* **OBP**: The on-base percentage of the team in that year
* **SLG**: The slugging percentage of the team in that year
* **BA**: The batting average of the team in that year
* **Playoffs**: Whether the team made the playoffs in that year (1 for yes, 0 for no)
* **RankSeason**: Among the playoff teams in that year, the ranking of their regular season records (1 is best)
* **RankPlayoffs**: Among the playoff teams in that year, how well they fared in the playoffs. The team winning the World Series gets a RankPlayoffs of 1.
* **G**: The number of games a team played in that year
* **OOBP**: The team's opponents' on-base percentage in that year
* **OSLG**: The team's opponents' slugging percentage in that year

**Week4: Trees**

**Supreme Court** – stevens.csv

http://wusct.wustl.edu/data.php

**D2Hawkeye** – ClaimsData.csv

<http://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html>

**Recitation** – Housing: boston.csv

<http://archive.ics.uci.edu/ml/datasets/Housing>

**Assignment – Vote**: gerber.csv

The researchers grouped about 344,000 voters into different groups randomly - about 191,000 voters were a "control" group, and the rest were categorized into one of four "treatment" groups. These five groups correspond to five binary variables in the dataset.

1. "Civic Duty" (variable **civicduty**) group members were sent a letter that simply said "DO YOUR CIVIC DUTY - VOTE!"
2. "Hawthorne Effect" (variable **hawthorne**) group members were sent a letter that had the "Civic Duty" message plus the additional message "YOU ARE BEING STUDIED" and they were informed that their voting behavior would be examined by means of public records.
3. "Self" (variable **self**) group members received the "Civic Duty" message as well as the recent voting record of everyone in that household and a message stating that another message would be sent after the election with updated records.
4. "Neighbors" (variable **neighbors**) group members were given the same message as that for the "Self" group, except the message not only had the household voting records but also that of neighbors - maximizing social pressure.
5. "Control" (variable **control**) group members were not sent anything, and represented the typical voting situation.

Additional variables include **sex**(0 for male, 1 for female), **yob** (year of birth), and the dependent variable **voting** (1 if they voted, 0 otherwise).

**Assignment – Letter Recognition**: letter\_ABPR.csv

http://archive.ics.uci.edu/ml/datasets/Letter+Recognition

* *letter =* the letter that the image corresponds to (A, B, P or R)
* *xbox =* the horizontal position of where the smallest box covering the letter shape begins.
* *ybox =* the vertical position of where the smallest box covering the letter shape begins.
* *width =* the width of this smallest box.
* *height =* the height of this smallest box.
* *onpix =* the total number of "on" pixels in the character image
* *xbar =* the mean horizontal position of all of the "on" pixels
* *ybar =* the mean vertical position of all of the "on" pixels
* *x2bar =* the mean squared horizontal position of all of the "on" pixels in the image
* *y2bar =* the mean squared vertical position of all of the "on" pixels in the image
* *xybar =* the mean of the product of the horizontal and vertical position of all of the "on" pixels in the image
* *x2ybar =* the mean of the product of the squared horizontal position and the vertical position of all of the "on" pixels
* *xy2bar =* the mean of the product of the horizontal position and the squared vertical position of all of the "on" pixels
* *xedge =* the mean number of edges (the number of times an "off" pixel is followed by an "on" pixel, or the image boundary is hit) as the image is scanned from left to right, along the whole vertical length of the image
* *xedgeycor =* the mean of the product of the number of horizontal edges at each vertical position and the vertical position
* *yedge =* the mean number of edges as the images is scanned from top to bottom, along the whole horizontal length of the image
* *yedgexcor =* the mean of the product of the number of vertical edges at each horizontal position and the horizontal position

**Assignment – Census**: census.csv

http://archive.ics.uci.edu/ml/datasets/Adult

* *age =* the age of the individual in years
* *workclass =* the classification of the individual's working status (does the person work for the federal government, work for the local government, work without pay, and so on)
* *education =* the level of education of the individual (e.g., 5th-6th grade, high school graduate, PhD, so on)
* *maritalstatus =* the marital status of the individual
* *occupation =* the type of work the individual does (e.g., administrative/clerical work, farming/fishing, sales and so on)
* *relationship =* relationship of individual to his/her household
* *race =* the individual's race
* *sex =* the individual's sex
* *capitalgain =* the capital gains of the individual in 1994 (from selling an asset such as a stock or bond for more than the original purchase price)
* *capitalloss =* the capital losses of the individual in 1994 (from selling an asset such as a stock or bond for less than the original purchase price)
* *hoursperweek =* the number of hours the individual works per week
* *nativecountry =* the native country of the individual
* *over50k =* whether or not the individual earned more than $50,000 in 1994

**Assignment – State**: state.csv

Same as Week2 state data

**Week5: Text Analytics**