## [CAP4611-21Spring](https://webcourses.ucf.edu/courses/1369384/calendar_events/2158980)

# Variant 1

# Day 3 (Tuesday, Jan 19):

Before lecture stuff:

Reminder:

Datacamp tutorials not required, but are just there to help you understand the software and libraries behind it.

The only thing required are:

Lecture notes and course content

Quizzes:

Going to be at end of the week when we actually talk about the material

…

Talk about programming languages

…

4:38 pm

…

More talk about programming languages

…

Webcourses:

* TAs have office hours
* They are there to help you
* In terms of your grade, datacamp is not strictly required
* For your assignments, you’ll need to write python code utilizing the modules discussed in Datacamp
* No test questions on datacamp, its free access, just use your email to register
* Updates:
  + Added recommended DataCamp courses

Schedule:

* Today you should know what Git is and how it works
* I strongly suggest you now read up on kaggle, pandas, python
* This week we’re talking about data,

Homework

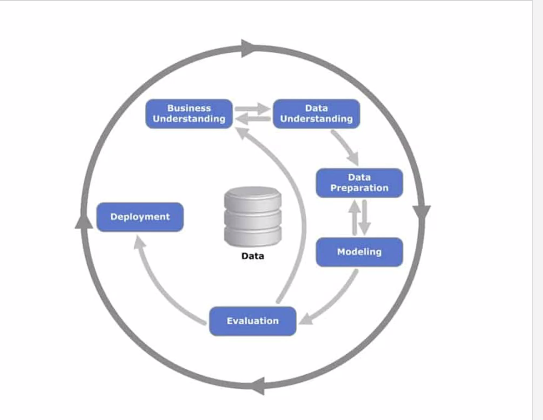
* After talk about decision trees and similarity based learning
  + You’ll get your first assignment
  + You’ll have to submit your assignment through kaggle, learn to manipulate, utilize machine learning etc etc
* Assignments must be turned in as a pair
* If you don’t have a partner a partner will be assigned to you

Schedule:

* Pretty standard stuff, another homework (?) and then test
* He mentioned homework at the “Evaluation” part of the schedule (2/8/21)

Different types of machine learning software:

* Premade ready to go systems
* So, there are a number of packages that exist where you can just drop your data and don’t have to write code for Machine Learning
* One of them is KNIME, which is free with some limitations
* KNIME is a free version of RapidMiner
* RapidMiner can be used free for 30 days,
* WEKA written in Java, has GUI to manage training
* **Scikit-learn** a fairly expansive ML toolkit, this is what we will be using for the majority of our networks
* **TensorFlow** what we will be using for our neural networks.
  + We will get into this when we talk about neural networks
* If you want to use R, there is a package called Carrot, that allows you to use ML with R.
  + It is a little bit lacking, but is possible to do with machine learning stuff…
* If you are a masochist, and you want to use C++, its difficult to find libraries for data that is useful.
  + There is mlpack
  + Software for NVIDIA, has a bunch of libraries written in C (?)
* You can do machine learning in C, R, Java, Python, etc etc
* There is a library called Julia
  + A typed python version wannabe (?)
* The reason python is popular is because it has a massive 3rd party module ecosystem
  + Libraries/modules for practically anything
  + If you can think it, there’s a module for that



Lets talk for a moment, about how do we convert data into decisions:

* So you remember from last time we talked about this idea called the CRISP-DM cycle.
  + In the CRISP-DM cycle, its basically just a software dev cycle for data science
* Side note: if you look at course material, you’ll see Dr. H’s notes, they’ll be done “after the fact”
* Remember we first have to understand the problem and the context around it
* A business problem is something higher in the layer of extraction
* Besides having a fraud investigation team, the insurance company is still having problems
  + Is there a way in which we can make detecting fraud easier?
  + Yes, Machine learning
* What are the goals, the problems and what business is trying to achieve
* This usually requires and understanding of what the business does
* If what you do does not help the business in terms of profit, then its not helpful (?) < It is not helpful in that it will not happen
* Once we have the ideas in terms of machine learning and how it can solve the business problems.
  + Evaluate for feasibility.
  + Is there a way to collect data?
  + What capabilities and resources does the company have
    - Do they have the processing power?
    - The servers to run it?
  + Is the data able to be legally obtained?
    - Europe requires explicit permission to collect some stuff
    - If you happen to circumvent constraints, you can be fined, terminated etc etc
* After finding the problem and planning out a solution,
* You need a (dataframe ?) a way to store your features
  + Columns are the features, rows are the entries (?)
  + One thing that you need to think about when designing this table and thinking about the data you need is the amount of effort in this process
  + Think of the analytical table as a centralized process
    - The data may come from 10 different sources
    - They will need to all be formatted in the same way,
    - Data cleaning/Data rangling describes what this process
  + After forming the rows and columns of data table:
    - Which of these columns are actually useful?
      * Some might be useful
      * Some might need to be combined or broken down
      * Think about temporal (?) constraints
      * You want to look at data before and after events
        + If you don’t have that, you can’t make predictions
      * In terms of the data itself, it can be
        + Numeric
        + Interval
        + Ordinal - nonZero
        + Categorical
        + Binary
        + Text
      * The different types of data should be familiar to stat students
      * There are different distributions for different types of data
  + Recap: Business, solutions, gather data, deploy,
  + Modelling is the fun part, data preparation is a necessary evil
  + Data understanding
    - Talking to people about the data and looking at the data until you understand the data and what you need
    - There are some software packages and special purpose tools for cleaning the data
  + Once you have the data and its cleaned up in a database, you can write the model
    - Use some software, modeling tools, etc etc
  + Once the model is created deployment is a whole other issue
  + Deployment
    - If you’re not in it for the money, go to a non-profit with your machine and you might get some money/position
  + Machine learning is used in various fields:
    - Where to put windmills
    - Medical field
    - …
  + Tumor detection with computer vision and ML (?)
    - MRI/XRAY picture passed to machine to see if it can deduce if tumor is inert/bad

Switching over to notebook:

Q: What is a deep learning network?

We’ll get to that, they aren’t new, just rediscovered

Q: Video game AI

Machine learning for enemy characters, (eventually the player will stop winning)

If you have a first person shooter game like HL, Halo, Control, and the ai are generated using simulated player data, the game will eventually not be fun

A few games have done this like FEAR (?)

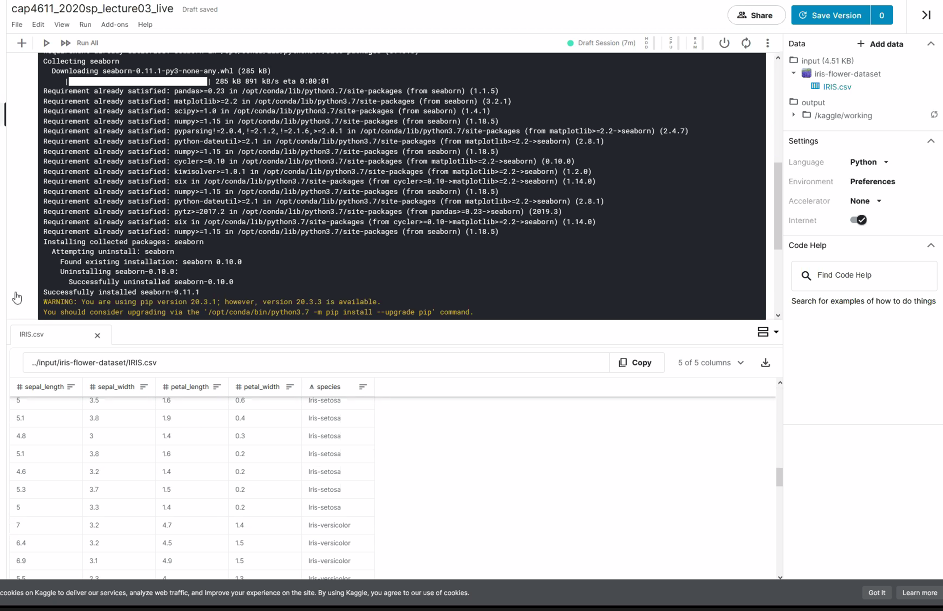
The problem with ML in videogames is that it will kill the FPS

Kaggle Notebook:

So, this notion of exploratory data analysis:

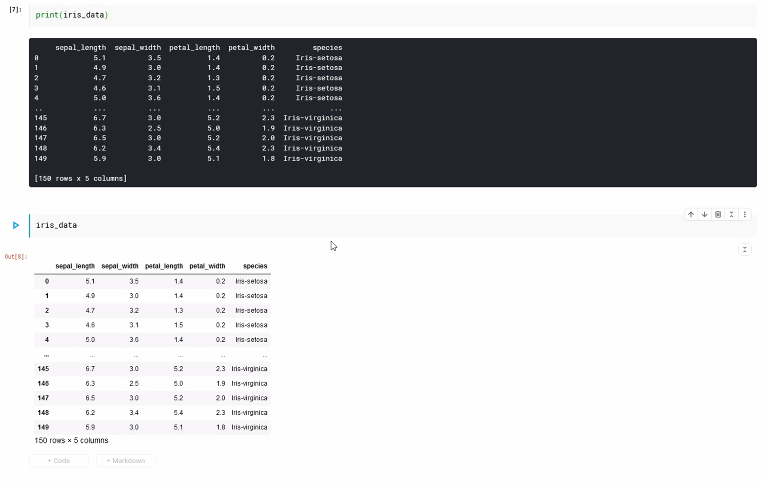
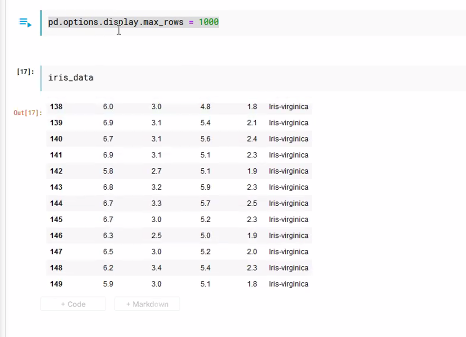
* Once you actually have your data,
* You would want to look in it and see what is in there
* What you want to find is some pattern or information you can gleam from the data.
* Kaggle is pretty good about updating libraries, but not yet with seaborn (used for plot generation)
* In addition to seaborn, you’ll notice by default kaggle imports numpy and pandas.
* Numpy is a linear algebra library, used for a lot of things
* Pandas is used for a lot of data collection, relies on the pandas.DataFrame() object

Getting data onto Kaggle:

* Import from online datasets on Kaggle or import your own
* In the online datasets...\
* SQLlite
  + A lightweight SQL database, really popular
  + But, SQL lite requires some different functions
* CSV = Comma Separated Values
* We import the data as a .CSV, it will appear in a dataset folder
* If we click on the CSV folder,
  + Down at the bottom, it will show a separate window that shows the data.
  + Inside the kaggle notebook environment, you can easily view the data within teh dataset
  + 
  + (Bottom left portion of screen)(
  + However, we are CS students, we will like the code to do this for us,
* /Kaggle is our root directory
  + Our datafiles are stored within the input directory

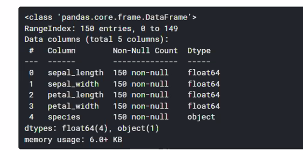
Coding stuff:

Iris\_data = pd.read\_csv(‘SOME.CSV’)

* Inside of a notebook environment, if we just refer to
  + Iris\_data
    - Prints the table
  + print(iris\_data)
    - Prints data in console
  + 
  + Inside a notebook environment, just type the variable and it will print
  + … (?) Explaining how to print out datasets using pandas
  + Notice that the data is truncated.
    - There are options to override it or trick it depending on the notebook environment
  + If you are not in a notebook setting:
    - Pd.options.display.max\_rows = someLargeValue
    - Iris\_data
    - 

What are some ways you can look at the data?

iris\_data.info()

* Displays the dataframe object
* Shows the number of rows
* Shows the number of columns
* 
* This is useful for when you have a dataset that has 100 different columns, its better just to see a summary
* **Every Column only contains one datatype, they are basically numpy arrays**
* So generally, numerical values are stored as floating points in numpy.
* We can see the summary of the columns,
* We might also want to see…

Statistical information of the data

iris\_data.describe()

* 
* Shows you various statistical elements of the data
* 25% == 25% of data is below and 75% of data is above
* **This function only returns information on the numerical columns**

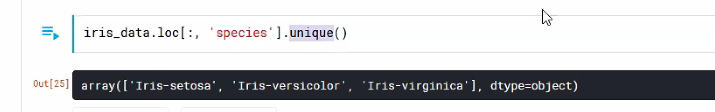
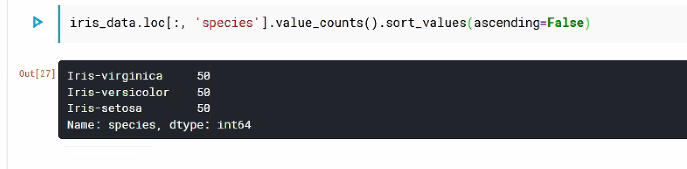
Statistical information about the categorical data:

Iris\_data.loc[:, ‘THING’].describe()

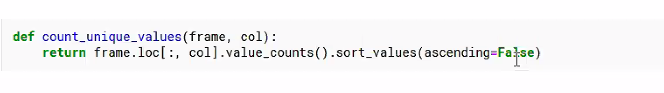
* [:, ‘THING’]
  + **The colon basically means “give me every row”**
  + The string is what to select
* 

It says there are 3 unique values, but how could we find the other values?

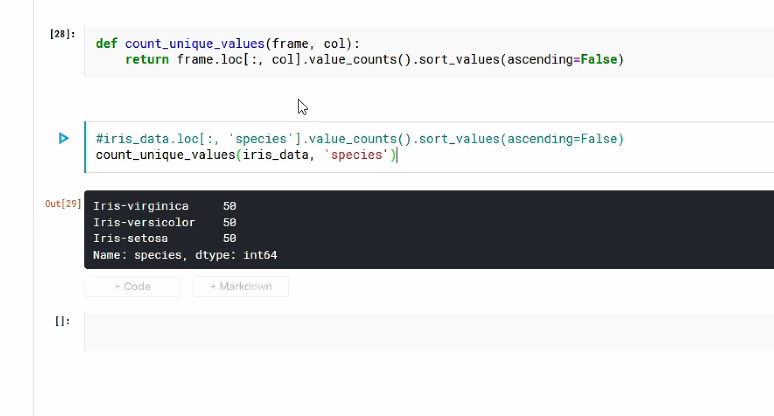
Iris\_data.loc[:, ‘species’].value\_counts()

* 
* We may want to sort them,
* Iris\_data.loc[:, ‘species’].value\_counts().sort\_values(ascending=False)
* 

Lets say we want a function that counts the unique values for a column:



**We first execute the cell that contains the function we defined,** and then call the function on some data



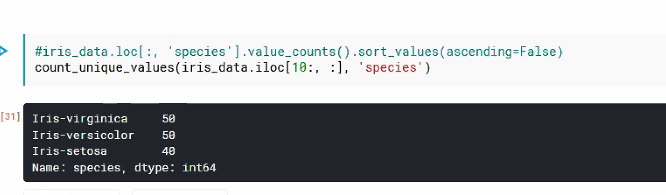
We can also chop off the first 140 rows (0-139) of the array:

count\_unique\_values(iris\_data.iloc[:140, :], ‘species’)

[start:Cut, :]

Start = first index

Cut = last index



If we want to find the actual percentages of data?

Divide by the total count.

For people coming from Java:

For species in iris\_data.loc[:, ‘species’].unique():

print(species / 150)

Note that for loops and numpy are slow, so if we want to find percentages, we would want to avoid for loops. To avoid writing for loops for finding the percentages,

* We take a frame
* We pull out the data we are interested in, and divide it by the count
* 
* We can test stuff things within the notebook as we go along, (its kinda like unit testing but not)
* Unit testing would be getting some software to test all your functions to make sure they have the correct output

How many elements are in each category and what is that as a percentage?

Is there an easy way to look at the data?

Lets make a chart:

* Importing the seaborn library as sns:
  + sns.countplot(data=DataFrame, x=’column’)
  + “It just works” -- Todd Howard

Note in Kaggle, you might need to do “magic” commands

* After importing
  + $mathlib (?)

At this point, we have a dataset, we can do a basic analysis of the categorical data, see what kind of data is occurring and whats be.

Tuesday of next week quiz first 15 minutes. Don’t need to be in Zoom to take it.

No official homework, just go through the tutorials for Python, etc etc

In terms of what you should know:

Everything we talk about in class.

The quiz is going to be abstract ideas,

Example:

What is the mean?

What is the median?

What is a scatterplot?

How would you make a graph using this…

Not super indepth, just there to see if your following along, should occur every Tuesday and discussed about the week prior.

Recording stopped

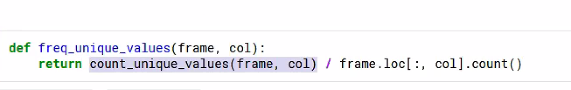
Questions/Things 2 note:

In Python, you’ll notice that things are stringed together:

Return thing.loc().vis.(lime).blablabla.blaba()

* This breaks some law because we are assuming return types

The unique values function



Divides a series of values by a single value, that returns a series of values that contain the percentages

(Better explained through the tutorials linked through webcourses)

Question:

“Can you elaborate on why we shouldn’t use for loops and instead use vectors?”

* So when you look at the documentation for the numpy library you’ll see something called vectorization.
  + Vectorization hides the for loops
  + Since numpy is written in C, there are some neat tricks you can do to optimize this
  + If you have a vector (array) in numpy, it will do something called Broadcast
    - Broadcast creates a copy, (???) somethin somethin somethin, returns the right thing
  + Think about numpy operations as vector operations,
  + Inside of Python ,there is something called a list comprehension:
    - A list comprehension is basically:
      * [x for x in range(10)]
      * -> [0,1,2,3,4,5,6,7,8,9]
  + **%%time shows you how long an execution takes**
  + 
  + You want to try avoid using for loops when your writing Python / Numpy
  + You want to think as things as vectors.
  + If you’ve taken linear algebra, you can add/subtract vectors and do a lot of different things on them
  + Tutorials will talk about this

Note:

* Use the build-in functions when you can, they are usually the most optimized
* I would suggest writing code that makes sense to you, then worry about optimizing

Syntax Question:

* 
* This is not a for loop.
* This line converts it directly into C code
* ByteCode
  + Java ecosystem is ontop of a VM
  + No assembly, it uses a high level version called bytecode
* Python takes the python code and turns it into ByteCode format
  + Is slow
  + But this will turn it into C code (circumventing ByteCode ?)
* That’s the difference

The main thing about Kaggle:

* Get used to using it and handling the environment

Data Scraping:

* So when you’re dealing with real data, it is often missing values, not formatted correctly, might need to derive data etc etc

GTG someone take over ]]]

GG man, it’s mostly personal student stuff from here out. Someone asked if they could ever send the professor their own python code from a side project and if he could look over it and he said if he had the time he would