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

# Variant 1

# Day 5 (Tuesday, Jan 26):

Recap:

* Left off on data analysis, going to go over more of that today+3
* Remember, features are the columns of the dataset
* Species are the target vector, the terminology surrounding ML is kinda loose

Quiz (This Thursday)

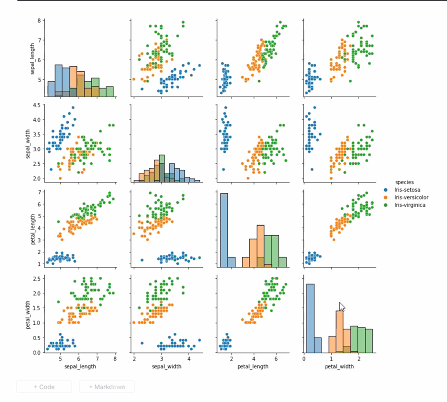
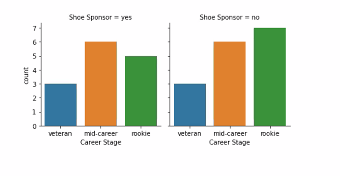
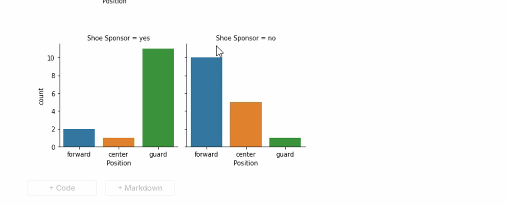
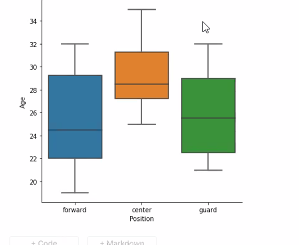
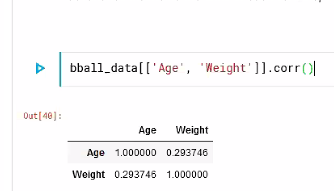
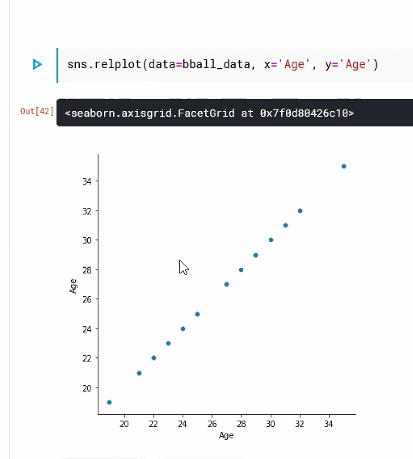
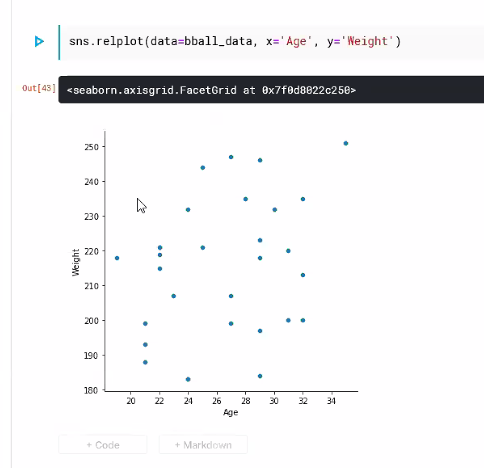
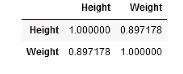
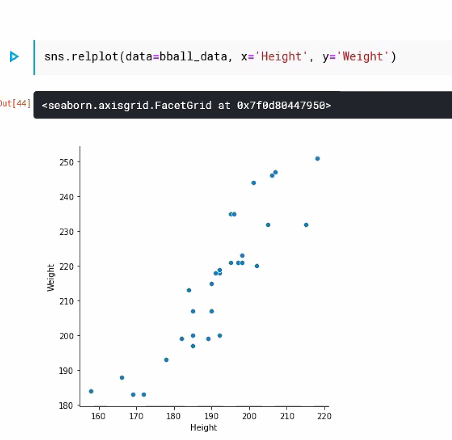
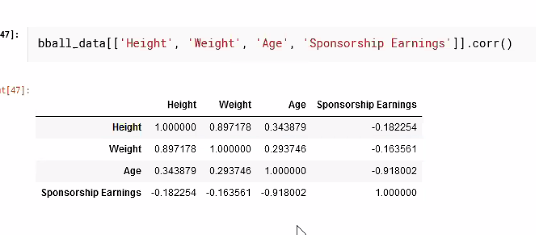
* Quiz Thursday, main focus is everything on exploratory data analysis, and what we’re talking about tonight, 15 minutes 10 questions. Answers are X w z e ?
* To study, look over the keggle notebooks and stuff on webcourses.
* No moderating software required
* There are a couple python questions in the context of Python examples.
* Test will not be like quizzes
* If you can understand everything in the notebook, you should be fine,
* Most answers will be on the first page of [notes?]

First assignment:

* 2 weeks from now, will cover decision trees (?) and stuff we will talk about after this week
* After similarity based learning
* Thinking turning homework into kaggle competitions,

Recording starts:

Recap (Continued):

* We left off talking about exploratory data analysis
* For knowledge of Python and Pandas, consult links on webcourses
* Starting after then end of this week, documentation on future stuff will be your lifeline
* If you remember, exploratory data analysis is getting a feel for the data before building the ML library
* Before doing anything on Kaggle, make sure you import the libraries you need (Pandas, Numpy)
* First thing you want to do is import your data
* In kaggle:
  + pd.read\_csv(“file”) [?]
* The idea is that from this status you are considering you want to find some notion of data quality. You can see the mean, median, mode, categorical data, count, % missing, unique values
* Account for distributions within the data. This will allow you to start making assumptions about the data
* dataframe.info() gives you information about the data in general, it gives you a quick view of what columns may have missing data
* dataframe.describe() shows various information about the columns of the data.
* A normal distribution, anything +- 3 std, we could (?) find outliers based on the standard deviation
* We can find the mode of the data…
* If we don’t like numbers and like pictures, we can use seaborn to make graphs of everything
* Regarding continuous values, there are functions for those
* [side-note] when I made a kaggle notebook yesterday, seaborn was up 2 date, but sometimes this is not the case, make sure to keep note of the library versions as updating it might break your code…
* If something isn’t working with kaggle, first thing you should try is reboot
* One of the very important things that we do in the exploratory data analysis process is using histograms
  + We can notice skewed data and find outliers.
  + There are different distributions, uniform, normal, exponential
  + There is a way you can write these distributions in a general form
  + Normally, as a practitioner, you’ll be the one looking at the data
* There are various tools for finding statistical information about the data. However, it is advisable that you look at the graphs and tables for the data, as some code may have a bug etc.
* Once we get a general notion of the shape and behavior, we look to see outliers and missing data.
  + Note that not all outliers are bad data, you’ll just have to figure out how to deal with them
* We left off right on Advanced data exploration (more tools to use on the data)
* Scatterplots
  + Used to find relationships within the data
  + Seaborn allows you to easily categorize ? your data
  + One thing that you want to do when your doing plotting, is to make sure you give some differential marking upon the data to see relations
  + When seaborn updated, they changed the names of some of the methods and things used within seaborn to make it not backwards compatible
  + The thing to note, is that when we pass in x and y values, we also need to pass in the name of the dataframe.
  + Example
    - plt.relplot(x=iris\_data[‘sepal\_width’], y=’pedal\_length’, hue=’species’)
    - (?)
    - With the hue, we can color the dots based on the categorical data within a specified column
  + A scatterplot by itself, gives you a notion of how two variables might be related to each other.
  + If we look at a mulipanel scatterplot
* MultiPanel scatterplot
  + Creates a bunch of scatterplots comparing every variable to every other variable
  + 
  + [Bottom right most graph] we can see there is a differentiation regarding the species and petal width
  + (?)...
  + If we go out and gather new data, we should be able to compare the new data with the scatterplots and be able to tell whether some plant is part of some species or not due to having close measurements.
  + The flower problem is a classification problem,
    - You have a number of categorical values (vectors)
    - You are trying to figure out given some input data, what is that category?
    - When we will cover probability, we will see that what we really want to know is:
      * What is the probability given some class has some output?
      * What is the probability that these inputs will correspond to this class?
    - In the case of the iris data,
      * What is the probability that the plant I have is the (?)
  + “Accuracy is a bad way to judge the success of your algorithm”
  + If you manage to create an algorithm thats 100% right on test data and real world data, you can make a lot of money
  + A side from histograms down the middle, looking at hte individual scatter plots
  + We can see some diagonal shapes:
    - Diagonal shapes usually indicate a direct/indirect (?) correlation between two values
  + Examining the data in terms of histograms and scatterplots, we can see…
    - If two variables directly change, we can eliminate one of them as the one will be a predictor for the other
  + [Sidenote] Textbook’s website has a bunch of datasets you can mess around with
  + Baseball data:
    - Given info about a player, can we see if they have a sponsor or what career stage they are in? (Categorical)
    - Can we figure out if (?) [Regression]
    - Career question:
      * We might want to investigate whether experienced players get sponsors more than rookie players.
      * Looking at the data at its face value, that will be hard to tell
      * What we can do is split the data using bar graphs
      * 
      * And you can see that there isn’t much of a relation
  + “Machine learning is basically applied statistics”
  + At the end of the day, ML is statistics with algorithms,
  + One of the things you’ll notice is that statistics don’t really deal with large amounts of data
  + 1000 is large in the statistics world
  + 1000 is nothing in the ML world
  + When you start dealing with 100,000,000,000 rows of data, you really need to think about the algorithms that will use that data.
  + We may need large scale hardware to tackle these problems
* Relationships:
  + You can take a bar chart with seaborn and do what is called an R factoring, splitting the data on a particular value
  + Taking your dataset, you can take subsets and see whether they differ based on some variable
    - Baseball example:
      * We see sponsor and career stage doesn’t correlate much
      * However, if we look at positions and sponsorships:
      * 
      * We can see that the positions correlate with shoe sponsorships
      * Additionally Age:
      * [insert picture here]
      * Not much to go off of here
  + Box plot:
    - 
    - Most of the data is found in the green/orange/blue area.
    - The line in the middle is usually the mean,
    - The horizontal lines (whiskers) outside of the colored box represents (?)
    - Occasionally, you will see diamonds, those represent outliers
    - With a box plot, we can compare different variables and see if they are truely different statistically.
  + Numerical representations
    - Are arguably better
    - Charts can be deceiving
    - **The way we define numerically the relationship between two variables is to look at covariance**
    - The formula for covariance:
      * 
      * So i in this formula, is the instance position (the row)
      * Ai is whatever variable A happens to represent at that row
      * Bi is whatever variable B happens to represent in that row as well
      * We do this process for all the variables in the A and B columns
    - [wrong] np.cov(bball\_data[‘Age’], bball\_data[‘Weight’])
    - bball\_data([‘Age’, ‘Weight’]).cov()
    - Shows you the covariance, but what is more powerful is to use
    - .corr()
    - 
    - This normalizes the data and allows us to better interpret the data.
    - Keep in mind that correlation by default, has a linear correspondence.
    - ….
    - If we actually plot this
    - sns.replot(data=bballl\_data, x=’Age’, y=’Weight’)
    - 
    - Perfect correlation (45 deg slope)
    - If we factor in weight..
    - 
    - Its a cloud, you can draw a line that doesn’t go straight vertical or horizontal, but it will have a low correlation (?)
    - The whole purpose of correlation is to tell you that there could be a relationship between two variables
    - If we look at height vs weight
    - 
    - 
    - All these points exist upon a line,
    - Its not a perfectly straight line, but we can fit it with some slope
  + Correlation with more than two variables:
    - [Note] when you are creating your note, try to avoid putting spaces in your data as those spaces might cause problems
    - Think as correlation as slope of a linear line. **Not all data is linearly correlated**
    - Any statistic is measurement that you make from data.
    - We can see from this:
    - 
    - There are negative and positive correlation
    - It seems that as Age increases, sponsorship seems to decrease
    - There are a lot of variables that use covariance instead of correlation
    - There are ways to correlate categorical method, (Consult your statistics book)
  + Question: personalized ads, can you tailor personal ads based on what they do online?
    - Yes, they build a profile of you based on the information that you give them to target you with ads
    - The sad thing about ML is that most of the work done with it goes toward marketing
  + Normalization is incredibly important for models as well as the linear regression
  + You can create bins with the data, which has a range of usefulness
  + The size of the bins can greatly impact the distribution of the data
  + (Best to maintain teh original distribution of the data)
  + Sampling we will talk about more later on
  + **The way we trim data can have massive consequences**
  + Sampling:
    - Top sampling- select flat % from top of data
    - Random sampling- select rows randomly from dataset
    - Stratified sampling - select x amount from each category
    - Under sampling -
    - Over sampling -
  + CAP4611\_lecture\_03 notebook pertains pretty much to the quiz
  + Remember:
    - Covariance is the measurement of the linear relationship between two variables, as one variable changes, the other one changes as well (as opposed to one changing and the other doing nothing)
    - Normalizing is to rescale the values between 0 and 1.
  + [GTG someone take over]
  + Make sure you understanding everything in the notebook
    - How do you deal with outlier?
    - Why do you use boxplots?
    - What does it mean to be correlated?
    - What is correlation?
    - What is bining?
    - etc...