Sian’s questions from class – I will attempt to answer any questions from class here that I needed to look into afterwards

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Lesson | Your question (as I understand it!) | Response including useful resources |
| 20-10-20 | Intro to Pandas | *If we are bringing two files to our data frame, with the same headers but they are in different orders, what will happen?* | My assumption on this would be that pandas/python would do the heavy lifting of matching column names for us… that has proven to be true in a couple of different variations I have tried. So as long as the column headers are the same strings, ie not lowercase /uppercase mismatched, or typos, the concat will stack the data under the appropriate columns regardless of the incoming order. If there are some new columns – eg in one file, not in another, those are added at the end of the column list. Which is quite logical too. So, no worries there  The other thing I noticed is that the order of the columns on the output is obedient to the order of the columns in the first file listed in the concat cell.  Here is a really nice tip for working with frames with lots of columns – reorder the ones that matter! https://stackoverflow.com/questions/41968732/set-order-of-columns-in-pandas-dataframe |
| 20-10-20 | Intro to Pandas | *Why do we need to say axis =0 or axis =1 … what difference does it make ?*  *Examples*  **data = pd.concat([data,file1], axis=0)**  **data = data.drop(['tcode'], axis =1)** | Ambiguity in Pandas Dataframe / Numpy Array "axis" definition - Stack  Overflow  A data frame in pandas looks like the image above. Axis 1 and 0 is pre defined. In our example concat, we want to create long thin data, like append the rows to the existing frame – hence referring to axis =0 (axis = index/rows)  In our drop example we want to search along the column headers and apply a drop, ie following axis =1. (axis = columns)  We are specifying the axis along which we compute the function.  Useful resource for where we are now and gives a hint of where this is going to come up later on :  <https://railsware.com/blog/python-for-machine-learning-pandas-axis-explained/> |
| 21.10.20 | Using seaborn and matplotlib | What are the black lines on the top of the bars in the seaborn barplot ? | I found a couple of great articles for you on this – and you may be glad to know that you aren’t the only one who asked!  <https://stackoverflow.com/questions/35193996/what-is-the-overlay-line-on-each-bar-in-a-bar-chart>  … they are vertical error bars  -“ Error bars are a graphical representation of the variability of data and are used on graphs to indicate the error, or uncertainty in a reported measurement.”  The confidence interval – 95% in our case is calculated for you based on the number of data points and diversity. Its worth understanding what it means :  The "95%" says that 95% of experiments like we just did will include the true mean, but **5% won't**.So there is a 1-in-20 chance (5%) that our Confidence Interval does NOT include the true mean.  You can modify what the error bars show – eg standard deviation, add caps to make it more obvious that’s whats showing  <https://seaborn.pydata.org/generated/seaborn.barplot.html>  You can also turn them off <https://stackoverflow.com/questions/40088585/turn-off-error-bars-in-seaborn-bar-plot> |
|  |  | How do you turn a boxplot horizontal with matplotlib? | Simple answer which we should have guessed! Just define inside the () your requirement that this is not vertical… as that’s default, we use vertical =False  **Data[[‘actual\_height’]].boxplot(vert=False)**  **Before :**    **After :** |
| 22.10 | Review | Things we want to revisit :   * Environments (Brecht/Andres) * Github merging branches * Useful python commands for processing data | TA will cover in labs in unit 2  Will cover in class in unit 2/3 as I have invited one of the Tas from web dev to guest lecture  I have added the screenshot to the class repo under Notes |
|  | Review | New things we want to see/do   * Unstructured data - cleaning and sentiment analysis * GitKraken * Talk about the ethics and moral aspects of ML | We will start introducing this from unit 2 onwards but under review how to cover it in more depth  We will start during unit 2  Definitely unit 7, if not before |
|  | Review | Why am I seeing this error with github?  *“remote: This repository moved. Please use the new location”* | You just need to update the origin with the updated url – this has happened because you shifted your repos around in github  From command line you can run  Git remote set-url origin <updated url>  https://stackoverflow.com/questions/30443333/error-with-renamed-repo-in-github-remote-this-repository-moved-please-use-th |
|  | Seaborn /Matplot lib | Is there a way to plot inline two different charts for the same dataset? | Yes, you will need to define the subplots in your query  I found an example online :  import seaborn as sns  import pandas as pd  import matplotlib.pyplot as plt  batData = ['a','b','c','a','c']  bowlData = ['b','a','d','d','a']  df=pd.DataFrame()  df['batting']=batData  df['bowling']=bowlData  fig, ax =plt.subplots(1,2)  sns.countplot(df['batting'], ax=ax[0])  sns.countplot(df['bowling'], ax=ax[1])  fig.show()  an alternative approach :  import matplotlib.pyplot as plt  l=['batting\_team', 'bowling\_team']  figure, axes = plt.subplots(1, 2)  index = 0  for axis in axes:  sns.countplot(high\_scores[index])  index = index+1  plt.show() |
|  | Seaborn /Matplot lib | Is there an easy way to plot a pie chart ? | <https://matplotlib.org/3.1.1/gallery/pie_and_polar_charts/pie_features.html>  this link contains some examples you can walk through : <https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.pie.html>  and I like this one has a section on legends  <https://pythonspot.com/matplotlib-pie-chart/> |
|  | Seaborn /Matplot lib | How can we find out the X intercept of a linear model ? | This one is hard to track down  You can try this approach  slope, intercept = np.polyfit(x, y, 1)  Where x is your x value list  This solution seems more sophisticated but much more in depth – one to try yourselves.  <https://www.includehelp.com/python/find-the-x-intercept-and-y-intercept-of-a-line-passing-through-the-given-point.aspx> |
| 23.10 | Pre processing | Can we get a process diagram for the steps involved in the ML workflow? | Its not pictographic but I like this one <https://towardsdatascience.com/a-data-science-workflow-26c3f05a010e>  Before we revisit machine learning I will take the time to draw our planned steps with lucidchart, as I think it would be useful for future students to accompany the case studies we will work on  I found this on https://github.com/glc12125/Machine-Learning-Workflow-with-Python    And I thought this one was helpful for the non machine learning view of the world  Data Preprocessing and Workflow of a Machine Learning Project | by Rishabh  Mall | Medium |
|  | Pre processing | How can we flatten a 2d array into a 1d array? | We should be able to employ **numpy.ndarray.flatten**  There are a few alternative approaches here  <https://www.geeksforgeeks.org/python-ways-to-flatten-a-2d-list/>  and this very colourful example of multi dimensional array reshaping  <https://towardsdatascience.com/reshaping-numpy-arrays-in-python-a-step-by-step-pictorial-tutorial-aed5f471cf0b> |
| 26.10 | Basic sql | How can I display a random set of data in the preview/output from my query? | <https://www.mysqltutorial.org/select-random-records-database-table.aspx/>  **SELECT** \* **FROM** table\_name  **ORDER** **BY** **RAND**()  **LIMIT 20 ;**  A much more in depth answer here  <https://www.mssqltips.com/sqlservertip/6347/selecting-a-simple-random-sample-from-a-sql-server-database/> |
| 27.10 | Sql regex | How is REGEXP used in the ‘real world’ | Im not suggesting to do these examples but I think there are some familiar concepts on here, eg how to pull the HTML tag out of a url , password validation  <https://dev.to/oahehc/master-regular-expression-through-real-world-examples-3164>  for more info on regex and to try it – I recommend these links  <https://www.rexegg.com/regex-quickstart.html>  <https://en.wikipedia.org/wiki/Regular_expression> - sometimes wikipedia is useful!  <https://www.freeformatter.com/regex-tester.html> |
| 30.10 | ERD group project | How would we capture opening times efficiently and sufficiently in a relational DB? | There are a number of different approaches to doing this. A denormalised approach would insert the relevant columns in the restaurant table – one of the groups created a field weekday close time, which would have answered the question set for the project with the minimal amount of redundancy – as I was only interested in restaurants open after class on a weekday. The only thing extra step would be to assert a slowly changing dimension, timestamp or other logic into the design to ensure that changes over time could be captured for that field.  However, for the full flexibility around opening times it is likely that a normalised solution would be required. Some discussions about the method and recommended constraints on such a table here:  <https://stackoverflow.com/questions/1036603/storing-business-hours-in-a-database>  I have mocked up 2 approaches in excel, saved to the class repo /QuestionsfromClass  Try importing either model into a database from the excel file to see how the joins would work in mysql.  note both of these designs, though adequate for our purposes, wouldn’t allow for special closing circumstances (ie the restaurant is closed for refurb), for exceptional holiday hours (we are assuming a standard 7 day pattern) or for breaks in the day (eg closed for lunch, reopens again for dinner) because that solution would get quite complex!  You could also use a bridge table to handle many to many relationships, such as the film\_category table shown in the Sakila database ERD. This table can be labelled as restaurant\_openinghours and would contain a unique row id as the PK, with links to both the restaurant table and a more generic opening hours table with all of the opening hours options listed  A final option to consider would be storing the times in a semi structured/ nested way – heres an example https://schema.org/hoursAvailable |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |