**POSTS:**

**POST 1: Posted on 3/2/2022 [7:03 PM] Saima Rahmanzai**

**What I learned in this course:**

I learned a lot in this course, statistical concepts like CDF, PMF, Pareto distributions, and many more concepts that I was not familiar with.  I also learned how to use pandas, NumPy, SciPy, Seaborn, Matplotlib, etc. In my prior courses, I had not used pandas so it was good work with pandas to perform statistical tests and data exploration, plots, etc.  I used many learning tools to learn these concepts and alternative ways/commands to perform calculations.  That included the use of our textbook, the exercises and solutions it provided, YouTube videos, Udemy courses I purchased on these specific topics, google articles, and access to all Data Scientists forums where any issues identified were already discussed and techniques provided to help a user like me when I got stuck.  My other great learning avenue was the team environment where I could reach Dr. Shankar and fellow students to ask questions where all the other resources were unable to help.  I was rescued by my fellow class mates several times.  Thank you all for your help!  In summary, a great learning experience by taking a class this semester!

**POST 2: Posted on 3/2/2022 [7:12 PM] Saima Rahmanzai**

**How I will apply these techniques in the real world:**

I am in internal audit and work in IT side of things.  I also run the data analytics department.  I will use these skills (Python and R) in my data analytics scripts.  These programming languages are very inexpensive compared to the other tools we are currently using.  It will cut down our department costs but the learning curve is high.  I started to learn these languages and practice with it and am urging our Data Analytics staff to start learning these skills for future use in case we replace our current tools.  The learning curve is high.  The current tools are in SQL language and in tools that have gui interface so if you choose not to write scripts, you can click and run commands so user friendly in that sense.

 Additionally, we audit the Enterprise Data Management group and they, in addition to running the data warehouse and creating adhoc data analytics reporting for management decision making, they also create models (could be traditional and AI (Artificial Intelligence)), and having training and test datasets and those models are then placed into production.  We are required to audit these models and since we in audit do not have the statistical skill sets, we hire external consultants to help us with the technical know-how if the model is working as expected, not biased, properly tested and variables selected were appropriate etc.  By going through the Data Science program, I am learning a lot on model creation starting with EDA, Data preparation, and other phases of model creation and hope to be able to do model validations internally and save the enterprise consulting expense.

**POST 3: Posted on 3/2/2022 [7:54 PM] Saima Rahmanzai**

**The areas I struggled the most:**

While I learned a lot in this course, I struggled with certain concepts.  Not just in this course, the use of gitbash to commit files and move to remote GitHub repository is always challenging to me as I have to remember the code and I do not always remember from one class to the next.  I keep my screenshots so that I can refer back to them since I do not use it at work.   Hopefully the more I use it, it will become second nature.

Like the posts I saw of my fellow classmates, I too struggled with the concepts of CDF, PMF, etc. and was hard to understand especially as taught in the book.  The instructor called on their own created code under ThinkStats package.  I did my own research trying to find a way to understand how to use alternative code and easy to understand videos to recreate the code.  Though I learned a lot, having a full-time job and performing the extra research and learning did translate into long study nights and less sleep.  However, it was worth it 😊

Another struggle was to find datasets that contained all the variables that would help solve the hypothesis.  That is a lot of data available but may not have all the fields we need.  Even merging datasets could be a struggle as common field used to merge may not exist in both datasets.  I would think it would be easy to use these concepts at work as data is readily available but for class purposes and finding data publicly available, it is hard to find sources where you can find all the data fields what you are looking for.

**Post 4: 3/2/2022 [7:58 PM] Saima Rahmanzai**

**Areas where I could spend more time understanding:**

I learned pandas and the command used to do different statistical or data manipulation techniques or create plots.  I will need to continue to work with pandas and also with other languages like python and R so I get used to the syntax and not always look for examples for simple commands.  I am also not very comfortable with the loop command and would need to be comfortable in its use and accurate application.  Another struggle of mine is creating methods and functions and then calling them into the program to be executed.  In general, I want to be proficient in these languages and it will come with practice. Oh, and I left out GitBash 😊.  Another one that I need to be proficient at.

**Post 5: 3/2/2022 [8:04 PM] Saima Rahmanzai**

**Areas I found to be most interesting:**

To me, creating the visual plots and how much is out there to make the graphs look so colorful and overlay google maps etc. was very interesting to me.  I came across a video where the instructor uses the plots to identify different species of trees and their age in the San Francisco area.

I was surprised how much you can do with plots and different styles and also so many tools to create these plots such as Seaborn, Matplotlib, and more.

**Replies:**

**Reply 1: 3/2/2022 [1:15 AM] Sameer Nepal**

I was wondering if anyone here has expertise on both python and R. It would be great to hear from you if you prefer one to another and how they are popular in data science world. Also, has  anyone used R as their coding language for their final project ?

​​

**3/2/2022 [6:13 PM] Saima Rahmanzai**

Hi Sameer Nepal.  I also am not an expert but I thought R was a better statistical and visualization tool than using pandas.  Comparing Python than with pandas, I preferred Python.  Pandas was a bit confusing to follow.  Python is good writing code outside of statistics in my opinion but for concepts like regression, etc. I think R is better.

**Reply 2: [3/1/2022 9:54 PM] Logan Quandt**

One thing I appreciated learning in this class was the use of Juypter Notebook. In our previous courses we just used Pycharm but Juypter Notebook is great for learning the data science materials in this course due to its ability for quick visualization of both the code and results. I also found it somewhat easier to correct errors than in an IDE. I'll be interested to see what future courses prefer whether its PyCharm/Spyder or Juypter Notebook.

​​

**3/2/2022: [6:18 PM] Saima Rahmanzai**

Hi Logan Quandt.  I liked Jupiter as well but I think there is limitation on size of the code as it gets to a certain size, the Markdown option to pdf does not work.  The option available in Pycharm is much more forgiving when it comes to the size of your coding project.

**Reply 3: [3/1/2022 10:11 AM] Madeleine Sharp**

Overall Thoughts on DSC530

Overall, I have enjoyed this course and have most certainly learned some new topics that I was not familiar with prior. More specifically, I gained additional experience with the ThinkStats/ThinkStats2 package, which really is not something I had used within Python in other courses as of yet. I also had not yet done any work with PMF, PDF, CDF, etc. – so pretty much each week I felt as though I were learning a new statistical concept, which I appreciated. I think the text and accompanying GitHub was extremely helpful, and I enjoyed the Teams discussions that were had throughout. Given that I have learned so many new statistical concepts, I truly feel as though I will be able to apply these to the real world in the sense that my data science knowledge will be bolstered, and I will be able to utilize these methodologies moving forward – whether in a future class or at my job.

On the flip side of that coin, I do wish there was perhaps some additional instruction on the ThinkStats package (or that we also incorporated other packages more frequently as well) – there is not a ton of documentation regarding this package, so sometimes it was difficult to figure out, even when I needed to do some additional Googling. The silver lining in that is it pushed me to really learn the package for what it is, and not necessarily rely on others that I was already familiar with. While I liked that the author used a single dataset throughout the book so that we could see all of these concepts utilized on the same dataset from start to finish, I do wish perhaps some other datasets were introduced so that we could see the concepts on a variety of different datasets (I know a couple others were thrown into the mix here or there, but the main focus felt as though it were on the pregnancy dataset - which can be both a pro and a con in its own right).

All in all, I really enjoyed this course. I continually learned something new, the resources were more helpful than not, and Dr. Shankar Parajulee always gave great feedback on my assignments or any questions that I had and reached out about (thank you!). I feel I also learned from all of you, my peers, so thank you for your contributions, bright ideas, help, etc. I am grateful for the people in this course as well as the subject matter!

​

**3/2/2022 [6:24 PM] Saima Rahmanzai**

Well said Madeleine Sharp.  I totally agree with all the comments you made.  Learnt a lot and learned even more trying to figure out the code in ThinkStats package and googling and searching YouTube for alternative easy ways to understand the concept as sometimes I did not understand the code behind ThinkStats.

**Reply 4: 3/2/2022 [12:31 AM] Sameer Nepal**

Now that we are at the last week of the course, I wanted to know if anyone had hard time with the coding done in the book. I personally found the codes tough to remember but this could be just because I am not using it at my day to day work. I hope it will be easier when I start using it more when I start using it to see the behavior of the data sets.

​

​

**3/2/2022 [6:30 PM] Saima Rahmanzai**

Hi Sameer Nepal.  I think we all felt the same way and that is why we were trying to understand e.g. how to calculate CDF, PDF, PMF, etc. using other sources on the internet or using my Udemy courses that I purchased to understand these concepts.  All in all though, I learned a lot due to looking at different sources.

**Reply 5: [3/2/2022 12:30 AM] Sameer Nepal**

I was wondering if anyone here was already familiar with the topics that we covered in this course. If so, is it that you have taken statistics class in the past or do you use it at your current work. I am asking this to know how useful these topics would be in our career path in data science as we know not everything we study in school is used later at the job we do today.

​

**3/2/2022 [6:42 PM] Saima Rahmanzai**

Hi Sameer Nepal.  Most of the topics were very unfamiliar to me so I did learn a lot in this course.  I prior course, I heard of using pandas and now I used it.  So far, I have learned three languages thus far including R, Python, and now pandas which is like a flavor or Python in my opinion in this Master's program.  I am by no means an expert and would need to practice these languages in my work in data analytics or auditing our Model group.  We do have our other data analytics tools but these tools are pretty much free but the learning curve is steep.  Practice makes perfect and I will continue to practice on off semesters by going through the Udemy courses I have practiced.  I also learned Markdown which also has a bit of programming involved and not to forget GitHub.  I forced myself to use the Gitbash vs. the gui which means more code.  I have saved my code and revisit it as hard to remember everything smile