**DSC 530 Assignment 12.1 Discussion (Summary of what was Learnt)**

**Summary of What was Learnt in this Course**

**Data Science Process and Preparing for Exploratory Data Analysis**

In the first week of Class, I learnt the importance of using “Anaconda” for exploratory data analysis. In Anaconda, you have the Pandas library which is best for data wrangling and data exploration. The Pandas library is organized into data frames and can be used to import all kinds of data from CSV, Excel, etc. Pandas can also be used to filter, sort, and display data with ease. For numerical modeling analysis there is NumPy library and scikit-learn which is one of the most powerful machine learning algorithms. It also has an interface that allows Data Scientist to tap all the power of machine learning without it’s many complexities.

**Distribution**

Another thing I learnt in this course is the use of the matplotlib for data visualization. Matplotlib is a very powerful data visualization tool in python and is incorporated into anaconda. From this I was able to plot histograms that showed the frequencies of the different variables available. The importance of data visualization to a data scientist cannot be over emphasized. Data visualization helps to see the central tendency, modes spread of the data; tails and outliers if present. I also learnt the different ways of handling outliers in the data.

**Probability Mass Function and Cumulative Distribution functions**

Under this topic, I learnt the definition of both the PMF and CDF. I further learnt what both tells us and how it can be used in exploratory data analysis in real life. I learnt how to plot both the PMF and the CDF, although I have not perfected it yet. It is still an ongoing improvement process. The Class size Paradox is a fascinating discovery that was found along the learning pathway.

**Modeling Distribution and Probability Density function**

I learnt the different types of distributions which can vary from exponential distribution to Pareto distribution. How to determine each distribution and in what situation each distribution can be applied. I learnt the definition of probability density function, when it can be used and what the results indicate. I also learnt about the KDE, the benefits of KDE, and what the interpretation of the results of KDE.

In the later part of the class I solidified my knowledge of the relationship between variables sampling bias and how it can occur. Furthermore, I solidified my knowledge on hypothesis testing and the importance of the p-value in hypothesis testing. I struggled a little bit in the final 2 topics which covered “Time Series” and “Survival Analysis & Data Modeling Basics. These areas where new and I struggled a little bit without prior knowledge. I feel confident that I will be able to apply most of these techniques to solving real world data science problems. I may need to reference the text and look for additional references with regards to the last 2 areas I struggled with. I found Regression analysis to be the most useful.

References

Bichler, M., Heinzl, A., & van der Aalst, W. M. (2017). Business analytics and data science: once again?.

Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big data analytics: Applications, prospects and challenges. In *Mobile big data* (pp. 3-20). Springer, Cham.

Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, *41*(4), 70-73.

Ertemel, A. V. (2015). Consumer insight as competitive advantage using big data and analytics. International Journal of Commerce and Finance, 1(1), 45-51.

Hamid, H. A., Rahmat, M. H., & Sapani, S. A. (2018, April). The classification of PM10 concentrations in Johor Based on Seasonal Monsoons. In *IOP Conference Series: Earth and Environmental Science* (Vol. 140, No. 1, p. 012028). IOP Publishing.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., ... & Yu, T. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, *395*(10223), 507-513.

Rosner, B. (2011). *Fundamentals of biostatistics* (7th ed.). Boston, MA: Brooks/Cole Cengage Learning.

Downey, A. B. (2015). *Think Stats: Exploratory data analysis*. Beijing: OReilly.