CS306 Data Analysis and Visualization

Lab. 3 Airline Data (Big) Analysis

1. In this lab practical we will work with Big data related to Airline. Download 2 files from the lecture/lab folder i.e. (1) 2008.csv.bz2 (2) Airline.desc. Unzip the files and you will get .csv file for further experiments.

1. Load the csv file in Excel/Calc utility available in your system. Are you able to load all the data in Calc? Explain the limitations of Calc in handling big data.
2. Now load the same csv file in Python/R. List the difference between the rowcount compare to file loaded in Calc?
3. We will use two features from the given csv file i.e. AirTime and Distance. Use Data cleaning techniques and find outliers if any. Normalize both the features to zero mean one standard deviation, find pdf and cdf of the normalized features. To build the pdf you have decide the number of bins and bin sizes.
4. Now use random number generator to get samples from Normal distribution N(0,1) approximately same number of samples. Obtain pdf and cdf of these samples simulated from a normal distribution. Verify that your simulated data is correct.
5. Use K.S. Test to find out that the Distance feature is Normally distributed or not. Comment on your results.
6. What α value you have used for K.S. Test ? What impact it has on your results.

Use Python/R for this exercise….

Optional suggestion for BRAVOS

For working online with python you may try to use Google Colab (with GPU) which is very fast and can be used to work remotely as well….

For tutorials of Google Colab you may try:-

1. Colab welcome notebook - https://colab.research.google.com/notebooks/welcome.ipynb
2. Medium Article:- https://medium.com/deep-learning-turkey/google-colab-free-gpu-tutorial-e113627b9f5d