# **ADTA 5410**

**Applications and Deployment of Advanced Analytics**

**Assignment-2**

1. This exercise relates to the colleges in the US, which can be found [here](https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?goToReportId=7&sid=25547b1a-aff9-4021-93cc-44172cb7dd10&rtid=7) (main Menu> Complete Data Files). It contains a number of variables for more than 6000 different universities and colleges in the US.
   1. Download the “Institutional Characteristics (HD)” from [here.](https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?goToReportId=7&sid=25547b1a-aff9-4021-93cc-44172cb7dd10&rtid=7) Unzip the file and save it as .xlsx. Pick the year you are assigned from the list below. This is NOT a group work. You must do your homework individually. If your name is not on the list, please let me know, so that I can add you to the list.

|  |  |
| --- | --- |
| **Name of the Student** | **Year** |
| Biniam Abebe, Cynthia Ani | 2022 |
| Akhila Chilukuri, Anthony Cisneros | 2021 |
| [Saicharan Goud Gunegari](https://unt.instructure.com/courses/106763/users/357805), Tharuni Dusakanti  .  Zyon Daniels, Tharuni Dusakan | 2020 |
| Keerthi Erram, Briaunna Fontaine-Deth | 2019 |
| Prathyusha Gogineni, Niranjan Guda  Prathyusha Gogineni, Niranjan Guda | 2018 |
| Kaladhar Gundla, Kristin Hassett | 2017 |
| Liz Jones, Gowtham Kavvadi | 2016 |
| Supriya Kolluru, Shravani Kuragayala | 2015 |
| Ian Lynch, Surya Sai Srikanth Machimchetty  Ian Lynch, Surya Sai Srikanth  ManchimcheVy | 2014 |
| Chloe Marshall, Goda Harshini Munagala Naga Venkata | 2013 |
| Shivani Muvva, Chariteash Narra | 2012 |
| Trevor Nolen, Nikhila Pinnapureddy | 2011 |
| Hima Sindhuja Pushadapu, Kayla Reynard | 2010 |
| Venkata Sanath Kumar Suram, Vaishnavi Thalla | 2009 |
| Krishna Vamsi Uppala, Hari Babu Uppari  Krishna Vamsi Uppala, Hari Babu Uppari    . | 2008 |

* 1. Use the pd.read\_xlsx(…., index\_col="…") function to read the data into Python. Call the loaded data college. Set the INSTNM as index, using index\_col="…". Make sure that you have the directory set to the correct location for the data.
  2. Check the first five rows of the data.
  3. How many rows are in this data? How many columns?
  4. Delete all the columns except the ones given below. Print the first 5 rows.

|  |  |  |
| --- | --- | --- |
| 1 | UNITID | Unit ID for institution |
| 2 | INSTNM | Institution name |
| 3 | CITY | City |
| 4 | CONTROL | Control: (1) Public (2) Private non-profit (3) private for profit |
| 5 | HLOFFER | Highest level of offering |
| 6 | STABBR | States |

* 1. Delete all the rows(universities/colleges) which do not have at least 4 “NULL” columns, using .dropna(axis=0,thresh=4) function and call the new dataframe filtered\_college. Print the first 5 rows of filtered\_college.
  2. How many rows and columns does filtered\_college have?
  3. Use the describe() method of to produce a numerical summary of the variables in filtered\_college data set.
  4. This part has two steps
     1. Delete all the rows which has HLOFFER= -3
     2. Using ‘import matplotlib.pyplot as plt’ and df.plot.scatter(x=`….`, y=`…`) function, produce a scatterplot matrix of the columns ['CONTROL','HLOFFER'].
  5. Find an appropriate visualization to show how the control of universities (CONTROL) is distributed across states (STABBR).

1. (BONUS- Up to 10% additional) Create your own question and provide the answer.