A screenshot of a computer

Description automatically generatedA table with numbers and letters

Description automatically generatedA table with numbers and letters

Description automatically generatedThe datasets contained in this report are the USA NPN A dataset, the USA NPN C (NPN, 2025) dataset, Higher education administrative characteristics dataset (IPEDS\_Education, 2025), Student tuition dataset (IPEDS\_Education, 2025) which includes a data dictionary two healthcare datasets delineating procedures covered by insurance (data.healthcare.gov, 2020) and zip codes (data.healthcare.gov, 2014). Some snippets of the datasets follow. (data dictionaries are linked in [appendix 1](#_Appendix_1:_Data_1) and datasets are linked in [appendix 2](#_Appendix_2:_Datasets)). The caption **beneath** each figure indicates the dataset.

A table with numbers and text

Description automatically generatedA table with numbers and letters

Description automatically generatedA screenshot of a computer

Description automatically generated

Student Tuition Dataset

NPN A

Zip codes

Healthcare procedures

Academic Characteristics

NPN C

The NPN A dataset has plain text, numeric, date, and geospatial coordinate data. The Zip Codes dataset is made up of numeric, text and null field types. The NPN C dataset is made up of the same field types as the other NPN datasets along with a siteID that is of the integer type. The Healthcare Procedures dataset contains string types, numeric types, and geospatial – numeric data. The Student Tuition dataset is made up primarily of numeric data, Boolean fields (2- or 4-year institution), with some string data. The Academic Characteristics dataset contains string, numeric, and Boolean fields.

Note that all analysis methods discussed below will be using supervised learning in that the data is already labelled and the target is known. In doing my research I found this definition of supervised versus unsupervised approaches, generated by Google’s AI:

‘… supervised learning uses labeled datasets where the desired outcome is known, allowing the model to learn relationships between input and output data to make predictions on new data, … unsupervised learning works with unlabeled data, aiming to discover hidden patterns and structures within the data without any predefined target outcome …’

NPN A lends itself to *categorization* since the target variable is whether or not the species will live in the particular microclimate (1-True or 0-False) and the features are numeric, Boolean, geospatial and string. One could use *regression* here as well, if the target is defined as say the number of inches of precipitation. Clustering is probably not well suited to this dataset in that it is used primarily to identify patterns in a dataset without labelled outcomes. In this case the target, whether the individual or species lives, is labelled. At this time other analytical options have not been investigated to any extent that would merit discussion here. Questions that could be answered with this dataset include, Which types of species survive best in this microclimate. This dataset has several columns/rows where the value has been set to -9999. In a lecture the professor suggested to us that this value was used by many data scientists as a place holder for either missing information or information that was out of bounds. Given the nature of the dataset and the placement of the -9999 values, it is most likely that these are non-existent values. In this dataset I will treat these values as nulls and calculate accordingly. Looking at the data, I can see right away that the dataset ID is set to ‘-9999’ and the patch is set to -9999. The dataset ID field, from the data dictionary, “"-9999" indicates one or more status records were submitted…” The value of -9999 in the patch field, according to the data dictionary, indicates that a value for this field was not recorded (the number one (1) indicates yes, but no other value is available). This dataset seems to have a significant amount of data that is represented by these placeholders (in other fields as well). I find this dataset interesting since I have been making these observations as a citizen scientist since 2008. The dataset is most interesting to the end-user/industry as a way to predict the ability of a specific species to survive in a given microclimate.

The USA NPN C (NPN, 2025) dataset is larger than either of the other NPN datasets listed in the journals. Because some of the features can be used to predict the target variable of whether the individual with live or not, this easily lends itself to classification. Like the NPN A dataset this one can be used with regression analysis if the target is changed. On the other hand, clustering is not an efficient method here since the features, by and large, and the target have been labelled. This dataset can answer the questions, using a broader sampling, of which individual species survive better in varying microclimates, How the Botanical Garden compares with the rest of the state, What species do best in protected (like a botanical garden) versus unprotected microclimates, Which specimens/types of specimens are most likely to experience problems due to environmental factors. Both the personal and industry interest are the same for this dataset as the NPN A dataset.

The Academic Characteristics dataset (IPEDS\_Education, 2025) may be best used for classification since the target would most likely be whether it is a 2- or 4-year institution. The most pressing question one might ask of this data would be whether cost per semester or credit hour is predictive of a 2 vs a 4-year institution. One also might ask if one could predict population density from the number of institutions in each area. I find this to be a more than interesting way to look at the plethora of academic institutions in a new light. The industry itself might find this helpful when it answers the questions regarding 2 vs 4-year institutions.

The Student Tuition dataset (IPEDS\_Education, 2025) is made up mostly of financial data types. The dataset lends itself to regression analysis since both the fields and target would be numeric and the target and features are related in a significant way. Although clustering could be used in this case, it would be mostly redundant since the features are already marked and well defined in the data dictionary (see [appendix 1](#_Appendix_1:_Data)). Questions one might ask using this data include: How does the in-state tuition compare with the out-of-state tuition? Is it better to be an in-state undergraduate student on a cost basis, or an instate graduate student? Which graduate program costs the least, or the most? Similarly to the other education dataset, this one is interesting mostly for how tuition may or may not affect the institution itself, in terms of enrollment or potential enrollment. This is also the reason the institutions might be interested.

The Healthcare Procedures dataset (data.healthcare.gov, 2020) includes a variety of medical procedures, whether they were covered by insurance, how much was covered, which healthcare plan was used, and where the procedure took place. The data itself almost asks the question of how likely a procedure is going to be covered given the place it is performed. One could also compare the coverages of each plan, and even which plans are more likely to cover procedures performed in a specified type of place. This is interesting to me in that one could ascertain which healthplans are most beneficial for the end user. The industry would probably be interested on the other end – which healthplans would be best, in terms of overall coverage, for any one institution to belong to.

The Healthcare Zip Codes dataset (data.healthcare.gov, 2014) is best used to augment the Healthcare Procedures dataset. It is made up of a list of zip codes for the various areas identified within it. The majority of the data is numeric, although there are some features which are string variables, such as the state and county name. By itself, this dataset is mostly useful to ascertain age statistics of residents in a given zip code. Questions might include asking about the number of residents of a zip code in a certain age range (since the dataset is broken down by age ranges). This dataset could be analyzed using regression analysis since the data itself is mostly numeric floats, but depending on the target, one could use classification. Clustering does not seem to be an efficient analysis method for this dataset, mostly because the fields are well labelled (if one uses the data dictionary).

Up until now I have considered the use of data in a singular way, how I could use it to visualize the data in such a way as to make it usable for clients to make more informed decisions regarding whatever the data represents. I had thought that datasets of several thousand data points (rows) were more than sufficient to understand a given situation. I am now looking at datasets of just over 1 million rows and sometimes finding too few datapoints to properly represent the data. Indeed, one of the sets I am using was truncated when it was opened using Microsoft Excel, because excel can only display just over a million lines of data in a single worksheet.

So far in this project I have been able to identify, retrieve, and process over 10 datasets. This was instructive as I needed to make decisions about the data as I was looking at (would this be useful, is it representative, is there enough usable data, etc.). For each dataset I have tried to ascertain not only the best way in which to analyze it, but also the features and a target that would get the best result. During this project I have developed the outlook about data that enables me to see each dataset in the light of how it could be used to predict an outcome that is both useful and interesting. For my team work in the Education industry, I believe this will stand me in good stead to be able to identify and analyze datasets with which I am not as familiar. I have also developed a better understanding of decision trees, which I can use to help analyze data our team may choose.

Probably the most difficult and challenging part of this project was writing the report. Dealing with the datasets has been relatively easier. Understanding the differences between classification, clustering, and regression has been extremely important, but not as straightforward as one might think. I found the Data Science for Business (Provost, 2013) readings were extremely helpful in both understanding these analysis methods and the various other modelling ideas throughout the course.

I have found the different required readings somewhat challenging, in that they sometimes assume a level of expertise in statistics and probability that I do not possess. Due to this I have ordered and started reading the Data Science for Dummies (Pierson & Pierson, 2021) and Statistics Workbook for Dummies (Rumsey, 2019) books to get up to speed. Because of my background, which did not include statistics and probability, I have needed to play catch-up almost the entire time. Luckily, I have extensive background in data visualization and have found those aspects of the course to be much easier. I feel that my biggest success so far has been the ability to explain what I am doing to my wife. She has always been my willing sounding board to see if I am explaining something in plain English, rather than technobabble. Amid the constant hubbub of this course, one person posted on Yellowdig about customizing your GitHub readme.md file. That sent me down a “rabbit hole” from which I haven’t yet emerged. I feel I have contributed a large amount to all students by putting all course resources in one place.

I still need to improve my coding skills, as I can get the homework for DX602 correct but still understand only about ½ to ¾ of the code I have written. I also need to get a better understanding of the various analysis methods listed here.

**Bibliography**

data.healthcare.gov. (2014). SLCSP - County-Zip Reference Data - Data.Healthcare.gov. https://data.healthcare.gov/home, https://data.healthcare.gov/dataset/yaaf-rjhy

files/178/yaaf-rjhy.html

data.healthcare.gov. (2020). Benefits and Cost Sharing PUF -2020 - Data.Healthcare.gov. https://data.healthcare.gov/home, https://data.healthcare.gov/dataset/kq37-29bw

files/176/kq37-29bw.html

IPEDS\_Education. (2025). IPEDS Data Center. https://nces.ed.gov/ipeds/datacenter/InstitutionList.aspx?goToReportId=1&sid=4964fa06-44c4-4b88-be68-47bbbce7e33f&rtid=1

files/145/InstitutionList.html

NPN, U. (2025). Phenology Observation Portal. https://www.usanpn.org/data/observational

files/5/get-started.html

Pierson, L., & Pierson, L. (2021). *Data science for dummies* (3rd edition. ed.). John Wiley & Sons, Inc. https://catalogue.solent.ac.uk/openurl/44SSU\_INST/44SSU\_INST:VU1?u.ignore\_date\_coverage=true&rft.mms\_id=9997437348304796

Provost, F. F., Tom. (2013). *Data Science for Business*. O’Reilly Media, Inc.

Rumsey, D. J. (2019). *Statistics workbook for dummies* (2nd ed.). Wiley Publishing, Inc.

# Appendix 1: Data Dictionaries

1. USA NPN dataset A
   1. [Link](https://docs.google.com/spreadsheets/d/1AS4hXy0uTIhZYt6Htagm5u1rkGHPqaS_/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true) (from BU Google Drive)
   2. [Search Parameters](https://docs.google.com/spreadsheets/d/1AWmfKyAQp8fvcaZTW5ZZjdsNm47lrMUH/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true)
2. USA NPN dataset C:
   1. [Link](https://docs.google.com/spreadsheets/d/15royrO-WT_DwzhC-Ek-ty7LFMzqyFILn/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true) (from BU Google Drive)
   2. [Search Parameters](https://drive.google.com/file/d/1683lpBKAYUW_eADZWWaPLeSzCN4NXEV8/view?usp=drive_link)
3. Education institutions
   1. [Link](https://docs.google.com/spreadsheets/d/1AblQkin8tOCazmnRedDGeyQoj9GHpCNt/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true) (from BU Google Drive)
4. Education tuition
   1. [Link](https://docs.google.com/spreadsheets/d/1AkxzBBwigOgw4WRG0wzP4M9r8eaGITJ4/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true) (from BU google drive)

# Appendix 2: Datasets

1. USA NPN dataset A
   1. [Link](https://docs.google.com/spreadsheets/d/1AN5KZLtM165bpBleNt_2Qd_lpUSnZYDV/edit?usp=sharing&ouid=103523606972182441044&rtpof=true&sd=true) (from BU Google Drive)
2. USA NPN dataset C:
   1. [Link](https://docs.google.com/spreadsheets/d/15gTvxQEm0rBkrk0TlmFi3VPCbrKFYlw0/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true) (from BU Google Drive)
3. Education institutions
   1. [Link](https://docs.google.com/spreadsheets/d/1AYuS4U62m1VWNwEzphKBgmSDN3aT4eJu/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true) (from BU google drive)
4. Education tuition
   1. [Link](https://docs.google.com/spreadsheets/d/1AgH55ta07k-afzBkC_0uMPD6dBltDe8y/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true) (from BU Google Drive)
5. Healthcare procedures
   1. [Link](https://docs.google.com/spreadsheets/d/1uB3_Fc9R08zJgovOFm-j4usxNIdn2u2r/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true) (from BU Google Drive)
6. Healthcare zip codes
   1. [Link](https://docs.google.com/spreadsheets/d/1uCNVFgmm5sW5c16AQhP5iGDX8cEqkCJZ/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true) (from BU Google Drive)

# Appendix 3: [PowerPoint from week 5 reflection](https://docs.google.com/presentation/d/1siVimktIrtphP6c9sqwHSI4wF70vzdOf/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true)

# Appendix 4: Reflection Journals

1. Week 2 [link](https://docs.google.com/document/d/16W24TnxLvppP3OI3oE-OPnQlWCUWLeWH/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true), journal
2. Week 3 [link](https://docs.google.com/document/d/16THL0NDWnM_tc-huOccDp8SLKsppyCZL/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true), journal
3. Week 4 [link](https://docs.google.com/document/d/16P7PkxIb5hf_b13OdnNiTW51ic_DU_mX/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true), journal
4. Week 5 [link](https://docs.google.com/document/d/16ERcfNPi5Rjo46cx4u8Pw5ve1EAqVqlw/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true), journal
5. Week 6 [link](https://docs.google.com/document/d/16L732eWlPXcMvudJe7gZuIswh3unxha5/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true), journal