In this report six datasets will be discussed. The first two datasets detail information on plant life in the USA; USA NPN A (NPN, 2025) dataset and the USA NPN C (NPN, 2025) dataset. These datasets both have plain text, numeric, date, and geospatial coordinate data. NPN C additionally contains a Site\_ID field that is of the integer type.

The next two datasets are to do with education. First is the higher education administrative characteristics dataset (IPEDS\_Education, 2025), which contains string, numeric, and Boolean fields. The second is a Student Tuition dataset (IPEDS\_Education, 2025) which is made up primarily of numeric data, boolean fields (2- or 4-year institution), with some string data.

The final two datasets contain information on healthcare. The first delineates procedures covered by insurance (data.healthcare.gov, 2020) and the second covers zip codes (data.healthcare.gov, 2014). These two contain string types, numeric types, and geospatial – numeric data and numeric, text and null field types respectively.

[A screenshot of a computer

Description automatically generated](#NPNC_dataset_appendix2)[A table with numbers and letters

Description automatically generated](#ED_Inst_dataset_appendix2)[A table with numbers and letters

Description automatically generated](#HC_Procs_dataset_Appendix2)[A screenshot of a computer

Description automatically generated](#NPNA_dataset_appendix2)[A table with numbers and text

Description automatically generated](#HC_Zips_dataset_Appendix2)[A table with numbers and letters

Description automatically generated](#ED_Tuition_dataset_appendix2) 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. (click the dataset image to see the appendix entry, or click the dictionary caption for the dictionary[[1]](#footnote-1)).

NPN C Dictionary

Healthcare procedures

NPN A Dictionary

Academic Characteristics Dictionary

Student Tuition Dataset Dictionary

Zip Codes

Note that all analysis methods discussed below will be using supervised learning in that the data is already labelled and the target is known. NPN A lends itself to *categorization* since the target variable is whether the plant 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? Which type of specimen is hit the hardest by changes in weather patterns (by genus/species)? What types of weather have been shown to have the most benefit for each of the different species represented? How is the climate affecting the manifestation of each phenophase? Has there been a change of when a specific phenophase is manifested by date?

This dataset has several columns/rows where the value has been set to -9999. Given the nature of the dataset and the placement of the -9999 values, it is most likely that these are non-existent values. According to the data dictionary, the value of -9999 in the patch field indicates that a value for this field was not recorded. This dataset seems to have a significant amount of data, in various fields, that is represented by these placeholders. In this dataset, I will treat these values as nulls and calculate accordingly. 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[[2]](#footnote-2). It contains a broader sampling of which individual plant species survive better in varying microclimates. Because some of the features can be used to predict the target variable of whether the individual plant species will live or not, this easily lends itself to classification. Like the NPN A dataset (NPN, 2025) 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, by and large, the features and the target have been labelled. This dataset could answer the following questions. 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 the 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. This will allow us to look at 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 in-state graduate student? Which type of 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.

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. Interesting questions arise from this data such as 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 health plans are most beneficial for the end user. The industry would probably be interested on the other end – which health plans 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, or any other 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, several of the sets I am using were truncated when they were opened using Microsoft Excel[[3]](#footnote-3) (Lumiere, 2023), 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[[4]](#footnote-4), I believe this will put 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 most of the course resources in one place.

**Bibliography**

1 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

2 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

3 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

4 Lumiere, V. (2023). *The Microsoft Office 365 Bible all-in-one for beginners : the complete step-by-step user guide for mastering Word, PowerPoint, OneDrive, Excel, Teams, OneNote, Outlook, Access, Publisher, SharePoint, and Exchange to help with productivity and completing common tasks on PC or MAC for business, office & personal use (computer/tech)* (First edition. ed.). publisher not identified.

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

files/5/get-started.html

6 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

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

8 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](https://docs.google.com/spreadsheets/d/1AS4hXy0uTIhZYt6Htagm5u1rkGHPqaS_/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true)
   2. [Search Parameters: https://docs.google.com/spreadsheets/d/1AWmfKyAQp8fvcaZTW5ZZjdsNm47lrMUH/edit?usp=drive\_link&ouid=103523606972182441044&rtpof=true&sd=true](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](https://docs.google.com/spreadsheets/d/15royrO-WT_DwzhC-Ek-ty7LFMzqyFILn/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true)
   2. [Search Parameters: https://drive.google.com/file/d/1683lpBKAYUW\_eADZWWaPLeSzCN4NXEV8/view?usp=drive\_link](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](https://docs.google.com/spreadsheets/d/1AblQkin8tOCazmnRedDGeyQoj9GHpCNt/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true)
4. Education tuition
   1. [Link: https://docs.google.com/spreadsheets/d/1AkxzBBwigOgw4WRG0wzP4M9r8eaGITJ4/edit?usp=drive\_link&ouid=103523606972182441044&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1AkxzBBwigOgw4WRG0wzP4M9r8eaGITJ4/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true)

# Appendix 2: Datasets

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

# 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 journal: <https://docs.google.com/document/d/16W24TnxLvppP3OI3oE-OPnQlWCUWLeWH/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true>
2. Week 3 journal: <https://docs.google.com/document/d/16THL0NDWnM_tc-huOccDp8SLKsppyCZL/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true>
3. Week 4 journal: <https://docs.google.com/document/d/16P7PkxIb5hf_b13OdnNiTW51ic_DU_mX/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true>
4. Week 5 journal: <https://docs.google.com/document/d/16ERcfNPi5Rjo46cx4u8Pw5ve1EAqVqlw/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true>
5. Week 6 journal: <https://docs.google.com/document/d/16L732eWlPXcMvudJe7gZuIswh3unxha5/edit?usp=drive_link&ouid=103523606972182441044&rtpof=true&sd=true>

1. Data dictionaries are not needed for the Zip Codes and Healthcare Procedures datasets [↑](#footnote-ref-1)
2. The Capstone Reflection Journals are listed and linked in [Appendix 4](#_Appendix_4:_Reflection) of this document [↑](#footnote-ref-2)
3. This refers to Microsoft 365, which includes a version of Excel [↑](#footnote-ref-3)
4. Team 38 [↑](#footnote-ref-4)