CEE 492 Data Science Project

This manuscript (permalink) was automatically generated from uiceds/cee-492-term-project-fall-2022-her@7fa3c6d on September 22, 2022.

Authors

- **Hadil Helaly**
  55555-55-5555 · hadilhelaly · johndoe
  Department of Civil and Environmental Engineering · Funded by Grant XXXXXXXX

- **Emma Golub**
  XXXX-XXXX-XXXX-XXXX · emmaagolub
  Department of Civil and Environmental Engineering

- **Riley Blasiak**
  XXXX-XXXX-XXXX-XXXX · blasiak2
  Department of Civil and Environmental Engineering

- **Rupesh Rokade**
  XXXX-XXXX-XXXX-XXXX · RupeshRokade16
  Department of Civil and Environmental Engineering

Abstract

Data Science is easy?

This is the abstract.

Yes

This manuscript is a template (aka “rootstock”) for Manubot, a tool for writing scholarly manuscripts. Use this template as a starting point for your manuscript.

The rest of this document is a full list of formatting elements/features supported by Manubot. Compare the input (`.md` files in the `/content` directory) to the output you see below.

Basic formatting

Bold text

Semi-bold text

Centered text

Right-aligned text

Italic text

Combined *italics and **bold***

~~Strikethrough~~

1. Ordered list item
2. Ordered list item
    a. Sub-item
    b. Sub-item
        i. Sub-sub-item
3. Ordered list item
    a. Sub-item

- List item
- List item
- List item

subscript: H_2O is a liquid

superscript: 2^{10} is 1024.

unicode superscripts^{0123456789}

unicode subscripts$_{0123456789}$

A long paragraph of text. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Putting each sentence on its own line has numerous benefits with regard to editing and version control.

Line break without starting a new paragraph by putting
two spaces at end of line.

Document organization

Document section headings:

Heading 1

Heading 2

Heading 3

Heading 4

Heading 5

Heading 6

A heading centered on its own printed page

Horizontal rule:

`Heading 1`'s are recommended to be reserved for the title of the manuscript.

`Heading 2`'s are recommended for broad sections such as *Abstract*, *Methods*, *Conclusion*, etc.

`Heading 3`'s and `Heading 4`'s are recommended for sub-sections.

Links

Bare URL link: https://manubot.org

Long link with lots of words and stuff and junk and bleep and blah and stuff and other stuff and more stuff yeah

Link with text

Link with hover text

Link by reference

Citations

Citation by DOI [1].

Citation by PubMed Central ID [2].

Citation by PubMed ID [3].

Citation by Wikidata ID [4].

Citation by ISBN [5].

Citation by URL [6].

Citation by alias [7].

Multiple citations can be put inside the same set of brackets [1,5,7]. Manubot plugins provide easier, more convenient visualization of and navigation between citations [2,3,7,8].

Citation tags (i.e. aliases) can be defined in their own paragraphs using Markdown's reference link syntax:

Referencing figures, tables, equations

Figure 1

Figure 2

Figure 3

Figure 4

Table 1

Equation 1

Equation 2

Quotes and code

> Quoted text

> Quoted block of text
>
> Two roads diverged in a wood, and I—
> I took the one less traveled by,
> And that has made all the difference.

Code `in the middle` of normal text, aka `inline code`.

Code block with Python syntax highlighting:

```python
from manubot.cite.doi import expand_short_doi

def test_expand_short_doi():
    doi = expand_short_doi("10/c3bp")
    # a string too long to fit within page:
    assert doi == "10.25313/2524-2695-2018-3-vliyanie-enhansera-copia-i-
        insulyatora-gypsy-na-sintez-ernk-modifikatsii-hromatina-i-
        svyazyvanie-insulyatornyh-belkov-vtransfetsirovannyh-geneticheskih-
        konstruktsiyah"
```

Code block with no syntax highlighting:

```
Exporting HTML manuscript
Exporting DOCX manuscript
Exporting PDF manuscript
```

Figures

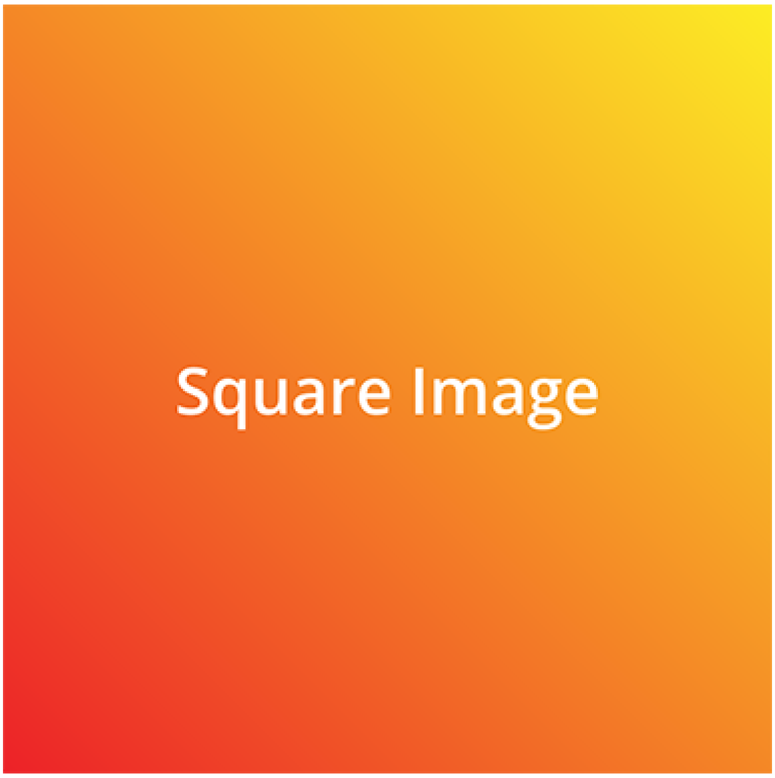

Figure 1: A square image at actual size and with a bottom caption. Loaded from the latest version of image on GitHub.

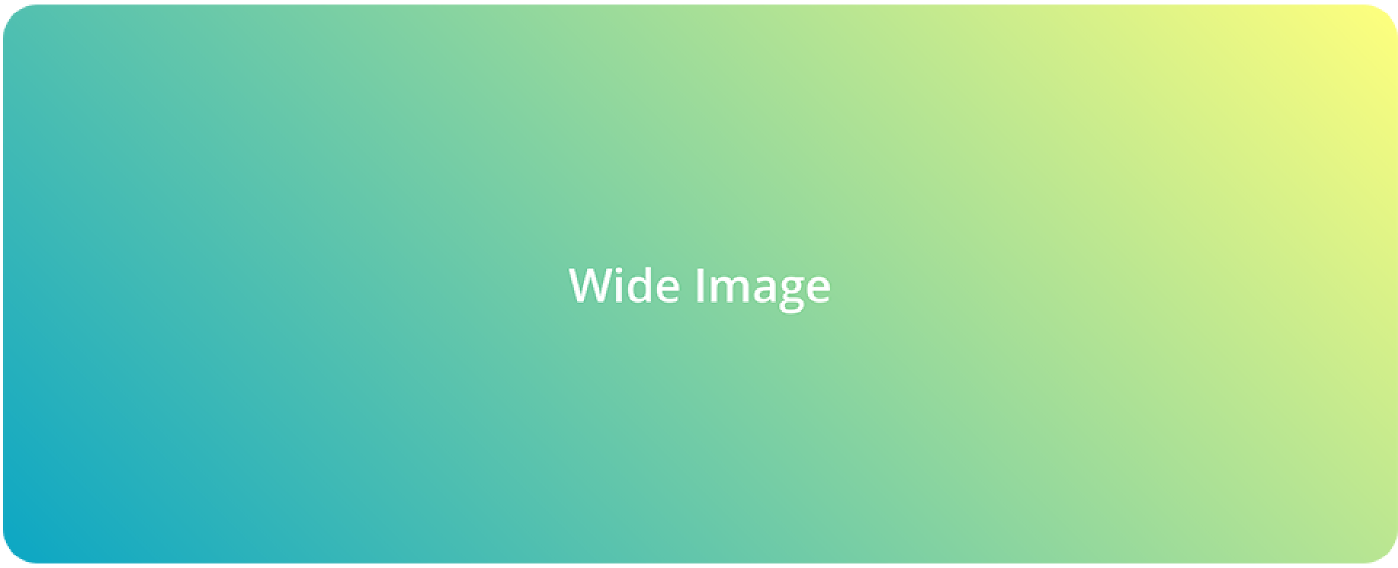

Figure 2: An image too wide to fit within page at full size. Loaded from a specific (hashed) version of the image on GitHub.

Figure 3: A tall image with a specified height. Loaded from a specific (hashed) version of the image on GitHub.

Figure 4: A vector `.svg` image loaded from GitHub. The parameter `sanitize=true` is necessary to properly load SVGs hosted via GitHub URLs. White background specified to serve as a backdrop for transparent sections of the image.

Tables

Table 1: A table with a top caption and specified relative column widths.

Bowling Scores	**Jane**	**John**	**Alice**	**Bob**
Game 1	150	187	210	105
Game 2	98	202	197	102
Game 3	123	180	238	134

Table 2: A table too wide to fit within page.

	Digits 1-33	**Digits 34-66**	**Digits 67-99**	**Ref.**
pi	3.14159265358979323 846264338327950	28841971693993751 0582097494459230	78164062862089986 2803482534211706	`piday.org`
e	2.71828182845904523 536028747135266	24977572470936999 5957496696762772	40766303535475945 7138217852516642	`nasa.gov`

Table 3: A table with merged cells using the `attributes` plugin.

	Colors	
Size	**Text Color**	**Background Color**
big	blue	orange
small	black	white

Equations

A LaTeX equation:

$$\int_0^\infty e^{-x^2} dx = \frac{\sqrt{\pi}}{2} \tag{1}$$

An equation too long to fit within page:

$$\begin{aligned} x = a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t \\ + u + v + w + x + y + z + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 \end{aligned} \tag{2}$$

Special

⚠ WARNING *The following features are only supported and intended for* `.html` *and* `.pdf` *exports. Journals are not likely to support them, and they may not display correctly when converted to other formats such as* `.docx`.

LINK STYLED AS A BUTTON

Adding arbitrary HTML attributes to an element using Pandoc's attribute syntax:

> Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Adding arbitrary HTML attributes to an element with the Manubot `attributes` plugin (more flexible than Pandoc's method in terms of which elements you can add attributes to):

> Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Available background colors for text, images, code, banners, etc:

`white` `lightgrey` `grey` `darkgrey` `black` `lightred` `lightyellow` `lightgreen` `lightblue` `lightpurple` `red` `orange` `yellow` `green` `blue` `purple`

Using the Font Awesome icon set:

✓ ? ★ 🔔 ⊗ ⋯

Light Grey Banner
useful for *general information* - manubot.org

Blue Banner
useful for *important information* - manubot.org

Light Red Banner
useful for *warnings* - manubot.org

Introduction

The Urban tree database, which was collected by the US Forest Service Research Archive of the US Department of Agriculture, includes data about tree growth in urban areas across 17 cities and 13 states over the span of 14-years (from 1998-2012). The states included in the study are: Arizona, California, Colorado, Florida, Hawaii, Idaho, Indiana, Minnesota, New Mexico, New York, North Carolina, Oregon, and South Carolina. The data come from measurements taken to over 14,000 street and urban park trees, and the data can be obtained by downloading the 1.08 MB compressed "data publication" file from here. Some measurements of interest include tree age, location, height, crown diameter, leaf area, foliar biomass, and utility line interference. Tree age, for example, was determined from interviews with residents, street construction dates, aerial and historical photos, the city's urban forester, and laboratory cores developed by the Lamont-Doherty Earth Observatory's Tree Ring Laboratory.

The downloaded folder includes 9 data sheets in CSV format. The most interesting data files are i)TS1_Regional_information.csv, ii) TS2_Regional_species_and_counts.csv, and iii) TS3_Raw_tree_data.csv. First, the "TS1_Regional_information.csv" file contains information about region code, city, state, airport codes, and collection year. Second, the "TS2_Regional_species_and_counts.csv" file contains information (columns) regarding region, scientific and common names of trees, tree type, and 9 columns of dbh_class, which represent a species diameter at breast height and are used to predict tree height, crown diameter, crown height, and leaf area. These 9 classes of dbh are stratified into the following groups: 3 inch and 6 inch classes: 0-3, 3-6, 6-12, 12-18, 18-24, 24-30, 30-36, 36-42, > 42 inches. The file contains a total of 347 rows. Finally, the "TS3_Raw_tree_data.csv" file includes 14487 observations (rows) of raw tree data. For each observation, 41 different variables were collected (columns). A detailed description of each of these 41 variables is as followed: 1. DbaseID = Unique id number for each tree. 2. Region = 16 U.S. climate regions, abbreviations are used - City = City/state names where data collected. - Source = Original *.xls filename (not available in this data publication). - TreeID = Number assigned to each tree in inventory by city. - Zone = Number/ID/name of the management area or zone that the tree is located in within a city; or nursery if young tree data collected there. - Park/Street = Data listed as Park, Street, Regional Big Tree, or Nursery (for young tree measurements). - SpCode = 4 to 6 letter code consisting of the first two letters of the genus name and the first two letters of the species name followed by two optional letters to distinguish two species with the same four-letter code (See _Regional_species_and_counts.csv for a list of the SpCodes and corresponding scientific names.) - ScientificName = Botanical name of species. - CommonName = Common name of species. - Tree Type = 3 letter code where first two letters refer to life form (BD=broadleaf deciduous, BE=broadleaf evergreen, CE=coniferous evergreen, PE=palm evergreen) and the third letter is mature height (S=small which is < 8 meters, M=medium which is 8-15 meters, and L=large which is > 15 meters). - Address = From inventory, street number of building where tree is located. - Street = From inventory, the name of the street the tree is located on. (NOTE: zero values denote data were not recorded in that city. These values were left unchanged because they originated from city inventories.) - Side = From inventory, side of building or lot tree is located on (F=front, M=median, S=side, P=park). (NOTE: zero values denote data were not recorded in that city. These values were left unchanged because they originated from city inventories.) - Cell = From inventory, the cell number (i.e., 1, 2, 3, ...), where protocol determines the order trees at same address are numbered (e.g., driving direction or as street number increases). - OnStreet = From inventory (omitted if not a field in city's inventory), for trees at corner addresses when tree is on cross street rather than addressed street. FromStreet = From inventory, the name of the first cross street that forms a boundary for trees lining un-addressed boulevards. Trees are typically numbered in order (1, 2, 3 ...) on boulevards that have no development adjacent to them, no obvious parcel addresses. - ToStreet = From inventory, the name of the last cross street that forms a boundary for trees lining un-addressed boulevards. - Age = Number of years since planted. (NOTE: zero values represent newly planted trees, < 1 year old.) - DBH (cm) = Diameter at breast height (1.37 meters [m]) measured to nearest 0.1 centimeters (tape). For multi-stemmed trees forking below 1.37 m measured above the butt flare and below the point where the stem begins forking, as per protocol. - TreeHt (m) = From ground level to tree top to nearest 0.5 m (omitting erratic leader). - CrnBase (m) = Average distance between ground and lowest foliage layer to nearest 0.5 m (omitting erratic branch). - CrnHt (m) = Calculated as TreeHT minus Crnbase to nearest 0.5 m. (NOTE: zero values indicate no live crown was present, hence no other tree dimension data were available.) - CdiaPar (m) = Crown diameter measurement taken to the nearest 0.5 m parallel to the street (omitting erratic branch). - CDiaPerp (m) = Crown diameter measurement taken to the nearest 0.5 m perpendicular to the street (omitting erratic branch). - AvgCdia (m) = The average of crown diameter measured parallel and perpendicular to the street. - Leaf (m2) = Estimated using digital imaging method to nearest 0.1 squared meter (m2). - Setback = Distance from tree to nearest air-conditioned/heated space (may not be same address as tree location): 1=0-8 m, 2=8.1-12 m, 3=12.1-18 m, 4=> 18 m. - TreeOr = Taken with compass, the coordinate of tree taken from imaginary lines extending from walls of the nearest conditioned space (may not be same address as tree location). - CarShade = Number of parked automotive vehicles with some part under the tree's drip line. Car must be present (0=no autos, 1=1 auto, etc.). - LandUse = Predominant land use type where tree is growing (1=single family residential, 2=multi-family residential [duplex, apartments, condos], 3=industrial/institutional/large commercial [schools, gov't, hospitals], 4=park/vacant/other [agric., unmanaged riparian areas of greenbelts], 5=small commercial [minimart, retail boutiques, etc.], 6=transportation corridor). - Shape = Visual estimate of crown shape verified from each side with actual measured dimensions of crown height and average crown diameter (1=cylinder [maintains same crown diameter in top and bottom thirds of tree], 2=ellipsoid, the tree's center [whether vertical or horizontal is the widest, includes spherical], 3=paraboloid [widest in bottom third of crown], 4=upside down paraboloid [widest in top third of crown]). - WireConf = Utility lines that interfere with or appear above tree (0=no lines, 1=present and no potential conflict, 2=present and conflicting, 3=present and potential for conflicting). (NOTE: -1 denotes data were not collected.) - dbh1 = Dbh (centimeters [cm]) for multi-stemmed trees; for non-multi-stemmed trees, dbh1 is same as Dbh (cm). - dbh2 = Dbh (cm) for second stem of multi-stemmed trees. - dbh3 = Dbh (cm) for third stem of multi-stemmed trees. - dbh4 = Dbh (cm) for fourth stem of multi-stemmed trees. - dbh5 = Dbh (cm) for fifth stem of multi-stemmed trees. - dbh6 = Dbh (cm) for sixth stem of multi-stemmed trees. - dbh7 = Dbh (cm) for seventh stem of multi-stemmed trees. - dbh8 = Dbh (cm) for eight stem of multi-stemmed trees.

Additionally, a fourth data set may be of later interest for estimating leaf area, species dominance at a spatial scale, and carbon storage estimates. The TS5_Foliar_biomass_leaf_samples.csv contains urban foliar samples data by species for 17 U.S. cities. The following variables (columns) are included, and a total of 261 rows are provided.

The breadth of this dataset allows for a myriad of problems to be explored. The primary data that will be utilized for this project is the "TS3_Raw_tree_data.csv" file, as this contains the most columns which will result in more feasible predictions during the machine learning portion of the project. This data can be used to analyze correlations between tree characteristics and their surroundings. One potential research question using the “TS3_Raw_tree_data.csv” file is: how does utility line interference affect the growth of a certain type of tree in one state versus a different state. the preliminery 14 variables that can be used in the proposed analysis include “Address”, “Age”, “Shape”, “WireConf”, “Setback”,“CarShade”, “DBH”, “TreeHt”, “CrnBas” “CrnHt”, “CdiaPar”, “CDiaPerp”, “AvgCdia”, “Leaf”.

After tidying the dataset, we can compare the effect of the WireConf, Setback, CarShade on the remaining variables of similar trees. Since we also contain the addresses of the trees, along with visualizing graphs from results of the comparisons, we can create maps to understand the variance of these effects across different cities. Further, a machine learning model can be created to possibly target and predict the : The above results for a city that is not mentioned in the dataset Predict the missing values in the dataset

Testing to see if commit completes - Emma

References

1. **Sci-Hub provides access to nearly all scholarly literature**
Daniel S Himmelstein, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, Casey S Greene
eLife (2018-03-01) https://doi.org/ckcj
DOI: 10.7554/elife.32822 · PMID: 29424689 · PMCID: PMC5832410

2. **Reproducibility of computational workflows is automated using continuous analysis**
Brett K Beaulieu-Jones, Casey S Greene
Nature biotechnology (2017-04) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/
DOI: 10.1038/nbt.3780 · PMID: 28288103 · PMCID: PMC6103790

3. **Bitcoin for the biological literature.**
Douglas Heaven
Nature (2019-02) https://www.ncbi.nlm.nih.gov/pubmed/30718888
DOI: 10.1038/d41586-019-00447-9 · PMID: 30718888

4. **Plan S: Accelerating the transition to full and immediate Open Access to scientific publications**
cOAlition S
(2018-09-04) https://www.wikidata.org/wiki/Q56458321

5. **Open access**
Peter Suber
MIT Press (2012)
ISBN: 9780262517638

6. **Open collaborative writing with Manubot**
Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter
Manubot (2020-05-25) https://greenelab.github.io/meta-review/

7. **Opportunities and obstacles for deep learning in biology and medicine**
Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, … Casey S Greene
Journal of The Royal Society Interface (2018-04) https://doi.org/gddkhn
DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

8. **Open collaborative writing with Manubot**
Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter
PLOS Computational Biology (2019-06-24) https://doi.org/c7np
DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653