**Final report**

* Project title with student’s names
* Problem definition (be specific)
* The primary challenge is informing the California Insect Barcoding Initiative where to collect data in order to efficiently utilize their limited sampling resources
* Description of background:
  + Why is this meaningful?
  + What does it solve?
  + Related work?
  + Contribution?
* Description of dataset
  + Quantity (adequate size?)
  + Quality
  + Explain important attributes
* Description of methods
  + Description
  + Why was it used? (fine if saying it is the one you know…)
* Experiment setup and analysis of results
  + Select and describe *the most important* result
* Observation and conclusion
  + Good or bad, describe your understanding of the conclusion
* References
  + Half page is enough
  + Follow some standard way
  + Should be referenced in text

**Final Report (65% of project)**  
Due on **Dec 4, 11:59 PM**  
Max 8 page report + codes Submit a single zip file with the name, Final-proj-“team#”, which include report an codes.  
Late penalty: 30% deduction each day   
**Project Final Report** should include the followings   
Project title with student’s names   
Problem definition  
Description of background (why is this meaningful? What to solve? Related work?   
Contribution?)  
Description of dataset (quantity, quality)   
Description of methods used  
Experiment: experiment setup and analysis results  
Observation and Conclusion   
References

Ideas:

* Linear regression between # of individuals and human population, show relationship and explain how this can be used to encourage exploration of a causal relationship (use correlation coefficient and r^2 metric) (basically want to see if there’s a sampling bias by actually investigating if mantis biodiversity is caused by human density) **ARIELLE**
* Decision tree based on features (one tree per species) (one tree introduced, one tree native) **CHARLES**
* Nearest Neighbor - start with complete linkage and assess: (cluster by introduced vs. native) **DYLAN**
* Color by native vs. introduced
* Compare centroids
* Try to figure out which features are most explanatory
* Null result (no clustering) could imply further research needs to be done into competition
* ER Diagram for dataset description **BRIAN**
* Most of the writing: **BRIAN**
* Find text and references to work with here: <https://docs.google.com/document/d/12VbrE-YWvrlu3IxT_5lekDCCcTvvDTN6IUaSIISdXRM/edit>
* Apocalypse citations and other citations: <https://docs.google.com/document/d/1_MGv_APSpMvFkHJlPXYtec0vtNR1i7vwHe1lByIBFok/edit>

Problem Definition: The primary challenge is informing the California Insect Barcoding Initiative where to collect data in order to efficiently utilize their limited sampling resources

**Progress report:**  
  
Biodiversity: Dylan

Human population density: Arielle

Climatological: Charles

Ecoregion: Brian

Have your data cleaned and a map done for your datatype by Thursday (10/26) 2-3pm so we can decide on next steps

To-do:

Biodiversity data:

* GBIF California Mantodea:
  + <https://www.gbif.org/occurrence/search?taxon_key=788&occurrence_status=present&gadm_gid=USA.5_1>
* Cleaning steps:
  + Remove samples without lat
  + Remove samples with lon
  + Remove samples without species name (where species name is “sp.” or “spp.”
  + Remove samples derived from iNaturalist that are not “Research Grade”
* Heatmap by county (if you can figure it out, Getis-Ord Gi\* is better than KDE for heatmaps)
  + County data can be found here (GADM): <https://geodata.ucdavis.edu/gadm/gadm4.1/shp/gadm41_USA_shp.zip>
    - Will need to clip for just California
* Produce a table with county as the primary key

Human population data:

* Look up YouTube video describing shapefiles
* Data source: census API: <https://www.youtube.com/watch?v=l47HptzM7ao>
  + I think 2020 is probably the year we should pull from, last decennial census
* Heatmap by county (if you can figure it out, Getis-Ord Gi\* is better than KDE for heatmaps)
  + County data can be found here (GADM): <https://geodata.ucdavis.edu/gadm/gadm4.1/shp/gadm41_USA_shp.zip>
    - Will need to clip for just California
* Produce a table with county as the primary key

Ecoregion data:

* Data source:
  + <https://www.epa.gov/eco-research/ecoregion-download-files-state-region-9#pane-04>
    - Choose between level 3 or level 4
* Heatmap by county (if you can figure it out, Getis-Ord Gi\* is better than KDE for heatmaps)

<https://data.census.gov/map?t=Populations+and+People&g=040XX00US06$0500000&tid=ACSDT5Y2020.B01003&cid=B01003_001E&layer=VT_2020_050_00_PY_D1&palette=Blues&mode=thematic&loc=37.4147,-120.1656,z5.0331>

* + County data can be found here (GADM): <https://geodata.ucdavis.edu/gadm/gadm4.1/shp/gadm41_USA_shp.zip>
    - Will need to clip for just California
* Produce a table with county as the primary key

Climatological data:

* ~~Citations~~
  + <https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=2131>
  + <https://daymet.ornl.gov/overview>
  + <https://www.unidata.ucar.edu/software/tds/?query_float=OPeNDAP>
  + <https://thredds.daac.ornl.gov/thredds/catalog/ornldaac/2131/catalog.html>

Analyses from class notes

* Make ER diagram of data
* Relational schema that is in 3NF

Project Proposal Report (20% of project)

- Due on Sep 29, 11:59 PM

- 2 page report

- Submit a PDF file with the name, Pre-proj-”team#”.pdf. Only one submission by

the team leader.

- Late penalty: 30% deduction each day

Project Proposal Report should include the followings:

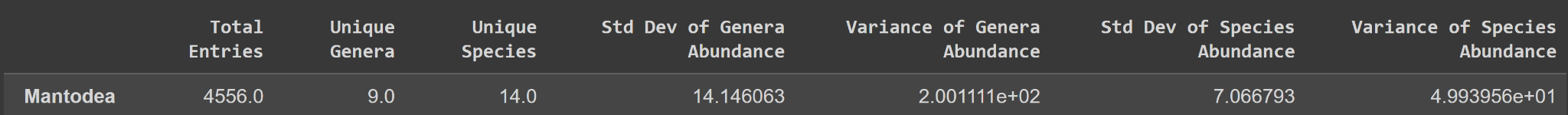
* ~~• Project title with student’s names~~
* ~~• Project idea: what you want to do and achieve. Make the description as clear as possible.~~
* ~~• Description of dataset (what to collect and how, size, content, why the dataset is appropriate)~~
* ~~• Project plan: how you do the project with the description of system, software, and methodology you will need to use.~~
* ~~• Some relevant articles or papers as reference. (No more than 5)~~
* ~~• Teammates and work division.~~

Moths

History / background

Sphingidae

Notes:

* Biodiversity data
  + <https://www.gbif.org/>
  + <https://www.gbif.org/what-is-gbif>
  + Mantises (members of order Mantodea)
    - Initially we selected 7 insect groups (Lepidopterans, Sphingids, Odonates, Orthopterans, Blattodeans, Mantises, and Carabids) that include species simple enough for amateur biologists to identify so that data is accurate. By comparing these groups through preliminary analyses, we selected mantises (Mantodea) for this project because records are relatively uniformly distributed geographically and between species.
  + <https://en.wikipedia.org/wiki/Mantis>
  + 
  + \
  + **We will also need a map of where CIBI has already sampled, I know Austin has this data in a form of a Shiny app**
* Population density data
  + <https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html>
  + <https://en.wikipedia.org/wiki/Topologically_Integrated_Geographic_Encoding_and_Referencing>
* Ecoregion data
  + <https://www.epa.gov/eco-research/ecoregion-download-files-state-region-9#pane-04>
  + Note from Austin “. I'm working with both level 3 and level 4 ecoregions (level 4 is nested within level 3), but you may want to choose just one for your project. ”
* Climatological data
  + Various sources? (temperature, humidity, solar radiation, elevation, distance from the ocean, precipitation, etc?)
  + <https://docs.opendata.aws/noaa-ghcn-pds/readme.html>

Other resources

* <https://nhm.org/stories/bugs-life#:~:text=NHM%20researchers%20and%20their%20collaborators,huge%20goal%2C%E2%80%9D%20says%20Dr>.
* <https://www.nytimes.com/2018/11/27/magazine/insect-apocalypse.html>

Notes:

* Data cleaning:
  + Remove records not identified to species level
  + Remove samples missing lat/long
  + Remove samples that are not iNaturalist research grade (<https://www.gbif.org/dataset/50c9509d-22c7-4a22-a47d-8c48425ef4a7#:~:text=They%20become%20%22Research%20Grade%22%20when,of%20the%20voters%20agree%20with>.)
  + Remove iNaturalist samples for groups we think amateurs could not identify with accuracy