CARNEGIE MELLON UNIVERSITY
BIG DATA SCIENCE (COURSE 18-899/K4)
ASSIGNMENT 3

INSTRUCTIONS

- Submissions should be made via canvas.
- **Single** Python/MATLAB code file(.ipynb or .m) **[Do not Submit checkpoints for .ipynb].** In addition, each line of code should be documented by text. This demonstrates that the code is unique and owned by the student
- Assignment report(.pdf) with full evidence that the assignment was completed by the student and demonstrate a full understanding of each step in the process including textual descriptions of each result (statistics, table, graph etc) represents and insights that can be gained
  - Indicate the libraries you have used in your code at the beginning of the report (After the title page)
- Data files (as given)

Submission process:

1. Put source code **file and data files** in a single folder
2. Name of the folder should be the same as your andrew ID
3. **Zip this folder and attach the zipped file on assignment submission page (CANVAS)**
4. After attaching zipped file, click on "Add Another File" from assignment submission page and **attach your report**
5. Submit your assignment

N.B. This new process will allow us to compile your reports in **Turnitin** to check for plagiarism.

Specific reasons for a submission being classified as incomplete include:

- Failure to correctly name your folder with your Andrew ID, report, and code file with andrewID-BDS-AssignmentNo. For example, mcsharry-BDS-Assignment1, mcsharry-BDS-Assignment2 and mcsharry-BDS-Assignment3.
- A missing report describing the steps, results, and insights
- A missing dataset required for running the code
- A missing code file such as .ipynb or .m file
- An error in the file path needed to run the code

The student is responsible for checking that their submission is complete. Students will lose 10% as for usual late submission even if the submission is repaired during the 24 hours after the deadline has passed, and receive 0 for the assignment if it is not repaired.

The submission deadline is **Rwandan Time (CAT) on Monday 02, May, 2022 / 23:59**.

Questions

1. Follow the instructions from CMU computing services to download and set up ArcGIS on your PC. https://www.heinz.cmu.edu/current-students/computing-services/software.

Alternatively, you could use the online version of ArcGIS that CMU offers by logging in with your Andrew account details at https://carnegiemellon.maps.arcgis.com/.

2. Download the required datasets

a. Download the nightlight map file containing night light data for Rwanda from the Google Drive link: https://drive.google.com/file/d/1Ii3RjUjOmwjL4AqpECXsRqGD8WVSaCKR/view?usp=sharing

b. Download the Rwanda country and sector administrative boundary maps in shapefile format from the Google Drive link:

https://canvas.cmu.edu/files/8009479/download?download_frd=1

c. Download MPIAssignment.xlsx containing population and MPI values for the sectors in Rwanda from the Google Drive link:

https://canvas.cmu.edu/files/8018006/download?download_frd=1

3. Open ArcGIS Pro and create a new blank project named (BigDataAssign3) and insert a new map in the project.

4. Load Rwanda admin level maps

a. Extract the country and sector shapefiles from the respective zip files downloaded in step 2.b.

b. Add the country and sector shapefiles as layers in the map created in step 3 (Map -> Add data)

5. Process nightlights

a. Extract the nightlight .tif raster data file from the .tgz compressed file downloaded in step 2.a.

b. Add the nightlight .tif file as a layer in the map created in step 3 (Map -> Add Data).

c. Clip the nightlight layer using the country layer to yield the nightlights for Rwanda. (Input Raster: nightlight layer and Output Extent: country boundary). Use the Analysis-> Tools->Geoprocessing-> and search for clip (data management) with the default No Data value. Also select the “Use Input Features for Clipping Geometry” box. Name the new layer

rw_nightlight_2015_07 (in the Output Raster Dataset). (Deselect the original nightlight and country boundary layers to make this new layer visible)

d. Set negative values to zero from the rw_nightlight_2015_07 raster layer created in step 5.c. above using the raster calculator in the geoprocessing, search for raster Calculator (Spatial Analyst Tools) and use the Map Algebra expression **("rw_nightlight_2015_07" >= 0)* "rw_nightlight_2015_07"**. Create a new layer and name it rw_nightlight_2015_07_nonneg.

e. Visually analyze the map. It may help to open google maps in your browser separately. **Note where the bright spots occur.** Do they make sense?

f. Calculate nightlight statistics for each sector in Rwanda. Use the rw_nightlight_2015_07_nonneg layer and Sector map as input to the "Zonal Statistics as Table" tool (Spatial Analyst Tools -> Zonal -> Zonal Statistics as Table). Use the FID field from the Sector map as the Zone Field. Name the output table rw_nightlight_2015_ZonalSt.

g. Export the nightlight statistics to excel. Use the Table to Excel tool (Conversion Tools -> Excel -> Table to Excel) from rw_nightlight_2015_ZonalSt. Save the output into rw_nightlight_2015_07.xls.

h. Copy the nightlight data and insert it into the nightlight sum column in MPIAssignment.xlsx

6. Analyze the data

a. Load the MPIAssignment.xlsx into your programming environment

b. Plot histograms of each of the features and the dependent variable

i. Are the variables normally distributed?

c. Create scatter plots of the mpi (dependent variable) vs each of the features

i. Are the relationships between the features and the dependent variable linear?

ii. Are there significant outliers?

d. Calculate the following correlations for each feature (Xi) with the MPI (y):

i. X vs y

ii. log X vs y

iii. X vs log y

iv. log X vs log y

v. Which are the strongest correlations for each feature?

7. Create final features.

a. We ensure that our features are intuitive. Since the MPI (y) has a population and an area component we create versions of our features that take these components into consideration. This prevents either the area or population overly influencing any feature. This will also serve to normalize the variables. Create the following features:

i. nightlight_per_capita: nightlight_sum / landscan_pop

ii. population_density: landscan_pop / area

b. Plot histograms of each of the features and the dependent variable

i. Are the features normally distributed?

c. Calculate the following correlations for each feature (Xi) with the MPI (y):

i. X vs y

ii. log X vs y

iii. X vs log y

iv. log X vs log y

v. Which are the strongest correlations for each feature?

8. Build the model

a. Using the strongest correlations from the previous question, check if the features we have selected are significant in explaining the MPI

i. Using backward-stepwise, ridge-regression and elastic nets:

- What is the p-value of each feature?**1point** Are all features significant? At what levels?**1point**
- What is the overall p-value of the model?**1point** At what levels is it significant?**1point**

[Building a mdl = **1point** as well total for each model is 5points x 3 = 15 points]

9. Evaluate the model

a. Using LASSO, calculate the estimated MPI (log yhat) for each sector

b. Calculate the correlation of log yhat to log y. What does this tell us about the model?

c. Calculate the R-squared of this result. What does the R-squared tell us about the model?

10. Visualize the results. Add the estimated MPI in the MPIAssignment.xlsx then in ArcGIS:

a. Load the “Sector_table$” dataset in the MPIAssignment.xlsx file as a layer

b. Create 2 sector layers one to display the original MPI and the other to display the estimated MPI

c. Style the two layers using quantile classification based on the respective MPI values. Use 10 quantiles for classification i.e. deciles.

d. Do you see similarities between the two maps?