BIG DATA ASSIGNMENT 3 REPORT

Nathalie Uwamahoro

CMU_AFRICA

BIG DATA ASSIGNMENT 3 REPORT

QUESTION 1 TO QUESTION 5 D

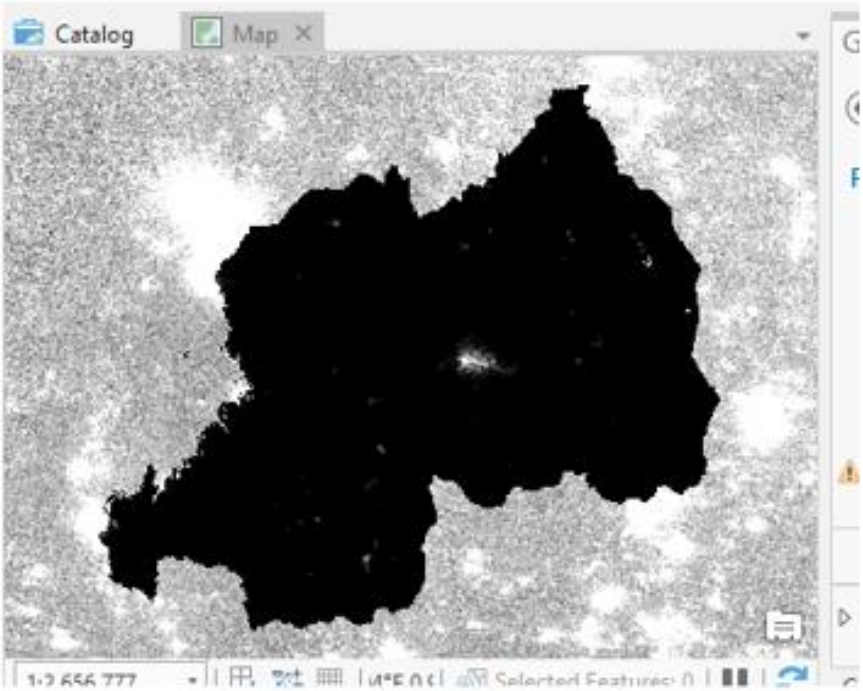
To solve question 1 until question 5 d. I followed the instruction to do each question as they instruct.

The Required created layers and tables were saved in ARCGIS in VMWARE horizon client.

For more proof, please check the uploaded ARCGIS folder.

QUESTION 5 e.

To visualize the map the following map was created in ARCGIS



Noting where the brightness occurs

By referring to Rwanda google map

Map of Rwanda with districts and sectors

The brightness occurs in the city of Kigali and nearby cities in the country side.

Do they make sense

Yes, the brightness make sense because the brightest spot occurs in Kigali city where the development and socio economic are high.

QUESTION 5.F, G, H

By following the instructions listed in the assignment. I have created the necessary table to use to the next questions. Please for more details, check the ARCGIS folder.

QUESTION 6: ANALYZING THE DATA

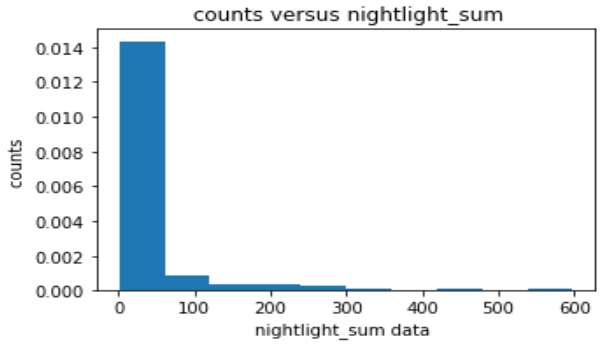
6.a Loading the MP assignment data

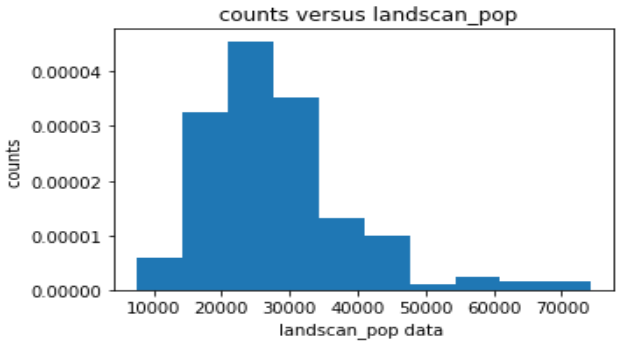
The following data was logged into Python.

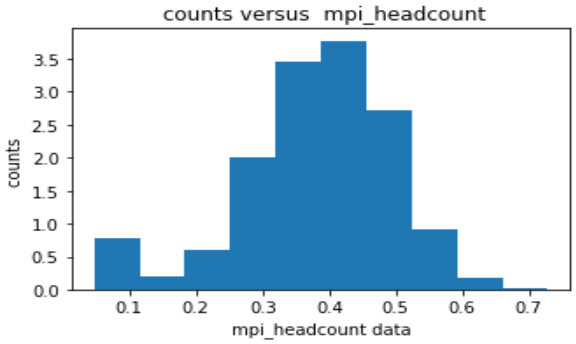
The following table is a sample of MPI assignment data:

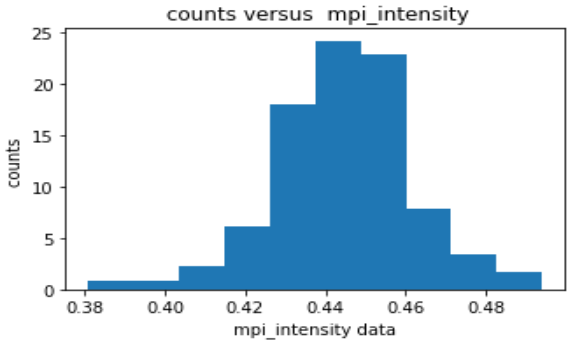
FID	Prov_ID	Province	Dist_ID	District	Sect_ID	Sector	nightlight_sum	landscan_pop	mpi_headcount	mpi_intensity	mpi
0	1	Kigali City	11	Nyarugenge	1101	Gitega	101.297202	30758	0.064	0.412	0.027
1	1	Kigali City	11	Nyarugenge	1102	Kanyinya	85.060422	19802	0.282	0.445	0.126
2	1	Kigali City	11	Nyarugenge	1103	Kigali	133.991142	26452	0.212	0.434	0.092
3	1	Kigali City	11	Nyarugenge	1104	Kimisagara	161.933015	62266	0.081	0.409	0.033
4	1	Kigali City	11	Nyarugenge	1105	Mageregere	49.415699	23144	0.369	0.430	0.159

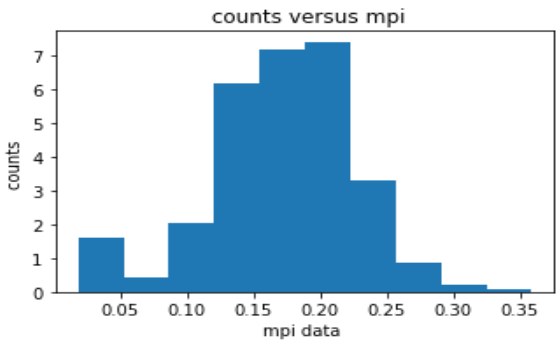
6.b plotting histogram of each features and the dependent variables.











6.i Are the variables normally distributed

Histogram of nightlight sum:

As seen from its graph, the data are skewed to the right, so it is not a normal distribution.

Histogram for landscan_pop:

It is not normal distribution since its distribution is skewed to the right.

Histogram for mpi_headcount is not also a normal distribution as there is not any symmetry in its graph.

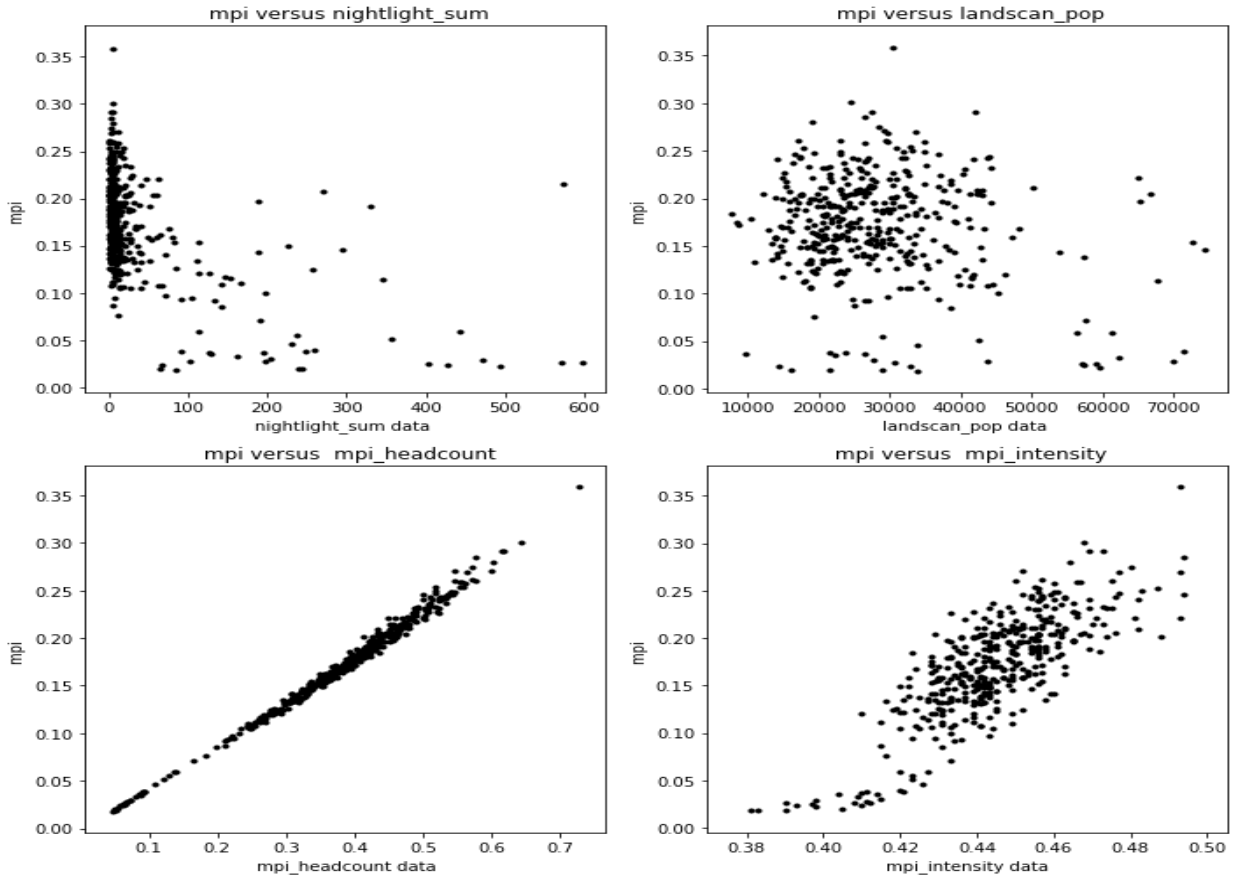
Histogram of mpi_intensity:

By vision the distribution is almost a normal distribution, but not a perfect bell curve hence this histogram is not normal distribution.

Histogram of mpi:

This histogram is not normal distribution because there is no symmetry in the data as the graph shows.

6.c. Create a scatter plots of the mpi vs each feature



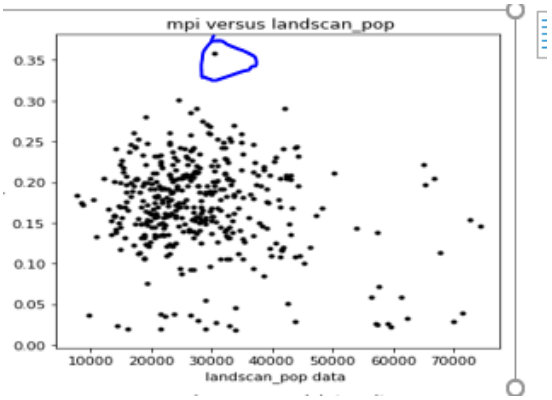
6.c.i. Relationship between each explanatory variable and dependent variable

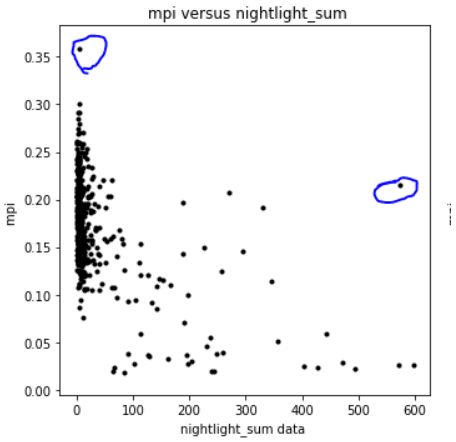
mpi with mpi_headcount scatter plot is the only linear relationship that is clearly visible from the graph. However, other scatter plot for other variables the relationship is not linear.

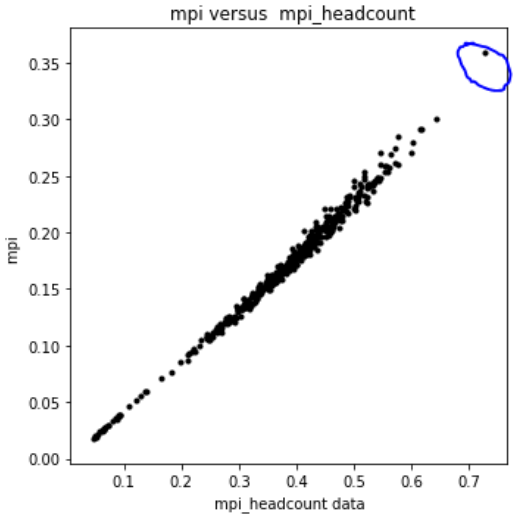
6.c. ii. Are there significant outliers

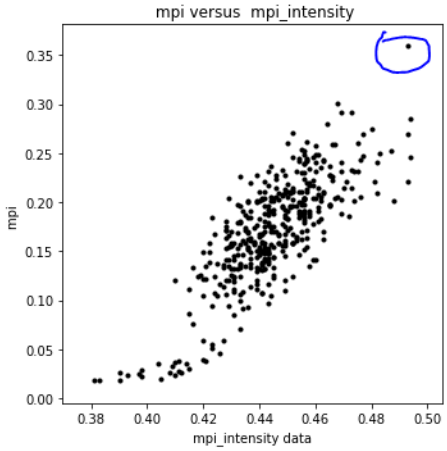
Yes, in every scatter plot. They are significant outliers.

Blue circles are the examples of the outliers in different plots, because they are away of other dots in the plotted graph.









6.d. Calculation of the correlation between:

- X vs y
- log X vs y
- X vs log y
- log X vs log y v

case i : correlation between x and y

	nightlight_sum	landscan_pop	mpi_headcount	mpi_intensity	mpi
nightlight_sum	1.000000	0.447506	-0.470580	-0.196437	-0.449116
landscan_pop	0.447506	1.000000	-0.050481	0.070575	-0.035350
mpi_headcount	-0.470580	-0.050481	1.000000	0.690300	0.994479
mpi_intensity	-0.196437	0.070575	0.690300	1.000000	0.754426
mpi	-0.449116	-0.035350	0.994479	0.754426	1.000000

Case ii: correlation between log x and y

	nightlight_sum	landscan_pop	mpi_headcount	mpi_intensity	mpi
nightlight_sum	1.000000	0.447506	-0.470580	-0.196437	-0.449116
landscan_pop	0.447506	1.000000	-0.050481	0.070575	-0.035350
mpi_headcount	-0.470580	-0.050481	1.000000	0.690300	0.994479
mpi_intensity	-0.196437	0.070575	0.690300	1.000000	0.754426
mpi	-0.449116	-0.035350	0.994479	0.754426	1.000000

Case iii: x vs log y

	nightlight_sum	landscan_pop	mpi_headcount	mpi_intensity	mpi
nightlight_sum	1.000000	0.494108	-0.552686	-0.373522	-0.638927
landscan_pop	0.494108	1.000000	-0.193782	-0.036433	-0.223342
mpi_headcount	-0.552686	-0.193782	1.000000	0.750575	0.942200
mpi_intensity	-0.373522	-0.036433	0.750575	1.000000	0.769113
mpi	-0.638927	-0.223342	0.942200	0.769113	1.000000

Case iv .log x vs log y

	nightlight_sum	landscan_pop	mpi_headcount	mpi_intensity	mpi
nightlight_sum	1.000000	0.447506	-0.470580	-0.196437	-0.449116
landscan_pop	0.447506	1.000000	-0.050481	0.070575	-0.035350
mpi_headcount	-0.470580	-0.050481	1.000000	0.690300	0.994479
mpi_intensity	-0.196437	0.070575	0.690300	1.000000	0.754426
mpi	-0.449116	-0.035350	0.994479	0.754426	1.000000

v. Which are the strongest correlations for each feature.

Original feature	**Selected feature from this case**	**Strongest correlation**
nightlight sum	x vs logy	**-0.638927**
landscan_pop	X vs log y	**-0.223342**
mpi_headcount	Log x vs logy, log x vs y, x and y	**0.994479**
mpi_intensity	x vs log y	**0.769113**

Question 7

Creation of final features

7.a.i creation of nightlight_per_capita

nightlight_per_capita = nightight_sum / landscan_pop

The following is the sample of the data created:

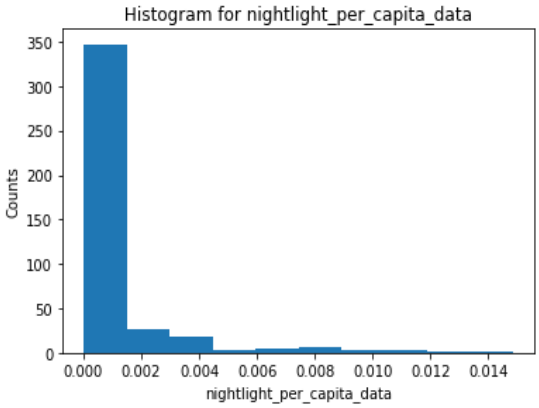
FID	nightlight_sum	landscan_pop	mpi_headcount	mpi_intensity	mpi	nightlight_per_capita	population_density
0	101.297202	30758	0.064	0.412	0.027	0.003293	3.556028e+08
1	85.060422	19802	0.282	0.445	0.126	0.004296	1.040623e+07
2	133.991142	26452	0.212	0.434	0.092	0.005065	1.116131e+07

7.a. ii. Creation of population density

population _density = landscan_pop / area

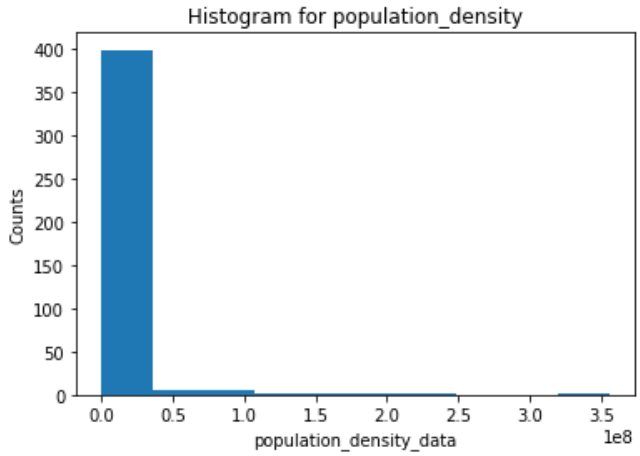
FID	nightlight_sum	landscan_pop	mpi_headcount	mpi_intensity	mpi	nightlight_per_capita	population_density
0	101.297202	30758	0.064	0.412	0.027	0.003293	3.556028e+08
1	85.060422	19802	0.282	0.445	0.126	0.004296	1.040623e+07
2	133.991142	26452	0.212	0.434	0.092	0.005065	1.116131e+07

7.b. Plotting histograms for each recreated feature and the dependent variables



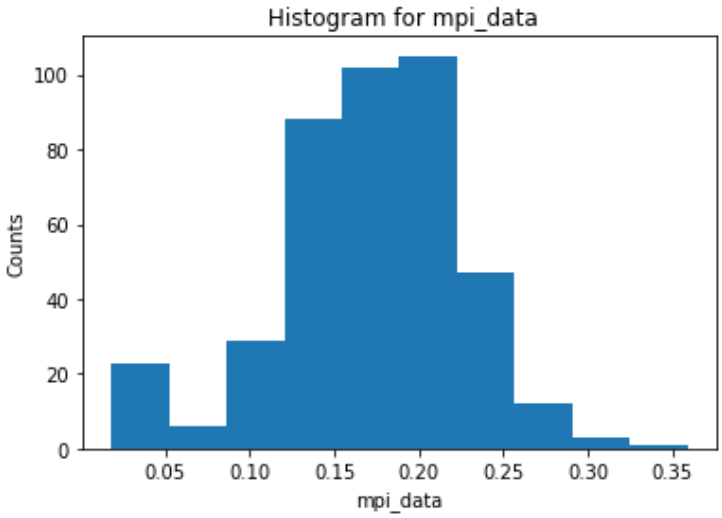
The above histogram shows that nightlight_per_capita data is not normally distributed. Instead, the graph is skewed to the right.

Histogram for population density



The histogram for population density is not normal distribution.

Histogram for dependent variable



This graph is fairly a normal distribution. Hence it is not a normal distribution because there is no symmetry in its graph.

7.c. calculation of the following correlation

i. X vs y

	nightlight_per_capita	population_density	mpi
nightlight_per_capita	1.000000	0.370426	-0.546978
population_density	0.370426	1.000000	-0.487136
mpi	-0.546978	-0.487136	1.000000

ii. log X vs y

:

	nightlight_per_capita	population_density	mpi
nightlight_per_capita	1.000000	0.431585	-0.605358
population_density	0.431585	1.000000	-0.617437
mpi	-0.605358	-0.617437	1.000000

iii. X vs log y

	nightlight_per_capita	population_density	mpi
nightlight_per_capita	1.000000	0.370426	-0.660497
population_density	0.370426	1.000000	-0.668281
mpi	-0.660497	-0.668281	1.000000

iv. log X vs log y

	nightlight_per_capita	population_density	mpi
nightlight_per_capita	1.000000	0.431585	-0.638304
population_density	0.431585	1.000000	-0.745331
mpi	-0.638304	-0.745331	1.000000

v. Which are the strongest correlations for each feature?

Original variables	Selected variables for each feature	Strongest Correlation coefficient
nightlight_per_capita	Logy_x_data, case iii	-0.660497
Population_density	log x and logy ,case iv	-0.745331

QUESTION 8

The strongest correlations from Q7 are:

- Nightlight _per capita
- Log of population density

Building model by using backwise-stepwise,ridge regression,and elastic nets

Back wise stepwise

Table that shows the P value for each feature

	coef	std err	t	P>|t|	[0.025	0.975]
population_density	0.0270	0.000	65.676	0.000	0.026	0.028
nightlight_per_capita	-14.8427	1.127	-13.174	0.000	-17.057	-12.628

Are the features significant

The levels where the variables are significant

Yes, since the p value of all features are less than significant level. Hence all features are statistically significant All features are significant at 0.05 significant level.

The p_value of the whole model

By using print(stats.coef_pval(Model,SelectedExplanarVars,y))

The overall p-value of the stepwise model is 2.22044605e-16 and it is significant level at 0.05.

Ridge regression

```
Coefficients:
            Estimate  Std. Error  t value  p value
_intercept  0.716943    0.040212  17.8291      0.0
x1         -0.078335    0.001128 -69.4202      0.0
x2         -8.602184    0.957748  -8.9817      0.0
---
R-squared:  0.48171,    Adjusted R-squared:  0.47920
F-statistic: 191.92 on 2 features
```

At 0.05 level

Are the values significant :

yes ,all values are significant since their values are zero which less than 0.05 significant level.

The P-value of the overall model

The overall p-value of the stepwise model is 2.22044605e-16 and it is significant level at 0.05.

Elastic nets

```
Coefficients:
            Estimate  Std. Error  t value   p value
_intercept  0.845178    0.042734  19.7776  0.000000
x1         -0.098164    0.001199 -81.8579  0.000000
x2         -1.574787    1.017821  -1.5472  0.122574
---
```

No, all variables are not significant at 0.05.

X2 which is **logarithm of population density** is not significant for elastic model, because its p values is greater than 0.05.

The P value for the elastic model

0.12257377. This value is not statistically significant.

QUESTION 9

Evaluate the model by using Lasso

By using Lasso cross validation estimates the log Yhat.

The sample values of the estimated log mpi is as follow:

```
array([-1.50953564, -0.99212808, -1.03392861, -1.40959851, -0.80215317,
       -1.59207851, -1.30565927, -1.21998273, -1.65835294, -1.43682069,
       -0.97001496, -1.21996007, -0.77751629, -1.33211044, -1.00494553,
       -0.91231141, -1.491587  , -1.66107454, -1.43798058, -1.28116993,
       -1.09778899, -0.84830447, -1.55690933, -0.9704822 , -0.7735689 ,
       -0.96805061, -1.26271746, -1.46305238, -1.39304392, -1.29463283,
       -1.42030768, -1.32657223, -0.94574101, -1.4584176 , -1.4497833 ,
       -0.96921713, -0.75079374, -0.71192333, -0.72435469, -0.78297225,
       -0.811974  , -0.74614513, -0.74601055, -0.71114087, -0.71398472,
       -0.74943065, -0.72344375, -0.72908673, -0.82544135, -0.7595134 ,
       -0.74686043, -0.7400909 , -0.73291238, -0.76050952, -0.7560647 ,
       -0.81240834, -0.72646613, -0.91200291, -0.72512394, -0.74253314,
       -0.76038427, -0.53839183, -0.62444884, -0.58988611, -0.71037935,
       -0.71336235, -0.76501676, -0.5583333 , -0.74944095, -0.67354689,
       -0.73603931, -0.71583483, -0.71323356, -0.78855373, -0.69879669,
```

9.b.Calculate the correlation of log yhat to logy

The correlation coefficient obtained is approximately 0.83792371.

What this value tell us :

Is that all explanatory variables used are the best candidates to build and predict the model.

The calculated correlation is good close to 1.

9.c. What does this R-squared tell us

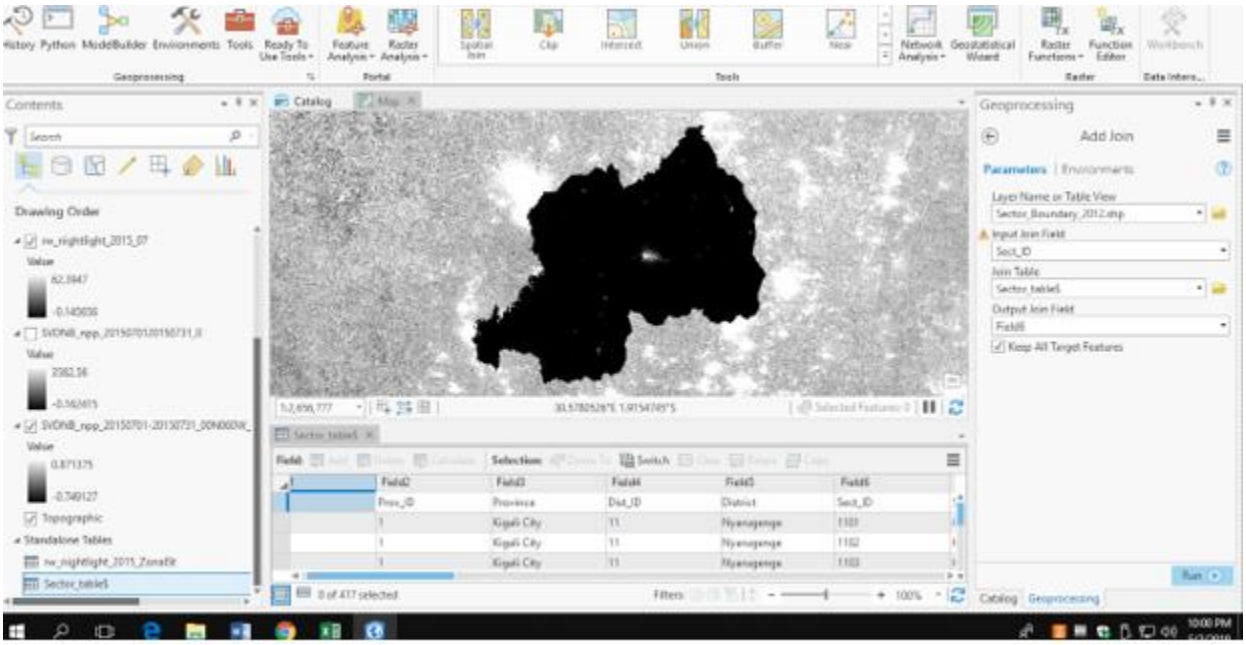
0.7 R-squared shows that 70% of the observed variance can be observed by the model's input. In addition, since the model R-squared is close to 1. Probably, the model is predicting the log mpi very well.

QUESTION 10

Add the estimated MPI in the MPIassignment.xlsx then in ARCGIS

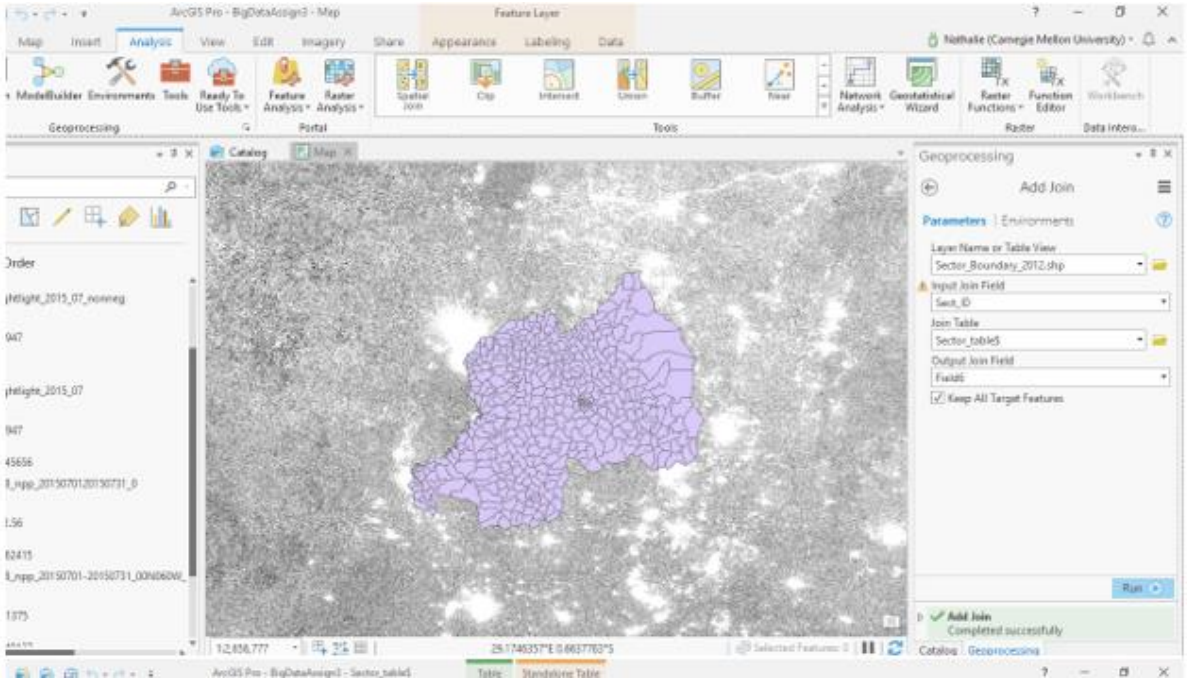
Please refer to the uploaded data to check for this step.

10.a. Loading sectors table$ in dataset MP Assignment.

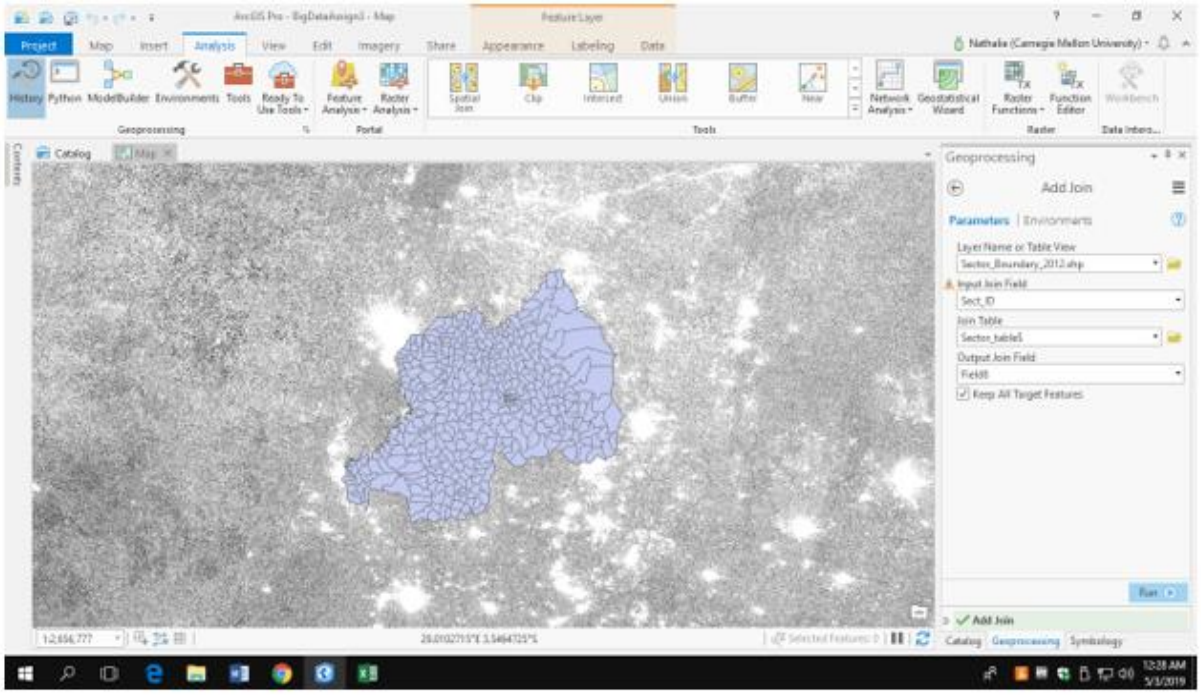


Selecting layer by following the instructions given on the Piazza.

Creation and selecting layer for sector_2012 actual mpi.



Creation and selecting layer for estimated values.

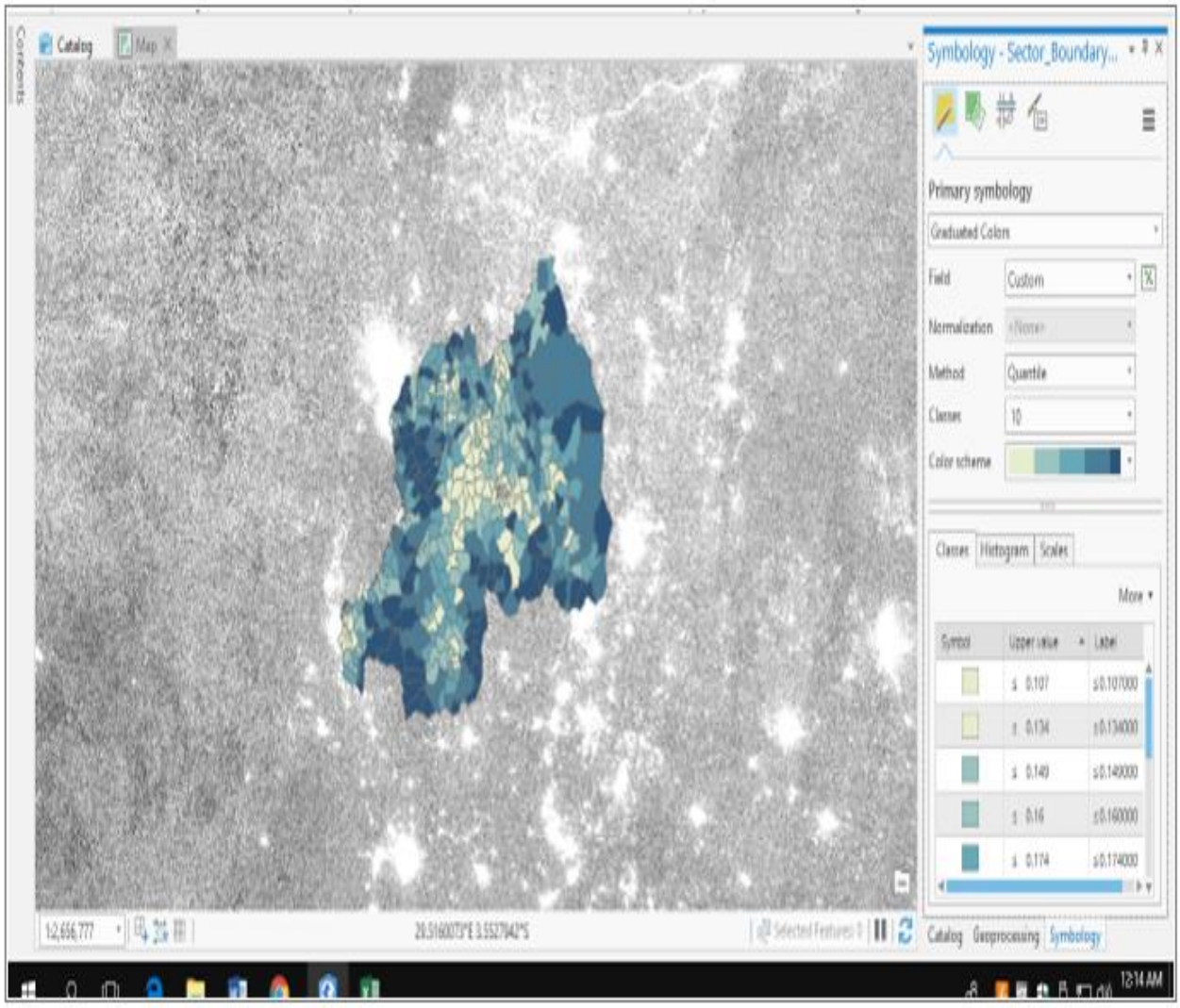


Question 10.c :

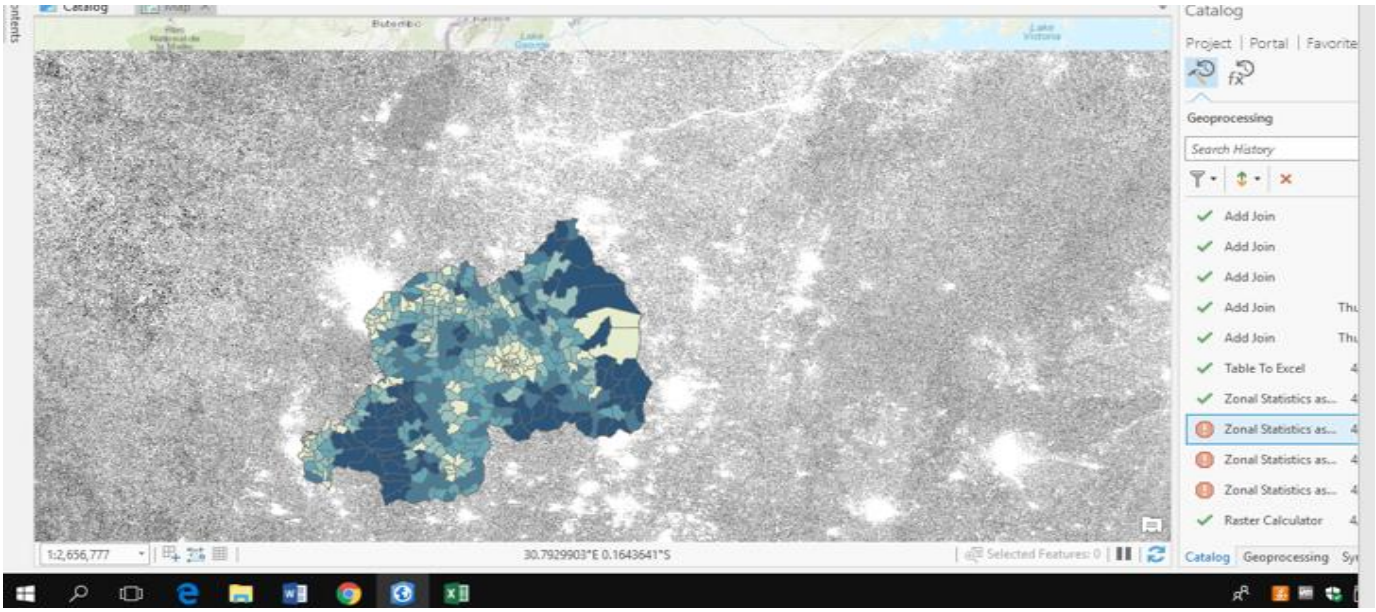
Style it by using symbology.

And select 10 Quantiles for exact sector layer

Case 1: for actual mpi values



Case2: Sector _boundaries for estimated mpi



10.d. Spot the similarities between two maps

They are some similarities, even though, they are not enough.

Following points represent the similarities between two maps:

- The west part of Rwanda near Cyangungu , Rusizi the nightlight is approximately zero.
- More light is in Kigali city due to many businesses and socio economic activities.
- There is no nightlight to places such as Akagera, National park of Rwanda.