BDS-Assignment3

Author: Jingyi Wu (jingyiw2)

Used Libraries:

Numpy

Pandas

Matplotlib

Regressors

Scipy

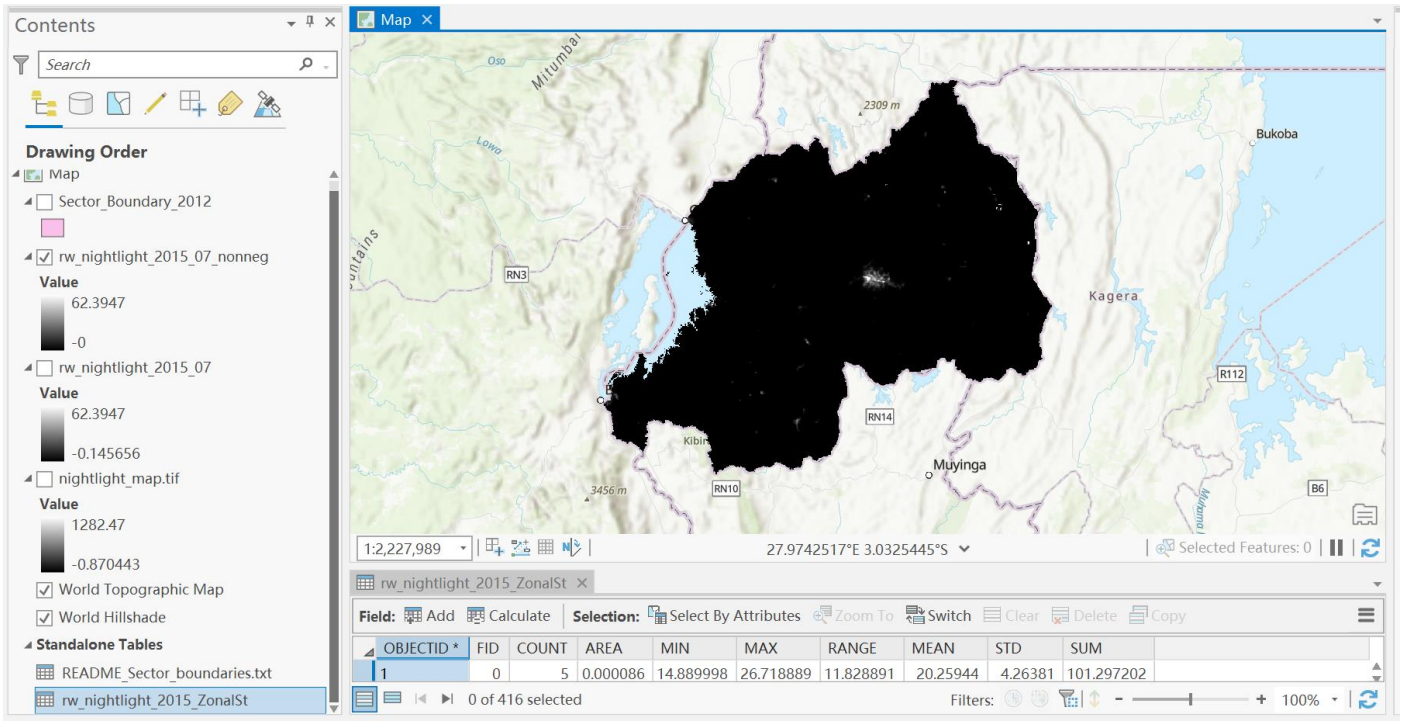
Statsmodel

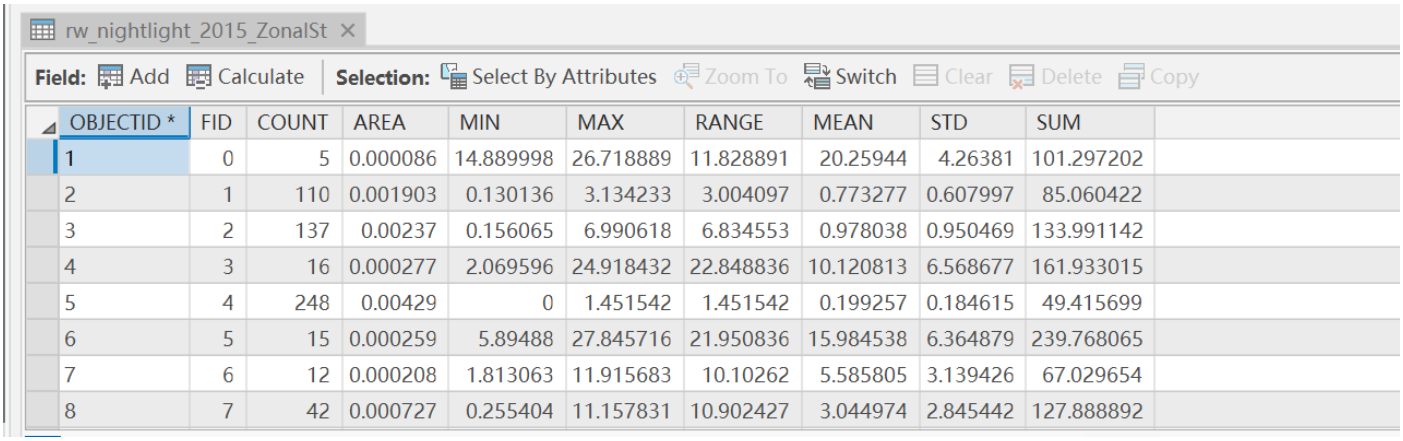
Sklearn

joblib

Q1-Q5:

Following the steps in the tutorials video, we can get nightlight distribution in Rwanda and night_light sum for all areas. The screenshot of the project is as below.



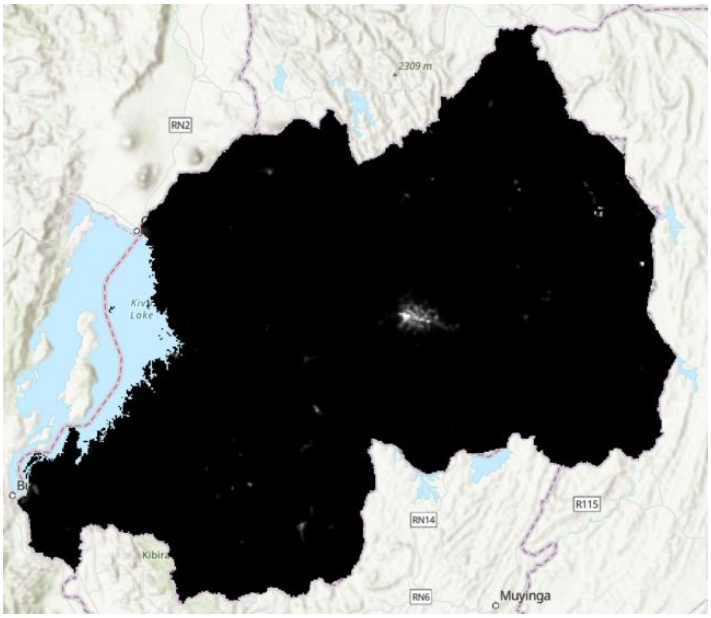


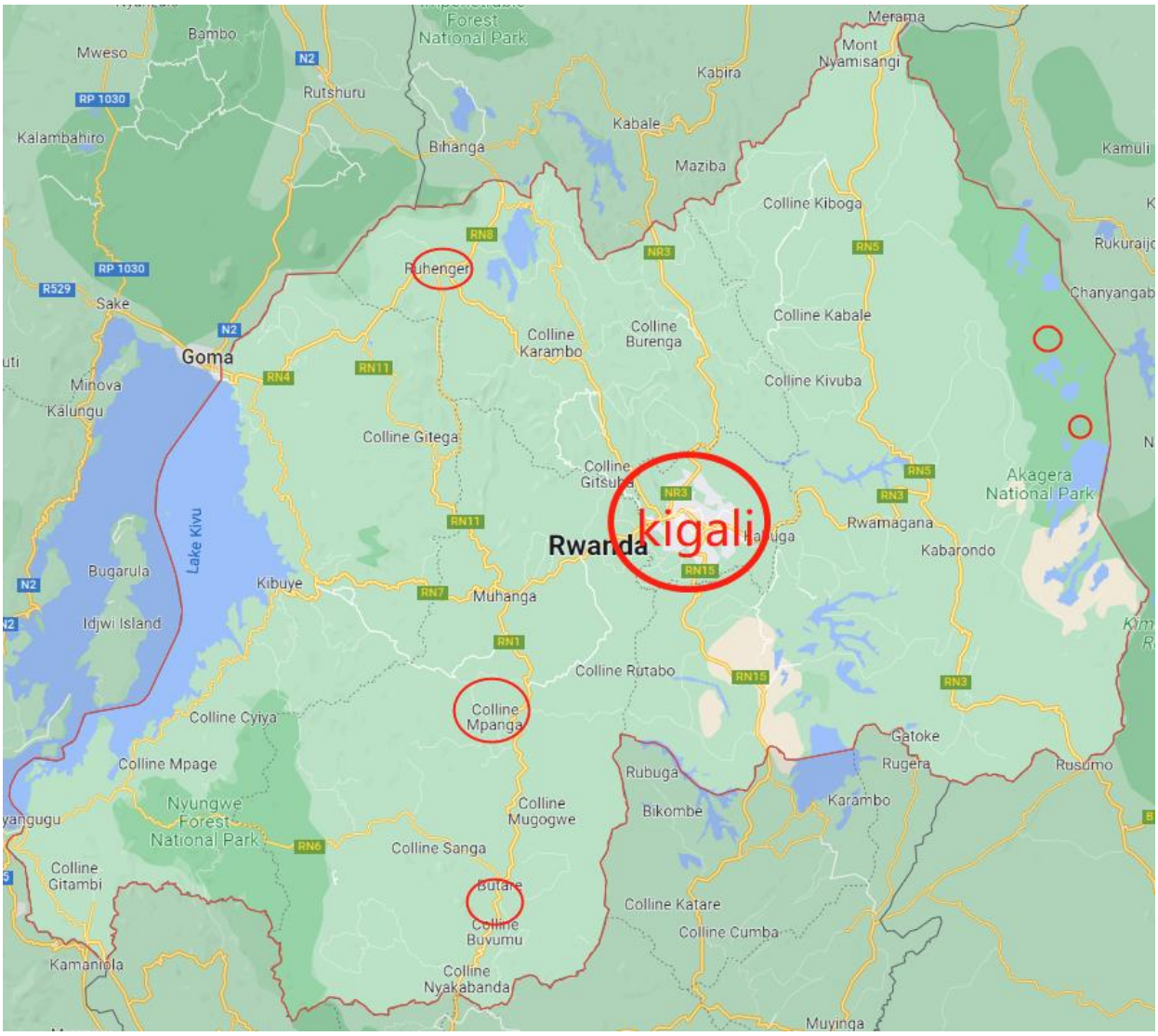
rw_nightlight_2015_ZonalSt

Field: Add Calculate | Selection: Select By Attributes Zoom To Switch Clear Delete Copy

OBJECTID *	FID	COUNT	AREA	MIN	MAX	RANGE	MEAN	STD	SUM
1	0	5	0.000086	14.889998	26.718889	11.828891	20.25944	4.26381	101.297202
2	1	110	0.001903	0.130136	3.134233	3.004097	0.773277	0.607997	85.060422
3	2	137	0.00237	0.156065	6.990618	6.834553	0.978038	0.950469	133.991142
4	3	16	0.000277	2.069596	24.918432	22.848836	10.120813	6.568677	161.933015
5	4	248	0.00429	0	1.451542	1.451542	0.199257	0.184615	49.415699
6	5	15	0.000259	5.89488	27.845716	21.950836	15.984538	6.364879	239.768065
7	6	12	0.000208	1.813063	11.915683	10.10262	5.585805	3.139426	67.029654
8	7	42	0.000727	0.255404	11.157831	10.902427	3.044974	2.845442	127.888892

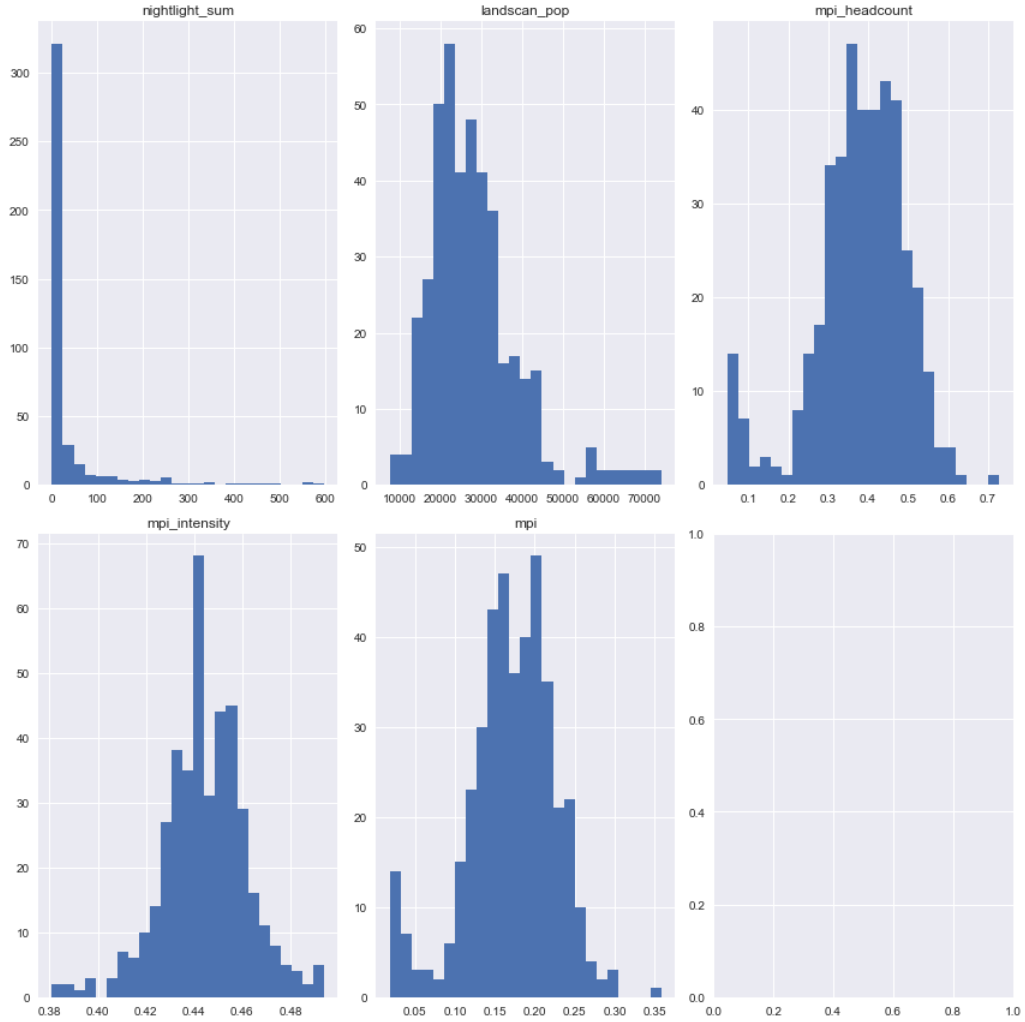
Compare the visualization of map and Google map, I highlighted the light areas with high nightlight values in google map and we can find out that Kigali has the highest nightlight value. Other areas include Ruhengeri (city and capital of Musanze District in the Northern Province), colline mpanga (a place in the province of Kigali), and Butare (a city in Southern province of Rwanda and the capital of Huye district). From the map visualization, the more prosperous the area is, the higher/denser the nightlight is.



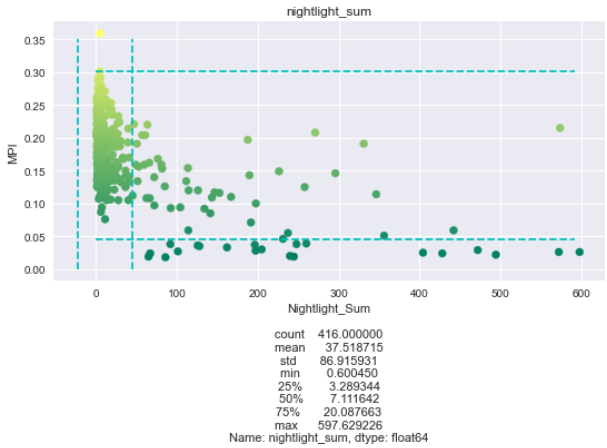


Q6:

a. Use read_excel function to load the data file.
b. The histograms of all features are as below. Features except nightlight_sum can be considered as normally distributed. Nightlight_sum is right skewed.

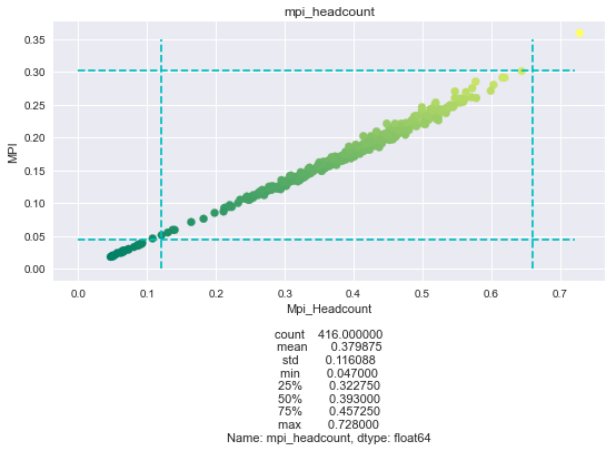


c. Scatterplots of all features to MPI are plotted as below.

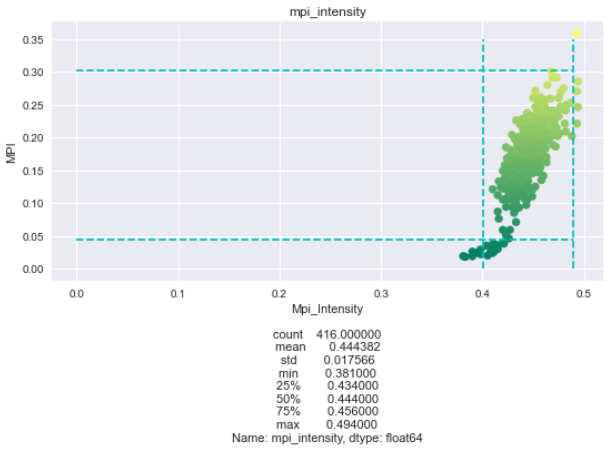


count 416.000000
mean 37.518715
std 86.915931
min 0.600450
25% 3.289344
50% 7.111642
75% 20.087663
max 597.629226
Name: nightlight_sum, dtype: float64

count 416.000000
mean 28189.257212
std 10998.628426
min 7649.000000
25% 20813.750000
50% 26402.000000
75% 32761.750000
max 74297.000000
Name: landscan_pop, dtype: float64



count 416.000000
mean 0.379875
std 0.116088
min 0.047000
25% 0.322750
50% 0.393000
75% 0.457250
max 0.728000
Name: mpi_headcount, dtype: float64



count 416.000000
mean 0.444382
std 0.017566
min 0.381000
25% 0.434000
50% 0.444000
75% 0.456000
max 0.494000
Name: mpi_intensity, dtype: float64

i. We can see that not all features are linear correlated to MPI. MPI shows a decreasing trend with nightlight increasing. No apparent relationship appears between landscan_pop and MPI. Other two features, mpi headcount and mpi_intensity shows linear increasing relationship with MPI.

ii. To find out significant outliers, I added upper bound threshold and lower bound threshold dotted lines to the scatter plot. The upper bound is 0.75Quantile+1.5IQR and the lower bound for each feature is 0.25Quantile-1.5IQR. Points outside of these ranges are potential outliers. Maybe these point should be removed when building the model.

d. Correlations for each feature with MPI:

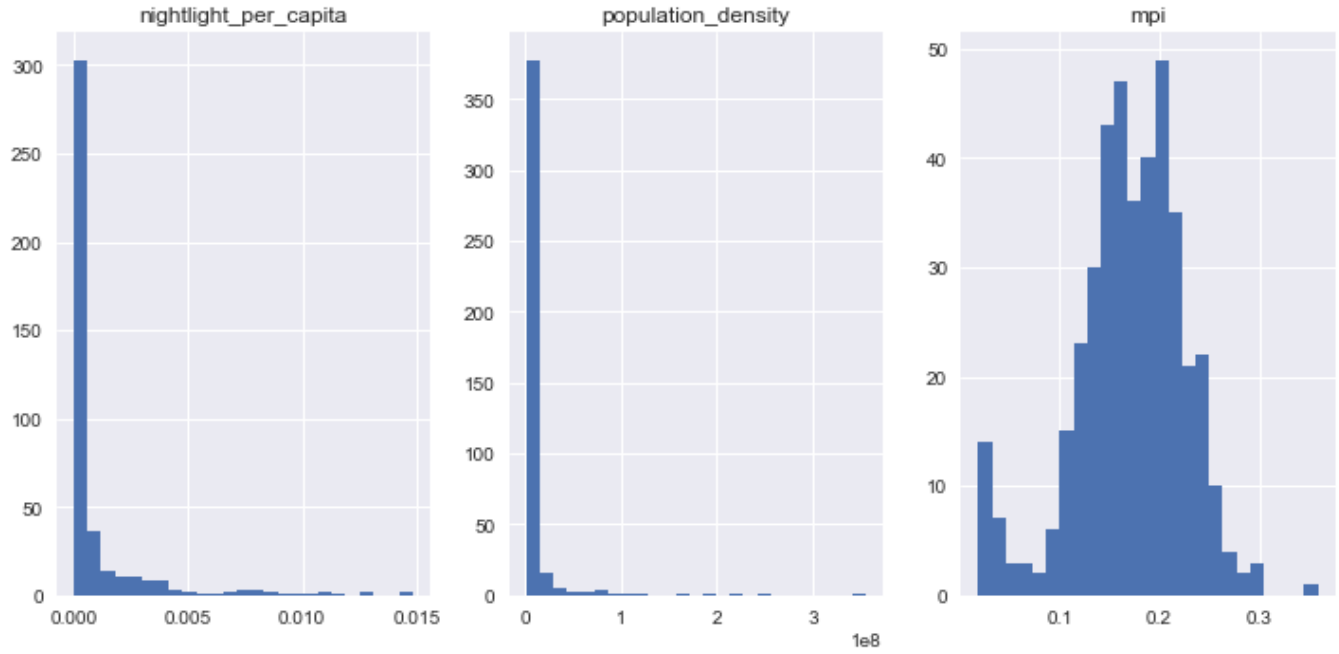
	X vs y	log_X vs y	X vs log_Y	log_X vs log_Y
nightlight_sum	-0.528349	-0.575816	-0.638927	-0.617078
landscan_pop	-0.172782	-0.113587	-0.223342	-0.160110
mpi_headcount	0.995378	0.922131	0.942200	0.998507
mpi_intensity	0.799883	0.803473	0.769113	0.781876
mpi	1.000000	1.000000	1.000000	1.000000

we can see that for nightlight_sum, it has strongest correlation with log of MPI.
For landscan_pop, log of it has strongest correlation with log of mpi.
Log of mpi_headcount has strongest correlation with log of mpi. Log of mpi_intensity has strongest correlation with mpi.

Q7:

After creating two new features, we plot the histograms for these new features as below.

b. From the charts below, we can see that only mpi is normally distributed and other two features are right skewed.



c. new features' correlations with MPI are as below.

	X vs y	log_X vs y	X vs log_Y	log_X vs log_Y
nightlight_per_capita	-0.546978	-0.605358	-0.660497	-0.638304
population_density	-0.487136	-0.617437	-0.668281	-0.745331
mpi	1.000000	1.000000	1.000000	1.000000

Night_light_per_capita is strongly correlated with log of MPI.

Log of population_density is strongly correlated with log of MPI.

Q8:

Using the strongest correlation result above, we build model between log of MPI and log of population_density and Night_light_per_capita to build three models.

<Backward stepwise>

Both features are significant in the model. The p-values for them are as below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                log_mpi   R-squared:                       0.702
Model:                            OLS   Adj. R-squared:                  0.701
Method:                 Least Squares   F-statistic:                     486.7
Date:                Mon, 02 May 2022   Prob (F-statistic):          2.46e-109
Time:                        09:44:14   Log-Likelihood:                -42.884
No. Observations:                 416   AIC:                             91.77
Df Residuals:                     413   BIC:                             103.9
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                  3.9886      0.297     13.450      0.000       3.406       4.572
nightlight_per_capita    -92.0618      6.458    -14.257      0.000    -104.755     -79.368
log_population_density    -0.3638      0.019    -19.199      0.000      -0.401      -0.327
==============================================================================
Omnibus:                        3.968   Durbin-Watson:                   1.506
Prob(Omnibus):                  0.138   Jarque-Bera (JB):                4.759
Skew:                          -0.049   Prob(JB):                       0.0926
Kurtosis:                       3.515   Cond. No.                     7.75e+03
==============================================================================
```

p-values here are all 0.000 and below 0.05, so all features are significant under the significance level α =0.05.
R-squared here is 0.7021 and Adjusted R-squared is 0.70067 for reference.
The p-value for this model here is 0.9999999999999776, not significant under the level α =0.05. Note what we use for testing the significance of the whole model is t-test ind. The result indicates the mean value of predicted log of MPI does not equal to mean of actual log of MPI.

<Ridge regression>
We use RidgeCV to find the best alpha and base our model on that alpha to find out the p-values for each feature.

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.8539 -0.1572 -0.0088 0.1572 0.8928


Coefficients:
             Estimate  Std. Error   t value  p value
_intercept   3.933190    0.272372   14.4405      0.0
x1         -91.501792    6.433691  -14.2223      0.0
x2          -0.360320    0.001916 -188.0905      0.0
---
R-squared:  0.70207,    Adjusted R-squared:  0.70063
F-statistic: 486.61 on 2 features
```

Under the significance level α =0.05, both features are significant (p-values are all 0.0).

p-value for the whole model is 0.9999999999999774, not significant under the level α =0.05.

<Elastic Net>
We tried two parameters here for elastic net and select the one with best R2 values.

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.8534 -0.1583 -0.0086 0.1577 0.8931


Coefficients:
             Estimate  Std. Error   t value  p value
_intercept   3.947197    0.272363   14.4924      0.0
x1         -91.586825    6.433484  -14.2360      0.0
x2          -0.361201    0.001916 -188.5565      0.0
---
R-squared:  0.70209,    Adjusted R-squared:  0.70064
F-statistic: 486.66 on 2 features
```

p-values for both features here are all 0.0 and indicate features are significant under the level α =0.05.
p-value for the whole model here is 0.9999999999999831, indicating the model is insignificant under the α =0.05.

Q9:
Try **lasso regression** with log of MPI and two features and the result is as in the picture.

p-values for two features are all 0.0, below 0.05, indicating two features are significant under the level α =0.05.

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.8538 -0.1593  -0.008  0.1585  0.8941


Coefficients:
             Estimate  Std. Error   t value  p value
_intercept   3.983160    0.272350   14.6251      0.0
x1         -91.948828    6.433185  -14.2929      0.0
x2          -0.363454    0.001916 -189.7413      0.0
---
R-squared:  0.70212,    Adjusted R-squared:  0.70067
F-statistic: 486.72 on 2 features
```

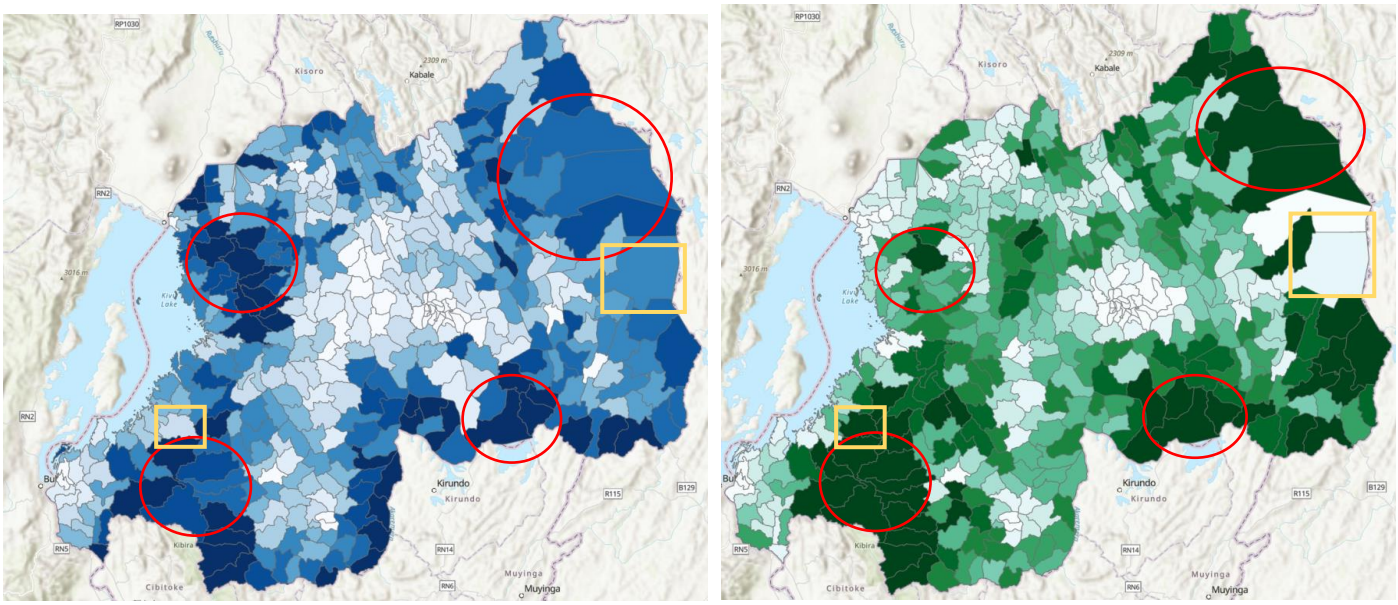
The correlation between predicted values and actual log of MPI is 0.83792, indicating the predictions are highly correlated with actual values. The model has a good predictability.

R-squared value here is 0.70212, and compared this value with the previous R2, we can see that lasso is slightly better than other models. The lasso model's predictability is good.

Q10:

In ArcGis, visualize the actual log of MPI and estimated log of MPI and the graphs are as below.

Actual (left), Estimated(right):



Compare the two maps, we can see they are somewhat similar but yet different in some sectors. In general, our estimation get the MPI pattern correctly. The general pattern here is central part with low MPI and some sectors highlighted in red circle have a quite high MPI. Kigali in both maps display rather low MPI value.

But some sectors in the central part of Rwanda have a higher estimated MPI. Some sectors, for example those highlighted in yellow, actually have a quite low MPI but are estimated quite high. Also, some sectors which have in fact have high MPI are estimated low.