# MPI ESTIMATION USING NIGHTLIGHT DATA

## MPI: Multidimensional Poverty Index

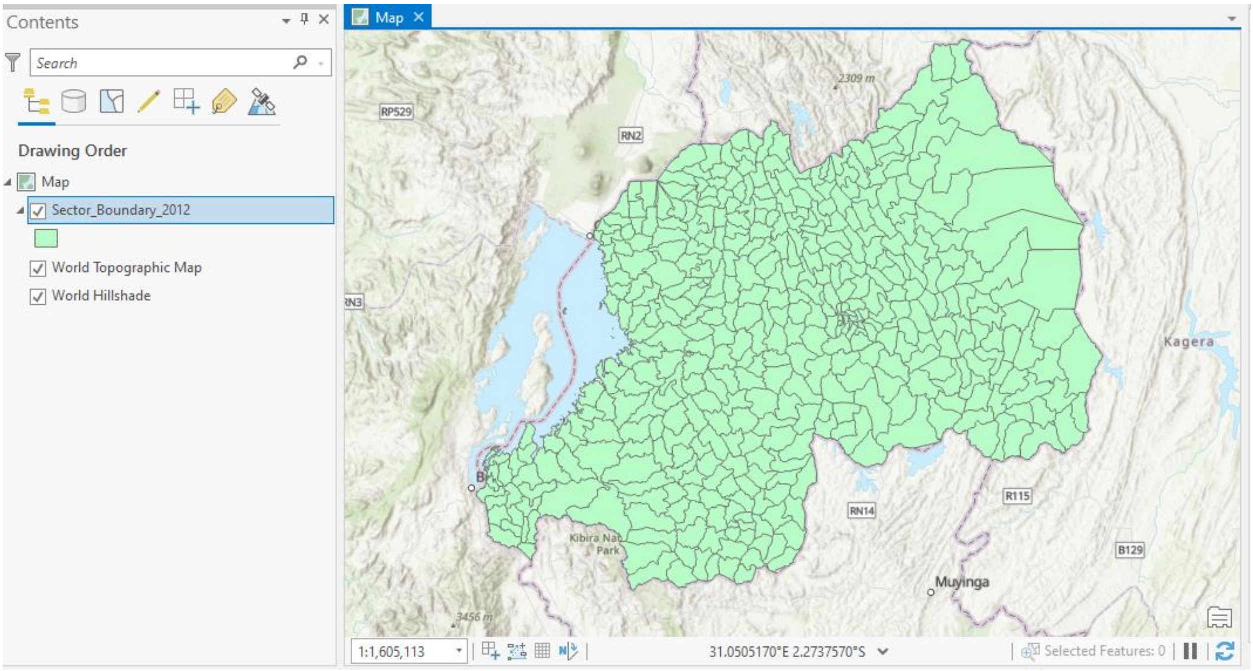
In this project, I worked on ArcGIS (Geographic Information System) software to analyze geospatial data and export the results from the GIS software into a programming environment for further analysis including visualizations and building models for feature selection and predictions among others. In this review, I have explored different machine learning models for regression like backward-stepwise, ridge regression and elastic nets. Therefore, after obtaining the results, I used those results from data analysis and reconstruct maps in GIS software and make comparisons with the original data I had. I have used night light data for Rwanda to make analysis and predictions.

## STEPS USED

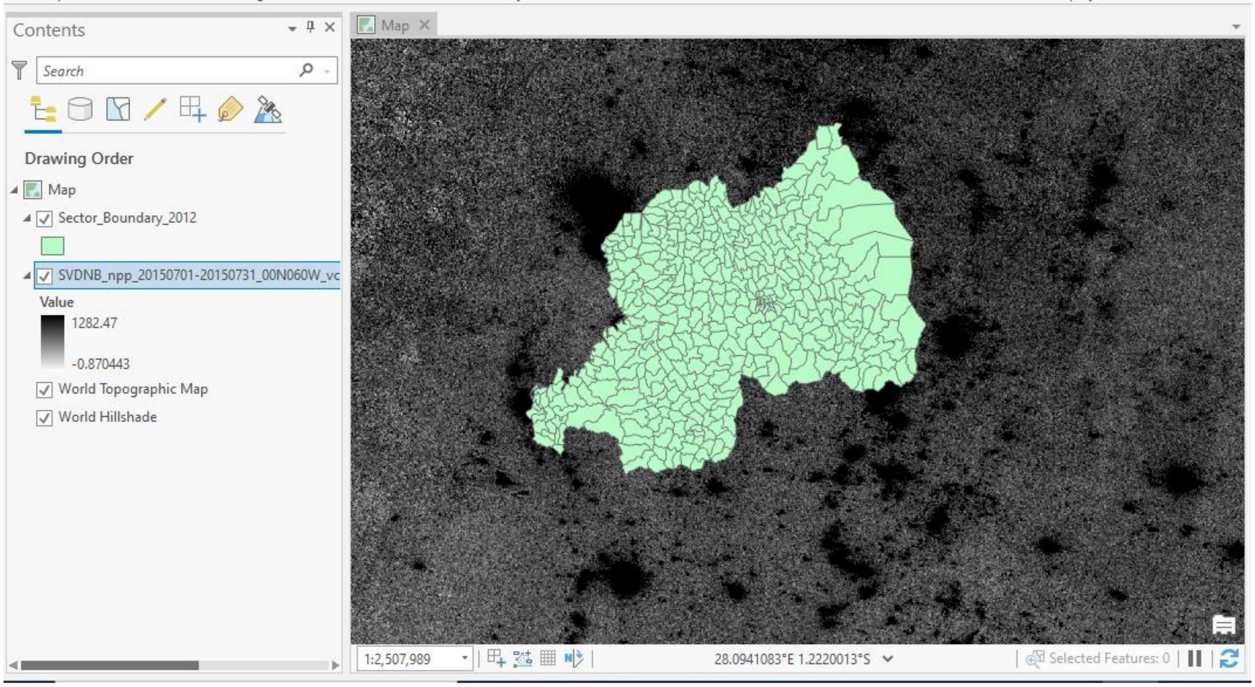
1. Downloaded the nightlight map file which contains night light data for Rwanda through this link
2. Downloaded the Rwanda country and sector administrative boundary maps in shapefile format here
3. Downloaded MPI excel file containing population and MPI values for the sectors in Rwanda through this link
4. Insert the new map on ArcGIS pro software
5. Load Rwanda admin level maps
6. Process nightlight data, visually analyze the map
7. Calculate nightlight statistics for each sector and export them into excel
8. Copy the nightlight data and insert it into the nightlight sum column in the MPI excel downloaded
9. And then, load this final excel file into the programming environment and do further analysis and make predictions.
10. And finally, load these predictions back into the GIS software, and compare its visualization with the original data.

Those are the steps I used during this project, and I am going to show and discuss on the results obtained for each step and finally will make a conclusion on the result obtained.

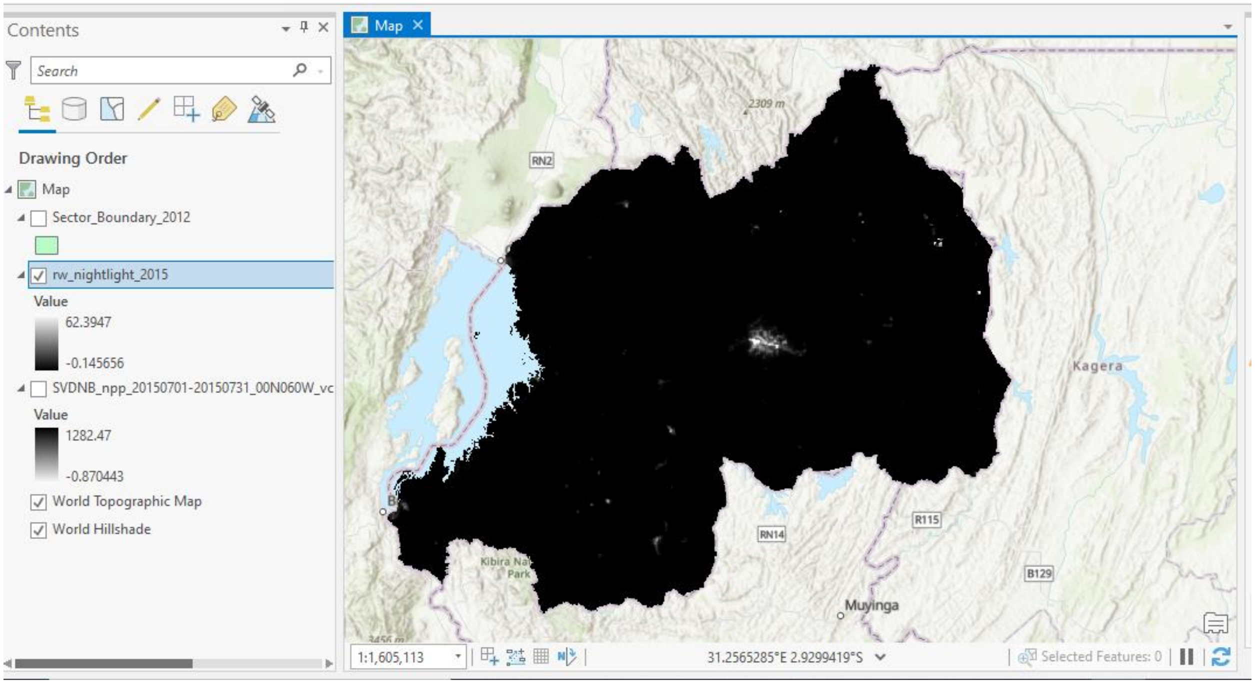
For instance, for step five, after loading the Rwanda sector boundaries, here is the map of Rwanda with sector boundaries I got



Then after, I inserted the nightlight data for Rwanda and got this map for Rwanda nightlight sectors



Then after, using the country layer for nightlight data and using clip Raster to load the nightlight data map for Rwanda, which gives this map



Contents

Search

Drawing Order

- ☑ Map
  - □ Sector_Boundary_2012
    - 🟩
  - ☑ rw_nightlight_2015
    - Value
    - 62.3947
    - -0.145656
  - □ SVDNB_npp_20150701-20150731_00N060W_vc
    - Value
    - 1282.47
    - -0.870443
  - ☑ World Topographic Map
  - ☑ World Hillshade

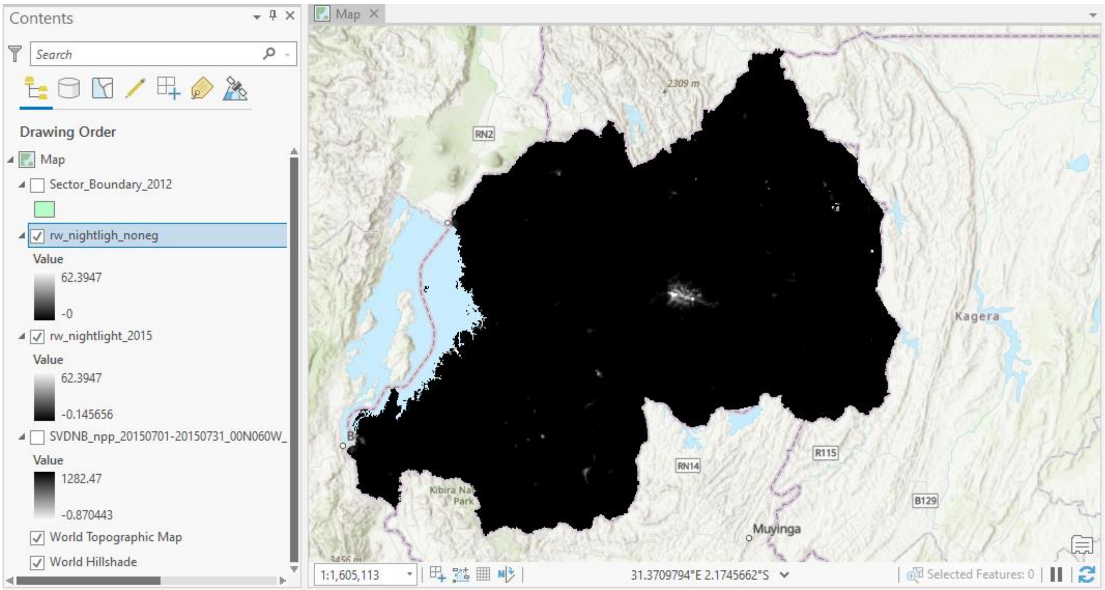
1:1,605,113

31.2565285°E 2.9299419°S

Selected Features: 0

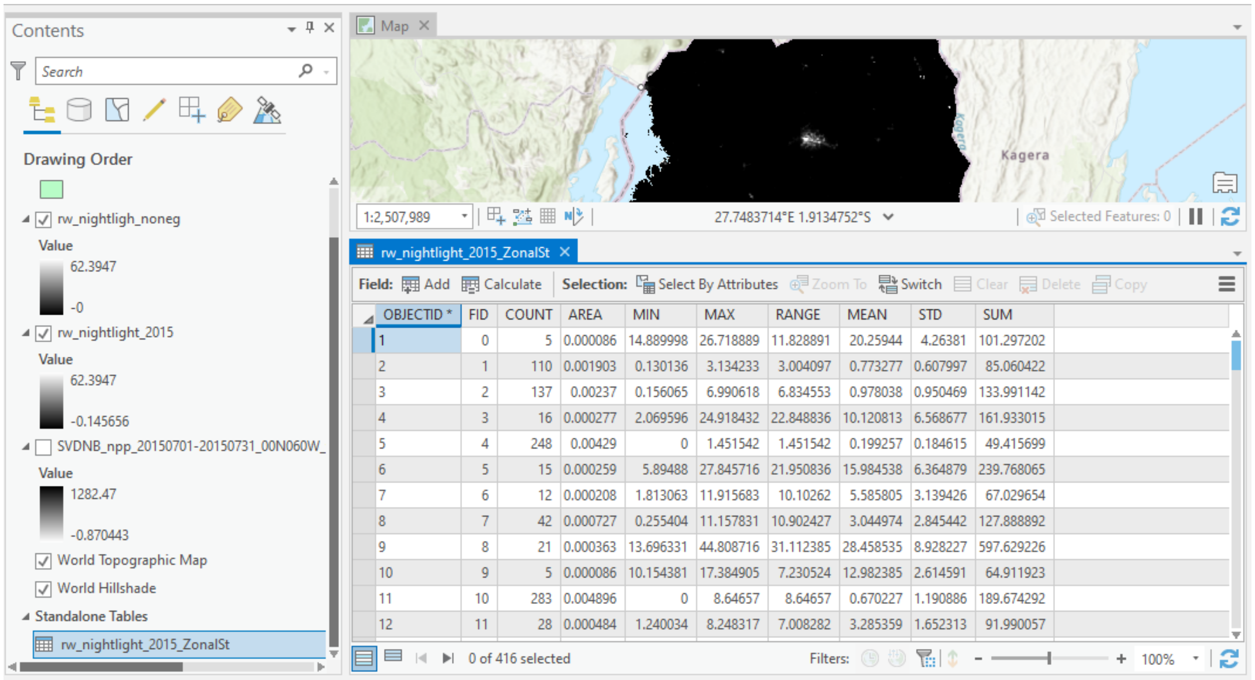
Overall, this is the nightlight map for Rwanda from the geospatial data I have. Therefore, we can immediately see that there is a huge difference between sectors in Kigali and others sector in the rest of the country as we can see that there a big shine in Kigali than in other sectors. Thus, Kigali sectors have more night light (more electricity) than other sectors. We can also see that there are also many sectors without any night light which indicates that there was no or little electricity usage in many of the sector in Rwanda.

Then after, I used Spatial Raster Calculator to set the negative values to zero and here is the map obtained with Rwanda-nightlight-nonnegative created



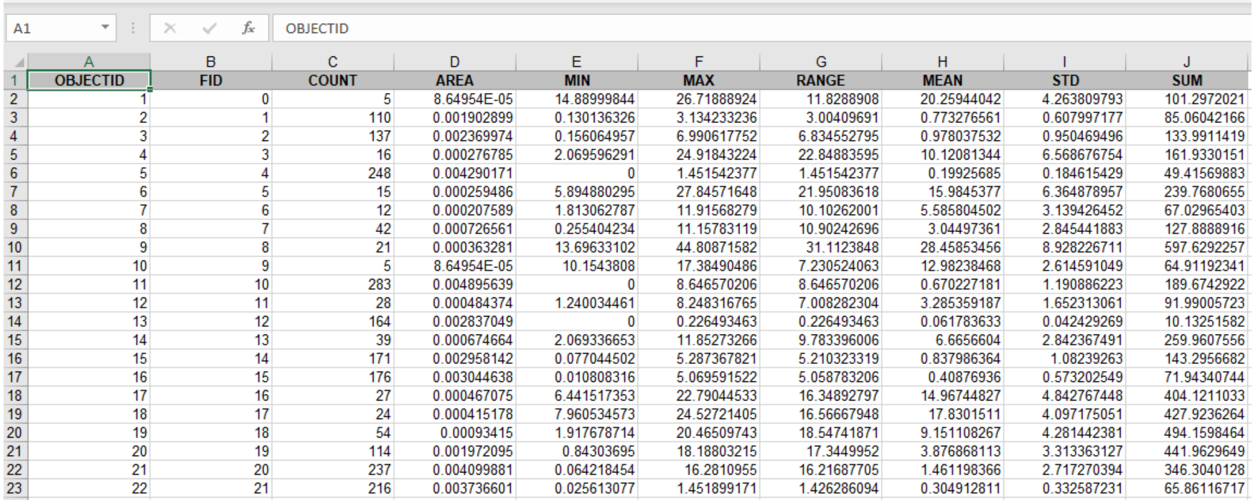
Therefore, by changing all the negative values to zero, there is a small change compared to the original map though it is very hard to see it using these visuals.

Then after, I calculated the summary statistics for each sector in Rwanda “using zonal statistics as table analysis tool” depending on their nightlight visual on the map and then after export this night light statistics to excel for further analysis. Here is the table of statistics obtained



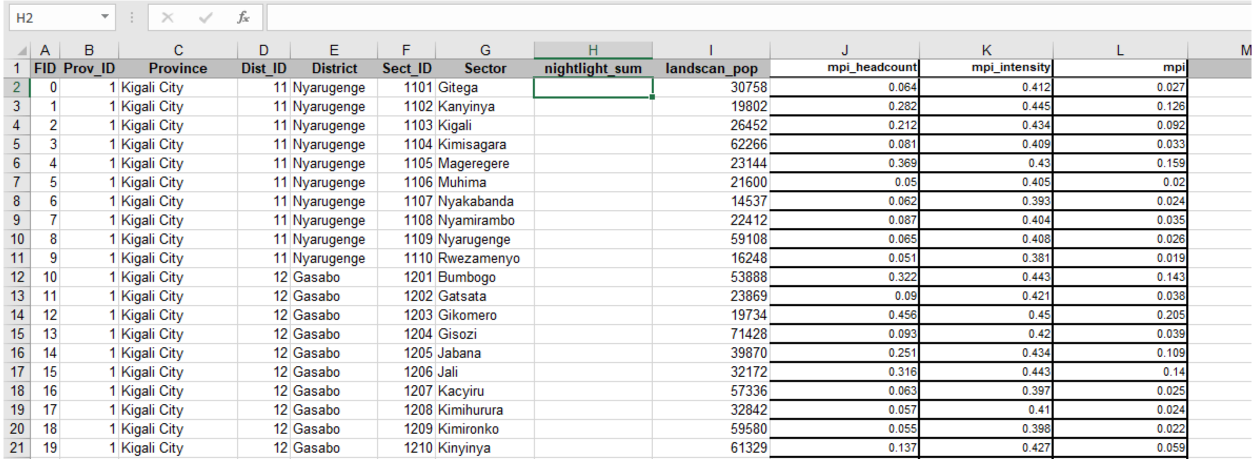
| OBJECTID * | FID | COUNT | AREA | MIN | MAX | RANGE | MEAN | STD | SUM |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 5 | 0.000086 | 14.889998 | 26.718889 | 11.828891 | 20.25944 | 4.26381 | 101.297202 |
| 2 | 1 | 110 | 0.001903 | 0.130136 | 3.134233 | 3.004097 | 0.773277 | 0.607997 | 85.060422 |
| 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 |
| 9 | 8 | 21 | 0.000363 | 13.696331 | 44.808716 | 31.112385 | 28.458535 | 8.928227 | 597.629226 |
| 10 | 9 | 5 | 0.000086 | 10.154381 | 17.384905 | 7.230524 | 12.982385 | 2.614591 | 64.911923 |
| 11 | 10 | 283 | 0.004896 | 0 | 8.64657 | 8.64657 | 0.670227 | 1.190886 | 189.674292 |
| 12 | 11 | 28 | 0.000484 | 1.240034 | 8.248317 | 7.008282 | 3.285359 | 1.652313 | 91.990057 |

And then the next step is to export those statistics table into excel file and export it. Here is the sample for the excel exported



|  | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | OBJECTID | FID | COUNT | AREA | MIN | MAX | RANGE | MEAN | STD | SUM |
| 2 | 1 | 0 | 5 | 8.64954E-05 | 14.88999844 | 26.71888924 | 11.8288908 | 20.25944042 | 4.263809793 | 101.2972021 |
| 3 | 2 | 1 | 110 | 0.001902899 | 0.130136326 | 3.134233236 | 3.00409691 | 0.773276561 | 0.607997177 | 85.06042166 |
| 4 | 3 | 2 | 137 | 0.002369974 | 0.156064957 | 6.990617752 | 6.834552795 | 0.978037532 | 0.950469496 | 133.9911419 |
| 5 | 4 | 3 | 16 | 0.000276785 | 2.069596291 | 24.91843224 | 22.84883595 | 10.12081344 | 6.568676754 | 161.9330151 |
| 6 | 5 | 4 | 248 | 0.004290171 | 0 | 1.451542377 | 1.451542377 | 0.19925685 | 0.184615429 | 49.41569883 |
| 7 | 6 | 5 | 15 | 0.000259486 | 5.894880295 | 27.84571648 | 21.95083618 | 15.9845377 | 6.364878957 | 239.7680655 |
| 8 | 7 | 6 | 12 | 0.000207589 | 1.813062787 | 11.91568279 | 10.10262001 | 5.585804502 | 3.139426452 | 67.02965403 |
| 9 | 8 | 7 | 42 | 0.000726561 | 0.255404234 | 11.15783119 | 10.90242696 | 3.04497361 | 2.845441883 | 127.8888916 |
| 10 | 9 | 8 | 21 | 0.000363281 | 13.69633102 | 44.80871582 | 31.1123848 | 28.45853456 | 8.928226711 | 597.6292257 |
| 11 | 10 | 9 | 5 | 8.64954E-05 | 10.1543808 | 17.38490486 | 7.230524063 | 12.98238468 | 2.614591049 | 64.91192341 |
| 12 | 11 | 10 | 283 | 0.004895639 | 0 | 8.646570206 | 8.646570206 | 0.670227181 | 1.190886223 | 189.6742922 |
| 13 | 12 | 11 | 28 | 0.000484374 | 1.240034461 | 8.248316765 | 7.008282304 | 3.285359187 | 1.652313061 | 91.99005723 |
| 14 | 13 | 12 | 164 | 0.002837049 | 0 | 0.226493463 | 0.226493463 | 0.061783633 | 0.042429269 | 10.13251582 |
| 15 | 14 | 13 | 39 | 0.000674664 | 2.069336653 | 11.85273266 | 9.783396006 | 6.6656604 | 2.842367491 | 259.9607556 |
| 16 | 15 | 14 | 171 | 0.002958142 | 0.077044502 | 5.287367821 | 5.210323319 | 0.837986364 | 1.08239263 | 143.2956682 |
| 17 | 16 | 15 | 176 | 0.003044638 | 0.010808316 | 5.069591522 | 5.058783206 | 0.40876936 | 0.573202549 | 71.94340744 |
| 18 | 17 | 16 | 27 | 0.000467075 | 6.441517353 | 22.79044533 | 16.34892797 | 14.96744827 | 4.842767448 | 404.1211033 |
| 19 | 18 | 17 | 24 | 0.000415178 | 7.960534573 | 24.52721405 | 16.56667948 | 17.8301511 | 4.097175051 | 427.9236264 |
| 20 | 19 | 18 | 54 | 0.00093415 | 1.917678714 | 20.46509743 | 18.54741871 | 9.151108267 | 4.281442381 | 494.1598464 |
| 21 | 20 | 19 | 114 | 0.001972095 | 0.84303695 | 18.18803215 | 17.3449952 | 3.876868113 | 3.313363127 | 441.9629649 |
| 22 | 21 | 20 | 237 | 0.004099881 | 0.064218454 | 16.2810955 | 16.21687705 | 1.461198366 | 2.717270394 | 346.3040128 |
| 23 | 22 | 21 | 216 | 0.003736601 | 0.025613077 | 1.451899171 | 1.426286094 | 0.304912811 | 0.332587231 | 65.86116717 |

Therefore, these are the summary statistics we have obtained for each sector depending on the nightlight data we have. The summary statistics calculated include MIN, MAX, RANGE, MEAN, STD and SUM. In this project, I was only interested in the sum of nightlight for each sector in Rwanda. Therefore, I copied the entire SUM column and pasted it into the MPI excel file given here



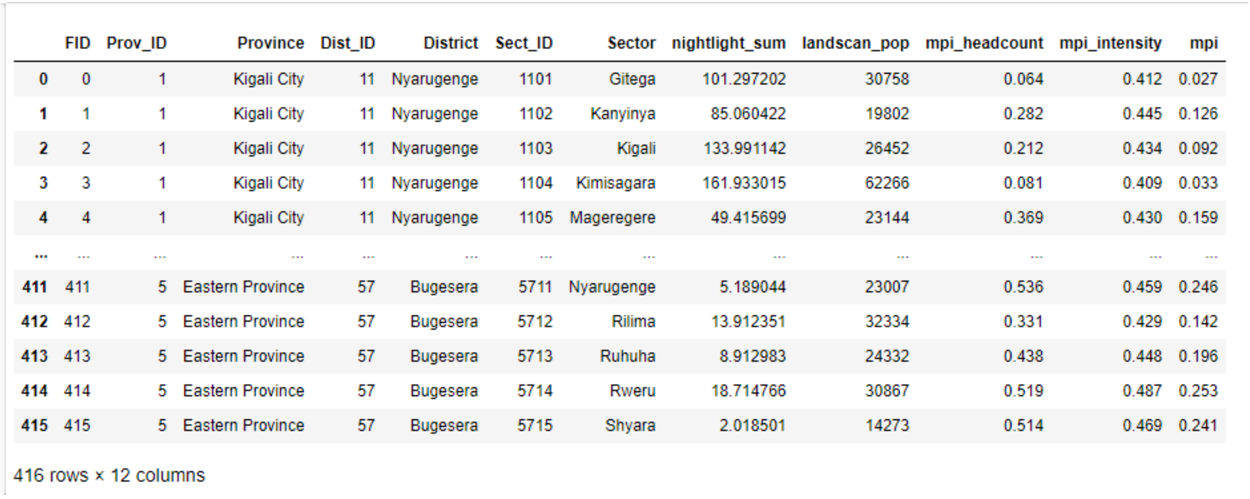
|  | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | FID | Prov_ID | Province | Dist_ID | District | Sect_ID | Sector | nightlight_sum | landscan_pop | mpi_headcount | mpi_intensity | mpi |  |
| 2 | 0 | 1 | Kigali City | 11 | Nyarugenge | 1101 | Gitega |  | 30758 | 0.064 | 0.412 | 0.027 |  |
| 3 | 1 | 1 | Kigali City | 11 | Nyarugenge | 1102 | Kanyinya |  | 19802 | 0.282 | 0.445 | 0.126 |  |
| 4 | 2 | 1 | Kigali City | 11 | Nyarugenge | 1103 | Kigali |  | 26452 | 0.212 | 0.434 | 0.092 |  |
| 5 | 3 | 1 | Kigali City | 11 | Nyarugenge | 1104 | Kimisagara |  | 62266 | 0.081 | 0.409 | 0.033 |  |
| 6 | 4 | 1 | Kigali City | 11 | Nyarugenge | 1105 | Mageregere |  | 23144 | 0.369 | 0.43 | 0.159 |  |
| 7 | 5 | 1 | Kigali City | 11 | Nyarugenge | 1106 | Muhima |  | 21600 | 0.05 | 0.405 | 0.02 |  |
| 8 | 6 | 1 | Kigali City | 11 | Nyarugenge | 1107 | Nyakabanda |  | 14537 | 0.062 | 0.393 | 0.024 |  |
| 9 | 7 | 1 | Kigali City | 11 | Nyarugenge | 1108 | Nyamirambo |  | 22412 | 0.087 | 0.404 | 0.035 |  |
| 10 | 8 | 1 | Kigali City | 11 | Nyarugenge | 1109 | Nyarugenge |  | 59108 | 0.065 | 0.408 | 0.026 |  |
| 11 | 9 | 1 | Kigali City | 11 | Nyarugenge | 1110 | Rwezamenyo |  | 16248 | 0.051 | 0.381 | 0.019 |  |
| 12 | 10 | 1 | Kigali City | 12 | Gasabo | 1201 | Bumbogo |  | 53888 | 0.322 | 0.443 | 0.143 |  |
| 13 | 11 | 1 | Kigali City | 12 | Gasabo | 1202 | Gatsata |  | 23869 | 0.09 | 0.421 | 0.038 |  |
| 14 | 12 | 1 | Kigali City | 12 | Gasabo | 1203 | Gikomero |  | 19734 | 0.456 | 0.45 | 0.205 |  |
| 15 | 13 | 1 | Kigali City | 12 | Gasabo | 1204 | Gisozi |  | 71428 | 0.093 | 0.42 | 0.039 |  |
| 16 | 14 | 1 | Kigali City | 12 | Gasabo | 1205 | Jabana |  | 39870 | 0.251 | 0.434 | 0.109 |  |
| 17 | 15 | 1 | Kigali City | 12 | Gasabo | 1206 | Jali |  | 32172 | 0.316 | 0.443 | 0.14 |  |
| 18 | 16 | 1 | Kigali City | 12 | Gasabo | 1207 | Kacyiru |  | 57336 | 0.063 | 0.397 | 0.025 |  |
| 19 | 17 | 1 | Kigali City | 12 | Gasabo | 1208 | Kimihurura |  | 32842 | 0.057 | 0.41 | 0.024 |  |
| 20 | 18 | 1 | Kigali City | 12 | Gasabo | 1209 | Kimironko |  | 59580 | 0.055 | 0.398 | 0.022 |  |
| 21 | 19 | 1 | Kigali City | 12 | Gasabo | 1210 | Kinyinya |  | 61329 | 0.137 | 0.427 | 0.059 |  |

The attributes/variables we have here for analysis include nightlight-sum, landscan-pop, mpi-headcount, mpi-intensity and mpi (which is the target output). Now that I have the complete excel file, the next step is to use Python programming language for further analysis and making predictions.

# ANALYZING THE DATA

1. **Load MPI excel into programming environment**

   Using pandas to load excel file into Python programming environment



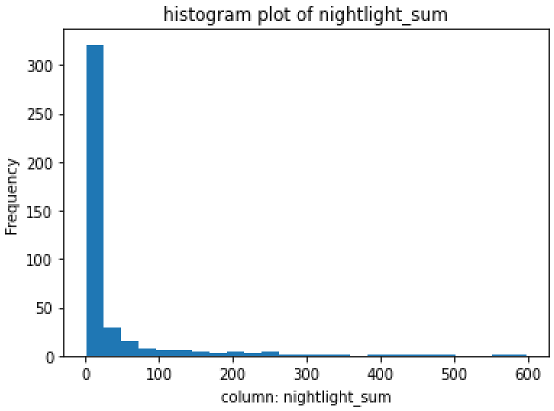
| | FID | Prov_ID | Province | Dist_ID | District | Sect_ID | Sector | nightlight_sum | landscan_pop | mpi_headcount | mpi_intensity | mpi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 1 | Kigali City | 11 | Nyarugenge | 1101 | Gitega | 101.297202 | 30758 | 0.064 | 0.412 | 0.027 |
| **1** | 1 | 1 | Kigali City | 11 | Nyarugenge | 1102 | Kanyinya | 85.060422 | 19802 | 0.282 | 0.445 | 0.126 |
| **2** | 2 | 1 | Kigali City | 11 | Nyarugenge | 1103 | Kigali | 133.991142 | 26452 | 0.212 | 0.434 | 0.092 |
| **3** | 3 | 1 | Kigali City | 11 | Nyarugenge | 1104 | Kimisagara | 161.933015 | 62266 | 0.081 | 0.409 | 0.033 |
| **4** | 4 | 1 | Kigali City | 11 | Nyarugenge | 1105 | Mageregere | 49.415699 | 23144 | 0.369 | 0.430 | 0.159 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **411** | 411 | 5 | Eastern Province | 57 | Bugesera | 5711 | Nyarugenge | 5.189044 | 23007 | 0.536 | 0.459 | 0.246 |
| **412** | 412 | 5 | Eastern Province | 57 | Bugesera | 5712 | Rilima | 13.912351 | 32334 | 0.331 | 0.429 | 0.142 |
| **413** | 413 | 5 | Eastern Province | 57 | Bugesera | 5713 | Ruhuha | 8.912983 | 24332 | 0.438 | 0.448 | 0.196 |
| **414** | 414 | 5 | Eastern Province | 57 | Bugesera | 5714 | Rweru | 18.714766 | 30867 | 0.519 | 0.487 | 0.253 |
| **415** | 415 | 5 | Eastern Province | 57 | Bugesera | 5715 | Shyara | 2.018501 | 14273 | 0.514 | 0.469 | 0.241 |

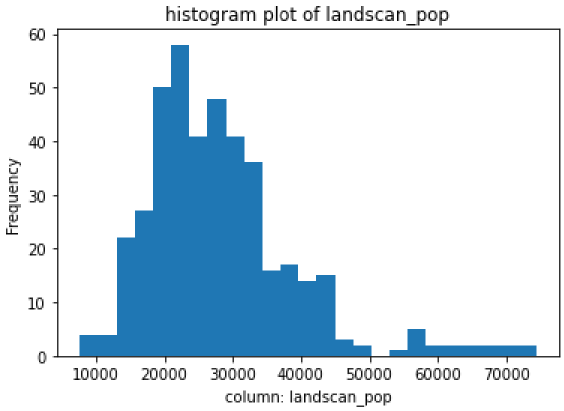
416 rows × 12 columns

We can see that the dataset has 416 rows (corresponding to the number of sectors in Rwanda) and 12 columns which corresponds to the features we have.

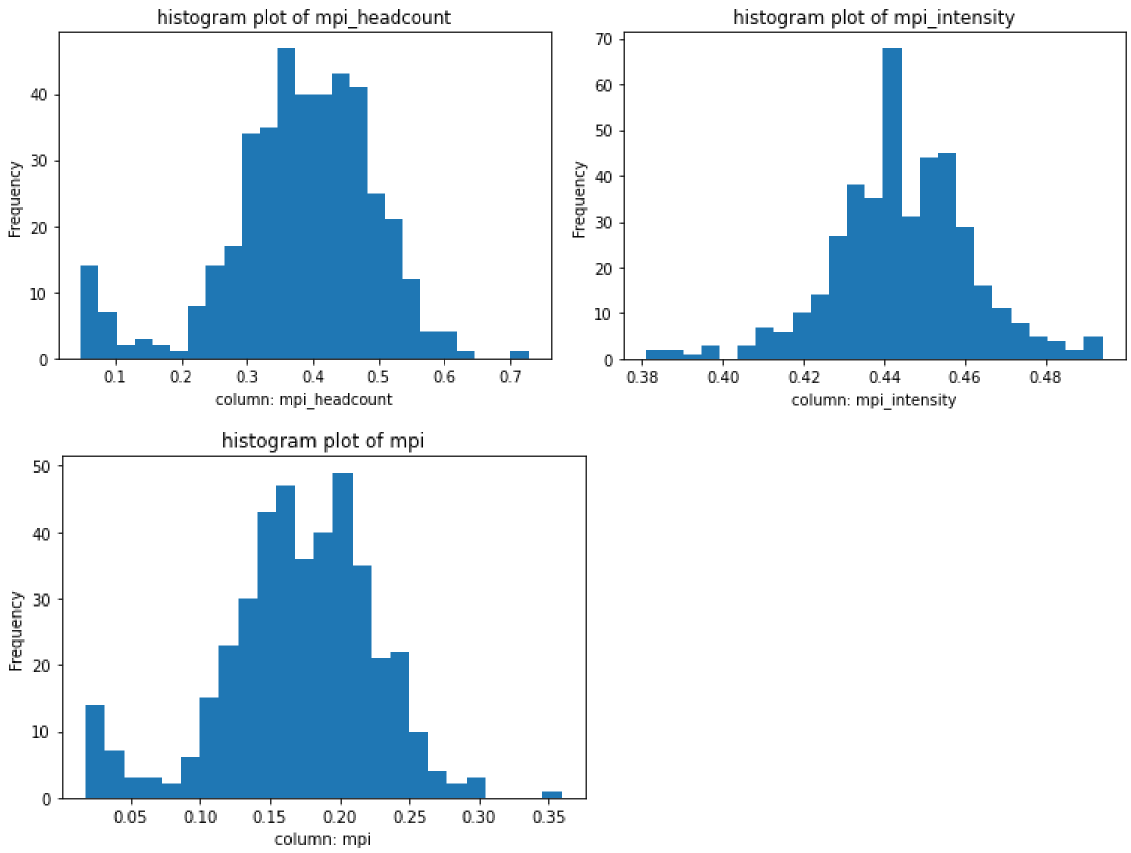
2. **Plot histograms of each of the features and dependent variable**
   - **Are the variables normally distributed?**

Here are the histograms plot for each feature





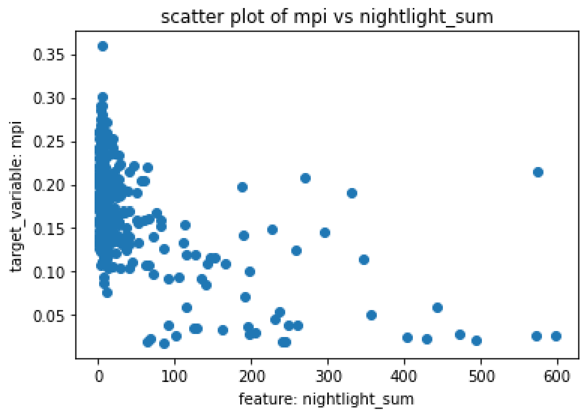
We can see from this graph that nightlight-sum is not normally distributed because it is asymmetric about its mean. In addition, it appears that half of landscan-pop's data are symmetric about its mean and the other half are not. As a result, this can be classified as semi-normally distributed.

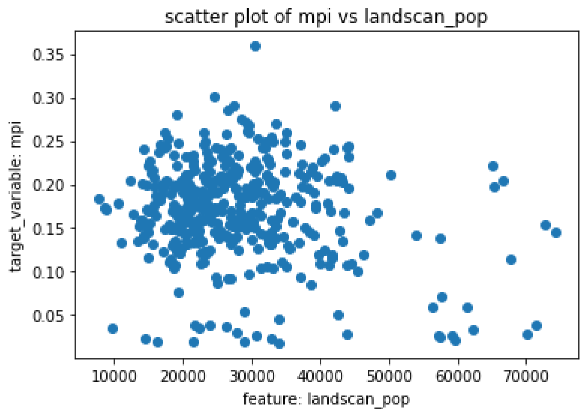


Furthermore, for mpi-headcount, mpi-intensity, and mpi, it appears that half of their data is symmetric about their mean and the other half is not. As a result, this can be classified as semi-normally distributed. As a result, none of those histogram plots are normally distributed.

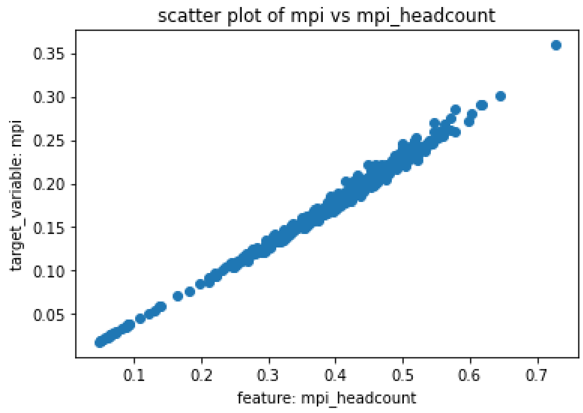
3. **Create scatter plots of the mpi (dependent/target variable) vs each of the features**
    - **Are the relationships between the features and the dependent variable linear?**
    - **Are there significant outliers**

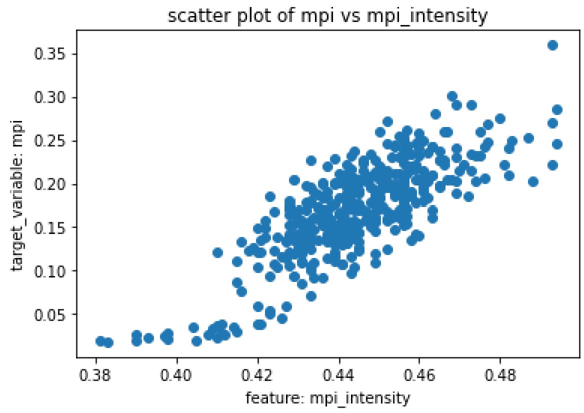
Here are the scatter plots





From those scatter plots, we can see there is a negative correlation between nightlight-sum and landscan-pop with the target variable (mpi). However, the relationship between those independent features are not linearly with the target i.e there is non-linear relationship between nightlight-sum and landscan-pop with the mpi



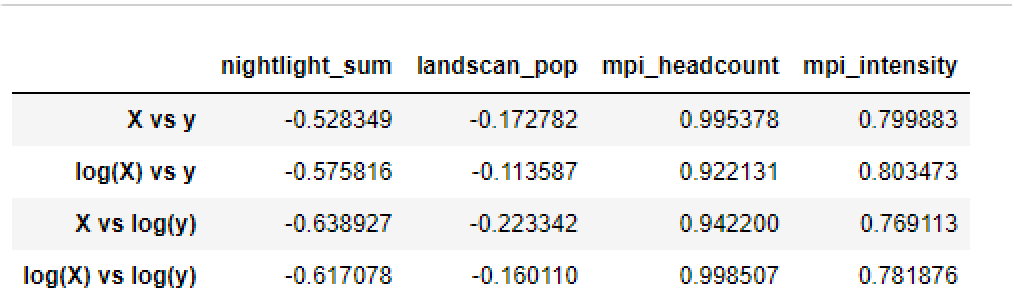


As we can see from the above scatter plots, there is a very strong positive correlation between the features (mpi-headcount and mpi-intensity) with the target output feature (mpi) and hence there is a very strong linear relationship between those features with the target variable mpi.

Moreover, as we can see, there are significant outliers for each of those scatter plots. Those outliers may be caused by improper reading, or incorrect observations for the data taken

4. **Calculate the following correlations for each feature (Xi) with the MPI (y)**
    - **X vs y**
    - **Log(X) vs y**
    - **X vs log(y)**
    - **Log(X) vs log(y)**
    - **Which are the strongest correlations for each feature**

Here is the table indicating all the correlations



| | nightlight_sum | landscan_pop | mpi_headcount | mpi_intensity |
|---|---|---|---|---|
| **X vs y** | -0.528349 | -0.172782 | 0.995378 | 0.799883 |
| **log(X) vs y** | -0.575816 | -0.113587 | 0.922131 | 0.803473 |
| **X vs log(y)** | -0.638927 | -0.223342 | 0.942200 | 0.769113 |
| **log(X) vs log(y)** | -0.617078 | -0.160110 | 0.998507 | 0.781876 |

Therefore, from the table of correlations we have, we can see the maximum/strongest correlation for each feature.

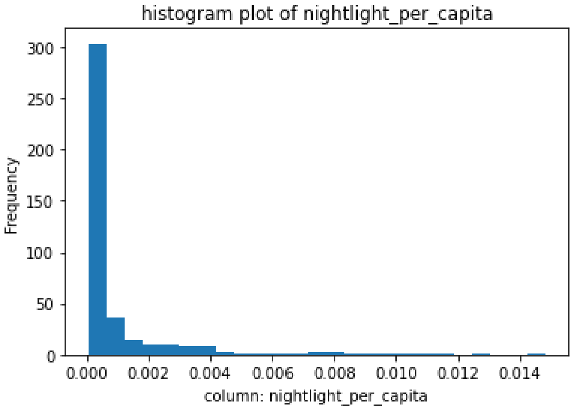
- For nightlight-sum we have maximum correlation at **X vs log(y)**
- For landscan-pop we have maximum correlation at **X vs log(y)**
- For mpi-headcount we have maximum correlation at **log(X) vs log(y)**
- For mpi-intensity we have maximum correlation at **log(X) vs y**

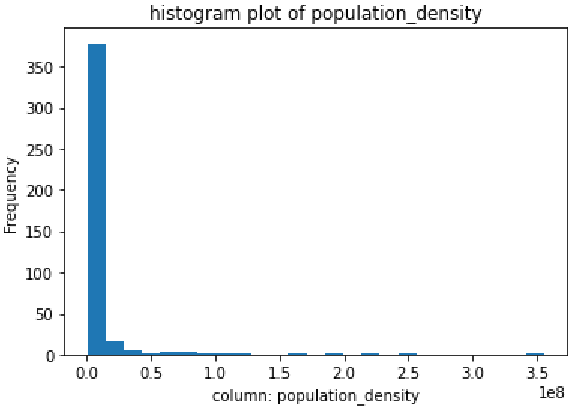
## CREATING FINAL FEATURES

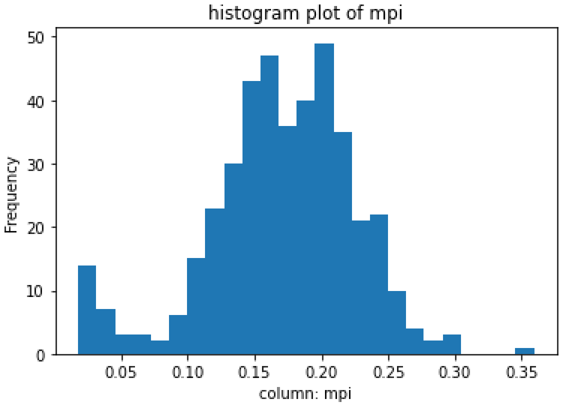
A. By ensuring that the features are intuitive since the MPI has a population and an Area component, then we create versions of the features that take these components into consideration. Thus, this will prevent either the area or population overly influencing any feature. This will also serve to normalize the variables. Then here we created the following features:
   - Nightlight-per-capita = nightlight-sum / landscan-pop
   - Population-density = landscan-pop / Area

B. Then, plot histograms of each of the features and the dependent variables
   - Are the features normally distributed?

Here are the histograms obtained





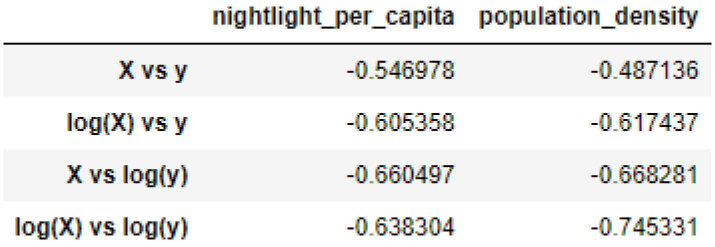


Therefore, from the graph above, we can see that the features are **not normally distributed.**

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

- X vs y
- Log(X) vs y
- X vs log(y)
- Log(X) vs log(y)
- Which are the strongest correlations for each feature

Here is the final table containing all correlations



| | nightlight_per_capita | population_density |
|---|---|---|
| **X vs y** | -0.546978 | -0.487136 |
| **log(X) vs y** | -0.605358 | -0.617437 |
| **X vs log(y)** | -0.660497 | -0.668281 |
| **log(X) vs log(y)** | -0.638304 | -0.745331 |

Therefore, from the table of correlations we have, we can see the maximum/strongest correlation for each feature.

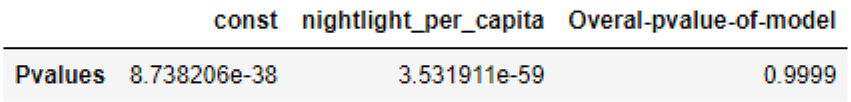
- For nightlight-per-capita we have maximum correlation at **X vs log(y)**
- For population-density we have maximum correlation at **log(X) vs log(y)**

## MODEL BUILDING

By using the strongest correlations from the previous question, check if the features we have selected are significant in explaining the MPI

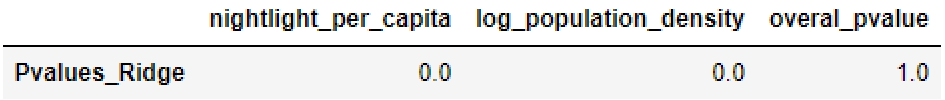
- ❖ Using the backward-stepwise, ridge-regression and elastic nets:
  - To calculate the p-value of each feature, are all features significant? At what levels
  - To find the overall p-value of the model, at what levels is it significant

By using **BACKWARD-STEPWISE Regression** model, here are the p-values obtained



| | const | nightlight_per_capita | Overal-pvalue-of-model |
|---|---|---|---|
| **Pvalues** | 8.738206e-38 | 3.531911e-59 | 0.9999 |

By using Ridge Regression model, here are the p-values obtained



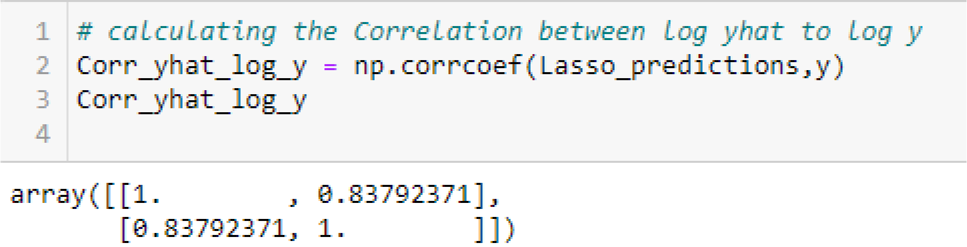
| | nightlight_per_capita | log_population_density | overal_pvalue |
|---|---|---|---|
| **Pvalues_Ridge** | 0.0 | 0.0 | 1.0 |

In this project, I have used the significant level known as alpha of 0.05. And thus, features with p-value less than alpha are statistically significant and those with p-value less than alpha are not statistically significant. Therefore, both features here have p-value which are less than alpha, so those features are statistically significant. Finally, the overall p-value if greater than alpha, then the overall p-value is not significant

## MODEL EVALUATION

A. By using LASSO, calculating the estimated MPI (log (yhat)) for each sector and then, calculate the correlation of log(yhat) to log(y) and to explain what does it tell us about the model

After building the model and generating MPI predictions, here is the correlation calculated between predictions and the log y



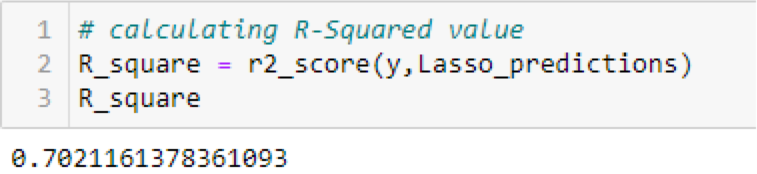
```
# calculating the Correlation between log yhat to log y
Corr_yhat_log_y = np.corrcoef(Lasso_predictions,y)
Corr_yhat_log_y

```

```
array([[1.        , 0.83792371],
       [0.83792371, 1.        ]])
```

This indicates that the correlation between the predicted MPI and the original MPI had is 0.838 which is a very strong positive correlation telling us that there is a strong linear relationship between the predicted MPI and the original MPI data. Therefore, we can trust the performance of the model.

B. To calculate the R-squared of the result and to explain what does the R-squared tell us about the model



```
# calculating R-Squared value
R_square = r2_score(y,Lasso_predictions)
R_square
```

```
0.7021161378361093
```

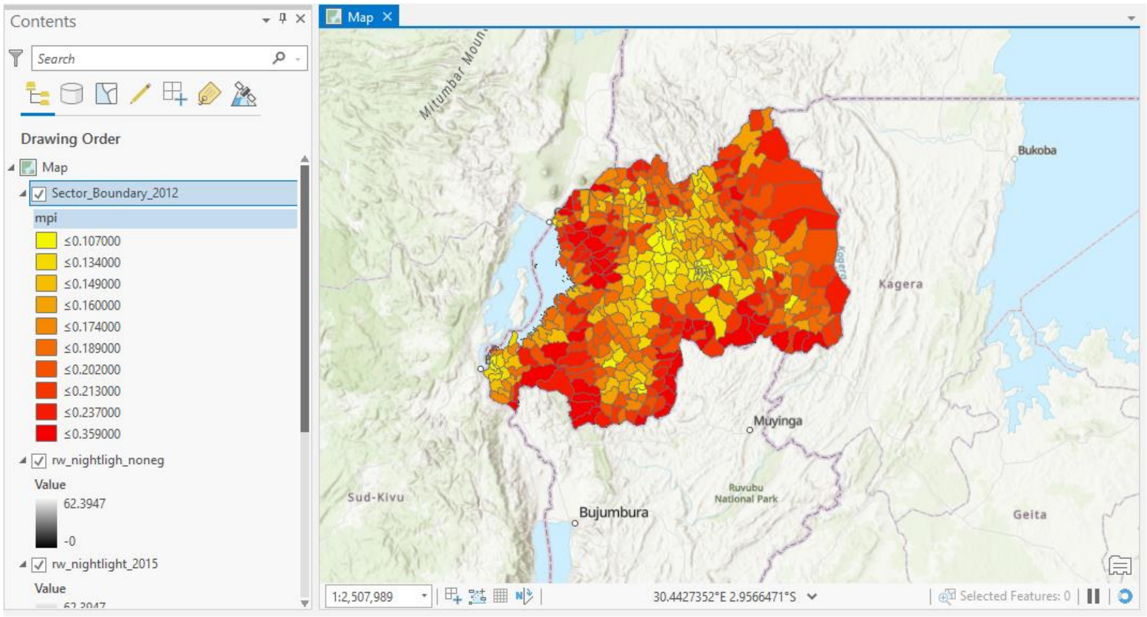
R-square of 0.7021 indicates that the predictors explained approximately 70.2% of the variation in our response variable. According to this viewpoint, the higher the R-square, the more variations are explained and thus the model is more accurate.

## VISUALIZE THE RESULTS

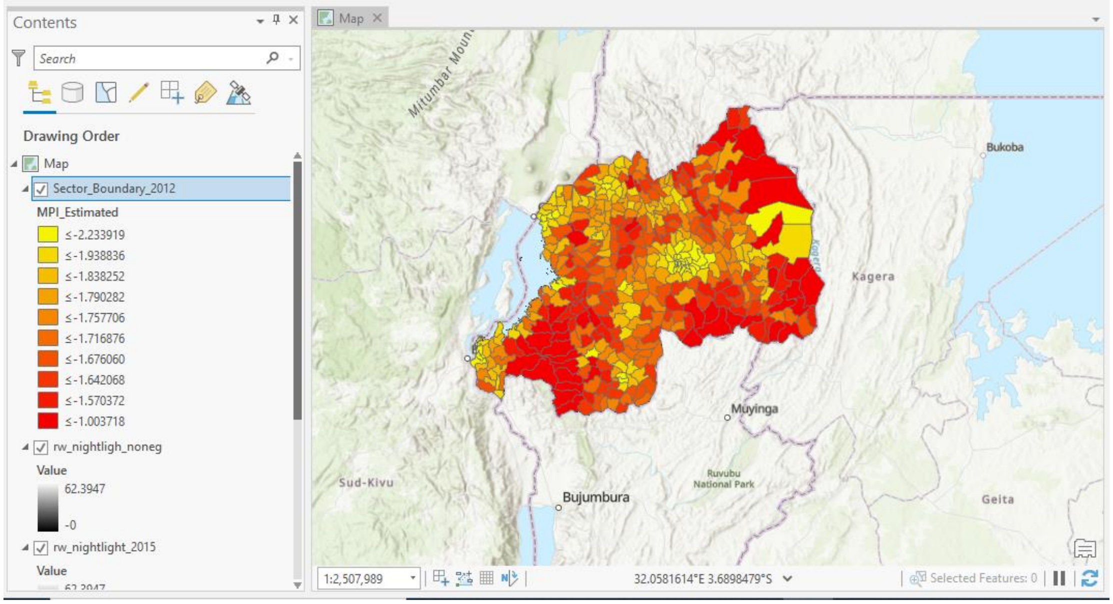
Add the estimated MPI in the MPI excel file, and then load it again in ArcGIS

Create two sector layers one to display the original MPI and the other to display the estimated MPI and give an insight if there is any similarity

Here is the figure obtained for the original mpi



And here is the figure obtained for estimated mpi



There is highly percentage of similarity between those graphs which means that the model used for prediction performed well and gave us high accuracy with good predictions.