MA615_Midterm_Project

Yiping Jiang

10/20/2019

Indicators:

GNI per capita (Atlas $):
It is the gross national income, converted to U.S. dollars using the World Bank Atlas method, divided by the midyear population. GNI is the sum of value added by all resident producers plus any product taxes (less subsidies) not included in the valuation of output plus net receipts of primary income (compensation of employees and property income) from abroad.

Access to electricity (% of population):
It is the percentage of population with access to electricity. Electrification data are collected from industry, national surveys and international sources.

Urban population (% of total population):
It refers to people living in urban areas as defined by national statistical offices. The data are collected and smoothed by United Nations Population Division.

Research Question:

Is there a correlation between GNI per capita and access to electricity? How strond is the correlation? And how about GNI per capita and urban population? How strong the correlation is?

1. Get Data

First we search and load updated lists of inidicators: GNI per capita, access to electricity and urban population.

2. Clean Up

First we assign a dataframe with combining the three indicators, and then we select their corresponding data and assign to each indicator name; next we use na.omit to returns the non-missing value objects to clean the raw data.

```
wdi_data <- WDI(indicator = c("NY.GNP.PCAP.CD",
                              "EG.ELC.ACCS.ZS",
                              "SP.URB.TOTL.IN.ZS"),
                start = 1960, end = 2019, extra = TRUE)
names(wdi_data)
```

```
##  [1] "iso2c"             "country"           "year"
##  [4] "NY.GNP.PCAP.CD"    "EG.ELC.ACCS.ZS"    "SP.URB.TOTL.IN.ZS"
##  [7] "iso3c"             "region"            "capital"
## [10] "longitude"         "latitude"          "income"
## [13] "lending"
```

```
wdi_data <- subset(wdi_data, region != "Aggregates")
names(wdi_data)[which(names(wdi_data) == "NY.GNP.PCAP.CD")] <- "GNI"
names(wdi_data)[which(names(wdi_data) == "EG.ELC.ACCS.ZS")] <- "Electricity"
```

```
names(wdi_data)[which(names(wdi_data) == "SP.URB.TOTL.IN.ZS")] <- "Urban_Population"
data = na.omit(wdi_data)
names(data)
```

```
##  [1] "iso2c"            "country"          "year"
##  [4] "GNI"              "Electricity"      "Urban_Population"
##  [7] "iso3c"            "region"           "capital"
## [10] "longitude"        "latitude"         "income"
## [13] "lending"
```

3. Plot Graphs

Observe the summary and distribution of these three indicators: the mode of GNI concentrates in 0 to 10,000; the mode of Electricity concentrates in 90 to 100; and the distribution of Urban Population looks comparatively more average than the other two.

```
pt1 = ggplot(data, aes(x = GNI)) +
  geom_histogram(bins = 100, fill="blue")
summary(data$GNI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     110    1100    3775   11424   15270  121890
```

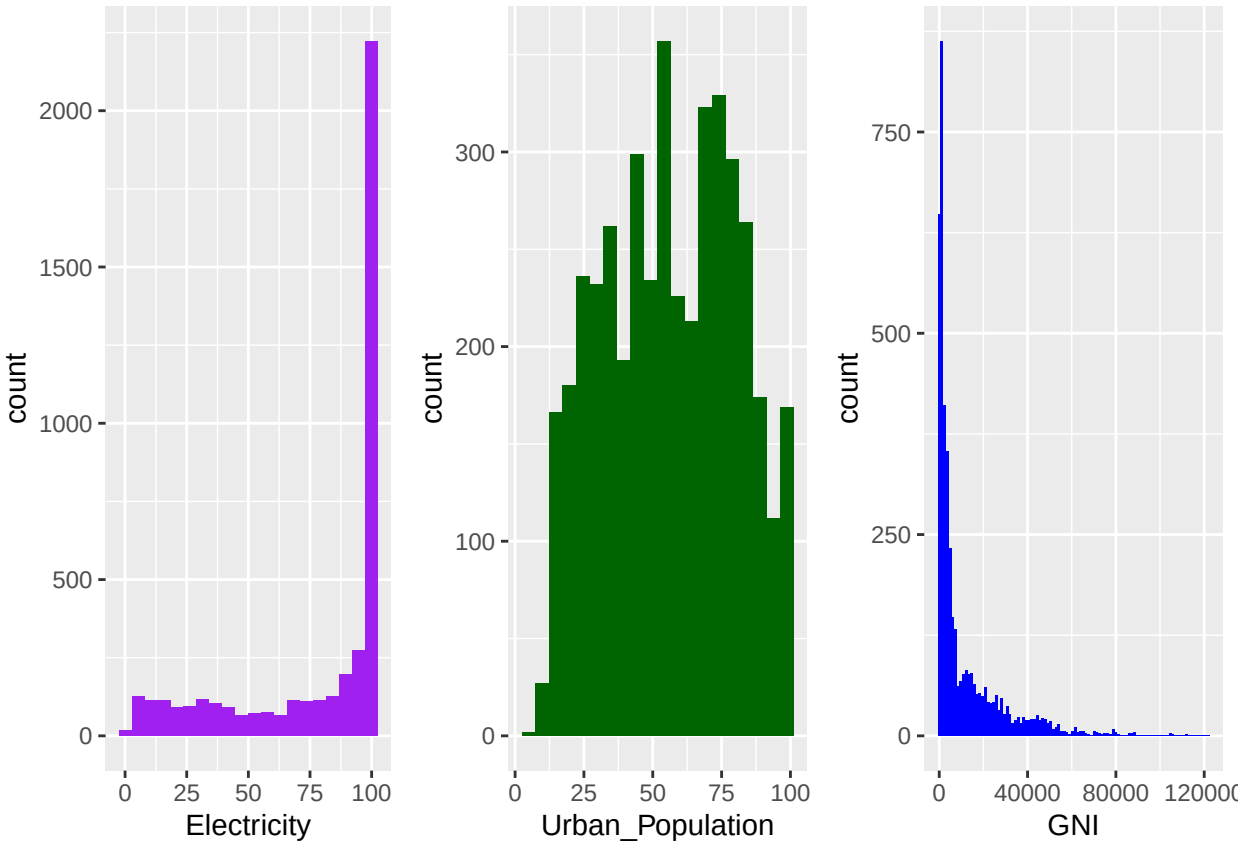
```
pt2 = ggplot(data, aes(x = Electricity)) +
  geom_histogram(bins = 20, fill="purple")
summary(data$Electricity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.01   60.42   98.04   78.08  100.00  100.00
```

```
pt3 = ggplot(data, aes(x = Urban_Population)) +
  geom_histogram(bins = 20, fill="darkgreen")
summary(data$Urban_Population)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.288  35.962  56.072  56.234  75.569 100.000
```

```
gridExtra::grid.arrange(pt2, pt3, pt1, ncol= 3)
```



Data Transformation

Using the Pipes, group_by and summarize functions to transform the dataframe into a table that only contains the mean of the three indicators in each year from 1990 to 2017. We can tell that from 1990 to 2003, both the Electricity and Urban Population decrease together with GNI; and then when GNI keep increasing from 2003 to 2013, the Electricity and Urban Population raise up as a result; while although the GNI decreases from 2014 to 2017, the Electricity and Urban Population still keep increasing.

```
develop <- data %>%
  group_by(year) %>%
  summarize(Electricity = mean(Electricity),
            Urban_Population = mean(Urban_Population),
            GNI = mean(GNI))
kable(develop, digits = 4, align = "c",
      booktabs = TRUE, caption = "Worldwide Development",
      col.names = c("Year", "Urban Population","Electricity","GNI" ))
```

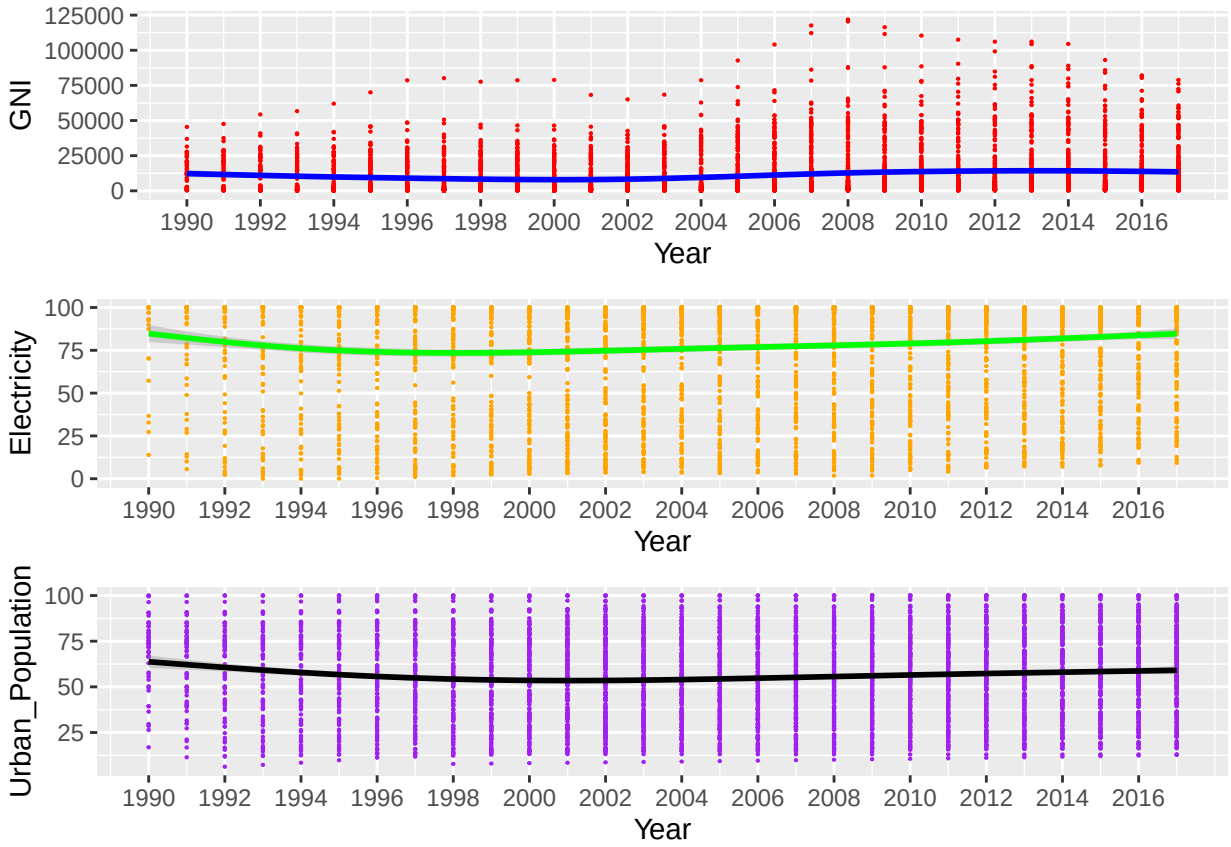
Scatter Plots

```
# scatter plot of GNI per Capita:
pt4 = ggplot(data, aes(x = year, y = GNI), title = "GNI") +
  geom_point(size = 0.1, color = "red") +
  scale_x_continuous("Year", breaks = seq(1990, 2017, 2)) +
  geom_smooth(color = "blue")

# scatter plot of Electricity:
pt5 = ggplot(data, aes(x = year, y = Electricity), title = "Electricity") +
```

Table 1: Worldwide Development

Year	Urban Population	Electricity	GNI
1990	91.4751	68.9220	13680.980
1991	84.2202	63.6047	11829.048
1992	78.4948	60.0153	10571.772
1993	74.3187	57.2752	9466.778
1994	74.5834	57.2027	9626.210
1995	74.9756	57.0669	9837.757
1996	73.1183	55.1188	9531.391
1997	74.2113	54.6744	9173.571
1998	73.5128	53.8863	8544.318
1999	73.6932	53.6627	8224.710
2000	73.8398	53.1118	7870.897
2001	73.8487	53.0274	7462.945
2002	75.2452	53.3400	7276.000
2003	75.6329	53.6597	7947.159
2004	76.1594	54.1972	9328.045
2005	76.3796	54.5735	10518.736
2006	76.9867	54.8677	11331.413
2007	77.6459	55.4673	12880.107
2008	77.7708	55.6885	13812.204
2009	78.3236	56.1350	13385.714
2010	78.8481	56.7448	13169.683
2011	79.2698	56.9063	13495.026
2012	80.5030	57.3321	14321.105
2013	81.1174	57.8071	15004.555
2014	81.8680	58.0694	14665.131
2015	82.6288	58.3189	14001.361
2016	83.8730	58.5311	13469.842
2017	84.9525	59.0878	13084.202

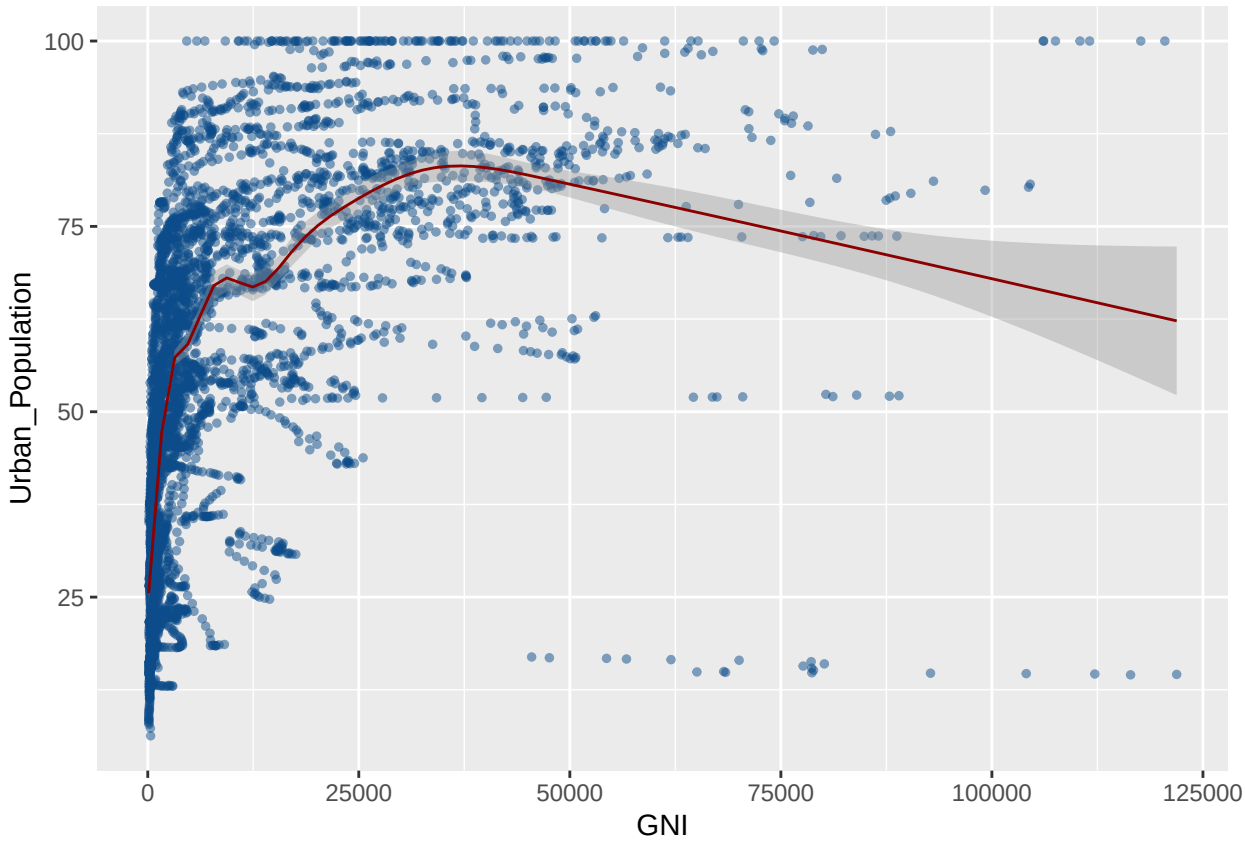
```
  geom_point(size = 0.1, color = "orange") +
  scale_x_continuous("Year", breaks = seq(1990, 2017, 2)) +
  geom_smooth(color = "green")

# scatter plot of Prevalence of Urban Population:
pt6 = ggplot(data, aes(x = year, y = Urban_Population), title = "Urban_Population") +
  geom_point(size = 0.1, color = "purple") +
  scale_x_continuous("Year", breaks = seq(1990, 2017, 2)) +
  geom_smooth(color = "black")

gridExtra::grid.arrange(pt4, pt5, pt6, ncol= 1)
```



From the above scatter plots, from 1990 to 2017, we can see there is no obvious flunctuation from GNI (mean), while the maximum fluctuates from 50,000 to 125,000; the electricity varies from 75% to 88%; and the urban population varies from 50% to 63%.

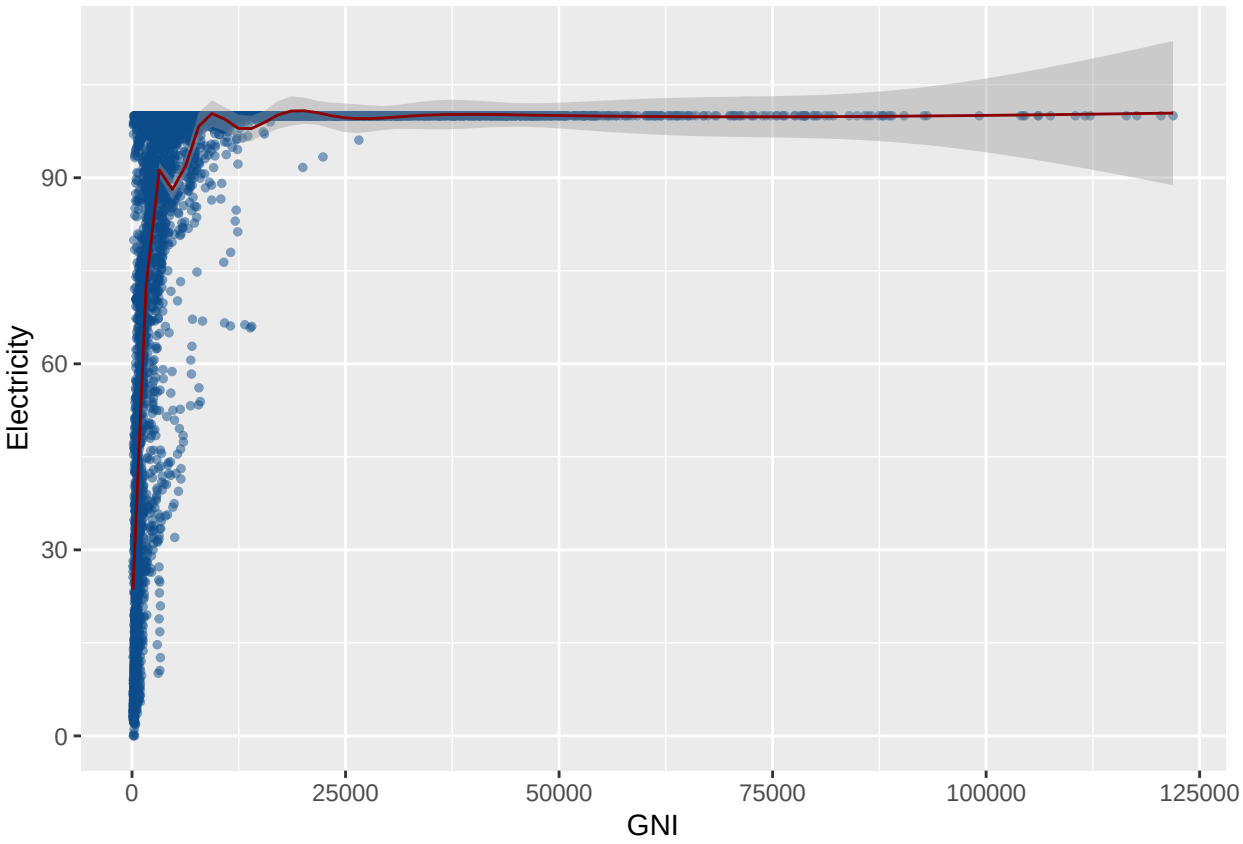
Linear Regression & Correlation.

```
#plot Urban Population vs. GNI per Capita.
pt8 = ggplot(data, aes(x = GNI, y = Urban_Population)) +
  geom_point(size = 1L, alpha = 0.5, colour = "#0c4c8a") +
  geom_smooth(color= "darkred", size = 0.5)
pt8
```



From the above plot we can tell that there is no direct correlation between GNI and Urban Population. While the urban population keep increasing with GNI increasing from 0 to 30,000, and it decreases when GNI gets larger than 30,000 (less than 125,000).

```
#plot Electricity vs. GNI per Capita.
pt9 = ggplot(data, aes(x = GNI, y = Electricity)) +
  geom_point(size = 1L, alpha = 0.5, colour = "#0c4c8a") +
  geom_smooth(color= "darkred", size = 0.5)
pt9
```



From above plot we can tell that before the GNI reaches around 8,500 (Atlas $), the percent of the access of electricity keeps increasing rapidly, and when GNI surpasses around 17,000 (Atlas $), the access of electricity increases to 100% and remains stable.

Conclusion:

Based on the plots and analysis of these three indicators, we can draw an initial conclusion that there is no clear linear regression relation between the GNI and urban population; while the access to electricity does correlated with the GNI, though the maximum access of electricity is 100%, there is a significant level of linear relationship between GNI and Electricity (in spite of the boundedness of the scale of electricity).