MrinalSoni

Mrinal Soni

10/28/2019

Problem 1: Group members posted on Piazza.

Problem 2:

Link to the Noah's solution: https://github.com/noahdemoes/DS5500_Homework1/blob/master/DS5500_Homework1.pdf

In terms of visualization, we both had similar visualization techniques (static plots) but our solution differs in a lot of ways. Noah's solution offers more details and insight into the asked quetion as he has captured country by country breakdown for all six regions from the 1850s to 2030 as compared to my one plot for four regions from 1960 to 2015. Noah's plots are more effective but since he has broken it down a continent by countries it is a little tough to read into the graph to comprehend which coloured line represents what country and how does the GDP for that particular country fluctuate, especially between the 1950s to early 2000s. Noah did a great job and it would've the perfect way to plot this breakdown using dynamic plots that would grow one by one for every continent.

Problem 3:

Link: https://github.com/akhil-vader/DS5500/blob/master/HW1.pdf

Akhil's visualizations for Life Expectancy, Child Mortality and GDP Per Capita over time are quite on point as all three plots seem to capture the essence of the question. As evident from our graphs, we both went for a similar approach (static line plots) where we tried to capture relationships between the three variables independently. In hindsight, it seems like a dynamic plot for this problem would've been a more interactive plot as it would have explined the growth and fall of the the three variables more effectively in a visually pleasing sense.

Problem 4:

Solution: To quantify the relationship between GDP Life Expectancy over Time, I decided to test Linear Regression and Multivariate Regression.

```
## Parsed with column specification:
## cols(
##   geo = col_character(),
##   time = col_integer(),
##   gdppercapita_us_inflation_adjusted = col_double()
## )
```

```
## Parsed with column specification:
## cols(
##   geo = col_character(),
##   time = col_integer(),
##   life_expectancy_years = col_double()
## )
```

```
## # A tibble: 6 x 4
##   geo    time gdppercapita_us_inflation_adjusted life_expectancy_years
##   <chr> <int>                              <dbl>                 <dbl>
## 1 abw    2010                             24272.                  75.1
## 2 afg    2002                               365.                  52.4
## 3 afg    2003                               377.                  53.0
## 4 afg    2004                               364.                  53.5
## 5 afg    2005                               389.                  53.9
## 6 afg    2006                               398.                  54.2
```

```
##
## Call:
## lm(formula = gdppercapita_us_inflation_adjusted ~ time, data = merged_data)
##
## Coefficients:
## (Intercept)         time
##   -325674.6        168.9
```

```
##
## Call:
## lm(formula = life_expectancy_years ~ time, data = merged_data)
##
## Coefficients:
## (Intercept)         time
##   -467.3042       0.2678
```

```
##
## Call:
## lm(formula = life_expectancy_years ~ gdppercapita_us_inflation_adjusted +
##     time, data = merged_data)
##
## Coefficients:
##                         (Intercept)  gdppercapita_us_inflation_adjusted
##                          -3.678e+02                           3.246e-04
##                                time
##                           2.161e-01
```

```
##
## Call:
## lm(formula = life_expectancy_years ~ log(gdppercapita_us_inflation_adjusted) +
##     time, data = merged_data)
##
## Coefficients:
##                               (Intercept)
##                                 -351.2779
## log(gdppercapita_us_inflation_adjusted)
##                                    4.7872
##                                      time
##                                    0.1898
```

Interpretation: The data doesnt appear to be in a linear format, but it seems like Linear Regression could be a good approach if the data was transformed, for example taking log of income affects the resuls. Logarithmic operation would transform the non-linear relationship to linear making linear regression a fit. When linear regression is performed, it's clear that for increase in time the life expectancy increases by a measure of 0.21 where as when we transform our variable and perform Log Regression we obtain a better relationship.

Problem 5:

Solution: For this problem, I decided to implement Simple Linear Regression and Multivariate Linear Regression to see if there is a relationship between Mortality Rate and GDP Income over Time.

```
## Parsed with column specification:
## cols(
##   geo = col_character(),
##   time = col_integer(),
##   gdppercapita_us_inflation_adjusted = col_double()
## )
```

```
## Parsed with column specification:
## cols(
##   geo = col_character(),
##   time = col_integer(),
##   child_mortality_0_5_year_olds_more_years_version_7 = col_double()
## )
```

```
## # A tibble: 6 x 4
##   geo    time gdppercapita_us_inflation~ child_mortality_0_5_year_olds_mor~
##   <chr> <int>                      <dbl>                                <dbl>
## 1 abw    2010                     24272.                                 NA
## 2 afg    2002                       365.                                129.
## 3 afg    2003                       377.                                126.
## 4 afg    2004                       364.                                122.
## 5 afg    2005                       389.                                119
## 6 afg    2006                       398.                                116.
```

```
##
## Call:
## lm(formula = gdppercapita_us_inflation_adjusted ~ child_mortality_0_5_year_olds_more_years_version_7
##     time, data = merged_data2)
##
## Coefficients:
##                                         (Intercept)
##                                           169408.97
## child_mortality_0_5_year_olds_more_years_version_7
##                                                -94.72
##                                                  time
##                                                -76.45
```

Interpretation: From the plot above (before implementing a model) its quite evident that there is a non-linear, negative correlation between the variable which implies that with increase in GDP there is a decrease in the Child Mortality Rate making it constant after a point. On performing Linear Regression we can observe from the coefficients values that with an increase in GDP, there is a decrease in child mortality rate implying that the model captures the relationship well but could have worked better with some transformations.

Autocorrelation: The autocorrelation in this data violates the assumption that linear regression makes and affects our analytics less reliable, which might lead to inaccurate results. hence, autocorrelation isn't ideal for our analysis.