IT 497 FINAL ASSIGNMENT

Eric Agyemang

library(tidyverse)

## -- Attaching packages ---------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readr)

# 1. obtain a dataset related to the research proposal:   
## Dataset is obtained from the World Development Indecators databese

# 2. Complete the Exploratory Data Analysis Checklist  
#.1. Formulate your question   
# . What key factors are statistically significant that affects the life insurance demand in Africa?   
  
# . Does evidence-based approach support this ?   
  
# . What are the implications for policymakers and the private sector wishing to scale up life insurance in Africa?

# .2. Read in your data   
library(readxl)  
Life\_Ins <- read\_excel("C:/Users/eagye/Desktop/Life\_Ins.xlsx")

# .3. Check the packaging   
  
# Checking number of roles  
nrow(Life\_Ins)

## [1] 110

# Checking number of columns  
ncol(Life\_Ins)

## [1] 27

# .4. Run str()   
str(Life\_Ins)

## tibble [110 x 27] (S3: tbl\_df/tbl/data.frame)  
## $ Country code: chr [1:110] "BEN" "B0TW" "BFA" "CIV" ...  
## $ Year : num [1:110] 2014 2014 2014 2014 2014 ...  
## $ POLSTAB : num [1:110] 0.412 0.365 0.38 0.354 0.332 ...  
## $ ADR : num [1:110] 6 6.1 4.7 5.2 8.2 6.4 5.1 3.7 5.1 9.5 ...  
## $ PSIZE : num [1:110] 10287 2089 17586 22648 90425 ...  
## $ LEB : num [1:110] 60.3 64.3 59.5 55.5 71.1 64.5 62.4 64.1 61 75.5 ...  
## $ PGR : num [1:110] 2.79 1.3 2.98 2.47 NA 2.81 2.34 2.61 2.83 1.39 ...  
## $ PPOOR : num [1:110] 40.1 19.3 40.1 46.3 NA 32.5 23.4 31.1 51.5 2.2 ...  
## $ PPLRA : num [1:110] 54.8 33.6 73.1 51 57.2 81 46.6 74.8 83.9 39.7 ...  
## $ MCTS : num [1:110] 84.2 163.3 71 97.6 105.4 ...  
## $ PIUI : num [1:110] 6 36.7 9.4 19.3 33.9 7.7 25.5 16.5 5.8 56.8 ...  
## $ EGROWTH : chr [1:110] "2975.9" "17264.400000000001" "1917.4" "4161.8999999999996" ...  
## $ EODB : num [1:110] 50.7 50.9 51.4 51.2 51.3 52.4 48.7 49.7 51.6 66.2 ...  
## $ NETIPC : num [1:110] -2.8 1.5 -0.2 0.2 9.8 2.4 -2 0.3 2 7.1 ...  
## $ GNIPC : num [1:110] 1460 1550 1460 1730 1890 ...  
## $ INFLAT : num [1:110] -3.1 -1.3 -0.6 4.7 2.4 NA 0.9 0.7 -0.4 11.9 ...  
## $ REALIR : num [1:110] 8.5 6.6 6 0.4 0 5.6 4.3 4.6 5.5 -2.6 ...  
## $ GDPOMT : num [1:110] 50.4 53.9 48.7 NA 54.1 ...  
## $ INSP : num [1:110] 0.9 0.9 0.8 1 1 0.8 0.9 1 0.9 2.7 ...  
## $ INSDENS : num [1:110] 5 6 6 7 8 8 7 8 8 203 ...  
## $ BUSFRD : num [1:110] 50 46 61 46 52 51 55 51 NA 69 ...  
## $ CPI : num [1:110] 38 42 38 42 41 39 37 36 39 63 ...  
## $ OPOE : num [1:110] 59 59 59 60 60 57 59 59 59 72 ...  
## $ FISFRED : num [1:110] 82 83 83 83 NA 68 68 68 69 81 ...  
## $ FINSF : num [1:110] 82 83 83 83 81 68 68 68 69 81 ...  
## $ PROPR : num [1:110] 25 30 30 38 42 30 30 30 36 70 ...  
## $ GOVSPEND : num [1:110] 2.42 2.68 2.7 3.18 3.77 1.35 NA 1.22 1.33 3.24 ...

# .5. Look at the top and the bottom of your data   
  
# Top of the data  
head(Life\_Ins)

## # A tibble: 6 x 27  
## `Country code` Year POLSTAB ADR PSIZE LEB PGR PPOOR PPLRA MCTS PIUI  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 BEN 2014 0.412 6 10287 60.3 2.79 40.1 54.8 84.2 6   
## 2 B0TW 2014 0.365 6.1 2089 64.3 1.3 19.3 33.6 163. 36.7  
## 3 BFA 2014 0.380 4.7 17586 59.5 2.98 40.1 73.1 71 9.4  
## 4 CIV 2014 0.354 5.2 22648 55.5 2.47 46.3 51 97.6 19.3  
## 5 EGY 2014 0.332 8.2 90425 71.1 NA NA 57.2 105. 33.9  
## 6 ETP 2014 0.508 6.4 98094 64.5 2.81 32.5 81 31.1 7.7  
## # ... with 16 more variables: EGROWTH <chr>, EODB <dbl>, NETIPC <dbl>,  
## # GNIPC <dbl>, INFLAT <dbl>, REALIR <dbl>, GDPOMT <dbl>, INSP <dbl>,  
## # INSDENS <dbl>, BUSFRD <dbl>, CPI <dbl>, OPOE <dbl>, FISFRED <dbl>,  
## # FINSF <dbl>, PROPR <dbl>, GOVSPEND <dbl>

tail(Life\_Ins)

## # A tibble: 6 x 27  
## `Country code` Year POLSTAB ADR PSIZE LEB PGR PPOOR PPLRA MCTS PIUI  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 ZAF 2018 0.522 8.1 57793 63.9 1.4 55.5 33.6 160. 56.2  
## 2 TOG 2018 0.410 5.1 7889 60.8 2.4 31.7 58.3 77.9 12.4  
## 3 TUNI 2018 0.394 12.3 11565 76.5 1.1 NA 31.1 128. 64.2  
## 4 UGA 2018 0.410 3.8 42729 63 3.7 40.9 76.2 57.3 23.7  
## 5 ZMB 2018 0.394 4 17352 63.5 1.4 64.4 56.5 89.2 14.3  
## 6 ZWE 2018 0.394 5.4 14439 61.2 1.4 70 67.8 89.4 27.1  
## # ... with 16 more variables: EGROWTH <chr>, EODB <dbl>, NETIPC <dbl>,  
## # GNIPC <dbl>, INFLAT <dbl>, REALIR <dbl>, GDPOMT <dbl>, INSP <dbl>,  
## # INSDENS <dbl>, BUSFRD <dbl>, CPI <dbl>, OPOE <dbl>, FISFRED <dbl>,  
## # FINSF <dbl>, PROPR <dbl>, GOVSPEND <dbl>

# .6. Check your "n"s   
# Looking at the the Life insurance density variable to see if all go well with that dependent variable  
table(Life\_Ins$INSDENS)

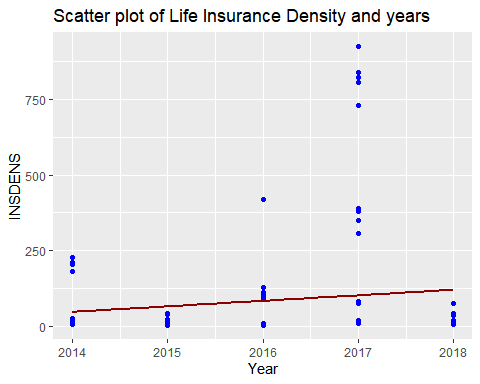
##   
## 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 19 20 21 22   
## 4 6 6 4 6 4 7 5 3 2 1 1 4 3 2 5 4 1 1 2   
## 23 26 36 37 38 39 40 42 43 74 75 81 89 98 101 112 127 182 203 207   
## 3 1 1 3 1 1 2 1 1 1 3 1 1 1 1 1 1 1 1 1   
## 211 226 308 350 379 390 418 730 806 821 840 925   
## 1 1 1 1 1 1 1 1 1 1 1 1

# # Looking at the the Economic Growth variable to see if all go well with that independent variable  
table(Life\_Ins$EGROWTH)

##   
## 1-37.5 10050.5 10216.4 1024.5999999999999   
## 1 1 1 1   
## 10266.700000000001 1027.2 10353.700000000001 1042.5   
## 1 2 1 1   
## 10487.5 10505.3 10519.7 10525   
## 1 1 1 1   
## 10570.4 10605.3 10763.8 10795.8   
## 1 1 1 1   
## 11014.5 1126.2 1130.5999999999999 11366.3   
## 1 1 1 1   
## 1149.7 1161.7 1196.5 1217.0999999999999   
## 1 1 1 1   
## 1262.5999999999999 12630.7 12703.4 12703.8   
## 1 1 1 1   
## 1273.8 1283.7 12840 12884.1   
## 1 1 1 1   
## 1289.5 1404.1 1447.2 1489.7   
## 1 1 1 1   
## 1516.5 1552.5 1656.6 16714.400000000001   
## 1 1 1 1   
## 17117.099999999999 17253.2 17264.400000000001 17634.099999999999   
## 1 1 1 1   
## 1779.1 1779.6 1888.8 1894.9   
## 1 1 1 1   
## 1917.4 1934.8 1950.3 1975.3   
## 1 1 1 1   
## 1991.1 2020.5 2021.6 2052.3000000000002   
## 1 1 1 1   
## 2053.5 2073.3000000000002 2074.9 2088.6   
## 1 1 1 1   
## 2103.5 2122.1 2131.5 2868.3   
## 1 1 1 1   
## 2934.7 2946 29461.4 2955.2   
## 1 1 1 1   
## 2958.2 2966.5 2975.9 3028.2   
## 1 1 1 1   
## 3044.5 3067.7 3130 3160.8   
## 1 1 1 1   
## 3203.9 3314.8 3443.5 3450.1   
## 1 1 1 1   
## 3467.9 3485 3521.5 3709.2   
## 1 1 1 1   
## 3824.8 3952.7 4046.2 4161.8999999999996   
## 1 1 1 1   
## 4203.8 4417.2 4615.7 4670.3999999999996   
## 1 1 1 1   
## 4675.8 4724.1000000000004 4830.8 4996.6000000000004   
## 1 1 1 1   
## 5028.8999999999996 5155.1000000000004 5190.3999999999996 5194.3999999999996   
## 1 1 1 1   
## 5284.9 5514.8 5516.4 6915.1   
## 1 1 1 1   
## 7109.2 7129.7 7314.2 7437.6   
## 1 1 1 1   
## 9932   
## 1

#.7. Validate with at least one external data source (for this you would normally look at existing data -- but for this assignment, just list 5 or more journals or conference proceedings that have research similar to your research proposal))   
  
# Validation could be made with the following journals publications  
  
# Journal of Service Science Research  
# The World Bank Economic Review  
# Journal of Policy Modeling  
# The Geneva papers on risk and insurance-issues and practice  
# Journal of Global Strategic Management  
# Household Survey Report on the Anidaso Policy of the Gemini Life Insurance Company

## .8. Try the easy solution first (create a simple graph that describes part of the data. This can be a scatter plot, box plot or histogram or anything else that you feel would be useful).  
  
# Plot a scatter plot of life insurance density against years  
  
library(ggplot2)  
ggplot(Life\_Ins, aes(x=Year, y= INSDENS)) + geom\_point(color="blue")+ geom\_smooth(method=lm, formula = y~x, se= FALSE, color="darkred", fill="blue")+ ggtitle("Scatter plot of Life Insurance Density and years")



# .9 Challenge your solution (for this part write a few sentences about what the graph in #8 shows -- why does the graph look the way it does)  
  
# Fitting a linear regression of all predictors against Life insurance density variable to determine the significant factors.  
  
Life\_Ins$EGROWTH<-as.numeric(Life\_Ins$EGROWTH)

## Warning: NAs introduced by coercion

good<-complete.cases(Life\_Ins)  
Life\_Ins1<-Life\_Ins[good,]  
Life\_Ins1<-select(Life\_Ins1, -1:-2)

lmLife<-lm(INSDENS~ . , data = Life\_Ins1, na.action = na.omit)  
summary(lmLife)

##   
## Call:  
## lm(formula = INSDENS ~ ., data = Life\_Ins1, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -66.413 -19.878 0.122 15.790 75.154   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.738e+02 1.258e+02 -1.381 0.17263   
## POLSTAB -5.432e+01 7.249e+01 -0.749 0.45669   
## ADR -9.511e-01 4.642e+00 -0.205 0.83840   
## PSIZE -3.492e-05 1.472e-04 -0.237 0.81334   
## LEB 1.120e+00 1.361e+00 0.823 0.41399   
## PGR 7.641e+00 9.963e+00 0.767 0.44627   
## PPOOR 4.967e-03 3.066e-01 0.016 0.98713   
## PPLRA 3.074e-02 6.348e-01 0.048 0.96155   
## MCTS 2.644e-01 3.241e-01 0.816 0.41792   
## PIUI -1.307e-03 5.496e-01 -0.002 0.99811   
## EGROWTH -5.408e-04 1.053e-03 -0.514 0.60952   
## EODB -2.445e+00 7.312e-01 -3.344 0.00146 \*\*   
## NETIPC -3.073e-01 9.774e-01 -0.314 0.75433   
## GNIPC 4.091e-03 1.717e-03 2.383 0.02055 \*   
## INFLAT 2.345e-01 8.575e-01 0.273 0.78552   
## REALIR 2.140e+00 8.257e-01 2.592 0.01211 \*   
## GDPOMT 3.028e-03 3.203e-01 0.009 0.99249   
## INSP 5.012e+01 3.412e+00 14.689 < 2e-16 \*\*\*  
## BUSFRD 3.737e-02 5.634e-01 0.066 0.94735   
## CPI 1.829e+00 9.544e-01 1.917 0.06032 .   
## OPOE 3.112e+00 1.482e+00 2.100 0.04016 \*   
## FISFRED -7.564e-01 4.132e+00 -0.183 0.85542   
## FINSF -7.598e-01 4.334e+00 -0.175 0.86146   
## PROPR 3.589e-01 4.971e-01 0.722 0.47326   
## GOVSPEND 1.762e+00 6.209e-01 2.838 0.00629 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.58 on 57 degrees of freedom  
## Multiple R-squared: 0.9827, Adjusted R-squared: 0.9754   
## F-statistic: 134.8 on 24 and 57 DF, p-value: < 2.2e-16

# The graph in 8 shows the relationship between the life insurance density variables and the years. Here, years are plotted on the x-axis against the life insurance density on the y-axis. The dark red line is the line of best fit which shows that most of the points are scatted far away from the line of best fit and this shows how heterogeneous the variables of life insurance in Africa are. These explains the reason why the graph looks like that.

#.10. Follow up (for this part, write a few sentences about the data -- could you use it to answer your research questions? If not, where might you find data that could be used to answer your research questions?   
   
# The data looks good though there are few NA's and graph in 8 shows that there are lots of outliers in the data points and the data points looks heterogeneous. This data could be used to answer the research questions.  
  
#From the above regression output, seven variables significantly affect life insurance density. These include the Ease of doing business, GNI per capita, Real Interest Rate, Insurance Penetration, Corruption Perception Index, Openness of Economy, and Government Spending.