hw07.Rmd

Homework 7

Automation Pipeline

First of all lets load all the required libraries.

```
suppressPackageStartupMessages(library(tidyr))
suppressPackageStartupMessages(library(stringr))
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(knitr))
suppressPackageStartupMessages(library(kableExtra))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(tidyverse))
```

Lets also define a function for formatting the tables.

```
tableFormat<-function(table,title=""){
  table %>%
    kable("latex",caption=title, align=c(rep('c', 5))) %>%
    kable_styling(bootstrap_options =
                    c("striped", "hover",  "responsive"),
                  position="center",font_size=14)
}
```

Lets take a look at the gapminder data downloaded from online.

```
input_data<- read.table(file = 'files/gapminder.tsv', sep = '\t', header = TRUE)
head(input_data)%>%
  tableFormat(title = "Downloaded Gapminder data")
```

This data has some problem. Lets take a look at that.

```
input_data %>%
  filter(str_detect(country, "Cote"))
```

```
##                                                                  country
## 1  Cote dIvoire\tAfrica\t1952\t40.477\t2977019\t1388.594732\nCote dIvoire
## 2   Cote dIvoire\tAfrica\t1962\t44.93\t3832408\t1728.869428\nCote dIvoire
## 3  Cote dIvoire\tAfrica\t1972\t49.801\t6071696\t2378.201111\nCote dIvoire
```

Table 1: Downloaded Gapminder data

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Afghanistan	Asia	1957	30.332	9240934	820.8530
Afghanistan	Asia	1962	31.997	10267083	853.1007
Afghanistan	Asia	1967	34.020	11537966	836.1971
Afghanistan	Asia	1972	36.088	13079460	739.9811
Afghanistan	Asia	1977	38.438	14880372	786.1134

Table 2: Cleaned Gapminder data

country	continent	year	lifeExp	pop	gdpPercap
Cote dIvoire	Africa	1952	40.477	2977019	1388.595
Cote dIvoire	Africa	1957	42.469	3300000	1500.896
Cote dIvoire	Africa	1962	44.930	3832408	1728.869
Cote dIvoire	Africa	1967	47.350	4744870	2052.050
Cote dIvoire	Africa	1972	49.801	6071696	2378.201
Cote dIvoire	Africa	1977	52.374	7459574	2517.737
Cote dIvoire	Africa	1982	53.983	9025951	2602.710
Cote dIvoire	Africa	1987	54.655	10761098	2156.956
Cote dIvoire	Africa	1992	52.044	12772596	1648.074
Cote dIvoire	Africa	1997	47.991	14625967	1786.265
Cote dIvoire	Africa	2002	46.832	16252726	1648.801
Cote dIvoire	Africa	2007	48.328	18013409	1544.750

```
## 4  Cote dIvoire\tAfrica\t1982\t53.983\t9025951\t2602.710169\nCote dIvoire
## 5 Cote dIvoire\tAfrica\t1992\t52.044\t12772596\t1648.073791\nCote dIvoire
## 6 Cote dIvoire\tAfrica\t2002\t46.832\t16252726\t1648.800823\nCote dIvoire
##   continent year lifeExp      pop gdpPercap
## 1    Africa 1957  42.469  3300000  1500.896
## 2    Africa 1967  47.350  4744870  2052.050
## 3    Africa 1977  52.374  7459574  2517.737
## 4    Africa 1987  54.655 10761098  2156.956
## 5    Africa 1997  47.991 14625967  1786.265
## 6    Africa 2007  48.328 18013409  1544.750
```

This shows that the data downloaded needs some cleaning up. This is done in the exploratory analysis file. Now lets source this file to check this dataset (gap_clean_data).

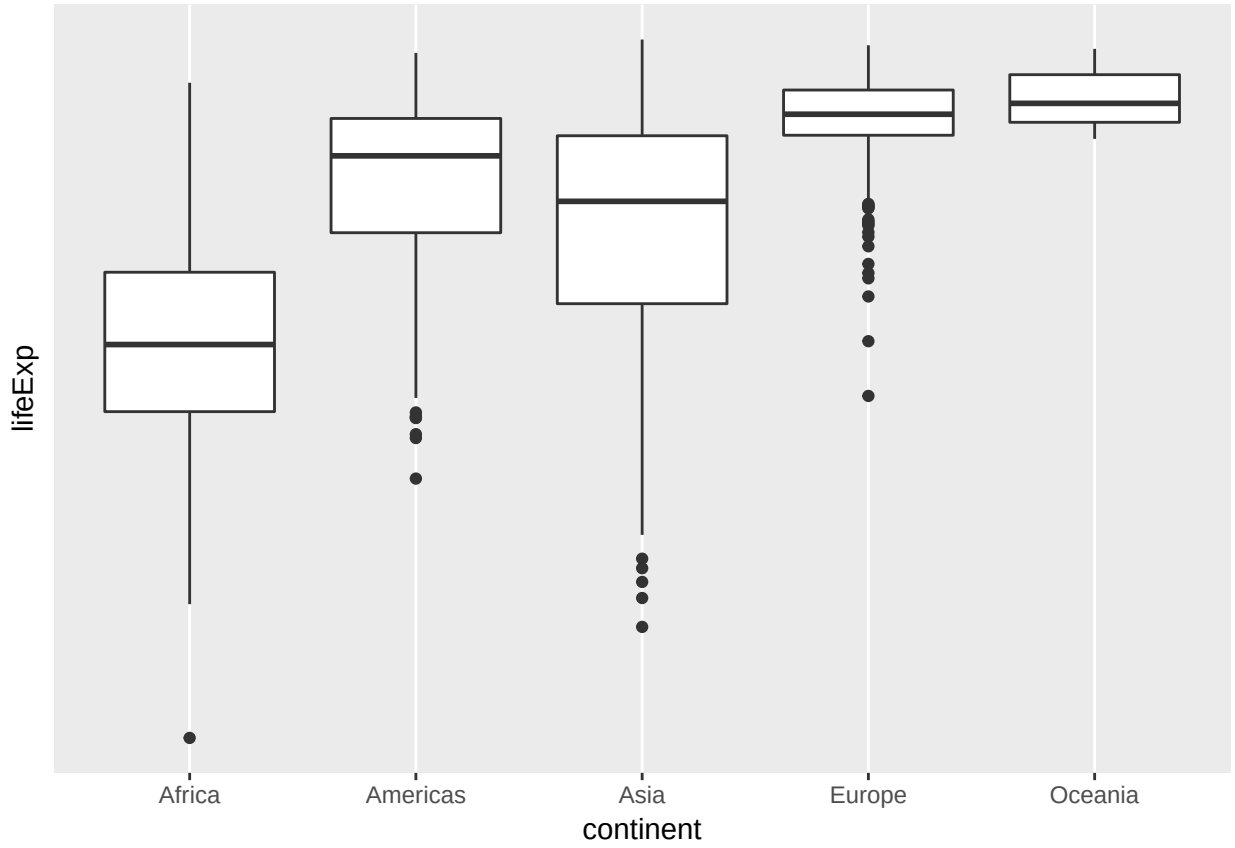
```
source('01_exploratory_analysis.R')
```

```
## 'data.frame':    1698 obs. of  6 variables:
##  $ country  : Factor w/ 147 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int  1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num  28.8 30.3 32 34 36.1 ...
##  $ pop      : int  8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 16317921 22
##  $ gdpPercap: num  779 821 853 836 740 ...
```

```
gap_clean_data %>%
  filter(str_detect(country, "Cote"))%>%
  tableFormat(title = "Cleaned Gapminder data")
```

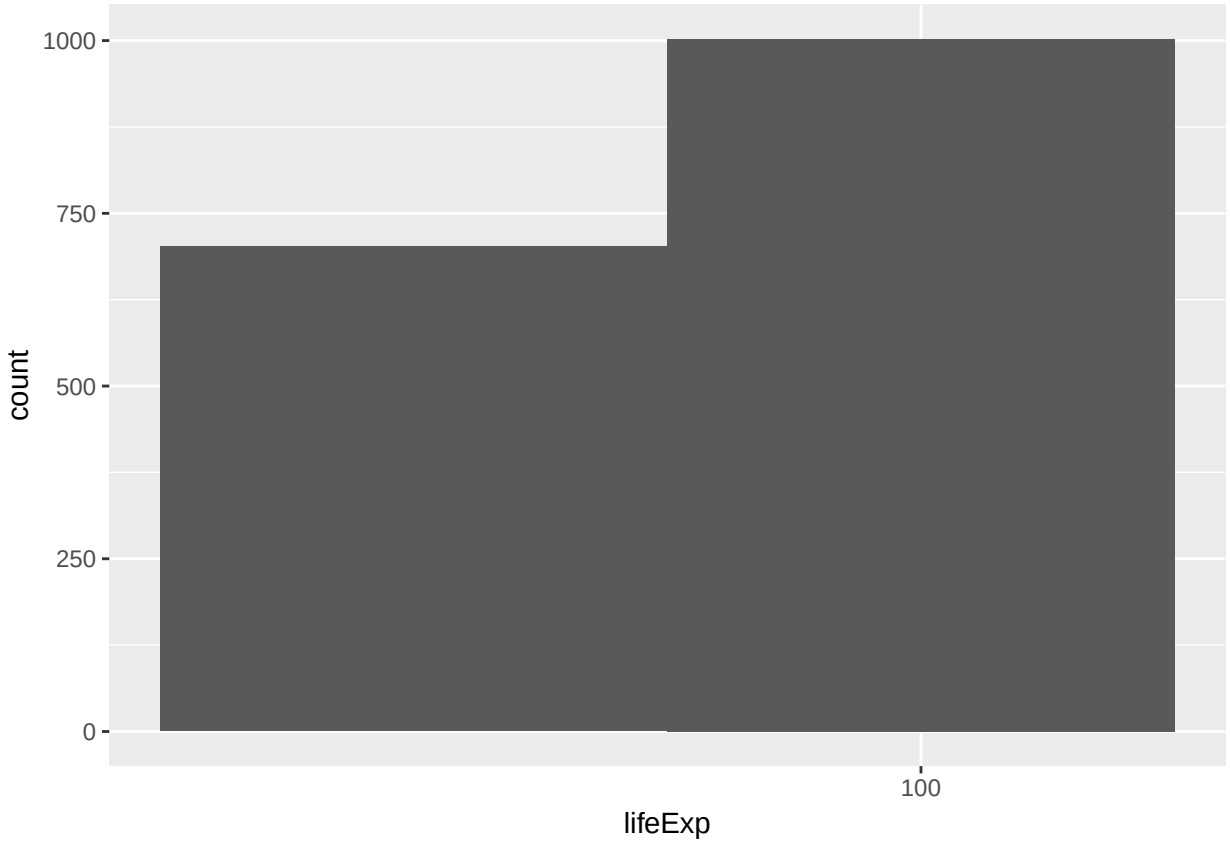
It looks the data is cleaned and the problem is solved. Now, lets look at the boxplot of lifeExp vs year.

```
boxplot
```

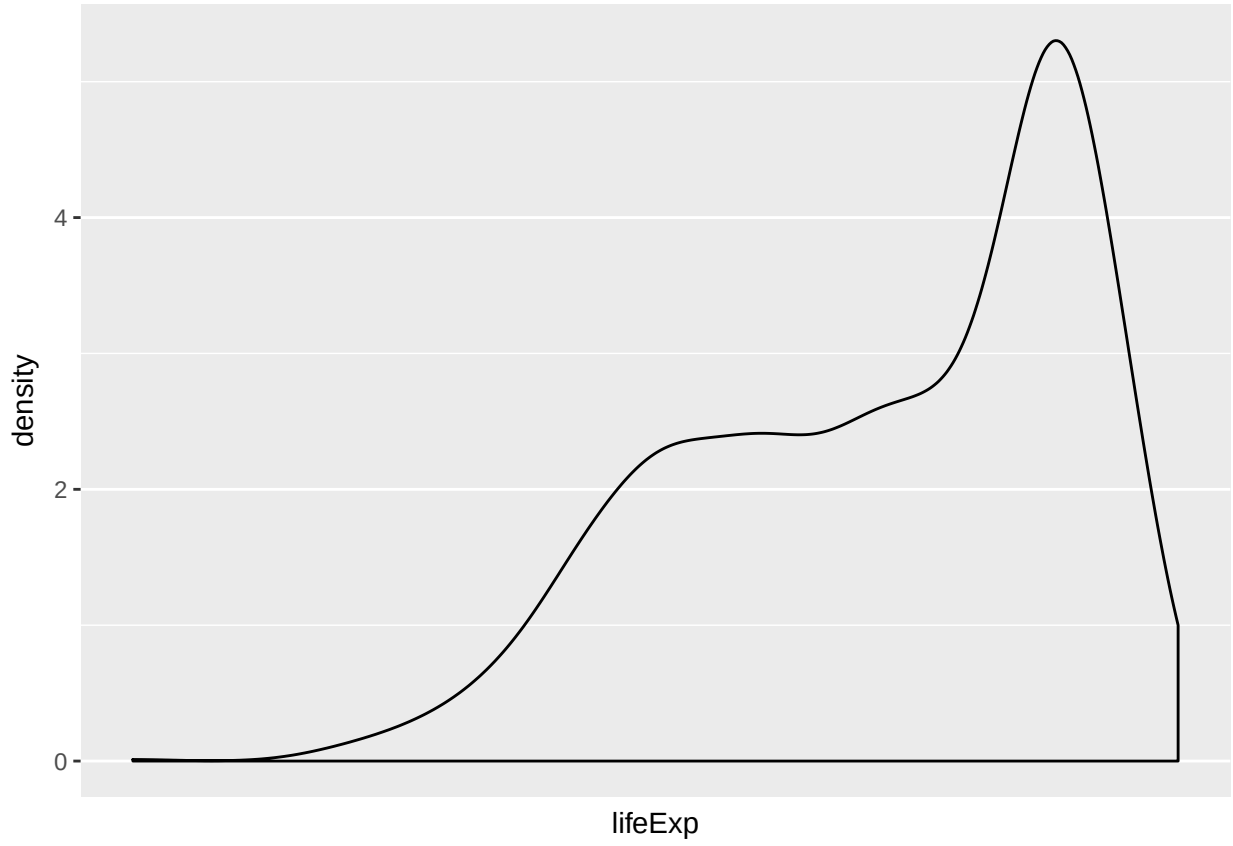


Now,lets look at few other plots of lifeExp such as histogram, density plot and frequency plot.

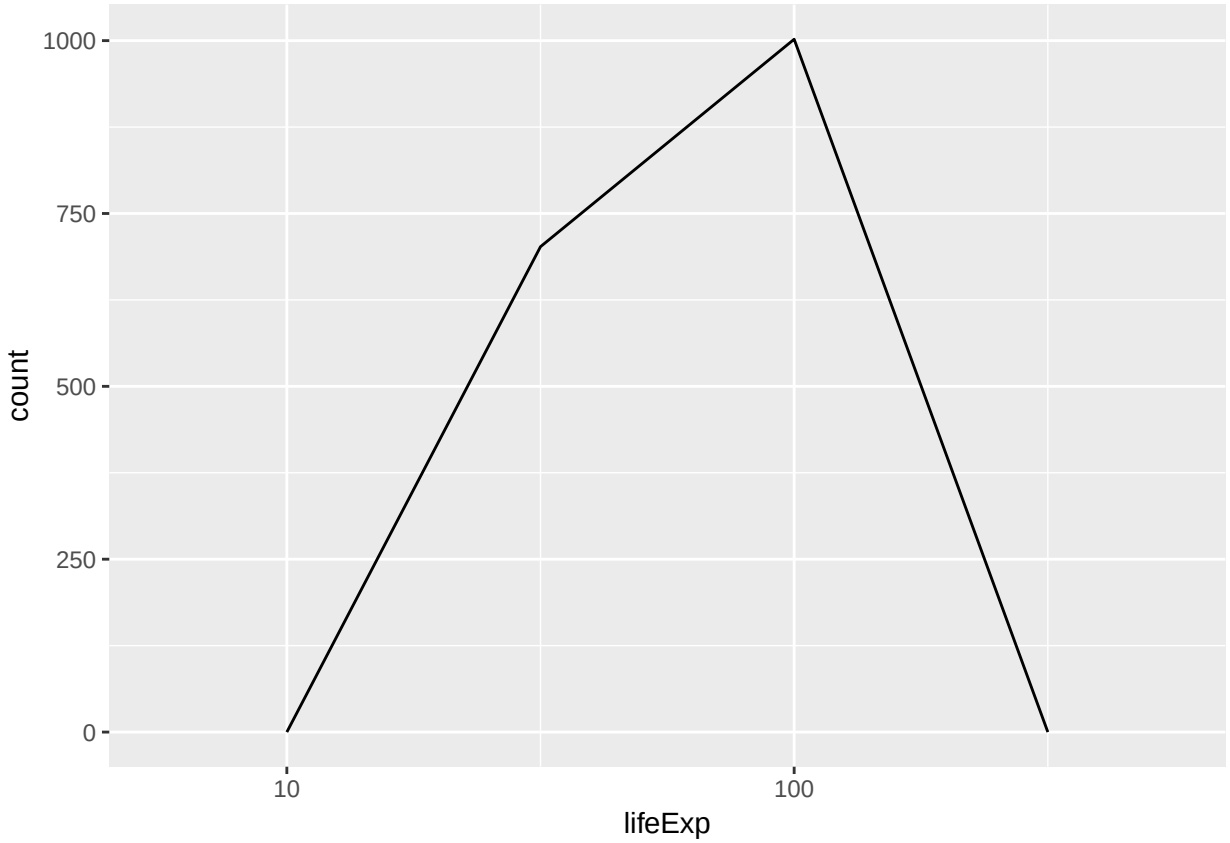
```
hist_plot
```




```
density_plot
```




```
freq_plot
```



Now lets take a look at the levels of continent factor variable before and after reordering.

```
#Before
gap_clean_data$continent%>%
  levels()
```

```
## [1] "Africa"   "Americas" "Asia"     "Europe"   "Oceania"
```

```
#After
gap_reordered$continent%>%
  levels()
```

```
## [1] "Oceania"  "Europe"   "Americas" "Asia"     "Africa"
```

Now, lets source the statistical analysis file.

```
source('02_statistical_analysis.R')
```

```
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
```

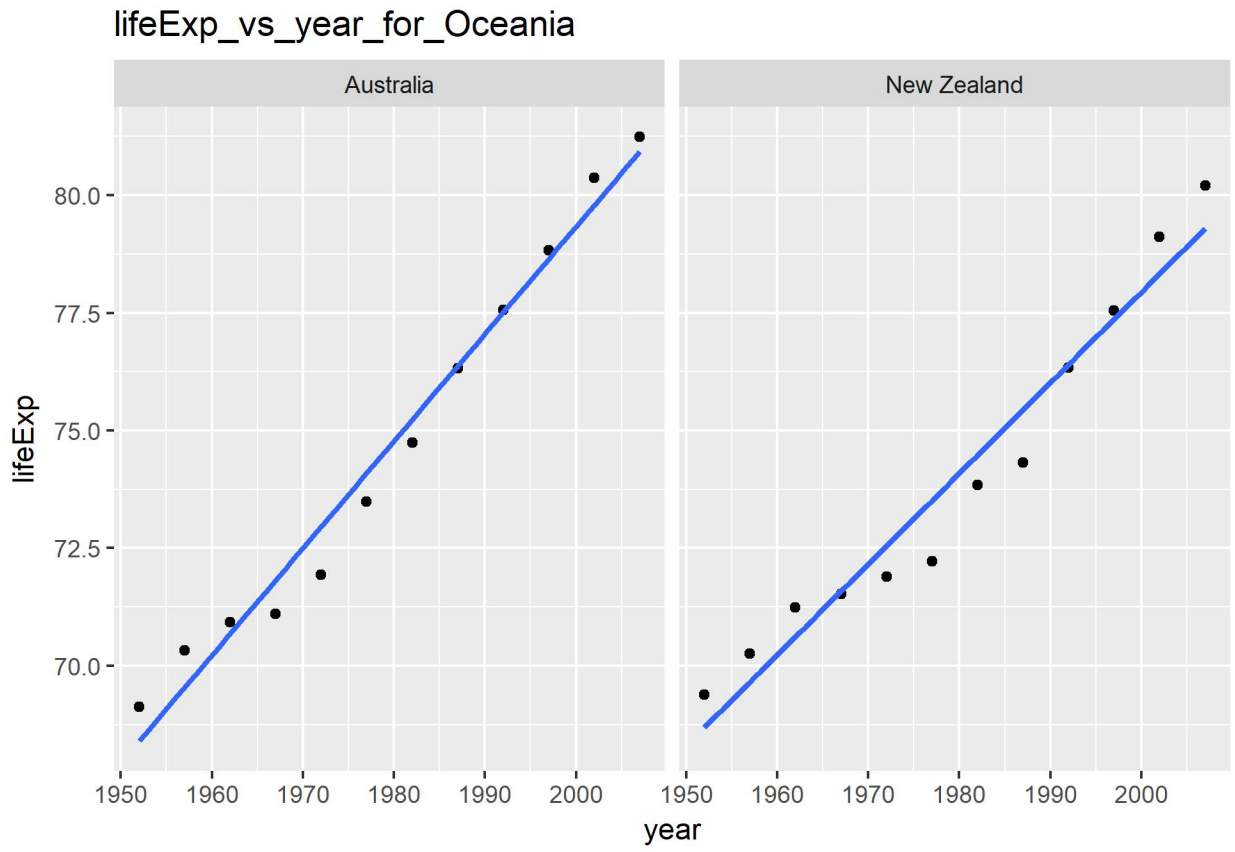
Now, lets look at the fitted result table.

```
fitted_result
```

```
## # A tibble: 140 x 7
## # Groups:   country, continent [140]
##        country continent  intercept     slope Res_Err_Std Res_Err_Variance
```

```
##          <fctr>    <fctr>      <dbl>     <dbl>       <dbl>           <dbl>
##  1 Afghanistan      Asia  -507.5343 0.2753287   1.2227880      1.49521045
##  2     Albania    Europe  -594.0725 0.3346832   1.9830615      3.93253302
##  3     Algeria    Africa -1067.8590 0.5692797   1.3230064      1.75034589
##  4      Angola    Africa  -376.5048 0.2093399   1.4070091      1.97967471
##  5   Argentina  Americas  -389.6063 0.2317084   0.2923072      0.08544349
##  6     Austria    Europe  -405.9205 0.2419923   0.4074094      0.16598240
##  7     Bahrain      Asia  -859.8258 0.4675077   1.6395865      2.68824402
##  8  Bangladesh      Asia  -936.2158 0.4981308   0.9766908      0.95392498
##  9     Belgium    Europe  -340.2412 0.2090846   0.2929025      0.08579187
## 10       Benin    Africa  -612.8340 0.3342329   1.1746910      1.37989891
## # ... with 130 more rows, and 1 more variables: R_squared <dbl>
```

Here is a look at one of the saved figures containing lifeExp vs year for each country in Oceania continent with regression line laid.



For plots of other continents, check out here.

Now, lets check the best 5 countries that fit our model perfectly in each continent except Oceania.

```
best_countries%>%
  select(country,continent,intercept,slope,R2_norm,std_norm)%>%
  tableFormat(title = "Best countries in Each Continent")
```

Now, lets check the worst 5 countries that didn't fit our model in each continent except Oceania.

```
worst_countries%>%
  select(country,continent,intercept,slope,R2_norm,std_norm)%>%
  tableFormat(title = "Worst countries in Each Continent")
```

Table 3: Best countries in Each Continent

country	continent	intercept	slope	R2_norm	std_norm
France	Europe	-397.7646	0.2385014	1.0000000	0.0799057
Sweden	Europe	-252.9239	0.1662545	0.9978525	0.0768994
Switzerland	Europe	-364.3421	0.2222315	0.9997657	0.0780409
Argentina	Americas	-389.6063	0.2317084	0.9975158	0.1421767
Brazil	Americas	-709.9427	0.3900895	1.0000000	0.1586794
Canada	Americas	-358.3489	0.2188692	0.9983348	0.1212330
Indonesia	Asia	-1201.9366	0.6346413	0.9998642	0.1146592
Iran	Asia	-924.4620	0.4966399	0.9977595	0.1180490
Israel	Asia	-455.0911	0.2671063	0.9975264	0.0649611
Pakistan	Asia	-748.3836	0.4057923	1.0000000	0.0715672
Comoros	Africa	-839.1671	0.4503909	0.9991745	0.0664286
Equatorial Guinea	Africa	-571.0228	0.3101706	0.9991925	0.0456170
Mali	Africa	-702.4815	0.3768098	0.9977719	0.0668388
Mauritania	Africa	-831.3813	0.4464175	1.0000000	0.0565590

Table 4: Worst countries in Each Continent

country	continent	intercept	slope	R2_norm	std_norm
Bulgaria	Europe	-218.64725	0.1456888	0.5478435	0.9111365
Montenegro	Europe	-509.69710	0.2930014	0.8037745	1.0000000
Poland	Europe	-318.23836	0.1962189	0.8416624	0.5887156
Romania	Europe	-243.28540	0.1574014	0.8074848	0.5309380
Jamaica	Americas	-369.50089	0.2213944	0.8072352	1.0000000
Trinidad and Tobago	Americas	-276.93502	0.1736615	0.7995687	0.8035163
Cambodia	Asia	-735.78684	0.3959028	0.6404537	1.0000000
Iraq	Asia	-409.01741	0.2352105	0.5472894	0.7206027
Korea, Dem. Rep.	Asia	-562.75907	0.3164266	0.7050021	0.6905999
Botswana	Africa	-65.49586	0.0606685	0.0341027	0.8482737
Rwanda	Africa	132.20498	-0.0458315	0.0171996	0.9101842
Zambia	Africa	165.60797	-0.0604252	0.0599759	0.6285138
Zimbabwe	Africa	236.79819	-0.0930210	0.0563630	1.0000000