MEMORY OF DATA SCIENCE FINAL PROJECT

Lorena Recio Cabeza

8 de Junio de 2018

INTRODUCTION

Due to we live in a world with a lot of inequality among countries that form it, it is interesting to see what factors cause this inequality. They can be political, economic, social or health factors. The main goal of this projects is to analyse these factors and determine which ones influence this inequality more.

DATASET DESCRIPTION

I have worked with WHO Public Dataset which you can download in the next website: https://data.world/resiport/who-dataset.

WHO dataset is in semicolon CSV format and it contains 202 countries in rows and 358 different features of this countries.

Within the 348 features, I have chosen the most striking to carry out the project. They are the following:

- Country
- CountryID
- Continent
- Adolescent.fertility.rate
- Gross.national.income.per.capita.ppp.international
- population in thousands total
- population annual growth rate
- population median age years
- total fertility rate per woman
- births attended by skilled health personnel
- tuberculosis detection rate under dots
- general_government_expenditure_on_health_as_percentage_of_total_government_expenditure
- hospital_beds_per_10_000_population
- number_of_dentistry_personnel
- total_expenditure_on_health_as_percentage_of_gross_domestic_product
- adult_mortality_rate_probability_of_dying_between_15_to_60_years_per_1000_population_both_sexes
- healthy_life_expectancy_hale_at_birth_years_both_sexes
- years_of_life_lost_to_communicable_diseases
- children_under_five_years_of_age_underweight_for_age
- per_capita_recorded_alcohol_consumption_litres_of_pure_alcohol_among_adults_gt_15_years
- population_with_sustainable_access_to_improved_drinking_water_sources_total
- population_with_sustainable_access_to_improved_sanitation_total
- prevalence_of_current_tobacco_use_among_adolescents_13_15_years_both_sexes
- agriculture_contribution_to_economy
- arms_imports
- broadband_subscribers
- co2_emissions
- capital_formation
- cell_phones_total
- consumer_price_index
- expenditure_per_student_primary

- expenditure_per_student_secondary
- expenditure_per_student_tertiary
- fixed_line_and_mobile_phone_subscribers
- exports_of_goods_and_services
- hiv_infected
- health_expenditure_per_person
- health_expenditure_total
- imports_of_goods_and_services
- income_per_person
- inequality_index
- inflation_gdp_deflator
- internet_users
- maternal_mortality
- military_expenditure
- personal_computers_total
- population_total
- sugar_per_person
- urban_population

METHODOLOGY

Firstly I've tried to clean the dataset and select the most interesting features to analyse. Once cleaned I've done some multiple or linear regression to see what are the most influence factors. Lately I've clusterized Finally I have visualized to compare and see these differences graphically.

RESEARCH FINDING

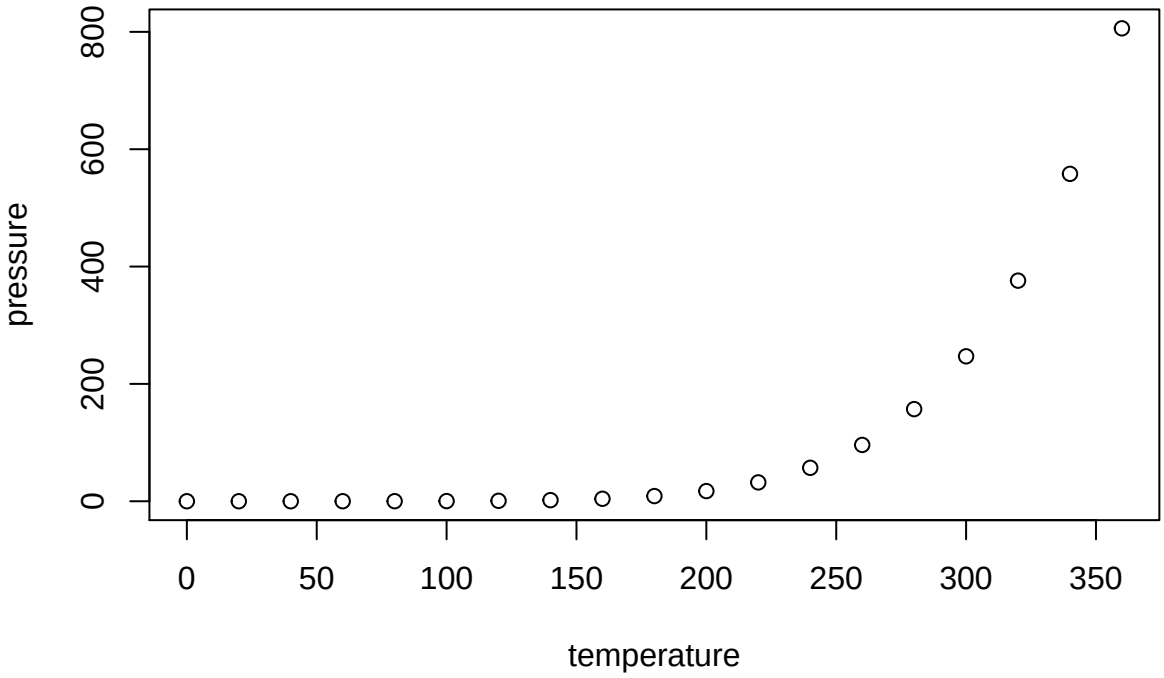
DESCRIPTION OF FRONTEND

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.