The quality score of the City of Toronto's open datasets is directly proportional to its metadata.

Dina Abu Ghosh,Anusha Champaneria,Ahmed Lugya

29 February 2020

Abstract

We use Catalogue Quality Scores dataset from Toronto's Open Data Portal to analyze the relationship between the quality score of datasets on the city of Toronto's open data portal and their metadata. The importance behind this analysis is that although more of the data is being published in various formats, data analysts, scientists, and students are on numerous occasions left with a time consuming and challenging task of identifying relevant data and understanding this data before any analysis can be done. Therefore it is necessary to publish datasets with satisfactory metadata to aid the usability of the datasets. We identified that as the quality score of the datasets increases, its metadata increases.

Introduction

Open data has several opportunities that it presents to society and among these include; increased transparency and enabling citizens to evaluate policies and hold governments accountable (European Open Data Portal,2020). In Canada, for example, the release of data drawn from charities' mandatory forms by the Canadian Charities Directorate helped uncover one of the biggest tax frauds in Canada's history. An analysis of this data that included sources of revenue, expenditures, assets and liabilities helped identify that in 2007, illegally operating charities alone carried out fraudulent donations amounting to over a billion dollars (David, 2010).

Metadata is the information that describes a dataset as well as its structure,and its aim is helping the users discover it. This information is usually released together with the dataset including collection methods, title, who the publisher and author of the dataset is, and when it was published or authored, what license is associated with the dataset and how often it is updated. Opening government data can help reduce the costs associated with acquiring data. However, concerns keep growing on whether the way such data are structured and made available to citizens makes the task of finding the required data easier.Furthermore, this raises questions with regard to the usability and quality of open data portals and the extent to which they fulfil the eight principles of open government data (Opengovdata, 2020).

The motivation behind this analysis of the Catalogue Quality Scores dataset was to help the City of Toronto in improving the overall quality of their open data portal. We were specifically interested in identifying the relationship between the quality score of datasets on the city of Toronto's open data portal and their metadata.Using simple linear regression, we observed that as the quality score of the datasets increases, its metadata increases.We also looked at the assumptions a linear regression model must hold true in order for the results to be trusted. In particular we looked at the mean of the residuals, whether the residuals have a constant variance and whether the residuals are normally distributed. We concluded the report by offering a discussion on the ethics and limitations of the dataset we used and pointing out a few recommendations the City of Toronto can follow to improve the quality of its datasets.

Data Analysis

The Catalogue Quality Score dataset shows the value of a dataset on a scale of bronze, silver, and gold. These are categorized by several different factors in terms of completeness, description, and how usable it is. There are a total of twelve parameters where all scores are ranged on a scale from 0 to 1.

Package is the name of the dataset. Accessibility is the score of how accessible a dataset is where 0 being hard to access and 1 being easily accessible. Completeness is the score of missing data where a score of 1 means very few missing values and a score of 0 means a lot of missing values. Freshness is determined by the gaps between collection and publication and between stated refresh rate and time last refreshed. The larger the gap the smaller the score, the smaller the gap the higher the score. Metadata is scored based on the completeness of optional metadata fields. These include collection method, limitations, topics, and email contact. The lower the score means less optional fields are filled and the higher the score the more optional fields are filled. Usability is scored based on ease of working with the data where 1 means easy and 0 means difficult. Score is the summary of accessibility, completeness, freshness, metadata, and usability into a single score. Score_norm is the score scaled from 0 to 1. Grade is a badge of bronze, silver and gold based on the score column. Grade_norm is the badge of bronze, silver, or gold based on the score_norm column. The value for grade_norm is the actual value displayed on the dataset page for data quality score. Recorded_at is the date and time the dataset was scored and version is the version of algorithm used to calculate the score.

Visualization

In this report, we used the simple linear regression with a goal of building a formula that defines quality scores as a function of the metadata. The general formula for the linear regression can be written as y = b0 + b1*x + e. Looking at the general formula, b0 and b1 are the regression beta coefficients where b0 is the predicted value when x = 0 hence its the intercept of the regression line. b1 is the slope of the regression line and e which is the error term, also known as the residual errors, is the part of y that can be explained by the regression model (Kassambara, 2018).

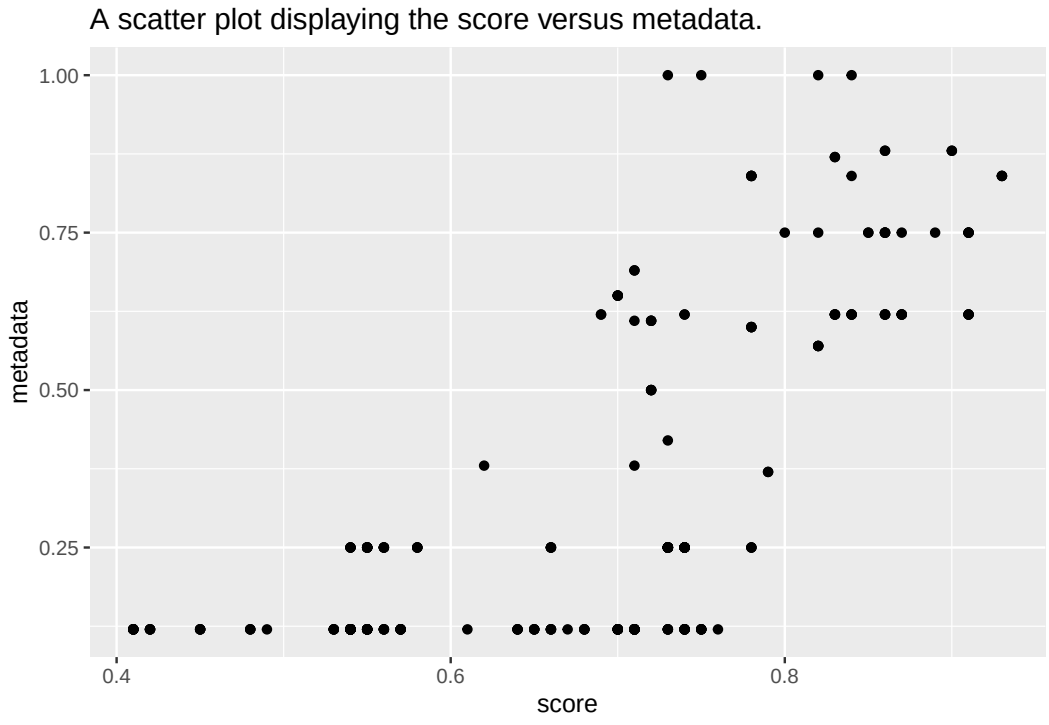

Figure 1:

Figure 1 above suggests a linearly increasing relationship between the quality score and the metadata variables. We went ahead and computed the correlation coefficient between the two variables to help us statistically measure the relationship between the relative movements of the two variables. Using the R function cor(); we got 0.7008119 hence the correlation is large enough.

Computation:

The simple linear regression is used to find the best line to predict the score on the basis of metadata.

```
##
## Call:
## lm(formula = score ~ metadata, data = score_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.197324 -0.067324  0.006226  0.080797  0.152676
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.564447   0.008989   62.80   <2e-16 ***
## metadata    0.357309   0.021843   16.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09228 on 265 degrees of freedom
## Multiple R-squared:  0.5024, Adjusted R-squared:  0.5005
## F-statistic: 267.6 on 1 and 265 DF,  p-value: < 2.2e-16
```

Figure 2.

The model summary outputs in Figure 2 above show components including Call, Residuals, Coefficients, Residual Standard Error (RSE), R-Squared (R2), and F-statisitcs. Call shows the function call used to compute the regression model. Residuals provide a quick view of the distribution of the residuals, which by definition have a mean zero. Therefore, the median should not be far from zero, and the minimum and maximum should be roughly equal in absolute value. Coefficients show the regression beta coefficients and their statistical significance. Predictor variables, that are significantly associated with the outcome variable, are marked by stars. RSE, R2and the F-statistic are metrics that are used to check how well the model fits our data (Kassambara, 2018).

Interpretation of Figure 2:

The estimated regression line equation can be written as follow: score = 0.5613 + 0.3610 * metadata. The intercept(b0) is 0.5613 which is the predicted score for metadata fields that were not filled. This means that, for every metadata field that is not filled, we can expect a quality score of 0.5613. The regression beta coefficient for the variable metadata, also known as the slope, is 0.3610. This means that for every increase in the score of metadata fields, we can expect an increase of 0.36190 in the quality score. R2 is 0.49 with p-value of 2.2e-16 which is less than 0.05 and three stars, therefore it is significant. All predictors explain 49% of the variance in the Dependent Variable.

Regression line:

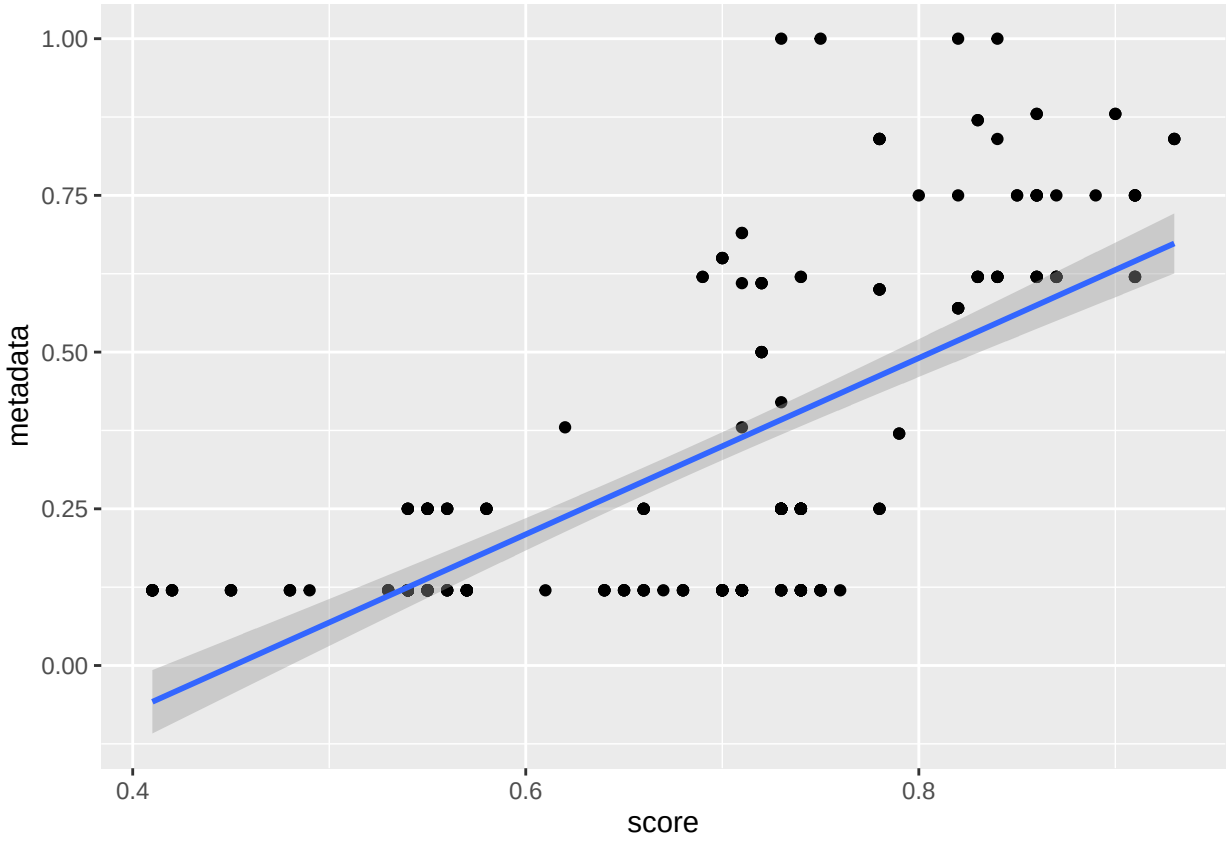

Figure 3.

Figure 3 shows that there is a positive linear relationship between the two variables. The scatter plot shows the more completed the metadata fields are the higher the quality of the dataset. However, many datasets with a low metadata score have a high overall score which can be due to others factors that make up the total quality score.

Despite the linear regression model fitting the data very well, there are various assumptions usually made about the data which must be satisfied inorder for our results to be trusted.In the next section we shall look at the assumptions such as the mean of the residuals, whether the residuals have a constant variance and whether the residuals are normally distributed.

Mean of the residuals.

For the assumption of mean of residuals to be held true for a regression model, it must either be zero or close to zero. Since 2.184647e-19, the mean of residuals for our model is approximately zero, we can say that this assumption holds true for our model.

Homoscedasticity of residuals.

Homoscedasticity of residuals is when the variance for all observations around the regression line are equal.

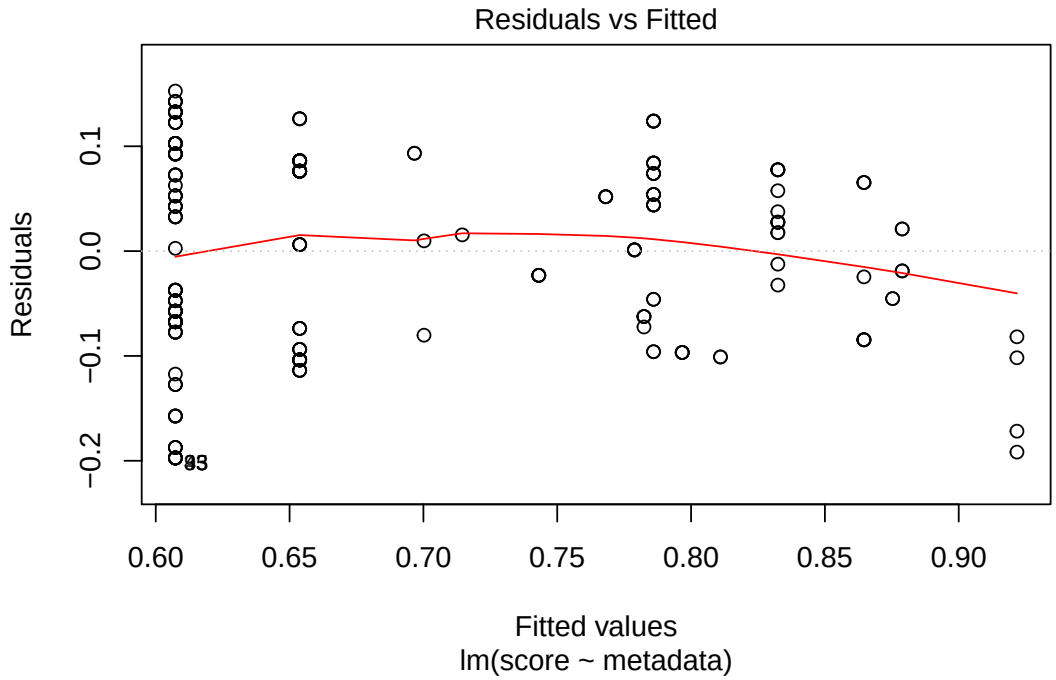

Figure 4

From Figure 4 above,points on the graph are random and the red line looks pretty flat, with no decreaing or increasing pattern.Hence, the assumption of homoscedasticity holds for our model.

Normality of residuals.

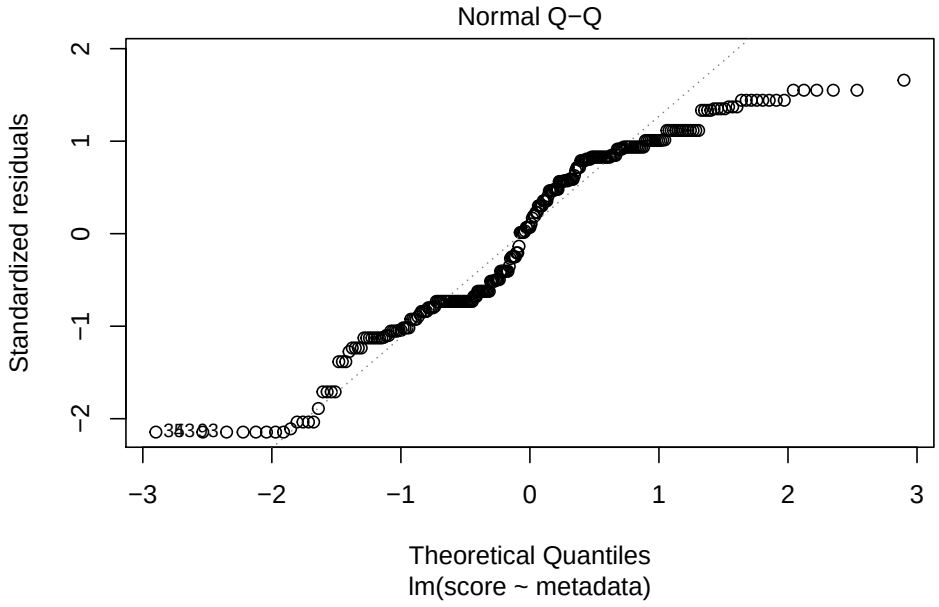

Figure 5.

The qqnorm() plot is used to check the assumption of normality of residuals. If the X and Y axis are normally distributed, the residuals should be normally distributed. Even though some deviation is expected, a perfectly normal distribution is when the points lie exactly on the line. Looking at Figure 5 above, this assumption does not hold for our model.

Weaknesses of the analysis method

Looking at Figure 5 above, there is a very large deviation away from the line as the standardized residuals increase. This might have been due to outliers which are known to have very huge effects on the linear regression method. Another weakness of our regression method is that it only works well for linear relationships between dependent and independent variables. Hence as a way forward,in order to obtain trusted answers to the researh questions we suggest using model validation with respect to outliers.

Ethics

The criteria that is used to assess the accessibility of the datasets is somewhat inadequate especially when compared to the 8 principles of open data portal (Opengovdata, 2020).The attributes used are not fully comprehensive and somewhat restrictive. None of the attributes listed touch upon biases or discriminatory biases which is a very important aspect of a dataset that should be included to the criteria when assessing the overall score of a dataset. According to the 8 principles of open data portal, having the dataset based on the public input is important as well but there is no mention of a similar attribute in the criteria that is adopted to assess the overall quality of the dataset. It also talks about datasets that are safe to open, meaning they do not contain any content that poses a threat on the person accessing it such as viruses or any cyber security threats associated with accessing the dataset.

Shortcomings

The format of datasets that are scored have to be in a SQL database, files such as excel or zip are not assessed because they are standardized. Both of the “read me” and “data” resources give an explanation of what the dataset is about. And data resource represents all the data. Most of the data sets on this portal have both a read me and data resource and when it came to calculating the quality score, they didn’t differentiate the two as they are both given the same weight.

Metadata attribute that is used to describe the data is not always an accurate depiction of the dataset and the quality of it. In most cases when it is accurate it is used as a way of misleading the readers. So, assessing the overall score based on the mentioned attributes can have its limitations as some are not an accurate reflection of the dataset so the score associated with the metadata attribute is inaccurate meaning the overall score will be flawed.

The weight of the attributes used to assess the datasets are all equal, however the weight should be allocated depending on the reliance of the attribute. If the more important attributes have higher weight, the overall score will not be skewed due to less significant ones in comparison to the more important ones. This can lead to a misrepresentation of the overall score of certain datasets. The significance of the attributes chosen for the criteria can be further looked in and based on that the weight can be adjusted. Even if the weights are changed that should be reflected and explicitly stated to inform the reader where the overall score came from and how it was calculated. Merely stating that the weight of the attributes is not the same while using the same scale for all of them is not a comprehensive indication.

APPENDIX

Source Code

Gather Catalogue Scorecard data

```
#### Contact details ####
# Title: Get data from Open data Toronto
# Purpose: This script gets data from Open Data Portal and saves it to inputs.
# Author: Dina, Anusha, Ahmed
# Last updated: 29 February 2020
# License: MIT License.

#### Set up workspace ####
library(rvest)
library(tidyverse)
library(opendatatoronto)
library(dplyr)

# get package
package <- show_package("473def30-1c87-45f1-95c4-06b9bf693fec")
package

# get all resources for this package
resources <- list_package_resources("473def30-1c87-45f1-95c4-06b9bf693fec")

# identify datastore resources; by default, Toronto Open Data sets datastore resource format to CSV for
datastore_resources <- filter(resources, tolower(format) %in% c('csv', 'geojson'))

# load the first datastore resource as a sample
data <- filter(datastore_resources, row_number()==1) %>% get_resource()
data


write.csv(data, "inputs/catalogue_data.csv")
```

Exploratory Data Analysis

```
#### Contact details ####
# Title: Graphs and Tables
# Purpose: This script draws graphs and tables out of the dataset.
# Author: Dina, Anusha, Ahmed
# Last updated: 29 February 2020
# License: MIT License.

#### Set up workspace ####
library(janitor)
library(tidyverse)
library(ggplot2)
```

```r
library(gvlma)
library(ggpubr)
theme_set(theme_pubr())

score_data <- read.csv("inputs/catalogue_data.csv")

#Scatterplot of score versus metadata
ggplot(score_data, aes(x = score_data$score, y =score_data$metadata)) +
geom_point()+
stat_smooth()+
ggtitle("A scatter plot displaying the score versus metadata."
)+
  xlab("score")+
  ylab("metadata")


#Correlation Coefficient
cor(score_data$score, score_data$metada)

#Computation
model <- lm(score ~ metadata, data = score_data)

#Model summary
summary(model)

#mean of residuals
set.seed(2)
mean(model$residuals)

#Homoscedasticity of residuals.
plot(model,1)

#Normality of residuals.
plot(model,2)

#Regression line
ggplot(score_data, aes(metadata, score)) +
  geom_point() +
  stat_smooth(method = lm)
```

References

- Yihui Xie (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.26.
- Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963
- Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595
- Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2019). skimr: Compact and Flexible Summaries of Data. R package version 2.0.2. https://CRAN.R-project.org/package=skimr
- Hadley Wickham and Lionel Henry (2019). tidyr: Tidy Messy Data. R package version 1.0.0. https://CRAN.R-project.org/package=tidyr
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- B. Hofner (2019). papeR: A Toolbox for Writing Pretty Papers and Reports, R package version 1.0-4, https://CRAN.R-project.org/package=papeR.
- Firke, Sam (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 1.2.1. https://CRAN.R-project.org/package=janitor.
- Müller, Kirill (2017). here: A Simpler Way to Find Your Files. R package version 0.1. https://CRAN.R-project.org/package=here.
- R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.
- Wickham, Hadley, et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686.
- Wickham, Hadley (2019). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.5. https://CRAN.R-project.org/package=rvest.
- Sharla Gelfand (2019). opendatatoronto: Access the City of Toronto Open Data Portal. R package version 0.1.1. https://CRAN.R-project.org/package=opendatatoronto
- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr
- European Open Data Portal (2020). The Benefits and value of open data. Retrieved from https://www.europeandataportal.eu/en/highlights/benefits-and-value-open-data
- David Eaves (April 14, 2010) Case Study: How Open data saved Canada $3.2 Billion. Retrieved from https://eaves.ca/2010/04/14/case-study-open-data-and-the-public-purse/
- Open Government Data (2020). The Annotated 8 Principles of Open Government Data. Retrieved from https://opengovdata.org/
- Thomas, D., Pastrana, S., Hutchings, A., Clayton, R. and Beresford, A. (2017). Ethical issues in research using datasets of illicit origin. Cambridge Cybercrime Centre, Computer Laboratory, p.2.
- Kassambara(March 10, 2018).Simple Linear Regression in R. Retrieved from http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/
- Edsel A. Pena and Elizabeth H. Slate (2019). gvlma: Global Validation of Linear Models Assumptions. R package version 1.0.0.3. https://CRAN.R-project.org/package=gvlma