WorldBankNew

2024-11-17

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

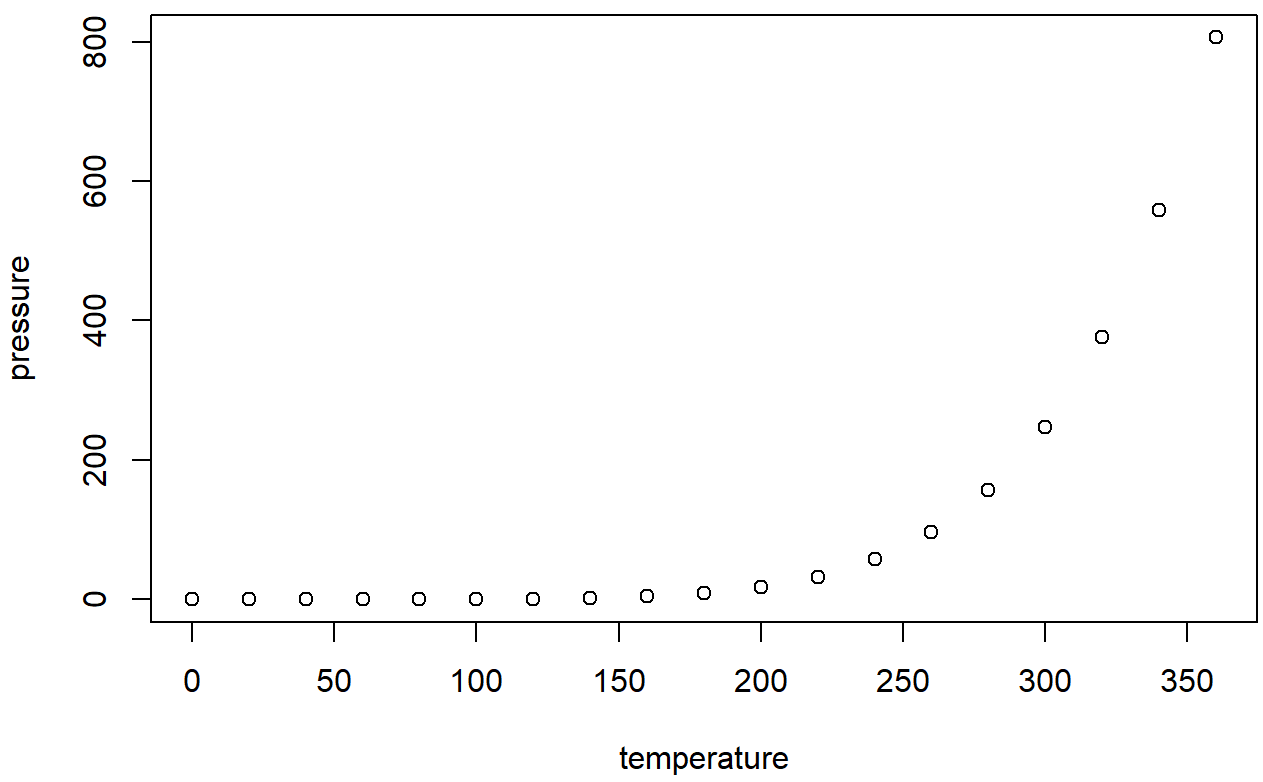
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
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.

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
##2. Importing data:
worldbank1 <- read.csv("C:/Users/conno/Downloads/WorldBankGR (1).csv")
```

```
head(worldbank1) ##shows first few rows of dataframe
```

```
##   Is.DRC Is.China Is.Russia Is.USA East.Asia...Pacific Europe.and.Central.Asia
## 1      0        0         0      0                   0                       1
## 2      0        0         0      0                   0                       1
## 3      0        0         0      0                   0                       1
## 4      0        0         0      0                   0                       1
## 5      0        0         0      0                   0                       1
## 6      0        0         0      0                   0                       1
##   Latin.America...Carrebian Middle.East...North.Africa North.America South.Asia
## 1                         0                           0             0          0
## 2                         0                           0             0          0
## 3                         0                           0             0          0
## 4                         0                           0             0          0
## 5                         0                           0             0          0
## 6                         0                           0             0          0
##   Sub.Saharan.Africa IncomeGroupRanking Year Birth.rate Death.rate
## 1                  0                  3 2014      12.26       7.22
## 2                  0                  3 2013      12.26       7.10
## 3                  0                  3 2012      12.20       7.00
## 4                  0                  3 2011      12.10       6.92
## 5                  0                  3 2010      12.00       6.84
## 6                  0                  3 2009      11.95       6.76
##   Electric.power.consumption         GDP GDP.per.capita
## 1                    2309.37 13228200000        4578.67
## 2                    2533.25 12776300000        4413.08
## 3                    2118.33 12319800000        4247.61
## 4                    2205.70 12890900000        4437.18
## 5                    1943.34 11927000000        4094.36
## 6                    1835.68 12044200000        4114.13
##   Individuals.using.the.Internet Infant.mortality.rate Life.expectancy
## 1                          60.10                   8.9           77.81
## 2                          57.20                   9.5           77.55
## 3                          54.66                  10.2           77.25
## 4                          49.00                  11.0           76.91
## 5                          45.00                  11.9           76.56
## 6                          41.20                  12.9           76.22
##   Population.density Unemployment..
## 1             105.44          17.49
## 2             105.66          15.87
## 3             105.85          13.38
## 4             106.03          13.48
## 5             106.32          14.09
## 6             106.84          13.67
```

```
nrow(worldbank1) ##shows number of rows in dataframe
```

```
## [1] 2775
```

```
##5. Partitioning between training, validation, and test
sample <- sample.int(n = nrow(worldbank1), size = nrow(worldbank1)*0.7, replace = F)
worldbank1_train <- worldbank1[sample, ] ##Yields training dataset
worldbank1_vt <- worldbank1[-sample, ] ##Yields validation & test portion

sample <- sample.int(n = nrow(worldbank1_vt), size = nrow(worldbank1_vt)*0.5, replace = F) ##Validation percentage = what pe
rcentage of this validation + test block should go into validation
worldbank1_validation <- worldbank1_vt[sample, ] ##Yields validation dataset
worldbank1_test <- worldbank1_vt[-sample, ] ##Yields test portion
```

```
##6. Train linear regression model
linear_regression_model <- lm(Life.expectancy ~ Is.DRC + Is.China + Is.Russia + Is.USA + East.Asia...Pacific + Europe.and.Ce
ntral.Asia + Latin.America...Carrebian + Middle.East...North.Africa + North.America + South.Asia + Sub.Saharan.Africa + Inco
meGroupRanking + Year + Electric.power.consumption + GDP + GDP.per.capita + Individuals.using.the.Internet + Population.dens
ity + Unemployment.., data=worldbank1_train) ##Or, can use all predictors except one using the ~ . -EXCLUDEDVARIABLE notatio
n
summary(linear_regression_model) ##Outputs summary of model & coefficients
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Is.DRC + Is.China + Is.Russia +
##     Is.USA + East.Asia...Pacific + Europe.and.Central.Asia +
##     Latin.America...Carrebian + Middle.East...North.Africa +
##     North.America + South.Asia + Sub.Saharan.Africa + IncomeGroupRanking +
##     Year + Electric.power.consumption + GDP + GDP.per.capita +
##     Individuals.using.the.Internet + Population.density + Unemployment..,
##     data = worldbank1_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0053  -1.9231   0.2278   2.0746  13.5340
##
## Coefficients: (1 not defined because of singularities)
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2.856e+02  4.100e+01  -6.966 4.46e-12 ***
## Is.DRC                         1.055e+00  1.054e+00   1.001   0.3170
## Is.China                      -5.221e-01  9.149e-01  -0.571   0.5683
## Is.Russia                     -8.591e+00  8.120e-01 -10.580  < 2e-16 ***
## Is.USA                        -8.976e+00  1.638e+00  -5.478 4.86e-08 ***
## East.Asia...Pacific            1.258e+01  3.541e-01  35.535  < 2e-16 ***
## Europe.and.Central.Asia        1.348e+01  2.937e-01  45.884  < 2e-16 ***
## Latin.America...Carrebian      1.373e+01  3.063e-01  44.814  < 2e-16 ***
## Middle.East...North.Africa     1.332e+01  3.204e-01  41.580  < 2e-16 ***
## North.America                  1.498e+01  8.639e-01  17.343  < 2e-16 ***
## South.Asia                     1.079e+01  4.628e-01  23.324  < 2e-16 ***
## Sub.Saharan.Africa                    NA         NA      NA       NA
## IncomeGroupRanking             2.859e+00  1.351e-01  21.156  < 2e-16 ***
## Year                           1.673e-01  2.042e-02   8.193 4.59e-16 ***
## Electric.power.consumption     4.197e-05  2.296e-05   1.828   0.0677 .
## GDP                            5.699e-13  1.135e-13   5.022 5.59e-07 ***
## GDP.per.capita                 6.378e-05  9.180e-06   6.948 5.04e-12 ***
## Individuals.using.the.Internet 1.126e-02  6.785e-03   1.659   0.0972 .
## Population.density             6.696e-04  1.329e-04   5.037 5.16e-07 ***
## Unemployment..                -3.382e-02  1.542e-02  -2.193   0.0284 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.411 on 1923 degrees of freedom
## Multiple R-squared:  0.8437, Adjusted R-squared:  0.8423
## F-statistic: 576.9 on 18 and 1923 DF,  p-value: < 2.2e-16
```

```
##7. Produce predictions on validation & test data using linear regression model
VALIDATION_PREDICTIONS <- predict(linear_regression_model, newdata=worldbank1_validation)
worldbank1_validation$LINEAR_PRED = VALIDATION_PREDICTIONS ##saves predictions into validation dataframe

TEST_PREDICTIONS <- predict(linear_regression_model, newdata=worldbank1_test)
worldbank1_test$LINEAR_PRED = TEST_PREDICTIONS ##saves prediction into test set dataframe
```

```
##8. Evaluate predictions on validation & test data: Caret package must be loaded to call this function!
postResample(pred = worldbank1_validation$LINEAR_PRED, obs = worldbank1_validation$Life.expectancy) ##evaluating validation
predictions
```

```
##      RMSE  Rsquared       MAE
## 3.2109552 0.8527866 2.4855484
```

```
postResample(pred = worldbank1_test$LINEAR_PRED, obs = worldbank1_test$Life.expectancy) ##evaluating test predictions
```

```
##      RMSE  Rsquared       MAE
## 3.4956621 0.8254692 2.6273678
```