MATH 6010 - Template RMarkdown

Wim R. M. Cardoen

8/19/2022

Data retrieval

You can retrieve your data set in different ways:

- log into kaggle and download the .csv file. (to be found in the **data** sub directory as well)
- use the kaggle Python-API. (requires a Python **pip install**)

Working with the downloaded data set

- Read the dataset
- Print the header of the data frame

```
mydata <- read.csv(file="./data/insurance.csv", header=TRUE)
mys <- sprintf("  Num. Rows:%d   Num. Columns:%d\n", dim(mydata)[1], dim(mydata)[2])
cat(mys)
```

```
##   Num. Rows:1338   Num. Columns:7
```

```
for(item in colnames(mydata)){
  mys <- sprintf("'%s'\n",item)
  cat("  Column:", mys)
}
```

```
##   Column: 'age'
##   Column: 'sex'
##   Column: 'bmi'
##   Column: 'children'
##   Column: 'smoker'
##   Column: 'region'
##   Column: 'charges'
```

```
head(mydata)
```

```
## # A tibble: 6 x 7
##     age sex      bmi children smoker region    charges
##   <int> <chr>  <dbl>    <int> <chr>  <chr>       <dbl>
## 1    19 female  27.9        0 yes    southwest  16885.
## 2    18 male    33.8        1 no     southeast   1726.
## 3    28 male    33          3 no     southeast   4449.
## 4    33 male    22.7        0 no     northwest  21984.
## 5    32 male    28.9        0 no     northwest   3867.
## 6    31 female  25.7        0 no     southeast   3757.
```

Making plots in R & Python

There are several options to generate plots, e.g.:

- R:
  - ggplot2
  - regular R plot function
- Python (see Jupyter Notebook)
  - matplotlib
  - seaborn

Perform Linear Regression

- R: use R's **lm()** (i.e. linear models)
- Python: use of the statsmodels module

Use of Latex