MATH 6010 - Template RMarkdown

Wim R. M. Cardoen

8/19/2022

1 Data retrieval

You can retrieve your data set in different ways:

- log into kaggle and download the .csv file from:

  https://www.kaggle.com/datasets/mirichoi0218/insurance

  The dataset can be found in the **data** sub directory.
- use the kaggle Python-API. (requires a Python **pip install**)

2 Using R

- R binaries exist for several OS (Windows, MacOS, Linux):
  - https://cran.r-project.org/
  - Microsoft R Open is an enhanced version (e.g. multi-threading)
- R source code
  - Use e.g. Intel MKL or OpenBlas to increase the speed of your R code.
- IDE:
  - RStudio (Windows, MacOS, Linux).
- Statistics/Linear Regression:
  - R's strength!
- Plotting in R:
  - ggplot2
  - regular R plot function

- Tutorials
  - The Art of R Programming
  - Advanced R

3 Using Python

- Python binaries can be obtained from anaconda.com.

  You can either choose anaconda or miniconda (light-weight version of the former).
- Python source code
  - Besides Python, you will need to install other packages such as numpy, scipy, pandas, matplotlib, statsmodels
  - Use the Intel MKL or OpenBlas libraries for NumPy, SciPy.

  - **Only** go this route if you **really** understand code compilation, linking and have time to perform the installation.

- IDE:
  - PyCharm
  - Jupyter

    You can use either jupyter notebook or jupyter lab.

- Statistics/Linear Regression:
  - statsmodels module

    Using R-style formulas and dataframes.

- Plotting
  - matplotlib
  - seaborn

- Tutorials
  - Python Tutorial
  - Intro to NumPy & SciPy
  - Matplotlib examples (including code)
  - Seaborn examples (including code)
  - Statsmodels examples (including code)

4 Using Latex

I have created *template* Latex and BibTeX files which may serve as a start.

- Binaries:
  - MikTex: available for Windows, MacOs and Linux
- Source Code:
  - CTAN
- Tutorials:
  - A Brief Introduction to TeX and LaTeX (Dr. Nelson Beebe)
  - Latex Tutorial

5 R

5.1 Exploration of the 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.
```