Modernizing the Curricula of Statistics courses through Statistical Learning

New Approaches to Statistical Learning in Developing Countries
3rd LISA 2020 Symposium

Marcus Nunes

29 October 2020

Federal University of Rio Grande do Norte

Motivation

Motivation

- It is easier than ever to fit complex models to data
- Many data repositories are available for free
- Free software and data can be used
- How statistical educators can take advantage of new technologies

Motivation

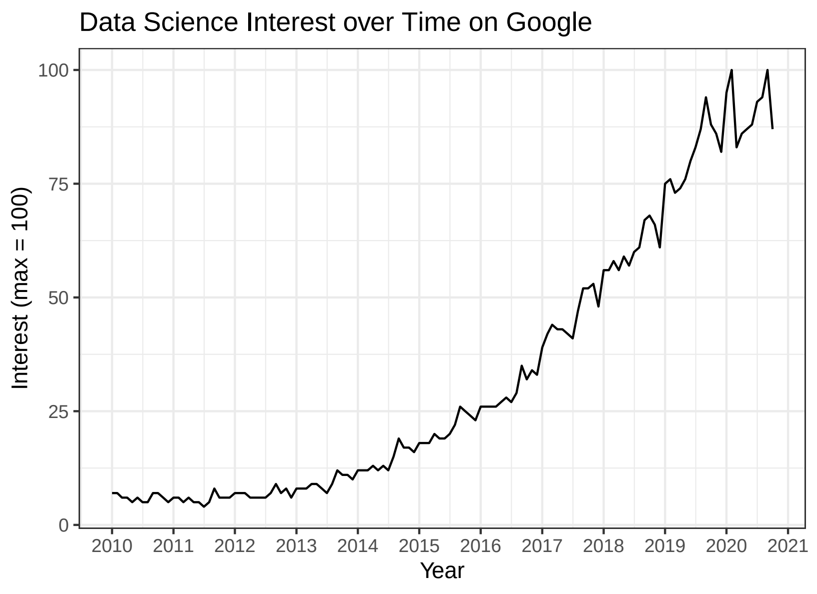
2014 ASA Guidelines:

- Increased importance of data science
- Real applications
- More diverse models and approaches
- Ability to communicate

Motivation

- These guidelines have been applied in a course called *Introduction to Big Data Modeling*
- Offered since 2015 at the Federal University of Rio Grande do Norte, Brazil
- It is offered regularly as an elective course to second-year students
- Pre-requisites: basic statistical inference (t-test, ANOVA, simple linear regression) and R programming

Increased Importance of Data Science



Real Applications

- One of the pillars of *Introduction to Big Data Modeling* is the use of real datasets
- According to Hicks and Irizarry (2016), students are more motivated when they see data collected from the real world
- Simple and complex datasets: Fisher's Iris dataset and FIFA Soccer

Real Applications

- As the course advances, the datasets become more complex
- There are many free great sources with interesting datasets
- US Government open data and Brazilian Institute of Geography and Statistics are two of them
- Kaggle and UC Irvine Machine Learning Repository are great sources too

More Diverse Models and Approaches

- Many courses in undergraduate level choose to show fewer modeling techniques to the students
- Proving results and going deep on the math behind them
- We prefer to present models focusing on their strengths and limitations
- The students are only required to intuitively know how the algorithms work

More Diverse Models and Approaches

- k-means
- Hierarchical clustering
- Principal components analysis
- Data acquisition
- Cross validation
- K nearest neighbor
- Support vector machine
- Classification and regression trees
- Random forests
- Model ensemble

Ability to Communicate

- The students are evaluated through midterms and a final project
- The final project has two parts: written report and live presentation
- While the default is to present slides, some students have built dashboards to present their results

Case Study: Web Scraping

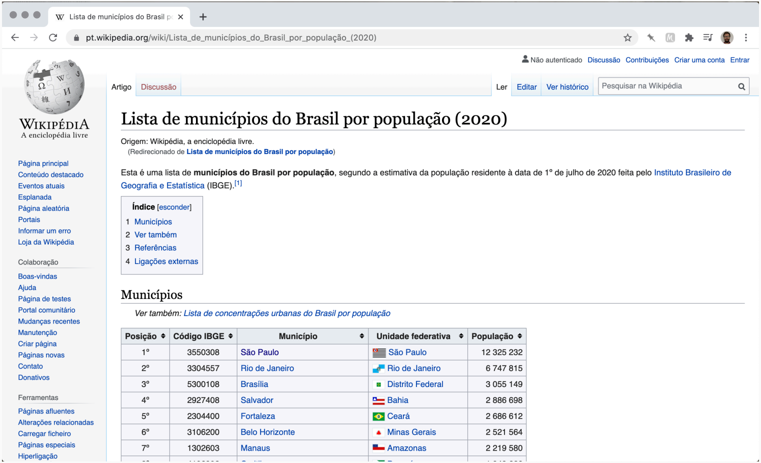
Case Study: Web Scraping

- This is the project the students have to complete on the web scraping module
- This is the fifth module of the course
- Dogucu and Çetinkaya-Rundel (2020) is a very good resource on this topic

Case Study: Web Scraping

- Extract data from websites
- Collect and organize data automatically
- Only open data can be reached this way

Case Study: Web Scraping



Lista de municípios do Brasil por população (2020)

Origem: Wikipédia, a enciclopédia livre.

(Redirecionado de **Lista de municípios do Brasil por população**)

Esta é uma lista de **municípios do Brasil por população**, segundo a estimativa da população residente à data de 1° de julho de 2020 feita pelo Instituto Brasileiro de Geografia e Estatística (IBGE).[1]

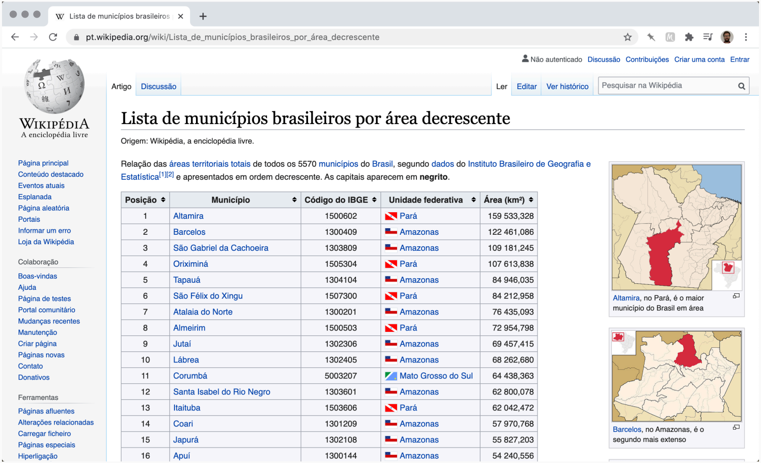
Índice [esconder]

Municípios

Ver também: Lista de concentrações urbanas do Brasil por população

Posição	Código IBGE	Município	Unidade federativa	População
1°	3550308	São Paulo	São Paulo	12 325 232
2°	3304557	Rio de Janeiro	Rio de Janeiro	6 747 815
3°	5300108	Brasília	Distrito Federal	3 055 149
4°	2927408	Salvador	Bahia	2 886 698
5°	2304400	Fortaleza	Ceará	2 686 612
6°	3106200	Belo Horizonte	Minas Gerais	2 521 564
7°	1302603	Manaus	Amazonas	2 219 580

Case Study: Web Scraping



Lista de municípios brasileiros por área decrescente

Origem: Wikipédia, a enciclopédia livre.

Relação das áreas territoriais totais de todos os 5570 municípios do Brasil, segundo dados do Instituto Brasileiro de Geografia e Estatística[1][2] e apresentados em ordem decrescente. As capitais aparecem em **negrito**.

Posição	Município	Código do IBGE	Unidade federativa	Área (km²)
1	Altamira	1500602	Pará	159 533,328
2	Barcelos	1300409	Amazonas	122 461,086
3	São Gabriel da Cachoeira	1303809	Amazonas	109 181,245
4	Oriximiná	1505304	Pará	107 613,838
5	Tapauá	1304104	Amazonas	84 946,035
6	São Félix do Xingu	1507300	Pará	84 212,958
7	Atalaia do Norte	1300201	Amazonas	76 435,093
8	Almeirim	1500503	Pará	72 954,798
9	Jutaí	1302306	Amazonas	69 457,415
10	Lábrea	1302405	Amazonas	68 262,680
11	Corumbá	5003207	Mato Grosso do Sul	64 438,363
12	Santa Isabel do Rio Negro	1303601	Amazonas	62 800,078
13	Itaituba	1503606	Pará	62 042,472
14	Coari	1301209	Amazonas	57 970,768
15	Japurá	1302108	Amazonas	55 827,203
16	Apuí	1300144	Amazonas	54 240,556

Altamira, no Pará, é o maior município do Brasil em área

Barcelos, no Amazonas, é o segundo mais extenso

Case Study: Web Scraping

```
> library(rvest)
> library(dplyr)
> library(ggplot2)
> theme_set(theme_bw())
> library(stringr)
> library(scales)
```

Case Study: Web Scraping

```
> url <- "https://pt.wikipedia.org/wiki/Lista_de_munic%C
>
> population <- url %>%
+     read_html()
>
> population <- population %>%
+     html_table(fill=TRUE)
>
> population <- population[[1]]
>
> names(population) <- c("Position", "IBGE.Code",
+   "City", "State", "Population")
```

Case Study: Web Scraping

```
> head(population)

##   Position IBGE.Code           City            State
## 1       1º   3550308      São Paulo        São Paulo
## 2       2º   3304557 Rio de Janeiro   Rio de Janeiro
## 3       3º   5300108       Brasília Distrito Federal
## 4       4º   2927408       Salvador            Bahia
## 5       5º   2304400      Fortaleza            Ceará
## 6       6º   3106200 Belo Horizonte     Minas Gerais
##   Population
## 1 12 325 232
## 2  6 747 815
## 3  3 055 149
## 4  2 886 698
## 5  2 686 612
## 6  2 521 564
```

Case Study: Web Scraping

```
> head(area)

##   Position                     City IBGE.Code    Stat
## 1        1                 Altamira   1500602     Par
## 2        2                 Barcelos   1300409 Amazona
## 3        3 São Gabriel da Cachoeira   1303809 Amazona
## 4        4                Oriximiná   1505304     Par
## 5        5                   Tapauá   1304104 Amazona
## 6        6       São Félix do Xingu   1507300     Par
##          Area
## 1 159 533,328
## 2 122 461,086
## 3 109 181,245
## 4 107 613,838
## 5  84 946,035
## 6  84 212,958
```

Case Study: Web Scraping

```
> brazil <- left_join(population, area,
+   by = "IBGE.Code")
>
> brazil <- brazil %>%
+     select(City.x, State.x, Area, Population)
>
> names(brazil) <- c("City", "State", "Area",
+   "Population")
```

Case Study: Web Scraping

```
> head(brazil)
```

```
##             City            State      Area Population
## 1      São Paulo        São Paulo 1 521,110 12 325 232
## 2 Rio de Janeiro   Rio de Janeiro 1 200,329  6 747 815
## 3       Brasília Distrito Federal 5 760,783  3 055 149
## 4       Salvador            Bahia   693,453  2 886 698
## 5      Fortaleza            Ceará   312,353  2 686 612
## 6 Belo Horizonte     Minas Gerais   331,354  2 521 564
```

Case Study: Web Scraping

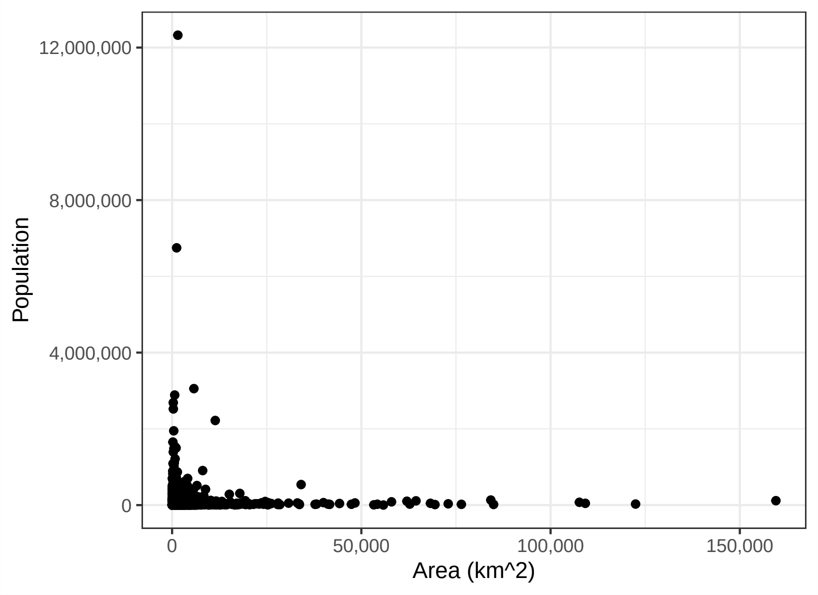
```
> brazil <- brazil %>%
+     mutate(Area = str_replace(Area,
+       "[[:space:]]", "")) %>%
+     mutate(Area = str_replace(Area, ",", ".")) %>%
+     mutate(Area = as.numeric(Area)) %>%
+     mutate(Population = str_replace_all(Population,
+       "[[:space:]]", "")) %>%
+     mutate(Population = as.numeric(Population))
```

Case Study: Web Scraping

```
> head(brazil)
```

```
##              City            State     Area Population
## 1      São Paulo        São Paulo 1521.110   12325232
## 2 Rio de Janeiro   Rio de Janeiro 1200.329    6747815
## 3       Brasília Distrito Federal 5760.783    3055149
## 4       Salvador            Bahia  693.453    2886698
## 5      Fortaleza            Ceará  312.353    2686612
## 6 Belo Horizonte     Minas Gerais  331.354    2521564
```

Case Study: Web Scraping



Final Remarks

- Student evaluations indicate students are satisfied with this course contents
- 2019 was the first year the course was offered for the students enrolled in the Actuarial Science Department
- Our future plans for this course include expanding it from a one-semester course to a two-semester course
- And everything is free!

References

- Dogucu, Mine and Çetinkaya-Rundel, Mine (2020) “Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities.” *Journal of Statistics Education* **0** (0): 1-11.
- Hicks, Stephanie C. and Rafael A. Irizarry (2016) “A Guide to Teaching Data Science.” *The American Statistician* **72** (4): 382-391.

Modernizing the Curricula of Statistics courses through Statistical Learning

New Approaches to Statistical Learning in Developing Countries
3rd LISA 2020 Symposium

Marcus Nunes

29 October 2020

Federal University of Rio Grande do Norte