Marcio Oliveira Silva Data Analytics Bootcamp May 13, 2020

ETL Project Report Work Description

Objective:

It is to make available information about monthly internet access of people aged ten years old or over in Brazil. More specifically, we want to extract, transform, and load two datasets covering both household income average and median per capita and their level of education in all 26 Brazilian States and the Federal District in 2015.

Work:

We extracted the data from the Brazilian Institute of Geography and Statistics (IBGE) website. IBGE is the agency responsible for the official collection of statistical, geographic, cartographic, geodetic, and environmental information in Brazil.

We used Jupyter Notebook as our primary environment to do the data cleaning and transformation. We first imported Pandas and NumPy to Jupyter, and then pulled out the data from the CSV files. The data manipulation occurred over consecutive steps until the two datasets were ready to be merged into a single dataset. We finally imported SQLAchemy to Jupyter notebook.

Outcome:

| Admin File > Object > Tools > Help > | Ξ. | | | | | | | | | | | | | | |
|---|--|---|-----------------------------|-------------------|----------------------|----------------------------|-------------------|----------------------|------------------|---------------------|------------------|-------------------------|-------------------------------|-------------------------------|--|
| | Dash | Databased Properties 50, Statistics Operationics Operat | | | | | | | | | | | | | |
| Servers (1) | ₽ | B ~ B 0 | Q ~ @ ~ E B B | ✓ T ✓ No limit | : > - 6 0 | 1 v 3 3 #v ± | | | | | | | | | |
| (a) localhost | ASS | ETI Project/marc | :iooliver@localhost | | | | | | | | | | | | |
| ▼ ■ Databases (9) | _ | | | | | | | | | | | | | | |
| SETL_Project DG Casts | Quer | y Editor Scratch | Pad Query History | | | | | | | | | | | | |
| > @ Casts > % Catalogs | 1 | SELECT * FROM | internet_access_income_ | vs_edu_level | | | | | | | | | | | |
| Catalogs Event Triggers | | | | | | | | | | | | | | | |
| > 10 Extensions | | | | | | | | | | | | | | | |
| > # Foreign Data Wrappers | Data Output Explain Messages Notifications | | | | | | | | | | | | | | |
| > = Languages | | Brazilian, States | , Income_Average_Per_Capita | . With Access Ave | . Without Access Ave | a Income_Median_Per_capita | _ With Access Med | . Without Access Med | a Total_Levels | Primary, incomplete | Primary_School | . Secondary, Incomplete | Secondary | Higher Incom | |
| ✓ 1 Schemas (1) | | text | text | 1est | test | test | text | test | double precision | double precision | double precision | double precision | Secondary double precision | Higher_incom double precis | |
| ✓ | - 1 | Rondônia | 947 | 1 190 | 728 | 692 | 800 | 552 | | 47.4 | 27.3 | 48.7 | 66.4 | 64.3 | |
| > 🚉 Collations | 2 | Acre | 754 | 1 074 | 503 | 431 | 642 | 335 | | 43.8 | 18.9 | 40.5 | 62.7 | 68.2 | |
| > & Domains | 3 | Amazonas | 734 | 970 | 501 | 451 | 593 | 352 | | 49.6 | 26.5 | 43 | 61.8 | 66.7 | |
| > [] FTS Configurations | 4 | Roraima | 974 | 1 163 | 643 | 581 | 684 | 453 | | 63.7 | 37.4 | 58.9 | 73.9 | 81.5 | |
| > (I), FTS Dictionaries | 5 | Pará | 675 | 876 | 521 | 454 | 556 | 394 | | 43.3 | 21.5 | 49.6 | 65.7 | 67.9 | |
| > An FTS Parsers | 6 | Arrepá | 822 | 1 004 | 609 | 496 | 636 | 403 | | 53.7 | 29.3 | 50.6 | 67.5 | 66.6 | |
| > (3) FTS Templates | | Tocertins | 919 | 1 260 | 660 | 575 | 753 | 500 | | 43.2 | 20.7 | 41.6 | | 57.7 | |
| > Iff Foreign Tables | | Mananhão | 616 | 947 | 439 | 393 | 524 | 312 | | 34.3 | 15.1 | 37 | | 59.2 | |
| > (i) Functions | | Plaul | 679 | 963 | 499 | 451 | 567 | 393 | | 38.8 | 17.3 | 48.1 | 64.9 | 67 | |
| Materialized Views (i) Procedures | | Ceaná | 660 | 874 | 496 | 449 | 556 | 394 | | 43.4 | 19.9 | 46.6 | | 68.6 | |
| > La Seguences | | | | | | 523 | 595 | 420 | | | | | | | |
| ▼ Im Tables (1) | | Rio Grande do Norte | | 1 004 | 564 | | | | | 52.8 | | 54.4 | | 79.2 | |
| ➤ ⊞internet_access_income_vs_edu_level | | Paraibe | 792 | 1 026 | 545 | 497 | 577 | 419 | | 51 | 27.5 | 56.6 | 75.1 | 78 | |
| > P Columns | | Pernambuoo | 745 | 997 | 531 | 468 | 546 | 399 | | 45.8 | 23 | 47 | | 69.6 | |
| > > Constraints | | Alagoas | 609 | 800 | 474 | 400 | 499 | 375 | | 41.3 | 21.6 | 46.4 | | 70.6 | |
| > 3 Indexes | 15 | Sergipe | 701 | 867 | 527 | 483 | 568 | 394 | | 51.2 | 30.4 | 56.4 | 74.1 | 77.8 | |
| > Thules | 16 | Bahia | 799 | 962 | 531 | 486 | 582 | 394 | | 48.3 | 25.1 | 54.2 | 71.3 | 74.6 | |
| > ‡+Triggers | 17 | Minas Gerals | 1 116 | 1 375 | 785 | 765 | 845 | 636 | | 56.2 | 29.7 | 60.9 | 79.2 | 79 | |
| > (ii) Trigger Functions | 18 | Espírito Santo | 1111 | 1 329 | 908 | 765 | 836 | 642 | | 58.5 | 32.6 | 56 | 78.7 | 78 | |
| > Types | 19 | Rio de Janeiro | 1380 | 1 611 | 932 | 789 | 864 | 763 | | 66.3 | 41.9 | 57 | 83.4 | 78.6 | |
| > News | 20 | São Paulo | 1.408 | 1 589 | 1 000 | 926 | 1 000 | 788 | | 69.5 | 40.8 | 65.1 | 84.8 | 83.5 | |
| ✓ HappinessData | 21 | Paraná | 1 327 | 1 612 | 893 | 874 | 1 050 | 767 | | 60.5 | 33 | 58.7 | 75.9 | 78.2 | |
| > @ Casts > % Catalogs | 22 | Santa Catarina | 1.454 | 1 663 | 1.111 | 1 059 | 1 198 | 843 | | 62.3 | 35 | 57.8 | 81.2 | 79 | |
| > Catalogs > Care Triggers | 23 | Rio Grande do Sul | 1367 | 1 582 | 1 033 | 943 | 1 033 | 798 | | 61.1 | 34.8 | 62.3 | 82.8 | 82.8 | |
| > 10 Extensions | | Mato Grosso do Sul | | 1 494 | 921 | 850 | 960 | 784 | | 63.7 | 40.7 | 65 | | 81.3 | |
| → agreements → ag | | Mato Grosso | 1120 | 1 359 | 848 | 788 | 903 | 742 | | 55.2 | 31.4 | 50.5 | 69 | 72 | |
| > = Languages | | Golds | 1104 | 1 250 | 869 | 788 | 864 | 728 | | 61.8 | | 61.3 | | 78.9 | |
| ✓ W Schemas (1) | 20 | OWN | 1.104 | 1 200 | 007 | 100 | 004 | 7.60 | | 41.4 | 94.7 | 91.0 | 90.0 | 140.7 | |

1

The final step allowed us to load the data on Jupyter Notebook to a schema and make it available to scripting languages like Python, JavaScript, and R.