Marcio Oliveira Silva

Data Analytics Bootcamp

May 13, 2020

ETL Project Report

Work Description

Objective:

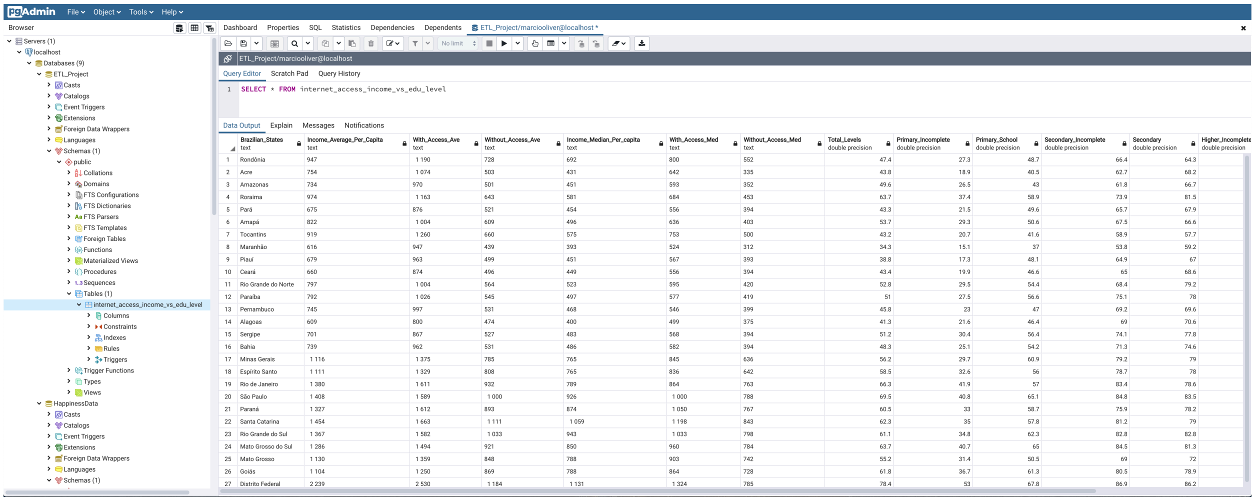
It is to make available information about monthly internet access of people aged ten years old or over in Brazil. More specifically, we want to extract, transform, and load two datasets covering both household income average and median per capita and their level of education in all 26 Brazilian States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



```
SELECT * FROM internet_access_income_vs_edu_level
```

	Brazilian_States	Income_Average_Per_Capita	With_Access_Ave	Without_Access_Ave	Income_Median_Per_capita	With_Access_Med	Without_Access_Med	Total_Levels	Primary_Incomplete	Primary_School	Secondary_Incomplete	Secondary	Higher_Incomplete
	text	text	text	text	text	text	text	double precision	double precision	double precision	double precision	double precision	double precision
1	Rondônia	947	1 190	728	692	800	552	47.4	27.3	48.7	66.4	64.3	
2	Acre	754	1 074	503	431	642	335	43.8	18.9	40.5	62.7	68.2	
3	Amazonas	734	970	501	451	593	352	49.6	26.5	43	61.8	66.7	
4	Roraima	974	1 163	643	581	684	453	63.7	37.4	58.9	73.9	81.5	
5	Pará	675	876	521	454	556	394	43.3	21.5	49.6	65.7	67.9	
6	Amapá	822	1 004	609	496	636	403	53.7	29.3	50.6	67.5	66.6	
7	Tocantins	919	1 260	660	575	753	500	43.2	20.7	41.6	58.9	57.7	
8	Maranhão	616	947	439	393	524	312	34.3	15.1	37	53.8	59.2	
9	Piauí	679	963	499	451	567	393	38.8	17.3	48.1	64.9	67	
10	Ceará	660	874	496	449	556	394	43.4	19.9	46.6	65	68.6	
11	Rio Grande do Norte	797	1 004	564	523	595	420	52.8	29.5	54.4	68.4	79.2	
12	Paraíba	792	1 026	545	497	577	419	51	27.5	56.6	75.1	78	
13	Pernambuco	745	997	531	468	546	399	45.8	23	47	69.2	69.6	
14	Alagoas	609	800	474	400	499	375	41.3	21.6	46.4	69	70.6	
15	Sergipe	701	867	527	483	568	394	51.2	30.4	56.4	74.1	77.8	
16	Bahia	739	962	531	486	582	394	48.3	25.1	54.2	71.3	74.6	
17	Minas Gerais	1 116	1 375	785	765	845	636	56.2	29.7	60.9	79.2	79	
18	Espírito Santo	1 111	1 329	808	765	836	642	58.5	32.6	56	78.7	78	
19	Rio de Janeiro	1 380	1 611	932	789	864	763	66.3	41.9	57	83.4	78.6	
20	São Paulo	1 408	1 589	1 000	926	1 000	788	69.5	40.8	65.1	84.8	83.5	
21	Paraná	1 327	1 612	893	874	1 050	767	60.5	33	58.7	75.9	78.2	
22	Santa Catarina	1 454	1 663	1 111	1 059	1 198	843	62.3	35	57.8	81.2	79	
23	Rio Grande do Sul	1 367	1 582	1 033	943	1 033	798	61.1	34.8	62.3	82.8	82.8	
24	Mato Grosso do Sul	1 286	1 494	921	850	960	784	63.7	40.7	65	84.5	81.3	
25	Mato Grosso	1 130	1 359	848	788	903	742	55.2	31.4	50.5	69	72	
26	Goiás	1 104	1 250	869	788	864	728	61.8	36.7	61.3	80.5	78.9	
27	Distrito Federal	2 239	2 530	1 184	1 131	1 324	785	78.4	53	67.8	86.9	86.2	

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.