Connecting to Data Part One and Part Two

Part Two of a series showing free and online resource to become a data scientist. The objective is to give you a practical online portfolio. The portfolio demonstrates your ability to work as a data scientist, even if you do not have experience.

The videos in this part focus on notebooks and connecting to data.

Helpful Links

Create your own notebook here: https://colab.research.google.com/

You can also fork the YAADS repo here: https://github.com/thedanindanger/yaads-examples I will explain github in an upcoming tutorial.

You may use BigQuery sandbox for free tier test data environment: https://cloud.google.com/bigquery/docs/sandbox

Project Challenge Summary

Because television media exposure is measured in GRP (a percent of the population), and digital media exposure is measured in impressions, we cannot compare them directly. We need to know the population of a region to evaluate either impressions as a percent of population or GRP as impressions.

Definitions and Abbreviations

DMA: “Designated Marketing Area”, geographic regions in which advertising is sold/targeted. For example, a television commercial may be sold in the New York DMA with Gross Rating Point of 1.5.

GRP: “Gross Rating Point(s)” a measure indicating 1% of the DMA population/household per each point within a given region – the region is typically a DMA. Note, GRP can exceed 100 as a household may be targeted multiple times.

DataFrames: The primary data object in Python Pandas and many other data focused code packages. A DataFrame is very similar to a SQL table or SQL Cube, in that the data is organized in rows and columns. Each column represents a type of data (e.g. ZIP code, Population, DMA Name) and each row represent a record of data containing values from each column (e.g. ‘12345’, 35,088, “Some DMA Name”).

“Cell Magics” Instruction for notebooks located at the top of a cell, usually indicated with “%%”. Can indicate how to process the contents of a cell, to run command line scripts, how to format cell output, etc.

DASK: An extension of Pandas for large data operations: https://www.dask.org/ will have follow up video to explain.

Zero Padding: Adding zeros to the beginning of a number to make all values the same length. Zero padding is important because most applications will treat, for example, ‘53’ and ‘0053’ as different numbers; however, applications will consider ‘00053’ and ‘00053’ as the same. We have to zero pad often because some applications will strip leading zeros (e.g. Excel(r)). We need numbers to match to combine data sets.

Zero Padding Examples

BigQuery SQL

Most SQL based data will have the ZIP already defined as a string/text value; however, some databases will have ZIP as varchar(5) which may be interpreted as numeric. Regardless, it will not cause an error to 'cast' the ZIP as a string if it is already a string, so we can use the following SQL pattern for almost any zero padding scenario.

```
right(concat('00000' ,cast(zipcode as string)),5) as zip_5
```

Python Pandas

When ZIP are correctly interpreted as string/text:

```
zip_pop_df['zip_5'] = zip_pop_df['zip_5'].str.zfill(5)
```

When ZIP are imported as numbers:

```
target_zips['Zip'] = target_zips['Zip'].astype(str).str.zfill(5)
```

Video Links

Video One:

Video Two: