Big Data and Analytics - Weatherlink
Personal Report
Matt Winchester
Milestone: Early Pipeline Exploration

I have been working on an early implementation of our pipeline at its most basic level. I wanted try out a few things to make sure we had a good approach:

I wanted to try and flesh out

- How to get data out of GBQ
- Google Looker
- Tableau
- Tooltips
- A database schema
- Thoughts about size of data and what is left to do

These last few weeks I have been wanting to get started on a basic version of what we are trying to do as a proof of concept and a starting point for iterations. My goal was:

1. Get useful census data out of a US Census dataset from Google Big Query
2. Store this data as a table and link it to a visualization

There ended up being two approaches I took to get a feel for this.
First, I created a Google Cloud project to host the extracted data. The Census Data is split by year and by granularity. I wanted to focus on County level data, but I did Zip Code data as well. This is because I was unsure how to get an actual location out of the government's format. The answer came after some searching, and it turns out this is a format that Google Looker (Data Studio) and Tableau could already interpret, a GeoLocation FIPS code. The 'GEO ID' column represents this FIPS code in the Census datasets. This code can represent various things, for the zip code based dataset, this was the postal code of the area the data was referring to. So I started with that, and then did the county later when I realized the country FIPS code was also something that Tableau and Looker could handle natively.

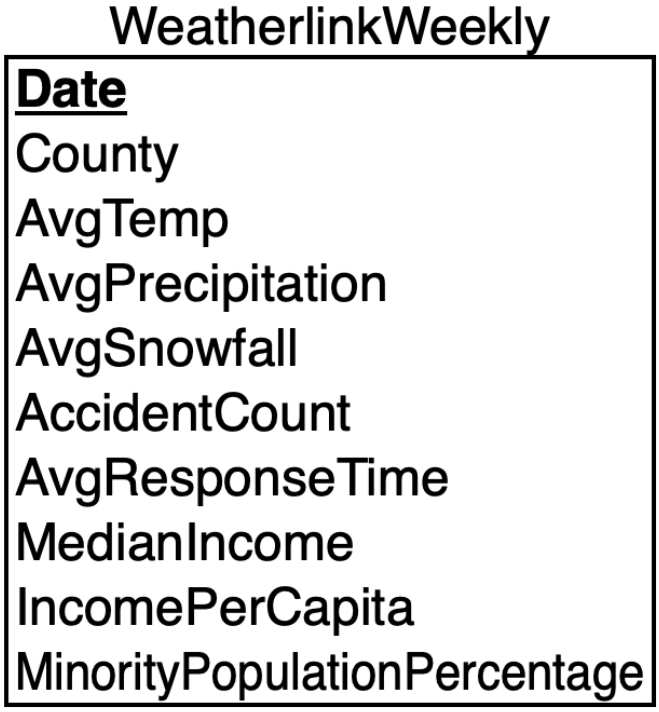
I grabbed a handful of columns from the 2021 datasets for the US Census, and created two visualizations. Initially, Google was showing me data all over the world, which I knew was wrong because this was United States specific data. I then found out how to specify in the data connection what kind of Geolocation FIPS code the ID was, and then it was showing correct data.

However, this level of data seemed to completely overwhelm Google Looker. As you can see, it is only filing in some of the counties. There are a lot of blank spots. I'm not sure if it is just choking on the amount of data, or if it is unable to handle nulls in the data. It would be nice to use Looker cause it implements very well with GBQ tables, and it is free.

Tableau is not free, but looks better in my opinion, has more powerful tooltips (which will be important for our visualizations). For a test of the free version of Tableau, I had to export the GBQ table into google sheets. This works for a test, but it won't work for the real system. The sheets export is limited to 50K rows, and our final dataset will have at least 20 million rows. Thankfully, the paid version of Tableau DOES have native google big query integration.

These two attempts helped me see our path forward, there is still a lot to do!

Something I feel was missing from our system design was a formal database schema. At a minimum, I think these columns are what we will need:



The SQL to Create such a table would look like this:

```
CREATE TABLE WeatherlinkWeekly (
Date DATE NOT NULL,
County VARCHAR(10) NOT NULL,
AvgTemp FLOAT NOT NULL,
AvgPrecipitation FLOAT NOT NULL,
AccidentCount FLOAT NOT NULL,
AvgResponseTime FLOAT NOT NULL,
MedianIncome FLOAT NOT NULL,
IncomePerCapita FLOAT NOT NULL,
MinorityPopulationPercentage INT NOT NULL,
AvgSnowfall INT NOT NULL,
PRIMARY KEY (Date) );
```

Here are some screenshots of the two attempts at visualizations:

Google Looker: https://lookerstudio.google.com/reporting/7f3efe9c-c5fe-429b-8557-2435dc6cefac

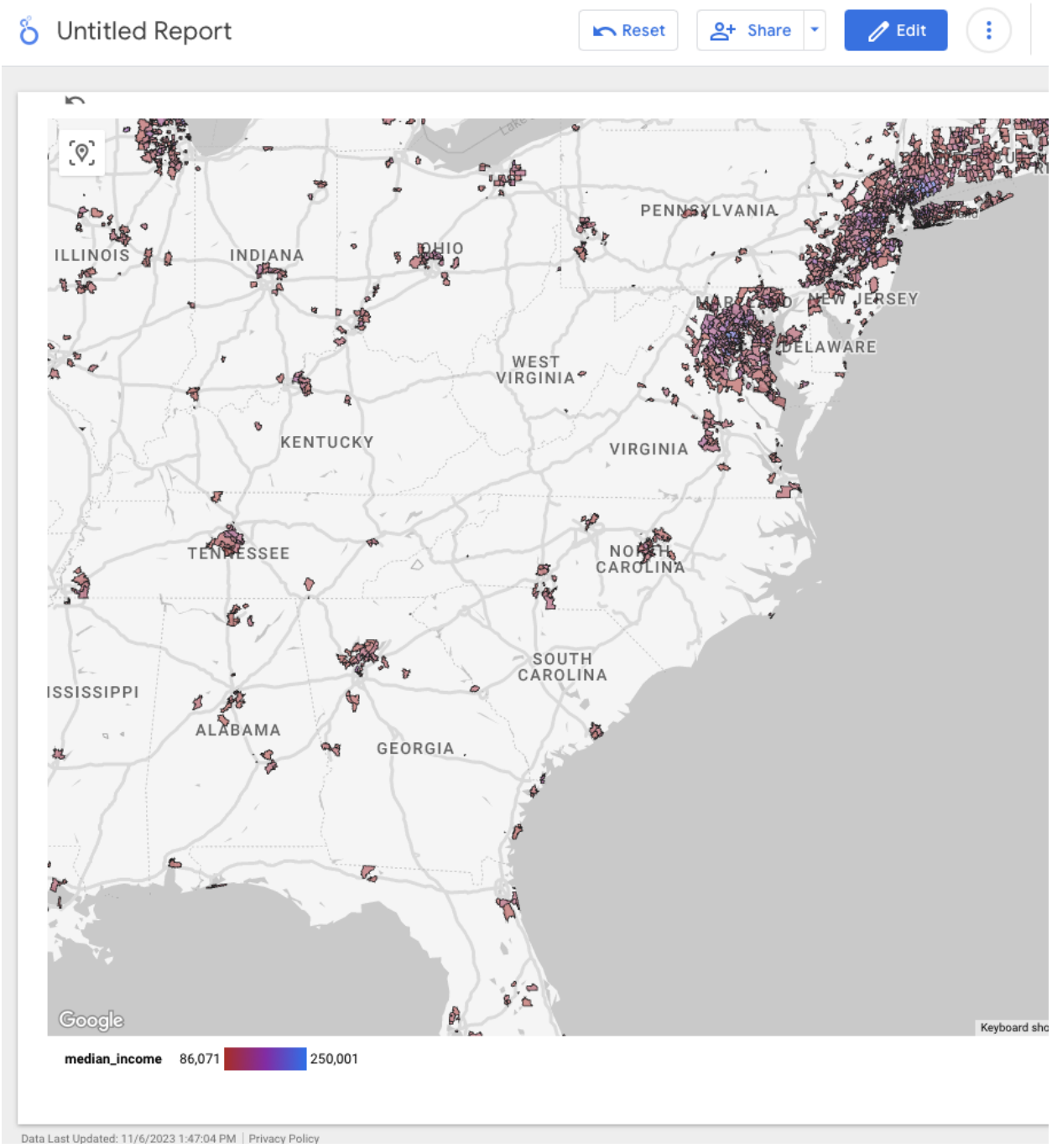


Tableau: https://public.tableau.com/app/profile/matthew.winchester/viz/CensusDataTestViz/Sheet1

Census Data Test Viz by Matthew Winchester

Sheet 1