Covid-19-Mask-Usage-Visualization-with-Google-BigQuery-and-Tableau-

After learning a considerable amount of SQL, I began looking for large, realworld databases where I could practice writing more complex queries.

I became aquainted with Google's Data Wherehouse BigQuery and it's vast array of public datasets. Given my background in Sociology, I'm no stranger to the American Community Survey ("ACS") the thorough and expansive demographic survey conducted by The United States Census Bureau. The ACS is comprised of numerous variables including: age, race/ethnicity, gender, employment, housing, and many others. It's an excellent source of information for community level demographic charectatistics. When evaluated alongside other variables of interest it becomes a treasure trove of explanatory power.

Enter BigQuery.

The ACS collects data at many different levels (state, county, zipcode, geographic coorrdinates). Since it's county level data uses the county fips code as it's unique primary key, it's easy to combine the ACS table with other tables that have county level data identified with the county fips code.

The New York Times conducted a survey on county level mask data in July:
https://www.nytimes.com/interactive/2020/07/17/upshot/coronavirus-face-mask-map.html

The New York Times also regularly collects county level COVID-19 data: https://www.nytimes.com/interactive/2021/us/covid-cases.html

Both of these datasets are housed in google's data warehouse and both use the county fips code as their unique primary key.

Using SQL, I wrote this query to retrieve a dataset combining these three data sources.

```
SELECT
  county,
  state_name,
  employed_pop/ total_pop AS employed_pop_pct,
  unemployed_pop/ total_pop AS unemployed_pop_pct,
  pop_in_labor_force/ total_pop AS pop_in_labor_force_pct,
  not_in_labor_force/ total_pop AS not_in_labor_force_pct,
  workers_16_and_over/ total_pop AS workers_16_and_over_pct,
  armed_forces/ total_pop AS armed_forces_pct,
  civilian_labor_force/ total_pop AS civilian_labor_force_pct,
  employed_agriculture_forestry_fishing_hunting_mining/ total_pop AS employed_agriculture_forestry_fishing_hunting_mining_pct,
  employed_arts_entertainment_recreation_accommodation_food/ total_pop AS employed_arts_entertainment_recreation_accommodation_food_pct,
  employed_construction/ total_pop AS employed_construction_pct,
  employed_education_health_social/ total_pop AS employed_education_health_social_pct,
  employed_finance_insurance_real_estate/ total_pop AS employed_finance_insurance_real_estate_pct,
  employed_information/ total_pop AS employed_information_pct,
  employed_manufacturing/ total_pop AS employed_manufacturing_pct,
  employed_other_services_not_public_admin/ total_pop AS employed_other_services_not_public_admin_pct,
  employed_public_administration/ total_pop AS employed_public_administration_pct,
  employed_retail_trade/ total_pop AS employed_retail_trade_pct,
  employed_science_management_admin_waste/ total_pop AS employed_science_management_admin_waste_pct,
  employed_transportation_warehousing_utilities/ total_pop AS employed_transportation_warehousing_utilities_pct,
  employed_wholesale_trade/ total_pop AS employed_wholesale_trade_pct,
  occupation_management_arts/ total_pop AS occupation_management_arts_pct,
  occupation_natural_resources_construction_maintenance/ total_pop AS occupation_natural_resources_construction_maintenance_pct,
  occupation_production_transportation_material/ total_pop AS occupation_production_transportation_material_pct,
  occupation_sales_office/ total_pop AS occupation_sales_office_pct,
  occupation_services/ total_pop AS occupation_services_pct,
  management_business_sci_arts_employed/ total_pop AS management_business_sci_arts_employed_pct,
  sales_office_employed/ total_pop AS sales_office_employed_pct,
  poverty/ total_pop AS poverty_pct,
  gini_index,
  median_income,
  median_rent,
  percent_income_spent_on_rent,
  million_dollar_housing_units,
  black_pop/ total_pop AS black_pop_pct ,
  hispanic_pop/ total_pop AS hispanic_pop_pct,
  asian_pop/ total_pop AS asian_pop_pct,
  white_pop/ total_pop AS white_pop_pct,
  other_race_pop/ total_pop AS other_race_pop_pct,
  amerindian_pop/ total_pop AS amerindian_pop_pct,
  total_cases / total_pop AS covid_cases_per_capita,
  total_deaths/ total_pop AS deaths_per_capita,
  total_pop,
  mask_score,
  mask_usage
--Selecting Variables of Interest
FROM
  `bigquery-public-data.census_bureau_acs.county_2018_5yr` acs
JOIN
--Joing the ACS data with a subquery that retrieves NYT covid-19 and mask use data
  (SELECT *,
   CASE
   WHEN mask_score < 0.4010546 THEN 'low'
   WHEN mask_score > 0.4010546 AND mask_score < 0.5788573 THEN 'med'
   WHEN mask_score > 0.5788573 THEN 'high'
   END AS mask_usage
--Catagorizing mask usage by splitting the data into thirds by calculating the probability distribution function in R
   FROM
      (SELECT covid.county, covid.state_name, covid.total_cases, covid.total_deaths, covid.county_fips_code,
      (never *-1)+ (rarely * -.5)+ (sometimes*0)+ (frequently*.5) +(always*1) AS mask_score
--Converting the five question survey data into a single statistic
        FROM
          (SELECT county, state_name, county_fips_code, SUM(confirmed_cases) as total_cases, SUM(deaths)as total_deaths
           From `bigquery-public-data.covid19_nyt.us_counties`
           GROUP BY county, state_name, county_fips_code) AS covid
           JOIN `bigquery-public-data.covid19_nyt.mask_use_by_county` AS mask
           ON covid.county_fips_code = mask.county_fips_code) AS dat) as dat1
--Joining the NYT Covid-19 data with the mask use data

ON
  dat1.county_fips_code = acs.geo_id
```

As you can see, I converted the mask usage survey results into a single statistic which I call "mask score".

I also wrote a second query to retrieve summary statistics

```
SELECT AVG(mask_score), STDDEV_POP(mask_score)
FROM
(SELECT covid.county, covid.state_name, covid.total_cases, covid.total_deaths, covid.county_fips_code,
  (never *-1)+ (rarely * -.5)+ (sometimes*0)+ (frequently*.5) +(always*1) AS mask_score

FROM    (SELECT county, state_name, county_fips_code, SUM(confirmed_cases) as total_cases, SUM(deaths)as total_deaths
         From `bigquery-public-data.covid19_nyt.us_counties`
         GROUP BY county, state_name, county_fips_code) AS covid

         JOIN `bigquery-public-data.covid19_nyt.mask_use_by_county` AS mask
         ON covid.county_fips_code = mask.county_fips_code

 ORDER BY mask_score) as Dat
```

Query results   SAVE RESULTS   EXPLORE DATA

Query complete (0.7 sec elapsed, 32.1 MB processed)

Job information   Results   JSON   Execution details

Row	f0_	f1_
1	0.4899559527609327	0.20639825812739163

Then I used the qnorm function in r to determine the 33.3 and 66.6 percentile markers.

```
> med_threshold <- qnorm(2/3, mean = 0.48995595276093235, sd = 0.2063
9825812739154)
>
> med_threshold
[1] 0.5788573
> low_threshold <- qnorm(1/3, mean = 0.48995595276093235, sd = 0.2063
9825812739154)
>
> low_threshold
[1] 0.4010546
```

I then took the percentintile markers and used a Case statement within my query to classify each county as "high", "med", or "low".

```
    CASE
    WHEN mask_score < 0.4010546 THEN 'low'
    WHEN mask_score > 0.4010546 AND mask_score < 0.5788573 THEN 'med'
    WHEN mask_score > 0.5788573 THEN 'high'
    END AS mask_usage
```

I ran my query and returned the following table:

Query results   SAVE RESULTS   EXPLORE DATA

Query complete (2.0 sec elapsed, 52.3 MB processed)

Job information   Results   JSON   Execution details

Row	county	state_name	employed_pop_pct	unemployed_pop_pct	pop_in_labor_force_pct	not_in_labor_force_pct	workers_16_and_over_pct	armed_forces_pct	civilian_labor_force_pct	employed_agric
1	Apache	Arizona	0.25677413942563126	0.035024188361623	0.2918123094991751	0.4638293112608708	0.25392187019378654	1.3981711920807584E-5	0.2917983277872543	
2	Cochise	Arizona	0.34128398229317625	0.02646520799182762	0.3970731475542252	0.40572858511708204	0.36537349836473204	0.029323957269221327	0.36774919028500386	
3	Coconino	Arizona	0.47290984688019283	0.03844041735310269	0.5122060805751086	0.30162533786915996	0.46366703038861196	8.55816341813047E-4	0.5113502642332955	
4	Gila	Arizona	0.3401123595505618	0.03842696629213483	0.3785393258426966	0.43735955056179776	0.32945692883895134	0.0	0.3785393258426966	
5	Maricopa	Arizona	0.4711511495416103	0.027517958171688044	0.4998212234241744	0.2829782837589767	0.46459318749584205	0.0011521157108760804	0.4986691077132983	
6	Mohave	Arizona	0.34573724668064293	0.033746797111576986	0.37951316095970183	0.46376853792996353	0.3382686932215234	2.9117167481947357E-5	0.3794840437922199	
7	Navajo	Arizona	0.30607607745733867	0.05327261855480429	0.359348696012143	0.396301918035049	0.3003449703325514	0.0	0.359348696012143	
8	Pima	Arizona	0.4320746242603376	0.03551850406287204	0.47321328754307546	0.33741254969491685	0.4277283416460565	0.0056201592198658066	0.46759312832320965	
9	Pinal	Arizona	0.358080725053071	0.03046071080551128	0.38899173498585965	0.4027627876613274	0.35108560210234896	4.502991272774057E-4	0.38854143585858225	
10	Santa Cruz	Arizona	0.36993388287824147	0.03359522582861068	0.40402284046024384	0.35660312553666496	0.3657049630774515	4.937317533917225E-4	0.40352910870685216	
11	Yavapai	Arizona	0.3745554096463309	0.027047118787420152	0.40189187384540054	0.45062209263504643	0.36670747178882235	2.8934541164949145E-4	0.40160252843375105	
12	Yuma	Arizona	0.36526182582796435	0.03995111365593829	0.4225108141789644	0.34915242819818215	0.37350417891632065	0.017297874695061805	0.40521293948390263	
13	Fairfield	Connecticut	0.5035357728295078	0.03763972603319963	0.5414105816923422	0.2572208550237836	0.4937618335613566	2.35082829634838E-4	0.5411754988627074	
14	Hartford	Connecticut	0.5005901221597577	0.035050797447274594	0.536132688073497	0.2781487152548814	0.49084304762330533	4.917684664647435E-4	0.5356409196070323	
15	Litchfield	Connecticut	0.532095655927138	0.029388464249225543	0.56172451661194	0.2756746124973365	0.5211248367762839	2.4039643557648705E-4	0.5614841201763636	

Then I downloaded the data as a csv file and imported it into Tableau for vizualization and analysis.

I began by examining the relationships between mask usage and several other variables related to income and employment. I did this by creating a parameter to allow easy adjustment between variables and a calculated field to implement the paremeter selection.

Edit Parameter [Select Variable]

Name: Select Variable   Comment >>

Properties

Data type: Integer

Current value: Income Inequality (Gini I...

Value when workbook opens: Current value

Display format: Automatic

Allowable values: All   List   Range

Listofvalues

Value	Display As
1	Income Inequa...
2	Median Income
3	Median Rent
4	Poverty Rate
5	Unemploymen...
Add	

Fixed   Add values from   When workbook opens   None   Clear All

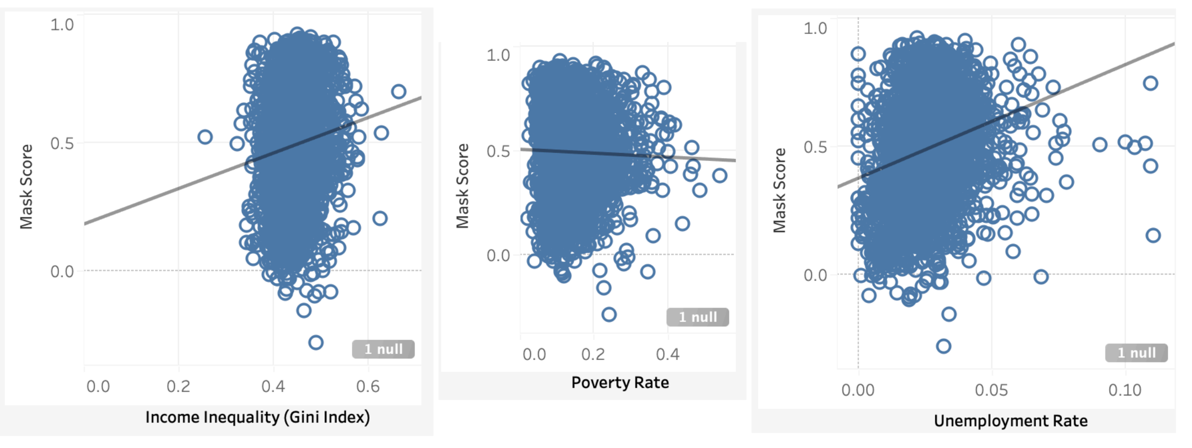
Cancel   OK

Selected Variable

```
IF [Select Variable]= 1
THEN [Gini Index]
ELSEIF [Select Variable] = 2
THEN [Median Income]
ELSEIF [Select Variable] = 3
THEN [Median Rent]
ELSEIF [Select Variable] = 4
THEN [Poverty Pct]
ELSEIF [Select Variable] = 5
THEN [Unemployed Pop Pct]
END
```

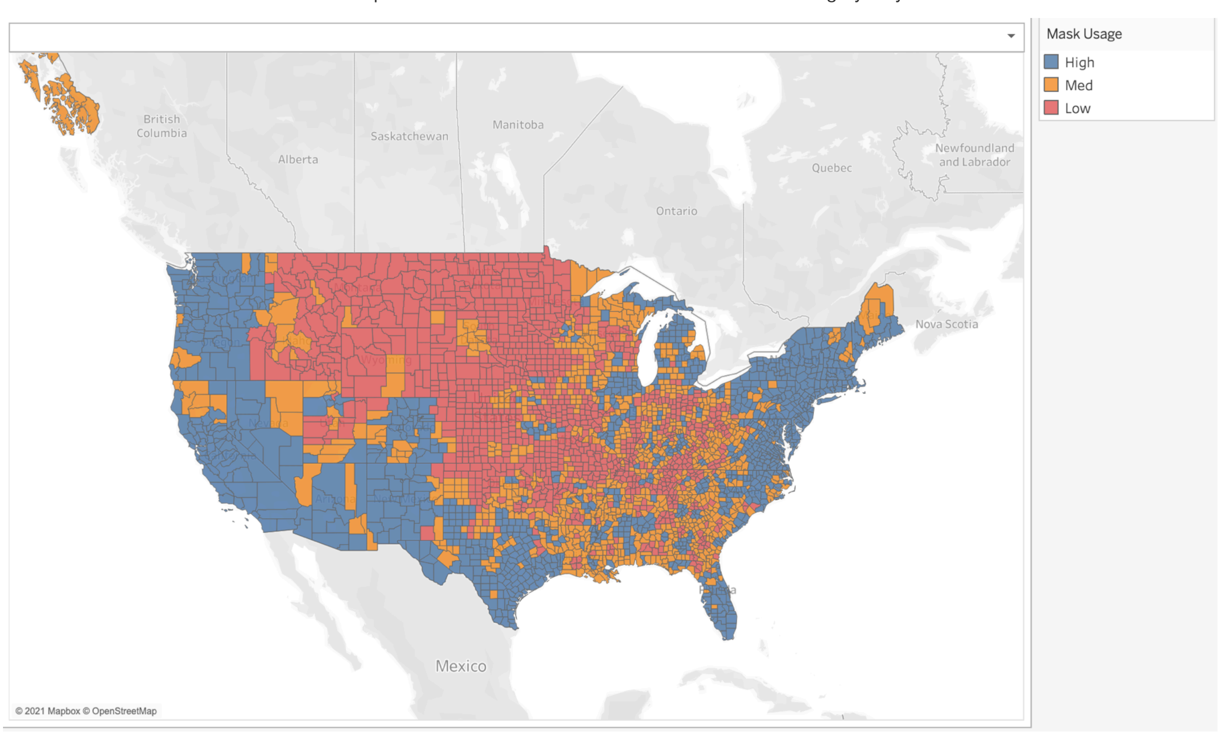
The calculation is valid.   2 Dependencies   Apply   OK

I discovered that there was little to no relationship between income inequality, poverty or unemployment and mask usage.



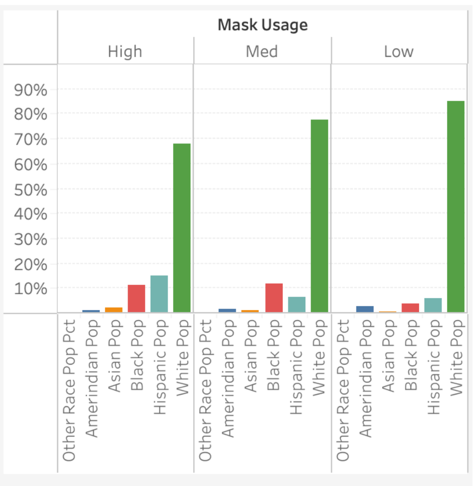
I found weak to moderate relationships between median rent and median income and mask usage.

Then I decided to visualize counties on a map and color code them based on which mask use catagory they fall into:



As you can see mask use is considerably stronger in the coastal areas than in the middle of the country.

I then turned my attention to diversity. When I saw that the low and medium mask counties had a considerably higher white populaltion percentage, I knew I needed to find away to quantify diversity.



In my day job working at a law firm, I've been tasked with measuring market concentration before and after a potential merger to assess its legality. We use a measure called the Herfindahl-Hirschman Index (HHI). The formula for HHI is HHI = S_1_2+S_2_2+S_3_2+…S_n_2 . Where S denotes a firm's market share. Each market share is the expressed as the percentage of total market revenue, and squaring each market share weights each firms market share to account for relative concentration. For example a market with one firm controlling 100% of the market share would have an HHI 100^2= 10,000 and while a market with two firms each controlling 50% would have an HHI of 50^2 + 50^2 = 5000 and so would be half as concentrated.

I adapted this formula to create an index that conceptualized how concentrated one racial group was in each county, which I labeled diversity index. I created a calculated field with this formula.

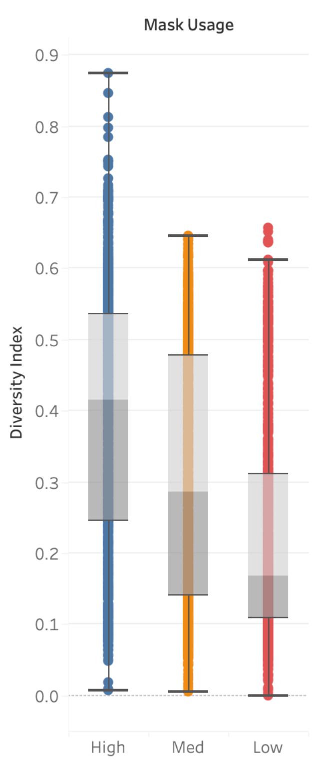
Diversity Index

```
1-
(([Amerindian Pop Pct]^2) +
([Black Pop Pct]^2) +
([Asian Pop Pct]^2) +
([White Pop Pct]^2)+
([Hispanic Pop Pct]^2))+
([Other Race Pop Pct]^2)
```

The calculation is valid.   2 Dependencies   Apply   OK

I subtracted the index from one so that higher values are more diverse and not more concentrated in one race since that is what the formula measured.

As you can see, the the high mask areas tend to be more diverse and all of the most diverse areas are all high mask.



I conducted exploratory data analysis on numerous other values, but did not find inferences significant enough to warrant space on the dashboard. My next project is to evaluate the relationship between mask wearing and COVID-19.

You can view the completed dashboard here:

https://public.tableau.com/profile/william.schulman#!/vizhome/PredictorsofCovid-19MaskWearing/Dashboard1