Analysing Vaccine Trends

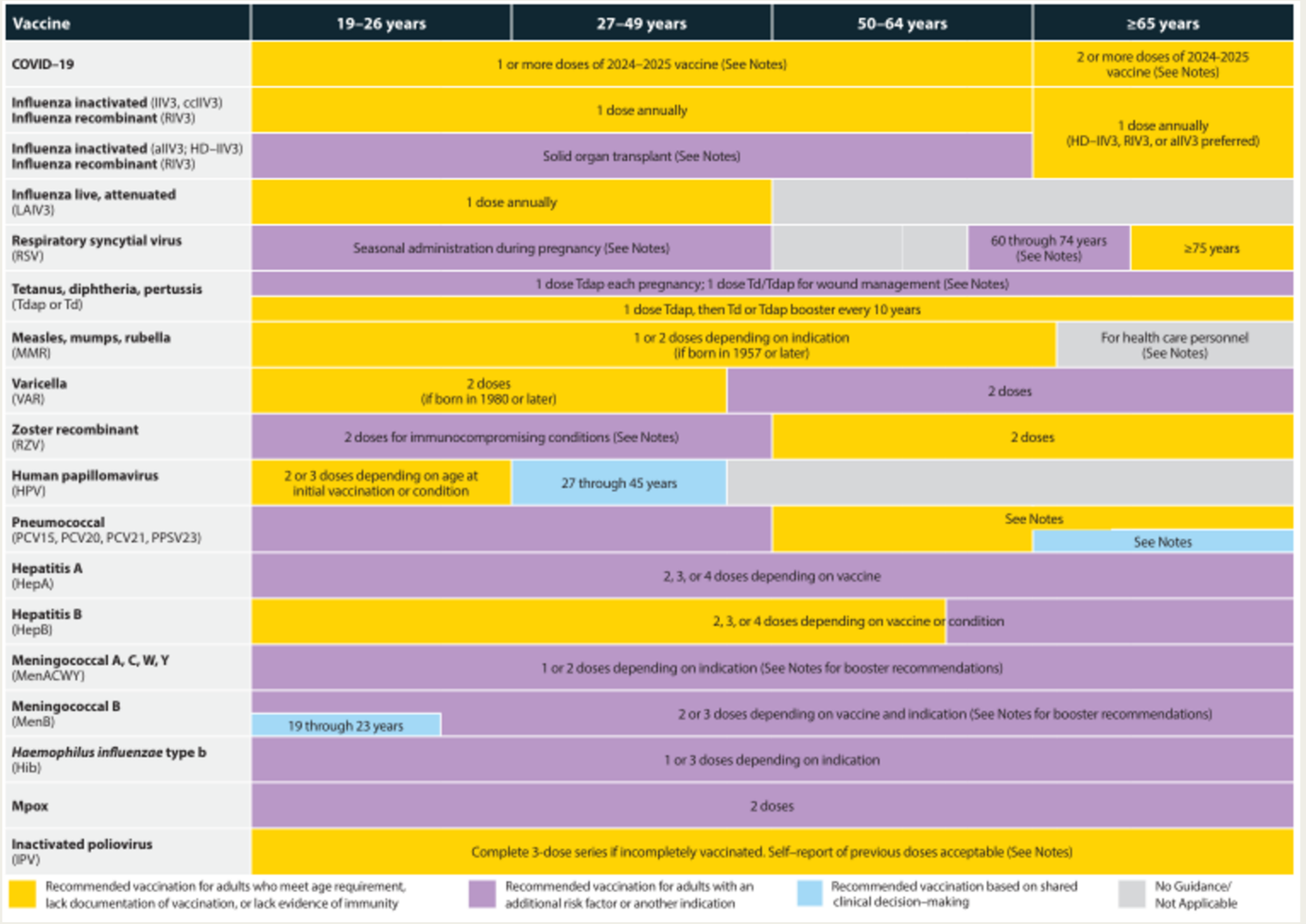
Ameera Adam

A G E N D A

- Scope
  - Motivation
  - Business Questions
  - Datasets Used
- Design
- ETL
- Analytics and Visualisation
- Conclusion
  - Future Work

MOTIVATON

- Vaccination plays a critical role in reducing disease severity
- In the 2023-2024 COVID-19 season, vaccines reduced the risk of critical illness by ~70% after the first 2 months of vaccination
- The CDC provides a comprehensive vaccination schedule for individuals across different age groups
  - Vaccinations administered at different ages for optimal protection
  - Certain vaccines are also recommended based on several risk factors (e.g. occupation, health conditions)



Source:
https://www.cdc.gov/covid/vaccines/benefits.html,
https://www.cdc.gov/vaccines/hcp/imz-schedules/adult-age.html
https://www.cdc.gov/vaccines/hcp/imz-schedules/child-adolescent-age.html

BUSINESS QUESTIONS

- How have vaccination rates changed year-over-year across different age groups?
- What impact does economic background have on influenza vaccination rates?
- Is there a correlation between receiving the influenza vaccine and receiving other recommended vaccines?
- How does vaccination coverage within each demographic group influence future disease trends?
    - In the event of a flu outbreak, can we predict which demographic will be most affected based on vaccination coverage?
- If we were to launch a campaign to increase vaccination rates, which demographic should we prioritise?

DATASETS USED

Vaccination Coverage among Young Children (0-35 months)

Source

- 128k rows, 10 columns
- Split according to birth year (transaction) and birth cohort (cumulative)
- Transactions split by Age
- Cohort split by dimensions:
  - Insurance Coverage
  - Poverty Status
  - Urbanicity
  - Race and Ethnicity
  - Overall: combines above dimensions
- Measures:
  - Estimated percentage
  - 95% Confidence Interval
  - Sample Size

Vaccination Coverage among Adolescents (13-17 Years)

Source

- 27K rows, 10 columns
- Split by survey year (transaction) and survey cohort(cumulative)
- Transactions split by Age groups
- Cohorts split by dimensions:
  - Insurance Coverage
  - Poverty Status
  - Ubranicity
  - Race and Ethnicity
  - Overall: combines above dimensions
- Measures:
  - Estimated percentage
  - 95% Confidence Interval
  - Sample Size

Vaccination Coverage among Pregnant Women

Source

- 4k rows, 9 columns
- Split by survey year/influenza season
- Transactional
- Split by:
  - Age groups
  - Race & Ethnicity
- Measures:
  - Estimated percentage
  - 95% Confidence Interval
  - Sample Size

Influenza Vaccination Coverage for All Ages (6Months +)

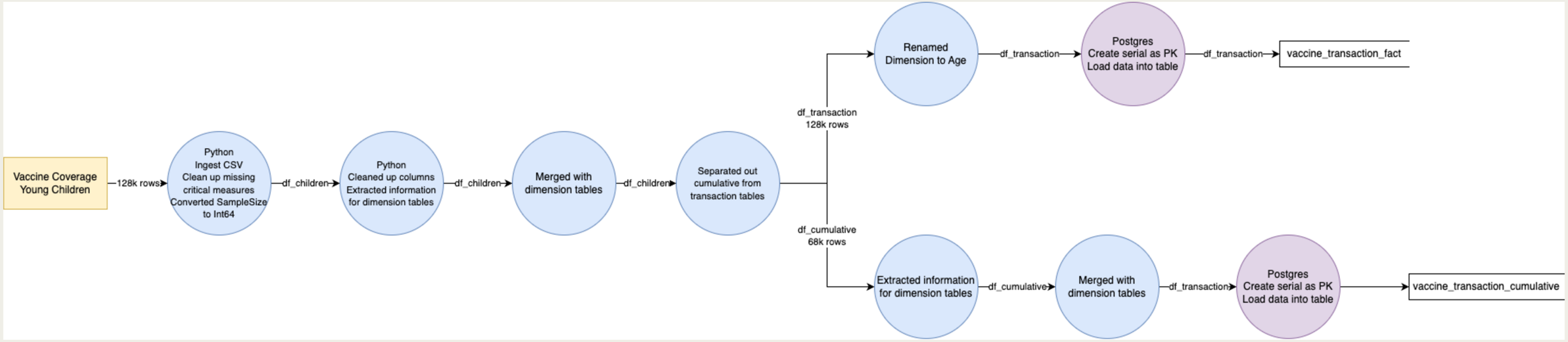
Source

- 220k rows, 11 columns
- Split by survey year (transaction) and survey cohort(cumulative)
- Focus on cohort only
- Split by:
  - Age groups
  - Race and Ethnicity
  - Vaccine Location

DATA FLOW DIAGRAM

Vaccination Coverage among Young Children (0-35 months)

	Vaccine	Dose	Geography Type	Geography	Birth Year/Birth Cohort	Dimension Type	Dimension	Estimate (%)	95% CI (%)	Sample Size
0	DTaP	≥3 Doses	States/Local Areas	North Dakota	2019	Age	19 Months	93.5	88.0 to 96.6	263.0
1	DTaP	≥3 Doses	States/Local Areas	North Dakota	2018	Age	19 Months	95.2	91.0 to 97.5	293.0
2	DTaP	≥3 Doses	States/Local Areas	North Dakota	2018-2019	Age	19 Months	91.8	88.3 to 94.3	556.0
3	Polio	≥3 Doses	States/Local Areas	North Dakota	2021	Age	19 Months	89.4	81.9 to 94.1	143.0
4	Polio	≥2 Doses	States/Local Areas	North Dakota	2021	Age	5 Months	79.3	69.0 to 86.8	143.0



WAREHOUSE DESIGN+ DELTA

- Slowly Changing Dimensions Type 0 (excluding datetime)

Table	Reason
Urbanicity	Urbanicity is assumed to be **static for each record** **Don't expect historical values to change**, and new records can be added without updating existing records
Dose	Number of doses already administered **does not change historically** Once recorded, the **value is considered final and should not be updated**
Gender	Gender is assumed to be **static for each record** **Don't expect historical values to change**, and new records can be added without modifying existing ones
Race and Ethnicity	Gender is assumed to be **static for each record** **Don't expect historical values to change**, and new records can be added without modifying existing ones
Vaccine Location	Locations where vaccine are administered are assumed to be **static for each past record** **Don't expect historical values to change**, and new records can be added without modifying existing ones
FIPS Location	FIPS location from US Census Bureau, mapped to state and region D**on't expect historical values to change** and want new FIPS to be allocated new IDs

WAREHOUSE DESIGN

- SCD Type1, 2, 3

Table	SCD Type	Reason
Geography	Type 1 Hierarchy	Names of municipalities and counties, along with their locations, do not change frequently However, when updates occur (e.g. boundary adjustments), the **data should reflect the latest information to maintain clarity and consistency**
Vaccine	Type2	Vaccines, particularly influenza vaccines, are **updated periodically** to target new strains (e.g., annually) Want to historical accuracy while capturing these changes, allowing us to expire outdated vaccine records and insert updated versions, **maintaining a full history of changes over time**
Insurance	Type2	Insurance coverage descriptions **may change over time regarding which vaccines are covered** Want to preserve historical accuracy and enable meaningful comparisons between past and current coverage policies Allows us to **maintain versioned records and track how insurance coverage has evolved**.
Poverty	Type3	**Poverty statuses themselves do not change frequently**, but there is **potential for redefinition or reclassification of some previously undefined statuses** Want to retain the most recent prior value alongside the current one to support comparisons and audits of limited changes

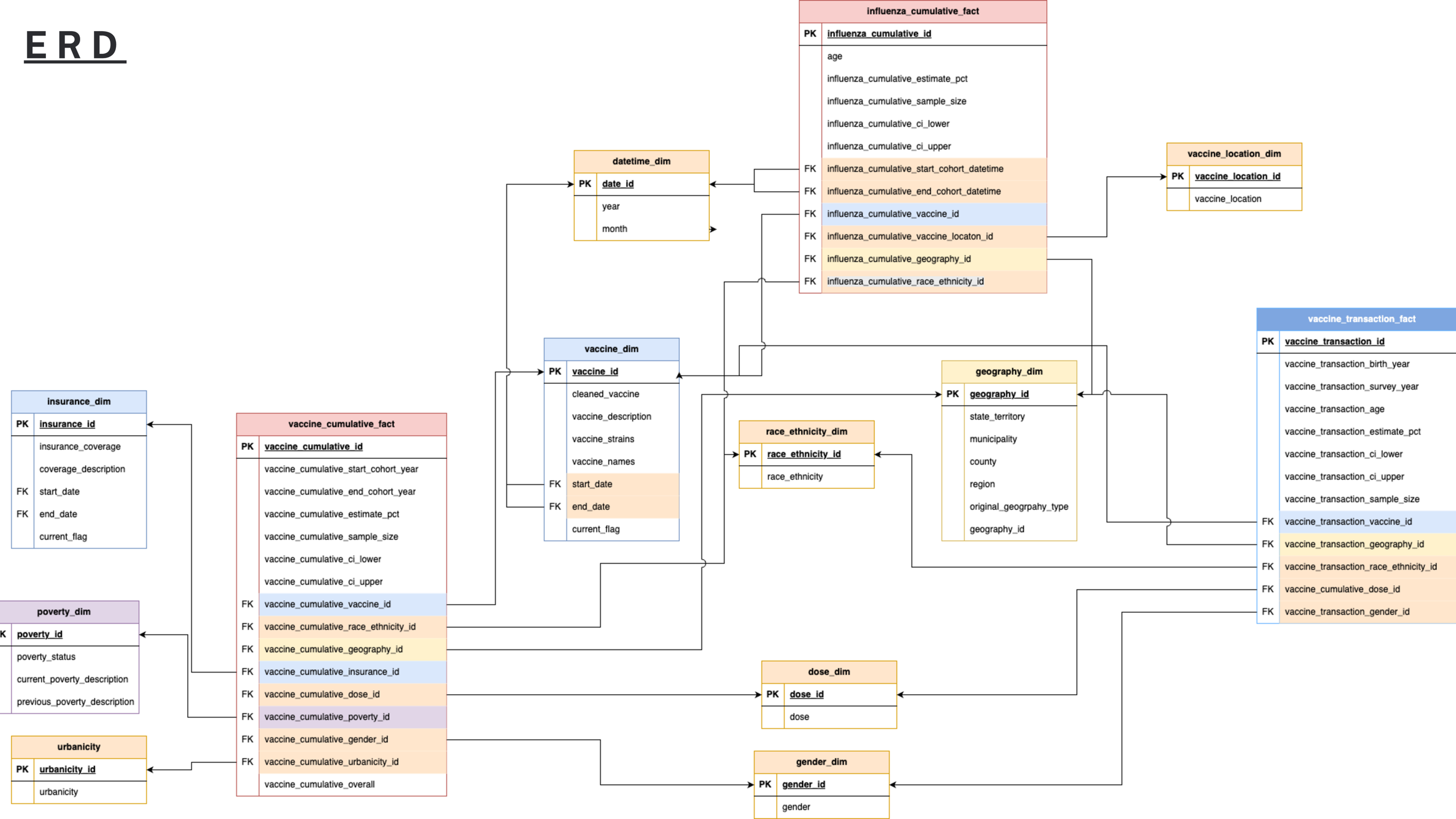
COMBINE DATASETS

- Cumulative Fact Table (vaccine_cumulative_fact)
  - Combine Vaccination Coverage for Young Children & Adolescents
  - By Birth Cohort/Survey Cohort
    - Kept separate in fact table
    - Birth Cohort: Young Children, Survey Cohort: Adolescents
  - Same dimensions (insurance coverage, poverty, race and ethnicity, urbanicity)
- Transaction Fact Table (vaccine_transaction_fact)
  - Combine Vaccination Coverage for Young Children, Adolescents & Pregnant Women
  - By Birth Year/Survey Year
    - Kept separate in fact table
    - Birth Cohort: Young Children, Survey Cohort: Adolescents, Pregnant Women

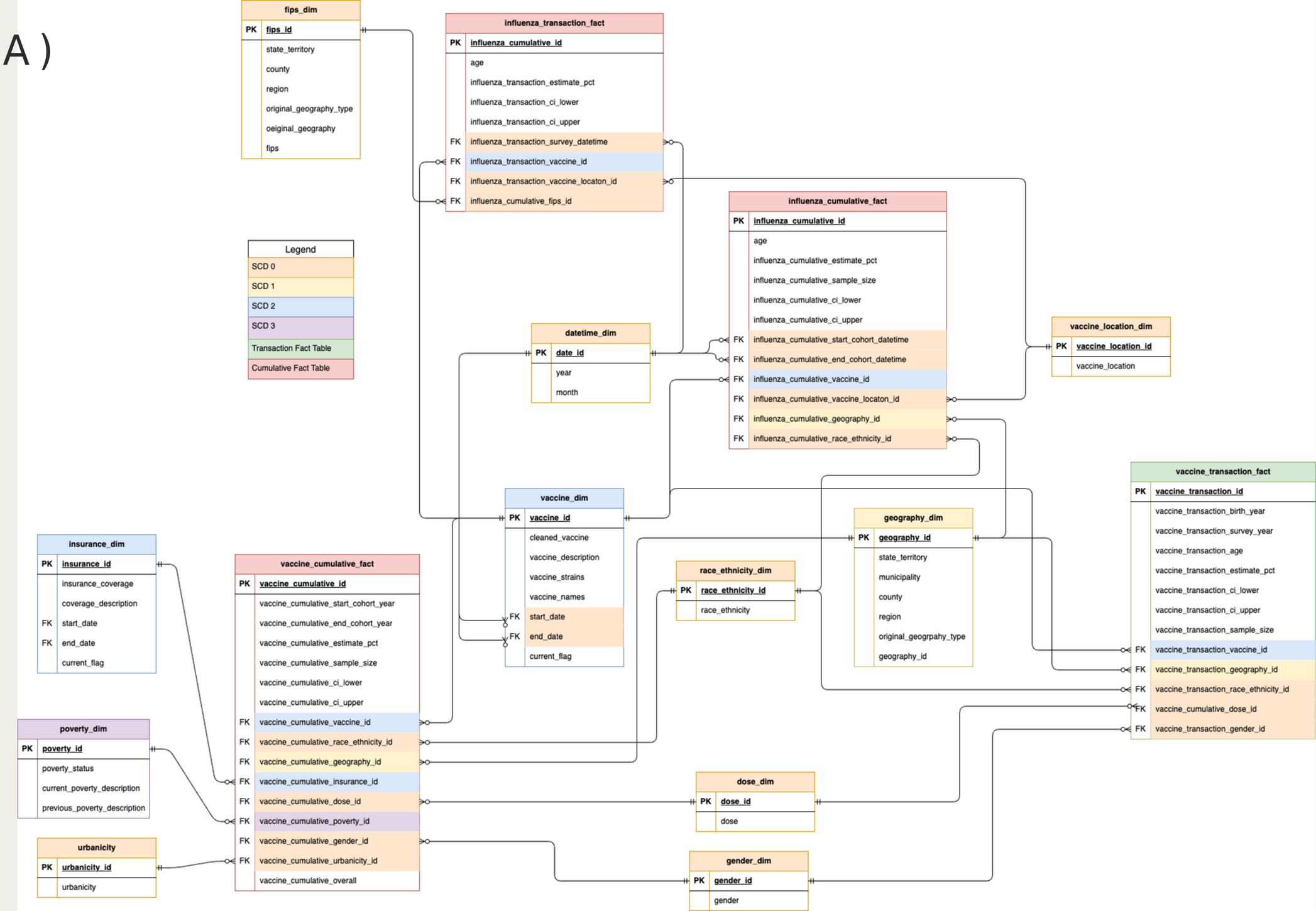
COMBINE DATASETS

- Cumulative Fact Table (influenza_cumulative_fact)
  - Survey Cohort Year
  - Doesn’t contain all the dimensions in vaccine_cumulative_fact
    - Only Race and Ethnicity
  - New dimensions
    - Vaccine_Location
  - Represents all age groups, not only children & adolescents
- No Transaction Fact Table for Influenza
  - No API key to map FIPS area to state
  - Census Bureau has only recently fulfilled my request 
    - Putting influenza_transaction_fact into the delta report

ERD


influenza_cumulative_fact
PK influenza_cumulative_id
age
influenza_cumulative_estimate_pct
influenza_cumulative_sample_size
influenza_cumulative_ci_lower
influenza_cumulative_ci_upper
FK influenza_cumulative_start_cohort_datetime
FK influenza_cumulative_end_cohort_datetime
FK influenza_cumulative_vaccine_id
FK influenza_cumulative_vaccine_locaton_id
FK influenza_cumulative_geography_id
FK influenza_cumulative_race_ethnicity_id
datetime_dim
PK date_id
year
month
vaccine_location_dim
PK vaccine_location_id
vaccine_location
vaccine_transaction_fact
PK vaccine_transaction_id
vaccine_transaction_birth_year
vaccine_transaction_survey_year
vaccine_transaction_age
vaccine_transaction_estimate_pct
vaccine_transaction_ci_lower
vaccine_transaction_ci_upper
vaccine_transaction_sample_size
FK vaccine_transaction_vaccine_id
FK vaccine_transaction_geography_id
FK vaccine_transaction_race_ethnicity_id
FK vaccine_cumulative_dose_id
FK vaccine_transaction_gender_id
vaccine_dim
PK vaccine_id
cleaned_vaccine
vaccine_description
vaccine_strains
vaccine_names
FK start_date
FK end_date
current_flag
geography_dim
PK geography_id
state_territory
municipality
county
region
original_geogrpahy_type
geography_id
race_ethnicity_dim
PK race_ethnicity_id
race_ethnicity
insurance_dim
PK insurance_id
insurance_coverage
coverage_description
FK start_date
FK end_date
current_flag
vaccine_cumulative_fact
PK vaccine_cumulative_id
vaccine_cumulative_start_cohort_year
vaccine_cumulative_end_cohort_year
vaccine_cumulative_estimate_pct
vaccine_cumulative_sample_size
vaccine_cumulative_ci_lower
vaccine_cumulative_ci_upper
FK vaccine_cumulative_vaccine_id
FK vaccine_cumulative_race_ethnicity_id
FK vaccine_cumulative_geography_id
FK vaccine_cumulative_insurance_id
FK vaccine_cumulative_dose_id
FK vaccine_cumulative_poverty_id
FK vaccine_cumulative_gender_id
FK vaccine_cumulative_urbanicity_id
vaccine_cumulative_overall
poverty_dim
K poverty_id
poverty_status
current_poverty_description
previous_poverty_description
dose_dim
PK dose_id
dose
urbanicity
PK urbanicity_id
urbanicity
gender_dim
PK gender_id
gender

ERD(DELTA)


fips_dim
PK fips_id
state_territory
county
region
original_geography_type
oeiginal_geography
fips
influenza_transaction_fact
PK influenza_cumulative_id
age
influenza_transaction_estimate_pct
influenza_transaction_ci_lower
influenza_transaction_ci_upper
FK influenza_transaction_survey_datetime
FK influenza_transaction_vaccine_id
FK influenza_transaction_vaccine_locaton_id
FK influenza_cumulative_fips_id
Legend
SCD 0
SCD 1
SCD 2
SCD 3
Transaction Fact Table
Cumulative Fact Table
influenza_cumulative_fact
PK influenza_cumulative_id
age
influenza_cumulative_estimate_pct
influenza_cumulative_sample_size
influenza_cumulative_ci_lower
influenza_cumulative_ci_upper
FK influenza_cumulative_start_cohort_datetime
FK influenza_cumulative_end_cohort_datetime
FK influenza_cumulative_vaccine_id
FK influenza_cumulative_vaccine_locaton_id
FK influenza_cumulative_geography_id
FK influenza_cumulative_race_ethnicity_id
datetime_dim
PK date_id
year
month
vaccine_location_dim
PK vaccine_location_id
vaccine_location
vaccine_dim
PK vaccine_id
cleaned_vaccine
vaccine_description
vaccine_strains
vaccine_names
FK start_date
FK end_date
current_flag
vaccine_transaction_fact
PK vaccine_transaction_id
vaccine_transaction_birth_year
vaccine_transaction_survey_year
vaccine_transaction_age
vaccine_transaction_estimate_pct
vaccine_transaction_ci_lower
vaccine_transaction_ci_upper
vaccine_transaction_sample_size
FK vaccine_transaction_vaccine_id
FK vaccine_transaction_geography_id
FK vaccine_transaction_race_ethnicity_id
FK vaccine_cumulative_dose_id
FK vaccine_transaction_gender_id
geography_dim
PK geography_id
state_territory
municipality
county
region
original_geogrpahy_type
geography_id
race_ethnicity_dim
PK race_ethnicity_id
race_ethnicity
insurance_dim
PK insurance_id
insurance_coverage
coverage_description
FK start_date
FK end_date
current_flag
vaccine_cumulative_fact
PK vaccine_cumulative_id
vaccine_cumulative_start_cohort_year
vaccine_cumulative_end_cohort_year
vaccine_cumulative_estimate_pct
vaccine_cumulative_sample_size
vaccine_cumulative_ci_lower
vaccine_cumulative_ci_upper
FK vaccine_cumulative_vaccine_id
FK vaccine_cumulative_race_ethnicity_id
FK vaccine_cumulative_geography_id
FK vaccine_cumulative_insurance_id
FK vaccine_cumulative_dose_id
FK vaccine_cumulative_poverty_id
FK vaccine_cumulative_gender_id
FK vaccine_cumulative_urbanicity_id
vaccine_cumulative_overall
poverty_dim
PK poverty_id
poverty_status
current_poverty_description
previous_poverty_description
dose_dim
PK dose_id
dose
urbanicity
PK urbanicity_id
urbanicity
gender_dim
PK gender_id
gender

ETL: CREATING DIMENSION TABLES

Creating and updating vaccine dimension (SCD2)

- Cleaned out vaccine column:
  - Removed '>= 1 Dose' in rows
  - Put into Dose column
  - Added in '1 Dose Only' for vaccines with only 1 Dose needed (Influenza, Rotavirus)

Before

	Vaccine	Dose	Geography Type	Geography	Birth Year/Birth Cohort	Dimension Type	Dimension	Estimate (%)	95% CI (%)	Sample Size
11	≥1 Dose MMR	NaN	States/Local Areas	North Dakota	2021	Age	19 Months	79.9	69.5 to 87.3	143.0
44	≥1 Dose MMR	NaN	States/Local Areas	North Dakota	2020-2021	Age	35 Months	91.0	85.0 to 95.2	391.0
65	≥1 Dose MMR	NaN	States/Local Areas	North Dakota	2020-2021	Age	19 Months	83.8	77.6 to 88.5	391.0
77	≥1 Dose MMR	NaN	States/Local Areas	North Dakota	2021	Age	13 Months	62.7	51.1 to 73.0	143.0
386	≥1 Dose MMR	NaN	States/Local Areas	North Dakota	2018	Age	13 Months	69.9	61.7 to 77.0	293.0

After

	Vaccine	Dose	Geography Type	Geography	Birth Year/Birth Cohort	Dimension Type	Dimension	Estimate (%)	95% CI (%)	Sample Size
11	≥1 Dose MMR	≥1 Dose	States/Local Areas	North Dakota	2021	Age	19 Months	79.9	69.5 to 87.3	143.0
44	≥1 Dose MMR	≥1 Dose	States/Local Areas	North Dakota	2020-2021	Age	35 Months	91.0	85.0 to 95.2	391.0
65	≥1 Dose MMR	≥1 Dose	States/Local Areas	North Dakota	2020-2021	Age	19 Months	83.8	77.6 to 88.5	391.0
77	≥1 Dose MMR	≥1 Dose	States/Local Areas	North Dakota	2021	Age	13 Months	62.7	51.1 to 73.0	143.0
386	≥1 Dose MMR	≥1 Dose	States/Local Areas	North Dakota	2018	Age	13 Months	69.9	61.7 to 77.0	293.0

- Assumptions:
  - Tetanus vaccine in Adolescent Dataset refers to Tdap: Type of tetanus vaccine for adolescents 7-18 years
  - Varicella has a 'History of Disease' as a dose as you don't need a vaccine or booster if you have the virus (i.e. you've had chickenpox previously). The body still has protection if you've tested positive for the virus
  - For the influenza table, assume seasonal influenza and influenza are the same
  - Not medical advice, just assumptions! IANAD!

ETL: CREATING DIMENSION TABLES

Creating and updating vaccine dimension (SCD2)

- Used Dr. ChatGPT to fill in the different strains and vaccine names starting from 2011
- Inserted into vaccine_dim table
- Updated influenza vaccines with new strains

```
update_query = """
UPDATE vaccine_dim
SET end_date     = %s - INTERVAL '1 month',
    current_flag = 'N'
WHERE cleaned_vaccine = %s
AND current_flag   = 'Y';
"""

insert_query = """
INSERT INTO vaccine_dim
(cleaned_vaccine,
vaccine_description,
vaccine_strains,
vaccine_names,
start_date,
end_date,
current_flag)
VALUES (%s, %s, %s, %s, %s, %s, %s)
ON CONFLICT (cleaned_vaccine, start_date) DO NOTHING;
"""
try:
    for _, row in df_update_vaccine.iterrows():
        cursor.execute(
            update_query,
            (row['start_date'], row['Cleaned Vaccine'])
        )
        cursor.execute(
            insert_query,
            (
                row['Cleaned Vaccine'],
                row['Vaccine Description'],
                row['Vaccine Strain'],
                row['Vaccine Name'],
                row['start_date'],
                row['end_date'],
                'Y'
            )
        )
    conn.commit()
```

Before

	cleaned_vaccine character varying (	vaccine_description character varying (200)	vaccine_strains character varying (200)	vaccine_names character varying (200)	start_date date	end_date date	current_flag character varying (1)	vaccine_id [PK] integer
1	DTaP	Protects against Diphthe...	Diphtheria, Tetanus, and Bordetella pertussis (whooping cough)	Daptacel, Infanrix	2011-09-01	2100-12-01	Y	1
2	Polio	Protects against Poliom...	Poliovirus Types 1, 2, and 3 (inactivated)	IPOL	2011-09-01	2100-12-01	Y	2
3	Hep B	Protects against Hepatit...	Hepatitis B virus (HBV)	Engerix-B, Recombivax HB	2011-09-01	2100-12-01	Y	3
4	PCV	Protects against Pneum...	Streptococcus pneumoniae — multiple serotypes (e.g., 13, 15, ...	Prevnar 13, Prevnar 15, Prevnar 20	2011-09-01	2100-12-01	Y	4
5	Varicella	Protects against chicken...	Varicella-zoster virus (chickenpox)	Varivax, ProQuad (MMRV combo)	2011-09-01	2100-12-01	Y	5
6	MMR	Protects against Measle...	Measles virus, Mumps virus, Rubella virus	M-M-R II, ProQuad (MMRV combo)	2011-09-01	2100-12-01	Y	6
7	Hib	Protects against Haemo...	Haemophilus influenzae type b	ActHIB, PedvaxHIB, Hiberix	2011-09-01	2100-12-01	Y	7
8	Hep A	Protects against Hepatit...	Hepatitis A virus	Havrix, Vaqta	2011-09-01	2100-12-01	Y	8
9	Influenza	Protects against season...	A/California/7/2009, A/Victoria/361/2011 (H3N2), B/Wisconsi...	Fluzone, Fluarix, FluMist (live), Flublok	2011-09-01	2100-12-01	Y	9
10	Combined 7 Se...	A combination of 7 reco...	Includes: DTaP, IPV, MMR, Hib, Hep B, Varicella, PCV	Includes: DTaP, IPV, MMR, Hib, Hep B...	2011-09-01	2100-12-01	Y	10
11	Rotavirus	Protects against Rotavir...	Rotavirus types G1, G2, G3, G4, G9 (Rotarix); G1–G4, G9, G12 (R...	Rotarix (monovalent), RotaTeq (pent...	2011-09-01	2100-12-01	Y	11

After

	cleaned_vaccine character varying (50)	vaccine_description character varying (200)	vaccine_strains character varying (200)	vaccine_names character varying (200)	start_date date	end_date date	current_flag character varying (1)	vaccine_id [PK] integer
1	Influenza	Protects against seasonal flu viruses	A/California/7/2009, A/Vic...	Fluzone, Fluarix, FluMist (live), Flublok	2011-09-01	2012-08-01	N	9
2	Influenza	Protects against seasonal flu viruses	A/California/7/2009, A/Vic...	Fluzone, Fluarix, FluMist (live), Flublok	2012-09-01	2013-08-01	N	12
3	Influenza	Protects against seasonal flu viruses	A/California/7/2009, A/Te...	Fluzone, Fluarix, FluMist (live), Flublok	2013-09-01	2014-08-01	N	13
4	Influenza	Protects against seasonal flu viruses	A/California/7/2009, A/Te...	Fluzone, Fluarix, FluMist (live), Flublok	2014-09-01	2015-08-01	N	14
5	Influenza	Protects against seasonal flu viruses	A/California/7/2009, A/Sw...	Fluzone, Fluarix, FluMist (live), Flublok	2015-09-01	2016-08-01	N	15
6	Influenza	Protects against seasonal flu viruses	A/California/7/2009, A/Ho...	Fluzone, Fluarix, FluMist (live), Flublok	2016-09-01	2017-08-01	N	16
7	Influenza	Protects against seasonal flu viruses	A/Michigan/45/2015, A/H...	Fluzone, Fluarix, FluMist (live), Flublok	2017-09-01	2018-08-01	N	17
8	Influenza	Protects against seasonal flu viruses	A/Michigan/45/2015, A/Si...	Fluzone, Fluarix, FluMist (live), Flublok	2018-09-01	2019-08-01	N	18
9	Influenza	Protects against seasonal flu viruses	A/Brisbane/02/2018, A/Ka...	Fluzone, Fluarix, FluMist (live), Flublok	2019-09-01	2020-08-01	N	19
10	Influenza	Protects against seasonal flu viruses	A/Guangdong-Maonan/S...	Fluzone, Fluarix, FluMist (live), Flublok	2020-09-01	2021-08-01	N	20
11	Influenza	Protects against seasonal flu viruses	A/Victoria/2570/2019, A/...	Fluzone, Fluarix, FluMist (live), Flublok	2021-09-01	2022-08-01	N	21
12	Influenza	Protects against seasonal flu viruses	A/Victoria/2570/2019, A/...	Fluzone, Fluarix, FluMist (live), Flublok	2022-09-01	2023-08-01	N	22
13	Influenza	Protects against seasonal flu viruses	A/Victoria/4897/2022, A/...	Fluzone, Fluarix, FluMist (live), Flublok	2023-09-01	2024-08-01	N	23

ETL: CREATING DIMENSION TABLES

Creating poverty dimension (SCD3)

- Poverty Levels don't really change often
- But there are undefined Poverty Statuses in the various datasets

```
array(['Below Poverty Level', 'Living At or Above Poverty Level'],
      dtype=object)
```

Undefined description

- Added in a dummy description and loaded it into SCD3 table
- If the definition gets updated, we can update the value with the new update and expire the old update as 'previous_poverty_description'

	poverty_status character varying (50)	current_poverty_description character varying (150)	previous_poverty_description character varying (150)	poverty_id [PK] integer
1	<133% FPL	Household income is less than 133% of the Federal Poverty Level	[null]	1
2	133% to <400% FPL	Income is between 133% and just under 400% of the Federal Poverty Level	[null]	2
3	>400% FPL	Income is above 400% of the Federal Poverty Level	[null]	3
4	Below Poverty Level	Living below poverty level. Undefined percentage.	[null]	4
5	Living At or Above Poverty Level	Living above poverty level. Undefined percentage.	[null]	5

```
cursor.execute("""
SELECT setval(
  'poverty_dim_poverty_id_seq',
  (SELECT MAX(poverty_id) FROM poverty_dim)
);
""")
conn.commit()
insert_query = '''
INSERT INTO poverty_dim (
    poverty_status,
    current_poverty_description,
    previous_poverty_description
)
VALUES (%s, %s, %s)
ON CONFLICT (poverty_status)
DO UPDATE SET
    previous_poverty_description = poverty_dim.current_poverty_description,
    current_poverty_description = EXCLUDED.current_poverty_description;
'''
for index, row in poverty_dim.iterrows():
    print(f"Processing poverty status: {row['Poverty']}")
    cursor.execute(
        insert_query,
        (
            row['Poverty'],
            row['current_poverty_description'],
            row.get('Previous Poverty Description', None)
        )
    )
conn.commit()
print(f"{len(poverty_dim)} records inserted into poverty_dim.")
```

ETL: CREATING DIMENSION TABLES

Creating geography dimension (SCD1 Hierarchy)

- The dataset has two geography tables
  - Geography: Either a ‘State/Local Area’ or ‘HHS Region/National’
  - Geography Type:
    - Municipality/County/Town if Geography=State/Local Area
      - Counties are in the form <state abbreviation>-<county> (e.g. TX-Dallas County)
    - Region Number or United States if Geography = National
- Map the municipality/county to the State/Local Area
- Map the State to the HHS Region
- Hierarchy: Municipality > County > State > HHS Region
- Created two dictionaries mapping the state to the county and state to region (Geographer ChatGPT helped with the state and region names and mapping)

```
array(['North Dakota', 'North Carolina', 'New Jersey', 'New Mexico',
       'Kansas', 'Nebraska', 'Pennsylvania', 'Oregon', 'Indiana',
       'IL-Rest of state', 'Iowa', 'Texas', 'Utah', 'Vermont',
       'U.S. Virgin Islands', 'Tennessee', 'Alaska', 'Alabama',
       'Oklahoma', 'Georgia', 'West Virginia', 'Wisconsin', 'Ohio',
       'Delaware', 'Connecticut', 'Colorado', 'South Dakota', 'Florida',
       'Kentucky', 'Virginia', 'Washington', 'Arizona', 'South Carolina',
       'Louisiana', 'Maine', 'Minnesota', 'Michigan', 'Hawaii', 'Montana',
       'Missouri', 'District of Columbia', 'Mississippi', 'Wyoming',
       'Maryland', 'Puerto Rico', 'Illinois', 'NY-Rest of state',
       'NY-City of New York', 'PA-Philadelphia', 'New Hampshire',
       'Nevada', 'IL-City of Chicago', 'Idaho', 'TX-City of Houston',
       'New York', 'Rhode Island', 'TX-Rest of state', 'Massachusetts',
       'Guam', 'Arkansas', 'California', 'TX-Travis County',
       'TX-Tarrant County', 'TX-Hidalgo County', 'TX-El Paso County',
       'TX-Dallas County', 'TX-Bexar County', 'PA-Rest of state'],
      dtype=object)
```

	state_territory	municipality	county	region	Geography Type	Geography	id
0	None	None	None	Region 6	HHS Regions/National	Region 6	1
1	None	None	None	Region 5	HHS Regions/National	Region 5	2
2	None	None	None	Region 10	HHS Regions/National	Region 10	3
3	None	None	None	Region 7	HHS Regions/National	Region 7	4
4	None	None	None	Region 3	HHS Regions/National	Region 3	5
...	...	...	...	...	...	...	...
74	Texas	None	Hidalgo County	Region 6	States/Local Areas	TX-Hidalgo County	75
75	Texas	None	El Paso County	Region 6	States/Local Areas	TX-El Paso County	76
76	Texas	None	Dallas County	Region 6	States/Local Areas	TX-Dallas County	77
77	Texas	None	Bexar County	Region 6	States/Local Areas	TX-Bexar County	78
78	Pennsylvania	Rest of state	None	Region 3	States/Local Areas	PA-Rest of state	79

ETL: CREATING DIMENSION TABLES

Creating Gender Dimension Table (SCD0)

- Gender is a dimension unique to Adolescent Dataset
- Specific to the HPV vaccine: Protect against STDs

```
df_adolescents['Vaccine/Sample'][df_adolescents['Dose']=='≥3 Doses, Males and Females'].unique()
```

Python

```
array(['HPV'], dtype=object)
```

- Can denormalise it into Males, Females, Males and Females
- Remove the Gender attribute in the Dose column

```
# since the HPV dose has additional male and female dimensions, we can denormalie it fu
gender = ['Males', 'Females', 'Males and Females']
def extract_gender(dose):
    # Check if there is a comma in the dose string
    if ',' in dose:
        # Extract the portion after the last comma and remove any extra whitespace.
        gender_part = dose.split(',')[-1].strip()
    else:
        gender_part = dose.strip()

    # Perform an exact match against our gender list
    if gender_part in gender:
        return gender_part
    else:
        return None
df_adolescents['Gender'] = df_adolescents['Dose'].apply(extract_gender)
```

	Dose	Gender
0	≥1 Dose, Males	Males
4	≥1 Dose, Males and Females	Males and Females
5	Up-to-Date, Males	Males
6	≥1 Dose, Females	Females
7	Up-to-Date, Males and Females	Males and Females

ETL: CREATING DIMENSION TABLES

Creating SCD0 dimension tables (e.g. Dose dimension table)

- Extract out unique values
- Loading them into the DB

Before

	Dose
0	≥1 Dose
1	≥1 Dose Tdap
2	≥1 Dose Td or Tdap
3	Up-to-Date
4	≥3 Dose
5	≥2 Dose
6	≥2 Doses or history of disease
7	≥2 Doses with no disease history
8	≥1 Dose with no disease history
9	History of disease
10	≥2 Doses
11	≥3 Doses
12	Series Completion (3 Dose) Among HPV Vaccinati...

```
cursor.execute("""
SELECT setval(
    'dose_dim_dose_id_seq',
    (SELECT MAX(dose_id) FROM dose_dim)
);
""")
conn.commit()

insert_query = '''
INSERT INTO dose_dim (dose)
VALUES (%s)
ON CONFLICT (dose) DO NOTHING;
'''
for index, row in dose_dim.iterrows():
    print(row['Dose'])
    cursor.execute(insert_query, (row['Dose'],))
conn.commit()
print(f"{len(dose_dim)} records inserted into dose_dim.")
```

After

	dose character varying (100)	dose_id [PK] integer
1	≥3 Doses	1
2	≥2 Doses	2
3	≥1 Dose	3
4	≥1 Dose, 2 Day	4
5	Full Series	5

ETL: CREATING DIMENSION TABLES(DELTA)

Creating SCD0 FIPS table

- I used the US Census Breuau API to get the FIPS data and map them to the state and US region

```
# base URL for the Census Bureau 2020 Decennial Census Public Law 94-171 Redistricting Data
url = "https://api.census.gov/data/2020/dec/pl"
## key = key

# Set up the parameters for the API call:
def chunk_dataframe(df, chunk_size):
    for i in range(0, len(df), chunk_size):
        yield df.iloc[i:i+chunk_size]

for chunk in tqdm(chunk_dataframe(df_counties, chunk_size), desc="Processing chunks"):
    for _, row in tqdm(chunk.iterrows(), total=len(chunk), leave=False, desc="Fetching FIPS"):
        fips = row['FIPS_clean']
        if fips in county_state:
            continue

        state = fips[:2]
        county = fips[2:]

        params = {
            "get": "NAME",
            "for": f"county:{county}",
            "in": f"state:{state}",
            "key": key

        }

        try:
            response = requests.get(url, params=params, timeout=10)

            if response.status_code == 200:
                data = response.json()
                county_info = data[1]
                county_state[fips] = county_info[0]
            else:
                print(f"Error for {fips}: {response.status_code} - {response.text}")


        except requests.exceptions.Timeout:
            print(f"Timeout for {fips}")
        except Exception as e:
            print(f"Request error for {fips}:", e)
```

```
df_county_list = df = pd.DataFrame([
    {'FIPS': k, 'County': v.rsplit(',', 1)[0], 'State': v.rsplit(',', 1)[1].strip()}
    for k, v in data.items()
])
```

✓ 0.0s

```
for idx, row in df_county_list.iterrows():
    state = row['State']
    region = state_to_region.get(state)
    df_county_list.loc[idx, 'Region'] = region
df_county_list.tail()
```

✓ 0.2s

	FIPS	County	State	Region
3150	09170	South Central Connecticut Planning Region	Connecticut	1.0
3151	09180	Southeastern Connecticut Planning Region	Connecticut	1.0
3152	09190	Western Connecticut Planning Region	Connecticut	1.0
3153	21179	Nelson County	Kentucky	4.0
3154	48219	Hockley County	Texas	6.0

ETL: LOADING IN FACT TABLES

Creating cumulative fact tables - Years are grouped together in cohorts

- Removed fields with missing measures: Estimated %, 95%CI, Sample Size
- Split Confidence Intervals to Upper and Lower Bounds
- Denormalise cohort year as Start Cohort Year and End Cohort Year
- Converted Sample Size to a Int64 - Sample Sizes represent actual humans, should be a whole number
- Age is kept as a character field:
  - Children Dataset: contains days, months
  - Adolescent Dataset: contains year range
- Filled in missing dimensions & cleaned out columns
  - Filled in Dose if the Dose is in the Vaccine column
  - Removed ‘Dose’ in Vaccine column
  - Extracted out gender for Adolescent dataset
- Left merge on dimension tables
  - Vaccine: vaccine, year
  - Geography: geography, geography type
  - Gender
- Extracted out each dimension from the Dimension column and perform left merge on them
  - Insurance
  - Poverty
  - Urbanicity
  - Race and Ethnicity
- Combined all into cumulative fact table
- Keep the Overall column

Estimate (%)	95% CI (%)	Sample Size	ci_lower	ci_upper	vaccine_id	dose_id	geography_id	gender_id	start_cohort_year	end_cohort_year	cumulative_id	insurance_id	poverty_id	race_ethnicity_id	urbanicity_id	Overall
88.3	68.1 to 96.4	36	68.1	96.4	26	13	66	2	2018	2022	1	4.0	NaN	NaN	NaN	NaN
91.0	87.0 to 93.8	365	87.0	93.8	26	13	66	2	2018	2022	2	2.0	NaN	NaN	NaN	NaN
89.2	86.4 to 91.4	897	86.4	91.4	26	13	66	2	2018	2022	3	3.0	NaN	NaN	NaN	NaN
92.6	90.8 to 94.0	1581	90.8	94.0	26	13	66	2	2018	2022	4	1.0	NaN	NaN	NaN	NaN
40.1	23.3 to 59.6	36	23.3	59.6	25	15	66	3	2018	2022	5	4.0	NaN	NaN	NaN	NaN

```sql
select
vaccine_cumulative_start_cohort_year as start_year,
vaccine_cumulative_end_cohort_year as end_year,
vaccine_cumulative_estimate_pct as estimated_ct,
vaccine_cumulative_sample_size as sample_size,
vaccine_cumulative_ci_lower as lower_ci,
vaccine_cumulative_ci_upper as upper_ci,
vaccine_cumulative_geography_id as geo_id,
vaccine_cumulative_vaccine_id as vac_id,
vaccine_cumulative_dose_id as dose_id
from vaccine_cumulative_fact
where vaccine_cumulative_overall is not null
limit 10
```


	start_year integer	end_year integer	estimated_ct double precision	sample_size integer	lower_ci double precision	upper_ci double precision	geo_id integer	vac_id integer	dose_id integer
1	2018	2019	91.8	556	88.3	94.3	12	1	1
2	2020	2021	91.7	391	87.4	94.7	12	2	2
3	2020	2021	92	391	87.8	94.9	12	3	1
4	2020	2021	86.9	391	81.6	90.9	12	3	4
5	2020	2021	72.9	391	66	78.9	12	7	5
6	2020	2021	72.1	391	65.4	78	12	4	1
7	2020	2021	89.5	391	84.7	92.8	12	1	1
8	2020	2021	86.5	391	74.5	94.6	12	7	5
9	2020	2021	78.9	391	72.6	84.1	12	3	1
10	2020	2021	90.5	391	85.8	93.7	12	4	1

ETL: LOADING IN FACT TABLES

Creating transaction fact tables - Years are separated

- Removed fields with missing measures: Estimated %, 95%CI, Sample Size
- Split Confidence Intervals to Upper and Lower Bounds
- Converted Sample Size to a Int64 - Sample Sizes represent actual humans, should be a whole number
- Age is kept as a character field:
  - Children Dataset: contains days, months
  - Adolescent Dataset: contains year range
- Filled in missing dimensions & cleaned out columns
  - Filled in Dose if the Dose is in the Vaccine column
  - Removed ‘Dose’ in Vaccine column
  - Extracted out gender for Adolescent dataset
- Left merge on dimension tables
  - Vaccine: vaccine, year
  - Geography: geography, geography type
  - Gender
- Included Race and Ethnicity dimension for Pregnant Women dataset

Adolescent Dataset

	Survey Year	Age	Estimate (%)	95% CI (%)	Sample Size	ci_lower	ci_upper	vaccine_id	dose_id	geography_id	gender_id
0	2023	13-17 Years	81.5	75.2 to 86.5	289	75.2	86.5	25	3	66	1
1	2023	13-17 Years	90.2	86.8 to 92.8	559	86.8	92.8	26	13	66	2
2	2023	13-17 Years	93.6	90.9 to 95.5	559	90.9	95.5	26	14	66	2
3	2023	13-17 Years	95.3	92.7 to 97.0	559	92.7	97.0	27	3	66	2
4	2023	13-17 Years	79.4	74.8 to 83.3	559	74.8	83.3	25	3	66	3

Survey Year is filled, Birth Year is empty

no race_ethnicity_id

	vaccine_ [PK] inte	vaccine_t integer	vaccine_trai integer	vaccine_transaction_ character varying (5	vaccine_transac double precisio	vaccine_t double pr	vaccine_transa double precisi	vaccine_trans integer	vaccine_trai integer	vaccine_tra integer	vaccine_trai integer	vaccine_tra integer	vaccine_transac integer
1	1	2023	[null]	13-17 Years	81.5	75.2	86.5	289	25	3	1	66	[null]
2	2	2023	[null]	13-17 Years	90.2	86.8	92.8	559	26	13	2	66	[null]
3	3	2023	[null]	13-17 Years	93.6	90.9	95.5	559	26	14	2	66	[null]
4	4	2023	[null]	13-17 Years	95.3	92.7	97	559	27	3	2	66	[null]
5	5	2023	[null]	13-17 Years	79.4	74.8	83.3	559	25	3	3	66	[null]
6	6	2023	[null]	13-17 Years	67.8	60.8	74.2	289	25	15	1	66	[null]
7	7	2023	[null]	13-17 Years	77.2	70.1	83	270	25	3	4	66	[null]
8	8	2023	[null]	13-17 Years	68.6	63.5	73.2	559	25	15	3	66	[null]
9	9	2023	[null]	13-17 Years	94.2	91.1	96.3	559	28	16	2	66	[null]
10	10	2023	[null]	13-17 Years	94.6	91.8	96.5	559	6	17	2	66	[null]

ETL: LOADING IN FACT TABLES

Creating influenza cumulative fact tables - Years are grouped together in cohorts

- Removed fields with missing measures: Estimated %, 95%CI, Sample Size
- Split Confidence Intervals to Upper and Lower Bounds
- Denormalise cohort year as Start Cohort Year and End Cohort Year
- There are fields where the estimates are unreliable: Setting as None, advised to discard
  - Sample Size <30
  - Standard Error >0.3
  - CI half-width >10
- Converted Sample Size to a Int64 - Sample Sizes represent actual humans, should be a whole number
- There is a specific vaccine for the 2009 H1N1 strain
  - Created and updated in vaccine dimension table
- Left merge on dimension tables
  - Vaccine: vaccine, year
  - Geography: geography, geography type
- Extracted out each dimension from the Dimension column and perform left merge on them
  - Vaccine Location
  - Race and Ethnicity

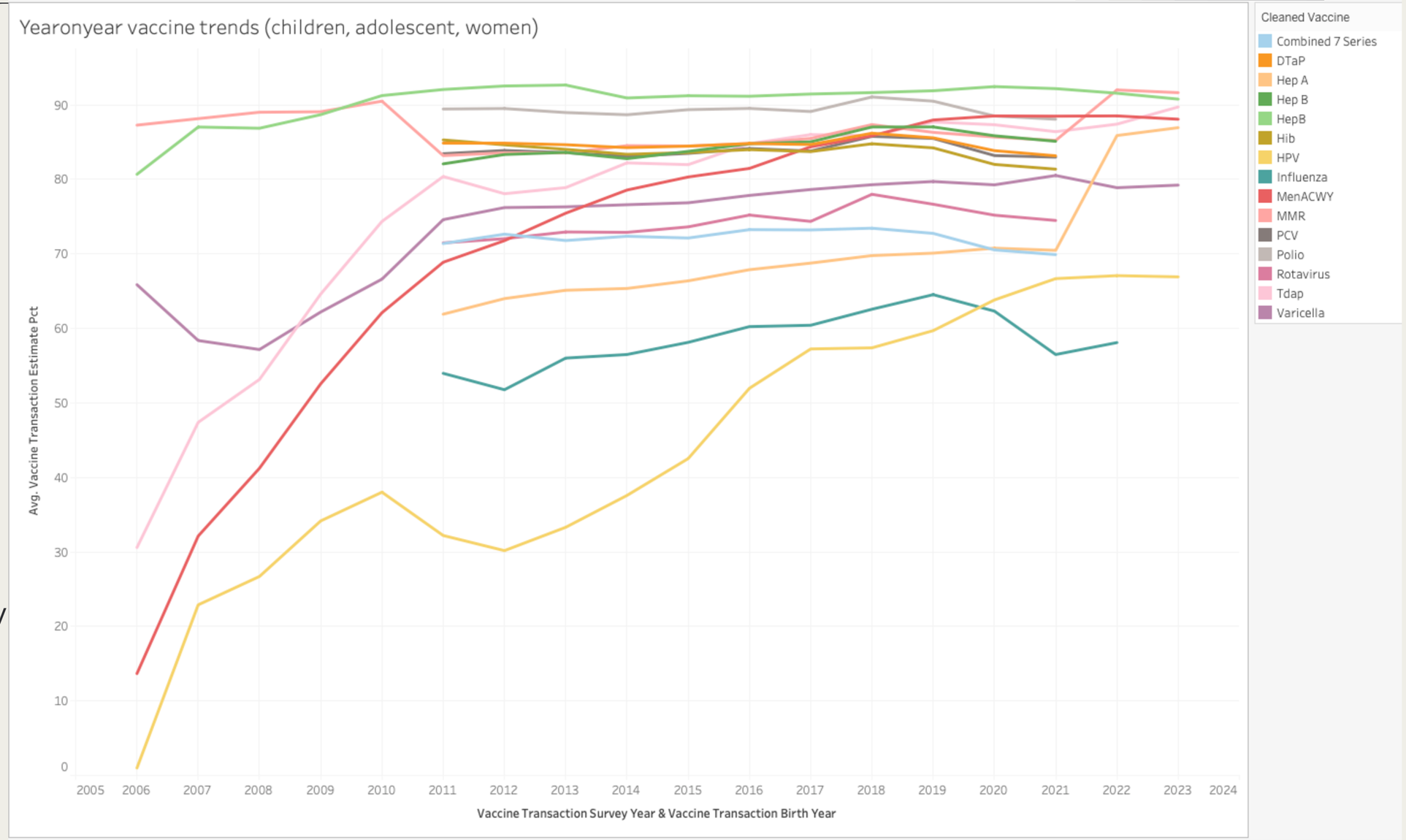
ETL: LOADING IN FACT TABLES (DELTA)

Creating influenza transaction fact tables

- Removed fields with missing measures: Estimated %, 95%CI, Sample Size
- Split Confidence Intervals to Upper and Lower Bounds
- Denormalise cohort year as Start Cohort Year and End Cohort Year
- There are fields where the estimates are unreliable: Setting as None, advised to discard
  - Sample Size <30
  - Standard Error >0.3
  - CI half-width >10
- Converted Sample Size to a Int64 - Sample Sizes represent actual humans, should be a whole number
- There is a specific vaccine for the 2009 H1N1 strain
  - Created and updated in vaccine dimension table
- Left merge on dimension tables
  - Vaccine: vaccine, year
  - Geography:FIPS
- Extracted out each dimension from the Dimension column and perform left merge on them
  - Vaccine Location

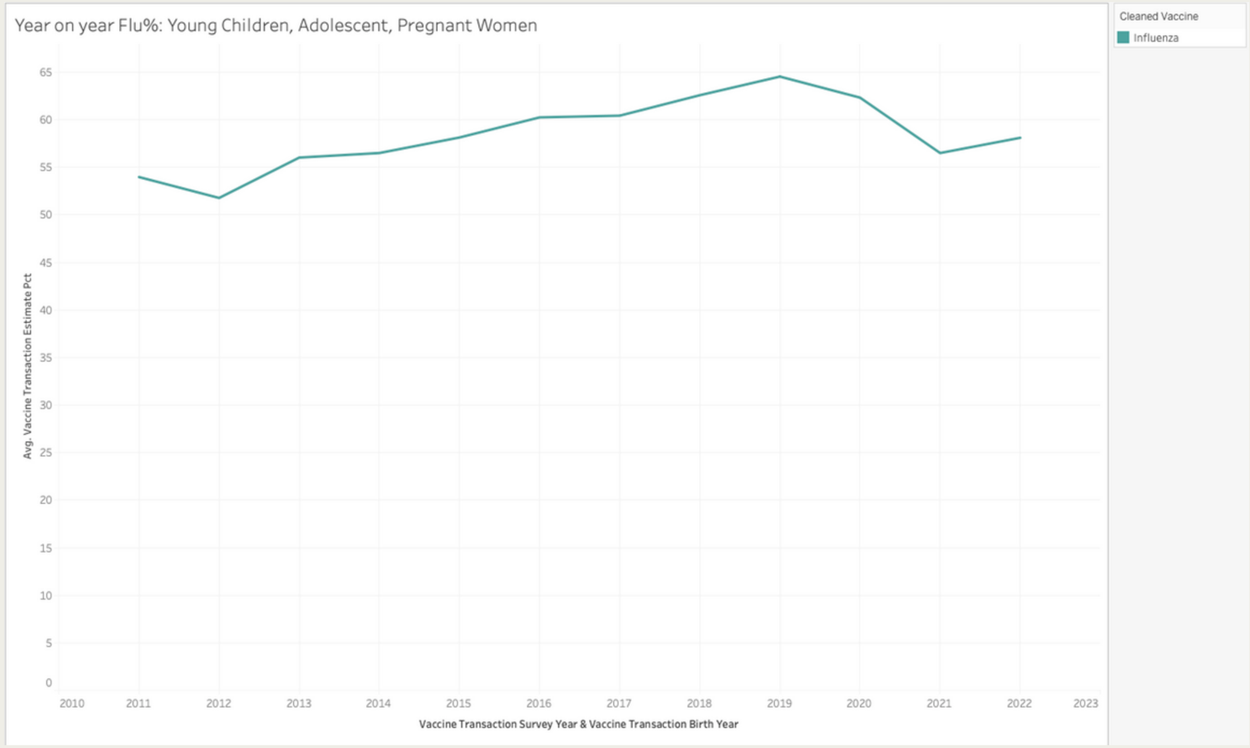
ANALYTICS & VISUALISATION

- **BQ: How have vaccination rates changed year-over-year across different age groups?**
  - Taking the average for a given year and vaccine
  - Combining both the Survey Year & Birth Year in Tableau
  - Children Dataset only starts from 2011



ANALYTICS & VISUALISATION

- Focusing on only influenza (flu) vaccine, there is an increase in vaccination rates in 2019
- The CDC launched a Flu Vaccine campaign in 2019 (source)



National Press Conference Kicks Off 2019-2020 Flu Vaccination Campaign

Español | Other Languages Print

On September 26, 2019, HHS Secretary Alex Azar, received a flu vaccine at the NFID 2019-2020 seasonal flu news conference. CDC Influenza Division Director Dan Jernigan, M.D., M.P.H., also attended the press conference. Dr. Jernigan took questions from the media and received his flu vaccine, emphasizing that now is the time for everyone 6 months and older to get their flu shot!

Thursday, September 26, 2019 — Today the Centers for Disease Control and Prevention (CDC) and the National Foundation for Infectious Diseases (NFID), along with other public health and medical groups, kicked off the 2019-2020 flu vaccine campaign at a press conference held at the National Press Club in Washington, D.C. Members of the public and health care professionals were urged to follow CDC's recommendation for everyone age six months and older to get an annual flu vaccine. Health and Human Services (HHS) Secretary Alex Azar II gave the keynote address. Secretary Azar reported CDC flu vaccination coverage estimates for 2018-2019 that showed 45% of Americans adults got a flu vaccine last season while nearly 63% of children were vaccinated.

ANALYTICS & VISUALISATION

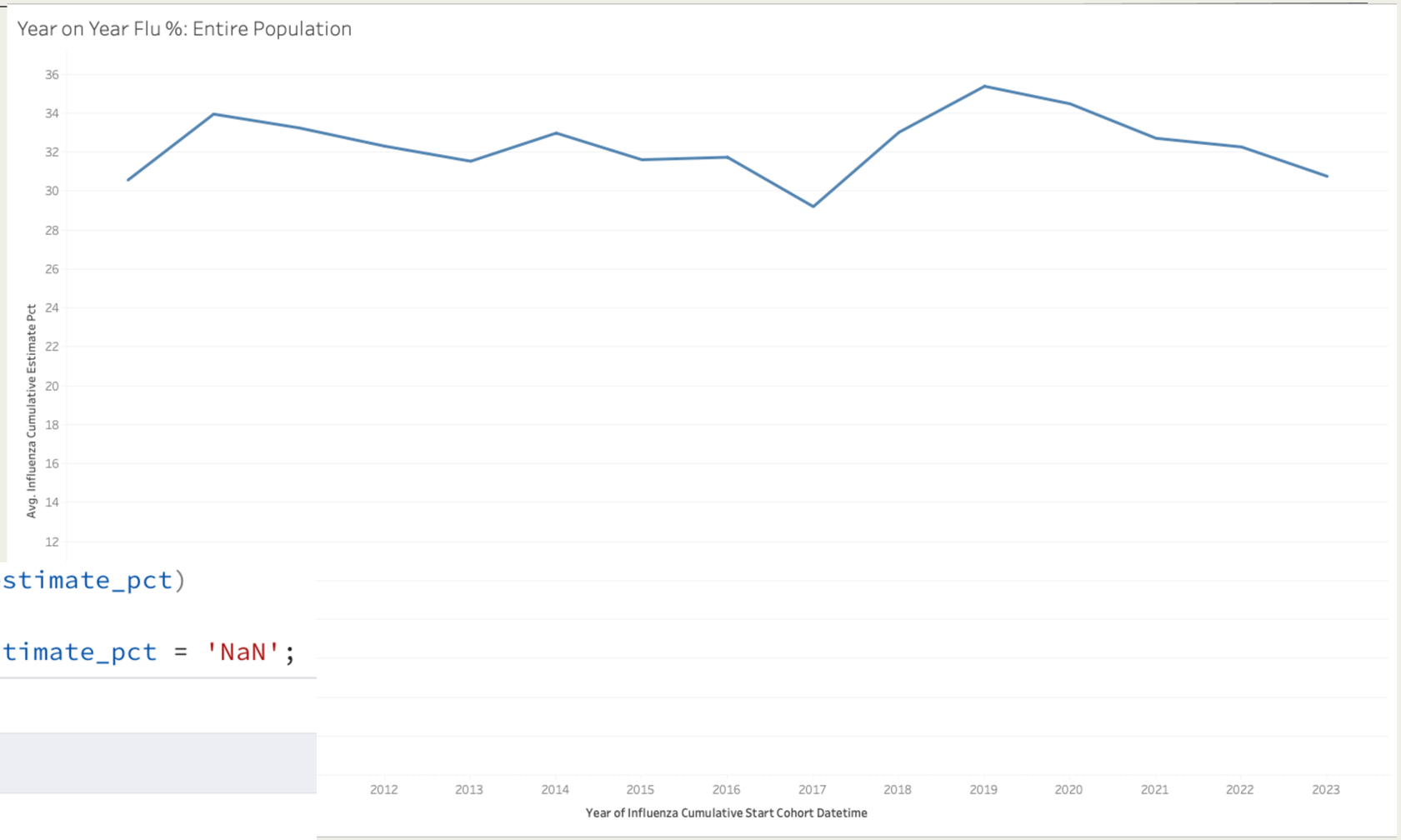
- Performing a ROLLUP on the overall vaccine (vaccine_cumulative_fact_overall is not null)
  - Average vaccine coverage for each vaccine per region over the years

```sql
-- getting the average vaccine coverage for each region
WITH aggregated_query AS (
  SELECT
    vaccine.cleaned_vaccine,
    vcf.vaccine_cumulative_end_cohort_year as year,
    vcf.vaccine_cumulative_estimate_pct AS estimated_vaccine_coverage,
    geography.region AS geography
  FROM vaccine_cumulative_fact vcf
  JOIN vaccine_dim vaccine ON vaccine.vaccine_id = vcf.vaccine_cumulative_vaccine_id
  JOIN geography_dim geography ON geography.geography_id = vcf.vaccine_cumulative_geography_id
  WHERE vaccine_cumulative_overall IS NOT NULL
)
SELECT
    year,
  geography,
  cleaned_vaccine,
  AVG(estimated_vaccine_coverage) AS avg_estimated_coverage
FROM aggregated_query
GROUP BY ROLLUP(geography, cleaned_vaccine, year)
ORDER BY geography, cleaned_vaccine, year
```

	year integer	geography character varying (20)	cleaned_vaccine character varying (50)	avg_estimated_coverage double precision
1	2014	Region 1	Combined 7 Series	77.24285714285713
2	2015	Region 1	Combined 7 Series	78.02857142857144
3	2016	Region 1	Combined 7 Series	79.12142857142858
4	2017	Region 1	Combined 7 Series	75.39659090909093
5	2018	Region 1	Combined 7 Series	81.90714285714286
6	2019	Region 1	Combined 7 Series	78.22921348314607
7	2020	Region 1	Combined 7 Series	81.04285714285713
8	2021	Region 1	Combined 7 Series	81.3
9	[null]	Region 1	Combined 7 Series	77.77126436781609
10	2014	Region 1	DTaP	89.83285714285714
11	2015	Region 1	DTaP	89.87000000000002
12	2016	Region 1	DTaP	89.3028571428571
13	2017	Region 1	DTaP	90.23807339449539
14	2018	Region 1	DTaP	90.37857142857142
15	2019	Region 1	DTaP	91.68227272727273
16	2020	Region 1	DTaP	90.31571428571431
17	2021	Region 1	DTaP	90.67142857142856
18	[null]	Region 1	DTaP	90.52214452214456
19	2022	Region 1	HPV	74.91623376623376
20	[null]	Region 1	HPV	74.91623376623376

ANALYTICS & VISUALISATION

- When we zoom out on the entire US population
- Flu vaccine rates are very low year-on-year
- Total cumulative average is 35.5%

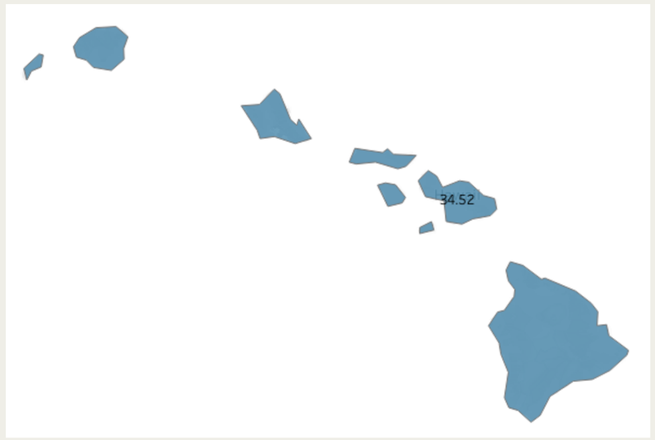


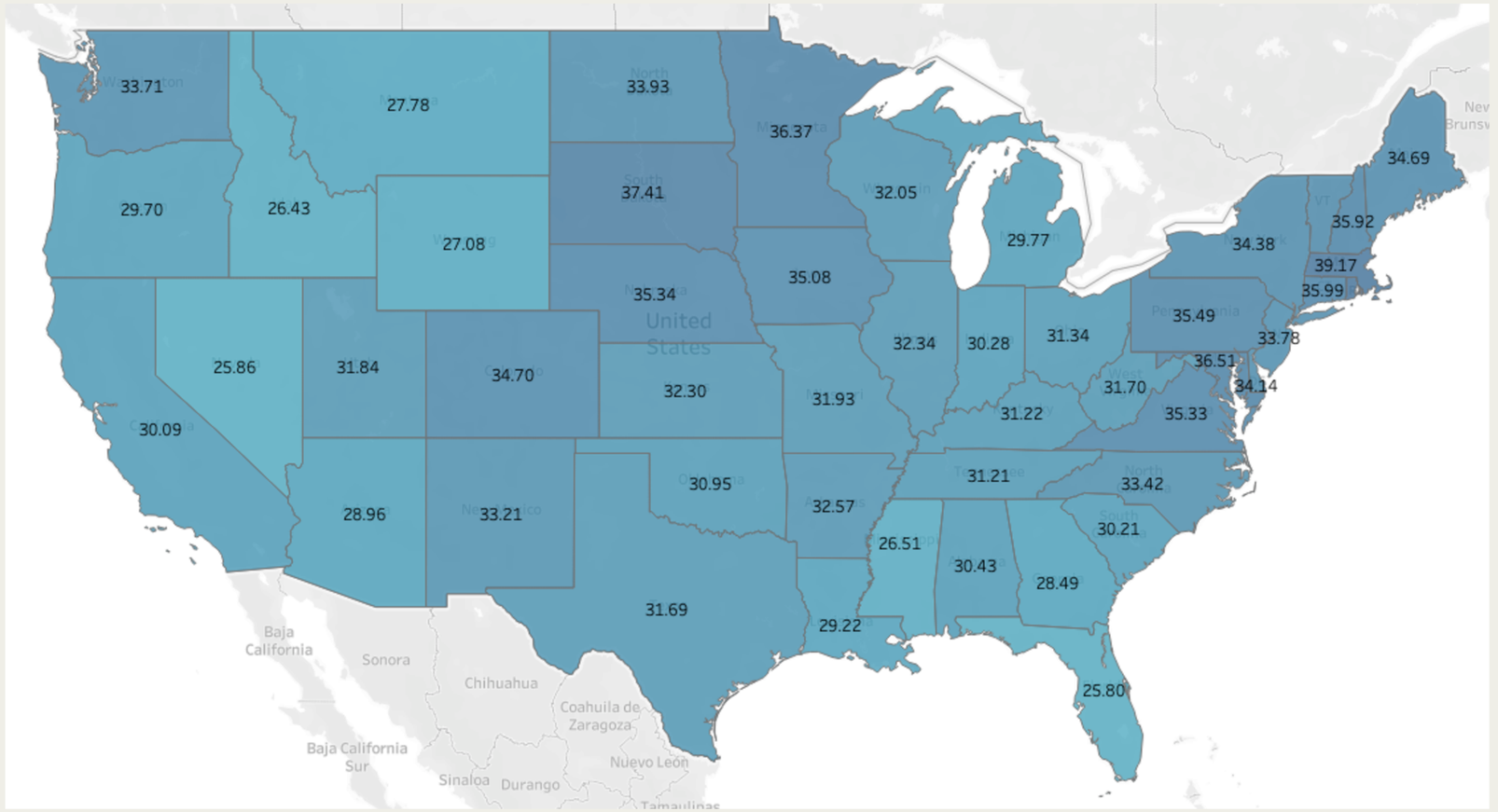

```
203  SELECT AVG(influenza_cumulative_estimate_pct)
204  FROM influenza_cumulative_fact
205  WHERE NOT influenza_cumulative_estimate_pct = 'NaN';
```

Data Output | Messages | Notifications

	avg double precision
1	35.51397322886093

Vaccination Rates by State





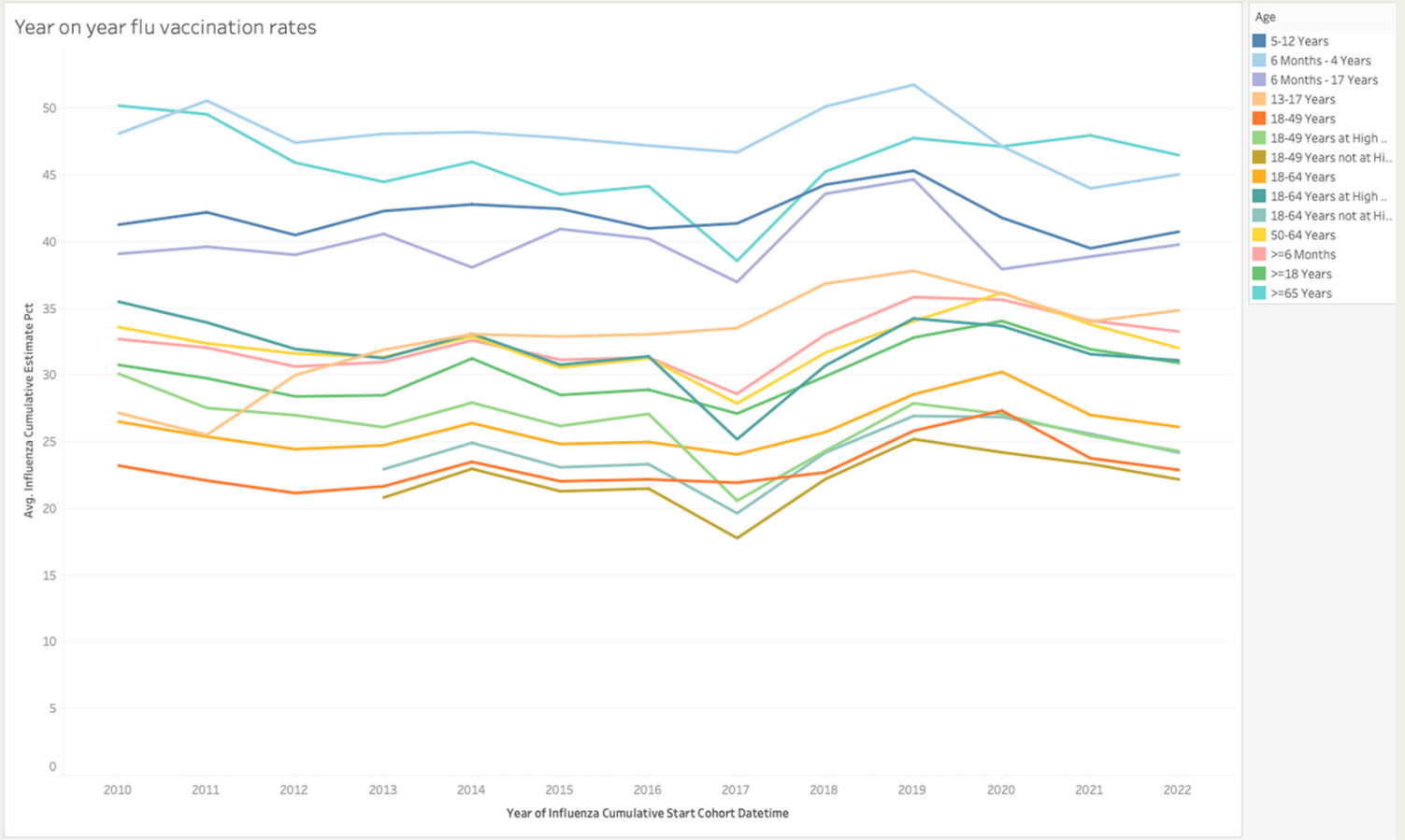
ANALYTICS & VISUALISATION

- Zooming out on the year on year for different age groups for the entire US population (Flu vaccine)

```
199  SELECT AVG(influenza_cumulative_estimate_pct), age,
200  rank() over(order by AVG(influenza_cumulative_estimate_pct)) as rank
201  FROM influenza_cumulative_fact
202  WHERE NOT influenza_cumulative_estimate_pct = 'NaN'
203  group by age
204  order by AVG(influenza_cumulative_estimate_pct);
```

Data Output | Messages | Notifications

	avg double precision	age character varying (100)	rank bigint
1	13.385656565656568	25-64 Years not in Initial Target Group	1
2	23.25627530364371	25-64 Years at High Risk	2
3	25.337107225822187	18-49 Years not at High Risk	3
4	25.622159090909065	18-49 Years	4
5	26.862331151650174	18-64 Years not at High Risk	5

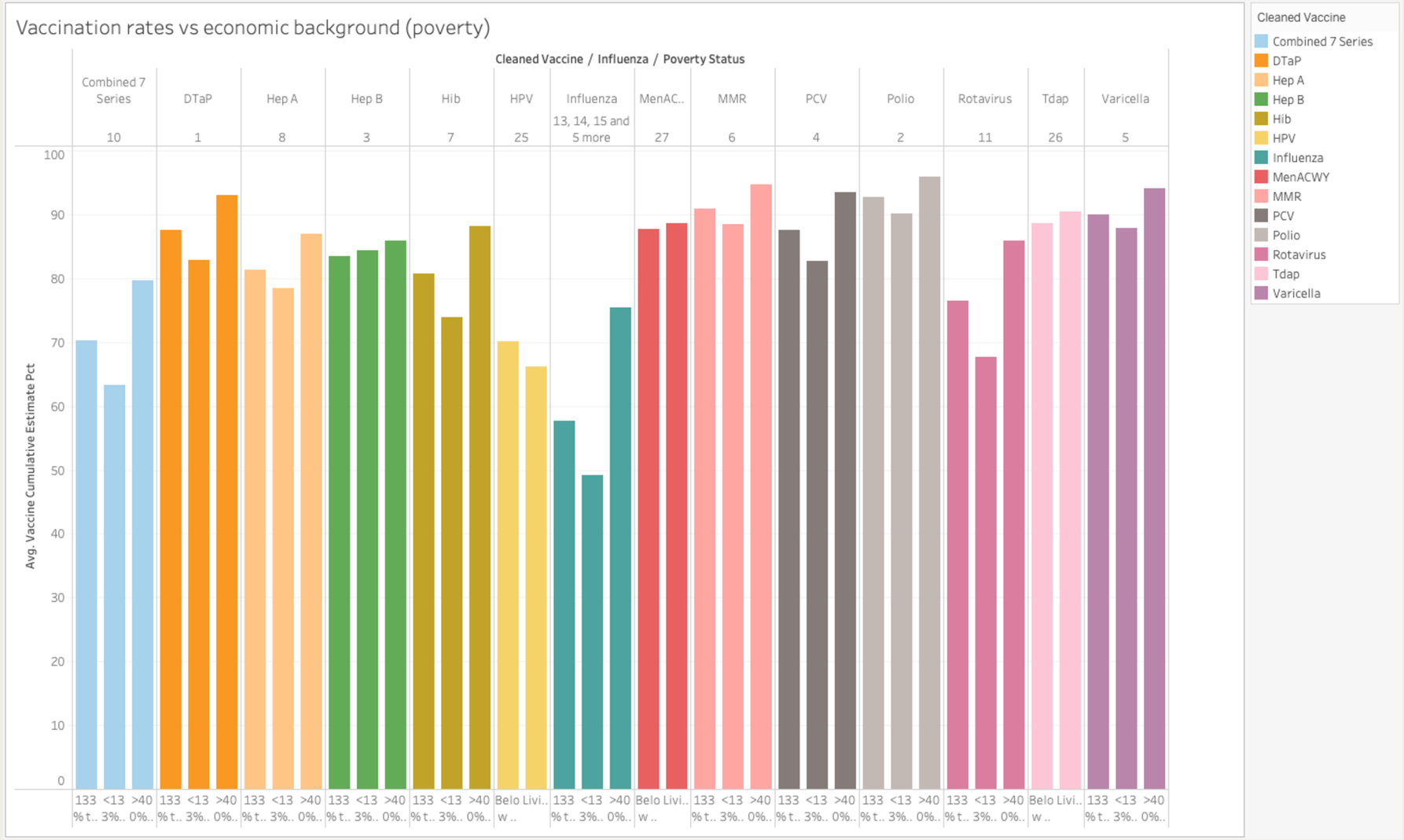


- The dip in 2017-2018 could be most likely due to ‘Vaccine Fatigue’ and the severity of the 2017-2018 flu season

ANALYTICS & VISUALISATION

- **BQ: What impact does economic background have on influenza vaccination rates?**

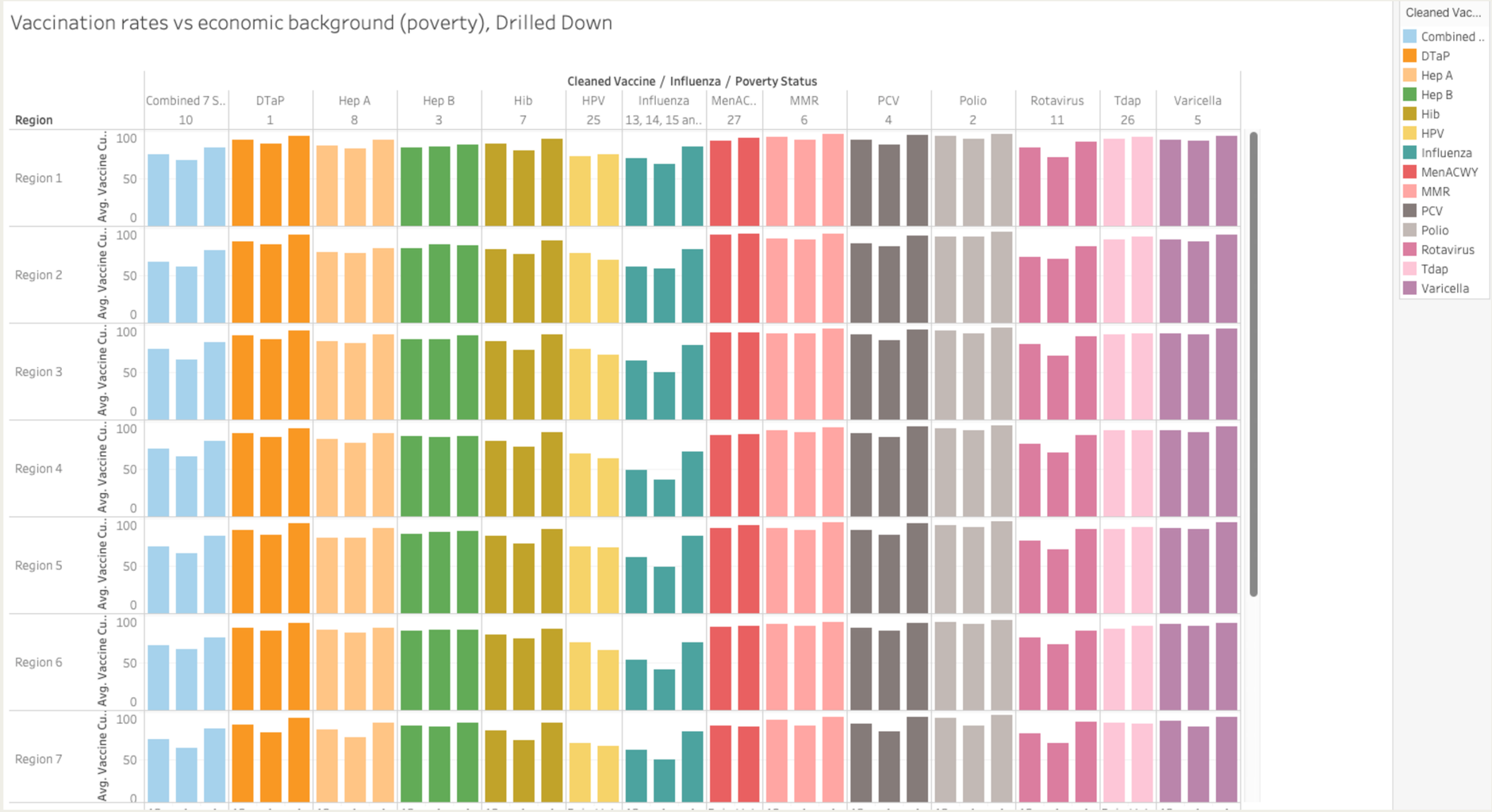
- The vaccination rates for children living below the poverty line (<133% FPL, Below Poverty Line) have lower vaccination rates than the ones above (>400% FPL)
- Those living around the poverty line <133%FPL<400% have lower vaccination rates for almost all vaccines (could be due to the ‘Coverage Gap’)
- Vaccination rates for school mandatory vaccines are higher than those not mandated
  - HepB, DTap/TDap, Polio, Varicella, MMR



ANALYTICS & VISUALISATION(DELTA)

- **BQ: What impact does economic background have on influenza vaccination rates?**

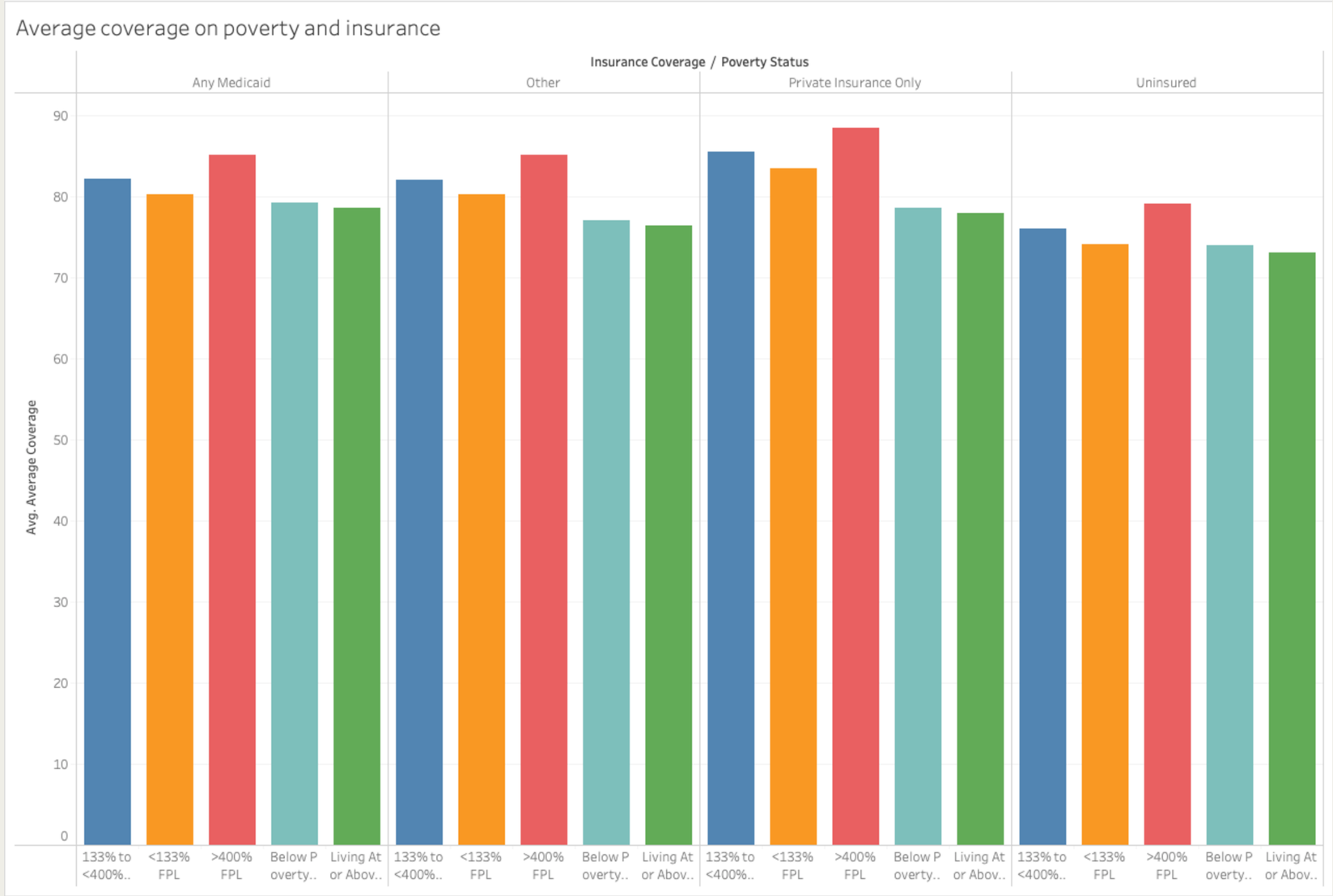
- Vaccination rates against poverty statuses, drilled down to region
- This highlights that at all regions, the vaccination rates for those below the poverty line is still above than those living around the poverty line



ANALYTICS & VISUALISATION

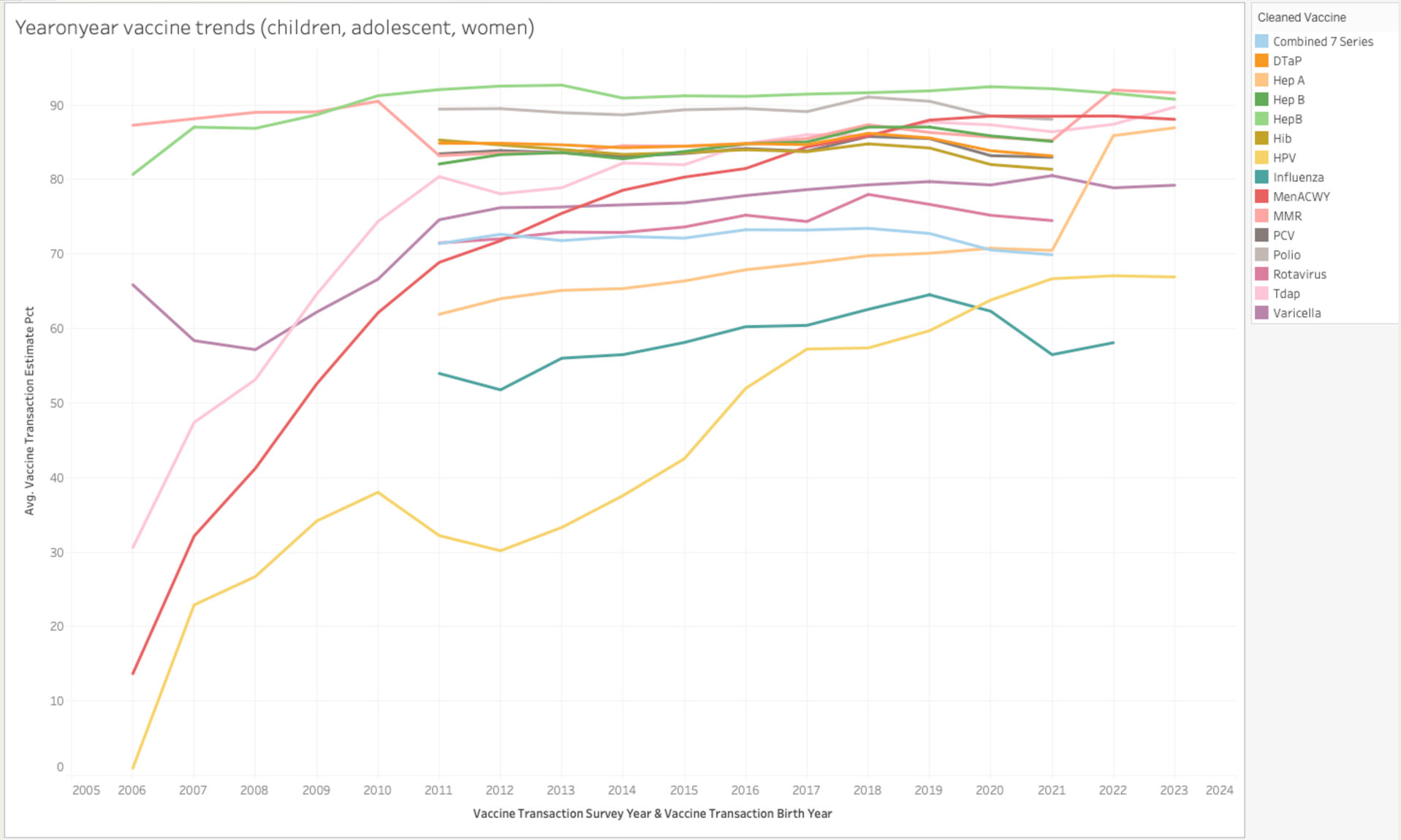
- **BQ: What impact does economic background have on influenza vaccination rates?**

- Combining both insurance and poverty as a view and taking the average vaccination rate for each vaccine, dose, and geography
- More people living around the poverty line are uninsured, leading to lower vaccination rates
    - Pay full price for a vaccine



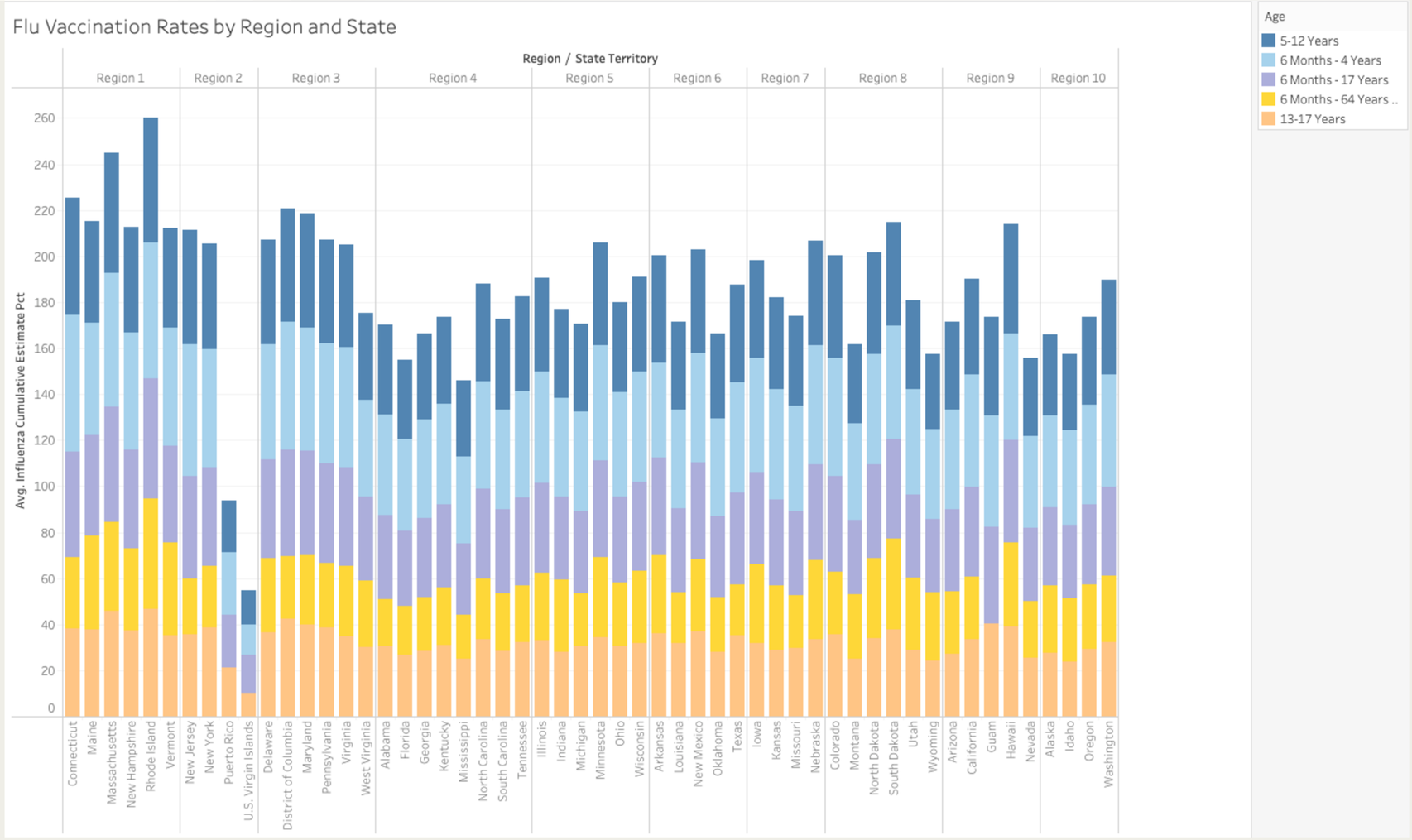
ANALYTICS & VISUALISATION

- **BQ: Is there a correlation between receiving the influenza vaccine and receiving other recommended vaccines?**
  - Looking at the graph, the flu vaccine rate is consistency lower than other vaccinations
  - We can infer that there may not be a correlation between getting the flu vaccine other vaccines



ANALYTICS & VISUALISATION

- **BQ: How does vaccination coverage within each demographic group influence future disease trends?**
  - **In the event of a flu outbreak, can we predict which demographic will be most affected based on vaccination coverage?**
- In the event of an outbreak <u>in school</u>, Puerto Rico and the Virgin Islands would most likely be affected
- In the mainland, Wyoming, Mississippi, Idaho would most likely be affected
- Mass is fine :)



ANALYTICS & VISUALISATION

- **BQ: If we were to launch a campaign to increase vaccination rates, which demographic should we prioritise?**
- Using Postgres for this, focusing on all vaccines for Children, Adolescent, and Pregnant Women

```sql
with aggregated_query as(
select vaccine.cleaned_vaccine,
dose.dose,
vtf.vaccine_transaction_estimate_pct as estimated_vaccine_coverage,
vtf.vaccine_transaction_age as age,
geography.original_geography as geography
from vaccine_transaction_fact vtf
join vaccine_dim vaccine on vaccine.vaccine_id = vtf.vaccine_transaction_vaccine_id
join dose_dim dose on dose.dose_id = vtf.vaccine_transaction_dose_id
join geography_dim geography on geography.geography_id = vtf.vaccine_transaction_geography_id
where geography.original_geography != 'United States' and dose.dose != 'History of disease'
and geography.original_geography not like 'Region%'
),
ranked_demographics as(
select cleaned_vaccine,
age,
geography,
estimated_vaccine_coverage,
dose,
rank() over(partition by cleaned_vaccine order by estimated_vaccine_coverage) as vaccine_rank
from aggregated_query)
select * from ranked_demographics where vaccine_rank <= 3

order by estimated_vaccine_coverage, vaccine_rank, geography
```

ANALYTICS & VISUALISATION

- **BQ: If we were to launch a campaign to increase vaccination rates, which demographic should we prioritise?**
- We should target the HPV first vaccine in Texas, Illinois, and DC

	cleaned_vaccine character varying (50)	age character varying (50)	geography character varying (50)	estimated_vaccine_coverage double precision	dose character varying (100)	vaccine_rank bigint
1	HPV	13-17 Years	Texas	2.4	≥3 Doses	1
2	HPV	13-17 Years	Illinois	3.9	≥2 Doses	2
3	HPV	13-17 Years	District of Columbia	4.8	≥3 Doses	3
4	HPV	13-15 Years	TX-Rest of state	4.8	≥3 Doses	3
5	Tdap	[null]	Delaware	5.2	1 Dose Only	1
6	Tdap	[null]	Delaware	5.2	1 Dose Only	1
7	Hib	7 Months	Alaska	6.3	≥3 Doses	1
8	Hib	7 Months	Rhode Island	6.6	≥3 Doses	2
9	Hib	7 Months	Alaska	7	≥3 Doses	3
10	Tdap	18-24 Years	Delaware	7.7	1 Dose Only	3
11	Tdap	[null]	Pennsylvania	7.7	1 Dose Only	3
12	Tdap	[null]	Pennsylvania	7.7	1 Dose Only	3
13	MenACWY	13-15 Years	South Dakota	9.6	≥1 Dose	1
14	Hep A	19 Months	Wyoming	9.7	≥2 Doses	1
15	Hep A	19 Months	Alaska	9.8	≥2 Doses	2
16	Hep A	19 Months	Puerto Rico	10.5	≥2 Doses	3

ANALYTICS & VISUALISATION

- **BQ: If we were to launch a campaign to increase vaccination rates, which demographic should we prioritise?**
  - **For each state, which age group should we focus on?**
- Focusing on flu vaccines for each state, I performed a GROUP BY ROLLUP() to get the aggregated vaccine percentage by age and by state

```sql
WITH aggregated_query AS (
  SELECT
    vaccine.cleaned_vaccine,
    icf.age AS age,
    geography.state_territory AS geography,
    icf.influenza_cumulative_estimate_pct AS estimated_vaccine_coverage
  FROM influenza_cumulative_fact icf
  JOIN vaccine_dim vaccine
    ON vaccine.vaccine_id = icf.influenza_cumulative_vaccine_id
  JOIN geography_dim geography
    ON geography.geography_id = icf.influenza_cumulative_geography_id
  WHERE
    icf.influenza_cumulative_estimate_pct != 'NaN'
    AND vaccine.cleaned_vaccine = 'Influenza'
)

  SELECT
    cleaned_vaccine,
    age,
    geography,
    AVG(estimated_vaccine_coverage) AS average_coverage
  FROM aggregated_query
  GROUP BY rollup(geography, cleaned_vaccine, age)
  ORDER BY
  geography NULLS LAST,
  cleaned_vaccine NULLS LAST,
  age NULLS LAST
```

	cleaned_vaccine character varying (50)	age character varying (100)	geography character varying (50)	average_coverage double precision
1	Influenza	13-17 Years	Alabama	34.51343283582091
2	Influenza	18-49 Years	Alabama	23.704255319148942
3	Influenza	18-49 Years at High Risk	Alabama	27.7123076923077
4	Influenza	18-49 Years not at High Ri...	Alabama	22.621238938053093
5	Influenza	18-64 Years	Alabama	26.1456375838926
6	Influenza	18-64 Years at High Risk	Alabama	33.86434108527133
7	Influenza	18-64 Years not at High Ri...	Alabama	24.237962962962957
8	Influenza	5-12 Years	Alabama	41.57902097902098
9	Influenza	50-64 Years	Alabama	35.170860927152326
10	Influenza	6 Months - 17 Years	Alabama	37.47894736842103
11	Influenza	6 Months - 4 Years	Alabama	47.413669064748184
12	Influenza	>=18 Years	Alabama	31.17532467532468
13	Influenza	>=6 Months	Alabama	31.133333333333347
14	Influenza	>=65 Years	Alabama	50.31020408163264
15	Influenza	Greater 65	Alabama	45.239999999999995
16	Influenza	Greater than 18 Years flu	Alabama	25.354545454545452
17	Influenza	Greater than 6 Months flu	Alabama	26.527272727272724
18	Influenza	NaN	Alabama	34.419831932773114
19	[null]	[null]	Alabama	33.887587006960516
20	Influenza	[null]	Alabama	33.887587006960516

ANALYTICS & VISUALISATION

- **BQ: If we were to launch a campaign to increase vaccination rates, which demographic should we prioritise?**
- After the ROLLUP(), I did a rank and listed out the top 3 age groups by state

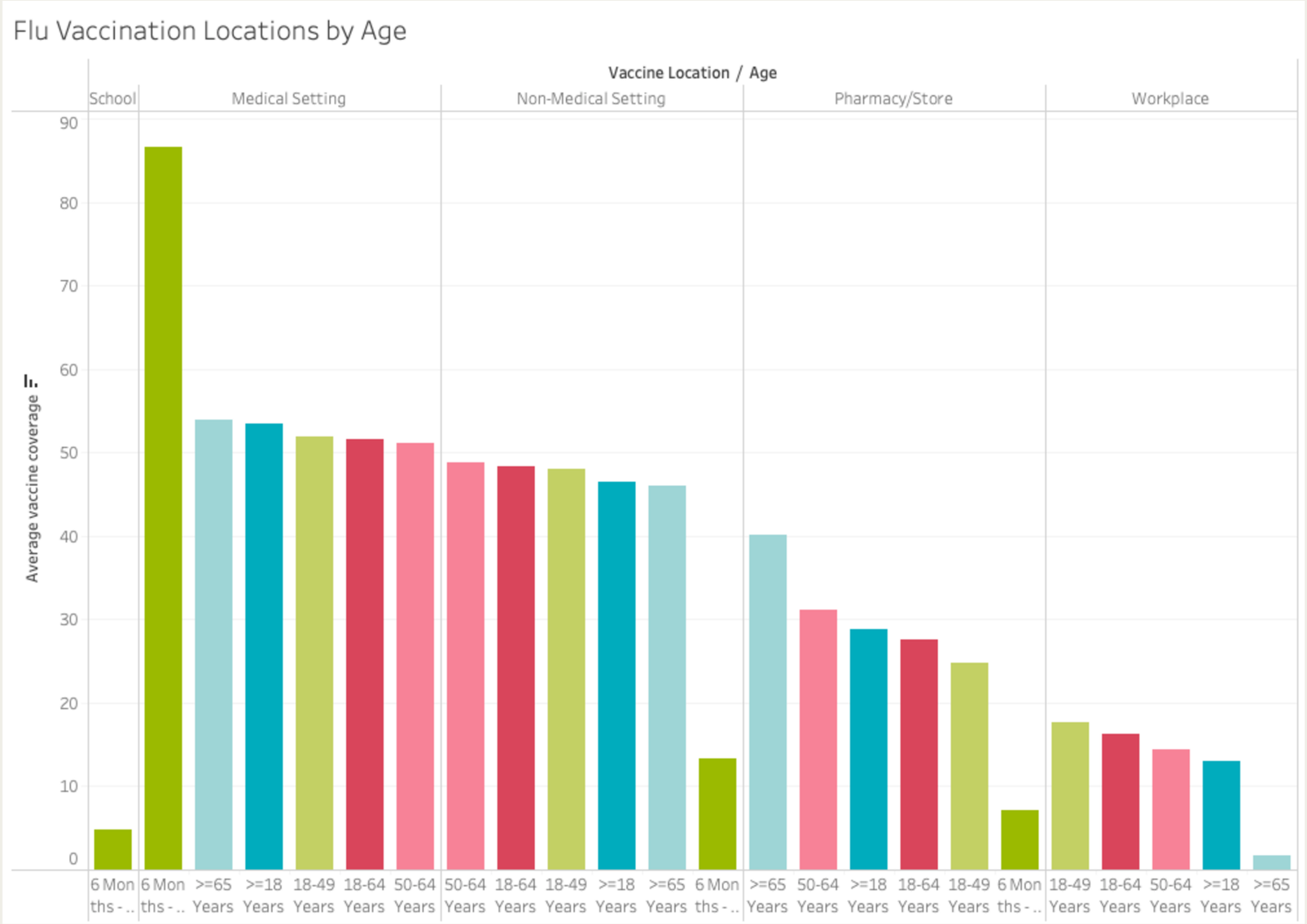
```
WITH aggregated_query AS (
  SELECT
    vaccine.cleaned_vaccine,
    icf.age AS age,
    geography.state_territory AS geography,
    icf.influenza_cumulative_estimate_pct AS estimated_vaccine_coverage
  FROM influenza_cumulative_fact icf
  JOIN vaccine_dim vaccine
    ON vaccine.vaccine_id = icf.influenza_cumulative_vaccine_id
  JOIN geography_dim geography
    ON geography.geography_id = icf.influenza_cumulative_geography_id
  WHERE
    icf.influenza_cumulative_estimate_pct != 'NaN'
    AND vaccine.cleaned_vaccine = 'Influenza'
),
rolledup_query AS (
  SELECT
    cleaned_vaccine,
    age,
    geography,
    AVG(estimated_vaccine_coverage) AS average_coverage
  FROM aggregated_query
  GROUP BY rollup(geography, cleaned_vaccine, age)
  HAVING geography IS NOT NULL AND cleaned_vaccine IS NOT NULL AND age IS NOT NULL
),
ranked_query AS (
  SELECT *,
         RANK() OVER (PARTITION BY geography ORDER BY average_coverage) AS rank
  FROM rolledup_query
)
SELECT *
FROM ranked_query
WHERE rank <= 3
ORDER BY average_coverage, geography;
```

	cleaned_vaccine character varying (50)	age character varying (100)	geography character varying (50)	average_coverage double precision	rank bigint
1	Influenza	18-49 Years at High Risk	U.S. Virgin Islands	0	1
2	Influenza	50-64 Years	U.S. Virgin Islands	8.018181818181818	2
3	Influenza	18-49 Years not at High Ri...	U.S. Virgin Islands	9.422222222222222	3
4	Influenza	18-49 Years	Puerto Rico	13.270769230769238	1
5	Influenza	18-49 Years not at High Ri...	Puerto Rico	13.685714285714292	2
6	Influenza	18-64 Years not at High Ri...	Puerto Rico	13.787692307692302	3
7	Influenza	18-49 Years not at High Ri...	Florida	17.61637931034482	1
8	Influenza	18-49 Years	Florida	17.74859154929578	2
9	Influenza	18-49 Years not at High Ri...	Nevada	18.96944444444444	1
10	Influenza	18-64 Years not at High Ri...	Florida	19.170370370370367	3
11	Influenza	18-49 Years not at High Ri...	Arizona	20.095689655172414	1
12	Influenza	18-49 Years	Nevada	20.52611940298508	2

- After the ROLLUP(), I did a rank and listed out the top 3 age groups by state
- We need to prioritise the Virgin Islands and Puerto Rico
- Mainland, we need to prioritise Floridia for 18-49 Years not at High Risk

ANALYTICS & VISUALISATION

- **BQ: If we were to launch a campaign to increase vaccination rates, which demographic should we prioritise?**
- On location, we need to target the workplace
  - Makes sense for targeting 18-49 Year Olds
- Not many people received the vaccine at their workplace



CONCLUSION

- Loaded 4 Vaccination Datasets into 3 tables:
  - 2 Cumulative, 1 Transaction
- 10 Dimension Tables
  - 6 SCD0
  - 1 SCD1
  - 2 SCD2
  - 1 SCD3
- What was done is just a small scale warehouse focusing on specific demographics and vaccines
- In reality, the Data Warehouse for the CDC is much MUCH bigger than this

FUTURE WORK

- API Key for FIPS code recently obtained
  - Work on getting the transaction data for Influenza Vaccine Coverage for All Ages
  - Map FIPS code to state
- Clean up code
  - Code is all over the place
  - Modularise
- Architecture diagram in AWS?

Thank you!