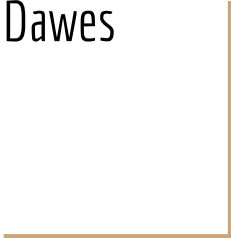




the-data-miners

Isabella Joseph, Rebecca Dawes



Introduction to The Behavioral Risk Factor Surveillance System (BRFSS)

- Yearly telephone survey by the Centers for Disease Control
- Measures behavioral and personal risk factors for disease
- 400,000 telephone interviews yearly with hundreds of questions
- Publicly accessible, anonymized public health data set

Research question: Does smoking status have a differing association with general mental and physical health differently in Texas and Massachusetts, states with different average educational attainment and average income?

erd_unified

Ernesto Duran, 1 April 2020

Texas

DEMOGRAPHIC_Beam_DF		
PK, FK	SEQNO	INTEGER
	_STATE	INTEGER
	SEX	INTEGER
	MARITAL	INTEGER
	EDUCA	INTEGER
	EMPLOY	INTEGER
	CHILDREN	INTEGER
	INCOME2	INTEGER
	WEIGHT2	INTEGER
	HEIGHT3	INTEGER

OVERALL_Health_Status_Beam_DF		
PK, FK	SEQNO	INTEGER
	_STATE	INTEGER
	GENHLTH	INTEGER
	PHYSHLTH	INTEGER
	MENTHLTH	INTEGER

TOCACCO_USE_Beam_DF		
PK, FK	SEQNO	INTEGER
	_STATE	INTEGER
	SMOKE100	INTEGER
	SMOKDAY2	INTEGER
	LASTSMK2	INTEGER
	USENOW3	INTEGER

Massachusetts

MASSACHUSETTS_DEMOGRAPHIC_Beam_DF		
PK, FK	SEQNO	INTEGER
	_STATE	INTEGER
	SEX	INTEGER
	MARITAL	INTEGER
	EDUCA	INTEGER
	EMPLOY	INTEGER
	CHILDREN	INTEGER
	INCOME2	INTEGER
	WEIGHT2	INTEGER
	HEIGHT3	INTEGER

MASSACHUSETTS_OVERALL_Health_Status_Beam_DF		
PK, FK	SEQNO	INTEGER
	_STATE	INTEGER
	GENHLTH	INTEGER
	PHYSHLTH	INTEGER
	MENTHLTH	INTEGER

MASSACHUSETTS_TOCACCO_USE_Beam_DF		
PK, FK	SEQNO	INTEGER
	_STATE	INTEGER
	SMOKE100	INTEGER
	SMOKDAY2	INTEGER
	LASTSMK2	INTEGER
	USENOW3	INTEGER

erd_unified

cdc_modeled

```
%%bigquery
CREATE TABLE cdc_modeled.DEMOGRAPHIC_2011 AS
SELECT * FROM(
SELECT distinct SEQNO, _STATE, SEX, MARITAL, EDUCA, EMPLOY, CHILDREN, INCOME2, WEIGHT2, HEIGHT3 from cdc_s
taging.2011
WHERE _STATE = 48
)
```

```
%%bigquery
CREATE TABLE cdc_modeled.DEMOGRAPHIC_2012 AS
SELECT * FROM(
SELECT distinct SEQNO, _STATE, SEX, MARITAL, EDUCA, EMPLOY, CHILDREN, INCOME2, WEIGHT2, HEIGHT3 from cdc_s
taging.2012
WHERE _STATE = 48
)
```

```
%%bigquery
CREATE TABLE cdc_modeled.OVERALL_Health_Status_2013 AS
SELECT * FROM(
SELECT SEQNO, _STATE, GENHLTH, PHYSHLTH, MENTHLTH from cdc_staging.2013
WHERE _STATE=48
)
```

```
%%bigquery
CREATE TABLE cdc_modeled.DEMOGRAPHIC_float AS
SELECT * FROM(
SELECT SEQNO, _STATE, SEX, MARITAL, EDUCA, EMPLOY, CHILDREN, INCOME2, WEIGHT2, HEIGHT3
FROM cdc_modeled.DEMOGRAPHIC_2011
UNION ALL
SELECT SEQNO, _STATE, SEX, MARITAL, EDUCA, EMPLOY, CHILDREN, INCOME2, WEIGHT2, HEIGHT3
FROM cdc_modeled.DEMOGRAPHIC_2012
UNION ALL
SELECT SEQNO, _STATE, SEX, MARITAL, EDUCA, EMPLOY1, CHILDREN, INCOME2, WEIGHT2, HEIGHT3
FROM cdc_modeled.DEMOGRAPHIC_2013)
```

```
%%bigquery
CREATE TABLE cdc_modeled.DEMOGRAPHIC AS
SELECT CAST(SEQNO AS INT64) SEQNO,
CAST(_STATE AS INT64) _STATE,
CAST(SEX AS INT64) SEX,
CAST (MARITAL AS INT64) MARITAL,
CAST (EDUCA AS INT64) EDUCA,
CAST (EMPLOY AS INT64) EMPLOY,
CAST (CHILDREN AS INT64) CHILDREN,
CAST (INCOME2 AS INT64) INCOME2,
CAST (WEIGHT2 AS INT64) WEIGHT2,
CAST (HEIGHT3 AS INT64) HEIGHT3
FROM cdc_modeled.DEMOGRAPHIC_float
```

%run DEMOGRAPHIC_beam.py

```
WARNING:apache_beam.runne
WARNING:apache_beam.runne
PCollection visualization
necessary dependencies to
/home/jupyter/venv/lib/py
g: options is deprecated
```

%run DEMOGRAPHIC_beam.py

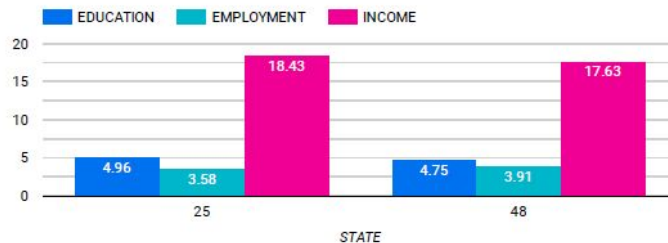
```
query_results PCollection[Re
/home/jupyter/venv/lib/pythc
g: options is deprecated sir
d
experiments = p.options.vi
INFO:apache beam.runners.dir
```

Beam Pipelines

Data cleansing tasks:

1. Replace “none” with 0
2. Remove outliers
3. Remove special codes
4. Unit standardization for height and weight

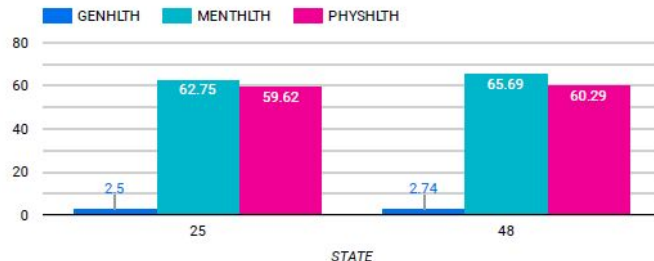
Average_demographics_in_TEXAS_and_MASSACHUSETTS



This column chart looks at the average Education, Employment and Income in the states of Texas (STATE = 48) and Massachusetts (STATE = 25).

Education and Income rates are slightly higher Massachusetts, however Employment rate is higher in Texas.

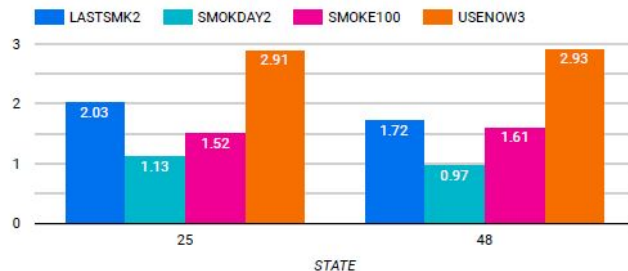
Average_Overall_Health_in_TEXAS_and_MASSACHUSETTS



This column chart looks at the average General Health, Mental Health and Physical Health in the states of Texas (STATE = 48) and Massachusetts (STATE = 25).

It is clear that Texas has a better overall health status than Massachusetts. Texas is higher than Massachusetts based on all the factors of General, Mental and Physical health.

Average_Tobacco_Use_in_TEXAS_and_MASSACHUSETTS



This column chart looks at the average Tobacco use in the states of Texas (STATE = 48) and Massachusetts (STATE = 25).

The graphs average values are unable to give a clear trend on whether Texas or Massachusetts has worse smoking rates. Texas ranks higher in terms of LASTSMK2 and SMOKEDAY2, however, Massachusetts has higher SMOKE100 and USENOW3 rates.

Combining Datasets and Cross-Dataset Queries

Demo

<https://121e03b141109d7d-dot-us-central1.notebooks.googleusercontent.com/lab/workspaces/auto-Y?authuser=0>

Workflow

Functionalities:

- Explicitly state functionality of cdc_staging and cdc_modeled to create staging and modeled datasets and to load in data
- Create and update tables after cleansing through referenced beam.py pipelines

<https://121e03b141109d7d-dot-us-central1.notebooks.googleusercontent.com/lab?authuser=0>

Future Improvements

- Expand number of responses and tables
- Conduct analysis for more years
- Expand to different states
- Look at more health related trends from the same datasets

Citations

1. Centers for Disease Control and Prevention. The Behavioral Risk Factor Surveillance System. Kaggle. Accessed: January 31, 2020.

<https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system/version/1>

2. 2018: ACS 1-Year Estimates Subject Tables - Educational Attainment. United States Census. Accessed May 7th.

<https://data.census.gov/cedsci/table?q=Educational%20Attainment&hidePreview=true&tid=ACST1Y2018.S1501&t=Education%20Attainment&vintage=2018&q=0400000US48.25>