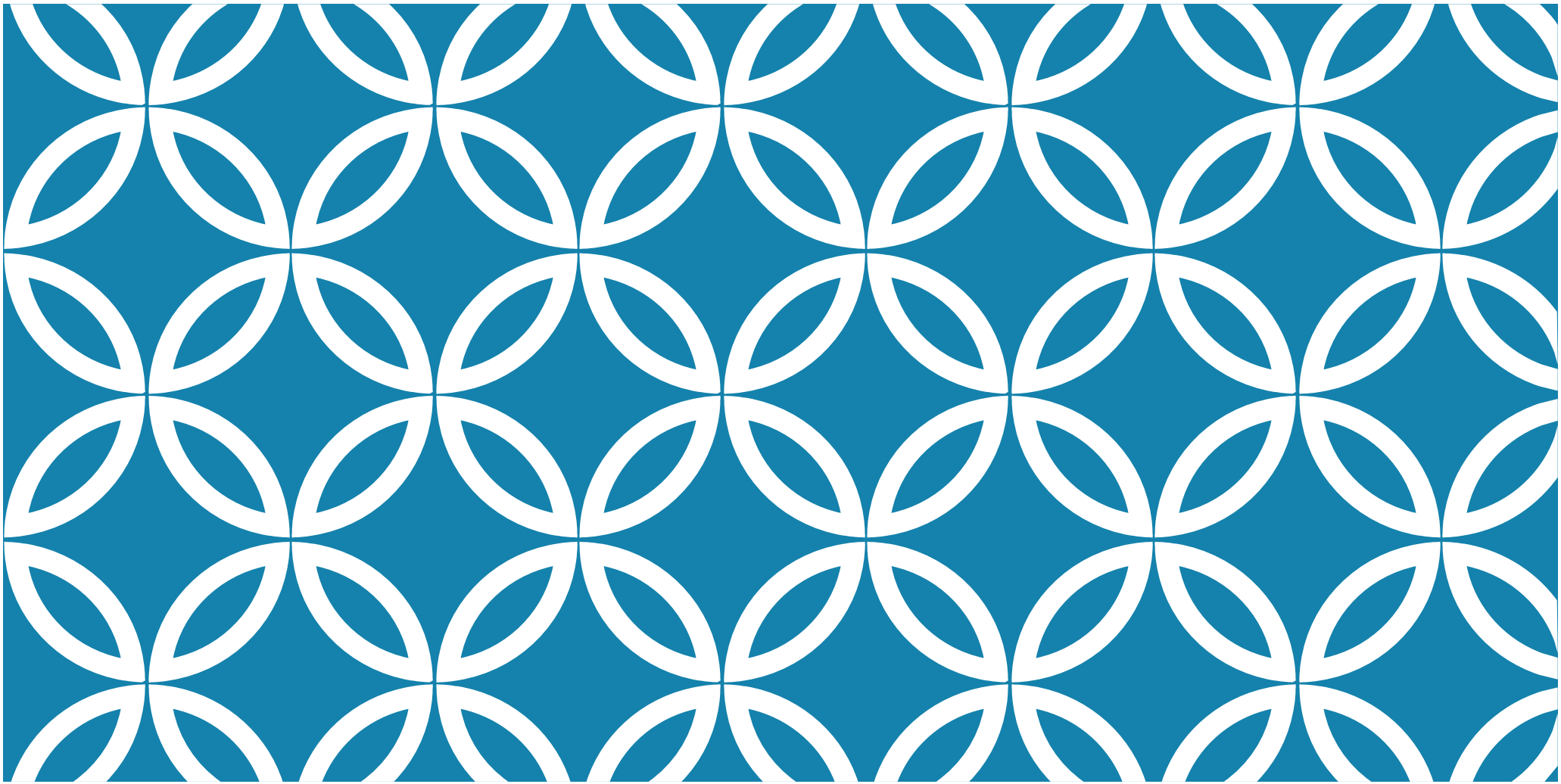
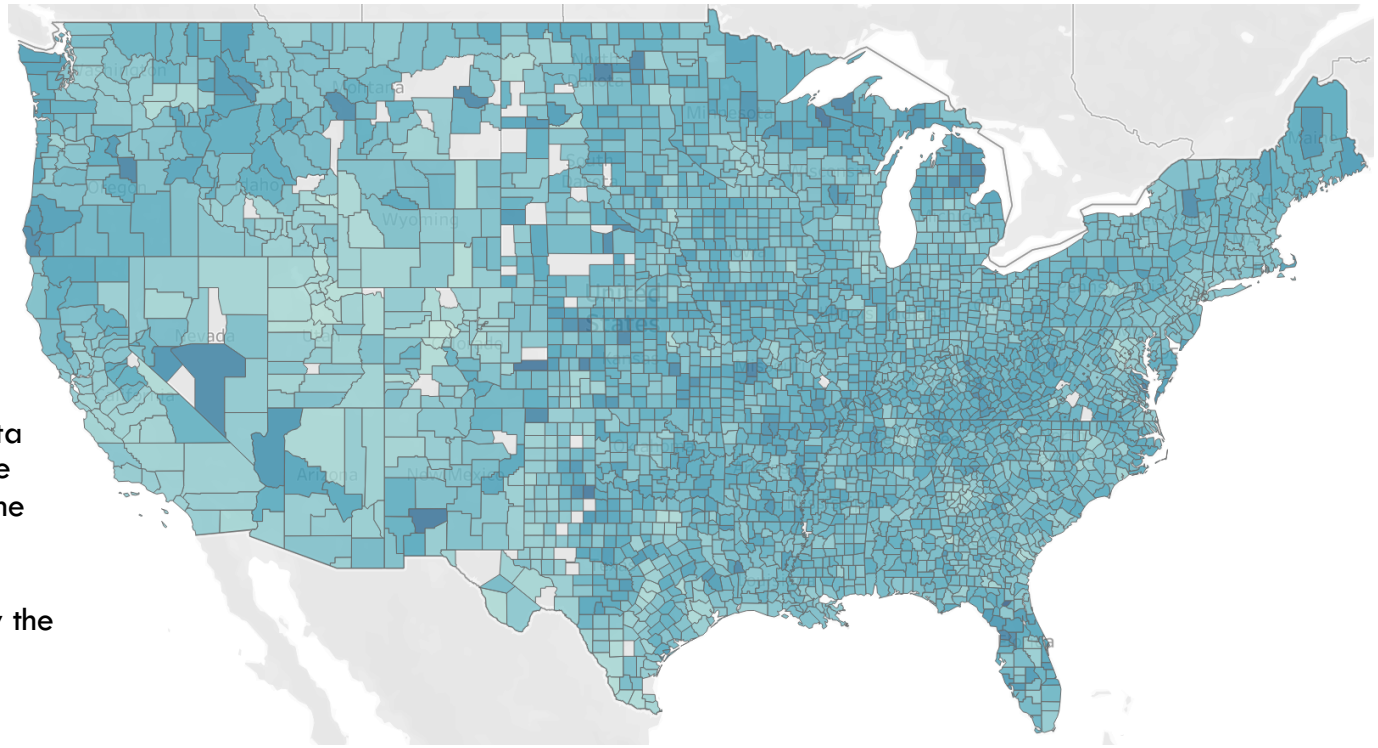


CANCER EPIDEMIOLOGY DATABASE

MSCA 31012 DATA ENGINEERING PLATFORMS

Friday Evening Group 2:
Aamer Hussain
Niharika Tyagi
Paul Whitson

EXECUTIVE SUMMARY



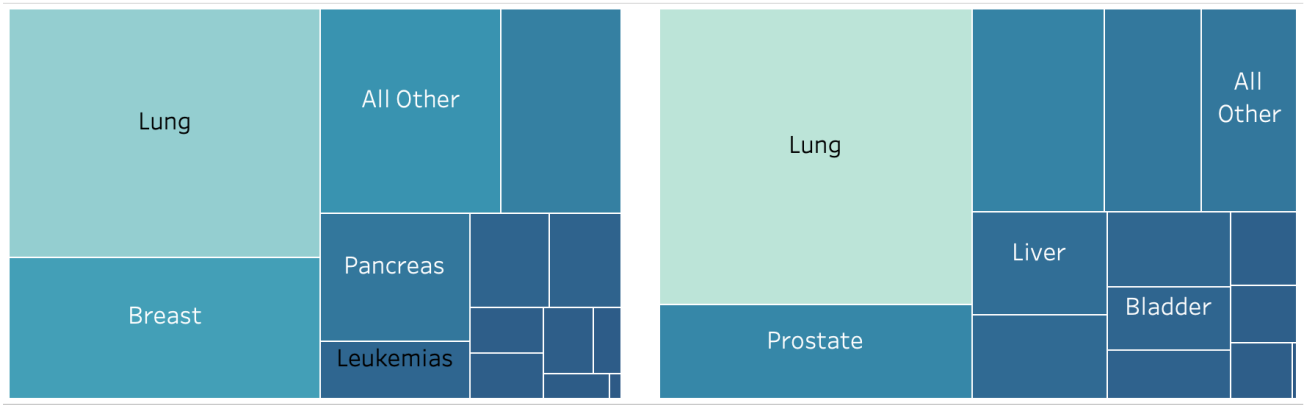
This project collated and analyzed data related to the rates of cancer incidence (new cases) and mortality (deaths) in the United States.

Visualizations were developed to show the relationships among these rates and

- Cancer Type
- Geography (state and county)
- Demographic Factors (gender / race)
- Selected other factors (income level, smoking, physical activity)

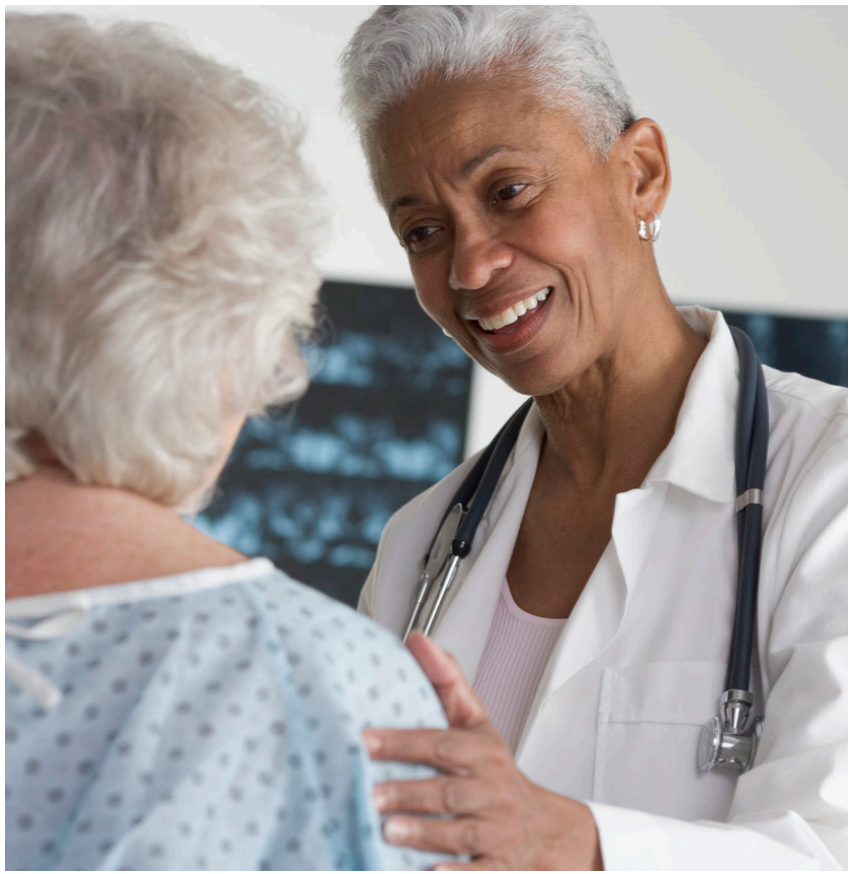
These analyses can be used by healthcare businesses, governments, and non-profits to support decisions on allocation of resources for cancer prevention and treatment

In high-income countries the public policy consensus is that costs of delivering high-quality equitable cancer care present an increasing challenge to national budgets. In the U.S. alone it is estimated cancer care expenditures in 2020 will be 157 billion dollars. The increase is being driven by a number of factors including technological innovation, rising costs of medical and hospital care, expensive therapeutics and an increase in the proportion of individuals susceptible to malignancy as the population ages. In this



AGENDA

1. Business Case
2. Data Pipeline & Database Design
3. ETL / Data Cleaning
4. Data Visualization and Analysis
5. Conclusions and Next Steps



BUSINESS CASE

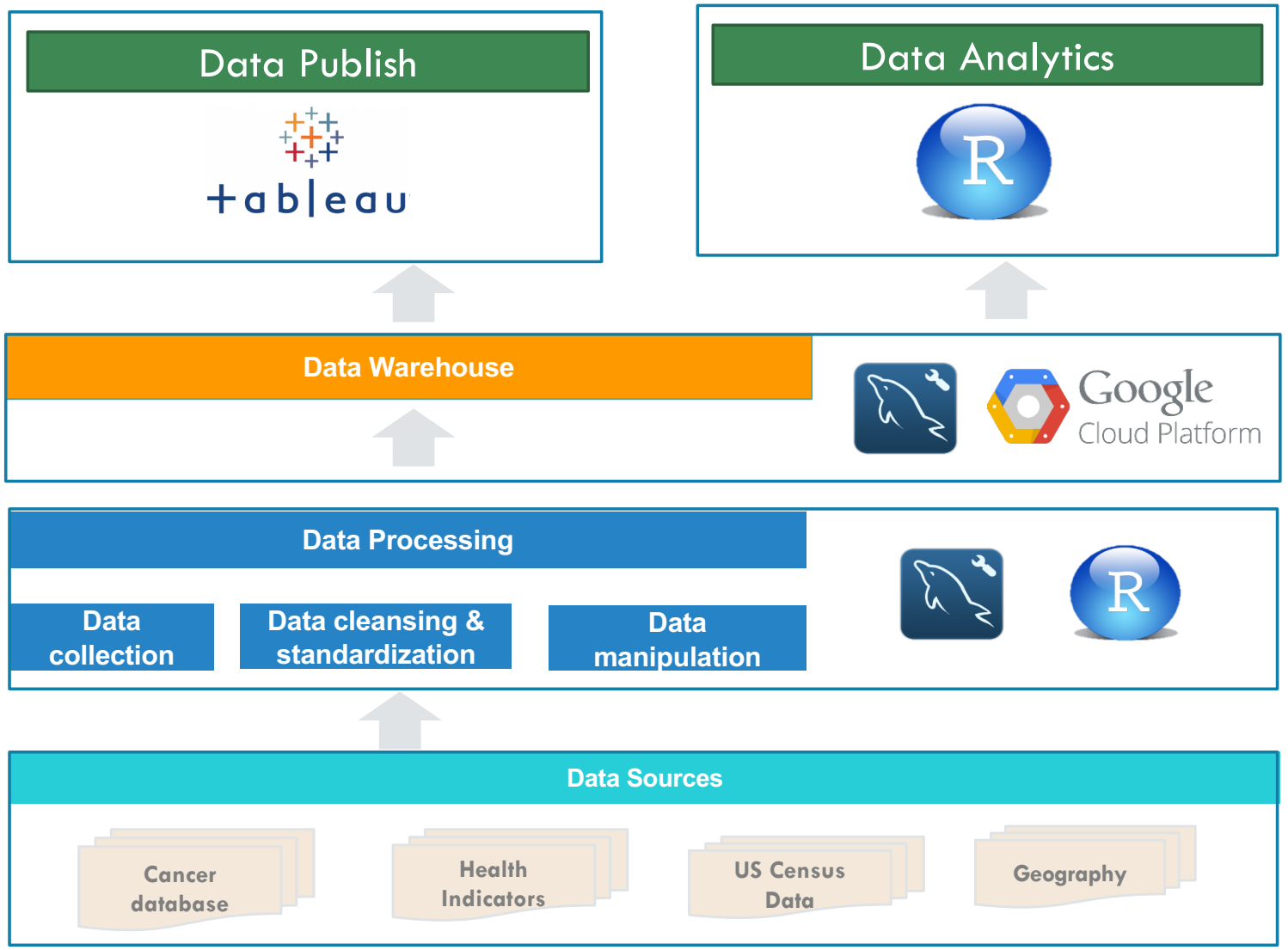
This data visualization and analytics can be used to understand the geographic variation in cancer rates across the United States and the correlation between cancer rates and various other factors (gender, income, etc.)

- Inform decisions about allocation of cancer care resources to various areas
- Understand where the cancer rates are worst and to design appropriate intervention and prevention programs

- Perform detailed analysis to determine market dynamics and understand target patient population
- Align business strategy with patient needs

- Provide research-oriented platform for analyzing cancer statistics
- Educate public on factors related to different types of cancers

DATA PIPELINE & TOOL STACK



DESIGN CONSIDERATIONS

Source data is structured (tabular) → **ER database**

- MYSQL instance in GCP

OLAP(analytical) approach:

- The data was not transactional and is fairly static, so speed of data intake is not a priority
- Focus of the project is on data analysis, so some de-normalization is appropriate

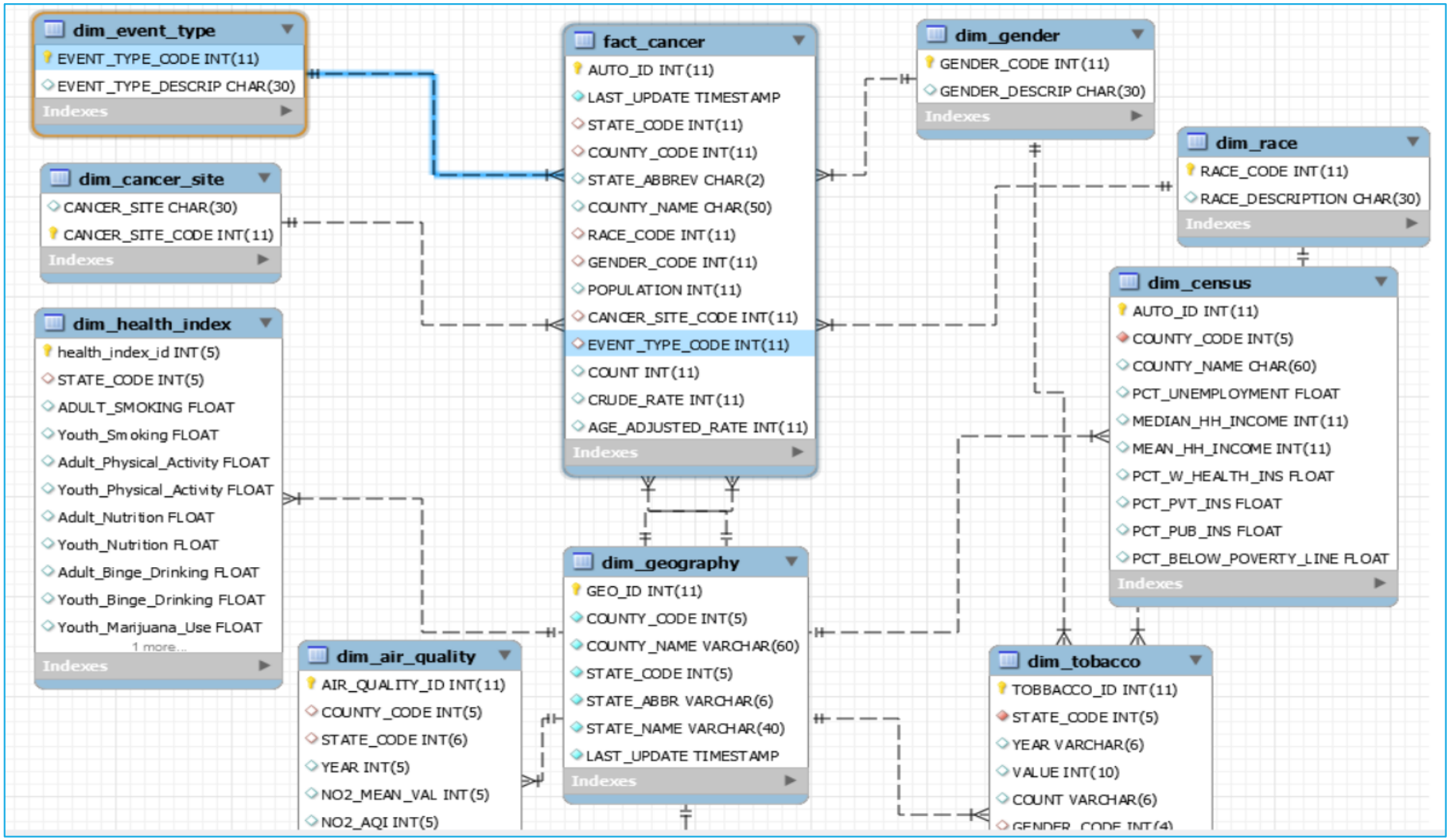
Dimensional model is a modified **star schema**

- Because we were working with existing data sets, some data sets were at different levels of aggregation (county/gender/race vs. state/age, etc.)
- This resulted in a few different fact tables, connected through the dimension of geography (state / county)

KEY DATA SOURCES

Data Source	Desciption	Data Table
U.S. Centers for Disease Control	Cancer incidence & mortality by county, race, gender, cancer type, 2012-2016	fact_cancer
US Census Bureau	2010 census: county-level statistics on median income, poverty, health insurance	dim_census
Kaggle	State-level data on nutrition, smoking, alcohol use, physical activity	dim_health_index
U.S. Centers for Disease Control	Obesity by gender and county (2012)	dim_obesity

ER DIAGRAM


dim_event_type
EVENT_TYPE_CODE INT(11)
EVENT_TYPE_DESCRIP CHAR(30)
Indexes
fact_cancer
AUTO_ID INT(11)
LAST_UPDATE TIMESTAMP
STATE_CODE INT(11)
COUNTY_CODE INT(11)
STATE_ABBREV CHAR(2)
COUNTY_NAME CHAR(50)
RACE_CODE INT(11)
GENDER_CODE INT(11)
POPULATION INT(11)
CANCER_SITE_CODE INT(11)
EVENT_TYPE_CODE INT(11)
COUNT INT(11)
CRUDE_RATE INT(11)
AGE_ADJUSTED_RATE INT(11)
Indexes
dim_gender
GENDER_CODE INT(11)
GENDER_DESCRIP CHAR(30)
Indexes
dim_race
RACE_CODE INT(11)
RACE_DESCRIPTION CHAR(30)
Indexes
dim_cancer_site
CANCER_SITE CHAR(30)
CANCER_SITE_CODE INT(11)
Indexes
dim_census
AUTO_ID INT(11)
COUNTY_CODE INT(5)
COUNTY_NAME CHAR(60)
PCT_UNEMPLOYMENT FLOAT
MEDIAN_HH_INCOME INT(11)
MEAN_HH_INCOME INT(11)
PCT_W_HEALTH_INS FLOAT
PCT_PVT_INS FLOAT
PCT_PUB_INS FLOAT
PCT_BELOW_POVERTY_LINE FLOAT
Indexes
dim_health_index
health_index_id INT(5)
STATE_CODE INT(5)
ADULT_SMOKING FLOAT
Youth_Smoking FLOAT
Adult_Physical_Activity FLOAT
Youth_Physical_Activity FLOAT
Adult_Nutrition FLOAT
Youth_Nutrition FLOAT
Adult_Binge_Drinking FLOAT
Youth_Binge_Drinking FLOAT
Youth_Marijuana_Use FLOAT
1 more...
Indexes
dim_geography
GEO_ID INT(11)
COUNTY_CODE INT(5)
COUNTY_NAME VARCHAR(60)
STATE_CODE INT(5)
STATE_ABBR VARCHAR(6)
STATE_NAME VARCHAR(40)
LAST_UPDATE TIMESTAMP
Indexes
dim_air_quality
AIR_QUALITY_ID INT(11)
COUNTY_CODE INT(5)
STATE_CODE INT(6)
YEAR INT(5)
NO2_MEAN_VAL INT(5)
NO2_AQI INT(5)
dim_tobacco
TOBBACCO_ID INT(11)
STATE_CODE INT(5)
YEAR VARCHAR(6)
VALUE INT(10)
COUNT VARCHAR(6)

ETL: DATA CLEANUP

- Data cleanup in R
- Primary fact table data from U.S. Centers for Disease Control
  - Large text file (2.7 MM lines): many empty/null rows and subtotals
  - Normalized Gender, Race, Event Type (Incidence vs. Mortality) and Cancer Site (Lung, Skin, etc.)
  - Aggregated 30 cancer types into top 15 + "All Other"
  - Extracted State & County Codes (FIPS) from text string for use as foreign key

```
COUNTY_CODE <- vector(mode = "character", length = length(cancer_data$AREA))
for(i in 1:length(cancer_data$AREA)) {
  COUNTY_CODE[i]<- substr(cancer_data$AREA[i],
                          start = regexpr("(", cancer_data$AREA[i], fixed = TRUE)[1]+1,
                          stop = regexpr(")", cancer_data$AREA[i], fixed = TRUE)[1]-1)
}
COUNTY_CODE <- as.factor(COUNTY_CODE)
```

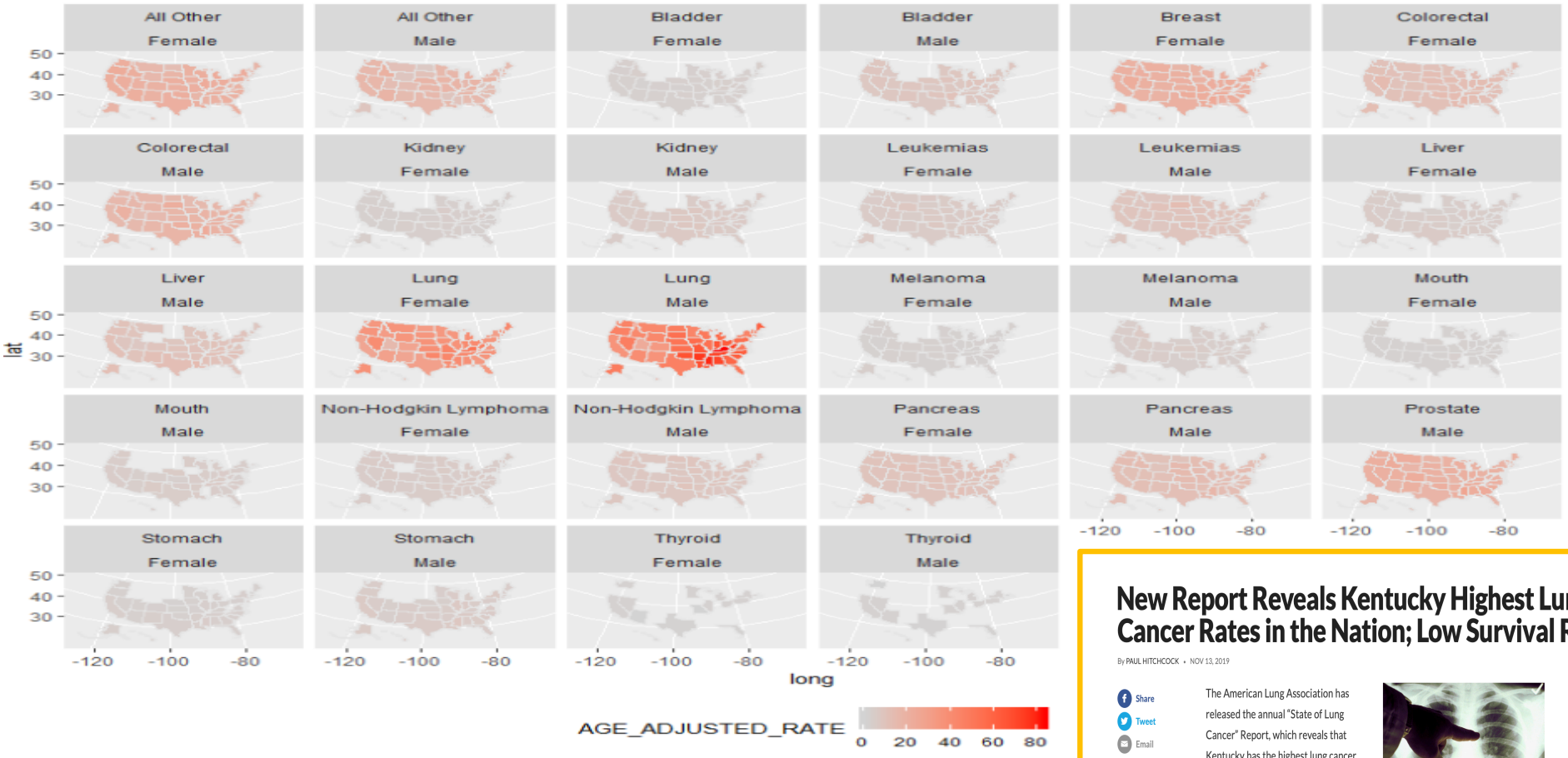
- Census data:
  - zipped JSON file with hundreds of columns
  - column descriptions in separate data dictionary

GEOGRAPHICAL DIFFERENCES

Cancer Incidence Rates by State, Cancer Type and Gender

GEOGRAPHICAL DIFFERENCES , CONTINUED

Cancer Mortality Rates by State, Cancer Type and Gender



New Report Reveals Kentucky Highest Lung Cancer Rates in the Nation; Low Survival Rates

By PAUL HITCHCOCK • NOV 13, 2019

Share
Tweet
Email

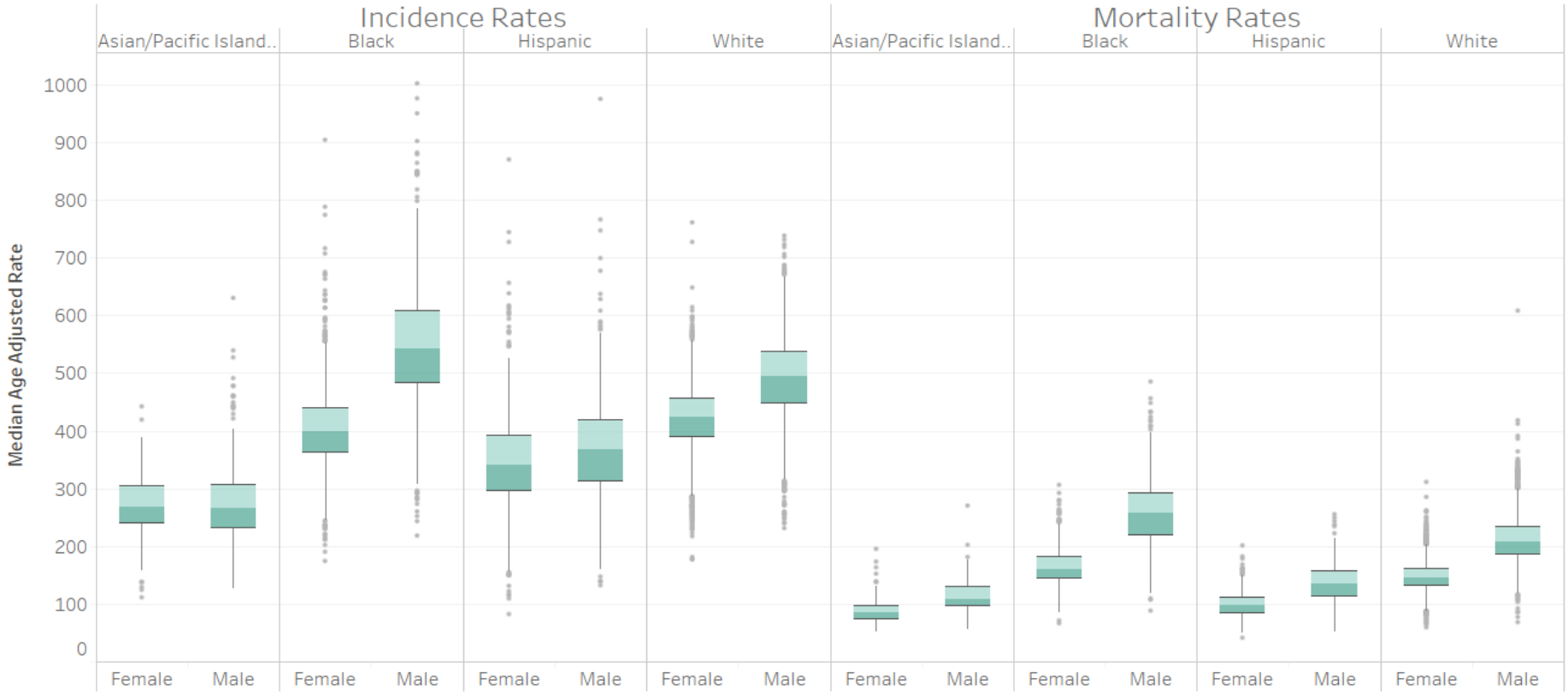
The American Lung Association has released the annual "State of Lung Cancer" Report, which reveals that Kentucky has the highest lung cancer incidence rates in the nation, as well as some of the lowest survival rates.

DISPARITIES BY RACE AND GENDER

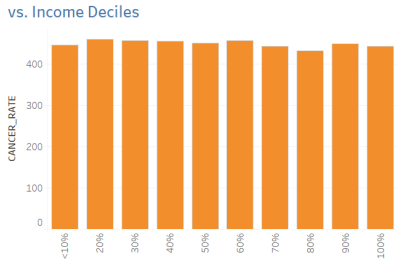
Cancer Site

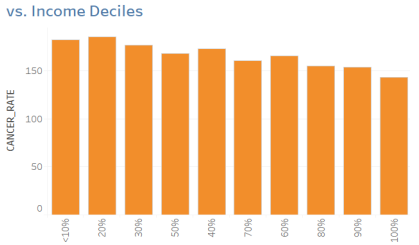
All Cancer Sites Combined

Rate Per 100,000 Population

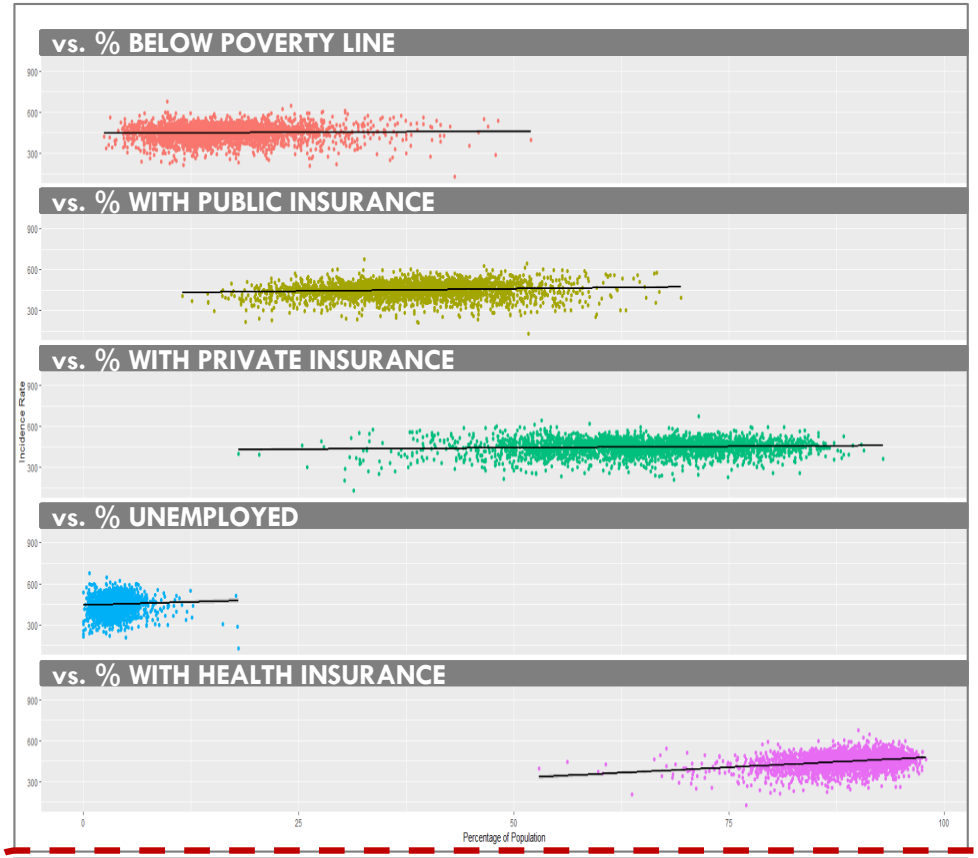


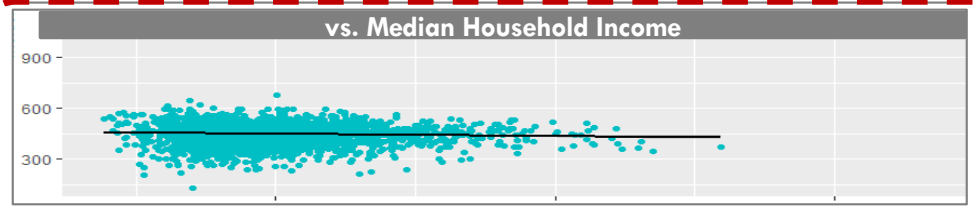
ECONOMIC FACTORS



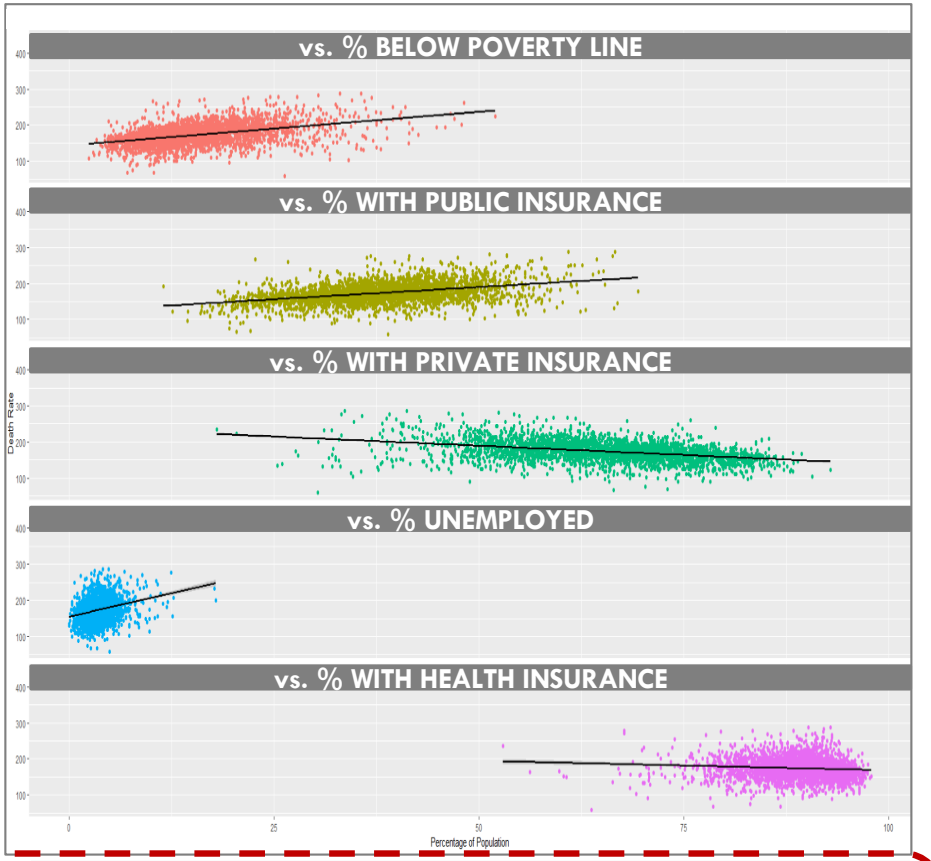


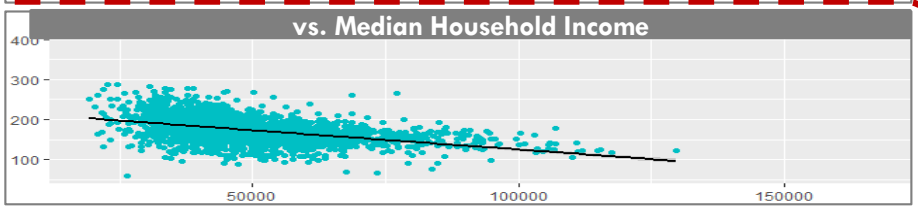
Cancer Incidence Rate





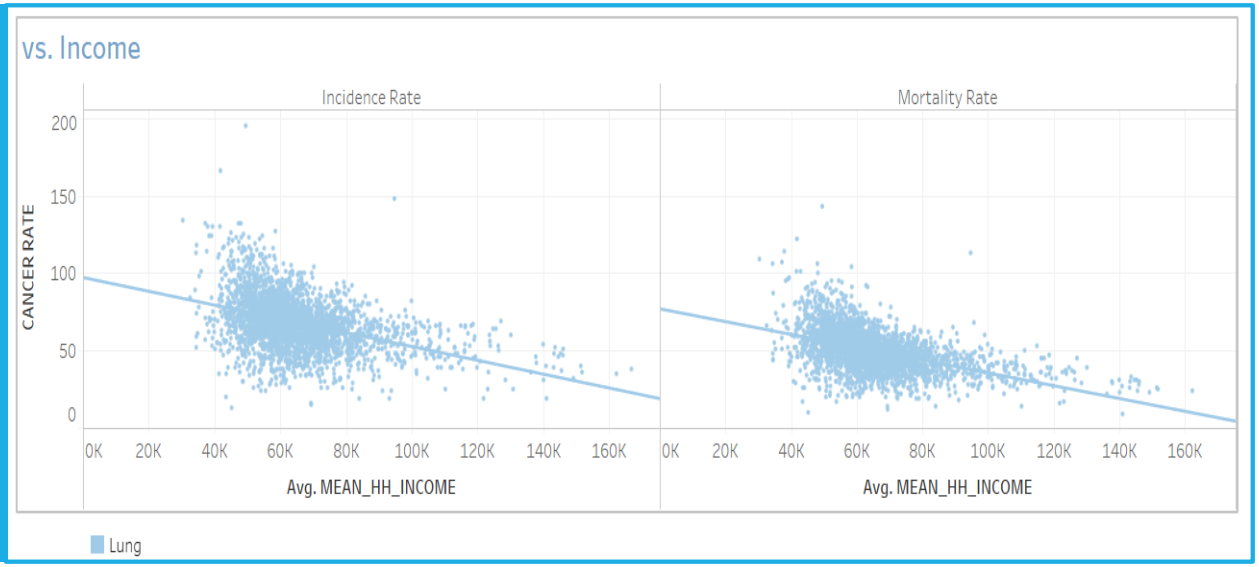
Cancer Mortality Rate

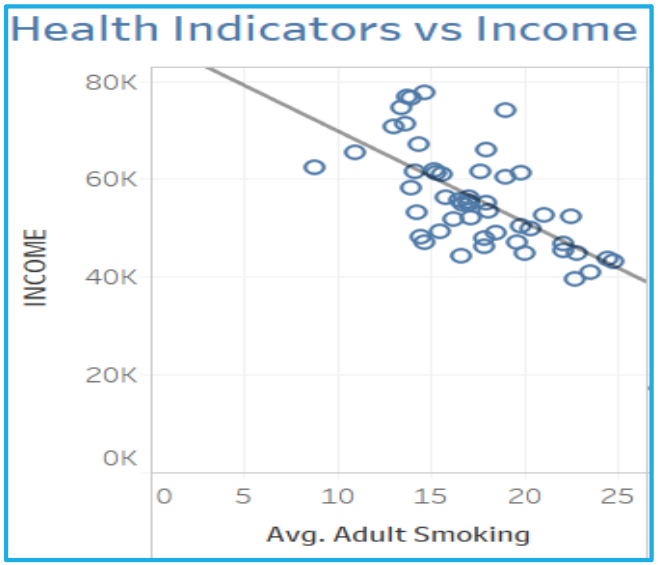




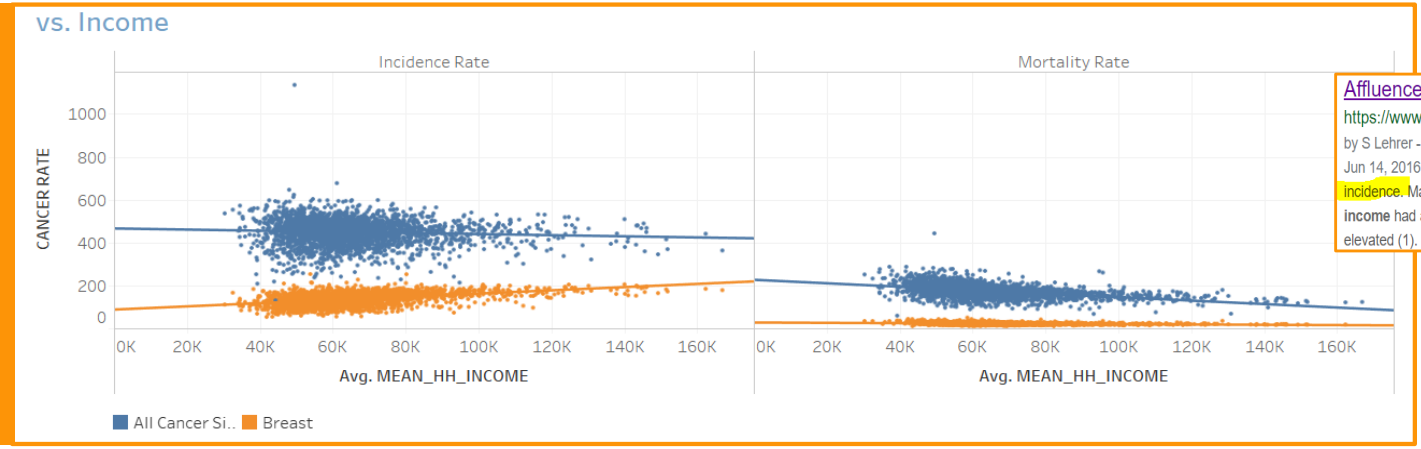
ECONOMIC FACTORS, CONTINUED

LUNG CANCER





BREAST CANCER



Affluence and Breast Cancer - NCBI - NIH
https://www.ncbi.nlm.nih.gov › pmc › articles › PMC6477537
by S Lehrer - 2016 - Cited by 10 - Related articles
Jun 14, 2016 - **High income**, **high** socioeconomic status, and affluence increase **breast cancer** incidence. Madigan et al. reported that **women** in the **upper** two-thirds of the USA population by **income** had an age-adjusted **breast cancer** risk of 1.7, while in the lower third their risk was not elevated (1).

LIFESTYLE FACTORS

Analysis of Variance Table: Mortality, All Cancer Types Combined

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
SMOKING	1	7634.6	7634.6	116.8521	4.288e-14	***
PHYSICAL_ACTIVITY	1	356.4	356.4	5.4542	0.02404	*
NUTRITION	1	36.1	36.1	0.5525	0.46116	
BINGE_DRINKING	1	26.7	26.7	0.4081	0.52618	
Residuals	45	2940.1	65.3			

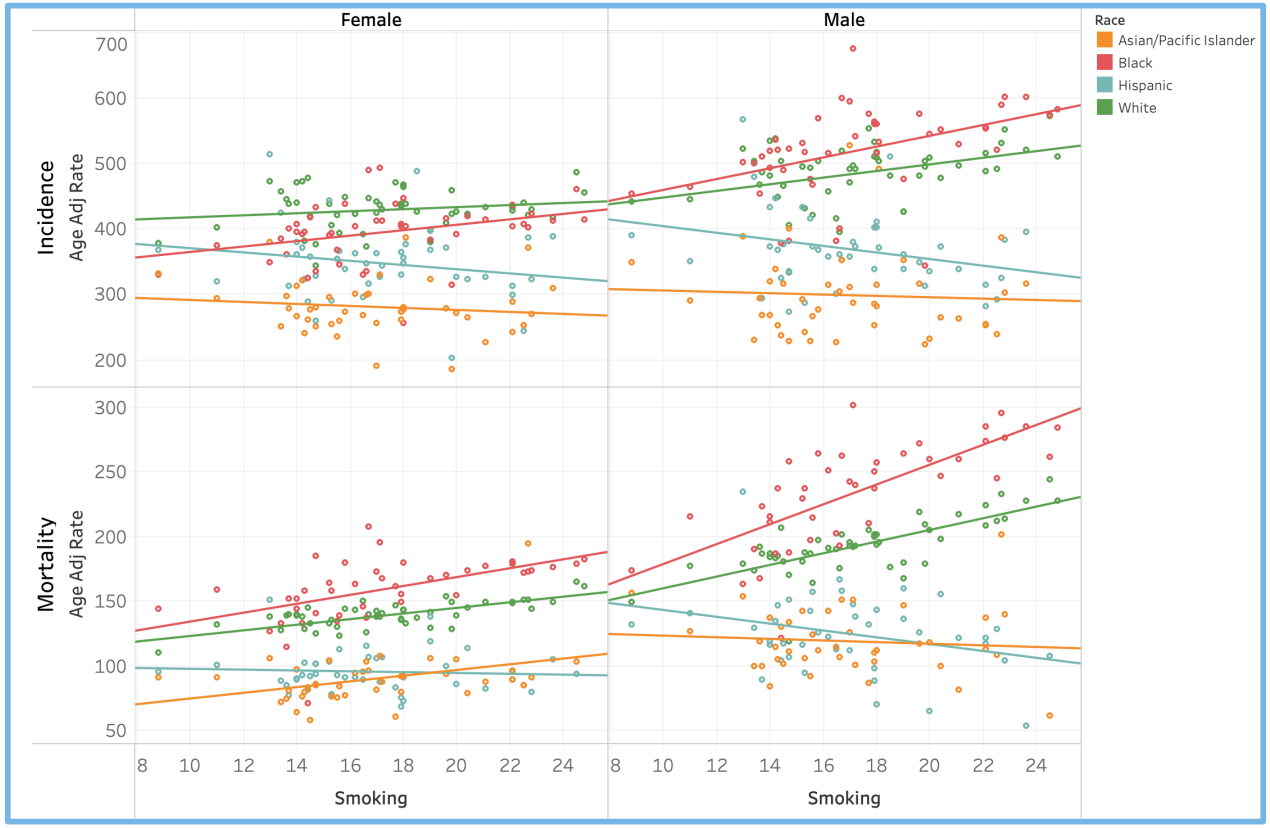
nature reviews clinical oncology

Research Highlight | Published: 01 May 2006

Risk of smoking-related lung cancer is affected by race and ethnicity

Pippa Murdie

Nature Clinical Practice Oncology **3**, 231–232(2006) | Cite this article



CONCLUSIONS AND RECOMMENDATIONS

By linking together publicly-available datasets and creating visualizations of the data, we were able to recognize many correlations that are validated in the medical literature

Applications for this database include:

- Identifying locations with unmet medical needs
- Anticipating future cancer "hotspots" based on current demographic trends
- Informing public policy on health insurance and related issues
- Educating the public on cancer risk factors

Recommendations for Future Work

- Correlation analysis on remaining cancer types
- Find datasets at consistent level of aggregation
- Finer detail if possible (individual or ZIP level)
- Time-series data for trend identification

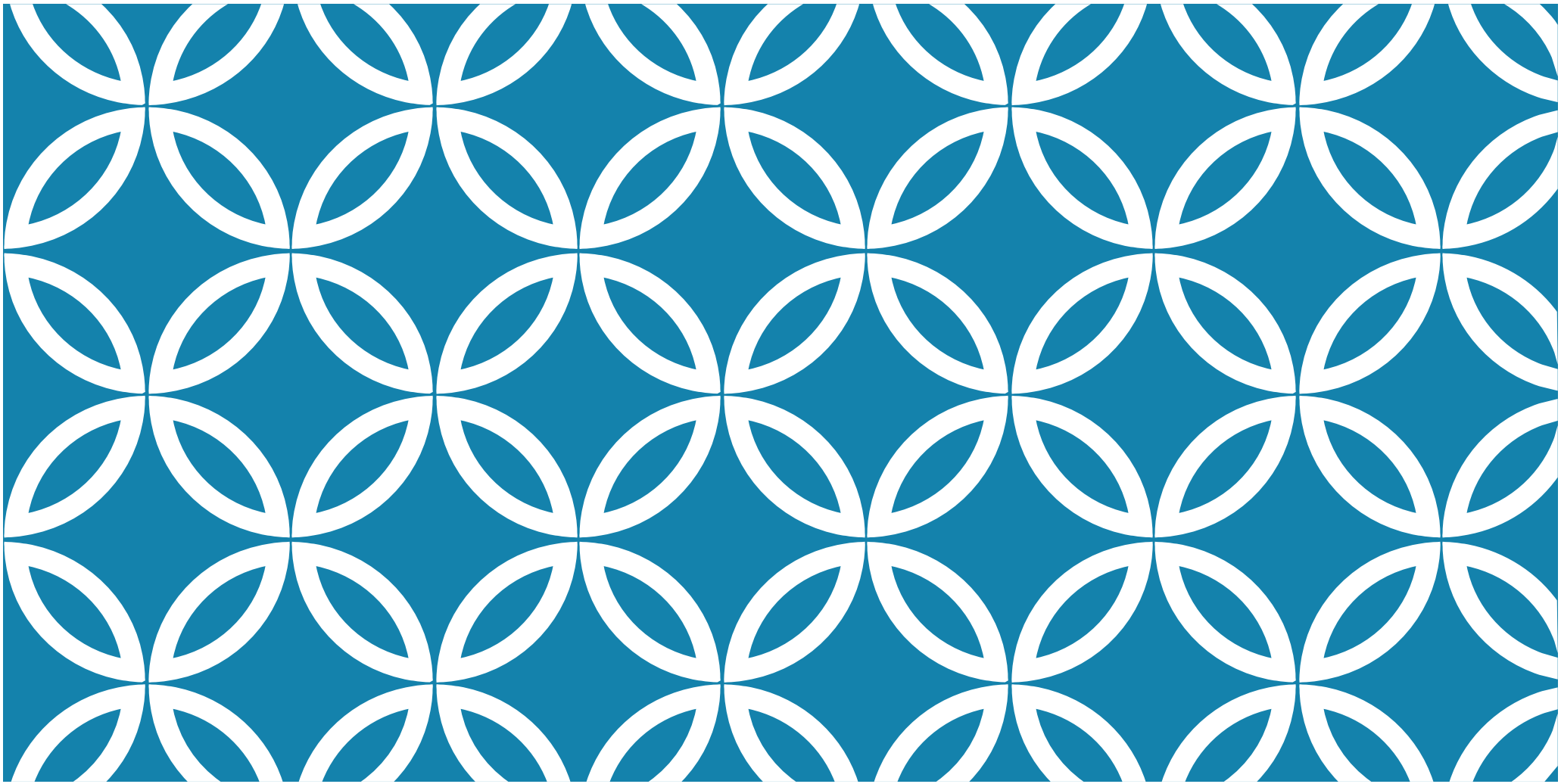


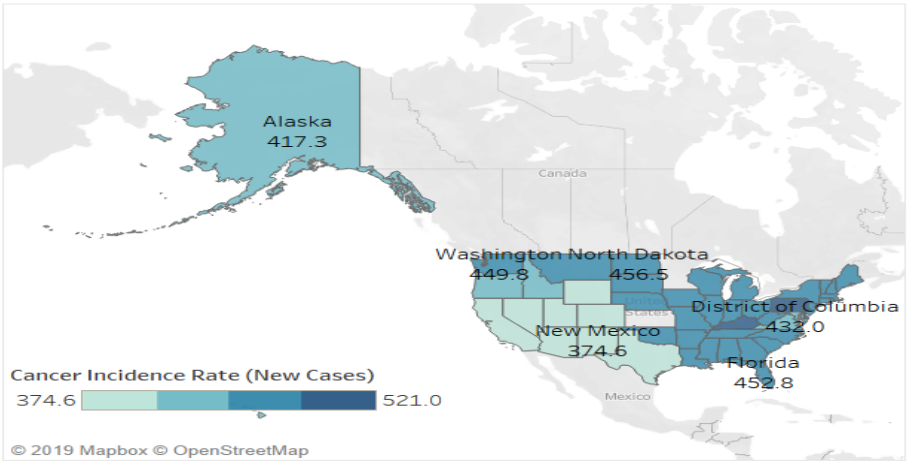
TABLEAU DASHBOARD

Friday Evening Group 2

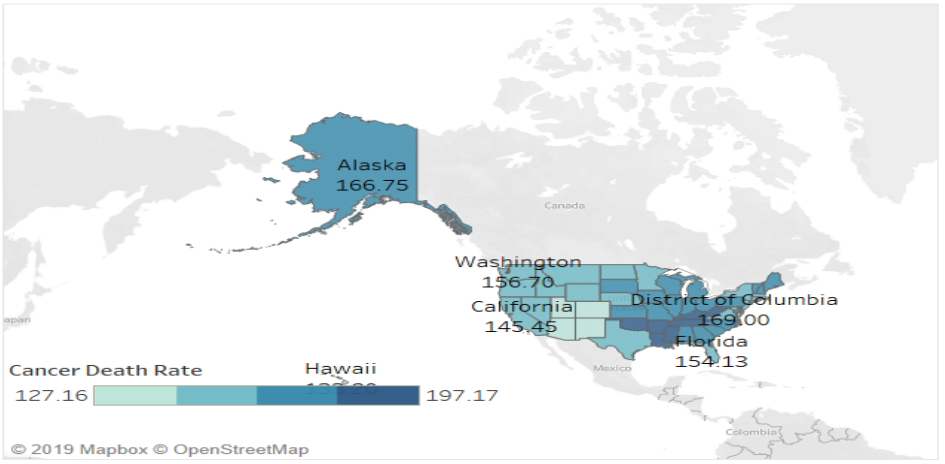
Cancer Rates- Summary Statistics

CANCER SITE: All Cancer Sites Combined | GENDER: Male and Female | RACE: All Races

Cancer Incidence Rates



Cancer Mortality Rates



Cancer Rates by Cancer Types

GENDER: (Multiple values)

CANCER_SITE	GENDER_DES..	Incidence Rate	Mortality Rate
All Cancer Sites Com..	Male	487.1	193.0
	Female	419.5	137.6
Breast	Female	155.2	20.8
Prostate	Male	104.7	19.7
Lung	Male	68.6	51.1
	Female	51.3	34.2
All Other	Female	60.3	21.4
	Male	35.3	16.1
Colorectal	Male	44.5	16.8
	Female	33.9	11.8
Melanoma	Male	27.8	3.5
	Female	17.2	1.5
Bladder	Male	34.7	7.4
	Female	8.7	2.2
Non-Hodgkin Lymph..	Male	23.3	7.3
	Female	16.1	4.3
Pancreas	Male	14.7	12.7
	Female	11.4	9.6
Thyroid	Female	22.2	0.9
	Male	7.8	0.9
Kidney	Male	22.6	5.3
	Female	11.6	2.3
Leukemias	Male	18.0	8.7
	Female	11.1	4.9
Liver	Male	13.8	10.1
	Female	4.8	4.2
Mouth	Male	17.5	3.8

0 50 100 150 200 250 300 350 400 450 500 550

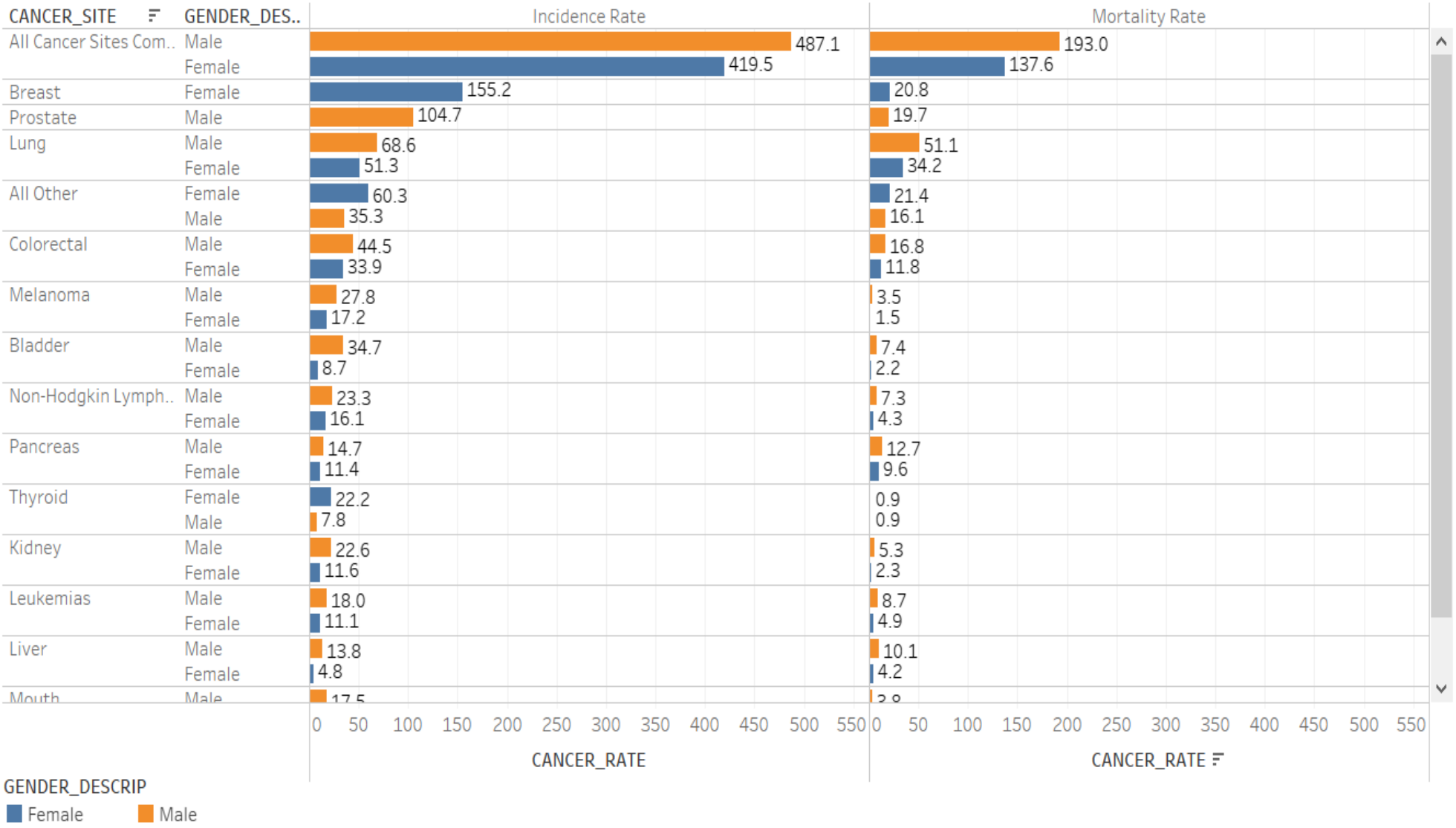
CANCER_RATE

GENDER_DESCRIP: Female, Male

Cancer Rates by Cancer Types

GENDER

(Multiple values)

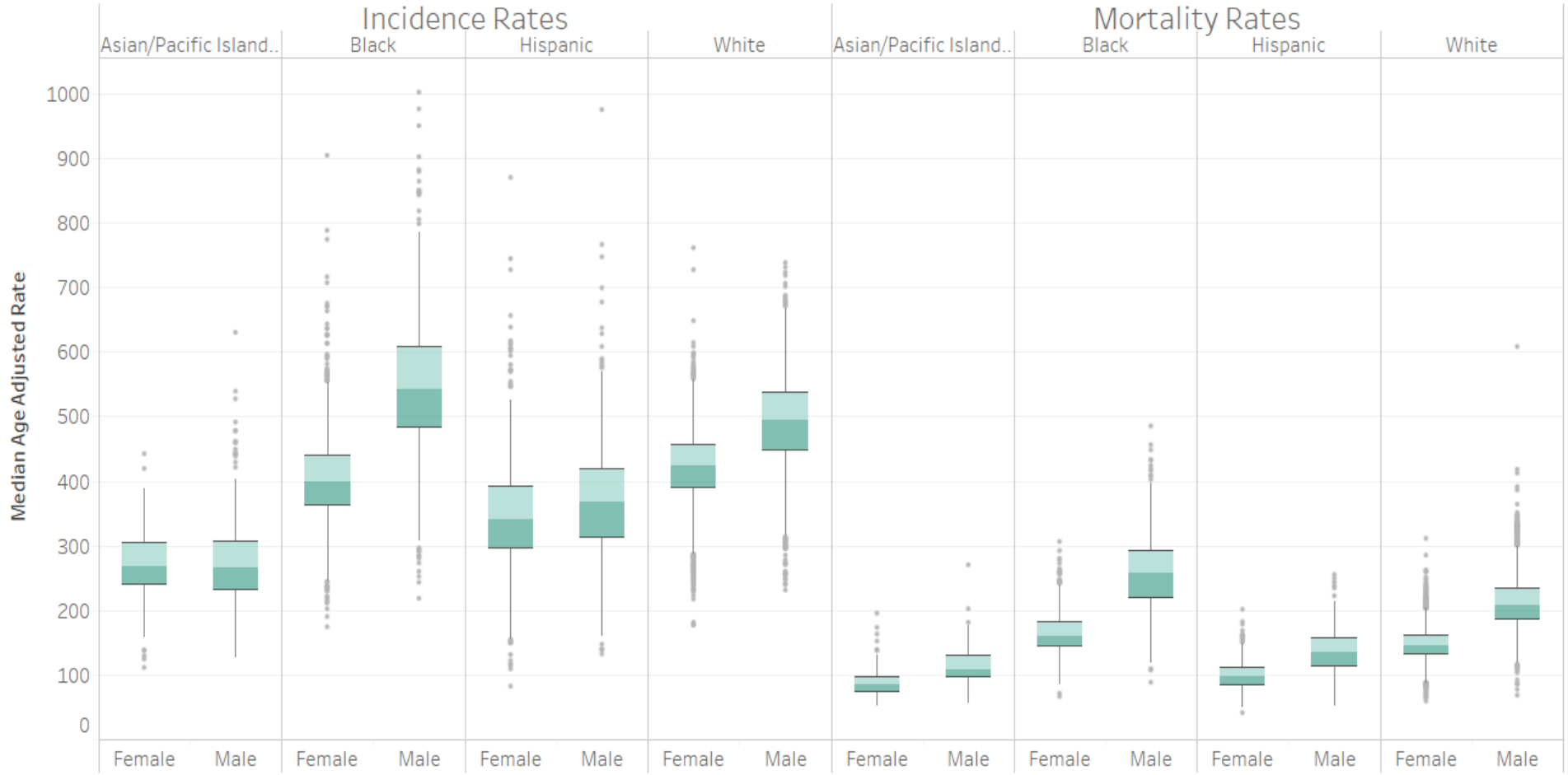


Cancer Rates by Gender and Race

Cancer Site

All Cancer Sites Combined

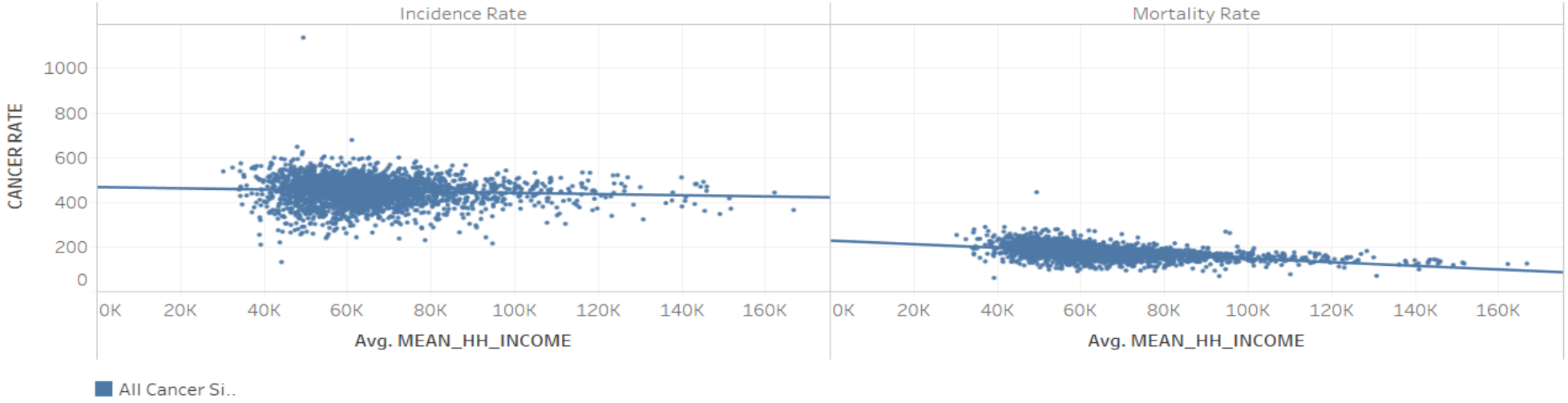
Rate Per 100,000 Population



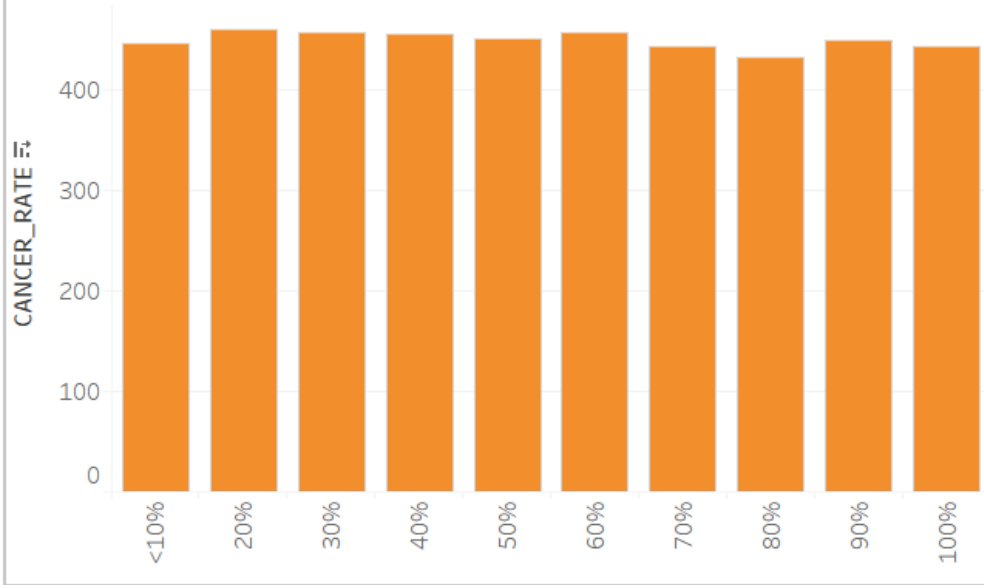
Cancer Rates vs Socioeconomic Status Indicators

CANCER SITE: All Cancer Sites Combined | GENDER: Male and Female | EVENT TYPE: Incidence Rate | RACE: All Races

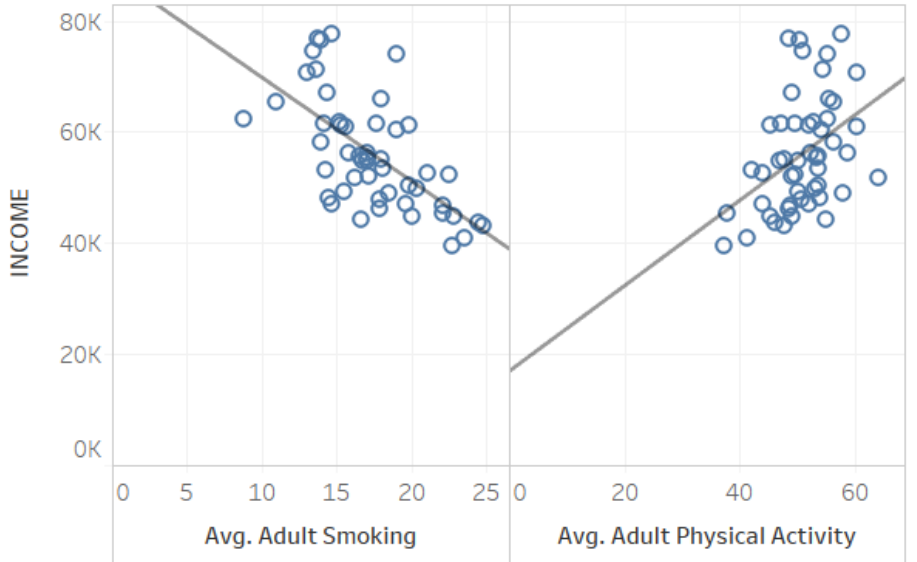
vs. Income



vs. Income Deciles



Health Indicators vs Income



Cancer Rates vs. Lifestyle Factors, by State

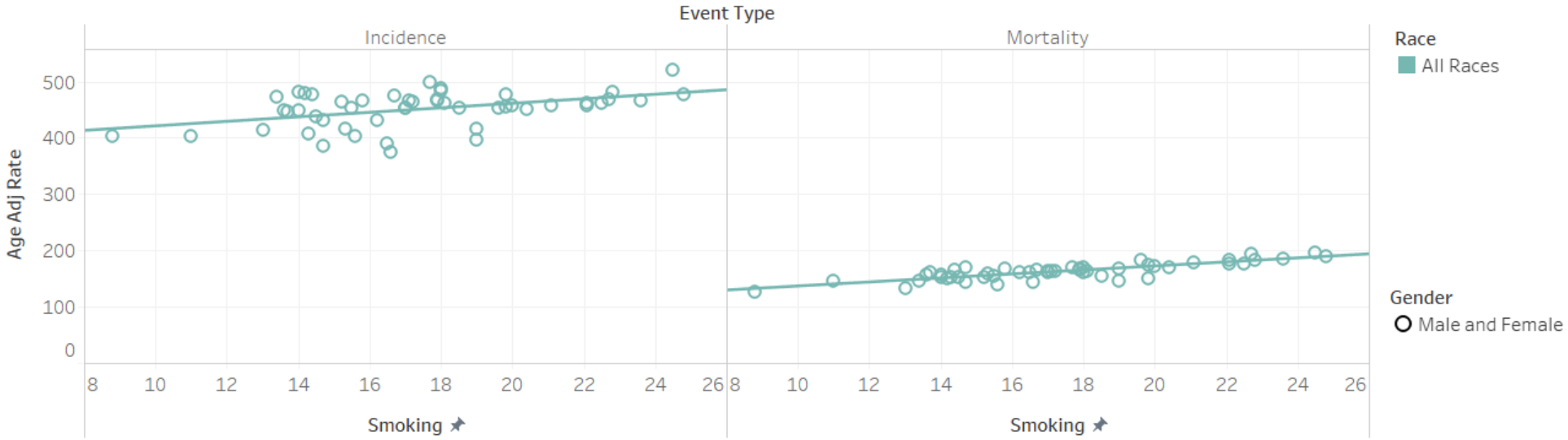
State Abbrev: (All)

Cancer Site: All Cancer Sites Combined

Gender: Male and Female

Race: All Races

vs. Smoking



vs. Physical Activity