# PREDICTING HEALTHCARE COST

*what factors influence healthcare spending in a county in the USA?*

# BACKGROUND

- Healthcare spending will be 20% of the USA economy by 2025[1]
- Where is that spending happening?
- Who is that spending on?
- What factors are most likely to influence spending in a county?

[1] *Sean P. Keehan, John A. Poisal, Gigi A. Cuckler, Andrea M. Sisko, Sheila D. Smith, Andrew J. Madison, Devin A. Stone, Christian J. Wolfe, and Joseph M. Lizonitz National Health Expenditure Projections, 2015–25: Economy, Prices, And Aging Expected To Shape Spending And Enrollment Health Affairs Vol 35 No. 7*

# DATASET

- USA Census bureau 2010 decennial census data
- Centers for Medicaid and Medicare outpatient spending data (2011-2014)
- US Dept of Housing and and Urban Development HUD USPS ZIP Code crosswalk Q2 2016

# THE PLAN?

- Originally wanted to use CMS ACS data, thought decennial census would be faster to implement for now

- Thought I would be able to predict year 2015 spending using regression, really only four data points on that scale though (2011-2014 CMS data)

- Better to look at per county characteristics then maybe extend to year on year.

# US CENSUS BUREAU DECENNIAL CENSUS DATA (2010)

- Data available through API only
- Example call:

```
http://api.census.gov/data/2010/sf1?key=[user
key]&get=PCT012A015,PCT012A119&for=state:01
```

- use preexisting api code to pull desired data:

  https://github.com/laurakurup/census-api

- provide user key (requested from census.gov)
- provide list of variables as CSV ( from census data dictionary)
- provides data out as CSV
- wait for it to iterate…

# CENSUS DATA EXAMPLE

| state | county | state_fips | county_fips | 2-person household H13. Household Size [8]_2010 | 3-person household H13. Household Size [8]_2010 |
|---|---|---|---|---|---|
| **Alabama** | Autauga | 1 | 1 | 6480 | 3841 |
| **Alabama** | Baldwin | 1 | 3 | 27641 | 11790 |
| **Alabama** | Barbour | 1 | 5 | 3289 | 1620 |
| **Alabama** | Bibb | 1 | 7 | 2634 | 1380 |
| **Alabama** | Blount | 1 | 9 | 7494 | 3852 |
| **Alabama** | Bullock | 1 | 11 | 1083 | 608 |
| **Alabama** | Butler | 1 | 13 | 2705 | 1520 |
| **Alabama** | Calhoun | 1 | 15 | 16293 | 8131 |
| **Alabama** | Chambers | 1 | 17 | 4674 | 2375 |
| **Alabama** | Cherokee | 1 | 19 | 4001 | 1782 |

# CENSUS DATA EXAMPLE

| state | county | state_fips | county_fips | 2-person household H13. Household Size [8]_2010 | 3-person household H13. Household Size [8]_2010 |
|---|---|---|---|---|---|
| **Alabama** | Autauga | 1 | 1 | 6480 | 3841 |
| **Alabama** | Baldwin | 1 | 3 | 27641 | 11790 |
| **Alabama** | Barbour | 1 | 5 | 3289 | 1620 |
| **Alabama** | Bibb | 1 | 7 | 2634 | 1380 |
| **Alabama** | Blount | 1 | 9 | 7494 | 3852 |
| **Alabama** | Bullock | 1 | 11 | 1083 | 608 |
| **Alabama** | Butler | 1 | 13 | 2705 | 1520 |
| **Alabama** | Calhoun | 1 | 15 | 16293 | 8131 |
| **Alabama** | Chambers | 1 | 17 | 4674 | 2375 |
| **Alabama** | Cherokee | 1 | 19 | 4001 | 1782 |

```
grep "<td>" Census_Data_API_data_2010_sf1_variables.htm | wc -l
    8969
```

# CENSUS DATA EXAMPLE

| state | county | state_fips | county_fips | 2-person household H13. Household Size [8]_2010 | 3-person household H13. Household Size [8]_2010 |
|---|---|---|---|---|---|
| **Alabama** | Autauga | 1 | 1 | 6480 | 3841 |
| **Alabama** | Baldwin | 1 | 3 | 27641 | 11790 |
| **Alabama** | Barbour | 1 | 5 | 3289 | 1620 |
| **Alabama** | Bibb | 1 | 7 | 2634 | 1380 |
| **Alabama** | Blount | 1 | 9 | 7494 | 3852 |
| **Alabama** | Bullock | 1 | 11 | 1083 | 608 |
| **Alabama** | Butler | 1 | 13 | 2705 | 1520 |
| **Alabama** | Calhoun | 1 | 15 | 16293 | 8131 |
| **Alabama** | Chambers | 1 | 17 | 4674 | 2375 |
| **Alabama** | Cherokee | 1 | 19 | 4001 | 1782 |

```
grep "<td>" Census_Data_API_data_2010_sf1_variables.htm | wc -l
    8969
```

# CMS OUTPATIENT PAYMENT DATA

- 2014 data available as download from CMS website
  - 1.7 mb ZIP
  - 4.5 mb CSV
- no data dictionary

| apc | provider_id | provider_name | Provider_Street_Address | Provider_City | Provider_State | Provider_Zip_Code |
|---|---|---|---|---|---|---|
| 0012 - Level I Debridement & Destruction | 40055 | SPARKS REGIONAL MEDICAL CENTER | 1001 TOWSON AVENUE | FORT SMITH | AR | 72902 |
| 0012 - Level I Debridement & Destruction | 50017 | MERCY GENERAL HOSPITAL | 4001 J ST | SACRAMENTO | CA | 95819 |
| 0012 - Level I Debridement & Destruction | 50320 | HIGHLAND HOSPITAL | 1411 E 31ST STREET | OAKLAND | CA | 94602 |
| 0012 - Level I Debridement & Destruction | 50357 | GOLETA VALLEY COTTAGE HOSPITAL | 351 S PATTERSON AVE | SANTA BARBARA | CA | 93111 |
| 0012 - Level I Debridement & Destruction | 30064 | BANNER-UNIVERSITY MEDICAL CENTER TUCSON CAMPUS | 1501 NORTH CAMPBELL AVENUE | TUCSON | AZ | 85724 |

| Hospital_Referral_Region | Outpatient_Services | Average_Estimated_Submitted_Charges | Average_Total_Payments |
|---|---|---|---|
| AR - Fort Smith | 279 | 208.986595 | 43.36767025 |
| CA - Sacramento | 25 | 239.6 | 77.5492 |
| CA - Alameda County | 58 | 123.6896552 | 71.48551724 |
| CA - Santa Barbara | 38 | 292.3863158 | 52.5 |
| AZ - Tucson | 132 | 267.625 | 53.11689394 |

# CMS OUT-PATIENT PAYMENT DATA

- Different level of categorization

| | Provider_Street_Address | Provider_City | Provider_State | Provider_Zip_Code |
|---|---|---|---|---|
| | 1001 TOWSON AVENUE | FORT SMITH | AR | 72902 |
| | 4001 J ST | SACRAMENTO | CA | 95819 |
| | 1411 E 31ST STREET | OAKLAND | CA | 94602 |
| | 351 S PATTERSON AVE | SANTA BARBARA | CA | 93111 |
| AMPUS | 1501 NORTH CAMPBELL AVENUE | TUCSON | AZ | 85724 |

# HUD ZIP CROSSWALK

- Available as direct download from HUD website
  - Download CSV - 3.2 mb

| ZIP | COUNTY | RES_RATIO | BUS_RATIO | OTH_RATIO | TOT_RATIO |
|---|---|---|---|---|---|
| **501** | 36103 | 0 | 1 | 0 | 1 |
| **601** | 72001 | 1 | 1 | 1 | 1 |
| **602** | 72003 | 1 | 1 | 1 | 1 |
| **603** | 72071 | 0.008104634 | 0.000948767 | 0.00681431 | 0.007679705 |
| **603** | 72005 | 0.991895366 | 0.999051233 | 0.99318569 | 0.992320295 |
| **604** | 72005 | 1 | 1 | 1 | 1 |
| **605** | 72005 | 1 | 1 | 1 | 1 |
| **606** | 72093 | 1 | 1 | 1 | 1 |
| **610** | 72011 | 1 | 1 | 1 | 1 |

# PUTTING IT ALL TOGETHER

- Goals is to create one data frame including
  - Features from Census data
  - label from CMS data
- appropriately aligned and integrated

# IMPORTING DATA – ZIP CROSSWALK

```
#CREATE CROSSWALK DATAFRAME
path = '/Users/Iain/DS-SEA-3/DS-SEA-3-Project-ILM/data/HUD/'
filename = 'zip_county_062016.csv'
zipCountyCrosswalk = pd.read_csv(path + filename, converters={'ZIP':lambda x: str(x), 'COUNTY':lambda x:str(x)})
```

# IMPORTING DATA – CMS PAYMENT DATA

```
#CREATE CMS 2014 OUTPAITENT PAYMENT INFO DATA FRAME - ENSURE LEADING ZEROS PRESENT ON ZIP CODES
path = '/Users/Iain/DS-SEA-3/DS-SEA-3-Project-ILM/data/cms/'
filename = 'Medicare_Provider_Charge_Outpatient_APC32_CY2014.csv'
rawPaymentData = pd.read_csv(path + filename, converters={'Provider_Zip_Code':lambda x: str(x)})
rawPaymentData['Provider_Zip_Code'] = rawPaymentData['Provider_Zip_Code'].apply(lambda x: x.zfill(5))
```

```
#SIMPLIFY PAYMENT INFO DOWN TO SUMMED PAYMENTS PER ZIPCODE
rawPaymentData['Total'] = rawPaymentData['Outpatient_Services'] * rawPaymentData['Average_Total_Payments']
paymentData = rawPaymentData[['Provider_Zip_Code','Total']].groupby(by='Provider_Zip_Code', as_index=False).sum()
```

# IMPORTING DATA – CENSUS DATA

```
#CREATE CENSUSDATA DATAFRAME
path = '/Users/Iain/DS-SEA-3/DS-SEA-3-Project-ILM/data/census/'
filename = 'census-data-by-county-2016.07.31-09.48PM.csv'
censusData = pd.read_csv(path + filename, converters={'state_fips':lambda x: str(x), 'county_fips':lambda x: str(x)})
```

```
censusData['state_fips'] = censusData['state_fips'].apply(lambda x: x.zfill(2))
censusData['county_fips'] = censusData['county_fips'].apply(lambda x: x.zfill(3))
censusData['FIPS'] = censusData['state_fips']+censusData['county_fips']
```

# MERGING THE DATAFRAMES

```
#CHANGE COLUMS NAMES TO MERGE BY THOSE COLUMN NAMES
zipCountyCrosswalk.rename(columns={'COUNTY':'FIPS'}, inplace=True)
paymentData.rename(columns={'Provider_Zip_Code':'ZIP'}, inplace=True)
```

```
#MERGE CENSUS AND ZIP CROSSWALK DATA FRAME
censusZip = pd.merge(censusData, zipCountyCrosswalk, on='FIPS', how='left')

#MEREGE THE CENSUSZIP DATAFRAME AND THE CMS PAYMENT DATA DATAFRAME
censusZipPayment = pd.merge(censusZip, paymentData, on='ZIP', how = 'left')
```

# NEXT STEPS

- Create column of proportioned labels
  - modified by zip code/county coverage (multiply CMS total payment by res_ratio column from xcrosswalk)

- eliminate rows where payment information NaN - deal with null values by exclusion

- Run regression with every census column as feature and proportional total payment column as label

- Visualize / Interpret results - heatmap?

- use CMS ACS data instead of decennial census data - (62720 variables)