**Citing Sources:**

Credit the source, author, url, name, version and whatever else is relevant for data, charts, programs used in this project. Data source #1 is an example.

**Data Source #1:** (gswider data.world 5 xlsx and ACA\_barplot.py)

* author- owner: gswider
* title: ACA State Data Set
* year of publication: 2016
* publisher: Data.World retains cleansed and modified data sets created from data set at https://aspe.hhs.gov/compilation-state-data-affordable-care-act.
* edition or version: Not sure
* access information: <https://data.world/gswider/aca-state-data>
  + Coverage Gains ACA state data 12\_2016.xlsx
  + Employer Coverage ACA state data 12\_2016 short headers.xlsx
  + Individual Market Coverage ACA state data 12\_2016 short headers.xlsx
  + Medicaid ACA state data 12\_2016 short headers.xlsx
  + Medicare ACA state data 12\_2016.xlsx
  + ACA\_barplot.py

**Data Source #2:** (Assistant Secretary for Planning and Evaluation HHS)

* <https://aspe.hhs.gov/compilation-state-data-affordable-care-act>
* <https://aspe.hhs.gov/system/files/pdf/253396/CompanionDecember2016StatebyState.pdf>

**Data Source #3:** (Kaiser Family Foundation 2017 Insurance Marketplaces)

<https://www.kff.org/health-reform/state-indicator/state-health-insurance-marketplace-types/?currentTimeframe=0&selectedDistributions=marketplace-type&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>

* **Raw data:** kff\_Ins\_Mkp\_raw\_data.csv
* KFF page has an interactive map - not sure how the map was made.
  + Make a map using <https://plot.ly/javascript/choropleth-maps/> or D3.

**Data Source #4:** (HHS<https://aspe.hhs.gov/compilation-state-data-affordable-care-act>)

* Medicaid\_Expansion\_by\_State.xlsx

**Data Source #5:** (Census.gov USA state abbreviations with FIPS and GNIS codes)

* usa\_state\_abv.csv

**Possible Data Source #1: Is there data about pre-existing conditions in 2010 - 2015?**

The goals of this dataset are to localize the previously estimated 24 million insured individuals covered by the Affordable Care Act who would stand either to lose coverage or see increased premiums under a repeal of the act. It only counts individuals under the age of 65 who are ineligible for Medicare. It has two csv files, one with a by state, by age bracket that is for under age 65, and another by congressional district, by age bracket that is also for those under age 65.

* <https://data.world/carlvlewis/pre-exisiting-conditions-by-state-congressional-district>

**Possible Data Source #2:**

* <https://data.world/hhs/2018-qhp-landscape-data> (<https://data.healthcare.gov/>)
  + These are health plans for 2018 and it looks like there are data sets for earlier years.

**Possible Data Source #3:**

* <https://www.kaggle.com/datasets>
* <https://www.kaggle.com/hhs/health-insurance>

**Possible Data Source #4:** (Governor-Legislature by Political Party)

* Governor\_Legislature\_by\_State\_2010-2016.xlsx

**Possible Data Source #5:** (Survey on Individual Market persons)

* <https://www.kff.org/health-costs/poll-finding/survey-of-people-who-purchase-their-own/>

**Possible Data Source #6:** (Summary ACA)

* <https://www.kff.org/health-reform/fact-sheet/summary-of-the-affordable-care-act/>

**Data to Find:**

* Data on persons with existing conditions 2010 - 2015.
  + <https://www.kff.org/tag/pre-existing-conditions/>
* Health insurance costs for urban versus rural populations.

**-----------------------------------------------------------------/------------------------------------------------------------------------**

**Groups Working on ACA Related Data:**

Sent a Data.World internal request for help with how to use data.world on 10/27 at 7:54 pm.

* **I have 5 spreadsheets of data from the gswider data.world link below.**

<https://data.world/gswider/aca-state-data>

* **This data set looks interesting because I want to see if persons with pre-existing conditions made gains by either getting insurance or having more affordable insurance.**

<https://data.world/carlvlewis/pre-exisiting-conditions-by-state-congressional-district>

* **kaggle.com (Health Insurance Coverage) Premium rates before and after the Affordable Care Act. There are some good charts at this link.**

<https://www.kaggle.com/hhs/health-insurance>

**Possible ACA Related Sources:**

**The home of the U.S. Government’s open data**

<https://www.data.gov/>

<https://www.healthdata.gov/>

<https://www.healthdata.gov/api>

<https://www.hhs.gov>

<https://opendata.socrata.com/profile/Christopher-Dennis-Guthrie/ybiw-7kvh/app_token/28764>

**Federal Government Health Care website:**

<https://www.healthcare.gov/>

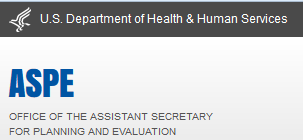
**Census Government Data:**

<https://www.census.gov/en.html>

<https://www.census.gov/data/data-tools.html>

**California HHS Open Data Portal:**

<https://data.chhs.ca.gov/>



**Interesting Links:**

**data.world explore the data.world python sdk - dataset by nrippner**

<https://data.world/nrippner/explore-the-data-world-python-sdk>

**awesome-public-datasets - A topic-centric list of high-quality open datasets in public domains**

<https://github.com/caesar0301/awesome-public-datasets>

**Excerpted from** <https://www.analyticsvidhya.com/blog/2016/09/this-machine-learning-project-on-imbalanced-data-can-add-value-to-your-resume/>

**The Problem Statement**

*Given various features, the aim is to build a predictive model to determine the income level for people in US. The income levels are binned at below 50K and above 50K.*

From the problem statement, it’s evident that this is a binary classification problem.

Generating hypothesis is the most crucial step in building models. Yet, most analysts tend to overlook this step. In simple words, this technique enlightens our way by indicating which direction (set of variables) to choose.

This step should be practiced before looking at the data. This is done to think broadly and not be constrained by what is available. In this step, we’ll create a laundry list of factors which we think could influence the prediction metrics. Read more about hypothesis generation [here](https://www.analyticsvidhya.com/blog/2015/09/hypothesis-testing-explained/).

Let’s think of some hypothesis which can influence the outcome. Here is a set of hypothesis to get you started:

Hò : There is no significant impact of the variables (below) on the dependent variable.

Ha : There exists a significant impact of the variables (below) on the dependent variable.

1. Age
2. Marital Status
3. Income
4. Family Members
5. No. of Dependents
6. Tax Paid
7. Investment (Mutual Fund, Stock)
8. Return from Investments
9. Education
10. Spouse Education
11. Nationality
12. Occupation
13. Region in US
14. Race
15. Occupation category

Remind you, this is not an exhaustive list. I’d suggest you not to limit your thoughts with the ones above, your aim should be to make your project as comprehensive & presentable as possible.

Also, every time you think of an hypothesis, try and think of what would be the relationship like and why would that hypothesis stand true. For example, when I say Education – what I really mean is this: “I think that with higher education, people would have higher chances of better employment and hence their income would have higher chances of being more than 50K.”

Similarly, spend some time thinking about your set of hypothesis, how would they be impacting income and hence what is the best way to capture the mathematical relationship.

**P.S.** *Do this before you move forward in the article*

**ACA - Stories to tell, Questions to ask, Relationships to find:**

In general, is there an impact of the ACA on uninsured, pre-existing condition, individual, group, Medicare, and Medicaid insured persons measured by percent uninsured, percent covered by Medicaid, premium costs, percent covered and/or costs for urban versus rural populations?

**Main Questions:**

1. Are there ACA related differences based upon whether an ACA website is state run, or Federally run (FFM or Federally Facilitated Marketplace)?
2. And in the Federally run sites, are there differences in the key indicators in the states that expanded Medicaid and those that didn't?
   1. Did the state's willingness to participate in the ACA have an impact?
   2. How does politics overlay the results?
   3. How to measure this last two parts? Use key indicators.

Not all questions need to be considered, if I end up with too much or can't find data to support research.

* Do I need to get into hypothesis testing, Z-value (also P and T) calculation?
* How statistical do I need to get?
* Am I trying to discover relationships for now?

**Key indicators - % uninsured, % per-existing conditions, % individual market, % group market.**

**Breakdown Questions:**

* Did whether a state set up their own ACA or used a federally managed website affect the key indicators?
  + State managed sites:
    - What differences are there between states with their own websites and those that were federally managed if all the members expanded Medicaid?
  + Federally managed sites:
    - Which of these states also expanded Medicaid?
    - Which had republican governors? State legislatures?
    - Which states Attorneys Generals sued the federal government?
    - FFM Fact Sheet:

<https://www.cms.gov/CCIIO/Resources/Fact-Sheets-and-FAQs/ffe.html>

* Is it possible to find out how many pre-existing condition persons got coverage and if they had coverage whether it was more affordable with the ACA tax credits?
  + Are there data sets that can tell me about the impact of the ACA on persons with pre-existing conditions?
* With the states that didn't expand Medicaid, were there any differences in uninsured rates, persons with pre-existing conditions being covered, or did insurance costs go up or down?
  + Of this group of states did any create and operate their own website?
  + Did whether a state expanded Medicaid or not affect the key indicators?
* Can I look at the urban and rural impact of the ACA on the key indicators? Is there data at Census.gov?
  + Did a state's mix of rural or urban population affect the key indicators?
* Did a state's per capita income affect the key indicators?
* Political:
  + What was the political party of the Governor and state assembly/senate of each state?
  + What was the political party of the Governor and state assemble/senate of the states the expanded Medicaid and those that didn't?
  + Is it possible that politics affected the key indicators? Did a state's mix of rural or urban population affect the key indicators?
* Are there data sets on the ACA impact on emergency room usage?

**Project Tools:**

**Database: sqlite or postgres or a combo?**

<https://sqlite.org/whentouse.html>

<https://sqlite.org>

**Charting:**

URL with some good state charts:

<http://www.vizwiz.com/2013/01/alberto-cairo-three-steps-to-become.html>

**Map of US Congressional districts D3**

<https://bl.ocks.org/mbostock/4657115>

**Census Map Link:**

<https://www.census.gov/geo/maps-data/data/cbf/cbf_state.html>

**D3 3d plots site:**

<http://christopheviau.com/d3list/>