Extraction, Transformation, and Load (ETL) Project

Team Members: Austin Little, Meryem ozcelik, kimberly gordon   
Summary

11/11/2018

## Topic for review

* <https://github.com/austinlittle1/ETLproject>
* How does the climate impact a person’s overall health?
* Expected Outcome – Air pollution level and Toxicity Levels based on the location is directly related to a person’s health. However, there are environmental issues that can positively / negatively change the results.
* Specific integration points will be needed to clearly understand the impacts.

## Project Scope

* For the team to find manageable files to review, analyze and share results after performing extraction, transformation and loading in database for review.
* Very quick project with less than two weeks to review and work with team.

## Approach

* Team decided on a general topic for analysis.
* Researched the availability of data.
* Analyzed possible share information that can be used as integration points.
* Performed various Python-Pandas analysis on the data to prepare the integration for formal review.

## Finding Data

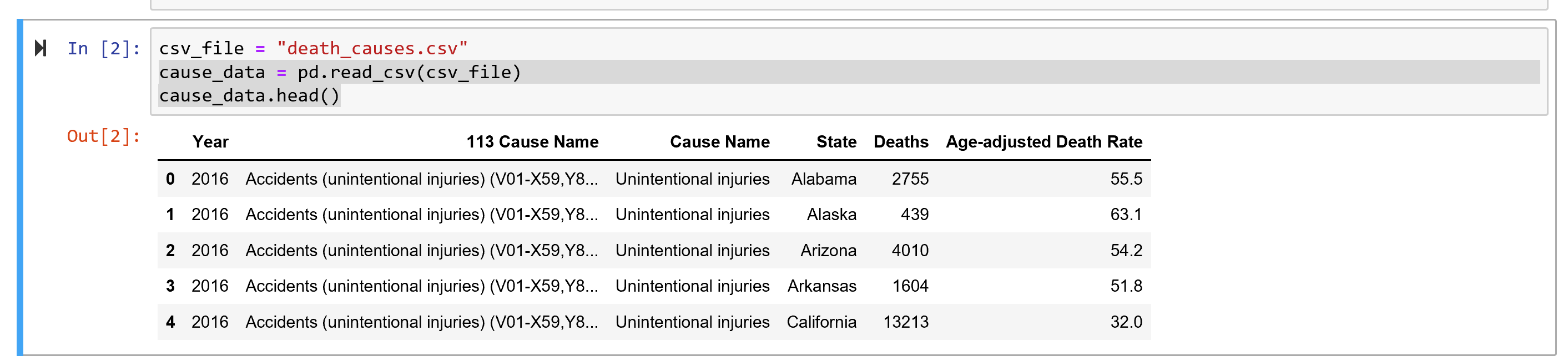
* Resources used to find the appropriate data considered and included
  + [data.world](https://data.world/)
  + [kaggle](<https://www.kaggle.com/>)
  + APIs
  + Data scraping
  + Large Data Sets from:
    - National Center for Health Statistics
    - Environmental Protection Agency
    - Centers for Disease Control and Prevention

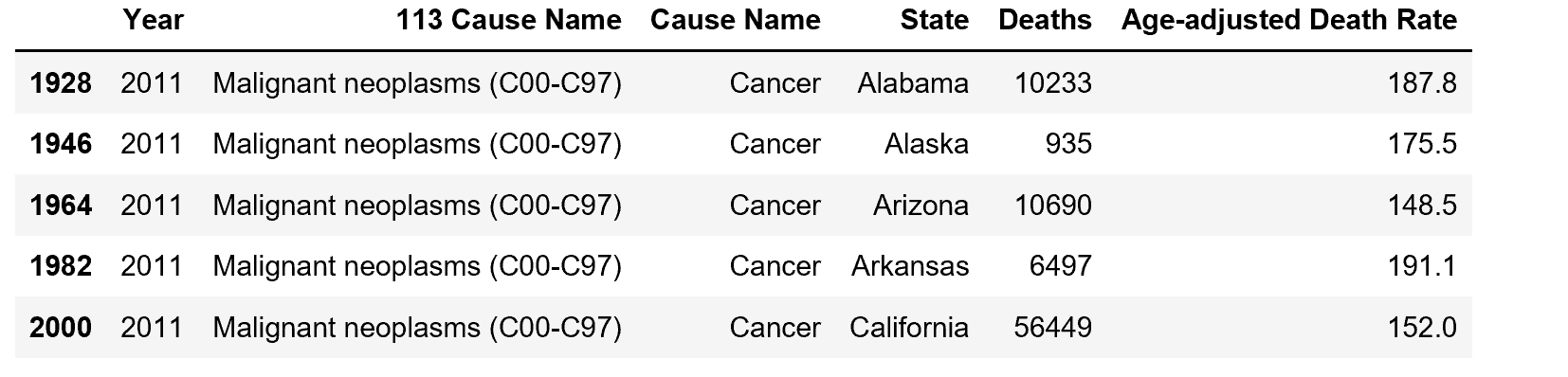
## Import – Extraction

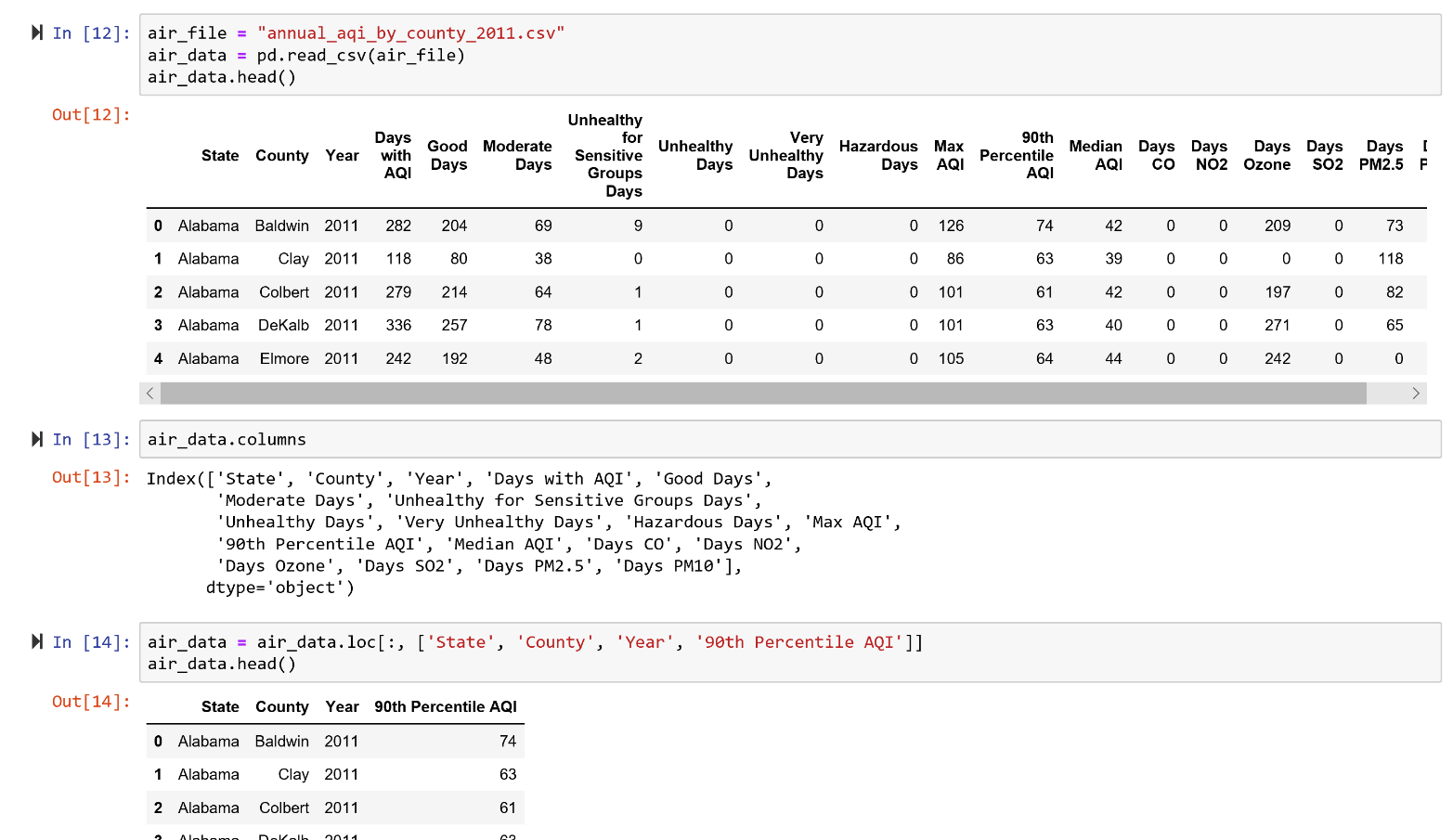
After finding the data needed, more research was done to find files needed for the research, most data was found in CSV format from the respective sites. To start the extraction – transformation, files were moved to Python-Pandas. All data was considered production data. Test data was not considered for this project review. All files had some type of design specification or data model. In some cases, fields have derived information. In those situations, the data was not used in the research. In the detail ETL for each file, handling of derived data is explained.

## Details of ETL work included identifying

* Documenting Files / Summary of type of Data / Source of Data
* Identifying specific parameters used in integration
* Confirming file format (fixed length, comma separated, etc.)
* Using the following ETL tasks before integrating all files
  1. Cleaning
  2. Joining
  3. Filtering
  4. Aggregating
* Structuring data files into relational tables
* Summary of ETL work
* [**NCHS - Leading Causes of Death: United States**](https://catalog.data.gov/dataset/age-adjusted-death-rates-for-the-top-10-leading-causes-of-death-united-states-2013) This dataset presents the age-adjusted death rates for the 10 leading causes of death in the United States beginning in 1999. Data are based on information from all resident death information.
* Source:
* Reviewed file concentrating on: data available in file, years available and any missing values…This resolved whether the data was robust for ETL
* Snapshot of the details. Over 10,000 rows of data was available.



* Cause of death parameters included:
  + 'Unintentional injuries'
  + 'All causes'
  + "Alzheimer's disease"
  + 'Cancer'
  + 'CLRD' (Chronic Lower Respiratory Disease)
  + 'Diabetes'
  + 'Heart disease'
  + 'Influenza and pneumonia'
  + 'Kidney disease'
  + 'Stroke'
  + 'Suicide'
* More Panda work to use data of 2011; survey other specific diseases to verify that data was available for each state
* Snapshot of 2011 for Cancer
* 
* Review possibility of an additional file with air quality data. A better file was found and this work was not used. Charts for the data can be found in Jupiter file etl\_project.ipynb in Github.



* **EPA Data** - ETL of Toxic Release Data:

Extraction: This data set is EPA’s data on the amount of toxic chemicals produced by a collection of sites they monitor throughout the United States. The data can be found on their website separated by years in csv format. The csv for year 2011 was read into a pandas DataFrame.

Transformation: The original dataset had 100+ columns describing location of the various sites as well as describing all the different disposal methods each site used for their respective chemicals. The main columns focused on for analysis were the ‘State’ and ‘Prod.\_Waste’.

Load: This toxic release data grouped by state was then pushed to sqlite database called ‘etlproject.sqlite’ within the ‘toxic\_release\_resources’ folder of the github repository.

**Air Pollution Data from Environmental Protection Agency (EPA)**

Extraction: This data set is EPA’s data on the air quality data collected at outdoor monitors throughout the United States. The data can be found on The link <https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual> separated by years in csv format. This report provides Air Quality Index annual summary information, including maximum AQI (Air Quality Index) values and the count of days in each AQI levels like good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy,and hazardous. And also, data listed information on factors contributing to the pollution like CO, NO2, Ozone, PM2.5, PM10. The csv for year 2011 was read into a pandas DataFrame.

Transformation: The original dataset had 20 columns describing different types of air pollutants, number of days with different quality levels. The main columns focused on for analysis were the ‘State’ and ’90th Percentile AQI’. The data was according to counties in states. But because our death causes information was according to the states, we had to take the average of the air pollution data of the counties in that state.

Load: The air quality data grouped by state was then pushed to sqlite database called ‘etlproject.sqlite’ within the ETLproject folder of the github repository.

|  |  |  |
| --- | --- | --- |
| **Air Quality Index Levels of Health Concern** | **Numerical Value** | **Meaning** |
| **Good** | **0 to 50** | **Air quality is considered satisfactory, and air pollution poses little or no risk.** |
| **Moderate** | **51 to 100** | **Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.** |
| **Unhealthy for Sensitive Groups** | **101 to 150** | **Members of sensitive groups may experience health effects. The general public is not likely to be affected.** |
| **Unhealthy** | **151 to 200** | **Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.** |
| **Very Unhealthy** | **201 to 300** | **Health alert: everyone may experience more serious health effects.** |
| **Hazardous** | **301 to 500** | **Health warnings of emergency conditions. The entire population is more likely to be affected.** |

## Possible Improvements

* Our data in Air Pollution had data entries for counties. But our Death Causes data was limited to state level info. So, we had to take the average of pollution information, and toxicity information for the stated. This made us loose sensitivity.

## Conclusion

* After searching correlation between Air pollution data and Death counts caused by Cancer, Heart Disease, and Influenza and Pneumonia, we couldn’t see a direct correlation in any of the data.
* Through visual analysis of scatter plots, there was no direct correlation between toxic production release and cancer, heart disease and influenza.

Supporting Details –

### NCHS – Leading Causes of Death: United States

## Metadata Updated: August 20, 2018

This dataset presents the age-adjusted death rates for the 10 leading causes of death in the United States beginning in 1999. Data are based on information from all resident death certificates filed in the 50 states and the District of Columbia using demographic and medical characteristics. Age-adjusted death rates (per 100,000 population) are based on the 2000 U.S. standard population. Populations used for computing death rates after 2010 are postcensal estimates based on the 2010 census, estimated as of July 1, 2010. Rates for census years are based on populations enumerated in the corresponding censuses. Rates for non-census years before 2010 are revised using updated intercensal population estimates and may differ from rates previously published. Causes of death classified by the International Classification of Diseases, Tenth Revision (ICD–10) are ranked according to the number of deaths assigned to rankable causes. Cause of death statistics are based on the underlying cause of death. SOURCES CDC/NCHS, National Vital Statistics System, mortality data (see <http://www.cdc.gov/nchs/deaths.htm>); and CDC WONDER (see [http://wonder.cdc.gov](http://wonder.cdc.gov/)).

REFERENCES

National Center for Health Statistics. Vital statistics data available. Mortality multiple cause files. Hyattsville, MD: National Center for Health Statistics. Available from: <https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm>.

Murphy SL, Xu JQ, Kochanek KD, Curtin SC, and Arias E. Deaths: Final data for 2015. National vital statistics reports; vol 66. no. 6. Hyattsville, MD: National Center for Health Statistics. 2017. Available from: <https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_06.pdf>.

URL: <https://data.cdc.gov/api/views/bi63-dtpu/rows.csv?accessType=DOWNLOAD>

**From the dataset abstract**

This dataset presents the age-adjusted death rates for the 10 leading causes of death in the United States beginning in 1999. Data are based on information from all resident death information.

Source: [NCHS - Leading Causes of Death: United States](https://catalog.data.gov/dataset/age-adjusted-death-rates-for-the-top-10-leading-causes-of-death-united-states-2013)