NetIDs: akh70, agb76, pnl8, kie3

Group Members: Adrian Hertel, Alan Balu, Chau Le, Kate Eisert

COSC-287: Intro to Data Science

Data Science Project 2: Analysis of CDC and EPA Data

Nov. 8, 2019

Exploratory Analysis

Basic Statistical Analysis and Data Cleaning Insight (min 10 attributes)

	mean	median	standard deviation
1. AgeAdjustedRate	472.86	475.4	34.0
2. CaseCount	29,891.24	20,892	31,494
3. Population	5,954,454	4,215,080	6,653,496
4. SUM_REL_EST	11321716.86	2291852.90	26712177.95
5.AVG_REL_EST	64610.40	18873.51	146971.05
6.MIN_REL_EST	168.19	0.022	3312.35
7.MAX_REL_EST	2560723.95	590000.0	6717419.24
8.STD_REL_EST	261056.38	72549.39	658185.35
9.VAR_REL_EST	501318891973.72	5263413888.68	5501990289083.74
10. (from CDC API)	68.313	21.6	107.97

Average Annual Age Adjusted Rate from 2010-2014			

For the CDC’s USCS data, we performed a basic statistical analysis on AgeAdjustedRate, CaseCount, and Population to find the mean, median, and standard deviation of these attributes; since these values were not categorical, we did not calculate mode. Additionally, we did some basic statistical analysis on four rates of cancer (average annual prevalence, average annual age adjusted rate, average annual crude rate, and crude prevalence) to see the overall rates for which cancer diagnosis methods were used and which types of cancer were diagnosed.

From the CDC’s USCS dataset, every numeric attribute aside from year (AgeAdjustedRate, lci, uci, CaseCount, and Population) had the potential to contain outliers although this aggregated data was already relatively clean when we first collected the data from the CDC. However, using a calculated z score, we were able to determine the presence of a small percentage of outliers in the three most important and relevant attributes we examined from this dataset: AgeAdjustedRate, CaseCount, and Population). After cleaning by using the pandas dropna() function, with a lower z score threshold of -2.5 and an upper z score bound of 2.5, the number of outliers in the aforementioned attributes were 12, 36, and 30 respectively out of 918 rows of data. Though outliers have the potential to negatively affect the conclusions of a statistical analysis, we chose not to remove these outliers as they are not necessarily abnormal or bad values; the greater deviation can be explained by differences in state size and population, which are natural.

-CDC API:

The CDC data we accessed via the API contained incidence and mortality rates for specific cancer types for each year and state combination. While this is more detailed and potentially interesting data than the USCS data, we realized upon closer inspection that the CDC API data only contained incidence mortality rates for a year range 2010-2014. The data rows for specific years didn't include rates on cancer types but rather on types of tests, which was not of interest to us for this study. We decided limited number of rows in this data set that were on cancer types were not specific enough because they weren't tied to a specific year, so we decided to use only the CDC USCS data for the rest of our analysis in this project.

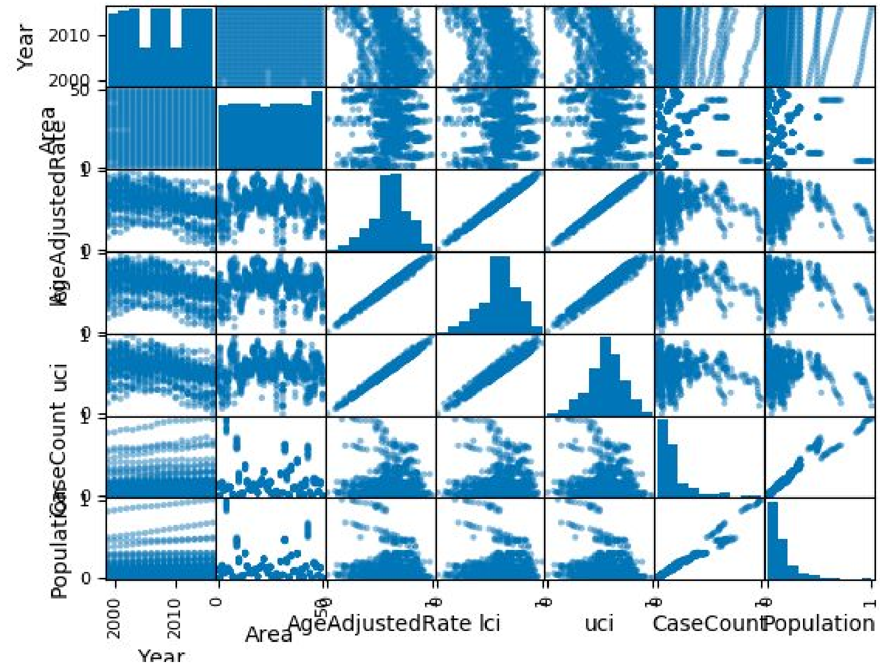
-EPA:

During Project 1 we downloaded two different EPA datasets. The larger one had estimates for every chemical, state, year, release type combination. We decided to use the small data set which included estimates for every state, year, release type combination, adding up all the individual chemicals release amounts. This smaller data was more manageable and still contained enough granularity to get meaningful insights. Our selected EPA dataset had several different statistics for the release estimate amount ranging from standard deviation to sum and average release. The average represented the best estimate of the actual amount released so we selected this as our only statistic of interest.. This gave us final EPA data set. We then normalized the average release estimate by calculating the z-score and then removing any with a score over 2.5 or under -2.5. We chose to delete the outliers rather than replace them because they could create a relationship when there isn't one.

When running analysis on the merged data we deleted any rows where the cancer rate was null and set all null values for the release estimates to zero since no estimate is equivalent to no pollutants released in that state in that year.

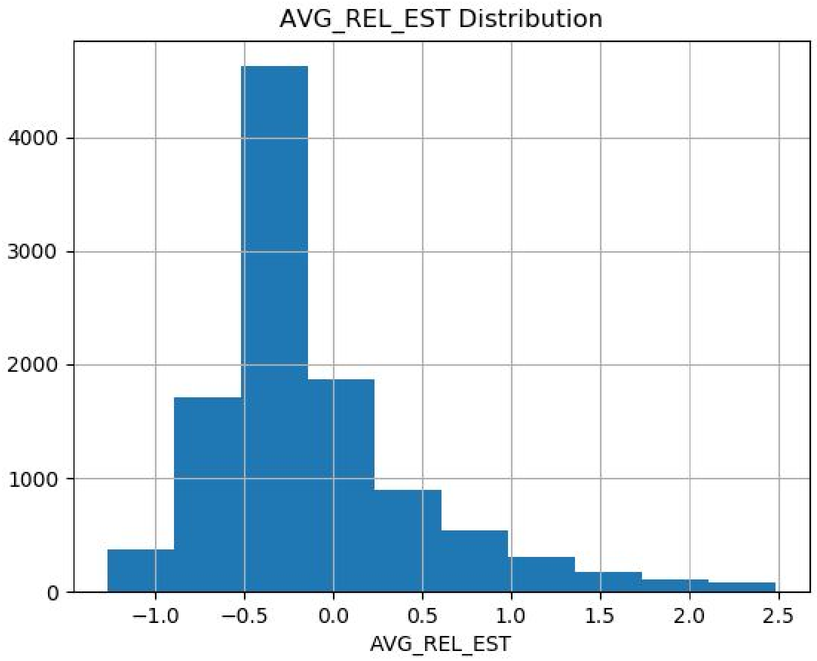
Histograms and Correlations (min 3 attributes)

CDC USCS: We generated histograms for the CDC USCS data to analyze the distribution of factors like AgeAdjustedRate, CaseCount, Population and others from the dataset. As expected, AgeAdjustedRate was the only attribute with a normal distribution around the mean, whereas other attributes were more skewed due to differences in size of states and population, for which the distributions for population and case count were very similar, with a strong left skew. The data was normalized prior to generating the histograms.

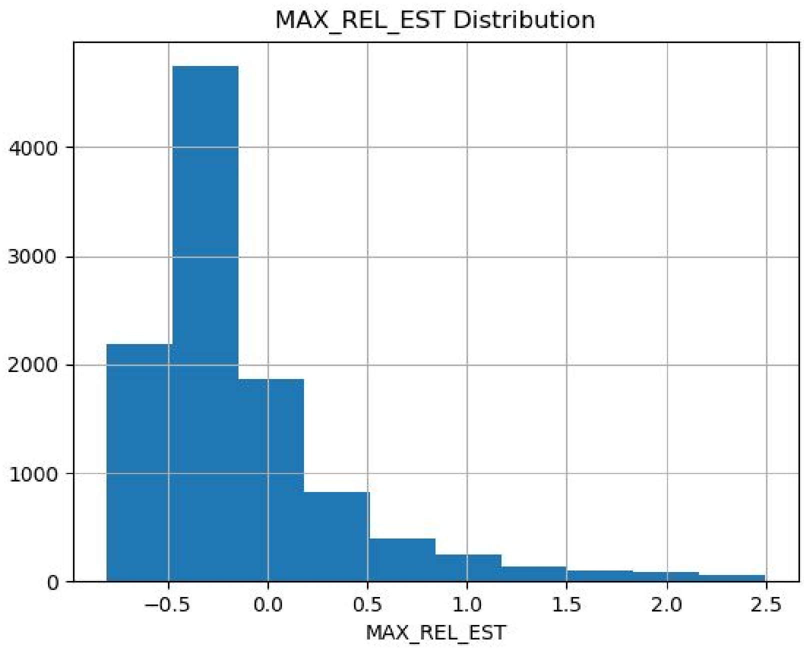


EPA Data Histograms

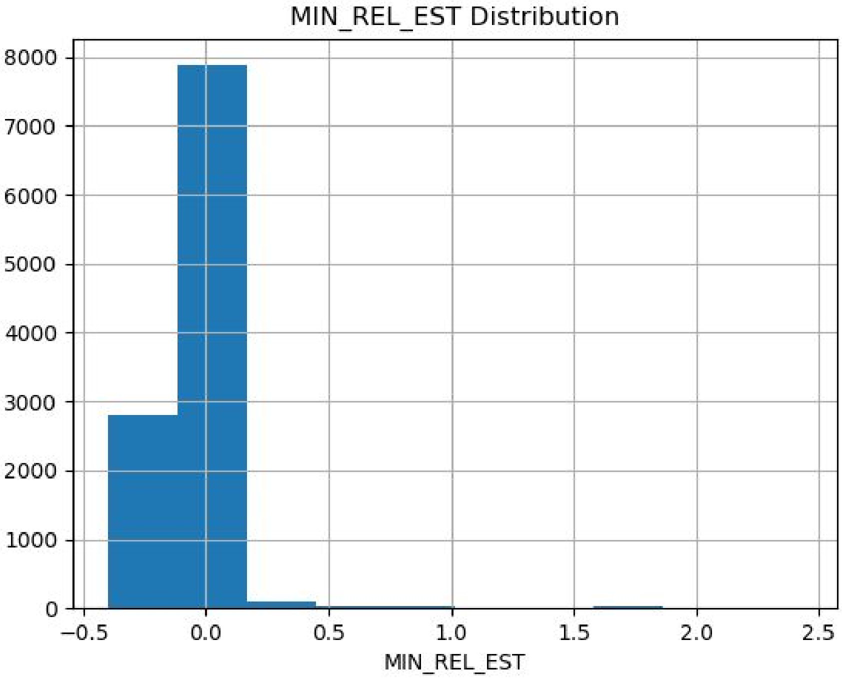
Histograms were created for numerical columns of the EPA data. Below are histograms that show the distribution of the release estimate columns (average release estimate/AVG_REL_EST, maximum release estimate/MAX_REL_EST, minimum release estimate/MIN_REL_EST, and sum of release estimates/SUM_REL_EST). Before creating histograms, the values were normalized based on the category of chemical release. For example, 'AIR STACK' was normalized against all other 'AIR STACK' values, 'WATER' was normalized against 'WATER' values, etc. This is because we noticed that some categories of chemical release had significantly higher or lower values than others. After the data was normalized, it was binned using equal width binning. We chose equal width binning because this was the best binning strategy to observe distribution. Whereas equal depth binning would split the data into bins with nearly equal numbers of data points per bin, equal width binning shows where the data seems to be concentrated.

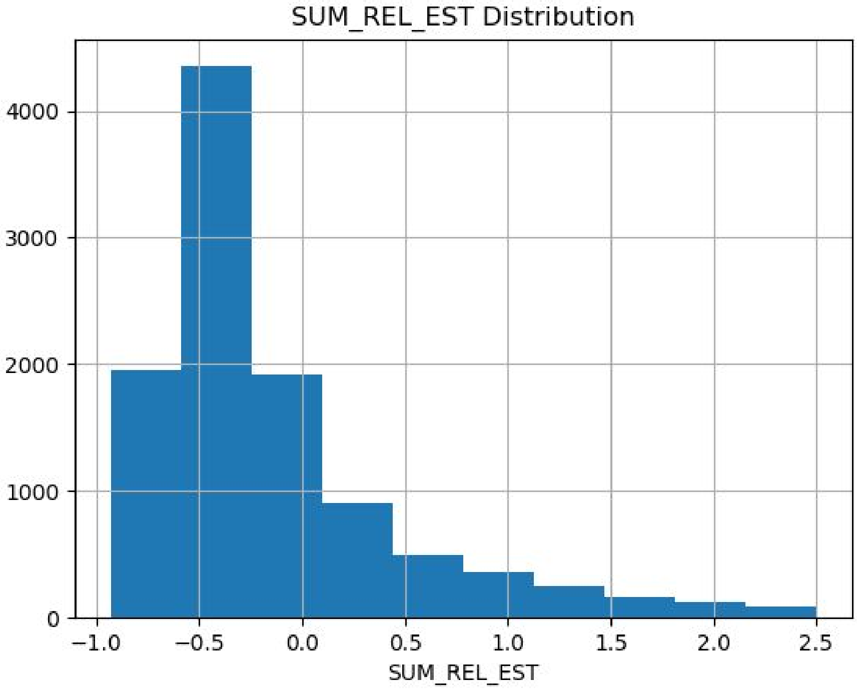


AVG_REL_EST Distribution



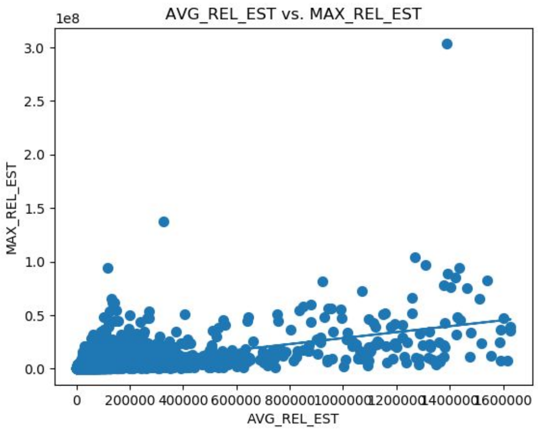
MAX_REL_EST Distribution

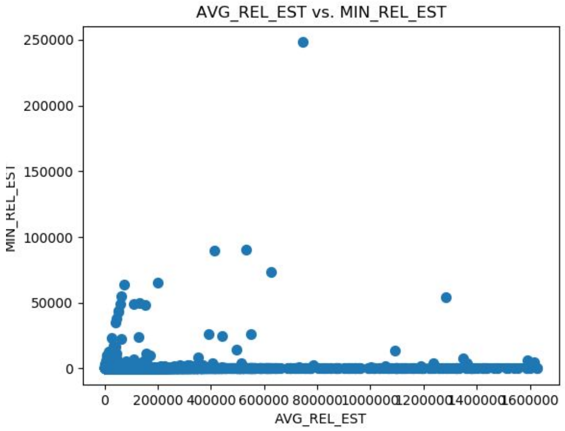


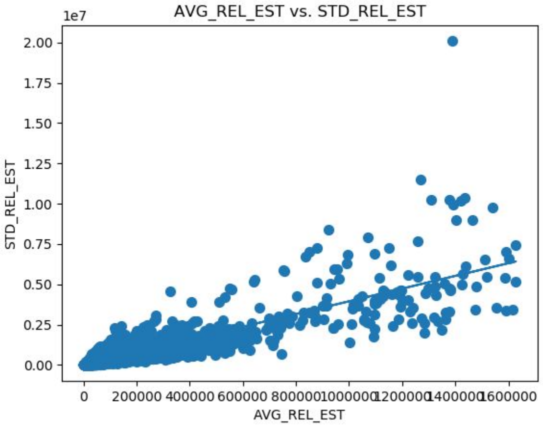


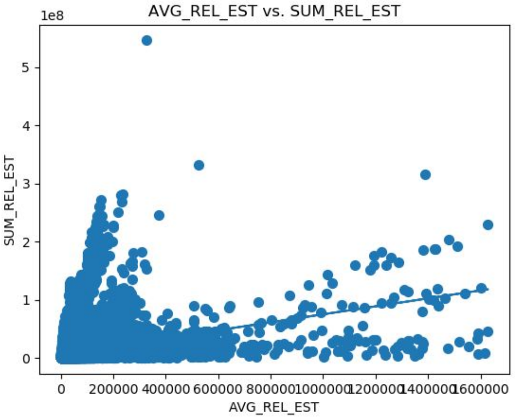
EPA Scatter Plots

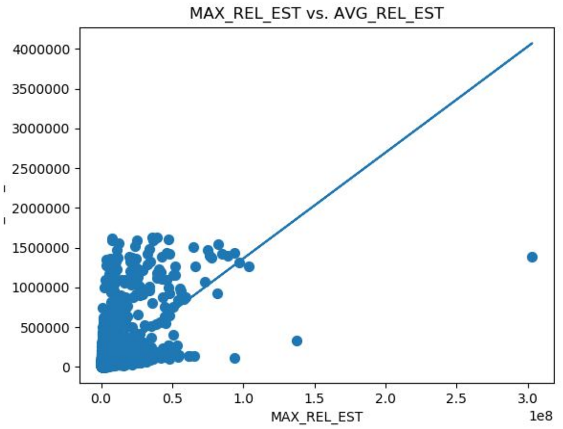
Scatter plots were made for all of the numerical EPA data. Each numerical feature was compared to every other numerical feature, and scatter plots were created. A linear regression test was performed on each pair of features and the linear regression line was added to each plot. Below the scatter plots given is a list of the feature pairs and their respective linear coefficients.

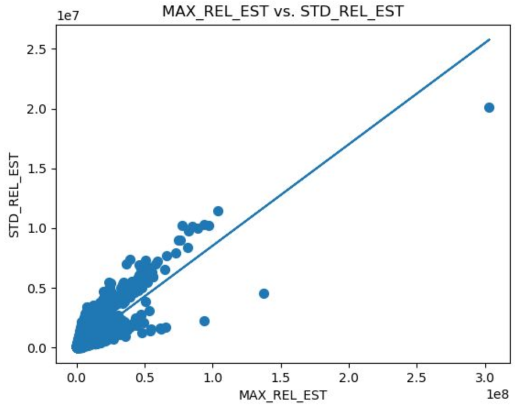


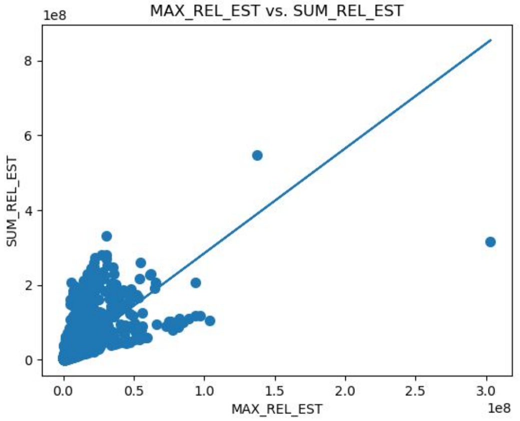


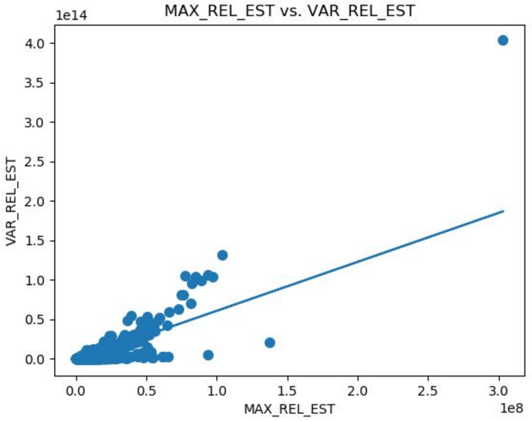






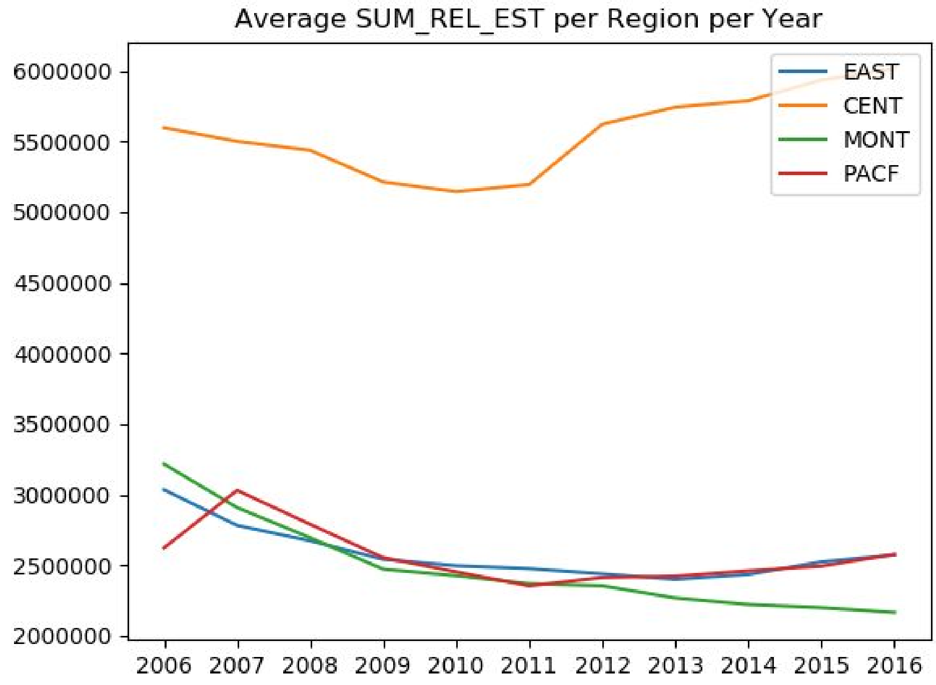






Coefficients of Determination for Each Linear Regression:
('var1, var2'): coefficient
('SUM_REL_EST', 'AVG_REL_EST'): 0.14231642268163192
('SUM_REL_EST', 'MIN_REL_EST'): 0.00037924426557700563
('SUM_REL_EST', 'MAX_REL_EST'): 0.49643711476225005
('SUM_REL_EST', 'STD_REL_EST'): 0.24282410702364185
('SUM_REL_EST', 'VAR_REL_EST'): 0.07244975998705028
('AVG_REL_EST', 'SUM_REL_EST'): 0.14231642268163192
('AVG_REL_EST', 'MIN_REL_EST'): 0.007584707266518387
('AVG_REL_EST', 'MAX_REL_EST'): 0.3708808352709746
('AVG_REL_EST', 'STD_REL_EST'): 0.7748507368263734
('AVG_REL_EST', 'VAR_REL_EST'): 0.2811732405387668
('MIN_REL_EST', 'SUM_REL_EST'): 0.0003792442655768946
('MIN_REL_EST', 'AVG_REL_EST'): 0.007584707266518165
('MIN_REL_EST', 'MAX_REL_EST'): 0.00016763703484112824
('MIN_REL_EST', 'STD_REL_EST'): 0.00013231474612351857
('MIN_REL_EST', 'VAR_REL_EST'): 1.7520325967801398e-08
('MAX_REL_EST', 'SUM_REL_EST'): 0.49643711476225016
('MAX_REL_EST', 'AVG_REL_EST'): 0.3708808352709746
('MAX_REL_EST', 'MIN_REL_EST'): 0.00016763703484112824
('MAX_REL_EST', 'STD_REL_EST'): 0.749044711642258
('MAX_REL_EST', 'VAR_REL_EST'): 0.5722349540784928

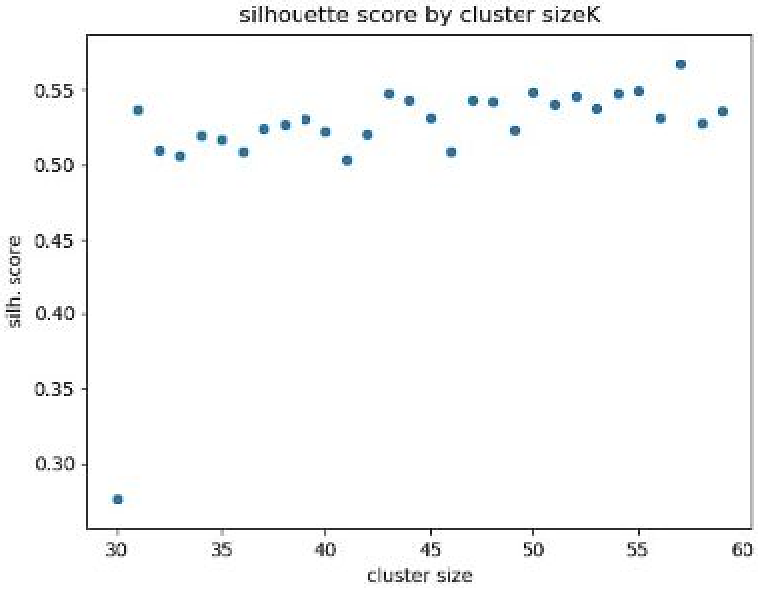
```
('STD_REL_EST', 'SUM_REL_EST'): 0.2428241070236421
('STD_REL_EST', 'AVG_REL_EST'): 0.7748507368263734
('STD_REL_EST', 'MIN_REL_EST'): 0.0001323147461236296
('STD_REL_EST', 'MAX_REL_EST'): 0.749044711642258
('STD_REL_EST', 'VAR_REL_EST'): 0.6124975670450109
('VAR_REL_EST', 'SUM_REL_EST'): 0.07244975998705039
('VAR_REL_EST', 'AVG_REL_EST'): 0.2811732405387669
('VAR_REL_EST', 'MIN_REL_EST'): 1.7520326189846003e-08
('VAR_REL_EST', 'MAX_REL_EST'): 0.5722349540784928
('VAR_REL_EST', 'STD_REL_EST'): 0.612497567045011
```



Cluster Analysis (min 3 attributes)

Results for USCS_CancerTrends_OverTime_ByState.csv

K means Clustering



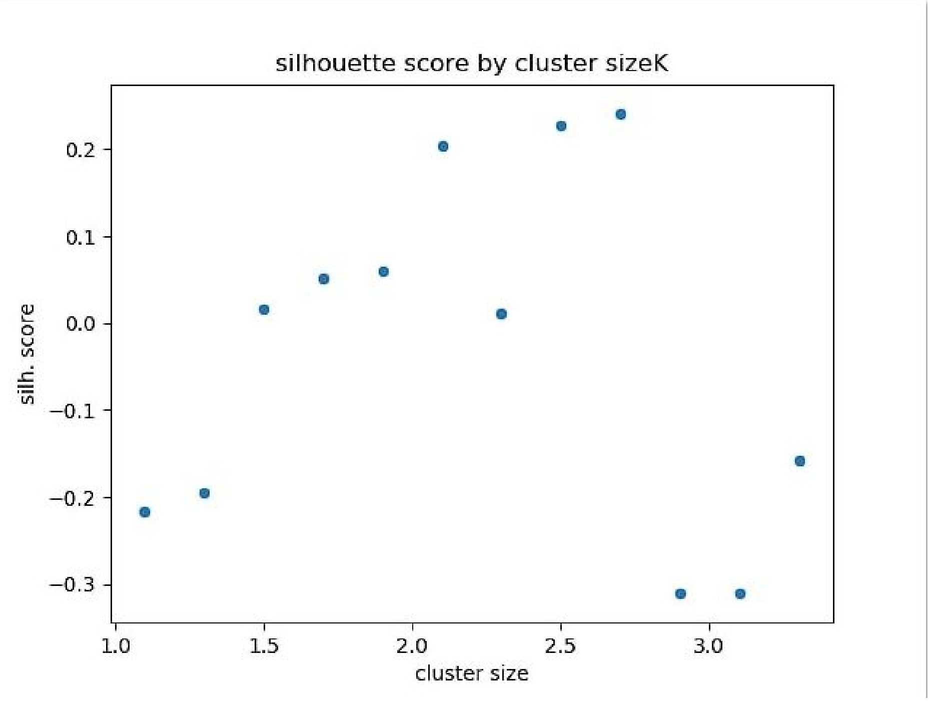
The K means clustering displays that from a range of k = 30 to k = 59, the silhouette score jumps at about 31, and then stabilizes to near ~0.53. This probably indicates that once the cluster size was large enough to include one feature that contained a lot of the data variation, the clustering was improved.

The “rate by rate bin”, “rate by region”, “rage by state”, and “rate by year” figures (from modifying lines 132-143 in CDC_USCS_clustering.py to comment in the right function call for the right type of clustering and commenting out the other lines) display the clustering based on the maximum silhouette score after scanning through different values of k. This value was 57 with a silhouette score of 0.567. Since the colors in the figures referenced above vary a lot for each of the features and do not indicate which feature the clustering was particularly based on. However, the “rate by state” figure seems to indicate that the clustering was heavily depended on the state since many states have cancer rates that are in the same cluster (same color).

Hierarchical Clustering

The Hierarchical Clustering gave very similar results to the K means clustering with a best clustering of 55 and a max silh. Score of 0.568. The figure that are produced after running this type of clustering, give very similar clusterings as the K means.

DBSCAN



The DBSCAN clustering worked differently and gave a different number of clusters as the best. This was 30 clusters with a silhouette score of 0.241. This silhouette score was significant lower than the other clustering methods, likely due to the lack of form to the data in a multidimensional space and the lack of very high density regions. This might indicate that DBSCAN produces a worse clustering, but looking at the "rate by state" figure displays that the clusters seem to be based on the states but not necessarily by the regions we created (seen in the “rate by region” figure).

Overall, this indicates that the clustering for K means and Hierarchical were similar, but differed from DBSCAN, likely due to the inherent differences in the algorithms.

Association Rules/Frequent Itemset Mining (min 3 different support levels)

Results for association rule mining at different support levels on CDC_API_Clean.csv data

```
support  0.1
association rules results:

Index(['yearstart', 'question', 'locationdesc', 'datavalue', 'region',
       'datavalue_level'],
      dtype='object')
"frozenset({'2010'}) support: 1.0"
"frozenset({'2010', 'CENT'}) support: 0.31683168316831684"
"frozenset({'EAST', '2010'}) support: 0.43564356435643564"
"frozenset({'2010', 'MONT'}) support: 0.13861386138613863"
"frozenset({'2010', 'PACF'}) support: 0.10891089108910891"
"frozenset({'high', '2010'}) support: 0.12252475247524752"
"frozenset({'2010', 'low'}) support: 0.10767326732673267"
"frozenset({'2010', 'medium'}) support: 0.7215346534653465"
"frozenset({'2010', 'medium', 'CENT'}) support: 0.2264851485148515"
"frozenset({'EAST', '2010', 'medium'}) support: 0.3378712871287129"
support  0.3
association rules results:

Index(['yearstart', 'question', 'locationdesc', 'datavalue', 'region',
       'datavalue_level'],
      dtype='object')
"frozenset({'2010'}) support: 1.0"
"frozenset({'CENT'}) support: 0.31683168316831684"
"frozenset({'EAST'}) support: 0.43564356435643564"
"frozenset({'medium'}) support: 0.7215346534653465"
"frozenset({'2010', 'CENT'}) support: 0.31683168316831684"
"frozenset({'EAST', '2010'}) support: 0.43564356435643564"
"frozenset({'2010', 'medium'}) support: 0.7215346534653465"
"frozenset({'EAST', 'medium'}) support: 0.3378712871287129"
"frozenset({'EAST', '2010', 'medium'}) support: 0.3378712871287129"
```

```
support  0.6
association rules results:

Index(['yearstart', 'question', 'locationdesc', 'datavalue', 'region',
       'datavalue_level'],
      dtype='object')
"frozenset({'2010'}) support: 1.0"
"frozenset({'medium'}) support: 0.7215346534653465"
"frozenset({'2010', 'medium'}) support: 0.7215346534653465"
```

Analysis:

The patterns that are most frequent should describe the most common association among terms within the data set. As shown above, the Region, Year, and Cancer Rate Level (very high, high, medium, low, very low) seems present in nearly all the associations involving multiple terms. The patterns like {'2010', 'medium', 'CENT'} and {'EAST', '2010', 'medium'} are common because most cancer rates for states in the data set are close to the mean for all the states, and thus have a z-score near 0. This gives that rate a “medium” bin label, so the fact that “medium” was present in many of the association rules is not surprising and very much expected. 2010 was also common in the association rules, but since all the data analyzed had 2010 as the “yearstart” column value, this was expected and again, not surprising. As a whole, the association rule mining displayed that most of the records have a medium value for the cancer rate, which is evident in the fairly “normal” and narrow distribution of the cancer rates. This is likely the case because the cancer rates used in the association rule mining analysis was averaged for the years 2010-2014 for each state, reducing variations in the data across years.

Predictive Analysis - Part 1

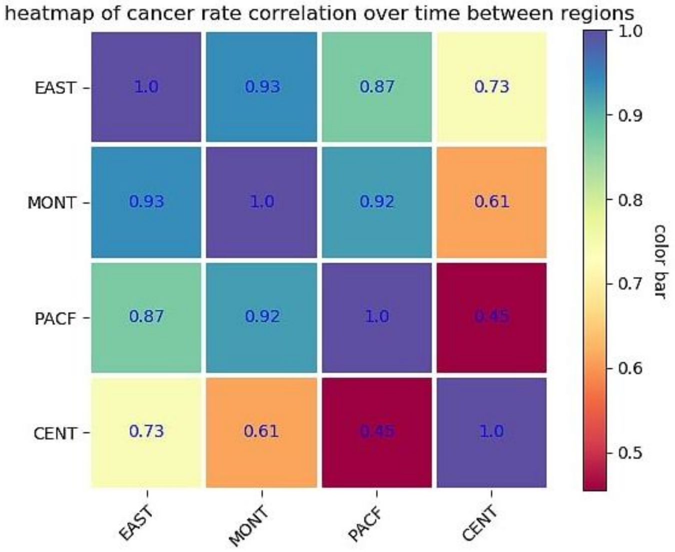
Hypothesis Testing and Classification (min 3 hypotheses and 6 required methods)

A one-way ANOVA test was conducted on the EPA data for each of the numerical release estimate values (sum release estimate, average release estimate, minimum release estimate, maximum release estimate, standard deviation of release estimate, variance of release estimate). For each ANOVA, we compared the variables across regions. We went through the list of state abbreviations. For each state, we summed the variable in question (regardless of year) and kept a count going. After we had gone through the data, we averaged each state's respective data (so AL had its own average value for the variable, AK had its own average value for the variable, etc. After state averages had been found, we appended each state's average value to a list for its corresponding category (AL's average would be appended to a list for central states, AK would be appended to a list for pacific states, etc.). Then, the lists of regional state averages (east, central, mountain, pacific) were used for the ANOVA. Below is a list of the f- and p-values of each such ANOVA for the numerical release estimate values.

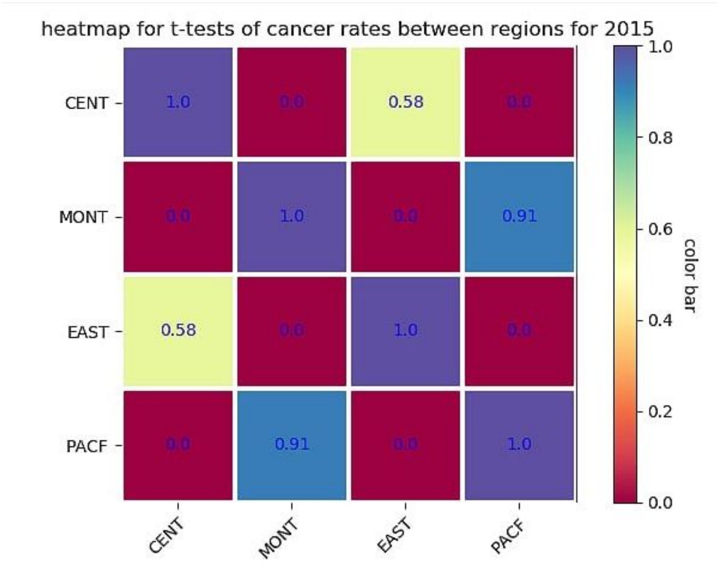
```
ANOVA:
 SUM_REL_EST:
  f = 2.2505161291555935, p = 0.09337170582578744
 AVG_REL_EST:
  f = 5.961170711843655, p = 0.0014265098898864354
 MIN_REL_EST:
  f = 1.3002492698140564, p = 0.28423153876045876
 MAX_REL_EST:
  f = 1.6730642974354004, p = 0.18414096083453033
 STD_REL_EST:
  f = 4.384545322213638, p = 0.007966714393700787
 VAR_REL_EST:
  f = 2.805689433051109, p = 0.04866841183331489
```

The null hypothesis for each EPA data ANOVA test was that there is no statistically significant difference between the regions' means for the variable in question. The alternative hypothesis for each test was that the means were, in fact, statistically significantly different.

We found that if we choose a significance level of 0.05, then our results for AVG_REL_EST, STD_REL_EST, and VAR_REL_EST are statistically significant, whereas those for SUM_REL_EST, MIN_REL_EST, and MAX_REL_EST are not. Therefore, for SUM_REL_EST, MIN_REL_EST, and MAX_REL_EST, we are unable to reject the null hypothesis. However, for AVG_REL_EST, STD_REL_EST, and VAR_REL_EST, we can reject the null hypothesis. This means that the regions' means for each of these variables are statistically significant, so there is significant variance of these variables across the different regions.



Heat map showing results of linear regressions between regions for cancer rate over time. Each box shows the r^2 value between the region on the x-axis and the region on the y-axis. R^2 values closer to 1 indicate higher correlation.

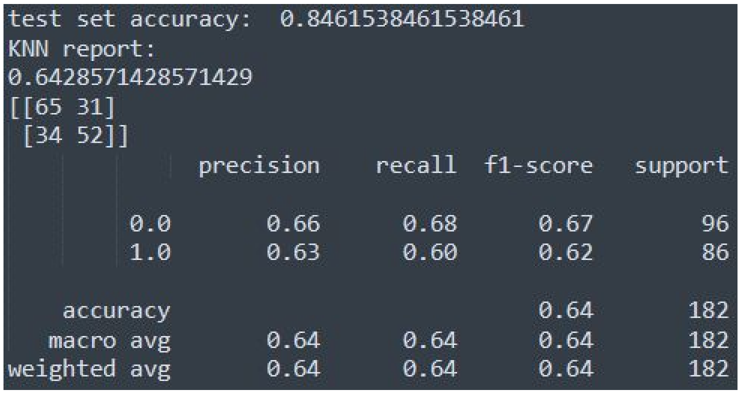


T-test between regional average cancer rate for the year 2015. Low p-values indicate that there was a significant difference between the samples. This indicates that regions differed in their cancer rates in a statistically significant manner.

Naive Bayes and Random Forest

These predictive models were used to determine whether or not there was a correlation between the rate of toxin release and rate of cancer diagnosis using the data we collected. For these two predictive models, we used year, state, toxin release rate and the accuracy of these factors in predicting if a state's age-adjusted rate of cancer is very low, low, medium, high, or very high. With these factors, the NB algorithm had an accuracy score of close to 0.93 using a test size of 20% of the total dataset, which we considered to be a relatively good performance on this multiclass classification. For the random forest, we received an accuracy very close to 1, which may have been an effect of the attribute selection. Consequently, we concluded that the naive bayes algorithm performed better for this dataset given the attributes we chose.

K Nearest Neighbors Classification:



```
test set accuracy:  0.8461538461538461
KNN report:
0.6428571428571429
[[65 31]
 [34 52]]
              precision    recall  f1-score   support

         0.0       0.66      0.68      0.67        96
         1.0       0.63      0.60      0.62        86

    accuracy                           0.64       182
   macro avg       0.64      0.64      0.64       182
weighted avg       0.64      0.64      0.64       182
```

The KNN classifier worked fairly well on the training set, but not so well on the test set with a test set of 80% of the data. This indicates, along with the confusion matrix and “ROC CURVE KNN” plot, that the KNN classifier did not predict the class label that well, but better than a random classifier, seen in the ROC curve plot. Importantly, in this case, we tried to classify whether the cancer rate was above or below the mean for all the states.