# Exploring Inhalation Toxicity of Chemical Compounds

## 1 Introduction

This section is where you describe your project to the general reader. Clearly explain what your project is about, what your main question(s) are, and what you will do to investigate your question(s).

Mention where you obtained your data set from (put the link in the References section at the bottom, and reference the link by number), the size of the data, what the outcome variable is and how it is measured. Explain what are the variables you will consider as predictors and what they measure. You can summarize if there are a lot of predictors. Mention if you plan to do "feature selection": the case where you have a lot of predictors, and you will use models to identify the most significant predictors.

> Our study on fragrance toxicity involves two datasets: one containing chemical compounds known to be toxic when inhaled, and another consisting of fragrance compounds with unknown toxicity. We sourced these datasets from PubChem, a chemical database maintained by the National Institutes of Health (NIH), and the Environmental Protection Agency (EPA). Initially, we obtained a list of chemicals from PubChem and then cross-referenced it with the EPA database to gather a comprehensive dataset that includes the relevant variables.
> The dataset of toxic compounds totals roughly 1500, while the unknown dataset totals roughly 3000. We aim to identify predictors for inhalation toxicity through analysis of both datasets. Due to the unknown nature of the non-toxic dataset, we plan to use k-means clustering as a method to explore the toxic dataset, with "Ames Mutagenicity Test Pred" as our variable of interest. In chemistry, the Ames Test is used to discern the capacity for a chemical to trigger cell mutations (mutagenesis), which may lead to cancer in humans. A positive Ames test value indicates higher mutagenicity, while a negative test value indicates no genotoxicity. Other variables we may consider for further study include Henry's Law Constant, vapor pressure, water solubility, and logP, as these properties are correlated with compound inhalation toxicity.

## 2 Data Processing

In this section, explain what steps you took to deal with any missing values and extreme outliers. Make sure you state how many instances you found and how many you removed from the data set.

> While reading in both datasets, we used the na.strings function to identify rows containing N/A in the dataset and convert them to NA's understood by R. Next, we removed extraneous columns irrelevant to our study and changed categorical columns (such as names of compounds) to factors. Finally, we removed NA values and extreme outliers for some variables, though as we are dealing with chemical compounds a wide variance in values still remained.

## 3 Data Exploration

In this section, you explore how the outcome variable is distributed. use the appropriate methods depending on the data type: descriptive stats (min,max, mean, median, sd) and histogram/boxplot for continuous, barplot and frequency table for categorical. You always want to include a visual display- a plot, and a quantitative display- a table with numbers for each variable's distribution.

IMPORTANT: For all plots, hide the R code, use the height and width parameters to adjust the appearance of the plot in the knitted file (you can to this with some trial and error), and be sure to add a fig.cap that describes the plot. ALL plots and tables in your report must be uniquely numbered and captioned.

> First, we analyzed the distribution of Ames test predictions in the dataset. The histogram plot and summary table below for the Ames Mutagenicity Test Predictions show that the data is right-skewed, with the largest portion of observations having a test prediction between 0-0.2. This indicates that the majority of observations in the toxic chemicals dataset have slight mutagenetic properties.

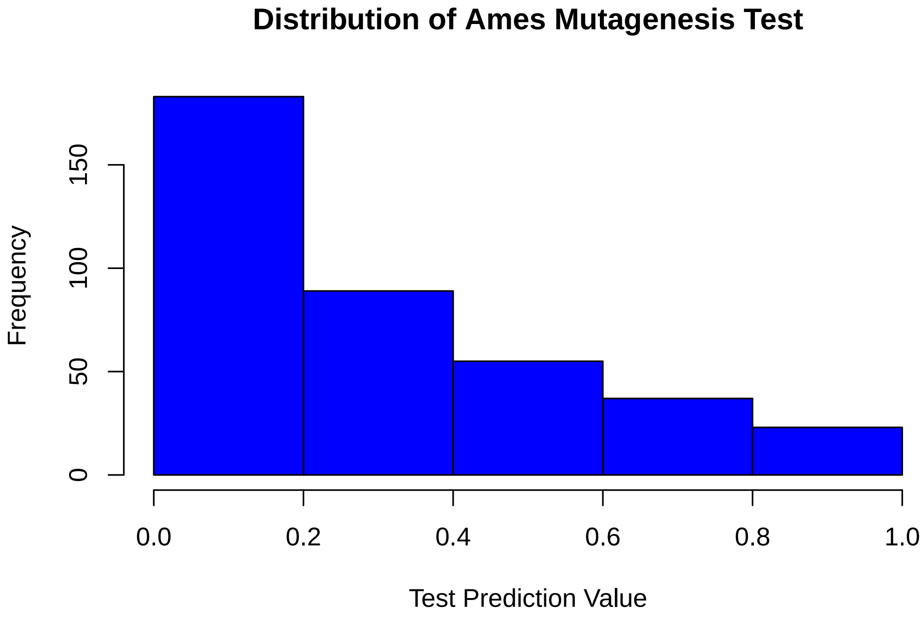


Figure 1: Distribution of Ames Mutagenesis Test

Make sure you describe in words what the plots and tables mean. You are telling the reader the "story" of how you carried out the analysis to investigate your question(s) you described in the beginning.

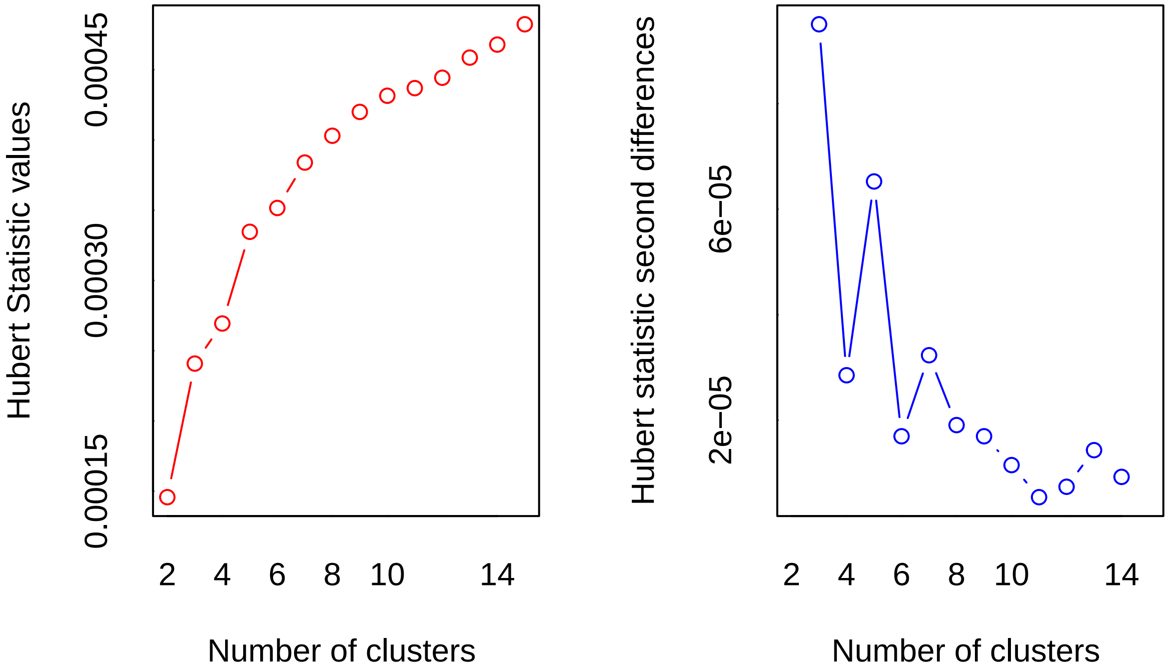
Remember: NO RAW R OUTPUT in your report. Instead, use plots, tables or text. Example: the output from the summary function in R should not be shown in the report. Use a table instead and put the values in manually.

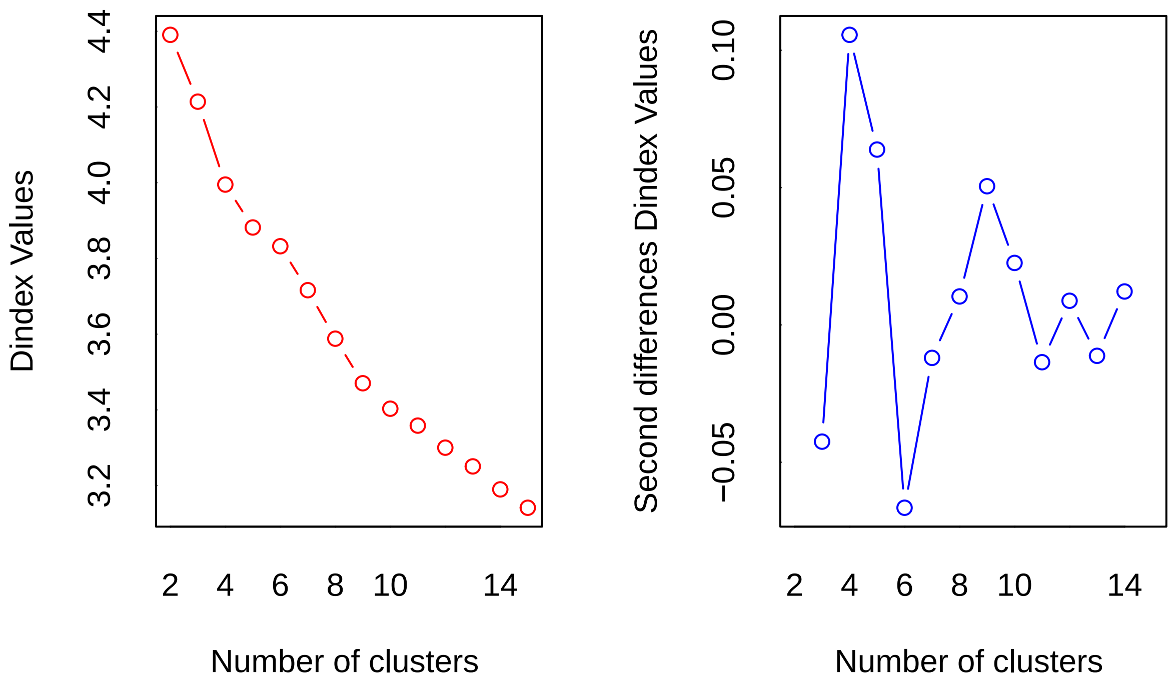
Any significant predictors that you find (via model selection) or that are of secil interest should also be explored with a plot and or table. You don't have to exlore every predictor, just those that are especially interesting.

# 4 Training and Testing Data Sets

Explain how you are making up the training and testing sets in this section. Mention the proportion of each and how you are creating them, e.g. by random sampling.

To create proper clusters, we first removed all categorical variables from the dataset, assigning the result to a new dataframe, inhalation2.data. Then, we scaled the results as there is wide variety between predictor values. To determine ideal parameters for clustering, used the NbClust package to automate the ideal kmeans cluster amount. The results below indicate an efficient cluster number would be around 6-8, where the graph comes to a “point,” or “elbow.”





# 5 Modeling

In this section, you describe the types of models you will fit and evaluate to investigate what they say about your question(s). You must use two different types of models to investigate each of your research questions.

## 5.1 K-Means Clustering

Include a sentence or two that explains what predictors and outcome you used to fit the model, and how you are training it: cross validation, how many folds, etc. and any tuning parameters as applicable.

> To fit a k-means cluster model, we passed in the scaled inhalation data, instructing the model to use 7 clusters. We decided to use a low nstart value of 5 due to the large variance in the dataset even after scaling, which distributed the data between clusters more meaningfully.

### 5.1.1 Model 1 Fit Results

Describe the fit output, and display or summarize the results of the fit. You may use a table or text to display or describe the variables that were important (significance/effect). If you did model selection, just describe the end model's fit, but include all code that you wrote to do the selection.

Include any plots that show the residuals were normally distributed, such as with linear regression.

The size of each cluster varies widely, with the smallest cluster having only 10 values, and the largest cluster with 126. The measure of distance within clusters is greater than the measure of distance between clusters, meaning there is likely no significant trend among values in the dataset as a whole.

Table 1: Size Of Each Cluster

| Cluster Amount | Cluster |
|:---|---:|
| 10 | 1 |
| 11 | 1 |
| 19 | 1 |
| 61 | 1 |
| 69 | 1 |
| 91 | 1 |
| 126 | 1 |

Table 2: Measure of Within Distance Of Each Cluster

| Measure | Freq |
|:---|---:|
| 6748.92906918926 | 1 |

Table 3: Measure of Between Distance Of Each Cluster

| Measure | Freq |
|:---|---:|
| 5217.07093081074 | 1 |

### 5.1.2 K-Means Clustering Prediction Performance

Explain how you are evaluating this model on the test data. What metric are you using to evaluate it's performance. Did you use a confidence or prediction interval? Categorical: confusion matrix. Show the results in a table and/or plot.

We then analyze the fit.inhalation model in terms of our variable of interest, Ames Mutagenicity Test Prediction. The table below analyzes the scaled means for each cluster, while the plot illustrates the clustering model itself. The table and plot appear to reflect the data above- there are certainly enough differences between values as a whole to create clusters, but there is still greater variety within clusters than between them. This may indicate the data itself is generally similar- as these are all toxic compounds with some degree of mutagenicity.

```
## Means for Each Cluster of Ames Test:
```

| Means | Cluster_Number |
|---:|---:|
| -0.2011762 | 1 |
| 0.0912903 | 2 |
| 0.4243295 | 3 |
| 0.0517261 | 4 |
| 0.1528265 | 5 |

| Means | Cluster_Number |
|---|---|
| -0.5350234 | 6 |
| -0.5355170 | 7 |

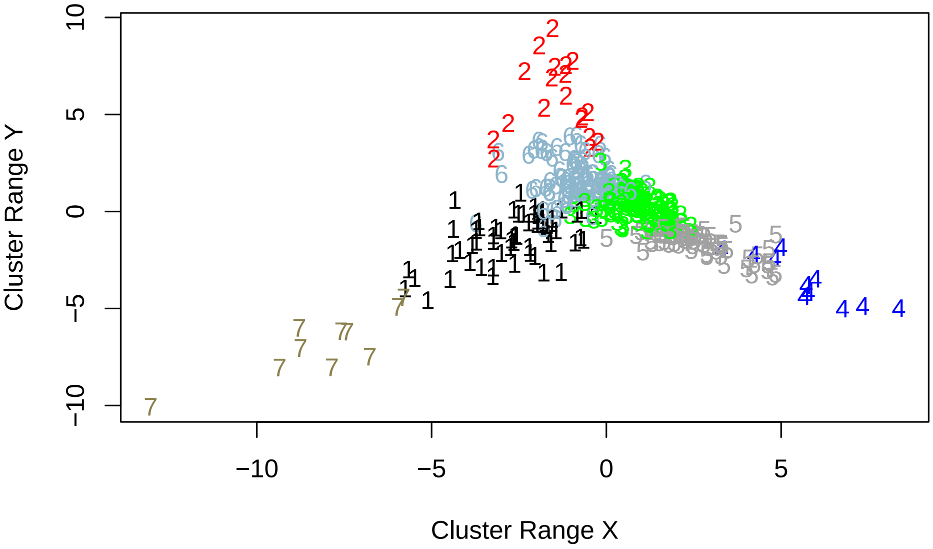


Figure 2: K-Means Clustering For Ames Test

Use tables with a caption and number.

### 5.2 Classification Model

We plan to use a logistic regression model once we identify which compounds in the unknown dataset are toxic when inhaled. After labeling, we will merge the known and labeled datasets, removing the 194 overlapping compounds. The combined dataset will then be shuffled randomly and split into training and testing sets. Our goal is to predict a binary outcome: “toxic” or “non-toxic.” We will evaluate the model’s performance based on its accuracy and misclassification rate, and identify which predictors are most influential in determining inhalation toxicity.

## 6 Conclusions

In this section, describe how the results informed your investigation of your research question(s). Did the models help you answer your questions? If not, why do you think that is? Did the results produce any unexpected results or aspects of the data? Did you get any insights from your work that suggest interesting directions to take? How would you go further if you were to continue with this analysis?

You can also mention any difficulties you faced and any threats to the validity o your conclusions.

# 7 References

This section contains a numbered list of any data sets or data sources you used, and any other references you accessed.

1. PubCHEM Datasets:
   https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72&search=Fragrances
   https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72&search=Inhalation+Risk

2. EPA Datasets:
   https://comptox.epa.gov/dashboard/batch-search