

# Basic Training 2: Data Identification & Analysis

## Teacher's Workbook



A Program of The Actuarial Foundation

**Modeling the Future  
Challenge**

## The Modeling The Future Challenge

As part of the Scenario Phase of the MTFC, teams will be demonstrating and applying their mathematical analysis skills to a scenario response paper as well as identifying a potential project and writing a proposal. The Actuarial Process Guide to be an invaluable resource.

- [The Actuarial Process Guide](#)

## How to Use this MTFC Data Identification & Analysis Scaffolding Guide

When all of the potential topics in the world are at your fingertips, identifying a topic, identifying possible risks, finding sources, mathematically modeling, and studying risks can seem overwhelming to begin. This guide will help scaffold the process and guide participants through the process of identifying data and completing data analysis (both self-directed and through guided questions) of the scenario response and subsequent MTFC Project. Each task refers to a specific section of The Actuarial Process Guide for more in-depth information.

**Content:** The process is scaffolded into 4 total tasks.

**Suggested pacing:** 1 task per week for 4 weeks.

## Common Core Standards for Mathematical Practice

The [Common Core Standards for Mathematical Practice](#). The MTFC Project Proposal specifically addresses the following standards:

- ❑ CCSS.MATH.PRACTICE.MP1 **Make sense of problems and persevere in solving them.**
- ❑ CCSS.MATH.PRACTICE.MP2 **Reason abstractly and quantitatively.**
- ❑ CCSS.MATH.PRACTICE.MP3 **Construct viable arguments and critique the reasoning of others.**
- ❑ CCSS.MATH.PRACTICE.MP4 **Model with mathematics.**



## 2.1 Categories of Data, Cleaning Data, and Finding Data

Refer to Sections [2.1](#), [2.2](#) and [2.3](#) of the [Actuarial Process Guide](#) for more detailed information on each of these areas of consideration.

Types of data to look for with MTFC projects:

1. **Defining historical trends** – These datasets may be used to explore a behavior or phenomenon. They may be descriptive of a once-occurring phenomena.
2. **Projecting future trends** – These datasets are similar to historical trends, captured over time, allowing for explanatory possibilities.
3. **Separating potential outcomes** – These datasets provide more detail about behaviors and phenomena, and may present an opportunity to aggregate outcomes by samples, populations, and products.
4. **Defining the severity of potential losses** – These datasets place a value on a good resource. These losses may be private, to an individual or organization, and/or social costs, which attempts to also consider the cost of externalities that are unaccounted for within the free market (e.g., carbon/climate datasets)
5. **Defining the frequency of potential outcomes** – These datasets offer an opportunity to understand how frequently an event occurs.

Points 1 and 2 can be quite similar in regards to the dataset types that are used to define historical trends and project future trends. Some helpful types of research study designs to consider in these situations when searching for data will be longitudinal or cross-sectional datasets. A cross-sectional study collects data from a specific population at a particular point of time whereas a longitudinal study is a correlational research study considering multiple variables over an extended period of time.

Considerations in Data Scope:

- 1) **Cleaning the data** - check for completeness and correctness. Watch for missing data which may need to be excluded or mismatched labeling which could lead to incorrect sorts and counts.
- 2) **Adjusting the data** - historical monetary data may need adjusted for inflation if compared to current or future trends. Consider the distribution of the data - could there be outliers that need to be excluded or accounted for?
- 3) **Data credibility** - how much data is there? Is the sample size large enough for the scope of the project?
- 4) **Efficacy** - is the source of the data reliable and trustworthy? Were the methods for obtaining the data described? Is there anything that would call into question the validity of the data upon further research?

Make sure to keep track of any modifications and adjustments as you go (make copies of the data set so the original is left untampered if the analysis goes awry).

## 2.2 Working with Datasets

Not all data is equal. One of the key skills for students in many fields of STEM is to understand how to work with datasets. This primer provides a few key points for students to consider when they are trying to find, organize, and manipulate the data for their Modeling the Future Challenge project.

- 1. Sorting and separating:** sometimes students who are not used to working with large amounts of data in a spreadsheet may get overwhelmed when they see a very big spreadsheet. The first thing you may want to do is to simply sort the data by relevant columns, and if it is a very large sheet, separate the data into different sheets based on categories of data. This can help the data feel less overwhelming.
- 2. Removing NULL values:** many datasets may have NULL values included where there was an error with data collection, or some other issue that caused the data to not be there. In this case, many datasets will set a NULL value that can be easily identified in the data. For example, a NULL value might be set to -9999. Before analyzing the data or doing any manipulation, you will need to make sure to remove these NULL values from consideration.
- 3. Cleaning data:** as with NULL values, sometimes datasets are not perfectly clean either, meaning that they may include errors from people mis-typing the data or accidentally moving things around. If the dataset is not too large, it may be useful to do a quick review of the spreadsheet and make sure that there are no obvious errors in the data.
- 4. Only export relevant data:** in working with large online datasets there is often a lot of data available, particularly with government data. There are often many variables and features that may or may not be useful for your project. In downloading this data, only download the variables and data that is immediately relevant to your project. Use filters in the data access portals wherever available to limit what you download.
- 5. Use Pivot Tables:** it can be very valuable to rearrange your data into different groups or categories. Pivot tables are a very useful way of doing this. Both Excel and Google Sheets have good Pivot Table functions. A good basic tutorial can be found online here:  
<https://support.office.com/en-us/article/create-a-pivortable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576>




## Task 2.1: Exploring and Identifying Types of Data

While some robust, longitudinal (panel and time-series) datasets may fit within all five categorizations, there are many single-study, cross-sectional datasets that contain data that may not have the same explanatory power, but may be useful in the exploratory phases of a project.

### Task topic: water quality in the USA

Along with your partner(s) for this exercise and using the datasets listed below, identify a few datasets that fit each of the 5 categories that either are aligned with this sample task topic (water quality in the USA) or a topic of their choosing for their project proposal. What ideas and stories might they be able to convey through these datasets? What is still lacking or what needs to be cleaned before moving forward?

- Modeling the Future Challenge website (<https://www.mtfchallenge.org/data-sources/>),
- or datasets from sites like data.world. (<https://data.world/datasets/insurance>),
- and Github (<https://github.com/awesomedata/awesome-public-datasets>)
- or data archives such as Inter-University Consortium for Political and Social Research, ICPSR, (<https://www.icpsr.umich.edu/web/pages/ICPSR/index.html>),



### Data That Defines a Historical Trend

Find and link some possible data/datasets.

What data story could be told with it?

What limitations are there to this data set?

What is the topic of the data set?

### Data That Projects a Future Trend

Find and link some possible data/datasets.

What data story could be told with it?

What limitations are there to this data set?

What is the topic of the data set?

### Data That Separates Outcomes

Find and link some possible data/datasets.

What data story could be told with it?

What limitations are there to this data set?

What is the topic of the data set?

### Data That Defines the Severity of Potential Losses

Find and link some possible data/datasets.

What data story could be told with it?

What limitations are there to this data set?

What is the topic of the data set?

### Data That Defines the Frequency of Outcomes

Find and link some possible data/datasets.

What data story could be told with it?

What limitations are there to this data set?

What is the topic of the data set?

## 2.3 Performing Initial Data Analysis

It can be daunting to know how to start data analysis once a data set is obtained. Some things to look for to get started and consider for an initial analysis could include:

- Look for meaningful measures:
  - measures of center (mean, median, mode, midrange)
  - measures of spread (range, variance, standard deviation)
- Look for distribution shapes or interesting trends
  - creating charts, graphs, summary tables
- Look for regression or trendlines
- Look for outliers or how to quantify them
  - Recall that outliers are often identified the following ways:
    - $> 1.5 \cdot IQR$  above Q3 or below Q1 (Recall that the IQR is the “interquartile range” or distance between Quartile 1 & Quartile 3 in the 5-number summary for a box-and-whiskers plot)
    - $>2$  (or 3) standard deviations above/below the mean
- Look for likelihoods or probabilities
  - Recall the conditional probability formula:  $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Look for Expected Value:  $EV = \sum P(x_i) \cdot x_i$ 
  - expected value is the “long term average” of a variable and is found by summing the product of each outcome and its associated probability



## Scenario Task 2.2: Water Quality

### Scenario description:

The U.S. Environmental Protection Agency is a department of the federal government committed to protecting human health and the environment in the United States. More frequently, the EPA has been dealing with issues of lead contamination in water that can cause a host of mental and physical health problems.

The EPA was recently called to perform a statistical analysis in five suburbs of Cincinnati, Ohio, over complaints of water pollution in the nearby Ohio River. 150 toxicology tests on the blood of randomly selected children living in each area are shown below, measuring the lead level in micrograms per deciliter of blood ( $\mu\text{g}/\text{dL}$ ). The water quality of each household was also tested by measuring the lead content of the water sources (sinks, showers, etc.).

### Data summary:

You are provided with 750 observations total with 150 from each of the five suburbs (Mariemont, Mason, Montgomery, Sherwood, and Terrace Park). The blood content was measured from a simple blood test from a child (2-17) selected from each township, and the lead content of each respective household is measured in parts per million (ppm). Each sample is identified by its suburb, its water lead level, and the lead content in the blood of a child living in the household.

Link for Dataset: [water\\_data](#)



### Task 2.2.1: Self-directed Exploratory Data Analysis

Given the data set, perform an initial data exploration and analysis. Describe what you computed, created and observed in the analysis and include meaningful values or tables.

Operation / Calculation / Chart	What is the value / interesting observations / potential usefulness?





### Task 2.2.2: Exploratory Data Questions for Further Analysis

Now that you have completed some initial analysis, it's time to think of some driving questions to explore further data analysis. Some questions to consider (not an exhaustive list):

- comparisons / outliers / observations / optimization between different categories
- cut-off values for guidelines and/or proportions of data falling into different categories
- likelihoods of events
- effects/sensitivity of adjustments to the computed data values
- considerations on whether the dataset answers the questions you have or if additional data is needed

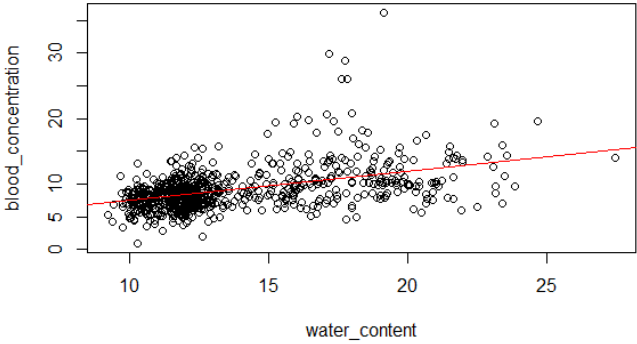
Exploratory Questions	Could it be found within this data set? Explain.

### Task 2.2.3: Guided Data Questions for Further Analysis

Answer the following analysis questions posed by an actuary who explored this dataset.

Question	Response																		
Which suburb has the highest average lead content in their water? Which suburb has the lowest average lead content in their blood?	Montgomery has the highest average lead content in their water (16.9841 ppm) and Terrace Park has the lowest average lead content in their blood (7.566373 $\mu\text{g/dL}$ )																		
What's the mean of the lead content in blood for each suburb?	<table border="1"> <thead> <tr> <th>Suburb</th> <th>Average Blood Lead Content</th> </tr> </thead> <tbody> <tr> <td>Mariemont</td> <td>8.160787447</td> </tr> <tr> <td>Mason</td> <td>11.08714336</td> </tr> <tr> <td>Montgomery</td> <td>10.97496727</td> </tr> <tr> <td>Sherwood</td> <td>8.082521721</td> </tr> <tr> <td>Terrace Park</td> <td>7.566373426</td> </tr> </tbody> </table>	Suburb	Average Blood Lead Content	Mariemont	8.160787447	Mason	11.08714336	Montgomery	10.97496727	Sherwood	8.082521721	Terrace Park	7.566373426						
Suburb	Average Blood Lead Content																		
Mariemont	8.160787447																		
Mason	11.08714336																		
Montgomery	10.97496727																		
Sherwood	8.082521721																		
Terrace Park	7.566373426																		
What's the variance and standard deviation for lead content in blood of each suburb?	<table border="1"> <thead> <tr> <th>Suburb</th> <th>Standard Deviation Water</th> <th>Variance Blood</th> </tr> </thead> <tbody> <tr> <td>Mariemont</td> <td>2.05533873</td> <td>4.224417294</td> </tr> <tr> <td>Mason</td> <td>3.291869049</td> <td>17.9807373</td> </tr> <tr> <td>Montgomery</td> <td>2.994850499</td> <td>13.86311485</td> </tr> <tr> <td>Sherwood</td> <td>0.5414646505</td> <td>3.115398026</td> </tr> <tr> <td>Terrace Park</td> <td>0.9786926122</td> <td>2.946042428</td> </tr> </tbody> </table>	Suburb	Standard Deviation Water	Variance Blood	Mariemont	2.05533873	4.224417294	Mason	3.291869049	17.9807373	Montgomery	2.994850499	13.86311485	Sherwood	0.5414646505	3.115398026	Terrace Park	0.9786926122	2.946042428
Suburb	Standard Deviation Water	Variance Blood																	
Mariemont	2.05533873	4.224417294																	
Mason	3.291869049	17.9807373																	
Montgomery	2.994850499	13.86311485																	
Sherwood	0.5414646505	3.115398026																	
Terrace Park	0.9786926122	2.946042428																	
What's the mean and standard deviation of the <b>variances</b> for lead content in blood of each suburb? (hint, use sample variance when calculating the standard deviation)	<p>First, find the variance of the data for each suburb:            Mariemont: 4.224417            Mason: 17.98074            Montgomery: 13.86311            Sherwood: 3.115398            Terrace Park: 2.946042</p> <p>The mean of these variances is 8.425942 and the variance of these variations is 49.18524. To find the standard deviation, take the square root of the variance.</p> <p>Mean: 8.425942            Variance: 7.013219</p>																		



<p>What is the probability that a randomly selected child from Mariemont has blood lead content within two standard deviations of the mean blood lead contents for Mariemont?</p>	<p>There is one below the range and 4 above the range. So, 145 out of 150 are within the range, or 0.967</p>
<p>If a lead content of 10 µg/dL or higher requires medical attention, how many kids in Sherwood need to see a doctor?</p>	<p>Put the 'blood' variable in decreasing order and count the number of data points above 10.</p> <p>21 children need medical attention.</p>
<p>What percentage of Mason households have a lead content between 8 and 10 ppm in their water?</p>	$P(8 \leq x \leq 10) = \frac{\sum_{i=1}^{750} x_i * I(\text{suburb}=\text{Mason}, 8 \leq \text{water lead content} \leq 10)}{\sum_{i=1}^{750} x_i * I(\text{suburb}=\text{Mason})}$ <p>I(.) is an indicator function which returns 1 when every condition in its interior is true, and 0 when its interior is false. Alternatively, students can find the probability that the water quality in Mason households is less than or equal to 10 ppm and subtract the probability that the water quality in Mason households is less than 8 ppm.</p>
<p>The EPA is interested in seeing how well they can predict lead content in a child's blood using water quality measurements.</p> <p>Create a scatterplot of the data distribution of water quality versus blood content and graph the least squares regression line.</p> <p>What is the expected lead content of a child living in a household that has 15ppm of lead in its water?</p>	 <p>This question involves simple linear regression utilizing the equation:</p> $\text{Blood content} = 3.00683 + 0.44817 * [\text{Water content}]$ <p>When 15 is plugged in for water content:</p> $\text{Predicted lead content} = 9.729364 \mu\text{g/dL}$