Exploratory Data Analysis Guide

For NPS Inventory and Monitoring Natural Resources Datasets

Table of Contents

# 1. Exploratory Data Analysis for NPS Inventory and Monitoring Data

## 1.1 What is Exploratory Data Analysis?

## 1.2 EDA for NPS Inventory and Monitoring Program datasets

## 1.3 Four EDA steps for exploring NPS natural resources data

# 2. Exploratory Data Analysis Step 0: Dataset Description and Preparation

## 2.1 EDA0 description

## 2.2 EDA0 qualitative investigation of the dataset

## 2.3 EDA0 techniques to illuminate overall data structure and issues

## 2.4 Document the initial EDA

# 3. Exploratory Data Analysis Step 1: Pattern Detection within Variables

## 3.1 EDA1 description

## 3.2 EDA1 techniques to illuminate within-variable patterns

## 3.3 Document patterns by variable

# 4. Exploratory Data Analysis Step 2: Pattern Detection between Variables

## 4.1 EDA2 description

## 4.2 EDA2 techniques to illuminate between-variable pairwise relationships

## 4.3 Document between-variable pairwise relationships

# 5. Exploratory Data Analysis Step 3: Exploration of Topic-specific Patterns

## 5.1 EDA3 description

## 5.2 EDA3 techniques to explore topic-specific patterns

## 5.3 Document topic-specific patterns

# 6. Best EDA Practices for I&M Datasets

List of Figures

## Figure 1: EDA in the NPS I&M workflow

## Figure 2: Example of EDA1 graphical investigation of a response variable

Appendices

## Appendix A: EDA0 Interview your Dataset

## Appendix B: Using autoEDA Packages in R to Explore NPS Natural Resources Datasets

# Exploratory Data Analysis for NPS Inventory and Monitoring Data

## 1.1 What is Exploratory Data Analysis?

Exploratory Data Analysis (EDA) is an approach to investigating a dataset before performing more-traditional analyses, such as hypothesis testing and trends detection. EDA was first widely championed by American mathematician John Tukey beginning in the 1960s and described in his 1977 book *Exploratory Data Analysis*. Tukey argued against focusing data analysis only on pre-planned targeted hypothesis testing; instead he suggested adding an exploratory approach to investigating datasets in order to suggest testable hypotheses that arise from the data themselves. Tukey advocated for the use of summary statistics and visual tools to explore the fundamental structure of a dataset before further analysis.

This EDA approach ideally results in 1) a more comprehensive understanding of the strengths/weaknesses of a specific dataset, 2) an improved understanding of patterns in and between the variables in the dataset and, 3) ideas for non-planned testable hypotheses that arise from the dataset itself.

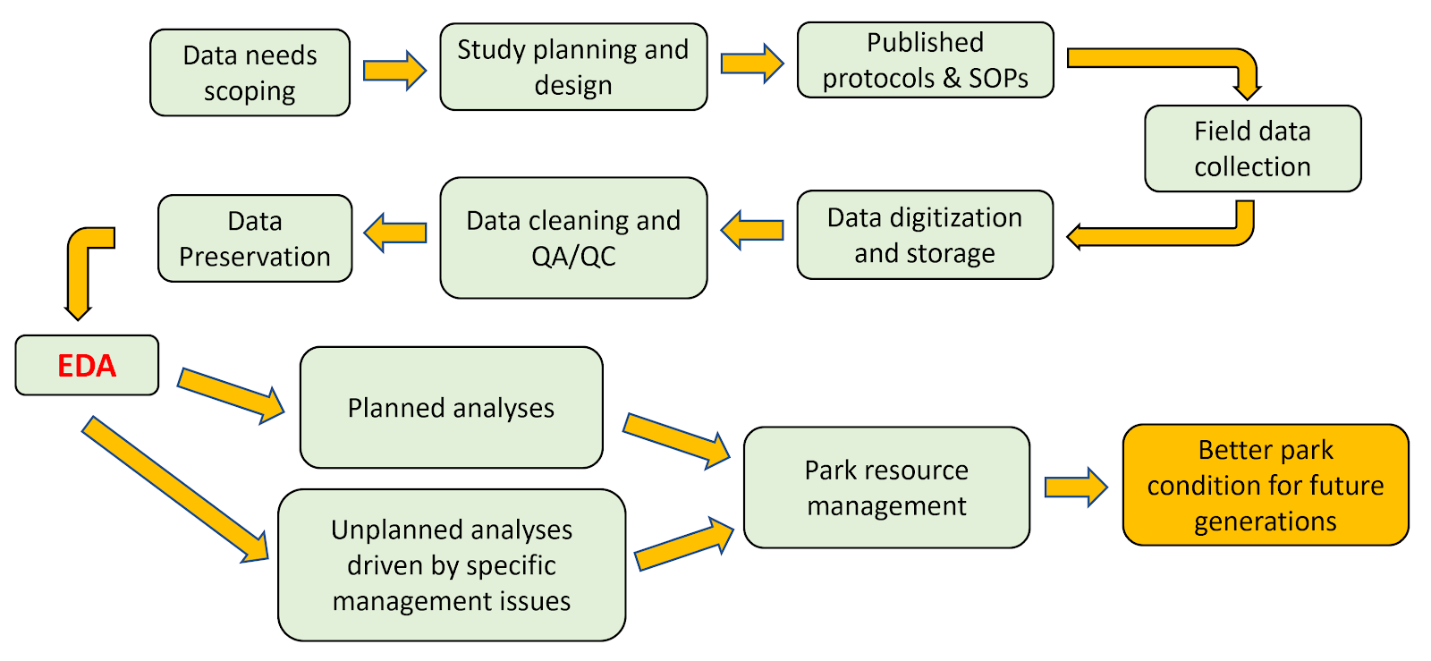
## 1.2 EDA for NPS Inventory and Monitoring datasets

The U.S. National Park Service (NPS) protects the natural and cultural resources of the country’s national parks “for the enjoyment of future generations” (NPS Organic Act, 1916). To help fulfill this mandate, the NPS established thirty-two Inventory and Monitoring (I&M) programs across the country - each dedicated to understanding ecological status and trends in a set of biogeographically similar park units.

In undertaking its mission, I&M program and park staff prioritized which resources should be monitored on an ongoing basis, commonly referred to as “vital signs”. For each of these vital signs the I&M program published protocols outlining monitoring goals and study design, along with associated standard operating procedures (SOPs) that detail data collection and data management procedures. Specific analysis procedures for monitoring data are rarely specified in I&M protocols; rather, they are typically left to be worked out only when targeted data summaries and analyses are scheduled to occur. The SOPs typically outline procedures for seasonal or annual quality assurance and quality control (QA/QC), including data range checks and other constraints built into data entry and storage systems.

However, there is a procedure gap between the steps of seasonal QA/QC checks and planned long-term trend analyses. This gap can be filled by formal EDA – as described by Tukey – to improve broad understanding of a dataset’s strengths/weaknesses, illuminate dataset-wide QA/QC issues to resolve, and suggest targeted analyses to extract meaningful information from the study.

Philippi (2020) placed EDA into the context of a typical I&M workflow for a vital sign monitoring program, which generally proceeds from scoping for information needs to study design and protocol publication, through to data collection, management, analysis, and use. In Philippi’s model (Figure 1), NPS staff would conduct EDA for a dataset before conducting formal trends analysis, in order to improve understanding of the utility of a dataset and provide a quality control function, and also before revising or updating a protocol.

**Figure 1:** EDA in the NPS I&M workflow (adapted from Philippi 2020)

This EDA guide is intended to facilitate a common understanding of EDA techniques among NPS ecologists and data managers and to promote use of EDA to improve the utility of agency datasets. Although this document primarily describes exploration of I&M monitoring data, information here may also be applied to other datasets created by the NPS and its partners.

## 1.3 Four EDA steps for exploring NPS natural resources data

Philippi’s 2020 model describes four steps for EDA for NPS datasets: EDA0 through EDA3. NPS staff would ideally complete these steps sequentially for a specific dataset (i.e., complete EDA0 then proceed to EDA1, etc.). However, staff can complete earlier steps without committing to later steps and still achieve improved understanding of their dataset to inform fieldwork, data management, and potential analyses. Furthermore, managers can complete each step and then pause to use products from the EDA to solicit feedback from colleagues before deciding if/how to proceed to later steps. This review can increase awareness of the strengths/weaknesses in the dataset, suggest a shared work plan for addressing issues with the dataset, and illuminate options for analysis and use of the data.

The four steps are listed below along with a brief description of their intent; subsequent chapters describe each of these in more detail.

### **EDA0: Dataset description and preparation**

Data managers characterize the purpose and status of the study, as well as identifying key response, structural, and supporting variables present in the dataset. In this step managers look for and resolve unexpected discontinuities in structural variables (e.g., site, year) to address before proceeding to EDA1.

### **EDA1: Pattern detection within variables**

Data managers begin basic pattern detection by selecting a set of key response and structural variables that may illuminate relevant patterns in the dataset, then investigating and describing patterns within these variables using graphical tools and summary statistics. This step can provide additional QC and also reveal associations between response and structural variables that may warrant further investigation in later EDA steps.

### **EDA2: Pattern detection between variables**

EDA2 pairs key response variables with each other to illuminate possible bivariate relationships. In this step data manager use graphical tools and summary statistics to explore and communicate about these relationships. As with EDA steps 1 and 0, EDA2 may reveal additional need for dataset QC or suggest avenues for further analysis if patterns are not as expected.

### **EDA3: Exploration of topic-specific patterns**

In this final EDA step, before potentially launching targeted trend detection or hypothesis testing, data managers explore topic-specific patterns and relationships within the dataset, ideally informed by improved knowledge of the intention and design of the study, as well as any weaknesses in the dataset that could create spurious results.

# Exploratory Data Analysis Step 0: Dataset Description and Preparation

## 2.1 EDA0 description

The goal of EDA0 is to gain a comprehensive understanding of the dataset before launching more rigorous analyses. EDA0 documents the purpose and status of the information through a narrative description of the study and associated data, then characterizes patterns in the dataset’s basic anatomy.

The first step in EDA0 – qualitatively exploring and documenting the study purpose and dataset status – may reveal flaws not discoverable within the dataset itself and provides a reference for communicating the state of the dataset with colleagues. While assembling this qualitative description managers identify the key *response*, *structural*, and *supporting* variables in the dataset. The second, data-investigation phase of EDA0 consists of a high-level review of patterns and structure of the dataset. Finally, EDA0 concludes with a subsetting of the dataset to only those data that will move forward for further analysis and documentation of the subsetting steps. These EDA0 steps are interrelated and it will often be beneficial to conduct them concurrently rather than strictly in sequence.

EDA0 is generally performed on whole-protocol or whole-study datasets, rather than data representing a limited subset of sites or years from the whole study. However, in some situations managers may find it useful to restrict the assessment to a subset of a study’s data.

After completing EDA0 data managers should be able to answer the questions: Does the dataset address the study objectives? What are the primary concerns with the dataset to either address or be aware of during further analyses? How well does the dataset meet expectations suggested by the protocol and its SOPs? EDA0 should end with a cleaned dataset ready for further exploration, with obvious errors in structural and response variables resolved.

#### EDA0 is complete when:

* Data managers have a clear understanding of how the data relate to the monitoring objectives, including identification of primary *response,* *structural*, and *supporting* variables
* Flaws in the dataset are flagged and either resolved *or* documented in order to understand how they may impact later analysis; data managers have documented how they reviewed, edited, and/or subsetted the original dataset for subsequent EDA steps
* Two products are completed: 1) a concise, narrative description of the purpose and status of the study, and 2) a dataset that has been reviewed for QA/QC and filtered for only the data that should go forward for further exploration or analysis

## 2.2 EDA0 qualitative assessment of the dataset

EDA0 starts with a qualitative investigation of the dataset in order to articulate the purpose of the study, provide a common reference document to communicate its current status, and document known issues that may not be described in writing elsewhere. We provided an example template for this qualitative assessment in **Appendix A: Interview Your Dataset**.

Some information documented through this investigation will already exist in protocols or reports, but managers can benefit from reviewing, confirming, and briefly consolidating this information before proceeding with further analyses. If managers choose to use the example template in Appendix A, they should feel free to only address the topics that illuminate the purpose and status of the study at hand and ignore topics that are not applicable or useful. Especially important in this step is to document known information about the data and data collection procedures that doesn’t exist in existing protocols, SOPs, and reports. In particular, document any issues that could bias the outcomes of subsequent analyses.

A key step in the qualitative assessment is identifying the dataset’s *response*, *structural*, and *supporting* variables, as described below:

* **Response variables** are metrics the study is designed to track, such as water temperature, native shrub cover, presence/absence of an animal in a habitat area. It is often immediately obvious which data represent a study’s response variables, but not always. For example, if a data logger is collecting water temperature values every fifteen minutes but the study is only interested in the daily high temperature values, the latter may be the true response variable. However, even in such a case it’s ideal to also identify the more granular variables as “response variables” (such as the 15-minute data in the example above) as well as to identify the study’s stated response variables in order to inform quality control investigations and suggest additional questions to ask of the raw dataset. The *response variables* can often be thought of as a study’s *dependent variables*.
* **Structural variables** represent data that indicate the time and place of data collected, as well as any key treatment information (e.g., burned vs. unburned plots). Although these are not of interest in and of themselves, the response variables lose their meaning when divorced from associated structural variables. Patterns within structural variable data are of interest in EDA for quality control investigations and also provide a means to “slice” or bin the response data to look for meaningful information and reduce noise in EDA1 through EDA3. *Structural variable* is often synonymous with *independent variable.*
* **Supporting variables** represent information collected anecdotally to the response and structural variables. These can include unconstrained text information in the form of comments or notes, and also categorical information such as data observer and recorder. These data are often useful when investigating quality control issues arising from unexpected values or patterns.

Some variables may fit into more than one of these three bins. For example, in a tree demography dataset “tree species” may be thought of as a response variable when tracking changing species composition within a monitoring plot; however that same variable may be structural when considering how many trees of each type have died in the past decade.

It may be helpful to only identify the *response*, *structural*, and *supporting* variables that are most key to the intent of the study, rather than exhaustively categorizing each variable in the dataset. Also, there may be an obvious hierarchy within that set of key variables of those that are most central to the purpose of the study. It may be useful to flag these as highest priority for investigation in subsequent analyses.

Finally, while identifying *response*, *structural*, and *supporting* variables is a useful step in obtaining a clear understanding of a dataset and its potential applications, it’s not critical to obtain one final “right” answer about which variables fit in which bins. In some cases, the distinctions will not be clear; more critical is the consideration and documentation of the roles of key variables within the dataset.

## 2.3 EDA0 techniques to illuminate overall data structure and issues

After (or, more likely, concurrent with) the documentation of qualitative information about the study, EDA0 continues with investigation of the dataset itself to better illuminate its basic structure. This step will almost certainly be conducted in a database, spreadsheet, statistical analysis package, or a combination of these tools.

Ideally, this step will be conducted on a dataset formatted to fit within a single table, rather than on a dataset represented by more than one related table. If a dataset exists as multiple tables (for example, stored in a relational database), managers may choose to combine tables to achieve a “flat” single-table structure of relevant response, structural, and support variables.

Alternatively, this step can be conducted on one portion of a larger, interrelated dataset. For example, a vegetation community monitoring program may collect multiple, non-comparable metrics such as shrub cover, total plant species richness, and tree trunk diameters at each observation location and store the associated data in a multi-table relational database. These datasets would ideally be treated separately for early EDA procedures; later EDA steps may be able to usefully compare metrics from separate but related data collection efforts.

After identifying the response and structural variables in the dataset, EDA0 can conduct a basic investigation of the distributions of response variables, through tools such as boxplots, histograms, and summary statistics, as well as an investigation into the patterns in structural variables. This step can provide dataset-wide quality control, illuminate any potential missing data or potential biases in the dataset, and also provide ideas for further investigations in later EDA procedures.

Primary questions to ask of the *response* variables in this step: are distributions as expected? If not, is this due to error or does it reflect patterns to investigate later? Do outliers represent true data (not errors) and if so, are there patterns in timing/location of those outliers that warrant further investigation? Are summary statistics (e.g., mean, median, measures of variability) as expected? How can these data best be “sliced” or binned based on structural variables (e.g., month, treatment, site) to illuminate patterns and minimize noise? Generating ideas about how to “slice” or bin data can start in EDA0; the actual investigation of the data grouped by these bins forms the core of EDA1.

Primary questions to ask of the *structural* variables: did observation take place as planned? If not, why not? Are there any unexpected discontinuities or grouping of the patterns in the structural variables? Do the patterns in data collection (by time, location, and/or treatment) potentially cause any issues for later data analysis? For example, if in early years most observations occurred in March and observations in later years mostly occurred in August, consider if this pattern may impact response data.

After categorizing the dataset’s variables and investigating their basic patterns, data managers should then consider which variables and data to bring forward for further analysis. A resulting subsetted dataset may exclude sites, years, and/or variables. For example, a data manager may exclude data from years or sites that were collected as part of a pilot study or data that exist in the dataset to help with field orientation for data collection. Subsetting procedures and the rationale(s) for subsetting should be described in the narrative, qualitative assessment document.

Regardless of the software employed to organize, summarize, and visualize the dataset for EDA0, there are a set of common descriptors and techniques most useful for this stage of EDA. We list these below to provide a common framework for NPS datasets. Generally, data managers will only perform a subset of these techniques on any specific dataset.

* **Describe dimensions of the dataset**
* Number of variables (usually represented in columns) and number of observations (usually represented in rows)
* **Categorize key variables by data type and by function**
* Numeric, boolean, categorical, character
* Response, structural, or supporting
* **For *structural* variables: investigate structural patterns of data collection**

These can often be investigated with the following plots:

* Histogram
* Frequency table
* Heat map
* **For *response* variables: investigate distributions and summary statistics**

These can often be investigated with the following summary statistics and plots:

* Mean, median, mode, measures of precision, measures of skew and normality
* QQplot
* Missing values
* Histogram
* Frequency table
* Bar plot
* Box or violin plot
* Kernel density estimation
* **For all variables: investigate unique values** in each variable to identify quality control concerns and **investigate patterns in null or missing values** in each variable to identify quality control concerns

Most of these techniques can be performed on a variety of software platforms. We reviewed a set of off-the-shelf EDA packages for R and RStudio that can facilitate these investigations, referred to collectively as “autoEDA” packages (see **Appendix B: Using autoEDA Packages in R to Explore NPS Natural Resources Datasets**). Data managers may find that it’s more efficient to conduct some EDA0 investigations without using automated EDA software. However, EDA software packages, such as those described in Appendix B, can provide overarching summary reports that are not easily produced otherwise. These include:

* Overview reports

These reports, generated by autoEDA packages, provide a high-level summary of an entire dataset from one line of code; the format of these reports differ substantially by package, see Appendix B for examples

* Interactive tables and charts

Some autoEDA packages generate interactive tables and charts from a single line of code that allow managers to investigate a dataset’s structure and potential issues

## 2.4 Document the initial EDA

After conducting initial exploratory analyses outlined in this section, data managers should document the work. The format of this documentation may vary based on the depth of the EDA0 analyses and the dataset, but it may be helpful for communication within the NPS to standardize this documentation to some extent.

Documentation of EDA0 may include, but is not limited to, the following components:

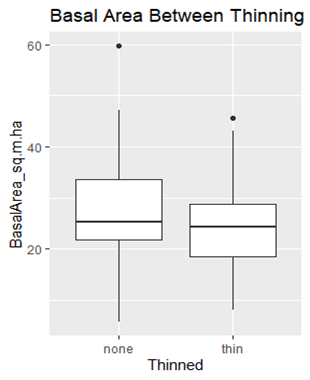
* Narrative discussing the **qualitative assessment of the dataset** (which may use all or part of the Interview Your Dataset template in Appendix A, or other format). This assessment may be updated in later EDA stages, but at the end of EDA0 data managers should have documented the purpose and structure of the dataset as well as known quality control issues and concerns. This assessment should include identification of key *response*, *structural,* and *supporting* variables.
* **Description of the procedures performed on the dataset** after export from its data storage system to prepare the data for further EDA. These procedures will likely be informed by the qualitative assessment and they may require excluding some data deemed as not-reliable or not-relevant for further analysis. In addition, some datasets have specific required QA/QC procedures that should be performed before proceeding; the narrative should confirm that these have been completed.
* A file of the dataset resulting from data-subsetting described in the previous bullet.

# Exploratory Data Analysis Step 1: Pattern Detection within Variables

## 3.1 EDA1 description

The goal of EDA1 is to further explore patterns within key response variables, in order to begin using the data to address study objectives, and to continue with broad QA/QC on the dataset. A first step is to determine which *response* variables best address the study objectives as informed by the EDA0 qualitative assessment. A second step is to select the most appropriate *structural* variables to “slice” or bin response data by.

Once a manager has selected these response and structural variables from the dataset, EDA1 investigates the distributions of response variables as binned by one or more structural variables. This can be performed graphically, such as in Figure 2 below, or through parameter estimation and comparison, such as comparing measures of central tendency (e.g., means, modes) and variability (e.g., confidence intervals around the mean). This step may suggest useful strategies for later comparison *between* variables (EDA2) and further pattern exploration (EDA3) and hypothesis testing.



**Figure 2:** Example of EDA1 graphical investigation of a response variable (basal area) grouped by a structural variable (thinned vs. not-thinned treatment area) (from Philippi 2020)

An additional step in EDA1 is to identify patterns of missing data, outliers, and anomalies that differ *by* structural variable. If differences in problematic data appear to be concentrated in certain times, locations, and/or treatment types, data managers should flag this information for additional quality control and to inform subsequent data analyses.

Finally, in EDA1 data managers may investigate relationships between patterns in response variables and associated categorical *support* variables in the dataset to provide additional quality control. For example, a data manager may look at distributions of response variables by field observer or equipment type to identify non-sampling bias. This step can incorporate information embedded in non-categorical support variables – such as comments – to further understand patterns in response variables.

EDA1 answers the questions: how do key response data act differently under different circumstances (e.g., site, year, season, treatment, etc.)? Do these patterns suggest further investigation in EDA2 and EDA3? Do the data vary unexpectedly by non-structural variables such as observer or equipment type, suggesting a quality control issue that should be addressed before further analysis?

#### EDA1 is complete when:

* Data managers have selected a set of key response variables from the dataset that best address the monitoring or study objectives, and selected a set of associated structural variables to “slice” or bin response variable data by
* Data managers have characterized distributions and summary statistics of the key response variables, grouped by target structural variables; these investigations can be documented with graphical tools and/or summary statistics
* Data managers have investigated how patterns of missing data, outliers, and anomalous data differ by structural variables
* Finally, data managers have considered if the patterns illuminated in EDA1 suggest pairs of response variables to compare in EDA2

## 3.2 EDA1 techniques to illuminate within-variable patterns

As with EDA0, there are a set of common techniques most useful for this stage of EDA. We list these below to provide a common framework for NPS datasets. Data managers will generally be conducting these in a spreadsheet or statistical analysis software package.

* **For key response variables plot data binned by relevant structural variables to investigate patterns** and conduct quality control, with charts such as
* Histogram
* Frequency table
* Box or violin plot
* Bar plot
* Scatter plot
* **For key response variables, create a table of summary statistics grouped by relevant structural variables** (such as mean, median, mode, measures of precision, measures of skew and normality) to investigate patterns and conduct quality control
* **Evaluate if there are unexpected patterns** **or missing information or** when data are grouped by relevant structural variables; these may suggest interesting avenues for further data exploration and/or need for additional quality control measures
* Consider whether the graphical and summary tools reveal patterns that suggest pairs of response variables to compare in EDA2

## 3.3 Document pattern detection by variable

As with EDA0, data managers should document the results of EDA1 investigations to inform subsequent EDA processes, record the work for future data managers working on this dataset, and communicate with colleagues.

Documentation of EDA1 may include the following components:

* **Graphs and summary statistics illustrating EDA1 investigations of key response variables by relevant structural variables**; this would ideally include a description of the potential meanings of these findings, in context with the study’s objectives
* **A description of quality control issues** revealed by comparing unexpected patterns and support data and a discussion of how the issues were addressed and/or corrected
* **A selection of pairs of response variables to compare in EDA2**, unless EDA will end with EDA1

# Exploratory Data Analysis Step 2: Pattern Detection between Variables

## 4.1 EDA2 description

EDA2 investigates bilateral relationships between pairs of response variables. During this EDA step, data managers first choose pairs of response variables to compare. Managers should include investigations of pairs of variables suggested by the study’s overall purpose, as described in EDA0, but in the spirit of Tukey’s 1977 *Exploratory Data Analysis*, managers should also be guided by their own curiosity and subject matter expertise in choosing variables to compare.

Managers should prioritize evaluation of pairs of variables that have an expected relationship. For example, in a water quality dataset, data managers would expect to see a strong negative relationship between water temperature and dissolved oxygen. EDA2 will ideally confirm or refute these expected relationships, providing additional avenues for quality control on the dataset and/or potentially illuminating meaningful unexpected patterns that warrant further investigation.

Managers may also consider which structural variables, if any, to slice or bin EDA2’s bilateral comparisons by. For example, if exploring the relationship between number of plant species per plot and number of non-native plant species per plot in a vegetation monitoring program, managers may choose to compare these data from the entire dataset and then also examine these data binned by treatment area or year of data collection to illuminate patterns not obvious from whole-dataset investigations.

In addition, during EDA2 data managers may consider using relevant structural data from outside the dataset for binning bilateral comparisons by. At times, the study purpose will immediately suggest external data relevant for EDA2 analyses; this may be documented in the EDA0 qualitative assessment. If not, managers may wish to brainstorm types and sources of potential relevant external data for EDA2 comparisons. If data managers do bring in external structural data these may require considerable manipulation before they are suitable for use in EDA2 analyses, so data managers may wish to be highly selective in choosing these data.

#### EDA2 is complete when:

* Data managers have selected pairs of *response* variables from the dataset to compare that align with monitoring or study objectives -or- that are suggested from the data themselves; also data managers considered which (if any) *structural* variables to slice or bin response variable pairs by
* With graphical tools and/or summary statistics, data managers have characterized the relationships between the paired variables, ideally both as whole-dataset pairs and also binned by key structural variables
* External datasets have been reviewed to evaluate if these can provide relevant structural data for investigating pairwise relationships between variables, and if so, incorporated into the working dataset and compared with project data
* Data managers have considered whether expected relationships between variables are confirmed or refuted by the dataset, and followed up on unexpected relationships with further investigation
* Finally, data managers have considered if the patterns illuminated in EDA2 suggest hypothesis tests, trends analyses, multivariate analyses, or other topic-specific patterns to investigate in EDA3

## 4.2 EDA2 techniques to illuminate pairwise patterns between variables

A relevant graphical tool for EDA2 is the scatterplot matrix.

## 4.3 EDA2 document pairwise patterns between variables

 placeholder

# Exploratory Data Analysis Step 3: Exploration of Topic Specific Patterns

## 5.1 EDA3 description

Multivariate, preliminary trend detection

#### EDA3 is complete when:

## 5.2 EDA3 techniques to explore multivariate patterns and targeted questions

placeholder

## 5.3 Document multivariate patterns and targeted questions

placeholder

# Best EDA Practices for I&M Datasets

*Tom’s text:*

But note that outliers, whether univariate or multivariate, and blocks of missing data, are “patterns” as well, so EDA functions as additional QA/QC.

Philippi added feedback arrows from EDA back to Data cleaning and QA/QC, which made the workflow complex.  An alternative would be to reverse the order of Data Preservation (including certification) and EDA, with at least parts of EDA acting as both QA/QC and dataset description before preservation.  A third alternative is based on his breaking EDA into several steps (EDA0 – EDA3, see Section 1.3 below).  Initial EDA steps could occur prior to data certification and preservation, with later aspects of EDA occurring after preservation.

## In a nutshell:

### EDA0: Dataset Description and Preparation

* Qualitative assessment
* Categorize variables: response, structural, supporting
* Prepared dataset ready for further analysis

### EDA1: Pattern Detection within Variables

* Identification of key response and structural variables from the dataset
* Investigation of patterns within key response variables, with all data and sliced by key structural variables

### EDA2: Pattern Detection between Variables

* Identification of pairs of variables for relationship investigation
* Investigation of patterns between these pairs, with all data and sliced by key structural variables

### EDA3: Exploration of Topic-specific Patterns

* Multivariate
* Trends

# EDA0 Interview Your Dataset

## Protocol and Study Design

1. Is there a **protocol, standard operating procedures**, and/or reports tied to the dataset? Where are all versions of the documents stored?
2. What is the **resource** represented in the dataset?
3. What are the **primary goals** of the data collection?
4. What is the **geographical span** of the dataset? Are there major divisions within that span (e.g., park units, lakes, watersheds, habitats)?
5. What is the **temporal span** of the dataset? Is there fundamental seasonality in that span (e.g., sites visited once a month, every summer, etc.)? Is there any rotating panel design in the observation schedule?
6. (Briefly) how were **sampling locations** chosen (e.g., subjectively, objectively) and do they represent a larger area/habitat (i.e., what is the sampling universe)?
7. Are there defined **thresholds to trigger management** tied to the dataset? Are there **other ecological thresholds** that might inform interpretation and utility of the data?
8. Were there **changes to the protocol/SOPs** during the span of data collection that could influence results (e.g., collection procedures, equipment, time-of-year for site visits, etc.)?

## Data Structure

1. Where are the **data stored** and in what type of system? How are they accessed/exported for use in EDA? What are the files called?
2. Are the data stored in **longform, shortform, or a hybrid**? Are there multiple data tables to be combined or related during EDA? If so, please describe the data tables.
3. **What constitutes an “observation”** for this dataset? By this definition, how many observations are in the dataset?
4. Do you need **information not in the export tables** to help contextualize analyses or outputs (e.g., a lookup table of full site or species names, or otherwise coded information)?
5. Are there **associated datasets**that could provide covariates for later stages of EDA?
6. What are the **response variables** in the dataset? What are their expected data types (e.g., character, categorical, boolean, integer, etc.) and ranges?
7. What are the **structural** variables in the dataset (i.e., information about how/where/when data were collected)? Unlike supporting variables, such as comments, without structural variable data the records cannot be parsed. What are their expected data types and ranges?
8. What are the **supporting variables** recorded in the dataset (e.g., comments, notes, accuracy assessments)? What are their expected data types and ranges?

## Ecological Drivers and Patterns

1. Were there **environmental changes** during the span of data collection that could influence results (changes in land use, wildfire, drought, etc.)?
2. Are any of the response variables in the dataset likely to be **correlated**?
3. Are there any broad **patterns or trends you expect** to emerge from the dataset during EDA? What **hypotheses or trends** would you like to investigate with this dataset?

## QA/QC Processes and Concerns

1. **What QA/QC has been completed**? Are there specific SOPs for QA/QC? Were these procedures followed each year?
2. What **additional QA/QC** needs to be completed to prepare the data for analysis?
3. **Who collected the data**? Is this information embedded in the dataset? Were there any discontinuities in staffing or field work leadership that might influence analyses?
4. Are there any **patterns in the broad relationships between structural and/or response variables** that could help inform additional QA/QC and EDA (e.g., expected elevations by park, or expected year by site)?

## Final Dataset Preparation for Further EDA

1. Are there **any other issues with the dataset that are not apparent from the data themselves** (e.g., gaps in data collection, potential misidentification of species, potential misidentification of sites, potential bias in study design, problematic datasheet design, potentially problematic equipment, etc.)? Are these (or should these) be flagged in the dataset?
2. Are there data that should not go forward to further exploration and analysis? Are there **filters** that should be applied to subset the data?
3. Is there **anything else** to be aware of that isn’t covered above when analyzing this dataset?
4. What are the specific **steps that need to be taken for data cleaning** to prepare the dataset for further EDA (e.g., removing sites or parameters that are not a relevant part of the dataset, subsetting, segmenting, block by attribute name, transforming dates into an appropriate format, etc.)?

## EDA0 Dataset Investigations

(*Note:* answers to some of these questions may come from later EDA steps)

1. Graph the **patterns in the *structural*variables**: are observations evenly spaced within/between the structural variables and if not, does the pattern line up with the expectations?
2. Graph the **patterns in the numeric and categorical *response*variables**, are the distributions of these as expected?
3. What is the **pattern of observations by key support variables** if this information is embedded in the dataset (e.g., observer, equipment)?
4. Are there any **patterns of missing values**?
5. Are there any **patterns in outliers and anomalies**?
6. Are there any **obvious trends or relationships** that appear from this initial exploration of the dataset?

# Appendix B: Review of useful EDA tools in R

Possible packages for EDA1:

* **DataExplorer**’s *plot\_boxplot* function and **funModeling**’s *plotar* function return boxplots for each variable grouped by a specified categorical variable.
* The *explore* function in explore produces a Shiny pop up in which customizable plots (including boxplots) can display a specified variable and target.
* ExPanDaR’s *ExPanD* function also allows the user to interact with the data in Shiny. In Shiny users can compare histograms of a numeric variable between the different levels of a categorical variable (e.g. compare the distributions of an observation between different sites). A violin plot displays a selected variable grouped by a target.
* Through **summarytools**, a table displaying variable metrics/stats grouped by a factor could help reveal a pattern of missing values.

Automated EDA packages in R are tools designed to facilitate exploration of a dataset and make EDA an easier and quicker process. However, users can be overwhelmed trying to choose which package(s) will work well for their data. After reviewing the autoEDA packages discussed in “The Landscape of R Packages for Automated Exploratory Data Analysis” (Staniak and Biecek 2019), we compared their relative strengths for EDA of NPS datasets and illustrated to what extent each package performs typical EDA techniques.

In this document, we briefly describe eleven packages that we believe have the most potential for the EDA of NPS natural resources data. autoEDA packages have the most utility for EDA0 and EDA1 techniques, which we illustrate below. We provide screenshots of data visualizations display the most useful outputs from each package, using an example dataset from the San Francisco Bay Area Network WQ monitoring program. The examples focus on two response and one structural variable from that dataset: *AirTemp\_Glass*, *DOpercent\_YSI85,* and *StationID*.

Although autoEDA packages can make EDA process simpler, some functions in off-the-shelf packages may result in errors. Inaccuracies are especially prevalent when reporting the percentage of NAs in variables with few unique values. If exploring a dataset with a high prevalence of NAs, data managers may prioritize working with a package that accurately handles NAs, as described below.

## Summary of the packages reviewed

Four of the packages highlighted (**dataMaid**, **explore**, **DataExplorer**, and **SmartEDA**) are able to generate a data overview report from one line of code; these are useful for overall dataset summaries as well as variable by variable plots and descriptions. Variable visualizations differ substantially between the packages and choosing between them is largely based on user preference.

**dataMaid**, **explore**, **DataExplorer**, and **SmartEDA** also contain separate functions for manual exploration and targeting variables. For example, **DataExplorer**’s *plot\_boxplot* function returns a boxplot of each continuous variable broken down by a selected target. Incorporating an alternative to the traditional data overview report, **explore** allows the user to interact with the data and customize variable plots in **Shiny**. Among the packages we reviewed, **dataMaid** is unique in its ability to focus on flagging problematic variables as well as detecting and describing data anomalies.

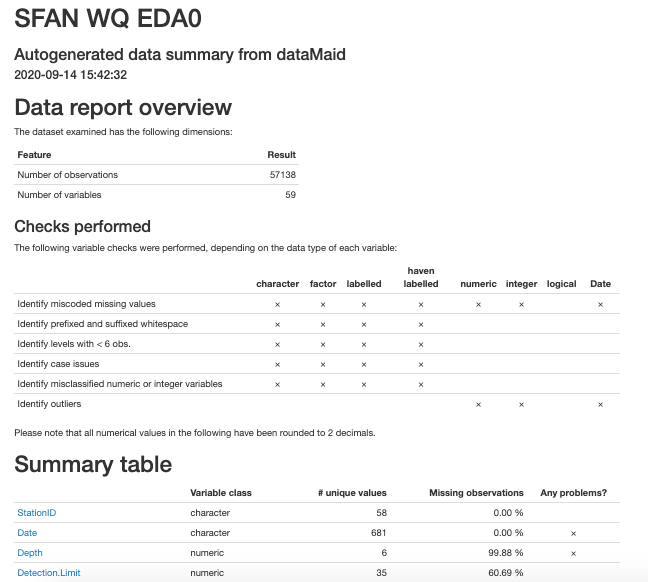
Although they do not include an option for data overview reports, **summarytools**, **Inspectdf**, **funModeling**, **visdat**, and **lubridate** are valuable for more manual exploration of a dataset. These packages provide analyses of missing values that can expose unexpected or problematic data quality characteristics of structural and response variables. **visdat**’s graphics help display the structure and important features of a data frame as a whole while **lubridate** takes a closer look at temporal variables, aiding the targeted exploration of dates.

**ggplot2** for plotting, **EnvStat** for data with non-detects

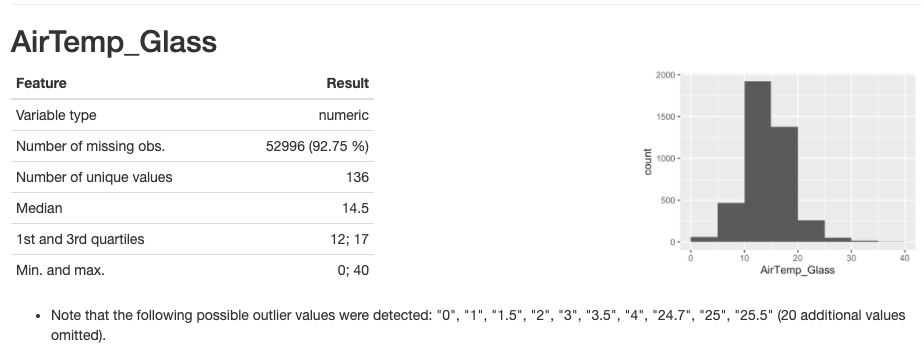
## dataMaid

**dataMaid**’s *makeDataReport* function creates a data overview report which includes a summary table and visualization (bar plot or histogram) of each variable. Unlike other autoEDA packages, **dataMaid** describes which variables are problematic and why they were flagged. A list of the variable checks performed during anomaly and outlier detection is on the first page of the document. To report only the problematic variables, specify onlyProblematic=TRUE. Users can manually explore a target variable through the separate *summarize*, *visualize*, and *check* functions.

For datasets with more complex categorical variables, text along the x axis of the bar plots can be difficult to read. In the summary table, columns with very few entries will be recorded as having 100% missing observations. *makeDataReport* will result in an invalid filename error if column headers include spaces or symbols.



**Figure B1:** Example of makeDataReport output from package **dataMaid**

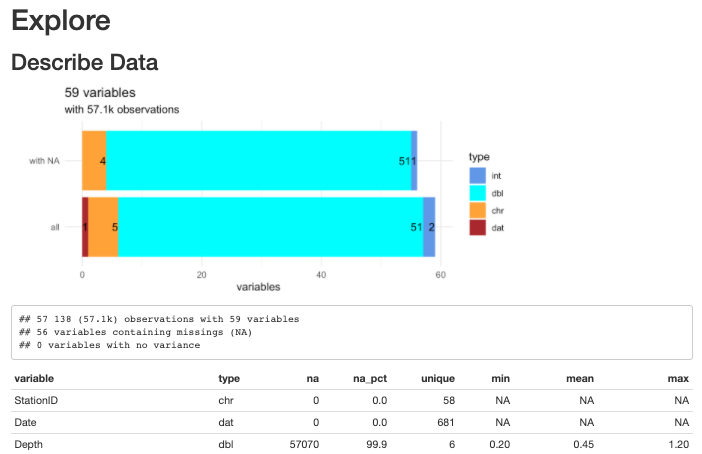


**Figure B2:** Example of makeDataReport output from package **dataMaid**

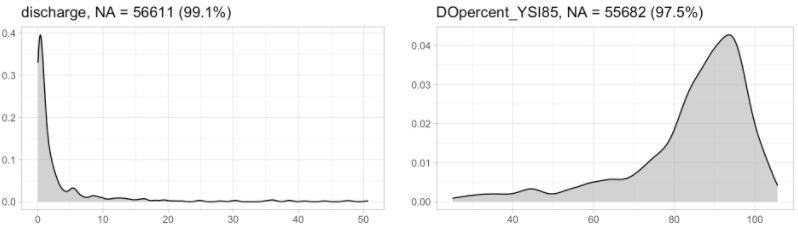
## explore

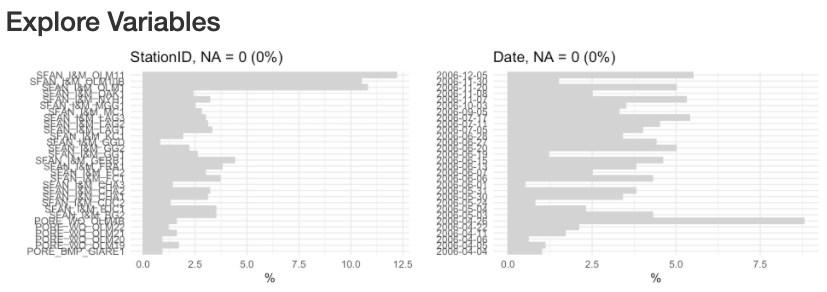
**explore** includes functions for both EDA0 as well as later EDA stages. Similar to **dataMaid**, **explore**’s *report* function generates a document which summarizes the dataset and includes variable visualizations. **explore**’s structural variable graphics may be preferable to those from **dataMaid** because the categorical characteristics are easier to distinguish along the y axis.

Furthermore, all variables in the report can be analyzed against a categorical structural variable (e.g., *StationID*). A column will be listed as having a NA\_pct of 100% if it does not have many unique values. Not only can **explore** create a data overview report from one line of code, but it also has the option to interact and look at the data in the package **Shiny** with the *explore* function. Within the **Shiny** pop up, customizable plots (e.g., boxplots) can display a specified variable and target for EDA1 investigations.

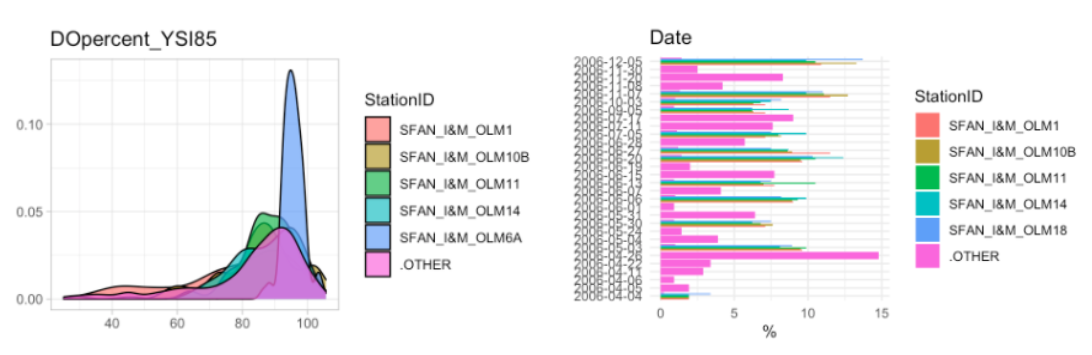


**Figure B3:** Example of report output from package **explore**

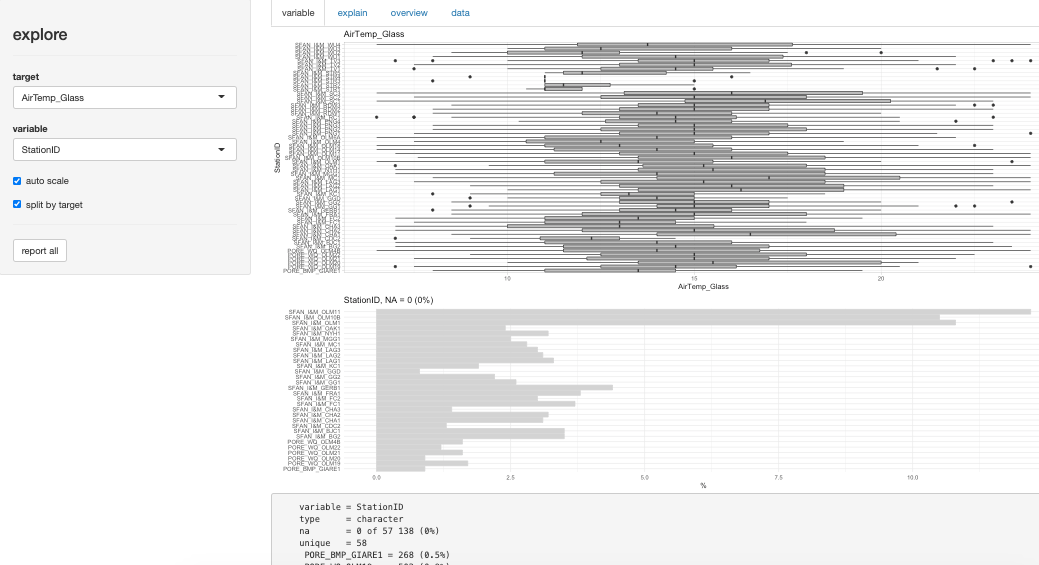




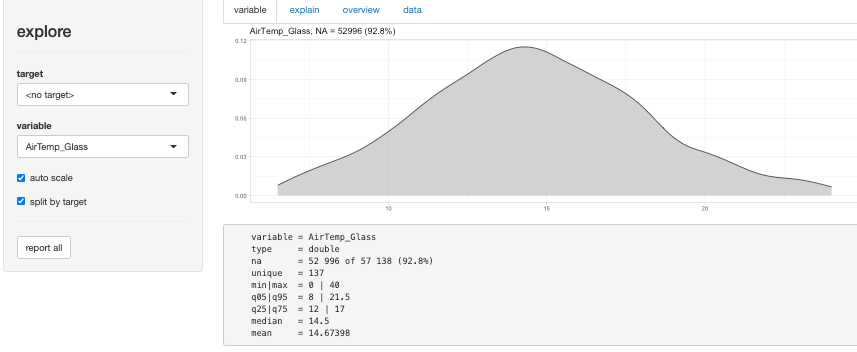
**Figure B4:** Examples of report output from package **explore**



**Figure B5:** Example of report output with target specified from package **explore**



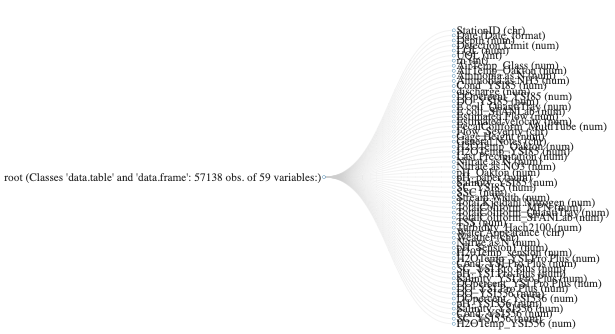
**Figure B6:** Example of interactive explore output from package **explore**



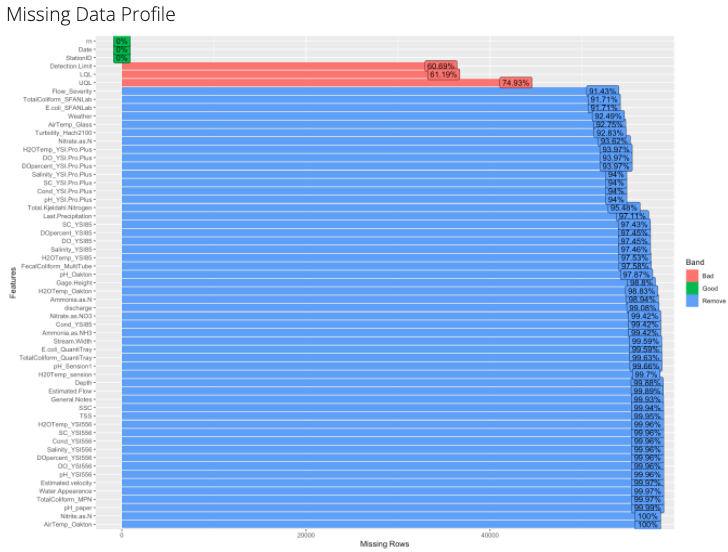
**Figure B7:** Example of interactive explore output from package **explore**

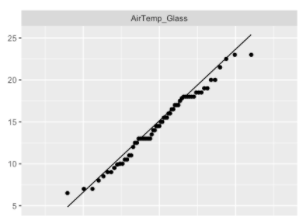
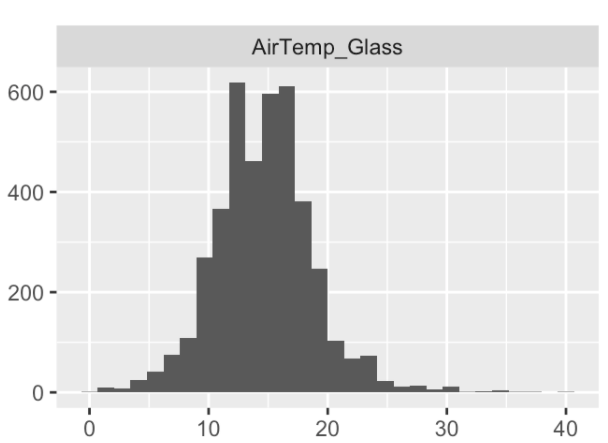
## DataExplorer

Relying more heavily on visualizations than the reports from **dataMaid** and **explore**, **DataExplorer**’s *create\_report* function produces a report with a brief summary of the dataset and detailed graphics for each variable (bar plot, histogram, QQplot) (see Figures B9 – B12). The data structure tree is difficult to interpret when analyzing a dataset with more than about 50 variables. Like **dataMaid**, the percentage of missing values in **DataExplorer**’s missing data profile is more accurate than **explore**’s NA\_pct. Nevertheless, variables with only a couple entries are listed as having 100% of their rows missing. The correlation analysis and PCA sections of the report, for later EDA stages, are not included if there are insufficient complete rows. Useful for EDA1, a separate *plot\_boxplot* function results in a boxplot of each continuous variable broken down by a selected target.

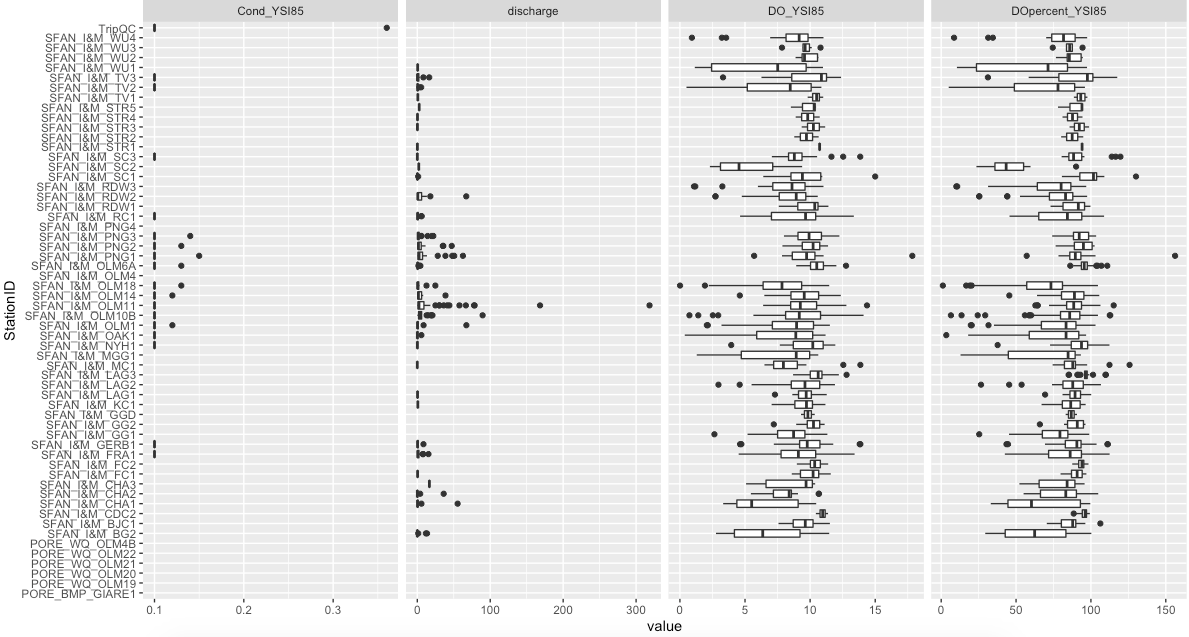


**Figure B8:** Example of create\_report output from package **DataExplorer**

**Figure B9:** Example of create\_report output from package **DataExplorer**



**Figure B10:** Examples of create\_report output from package **DataExplorer**

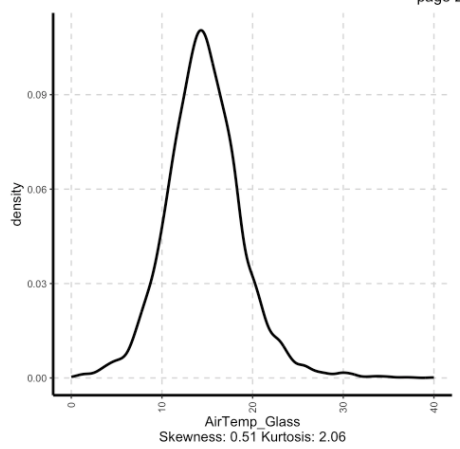
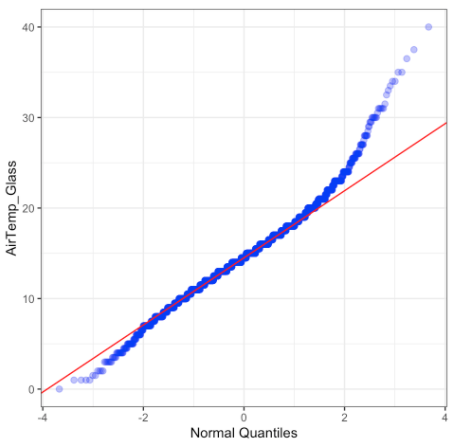


**Figure B11:** Example of plot\_boxplot output from package **DataExplorer**

## SmartEDA

The data overview report created by **SmartEDA**’s *ExpReport* function displays detailed summaries and graphics (bar plot, kernel density, QQplot) for each variable in a dataset. The ggplot visuals in the report are customizable through the ggthemes package. Along with graphical representations of structural variables, *ExpReport* produces a frequency table for all categorical variables. The accuracy of **SmartEDA**’s missing data description is superior to ones from **dataMaid**, **explore**, and **DataExplorer**.

Unlike other packages that skip correlation analysis because of "insufficient complete rows", **SmartEDA**’s report includes bivariate scatter plots, for each possible variable pairing, which can be used during EDA2. Consequently, **SmartEDA** takes the longest processing time to produce a report when compared to the other one-line report packages. Like **dataMaid**, when using **SmartEDA** a report will not be generated due to an unexpected symbol error if the column headers include symbols or spaces.



**Figure B12:** Example of ExpReport output from package **SmartEDA**

## summarytools

**summarytools**’ *dfsummary* function creates a dataset summary table with variable descriptions and simple plots. The option to display statistics for variables grouped by a specified parameter could be useful when trying to detect if there is a pattern of missing values in the dataframe. The *freq* function prints out a frequency table which displays missing value details for each column.

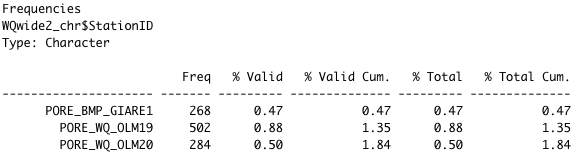


Figure B13: Example of freq output from package **summarytools**

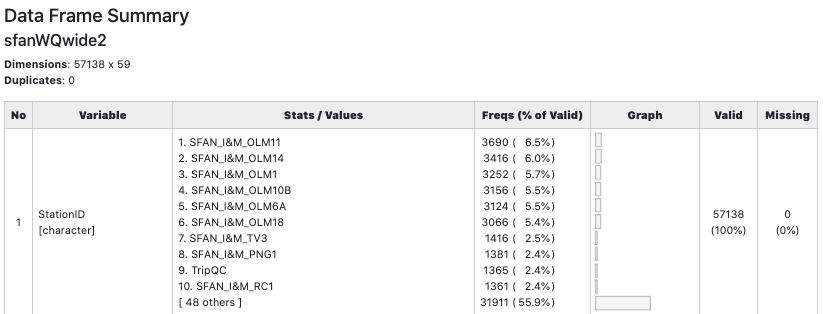
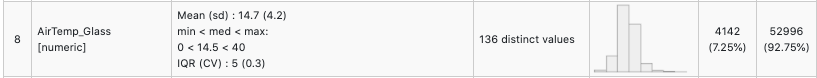
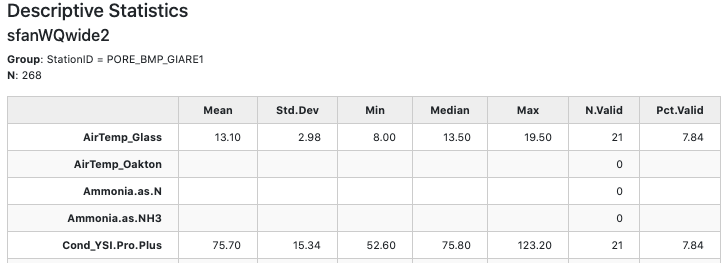


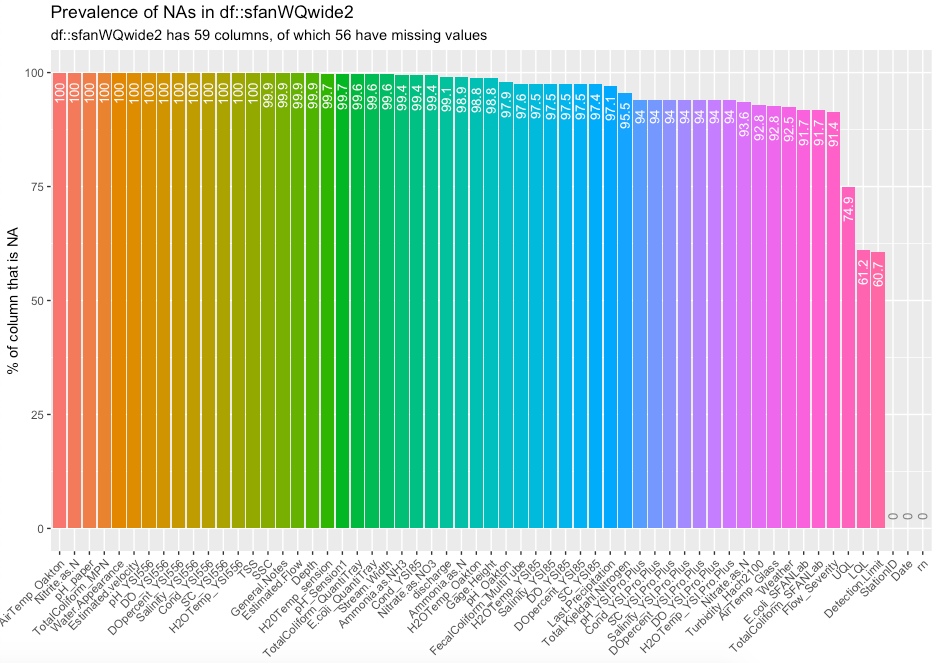
Figure B14: Example of dfSummary output from package **summarytools**



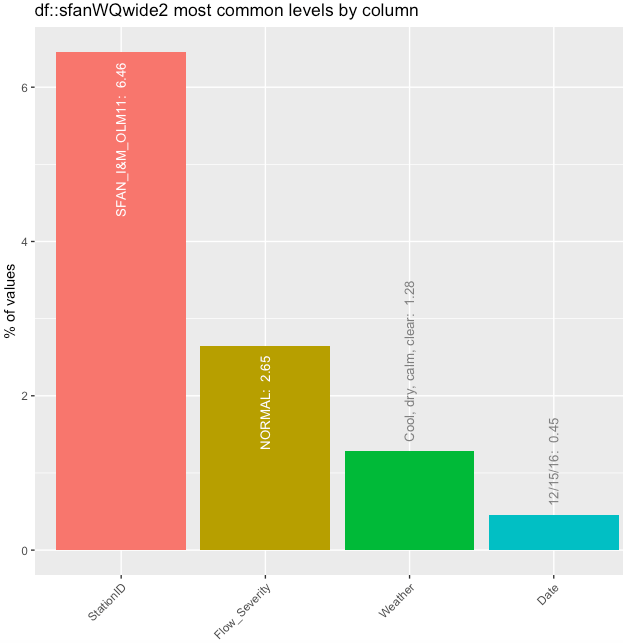
**Figure B15:** Example of stby output from package **summarytools**

## Inspectdf

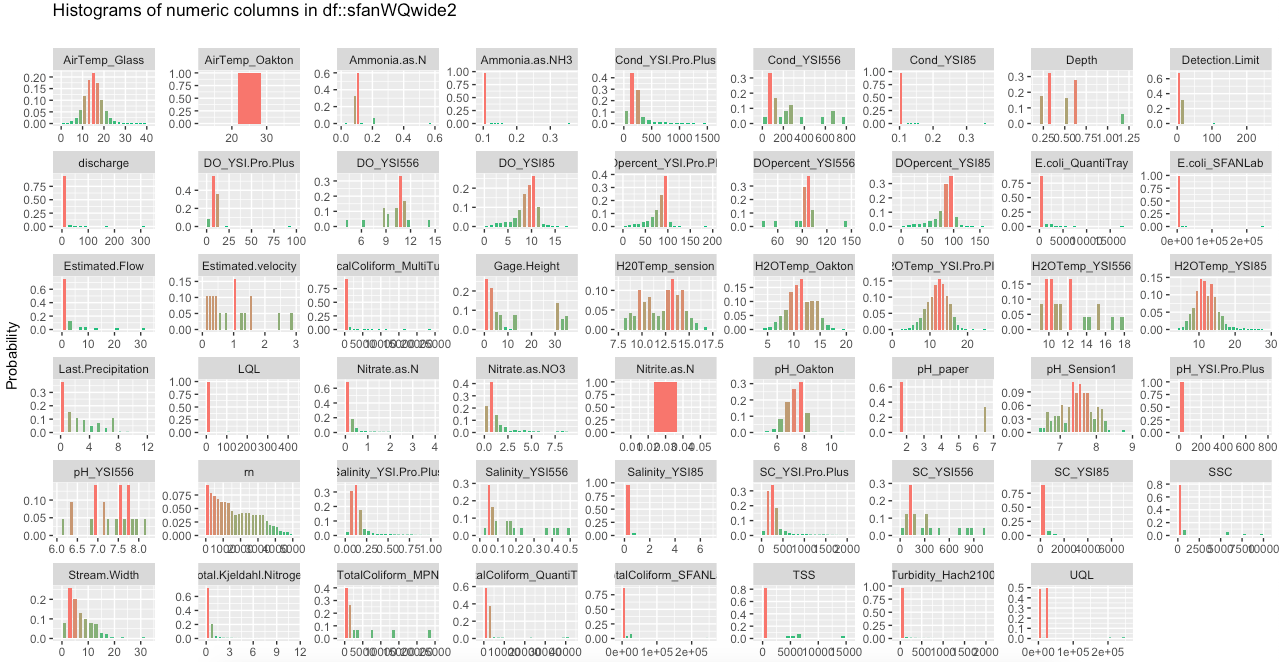
With the purpose of summarizing a dataset, **Inspectdf** comprises multiple functions that offer clear descriptions and visualizations (with the *show\_plot* function) of data types, missing values, and statistics. Notably, *inspect\_NA* describes columns with few values as consisting of 100% NAs. Even though **Inspectdf** does not provide a separate bar plot for each categorical variable, the *inspect\_cat* and *inspect\_imb* plots display the frequency of categorical levels and the most common level of each categorical variable in the dataframe. One significant feature of **Inspectdf** is its ability to compare two datasets in each function. During EDA2, users can explore bivariate relationships with the *inspect\_cor* function. Colors in **Inspectdf**’s plots do not reflect patterns in the dataset, they are aesthetic only.



**Figure B16:** Example of inspect\_NA output from package **Inspectdf**



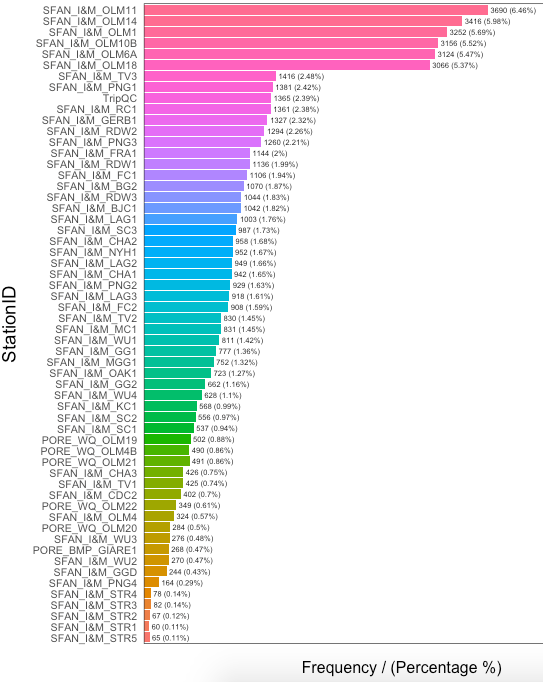
**Figure B16:** Example of inspect\_imb output from package **Inspectdf**

  
**Figure B17:** Example of inspect\_num output from package **Inspectdf**

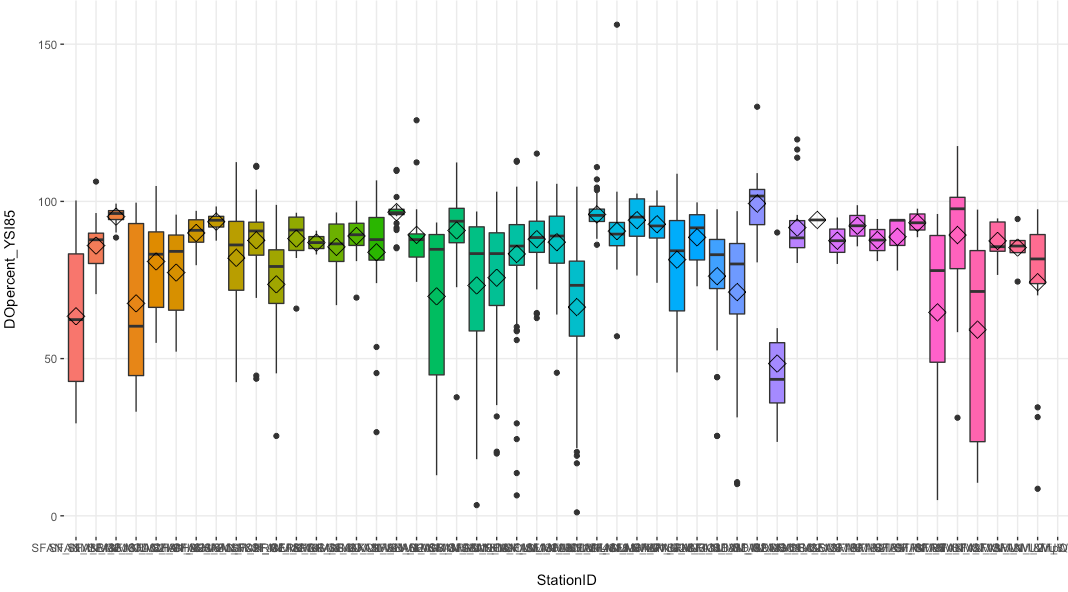
## funModeling

**funModeling** offers a wide range of EDA techniques for understanding a dataset and its variables. *df\_status* summarizes the data frame by including information on missing values, variable data types, and unique values. However, it seems to have an error in the way it reports NAs for columns with few unique values. The *freq* function outputs a frequency table and bar plot for each categorical variable while the *plotar* function generates detailed boxplots of each predictor variable broken down by a target parameter. Useful for EDA1 and similar to **DataExplorer**’s *plot\_boxplot* function, **funModeling**’s *plotar* function returns boxplots for each variable grouped by a specified categorical variable.

*profiling\_num* and *plot\_num* functions result in a metrics table and a histogram for each numeric variable. One potential issue with the *plot\_num* function is that the same scale is used for each histogram, making it difficult to visualize measurements of different size ranges. To plot variables that require the same scale together, the data could be subset prior to running the function. Also resulting in a single document with a histogram for every variable, **Inspectdf**’s *inspect\_num* function might be preferable to *plot\_num* because all variables are plotted against their own scale. For EDA2 bivariate analyses, **funModeling** will not produce a *correlation\_table* if numeric columns are incomplete. As with **Inspectdf**, colors in **funModeling**’s plots do not reflect patterns in the dataset, they are aesthetic only.



**Figure B18:** Example of freq output from package **funModeling**



**Figure B19:** Example of plotar output from package **funModeling**

## visdat

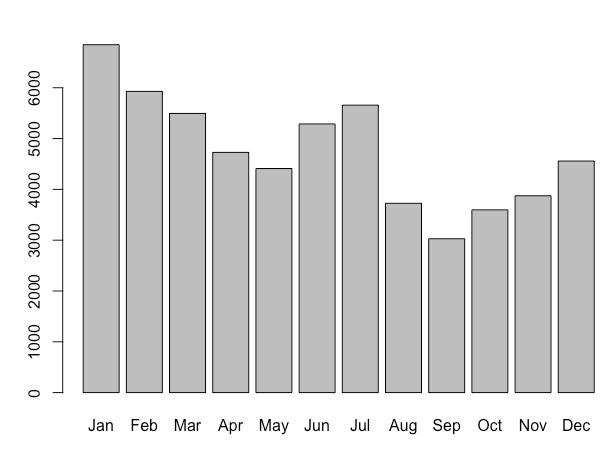
Depending solely on visual EDA techniques, **visdat** is a unique and helpful tool for describing a dataset. Its functional capabilities include visualizing data types, value types, and missing data. Furthermore, the general difference between multiple datasets can be shown through the *vis\_compare* function. *vis\_expect* detects whether a value is in a dataset or where a specified condition is met. For later EDA stages, *vis\_cor* creates a correlation heatmap. For large datasets that exceed the recommended size for visualizations, set warn\_large\_data = FALSE.



**Figure B20:** Example of vis\_dat output from package **visdat**

## lubridate

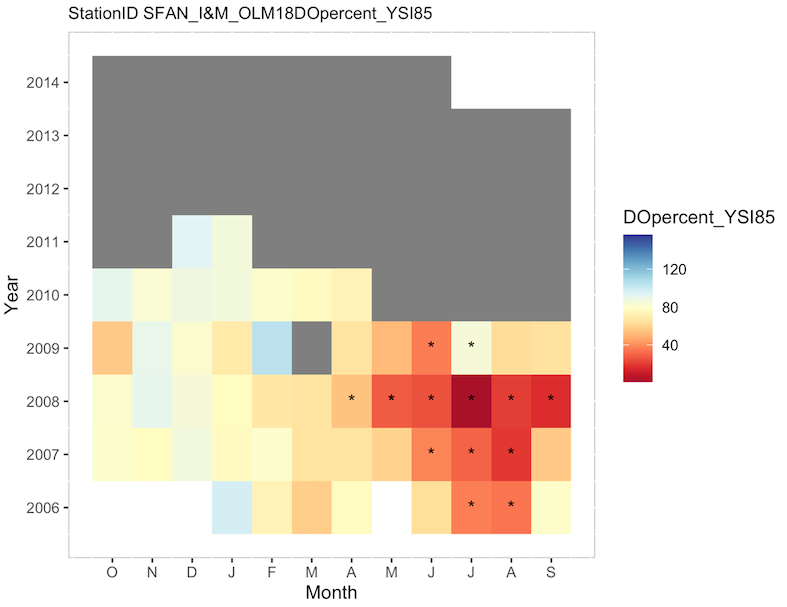
**lubridate** is a package that works with temporal data and can be utilized to perform EDA0 quality control reviews on the dates of a dataset. In addition to being able to ensure that the data type of temporal structural variables are classified as “date”, **lubridate** produces tables and plots to show the frequency of days, months, and years in a dataset. Possibly its most notable function, as.*duration* can be used to compare the intervals between dates in order to determine if they are evenly spread.



**Figure B21:** Example of month output from package **lubridate**

## ggplot2

A powerful tool for data visualizations, **ggplot2** encompasses many techniques that assist with completing EDA. Even though EDA is not automated in **ggplot2** and users must become oriented to the package’s basic grammar, the package can produce a wide range of charts to facilitate EDA techniques. One function of **ggplot2** for EDA is its ability to yield raster charts, or heat maps, to display data that are difficult to reduce into a tabular format. Different from the correlation heat maps that packages like **visdat** produce (*vis\_cor*), **ggplot2**’s heat maps display raster data in a gridded format. Rasters can be used for pattern detection (e.g., abnormal values and data gaps) and quality control for complex datasets.



**Figure B22:** Example of raster output from package **ggplot2**

## EnvStats

placeholder

Table 1: Comparison of autoEDA packages included in “The Landscape of R Packages for Automated Exploratory Data Analysis” (Staniak and Biecek 2019) across common EDA techniques. **Lubridate** and **ggplot2** were not mentioned as autoEDA packages in the paper.