Exploratory Data Analysis

Educators and Employers Survey Data Exploration

Brock Akerman Hanan Ali Taylor Cesarski

2025-06-15

Table of contents

1 Abstract

2 Introduction

Our researcher, Dr. Ross-Estrada has distributed a survey to two distinct participant groups:

- Educators who teach in the field of dental veterinary medicine (DVM), and
- Employers who have recently hired graduates from DVM programs.

These two respondent groups provide us with two separate datasets—educators and employers—each with its own structure and variables. While there is some overlap between them, differences in content and context mean that we will treat these datasets separately in most of our analysis. Dr. Ross-Estrada wishes to extract insights about the two groups and their perspectives concerning training and capabilities of new graduates of dental veterinarian medicine programs. The survey was conducted during the summer of 2024 using Qualtrics–an experience management software service. Selection of the participants was not conducted randomly; instead our researchers network was used.

3 Initial Data Inspection

The best approach is to first get a large look from the top down on these two datasets to see what we are working with. Let us first take a look at the Education dataset.

```
Educator_Data <- read_csv("PCVE_Dentistry_Survey.csv", show_col_types = FALSE)
```

```
New names:
* `Q46` -> `Q46...18`
* `Q45` -> `Q45...37`
* `Q46` -> `Q46...38`
* `Q45` -> `Q45...159`
```

```
dim(Educator_Data)
```

```
[1]  45 171
```

```
head(Educator_Data)
```

```
# A tibble: 6 x 171
  StartDate    EndDate Status IPAddress Progress Duration (in seconds~1 Finished
  <chr>        <chr>   <chr>  <chr>     <chr>    <chr>                  <chr>
1 "Start Date" "End D~ "Resp~ "IP Addr~ "Progre~ "Duration (in seconds~ "Finish~
2 "{\"ImportI~ "{\"Im~ "{\"I~ "{\"Impo~ "{\"Imp~ "{\"ImportId\":\"dura~ "{\"Imp~
3 "8/8/2024 1~ "8/8/2~ "0"    "128.84.~ "100"    "600"                  "1"
4 "8/8/2024 1~ "8/8/2~ "0"    "192.0.2~ "100"    "29"                   "1"
```

```
5 "8/8/2024 1~ "8/8/2~ "0"    "38.175.~ "100"    "801"                  "1"
6 "8/8/2024 1~ "8/8/2~ "0"    "152.1.8~ "100"    "3375"                 "1"
# i abbreviated name: 1: `Duration (in seconds)`
# i 164 more variables: RecordedDate <chr>, ResponseId <chr>,
#   RecipientLastName <chr>, RecipientFirstName <chr>, RecipientEmail <chr>,
#   ExternalReference <chr>, LocationLatitude <chr>, LocationLongitude <chr>,
#   DistributionChannel <chr>, UserLanguage <chr>, Q46...18 <chr>, Q44 <chr>,
#   Q3 <chr>, Q3_13_TEXT <chr>, Q4_1 <chr>, Q4_2 <chr>, Q4_3 <chr>, Q4_4 <chr>,
#   Q4_5 <chr>, Q4_6 <chr>, Q4_7 <chr>, Q4_8 <chr>, Q4_9 <chr>, ...
```

The results show a table with 45 rows and 171 columns. Retrieving a sample of this table reveals that the top two rows are subheaders likely included as part of the Qualtric survey output and not data used to measure the sentiment about dental veterinarian students knowledge.

We will do the same for the Employers dataset. -Load the data and inspect dimensions (n rows × p columns) -Check variable names and types (numeric, factor, character, date, etc.) -Print first few rows (head()) and summary statistics -Identify response (outcome) and predictor (explanatory) variables

```
Employer_Data <- read_csv("Employer_Dentistry_Survey.csv", show_col_types = FALSE)
dim(Employer_Data)
```

```
[1]  31 176
```

4 Missing Data

Count missing values per column and per row Visualize missingness patterns (e.g., heatmap or naniar/VIM plots) Check for systematic missingness (by group, time, etc.) Decide: drop, impute, flag, or model missingness?

5 Univariate Analysis

- For each variable: Compute summary stats: mean, median, SD, range, IQR Plot distribution: histograms (numeric), bar plots (categorical) Identify outliers or skewness Check for unusual values or coding errors

6 Bivariate Analysis

- Numeric vs Numeric Scatterplots with smoothing (e.g., LOESS) Correlation coefficients (Pearson/Spearman)
- Categorical vs Numeric Boxplots or violin plots Group means + CI/error bars ANOVA or t-tests (exploratory, not confirmatory)
- Categorical vs Categorical Cross-tabulations Chi-square or Fisher's tests (exploratory)

7 Multivariate Structure

Correlation heatmap (numeric variables) Principal Component Analysis (PCA) or t-SNE (if high-dimensional) Pair plots / scatterplot matrix Check multicollinearity (VIFs, condition index)

7.1 Time/Spatial Data (if needed)

Time series plots Trends, seasonality, anomalies Autocorrelation, lag plots Maps or geospatial distribution

8 Data Integrity Checks

Check for duplicates (rows, IDs) Validate ranges against expected values Consistency across related variables (e.g., start_date < end_date) Confirm units and scales are consistent

9 Documentation & client Communication

Create a clean report with summary tables and visualizations Highlight any data issues that could affect modeling Document assumptions, decisions (e.g., handling of missing or outliers) Make notes on variables of interest for modeling