# An Introduction to the Data Set

The data provided was generated by a survey emailed to members of the Women’s Flat Track Derby Association. Roller Derby is a fast-paced contact sport that is primarily played by women. The purpose of the survey was to quantitatively find relations between the sport as serious leisure and sense of identity among its female players.

# Fundamental Cleaning

Upon first inspection of the raw data, it was noted that there were a few rows at the top that needed to be dropped to create a usable dataframe:

Graphical user interface, application, table, Excel

Description automatically generated

Figure 1, original raw data.

Noted also is that there were several unused columns labeled “Custom Variable” followed by a number: 1, 2, etc. Moreover, the columns regarding email, “Respondent Email” and “Email List” were completely empty of data. Because the first 17 columns are not part of the survey, these two columns regarding email could be dropped if empty. Also, the “External Reference” variable only held data in the first 3 entries, which indicated these 3 rows were for testing and can be removed, along with the “External Reference” column itself.

There were three survey questions in particular that offered the choice of “Other”:

Graphical user interface, application

Description automatically generated

Figure 2, example of survey question with "Other".

Normally the survey data maps the selection to a number (“USA” giving 1, “Australia” giving 2, etc.). When “Other” is selected, a text entry must be made by the surveyee. In the raw data, the questions or requests, which are variables in the data set, still maps a selection of Other to a number (in the above case 8). However, questions/requests with “Other” as an option have an adjacent column to the right for handling a text entry:

Application

Description automatically generated with medium confidence

Figure 3, the "Please select your country: " variable and the adjacent column.

Note that “Other” in the above picture needs to be moved upward to become part of the header column. Three variables in the data set have an “Other” option and thus an “Other” variable following them to handle text entry.

Regarding the variables, many of our column names are quite long, as they are either questions or requests to the surveyee. While this makes for inconvenient column names, for now, these are left as is. One column name in particular was too long; it disrupted the layout of the dataframe:

A picture containing table

Description automatically generated

Figure 4, the first survey column which includes the introduction.

This was shortened to that last sentenced of the introduction: “Select a response that represents your participation in this survey”. This is still long, but for now we will leave it to stay consistent with the other column names. We now have a data set of 88 variables and 370 entries.

# Visualizing Missing Data

Now that we have gotten the dataframe, it would be nice to see what values we are missing. Using the missingno library, a bar chart can be generated where each bar represents the number of non-missing values in that column. The left side y-axis shows the column names, while the top side x-axis shows the values 0.0 to 1.0 representing percentage.

Chart, bar chart

Description automatically generated

Figure , bar chart displaying data present for each column.

The first variables of this chart, starting from the top down, are the initial data that is automatically generated for each survey. They include the columns from “Response ID” to “Country Code”. Note that these variables contain no missing values.

As questions and requests to the surveyee come in, a growth of missing values follows. See that the number of non-missing values shrink as we move further into survey.

This implies that the further into the survey, the more likely we have missing data. Likely this means that the participant quit the survey, so the following variables after a missing value should be missing as well. If we visualize the missing data in a tabular manner we can see where surveys may have stopped:

A picture containing text

Description automatically generated

Figure , a matrix displaying the tabular data set with whitespace representing missing values.

Recall the columns to the right represent pre-survey data and thus are full. We can see some of the entry rows drop off throughout the middle, indicating the participant exited the survey without finishing. Using the “Response Status” column, we can count the number of “Completed” survey entries; 198 completed out of 370 entries.

# Upcoming Challenges

Recall that question/request variables with an “Other” option as a selection have a following variable for handling text entry, labeled Other. Because three variables have “Other” as a selection option, there will be three variables sharing the name Other. This is not desirable, as it could interfere with mapping and manipulating the data properly.

Another issue that may need to be addressed is the length of the column names themselves. Some contain over 40 characters, and depending on the amount of data manipulation needed, these names may need to be changed.

Finally, many of the numerical values in the data are keys that map to text data. Dictionaries need to be created for each of these questions or request variable to successfully map these values to text entries. Each of these challenges will be attempted in the coming week.