# DATA CLEANING AND DOCUMENTATION

The New Coder dataset [freeCodeCamp. (2016). 2016-FCC-New-Coders-Survey-Data [Data set CSV file]. Retrieved from https://github.com/FreeCodeCamp/2016-new-coder-survey September 13, 2016] is a survey crafted for people who had recently started to learn to code through various platforms. The respondents were asked various questions such as their country of citizenship, the subjects they studied in university as well as their current and expected employment status. One interesting aspect of this dataset is that it caters to people from all backgrounds irrespective of their career field and grants researchers a glimpse of how adults learn to code. There can exist a wide range of potential data users and decision makers for this data set. It would include researchers, professors, MOOCs and even MIM program director or committee.

The given dataset is made available under the Open Database License (ODbL) v1.0: http://opendatacommons.org/licenses/odbl/1.0/. Any rights in individual contents of the database are licensed under the Database Contents License: http://opendatacommons.org/licenses/dbcl/1.0/ The **Open Database License** is a copyleft (If you publicly use any adapted version of this dataset or works produced from an adapted database, you must also offer that adapted database under the ODbL) license agreement intended to allow users to freely share, modify, and use a dataset while maintaining this same freedom for others.

## Metadata:

The data source consisted of a Survey Data Dictionary along with the data set that had a list of all columns and their corresponding questions with description. This information helped us understand the type of questions that were asked, the scale of measurement of the answers along with its possible range. The questions were divided into three types:

* Free text where user inputs values
* Multiple choice in choosing one answer
* Multiple choice allowing for multiple selections

Depending on the above, we categorized the fields into nominal, ordinal, interval or ratio scales which would help us in analyzing data later.

According to the five research questions, we performed data cleaning. The columns on the left of the dataset consist of people's basic information such as age, citizenship, school degree, school major, income, employment field, expected earning, job interests. By analyzing this, we can have a better understanding of individuals that took part in our survey. We also retained the columns related to online learning, such as high-speed internet availability, learning hours. We can also learn more about the efficiency of online learning platforms through these questions. In addition to this, all the questions about the online resources like Coursera, Lynda, DevTips are also there in our dataset which can help us check the popularity of different course resources.

## Rationale for Remediation:

The first step that we took to clean data was to remove all the columns that were not relevant to our research questions for eg. Based on your experience, would you recommend this boot camp to your friends?, which language do you speak at home with your family, which coding-related podcasts have you found helpful?. The rationale behind this step was to reduce the number of columns to a minimum and now we have 41 columns.

After getting the columns that were required, we removed a few rows that had outliers or inconsistencies for eg, the column Age had values above 100 which is clearly incorrect as the question specified the age to be entered in years.

## Challenges Faced:

Some of the issues that we faced while cleaning data:

* Outliers: The Age column which had to be answered in years had values above 100 which was clearly not practically possible.
* Information Overload: A lot of questions were either irrelevant or redundant and were thus removed. For e.g.:

1. Question: Based on your experience, would you recommend this boot camp to your friends?

Irrelevant question. This is more of a feedback rather than a survey question and it cannot be used for any of the research questions in our project.

1. Question: What's the highest degree of level of school you have completed?

Redundant question. This information can be better accessed by asking the respondents their highest degree attained instead of giving them options such as no high school, some high school which is insufficient to extract useful information that is relevant to research questions under study.

* Missing Values: All the questions were not mandatory so many people did not answer which created a lot of missing values.
* Scale of measurement used for a few questions: In the question 'Which field do you work in?’ one of the answer options is ‘arts, entertainment, sports, or media' which has many different fields grouped into one that cannot be used easily or any research question. Similarly, the answer option 'law enforcement and fire and rescue' seemed to be a very vague option to us.

## Data Cleaning Process:

The steps followed for the data cleaning process were:

|  |  |
| --- | --- |
| Step 1 | Identify field columns useful for the questions under study as well as redundant and irrelevant questions |
| Step 2 | Remove the columns that are not required |
| Step 3 | Apply filter on the remaining columns to get rows with impractical or false values |
| Step 4 | Remove rows that do not have relevant data. |
| Step 5 | The remaining 1372 rows is the final data set on which we will work |

Word count: 884 Words

## **References:**

### Dataset:

freeCodeCamp. (2016). 2016-FCC-New-Coders-Survey-Data [Data set CSV file]. Retrieved from https://github.com/FreeCodeCamp/2016-new-coder-survey September 13, 2016

### Terms of Use:

Given data set is released under Open Database License (ODbL) v1.0 at http://opendatacommons.org/licenses/odbl/1.0/