# THE DATA CLEANING AND TIDYING

Su Aladağ

**SUMMARY:**

If I needed to summarize what I did for the project briefly, I started by examining what the given data set is, what purpose it can be used for, and whether it has missing or incorrect values. My goal was to obtain a usable and inferable data set due to the changes I made by following the appropriate steps.

In my data set, which I organized by following the steps I will explain later, I visualized the relationships between the variables or the values I wanted to examine about the variables with the help of tables.

As a result of my work, I cleaned my data set and followed the data visualization steps for the research questions I created about the variables.

**DATA CLEANING AND EDITING STEPS:**

First, I started by calling any libraries that might be needed.

**1.** I imported my data file and read the Excel file using the "reading" command. When I examine variables and data types, I have "gender," "race," "class," "studying hour," "school ranking," "math score," and "reading score. I saw that I have a dataset with "writing score" variables.

**2.** I examined the head and tail of the dataset. By using the "df.info()" command, I saw the column names, the "null" values, and the data types of the columns. Using the ".title" command, I edited the column headings with the capitalized first letter. Meanwhile, I checked whether there were spaces between two words, "\_," or punctuation while the titles were being written. There was no problem with this step; there was a "\_" between the two-word titles and no punctuation errors.

**3.** I reprinted the data table whose headers I had edited using .head() and saw if it had changed. After the first step, where I corrected the headers, as a second step, I found whether there were duplicate or stored values, and if so, how many in total with the ".duplicated().sum()" command. Since I have one duplicate line, I removed it with the “.drop\_duplicates()” command. When I reprinted the table, I saw that the table that previously contained 200 rows had decreased to 199 rows. Thus, I have removed the duplicate line.

**4.** To see the line names with typos "for i in su.select\_dtypes(include = 'object').columns: I used the print(su[i].value\_counts())" command. I corrected the misspellings in the categorical variables by typing the correct one with the .loc command and changed them to capital letters with the .title command.

**5.** In the next step, I got rid of the white space between strings with the ".str.strip()" command.

**6**. In step 5, I worked on examining my numerical data and finding and replacing any unusual behavior. I removed the outlier in the "max" column for all my numeric values. Since the mean of my data is not greater than the median, and the standard deviations vary at normal rates, I decided that what was written in the "max" column, which did not have any outliers in my data, was wrong and calculated the "mean" values to fill this column. Then I changed the "max" column by placing the "mean" values that I calculated into the "fillna()" method.

**7.** The "mean," "median," and "standard deviation" values of my data match each other proportionally. I did not make any other changes because it has reasonable rates, such as 12%, 25% 32.3%.

**8.** I examined my lost data in step 6 using the “. isnull()” command. Then I saw the total "null" values with the “. isnul (). sum()” command. Additionally, I presented the percentage of my missing values and created a simple graph showing the location of the missing values using the "seaborn" library.

**9.** As the next step, I needed to fill in my missing values in a logical way. I calculated the values I needed to write using "median" for my numeric values and "mode" for my categorical values. I used the median because it was more resistant to variables than the mean. Again, using the ".fillna()" command, I wrote the values I obtained in parentheses and added them to the table.

**10.** I made my last change on the "anxiety level." They had negative values. I got the absolute value with the ".abs()" command. I presented my cleaned data again with the .head() command and saved the Excel file as "su\_clean.xlsx."

**GRAPHS:**

I wanted to create graphic codes using "Google Colab" to examine variables and their relationships with each other. First, I called up the libraries I might need while making the chart. These are "NumPy," "pandas," and "matplotlib.pyplot". I also used the "seaborn" library in the following steps. I read the excel file that I cleaned and saved and decided to proceed by giving it the name "P." I examined my data again by calling it with the ".head()" and ".info()" commands. My first chart was a graph showing the relationship between "School Ranking" and "Math Score." I got a diagram where there is a big difference between the two groups that do not provide a value for "Top Level" while the "Middle Level" values are high.

2. In my table, I wanted to examine the distribution of men and women in my data set, and I used a bar chart. I have seen the gender distribution presenting almost half the relationship.

3. In my chart, I used a bar chart to clearly see the distribution of "School Ranking" values and their differences. While the "Middle Level" value was more than 120, the "Top Level" and "Low Level" values that were close to each other were between 40-45. Here, I can say that the candidates included in the evaluation were asked to be at an average level, and the "top" and "low" values almost acted as outliers. Still, their similar numbers did not balance their effects in opposite directions.

4. In my chart, I used a bar chart to compare the distribution of 3 groups of candidates in terms of numbers and compare them with each other. "Group1" features the most significant number of groups. I have observed a situation similar to the above. Group1 has values Group2 and Group3, which have almost 1 in 7. However, I cannot comment on the opposite effect I mentioned above. In this case, excessive redundancy in Group1 will prevent reaching a clear result in the overall evaluation.

5. Using "seaborn" in my chart, the relationship between genders and monthly income, I wanted to make a boxplot where we can quickly examine the mean value, median value, and skewness, but it did not create a graph. Two found values represent outliers. There are outliers of the same value for the male and female gender.

6. I wanted to examine the change in writing scores by four classes by creating a "subset" value on my chart. Again, I thought it would be easier to interpret the data I have over a box plot, but it did not present me with the table I wanted. I got a graph with outliers. There are positive outliers for Class1, negative for Class3, and positive and negative outliers for Class2. Finally, I wanted to evaluate the distribution of the candidates' math scores with my 7th chart. There was no uniform distribution of all values. Too large an intensity occurred for a value of 2. Except for the outliers between 14-16, he saw that the mathematics scores of the participants were relatively low, and this low value was intense. Looking at our previous studies, we can say that a mathematics score of 2 in a data set with a high "Middle Level" education level indicates a low level of mathematics learning gained at this education level.

**CONCLUSION:**

Although I have a data set that is easy to clean, I had difficulty presenting the relationship between them with the help of graphs. The presence of some values at a much higher rate prevented the contribution of other values to the chart. So much so that when I wanted to make a "scatter" graph, I couldn't explain the net relationship of the variables we have, except for a few mountains and small and large values. However, there are rumors that success in reading affects mathematics. However, I have seen from this data set that this is a relationship that can create very diverse answers depending on what kind of group you are examining, how many people you are exploring, gender distribution, differences in education levels, and if you have grouped these people, how you group them.