Chris Arancio

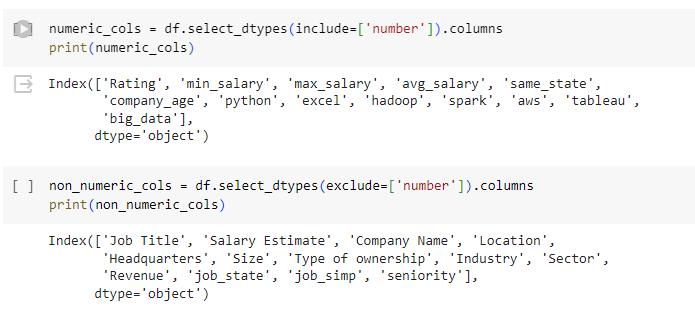
**Data Cleaning Report for DataScienceJobs.csv Dataset**

This report will document the steps taken to clean the data and will also explain why certain steps were taken. There are seven total steps described.

**Step #1:** Divide Columns into Numerical and Categorical DataFrames

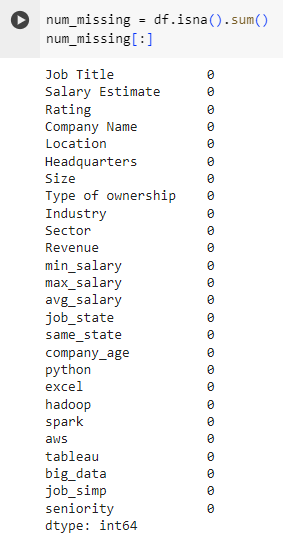
Certain data cleaning techniques depends on the type of data in a column, so the columns were divided into numerical and categorical variables.

Code used:



**Step #2:** Check for missing data

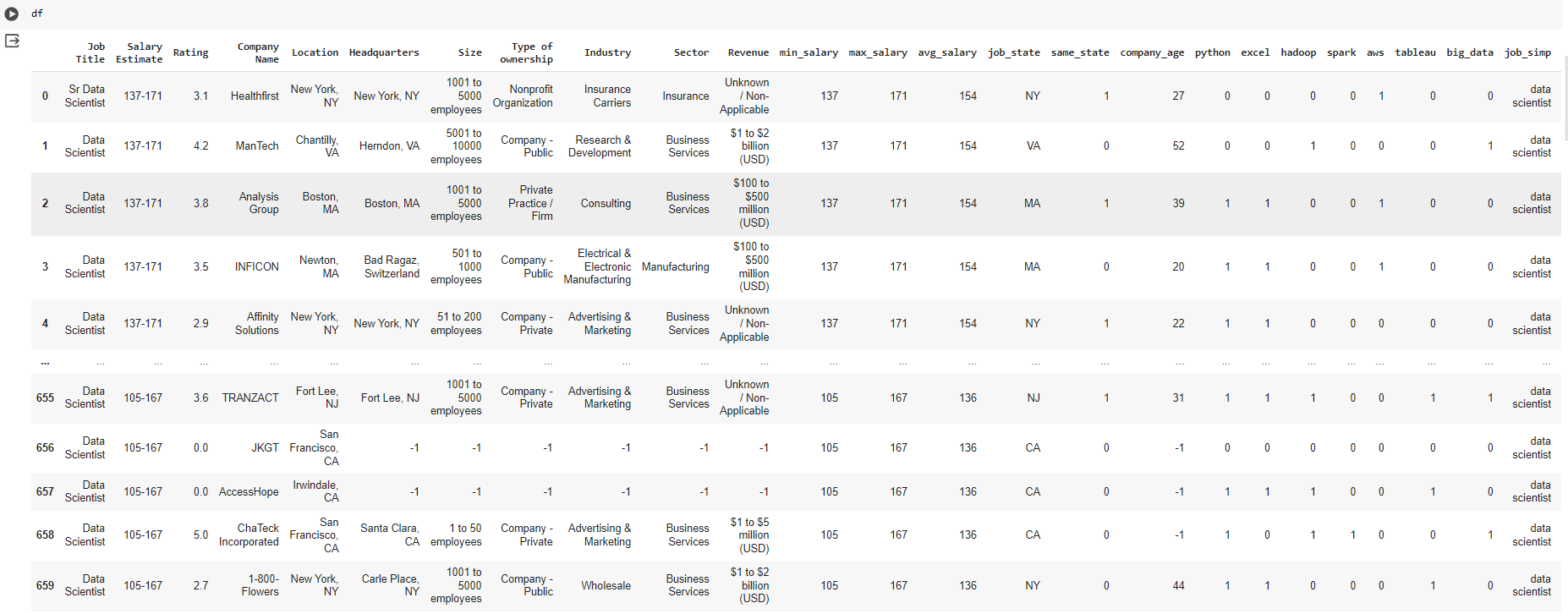
**Although it appears that no data is missing, there are some default data values that are inspected and changed later because they do not represent an actual data value. This could greatly skew (high Kurtosis) numerical columns and lead to less precise statistical analyses on these columns.**

****

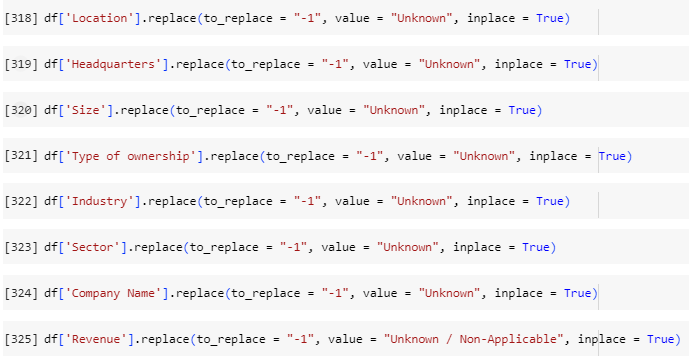
**Step #3:** Replace “-1” with “Unknown” in Categorical Variables

A numerical value of -1 does not make sense for any of the categorical variables, so I replaced this value with Unknown. Note that the Revenue column uses “Unknown / Non-Applicable” as its replacement value because this is category already exists for this column.

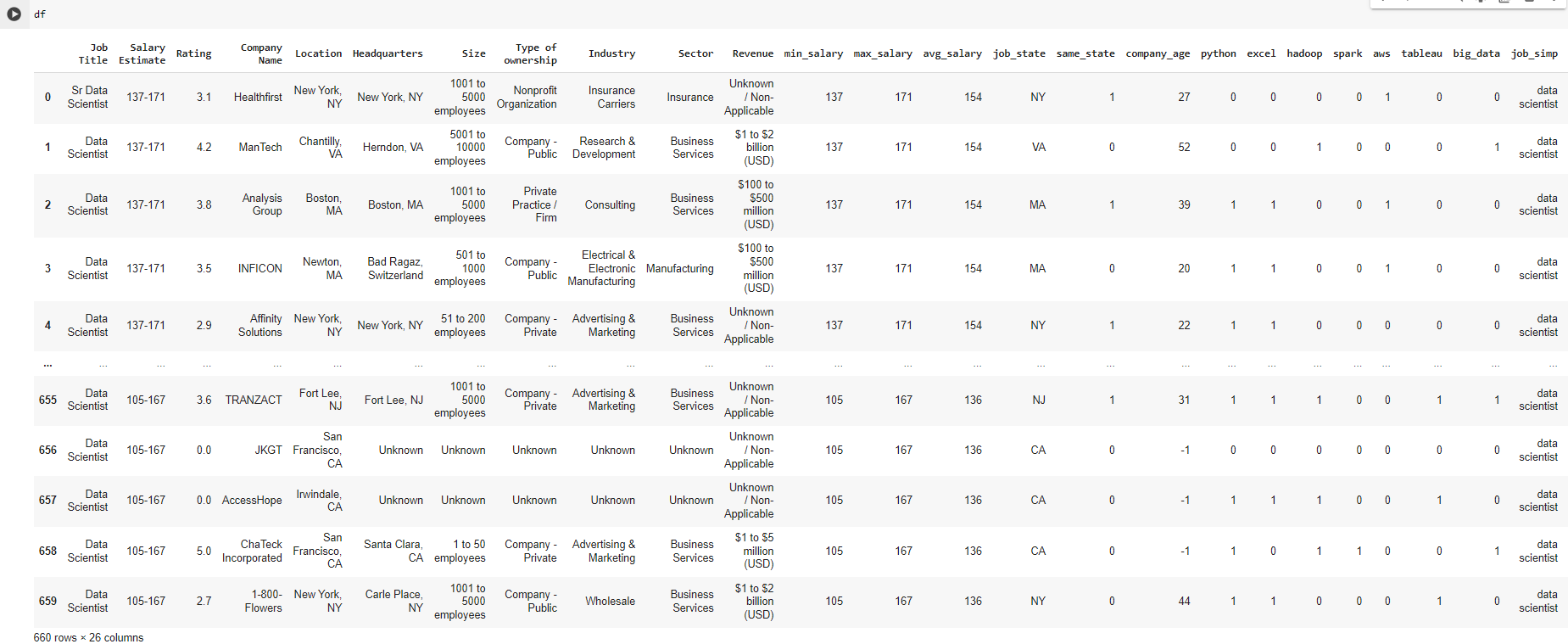
Before:



Code:

****

After:

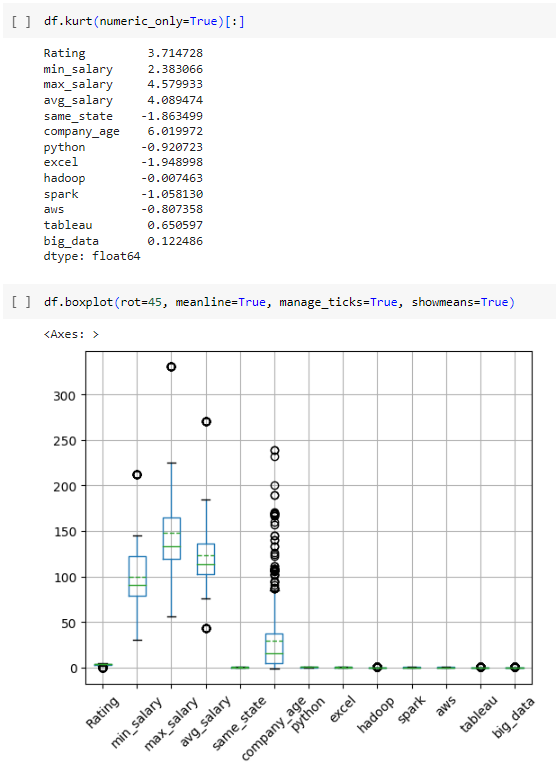
****

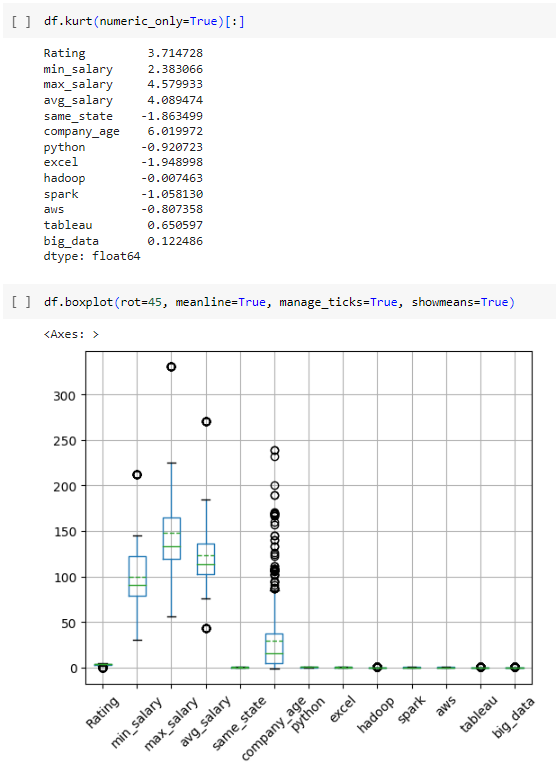
**Step #4:** Check for Outliers

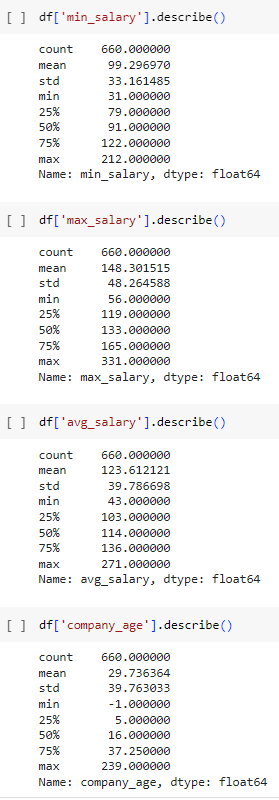
Outliers in a data set are not statically consistent with the rest of the data. This could indicate that they are inaccurate data. Specifically, I decided to investigate and remove outliers in the **min\_salary, max\_salary, avg\_salary, and company\_age columns.**

**After using the describe() function on each column, it appears that “company\_age” has a minimum of -1 and that does not make sense at all. This will be addressed later in step 5.**

This is the code used to check for any outliers:





****

# Step #5: Eliminating Outliers in Min Salary, Max Salary, Average Salary, and Rating in an attempt to remove inaccurate data and to reduce their high Kurtosis values (as shown in step 3).

# Outliers are not always inaccurate, but I have decided to remove them in this data set because I am not able to verify their authenticity.

# Min Salary: Based on the boxplot in step 3, there appears to be an outlier above 200. This will be dropped.

# 

# Max Salary: Based on the boxplot in step 3, there appears to be an outlier above 300. This will be dropped.

# 

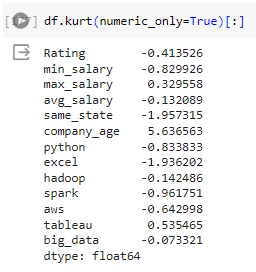
# Average Salary: Based on the boxplot in step 3, there appears to be an outlier above 250 and an outlier below 50. These will be dropped.

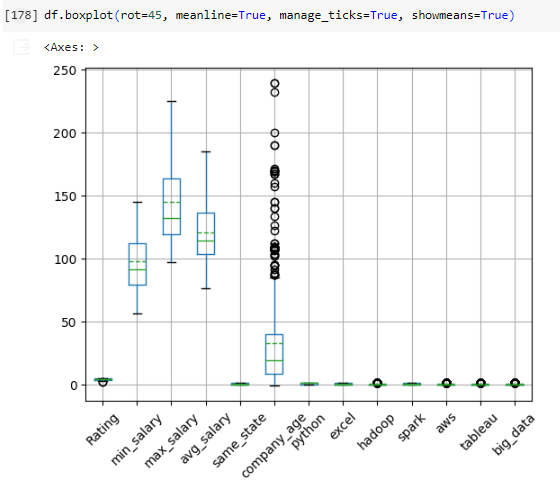
# 

# Rating: Based on the boxplot in step 3, there appears to be an outlier/default value of 0. I also checked for values above 5, but there were none.

# 

Checking the results of my work:

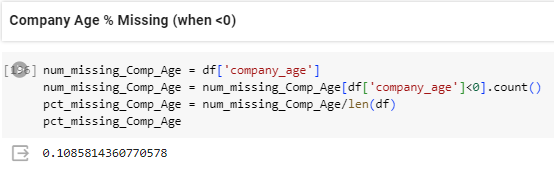




**Step #6**: Investigating the Outliers in Company Age

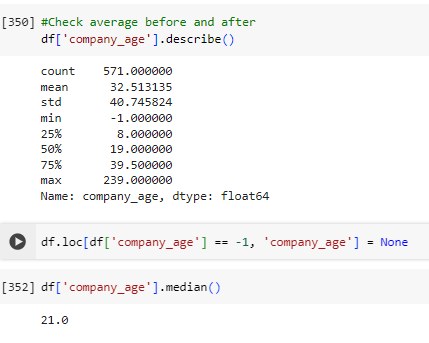
Obviously, a company age value less than zero does not make sense. I suspect that -1 was used as a default value for missing data. First, I check how many rows had an age below 0 (or used -1 as a default). I also decided not to eliminate large outliers in this column because there were many outliers and it was not clear what value to use as an upper limit when dropping rows. I also reasoned that having outliers in company age was certainly possible and were not obviously inaccurate values.

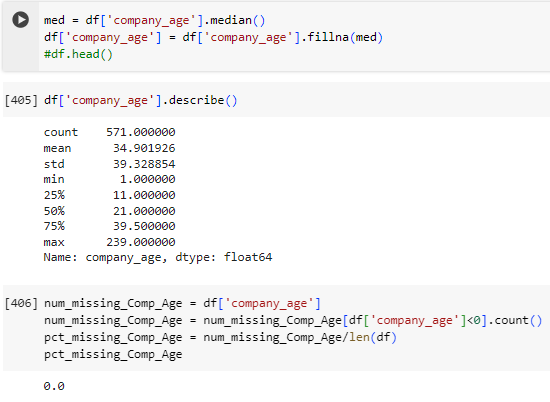
Code:



Since more than 10.8% of rows, dropping these rows would eliminate too much data. Instead, I opted to replace company\_age values of -1 with the median of the data set. This strategy is used because this numerical data is being skewed by ages of -1.

Code:





The describe() function now displays more accurate statistics because ages below zero are no longer considered.

# Step #7: Adjusting Categorical Variables to see if making them all uppercase will eliminate false unique values based on differences in capitalization

# I tried seeing if eliminating differences in capitalization would eliminate non-unique values. This was ultimately unsuccessful, but worth a try with categorical variables. I tried this with the Company Name, Job Title, Industry, and Type of Ownership columns.

# Company Name:

# 

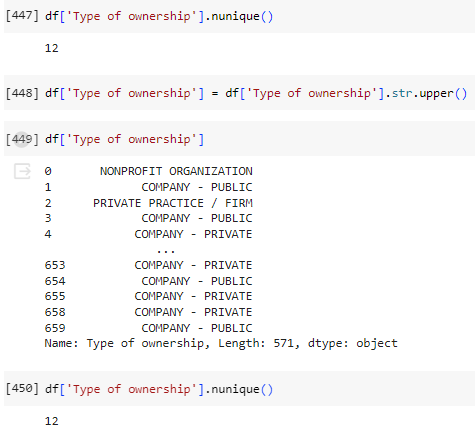
# Job Title:

# 

# Industry:

# 

# Type of Ownership:



**Step #8:** Check Remaining Outliers

The Hadoop, Tableau, and Big\_Data columns also appeared to have outliers according to the boxplot in step 3, but this is actually not the case. After examining them individually, I determined that the values 0 and 1 are valid to represent true and false values. The outliers of 1 are therefore not outliers and were not removed.

Hadoop:

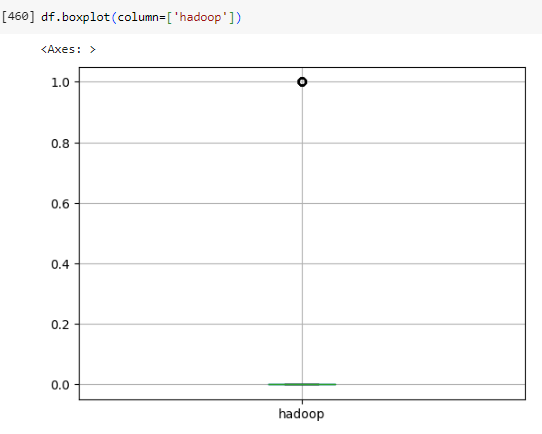
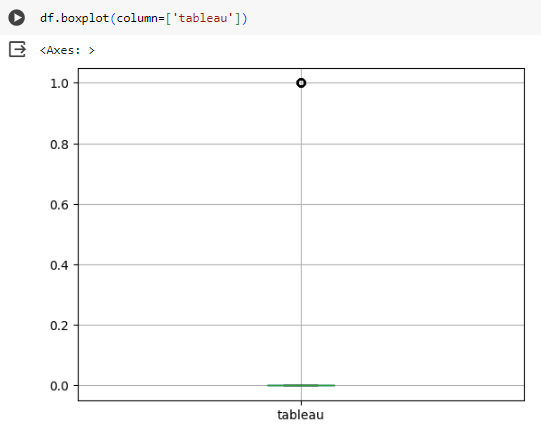


Tableau:



Big\_Data:

