ETL Project

By Sarah Klein, Luis Rojas, Jimmy White

# Extraction

We used two datasets from Kaggle, “How much do data scientists earn in 2017-2020,” and “HR Analytics.”

* **How much do data scientists earn in 2017-2020** – evaluates the impact different variables have on the expected salary of data scientists in the international market.
* **HR Analytics** – gauges the factors that lead an employee to leave a current job, including education, experience, and company loyalty.

# Transformation

Table

Description automatically generatedOur first steps in cleaning up the datasets included learning the definition of each variable and deciding whether it was relevant. An example of this was in the HR Analytics dataset, where “City” was coded by city, but the code was not defined (Figure 1).

Table

Description automatically generatedTable

Description automatically generated

**Figure 1:** HR Analytics – City Column

**Figure 2:** HR Analytics - Renaming values in columns

We found that in the database HR Analytics the column for length of time at their current job was labeled “never” if it was a new job, less than a year. We changed “never” to “0.”

Additionally, in the actively searching for a new job column, 1.0 and 0.0 was used to represent “Yes” and “No” respectively. We changed the values of the “Active” column for ease of readability (Figure 2).

Similarly, the Data Scientist Salary dataset was transformed in several steps. First, any null values were deleted, which still left a more than sizable dataset to analyze at a later point. Then, certain columns were renamed, both for clarity, and the “Education Level” column that occurred in both datasets were renamed to match each other. Then, steps were taken to change the values in the column “Education Level” in the Salary dataset to match those of the HR Analytics dataset. Thus, “Bachelor’s degree” in the Salary dataset became “Graduate,” and so on (Figure 3).

Text

Description automatically generated

**Figure 4:** Data Scientists Earnings – null values deleted and “Education Level” renamed

Graphical user interface

Description automatically generated with medium confidence

**Figure 3:** Data Scientists Earnings – null values deleted and “Education Level” renamed

# Load