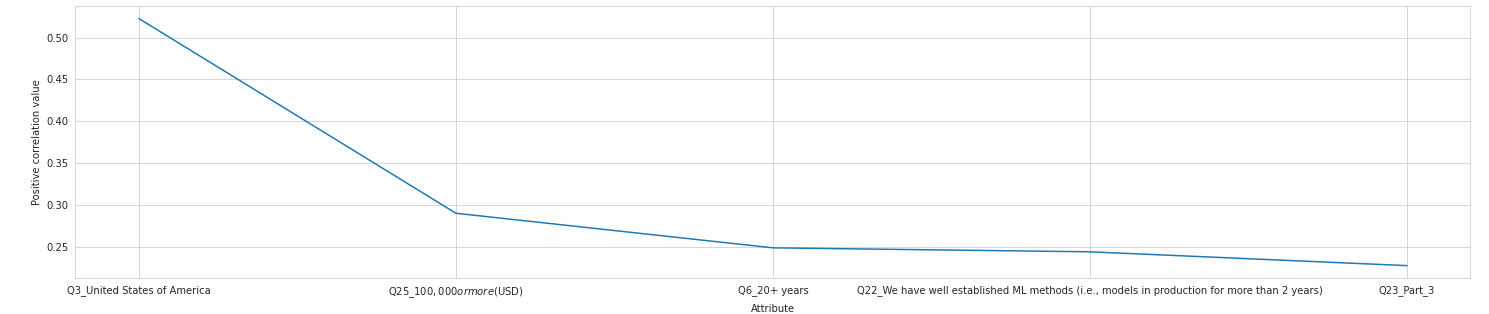
**Question 1:** Data Cleaning

The data contains many null values, by performing a count of null values in the data, I can see which columns to clean and handle them. I focused on columns without multiple parts since the data will be separated into separate columns using dummy encoding. 'Q8', 'Q11', 'Q13', 'Q15', 'Q25', 'Q30', 'Q32', 'Q38' have null values. Q8, 11, and 13 had their missing values filled using the mode because one of the values dominated the column. Q15 has the value of ‘Under 1 year’ and ‘Not using ML’ may be similar based on how you interpret the question. Since Q25 doesn’t have many null values (~1-2%), I dropped the null values. I dropped columns 30, 32, and 38 because they contain too many missing values.

I then converted the categorical data into numerical data by dummy encoding. Dummy encoding allows us to create a new column for each unique value of a feature, depending on the column, this will create many extra columns. The dummy variable sets the value as 1 and all other values as 0 for that column. This encoding was also used because there are questions that are answered as multiple parts and placed into different columns (Ex. Q7 Part 1, etc.). By using this method, I do not have to find a way to add up the columns or fill their missing values. By converting the data into numerical data, it is easier to work and process with.

**Question 2:** Exploratory Data Analysis and Feature Selection

There are multiple methods on feature selection and determining the most important features. After cleaning the data, I used the correlation values between the salary buckets and the encoded features to see which features are the most important on determining a higher salary.



**Question 3:** Estimating the difference between average salary between education level.