# Analysis Report

## Data Summary

The dataset refers to pay rates, geographical locations, job titles and job categories of occupations around the globe. It has 19 unique variables and 1655 observations. The dataset includes both categorical and continuous variables. The focus of this report is mainly on : ***location_country*** : country where the job is located, ***annual_base_pay*** : The annual base pay of job (min : 0, max :10.28m, mean : 136406, median : 100000). This report is formulated to compare the trends of annual base pay in the ***United States*** versus ***Other Countries***.

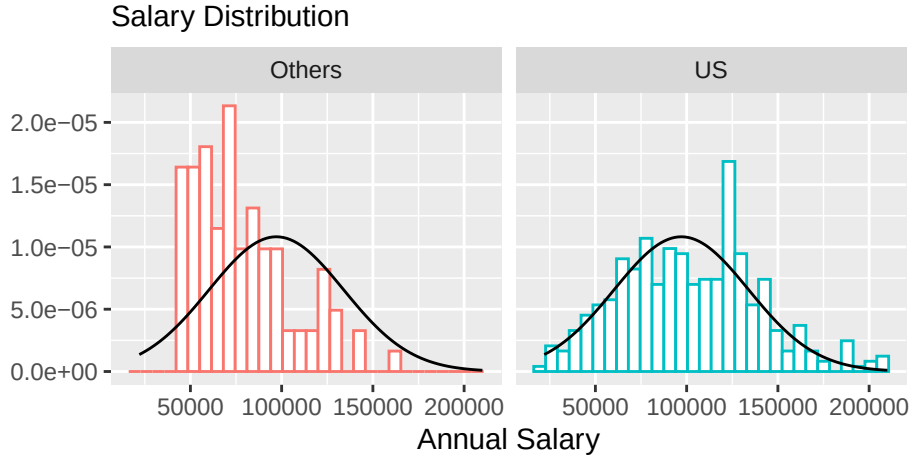
## Data Cleaning and Pre-processing

After closely inspecting the data, observations were found to have "NA" values, as these values may pose a bias in the results, for the scope of this analysis, NA values are omitted leaving 561 observations. The analysis is performed on salaries in the US vs Other Countries so ***annual_base_pay*** and ***location_country*** variables are selected, rest of the redundant data is dropped. Furthermore, annual_base_pay variable is of category "num", which is changed to "double", since this variable is used for mathematical calibrations and double is the preferred datatype for such calculations. The variable location_name is of type "chr", and is converted to "factor" as factors take up less space in memory and comparison computation is also quicker. After converting the type of variables, the data is tested for any missing values of salaries, which turned out to be nil.

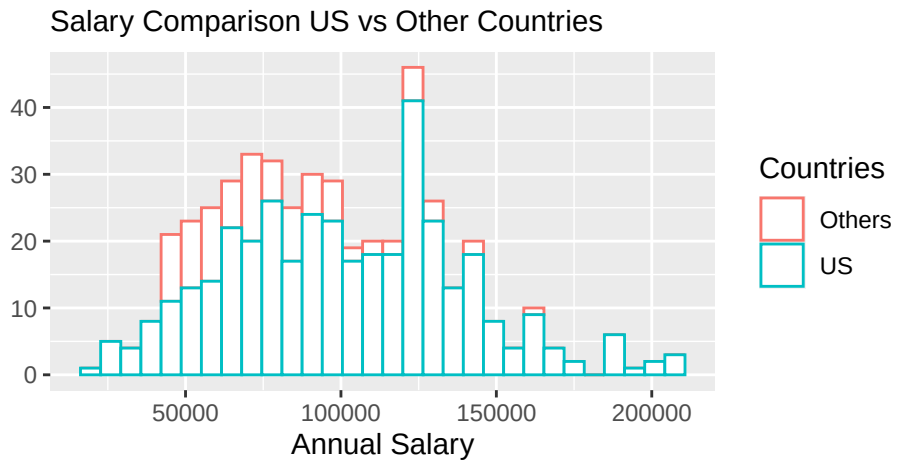
## Planning And Analysis

After initial observations, it is observed that the salaries in the US are on a higher end when compared to rest of the world in the dataset. The hypothesis formulated from this observation is **"Annual base pay rate is more in the US compared to Other Countries"**. Analysis is performed to test the hypothesis that includes plotting the data, for plotting, it is suitable if there is a variable to discriminate between the US and rest of the countries, thus the data is mutated to add a new variable ***territory*** of type "factor". The plotted data outlines a few outliers. When the data is analysed, some of the salaries between 0 and 200 are detected, these values are erroneous because yearly salary cannot be in this range and are removed. A few of the values are close to 10m which in turn are shifting the mean of the salaries, it is therefore justified to remove the extreme values and maintain an even distribution of pay rates. To remove such outliers Interquartile Range Method[1] is used in this analysis. Further interesting fact is that the US has the maximum number of tech jobs in the dataset.

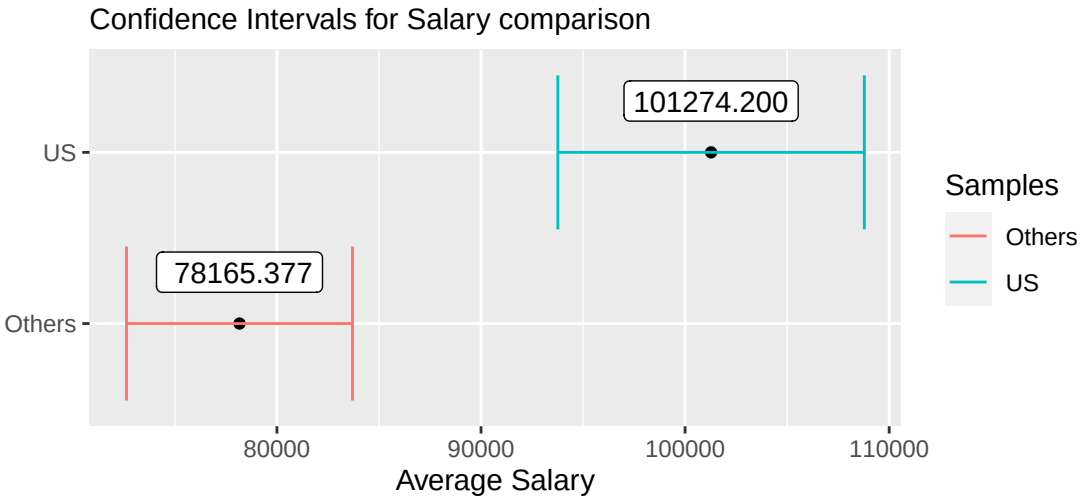
Once, the outliers are removed, the data is examined for normal distribution by plotting the Q-Q graphs, which showcased that the salary distributions of the US and Other Countries follow a near normal distribution. After closely examining the data it is detected that salaries in the US are spread throughout ranging from 50k to above 200k, whereas for Other Countries, the density is prevalent towards the lower end and only a few are between the range of 150k and 160K. This in turn supports the hypothesis stated in the analysis. Closely scrutinizing the distribution graphs, it infers that the mean salary for the US falls closely near the 100k mark as compared to Other Countries which is below 100K thereby complementing the hypothesis.




[1]Q1 = First 25% of observations, Q3= First 75% of observations, IQR = Q3-Q1 (middle 50% of observations). Any observation falling outside the fence of 1.5IQR+Q3 and 1.5IQR-Q1 is removed




The salary distribution so far is supporting the hypothesis but to confirm further, this analysis incorporates the 95% Confidence Intervals[2]. The datapoints for both the US and Other Countries $> 30$, hence by invoking the Central Limit Theorem[3], the distribution of salaries is treated to be normal. The plot for Confidence Intervals[4] represents the 95%CI for the sample of the US and Other Country's salaries. The CI interval for the US ranges between (97953.3 , 105576.6) and for the Other Countries between (69841.8 , 78964.6) .It is evident from the graph that the CIs of the samples do not overlap, which proves that the samples come from different set of populations and one population has interval higher than the other. These findings supplement the hypothesis even further that the salaries in the US are more when compared to the Other Countries.




## Conclusion

In order to test the hypothesis **Annual base pay rate is more in the US compared to Other Countries**, multifarious steps were performed. Primitively cleaning and refining the data, mutating the variables, visualizing using graphs and discovering the outliers and anomalies, removing the outliers using Interquartile Range Method, checking for the data to be normally distributed, invoking Central Limit Theorem to calculate 95% Confidence Intervals for samples in the process. The average salaries for 95%CI falls under the range (97953.3 , 105576.6) for the US and (69841.8 , 78964.6) for the Other Countries. The stated measures provided a step by step approach towards confirming and supporting the hypothesis. The above mentioned analysis clearly outlines that the salaries in the US are far greater than the Other Countries. Thereby, proving the hypothesis "Annual base pay rate is more in the US compared to Other Countries".

[2]95% probability of output to fall between the two intervals.
[3]Central Limit Theorem states that if the random sample size is sufficiently large than the sample distribution means follow an approximate normal distribution
[4]Confidence Interval states that the probability of output parameter to fall between set of two values.