Tech Industry Salary 2016

1. Data

The dataset contains survey data on tech industry salaries in 2016, including job titles, experience, location, and bonus information. **The study aims to investigate two key hypotheses: first, impact of geographical location on salary, and second, influence of experience on salary.** This dataset contains 1655 observations and 19 variables of various types, but for the purpose of our analysis, the relevant columns include salary_id (integer) which are unique identifiers, location_latitude (numeric) and location_longitude (numeric) in degrees which provide geographical location of the job, total_experience_years (numeric), employer_experience_years (numeric) and annual_base_pay (numeric) which is the annual base salary in USD and it is used as a basis for comparing salaries as it has the least missing values. Refer to appendix-1 for more on missing values. Since the dataset is a result of a survey it made sense to remove the outliers which could lead to distortion of results. They were removed using 1.5*IQR (Inter-Quantile Range).

2. Planning and Analysis

For our analysis, we use the cut() function, which splits the “annual_base_pay” column into 4 groups based on the range: low: <50k, medium: <75k, high: <100k, very high: >100k. This makes it easier to analyze how pay is related to different salary groups and spot any trends or patterns.

Geographical Locations vs Salary

The first hypothesis examines the relationship between salary and geographical location using a map visualization. The longitude and latitude of job locations are plotted to identify patterns or clusters of high and low salaries. Figure 1 illustrates the global distribution of salaries, with most high earners located in North America and a minimal representation in Europe. Asia and South America have primarily low and medium earners. Figure 2 shows the distribution within the USA, with clusters of high and very high earners in cities like San Francisco, San Jose, Los Angeles, and San Diego. The east coast, including cities such as New York, Boston, Washington, and Baltimore, have a mix of high, very high, and medium earners. The San Francisco area displays a notable contrast, with a lack of medium and low earners compared to the diverse distribution in New York. Further details can be found in the appendix-2.

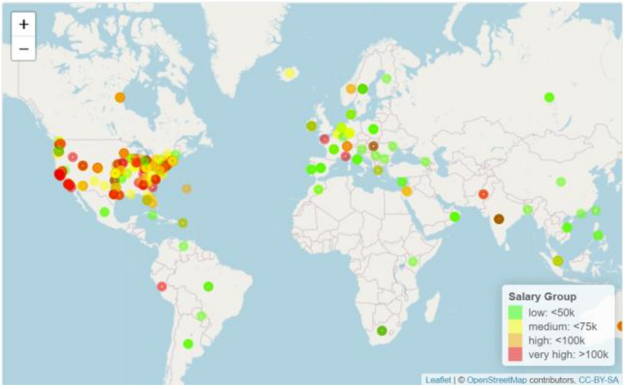

Figure 1

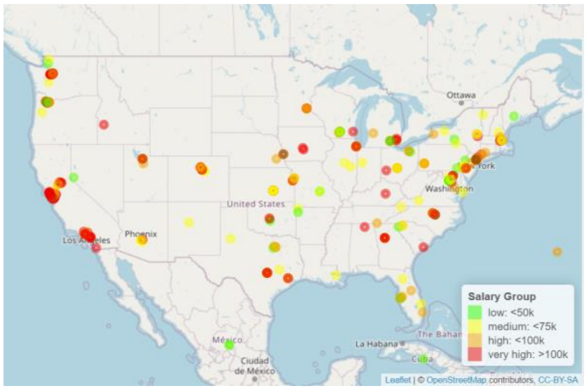

Figure 2

Experience vs Salary

The Shapiro Wilk, normality test shows that at 0.05 significance level the “total_experence_years”, “employer_experience_years” and “annual_base_pay” do not follow a normal distribution and visualizing their Q-Q plots (appendix-3) we can confirm the same. Since the assumption of normality is violated, we use Spearman's rank correlation a non-parametric test for calculating correlation.

```
Spearman's rank correlation rho
```

Total exp years vs Annual base pay	Employer exp years vs Annual base pay

S = 402076136, p-value < 2.2e-16,
rho = 0.2951122
	S = 512833984, p-value = 8.662e-05,
	rho 0.1009404

The Spearman's rank correlation results show a positive correlation between the two variables "total_experience_years" and "annual_base_pay" with a correlation coefficient (rho) of 0.2951122 and a p-value < 2.2e-16, suggesting a weak correlation. It also indicates a weaker correlation between "employer_experience_years" and "annual_base_pay”. Figure 3,4 show the correlation plots of the same.

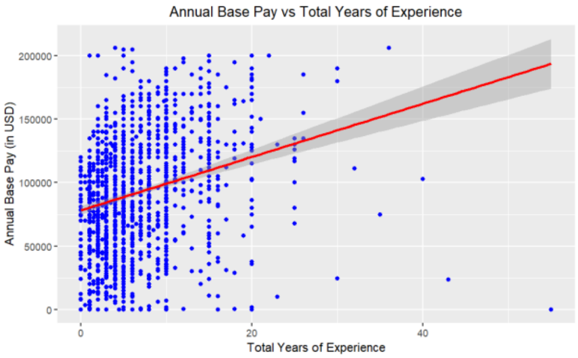

Figure 3

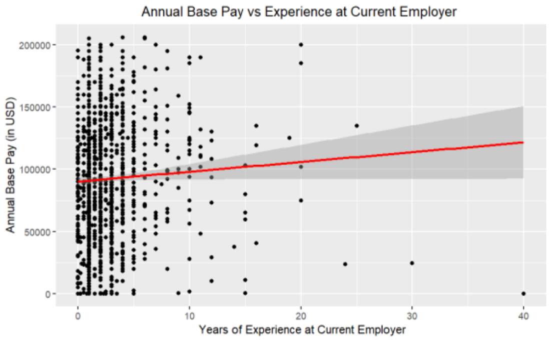

Figure 4

To investigate this further, we create a new column "experience_group" which groups the values in the "total_experience_years" column into different categories. Figure 5 illustrates the distribution of annual base pay across various experience levels using violin plots. The quantile lines for each group further reinforce the idea that each experience group possesses a distinct distribution. We use the Kruskal-Wallis rank sum test which is a non-parametric method for comparing the central tendency of two or more groups.

```
Kruskal-Wallis chi-squared = 117.57, df = 4, p-value < 2.2e-16
```

The low p-value indicates that there is a statistically significant difference in the annual base pay among the different experience groups.

3. Conclusion

1. The geographical location of a job plays a significant role in determining salary, with high earners primarily located in North America and low earners in Asia and South America.
2. Within the USA, cities such as SF, San Jose, Los Angeles, and San Diego have a high concentration of high and very high earners, while the east coast has a mix of high, very high, and medium earners.
3. There is a statistically significant difference in the annual base pay among the different experience groups.

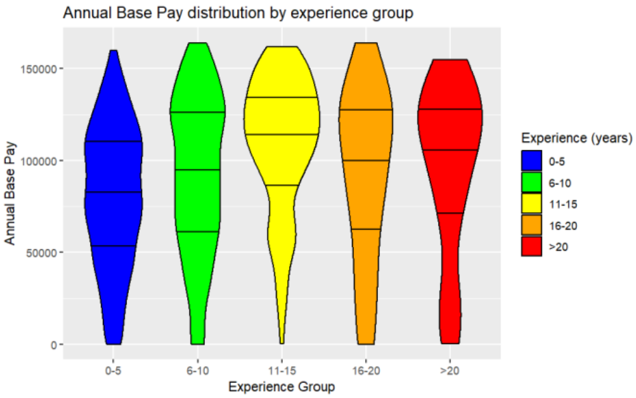

Figure 5