Jeff Pinegar
jeffpinegar1@gmail.com
717-982-0516

Assignment: SQL - Challenge
Due Date: Nov. 30 2022

SQL – Challenge: Pewlett Hackard

Bonus Question

As you examine the data, you suspect the dataset is fake. Maybe your boss gave you spurious data to test the data engineering skills of a new employee. To confirm your hunch, you decide to create a visualization of the data to present to your boss.

Test 1: Histogram of Salaries

Examining the distribution of Salaries for the approximately 300,000 employees, we see Figure 1. The distribution is very suspect for the following reasons:

- A very large number of employees in the $40,000 to $45,000
- The absence of salaries below $40,000
- The long tail to the right like does not go as high as I would anticipate

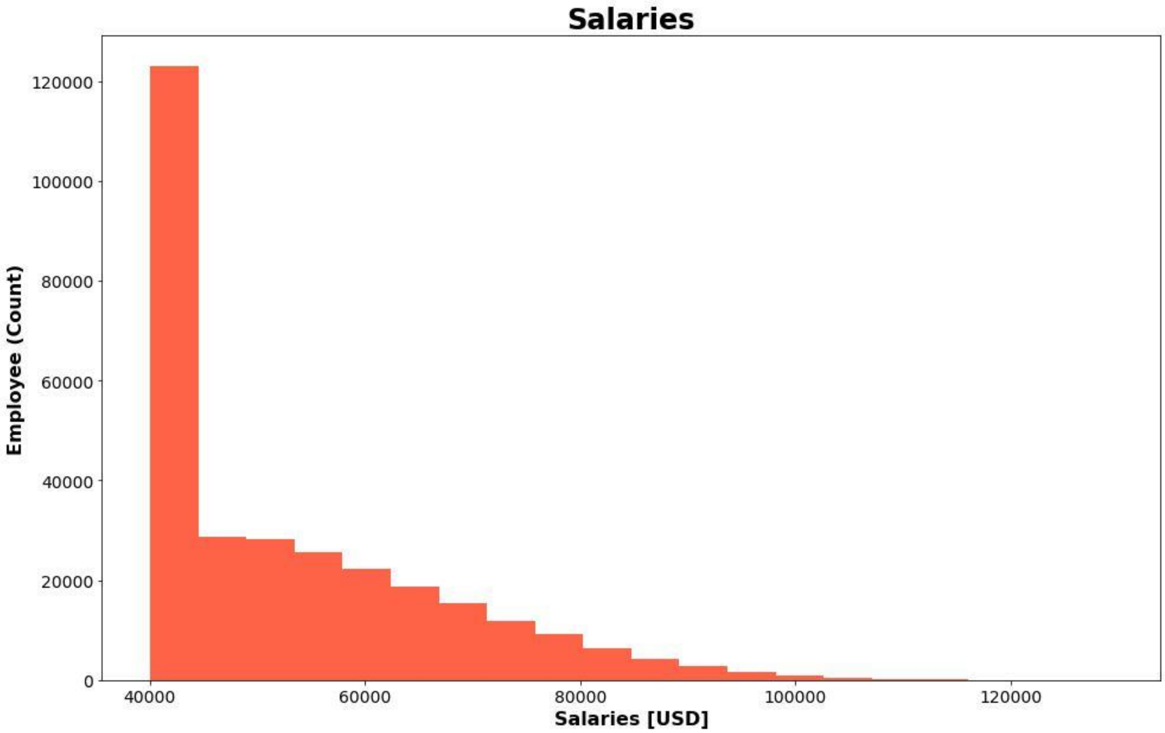

Figure 1: Salary Histogram for Employees

Test 2: Average Salary by Position

Examining the average salary by position, we see Figure 2. This reveals obvious issues with the data:

- Assistant Engineer's salary is greater than a Senior Engineer and Engineer
- The salary of a Senior Engineer is only slightly above that of an engineer
- The salary of staff is equal to the salary of Senior Staff
- The average salary shows very little variation by position, much less than should be expected.

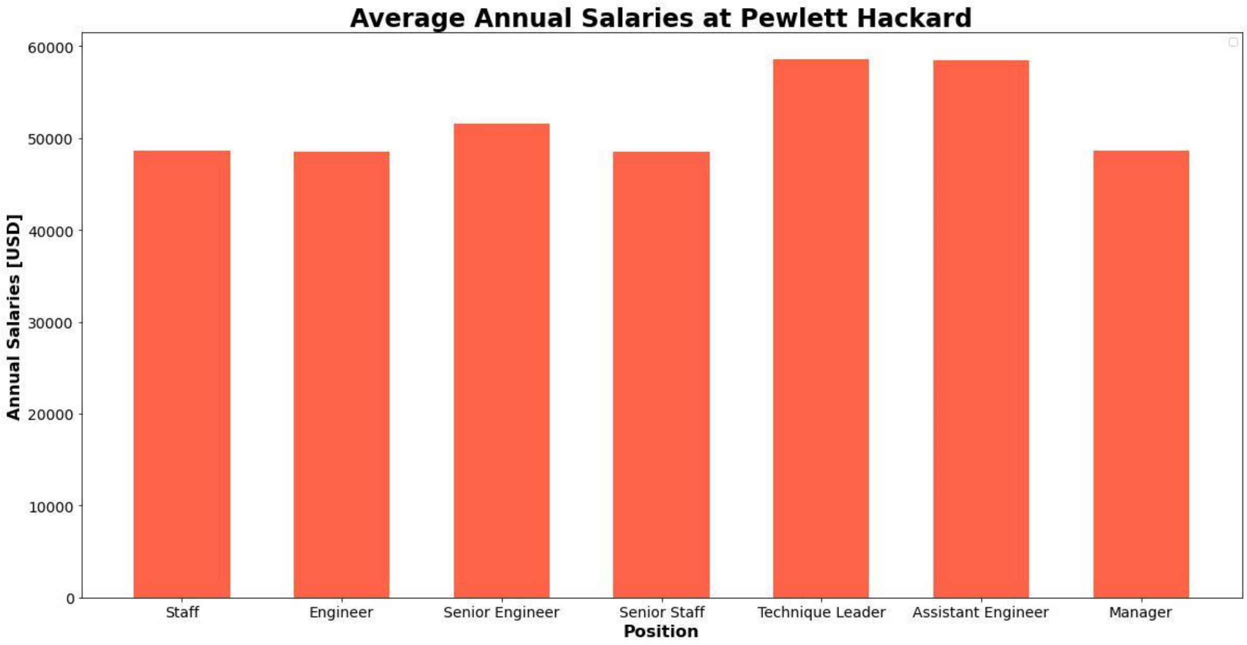

Figure 2: Average Salary by Position

Conclusion:

This data is cooked.

This data could be only accurate data if the company were very unusual. You would need an extremely high number of people with the same position (call center, retail, etc.) to explain Figure 1. For figure 2 to be true you would need a business with a salary plan that was heavily bonus-based, where everyone gets a minimum base salary and then bonuses to make up the bulk of their income and differentiate the income by position. If this was the case even a new analyst would know this.