Rithwik Statistical and EDA Analysis observations:

- Shape of the data: (100000, 18)
- education and skills are empty in every row so have to remove those columns
- experience_level and employment_type have some null values
- Duplicate Records: 10100
- Thought of mapping unknown experience_level to the labels junior, mid, senior and lead as these are ordinal.
- And also same for employment_type based on experience_level.
- But found all have same median and also identical 25th, 50th, 75th percentiles and a very similar mean which doesn't tell anything about relationship to classify the unknown experience_level based on years_experience
- But found all have same median and also identical 25th, 50th, 75th percentiles and a very similar mean which doesn't tell anything about relationship to classify the unknown employment_type based on years_experience
- so if we remove 20% data will be removed. so better to replace unknown with mode.
- experience_level mode used: Mid
- employment_type mode used: Part-Time
- software engineer and data scientist are not spelled exactly in each row causing different labeling. So have to label same.
- Distribution of base_salary, salary_in_usd, conversion_rate, adjusted_total_usd and remote_ratio are not normally distributed
- Box plot of base_salary and adjusted_total_usd have outliers and have to treat them.
- base_salary is highly correlated to adjusted_total_usd