**EDA\_Presentation\_Points:**

We conducted EDA, to identify the relationship between the variable of interest which is early career pay and other variables. We listed several questions to begin the EDA analysis and are here in the slide. For this analysis, we used both R and python. The very elegant and versatile ggplot2 package was used for EDA in R. Equally well known Matplotlib and seaborn libraries were used for EDA in python.

**-> (Next slide please) (Slide:9 Pair plot)**

We used the pair plot function from seaborn to understand the relationship between Early career pay with each variable (Region, Division, Make world Better and Stem Percent). Based on this, we used specific plot types to enhance the visualization.

**-> (Next slide please) (Slide:10 Box plot)**

Then box plot was used, when the independent variable was categorical. In this fig, we looked at the relationship between each States and Early career pay. We can see Mississippi has the lowest while New York has the highest average early career pay.

**-> (Next slide please) (Slide:11 Box plot)**

This fig shows the Early career pay verses Division. Here, we observed East south-central division which includes Mississippi has the lowest average early career pay while the Middle Atlantic division that includes New York has the highest average early career pay.

**-> (Next slide please) (Slide:12 Box plot)**

This is one of the box plots using ggplot in R. we used faceting function to further influence the distribution. When we applied faceting function, we observed private colleges from public colleges.

**-> (Next slide please) (Slide:13 Scatter plot)**

Scatter plot was used when the variable is continuous. Here are the couple of scatter plots we used in our analysis. In this fig, we can see the positive correlation between Early career pay and stem percent and don’t see the same between Early career pay and make world better.

**-> (Next slide please) (Slide:14 Scatter plot)**

To better understand the relationship between multiple variables, we used the 'hue' function in seaborn and 'facet wrap' in ggplot2 on top of coloring the variables. Here, we can clearly see the relationship between Early career pay and In-state is totally influenced by the college type.

Thus, our EDA analysis provided us an understanding on, which are the factors influencing the Early career pay. We used these insights from EDA to perform Machine learning analysis.

Next, **Jumoke** will continue with **Machine learning part**.