**Analysis for the project – Q&A’s as per code:**

Summary:

Methods used: Data Preprocessing, Clustering, Cluster Interpretation, Cluster Analysis and Business Insights, Critical Thinking and Problem Solving, Statistical Analysis

1. Column salary\_in\_usdis the target column, and the rest of the ten features are considered this project's feature set. From now on, we refer to the feature set as X and the target column as y. What is the average value of column 'experience\_level' in X? Enter the answer with two digits after the decimal point..

* 1.82

2. Now scale the feature set with MinMaxScaler. Store the scaled dataset in a data frame named X\_scaled with the same column names as X. What is the average value of column 'experience\_level' in X\_scaled? Enter the answer with two digits after the decimal point.

* 0.61

3. Now, train a KMeans clustering algorithm on the scaled feature set with 3 clusters and random\_state = 0. Assume the default values for the rest of the parameters. Enter the number of instances in the largest cluster..

* 329

4. As you can see in the results, one of the KMeans clusters is significantly smaller than the other clusters. This may raise the question that if the number of clusters = 3 is a good choice. Fortunately, there is an effective technique that helps us to choose the best number of clusters. For this purpose, we use a function called silhouette\_score. Please take some time to read about [this functionLinks to an external site.](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html). Now run this function on KMeans model with random\_state = 0 and n\_clusters in the range(2,10). Which number of clusters has the highest silhouette\_score. Enter the highest silhouette\_score with two digits after the decimal point.

* 0.41

5. What are the best n\_clusters for the scaled dataset?

* 9

6. (Regression) Now, train a KMeans clustering algorithm on the scaled feature set with n\_clusters you found using silhouette\_score and random\_state = 0. Assume the default values for the rest of the parameters. Enter the number of instances in the largest cluster.

* 143