

Bits, pilani work integrated learning program



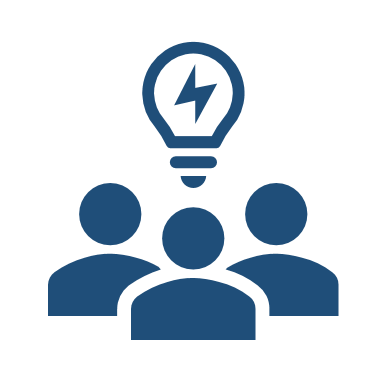
Introduction to Data Science

Analysis on New York Job Posting Data

Submission on 16 Dec 2019

**Introduction to Data Science**

**New York job posting data analysis (2019\_Cluster-DSE-IDS\_A1\_PS2)**



**Venkataramanan K | 2018AC04529**

**Balakavin Pon | 2018AC04531**

**Ponvani | 2018AC04559**

**Poornima J |2018AC04550**

Analysed and created in collaboration of

**IDS\_GROUP006**

**2019\_Cluster-DSE-IDS\_A1\_PS2**

Group information

# Table of content

[Table of content 3](#_Toc27431326)

[1 | Problem statement 4](#_Toc27431327)

[Business Problem 4](#_Toc27431328)

[Summary of Analysis 4](#_Toc27431329)

[Data Exploration 4](#_Toc27431330)

[Feature Engineering 4](#_Toc27431331)

[Analysis Done 4](#_Toc27431332)

[2 | Data Preparation 5](#_Toc27431333)

[3 | Identification of variables 6](#_Toc27431334)

[4 | Missing values and Variable selection 7](#_Toc27431335)

[5 | Feature Engineering 8](#_Toc27431336)

[Salary Range 8](#_Toc27431337)

[Years of Experience 9](#_Toc27431338)

[Preferred Skills 10](#_Toc27431339)

[6 | Analysis 12](#_Toc27431340)

[6.a Highest Paid Skills in US Market 12](#_Toc27431341)

[6.b Job categories involve the mentioned Niche Skills 12](#_Toc27431342)

[6.c Clustering on the data 13](#_Toc27431343)

[Normalize the data 13](#_Toc27431344)

[Building Cluster Model 13](#_Toc27431345)

# 1 | Problem statement

## Business Problem

Determine the below from the given data set for New York City Current Job Posting data.

1. What are the highest paid Skills in the US market?
2. What are the job categories, which involve above mentioned niche skills?
3. Applying clustering concepts, please depict visually what are the different salary ranges based on job category and years of experience.

## Summary of Analysis

### Data Exploration

The data is analyzed to summarize their main characteristics. Steps taken

* The percentage of null values are computed for each variable. If there are more than 30% null values, that variable is not considered for analysis.
* Handling missing values – Has been handled by replacing with most frequent occurring data (Mode strategy)

### Feature Engineering

Feature engineering is done to make the input dataset compatible with the machine learning algorithm requirements and improve the performance of machine learning models. Steps taken

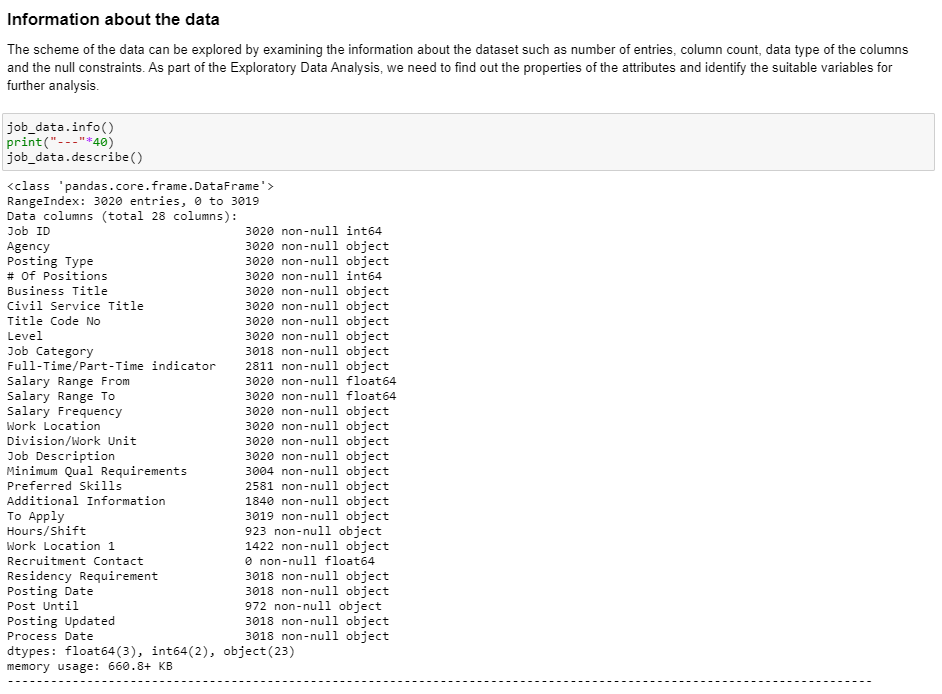
* Remove unwanted characters such as special characters, unwanted whitespaces and punctuation.
* Remove Stop Words
* Extracting some specific attributes

### Analysis Done

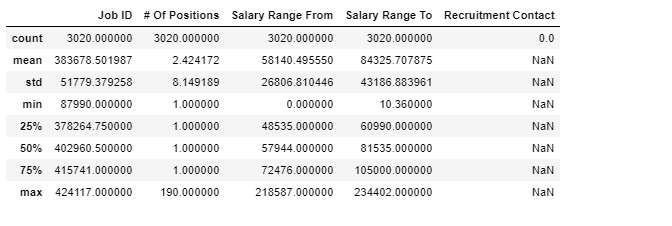
* Highest paid skills was derived by the grouping by job category, preferred skills with salary range mean and picking the top 10 in this list.
* For clustering , bag of words transformer, word vectorizer was explored. K means clustering mechanism was used.

# 2 | Data Preparation

Data preparation is the for most important step we could explore the dataset to understand the features and their relationship using both graphical and non-graphical quantitative analysis. The first step is to get to know about the size of the dataset such as the number of records given, number of columns and their data types.



Then followed by the overall descriptive analysis of the dataset, which gives us an insight on the statistical view of the numerical features.



# 3 | Identification of variables

Selection of appropriate variables for analysis plays an important role. We need to explore the data set and see the importance and the relationship of the variables. That will enable us to select the features more suitably for the analysis.

After we performed the preliminary analysis on the variables data type and the their values the decision was made that there are about 5 features would contribute enough for the further requirements in the analysis. The features are such as,

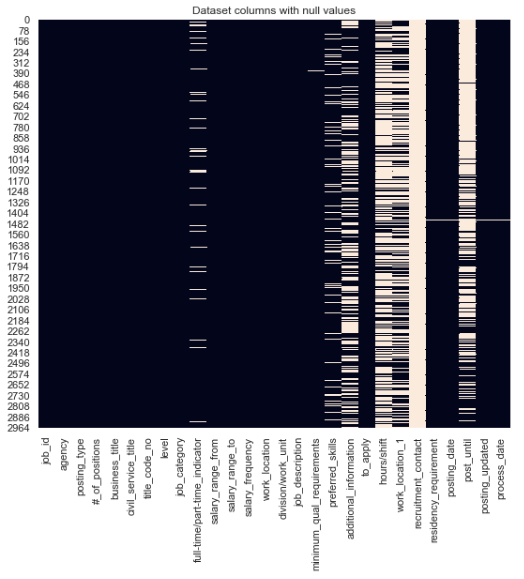
1. Salary Range From
2. Salary Range To
3. Salary Frequency
4. Preferred Skills
5. Job category

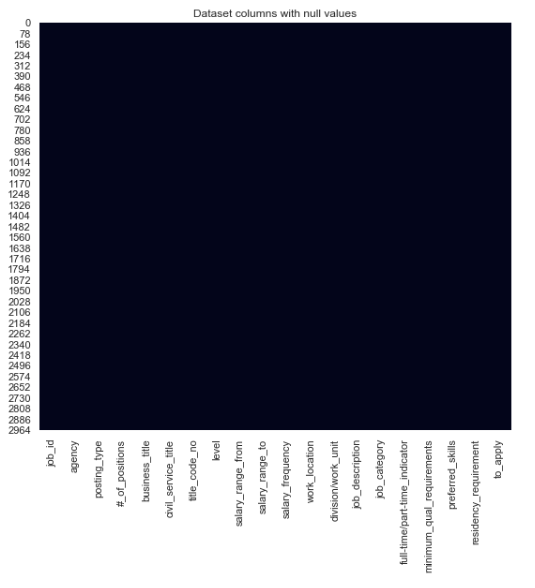
The derived fields are,

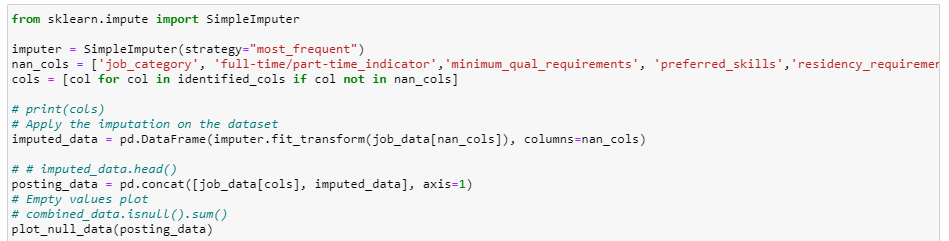
1. Years of experience
2. Projected salary from
3. Projected salary to

# 4 | Missing values and Variable selection

Imputing the missing values is a crucial step in data preprocessing. The null or empty data in the data fields will not give an expected result in the performance of the built model. We need to impute or fill an appropriate value for those missing fields.

The following type of heatmap depicts the null values of variables against the rows.

After the imputing the fields with most frequent and appropriate values the fields null or NaN data are removed. The plot of after imputed data is given below, where there are no white spaces in the any of the rows of the variables.



# 5 | Feature Engineering

Feature engineering is concerned with the process of creating features or manipulating the features to make more meaningful for the analysis.

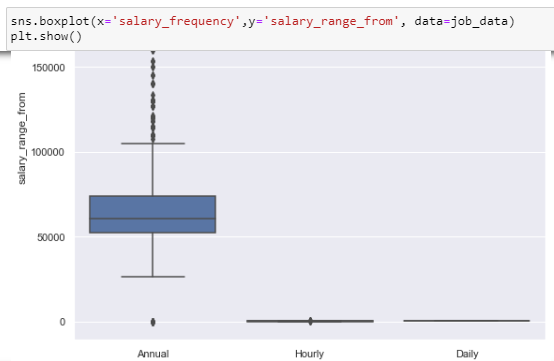
Here in the given dataset, most of the feature data are text based in nature. In order to indulge them in the analysis we need to perform certain feature engineering mechanism such as text parsing, n-gram phrase generation, field extraction, manipulation and deduplication. The 3 import features were taken for feature engineering for this analysis,

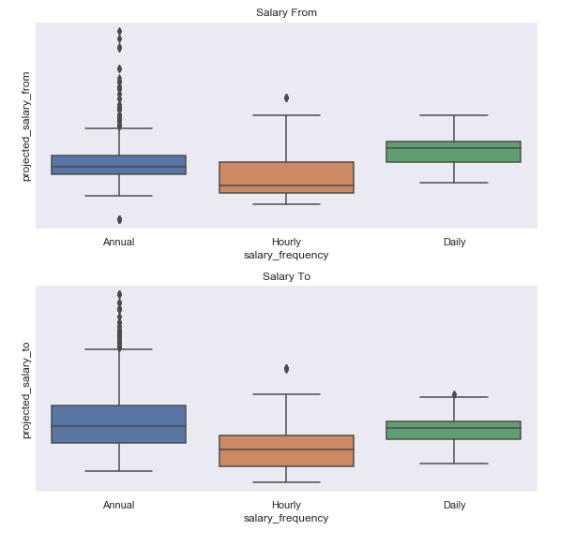
1. Salary Range From and Salary Range To
2. Years of Experience
3. Preferred skills

### Salary Range

The fields salary range from and to are the important fields with respect to arrive at the results for the analysis, because they are necessary to compare and group the niches skills and job category. However, these fields data are not properly given in terms of the value units. Some records have the money value in hours and some of them are in daily and annullay. It is important we need to standardize them based on the common frequency.

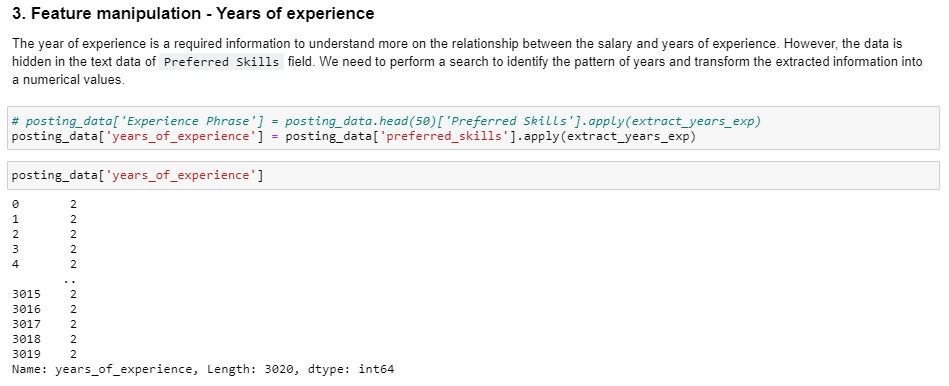
Therefore, the all the records salary information will be commonly treated based on the one frequency either daily or annually. We took the frequency as annually for further steps in the analysis.





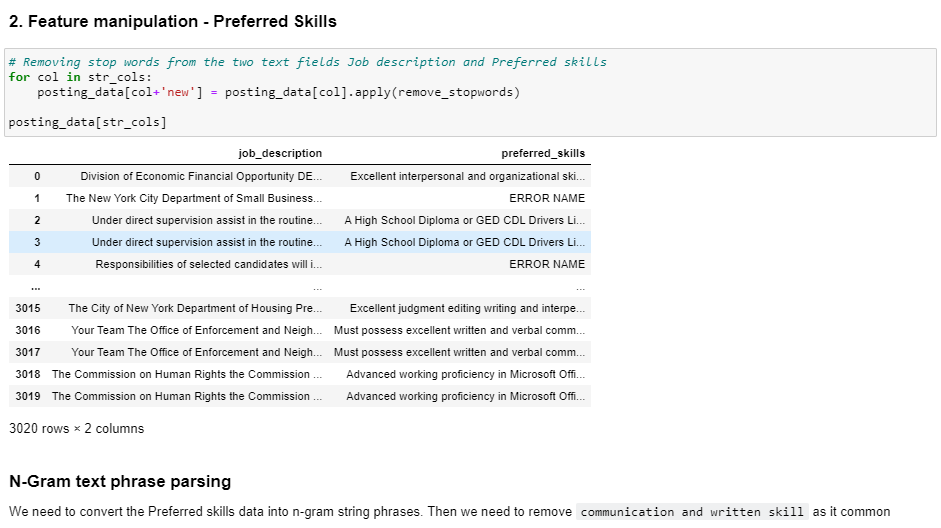
Once the salary range fields are standardized using the salary frequency field, they are plotted in whisker or box plot to visualize the data spread across different quartile.

### Years of Experience



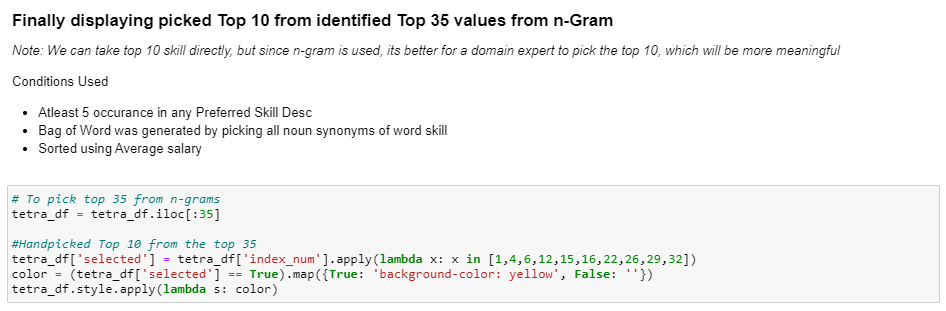
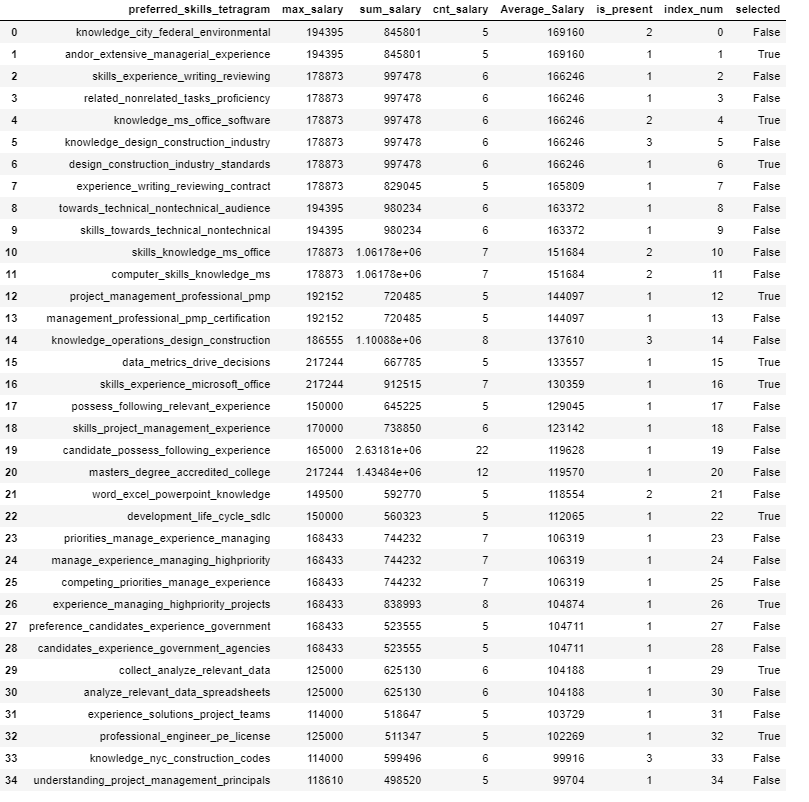
### Preferred Skills

The field preferred skills are the most important feature of the given job posting dataset. Because it had information for skills as well as the years of experience. Based on these only the entire analysis was involved. We need to implement many text manipulation mechanisms to extract meaningful information.



Algorithms used for text manipulations,

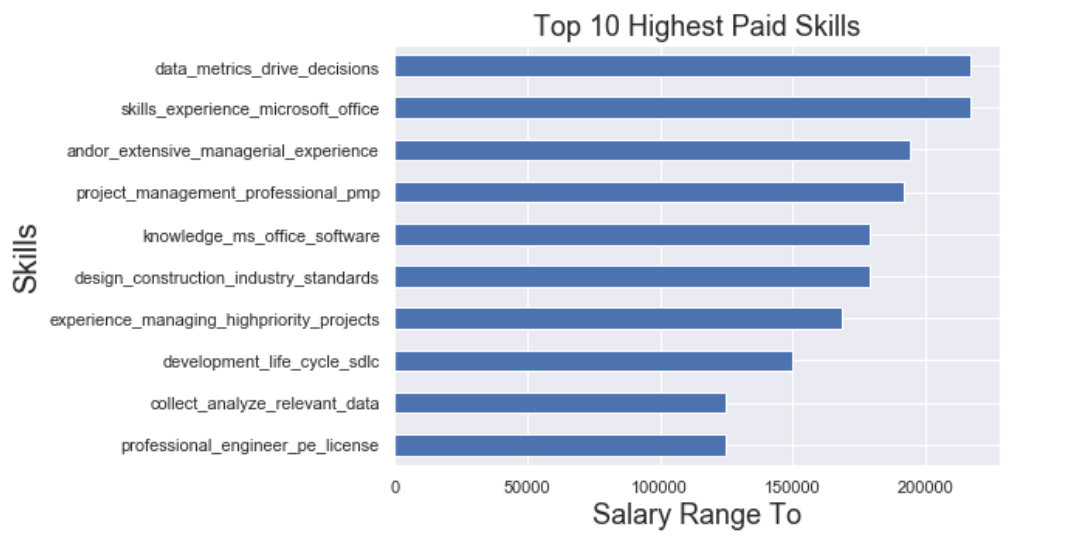
1. Bag of words
2. Bigram and N-Gram phrase detection
3. Pattern recognition
4. Data Aggregation with min and max functions



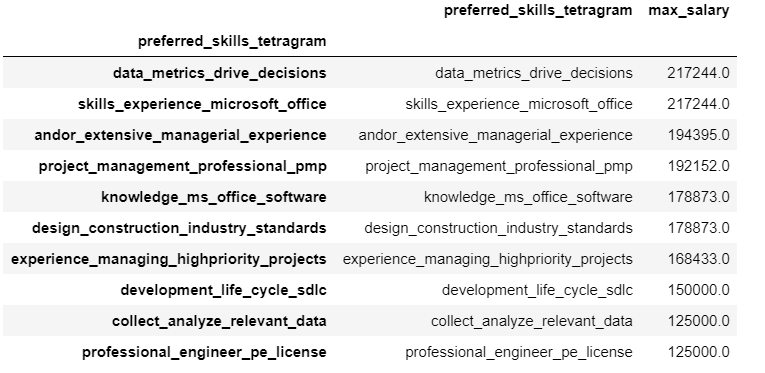
# 6 | Analysis

## 6.a Highest Paid Skills in US Market

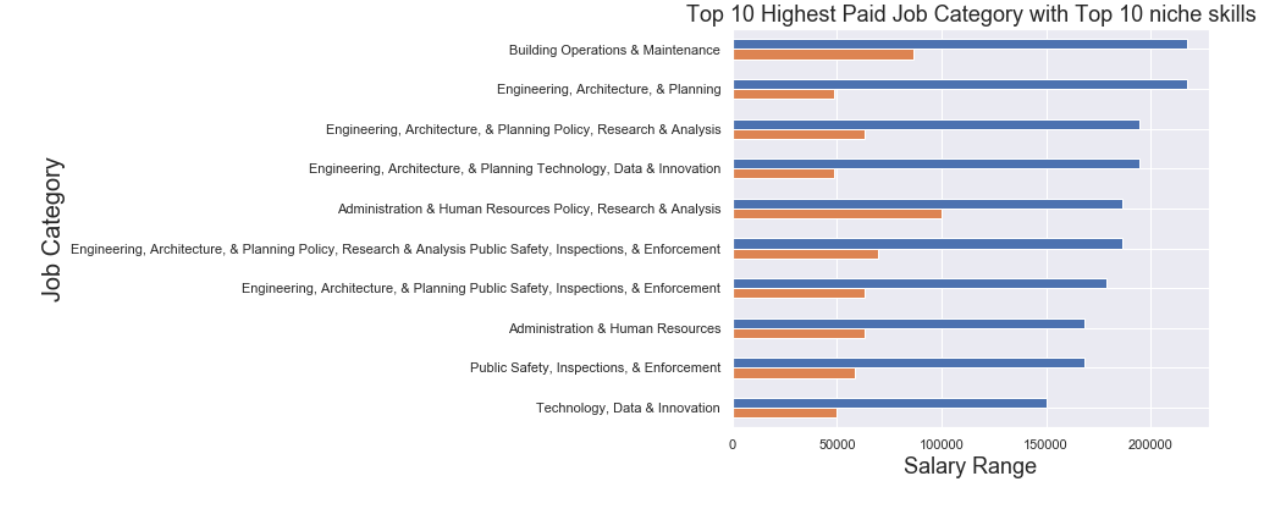
The skillset information was present in the attribute ‘Preferred Skills’. The text corpus had various combination of the string phrase to denote the skills. The analysis employed text parsing, and extraction techniques to identify and sort them out from the unstructured data. The below horizontal bar chart depicts the Top 10 highest paid skills in the US market against the salary range.



## 6.b Job categories involve the mentioned Niche Skills

The second question in the analysis is to look for the top 10 job categories with a niche skills in the US Market. The niche skills were identified from the extracted preferred skills and sorted by the calculated salary range.

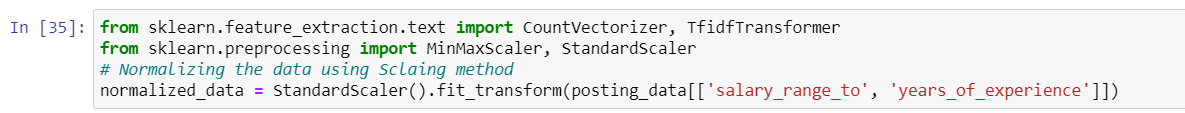
The column `max\_salary` denotes the values of normalized salary from the given range of salary from and salary to.

The graph below depicts the job category wise highest paid skills in the US market. The skills are niches skills which are rare in market but high in demand.

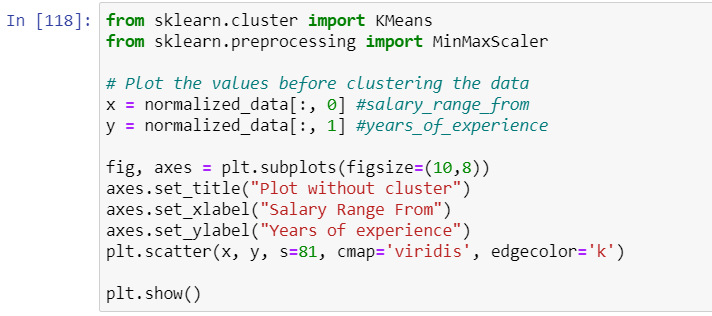
## 6.c Clustering on the data

Clustering technique is usually applied on the unstructured data to see the affinity of the data between each datapoints and their arrangements in the given dimensional space. Here, the clustering techniques were applied to see the salary and years of experience data points grouping based on the job category. In the clustering technique, first and foremost step is to standardize the feature values. Standardization is useful for data which has negative values. It arranges the data in normal distribution.

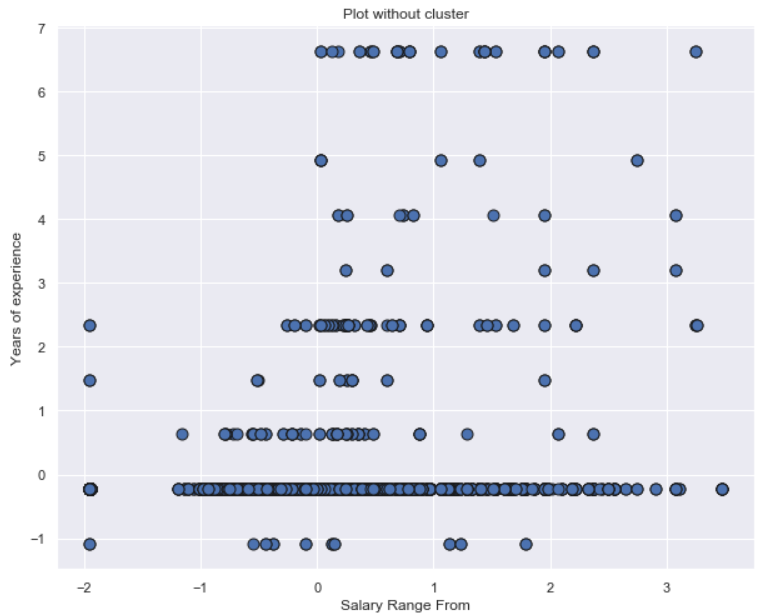
### Normalize the data

The StandardScalar class object is used to fit and transform the feature values into a normalized form.

### Building Cluster Model

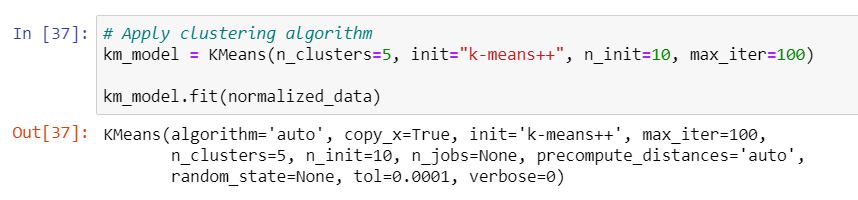
K-Means cluster algorithm was used to build the cluster model. The normalized data was used to fit into the k-means algorithm.

K-Means is a centroid based clustering algorithm, which forms a cluster based on the calculated centroids.



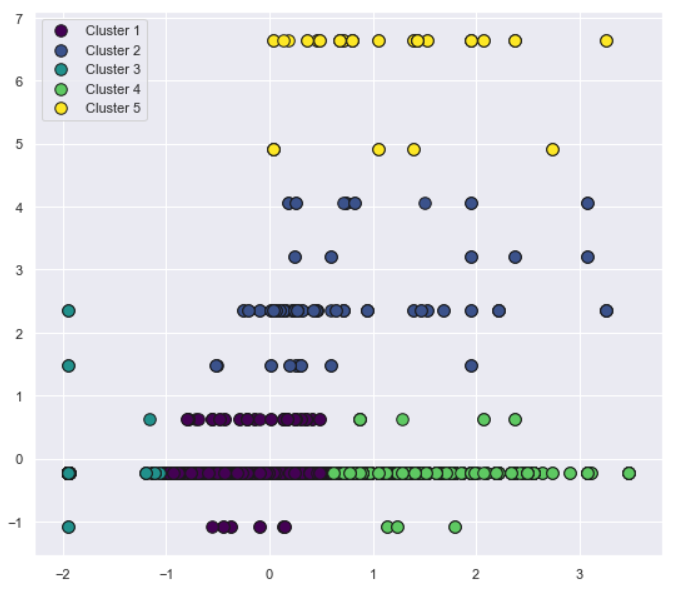
The scaled data of years of experience and salary range are plotted in a scatter graph.

The normalized data are fitted into the K-Means model object with basic set of parameters.



Once the cluster is fitted with a data, the labels of the cluster were taken from the properties of the model object and used to plot the data with different color code for each cluster.

The following plot depicts the data points of years of experience and salary range in 2-D space. The data points based on the clusters were colored appropriately with different color codes.



--------------------------------------------------------------------------------------------------------------------------