What can we Say and Forecast about the Contemporary Data Science Industry?

Peeyush Kumar
Boston University
peeyush@bu.edu

ABSTRACT

This Project aims to implement the latest Data Science approaches to extract knowledge from the latest survey conducted by Kaggle at international level, where people from every industry participated. The main goal of the project is to extract knowledge and analyze trends about the present data science industry. The complete process of data analysis is followed in this project, from data cleaning to prediction. Using the survey and data science tools, some of the main questions about today's industry were answered with proof. It was also identified through clustering that there are small subgroups within the industry, in which everyone has similar characteristics. At last, it was also identified that there is a pattern in the industry, which can be captured through machine learning algorithms, thus can be used to predict future trends in the industry.

1. INTRODUCTION

Data science is a branch which extracts useful information from noisy data, structured and unstructured, by means of scientific methods, algorithms, and processes. Forecasting refers to the task of estimating future events on the basis of past and present events. The whole field encompasses preparing data for analysis, formulating data science problems, analyzing data, developing data-driven solutions, and presenting findings to inform high-level decisions in a broad range of application domains. As such, it incorporates skills from computer science, statistics, information science, mathematics, information visualization, data integration, graphic design, complex systems, communication and business.

More and more companies are coming to realize the importance of data science, AI, and machine learning. Regardless of industry or size, organizations that wish to remain competitive in the age of big data need to efficiently develop and implement data science capabilities or risk being left behind.

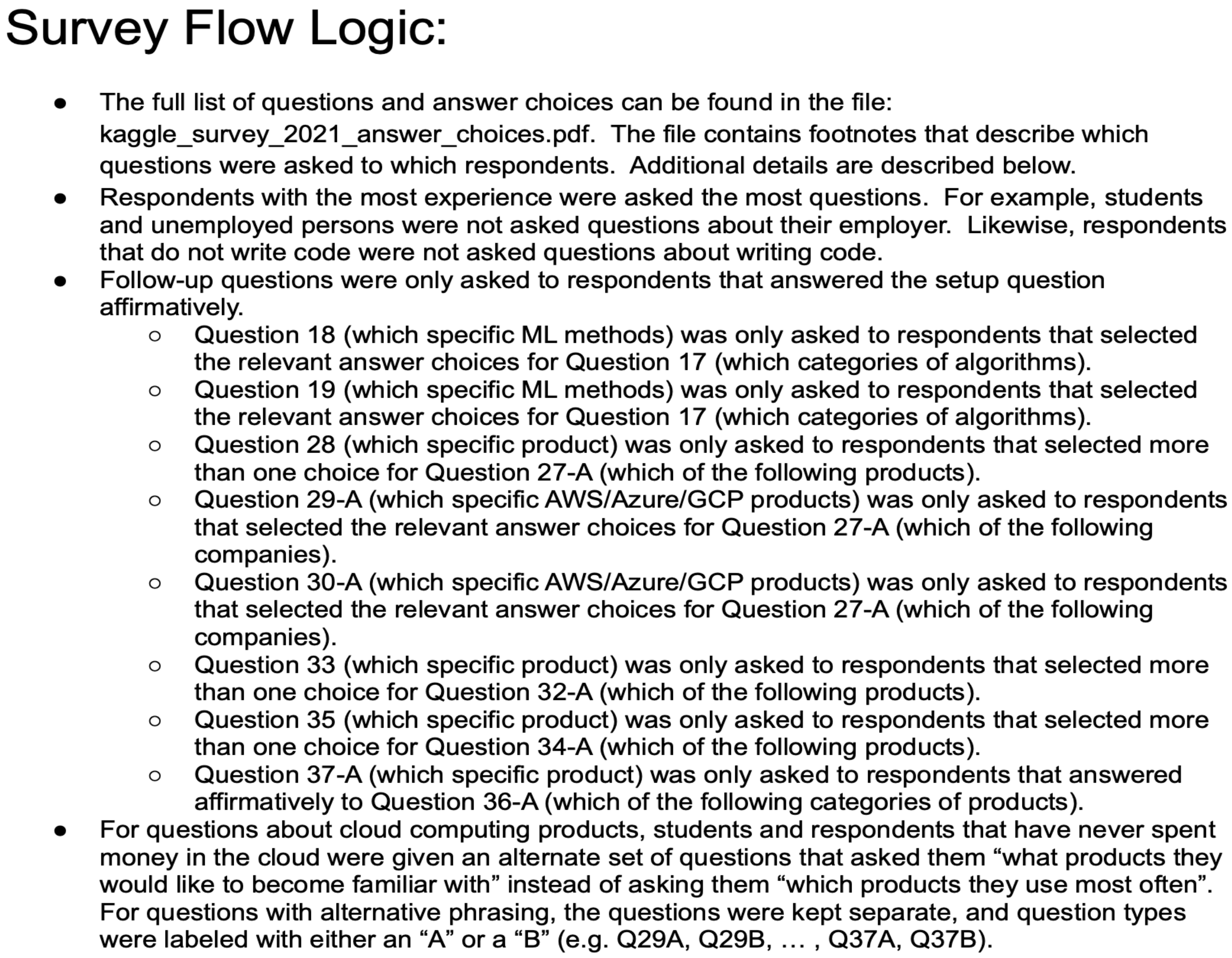
This project was done in order to not only understand and implement machine learning algorithms but also to get used to performing tasks like preparing data for analysis, constructing data science problems, analyzing data, developing data-driven solutions, and presenting findings to inform high-level decisions in a broad range of application domains.

2. APPROACH

This section describes the complete Data Science process followed in this entire project, along with the structure of the dataset used in this project.

` i. Dataset

The Dataset for this project is taken from an ongoing machine learning competition on Kaggle : "*2021 Kaggle Machine Learning & Data Science Survey*". The dataset is claimed to be one of the most challenging and comprehensive dataset available on the state of the Data Science Industry in today's world. The dataset consists of 25,000 Data Examples and over 320 Overlapping features. What makes the dataset most challenging is the fact that not everyone was asked the same set of questions. Everyone is asked questions on the basis of answers they have given for the previous question. As the data is newly collected and is made available to us in its rawest form, thus using proper data analysis tools we can derive thousands of industry's insights from it, which can help to answer some of the most important questions asked by people in today's world.



Survey Flow Logic:

- The full list of questions and answer choices can be found in the file: kaggle_survey_2021_answer_choices.pdf. The file contains footnotes that describe which questions were asked to which respondents. Additional details are described below.
- Respondents with the most experience were asked the most questions. For example, students and unemployed persons were not asked questions about their employer. Likewise, respondents that do not write code were not asked questions about writing code.
- Follow-up questions were only asked to respondents that answered the setup question affirmatively.
  - Question 18 (which specific ML methods) was only asked to respondents that selected the relevant answer choices for Question 17 (which categories of algorithms).
  - Question 19 (which specific ML methods) was only asked to respondents that selected the relevant answer choices for Question 17 (which categories of algorithms).
  - Question 28 (which specific product) was only asked to respondents that selected more than one choice for Question 27-A (which of the following products).
  - Question 29-A (which specific AWS/Azure/GCP products) was only asked to respondents that selected the relevant answer choices for Question 27-A (which of the following companies).
  - Question 30-A (which specific AWS/Azure/GCP products) was only asked to respondents that selected the relevant answer choices for Question 27-A (which of the following companies).
  - Question 33 (which specific product) was only asked to respondents that selected more than one choice for Question 32-A (which of the following products).
  - Question 35 (which specific product) was only asked to respondents that selected more than one choice for Question 34-A (which of the following products).
  - Question 37-A (which specific product) was only asked to respondents that answered affirmatively to Question 36-A (which of the following categories of products).
- For questions about cloud computing products, students and respondents that have never spent money in the cloud were given an alternate set of questions that asked them "what products they would like to become familiar with" instead of asking them "which products they use most often". For questions with alternative phrasing, the questions were kept separate, and question types were labeled with either an "A" or a "B" (e.g. Q29A, Q29B, … , Q37A, Q37B).

Figure 1. This figure shows the complete Survey flow logic followed during data collection.

Given the fact that not everyone was asked the same set of questions, the complexity of the dataset increased significantly. The Flow Logic followed during the survey is given in Figure 1. The figure lucidly describes how the questions varied from people to people. Due to following this specific flow logic, a meaning-full data representation technique is required before we can apply any algorithm over it.

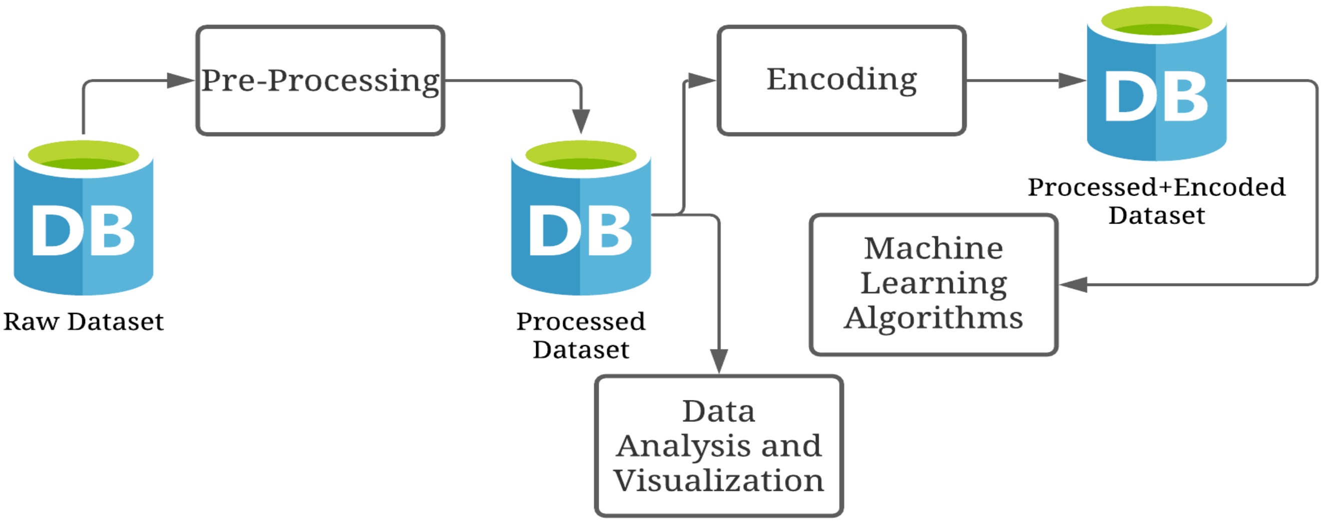

Raw Dataset
DB
Pre-Processing
DB
Processed Dataset
Encoding
DB
Processed+Encoded Dataset
Machine Learning Algorithms
Data Analysis and Visualization

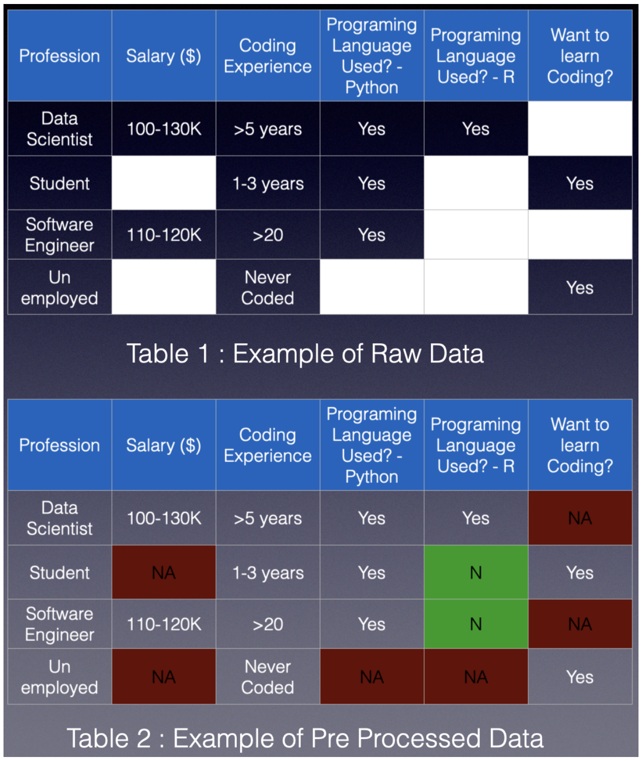
Figure 2. This figure shows the flow diagram of the process followed during the project,

ii. Process Followed

The process followed in the project mainly consists of 4 main parts. First one is Data Pre-Processing, in which we feed raw data and a meaning-full processed form is obtained as the output. Second one is Data Analysis and Visualization, in which the pre-processed data is used for analysis purposes. Third is Data Encoding, this step is done prior to applying machine learning algorithms. And at last, is the fourth step we apply machine learning algorithms like, clustering and decision tree on the encoded data.

a. Pre-Processing

The data preprocessing is one of the most challenging aspects of this project, and what makes it most challenging are the facts that not every user is asked the same set of questions, majority of questions were multiple answer questions and there were few questions that were not asked to certain users. So now the question is how do we preprocess the data such that the



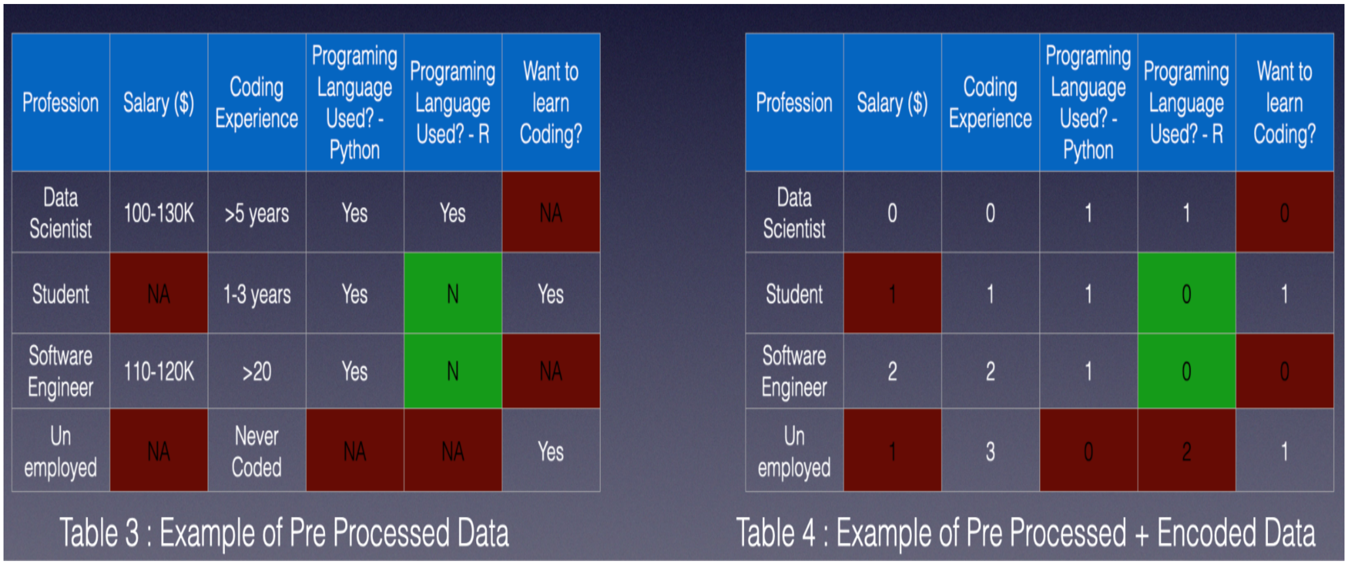
Profession	Salary ($)	Coding Experience	Programing Language Used? - Python	Programing Language Used? - R	Want to learn Coding?
Data Scientist	100-130K	>5 years	Yes	Yes	
Student		1-3 years	Yes		Yes
Software Engineer	110-120K	>20	Yes		
Un employed		Never Coded			Yes

Table 1 : Example of Raw Data

Profession	Salary ($)	Coding Experience	Programing Language Used? - Python	Programing Language Used? - R	Want to learn Coding?
Data Scientist	100-130K	>5 years	Yes	Yes	NA
Student	NA	1-3 years	Yes	N	Yes
Software Engineer	110-120K	>20	Yes	N	NA
Un employed	NA	Never Coded	NA	NA	Yes

Table 2 : Example of Pre Processed Data

Figure 3. Pre and Post Scenarios of Data Preprocessing.



Profession	Salary ($)	Coding Experience	Programing Language Used? - Python	Programing Language Used? - R	Want to learn Coding?
Data Scientist	100-130K	>5 years	Yes	Yes	NA
Student	NA	1-3 years	Yes	N	Yes
Software Engineer	110-120K	>20	Yes	N	NA
Un employed	NA	Never Coded	NA	NA	Yes

Table 3 : Example of Pre Processed Data

Profession	Salary ($)	Coding Experience	Programing Language Used? - Python	Programing Language Used? - R	Want to learn Coding?
Data Scientist	0	0	1	1	0
Student	1	1	1	0	1
Software Engineer	2	2	1	0	0
Un employed	1	3	0	2	1

Table 4 : Example of Pre Processed + Encoded Data

Figure 4. Pre and Post Scenarios of Encoding the Pre-processed Data

processed form gives the algorithms a sense that not everyone is asked the same set of questions.

As you can see in Figure 3, the table 1 one represents the raw form of data available to us. The white spaces represented the null values in the data. We can just fill all the null values with the same entity but that will not give the algorithms the sense that not everyone was asked the same set of questions. So the solution is to use 'N' : Not chosen. 'NA' : Not Applicable or Not Asked, as shown in table 2 of Figure 3.

b. Encoding

Now, before apply machine learning algorithms to the dataset we first have to encode the categorical data available to us into numerical form. For this, Categorical Encoding was used. In Figure 4, table 3 is the preprocessed form of data and table 4 represents the encoded form the same

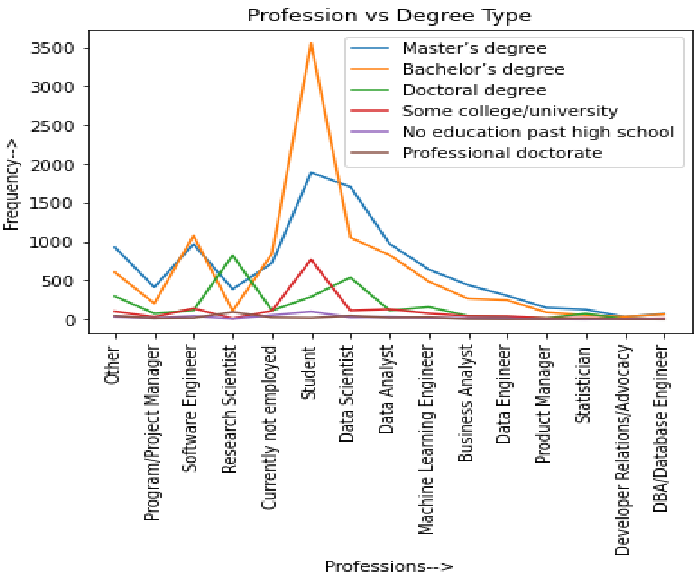


Figure 5. Answer for question : *What kind of degree is mostly needed for Professions in the Data Science Industry? Ans : Master's Degree is a must now.*

3. CONCLUSION USING DATA ANALYSIS

Through Data Visualization, some important questions regarding the industry were answered. All the questions were in correlation to Salary and Profession in the Data Science Industry. The questions which were answered are as follows :

1. *What kind of degree is mostly needed for Professions in the Data Science Industry?*
2. *Which profession is highly paid in the Data Science Industry?*
3. *Which Programming Language is Mostly used in Different Professions?*
4. *How does salary vary with an education Degree in this Industry?*
5. *What kind of coding Experience does Highly Paid Professionals of the Data Science Industry have?*
6. *What is the Correlation of Gender with Profession and Salary?*
7. *Do you need to work In Big- Company to get Paid more?*

The answers to all the questions were answered through visualization of graphs given in Figure 5 to 12.

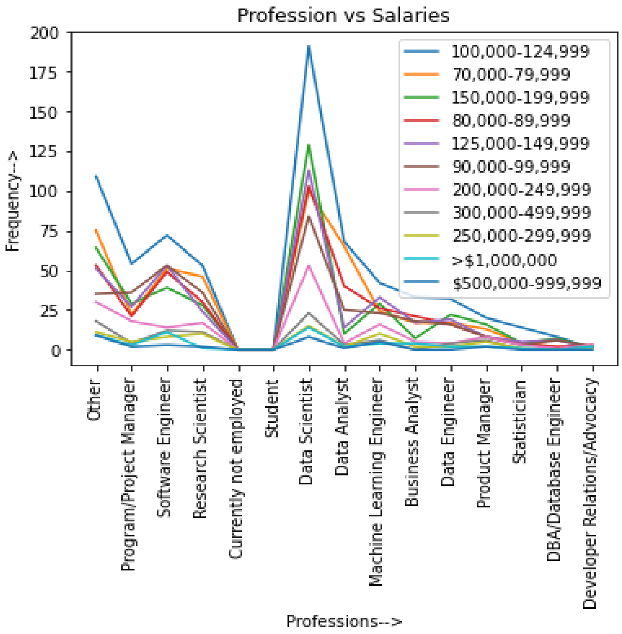


Figure 6. Answer for question : *Which profession is highly paid in the Data Science Industry? Ans : Data Scientist.*

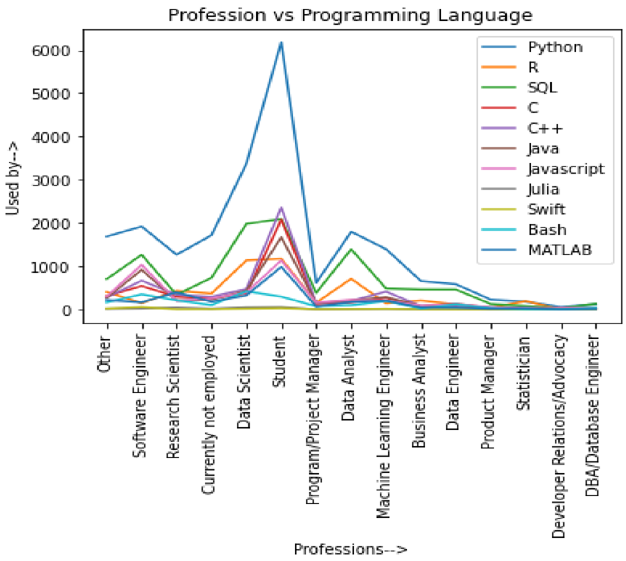


Figure 7. Answer for question : *Which Programming Language is Mostly used in Different Professions? Ans : Python.*

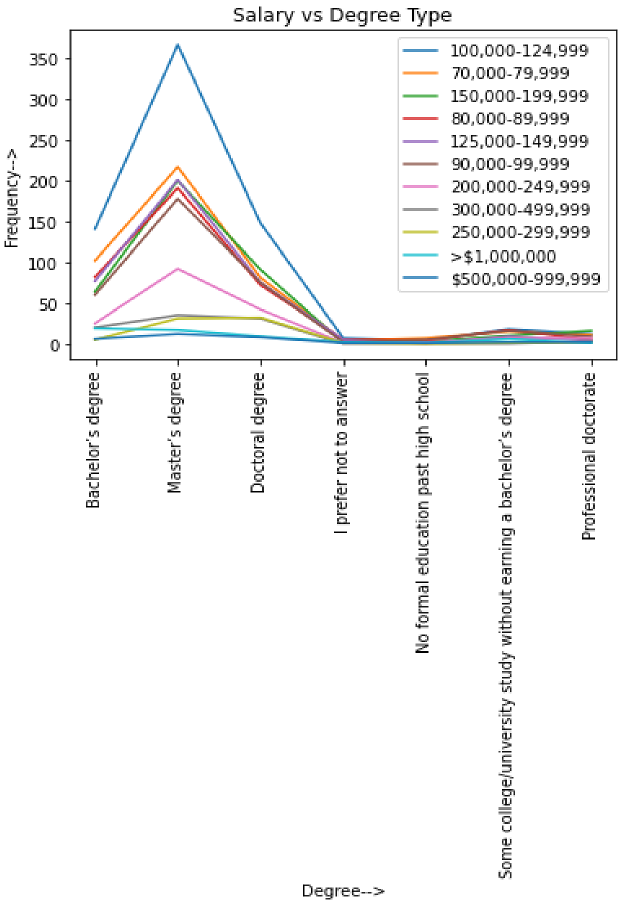


Figure 8. Answer for question : *How does salary vary with an education Degree in this Industry? Ans : Master's degree makes the most difference..*

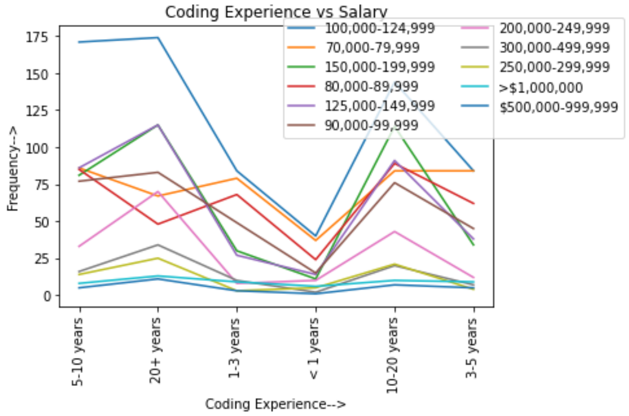


Figure 9. Answer for question : *What kind of coding Experience does Highly Paid Professionals of the Data Science Industry have? Ans: Coding Experience Doesn't matter! Quality does.*

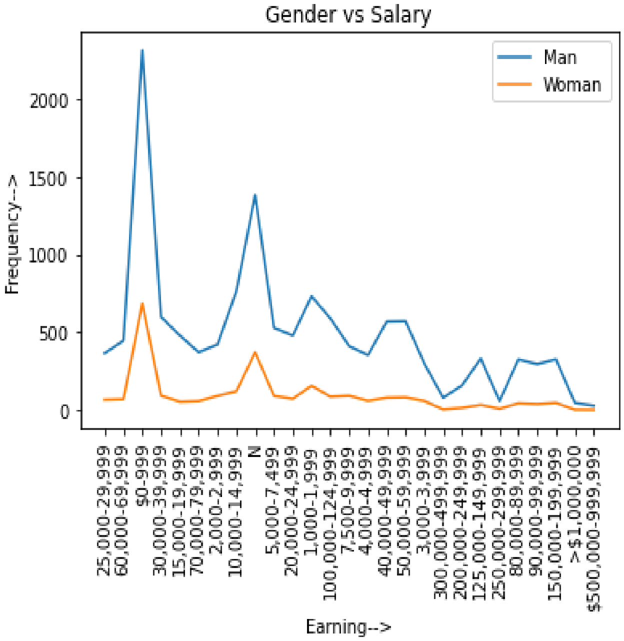


Figure 10. Answer for question : *What is the Correlation of Gender with Salary? Ans: Large numbers of males have higher salaries..*

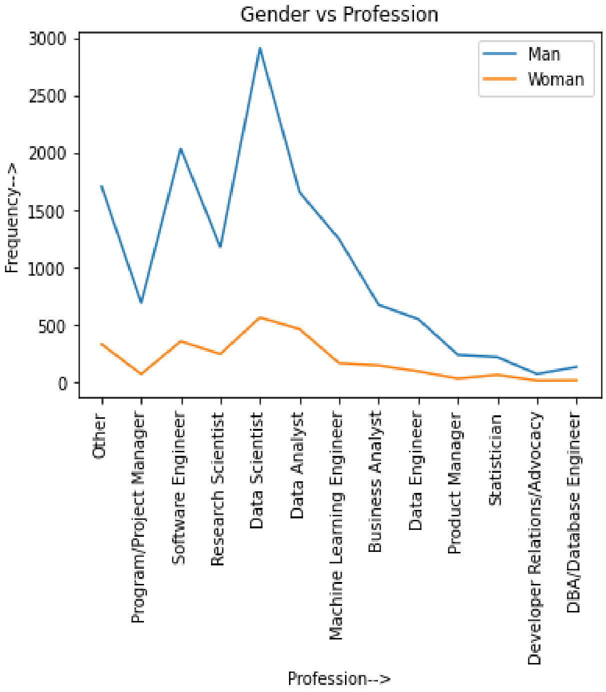


Figure 11. Answer for question : *What is the Correlation of Gender with Profession? Ans: The ratio of male is much higher in each profession..*

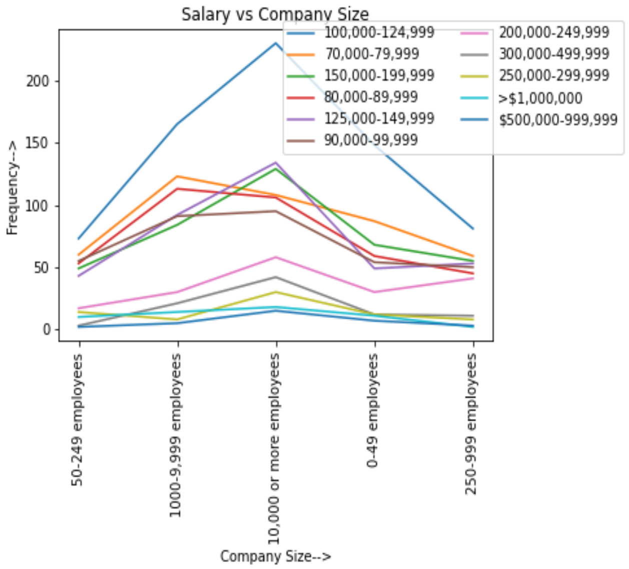


Figure 12. Answer for question : *Do you need to work In Big- Company to get Paid more? Ans: The size of the company doesn't matter.*

4.CONCLUSION via MACHINE LEARNING

After Encoding the pre-processed data, both unsupervised and supervised machine learning algorithms were applied to the data.

In Unsupervised Learning, the main clustering algorithms like KMeans, DBSCAN, BIRCH and OPTICS were applied. The figure 13 shows the result obtained after applying these clustering algorithms. The plots are plotted via applying Incremental-PCA algorithms to reduce the dimensionality of the data.

Through K-Means it was determined that there are clusters in the datasets and which indicates that

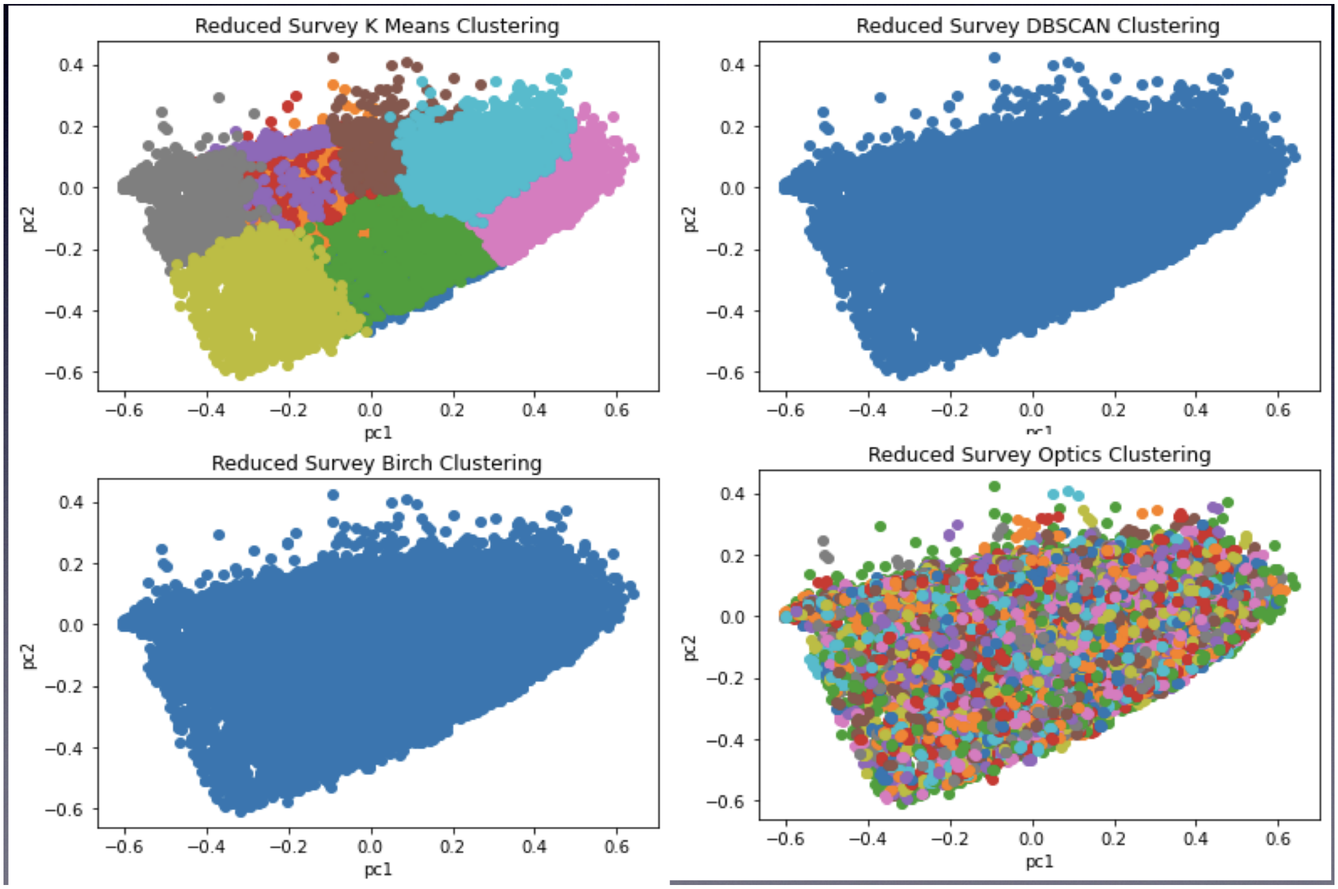


Figure 13. The output plot for different clustering algorithms.

there are specific groups in the industry that share the same characteristics. The results of DBSCAN and BIRCH were inconclusive, as the Algorithms were not able to detect and clusters.

In Supervised learning, Decision Tree was applied as it is very equipped for handling categorical data. Through supervised learning it was determined that there are patterns in the data which can be exploited to forecast future trends in the industry.

For measuring how one feature is dependent on other features a new metric was developed called *Dependency Value*. The Dependency Value for a Dataset D , with n examples and m features, for a given threshold θ is defined as :

$$\frac{|P(X_i|X_{(1->n)-\{i\}})|^{Accuracy\geq\theta}}{n}$$

As mentioned before, the decision tree was used to calculate dependency value. Figure 14. shows the dependency value vs Threshold Tradeoff.

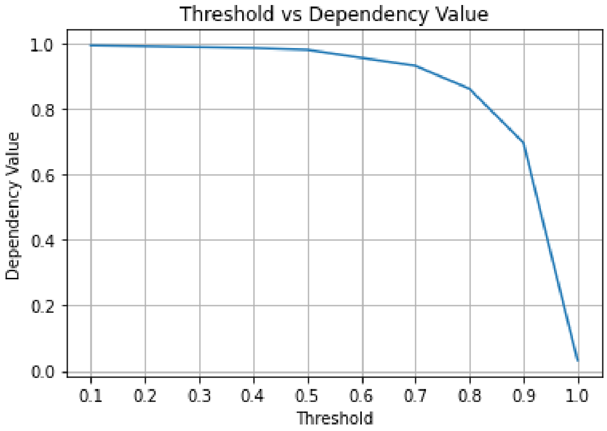


Figure 14. Graph showing the Tradeoff between The dependency value and Threshold