PROJECT SUBMISSION 1

Team Visualytics

GROUP 07

Submission 1

Part 1

- Proposal:
- -What is the problem you are solving?
- -Who benefits from this project?
- -Why is the project important?
- -What are the business questions you are looking to answer or objectives you are looking to achieve from this project?
- -Describe the data at a high level, explain the data collection process, source of data, etc

Submission 1

Part 2

- -Describe the data - Summary statistics, breakdown of variables (numeric, ordinal, categorical)
- -Check the validity of data - 1) Define schema; 2) Understand the data;
- -Answer the following questions:
- -Does the data include missing, incomplete or invalid records?
- -Does your data include outliers?
- -Is the data segmented into groups?
- -Is the data imbalanced (a large number of the records represent a majority class and very few records represent the minority class)?
- -Are some data elements highly correlated with each other?
- -How was the data collected?
- -What are the inclusion criteria for your data?
- -Can you generate preliminary visualizations for individual features?

Submission 1 – (Part 1)

(Proposal)

- What is the problem you are solving?

We are trying to find answers to some of the questions which are imperative for both, individual and also the IT organizations. If an individual wants to find a career in an IT field, he/she is usually struck by so many common, yet not easily available questions. For example, what is the lifestyle of a programmer, do I need to devote 10-12 hours in a day to be a part of this competitive industry? Which country has the biggest organizations? How much of expected pay can I think of?

While, on the other hand, if an organization wants to boost the productivity, wants to see what are the latest market trends on technologies, on an average how long does an employee work in a specific organization etc., they can also use this analysis and find the answers to a lot of common questions with some real responses from the survey.

- Who benefits from this project?

Both the organizations and individuals can benefit out of this project on various different aspects of IT field job market as outlined in the point above.

- Why is the project important?

This project is important as usually there are no reliable source to see the vast variety of commonly asked questions. Instead of going on blogs etc, spending hours of valuable time and still not being sure, an individual can actually just take a glimpse of the analysis and easily get answers to a lot of questions.

From an organization's perspective, many start ups can get a reliable and UpToDate information on various questions that are important for an organization to run the operations smoothly and effectively.

So, to summarize, this analysis can actually act as a reliable, one stop shop for all the queries/ doubts and decision making points.

- What are the business questions you are looking to answer or objectives you are looking to achieve from this project?

Our objective is to help the organizations and individuals to establish a clear line of sight about various attributes of different IT roles, as mentioned below.

Financial aspect of Various Analyst roles across multiple countries. For example, if an individual wants to see what is the expected salary amount for a specific role in IT industry, in a specific country, he can easily see the average results from the analysis to give him a better understanding of the data

An individual can also see if ***Job satisfaction is related to size of an organization***. The survey data provides details on how many employees are satisfied with their work basis the size of their respective organization (like 100 to 499 employees or 10,000 or more employees)

On the other hand, organizations can find the ***market trends on attrition rates***. For instance, if a start-up wants to hire employees, planning an expected attrition rate will be an integral part of the hiring process. Using this dataset, they can easily find out the average attrition rate during the recent time.

Another important aspect is the ***relationship between salary packages and education level***, where we can see if getting formal education is compulsory to get high salary package

- Describe the data at a high level, explain data collection process, source of data, etc

The dataset used is the survey result that was conducted amongst number of individuals to get information about IT/Data/Analyst related roles. It also contains various important points of information about IT fields related jobs, technologies preferred, education level, lifestyle etc.

This data was collected in a survey conducted by Stack Overflow across various countries.

Source of the data is as below:

https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey?select=survey_results_public.csv

File Name: survey_results_public.csv

Submission 1 – (Part 2)

Descriptive Statistics and Exploratory Data Analysis (EDA)

- Defining Schema and understanding the data

Schema:

A schema is a logical representation of the data which helps us interpret the data. Schemas are useful because they enable us to understand the large amount of data. They are often used to define the structure of various sorts of data.

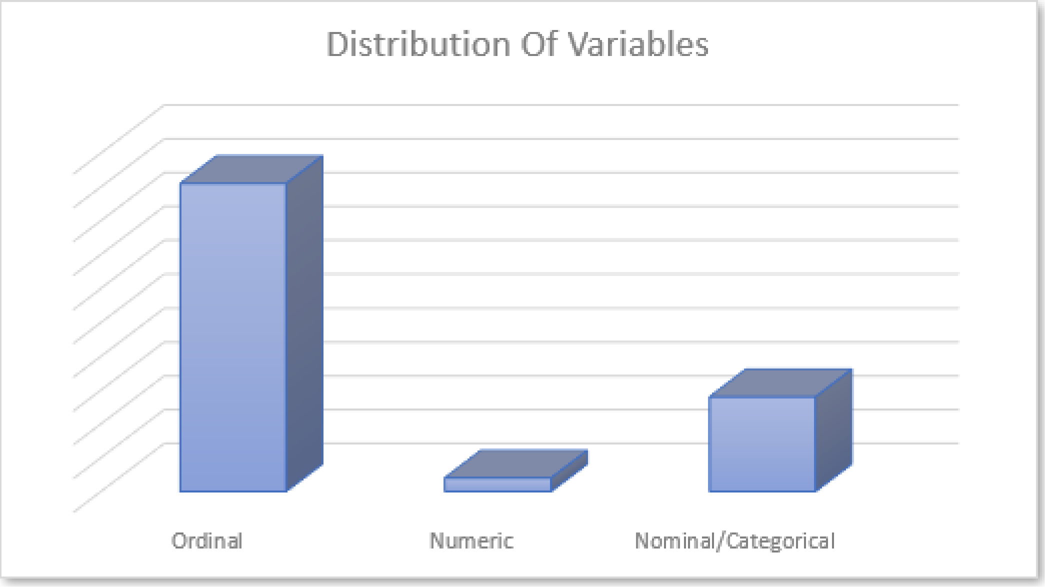
Understanding the data:

The data involves two files namely survey results public and survey results schema. First file represents the main result in which one respondent per row and one column per question. However, second file (survey results schema) constitutes the name of the columns along with the questions text corresponding to that column.

Summary: The data has 98856 rows and 129 columns.

- Describe the data - Summary statistics, breakdown of variables (numeric, ordinal, categorical)

The below graph depicts the distribution of the data. Since the data is related to surveys hence it is majorly dominated by Ordinal variables.

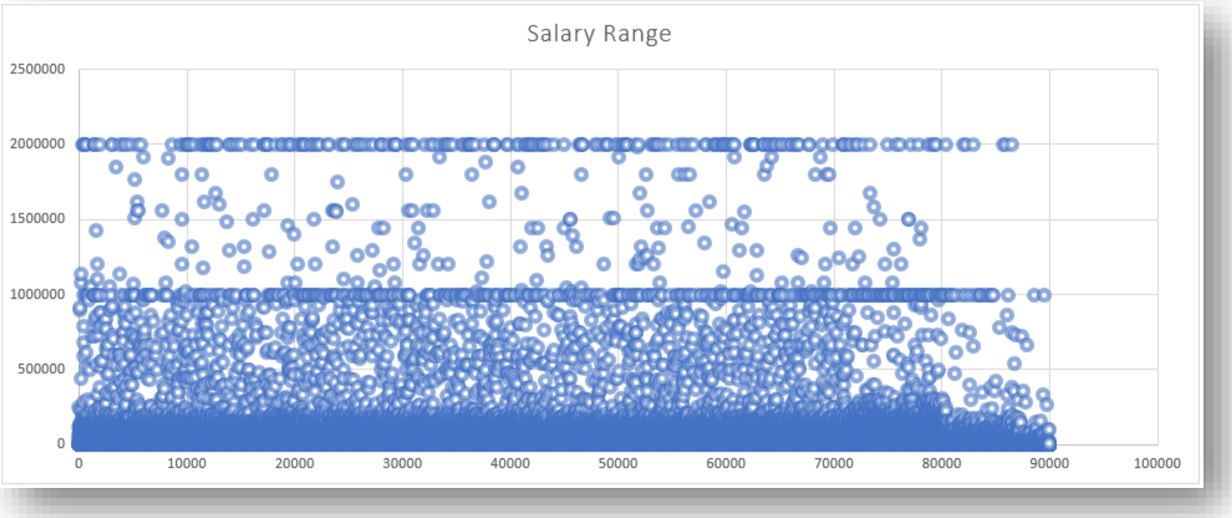


- Does the data include missing, incomplete or invalid records?

Yes, there are missing/Incomplete records in the data in the form of NA Values and in some cases, records with special characters are also included. 18 out of 130 columns have more then 50% NA value within themselves.

- Does your data include outliers?

No, the data doesn't include outliers. For instance, if we plot a scatterplot for Salaries (Converted in USD), we don't see any outlier.

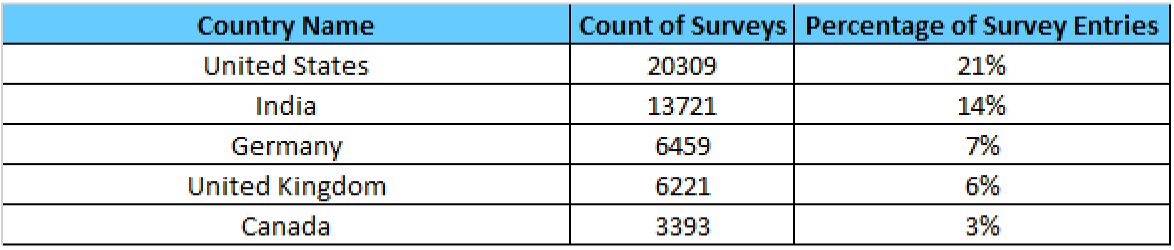


- ## Is the data segmented into groups?

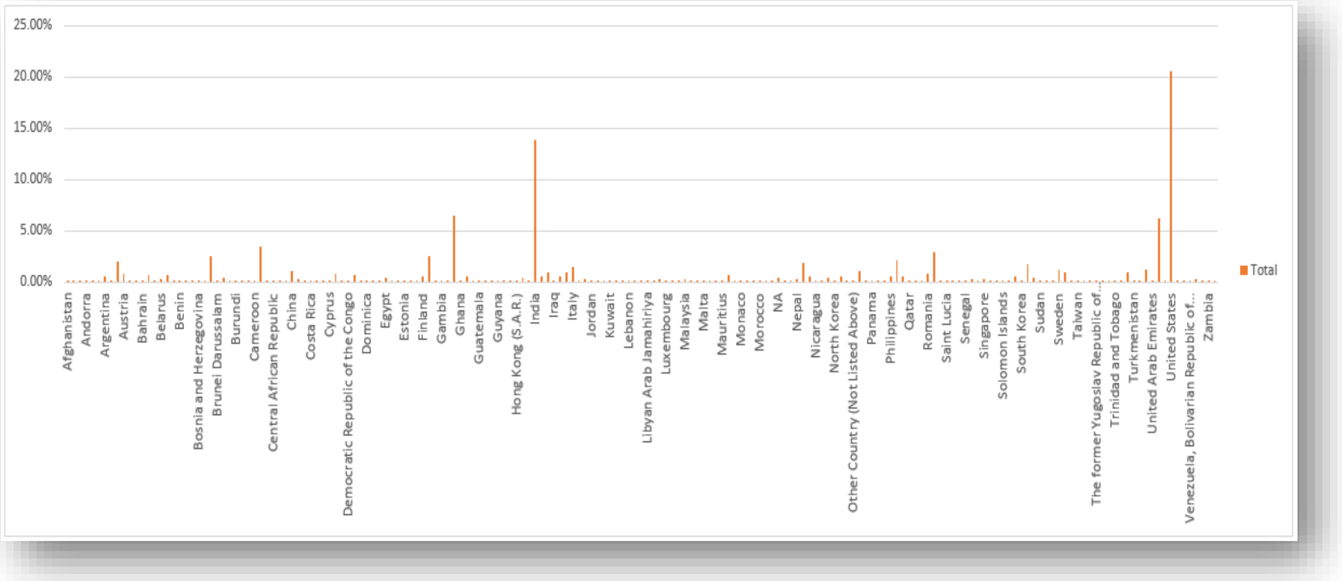
No, our data is not segmented into groups. Also, we will be analysing the overall view of the data in terms of individuals and not basis certain groups like male/female etc..

- ## Is the data imbalanced (a large number of the records represent a majority class and very few records represent the minority class)?

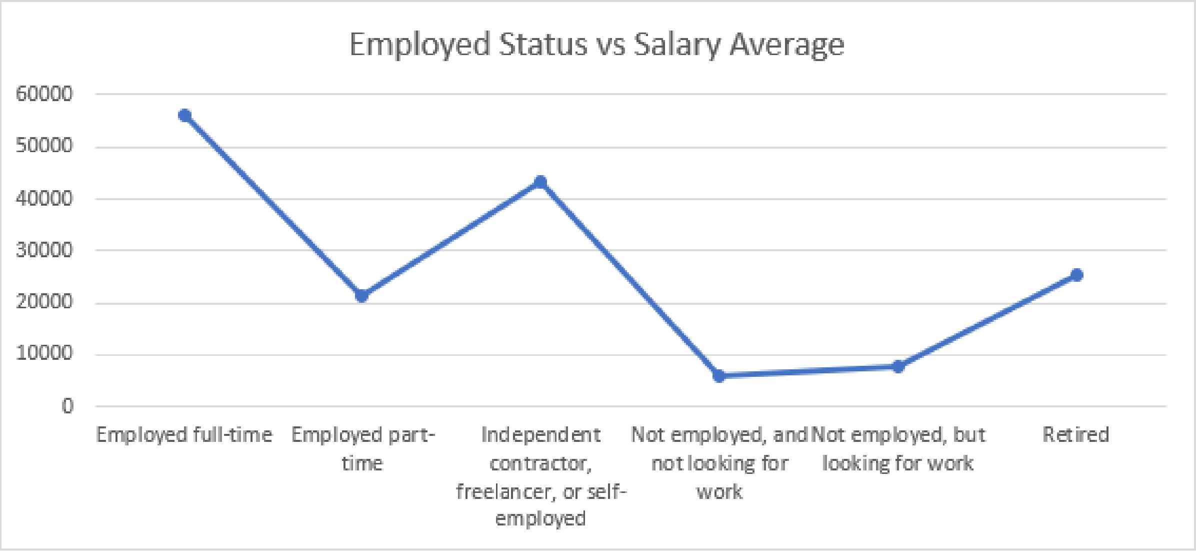
Yes, the data seems to be imbalanced where more than 50% of the records presented are just from 5 countries only. Please see the details below.



Country Name	Count of Surveys	Percentage of Survey Entries
United States	20309	21%
India	13721	14%
Germany	6459	7%
United Kingdom	6221	6%
Canada	3393	3%



- Are some data elements highly correlated with each other?



- How was the data collected?

- The dataset used is the survey result that was conducted amongst number of individuals to get information about IT/Data/Analyst related roles. It also contains various important points of information about IT fields related jobs, technologies preferred, education level, lifestyle etc.
- This data was collected in a survey conducted by Stack Overflow across various countries.

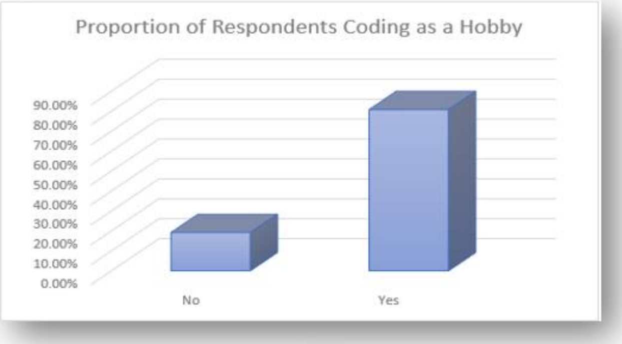
- Source of the data is as below:
- https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey?select=survey_results_public.csv
- File Name: survey_results_public.csv

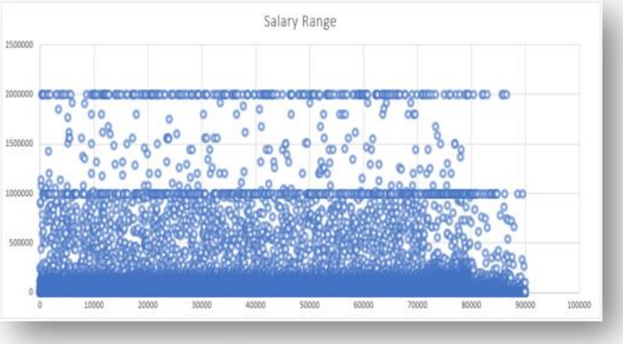
- What are the inclusion criteria for your data?

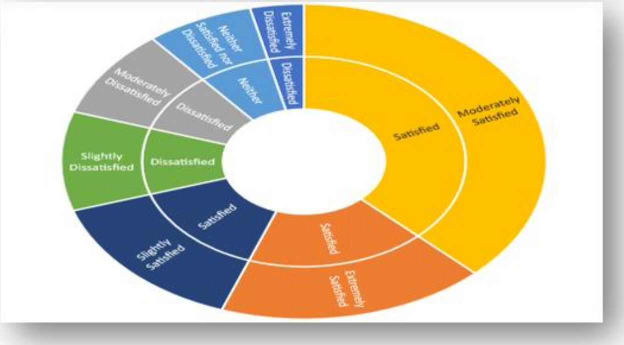
The inclusion criteria for our dataset will the class the future international students, job seekers in IT field, established organizations and start up organizations.

- Can you generate preliminary visualizations for individual features?

Some of the preliminary visualizations are as below:


Proportion of Respondents Coding as a Hobby
No
Yes


Salary Range


Extremely Dissatisfied
Neither Satisfied nor Dissatisfied
Moderately Dissatisfied
Slightly Dissatisfied
Slightly Satisfied
Extremely Satisfied
Moderately Satisfied
Dissatisfied
Neither
Dissatisfied
Dissatisfied
Satisfied
Satisfied
Satisfied