**Capstone Project**

**Data Scientist Jobs in India**

**Context:**

I am developing my skills into data science. Since my goal is to move to India, I would like to know everything about data scientist jobs in India. This capstone project will serve to my personal preparation into planning in advance for a data scientist role.

**Client:**

I am the client and I do care because I need to understand that where in India will give me more opportunities regarding a data scientist role. This will help me in making the right decision and take calculated risks.

**Problem**

When people want to move from overseas back to India, they do not do proper research into the job scene in regards with the role they are seeking a career in. People do not do enough research into, which companies are hiring for a data scientist position; what locations have the highest density of data scientists; what is the level of position the company is hiring at; what companies are hiring; what cities will likely hire in the future etc. When the above questions are not researched properly, immigrants are usually stuck at a locations that does not hire for the respective roles, or do not needs senior data scientists.

**Objectives:**

1. Develop in-depth analysis of:
   1. What cities are hiring data scientist jobs?
   2. Which states are hiring for data scientist roles?
   3. Which are the top 5 companies hiring for data scientist role?
   4. Which companies require all the three machine learning, python, sql in data scientist roles?
   5. Text analysis of which skills in the data analysis jobs are most sought after?

**Datasets:**

* The data set is acquired by getting an open API from indeed (job search website), searching for data scientist jobs from all over India. The data set has 10 variables and has over 1200 search results.

**Outline:**

Reading the XML data in R studio:

The first step is make sure that XML data format is read properly in R studio. For that I had to first create the file in R. Then the XMl FILE is parsed.

Converting the XML to dataframe:

After the data is read in the R studio. It is important to convert it in the data frame for analysis. Due to the HTML tagging, separate nodes were extracted using “getnodes”. After the nodes were extracted, they were combined in a single file.

Further data wrangling is required to analyze various variables.

Data Modeling

Data visualization

**Deliverables:**

Code

Paper/slide deck