# Title: San Francisco Employee Data

|  |  |  |
| --- | --- | --- |
| First name | Last Name | IIT Email |
| Utkarsha | Vidhale | [uvidhale@hawk.iit.edu](mailto:uvidhale@hawk.iit.edu) |
| Vaidehi | Rathkanthiwar | vrathkanthiwar@hawk.iit.edu |
| Prashant | Raina | [praina@hawk.iit.edu](mailto:praina@hawk.iit.edu) |

**Type of Your Projects (select one or more):** Multiple linear regression

1. **Introduction**

Our application is based on San Francisco Employee compensation data which describes the various features related to employee department, organization, job profile, salary and benefits.

In a corporate structure, employees are the integral part of the organization. No matter your company size, your people are your most important asset. They are the backbone of your business. So, one of the most important aspects of running your business is keeping your employees happy by offering them high-quality employee benefits and compensation.

Employee benefits refer to all non-wage compensation or bonus provided to employees in addition to their salaries. The type of benefits your company decides to offer will vary based on the organization and job profile.

But, sometimes many companies don’t realize how much time and money ineffective HR processes are costing them. Providing benefits to those type of job profile which have a very low productivity has often come into wrong consideration. So, there must be some solution in which company can know in advance about the compensation structure based on job profile and organization. This provided us the opportunity to develop model which can predict compensation and benefits based on different factors. Employers can use this model to imbibe some knowledge regarding the compensation factors and employees can use it to decide which job profiles are receiving maximum benefits

1. **Data Sets**

Briefly introduce your data sets, such as which application or domain the data belongs to, where did you collect it, how large it is, how many features there are, and so forth.

* The dataset hosted by the city of San Francisco. The organization has an open data platform and they update their information according the amount of data that is brought in. The San Francisco Controller's Office maintains a database of the salary and benefits paid to City employees since fiscal year 2013.
* This dataset is updated annually. New data is added on a bi-annual basis when available for each fiscal and calendar year. It has been collected from kaggle.com (<https://www.kaggle.com/san-francisco/sf-employee-compensation>) and is available in csv format (170 MB). There are 8,35,308 instances(records) and 22 attributes(columns) in the dataset. Out of 22 attributes, 13 are numerical variables and 9 are categorical variables.

Following are the attributes in this dataset:

* + - Year Type: (Nominal/Categorical variable)
    - Year: (Numerical)
    - Organization Group Code: (Numerical)
    - Organization Group: (Nominal/Categorical variable)
    - Department Code: (Nominal/Categorical variable)
    - Department: (Nominal/Categorical variable)
    - Union Code: (Numerical)
    - Union: (Nominal/Categorical variable)
    - Job Family Code: (Nominal/Categorical variable)
    - Job Family: (Nominal/Categorical variable)
    - Job Code: (Nominal/Categorical variable)
    - Job (Nominal/Categorical variable)
    - Employee Identifier: (Numerical)
    - Salaries: (Numerical)
    - Overtime: (Numerical)
    - Other Salaries: (Numerical)
    - Total Salary: (Numerical)
    - Retirement: (Numerical)
    - Health/Dental: (Numerical)
    - Other Benefits: (Numerical)
    - Total Benefits: (Numerical)
    - Total Compensation: (Numerical)

1. **Research Problems**

List your research problems, that is, what kinds of the problems you want to solve.  
You cannot simply say I want to explore the data and find the patterns  
You should provide finer-grained research problems that can be solved by statistical techniques.

* Based on the dataset, the interested research problems are:

1. Predicting the total compensation of the employee based on various factors that will help the employers to decide what compensation should be given to employee in advance in order to keep tabs on their financial section.
2. Predicting the total salary of the employee based on benefits, compensation and job profile that will help the employees to aim for better job profiles based on high benefits.
3. Are the average salaries of all the employees same or different for various

organizations or job profiles?

1. **Potential Solutions**

For each problem you list above, figure out feasible solutions, and introduce your plan to perform experiments

* Feasible solutions:

1. We will use Multiple linear regression to predict the compensation and benefits given to the employee based on salary, organization and job profile.
2. We will use Multiple linear regression to predict the total salary given to the employee based on organization and job profile and other factors.
3. We will use ANOVA to compare average salaries of different employees based on job profiles and organization.
4. **Evaluations**

There could be multiple solutions for a same problem, You must figure out how to evaluate them and the details about your evaluations, for example, hold-out or N-folds evaluation? which metrics you will use for evaluations.

We will take following steps in evaluating our model:

1. Since our database contains 0.8 million records, we will use Hold-out evaluation for our model. We will split the data into 80-20 ratio i.e training set (80%) and test set (20%) and will start building the model using training data.
2. We will build different models and evaluate them using coefficient of determination on training dataset to check the accuracy of model. We will then apply the prediction methods on test dataset and evaluate the results by comparing all the models and choose the best out of them.
3. **Expected Outcomes**

Introduce your expected outcomes for your project

We will get predicted values for total compensation that will help the employers to decide how much compensation should be given to employee based on job profile and organization.

We will get the predicted values for total salary based on job profile that will help the employees to decide what job profiles to aim for in order to receive high benefits.

We will get to know those job profiles and organizations which provide higher compensation and have higher benefits.