**Group Project Data Summary Report**

1. **Description of the Problem**

A US corporation must apply to the US government to obtain a green card or visa for the foreign candidate when it wishes to hire someone from outside the country for a technical role. With the help of these programs, the US government can keep tabs on who is traveling for business purposes and who is leaving the country, ensuring that neither immigrants are being exploited nor hurting American employees. Companies are required to disclose the employee's anticipated wage each time they file a visa or green card application to ensure pay fairness for US and non-US workers. Additionally, they must disclose the normal salary for a worker with comparable qualifications.

1. **Description of the Data**

The original data was compiled by the US Department of Labor’s Office of Foreign Labor Certification. This data includes 167,278 observations for the case status of permanent resident applications from 2008 to 2015. There also have 27 columns, which represent 27 variables that have been statistically listed: Case number, Case Status, Case Received Date, Decision date, Employer Name, Prevailing Wage Submitted, Prevailing Wage Submitted, Unit, Paid Wage Submitted, Paid Wage Submitted Unit, Job Title, Education Level Required, College Major Required, Experience Required Y/N, Experience Required Num Months, Country of Citizenship, Prevailing Wage Soc Code, Prevailing Wage Soc Title, Work City, Work State, Work Postal Code, Full Time Position Y N, Visa Class, Prevailing Wage Per Year, Paid Wage Per Year and Job Title Subgroup.

1. **Supervised or unsupervised learning**

This data can be supervised learning, which is a regression problem when we can use the data to predict the salary based on the Job Title and the working place.

1. **Interesting factors**

* **Technique Fun fact: they are not integers or float as type but objects, so the main method doesn’t work we have to wash our data So fun!!!!**

**PAID\_WAGE\_SUBMITTED**

**PREVAILING\_WAGE\_SUBMITTED**

1. There are more than 12,000 job titles in this data, however, all the job titles can be separated into only 8 group, which is: Software Engineer, Assistant professor, Teacher, Business analyst, Management consultant, Data analyst, Attorney, and Data Scientist.
2. An attorney seems to be the job with the highest mean salary (146,000) while a Teacher is a job with the lowest mean salary (46,000)
3. Most “Software Engineer” and “Data Scientist” works in California (38.9% and 53%, respectively)
4. India has the greatest number of citizenships that submit the case number in this data, which is 59.4%. /// be careful: a big majority of NA for this column, so the statistic is not working
5. The decision time for certified case status seemed to be quick. When the time is longer, the case status seemed to be denied.