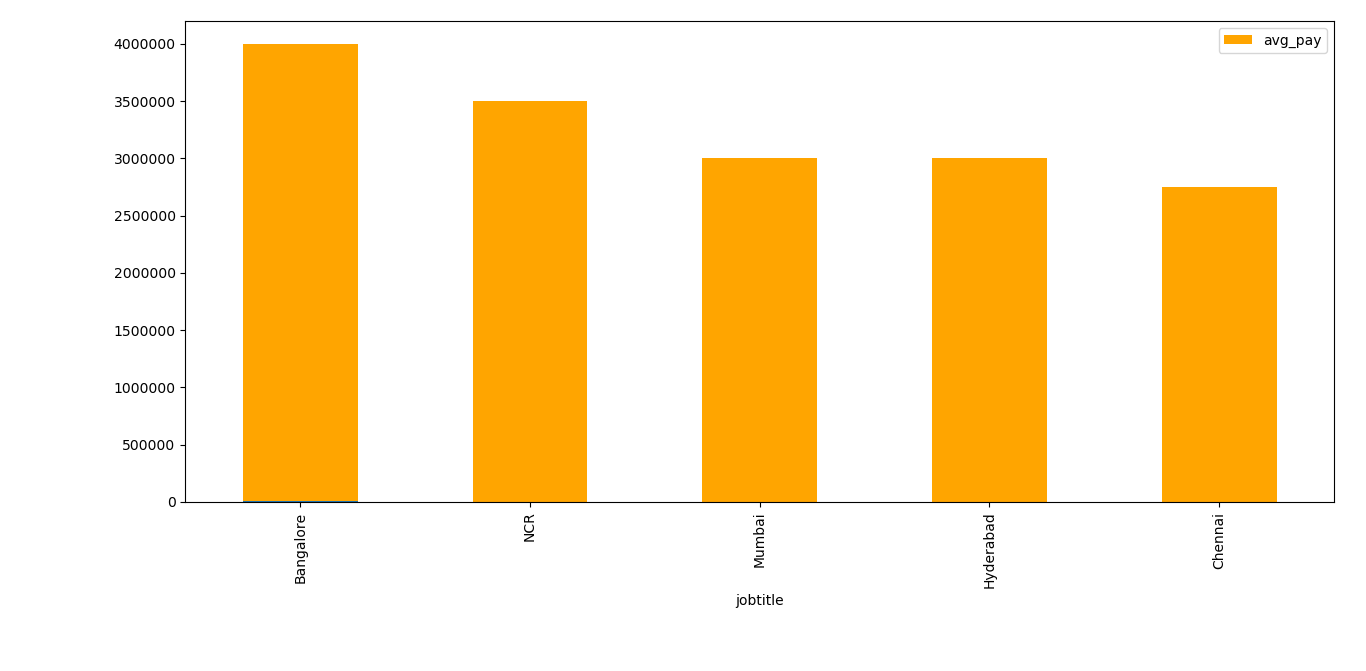
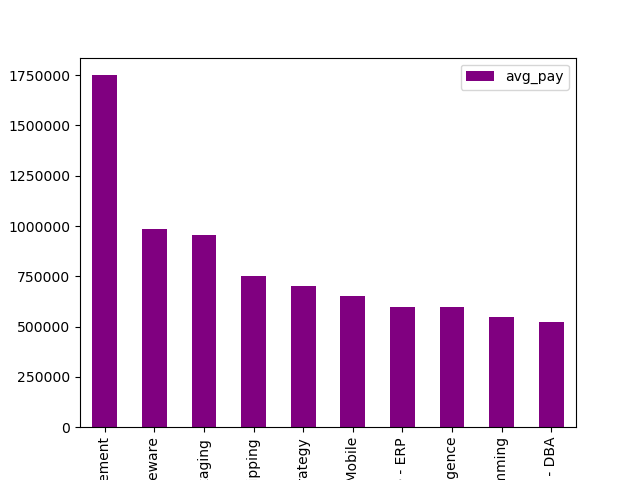
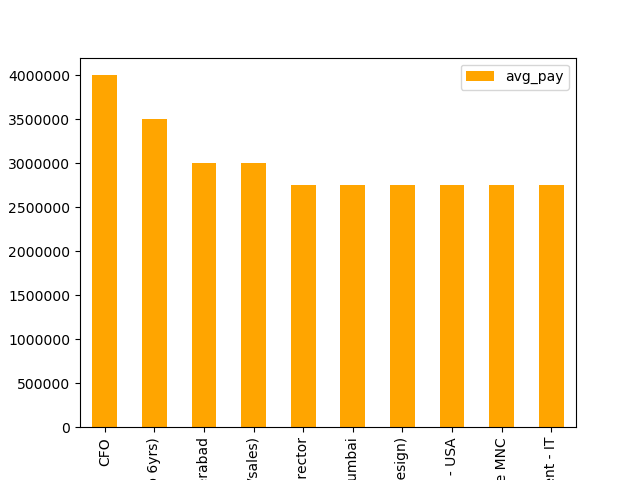
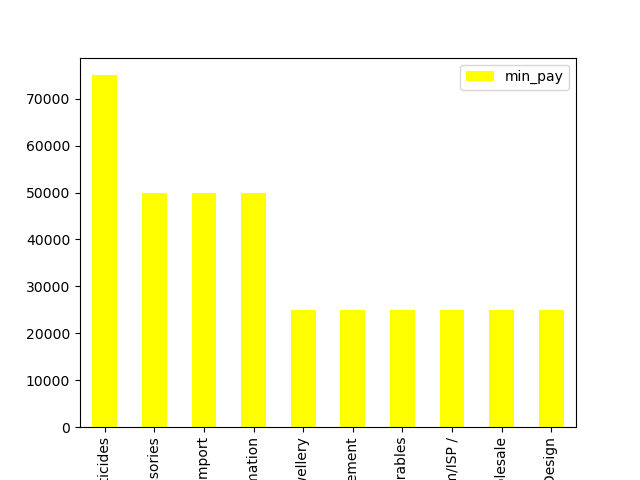
1. **Steps for data clean up:**
2. Duplicity withing the raw data was checked.
3. Experience and pay rate columns were divided into Minimum and maximum rates/experience columns while removing all the other extra characters within them.
4. Average rates and experience were also created.
5. Replacing the joblocation\_address with the common names, such as Delhi/Noida replaced with NCR and Mumbai/Navi Mumbai with just Mumbai throughout the data etc.
6. Checked the null values.
7. The null spaces for the companies show no relevancy so we dropped these 4 null companies’ complete row.
8. After reviewing the industry null values, there was only 1 row which was related to accounts and not software development so it was no use for our analysis, so we can drop it rather than imputing it
9. Dropped the unwanted columns which won’t be of any use in our model making
10. Filtered the data (As we are only interested in skills which are related to software development only, so we will make a different data set of the same and apply our model)
11. Filled the null places in the new data set with proper imputations such as mean, median, mode wherever it was necessary.
12. Converted the textual data into machine readable numbers.
13. Label Encoding was performed to do the 11th step.
14. **Data Visualization and**  **Insights:**
15. Below Figure shows that Bangalore has the maximum job positions followed by NCR.
16. Below Fig shows the average pay with respect to skills



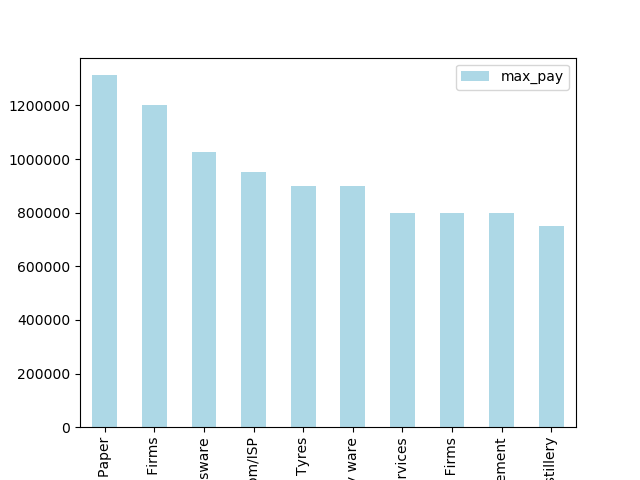
1. Average Pay with respect to job title



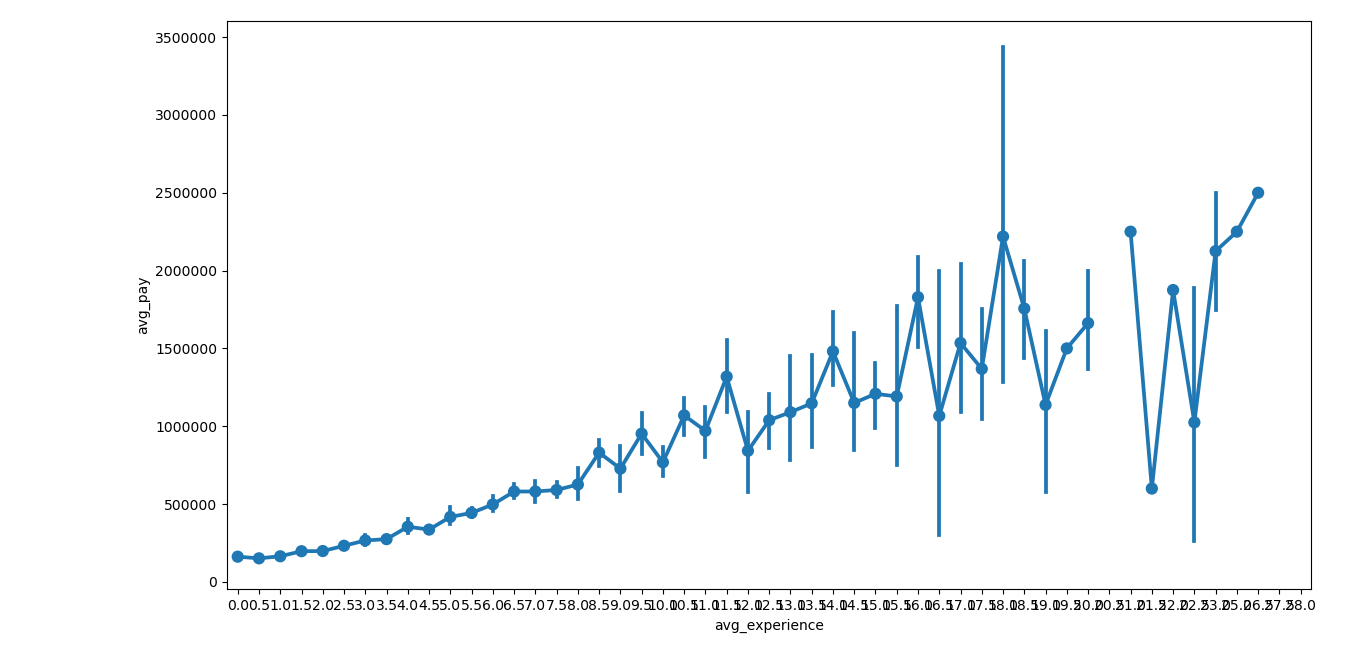
1. Minimum pay with respect to industry.

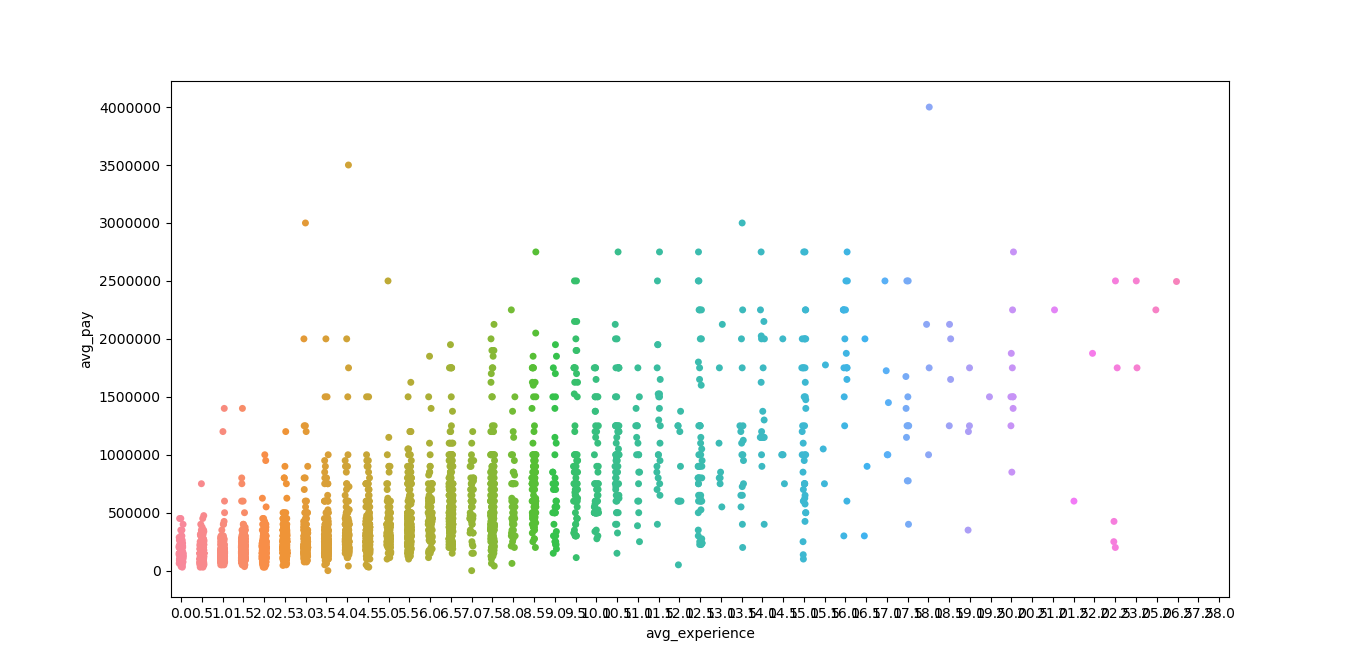


1. Max pay with respect to industry

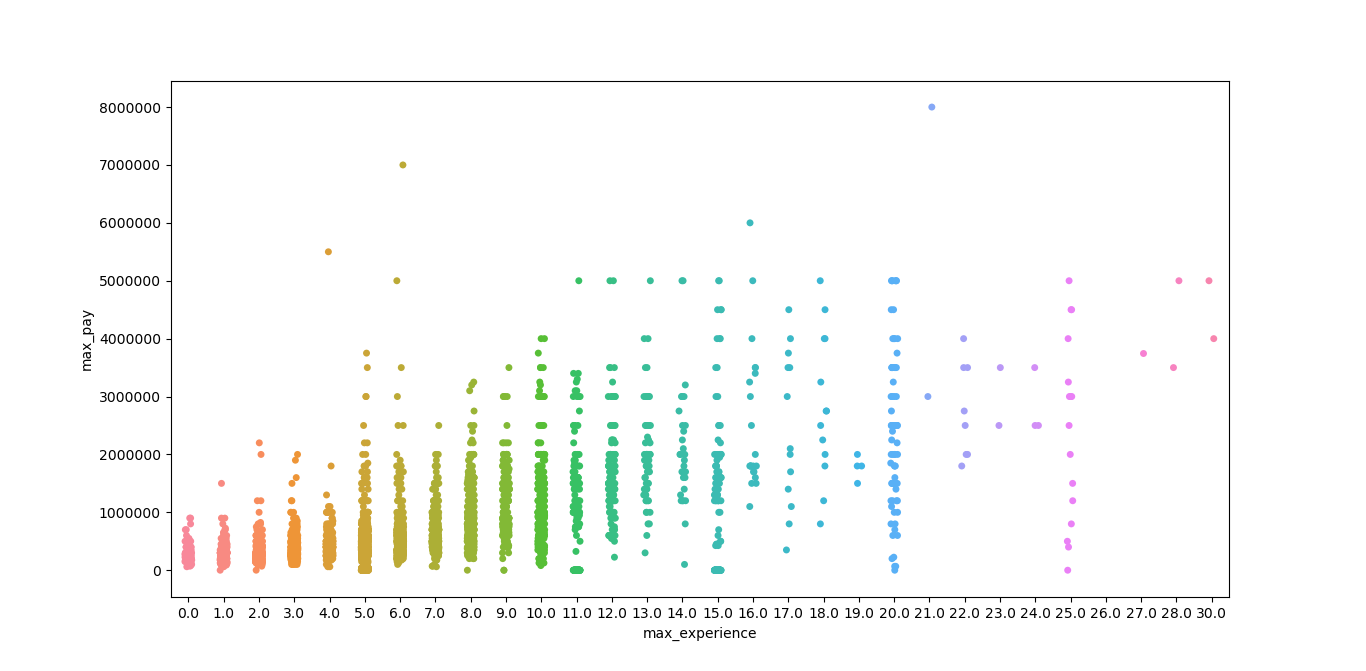
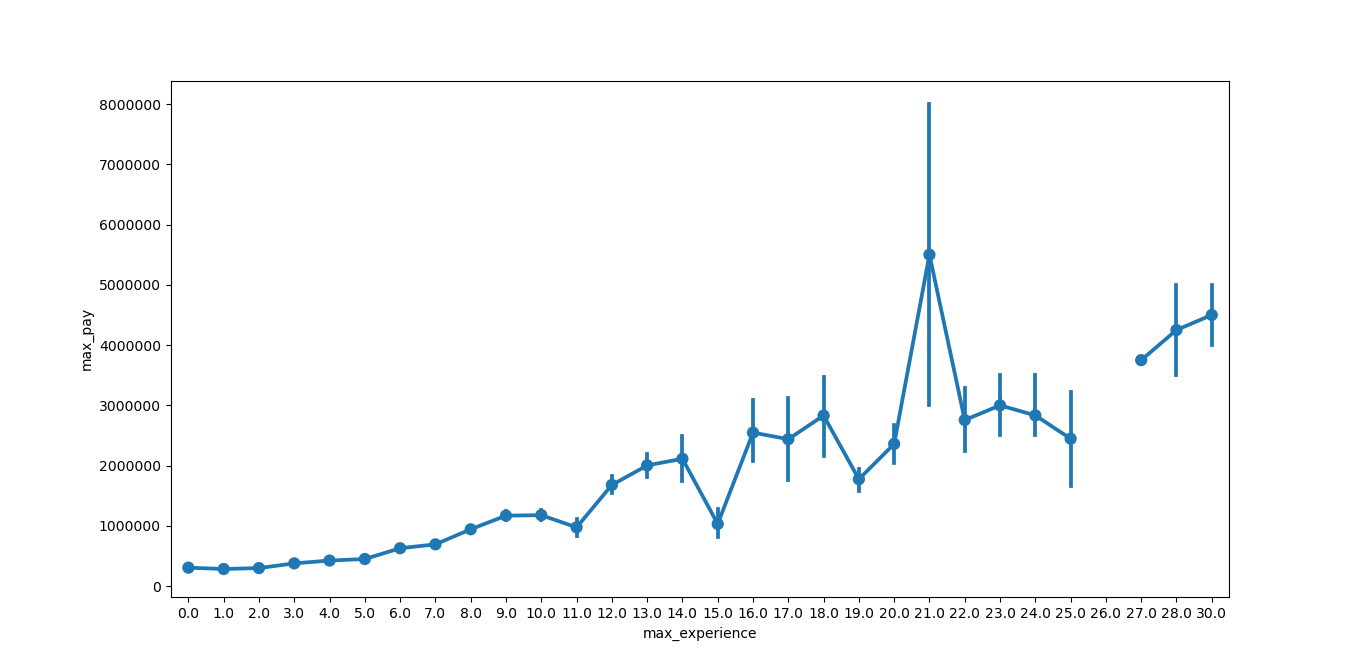


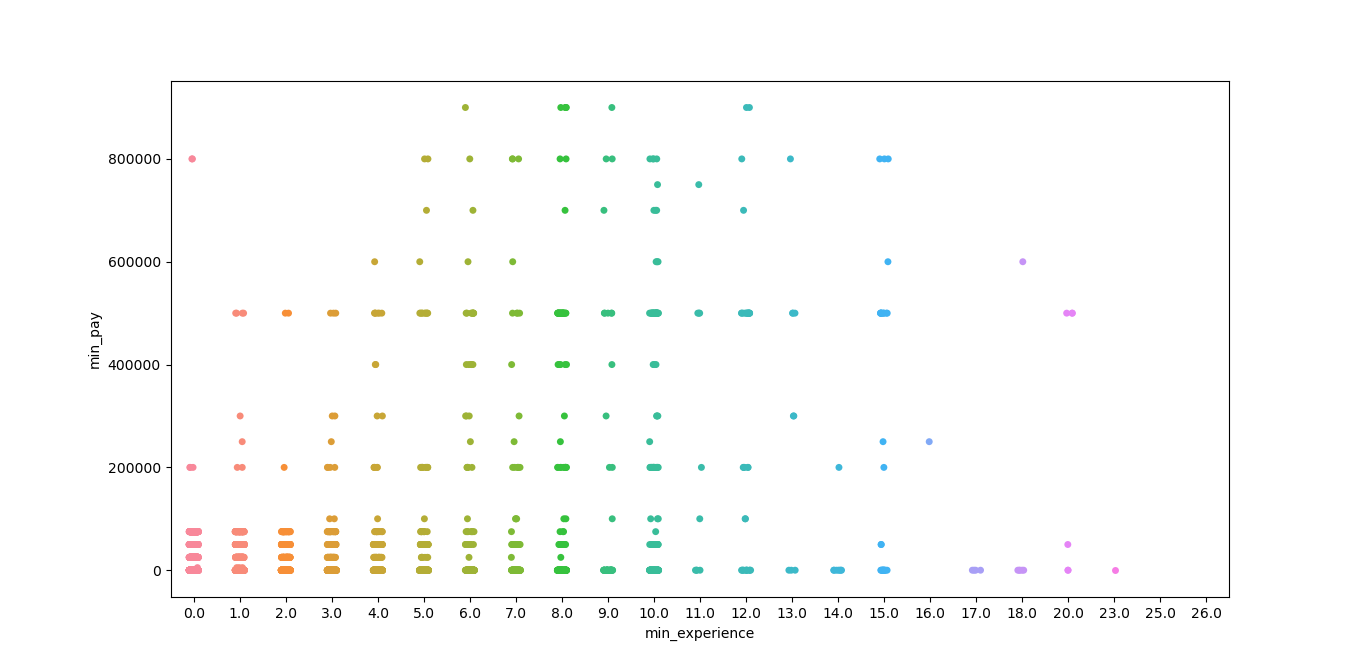
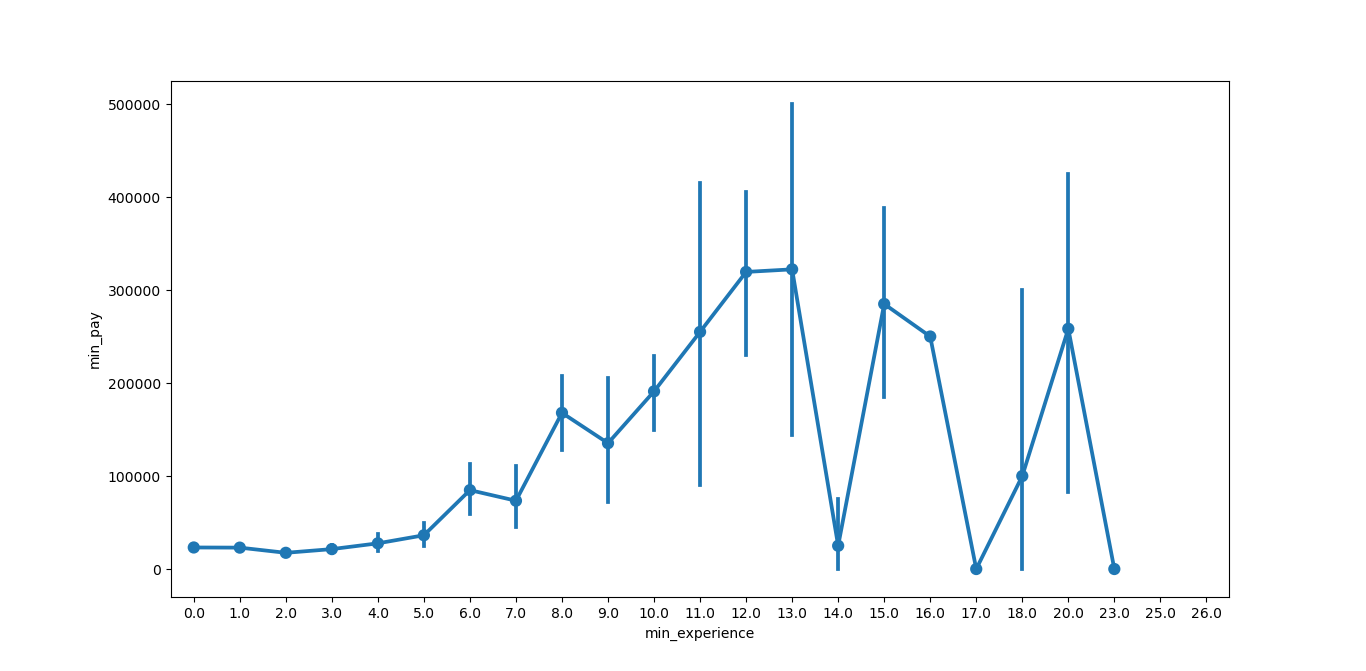
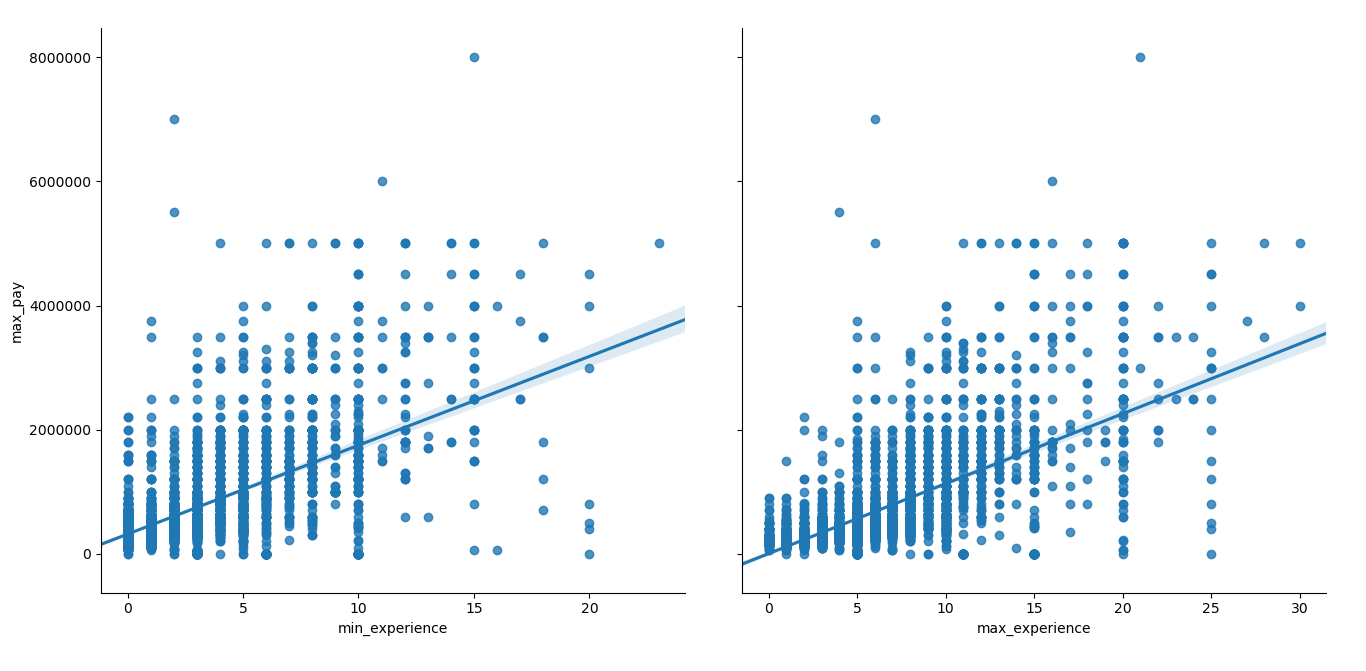
1. Average pay Vs. Average Experience

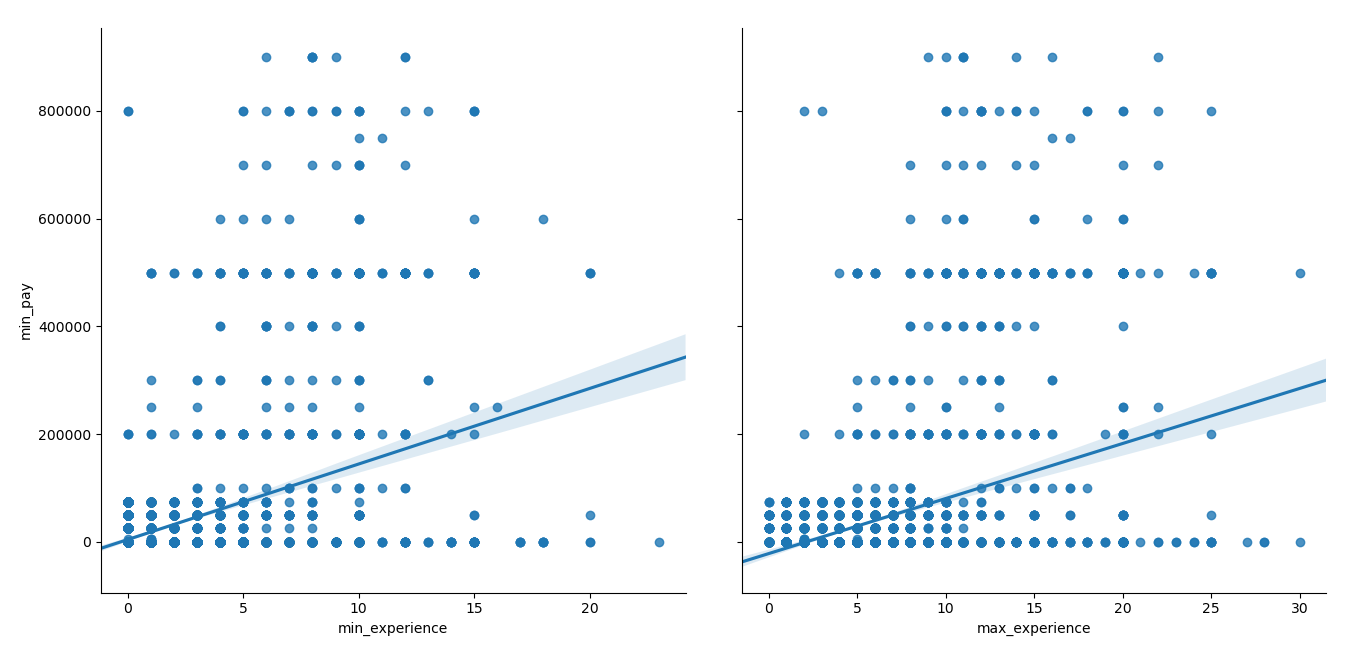




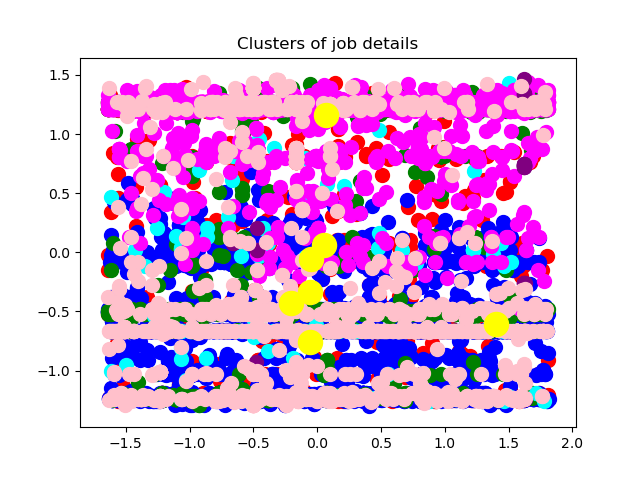
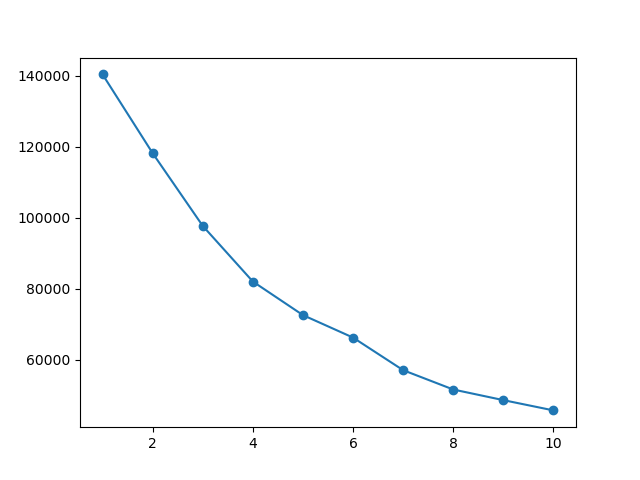
1. Max Pay Vs. Max Experience



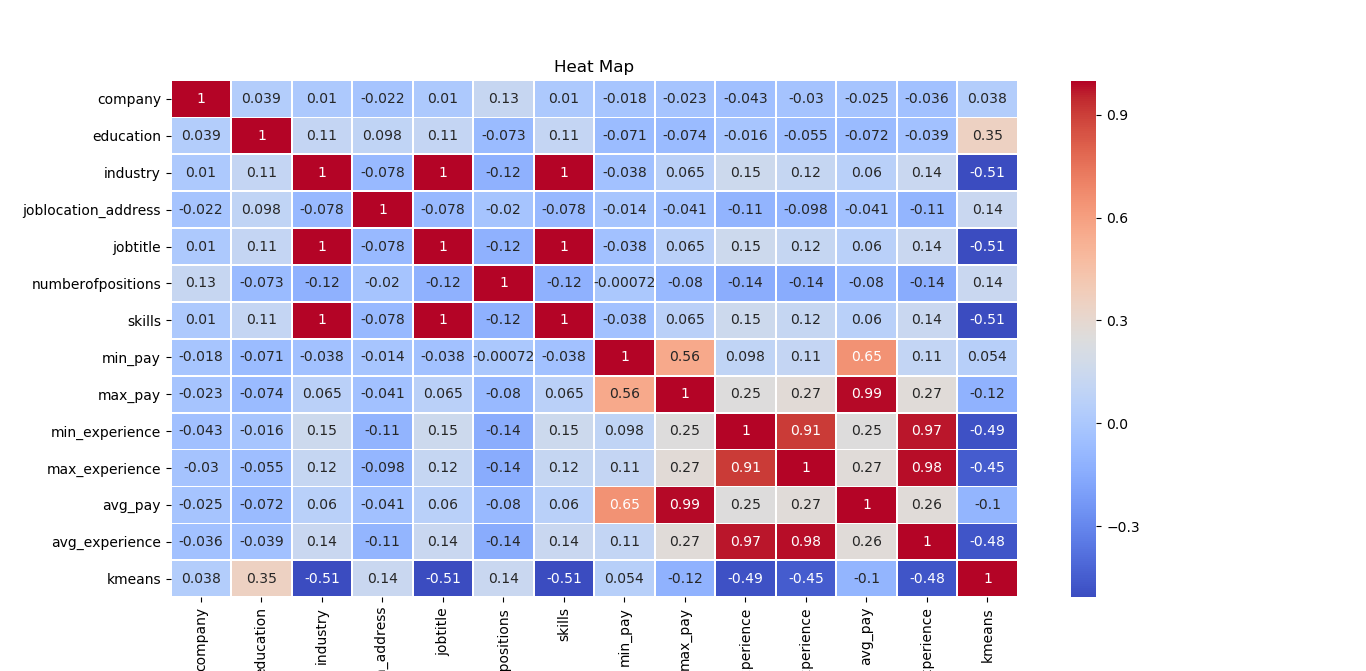
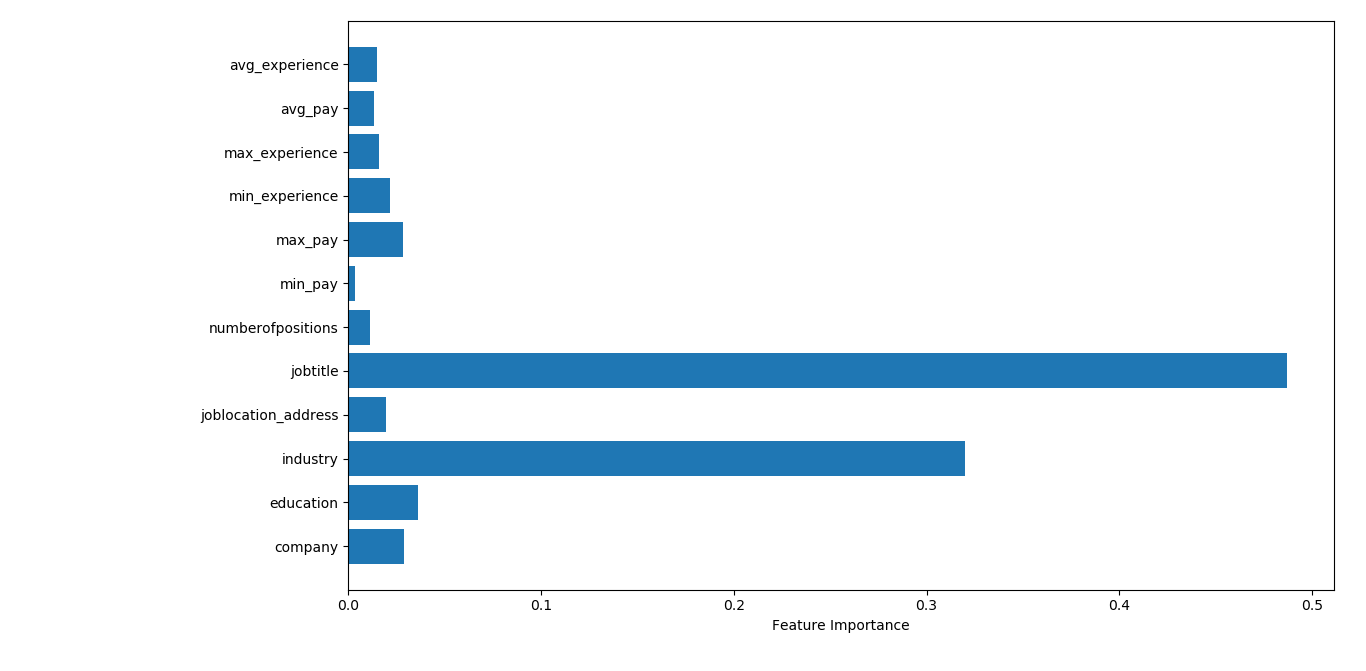
1. Min pay vs. min exp.
2. Regression plots



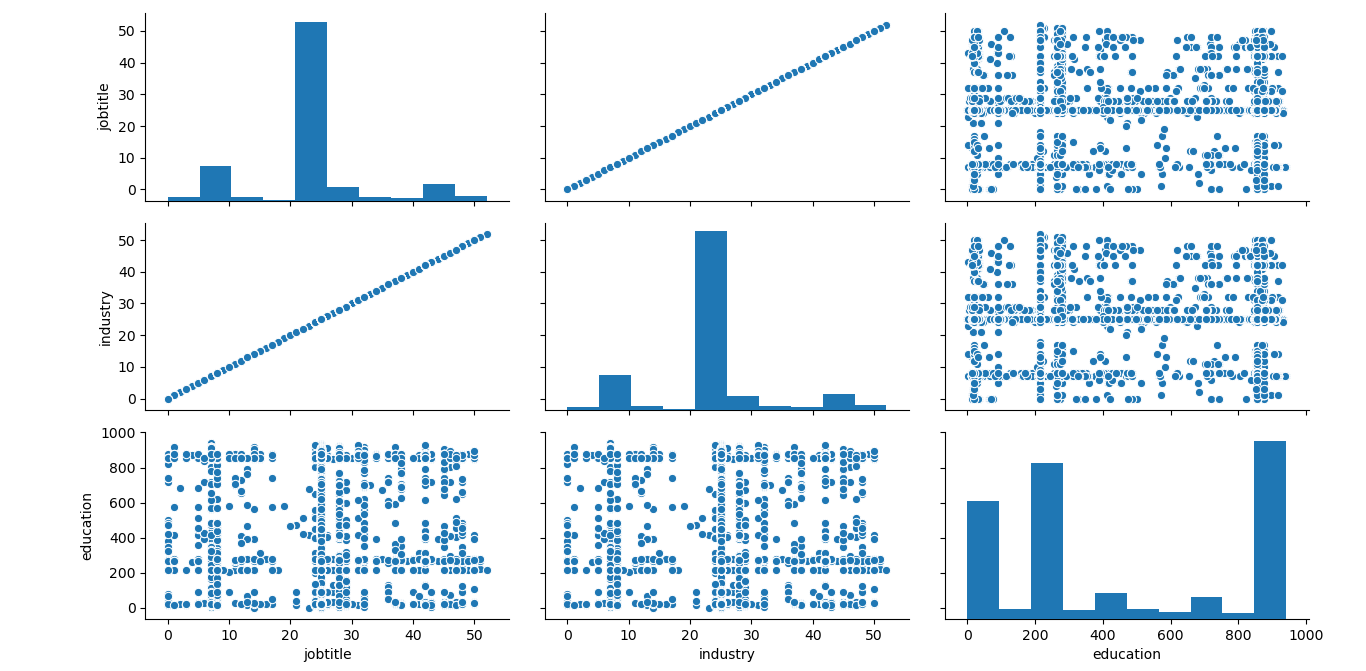
1. K Means Elbow and Clustering:



1. Heat Map and Feature importance



1. Pair Plot



1. **Models and There Performance:**

Linear Regression, Naive Bayes and Decision Tree were leading to over fitting

The testing result of all are mentioned below:  
 1. Logistic Regression — 82%  
 2. Nearest Neighbor — 85.1%  
 3. Support Vector Machines — 95.4%  
 4. Kernel SVM — 89.2%  
 5. Naive Bayes — 99.8% (**Overfitting**)  
 6. Decision Tree Algorithm — 99.9% (**Overfitting**)  
 7. Random Forest Classification — 95.2%