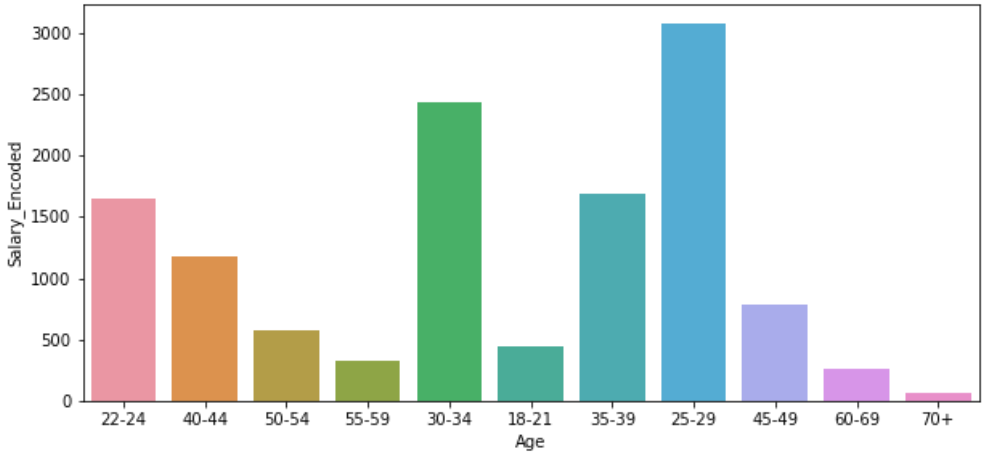
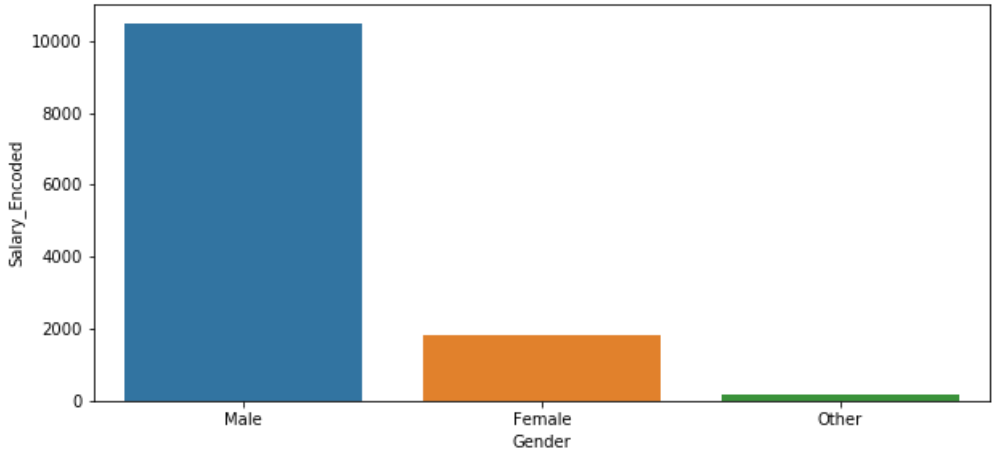
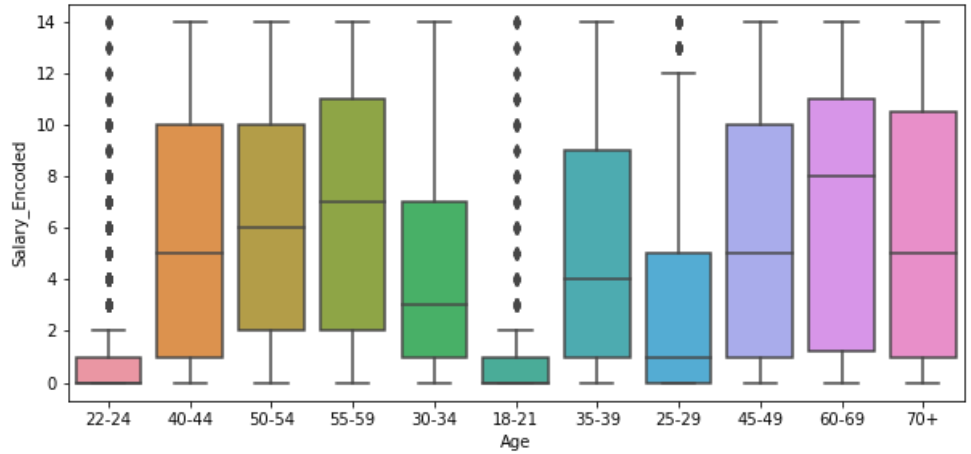
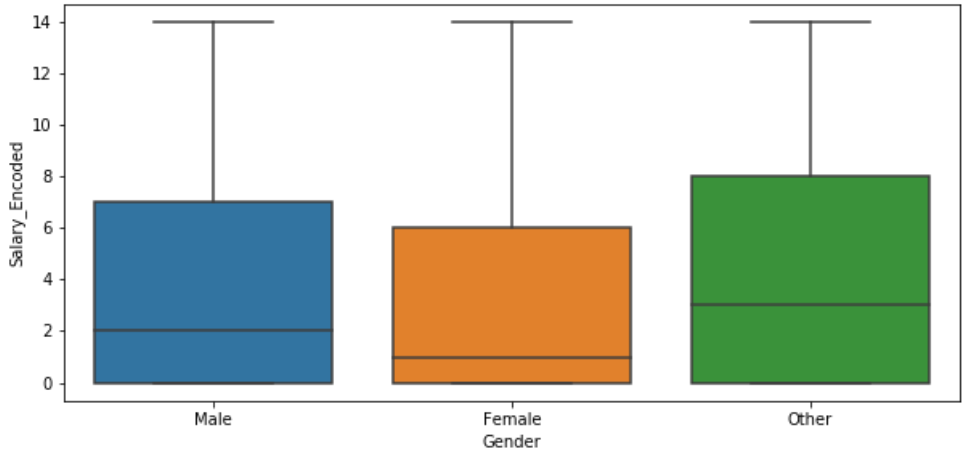
Exploratory Data Analysis



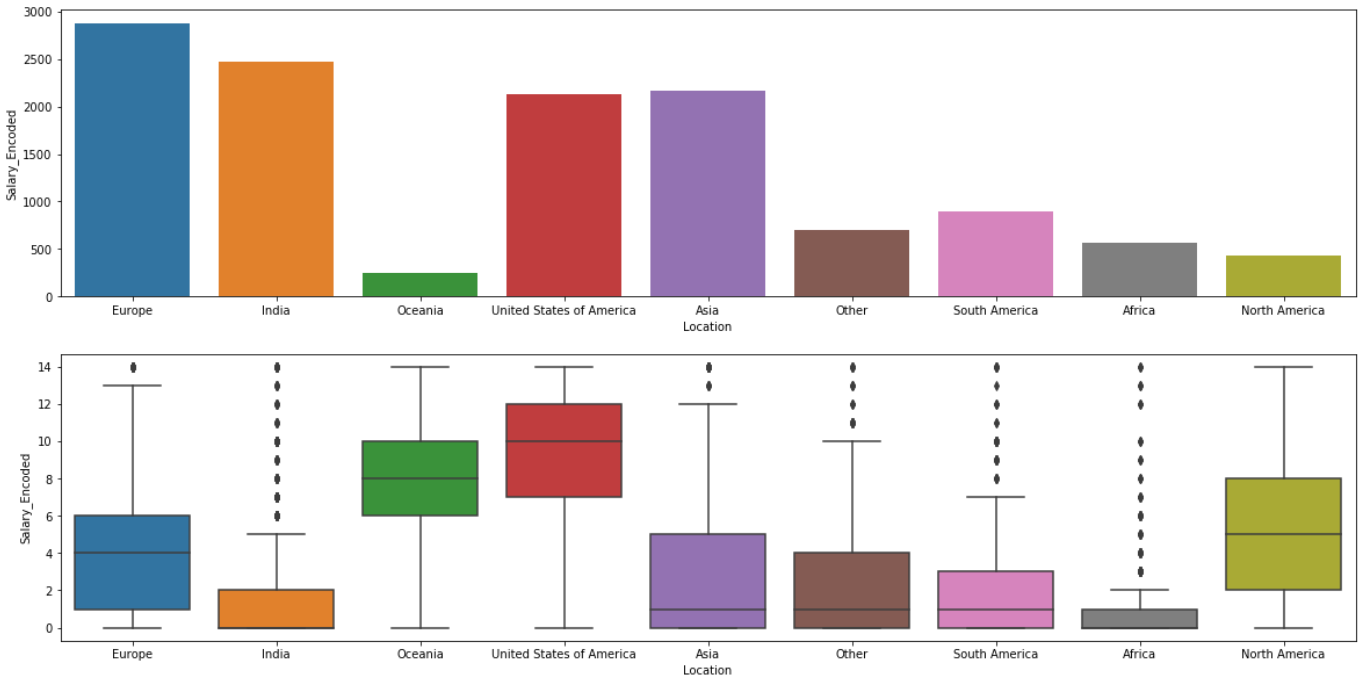


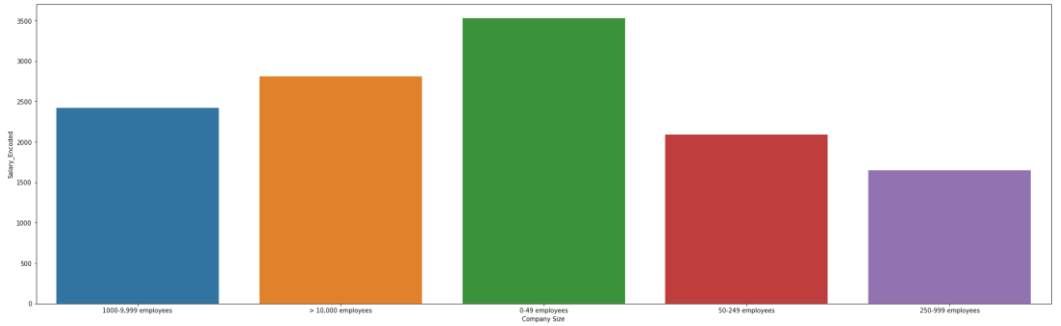


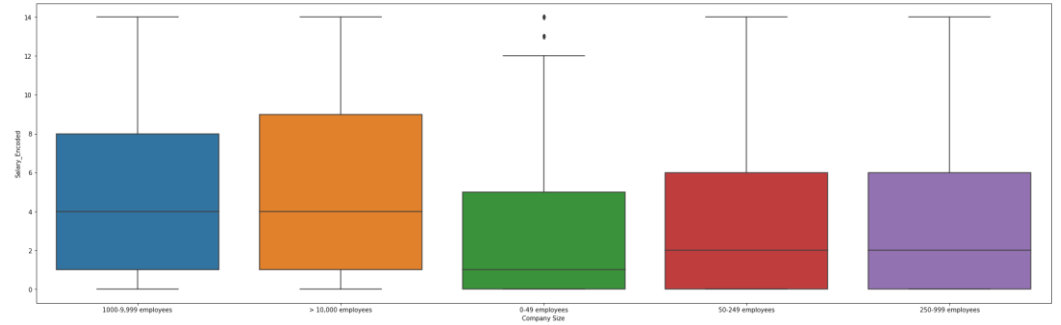


- Exploratory data analysis was carried out using box plot and countplot
- We see that age of the survey respondent has a significant influence on their salary with people having higher ages commanding higher salaries.
- We don't see a significant difference in salary between males & females, while those who have chosen not to identify their gender seem to command a higher salary in this survey. However, this is something that cannot be generalized since the number of respondents is 'other' category of gender is very low as evidenced by the frequency graph.

Exploratory Data Analysis (continued)







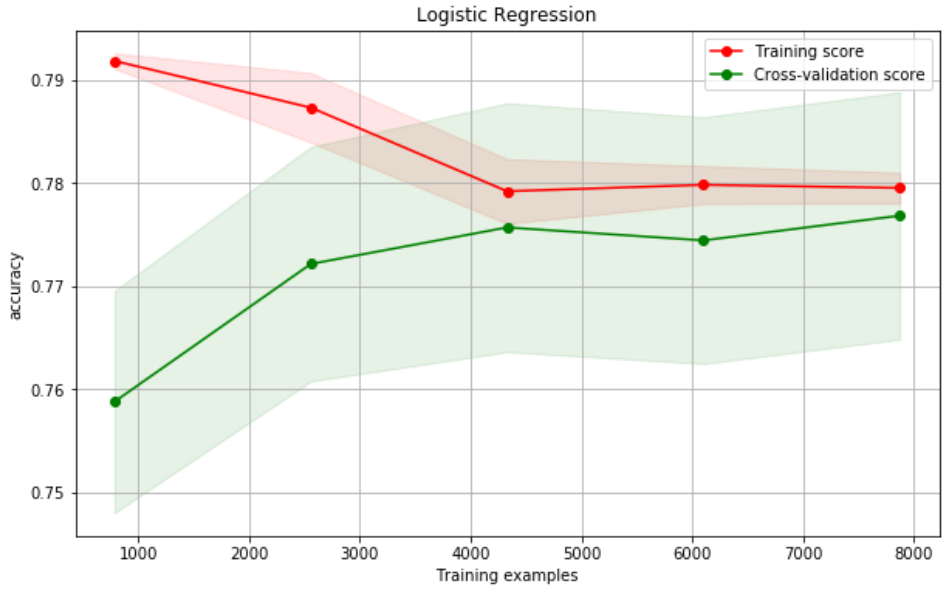
- The location graph is indeed a very crowded one and necessitates categorization into continents. I have proceeded to categorize each country into a continent but leave out US and India since they represent a significantly high number of respondents & so could potentially skew each continent in their direction. Also, leaving these two countries out could potentially help the logistic regression model to better predict the salary responses of data scientists from these countries.
- From the second chart, we see there exists a relationship between size of company and the salary received by the respondents with companies having greater than 10000 employees making the most on average.
- These are just samples & more analysis is carried out in Jupyter notebook.

Feature Selection & Engineering

The Cramers-V heatmap gives us a good idea of association between features. We can see that some features exhibit strong correlation and this can be exploited for feature engineering & reducing complexity.

- We see that Data Science Programming Language is strongly correlated (85%) with coding experience and fairly correlated with primary data analysis tool (55%) and ML experience (49%). It is safe to say that this feature is well-represented by these other features, so we can drop this column.
- RDBMS producs are fairly strongly correlated with ML Spend & Cloud computing producs, so we can drop this column as well.
- Computer Vision is srongly correlated with ML_algorithm categories & so we can drop his column.
- There seems to be a strong inter-dependence relationship between the last few features, namely cloud-computing platforms, data anaytics products, ML products. So, it makes sense to leverage this to reduce complexity by adding one feature that combines all of these features together: ML/Analytics technology which will be 1 if any of these products in use and 0 if none of them are in use. We will drop the features after creating this new column.
- We also note that the use of NLP is better captured by the use of ML algorithm categories feature (49% correlation). So, we can drop this. We also note a very weak correlation between ML automated algorithms & response. We also need to take into account that use of automated ML algorithms is characterized by our new ML/Analytics feature. So, we can drop. Time taken to fill survey is of no significance to the model, so we drop that too.

Implementing Ordinal Logistic Regression

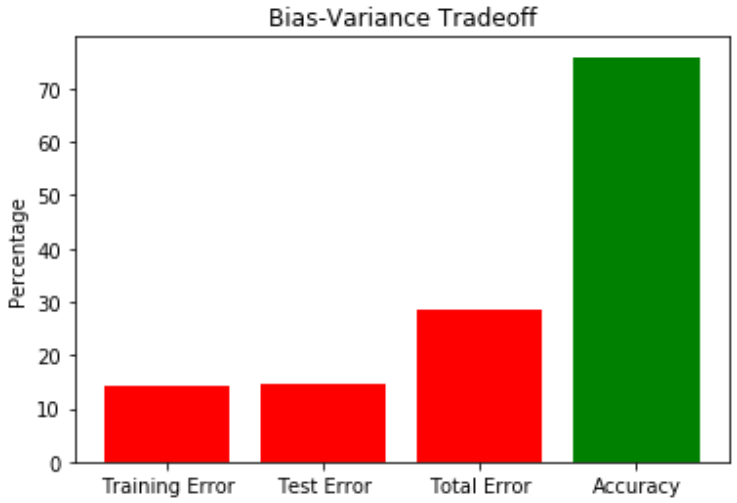


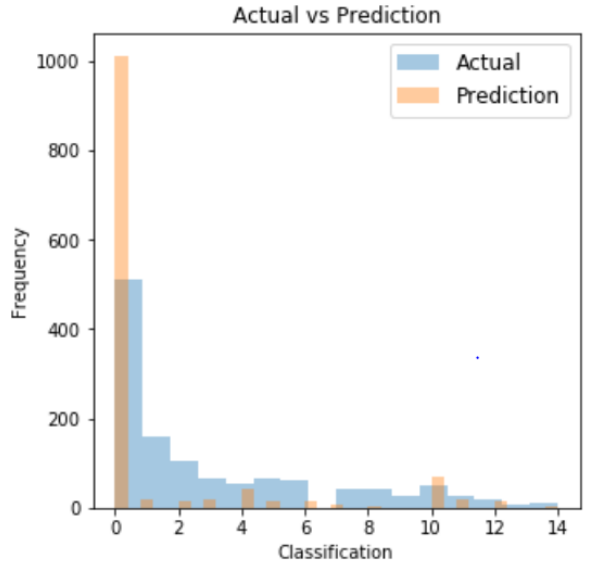
- Data was split into training set & test set on a 70:30 ratio
- Ordinal Logistic Regression was implemented and a 10-fold cross validation as performed on the training set (see learning curve to the left) with an average accuracy of 75.186%
- The table below represents the probability of prediction for each salary class after ordinal logistic regression (only 5 rows displayed)

```
Fold 1: Accuracy: 0.705%
Fold 2: Accuracy: 0.759%
Fold 3: Accuracy: 0.727%
Fold 4: Accuracy: 0.749%
Fold 5: Accuracy: 0.74%
Fold 6: Accuracy: 0.792%
Fold 7: Accuracy: 0.733%
Fold 8: Accuracy: 0.765%
Fold 9: Accuracy: 0.787%
Fold 10: Accuracy: 0.76%
Average Score: 75.186%(2.517%)
```

	Salary Label 0	Salary Label 1	Salary Label 2	Salary Label 3	Salary Label 4	Salary Label 5	Salary Label 6	Salary Label 7	Salary Label 8	Salary Label 9	Salary Label 10	Salary Label 11	Salary Label 12	Salary Label 13	Salary Label 14
0	0.612080	0.144341	0.685481	0.209219	0.713902	0.217665	0.741946	0.218999	0.752200	0.224445	0.757770	0.233542	0.763788	0.234828	0.765172
1	0.028820	0.008172	0.047078	0.021217	0.073452	0.038745	0.114038	0.104435	0.213525	0.203997	0.438100	0.407159	0.538862	0.449103	0.550897
2	0.732025	0.132619	0.789102	0.151719	0.801450	0.163593	0.805165	0.171684	0.806088	0.173334	0.811635	0.170556	0.811611	0.178502	0.821498
3	0.469288	0.250789	0.590389	0.274305	0.630302	0.298887	0.647044	0.316973	0.651665	0.322391	0.660580	0.322244	0.665928	0.326176	0.673824
4	0.014905	0.007263	0.026337	0.014185	0.035436	0.023466	0.068587	0.055005	0.164648	0.145561	0.495319	0.370601	0.565860	0.410065	0.589935

Model Evaluation, Bias-Variance Tradeoff & Hyperparameter Tuning





- The model got an accuracy of 75.9 percent on the test set.
- From the training error & test error, we see that while the bias and variance are fairly comparable and it can't be said that one dominates the other. Having said that, the model does lean towards under-fitting because the bias is only moderately low compared to the variance, and so there is room for improvement on the bias front.
- After performing the Hyper Parameter Tuning, we can see that optimal model uses C of 0.05 & a newton-cg solver. It has resulted in an improvement in accuracy of 0.019 percent.
- From the Actual vs Prediction plot, we see that the model has classified many labels in the <10000 USD class. Perhaps, this is due to the skewness present in the test data.