Bhavan's Vivekananda College
Sainikpuri
Autonomous College

Capstone Project

Analysis on Engineering Graduate Salary

Group-6
Team Members

1.UudhhayKiirran
2.Bhavana Reddy
3.Jashwanth
4.SaiPrasanna

Introduction

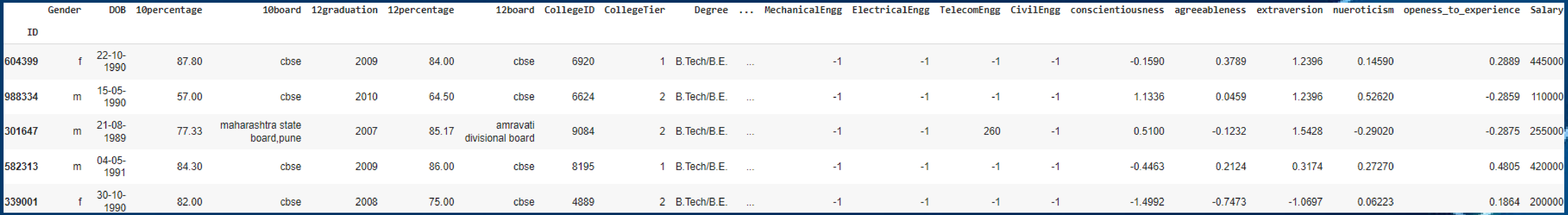
Dataset : Engineering_Graduate Salary

Objective : Our objective is to determine the salary of an engineering graduate .

Technical contents:

- **Data importing**
- **Data exploration**
- **Data Preprocessing**
- **Data Modelling**

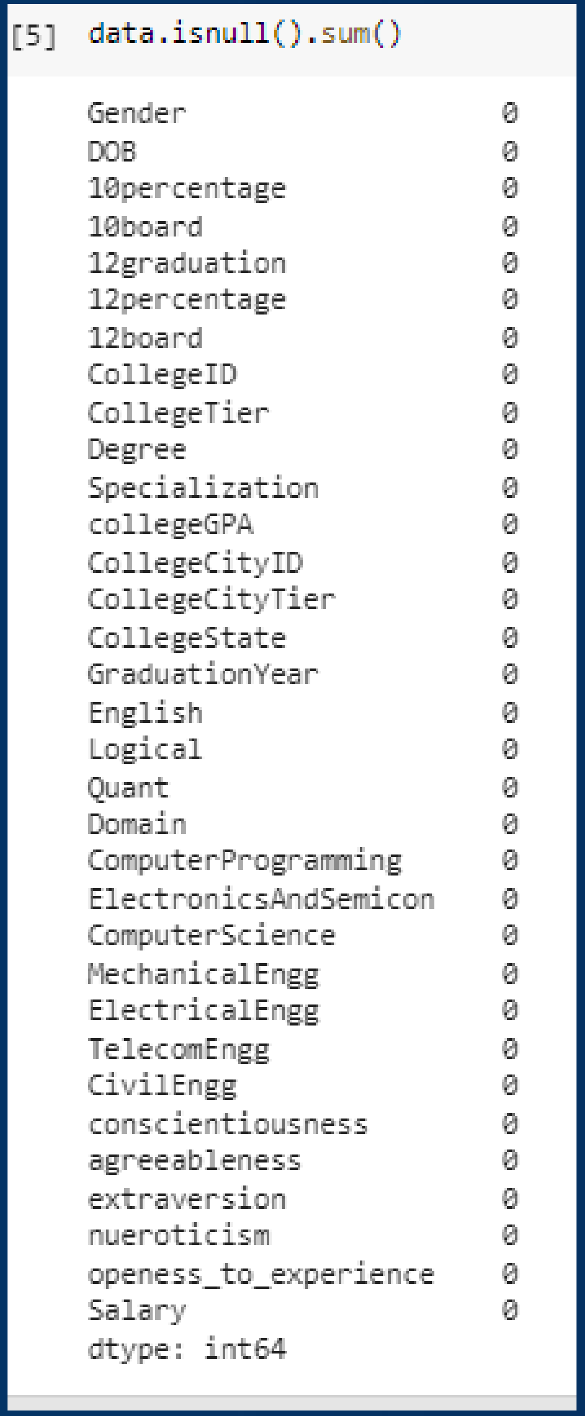
Data Imported:



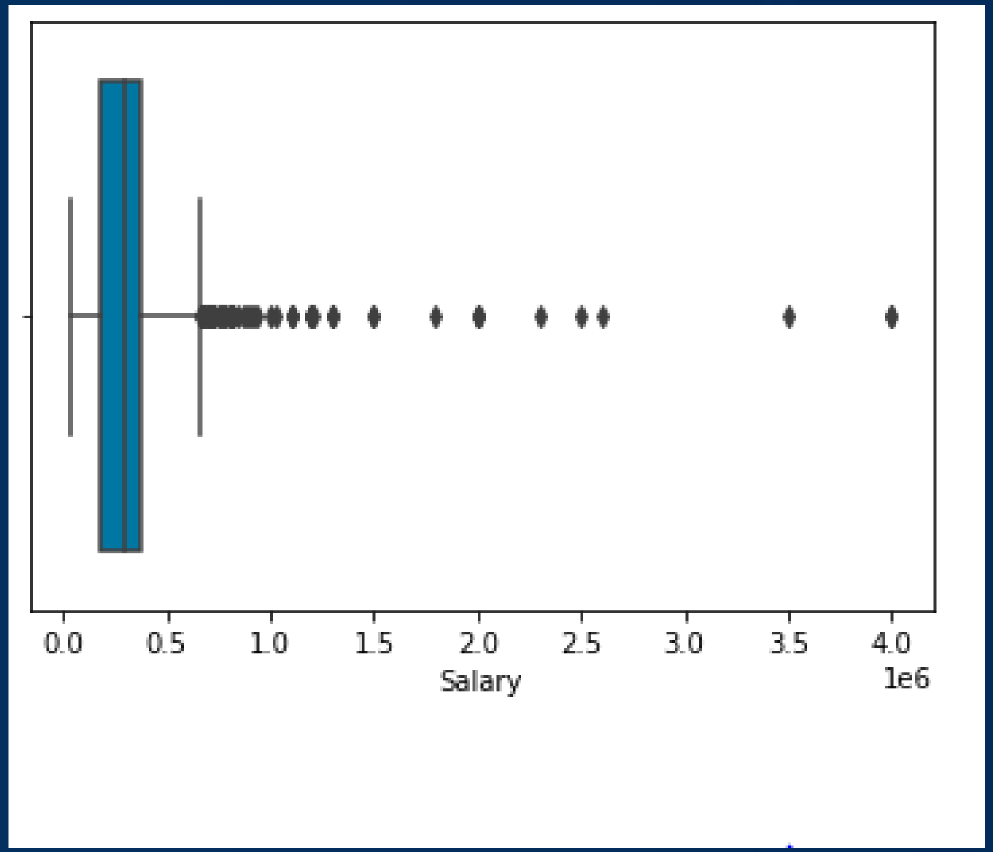
	Gender	DOB	10percentage	10board	12graduation	12percentage	12board	CollegeID	CollegeTier	Degree	...	MechanicalEngg	ElectricalEngg	TelecomEngg	CivilEngg	conscientiousness	agreeableness	extraversion	nueroticism	openess_to_experience	Salary
ID																					
604399	f	22-10-1990	87.80	cbse	2009	84.00	cbse	6920	1	B.Tech/B.E.	...	-1	-1	-1	-1	-0.1590	0.3789	1.2396	0.14590	0.2889	445000
988334	m	15-05-1990	57.00	cbse	2010	64.50	cbse	6624	2	B.Tech/B.E.	...	-1	-1	-1	-1	1.1336	0.0459	1.2396	0.52620	-0.2859	110000
301647	m	21-08-1989	77.33	maharashtra state board,pune	2007	85.17	amravati divisional board	9084	2	B.Tech/B.E.	...	-1	-1	260	-1	0.5100	-0.1232	1.5428	-0.29020	-0.2875	255000
582313	m	04-05-1991	84.30	cbse	2009	86.00	cbse	8195	1	B.Tech/B.E.	...	-1	-1	-1	-1	-0.4463	0.2124	0.3174	0.27270	0.4805	420000
339001	f	30-10-1990	82.00	cbse	2008	75.00	cbse	4889	2	B.Tech/B.E.	...	-1	-1	-1	-1	-1.4992	-0.7473	-1.0697	0.06223	0.1864	200000

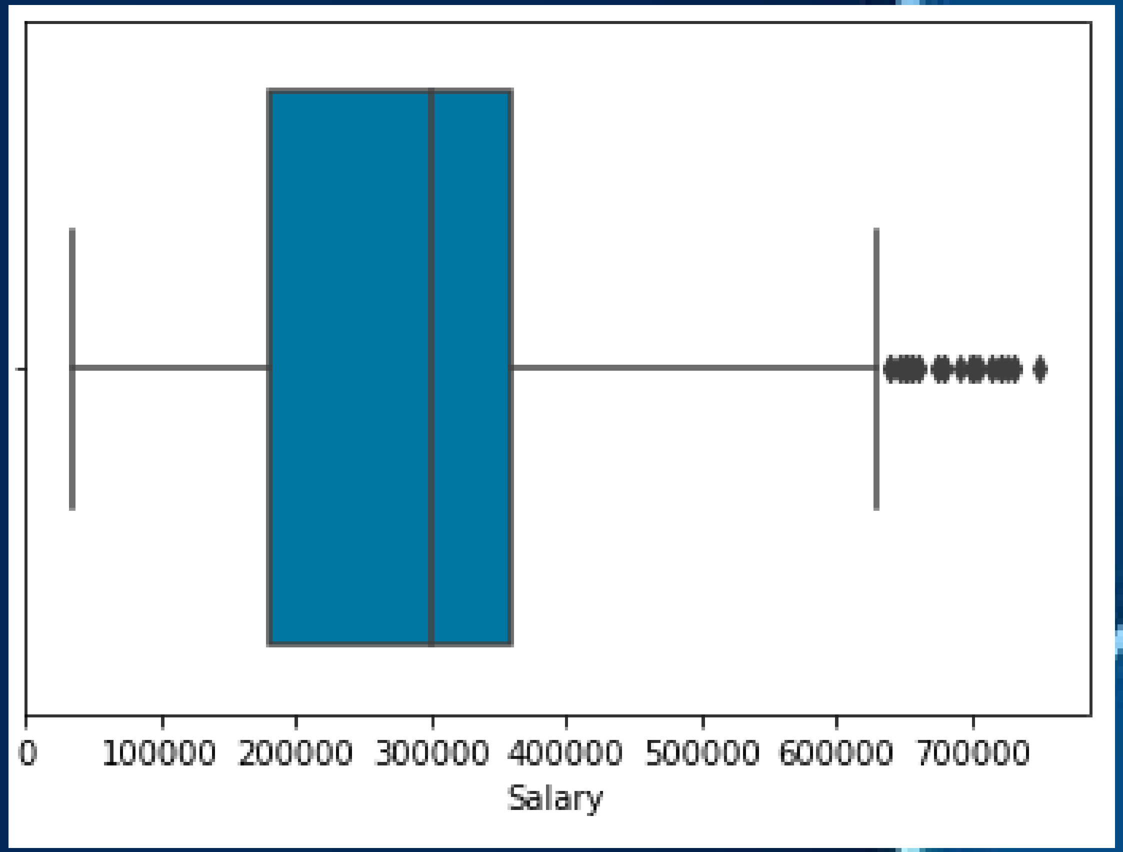
- This dataset contains 2,998 rows and 33 columns.
- The target variable is salary.
- This dataset contians numerical values.
- The columns of the dataset describe about Engineering graduates.

Checking for Null values



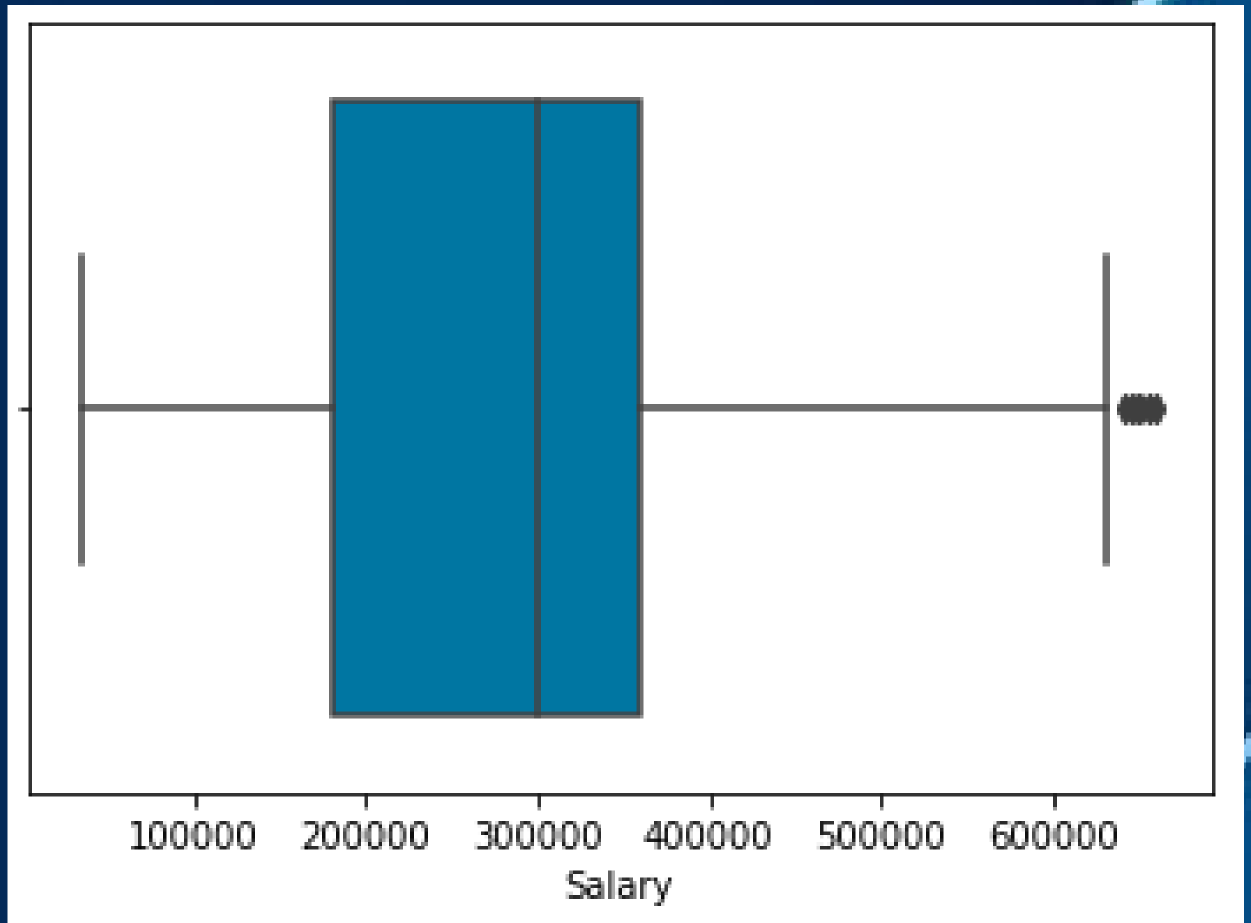
```
[5]  data.isnull().sum()

     Gender                    0
     DOB                       0
     10percentage              0
     10board                   0
     12graduation              0
     12percentage              0
     12board                   0
     CollegeID                 0
     CollegeTier               0
     Degree                    0
     Specialization            0
     collegeGPA                0
     CollegeCityID             0
     CollegeCityTier           0
     CollegeState              0
     GraduationYear            0
     English                   0
     Logical                   0
     Quant                     0
     Domain                    0
     ComputerProgramming       0
     ElectronicsAndSemicon     0
     ComputerScience           0
     MechanicalEngg            0
     ElectricalEngg            0
     TelecomEngg               0
     CivilEngg                 0
     conscientiousness         0
     agreeableness             0
     extraversion              0
     nueroticism               0
     openess_to_experience     0
     Salary                    0
     dtype: int64
```





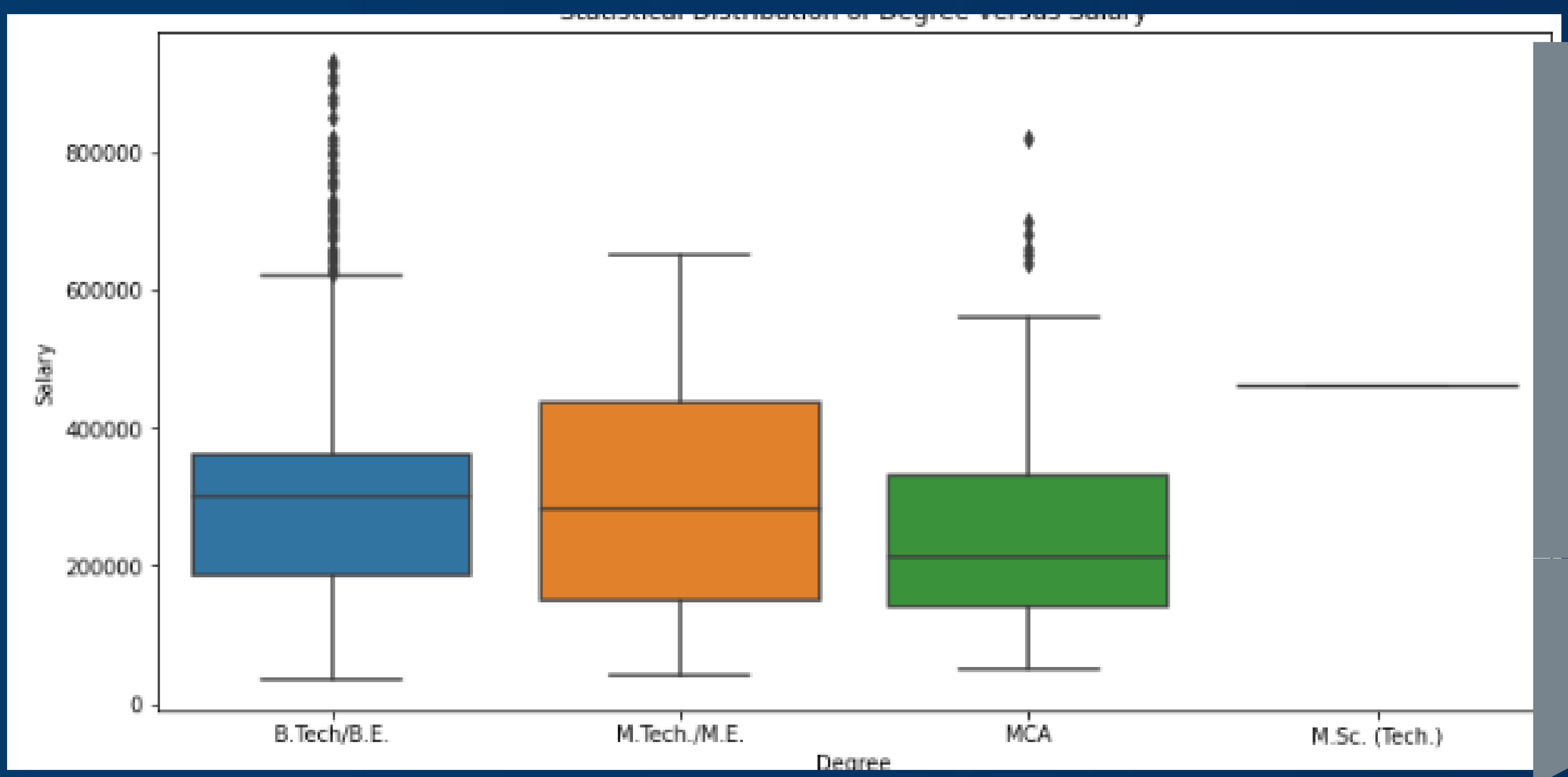
- The first box plot represents the outlers in dataset.
- We have used hample method to remove the outliers.

We used again zscore method to decrease the number of outliers.



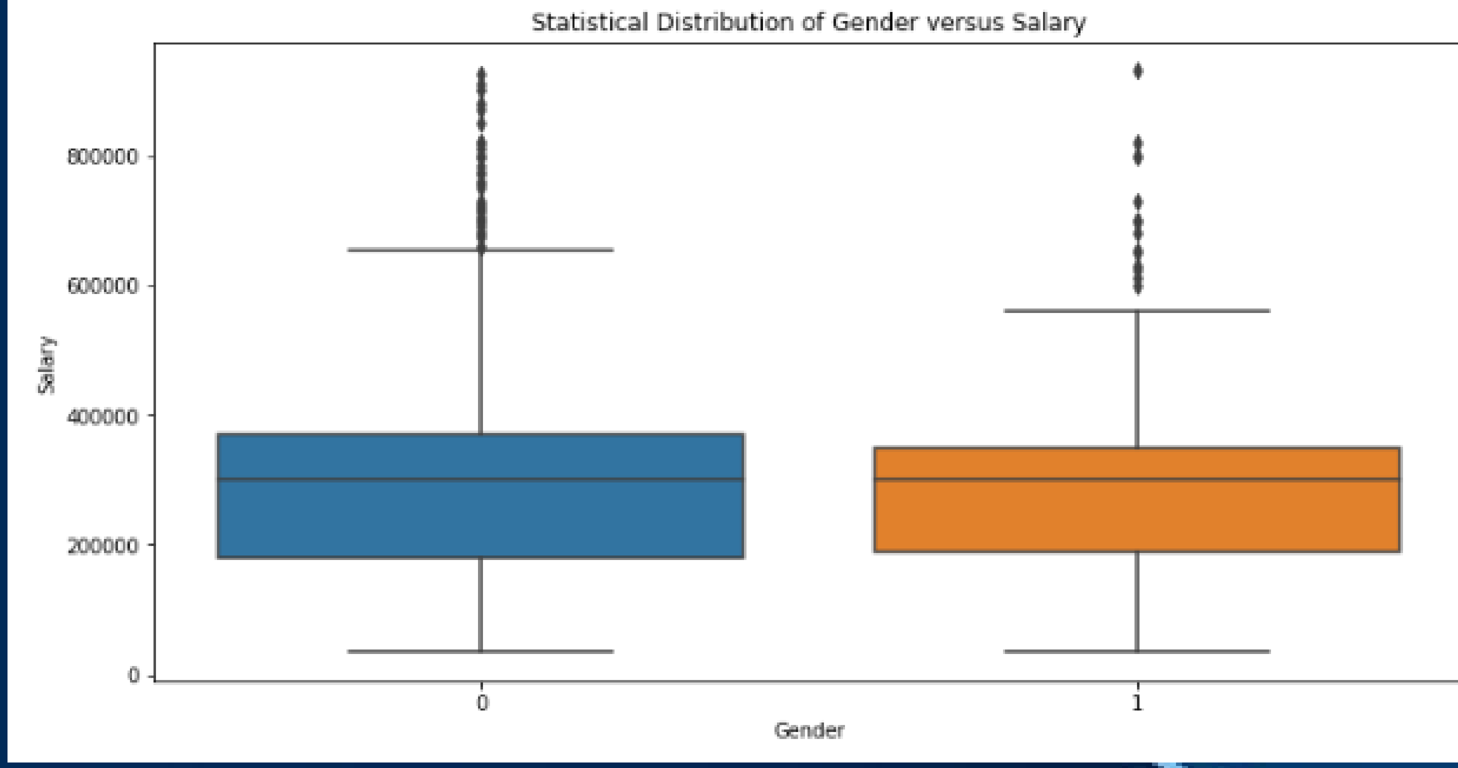
Exploratory Data Analysis:

Analysis of Salary with Degree

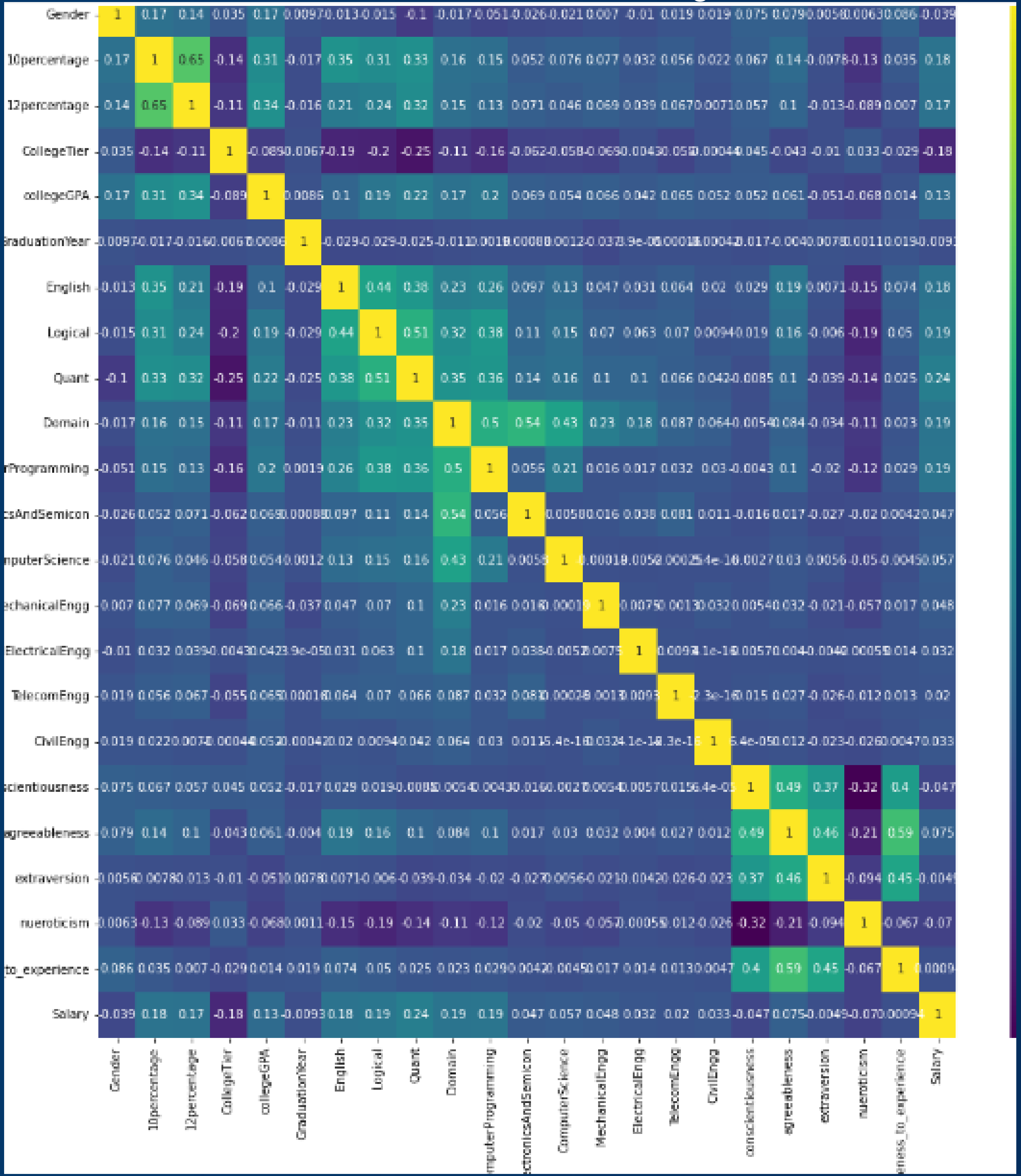


Average salary is highest for BE/B.tech graduates as compared to any other degree graduates.

Analysis of Salary with Gender

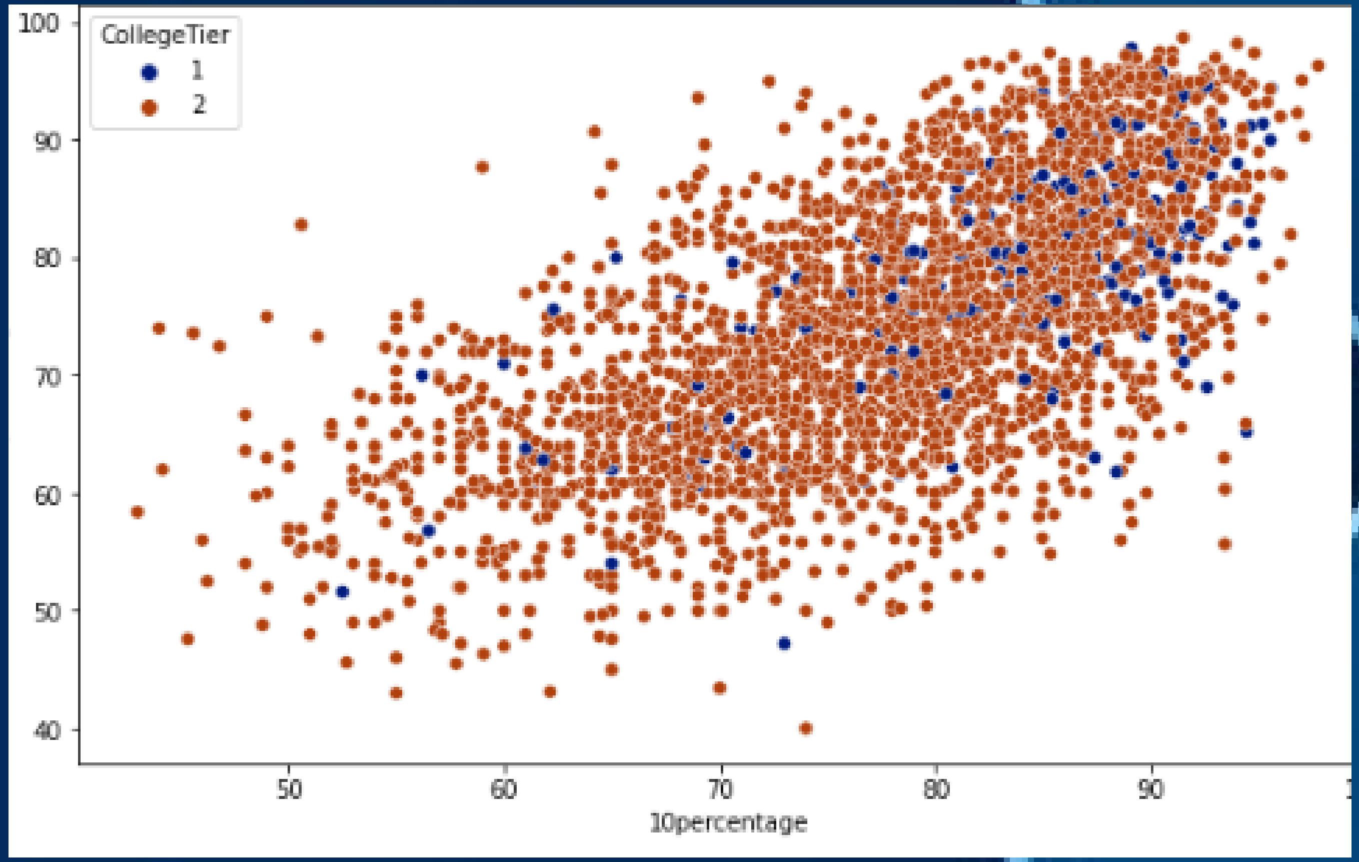


Correlation Analysis



	Gender	10percentage	12percentage	CollegeTier	collegeGPA	GraduationYear	English	Logical	Quant	Domain	mputerProgramming	ctronicsAndSemicon	ComputerScience	MechanicalEngg	ElectricalEngg	TelecomEngg	CivilEngg	conscientiousness	agreeableness	extraversion	nueroticism	eness_to_experience	Salary
Gender	1	0.17	0.14	0.035	0.17	0.0097	-0.013	-0.015	-0.1	-0.017	-0.051	-0.026	-0.021	0.007	-0.01	0.019	0.019	0.075	0.079	0.0056	0.0063	0.086	-0.039
10percentage	0.17	1	0.65	-0.14	0.31	-0.017	0.35	0.31	0.33	0.16	0.15	0.052	0.076	0.077	0.032	0.056	0.022	0.067	0.14	-0.0078	-0.13	0.035	0.18
12percentage	0.14	0.65	1	-0.11	0.34	-0.016	0.21	0.24	0.32	0.15	0.13	0.071	0.046	0.069	0.039	0.067	0.0071	0.057	0.1	-0.013	-0.089	0.007	0.17
CollegeTier	-0.035	-0.14	-0.11	1	-0.089	0.0067	-0.19	-0.2	-0.25	-0.11	-0.16	-0.062	-0.058	-0.069	0.0043	-0.055	0.00044	0.045	-0.043	-0.01	0.033	-0.029	-0.18
collegeGPA	0.17	0.31	0.34	-0.089	1	0.0086	0.1	0.19	0.22	0.17	0.2	0.069	0.054	0.066	0.042	0.065	0.052	0.052	0.061	-0.051	-0.068	0.014	0.13
raduationYear	-0.0097	-0.017	-0.016	0.0067	0.0086	1	-0.029	-0.029	-0.025	-0.011	0.0019	0.00088	0.0012	-0.037	3.9e-05	0.00018	0.00042	-0.017	-0.004	0.0078	0.0011	0.019	-0.0093
English	-0.013	0.35	0.21	-0.19	0.1	-0.029	1	0.44	0.38	0.23	0.26	0.097	0.13	0.047	0.031	0.064	0.02	0.029	0.19	0.0071	-0.15	0.074	0.18
Logical	-0.015	0.31	0.24	-0.2	0.19	-0.029	0.44	1	0.51	0.32	0.38	0.11	0.15	0.07	0.063	0.07	0.0094	0.019	0.16	-0.006	-0.19	0.05	0.19
Quant	-0.1	0.33	0.32	-0.25	0.22	-0.025	0.38	0.51	1	0.35	0.36	0.14	0.16	0.1	0.1	0.066	0.042	-0.0085	0.1	-0.039	-0.14	0.025	0.24
Domain	-0.017	0.16	0.15	-0.11	0.17	-0.011	0.23	0.32	0.35	1	0.5	0.54	0.43	0.23	0.18	0.087	0.064	-0.0054	0.084	-0.034	-0.11	0.023	0.19
rProgramming	-0.051	0.15	0.13	-0.16	0.2	0.0019	0.26	0.38	0.36	0.5	1	0.056	0.21	0.016	0.017	0.032	0.03	-0.0043	0.1	-0.02	-0.12	0.029	0.19
csAndSemicon	-0.026	0.052	0.071	-0.062	0.069	0.00088	0.097	0.11	0.14	0.54	0.056	1	0.0058	0.016	0.038	0.081	0.011	-0.016	0.017	-0.027	-0.02	0.0042	0.047
nputerScience	-0.021	0.076	0.046	-0.058	0.054	0.0012	0.13	0.15	0.16	0.43	0.21	0.0058	1	-0.00018	-0.0052	-0.00025	5.4e-16	-0.0027	0.03	0.0056	-0.05	-0.0045	0.057
echanicalEngg	0.007	0.077	0.069	-0.069	0.066	-0.037	0.047	0.07	0.1	0.23	0.016	0.016	-0.00018	1	-0.0075	-0.0013	-0.032	0.0054	0.032	-0.021	-0.057	0.017	0.048
ElectricalEngg	-0.01	0.032	0.039	-0.0043	0.042	3.9e-05	0.031	0.063	0.1	0.18	0.017	0.038	-0.0052	-0.0075	1	-0.0093	4.1e-16	0.0057	0.004	-0.0049	0.00055	0.014	0.032
TelecomEngg	0.019	0.056	0.067	-0.055	0.065	0.00018	0.064	0.07	0.066	0.087	0.032	0.081	-0.00025	-0.0013	-0.0093	1	-2.3e-16	0.015	0.027	-0.026	-0.012	0.013	0.02
CivilEngg	0.019	0.022	0.0071	0.00044	0.052	0.00042	0.02	0.0094	0.042	0.064	0.03	0.011	5.4e-16	-0.032	4.1e-16	-2.3e-16	1	6.4e-05	0.012	-0.023	-0.026	0.0047	0.033
scientiousness	0.075	0.067	0.057	0.045	0.052	-0.017	0.029	0.019	-0.0085	-0.0054	-0.0043	-0.016	-0.0027	0.0054	0.0057	0.015	6.4e-05	1	0.49	0.37	-0.32	0.4	-0.047
agreeableness	0.079	0.14	0.1	-0.043	0.061	-0.004	0.19	0.16	0.1	0.084	0.1	0.017	0.03	0.032	0.004	0.027	0.012	0.49	1	0.46	-0.21	0.59	0.075
extraversion	0.0056	-0.0078	-0.013	-0.01	-0.051	0.0078	0.0071	-0.006	-0.039	-0.034	-0.02	-0.027	0.0056	-0.021	-0.0042	-0.026	-0.023	0.37	0.46	1	-0.094	0.45	-0.0049
nueroticism	0.0063	-0.13	-0.089	0.033	-0.068	0.0011	-0.15	-0.19	-0.14	-0.11	-0.12	-0.02	-0.05	-0.057	0.00055	-0.012	-0.026	-0.32	-0.21	-0.094	1	-0.067	-0.07
to_experience	0.086	0.035	0.007	-0.029	0.014	0.019	0.074	0.05	0.025	0.023	0.029	0.0042	-0.0045	0.017	0.014	0.013	0.0047	0.4	0.59	0.45	-0.067	1	0.00094
Salary	-0.039	0.18	0.17	-0.18	0.13	-0.0093	0.18	0.19	0.24	0.19	0.19	0.047	0.057	0.048	0.032	0.02	0.033	-0.047	0.075	-0.0049	-0.07	0.00094	1

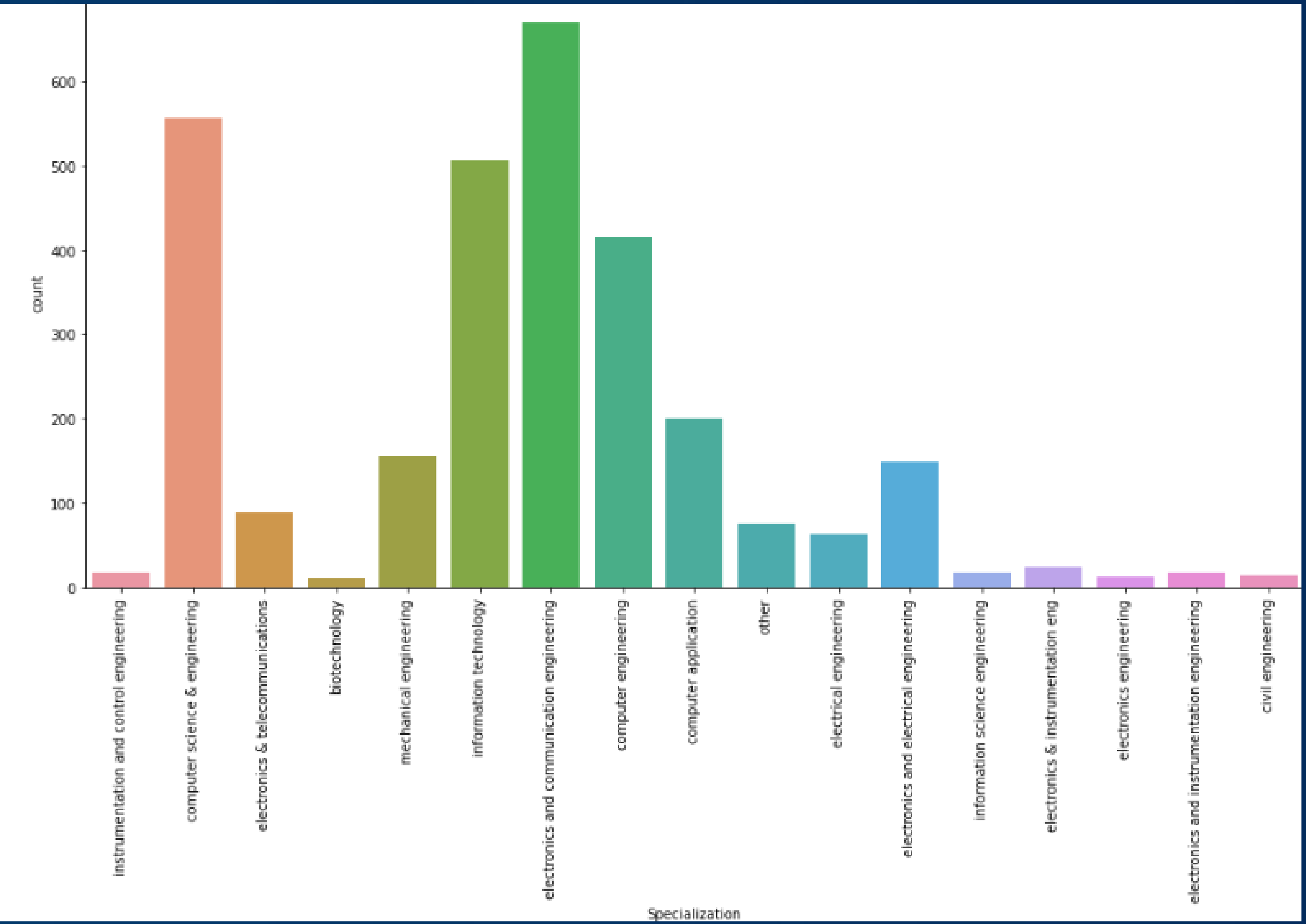
Analysis of 10th and 12th percentage by collage tier

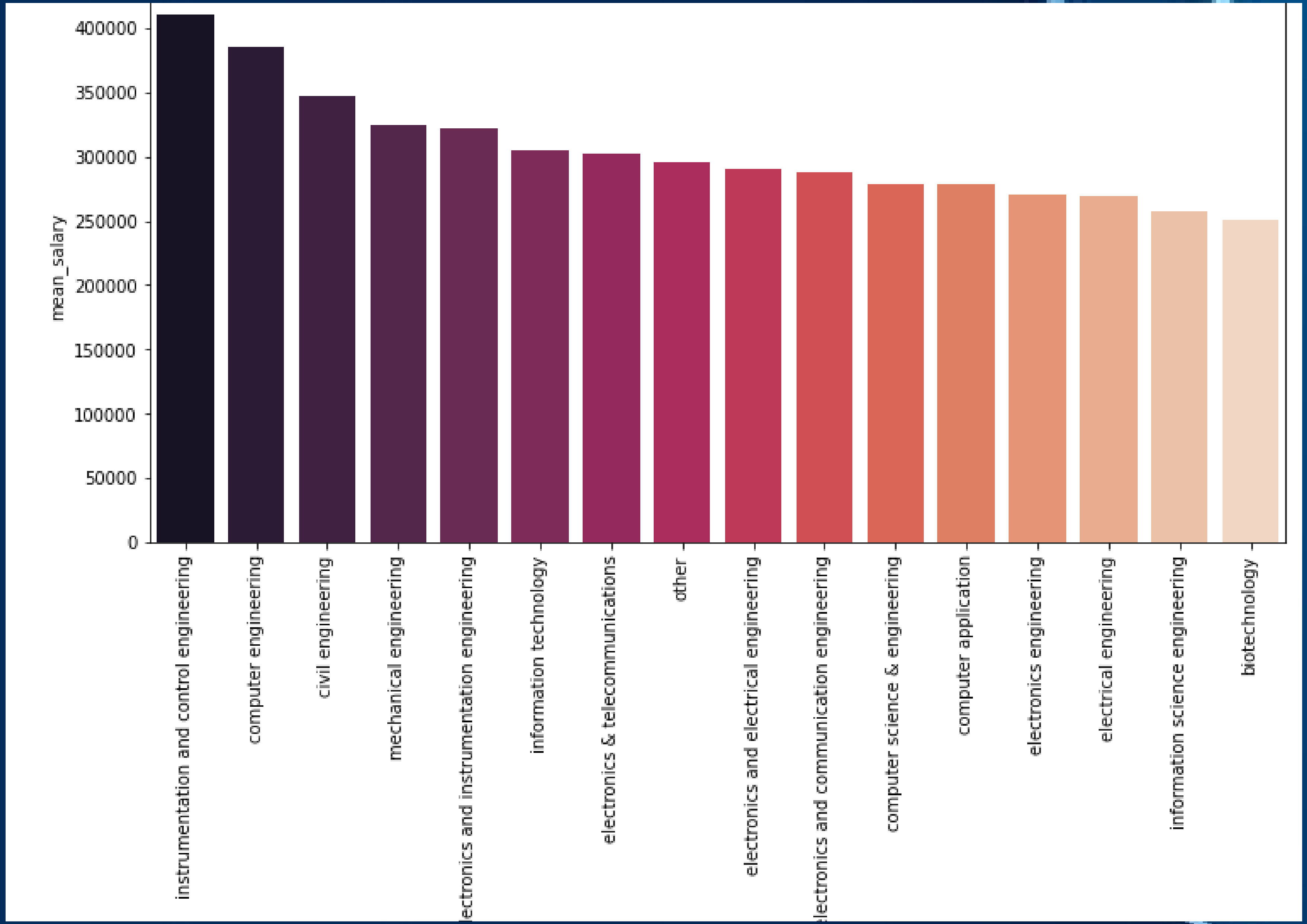


According to correlation and scatterplot we can see 10th and 12th are positively correlated and this is the case of multicollinearity.

So we decide to keep only one

Analysation of Salary with Specialisation

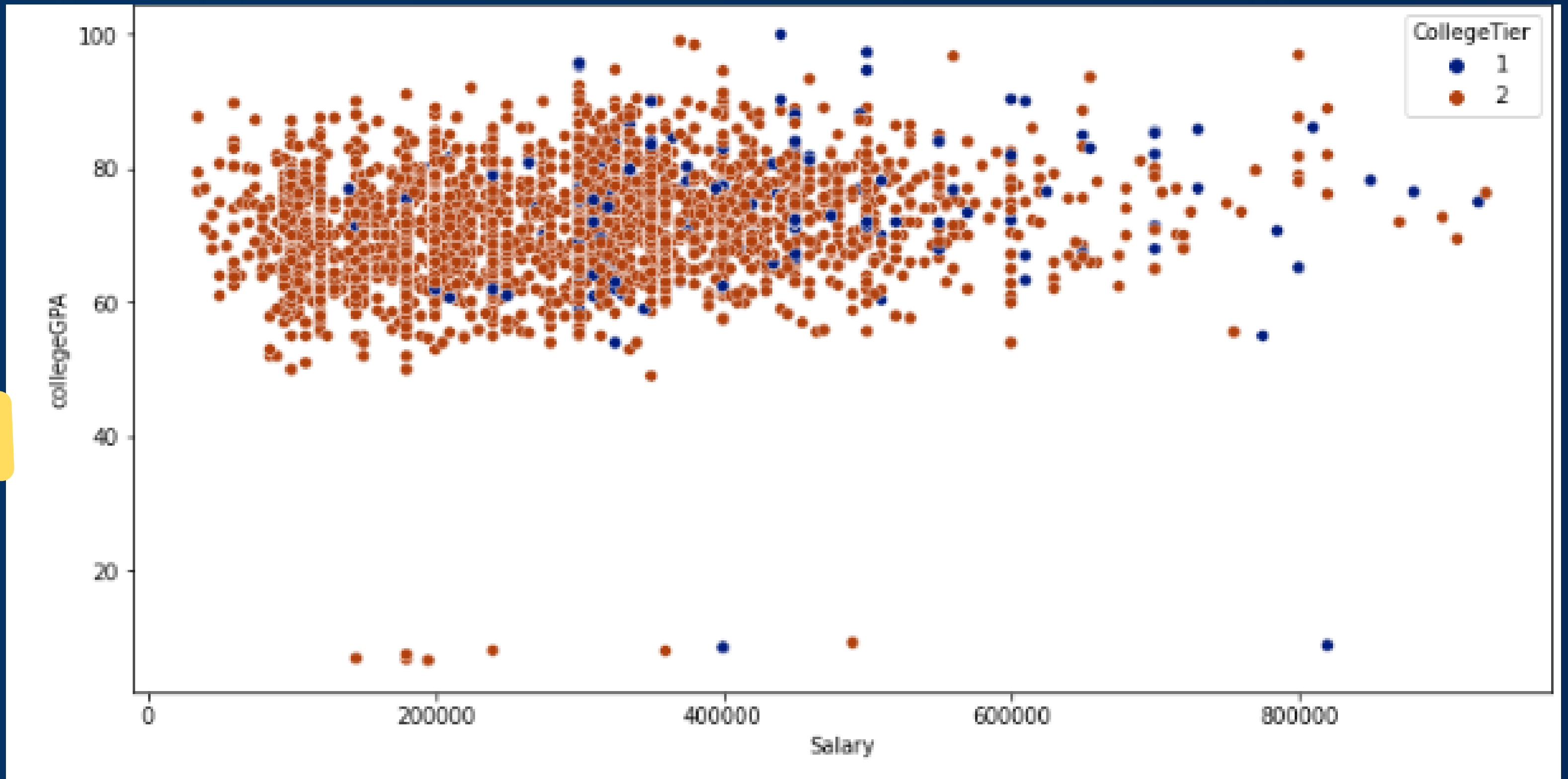




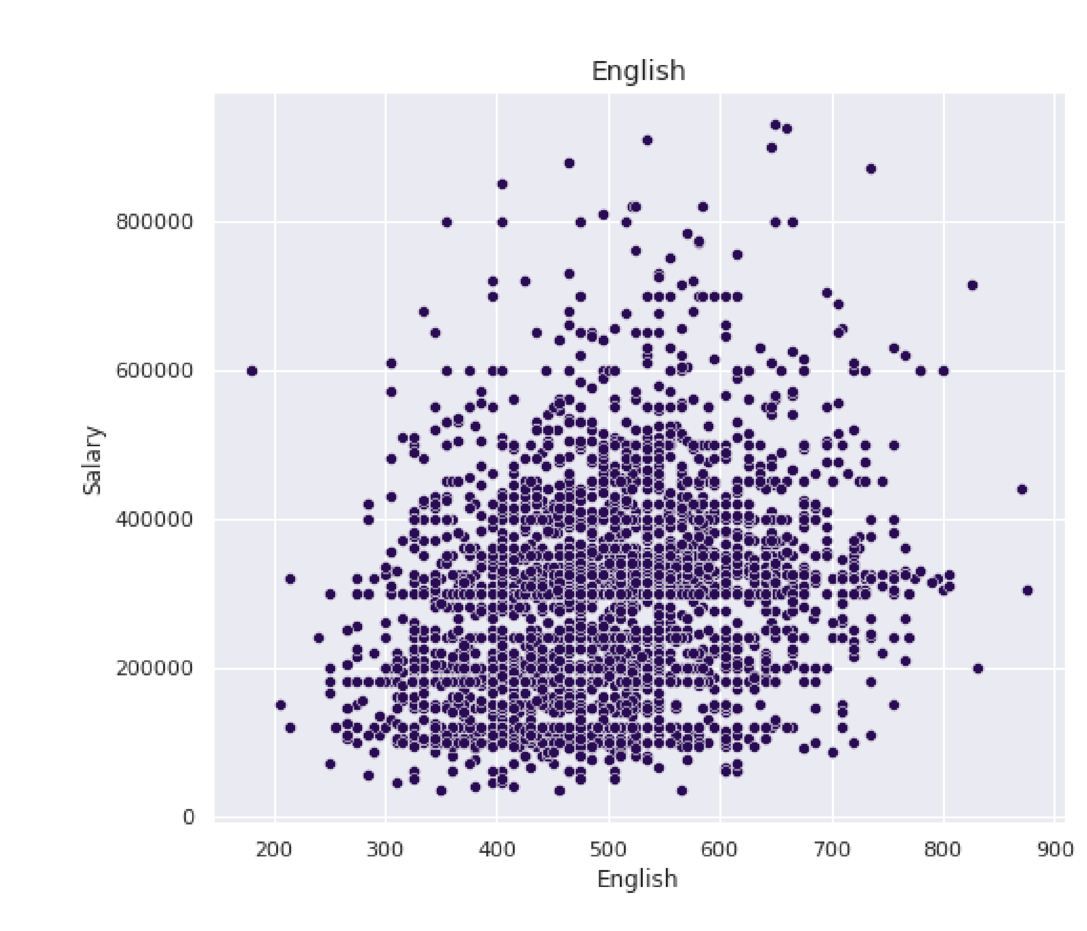
ICE Engineer, Computer Engineer and Electronics Engineer having highest mean salary

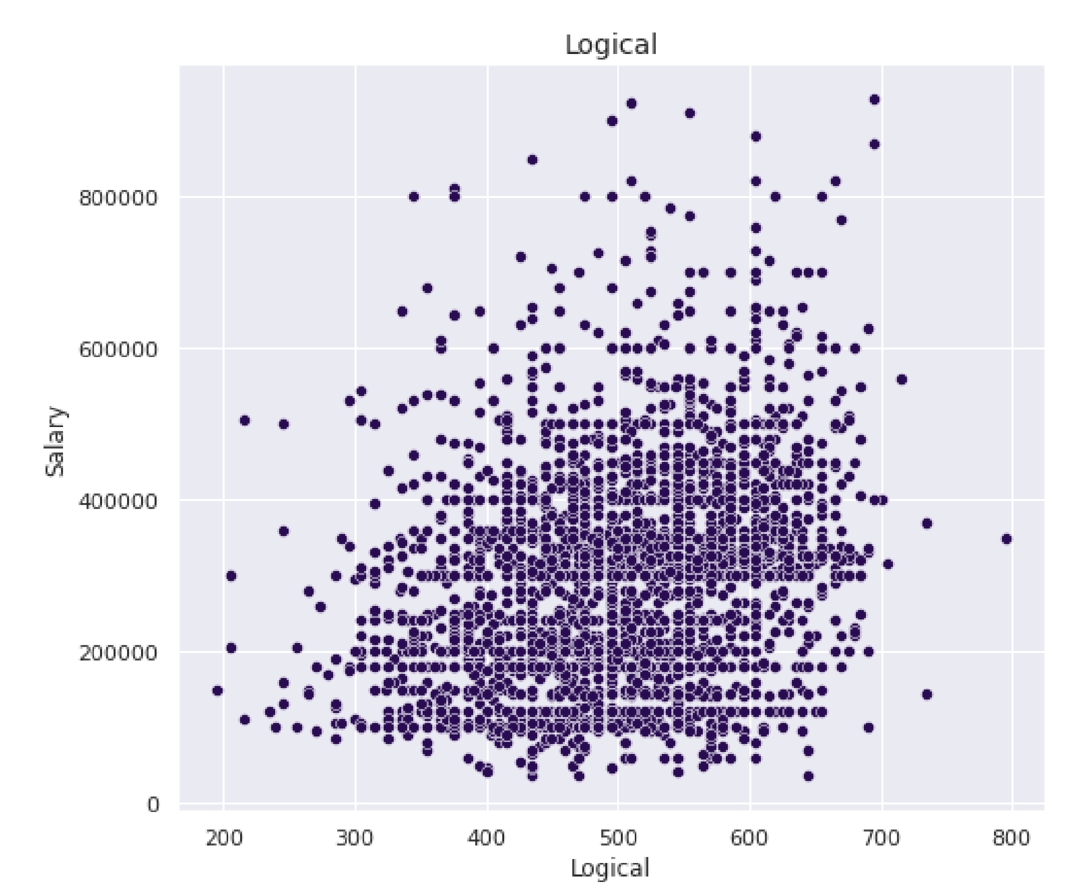
Analysis of Collage GPA and Salary

We have to remove the outliers present as we can view them in the plot .

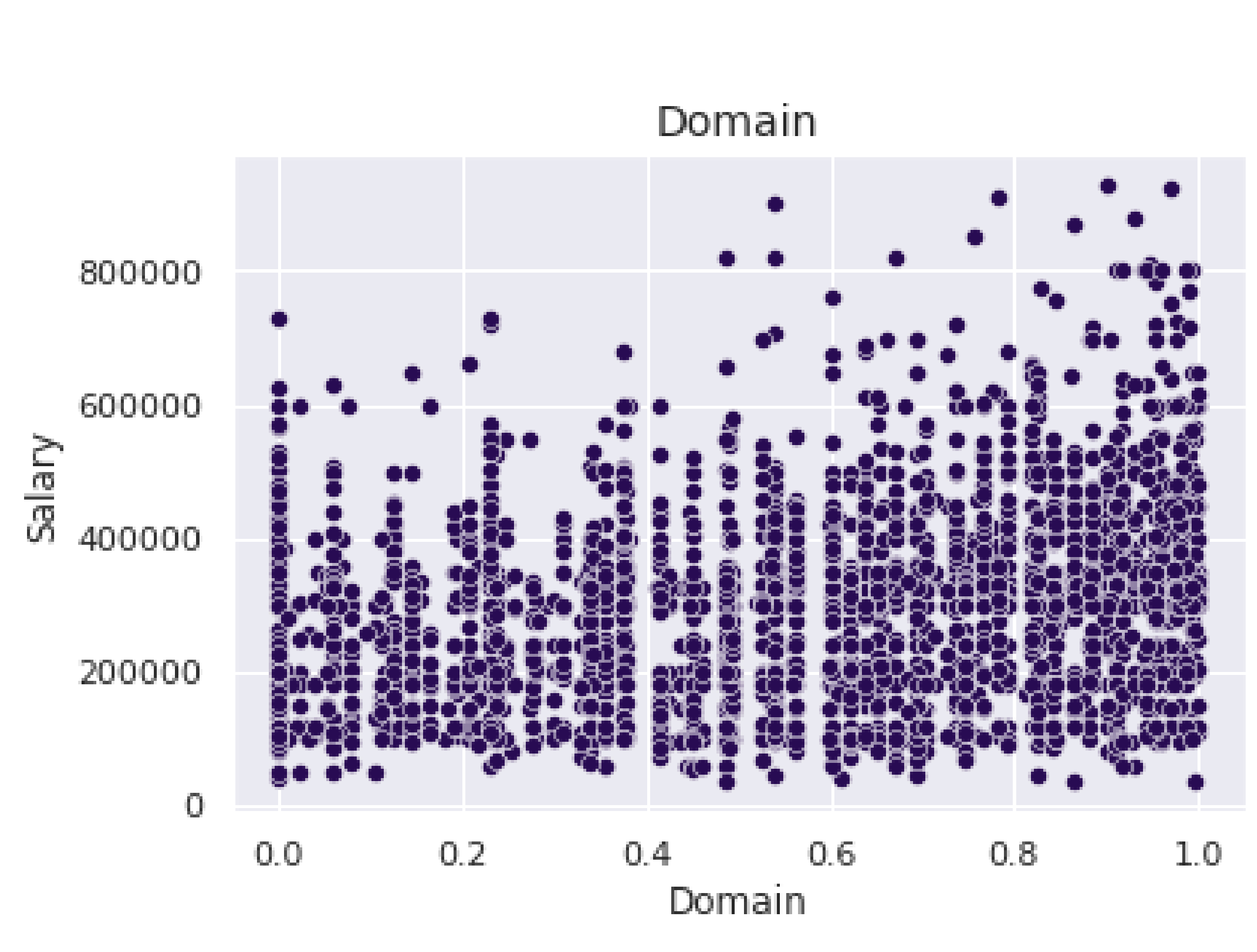


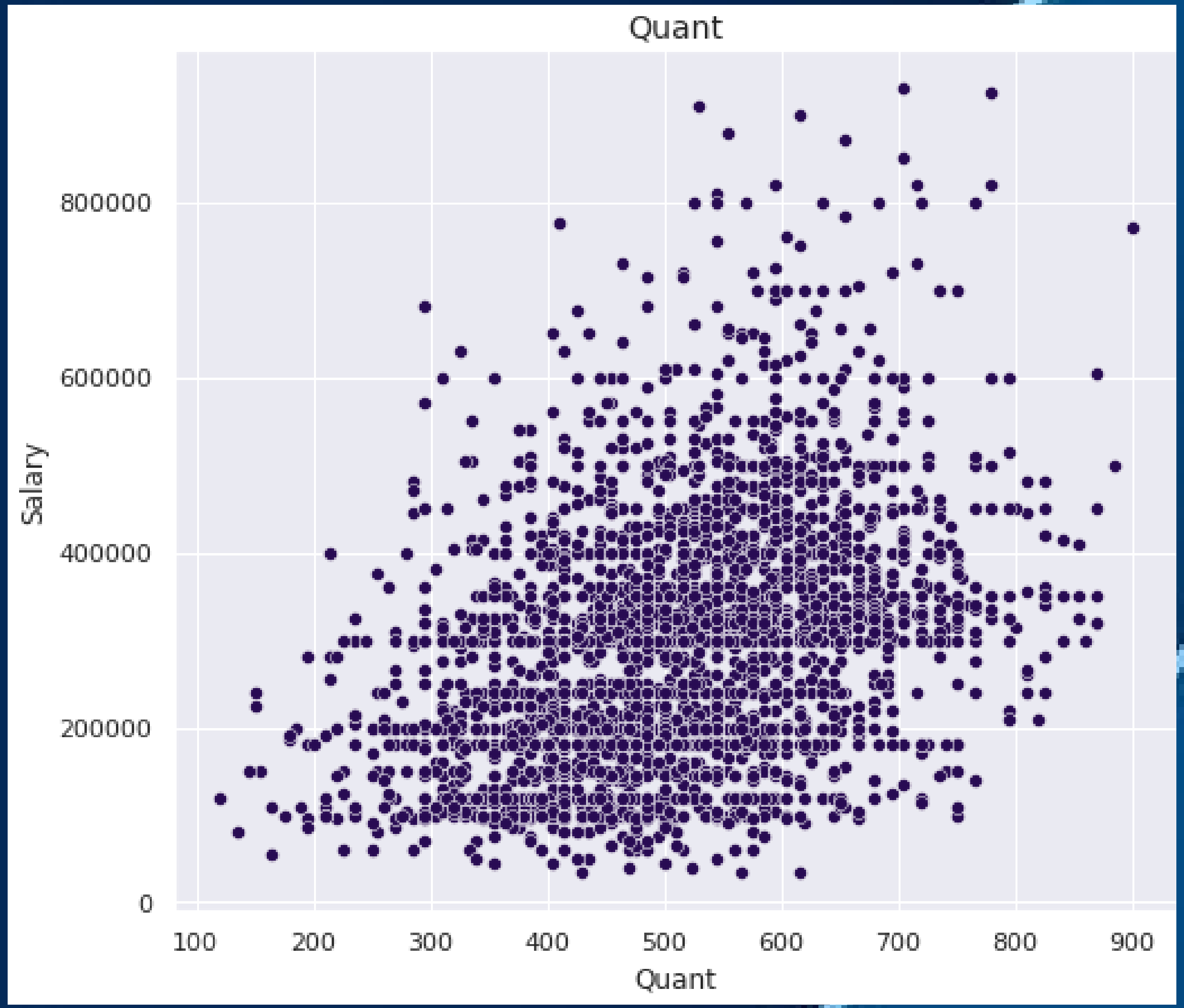
Analysis of AMCAT subjects and Salary





Here the on the basis of the above data visualised we observe that gender,degree, collage GPA,Specialisation are useful for analysing meaningful insight for the variable salary.



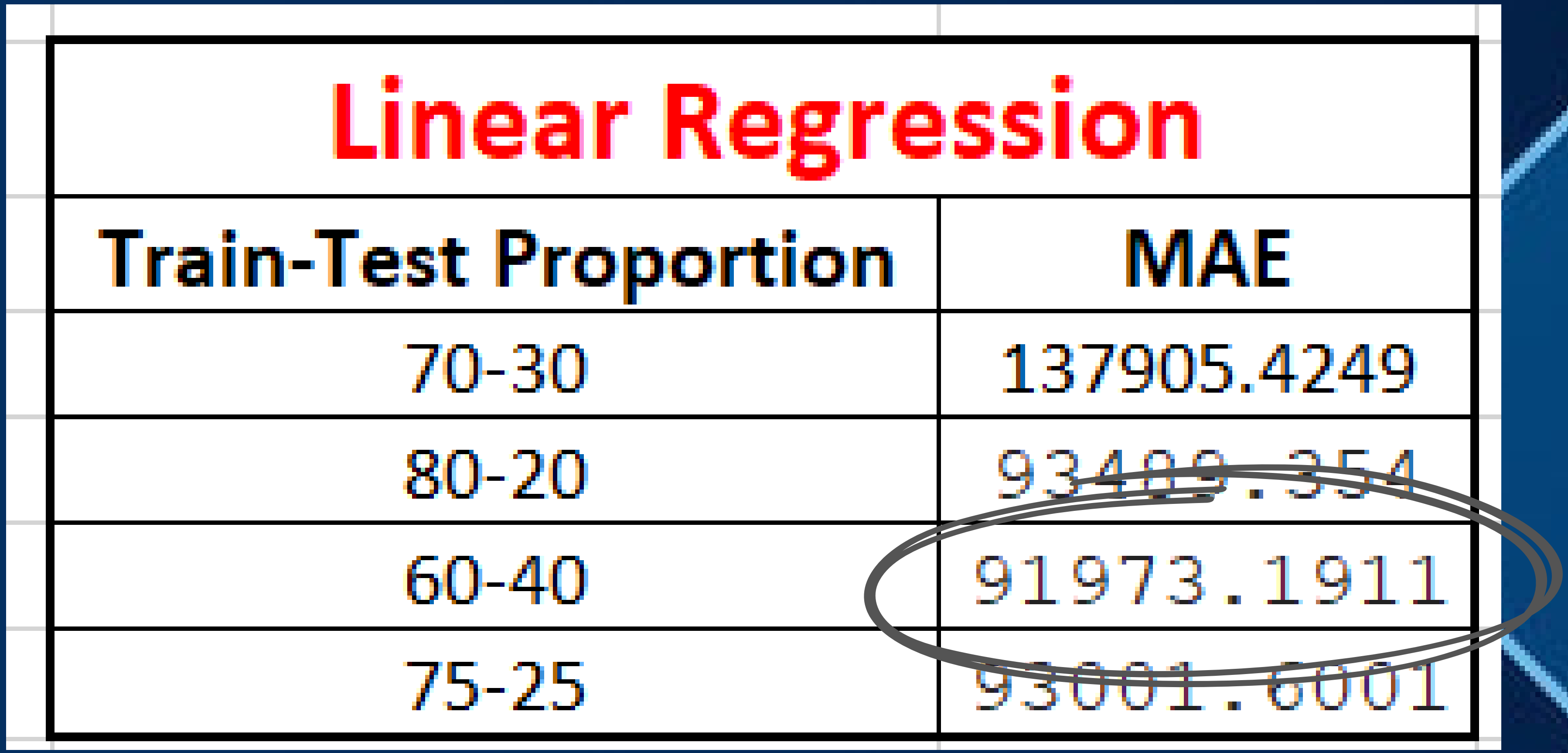


We have visulaised the Salary relation with AMCAT subjects through scatterplot above .
As AMCAT exams are conducted as an entrance for the jobs and it plays a key role but we found from the above graphs there is no impact on the the salary of the graduate.

Linear regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable.

- For 60-40 ratio we got the least MAE value i.e: 91973.1911

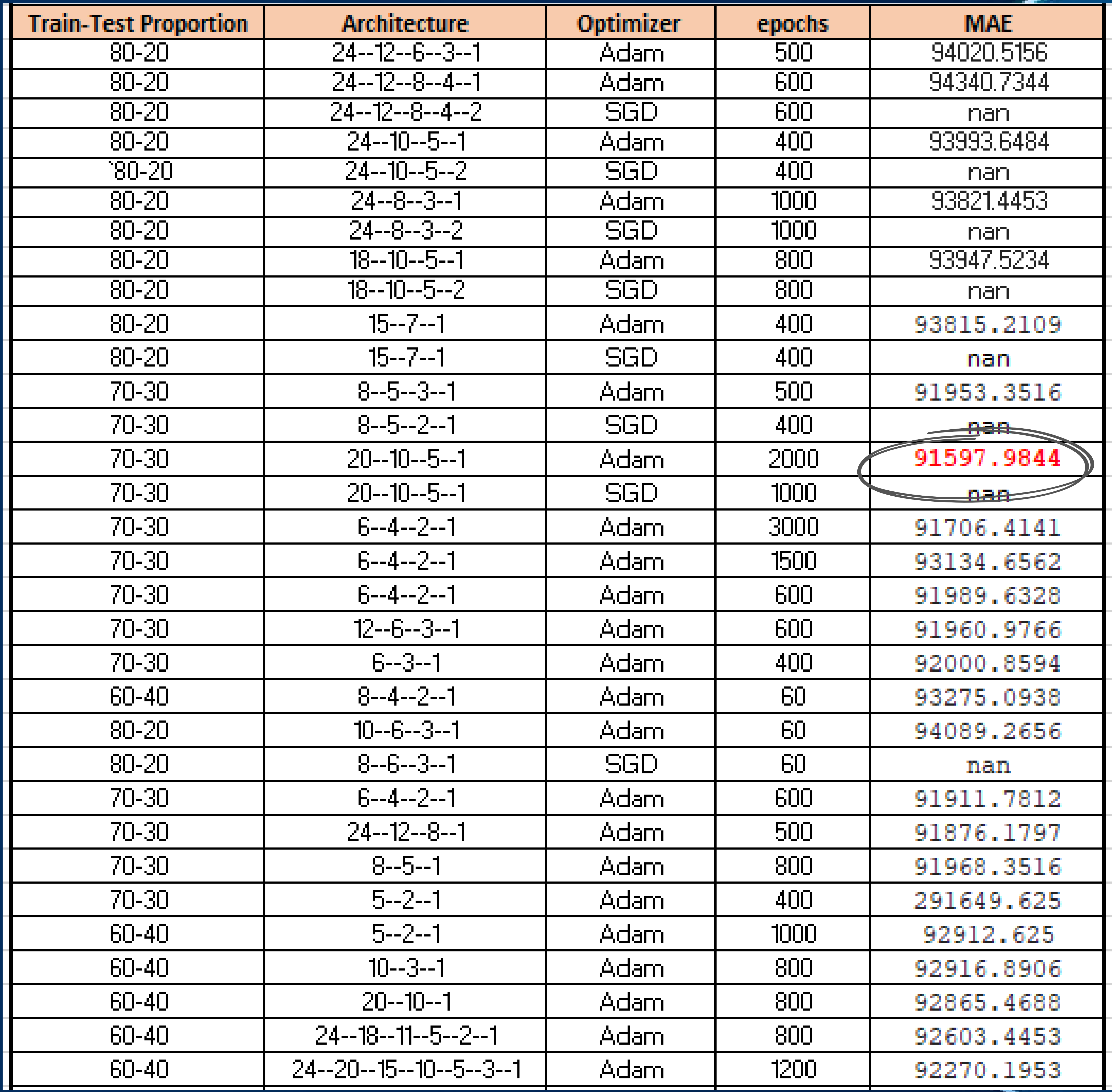


Linear Regression	
Train-Test Proportion	MAE
70-30	137905.4249
80-20	93409.354
60-40	91973.1911
75-25	93001.6001

Neural Networks

In the neural networks we are working with respect to two optimisers : Adam ,SGD So we applied it with various experiments and architectures

We find for 70-30 ratioat 20--10--5--1 architecture and Adam as optimiser at 2000epoch we have least MAE value: 91597.9844

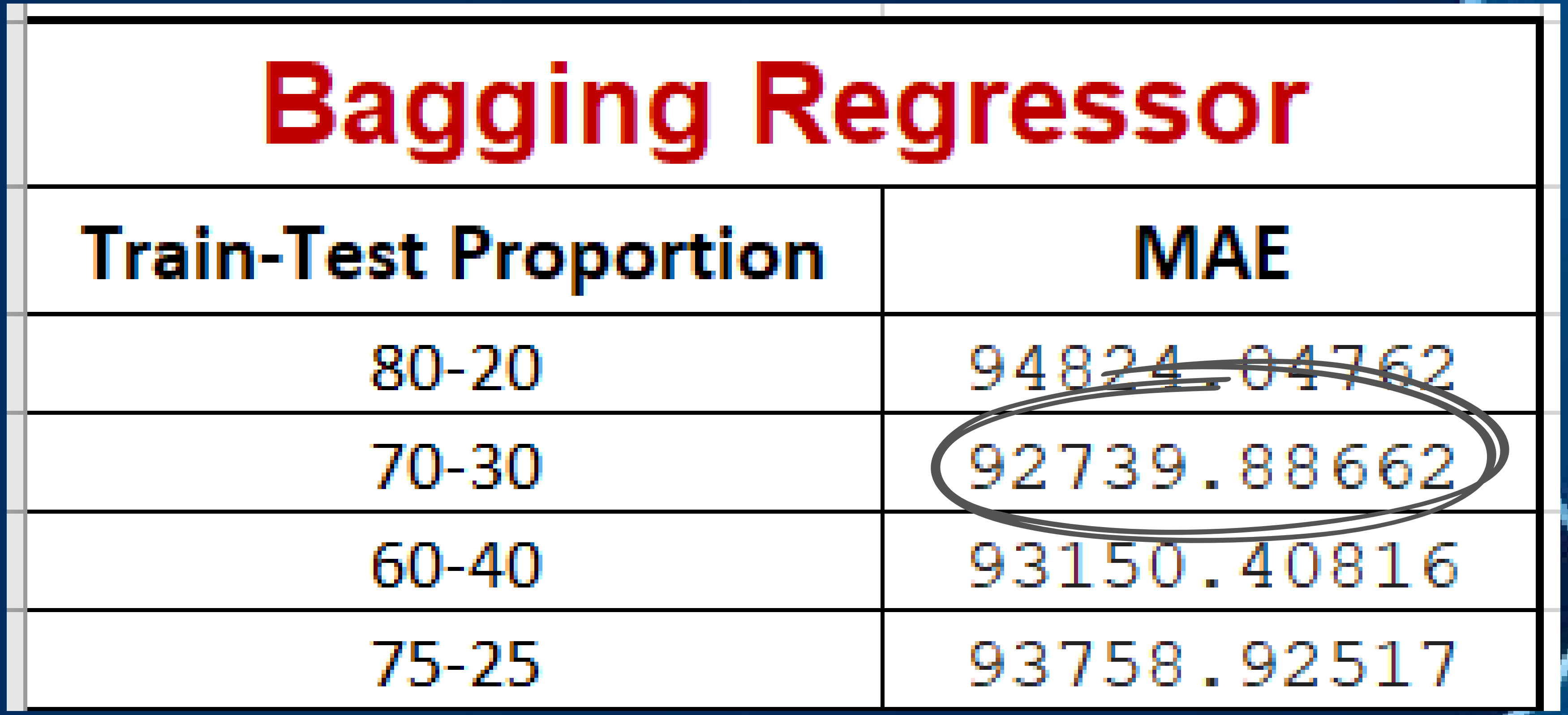


Train-Test Proportion	Architecture	Optimizer	epochs	MAE
80-20	24--12--6--3--1	Adam	500	94020.5156
80-20	24--12--8--4--1	Adam	600	94340.7344
80-20	24--12--8--4--2	SGD	600	nan
80-20	24--10--5--1	Adam	400	93993.6484
`80-20	24--10--5--2	SGD	400	nan
80-20	24--8--3--1	Adam	1000	93821.4453
80-20	24--8--3--2	SGD	1000	nan
80-20	18--10--5--1	Adam	800	93947.5234
80-20	18--10--5--2	SGD	800	nan
80-20	15--7--1	Adam	400	93815.2109
80-20	15--7--1	SGD	400	nan
70-30	8--5--3--1	Adam	500	91953.3516
70-30	8--5--2--1	SGD	400	nan
70-30	20--10--5--1	Adam	2000	91597.9844
70-30	20--10--5--1	SGD	1000	nan
70-30	6--4--2--1	Adam	3000	91706.4141
70-30	6--4--2--1	Adam	1500	93134.6562
70-30	6--4--2--1	Adam	600	91989.6328
70-30	12--6--3--1	Adam	600	91960.9766
70-30	6--3--1	Adam	400	92000.8594
60-40	8--4--2--1	Adam	60	93275.0938
80-20	10--6--3--1	Adam	60	94089.2656
80-20	8--6--3--1	SGD	60	nan
70-30	6--4--2--1	Adam	600	91911.7812
70-30	24--12--8--1	Adam	500	91876.1797
70-30	8--5--1	Adam	800	91968.3516
70-30	5--2--1	Adam	400	291649.625
60-40	5--2--1	Adam	1000	92912.625
60-40	10--3--1	Adam	800	92916.8906
60-40	20--10--1	Adam	800	92865.4688
60-40	24--18--11--5--2--1	Adam	800	92603.4453
60-40	24--20--15--10--5--3--1	Adam	1200	92270.1953

Bagging

Bagging Regressor

For 70-30 ratio we got the least MAE value i.e: 92739.88662

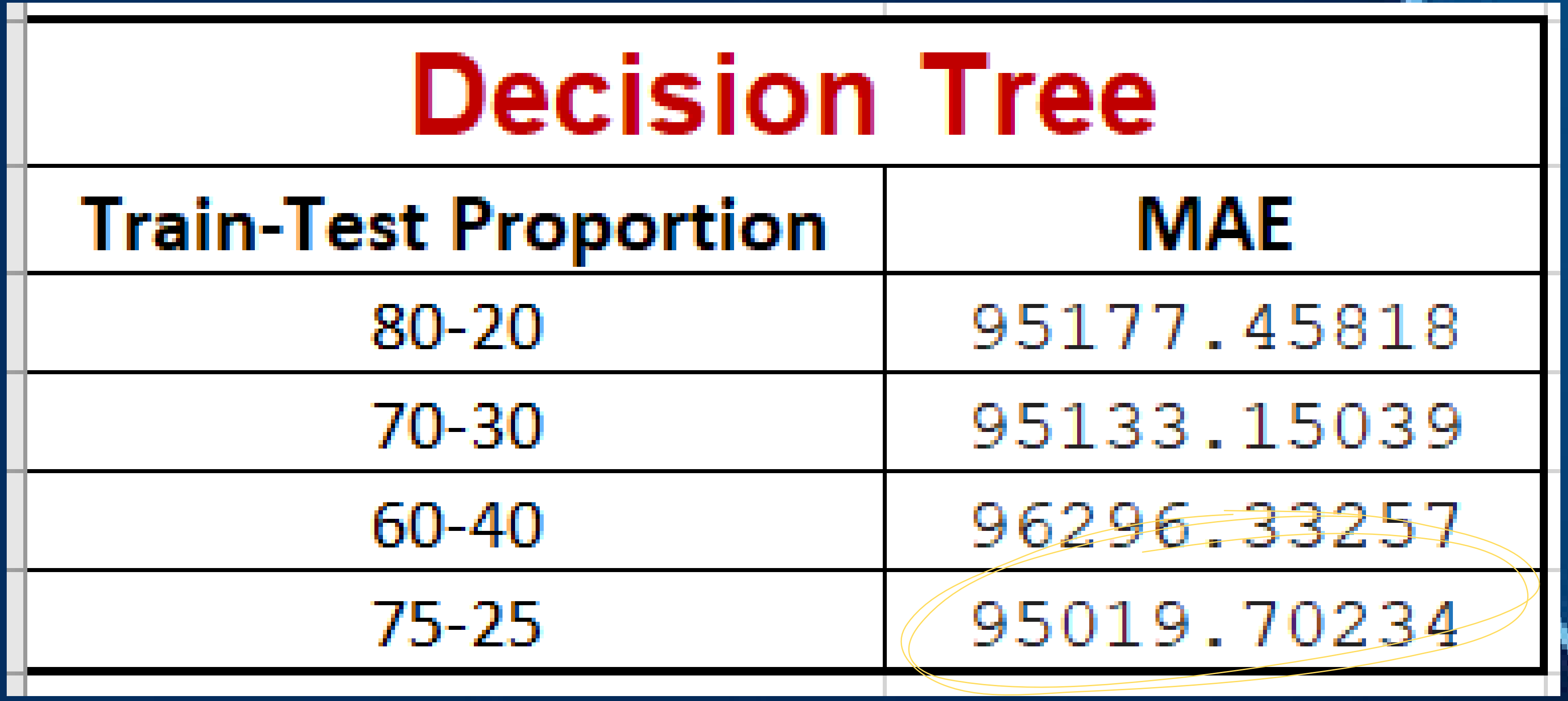


Bagging Regressor

Train-Test Proportion	MAE
80-20	94824.04762
70-30	92739.88662
60-40	93150.40816
75-25	93758.92517

Decision Tree

For 75-25 ratio we got the least MAE value i.e: 95019.70234

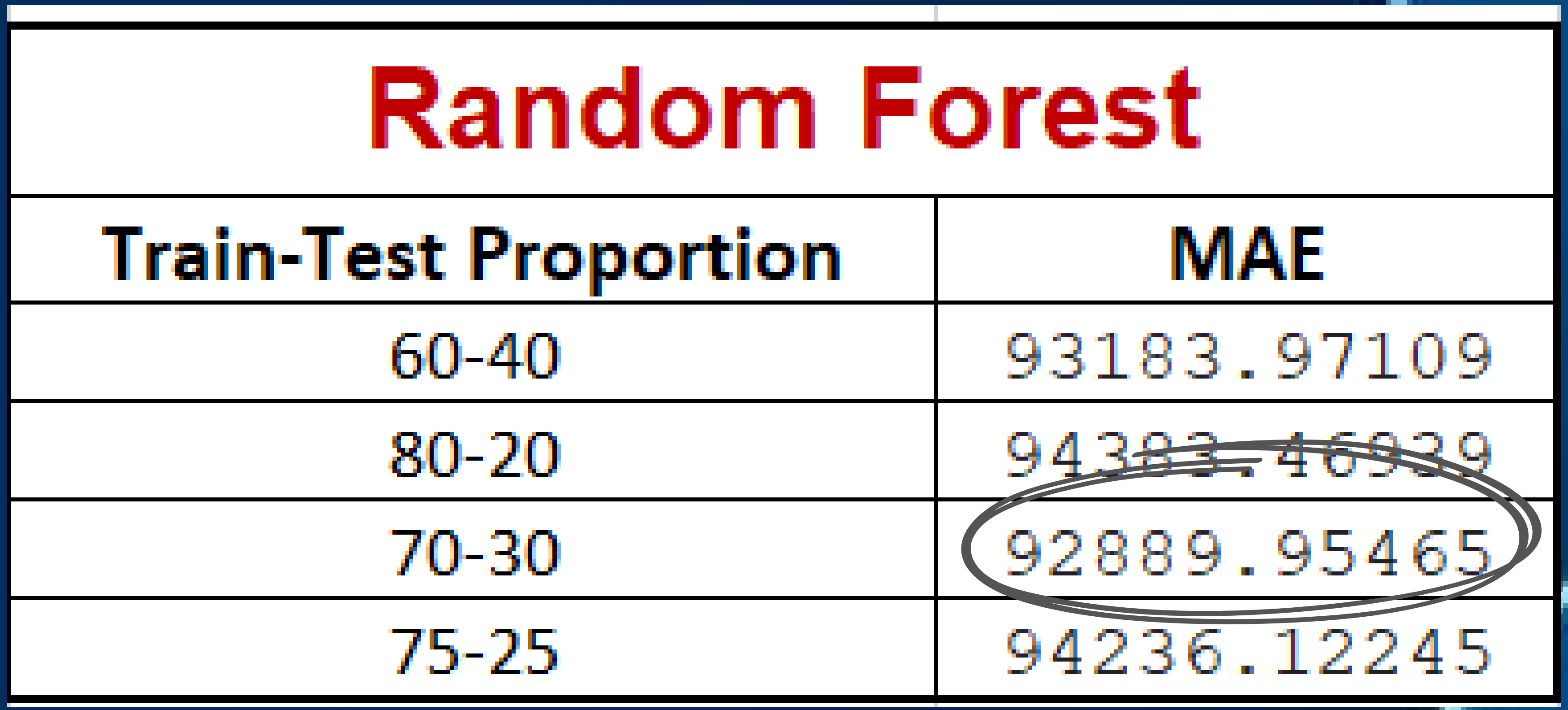


Decision Tree

Train-Test Proportion	MAE
80-20	95177.45818
70-30	95133.15039
60-40	96296.33257
75-25	95019.70234

Random Forest

For 70-30 ratio we got the least MAE value i.e: 92889.95465

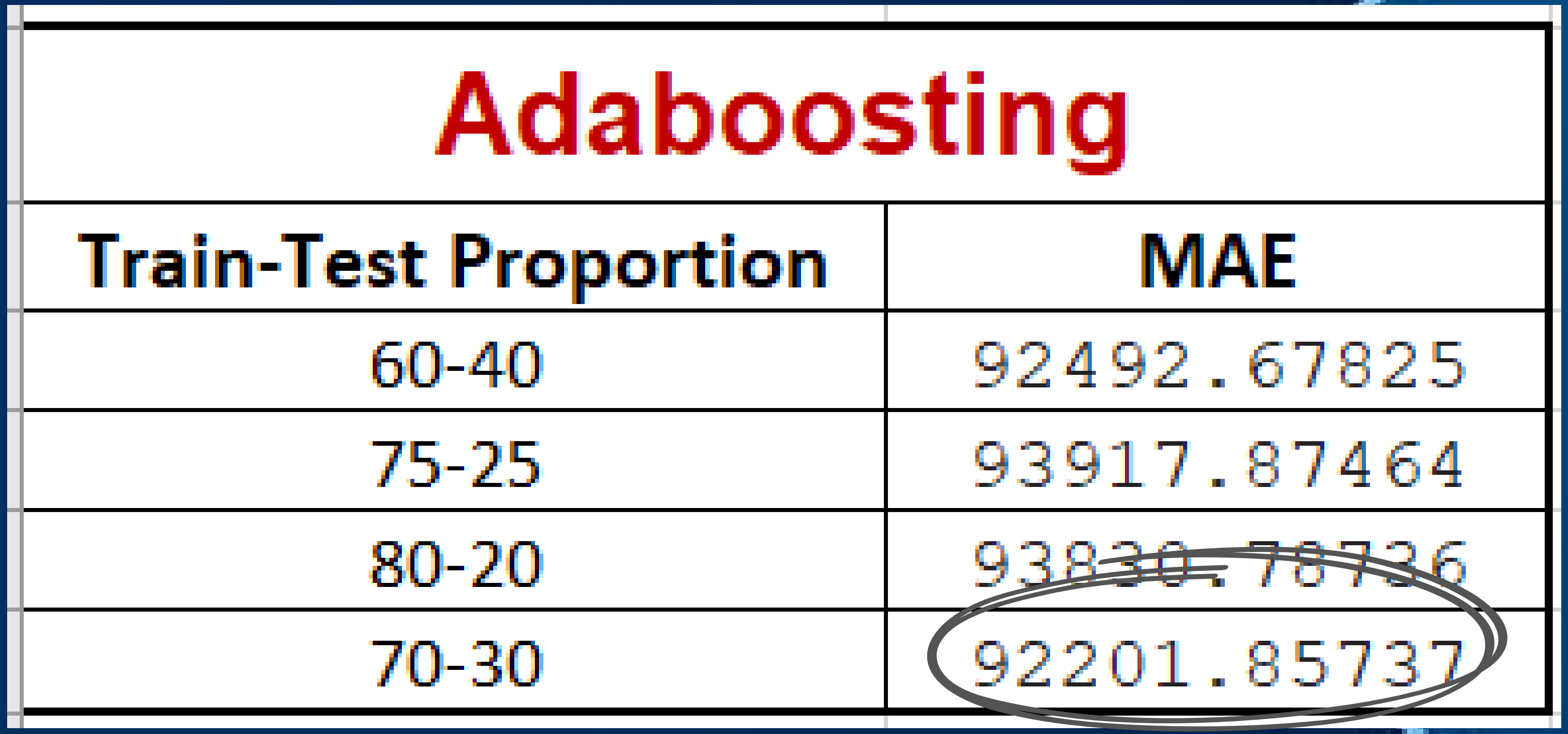


Random Forest	
Train-Test Proportion	**MAE**
60-40	93183.97109
80-20	94383.46939
70-30	92889.95465
75-25	94236.12245

Boosting

Adaboost

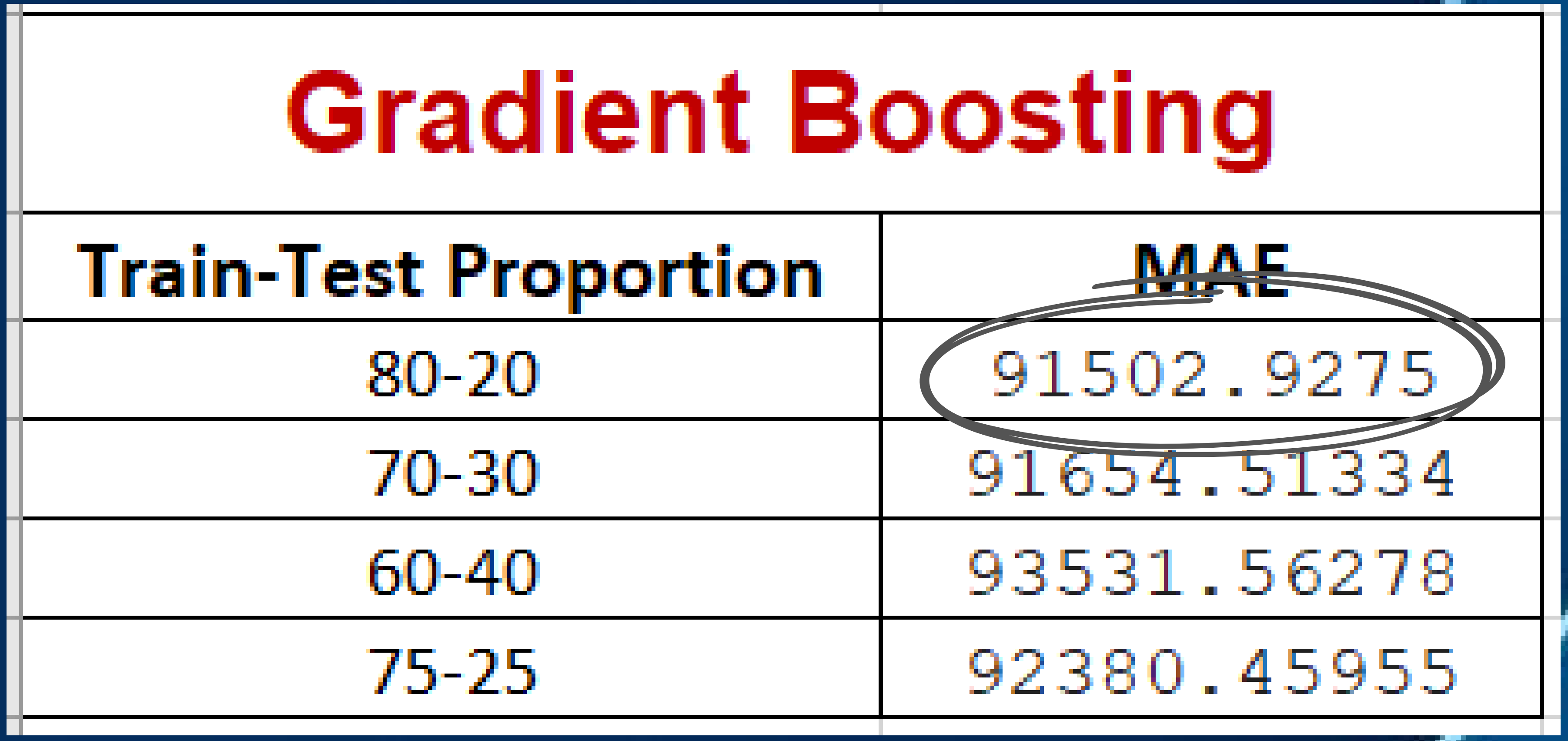
For 70-30 ratio we got the least MAE value i.e: 92201.85737



Adaboosting	
Train-Test Proportion	**MAE**
60-40	92492.67825
75-25	93917.87464
80-20	93830.78736
70-30	92201.85737

Gradient Boost

For 80-20 ratio we got the least MAE value i.e: 91502.9275

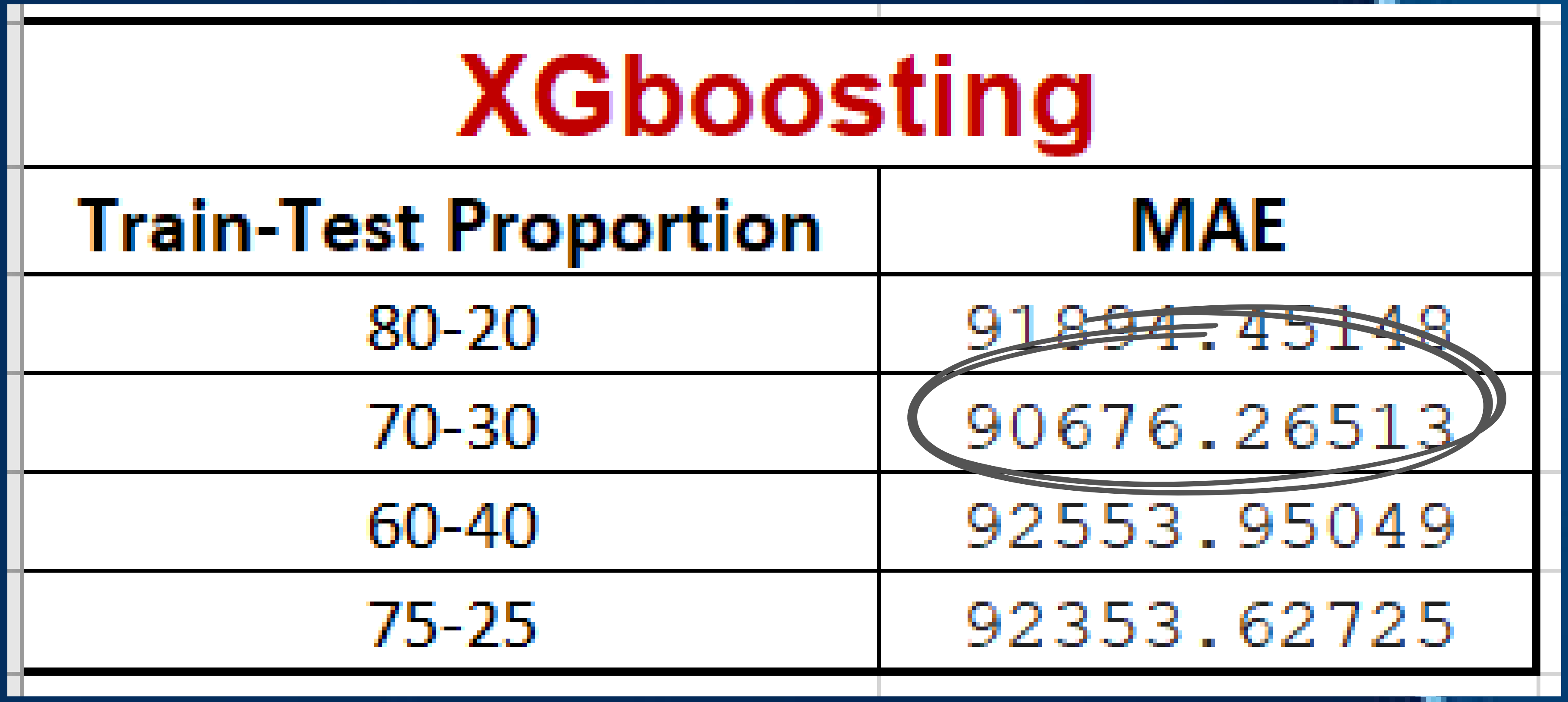


Gradient Boosting

Train-Test Proportion	MAE
80-20	91502.9275
70-30	91654.51334
60-40	93531.56278
75-25	92380.45955

XGboost

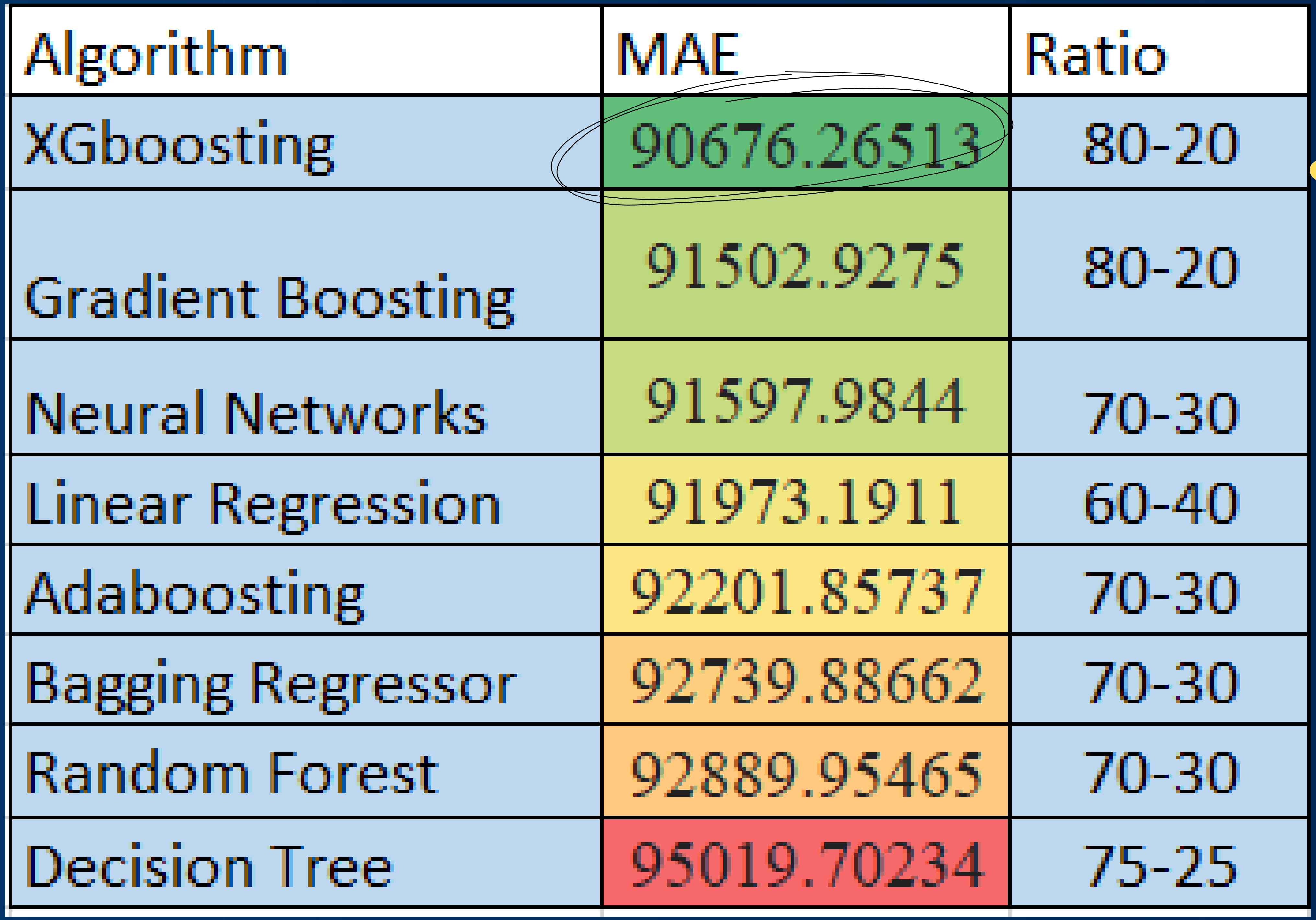
For 70-30 ratio we got the least MAE value i.e: 90676.26513



XGboosting

Train-Test Proportion	MAE
80-20	91894.45148
70-30	90676.26513
60-40	92553.95049
75-25	92353.62725

Comparision Table



Algorithm	MAE	Ratio
XGboosting	90676.26513	80-20
Gradient Boosting	91502.9275	80-20
Neural Networks	91597.9844	70-30
Linear Regression	91973.1911	60-40
Adaboosting	92201.85737	70-30
Bagging Regressor	92739.88662	70-30
Random Forest	92889.95465	70-30
Decision Tree	95019.70234	75-25

Conclusion

The Engineering Graduate salary dataset is a regression based data. So we have applied all the algorithms of machine learning with respect to regression such as linear regression, Neural Network, Decision tree, Random forest and Boosting.

Among all the algorithms in boosting XGboost techinque with 80-20 performed well an gave least MAE value when comapred to others.

So we can conclude boosting algorithm fits good and can be used for further usage of model .

Team members and their Roles

- N.Bhavana Reddy-Coding,PPT ,EDA
- UudhhayKiirran-Coding,PPT
- Sai Prasanna-PPT
- Jashwanth-PPT

Click on the icon of github and colab for
more details of the project .

Thank
You