ENGINEERING GRADUATE SALARY PREDICTION

SUBMITTING TO: **MR. MAYUR DEV SEWAK**

GENERAL MANAGER, OPERATIONS

EISYSTEMS SERVICES

MS. MALIKA SRIVASTAVA

TRAINER, DATA SCIENCE & ANALYTICS DOMAIN**,**

EISYSTEMS SERVICES

SUBMITTED BY: RAKHI KUMARI

CONTENT TABLE

S.no	Title	Page no.
1.	Abstract of project	3
2.	Project Summary	4
3.	Objective of project	5
4.	Details Of Process	6
5	System Requirements	7
6.	Input/Output Datasets / Screenshots	8-23
7.	End Page	24

ABSTRACT OF PROJECT

" THE INCREASE IN TECHNOLOGY OVER THE PAST FEW YEARS HAS INCREASED THE NUMBER OF STUDENTS PURSUING ENGINEERING THESE DAYS . EVERY ENGINEER GRADUATE ASPIRES A DECENT PLACEMENT . SINCE THE CROWD IN THIS FIELD HAS TREMENDOUSLY INCREASED , THIS PROJECT PREDICTS THE SALARY OF A ENGINEER GRADUATE BASED ON SOME FEATURES, AND I WOULD LIKE TO EXPRESS MY GRATITUDE TO MY MENTOR MS. MALIKA MAM TO SUPPORT ME THROUGH THIS ENTIRE LEARNING PROCESS".

PROJECT SUMMARY

We all are aware of the job scenario and salary that an engineering student in India get just after graduating. But we all are not aware of what are those different factors that affect the salary of Indian Engineering graduates. This project predicts the salary of an engineer based on parameters like percentage of marks scored in class 10^{th} and 12^{th}, college tier , marks scored in different subjects, overall gpa , logical reasoning and year of graduation. This project includes an ML model which uses different algorithm to predict the salary of the graduates.

OBJECTIVE OF PROJECT

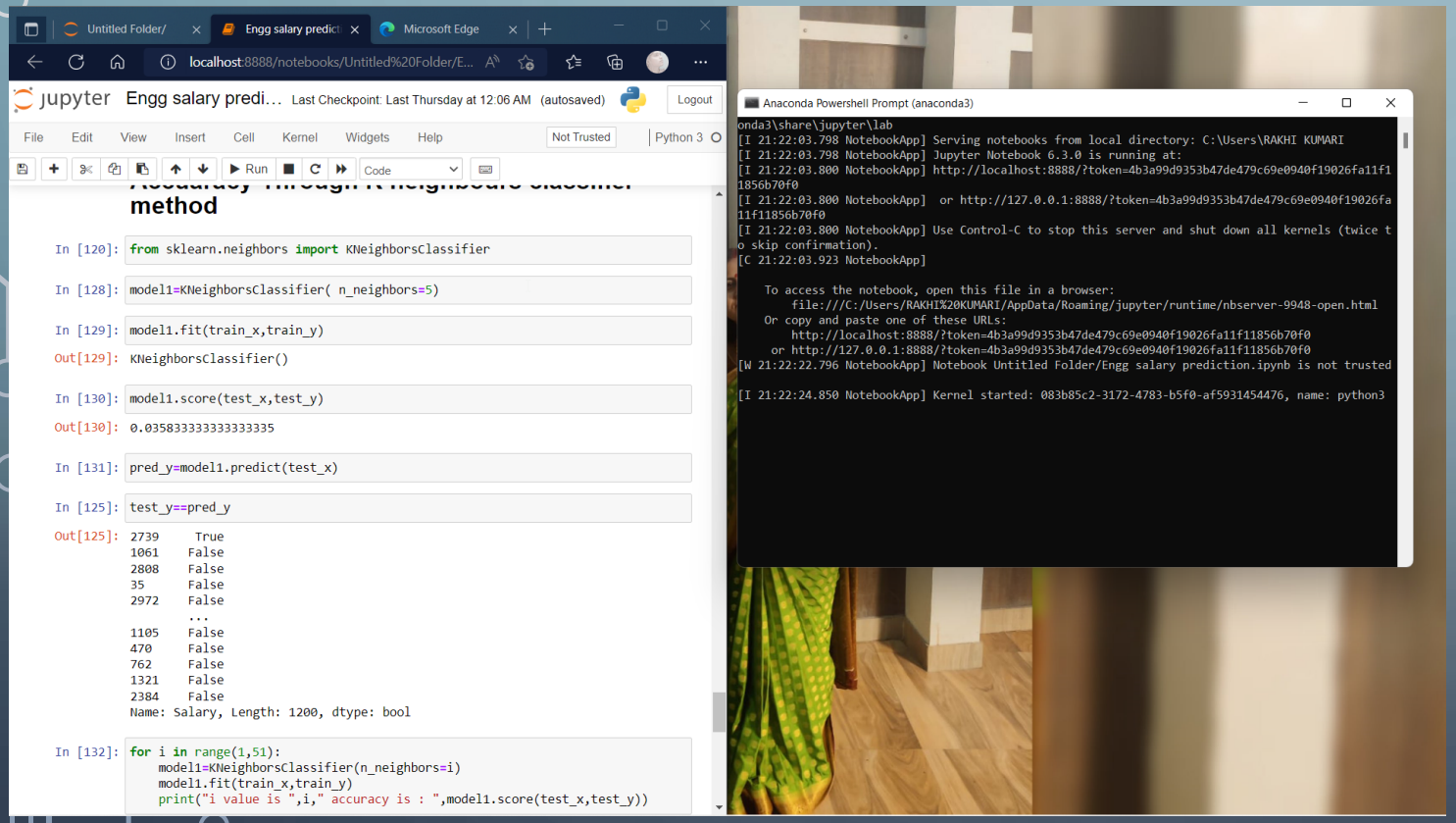
With the increase in uncertainty of job scenario of an engineering graduate in India and majority of students opting this field for higher education and being an undergraduate engineer I'm determined to make this project model which predicts the salary of engineers based on their performance while pursuing this course. Implementing the knowledge I acquired in this due course of time I have made a model which predicts the salary of a engineer based on some features described on the next slide.

DETAILS OF PROCESS

This model includes the following processes:

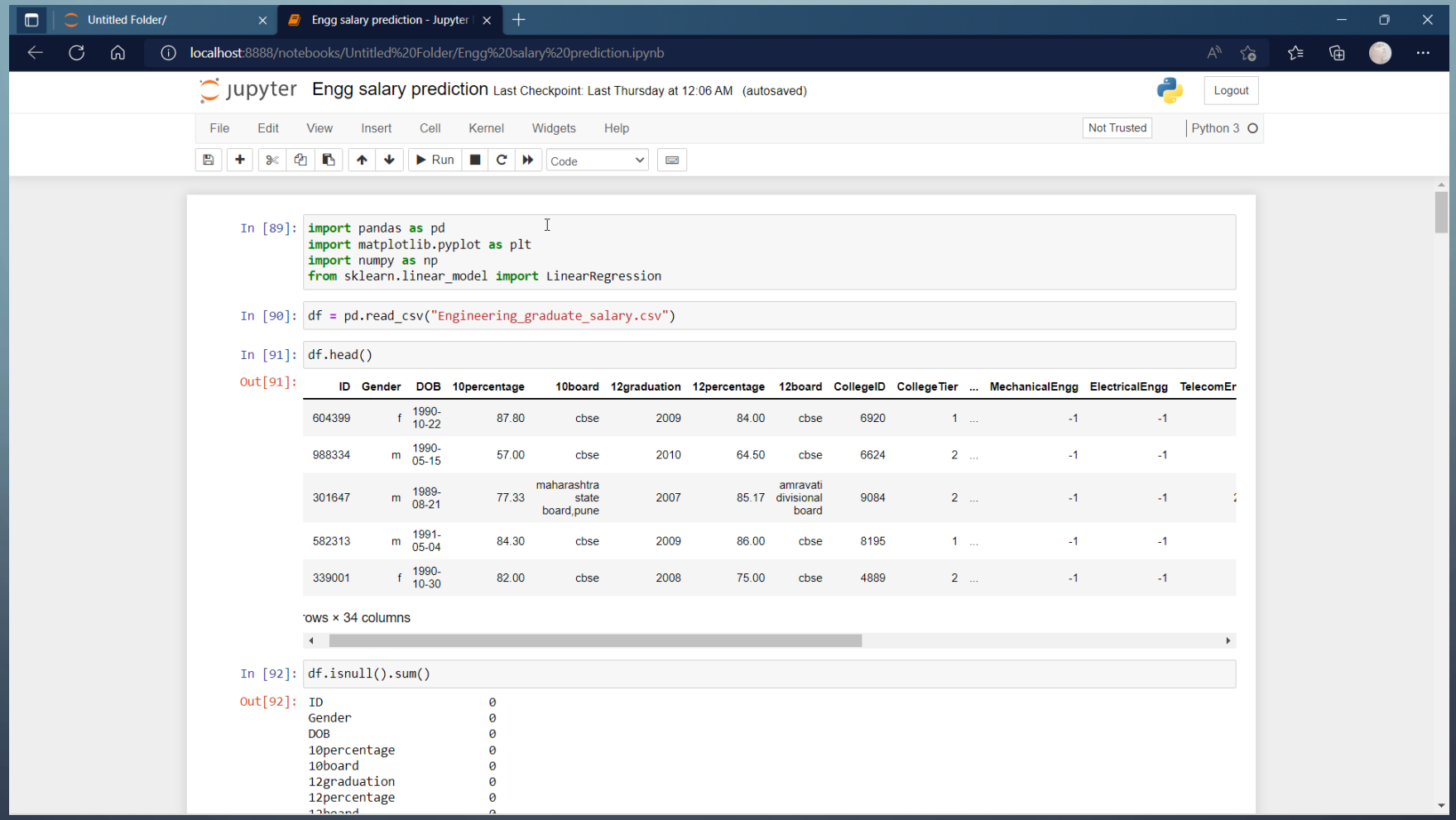
- Importing dataset and python libraries like pandas, matplotlib , NumPy ,sklearn , linear regression.
- Checking for null values and data cleaning.
- Data visualization including plotting of bar graphs, scatter plot.
- Predicting salary through Multi Linear Regression and training the model
- Predicting the salary through K Neighbors Classifier and training the model.

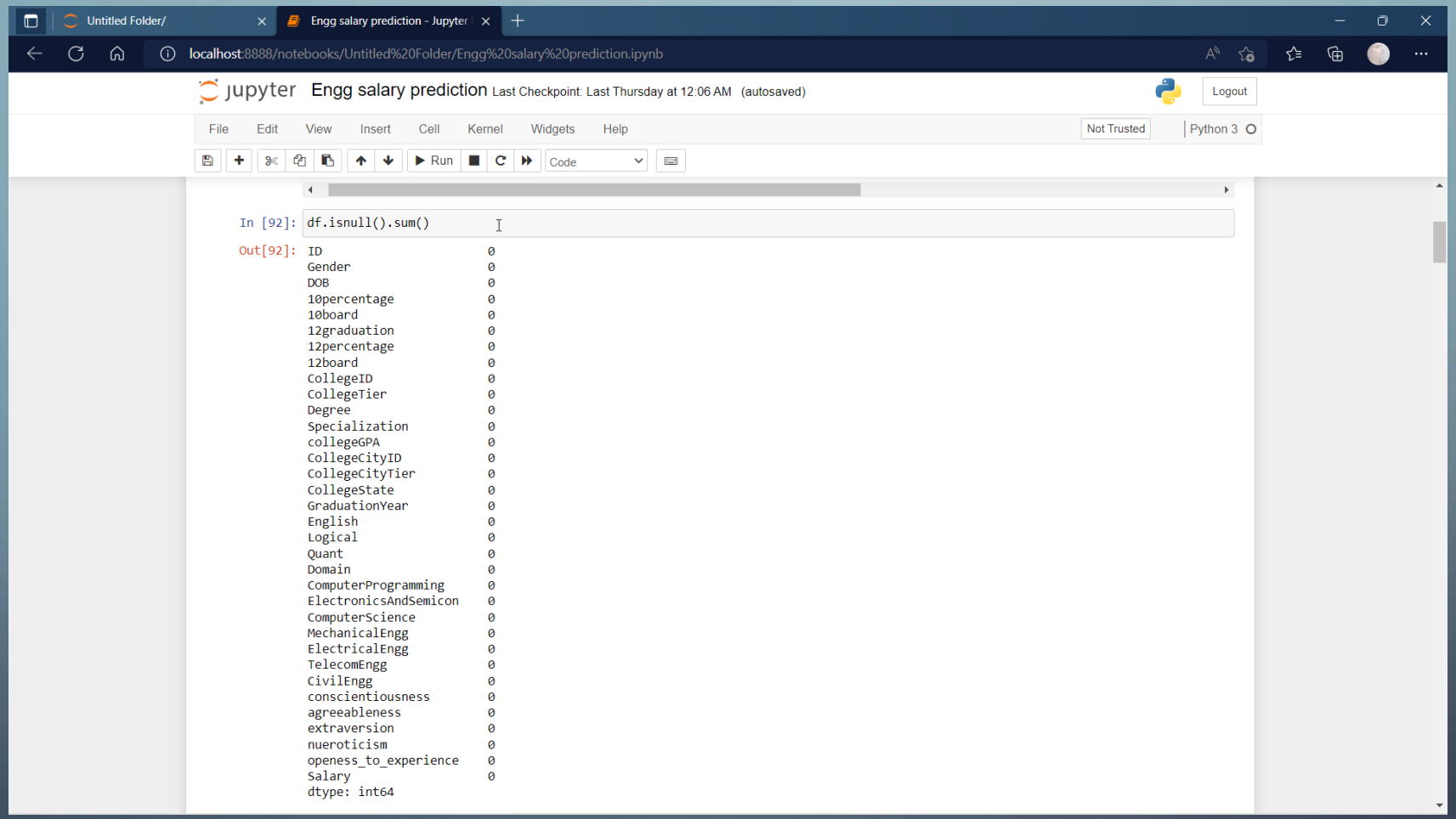
SYSTEM REQUIREMENTS

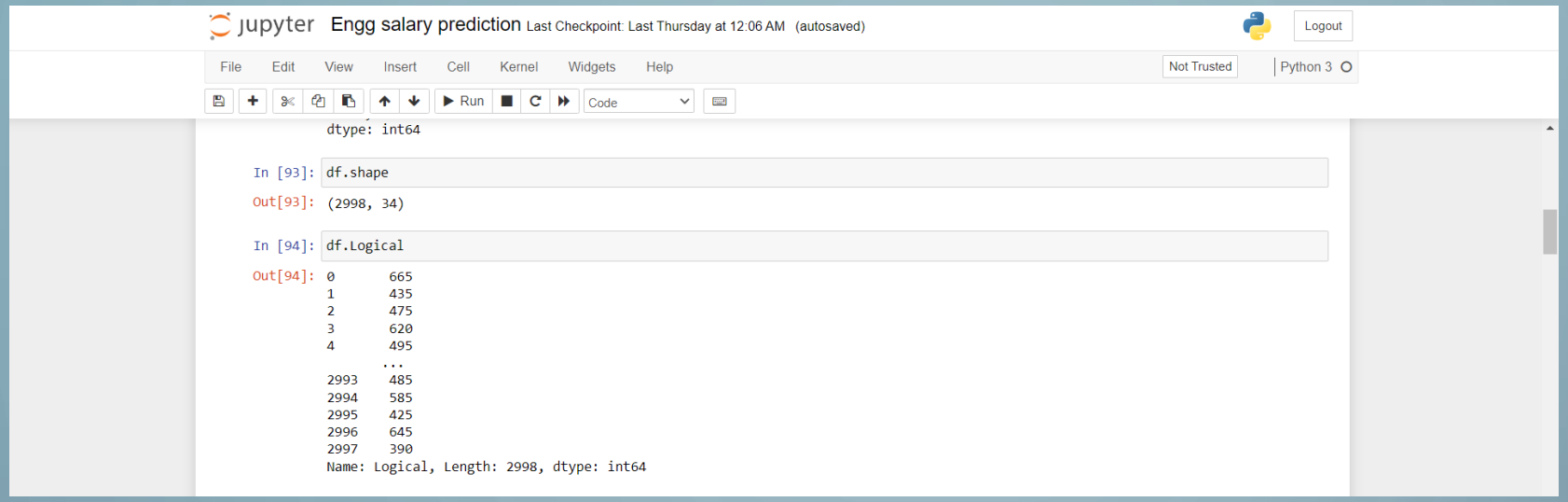


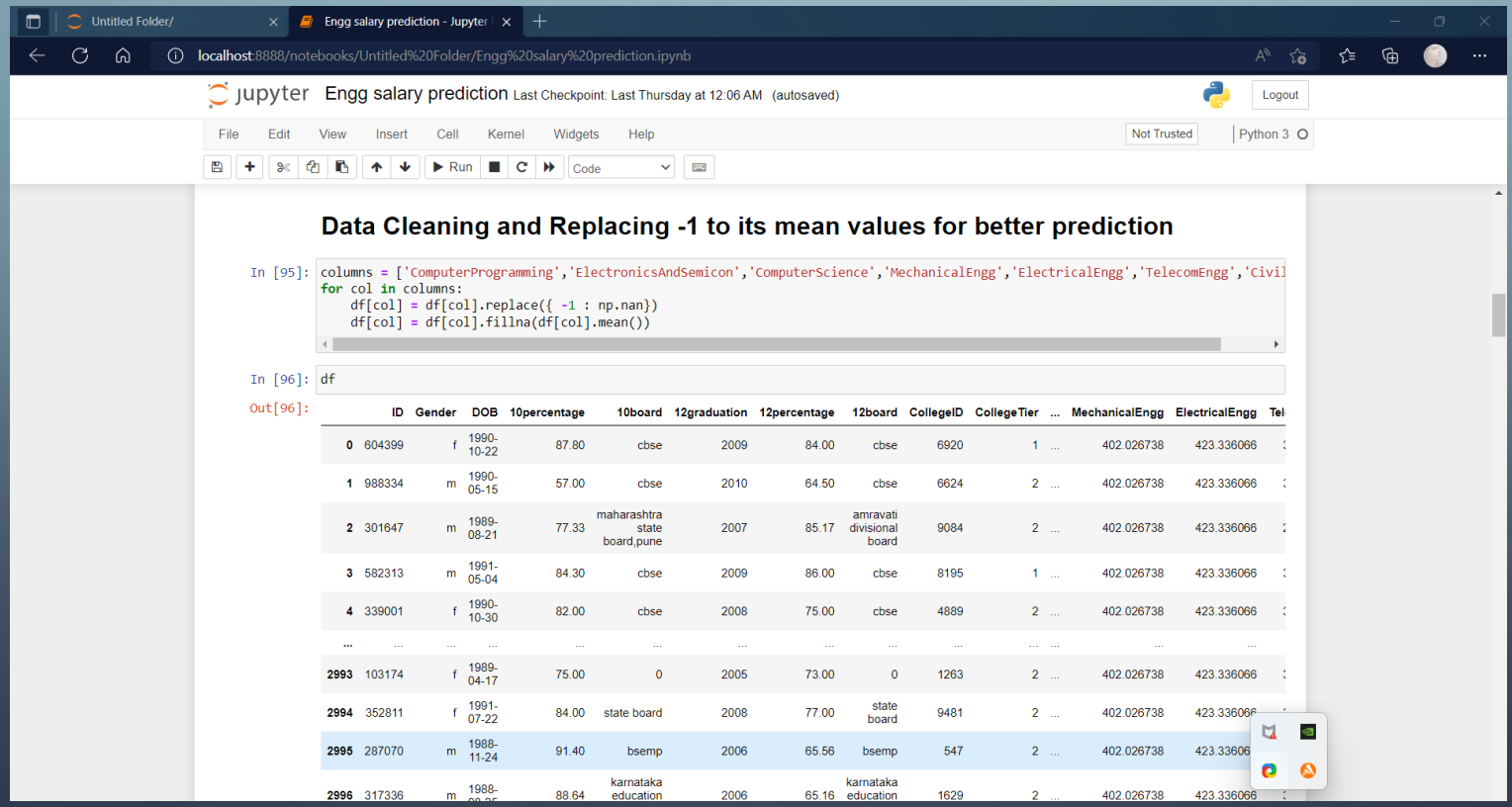
- Laptop
- Network connection
- Command prompt
- Jupyter Notebook

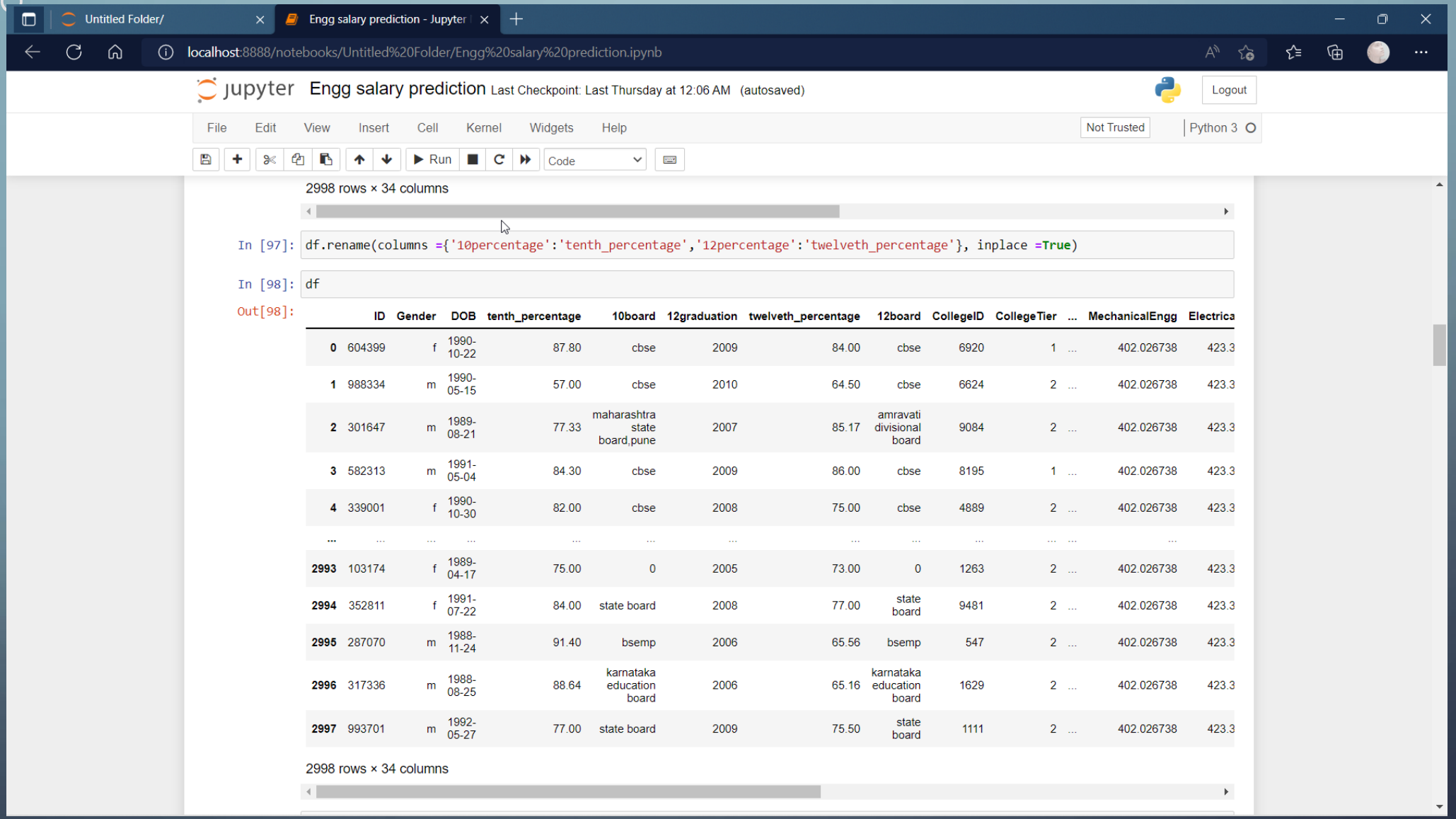
INPUT / OUTPUT / SCREENSHOTS:



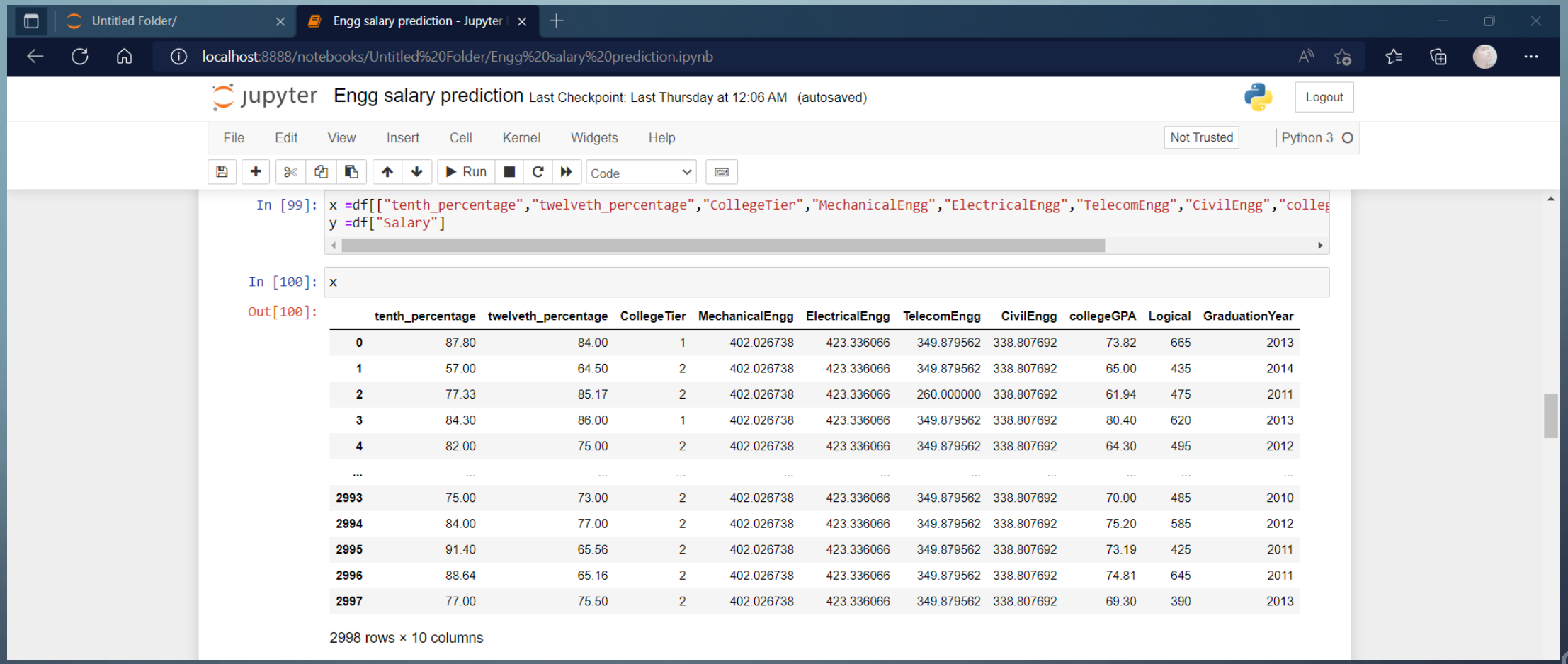

Untitled Folder/
Engg salary prediction - Jupyter
localhost:8888/notebooks/Untitled%20Folder/Engg%20salary%20prediction.ipynb
jupyter Engg salary prediction Last Checkpoint: Last Thursday at 12:06 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
In [92]: df.isnull().sum()
Out[92]:
ID 0
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


jupyter Engg salary prediction Last Checkpoint: Last Thursday at 12:06 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
dtype: int64
In [93]: df.shape
Out[93]: (2998, 34)
In [94]: df.Logical
Out[94]: 0 665
1 435
2 475
3 620
4 495
...
2993 485
2994 585
2995 425
2996 645
2997 390
Name: Logical, Length: 2998, dtype: int64


Untitled Folder/
Engg salary prediction - Jupyter
localhost:8888/notebooks/Untitled%20Folder/Engg%20salary%20prediction.ipynb
jupyter Engg salary prediction Last Checkpoint: Last Thursday at 12:06 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
Data Cleaning and Replacing -1 to its mean values for better prediction
In [95]:
columns = ['ComputerProgramming','ElectronicsAndSemicon','ComputerScience','MechanicalEngg','ElectricalEngg','TelecomEngg','Civil
for col in columns:
df[col] = df[col].replace({ -1 : np.nan})
df[col] = df[col].fillna(df[col].mean())
In [96]: df
Out[96]:
ID Gender DOB 10percentage 10board 12graduation 12percentage 12board CollegeID CollegeTier ... MechanicalEngg ElectricalEngg
0 604399 f 1990-10-22 87.80 cbse 2009 84.00 cbse 6920 1 ... 402.026738 423.336066
1 988334 m 1990-05-15 57.00 cbse 2010 64.50 cbse 6624 2 ... 402.026738 423.336066
2 301647 m 1989-08-21 77.33 maharashtra state board,pune 2007 85.17 amravati divisional board 9084 2 ... 402.026738 423.336066
3 582313 m 1991-05-04 84.30 cbse 2009 86.00 cbse 8195 1 ... 402.026738 423.336066
4 339001 f 1990-10-30 82.00 cbse 2008 75.00 cbse 4889 2 ... 402.026738 423.336066
2993 103174 f 1989-04-17 75.00 0 2005 73.00 0 1263 2 ... 402.026738 423.336066
2994 352811 f 1991-07-22 84.00 state board 2008 77.00 state board 9481 2 ... 402.026738 423.336066
2995 287070 m 1988-11-24 91.40 bsemp 2006 65.56 bsemp 547 2 ... 402.026738 423.336066
2996 317336 m 1988- 88.64 karnataka education 2006 65.16 karnataka education 1629 2 ... 402.026738 423.336066



	ID	Gender	DOB	tenth_percentage	10board	12graduation	twelveth_percentage	12board	CollegeID	CollegeTier	...	MechanicalEngg	Electrica
0	604399	f	1990-10-22	87.80	cbse	2009	84.00	cbse	6920	1	...	402.026738	423.3
1	988334	m	1990-05-15	57.00	cbse	2010	64.50	cbse	6624	2	...	402.026738	423.3
2	301647	m	1989-08-21	77.33	maharashtra state board,pune	2007	85.17	amravati divisional board	9084	2	...	402.026738	423.3
3	582313	m	1991-05-04	84.30	cbse	2009	86.00	cbse	8195	1	...	402.026738	423.3
4	339001	f	1990-10-30	82.00	cbse	2008	75.00	cbse	4889	2	...	402.026738	423.3
...	...	...	...	...	...	...	...	...	...	...	...	...	
2993	103174	f	1989-04-17	75.00	0	2005	73.00	0	1263	2	...	402.026738	423.3
2994	352811	f	1991-07-22	84.00	state board	2008	77.00	state board	9481	2	...	402.026738	423.3
2995	287070	m	1988-11-24	91.40	bsemp	2006	65.56	bsemp	547	2	...	402.026738	423.3
2996	317336	m	1988-08-25	88.64	karnataka education board	2006	65.16	karnataka education board	1629	2	...	402.026738	423.3
2997	993701	m	1992-05-27	77.00	state board	2009	75.50	state board	1111	2	...	402.026738	423.3



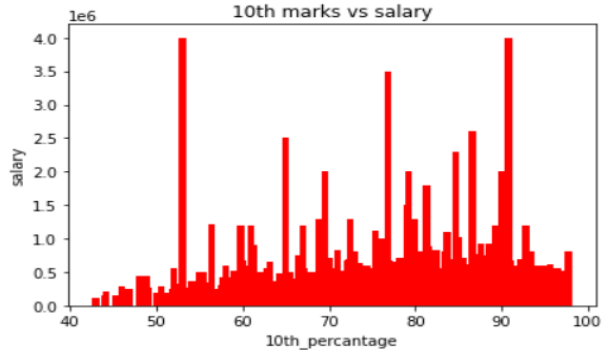
	tenth_percentage	twelveth_percentage	CollegeTier	MechanicalEngg	ElectricalEngg	TelecomEngg	CivilEngg	collegeGPA	Logical	GraduationYear
0	87.80	84.00	1	402.026738	423.336066	349.879562	338.807692	73.82	665	2013
1	57.00	64.50	2	402.026738	423.336066	349.879562	338.807692	65.00	435	2014
2	77.33	85.17	2	402.026738	423.336066	260.000000	338.807692	61.94	475	2011
3	84.30	86.00	1	402.026738	423.336066	349.879562	338.807692	80.40	620	2013
4	82.00	75.00	2	402.026738	423.336066	349.879562	338.807692	64.30	495	2012
...	...	...	...	...	...	...	...	...	...	...
2993	75.00	73.00	2	402.026738	423.336066	349.879562	338.807692	70.00	485	2010
2994	84.00	77.00	2	402.026738	423.336066	349.879562	338.807692	75.20	585	2012
2995	91.40	65.56	2	402.026738	423.336066	349.879562	338.807692	73.19	425	2011
2996	88.64	65.16	2	402.026738	423.336066	349.879562	338.807692	74.81	645	2011
2997	77.00	75.50	2	402.026738	423.336066	349.879562	338.807692	69.30	390	2013

2998 rows × 10 columns

Data Visulization

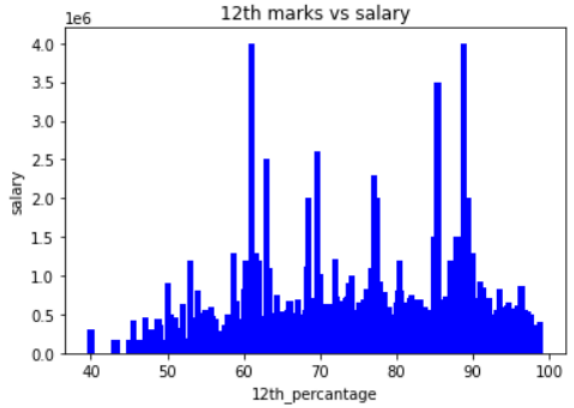
```
In [102]: plt.bar(x.tenth_percentage,y,color ="red")
          plt.xlabel("10th_percantage")
          plt.ylabel("salary")
          plt.title("10th marks vs salary")
```

```
Out[102]: Text(0.5, 1.0, '10th marks vs salary')
```



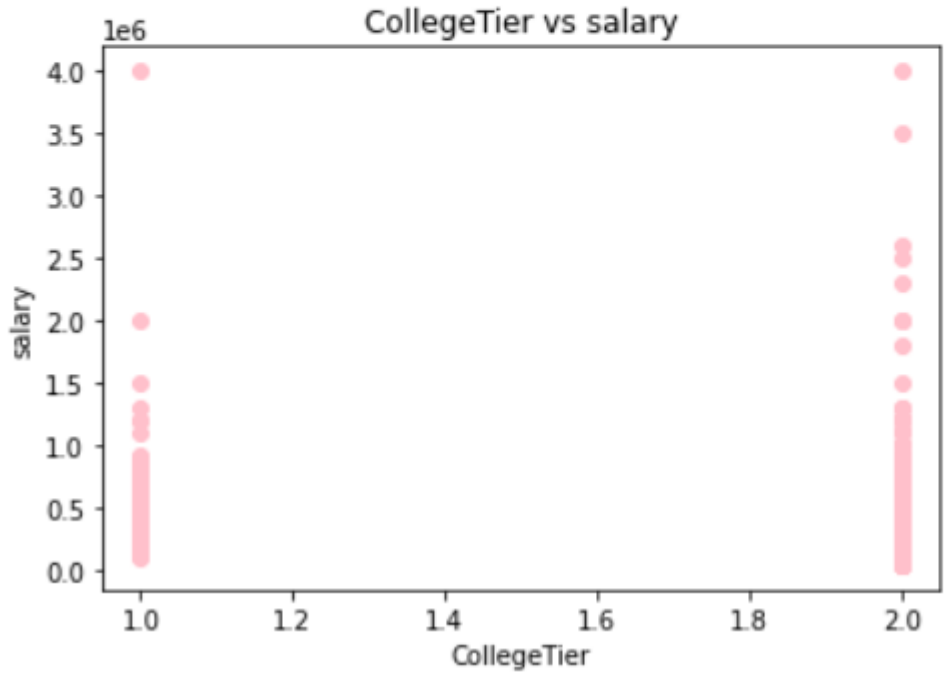

```
In [103]: plt.bar(x.twelveth_percentage,y,color ="blue")
          plt.xlabel("12th_percantage")
          plt.ylabel("salary")
          plt.title("12th marks vs salary")
```

```
Out[103]: Text(0.5, 1.0, '12th marks vs salary')
```



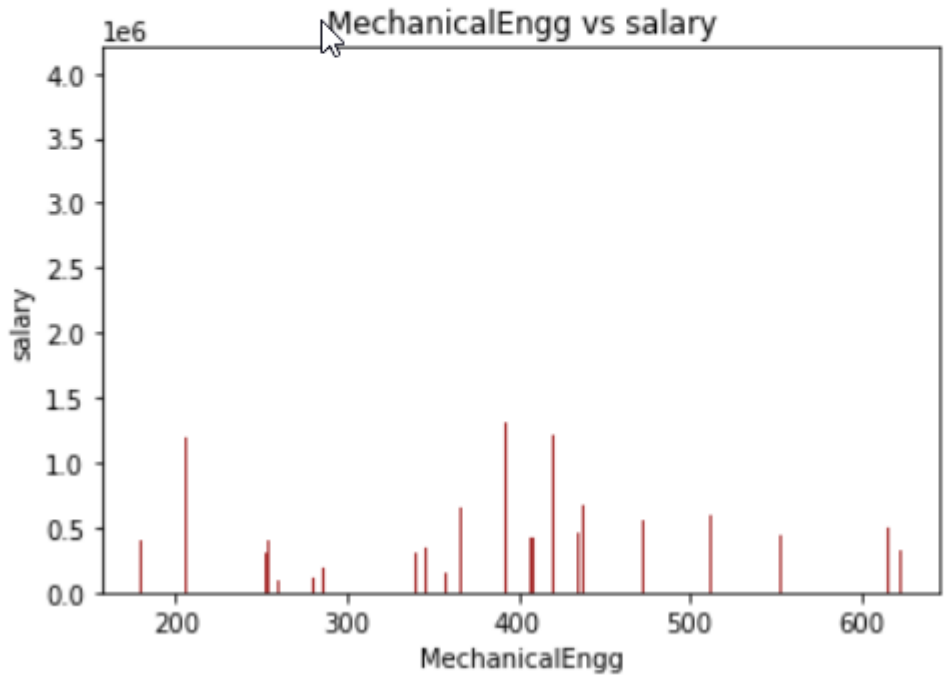
In [104]:
```
plt.scatter(x.CollegeTier,y,color ="pink")
plt.xlabel("CollegeTier")
plt.ylabel("salary")
plt.title("CollegeTier vs salary")
```

Out[104]: Text(0.5, 1.0, 'CollegeTier vs salary')



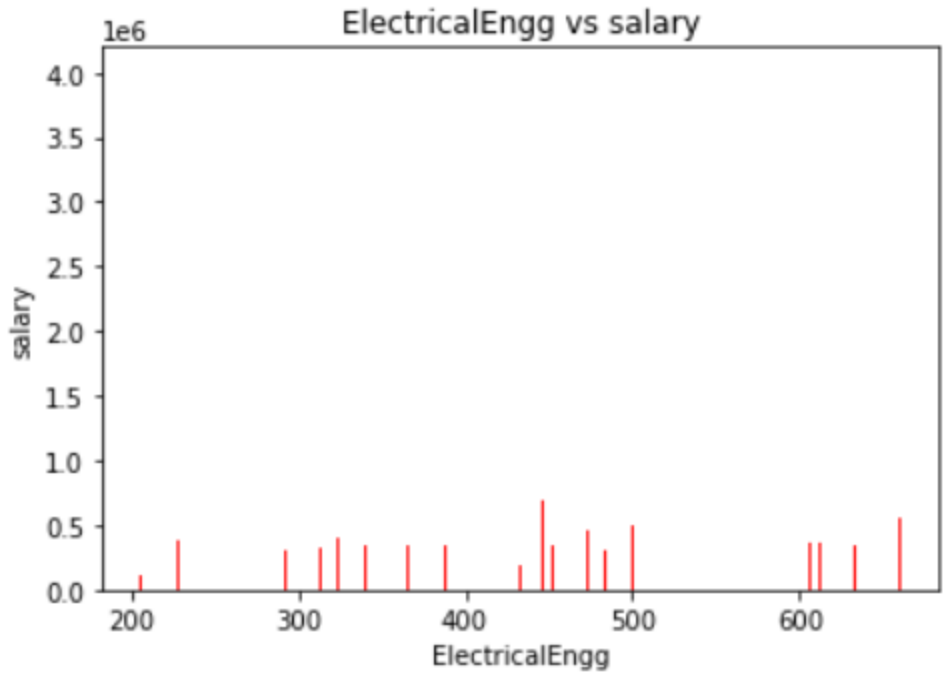
In [105]:
```
plt.bar(x.MechanicalEngg,y,color ="brown")
plt.xlabel("MechanicalEngg")
plt.ylabel("salary")
plt.title("MechanicalEngg vs salary")
```

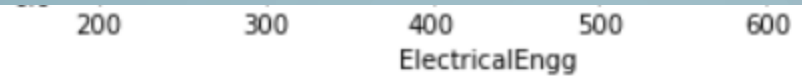
Out[105]:
```
Text(0.5, 1.0, 'MechanicalEngg vs salary')
```



In [106]:
```
plt.bar(x.ElectricalEngg,y,color ="red")
plt.xlabel("ElectricalEngg")
plt.ylabel("salary")
plt.title("ElectricalEngg vs salary")
```

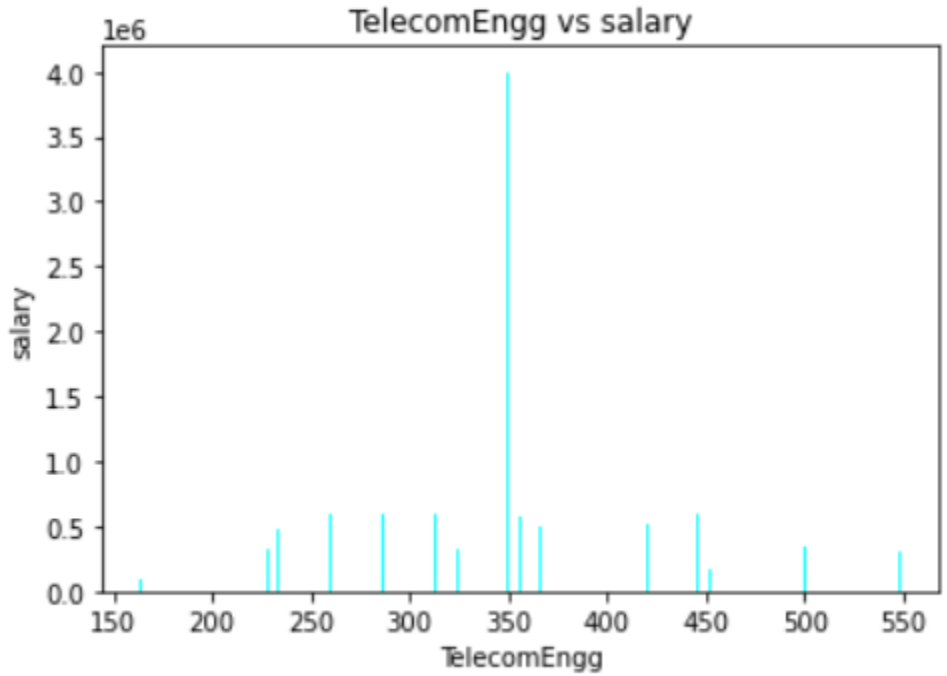
Out[106]:
```
Text(0.5, 1.0, 'ElectricalEngg vs salary')
```





In [107]:
```
plt.bar(x.TelecomEngg,y,color ="cyan")
plt.xlabel("TelecomEngg")
plt.ylabel("salary")
plt.title("TelecomEngg vs salary")
```

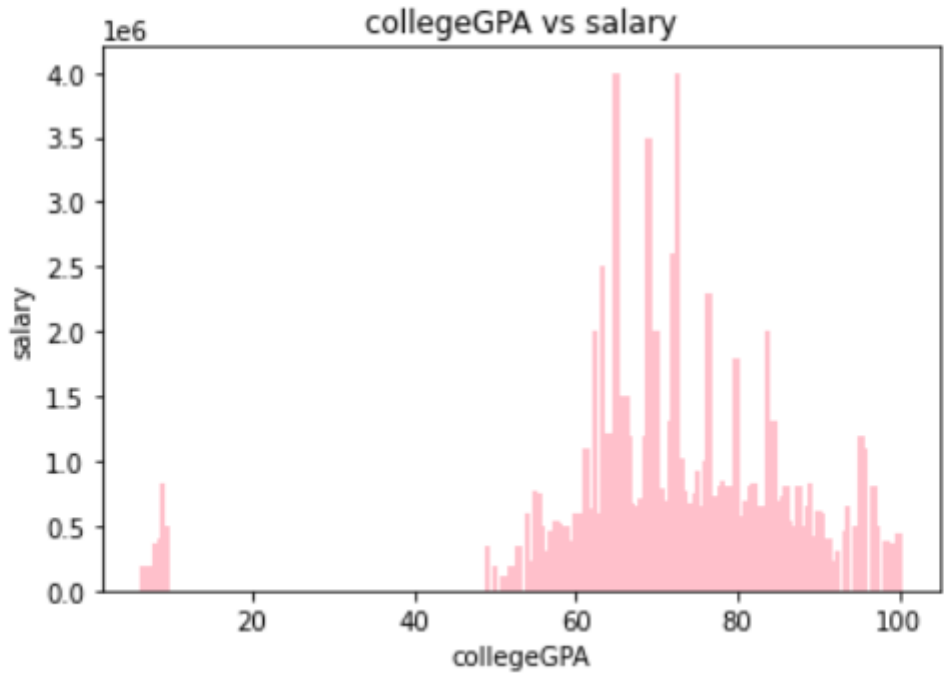
Out[107]: Text(0.5, 1.0, 'TelecomEngg vs salary')



TelecomEngg

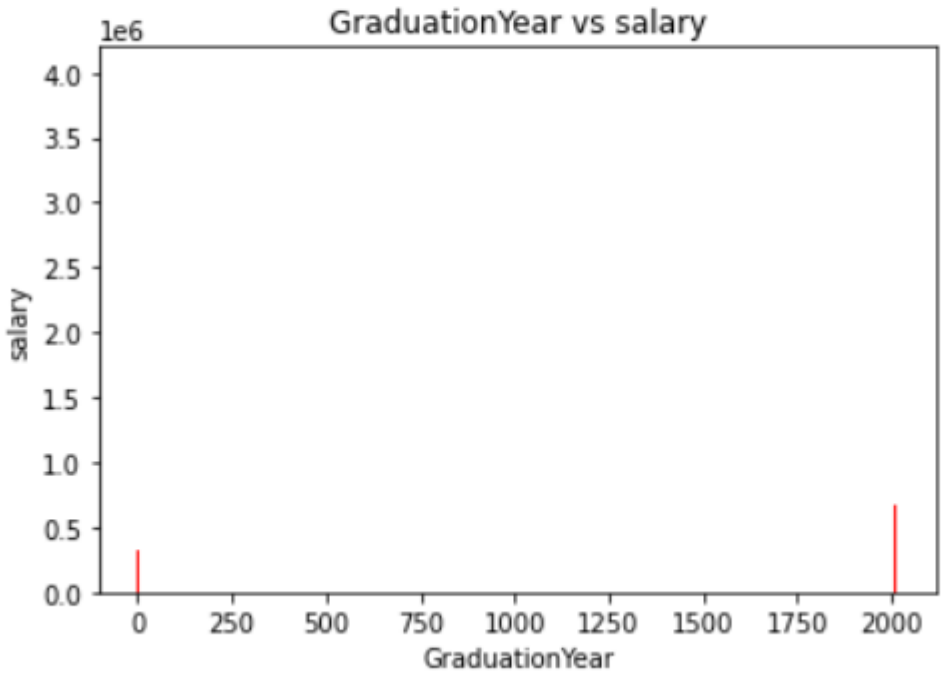
In [108]:
```
plt.bar(x.collegeGPA,y,color ="pink")
plt.xlabel("collegeGPA")
plt.ylabel("salary")
plt.title("collegeGPA vs salary")
```

Out[108]: Text(0.5, 1.0, 'collegeGPA vs salary')



In [109]:
```
plt.bar(x.GraduationYear,y,color ="red")
plt.xlabel("GraduationYear")
plt.ylabel("salary")
plt.title("GraduationYear vs salary")
```

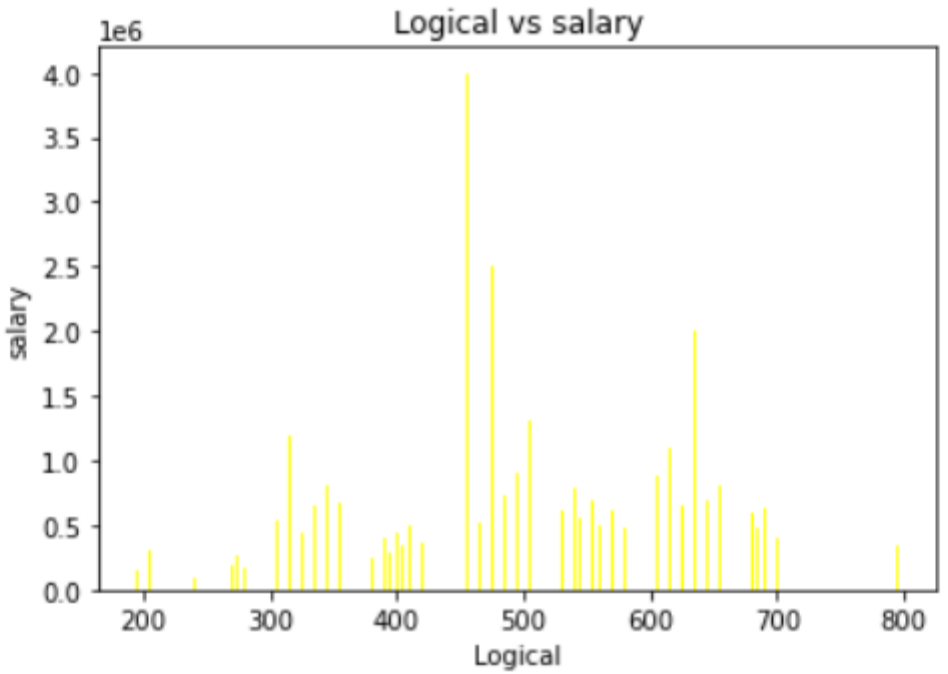
Out[109]: Text(0.5, 1.0, 'GraduationYear vs salary')

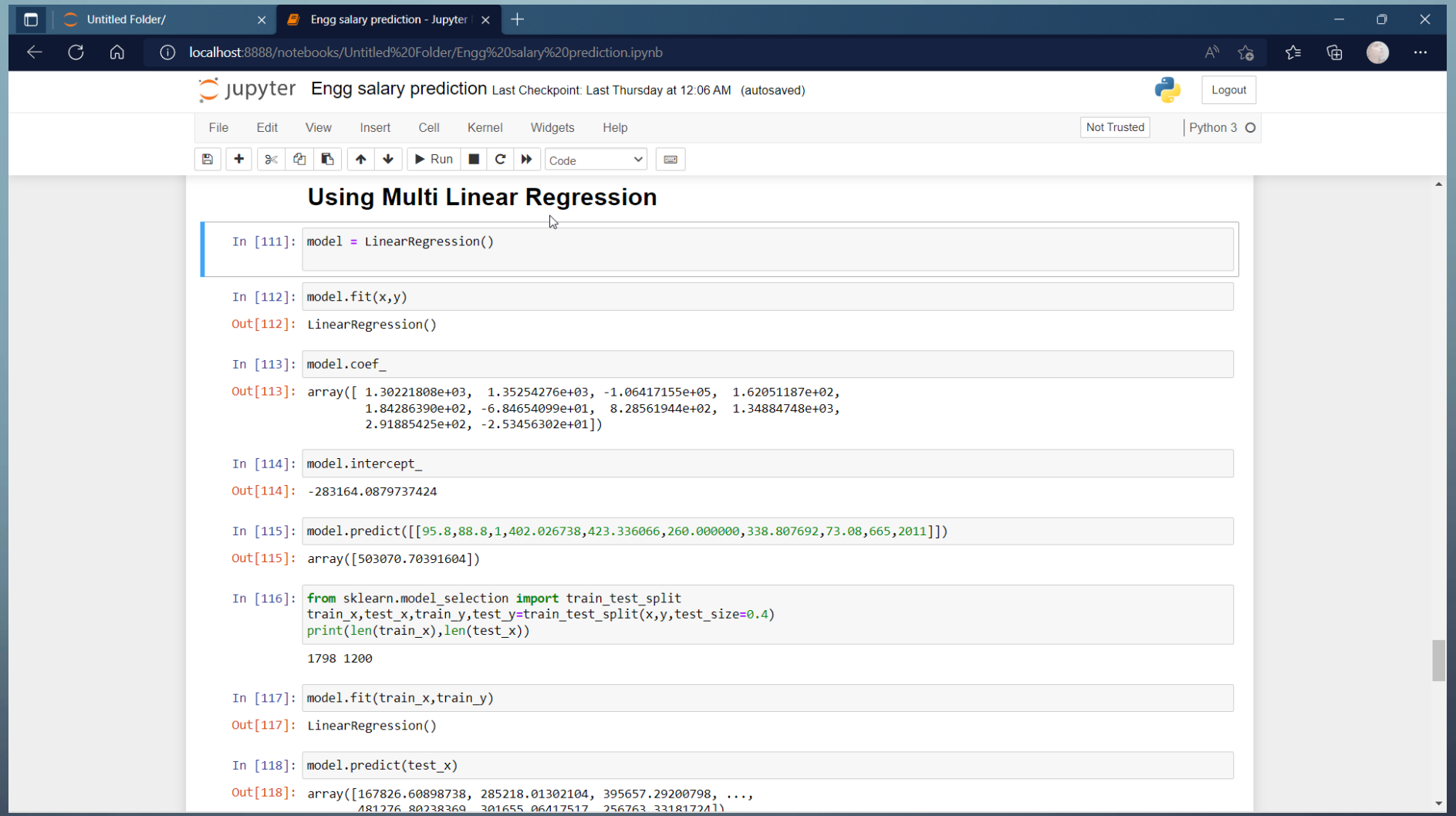


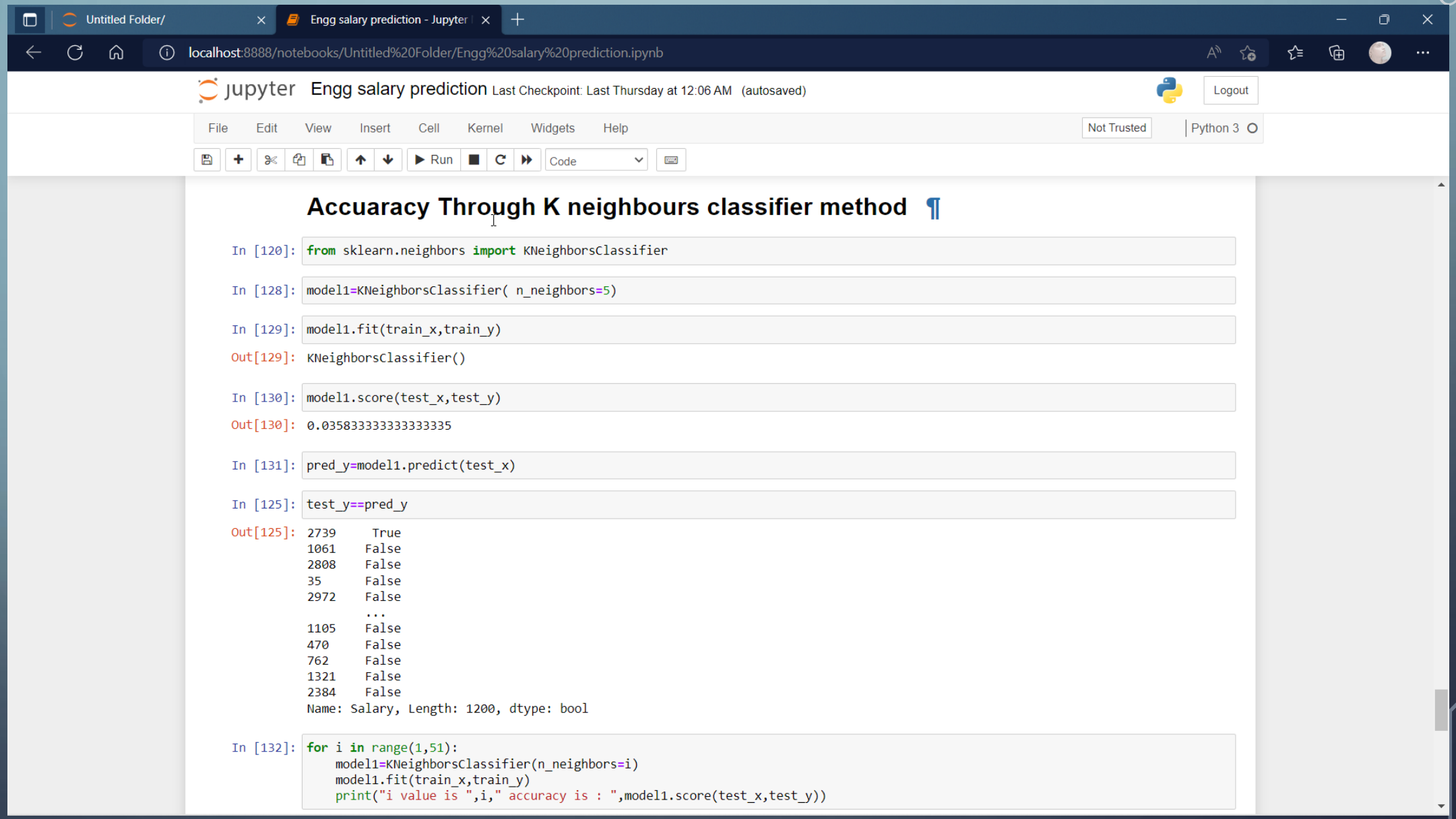
GraduationYear

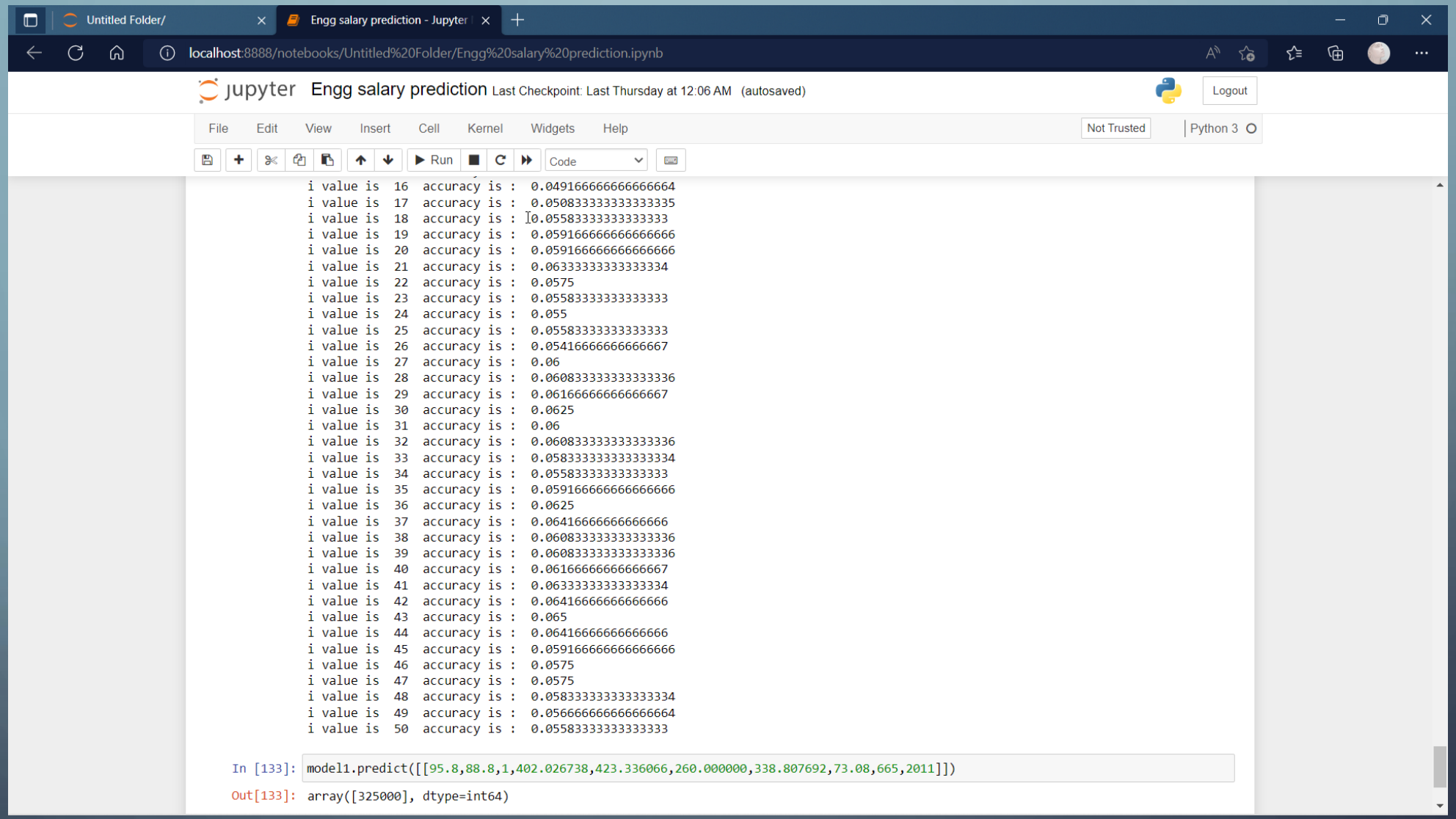
In [110]:
```
plt.bar(x.Logical,y,color ="yellow")
plt.xlabel("Logical")
plt.ylabel("salary")
plt.title("Logical vs salary")
```

Out[110]: Text(0.5, 1.0, 'Logical vs salary')




Untitled Folder/
Engg salary prediction - Jupyter
localhost:8888/notebooks/Untitled%20Folder/Engg%20salary%20prediction.ipynb
jupyter Engg salary prediction Last Checkpoint: Last Thursday at 12:06 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
Using Multi Linear Regression
In [111]: model = LinearRegression()
In [112]: model.fit(x,y)
Out[112]: LinearRegression()
In [113]: model.coef_
Out[113]: array([1.30221808e+03, 1.35254276e+03, -1.06417155e+05, 1.62051187e+02,
1.84286390e+02, -6.84654099e+01, 8.28561944e+02, 1.34884748e+03,
2.91885425e+02, -2.53456302e+01])
In [114]: model.intercept_
Out[114]: -283164.0879737424
In [115]: model.predict([[95.8,88.8,1,402.026738,423.336066,260.000000,338.807692,73.08,665,2011]])
Out[115]: array([503070.70391604])
In [116]: from sklearn.model_selection import train_test_split
train_x,test_x,train_y,test_y=train_test_split(x,y,test_size=0.4)
print(len(train_x),len(test_x))
1798 1200
In [117]: model.fit(train_x,train_y)
Out[117]: LinearRegression()
In [118]: model.predict(test_x)
Out[118]: array([167826.60898738, 285218.01302104, 395657.29200798, ...,


Untitled Folder/
Engg salary prediction - Jupyter
localhost:8888/notebooks/Untitled%20Folder/Engg%20salary%20prediction.ipynb
jupyter Engg salary prediction Last Checkpoint: Last Thursday at 12:06 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
Accuaracy Through K neighbours classifier method ¶
In [120]: from sklearn.neighbors import KNeighborsClassifier
In [128]: model1=KNeighborsClassifier(n_neighbors=5)
In [129]: model1.fit(train_x,train_y)
Out[129]: KNeighborsClassifier()
In [130]: model1.score(test_x,test_y)
Out[130]: 0.035833333333333335
In [131]: pred_y=model1.predict(test_x)
In [125]: test_y==pred_y
Out[125]: 2739 True
1061 False
2808 False
35 False
2972 False
...
1105 False
470 False
762 False
1321 False
2384 False
Name: Salary, Length: 1200, dtype: bool
In [132]: for i in range(1,51):
model1=KNeighborsClassifier(n_neighbors=i)
model1.fit(train_x,train_y)
print("i value is ",i," accuracy is : ",model1.score(test_x,test_y))


Untitled Folder/
Engg salary prediction - Jupyter
localhost:8888/notebooks/Untitled%20Folder/Engg%20salary%20prediction.ipynb
jupyter Engg salary prediction Last Checkpoint: Last Thursday at 12:06 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
i value is 16 accuracy is : 0.049166666666666664
i value is 17 accuracy is : 0.050833333333333335
i value is 18 accuracy is : 0.05583333333333333
i value is 19 accuracy is : 0.059166666666666666
i value is 20 accuracy is : 0.059166666666666666
i value is 21 accuracy is : 0.06333333333333334
i value is 22 accuracy is : 0.0575
i value is 23 accuracy is : 0.05583333333333333
i value is 24 accuracy is : 0.055
i value is 25 accuracy is : 0.05583333333333333
i value is 26 accuracy is : 0.05416666666666667
i value is 27 accuracy is : 0.06
i value is 28 accuracy is : 0.060833333333333336
i value is 29 accuracy is : 0.06166666666666667
i value is 30 accuracy is : 0.0625
i value is 31 accuracy is : 0.06
i value is 32 accuracy is : 0.060833333333333336
i value is 33 accuracy is : 0.058333333333333334
i value is 34 accuracy is : 0.05583333333333333
i value is 35 accuracy is : 0.059166666666666666
i value is 36 accuracy is : 0.0625
i value is 37 accuracy is : 0.06416666666666666
i value is 38 accuracy is : 0.060833333333333336
i value is 39 accuracy is : 0.060833333333333336
i value is 40 accuracy is : 0.06166666666666667
i value is 41 accuracy is : 0.06333333333333334
i value is 42 accuracy is : 0.06416666666666666
i value is 43 accuracy is : 0.065
i value is 44 accuracy is : 0.06416666666666666
i value is 45 accuracy is : 0.059166666666666666
i value is 46 accuracy is : 0.0575
i value is 47 accuracy is : 0.0575
i value is 48 accuracy is : 0.058333333333333334
i value is 49 accuracy is : 0.056666666666666664
i value is 50 accuracy is : 0.05583333333333333
In [133]: model1.predict([[95.8,88.8,1,402.026738,423.336066,260.000000,338.807692,73.08,665,2011]])
Out[133]: array([325000], dtype=int64)

THE END