Source: [Sloan Digital Sky Survey (SDSS) dataset](https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey) on Kaggle

Predicting Stars, Galaxies & Quasars using ML



In this blog-post, I will be using the power of ML to explore our space. Yes, you read it right, we will be predicting the class of some celestial bodies present in our universe with the help of various measurement attributes available in this dataset. ML has made it possible to learn about space as well now. Isn’t it interesting!

**About Space/ Universe**

As defined in Wikipedia- https://en.wikipedia.org/wiki/Space; “Space is the boundless three-dimensional extent in which objects and events have relative position and direction. Physical space is often conceived in three linear dimensions, although modern physicists usually consider it, with time, to be part of a boundless four-dimensional continuum known as space time.” Space contains various types of celestial bodies present in them. Some of them have been identified, however, very large portion of the space is still a mystery to us.

**PROBLEM STATEMENT**

Scientists use many techniques to find out the type of celestial body and to explore our universe. The most useful techniques they use are through measuring the distances by applying various mathematical operations. We won’t be going much deep into that as it may take a lot time to describe those space science and also, I am not a space scientist just an aspiring data scientist ;)

**Overview of the Data and its features**

The Sloan Digital Sky Survey (SDSS) which offers public data of space observations. Found this data to be very insightful and interesting to explore and to make much use of our ML techniques on it.

Data link: <https://raw.githubusercontent.com/dsrscientist/dataset1/master/Skyserver.csv>

**Some of the important features/ attributes are explained below, for more knowledge about the features, you may check on kaggle dataset overview @** [**brilliant Kaggle kernel**](https://www.kaggle.com/farazrahman/predicting-star-galaxy-quasar-with-svm/)**well explained by Faraz Rahman:**

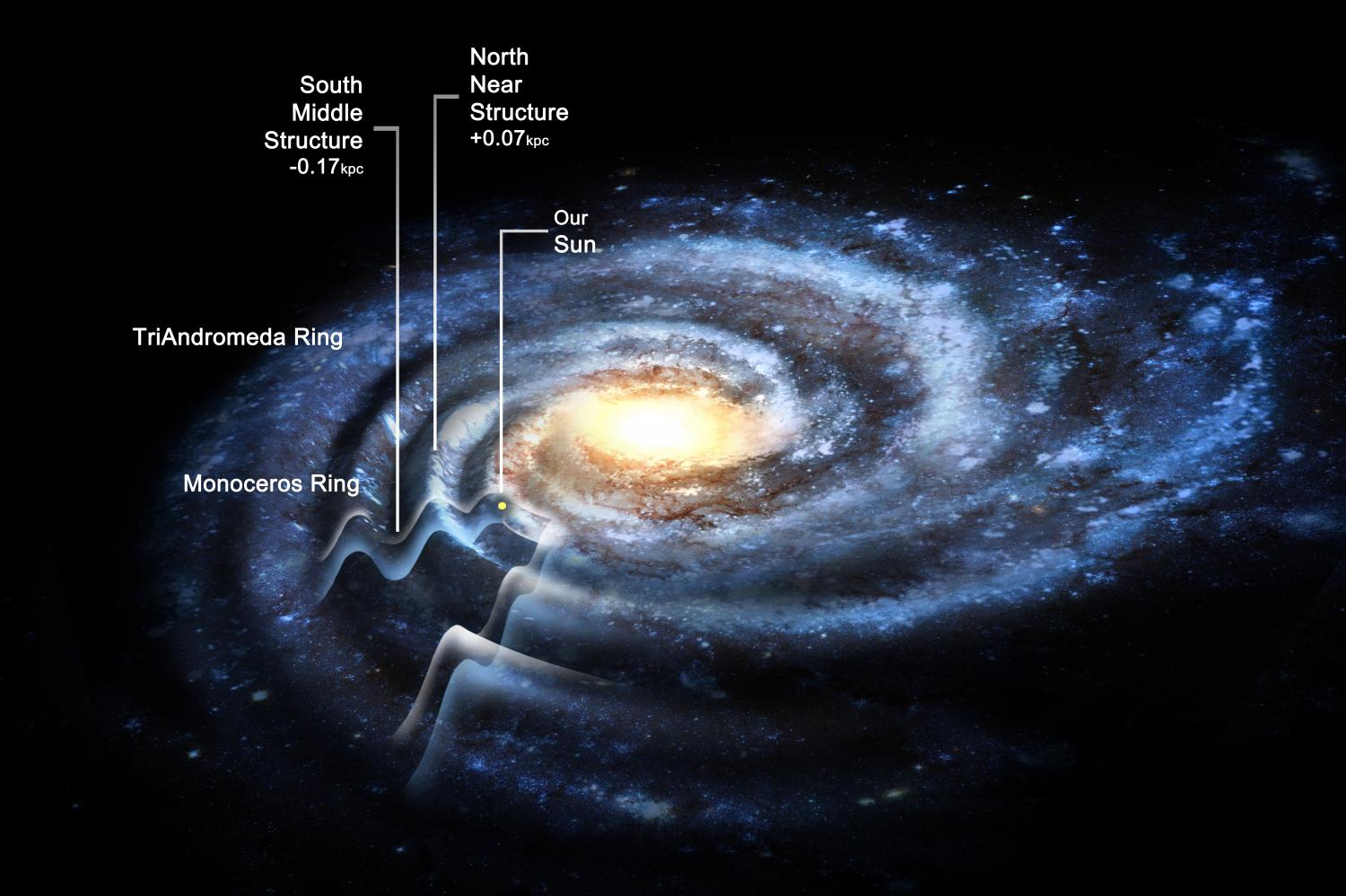
* ra, dec 🡪 right ascension and declination respectively
* u, g, r, i, z 🡪 filter bands (a.k.a. photometric system or astronomical magnitudes)
* run, rerun, camcol, field 🡪 descriptors of fields (i.e. 2048 x 1489 pixels) within image
* redshift 🡪 increase in wavelength due to motion of astronomical object
* plate 🡪 plate number
* mjd 🡪 modified Julian date of observation
* fiberid 🡪 optic fiber ID

**Here, we are predicting the class which includes values like star, galaxy and quasar with the help of various variables. We will be using classification model as our target variable has categorical data with more than 2 categories**

* Target 🡪 class = object class (galaxy, star or quasar)

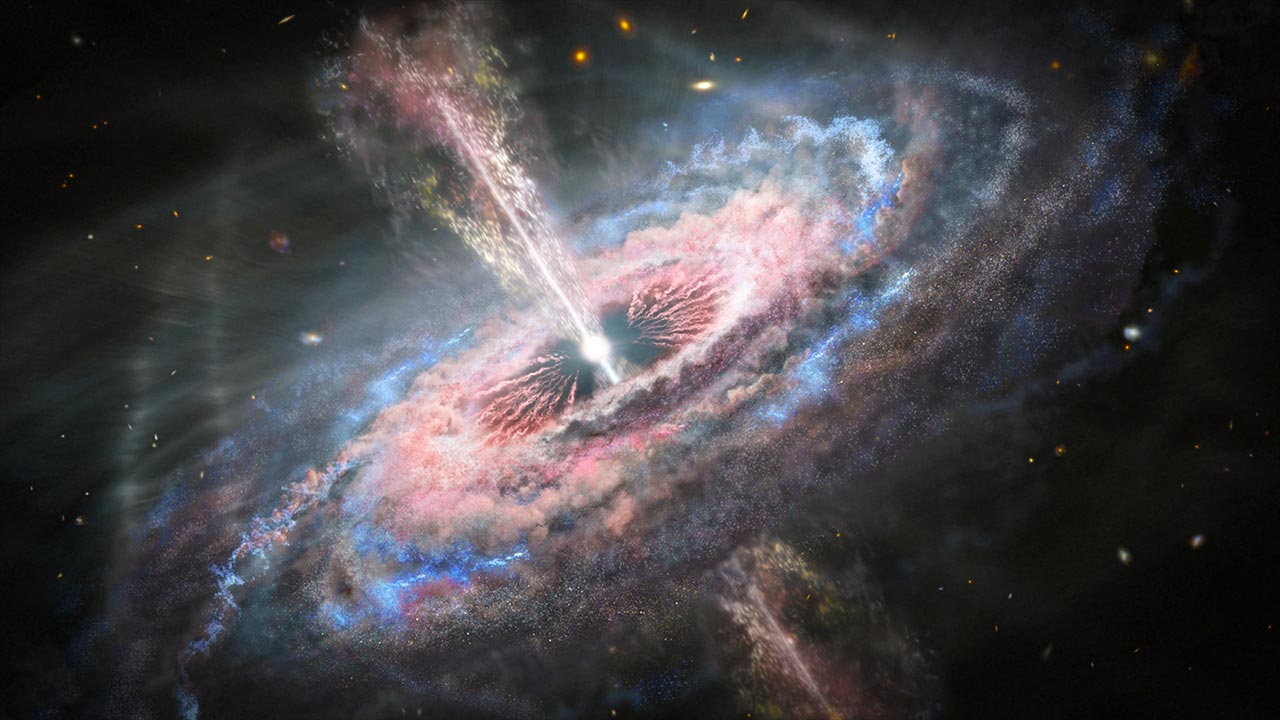
**Let’s now explore our class/ Target:**

As per the description available with the dataset, galaxy, star and quasar are explained as below:

* *A***GALAXY***is a gravitationally bound system of stars, stellar remnants, interstellar gas, dust, and dark matter. Galaxies are categorised according to their visual morphology as elliptical, spiral, or irregular. Many galaxies are thought to have supermassive black holes at their active centers.* 
* *A***STAR***is a type of astronomical object consisting of a luminous spheroid of plasma held together by its own gravity. The nearest star to Earth is the Sun.*



* *A***QUASAR***, also known as quasi-stellar object, is an extremely luminous active galactic nucleus (AGN). The power radiated by quasars is enormous. The most powerful quasars have luminosities exceeding 1041 watts, thousands of times greater than an ordinary large galaxy such as the Milky Way.*

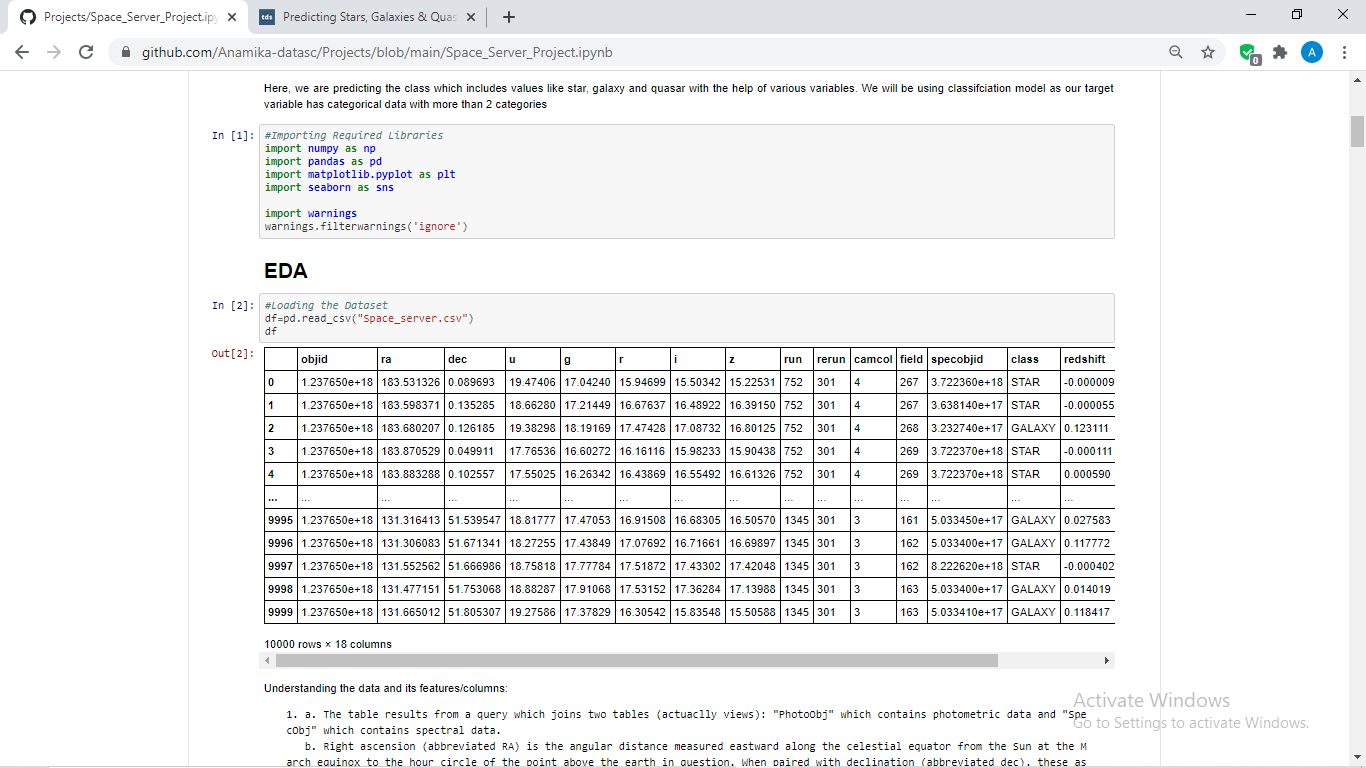


**EXPLORATORY DATA ANALYSIS**

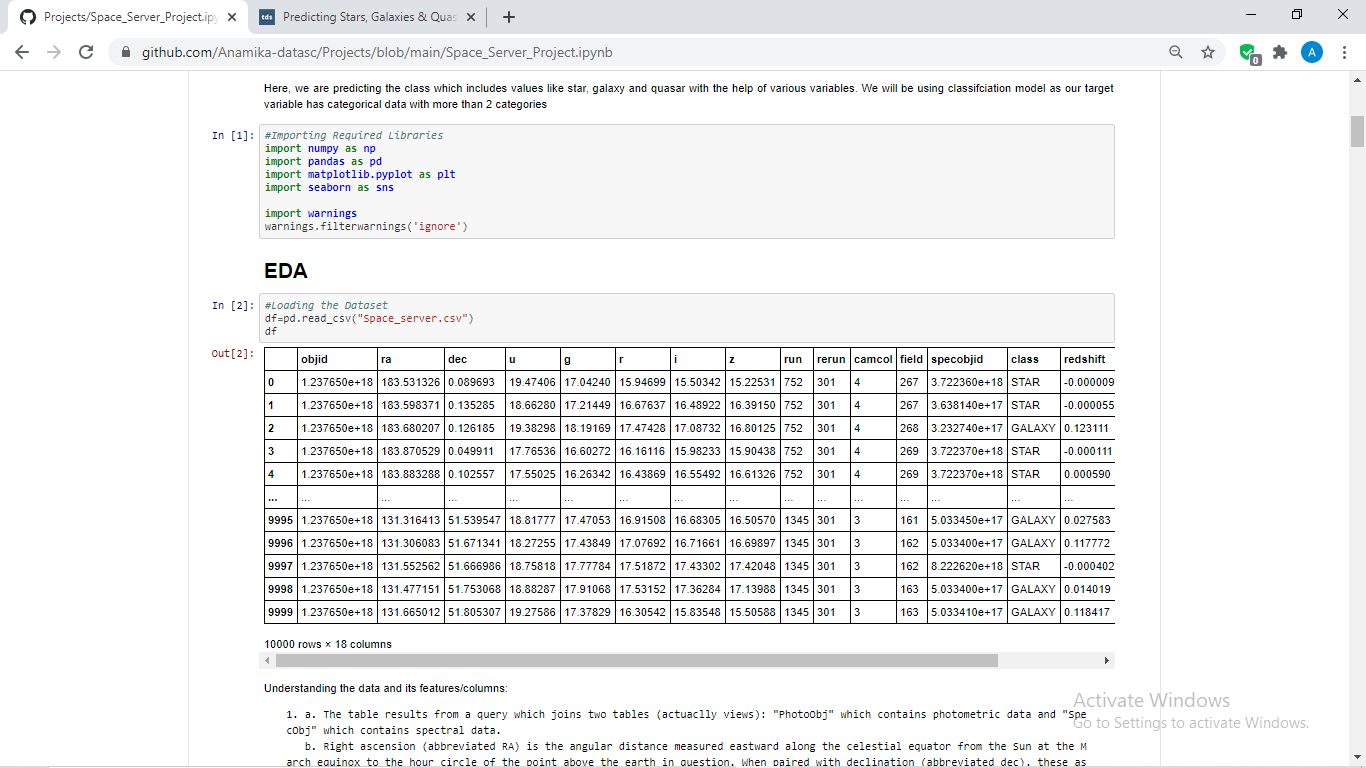
Let’s explore the data to know its types, nature, presence of missing values, outliers or skewness, knowing the correlation of the variables with target variables, range of each variables, its distribution etc. This analysis is done to know the pre-processing which needs to be adapted before sending the data to the model.

**I have done the analysis as per below steps:**

* **Importing Required Libraries and Data**



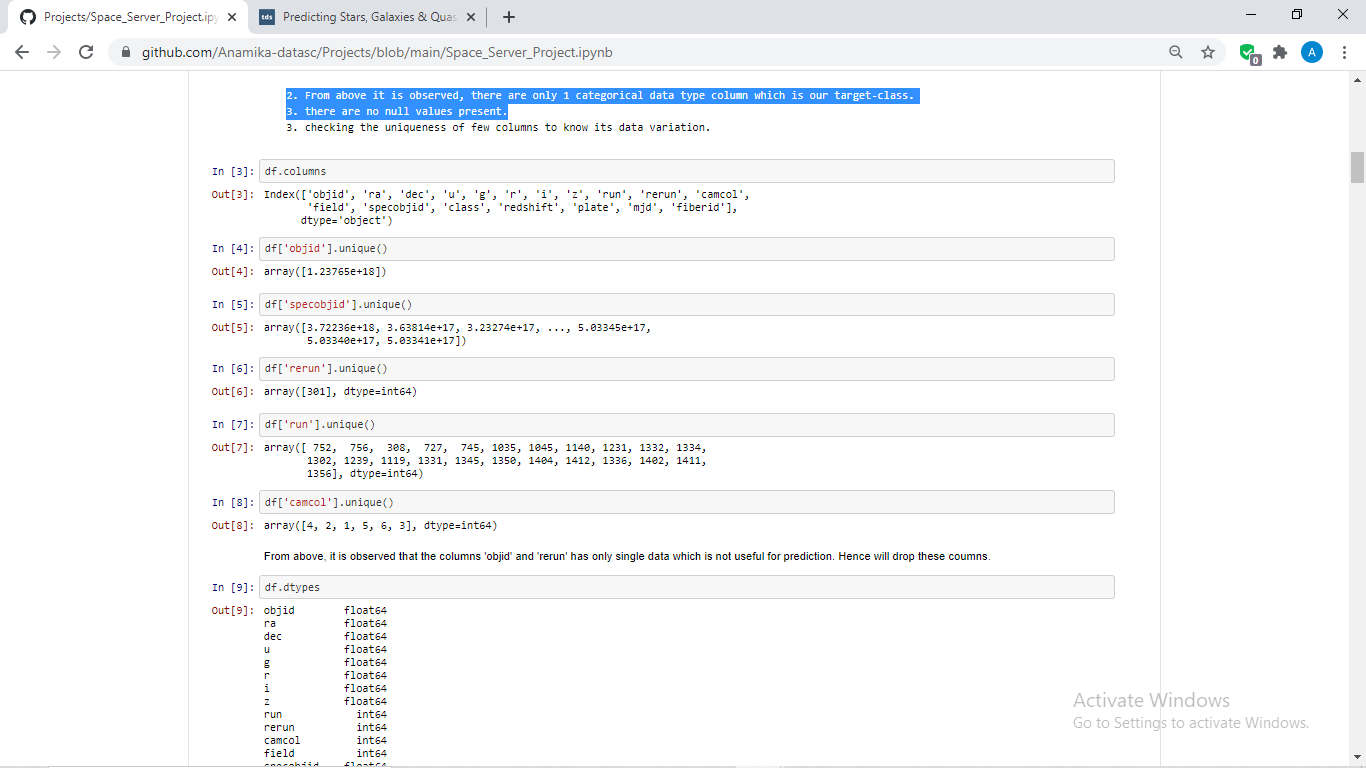
We have imported statistics libraries like **numpy** and **pandas** and visualization libraries like **matplotlib** and **seaborn**. Also, we have imported warnings to avoid warnings that comes after particular codes (However, it’s an optional step).



After loading the data, we have observed the following about the data:

* There are 10,000 rows and 18 columns
* There is only 1 categorical data type column which is our target-class.
* By looking at the data, no null values have been observed, however, will check the same by our standard method of **isnull().sum()**.
* It is also being seen that some of the columns have unique and single data in them.
* **Checking Uniqueness:**

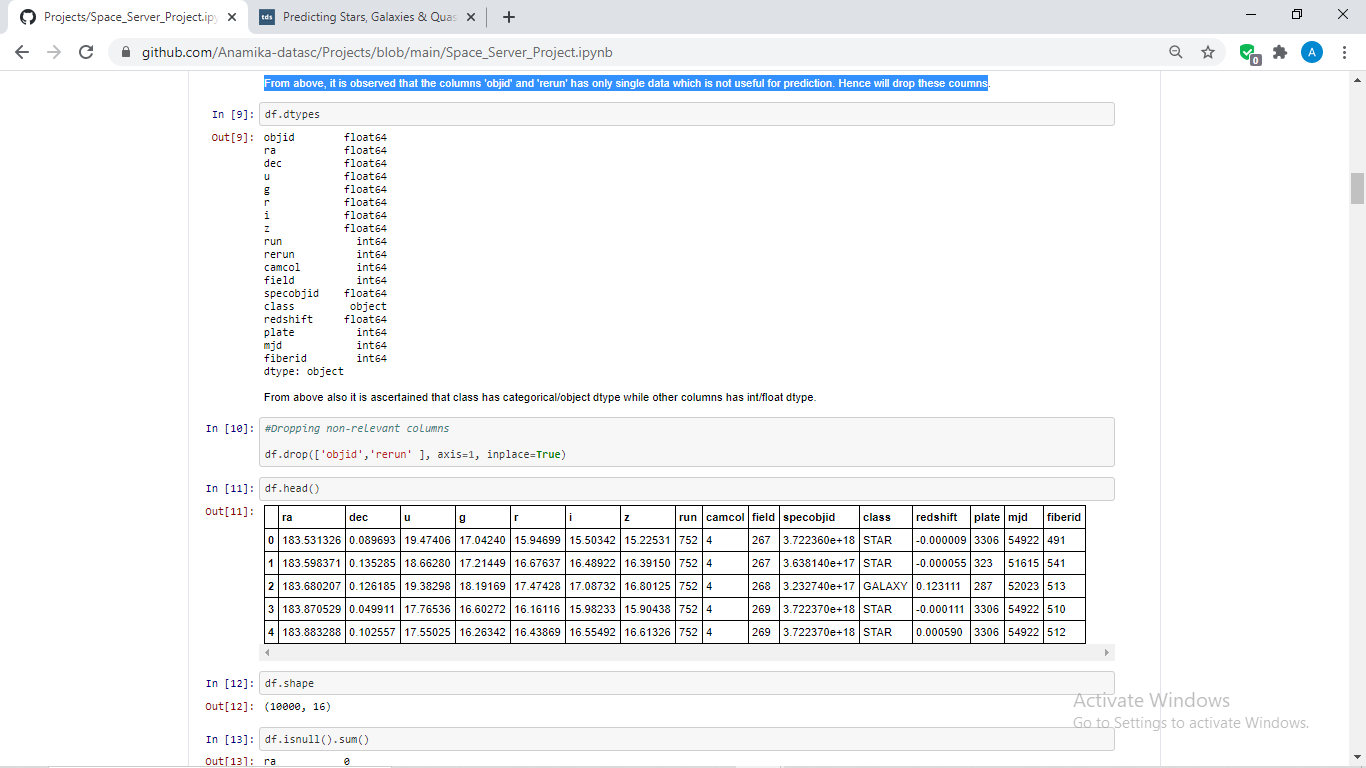
Now, I will be checking the uniqueness of few columns as per our above observations to know the variation of the data as single and unique data in all rows is not at all useful in any predictions. We will be removing such columns further.



From above, it is observed that the columns 'objid' and 'rerun' has only single data which is not useful for prediction. Hence will drop these columns.

* **Checking Datatype:**

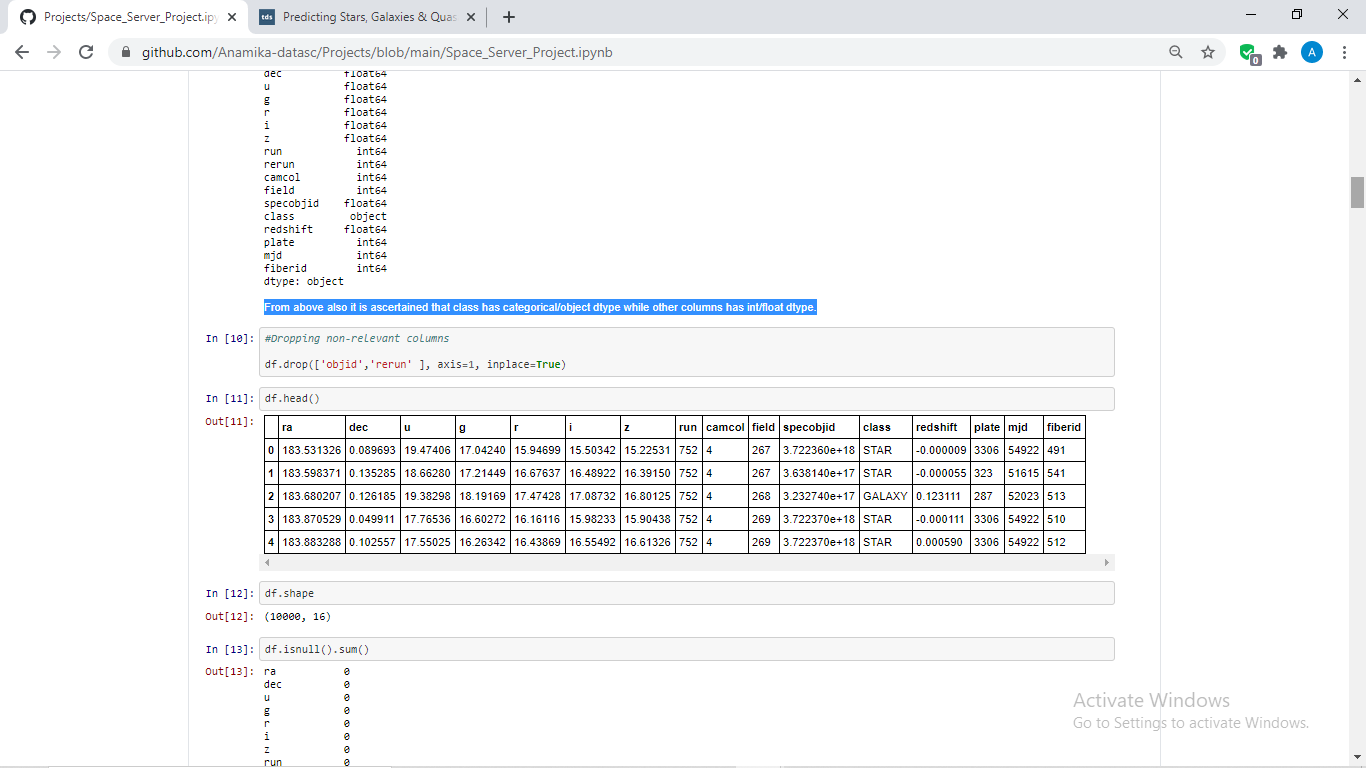
Will check the data type to know whether there is a need of encoding or data type conversion or not.



From above also it is ascertained that class has categorical/object data type while other columns has int/float data type. Will be performing label encoding for the ‘class’ column.

* **Dropping Non-relevant Columns:**

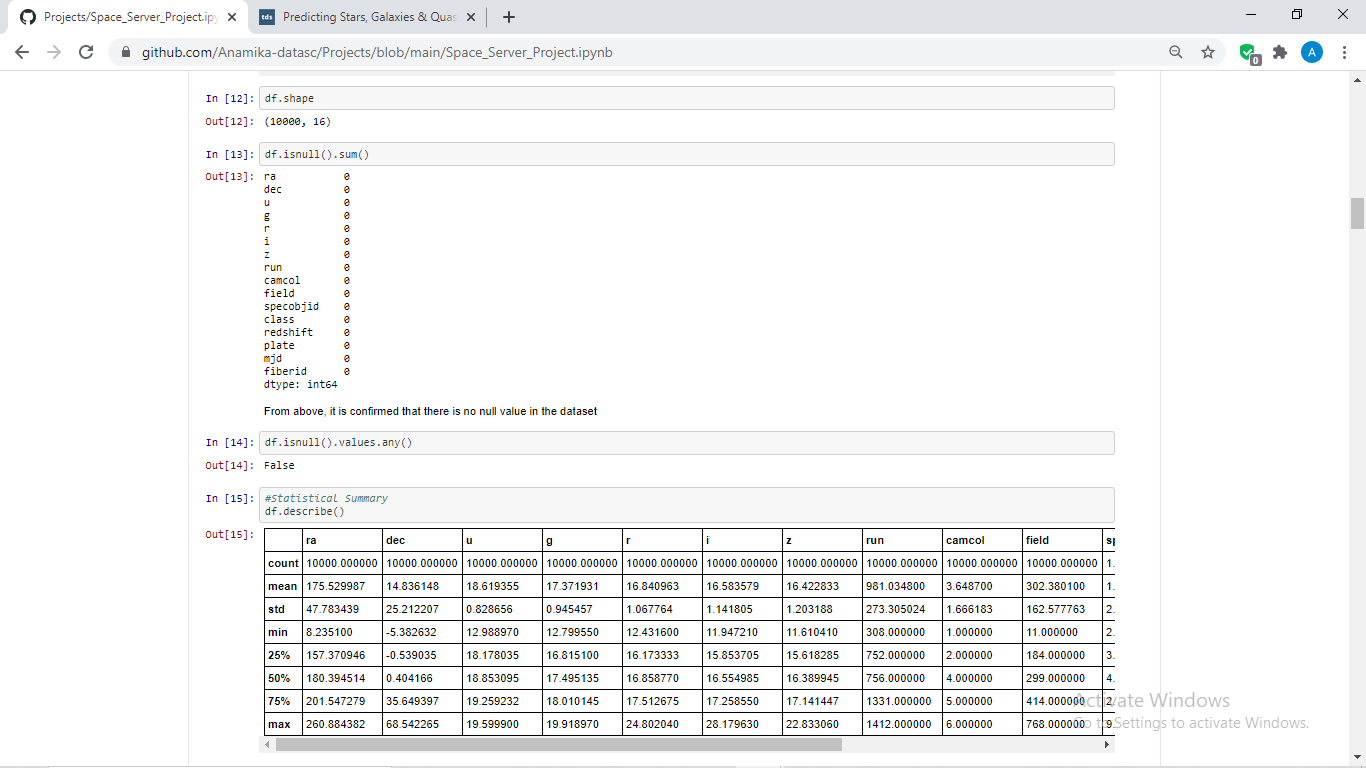
Here, we have dropped the 'objid' and 'rerun' columns as they had single data in them which was seen above.



With df.head() checking the first five rows of the data to know the available columns after dropping. (optional step)

* **Checking Shape and Nan Values:**

As now, we have dropped 2 columns will check the shape of the data and will also check the Nan values, if any in the dataset.

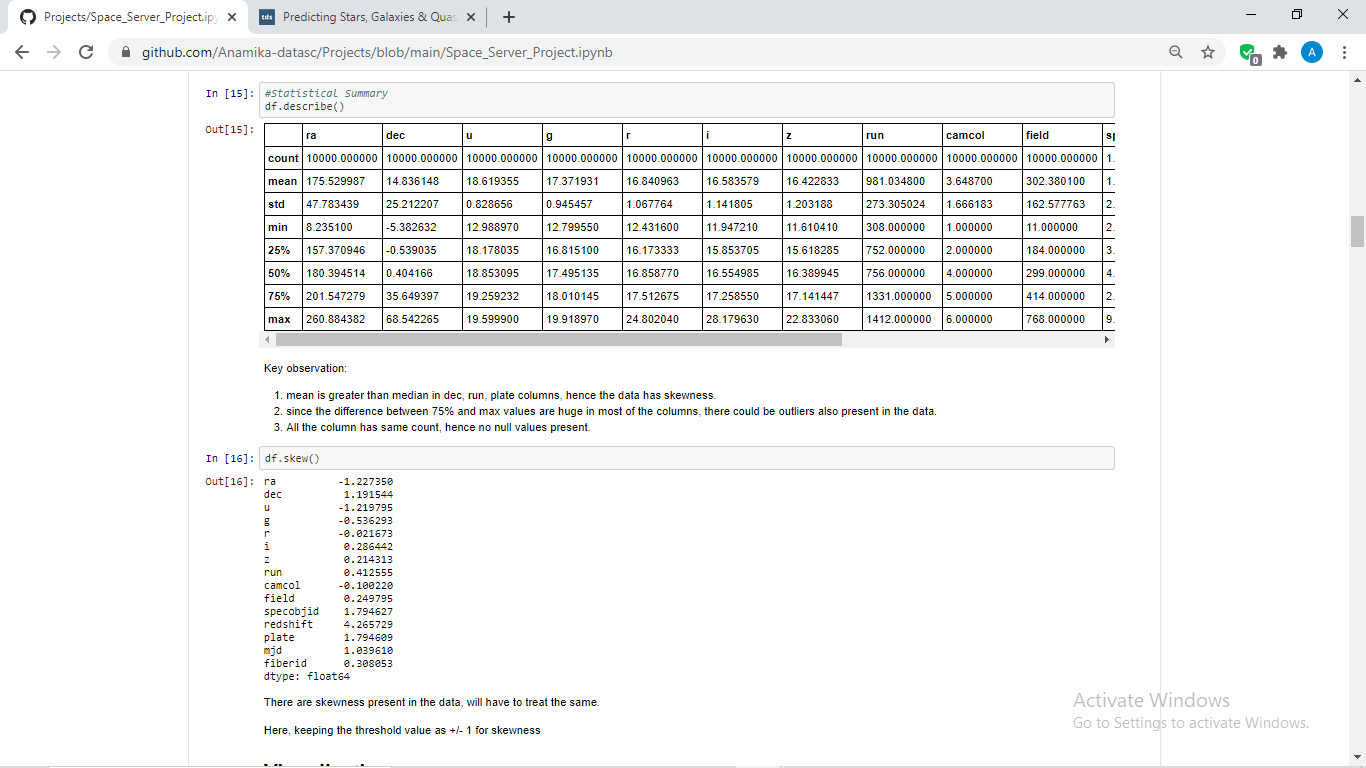


As seen above, shape of the data now changed to 10,000 rows and 16 columns

Also, there is no null values present in the data as I have indicated the same above itself.

* **Statistical Summary:**

With the help of statistical summary or as we call it in our python language **df.describe(),** we can know various aspects about the data.

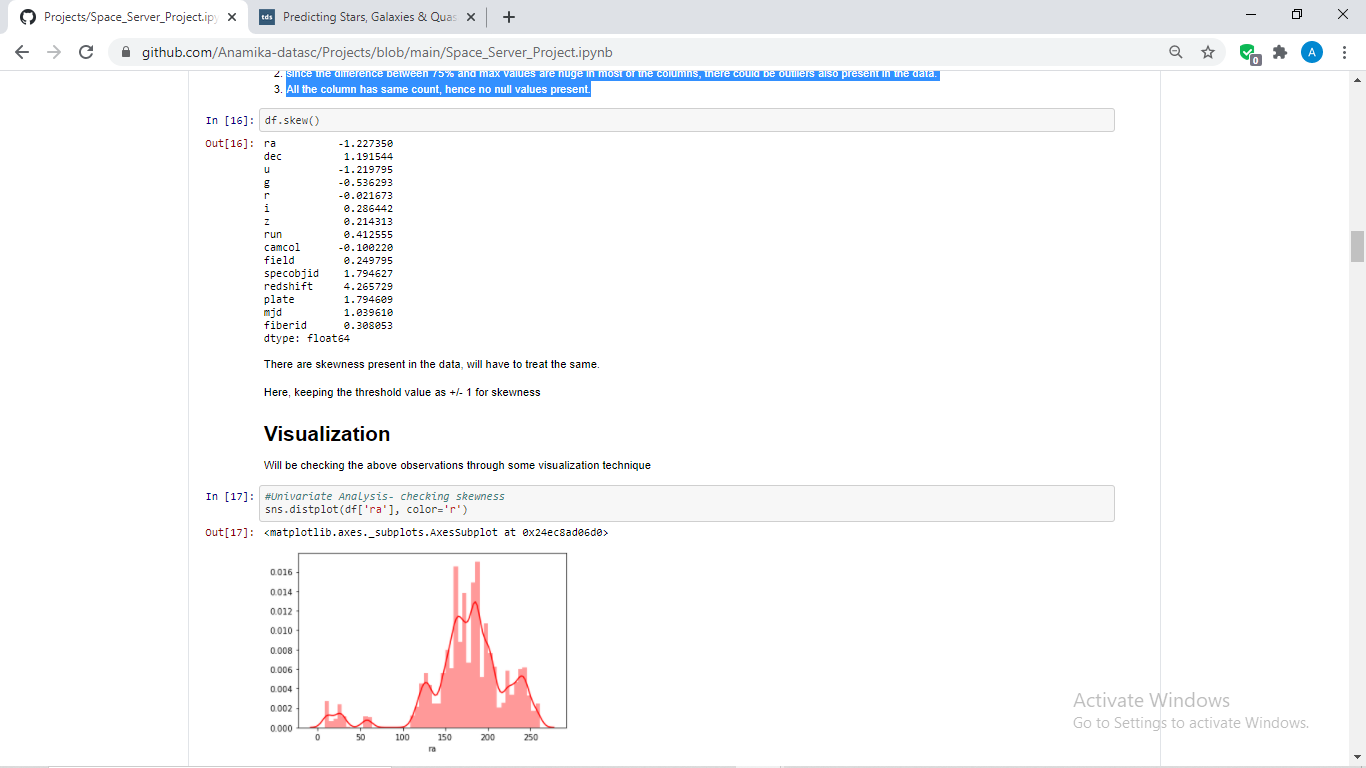


Key observations from summary statistics:

1. Mean is greater than median in ‘dec’, ‘run’, ‘plate’ columns, hence the data has skewness in them..
2. Since the difference between 75% and max values are huge in most of the columns, there could be outliers also present in the data.
3. All the column has same count, hence no null values present.

* **Checking Skewness:**

Let’s check the skewness with the dedicated method for checking the same- **df.skew()** which will show us the numerical values of the skewness in the data.

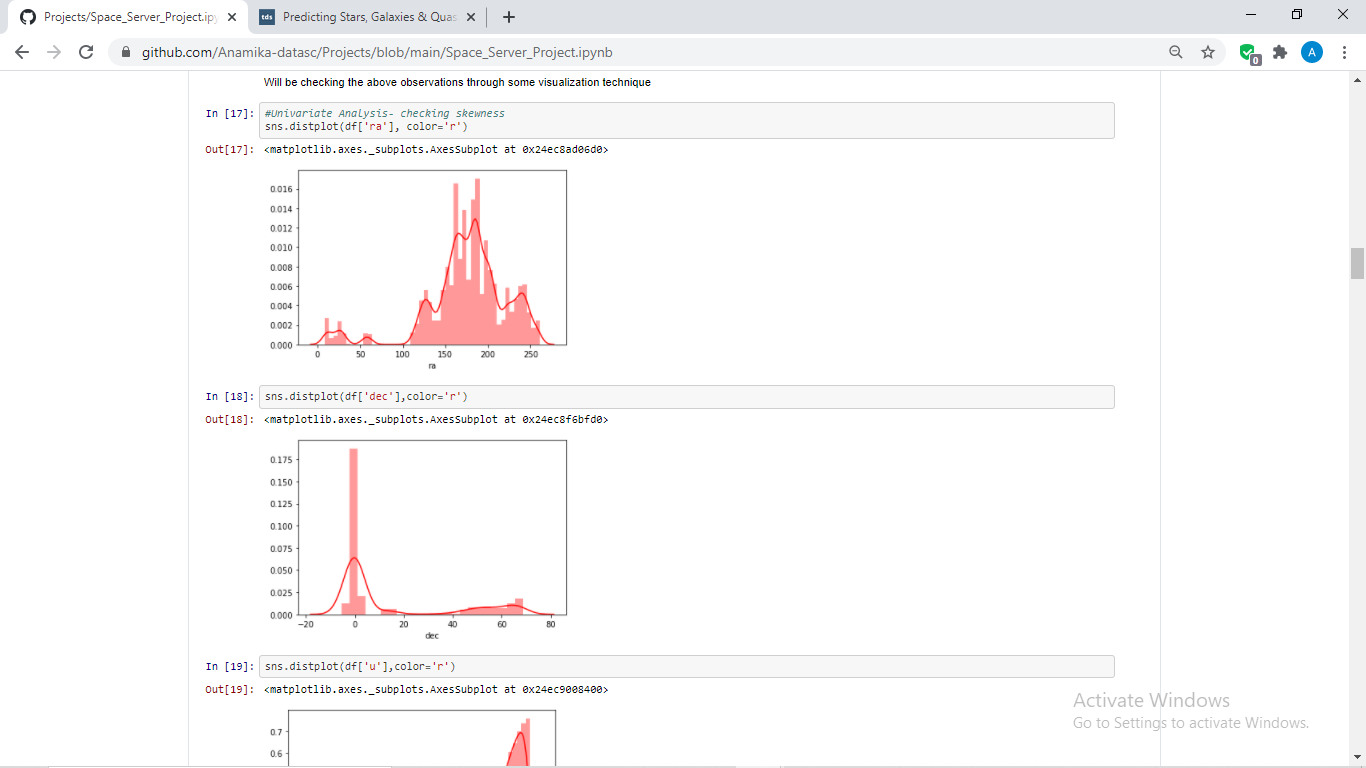
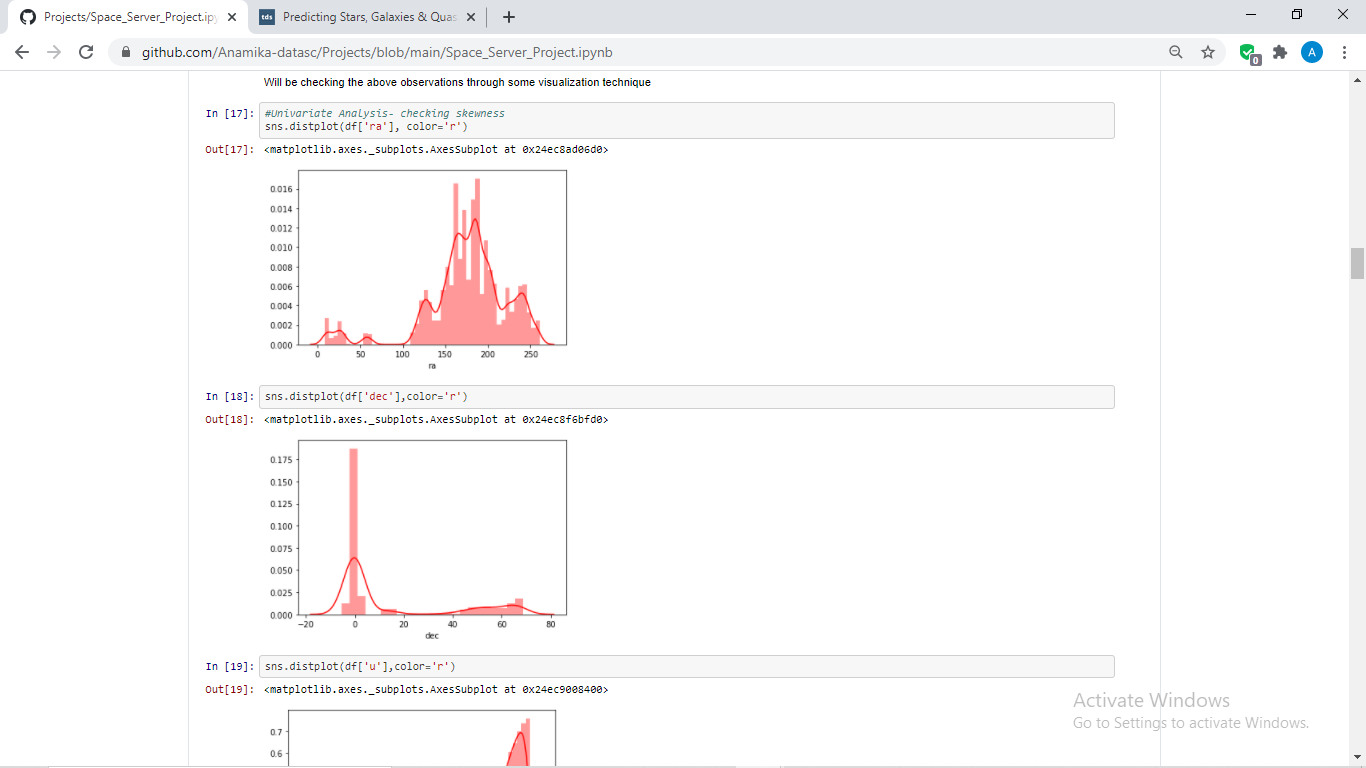


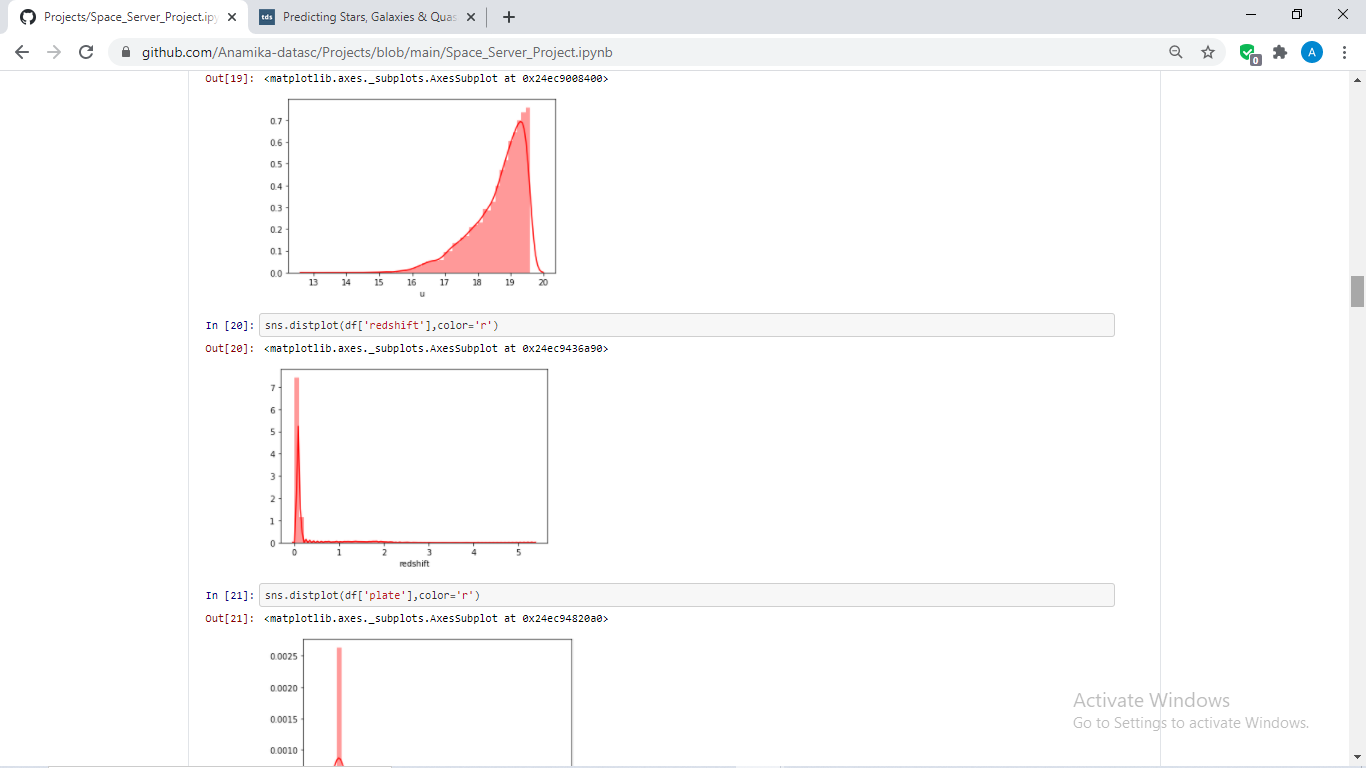
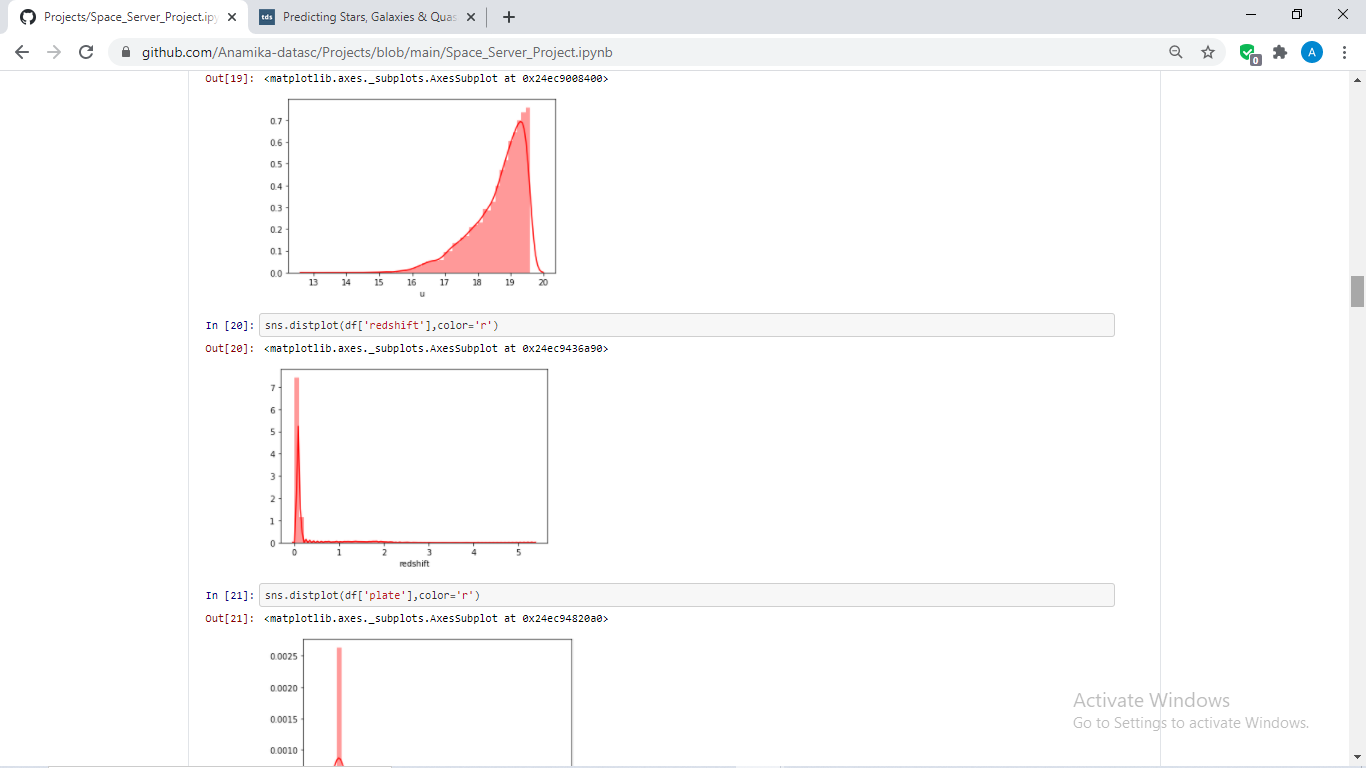
It is observed from above, there are skewness in various columns like ‘ra’, ‘dec’, ‘u’, ‘specobjid’ etc. This skewness needs to get treated.

* **Visualization:**

Now, will check the above observations and analysis about the data through some visualizations tools which are shown below;

* **Checking Skewness (Univariate Analysis)**

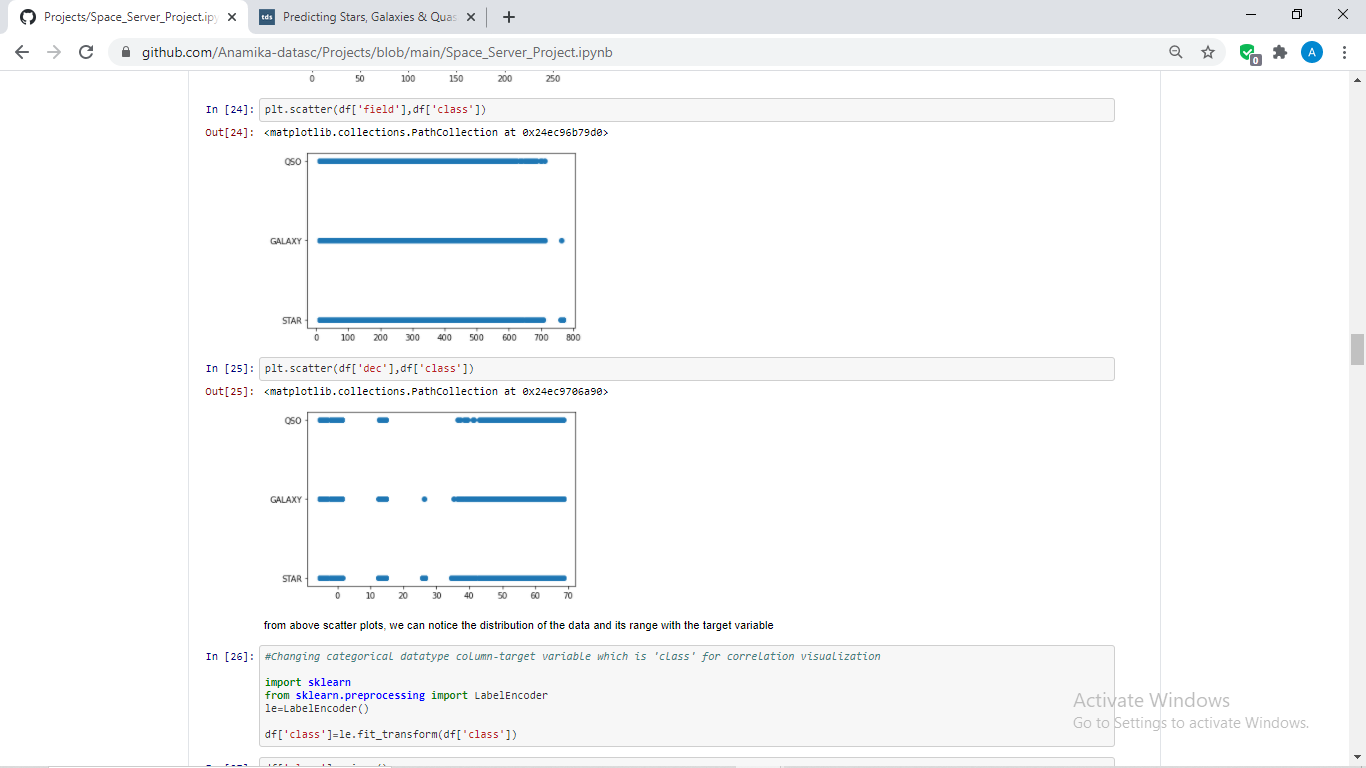
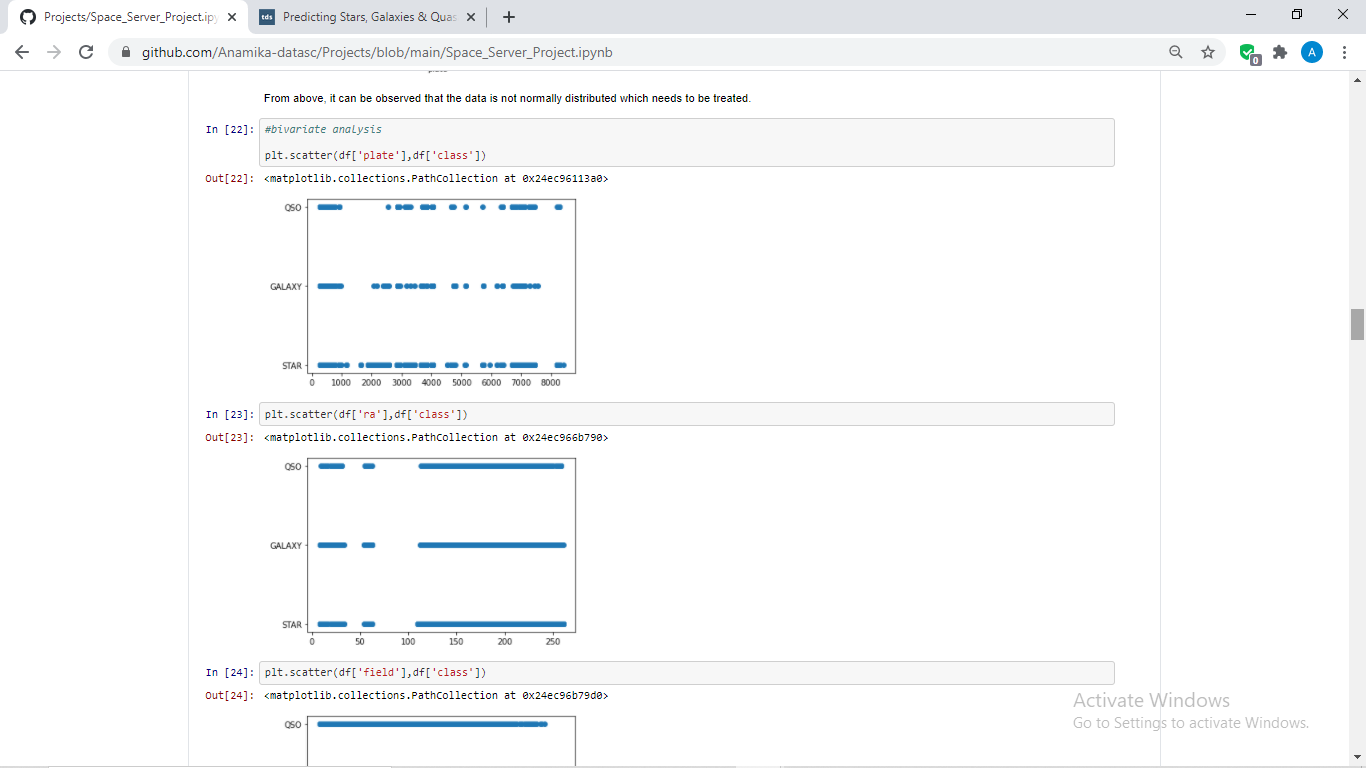




From above, it can be seen that the data does not have normal curve and has skewness in them which needs to be treated.

* **Checking Distribution(Bivariate Analysis):**

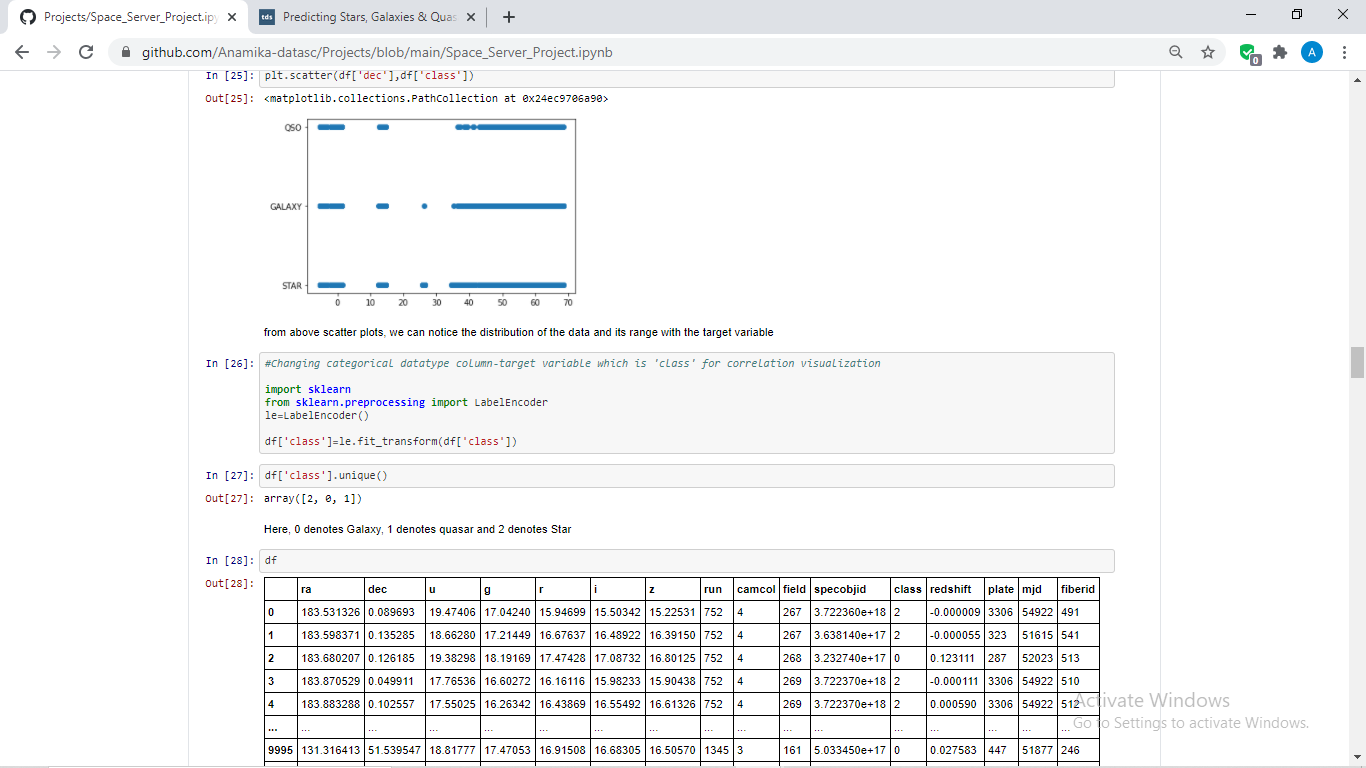
I have used here scatter plot to check the distribution and the range of the columns along with the target variables.



You can check the above distribution for all the columns, if wish to.

* **Label-Encoding:**

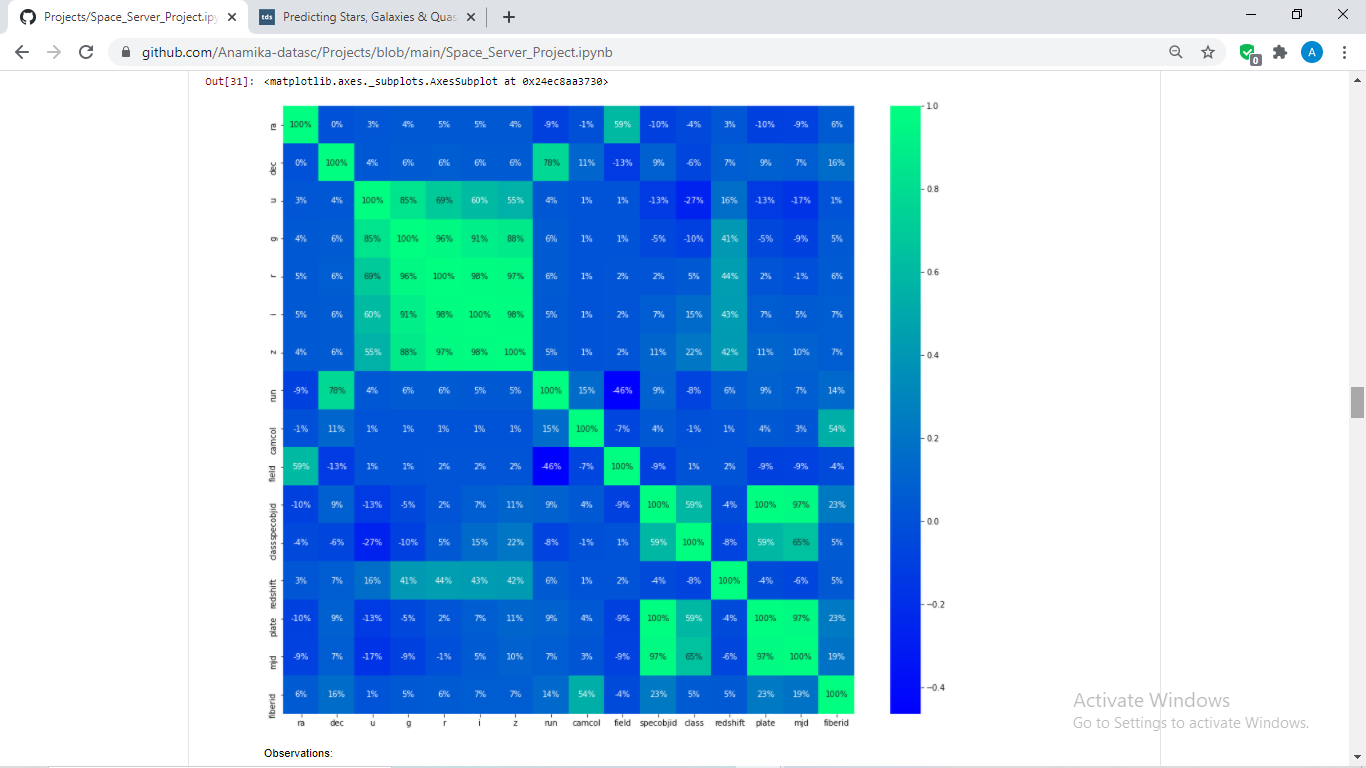
We will have to change the class column data type in this step itself, as heatmap and correlation syntax could not read the object data .



**Here, 0 denotes Galaxy, 1 denotes quasar and 2 denotes Star.**

**To Note here**, data science is not ruled based as many of us says that label encoding should be done at the pre-processing stage but as per the requirements and needs of our data we have done the label-encoding in between the EDA itself :D

* **Checking Correlation:**



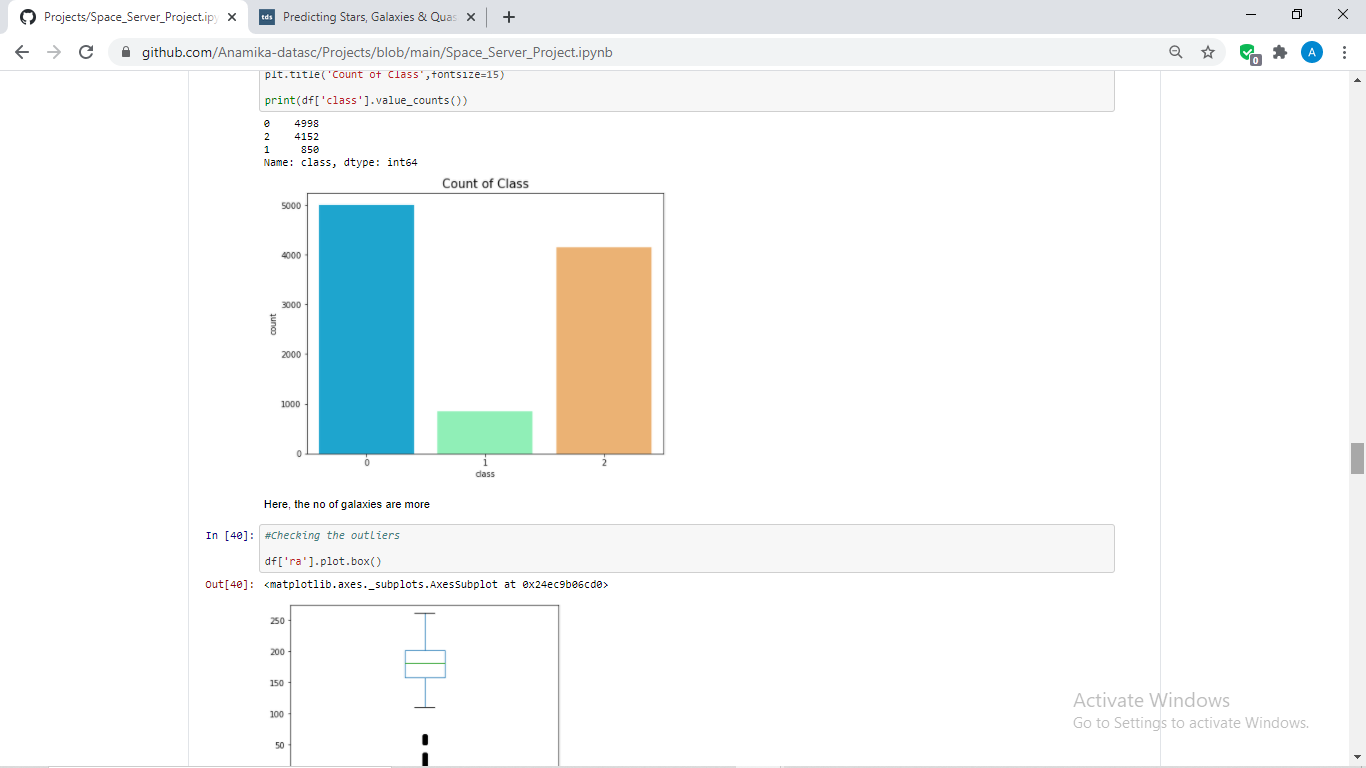
From the above heatmap, following observations has been noted:

1. ‘mjd’ has highest correlation of 65% with the target-class

2. ‘plate’ & ‘specobjid’ has 59%, other columns has distributed correlation with target.

3. ‘field’ and ‘camcol’ has very less correlation with target.

* **Checking Count of Target Variable:**

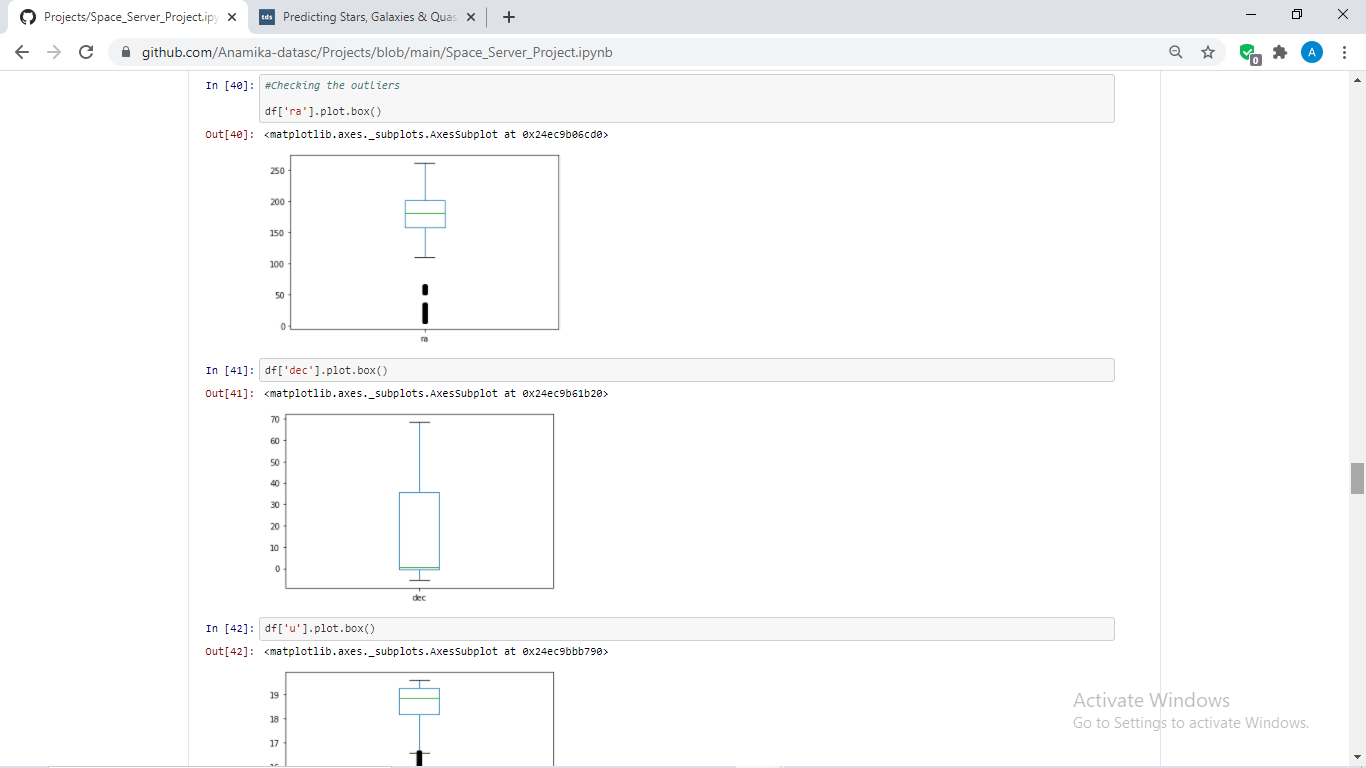
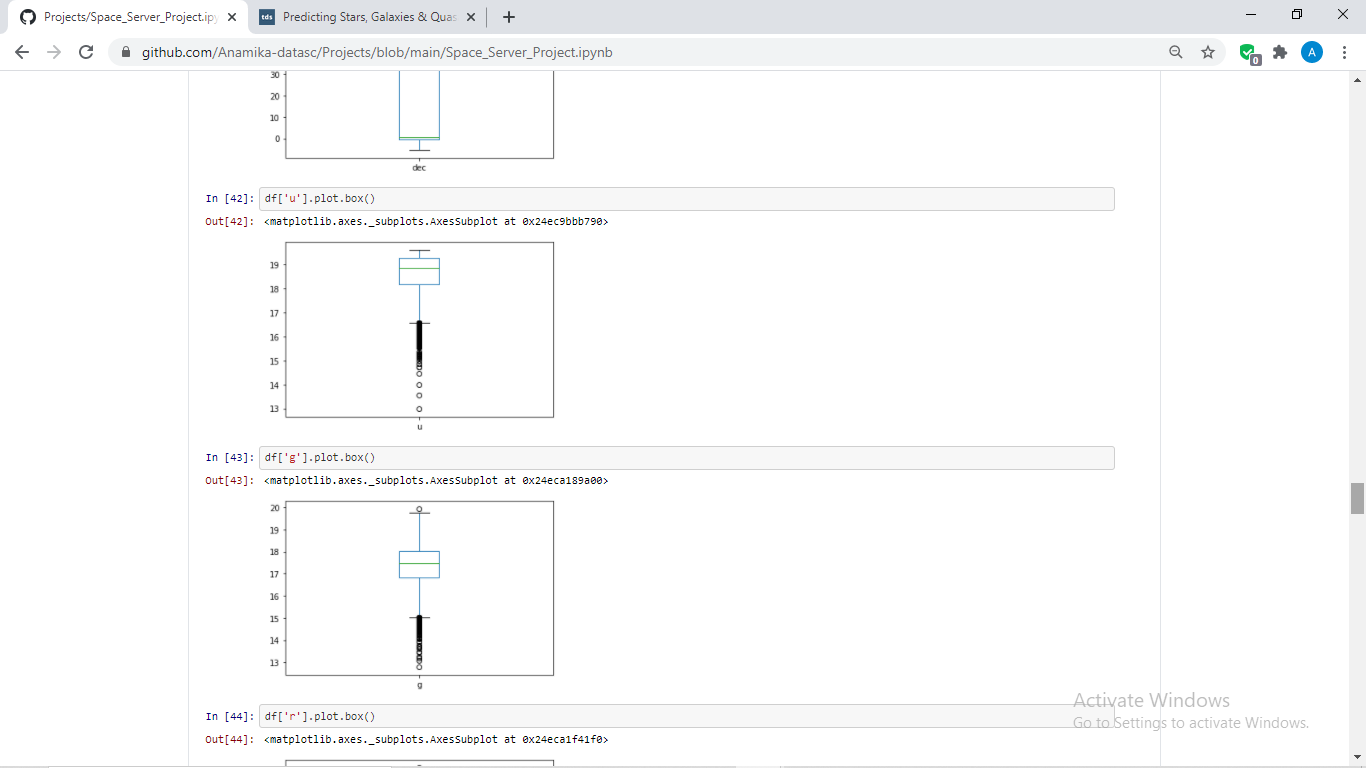
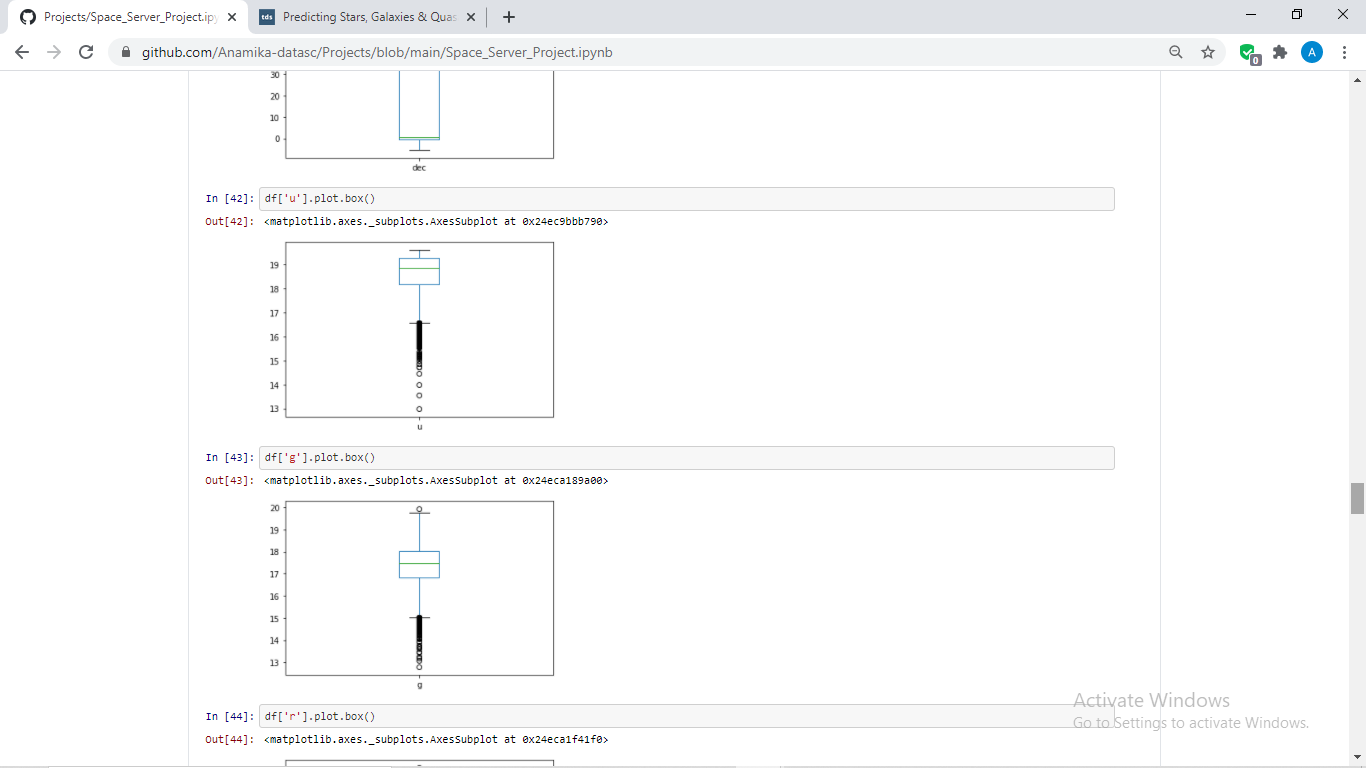
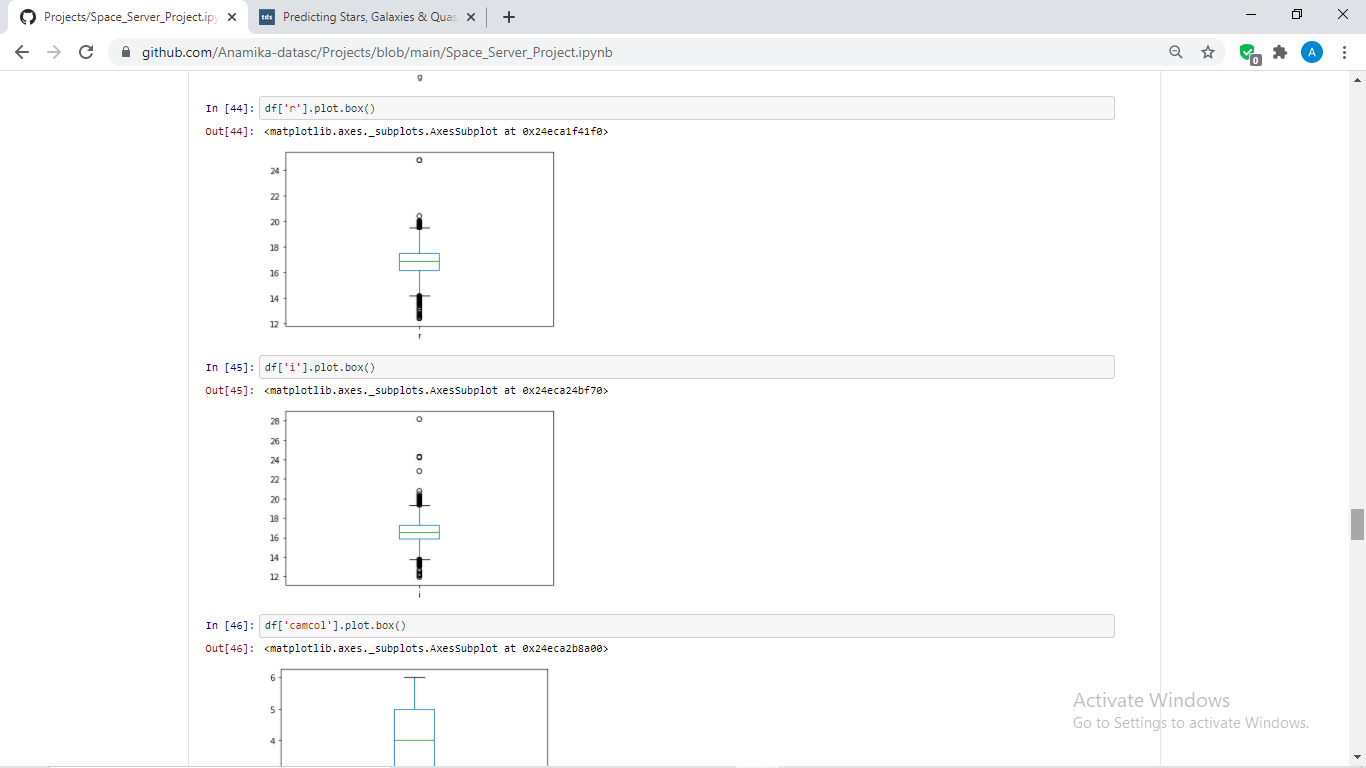


Here, number of galaxies and star are more whereas have very less amount of quasar.

Note: We cannot check the count of other variables as the range is very high. Tried and tested !

* **Checking Outliers:**

I have checked the outliers in the columns 1 by 1 as applying for loop has not given better view since the columns are more. Showing few of them as below:

As per the observations drawn out of these boxplots, found outliers in huge numbers which I will treat on the later stage.

**EDA CONCLUDING REMARKS**

From the above EDA and our data analysis done we have extracted the followings insights and knowledge about the data which will be useful in our data preparation:

1. 'objid' and 'rerun' columns has only single data in them and hence dropped them.
2. There is no Null values found in the dataset.
3. Target column is the only column which had categorical data in it and hence changed the same using label encoding.

**Here, 0 denotes Galaxy, 1 denotes quasar and 2 denotes Star.**

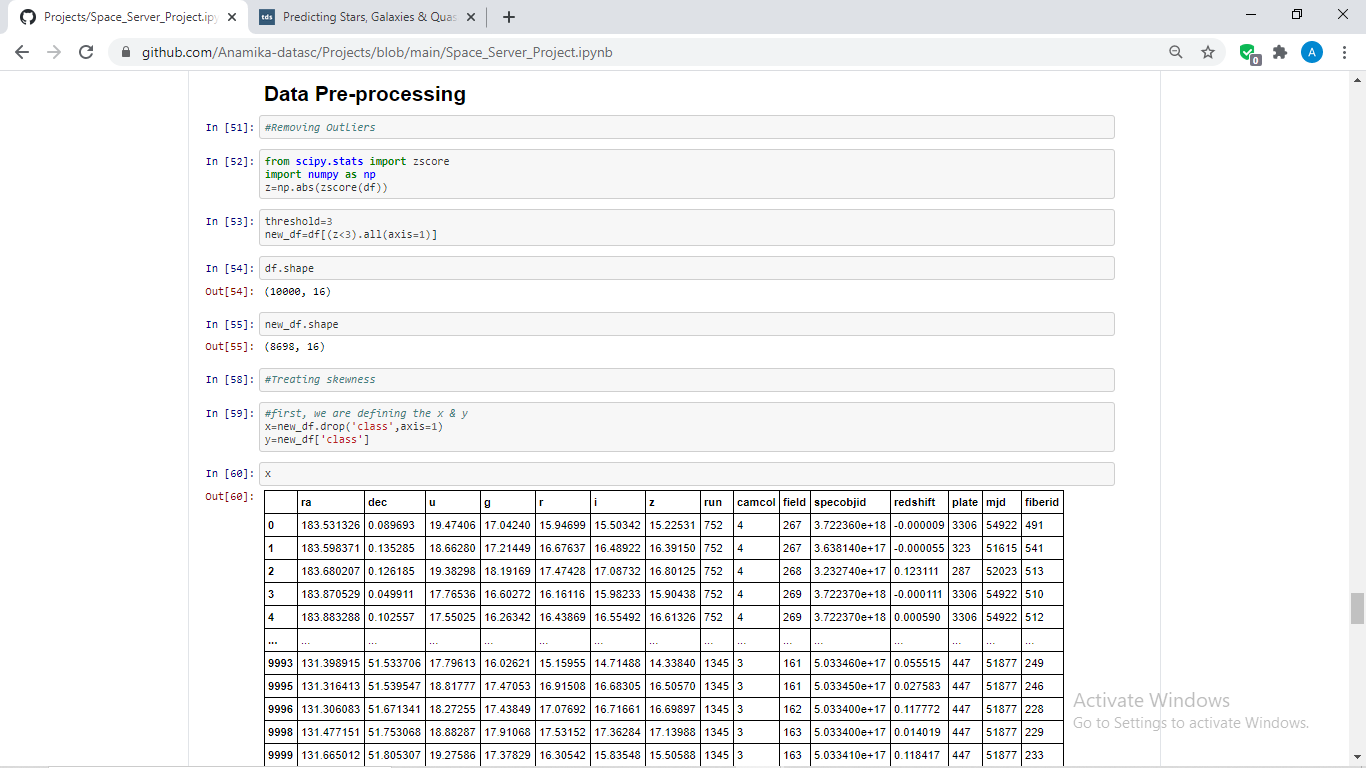
1. Data has some skewness presence which has been computed using **describe() and visualizations** method.
2. There is also a presence of outliers which has been observed using describe() and visualizations methods.
3. ‘mjd’ has highest correlation of 65% with the target-‘class’
4. ‘plate’ & ‘specobjid’ has 59%, other columns has distributed correlation with target.
5. ‘field’ and ‘camcol’ has very less correlation with target.
6. Number of galaxies and star are more whereas have very less amount of quasar.

With the help of above EDA we are now at a better position to do the data preparation.

**DATA-PREPARATION**

We will now perform some data pre-processing which involves data cleaning and preparation of the data. Steps of the same are listed below:

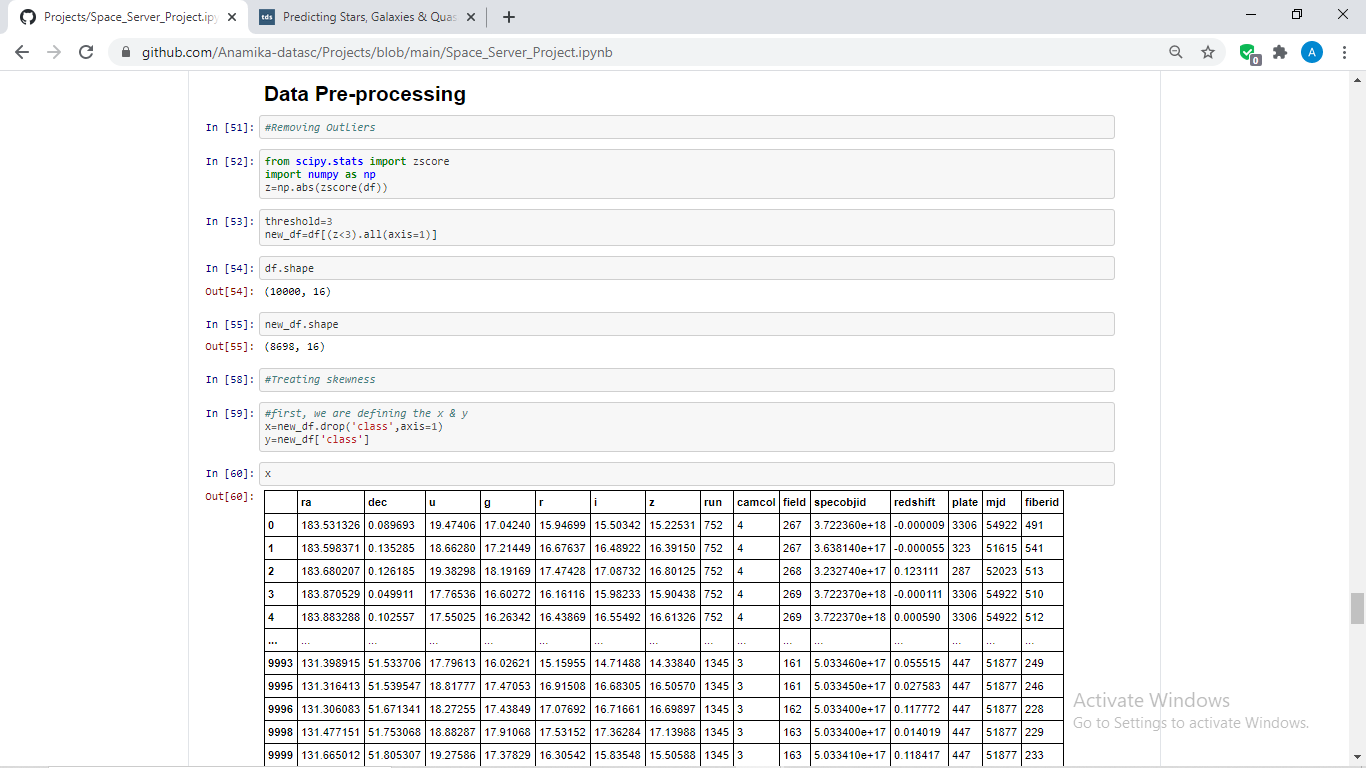
1. Some of the data cleaning has already been performed at the time of data analysis as it was the requirements for further processing dropping non-useful and categorical column like 'objid' and 'rerun'.
2. Chnaged the target column-‘class’ as it was having object datas in it, other columns has continuous and int/float data in them.
3. Since the correlation is distributed among the columns we have not dropped any columns as all the columns has some amount of correlation with the target variable. Also, dropping column is not advisable.
4. Some columns have outliers in them and hence will treat the outliers using **z score method** as below:

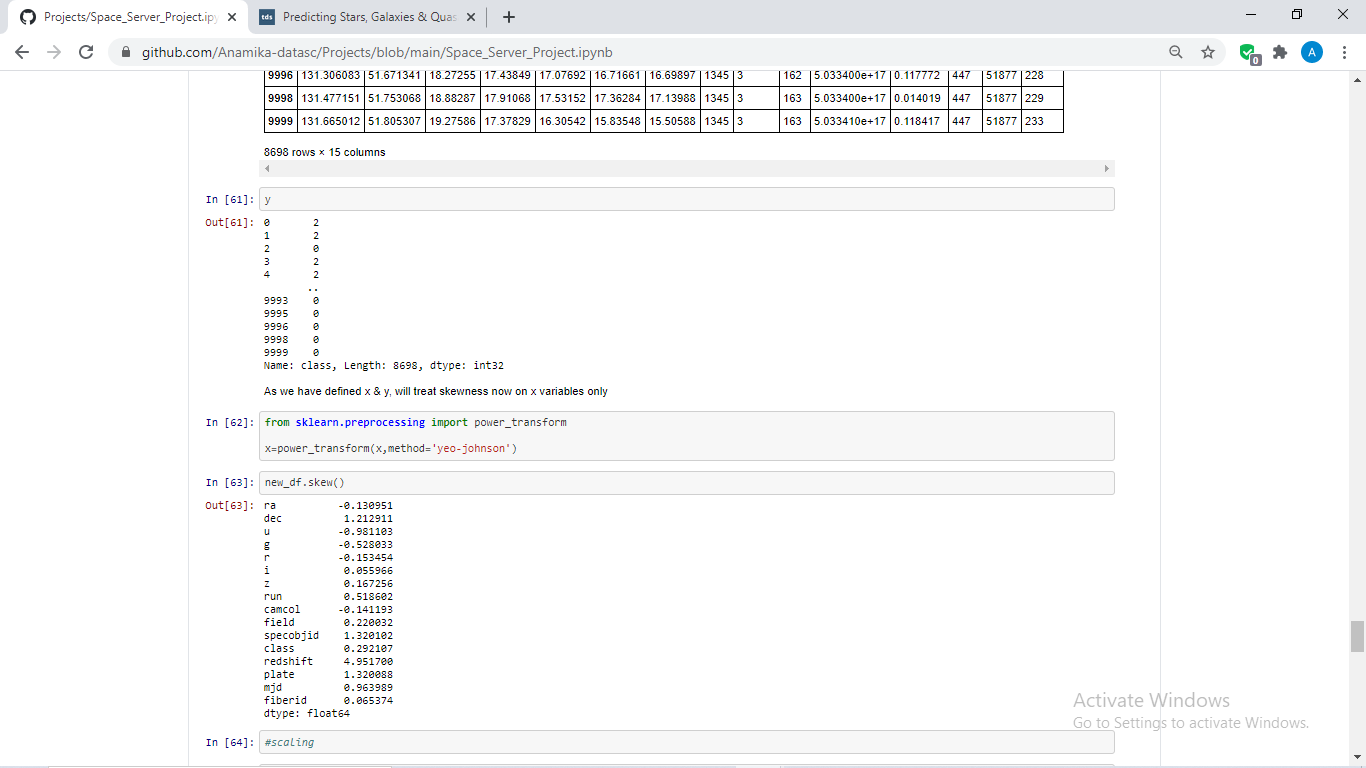


New shape of the data after removal of outliers is (8698,16) can bear this loss of data as we still have huge amount of data **but in a real case scenario, you may have to use certain different methods to remove outliers where minimal loss of data is observed as data loss of this amount is not acceptable by few clients.**

1. Since the data has skewness, will have to treat the same but we will remove skewness from the x variables only, as removing data from y variable could hamper our predictions.

For the same, will have to define the x and y variable first.

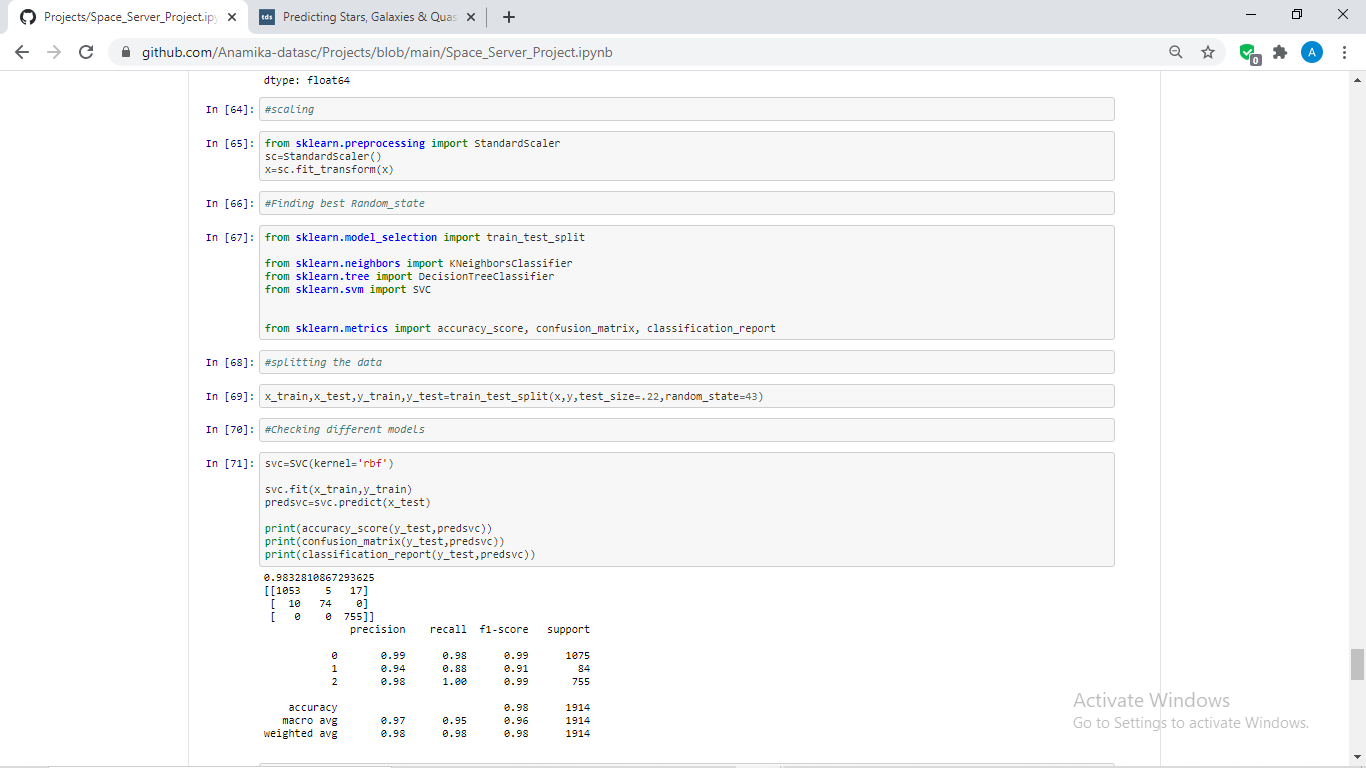




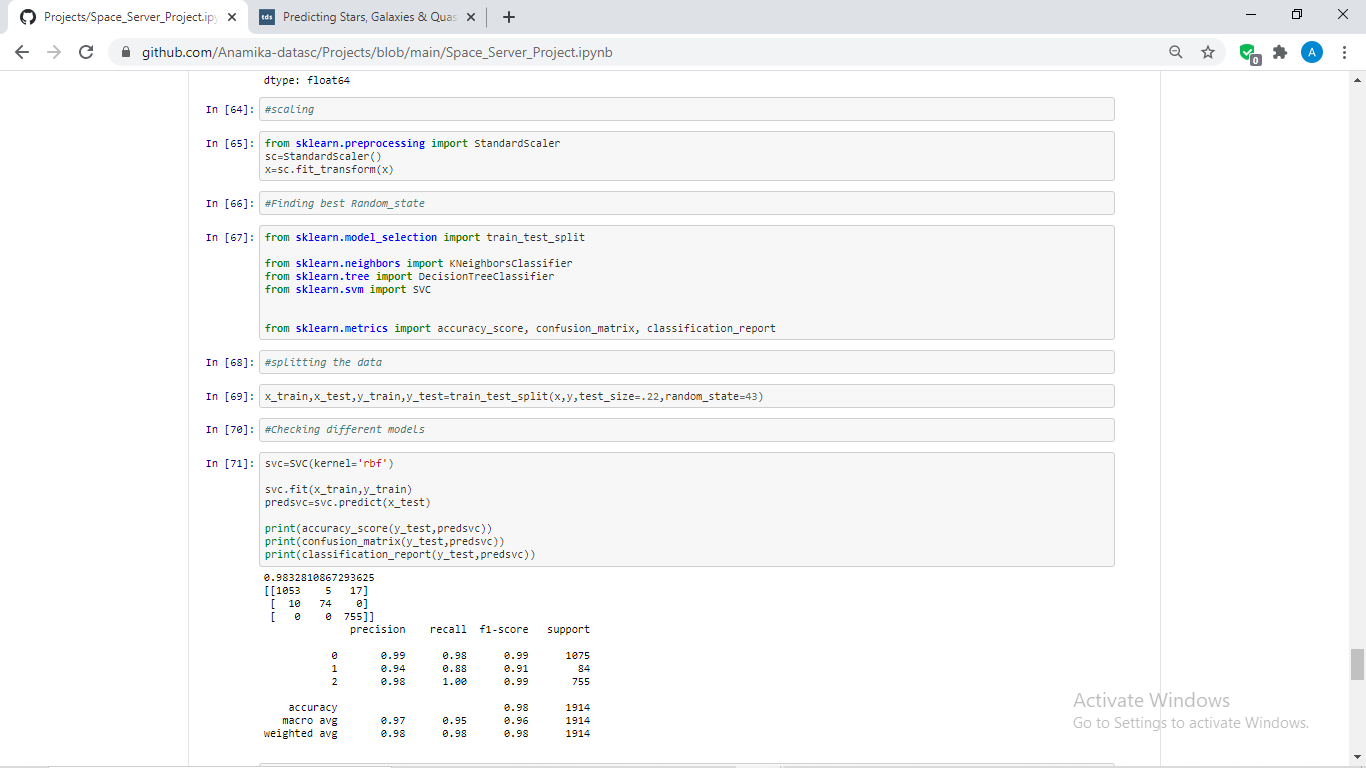
Treated the skewness using **yeo-johnson** method as this is the most affective for skewness removal.

**MODEL BUILDING**

* First, will sclae the data as this data has varied and very odd ranges, scaling will increase the algortithms performance to some extent.



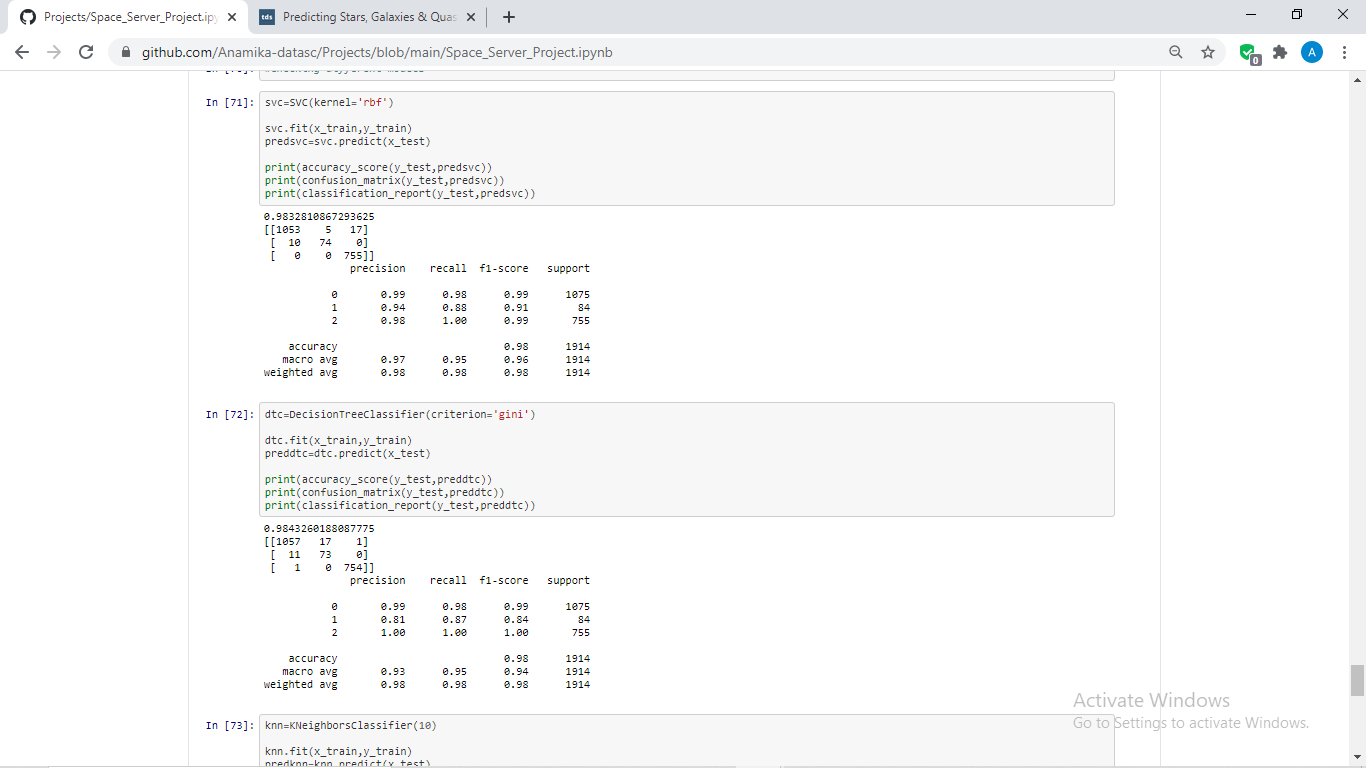
* Now, will import the required libraries for model building

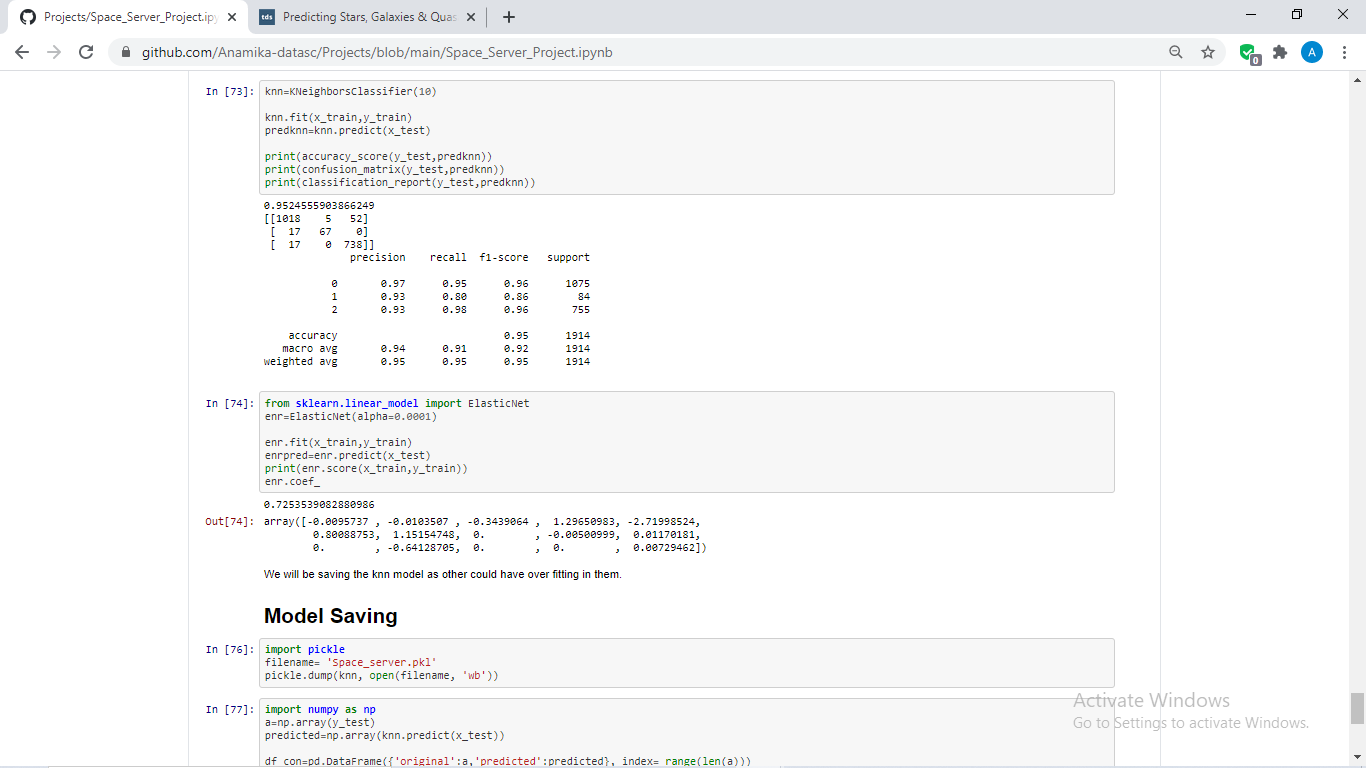


* Splitting the data as below:

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=.22,random\_state=43)

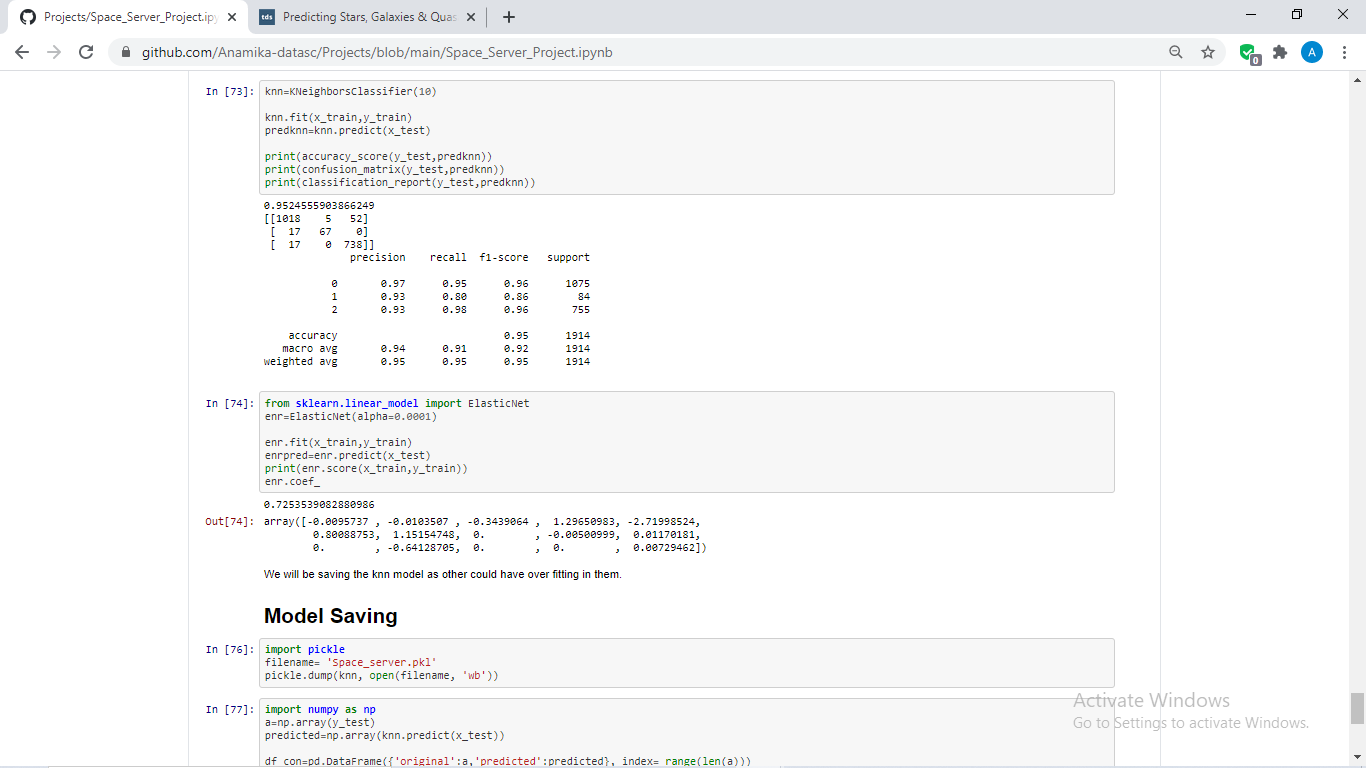
* Now, we are passing our train and test data to different models for the training and the testing purpose and we will check the scores of them with the help of matrices like accuracy\_score, confusion\_matrix and classification\_report. You can check the codes and the model performances as below:





From above, we can notice that all the three models are performing quite well but this could be the case of **over-fitting** as the data range is quite large.

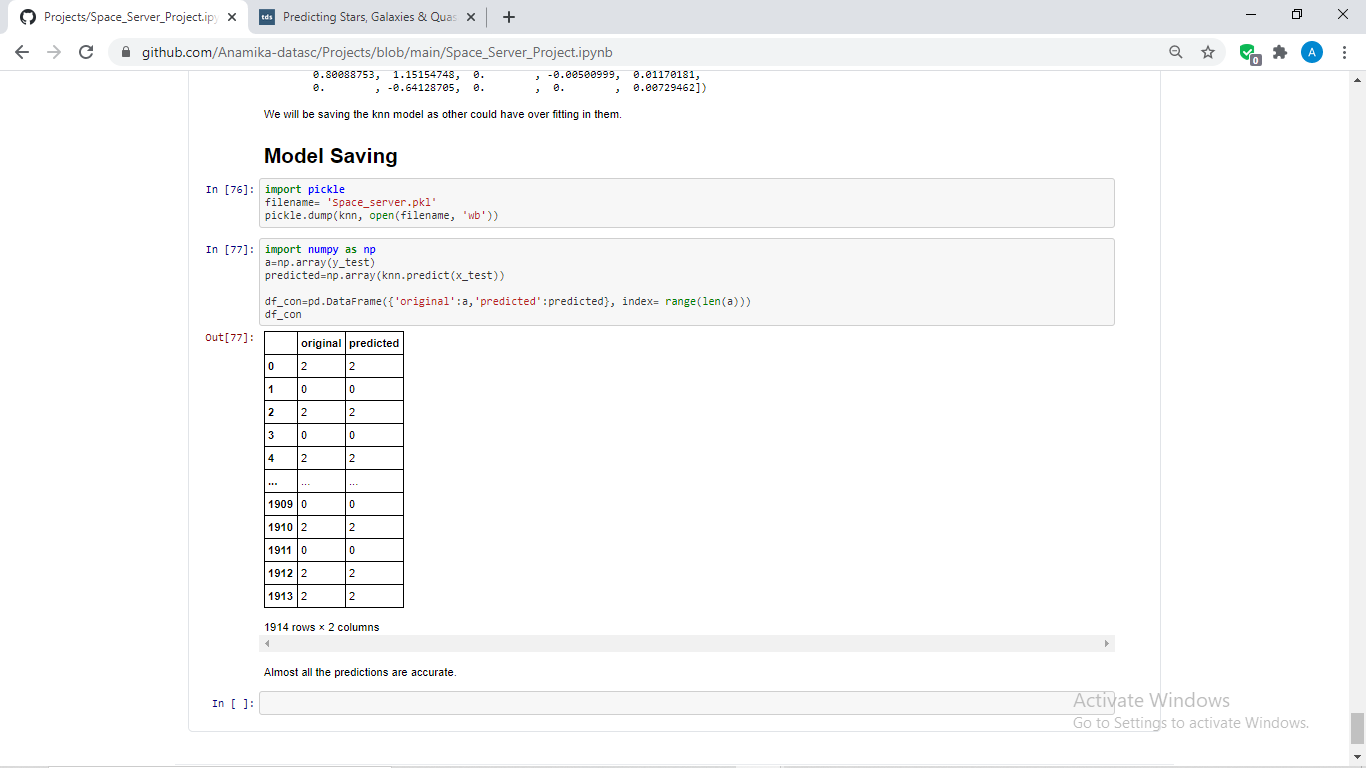
To overcome this over-fitting we will use one of the **regularization** method called **ElasticNet**. Will compare the score of the same**.**



**Here, accuracy comes to 72% which means the models are over-fitted, hence will save the KNN model as it has comparatively less accuracy than other models which is of 95%**

**CONCLUDING REMARKS AND MODEL SAVING**

After model saving will check the accuracy of the model by comparing original which is y\_test and the predicted data, which is done as below:



**We can observe, the model is performing very well with almost 95% accuracy** accuracy but there are still some scope of improvement in prediction which can be achieved by applying certain hypertuning methods, checking the best random state, with some boosting and ensemble methods and testing more models.

In summary, we have demonstrated a very interesting Space Server dataset which can be used for applying various ML tools as it has huge data in it. We have seen how ML is useful in detecting celestial bodies turning us to a space scientist ;)

You might have different way of approaching this data and may use better visualization tools, model building techniques etc. Please feel free to share your insights on this data.

Author

(Anamika. Jha)