**Classification of the class object based on several variables to predict whether the celestial object is a star or a galaxy or a quasar.**

**Abstract:** *The Sloan Digital Sky Survey has created the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one-third of the sky, and spectra for more than three million astronomical objects. It allows to learn and explore all phases and surveys - past, present, and future - of the SDSS.* *Data is obtained using Sky Server SQL Search, it contains 17 total variables with one dependent variable named “class”, class variable has three categorical values named “star”, “galaxy “and “quasar”. We tried to predict the classification of these variables depending upon the other independent variables obtained from the survey. We followed the CRISP-DM approach for this project.*

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **No.** | **Name** | **Page.No** |
| **1** | **Introduction** | **3** |
| **1.1** | **Questions & Findings** | **4** |
| **1.2** | **Data Description** | **4** |
| **2** | **Data Preparation & Exploratory Analysis** | **5** |
| **2.1** | **PCA** | **8** |
| **3** | **Model Creation and Implementation of Classification Models** | **9** |
| **4** | **Feature Importance** | **12** |
| **5** | **Conclusion** | **12** |
| **6** | **References** | **12** |

# 1. INTRODUCTION

**Dataset Link :** <https://www.kaggle.com/muhakabartay/starter-sloan-digital-sky-survey-dr16>

**Data Type:** Sloan Digital Sky Survey with [DR16](https://www.sdss.org/dr16) Server Data about Galaxies, Stars and Quasars.

# Prediction Models: Classification of celestial object ‘class’ (galaxy, star, quasar) utilizing Decision Tree, Decision trees, Linear Discriminant Analysis, Support Vector Machine, Random Forest, Gaussian Naïve Bayes, K-Nearest Neighbors and Logistic regression models.

**Parameters:** The models are compared based on the accuracies of different model classifier test and training data set.

The Sloan Digital Sky Survey or SDSS is a major multi-spectral imaging and spectroscopic redshift survey using a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States. The project was named after the Alfred P. Sloan Foundation, which contributed significant funding.

Data collection began in 2000 the final imaging data release (DR9) covers over 35% of the sky, with photometric observations of around nearly 1 billion objects, while the survey continues to acquire spectra, having so far taken spectra of over 4 million objects. The main galaxy sample has a median redshift of z = 0.1; there are redshifts for luminous red galaxies as far as z = 0.7, and for quasars as far as z = 5; and the imaging survey has been involved in the detection of quasars beyond a redshift z = 6.

Sloan Digital Sky Survey current DR16 Server Data release with Galaxies, Stars and Quasars.

Data is obtained using  SkyServer SQL Search  with the command below:  
This query does a table JOIN between the imaging (PhotoObj) and spectra  
(SpecObj) tables and includes the necessary columns in the SELECT to upload  
the results to the SAS (Science Archive Server) for FITS file retrieval.

*SELECT TOP 100000*

*p.objid,p.ra,p.dec,p.u,p.g,p.r,p.i,p.z,p.run,p.rerun,p.camcol,p.field,*

*s.specobjid,s.class,s.z as redshift, s.plate,s.mjd,s.fiberid*

*FROM PhotoObj AS p*

*JOIN SpecObj AS s ON s.bestobjid = p.objid*

*WHERE*

*p.u BETWEEN 0 AND 19.6*

*AND g BETWEEN 0 AND 20*

Thus, the table results from the above which joins two tables, “PhotoObj”- contains photometric data and “SpecObj” which contains spectral data.

17 variables(double) and 1 additional variable (char) ‘class’ is obtained and used for the project.

The ‘class’ can be predicted from the other 17 variables, with some data cleaning, visualizations, preprocessing and finally using different models to check the accuracy of the prediction.

* 1. **QUESTIONS AND FINDINGS**
* What is the good model to classify the celestial objects?
* What variables are useful is classifying the celestial objects?
* Which is the most important Independent variable in the study?
  1. **DATA DESCRIPTION**

A class of object can be predicted from the following 17 variables. A four-color UVGR is an intermediate-band photometric system, The Sloan Digital Sky Survey (SDSS) photometric system has a new five-color (u′ g′ r′ i′ z′) wide-band CCD system. The variables 'run', 'rerun', 'camcol' and 'field' are features which describe a field within an image taken by the SDSS. A field is basically a part of the entire image corresponding to 2048 by 1489 pixels. A field can be identified by: - run number, which identifies the specific scan, - the camera column, or "camcol," a number from 1 to 6, identifying the scanline within the run, and - the field number. The field number typically starts at 11 (after an initial rampup time) and can be as large as 800 for particularly long runs. - An additional number, rerun, specifies how the image was processed.

1. **Objid:** object id
2. **Ra:** angular distance measured eastward along the celestial equator from the Sun at the March equinox to the hour circle of the point above the earth.
3. **Dec:** declination
4. **U, g, r, i, z:** response of the 5 bands of the telescope
5. **Run:** run number
6. **Rereun:** rerun number
7. **Camcol:** camera column
8. **Field:** field number
9. **Specobjid:** object id
10. **Class:** object class (galaxy, star or quasar object)
11. **Redshift:** final redshift (a phenomenon where electromagnetic radiation from an object undergoes an increase in wavelength)
12. **Plate:** plate number
13. **Mjd:** modified Julian date of observation
14. **Fiberid:** fiber id
15. **DATA PREPARATION & EXPLORATORY ANALYSIS**

Information about the variables including the data types of each column and memory usage of the entire data.

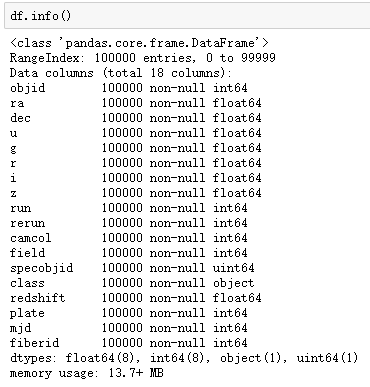


Fig.1: Information about all the variables

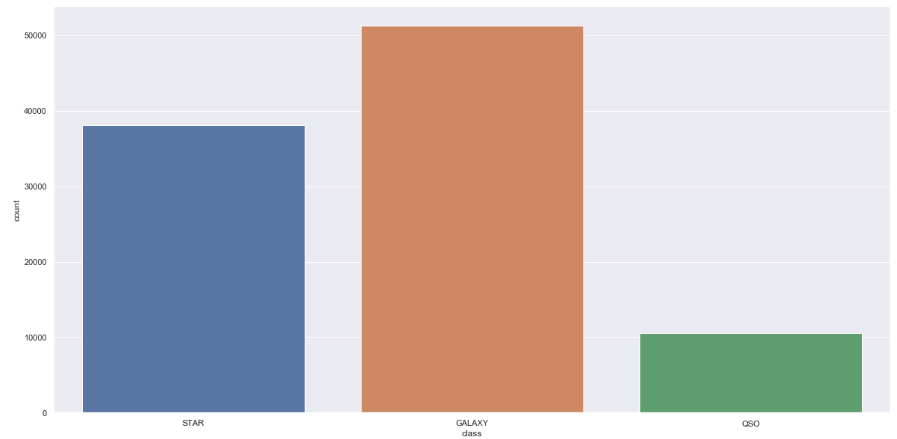


Fig 2: Count plot of the objects in dependent variable

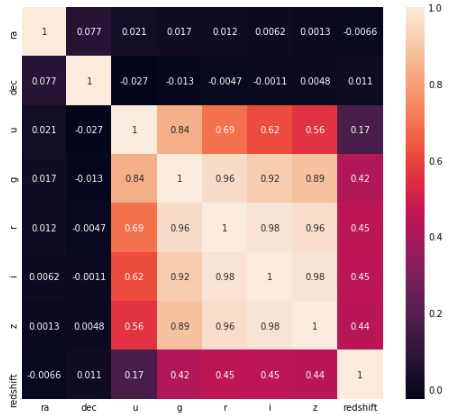


Fig 3: Heat Map

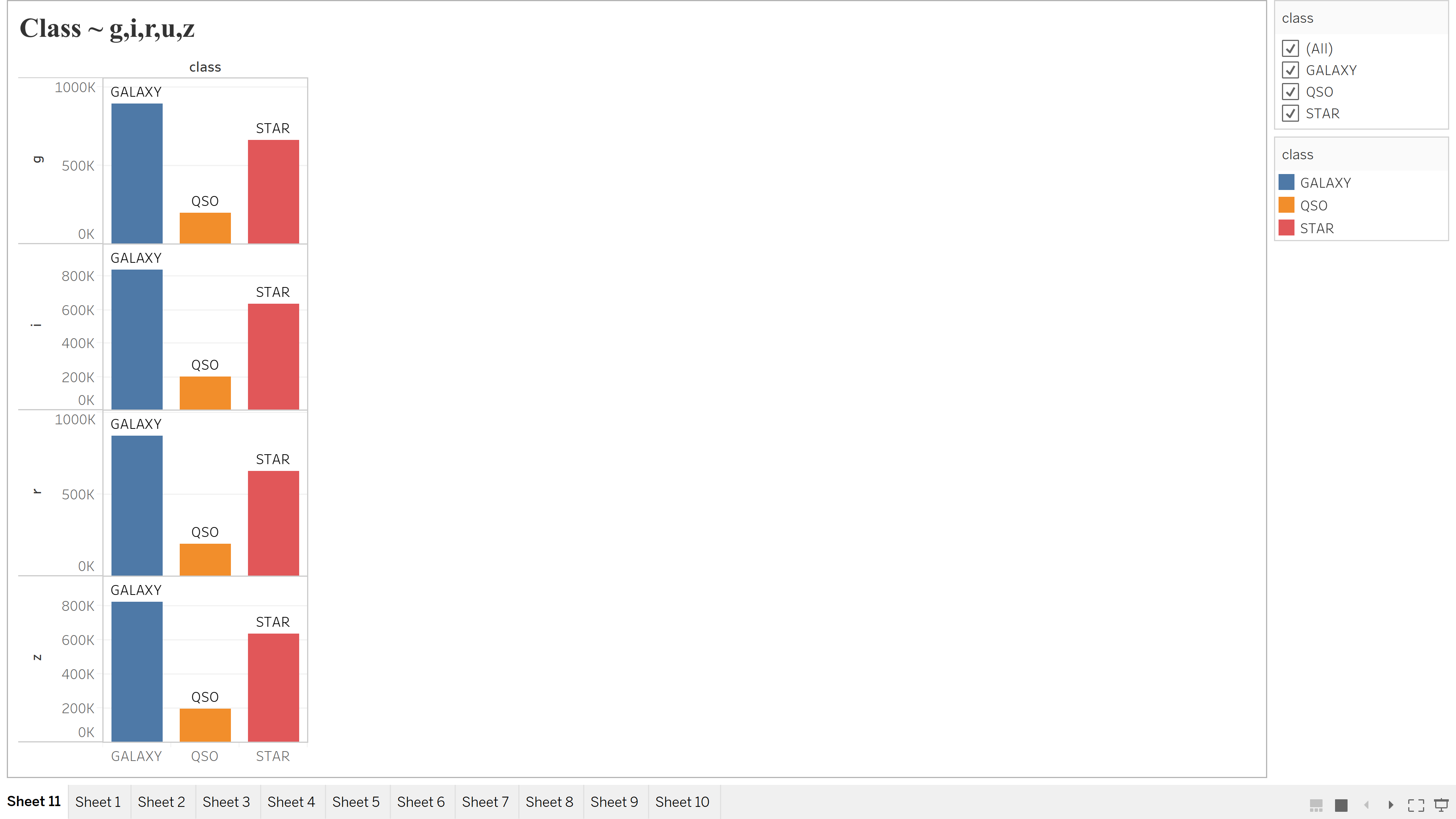


Fig 4: Class as the response variable for g,i,r,u,z as the predictors.

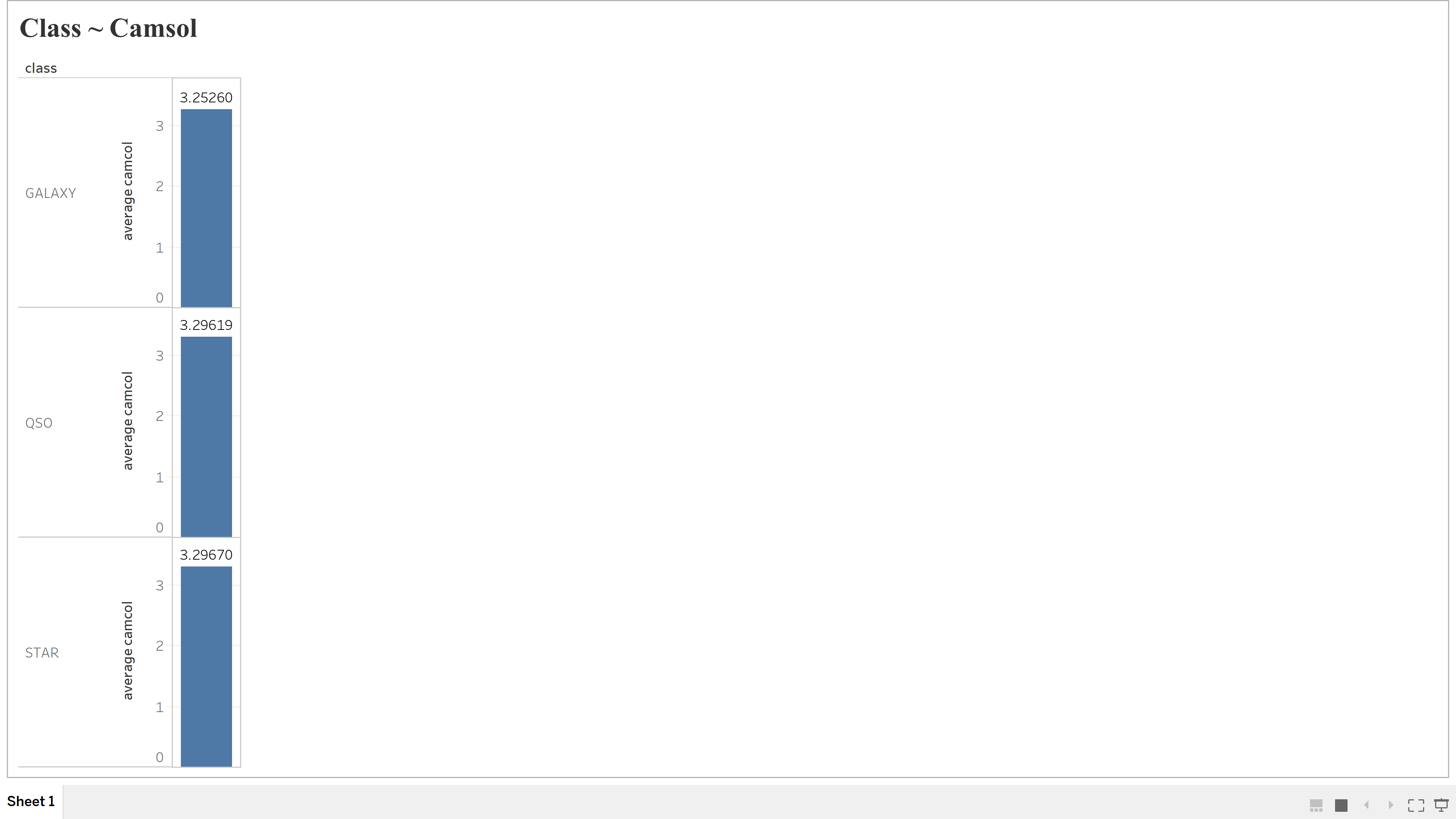


Fig 5: Box Plot of Class as response variable for Camsol predictor variable.

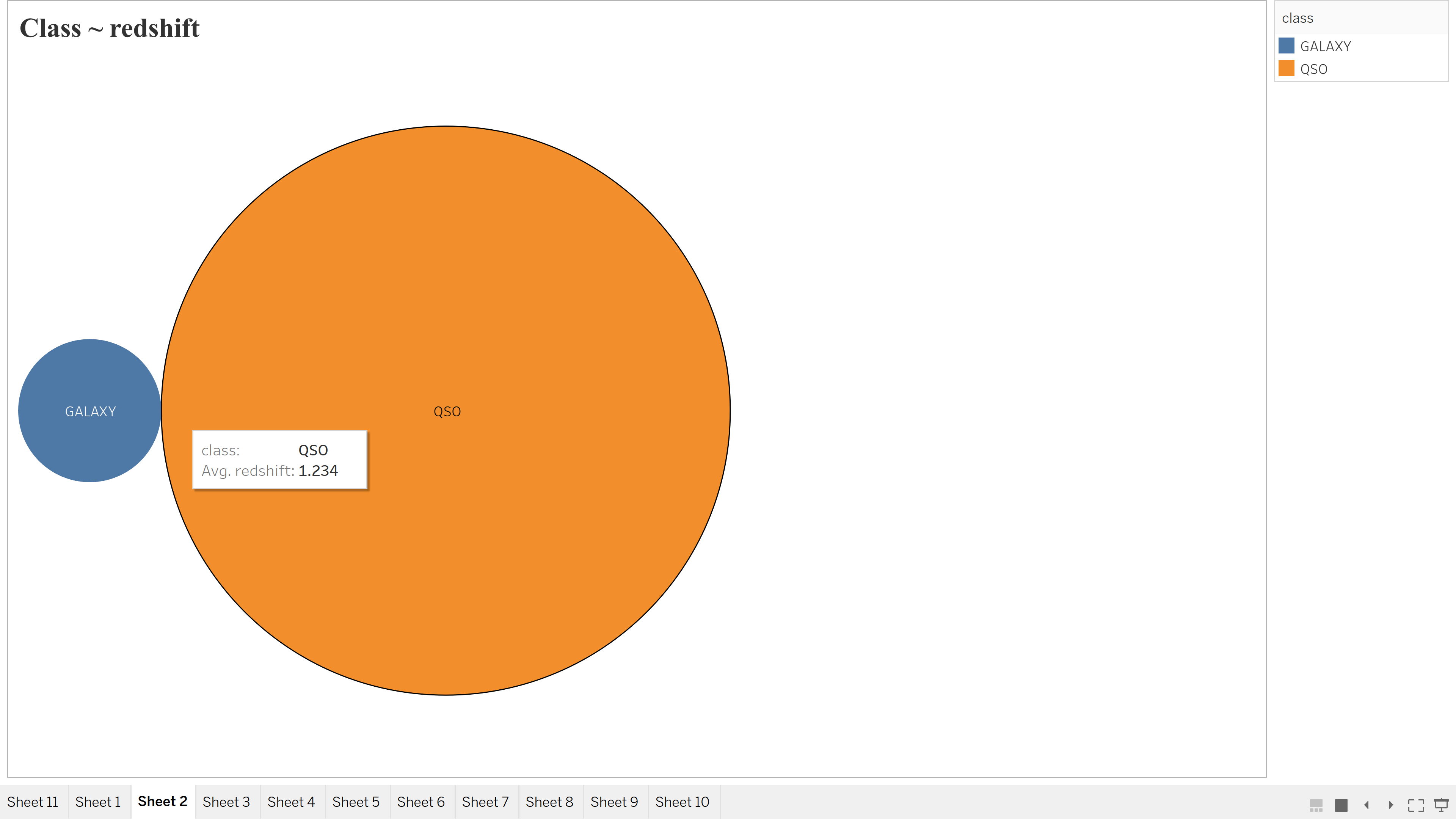


Fig 6: Understanding Class with redshift variable for the galaxy and Quasar category.

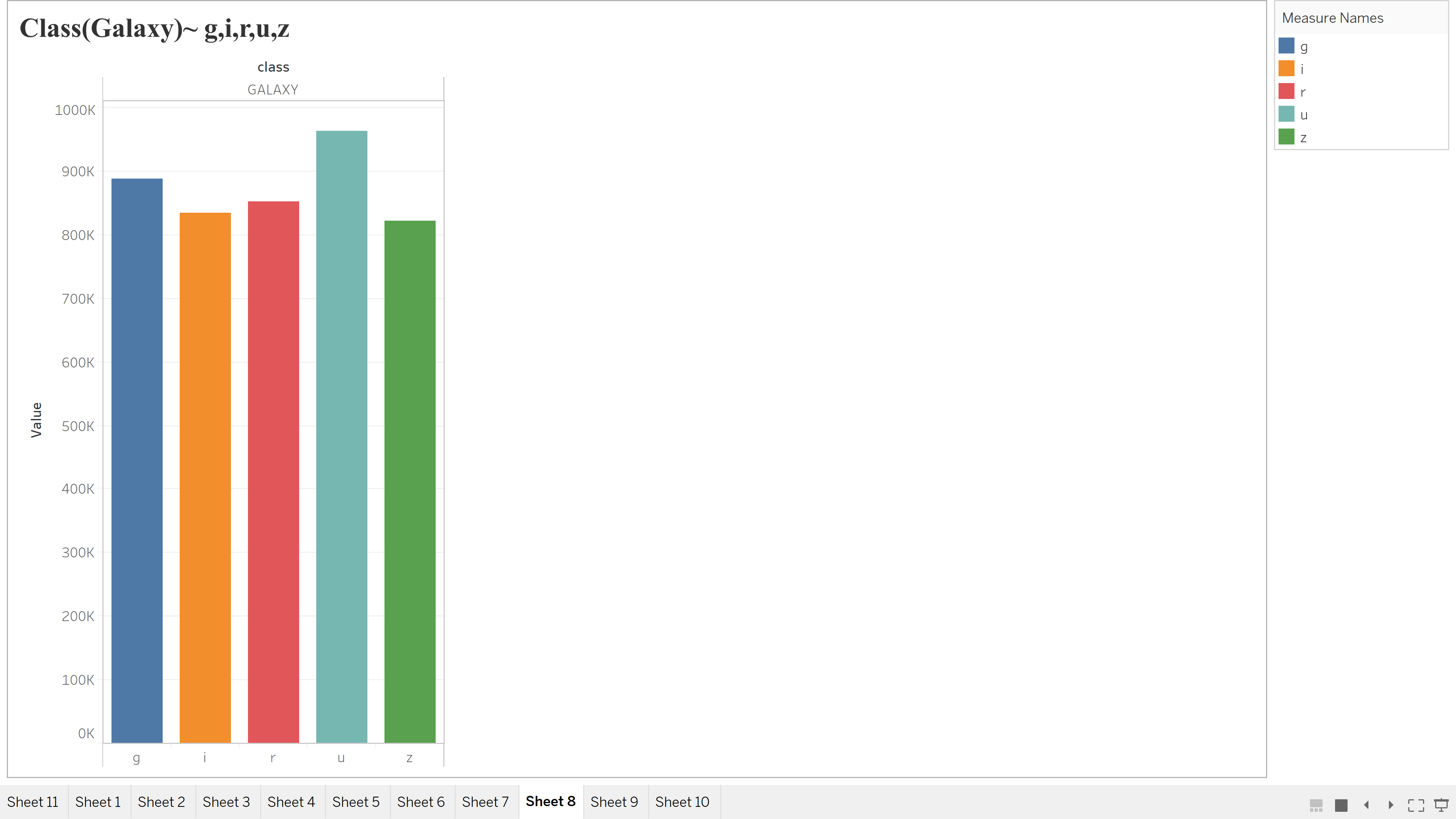


Fig 7: Bar chart explaining the range of predictor variable viz g,i,r,u,z for galaxy class category.

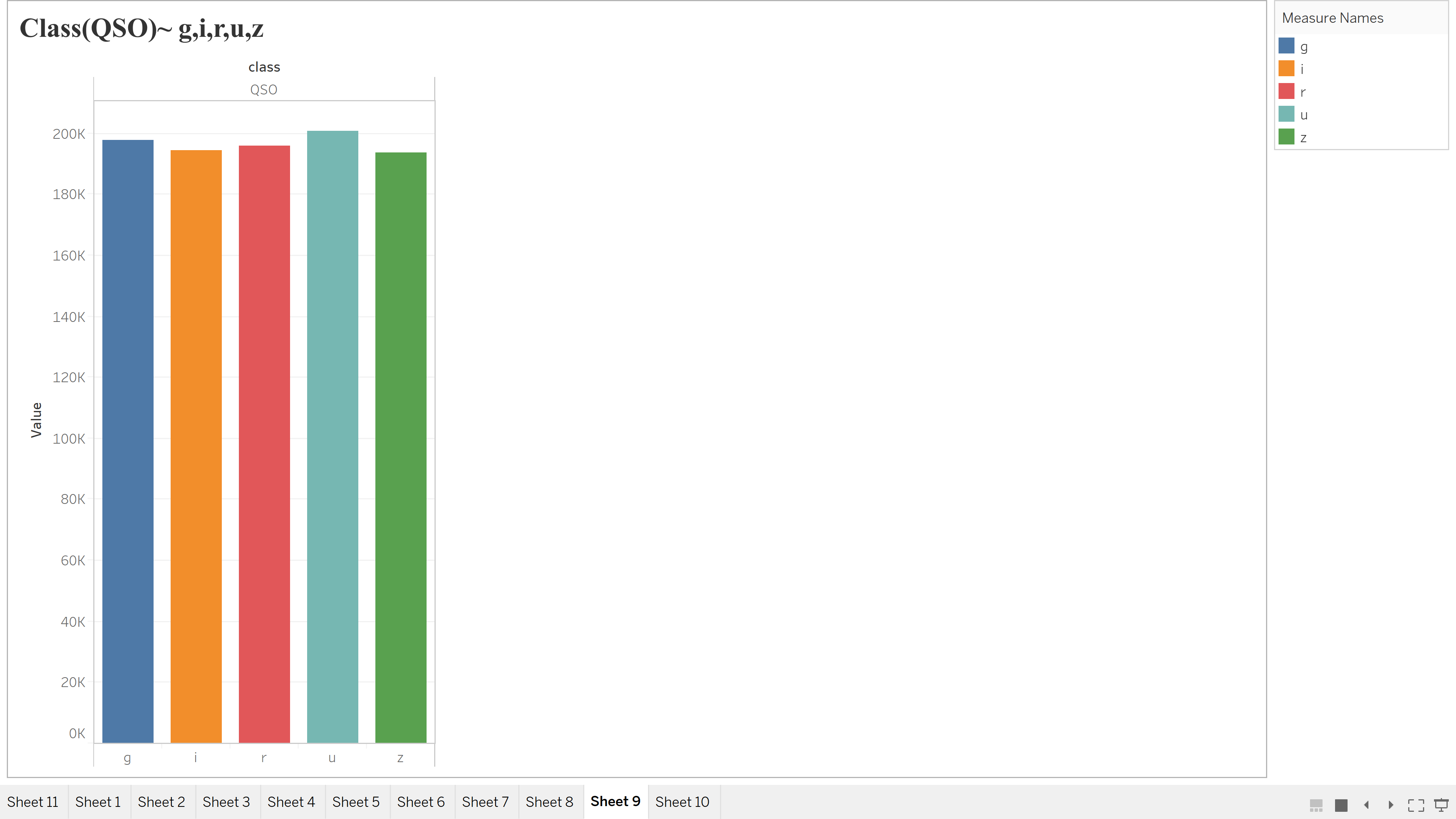


Fig 8: Bar chart explaining the range of predictor variable viz g,i,r,u,z for quasar class category.

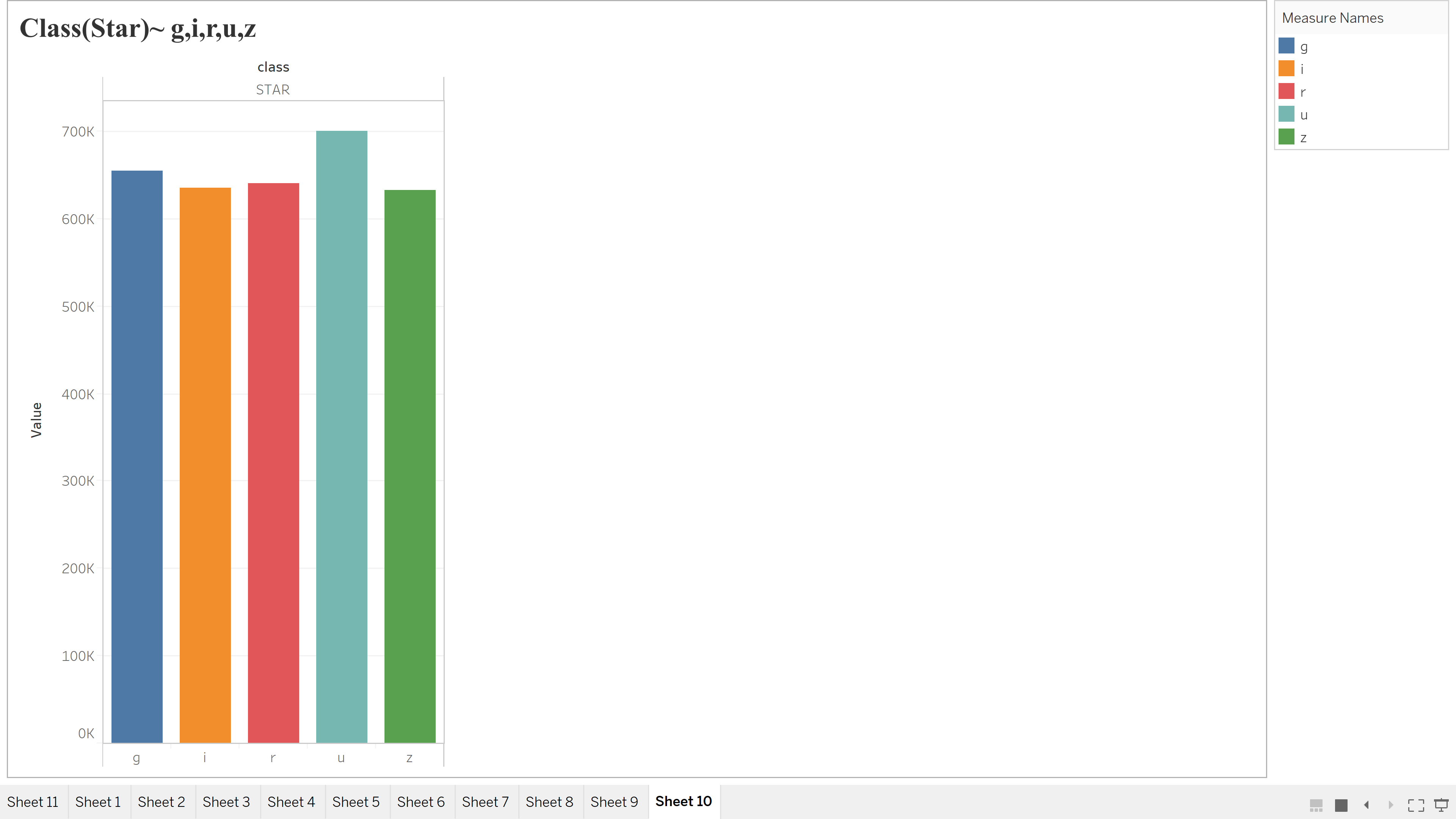


Fig 9: Bar chart explaining the range of predictor variable viz g,i,r,u,z for star class category.



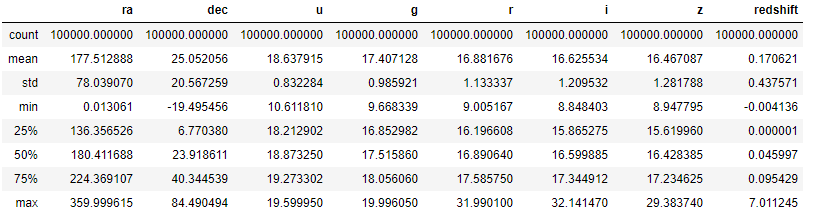


Fig 10: After data cleaning, predictor variables used.

* 1. **PRINCIPAL COMPONENT ANALYSIS (PCA)**

This method combines highly correlated variables together to form a smaller number of an artificial set of variables which is called “principal components” that account for most variance in the data. In order to handle “curse of dimensionality” and avoid issues like over-fitting in high dimensional space Principal Component analysis is used. (2) Variables ‘g’, ‘r’, ‘i’, ‘z’ is highly correlated thus PCA is used to reduce the four dimensions into one dimension.

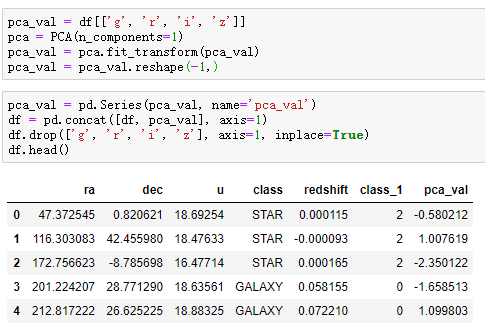


Fig 11: PCA analysis

1. **MODEL CREATION AND IMPLEMENTATION OF CLASSIFICATION MODELS**
2. **Logistic Regression (LogReg):**This model is used when predicting a multi-class target.

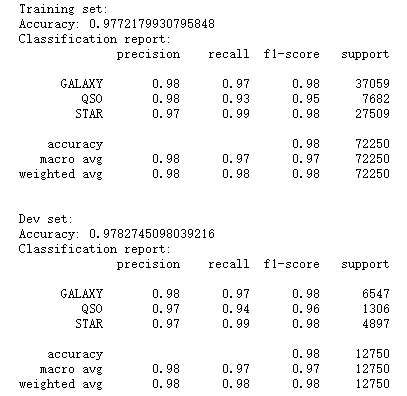
 



Fig 12: Logistic Regression for Classification with Accuracy.

1. **Decision Tree:** The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy.

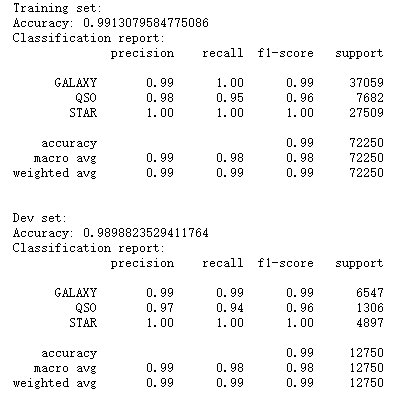
 

Fig 13: Decision Tree model for classication with accuracy.

1. **K-Nearest Neighbor:** This model classifies the targets based on the number of nearest neighbors to a specific class.

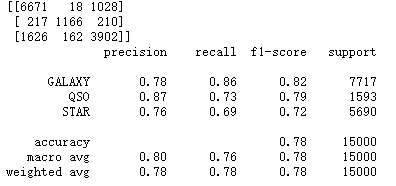
**** 

Fig 14: K- Nearest Neighbor Model for classification with accuracy.

1. **LDA:** Linear Discriminant Analysis (LDA) is a dimensionality reduction technique. As the name implies dimensionality reduction techniques reduce the number of dimensions (i.e. variables) in a dataset while retaining as much information as possible.



Fig 15: LDA Model with accuracy.

1. **Naïve Bayes:** It provides a way that we can calculate the probability of a piece of data belonging to a given class, given our prior knowledge.



Fig 16: Naïve Bayes Model for classification with accuracy.

1. **Support Vector Machine (SVM):** An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.



Fig 17: SVM model with accuracy.

1. **Random Forest:** The random forest is an ensemble learning method, composed of multiple decision trees. By averaging out the impact of several decision trees, random forests tend to improve prediction.

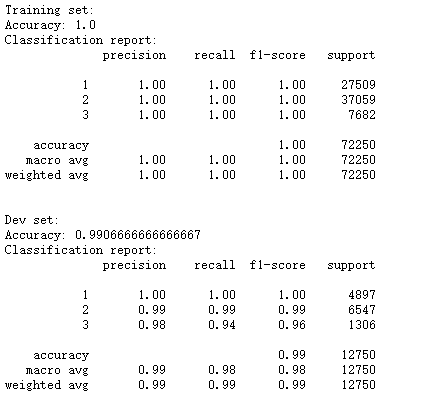
 

Fig 18: Random Forest Model with accuracy.

1. **FEATURE IMPORTANCE**

Feature importance gives you a score for each feature of your data, the higher the score more important or relevant is the feature towards your output variable. (1)

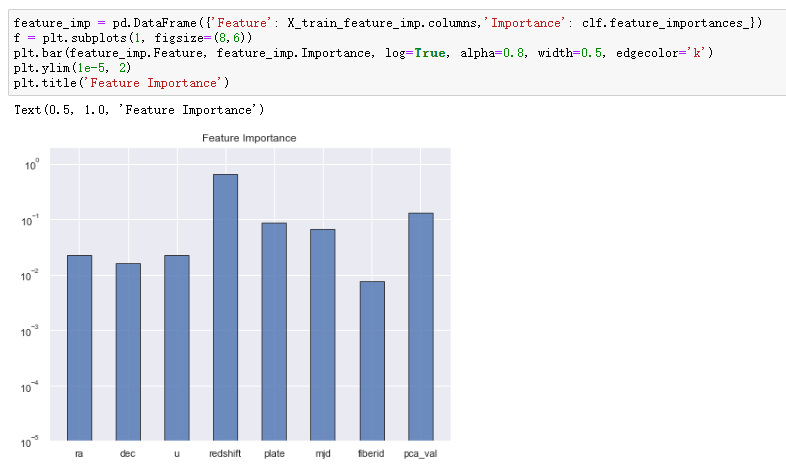


Fig 19: Feature Importance

So, we can see that the redshift is the most important feature.

1. **CONCLUSION**
2. Random forest and Decision Tree are two best supervised learning methods to predict the class of celestial object.
3. Galaxy is the most common celestial object in the universe.
4. The redshift is the most important feature to decide which class the object belongs to.
5. Z, g, r, i - four different types of magnitude fit have a strong relationship to each other.

**6. REFERENCES**

1. Shaikh, R. (2018, October 28). Feature Selection Techniques in Machine Learning with Python. Retrieved from <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>
2. Nagpal, A. (2017, November 22). Principal Component Analysis- Intro. Retrieved from <https://towardsdatascience.com/principal-component-analysis-intro-61f236064b38>
3. Minaie, N. (2019, August 2). A Beginner's Guide to Selecting Machine Learning Predictive Models in Python. Retrieved from <https://towardsdatascience.com/the-beginners-guide-to-selecting-machine-learning-predictive-models-in-python-f2eb594e4ddc>
4. Data Access for SDSS DR16 Overview. (n.d.). Retrieved from <https://www.sdss.org/dr16/data_access/>
5. Algorithms Overview. (n.d.). Retrieved from <https://www.sdss.org/dr16/algorithms/>