AstroChallenge 2021
Senior Data Analysis Round

Saturday 3rd April 2021

PLEASE READ THESE INSTRUCTIONS CAREFULLY.

In this part of **AstroChallenge 2021**, you will work with a large (approx. 30 000 points) data set. You will process this data set, analyse it, observe trends, and draw conclusions. **There are no right or wrong answers**; you will be marked solely on the quality of your analysis, even if your statistical methods are incorrect.

We **strongly** recommend you use industry-standard tools like `Microsoft Excel`™, `RStudio` or various `Python` libraries to process the data.

1 An Introduction to Data Analysis

Astronomy and Data

Astronomy has always been a data-driven discipline. From the earliest civilisations to the very bleeding edge of 21st-century humanity, data has been the core of the matter. Humans have **obtained** and **processed** data; we have **analysed** and **made conclusions** from said data to learn more about the world and the universe we live in.

In other words, astronomy is driven by new observations from data obtained by observing the universe. Scientists employ a very wide variety of techniques and tools to make sense of the data they have obtained. For instance, the CHANDRA X-Ray Observatory observes in X-ray wavelengths, but the data is clearly not recorded as X-ray photons! Instead, a computer on board the satellite encodes the data from the sensor as bits on some storage, which is then sent to ground-based stations.

This data can then be output into an image, or represented on a graph, or something else. Which representations of data are used, is determined by scientists with respect to the context: whether it be a paper in a scientific journal, or a news article in the morning newspaper, in a popular science magazine, or even as a video on some online platform.

At each step of data analysis, however, lies the possibility for bias to creep in: from the very beginning of data acquisition, to the presentation of data. This section of **AstroChallenge 2021** will lead you through the four steps mentioned above, using a real data source that can be freely accessed. There are four sections in the paper below. However, this does not mean each section corresponds to each step above.

About the Data Analysis Question

A question completely dedicated to data analysis in this form was introduced to **AstroChallenge 2020**. In the continued spirit of learning, you might find some parts to be rather guided—this is intentional. However, others will require you to think out of the box, and be resourceful with the data source.

This question is meant to replace one of the five questions in the Team Round, so the maximum number of marks in this entire round will be **20**. Exact mark allocations will be given within the right margin next to each part orsub-part, in boldface and in square brackets, like such: **[2]**.

Deliverables

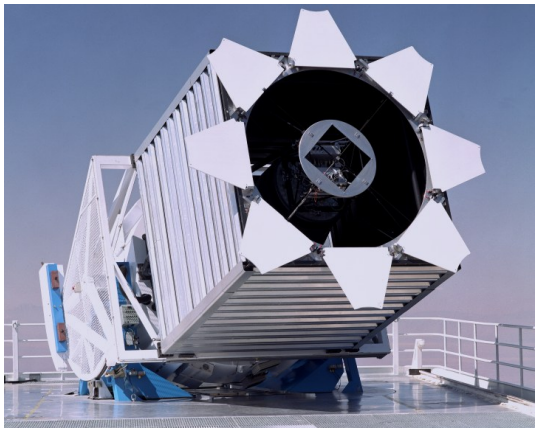
This section outlines the deliverables to be submitted for the data analysis question. **Everything** in the following list is to be submitted—any omissions will lead to your team losing marks.

- **One** `.xlsx` file containing sheets of *all* your responses which require data processing named ⟨`Your school abbreviation`⟩`_T`⟨`your team number`⟩`_data.xlsx` **OR two** `.csv`[1] files for all the parts you have attempted each labelled ⟨`Your school abbreviation`⟩`_T`⟨`your team number`⟩`_`⟨`question and part number`⟩`.csv` (In other words, if you are in Team 1 from Astro Challenge Secondary School, please label your response to Question 1 Part a as `ACSS_T1_1a.csv`, the same goes for the `Excel` document, omitting the question parts, i.e. `ACSS_T1_data.xlsx`);
- A *type-written* report in `.pdf` format named ⟨`Your school abbreviation`⟩`_T`⟨`your team number`⟩`_Report.pdf`, which should include:
  - Your team's attempts for all questions;
  - A bibliography, if any external sources were used and cited;
  - An appendix, with all your graphs and images

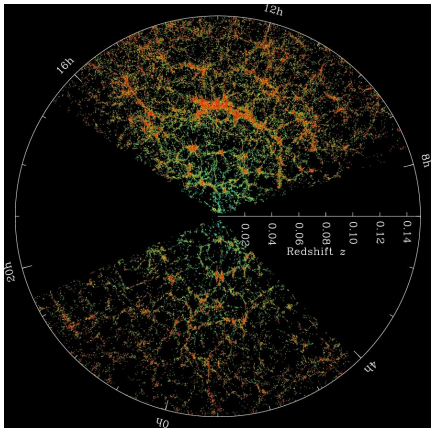
[1] `.csv` stands for Comma-Separated Values. Such files can actually be opened and manipulated using ordinary text editors like Notepad, TextEdit, Visual Studio Code, Sublime Text, etc. The data are arranged in columns, and separated by commas, which gives rise to the file name.

2 Colours of Galaxies

The Sloan Digital Sky Survey is one of the most successful and productive deep sky surveys, generating discoveries like Sloan's Great Wall and 3 dimensional maps of local surrounding galaxies. With its 2.5m Sloan Foundation Telescope and a 3-degree true field of view, it has captured the spectra and imaged close to one third of the night sky, allowing astronomers to create an in-depth view of the universe. Let us explore just a little of SDSS's plethora of data.



(a) The 2.5m Sloan Foundation Telescope[2]



(b) A 3D map of the neighbouring galaxies[3]

Before we jump into analysing SDSS's data, we need a little knowledge about spectroscopy and colours of celestial objects. Stars do not emit light in only one wavelength, they emit light in all wavelengths, following a model known as Blackbody Radiation. The reason why stars appear different colour to us is because they emit different proportions of the wavelengths of light, peaking at some wavelength and tapering towards lower and higher wavelengths. For example, red stars emit most of their light in red wavelengths and blue stars emit the most light in blue wavelengths. Figure 2 shows a plot of the wavelengths of light emitted by our Sun and their respective irradiance.

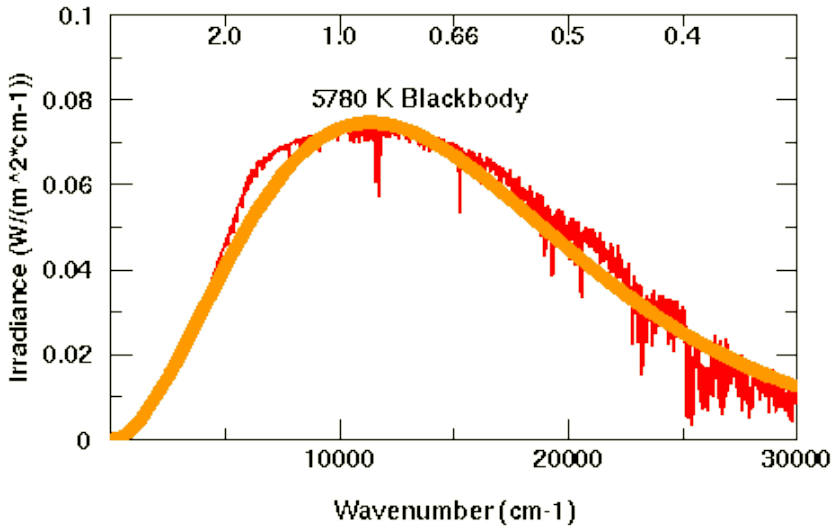


Figure 2: A comparison of the radiation of the Sun with a "black-body" radiator of the same temperature[4]

The collective visible wavelengths emitted by our Sun make it white (ignoring any atmospheric effects), but the Sun also emits light in the ultraviolet, infrared and radio wavelengths that are invisible to us.

Electrons in atoms in stars and gas clouds can transition between energy levels and when they do, they emit or absorb energy, in terms of photons, from higher to lower and lower to higher energy levels respectively. Due to the fixed and

[1]The Sloan Digital Sky Survey. The sloan foundation 2.5m telescope at apache point observatory, 2014b. URL https://www.sdss.org/wp-content/uploads/2014/11/SDSS_telescope_new.jpg. [Online; accessed 20 March, 2021]

[3]The Sloan Digital Sky Survey. Sdss map of the universe, 2014a. URL https://dev.sdss.org/wp-content/uploads/2014/06/orangepie.jpg. [Online; accessed 20 March, 2021]

[4]Courtney Seligman. A comparison of the radiation of the sun with a "black-body" radiator of the same temperature, 2014. URL https://cseligman.com/text/sun/sunbb.gif. [Online; accessed 20 March, 2021]

discrete energy levels in atoms, the energy (photons) absorbed are the same for the same type of transition in the same type of atoms. Hence, different atoms in stars emit or absorb characteristic wavelengths (energies) of light that reveal what type of atoms are present in the star. As shown in the spectra of the Sun (**??**)s, you can see various dips from the theoretical smooth curve that indicate specific wavelengths being absorbed by atoms in the Sun. Figure 3 below shows the photons emitted by electron transitions in a Hydrogen atom. This leads to spectrums of celestial objects that look more like those shown in Figure 4 rather than smooth Blackbody curves.

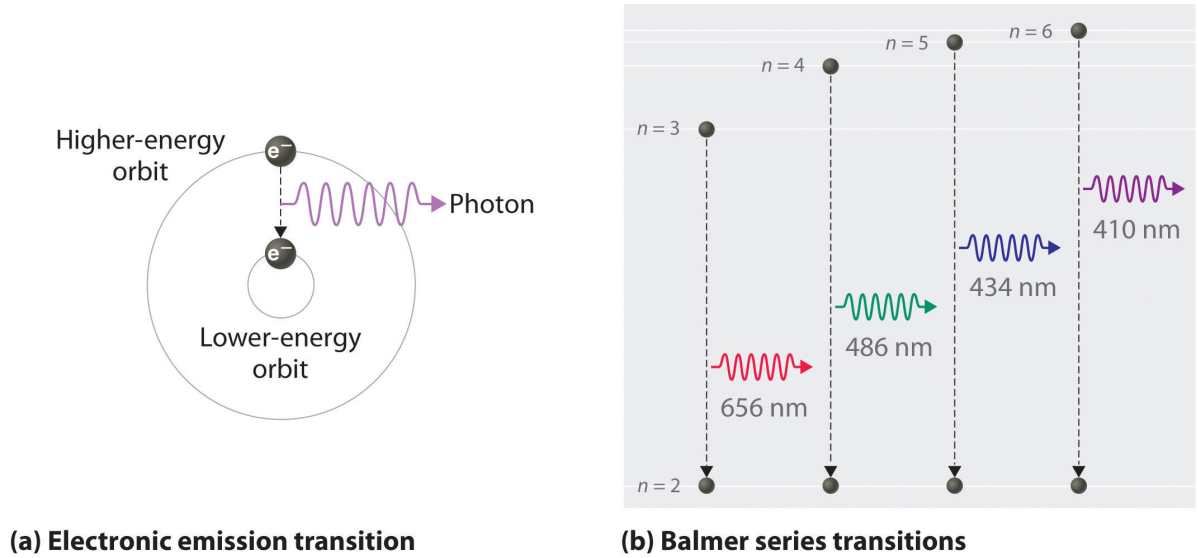


Figure 3: a) Transition of electron from higher to lower energy orbit releases energy in terms of a photon, b) A group of transitions in a hydrogen atom that produce photons of visible wavelengths[5]

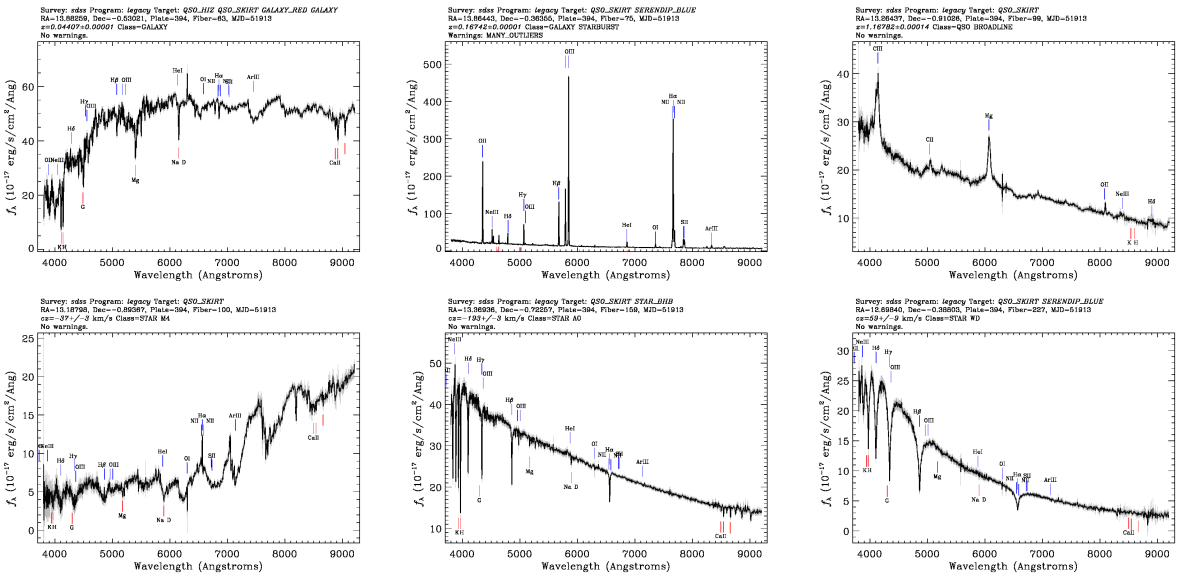


Figure 4: Examples of spectra obtained by SDSS[6]

The SDSS telescope can collect the spectrum of up to 1000 objects in the field of view simultaneously with various aluminium plug plates on the focal plane holding optical fibres which carries the light from each source to the spectrograph[7]. The spectrograph contains devices, such as a diffraction grating, that then split light into its constituent wavelengths (Figure 5).

[5]Saylor Academy. The emission of light by a hydrogen atom in an excited state, 2021. URL https://saylordotorg.github.io/text_general-chemistry-principles-patterns-and-applications-v1.0/section_10/191f6551113db14ecde90a88d9f13d9a.jpg. [Online; accessed 21 March, 2021]

[6]The Sloan Digital Sky Survey. Six example sdss spectra of stars, galaxies, and quasars., 2019. URL http://voyages.sdss.org/wp-content/uploads/2019/05/spectra_examples.png. [Online; accessed 21 March, 2021]

[7]http://voyages.sdss.org/preflight/sdss-plates/

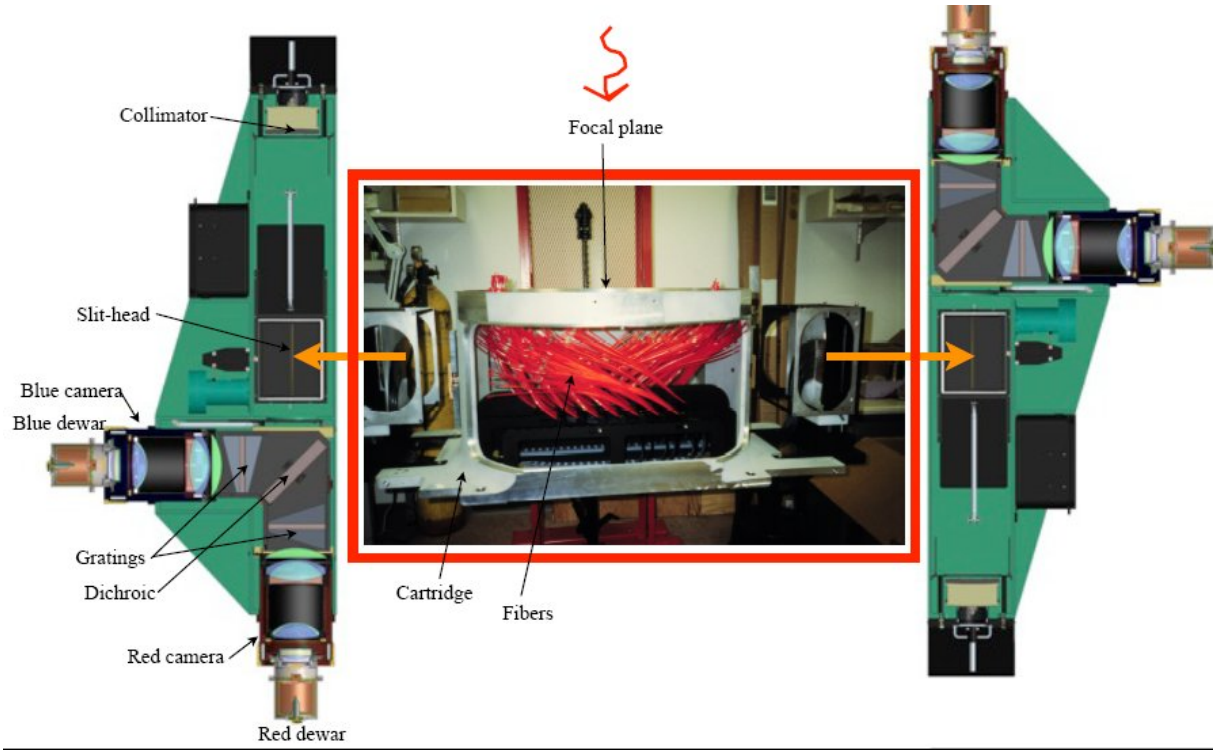


Figure 5: Spectrography setup of SDSS BOSS survey

In the first question, we will be dealing with a much more simplified representation of the spectrum of objects: only the brightness/intensity at specific wavelengths out of the entire spectrum collected by the SDSS telescopes through colour filters that only allow wavelengths in a narrow range about a specified wavelength (e.g. red) through to the sensors. Read more about SDSS's colour filtering system[8].

[8] https://skyserver.sdss.org/dr1/en/proj/advanced/color/sdssfilters.asp

2.1 Kinematics

1. Velocity dispersion measures the dispersion of velocities about the mean velocity of a cluster of astronomical objects. The parameter is often measured from the broadening of spectral lines coming from the distant cluster, arising from the superposition of the doppler shifts of faster and slower moving members of the cluster. In this case, we will be looking at the velocity dispersion of stars in a galaxy.

   Due to the complexity of the kinematics of galaxies with different components (e.g. the halo, bulge, and disk of spiral galaxies), the velocity dispersion is only determined for spheroidal systems whose spectra are dominated by the light of red giant stars.

   (a) Can you think of other reasons that the velocity dispersion of disk-shaped galaxies would be hard to determine? Would the measured velocity dispersion be an accurate indication of the rotation of the galaxy's stars? [2]

   (b) Why would the luminosity of these spheroidal galaxy systems that SDSS imaged be dominated by red giant stars? [2]

   (c) Use the following queries to retrieve 100 objects from the ***specObj*** table with and without ***velDisp*** calculations respectively as `.csv` files and find the mean $\boldsymbol{spectroFlux_r}$ / $\boldsymbol{spectroFlux_u}$, which group of galaxies are redder? Does this agree with the context? Combine the two tables into `⟨Your school abbreviation⟩_T⟨your team number⟩_1c.csv` or as a new sheet `1c` in your `Excel` document.

   Note: It is a good time to be familiar with SDSS's SQL tool in the Appendix, SDSS's colour classification scheme as detailed before[9], and the meaning of ***cModelMag*** under galaxy's schema browser[10]. [2]

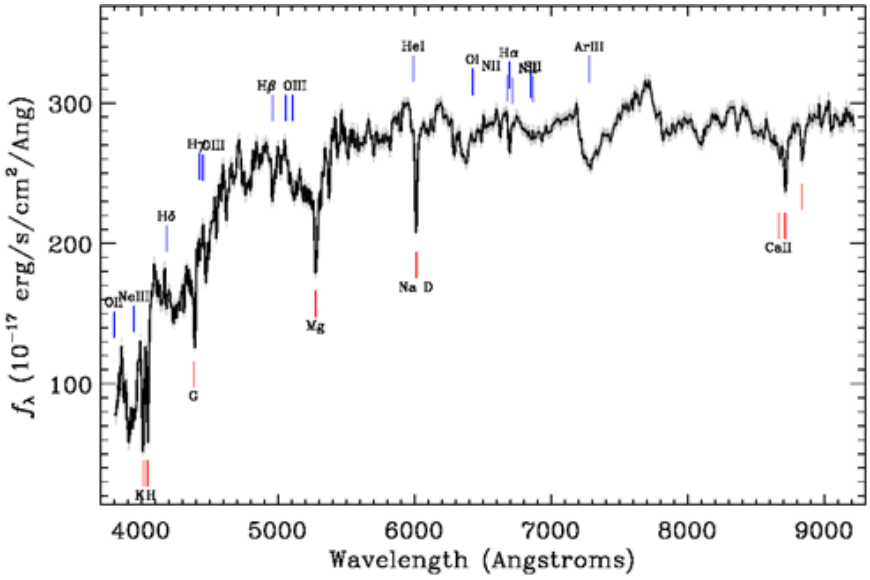
```
SELECT top 100
    specObjID, dec, ra, velDisp, spectroFlux_u, spectroFlux_r, spectroFlux_g
FROM
    specObj
WHERE
    class="GALAXY"
    AND velDisp>0
    AND velDispErr BETWEEN 0 and 5
```

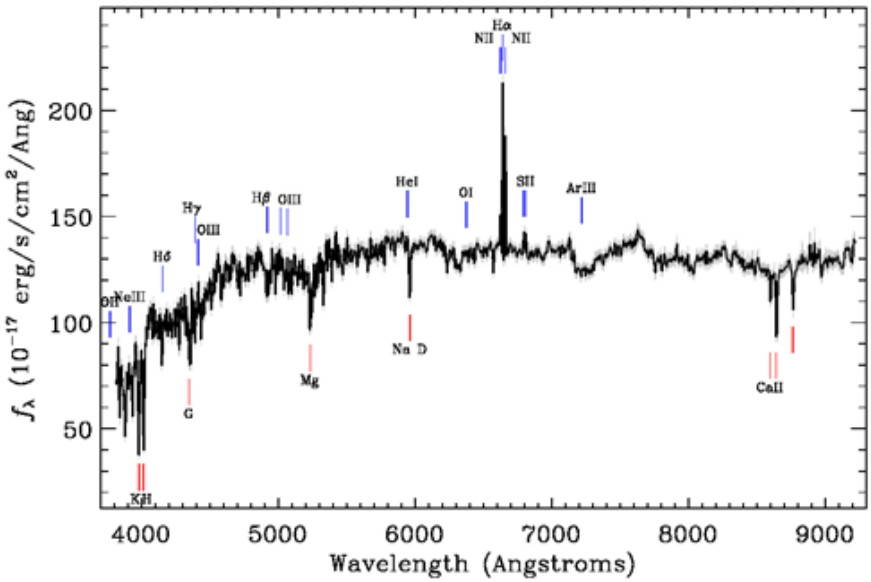
```
SELECT top 100
    specObjID, dec, ra, velDisp, spectroFlux_u, spectroFlux_r, spectroFlux_g
FROM
    specObj
WHERE
    class="GALAXY"
    AND velDisp=0
    AND spectroFlux_u>0
```

[9] https://skyserver.sdss.org/dr1/en/proj/advanced/color/sdssfilters.asp
[10] http://skyserver.sdss.org/dr16/en/help/browser/browser.aspx#&&history=description+Galaxy+V

2.2 Colours

2. Below shows the spectra of two galaxies. Red lines indicate absorption lines and blue lines indicate emission lines.





(a) Which galaxy has a higher luminosity? Provide sufficient evidence from the spectra shown.
HINT: the formula for redshift may be useful here:

$$z = \frac{\lambda_{\text{observed}} - \lambda_{\text{emitted}}}{\lambda_{\text{emitted}}}$$

[2]

(b) Why do the two galaxies seem to share the same strong absorption lines despite being so different in their spectrum? [1]

(c) Which galaxy has more active star formation? Provide two evidence from the spectra.

HINT: remember SDSS's colour classification scheme and the colour of galaxies (since the spectrum is limited, you may use the shortest wavelength measured in the spectrum as a proxy for ultraviolet light). [2]

2.3 Map

3. There are a lot more we can learn about galaxies from their spectra and colours. Like a colour-magnitude diagram for stars, a colour-magnitude diagram can be drawn for galaxies too. Run the following SQL query command and download the data as a `.csv` file, name this `⟨Your school abbreviation⟩_T⟨your team number⟩_3a.csv` or add a new sheet `3a` to your `Excel` document.

   (a) Using Hubble's law (see Appendix) and the Hubble's constant of 69.3 km s^{-1} Mpc^{-1}, calculate the absolute magnitudes under the various filters, create three new columns named M_u, M_r, and M_g beside your data for the absolute magnitudes.
   Note: Absolute and apparent magnitudes are related by the following formula:

$$m - M = 5 \lg \left(\frac{d}{10\text{pc}} \right)$$

   [2]

SELECT g.objID, g.cModelMag_u, g.cModelMag_r, g.cModelMag_g, s.z FROM galaxy AS g JOIN specObj AS s ON s.bestObjID = g.objID WHERE s.class="galaxy" AND s.zWarning = 0 AND s.z BETWEEN 0.00 AND 0.1 AND g.clean = 1 AND g.cModelMag_u>0 AND g.htmid*37 & 0x000000000000FFFF < (650 * 10)	This query randomly galaxies which are closer to us and which have spectroscopic data available.

   (b) Look at the above SQL query, why must the redshift values of galaxies queried be limited in such a way? [1]

   (c) Plot a graph M_u - M_r against M_g using a scatter plot. Adjust the horizontal scale to be between -25.0 and -15.0 and the vertical scale to be between 0.0 and 4.0 to capture the main plot of your data points. Next, we need to visualise our data better. If you are using `Excel`, right-click on the data points on your plot and select "Format Data Series..." in the drop-down menu. Choose the paint bucket icon in the Format Data Series Menu as shown in Figure 6 and adjust the data point markers to the following settings. You can also scale up your graph to see the data points better. [2]

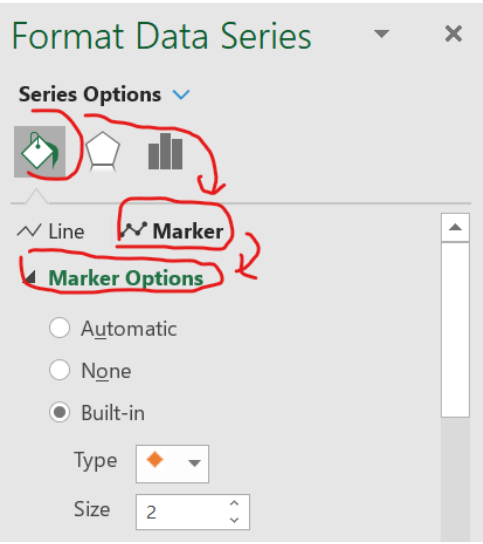


Figure 6: Settings for Markers in Excel Graph Plot

(d) From the graph you have plotted, do galaxies with higher visual luminosities seem to be bluer or redder? Suggest two reasons why this might be the case. [2]

(e) Can you identify different groups of galaxies within the cluster of data points, what are they? Draw red circles on your plot to identify the groups. [2]

A SDSS SQL

You will not be required to generate your own SQL Query command in this paper. However, to make the best use of SDSS's data, let us be familiar with SDSS's SQL query tool:
http://skyserver.sdss.org/dr16/en/tools/search/sql.aspx
An SQL query is a segment of instructions sent to the SDSS database to retrieve information from its tables. Each SQL query consists of the following segments:

SELECT specObjID, ra, dec FROM specObj WHERE ra BETWEEN 257.14 AND 259.14 AND dec BETWEEN 63.07 AND 65.07 AND class = "galaxy"	**SELECT:** the names of data columns that you wish to retrieve from the table in the database (a list of column names of the table *specObj* referenced in this query are found in SDSS SpecObj Schema Browser). Here, the columns *specObjID*, *ra*, and *dec* are selected. **FROM:** the name of the table that you wish to retrieve data from. The table *specObj* contains a list of all the spectrographs that SDSS has taken of objects to date. **WHERE:** conditions that the data must meet. Here, we only want to search for objects in the region where Right Ascension is between 257.14 and 259.14 and Declination between -63.07 and 65.07, and we only want galaxies
*Note you don't have to write keywords like SELECT, FROM, WHERE, BETWEEN, and AND in capital letters, but it is just good practice	

Now key in the query into the white box, select "CSV" as Output Format, and click "submit query". Once you have the .csv file, you can open it with spreadsheet software[11]. Use the SDSS Navigation tool to get a glimpse of the image of the area, are the data you retrieved expected?

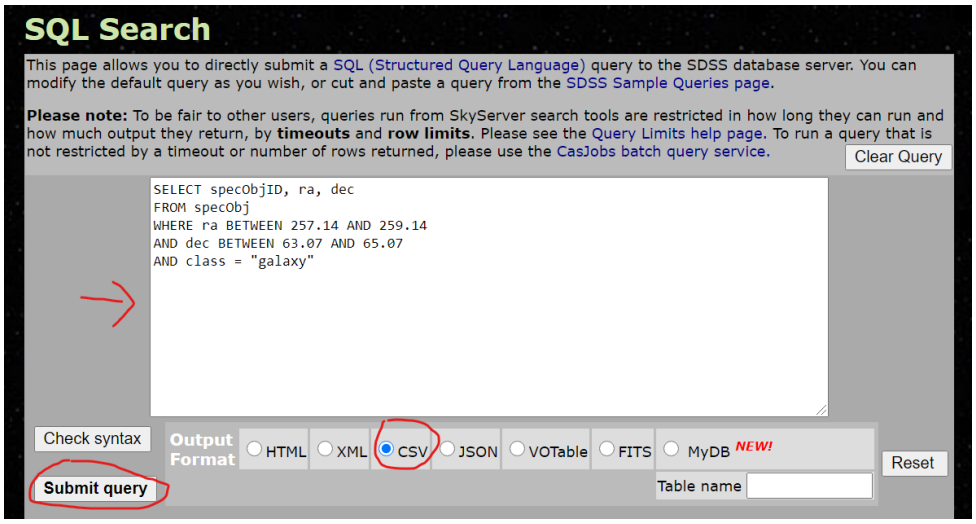


Figure 7: SDSS SQL Tool interface

To learn more about SDSS's SQL Query, go through the SQL exercises and the guide[12] and refer to some sample commands[13].

[11] such as Microsoft Excel, LibreOffice Calc, Apple Pages, Google Sheets
[12] http://skyserver.sdss.org/dr16/en/help/howto/search/searchhowtohome.aspx
[13] http://skyserver.sdss.org/dr16/en/help/docs/realquery.aspx

B Hubble's Law

In 1929, Edwin Hubble published his now famous work on the relationship between the distance and recession velocities of distant galaxies[14]. The relationship between the distance and recession velocities of galaxies are best summarised by the following equation, also known as Hubble's Law:

$$v = H_0 d,$$

where H_0 is known as the Hubble Constant.

[14]Edwin Hubble. A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15(3):168–173, 1929. ISSN 0027-8424. doi: 10.1073/pnas.15.3.168. URL https://www.pnas.org/content/15/3/168