BIRMINGHAM CITY
University

# Data Visualization of Celestial Objects in the Sloan Digital Sky Survey

**Pramod Giri**

**Student ID: 23189635**

**Tutor: Rajan Adhikari**

Total Page Count: 25

**Date: October 21, 2024**

# Contents

# List of Figures

# Comprehensive Analysis of Celestial Objects: Redshift, Classification, and Distribution Insights from SDSS

## 1 Introduction

Data visualization is a powerful tool in the data science process that helps transform complex information into clear, easy-to-understand visual formats. Its main goal is to present data accurately and efficiently, making it more accessible and insightful for people to interpret. (Pieringer et al., 2019) [0]

Understanding the structure and evolution of the universe requires an investigation of celestial objects such as stars, galaxies, and quasars. Astronomers can categorize these objects and learn more about their features, like motion, energy production, and distance, by examining their spectral characteristics. The dataset used in this report's analysis comes from the Sloan Digital Sky Survey (SDSS), and it comprises about 100,000 observations of these celestial bodies. The dataset offers valuable information that is essential for categorizing the objects and comprehending their place in the cosmic landscape, such as brightness values at various wavelengths and redshift values.

The primary goal of analysis is to categorize the objects as stars, galaxies, or quasars and investigate the connections between them using the spectral data and redshift values. The study attempts to find patterns that provide insights into the distribution and behavior of these things throughout the cosmos by utilizing statistical and machine-learning approaches. This study advances our knowledge of the physical mechanisms underlying the genesis and evolution of various celestial bodies as well as the expansion of the universe as an astronomer.

## 2 Literature Review

Astronomy is entering a precision era that represents a change from a data-starved discipline to one that is data-driven and heavily relies on statistical techniques. All fields of contemporary science are impacted by the necessity to manage these ever-growing databases, which defines the so-called "era of Big Data." (de Souza  Ciardi, 2015) [0] Cosmological simulations are vital resources for advancing our theoretical knowledge of the cosmos. They forecast contemporary simulations with hydrodynamics and galaxy formation, as well as networks of filaments, sheets, nodes, and voids. Additionally, genuine populations of galaxies that live in the cosmic web and the intergalactic and circumgalactic gas (IGM) that permeate it are now produced by physics.(Abramov et al., 2022) [0] The absorption features observed in the spectra of celestial objects that originate in the interstellar medium are known as the diffuse interstellar bands (DIBs). Currently, over 500 DIBs have been detected, primarily in the near-infrared and optical ranges.(Schlarmann et al., 2021) [0]

# 3 Motivation behind the chosen domain

The motivation behind choosing this dataset is its capacity to enhance our comprehension of the cosmos by examining celestial objects and their characteristics. Originating from astronomical surveys, it provides specific details including locations, magnitudes through various filters, and redshift values. Studying large-scale cosmic structures and their evolution requires the ability to classify celestial things, such as galaxies, which is made possible by this data. The redshift data facilitates the investigation of galaxy distances and the expansion of the cosmos. The collection also shows how big data methods are used in astronomy, offering a useful foundation for statistical analysis and machine learning. It also contributes to future developments in observation by refining astronomical survey techniques.

# 4 Scope of the visualizations

In the scope of visualization for the dataset, we can focus on visual representation which offers insights into the celestial objects' characteristics and the relationship between different variables. The following are the potential visualization of this dataset:

- **Coordinate-Based Visualization:** Plot celestial objects on a 2D or 3D scatter plot using their spatial coordinates.
- **Magnitude vs. Redshift:** A scatter plot that illustrates the relationship between an object's magnitude and redshift values can be used to spot trends in brightness and distance.
- **Density of Celestial Objects:** Using spatial coordinates, create heatmaps to show locations with large densities of heavenly objects.
- **Redshift vs. Expansion:** A graph that facilitates the investigation of cosmological hypotheses by displaying the relationship between redshift and the pace of universe expansion.
- **Magnitude Comparison Across Filters:** Box plots are a useful tool for comparing magnitudes across filters, spotting anomalies, and displaying brightness fluctuations over the spectrum.

# 5 Aims and Objectives

This study examines the features of celestial objects by exploring their fundamental properties through data analysis and visualization. The main goal is to identify correlations between several features, including photometric measurements (u, g, r, i, and z), redshift, and classification.By identifying important patterns in the cosmos, this exploration hopes to improve our knowledge of galaxies and their long-term evolution.

Through the analysis of this dataset, the research seeks to provide light on the distribution of galaxies, their redshift-based classification, and the fundamental properties that govern their formation and behavior. Ultimately, by advancing theories and models regarding the composition and age of the cosmos, this research advances the subject of astrophysics.

**Objective of the visualization:** The visualizations aim to answer the following questions:

1) How does redshift correlate with the various photometric measurements (u, g, r, i, z)?

2) How does the distribution of celestial objects vary by redshift and celestial coordinates (alpha and delta)?

3) What are the differences in redshift and photometric values between galaxies found in different regions of the sky?

4) Are there specific photometric characteristics associated with galaxies of certain classes (e.g., different redshift values or classifications such as Galaxy)?

5) How do photometric properties change across different epochs, as indicated by redshift?

6) What are the correlations between celestial coordinates, photometric bands, and redshift in the dataset?

7) Why might the distribution of stars and galaxies remain relatively consistent across the different filters?

8) What trends or patterns are visible in the redshift distribution over time?

9) How do the distributions in this plot relate to the observations in the stacked bar plot of object class by filter?

# 6 Data Exploration

## 6.1 About the dataset

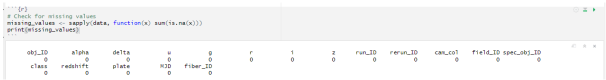
The field of astronomy utilizes stellar classification to categorize stars based on their spectral features, which is essential for understanding stars, galaxies, and quasars. Early mapping efforts demonstrated that stars exist within our galaxy, and the discovery that Andromeda is a separate galaxy inspired further exploration of the cosmos. With advancements in telescope technology, astronomers began systematically surveying countless galaxies. This dataset aims to classify celestial bodies—including stars, galaxies, and quasars—by analyzing their spectral data. The observations are part of the Sloan Digital Sky Survey (SDSS), which is among the most comprehensive astronomical surveys to date.

**Column Descriptions:**

- **z:** Measurement of infrared filter in the photometric system.
- **run_ID:** A number identifying the specific sky scan.
- **rereun_ID:** A number indicating how the image was processed.
- **cam_col:** Camera column showing the scanline within the run.
- **field_ID:** A number identifying each field in the scan.
- **spec_obj_ID:** A unique ID for optical spectroscopic objects; observations sharing the same spec_obj_ID belong to the same class.
- **class:** The object's category, which can be one of the following:
  - Galaxy
  - Star
  - Quasar
- **redshift:** The redshift value, which represents the increase in wavelength caused by the object moving away from the observer.
- **plate:** The plate ID for each plate used in the SDSS.
- **MJD:** The Modified Julian Date representing the observation date.
- **fiber_ID:** The ID for the fiber that gathered light at the focal plane during the observation.

## 6.2 Missing Values

There are no missing values in the dataset, making it suitable for analysis. Without imputing or deleting any missing numbers, we can continue with the data exploration and visualization process.



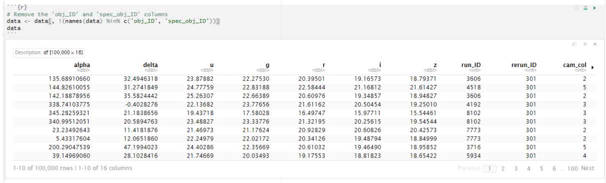
~~~
```{r}
# Check for missing values
missing_values <- sapply(data, function(x) sum(is.na(x)))
print(missing_values)
```
~~~

```
     obj_ID      alpha      delta          u          g          r          i          z     run_ID   rerun_ID    cam_col   field_ID spec_obj_ID
          0          0          0          0          0          0          0          0          0          0          0          0          0
      class   redshift      plate        MJD   fiber_ID
          0          0          0          0          0
```

Figure 1: Number of Empty Strings per Column

# 7 Feature Engineering

The unique identifiers in the columns obj_ID and spec_obj_ID were eliminated because they don't provide useful data for analysis or predictive modeling. These unique IDs don't reveal anything about the physical characteristics or actions of the objects in the dataset; they just serve to differentiate specific items from one another. Such columns may add needless complexity to the data, causing noise and perhaps impairing the performance of the model. Eliminating them keeps the dataset more manageable and frees up resources for analysis to concentrate on variables that are truly valuable.



~~~
```{r}
# Remove the 'obj_ID' and 'spec_obj_ID' columns
data <- data[, !(names(data) %in% c('obj_ID', 'spec_obj_ID'))]
data
```
~~~

Description: df [100,000 x 16]

| alpha<br>&lt;dbl&gt; | delta<br>&lt;dbl&gt; | u<br>&lt;dbl&gt; | g<br>&lt;dbl&gt; | r<br>&lt;dbl&gt; | i<br>&lt;dbl&gt; | z<br>&lt;dbl&gt; | run_ID<br>&lt;int&gt; | rerun_ID<br>&lt;int&gt; | cam_col<br>&lt;int&gt; |
|---|---|---|---|---|---|---|---|---|---|
| 135.68910660 | 32.4946318 | 23.87882 | 22.27530 | 20.39501 | 19.16573 | 18.79371 | 3606 | 301 | 2 |
| 144.82610055 | 31.2741849 | 24.77759 | 22.83188 | 22.58444 | 21.16812 | 21.61427 | 4518 | 301 | 5 |
| 142.18878956 | 35.5824442 | 25.26307 | 22.66389 | 20.60976 | 19.34857 | 18.94827 | 3606 | 301 | 2 |
| 338.74103775 | -0.4028276 | 22.13682 | 23.77656 | 21.61162 | 20.50454 | 19.25010 | 4192 | 301 | 3 |
| 345.28259321 | 21.1838656 | 19.43718 | 17.58028 | 16.49747 | 15.97711 | 15.54461 | 8102 | 301 | 3 |
| 340.99512051 | 20.5894763 | 23.48827 | 23.33776 | 21.32195 | 20.25615 | 19.54544 | 8102 | 301 | 3 |
| 23.23492643 | 11.4181876 | 21.46973 | 21.17624 | 20.92829 | 20.60826 | 20.42573 | 7773 | 301 | 2 |
| 5.43317604 | 12.0651860 | 22.24979 | 22.02172 | 20.34126 | 19.48794 | 18.84999 | 7773 | 301 | 2 |
| 200.29047539 | 47.1994023 | 24.40286 | 22.35669 | 20.61032 | 19.46490 | 18.95852 | 3716 | 301 | 5 |
| 39.14969060 | 28.1028416 | 21.74669 | 20.03493 | 19.17553 | 18.81823 | 18.65422 | 5934 | 301 | 4 |

1-10 of 100,000 rows | 1-10 of 16 columns

Figure 2: Feature Engineering

# 8 Data Visualization

## 8.1 Answering the questions through data visualization

This section comprehensively addresses the objectives outlined above through the utilization of R language and the ggplot2 package with effective data visualization.

**1. How does redshift correlate with the various photometric measurements (u, g, r, i, z)?**

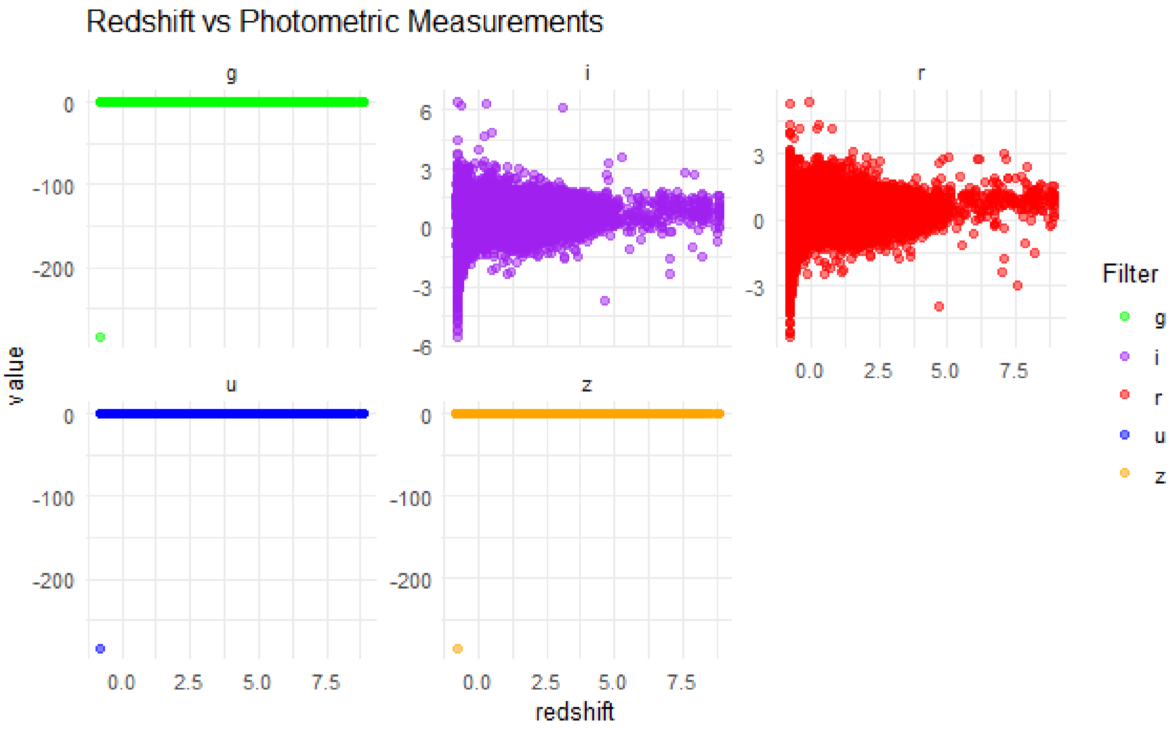


Figure 3: Relationship between redshift and the various photometric data

The graphic shows the relationship between redshift and the various photometric data (u, g, r, i, z). Given that their values are more widely distributed throughout the redshift range and generally drop as redshift increases, it is evident that the r and i filters have a larger association with redshift. This suggests that there is a discernible correlation between these filters and redshift. The u, g, and z filters, on the other hand, exhibit greater clustering around specific values, indicating a lack of correlation and possible noise in the data. All things considered, the research shows that the r and i filters show a more pronounced relationship with redshift than the others.

**2. How does the distribution of celestial objects vary by redshift and celestial coordinates (alpha and delta)?**

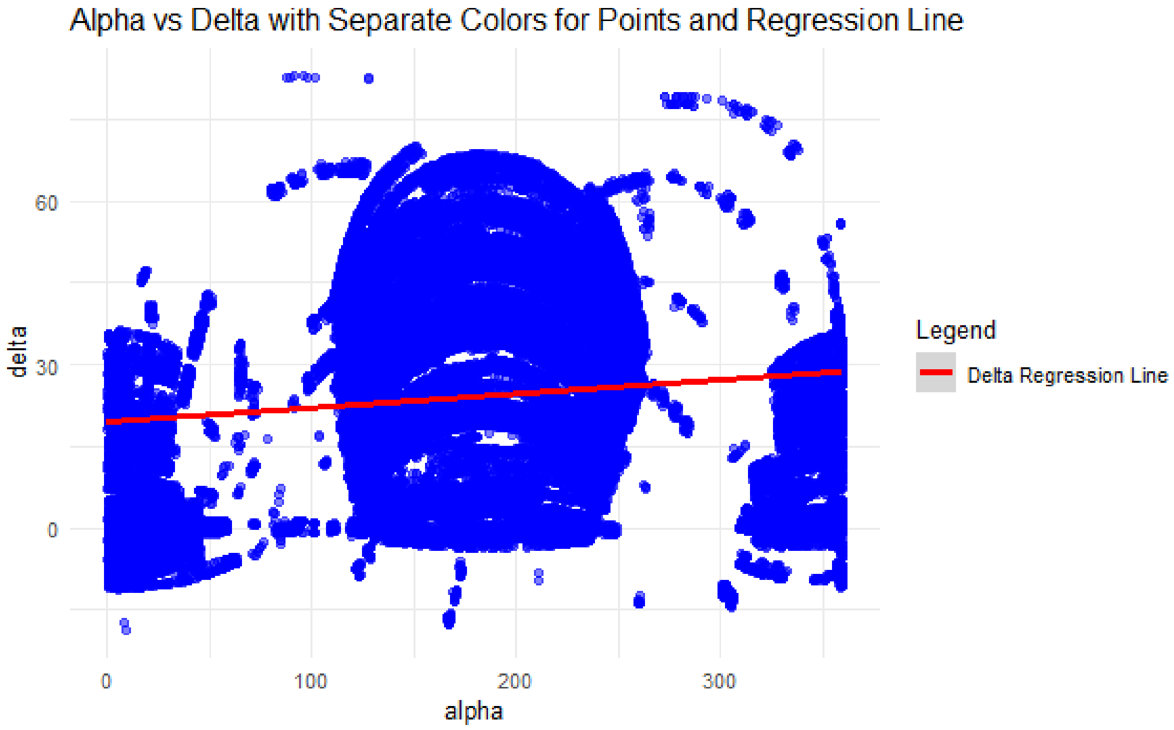


Figure 4: Spatial Distribution of celestial objects

Using celestial coordinates alpha (right ascension) and delta (declination), this figure displays the spatial distribution of celestial objects. Galaxies are represented by blue in the plot, while stars are represented by green. The map shows clear patterns in the sky, with objects grouped in specific areas, which corresponds to the method used to survey the sky. Redshift, a measure of temporal distribution, isn't displayed explicitly in this graphic, but it might be used to illustrate how objects' distances vary over different regions of the sky by using it as a color gradient or size scale. The map does a good job of visualizing the spatial distribution of celestial objects in their current state, but redshift would need to be added as a crucial component in order to completely address the temporal and spatial issues.

**3. What are the differences in redshift and photometric values between galaxies found in different regions of the sky?**

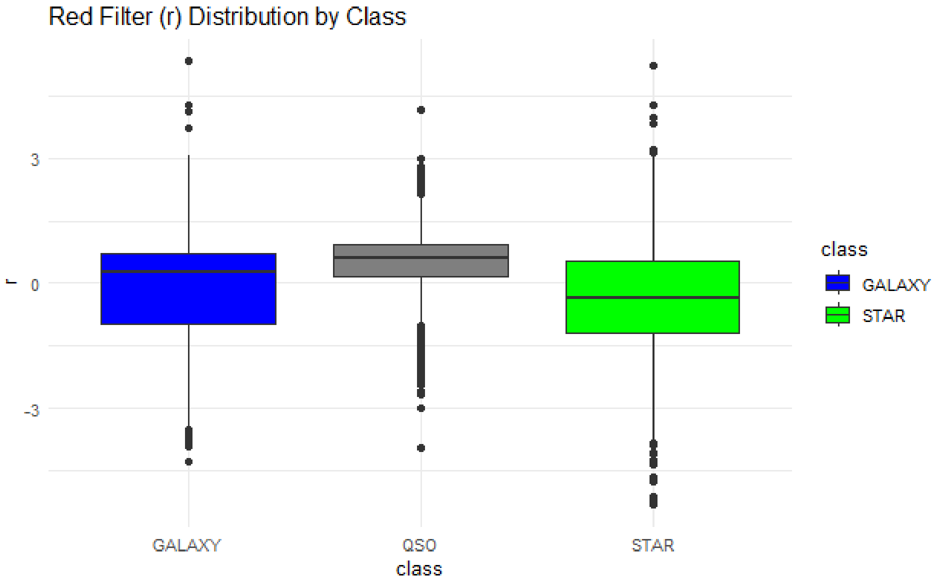


Figure 5: Boxplot for Red Filter (r) Distribution by Class.

The distribution of red filter (r) values for three classes—galaxy, QSO (quasar), and star—is compared using a box plot. It demonstrates that quasars have a wider range around zero, implying greater variability, whereas galaxies have more concentrated r values. Galaxies and stars share a similar distribution, with the exception of a somewhat higher median and more severe outliers. This indicates that different classes have different red filter values, with stars and quasars appearing more dispersed than galaxies.

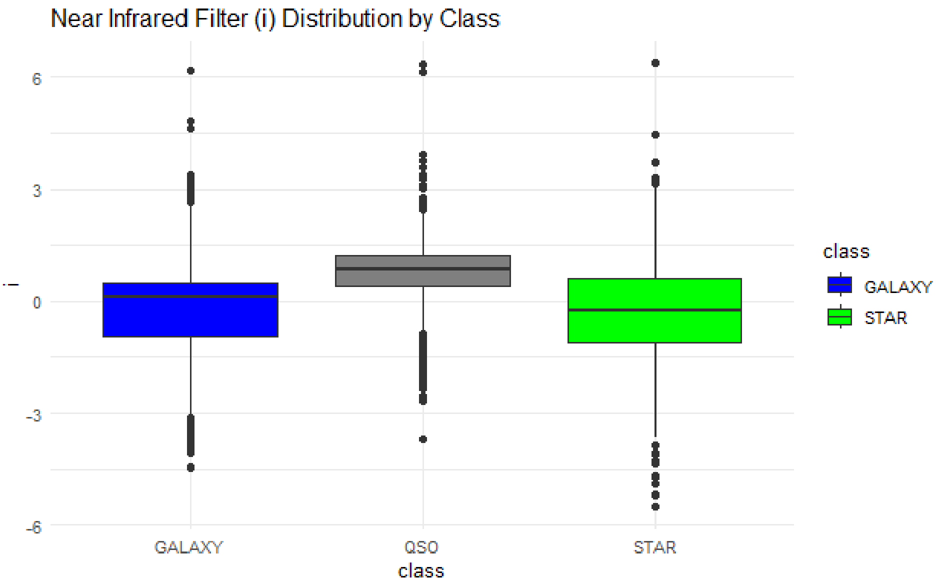


Figure 6: Boxplot for Near Infrared Filter (i) Distribution by Class.

The distribution of the near-infrared filter (i) values among the three classes—GALAXY, QSO, and STAR—is depicted in the box plot. The median value is displayed by the middle line, which indicates the interquartile range (IQR) for each box. The blue-colored GALAXY class has a symmetrical distribution with a few outliers on either end and a median that is marginally below zero. In comparison to the other classes, the QSO class, shown in gray, has a smaller IQR and a higher median around zero, suggesting less fluctuation. Similar to GALAXY, the STAR class, shown in green, has a median around zero as well as a broader spread, suggesting greater unpredictability and a number of outliers, particularly on the lower end. Overall, the plot illustrates variations in the distribution of near-infrared filter values and their variability among these astronomical classes.

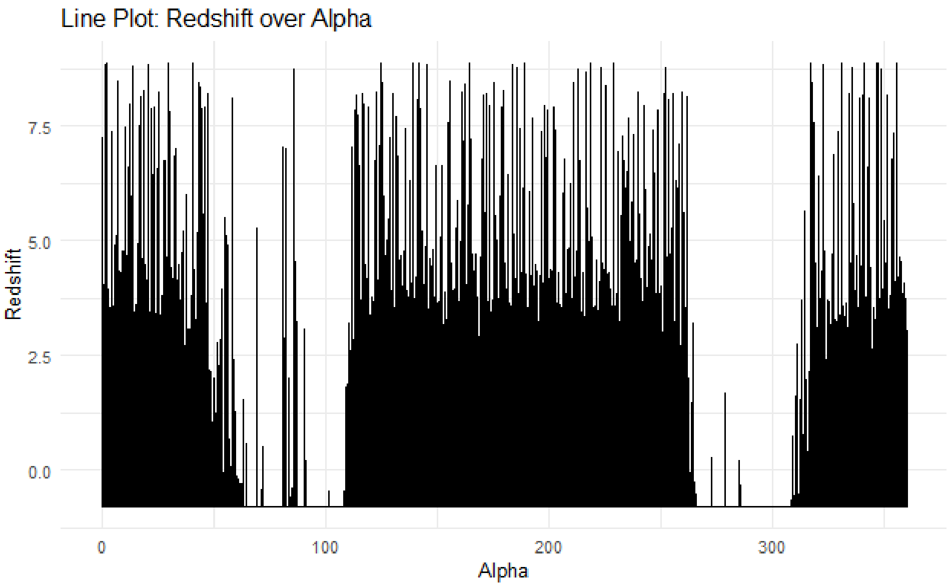


Figure 7: Line plot showing ”Redshift over Alpha.

The redshift fluctuation over the range of alpha values is shown in the line plot. Redshift fluctuates frequently and sharply, with values spanning from roughly 0 to above 7.5. The plot shows a very erratic relationship between redshift and alpha, with several peaks and troughs spread out along the alpha range and abrupt changes.

**4. Are there specific photometric characteristics associated with galaxies of certain classes (e.g., different redshift values or classifications such as Galaxy)?**

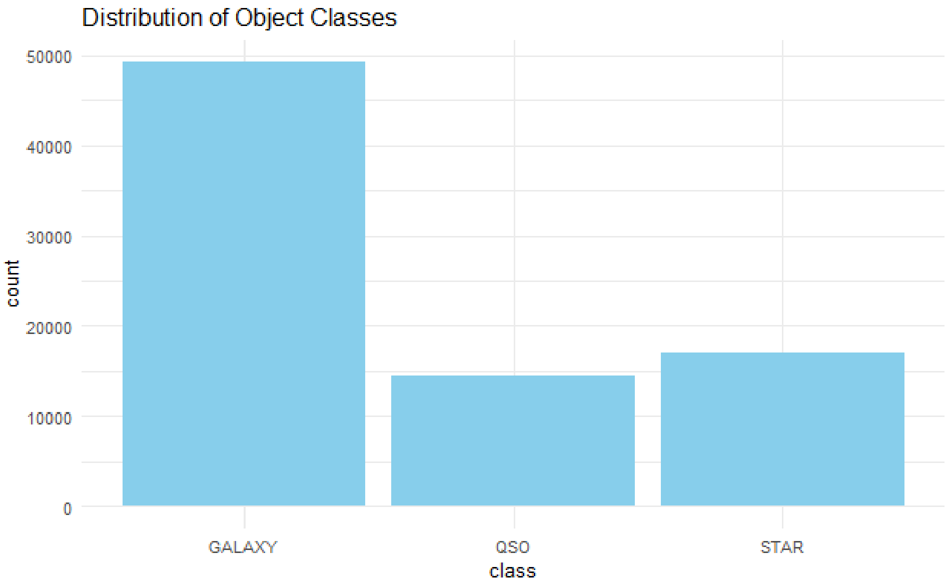


Figure 8: Distribution of Object Classes

The distribution of the GALAXY, QSO, and STAR object classes is illustrated in the above figure as a bar chart. For every class, the number of objects is represented by the height of the bar. Around 50,000 objects make up the GALAXY class, which has the greatest number. Around 15,000 is the lowest count for QSO, and roughly 25,000 is for STAR. In this dataset, galaxies are the most common object class, followed by stars and quasars (QSOs), as the chart graphically illustrates.

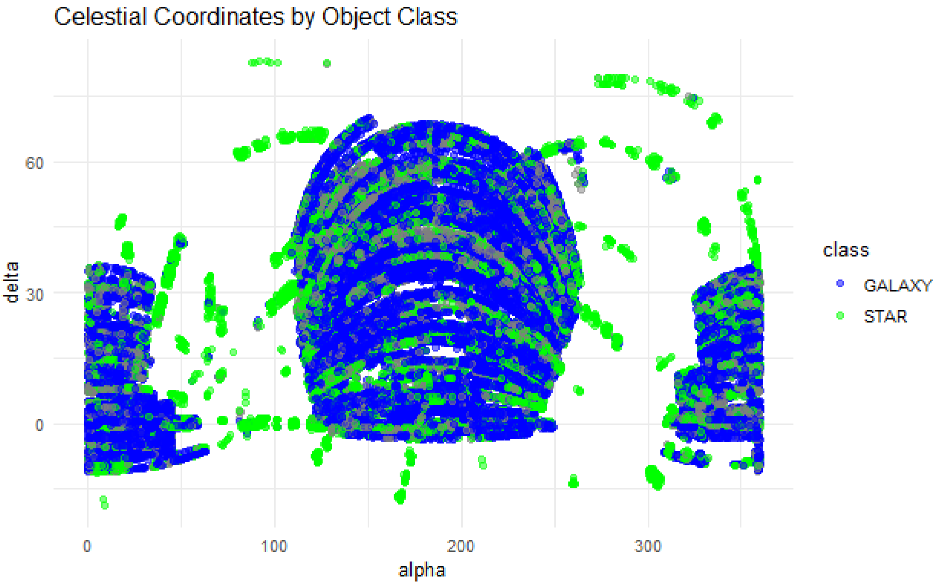


Figure 9: Celestial Coordinates by Object Class

The above scatter plot illustrates the celestial coordinates (alpha and delta) for two object classes: stars (green) and galaxies (blue). Alpha, or likely right ascension, is shown by the horizontal axis, and delta, or likely declination, is represented by the vertical axis. Clusters of data points that represent the locations of these objects in celestial coordinates are dispersed around the plot. There may be variations in the spatial distribution of stars (green points) and galaxies (blue points) in the sky as indicated by the overlapping and distinct zones between them.

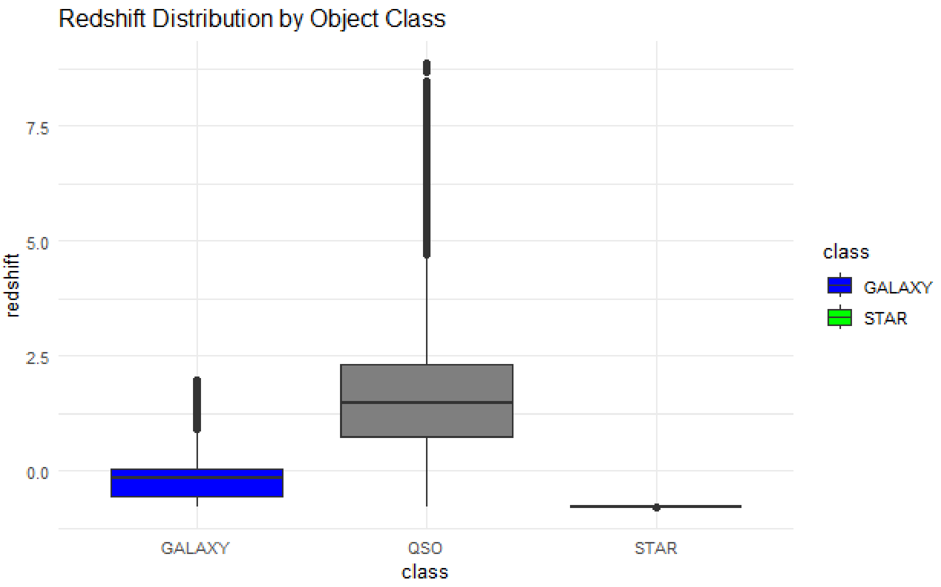


Figure 10: Boxplot for redshift by class

The redshift distributions for the three object classes—GALAXY, QSO, and STAR—are displayed in this box plot. Galaxies are quite close to Earth because of their low redshift values. With a median redshift of about 2, quasars (QSO) have the greatest and most diverse redshift distribution, indicating that they are generally significantly farther away. Stars are relatively close in comparison, as shown by their redshift close to zero. This shows how far apart these objects are from one another in the universe, with quasars being the furthest apart.

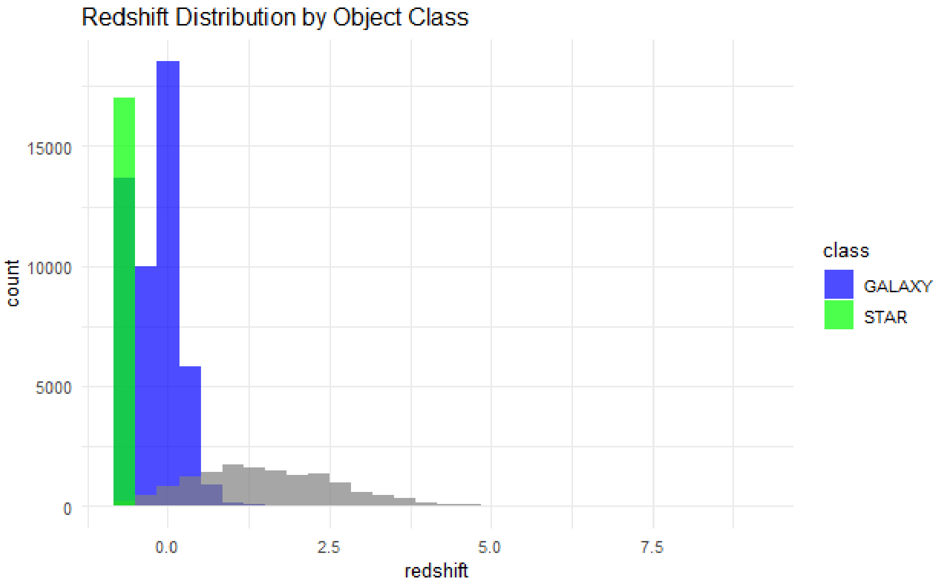


Figure 11: Histogram of redshift across object classes

The redshift distributions for QSO (gray), STAR (green), and GALAXY (blue) are presented in this histogram. Near zero redshift indicates that most stars and galaxies are close to Earth. The dispersion of quasars (QSOs) is wider; some have significantly greater redshifts, indicating that they are farther away. The graph illustrates how close stars and galaxies are to each other in relation to the farther-off quasars.

**5. How do photometric properties change across different epochs, as indicated by redshift?**

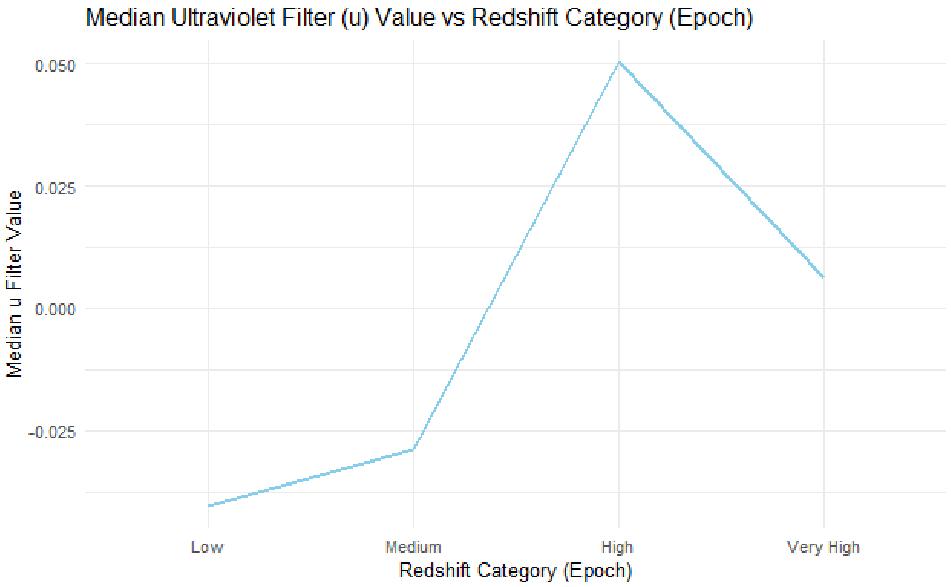


Figure 12: Line plot for median u filter value vs redshift bin

The link between the redshift categories (low, medium, high, and very high) and the median ultraviolet (u) filter value is depicted in this line graph. As redshift increases from low to high, the median u filter value rises and peaks in the high redshift group. In the extremely high redshift group, the median u filter value decreases after the peak. According to this, objects in the high redshift range have the largest UV filter value, but objects at extremely high redshift have lower UV filter values.

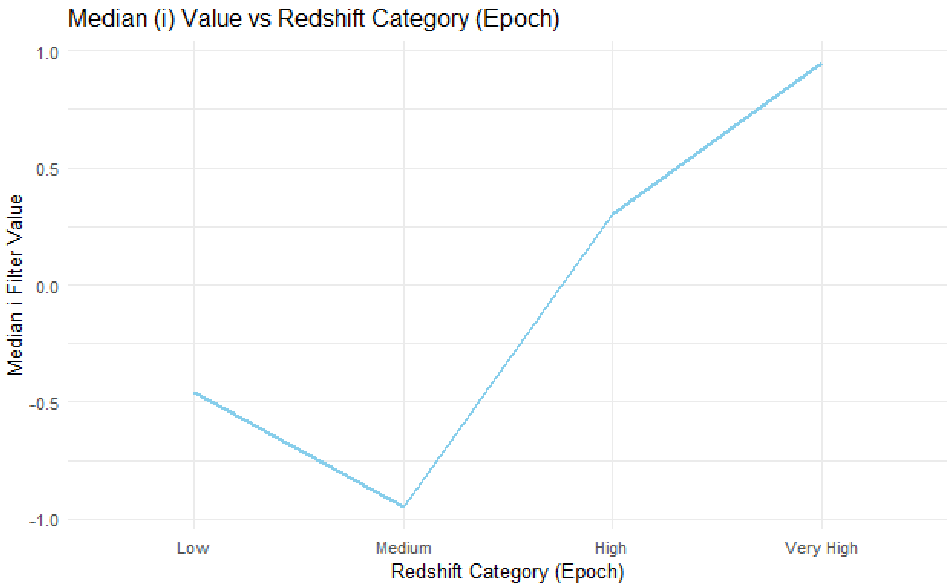


Figure 13: Line plot for median i filter value vs redshift bin

The median (i) filter value for each of the four redshift categories (low, medium, high, and very high) is displayed on this line graph. At low redshift, the median I filter value is negative; at medium redshift, it lowers even more; and finally, it starts to climb. In the high redshift category, it turns positive, and in the very high redshift category, it keeps rising. This shows a trend where the median i filter value first drops and then rises as the redshift grows, indicating a change in the properties of the objects that are observed as the redshift increases.

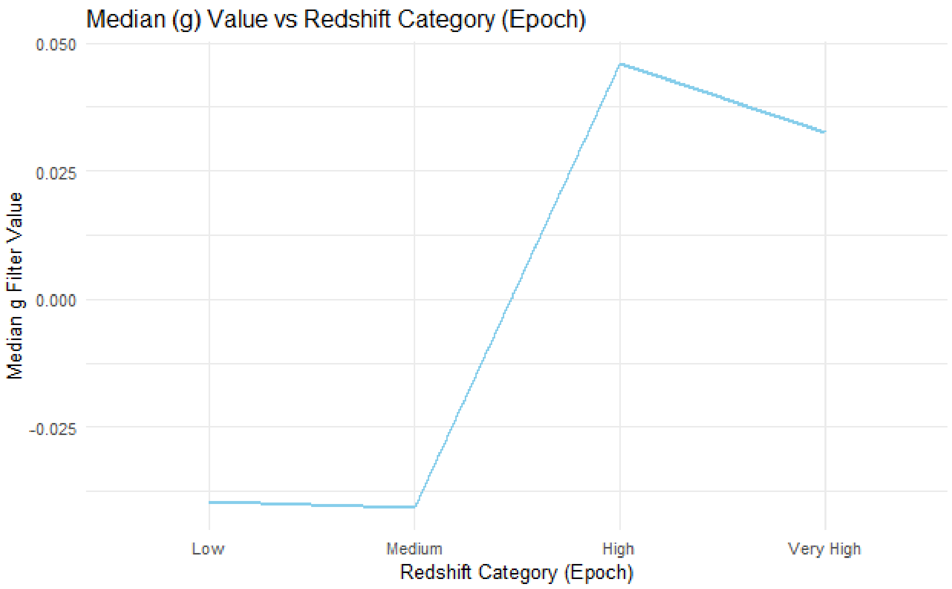


Figure 14: Line plot for median g filter value vs redshift bin

The median (g) filter value is presented on this line graph for each of the four redshift categories (low, medium, high, and very high). In the low and medium redshift groups, the median g filter value is remarkably constant, almost at zero. After that, it peaks at high redshift and then rapidly grows before slightly declining in the extremely high redshift range. According to this, the g filter value is steady at lower redshifts but rises noticeably for objects at higher redshifts before somewhat decreasing at extremely high redshifts.

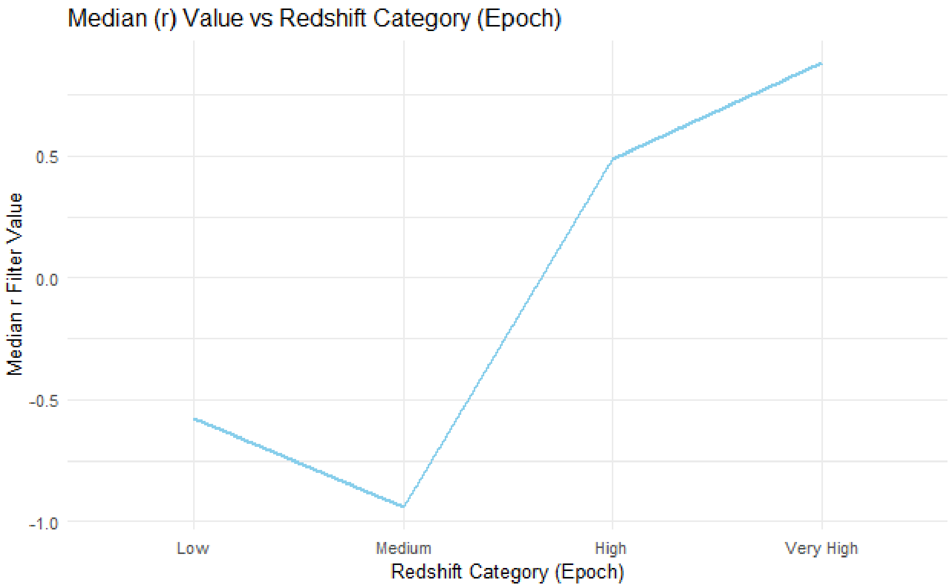


Figure 15: Line plot for median r filter value vs redshift bin

The median (r) filter value is represented on this line graph for each of the four redshift categories (low, medium, high, and very high). In the low redshift group, the median r filter value is negative; it then falls further in the medium category; finally, in the high redshift category, it starts to rise noticeably and turns positive. In the very high redshift range, it keeps becoming higher. This suggests that the median r filter value has a positive trend at higher redshifts and declines at lower redshifts before rising continuously as the redshift increases.

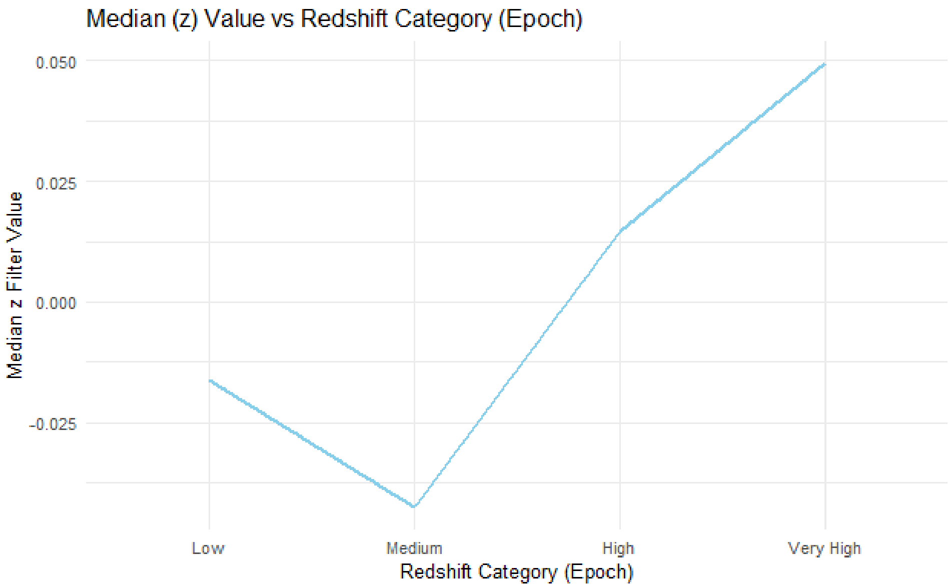


Figure 16: Line plot for median z filter value vs redshift bin

The above line graph illustrates the redshift categories' median (z) filter value (low, medium, high, extremely high). In the low redshift category, the median z filter value is first slightly negative, drops further at medium redshift, and then starts to rise. In the very high redshift category, it continues to rise after turning positive in the high redshift group. This shows a decreasing tendency at lower redshifts and a continuous increase in the filter value as the redshift increases.

**6. What are the correlations between celestial coordinates, photometric bands, and redshift in the dataset?**

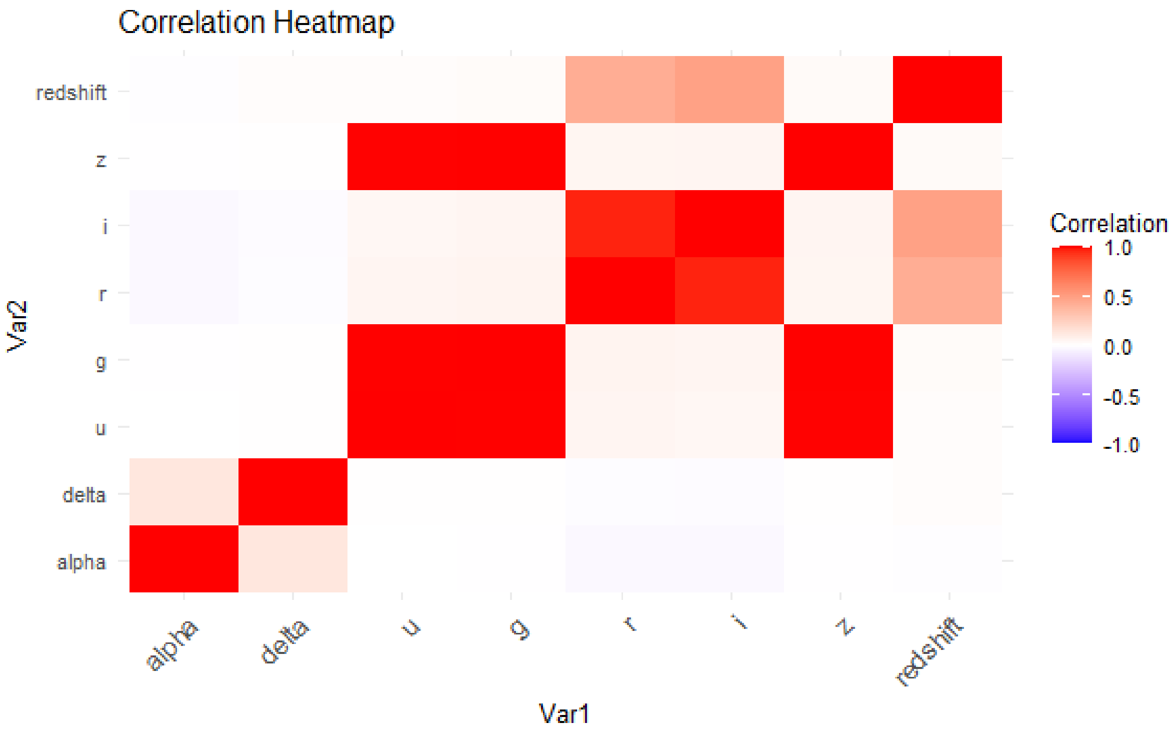


Figure 17: Correlation Heatmap

A correlation heatmap illustrating the associations between various variables (alpha, delta, u, g, r, i, z, and redshift) is shown in this picture. Red denotes strong positive correlations, whereas lighter colors indicate weaker or no connections. The colors indicate the direction and intensity of the correlations. White or light colors of red imply little to no correlation, while darker hues of red around 1.0 suggest a significant positive link. As an illustration, alpha and delta have little to no association with the majority of the other variables, however redshift has high positive correlations with the z, i, r, and g filters.

**7. Why might the distribution of stars and galaxies remain relatively consistent across the different filters?**

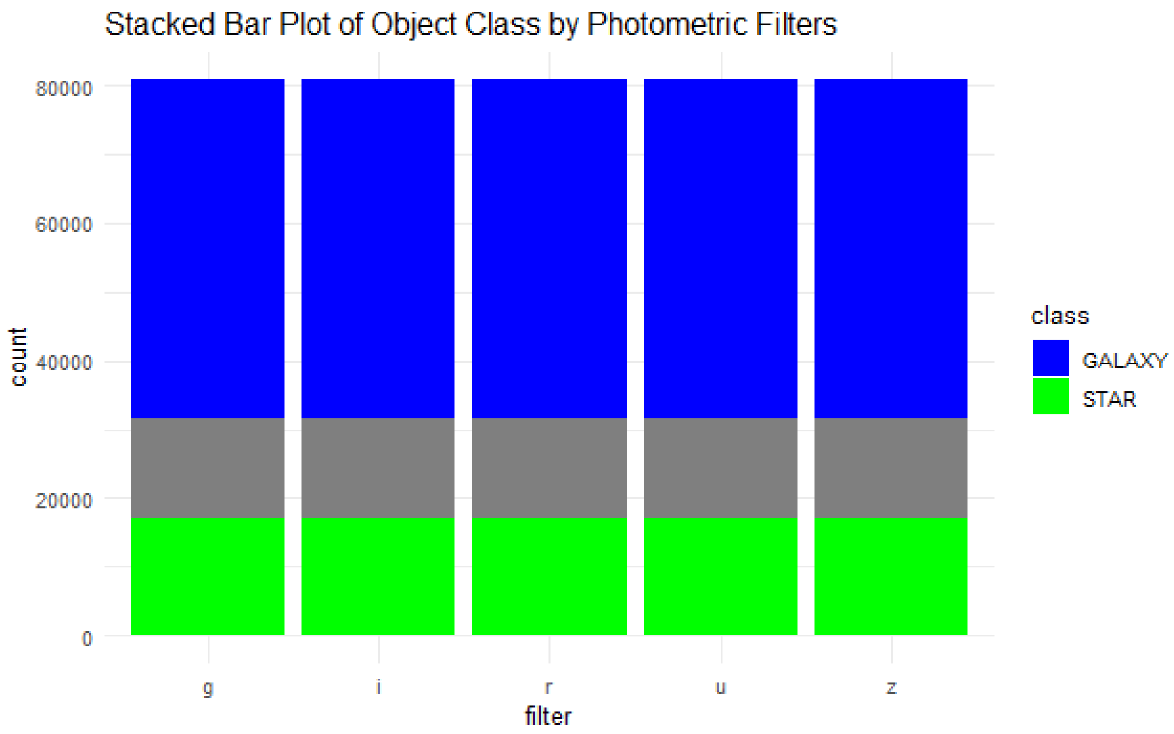


Figure 18: Stacked bar plot for object class distribution by filters

The above-stacked bar plot illustrates how the object classes (STAR and GALAXY) are distributed among the various photometric filters (g, i, r, u, z). Stars are displayed in green and galaxies in blue on each bar, which indicates the total number of observations for a particular filter. Every filter has the same distribution: galaxies are substantially more numerous than stars in every filter. This implies that throughout all photometric bands, galaxies predominate in the dataset.

**8. What trends or patterns are visible in the redshift distribution over time?**

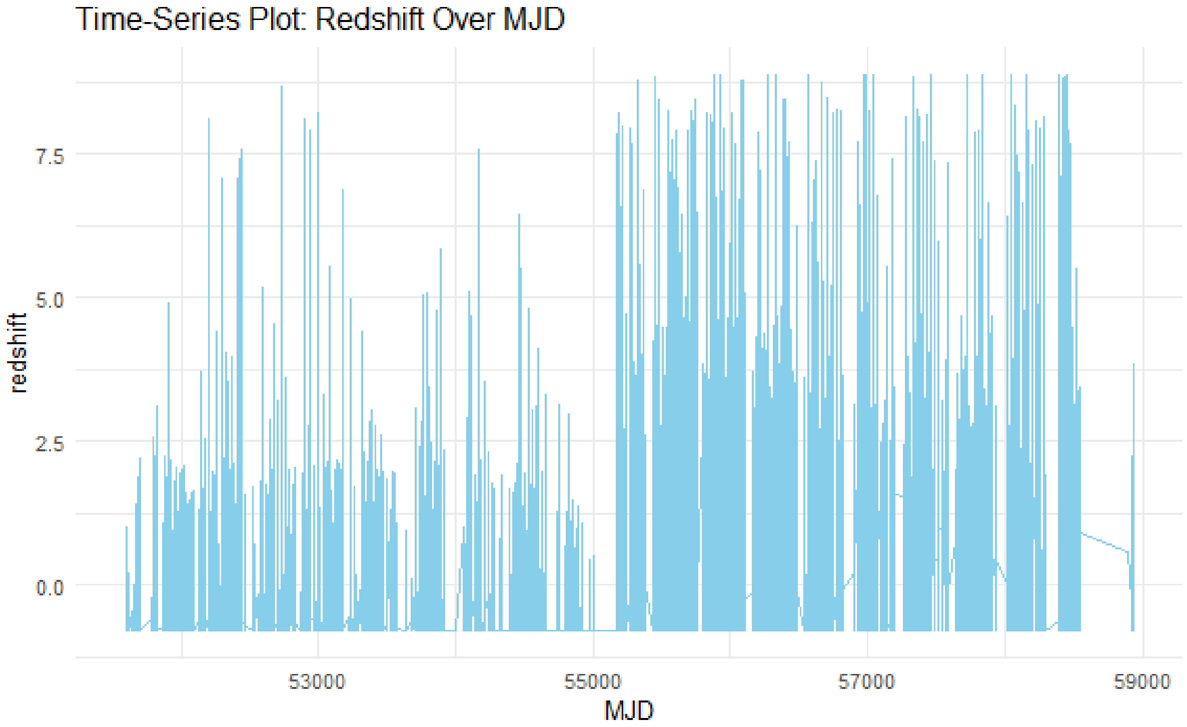


Figure 19: Time-series plot of redshift over MJD

This Modified Julian Date (MJD) time-series diagram illustrates how redshift changes throughout time. Throughout the time span, there is a noticeable variation in the redshift values, which can reach values as high as 7.5. While other periods have fewer or lower redshift values, there are periods of dense observations with significant redshift variability, indicating times when many objects were spotted. The map illustrates patterns in the observation of far-off objects throughout time, displaying objects at different redshifts and at high and low redshifts being photographed at different times.

**9. How do the distributions in this plot relate to the observations in the stacked bar plot of object class by filter?**

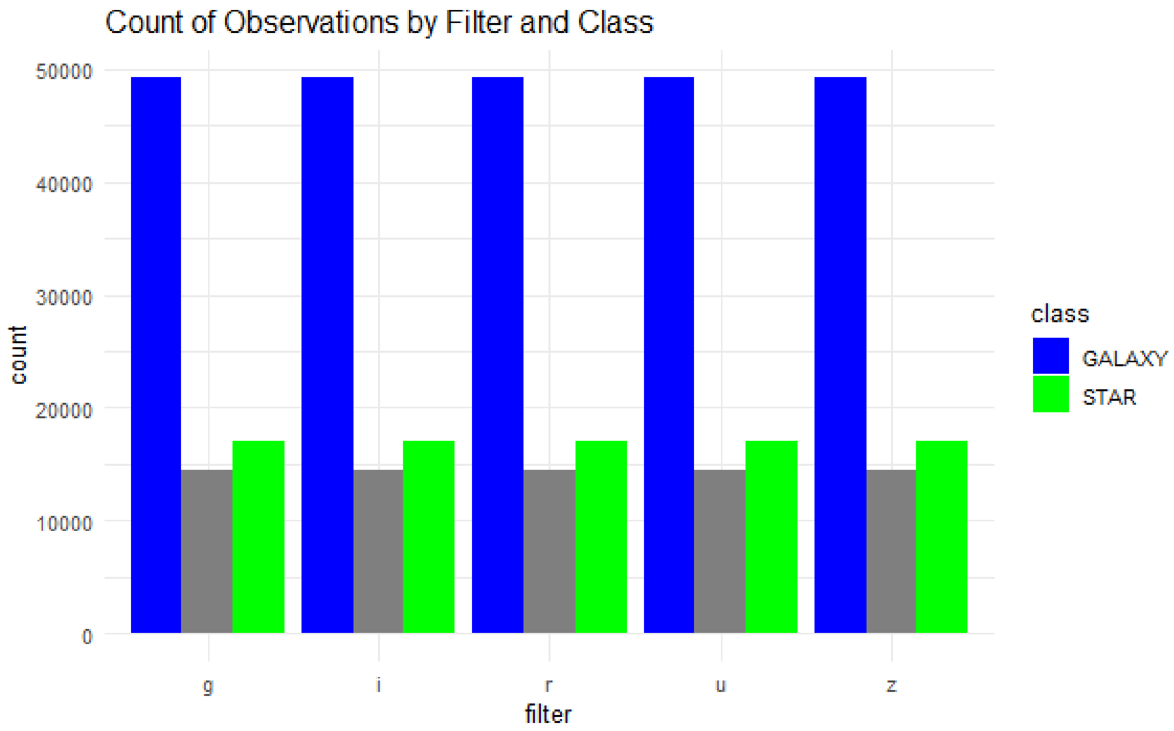


Figure 20: Count of observations by filter and class

The above bar plot depicts the number of observations for STAR (green) and GALAXY (blue) for each of the photometric filters (g, i, r, u, z). Galaxies are significantly more numerous than stars for every filter, and their distribution is constant. The stacked bar plot of object class by the filter, in which stars made up a lesser percentage and galaxies dominated the total count in all filters, closely resembles this plot. The fact that galaxies are spotted much more frequently than stars across all filters is supported by both graphs, which further indicates that galaxies comprise the majority of the sample throughout all photometric bands.

# 9 Summary

The examination of the data visualization concentrated on using information from the Sloan Digital Sky Survey (SDSS) to comprehend the properties and connections of celestial objects, such as stars, galaxies, and quasars. The study examined essential characteristics, including redshift and photometric measurements across multiple filters (u, g, r, i, and z), examining correlations and spatial distributions using a variety of visualizations. It was discovered that galaxies dominated the dataset in terms of both quantity and redshift range and that quasars, which are extremely distant objects, had the greatest redshift values. The work provided a clear explanation of how photometric characteristics change with redshift and how the distribution of celestial objects varies across time and space. It also demonstrated how the responses of stars and galaxies to different filters vary, with stars responding to most filters more frequently than galaxies.

# 10 Future Work

More research could focus on how redshift changes over time and how it affects the properties of celestial objects that are detected. More advanced machine learning models could also be used to enhance object classification, with an emphasis on hidden patterns and outlier identification that are not shown in traditional representations. Additional insights may be obtained by enlarging the dataset by including more comprehensive spectral data or by combining additional astronomical studies. Lastly, examining how various cosmological models affect the data may aid in improving our understanding of how the universe is expanding.

# 11 Conclusion

To sum up, the distribution and properties of galaxies, stars, and quasars have all been better understood because of this analysis of SDSS data. Redshift is crucial for categorizing and comprehending these objects, as the visualizations are made clear by highlighting the connections between redshift and photometric characteristics. The dataset revealed that galaxies dominated it and that quasars, with their larger distance from Earth, had higher redshift values. The study lays the foundation for further astronomical research by expanding our knowledge of the structure and evolution of the cosmos.

# 12 References

Abramov, D. et al. (2022) ‘CosmoVis: An interactive visual analysis tool for exploring hydrodynamic cosmological simulations’, IEEE Transactions on Visualization and Computer Graphics, 28(8), pp. 2909–2925. doi:10.1109/tvcg.2022.3159630.

de Souza, R.S. and Ciardi, B. (2015) ‘Amada—analysis of multidimensional astronomical datasets’, Astronomy and Computing, 12, pp. 100–108. doi:10.1016/j.ascom.2015.06.006.

Pieringer, C. et al. (2019) ‘An algorithm for the visualization of relevant patterns in astronomical light curves’, Monthly Notices of the Royal Astronomical Society, 484(3), pp. 3071–3077. doi:10.1093/mnras/stz106.

Schlarmann, L. et al. (2021) ‘C60+ diffuse interstellar band correlations and environmental variations’, Astronomy amp; Astrophysics, 656. doi:10.1051/0004-6361/202142669.