Leonardo Bocchi, Pol Riba

**EXPLORATORY DATA ANALYSIS:**

Presentation and Visualization - 2022/2023

**SUMMARY**

[**1. Dataset introduction**](#_mxniafq17o1s) **3**

[**2. Variable choice**](#_byxupqm0nax4) **3**

[**3. Variable introduction**](#_hg904lqo6yu7) **4**

[3.1 Categorical/Class Features](#_dabqicd5f43l) 4

[3.2 Numerical/Measure Features](#_llxaqhjve5qr) 4

[**4. Dataset cleaning**](#_ktcn1v4bflty) **5**

[**5. Variable creation**](#_dwqwh4jjgspe) **6**

[**6. Distribution of the data**](#_er0namiwj5ep) **7**

[6.1 Distribution of the samples by class](#_t2siric60lc5) 7

[6.2 Measure variables by class](#_smmbsqerdm00) 8

[**7. Outlier detection**](#_5xlb95goot7i) **10**

[**8. Scope and audience**](#_snfj7xq5wfci) **13**

[**9. Idea of the final result**](#_x9sdper9eake) **14**

[**10. Charts selection and encoding**](#_5fdfbvlx9r4p) **15**

[10.1 Chart selection](#_7twhi06fhldy) 15

[10.2 Encoding and perception](#_2etzpaec46qf) 16

[**11. Perception and best practices**](#_ipoahpehn21i) **16**

[11.1 Gestalt principles and application](#_dp935vx9zl0o) 16

[**12. Interaction**](#_u9wnumji3x11) **17**

[12.1 Filtering](#_qjxh346ojgqa) 17

[**13. Layout and content of the dashboards**](#_3woh1e52iu93) **17**

[**14. Implementation of the Dashboards**](#_5krt6gqkbzqx) **19**

[14.1 Design of the Dashboards](#_ivnib37e8kqu) 19

[14.2 First implementation of the Dashboards](#_t5pmcnfm2ojf) 20

[14.3 Further implementations](#_ht34u2jwe7zy) 21

# **1. Dataset introduction**

The dataset that will be studied in this project is a star description dataframe which gathers information on observation relying on some astrophysics equations. These are given below:

1. Stefan-Boltzmann's law of Black body radiation (To find the luminosity of a star)
2. Wienn's Displacement law (for finding surface temperature of a star using wavelength)
3. Absolute magnitude relation
4. Radius of a star using parallax

The dataset took 3 weeks to collect for a total of 240 samples, and it is a balanced dataset, which means that it has the same representation for each kind of star. This implies that we have the same amount of samples for each type of star. We chose to work with this dataset because of our personal interest on the subject as well as some specific ideas for visually exploring the dataset.

Each observation in the dataset can either be a star or a cluster of stars. The time span of the dataset is not particularly dynamic, therefore, it does not require frequent updates, in order to be veridical.

The dataset used in this visualization project can be also found at the following link: <https://www.kaggle.com/datasets/deepu1109/star-dataset>

# **2. Variable choice**

Initially, exploring the dataset we had to decide which variables or features we were interested in. As the number of variables was not particularly large and the dataset was already in good conditions for posterior analysis, we decided to keep all of them.

Every variable in the dataset may have meaning and it is interesting to us; having a larger number of features will help our visualization to be clear and transversal.

To be more precise, if we had found in the dataset variables of small interest or with a big percentage of missing values (greater than 75%) we would have thought about dropping those features.

In the following section, the different variables which will be considered in this study are introduced.

# **3. Variable introduction**

## **3.1 Categorical/Class Features**

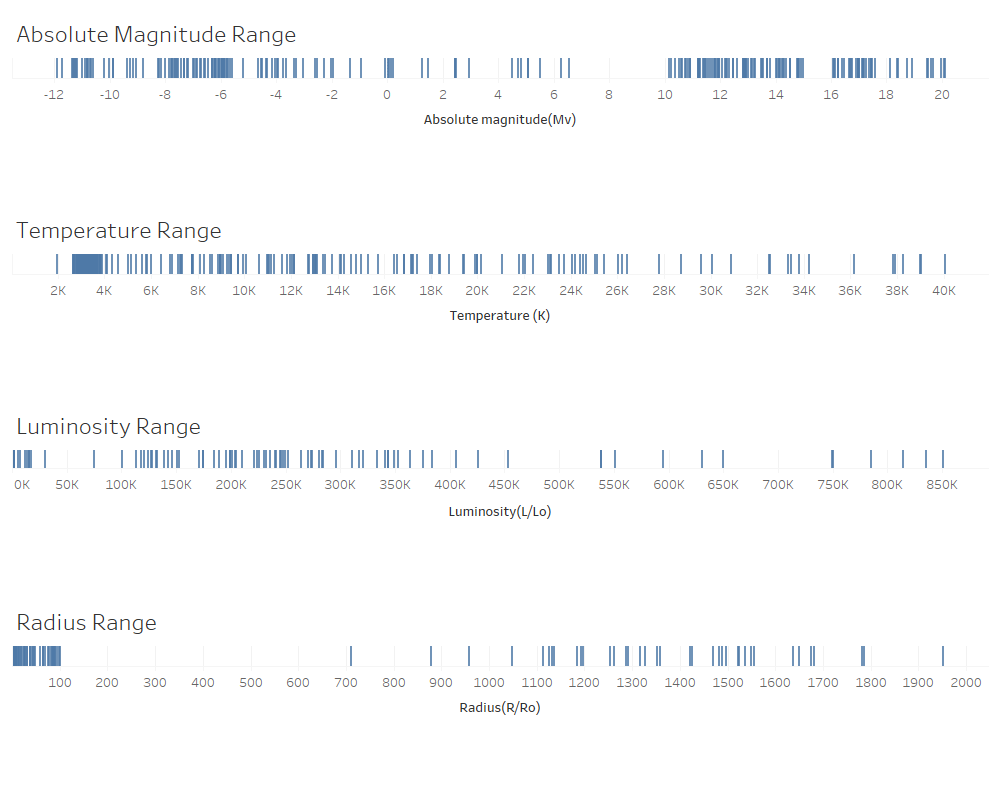
* **Star Color:** Color of the star after Spectral Analysis. These are the classes of the spectrum considered: Blue, Blue White, White, Whitish, Yellow White, Yellowish, Orange, Orange Red, Red
* **Spectral Class:** Spectral class of each star. The classes that can be found in the dataset are the following: O, B, A, F, G, K, M.
* **Star Type:** Output class, dependent on the star classification. These are the classes available: “Brown Dwarf”, ”Red Dwarf”, “White Dwarf”, “Main Sequence”, “SuperGiant”, “HyperGiant”.

## 3.2 Numerical/Measure Features

* **Absolute Magnitude (Mv):** Absolute Visual magnitude (Mv) of several stars. Range: [-12, 20].
* **Absolute Temperature (K):** Surface temperature of the star in Kelvin. Range: [1900, 40000].
* **Relative Luminosity (L/Lo):** Luminosity of the star with respect to sun (L/Lo). Range: [0, 850000].
* **Relative Radius (R/Ro):** Radius of the stars calculated with respect to sun (R/Ro). Range: [0, 2000].

It should be noted that Luminosity and Radius are relative to the sun values, which explains why their value ranges even close to 0. This aspect of the dataset could distort the visualization; on the other hand, using the absolute values could require the use of considerably different orders of magnitude, making the samples difficult to compare. For this reason, as it will be noted in section *5. Variable creation*, we will also create the corresponding absolute values in order to be able to choose the best suitable approach depending on the needs of the visualization considered.

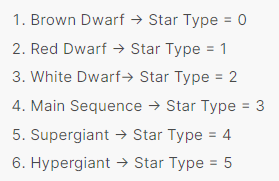
It follows a visual exploration of the values' ranges:



# 4**. Dataset cleaning**

Even though the dataset was already in good conditions for an analysis, meaning there were no features in which there could be found any missing variable, unwanted variables, etc, some values were not registered in the most optimal way.

The changes that were made to the dataset in order to make it better suited for visualization and study are the following:

* **Star Color:** some values were repeating but with different labels. (e.g. White-Yellow, White Yellow, Yellow-White, Wellow White) For this class the different labels indicating the same value were aligned with a common label
* **Star Type:** the star type value was encoded as a numerical value, ranging from 0 to 5. Each value refers to a different star type classification, according to a map given with the dataset. Here we include said map for easier consultation:

Considering the categorical nature of this feature, we decided to convert all the values to their corresponding label.

* **Spectral class:** another type of star classification, this is labeled using the letters O, B, A, F, G, K, and M. We didn’t change this feature of the dataset. However, we are expecting some issues with it, since it is categorical and ordered, but the way it is encoded does not preserve the order. Therefore, we will need to cosider this when visualizing the data.

As there was nothing else which could be an issue for the study, the dataset was deemed clean enough for our study and we proceeded with the following exploratory analysis.

# 5. Variable creation

* **Absolute Luminosity:** since the given values are relative to the Sun, it may be useful for visualization purposes to also have the absolute values of the Luminosity. These values were added under a new feature named “Absolute Luminosity”, which is computed using the following formula:
* **Absolute Radius:** since the given values are relative to the Sun, it may be useful for visualization purposes to also have the absolute values of the Radius. These values were added under a new feature named “Absolute Radius”, which is computed using the following formula:

From the Absolute Radius we decided to derive two more variables, describing the Area of the Section and the Volume. This was done because in some types of visualization it may be useful to have such measures, as they may be more meaningful than the Radius.

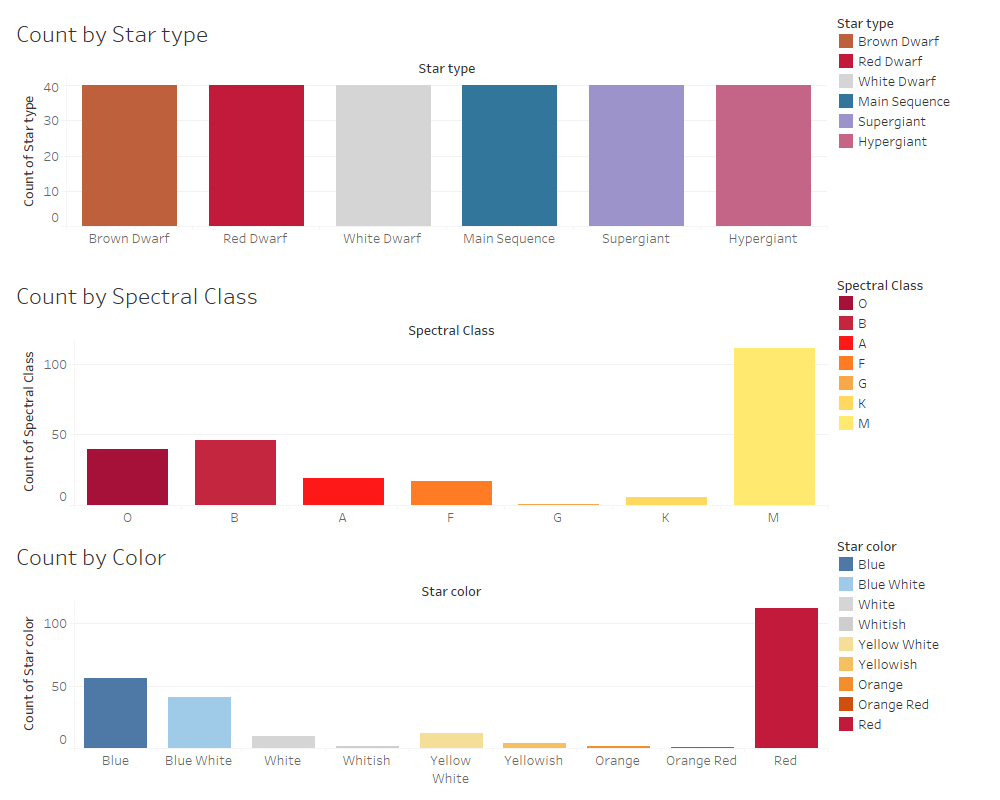
* **Area of Section:** the area of the section of a star, approximating the body to a sphere. It is computed using the formula for the area of a circle:
* **Volume:** the volume of a star, approximating the body to a sphere. It is computed using the formula for the volume of a sphere:

These last two values are computed as absolute values. The relative values were not computed since they would have had few information and uses. Firstly, they would have been more difficult to interpret, since they would have been relative to a large value. Also, the computation would have brought even more distortion to the measurement, since the formulas are not linear. The two relative variables would have been of hard interpretation and of less ‘truthfulness’ than their corresponding absolute variables.

# 6. Distribution of the data

## 6.1 Distribution of the samples by class

The distribution of the variables used in this project is not generated by a random variable or by sampling a population, but from selected empirical observations. By class, the distribution of the samples are the following:

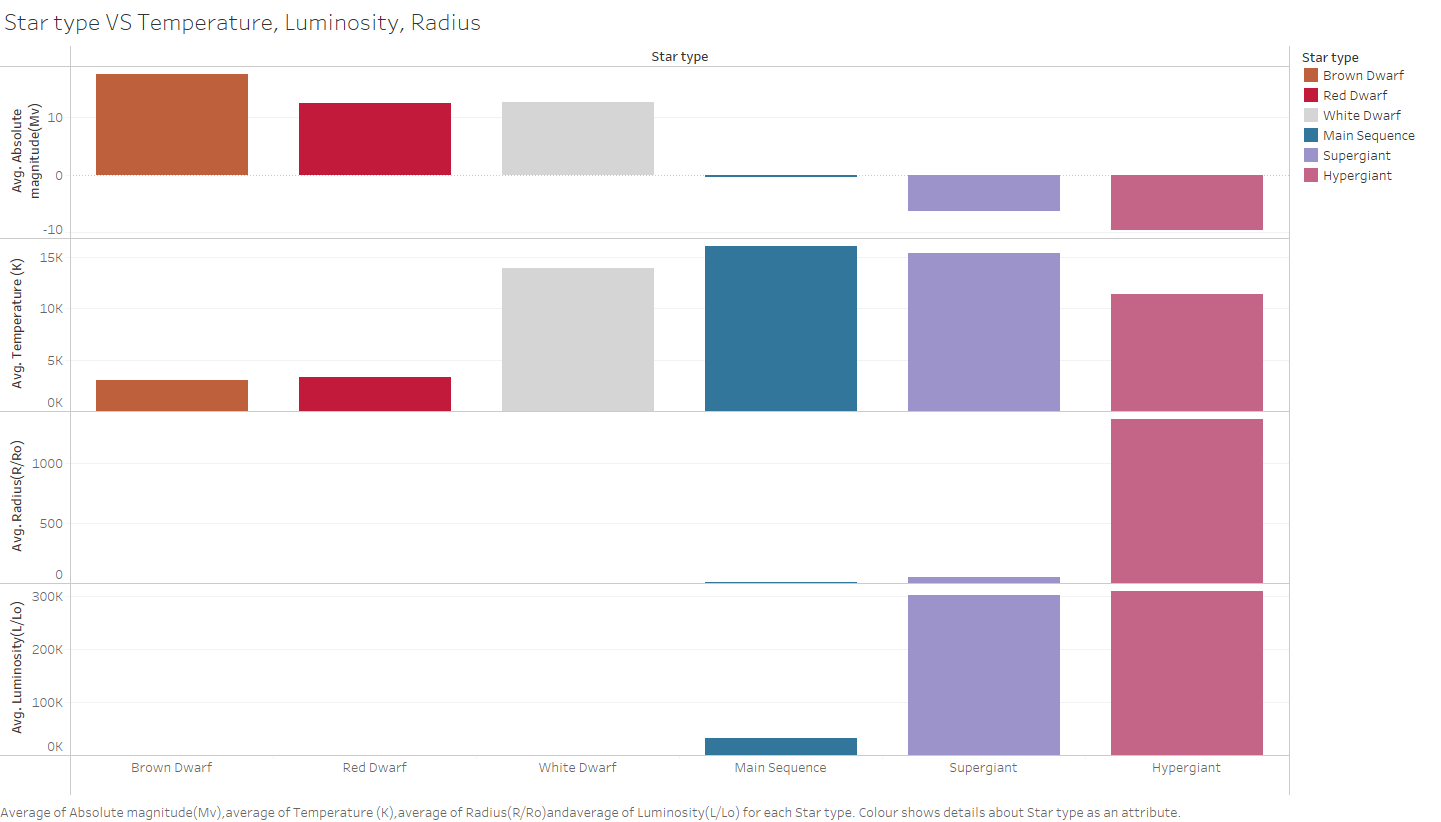


* **Count by Star type:** as previously noted, this dataset contains the same number of samples for each Star type taken into consideration
* **Count by Spectral Class:** as spectral class is a different kind of classification, the same uniformity of number of samples for each class is not present for this feature
* **Count by color:** this category shows a larger number of samples for some values. Furthermore, considering the nature of the value assigned to the observations, it is not an exact measure, and, for this reason, it strictly depends on the specificity of the scale of colors considered

## 6.2 Measure variables by class

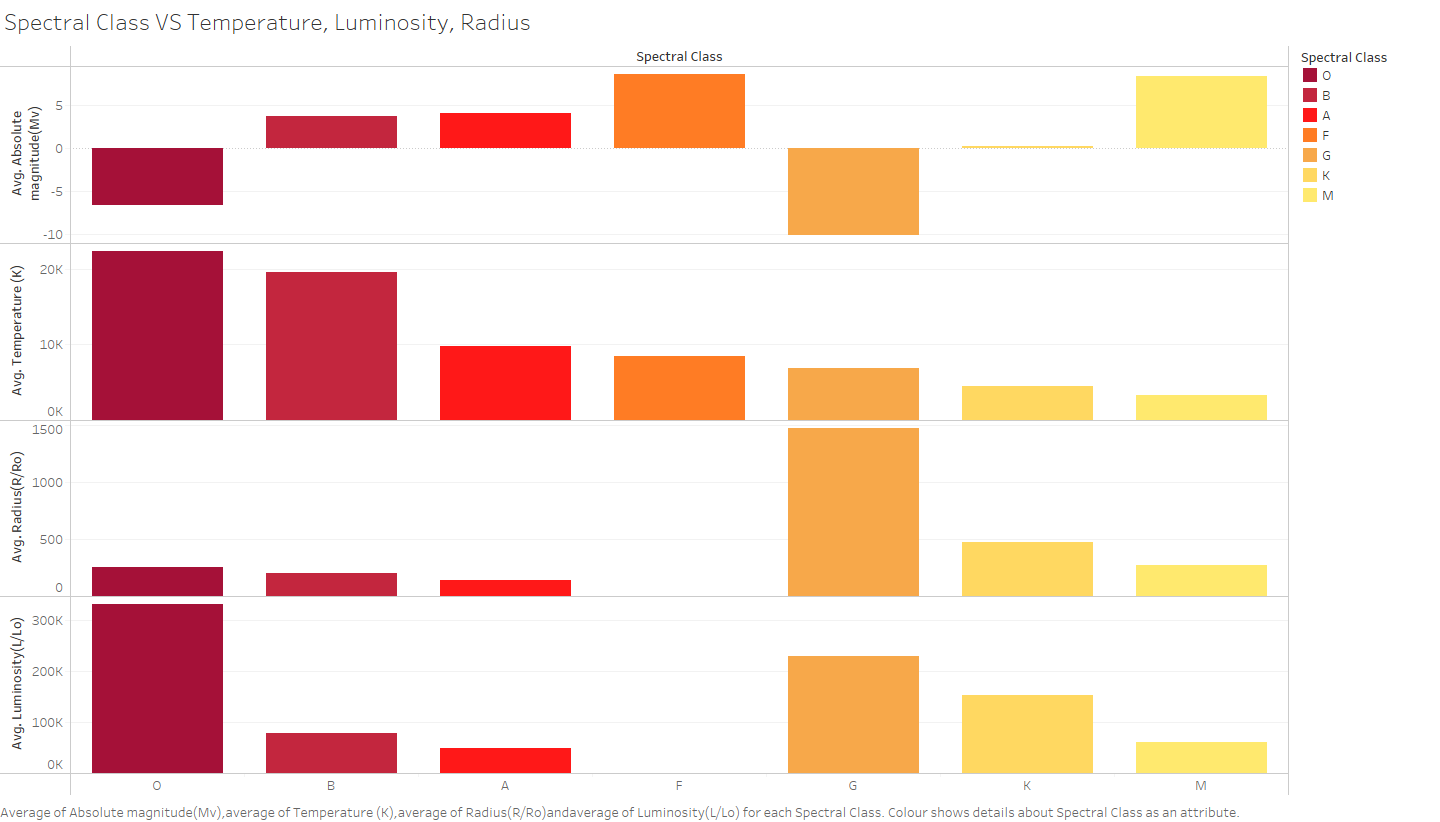
In this section we study the average of the numerical features grouped by categorical features. This is done in order to get some intuition over what the relationships within the variables and what the criteria used for classification might be.

* **Star type:**



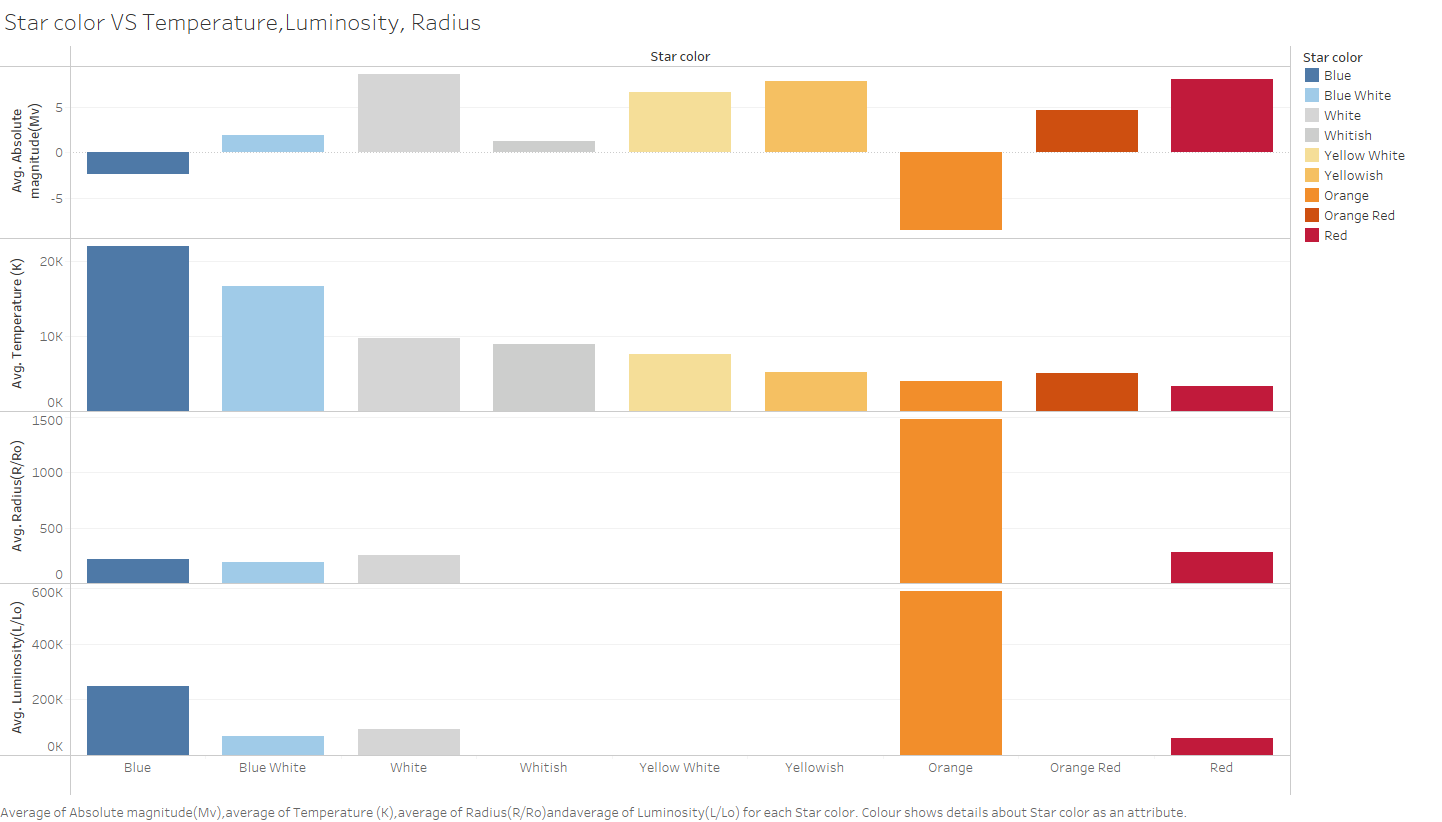
The plot shows a clear trend in the average Absolute Magnitude over the Star type. Other variables show similar trends as well but none as consistently. Therefore, this suggests that the main criteria for classifying stars by Star type is their Absolute Magnitude.

* **Spectral Class:**



The plot shows a clear trend in the average Temperature over the Spectral Class. Other variables show some trends as well but none of them is consistent on the whole classification spectrum. Therefore, we can gather that the main criteria for classifying stars by Spectral Class is their Temperature.

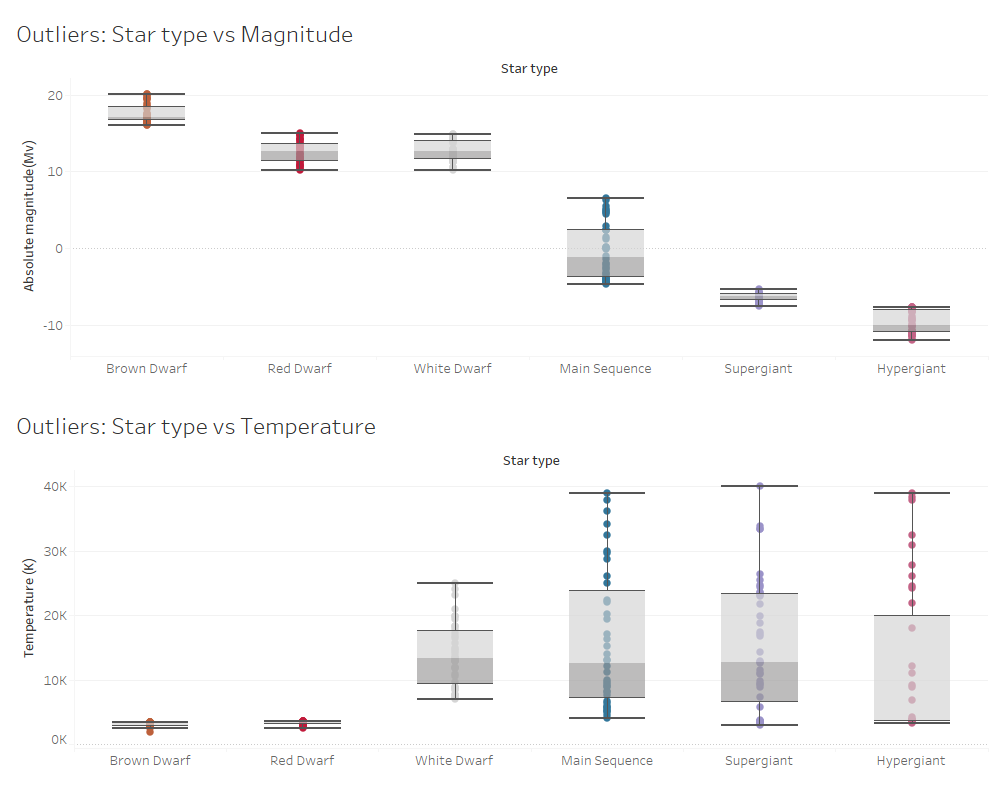
* **Star color:**

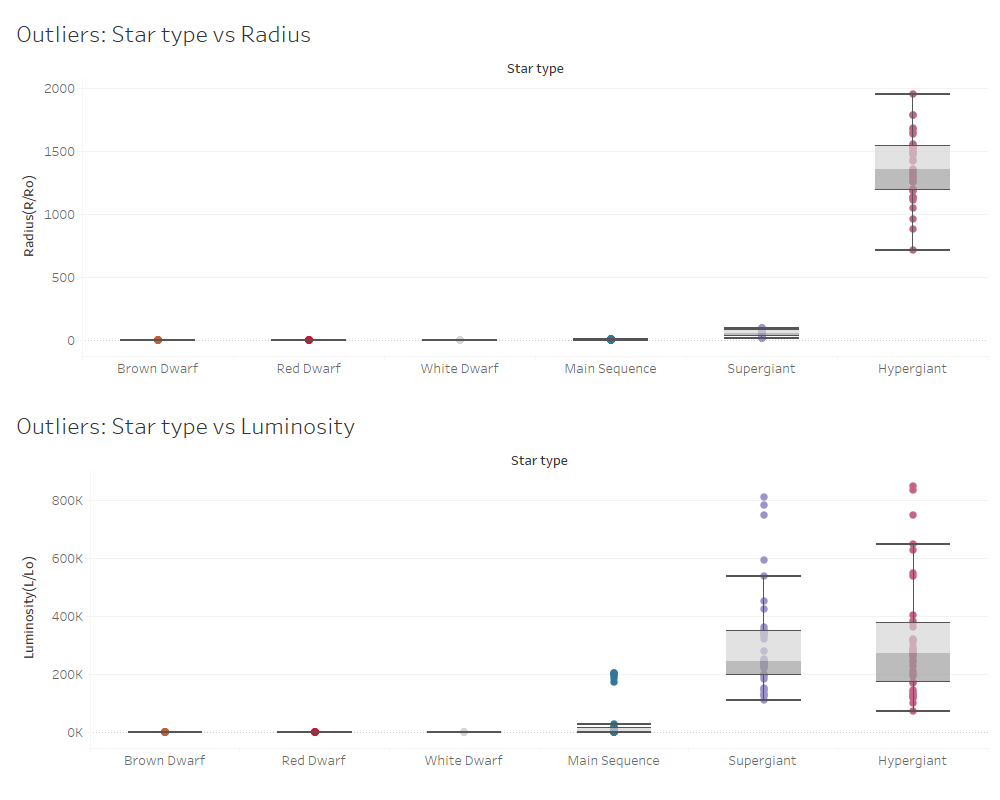


The plot shows a clear trend in the average Temperature over the Star color. Other variables do not show remarkable trends. The Star color classification is based on direct observation, which means it does not rely on a measurable variable. However, this allows us to form the hypothesis that the Star color and the Temperature of the star may be strictly correlated.

# 7**. Outlier detection**

As star types are classified by their features, every star needs to present certain qualities to be labeled by star type or by its spectral class. If a star does not fit the requirements of a particular class, it will not be classified as part of that class, therefore, allowing for no possible outliers. This can be noticed in the following representation:

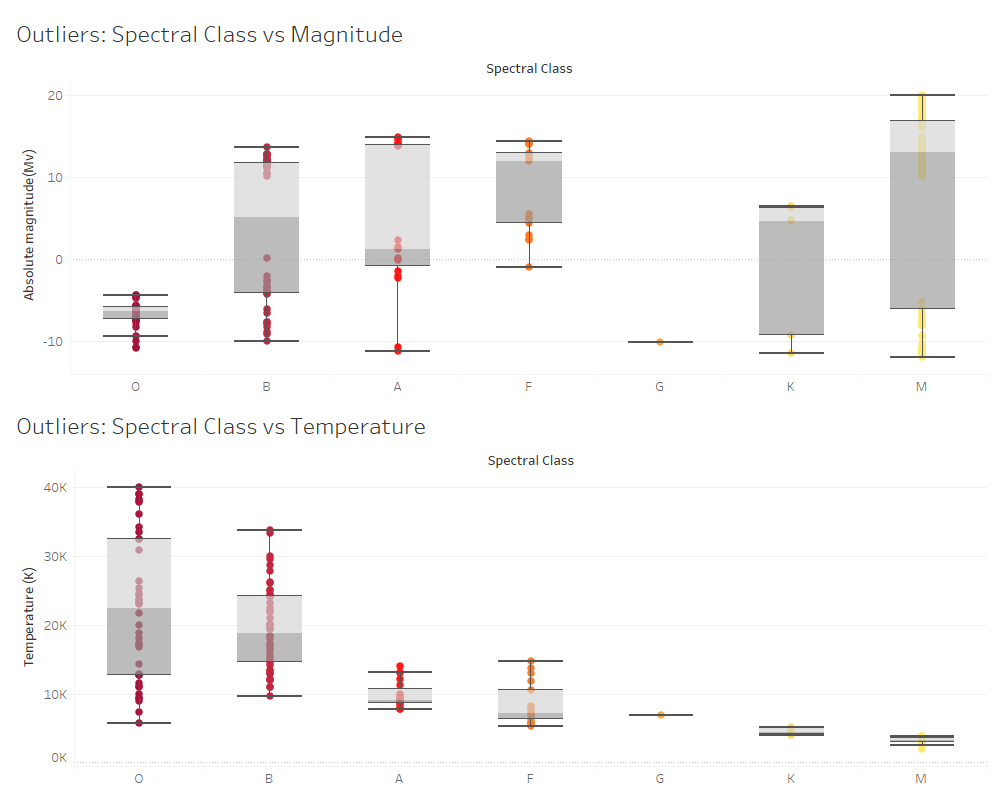


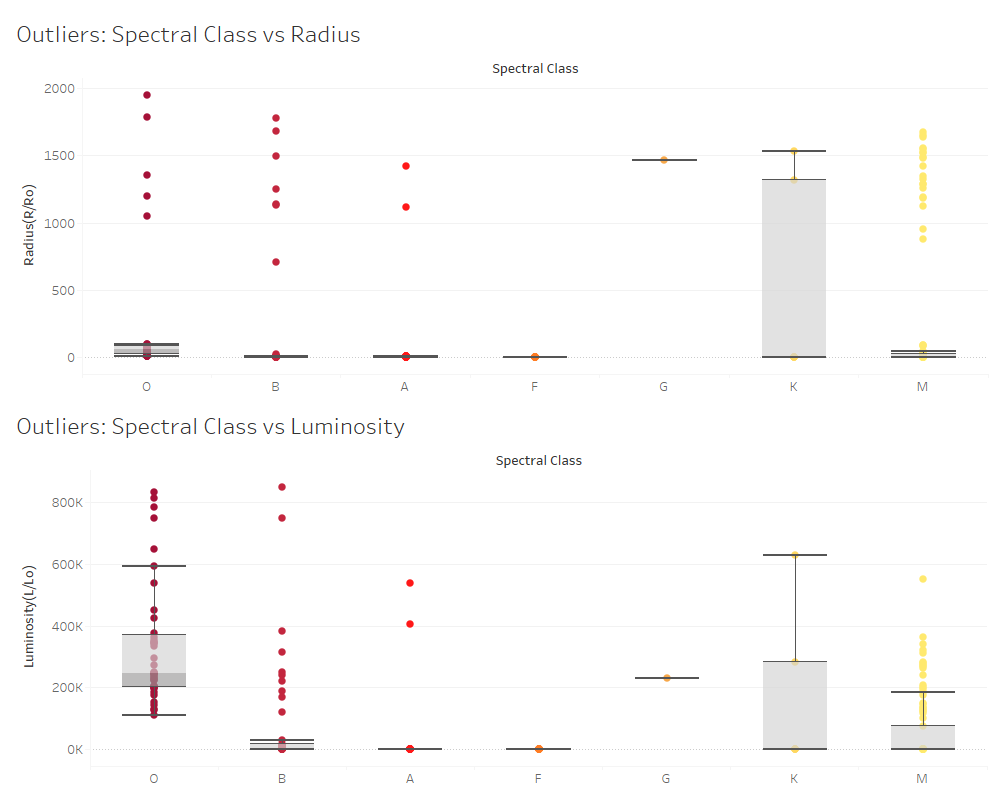


Considering the first two plots, it can be observed that there are no outliers when plotting the Star type against its Magnitude or its Temperature, as we were expecting from our previous considerations. This is congruent with the criteria used for classification; in fact, the star type of a particular observation is determined mainly by its Magnitude and its Temperature.

On the other hand, the plots of the Star type against the Radius and the Luminosity show some samples with values outside the class’ distribution range. The Radius values seem to almost follow the classification, probably because of its correlation with Magnitude and Temperature. However, the Luminosity measure shows many outliers inside different classes.

In the following representation we take into consideration the Spectral Class as the classification feature:





Similarly, we can observe that for the values of Magnitude and Temperature no outliers can be found over each Spectral Class. The reason is the same: the classification criteria for the Spectral Class of a star take into consideration mainly Magnitude and Temperature, which means most of the values in one class fall into the same range. On the other hand, for both Radius and Luminosity we find many different outliers, suggesting large variability over those two measures in each Spectral Class.

Since the dataset is gathered from observation and measurements that to us seem reasonably accurate, no samples are dropped in this study, in order to be able to study those samples and their variability as well.

# 8. Scope and audience

The audience of the visualization of a dataset like this would normally be someone from a small audience of science-related figures and people interested in Astronomy. However, to understand the visualization presented in this project it is not needed to have a wide background knowledge in Astronomy, because the variables considered are easy to understand and the bivariate or multivariate charts that will be used are intuitive.

Clearly, it is helpful to have some prior knowledge about the related phenomena, like understanding why the luminosity and color depends on the distance, etc. but it is intentionally not a requirement in order to understand this study.

Our audience does not need advanced knowledge in the field, since the visualization proposed will be focused on understanding the relationships of the main variables in the field of Astronomy.

The purpose of this visual exploration and presentation is to give a basic but solid and easy-to-approach understanding of the features Astronomy deals with.

We can imagine a researcher is working on a project about some phenomena regarding stars, willing to understand the differences and relationships between different types of stars (for example, why a particular type of star shows a specific color). This person also wants to have an easy to understand output, but with precise and scientific data supporting it. In order to do this, he wants to introduce some relationships analytically and describe them visually in a way everyone can understand them.

He could be interested in consulting our project, and using our dashboards.

In an eventuality like this one, the visualization could be consulted whenever there is any doubt on how the features are distributed and how they are related to each other.

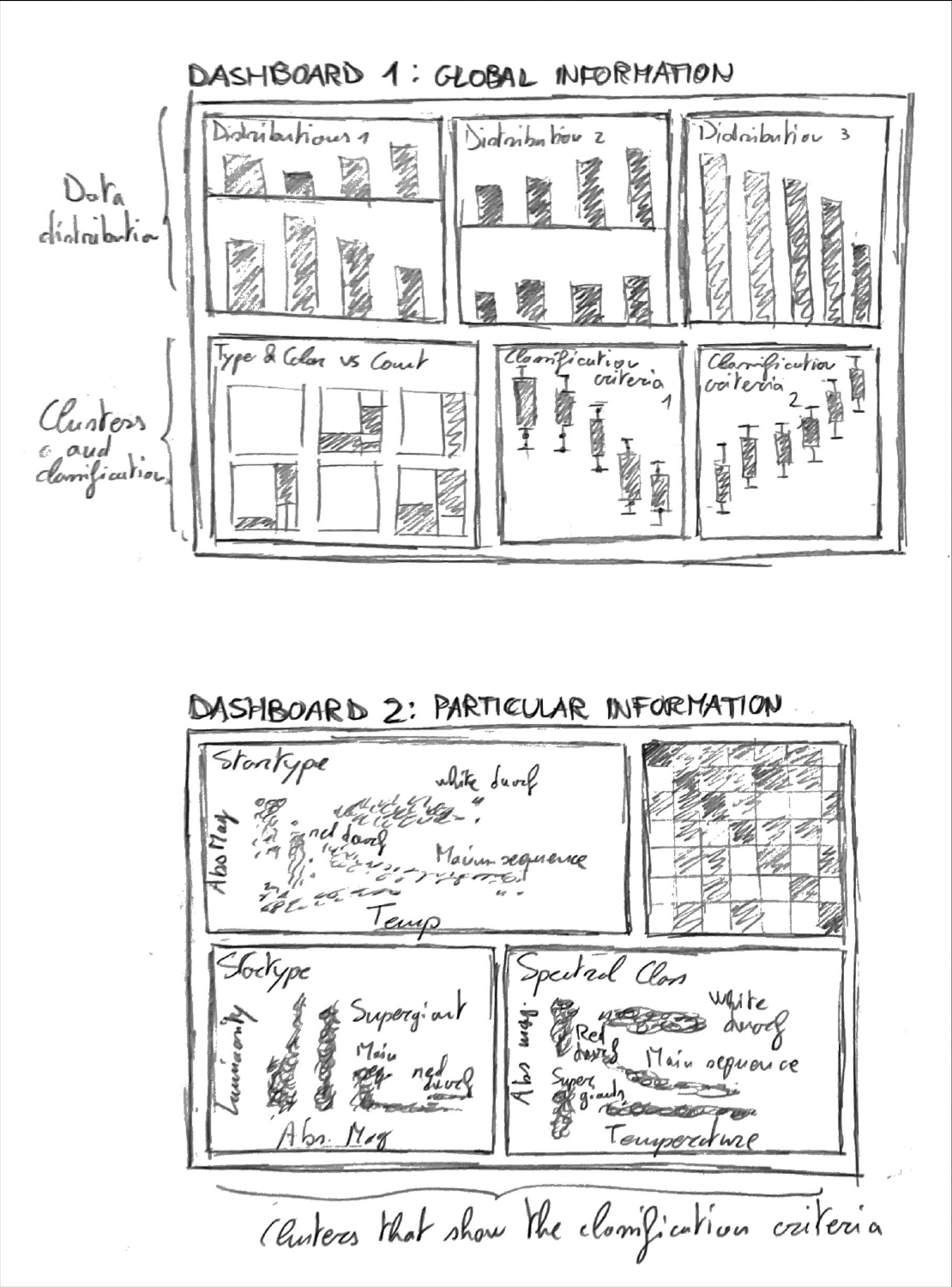
This project would be very enriching as a side information source. Some of the charts could even be used to complement technical explanations.

We are not experts in Astronomy, so we obviously could not post a scientific article on the results obtained in this visualization project. However, our results will be available on this pdf, as well as on a GitHub account, where any interested audience could consult it.

Astronomy is a very interesting field of research, which most people do not know much about but find it fascinating. This project helps to understand more about the subject without getting too technical.

# 9. Idea of the final result

Here it follows a “napkin” design to show the idea of what the final result will be like.



This is structured in order to satisfy the requirements given by the scope of the project, which was just stated, allowing for an easy understanding of the visualization methods used. The first dashboard will convey information regarding the dataset itself, with the distribution of the data and how it is divided within the different classes of the classifications. The second dashboard will make use of more structured visualization methods in order to present the different clusters, criteria of classification, and patterns that arise from this type of classification methods.

# 10. Charts selection and encoding

## 10.1 Chart selection

In order to reach the objective of this project, which was previously stated, the following choices were made.

* **Dashboard 1:** The first dashboard is going to focus on presenting information regarding the main quantitative variables, such as Radius, Luminosity, Temperature. It will do so in order to allow the audience to explore the relationship of the measure variables and the classifications, Star type, Spectral class and Star color.
* **Dashboard 2:** the second dashboard is going to be a more detailed and technical dashboard, focusing on more relevant quantitative variables such as Temperature and Absolute Magnitude. This will allow the audience to gather specific information and values regarding particular subsets of samples that may be of interest.

Given the nature of the data, the following charts were chosen as the most suitable to present and visualize the data.

* **Bar charts:** one of the more ‘natural’ choices given the nature of our data, these allow for a clear and unbiased visualization of quantitative variables measures, such as the average, over given groupings, thus allowing to divide the chart by the classes in the dataset.
* **Tree Charts:** this is well suited for understanding the subsets inside a determined class. In particular, this is well-suited in order to show, for every Star type, which classes of Star color are present, and how many samples for each class.
* **Scatterplots:** these are used in order to visualize where each sample is positioned in terms of 2 different quantitative values. For this reason this is well-suited in order to show a map of the samples, which allows to understand how the quantitative variables induce a determined classification
* **Packed bubbles:** this chart can normally give a biased understanding of the data. However, in this case, we were able to compute the value of the area of the section of the star for each sample. Therefore, this allows us to visualize as areas of the bubbles the areas of the sections of the stars, thus giving a more robust and less biased understanding of the data.

## 10.2 Encoding and perception

In this project a lot of encoding was taken in consideration in order to make the class-rich dataset more easily understandable. When presenting the data we found that the large number of classes made retaining the information more difficult, hence we decided to try to find ways to aid the retention of the groups and classes used.

* **Color econding:** we used consistent color scales for the three categorical features, Star type, Spectral Class and Star color, as well as color gradients for the quantitative features. However, in order not to work against the aim of this encoding and confuse the user, we had to select some main color encodings on the dashboards. The general rule we used is to choose the encoding that brought the most value in terms of perception of what we considered more important. For example, in the first dashboard, where the aim was to see the importance of the quantitative values, we used color gradients on the bar charts to encode the magnitude of the values. On the other hand, in the second dashboard, where the aim is more technical and focuses on the Star type, the color encoding is consistent, and it highlights the group division based on the Star type.
* **Shape encoding:** we took into consideration using shape encoding, mainly in the scatterplots in dashboard two. We considered it in order to also group the data by Spectral class. However, the visualization resulted in being too crowded and noisy in order for the user to take advantage of this additional grouping. For this reason we chose not to include it in the visualization.

# 11. Perception and best practices

## 11.1 Gestalt principles and application

As we were asked to apply the gestalt principles for data presentation and visualization, we studied its main points and how to apply them to our data visualization.

Gestalt principles are a list of best practices to be used when presenting data. It has been demonstrated by psychologists that, depending on how we visualize chartical outputs, our mind is able to collect the information more or less effectively.

We will start with the closure element:

* **Closure:** It is a psychological fact that human minds prefer to see complete shapes, where the figure is completely close. For example, we can use colors to create non complete figures or elements whose shape will be completed by these colors.
* **Common region and proximity:** When grouping similar elements, with similar characteristics, visualizing them together, and separated from other groups, stimulate the viewers understanding the differences among these groups. We will use this principle a lot when differentiating star types.
* **Figure/Ground:** When analyzing chartical results, it is important to see contrast between the important information and the background. Therefore, it will be needed to bring the interesting information to the front of the visualization while we leave the complementary information in the background with lighter colors which don’t take our focus from the real interest. We will apply this principle in all of our visualizations.

When creating our dashboard, we took into account the Gestalt principles in all of the charts.

For example, in the scatterplots, we group data by similarity to make it more visually comprehensible and easy to understand for all kinds of audience.

We also plotted bar charts, which are very easy to understand, as they work with single magnitudes and help understanding differences among groups.

It is also important to comment that we worked with colors that can be associated with the real values, and that at the same time can be used to understand the magnitudes. For example, the star's colors go from cold to warm colors as the temperature increases.

# 12. Interaction

## 12.1 Filtering

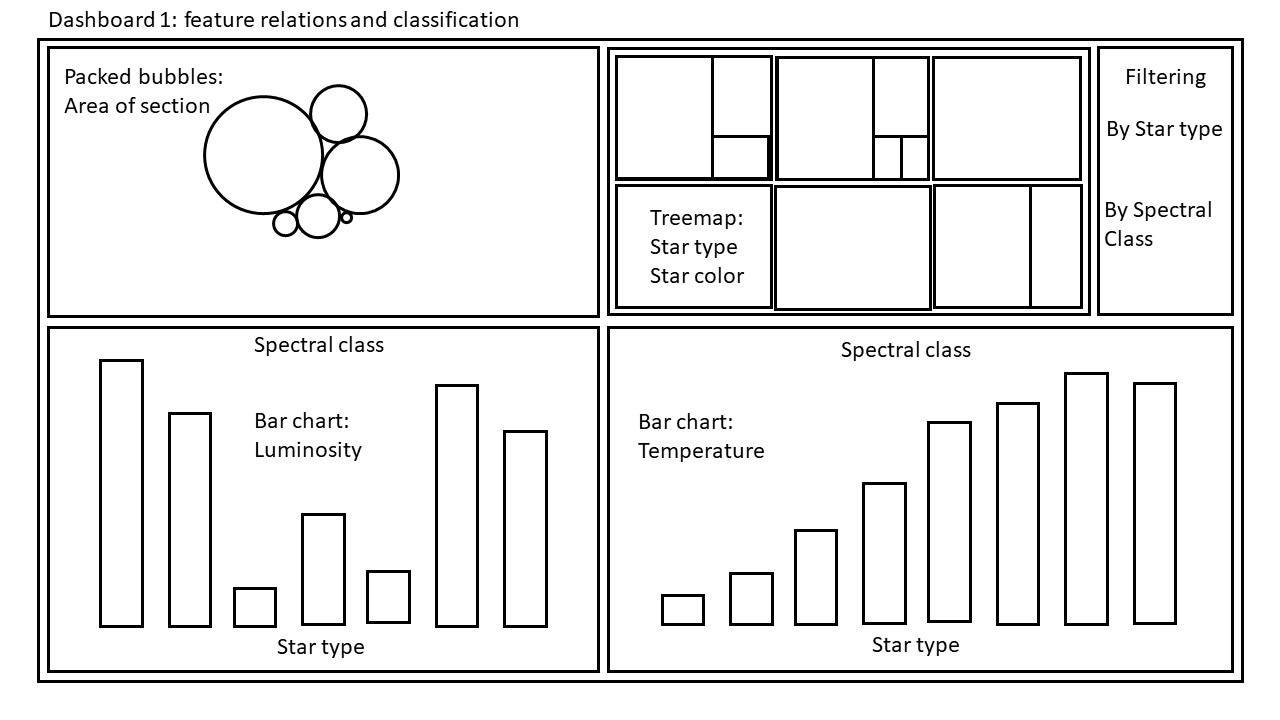
For this dataset, the ability to filter the data visualized is crucial. The main issue in visualizing the quantitative data is the extremely large difference in magnitude of the values. Because of this, a lot of samples cannot be presented and compared with others when this difference is too large, thus resulting in not comparable. We considered the use of logarithmic scales, but this is not a good way of understanding magnitudes of the values and it gives a lot of space for cognitive biases. With filtering this can be easily solved; allowing for the presented samples to change and restricting the data to subsets of interest, the axis can be rescaled in order to compare smaller values of the quantitative variables.

Considering this, we made sure that in both dashboards the user is able to filter and group for any of the three classes present in the dataset.

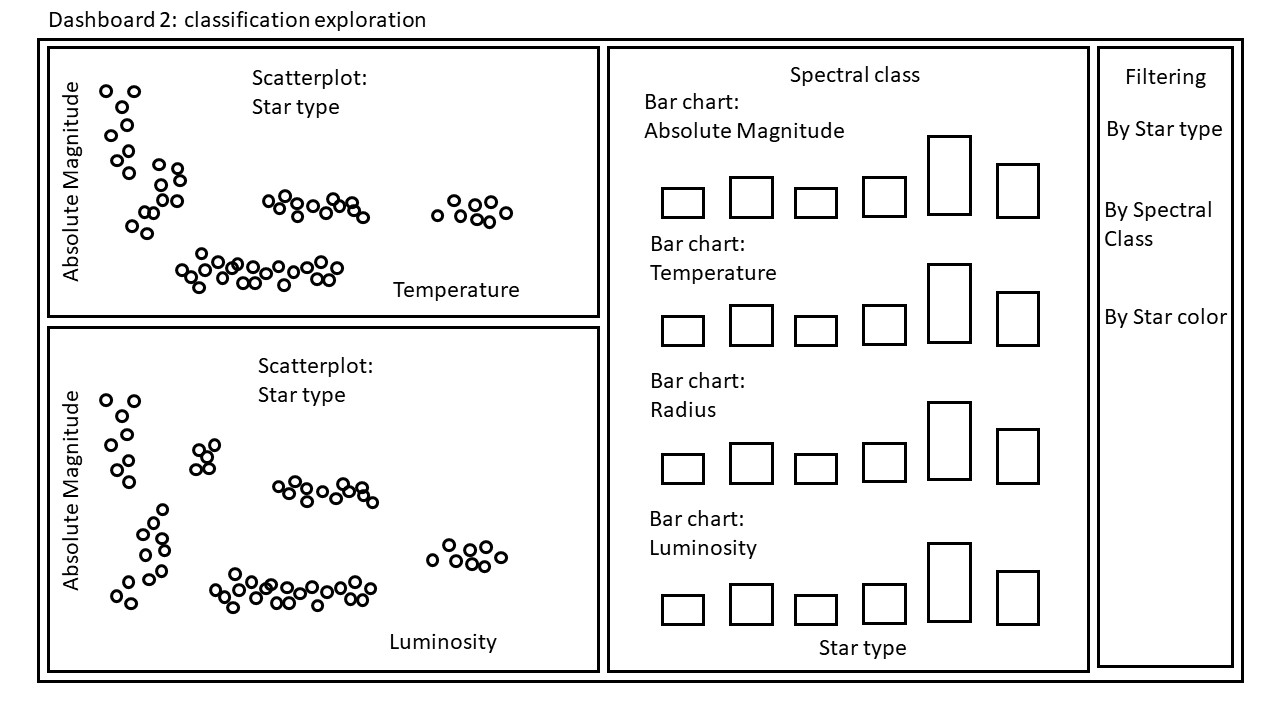
# 13. Layout and content of the dashboards

We decided to make some modifications to the previous design of the two dashboards, in order to make them more interesting and bring more value to the user with the data presented. Here it follows a second napkin design of the two dashboards.

**DASHBOARD 1**



**DASHBOARD 2**

****

# 14. Implementation of the Dashboards

## 14.1 Design of the Dashboards

**DASHBOARD 1**

In this dashboard, we present most of the information regarding the quantitative values, grouped using the classes present in the dataset. This is done in order to allow our audience to understand which quantitative values are most relevant when classifying the samples.

Starting from the top left chart, we use packed bubbles in order to show the dimensions of each star. We do this using the computed field of Area of section, this way the bias that is often present with this kind of charts is heavily mitigated. Groups are easily differentiated because of the colors; in this case the color is encoding the Star color.

The treemap shows for every type of star, the absolute frequency in which every star color appears, while grouping by Star color.

On the lower part of the dashboard we included two bar charts with two important quantitative features: Luminosity and Temperature. These two charts are pretty elementary and easy to understand, as it is one of the purposes of this first dashboard. They are grouped by both Star type and Spectral class, on two different axes. Also, we used a scale of colors that represents with a gradient the presented quantitative features, respectively, Luminosity and Temperature. This color encoding considerably helps with understanding the data, since it allows for an understanding of the magnitude of the value, not only by the height of the bar, but also with a color gradient.

**DASHBOARD 2**

In this dashboard, we have all the information needed when considering the classification of stars based on Star type and Spectral class.

Starting from the top left chart, we can see how every star type behaves when plotting Absolute Magnitude vs Temperature. Groups are easily differentiated because of the colors and their position. So in this case the color is encoding the Star type in order to make the most important grouping criteria more evident. Also, encoding the different classes with colors helps remembering the results relative to one group more easily. This is an example of the common region and proximity principles from Gestalt.

All this procedures make the chart much easier to understand.

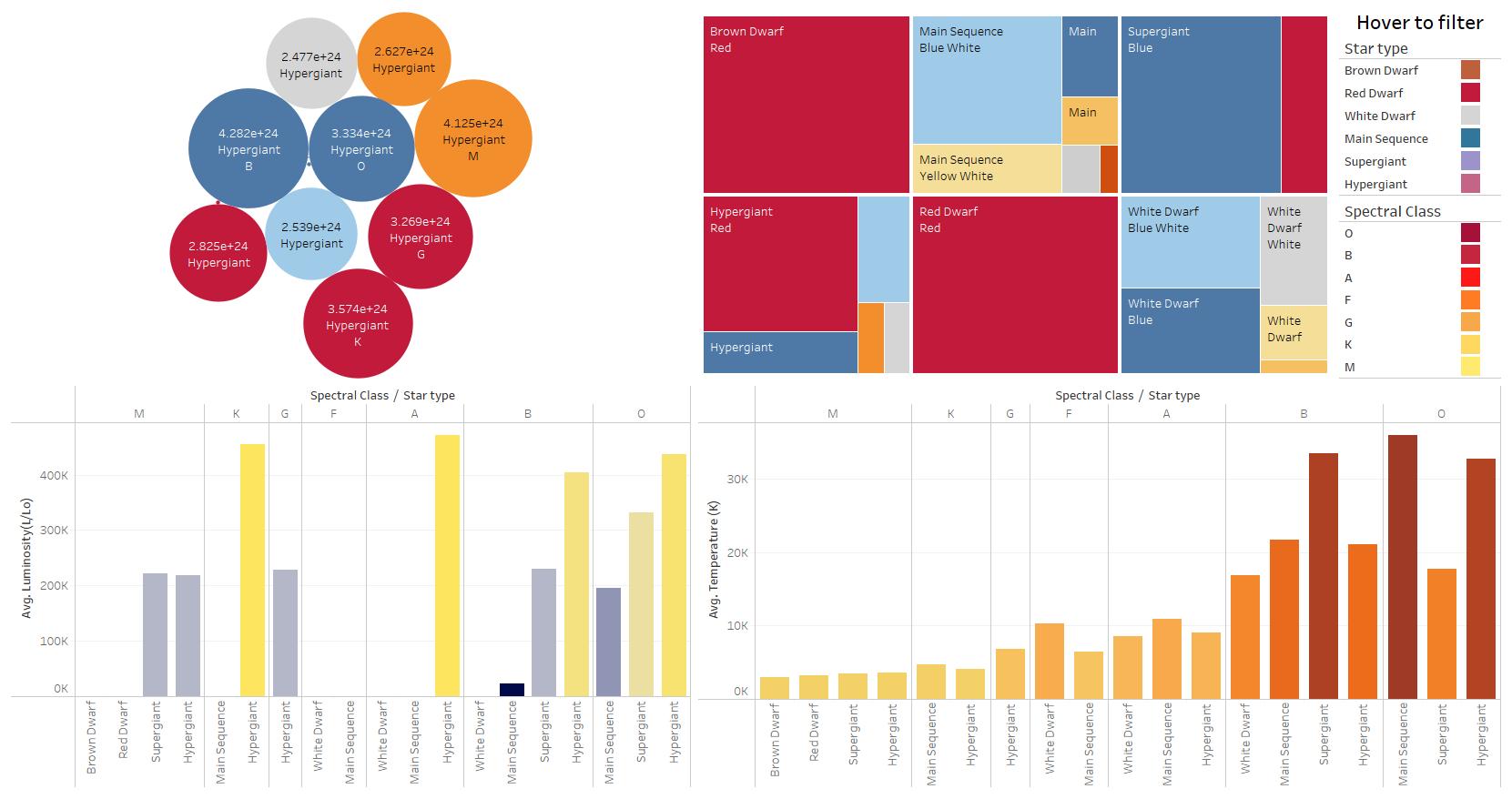
The left bottom scatterplot shows the relation every star type has when plotting Absolute Magnitude vs Luminosity. This chart is built following the same criteria used for the frist one.

Lastly, the right part of the dashboard is dedicated to a bar plot with the information regarding Average Absolute Magnitude, Average Temperature, Average Radius, and Average Luminosity. We used the average to compare magnitudes because as we have many stars in each group, it was the statistical parameter which made more sense (we could have also used the median, but as we did not have outliers, mean was a better choice). This becomes very useful when filtering the data, since it adapts to show the values for any selected group of samples.

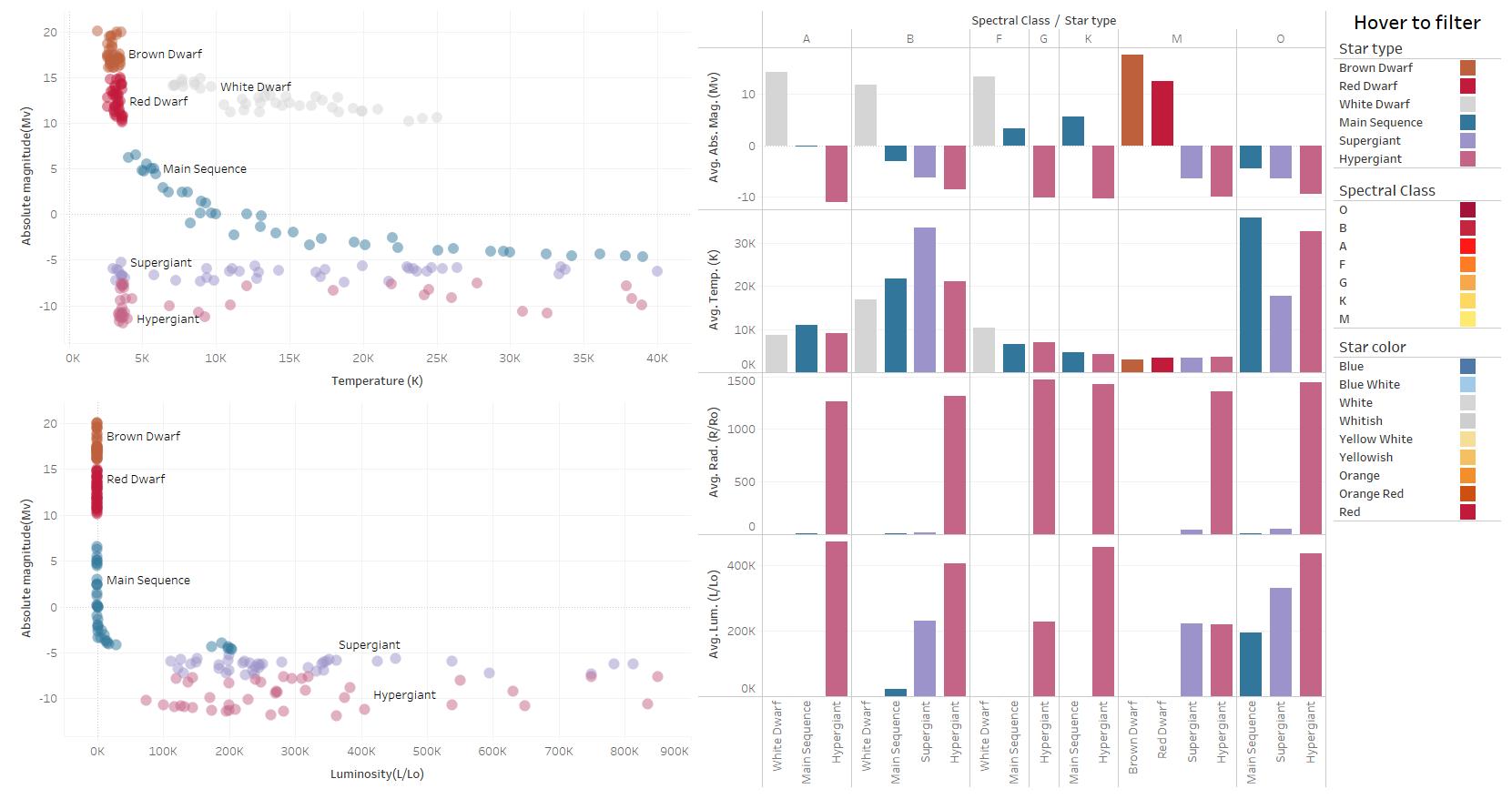
## 14.2 First implementation of the Dashboards

Here two screenshots of the first implementation of the two Dashboards following. The source Tableau Workbook is submitted as well with this document, but we insert the images here as well for better consultation.

**DASHBOARD 2**



**DASHBOARD 2**



## 14.3 Further implementations

There are some further implementations that we would like to apply to the dashboards, but we were not able to due to some technicalities. In this section we present some of them in the hope of finding a suitable way to implement them as well.

* **Chart description:** we realized some charts need some sort of description or title in order to be more easily understandable. Unfortunately, we couldn’t find an aesthetically appealing way of presenting it, thus we decided not to include it for the moment. Differently from the following two features, we realize this is much more necessary, so, for this reason, it will be prioritized.
* **Key values:** we would like to visualize some key values for the quantitative features that are responsive to the filtering choices of the user
* **Filtering by quantitative values:** for the moment our dashboard allows for filtering by the different categorical features. We would like to implement the option to filter the data by the quantitative features as well, presumably with some sliders that allow cumulatively to include more samples following criterias on the quantitative values.