LDA and Results

Logan Crandall

2024-04-09

An Analysis and Statistical Prediction of Star Color

Data History

This data was compiled by Devendra Singh Shekhawat on Kaggle. The data was sourced from Spartificial, a research institute in India. The data was cleaned by the original compiler, and required no further cleaning. This data includes observational variables based upon 240 stars in the known universe.

Summary

In this paper, I examine the research question: can we reliably predict star color from other observational variables. The other observational variables used in this study are Temperature, Luminosity, Radius, Absolute Magnitude, and Star Type. These variables are defined below. By the conclusion of this article, I was able to finalize a linear discriminant analysis model that predicted star color from the aforementioned variables with an accuracy of 88.7%.

A Quick Overall Look at the Data

Here, I have added a short overview of the variables for reference. They will be further examined below. Variables:

- Temperature (K) - Temperature of the star in Kelvin
- Luminosity(L/Lo) - Relative luminosity of the star to our sun, with our sun being equal to one Solar Luminosity
- Radius (R/Ro) - Relative Radius of the star, with our star being equal to one Solar Radius
- Absolute magnitude (Mv) - The magnitude of a star if it were viewed from a constant distance (10 Parsecs)
- Star type- numerical variable representing star classes (0=Brown, 1=Red Dwarf, 2=White Dwarf, 3=Main Sequence, 4= Supergiants, 5=HyperGiants)
- Star color - variable for the color of the star
- Spectral Class - Variable for the Spectral Class of the star, delineated by letters (O, B, A, F, G, K, M). These are based on the absorption lines of a star.

Table 1: Data summary

Name	stardata
Number of rows	240
Number of columns	7
Column type frequency:	
character	3
numeric	4

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Star type	0	1	1	1	0	6	0
Star color	0	1	3	12	0	5	0
Spectral Class	0	1	1	1	0	7	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Temperature (K)	0	1	10497.5	9552.4	1939.0	3344.2	5776.0	15055.5	40000.0	
Luminosity(L/Lo)	0	1	107188.4	179432.2	0.0	0.0	0.1	198050.0	849420.0	
Radius(R/Ro)	0	1	237.2	517.2	0.0	0.1	0.8	42.8	1948.5	
Absolute magnitude(Mv)	0	1	4.4	10.5	-11.9	-6.2	8.3	13.7	20.1	

Statistical Overview

Quantitative Variables

Table 4: Summary Statistics for Quantitative Star Data

Temperature (K)	Luminosity(L/Lo)	Radius(R/Ro)	Absolute magnitude(Mv)
Min. : 1939	Min. : 0.0	Min. : 0.0084	Min. :-11.920
1st Qu.: 3344	1st Qu.: 0.0	1st Qu.: 0.1027	1st Qu.: -6.232
Median : 5776	Median : 0.1	Median : 0.7625	Median : 8.313
Mean :10497	Mean :107188.4	Mean : 237.1578	Mean : 4.382
3rd Qu.:15056	3rd Qu.:198050.0	3rd Qu.: 42.7500	3rd Qu.: 13.697
Max. :40000	Max. :849420.0	Max. :1948.5000	Max. : 20.060

Table 5: Data summary

Name	quant
Number of rows	240
Number of columns	4
Column type frequency:	
numeric	4
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Temperature (K)	0	1	10497.5	9552.4	1939.0	3344.2	5776.0	15055.5	40000.0	
Luminosity(L/Lo)	0	1	107188.4	179432.2	0.0	0.0	0.1	198050.0	849420.0	
Radius(R/Ro)	0	1	237.2	517.2	0.0	0.1	0.8	42.8	1948.5	
Absolute magnitude(Mv)	0	1	4.4	10.5	-11.9	-6.2	8.3	13.7	20.1	

Here, you can see some summary statistics for the quantitative variables in my data set. A few things are worth noting upon first inspection. Some of the variables, such as temperature, vary widely (from 1939K to 40000K). This is to be expected, as stars in the universe vary widely. These large variations are a good example of why categories of stars have to be created, and the properties of different kinds of stars have to be analyzed to place them. Furthermore, the minimum Luminosity appears to be zero; these points are actually very, very close to zero. This does not mean that they emit no energy at all. Rather, it is relative luminosity, with our sun as the benchmark at one. So, a star with luminosity close to zero would emit fractions of the light that our sun emits. Descriptive statistics can be found in the tables above, and there is no missing data.

Categorical Variables

Table 7: Proportion of Star Types

Star Type	FREQ
0	16.67
1	16.67
2	16.67
3	16.67
4	16.67
5	16.67

Table 8: Proportion of Star Colors

Star Color	FREQ
Blue	23.33
Blue-White	17.08
Red	48.33
White	5.00
Yellow-White	6.25

Table 9: Proportion of Spectral Classes

Spectral Class	FREQ
A	7.92
B	19.17
F	7.08
G	0.42
K	2.50
M	46.25

Spectral Class	FREQ
O	16.67

Table 10: Data summary

Name	cat
Number of rows	240
Number of columns	3
Column type frequency:	
character	3
Group variables	None

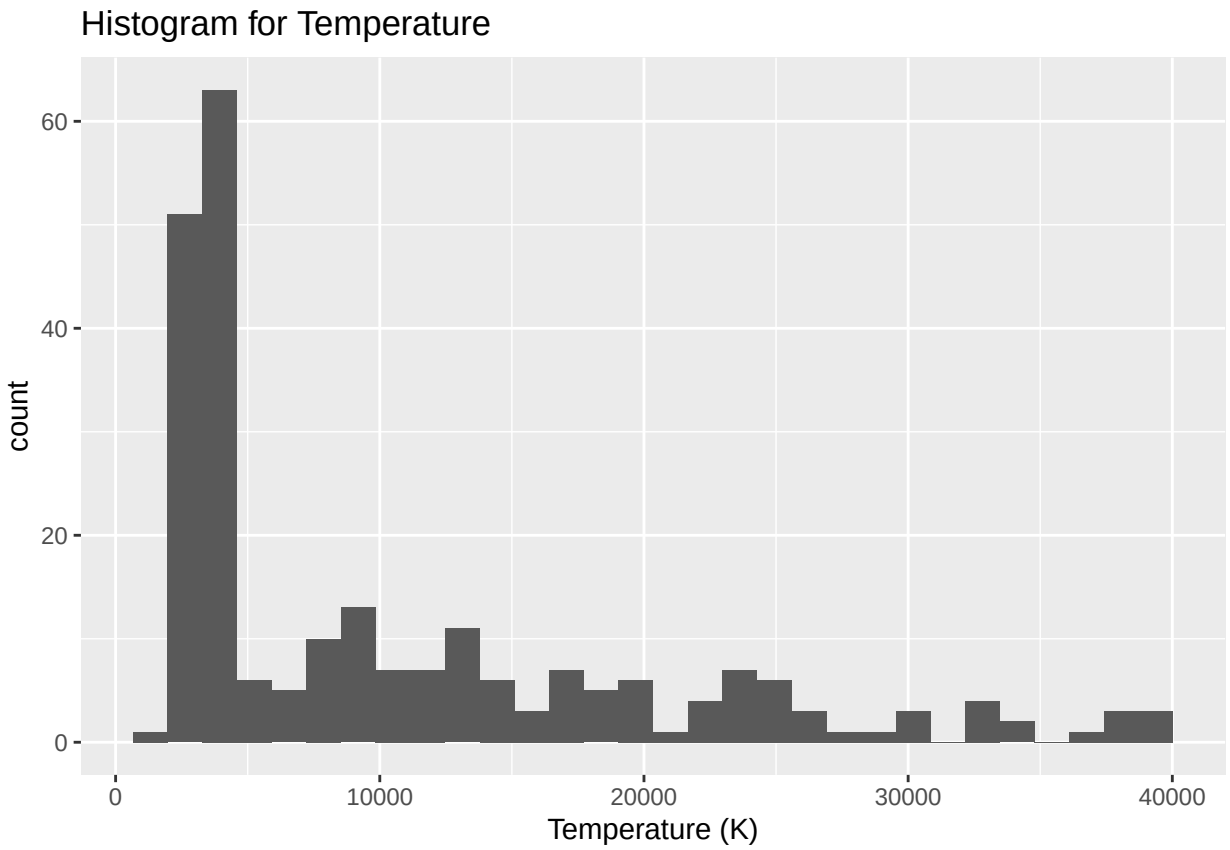
Variable type: character

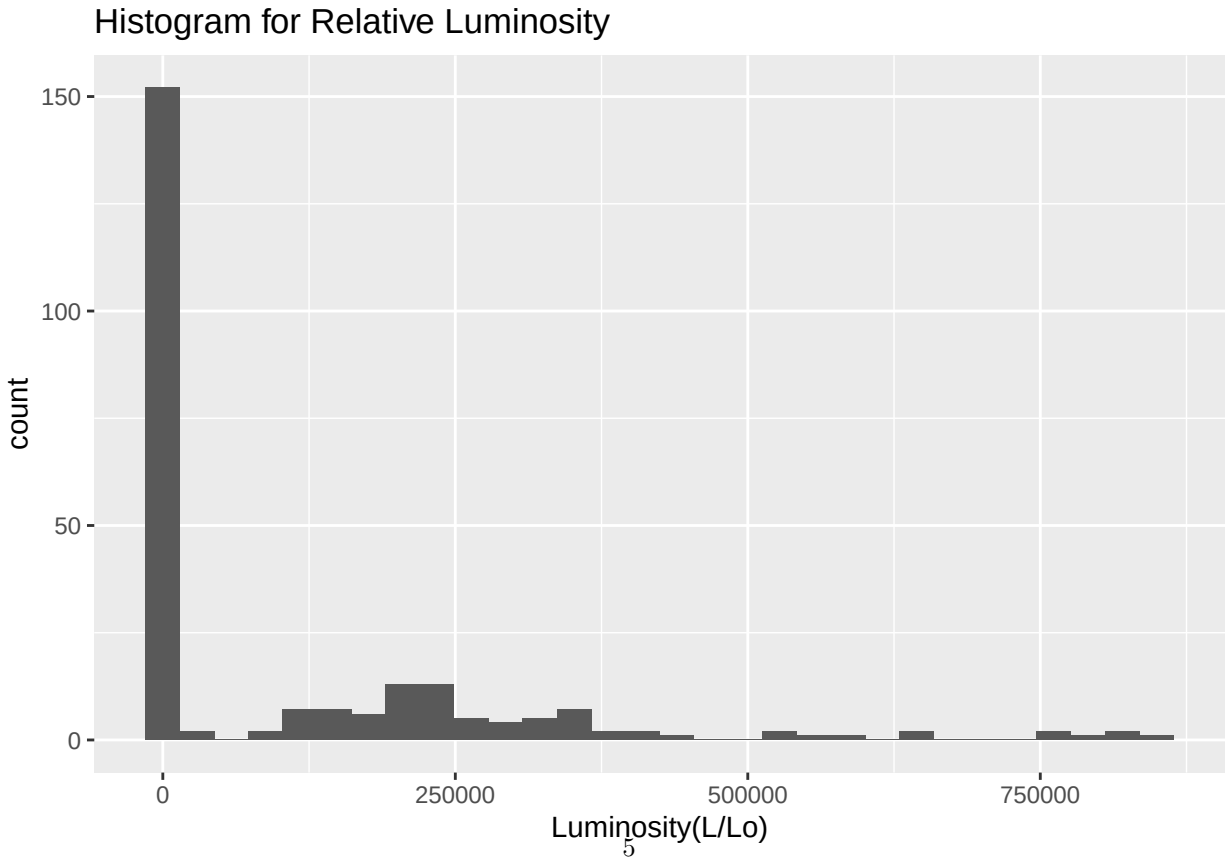
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Star type	0	1	1	1	0	6	0
Star color	0	1	3	12	0	5	0
Spectral Class	0	1	1	1	0	7	0

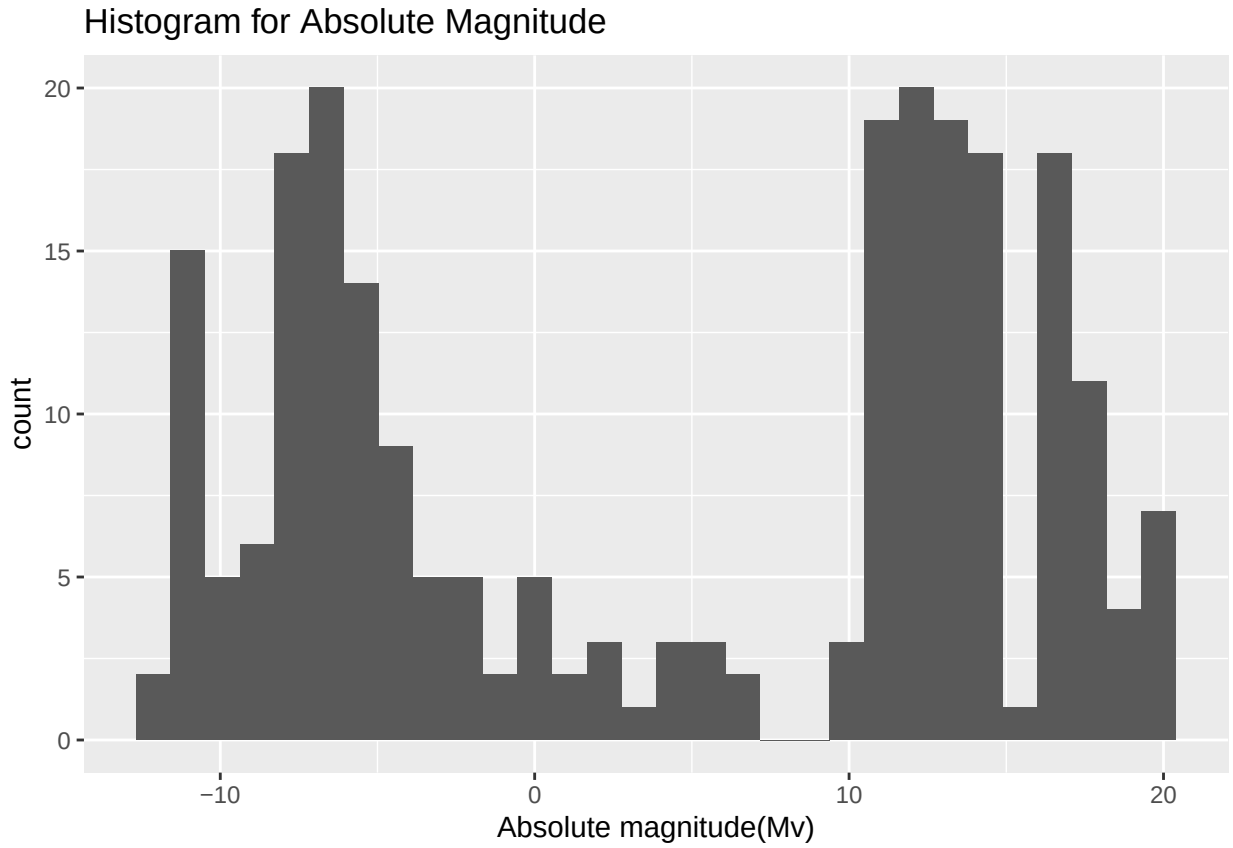
There are five star colors, seven spectral classes, and 6 star types in this data set, and there is no missing data on the stars. Star Types are distributed completely evenly. In terms of star colors, red makes up almost half of the recorded stars, followed by fairly large chunks of blue and blue-white stars. White and yellow-white stars, the remaining two categories, only make up about 11% of the data. Spectral class M makes up almost half the recorded stars. Conversely, the smallest class, G, makes up only 0.42% of the data.

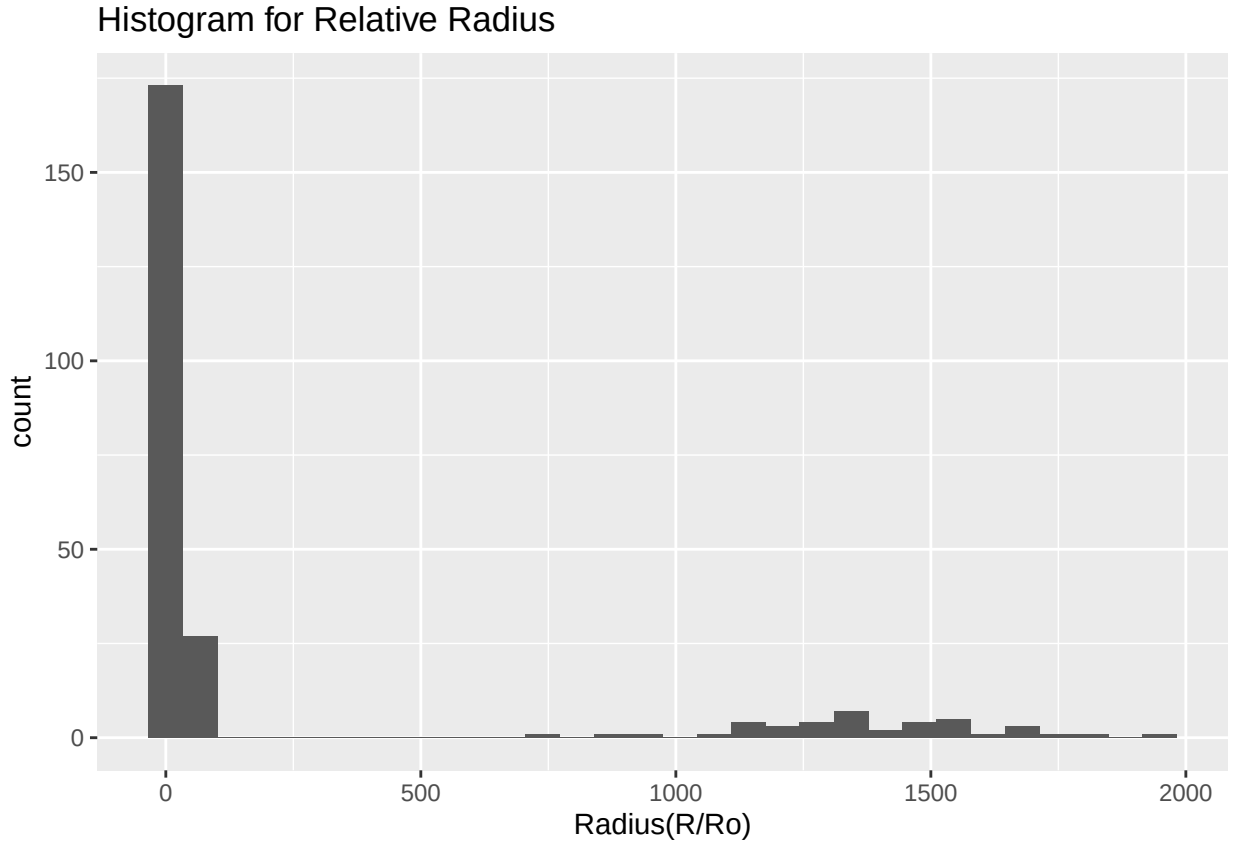
Graphical Examination

Quantitative variables





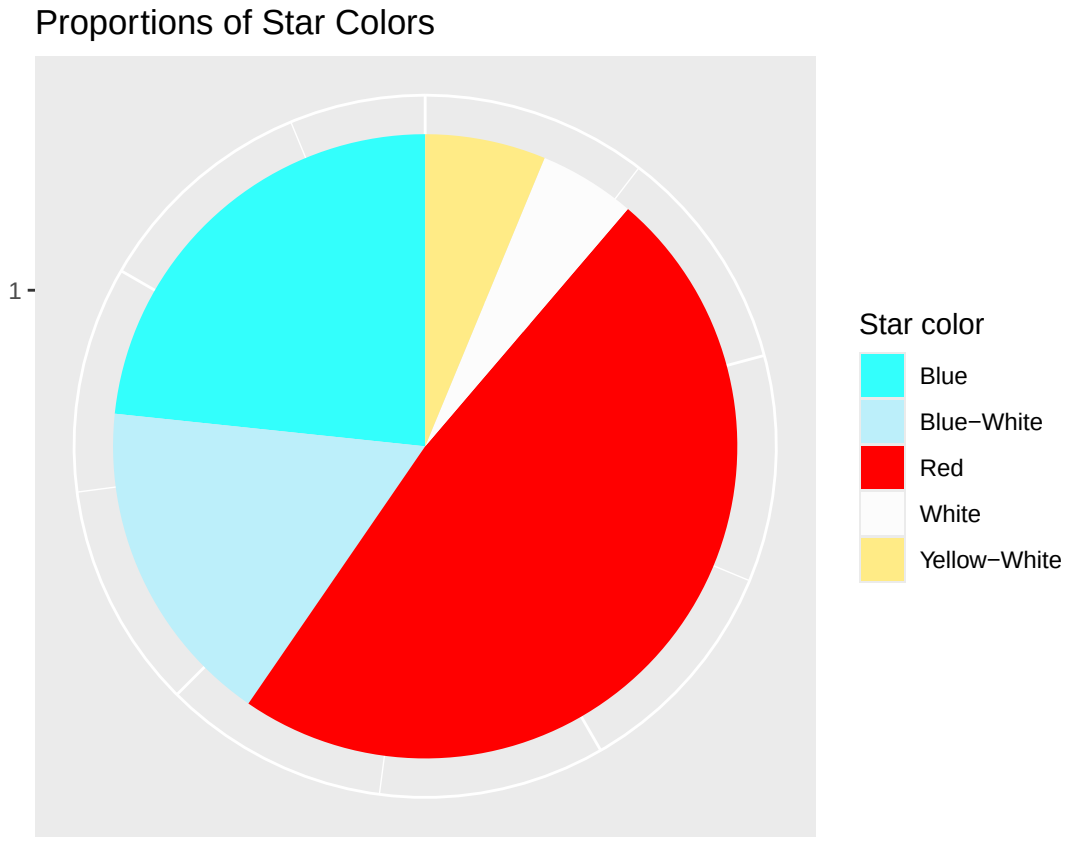


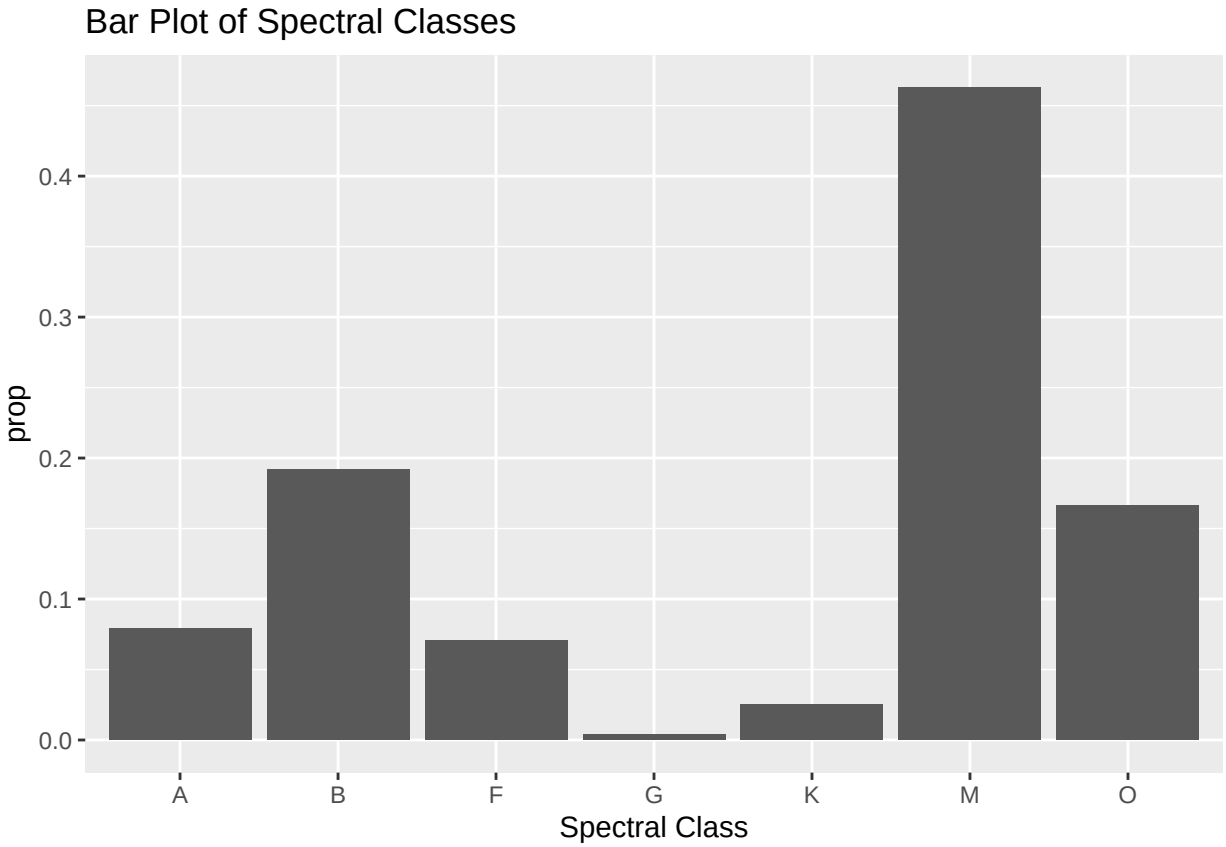


Above are histograms for each of the numeric variables. It appears that the histograms for both of the

relative variables (Radius, and Luminosity) are heavily right-skewed. This is to be expected, as the measurements are relative to our sun, and expected to hover around one. Temperature is also heavily right skewed, from some very high temperature outliers, while the bulk of the data lies closer to our sun (which is around 5,700K). Interestingly enough, the data for absolute magnitude is not skewed, but it does follow a bi-modal distribution. I hope that the reason becomes apparent through analysis of the data.

Categorical Variables



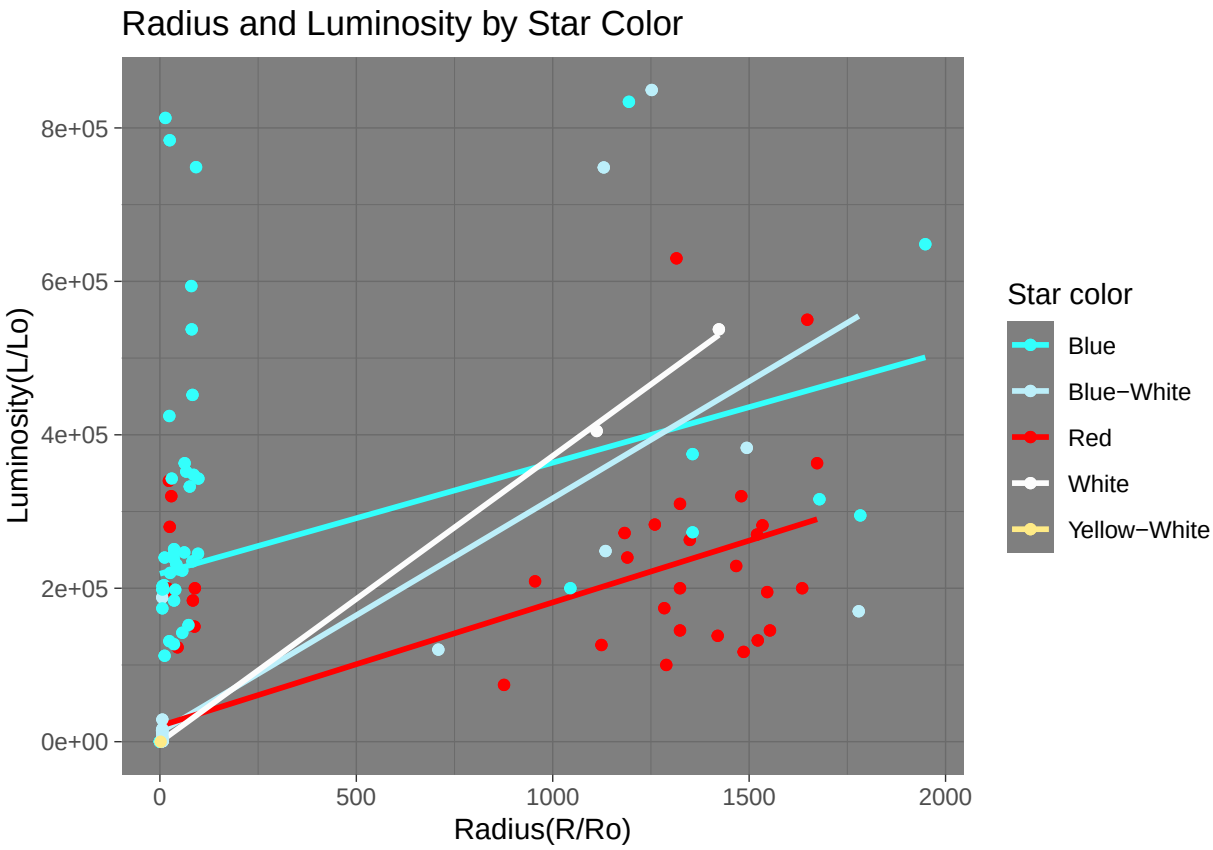


Again we are able to confirm the spreads of spectral class and star type to show that the largest groups, by far, are spectral class M and color Red.

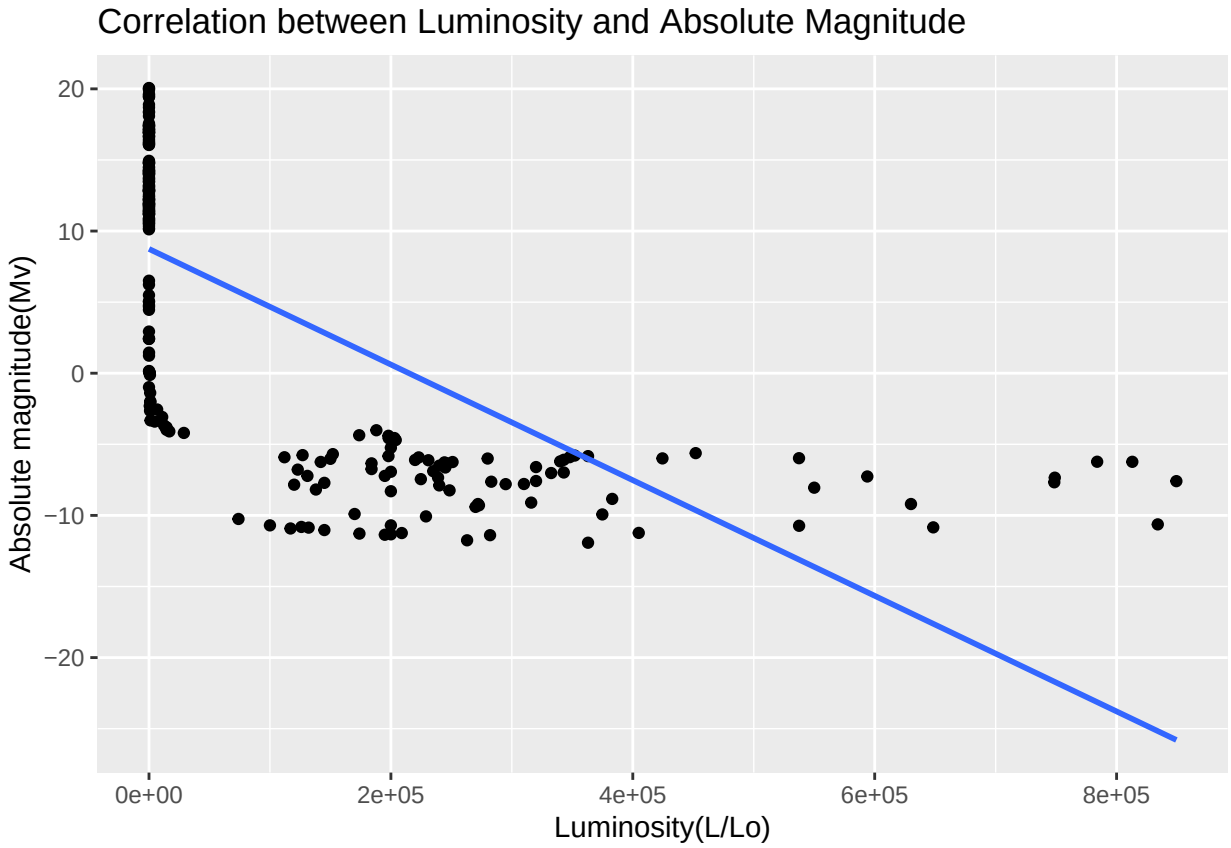
Multivariate Analysis

Table 12: Pearson Correlations for Quantitative Star Variables

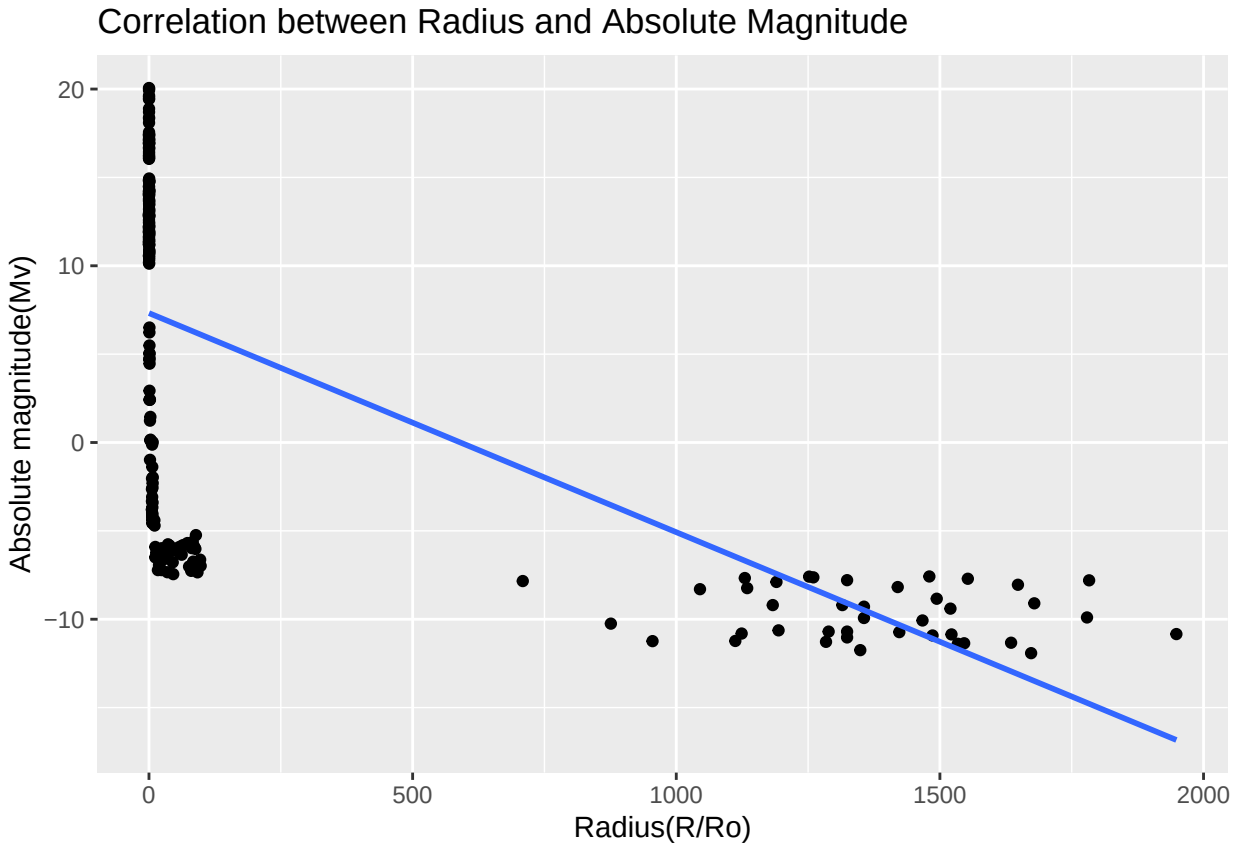
	Temperature (K)	Luminosity(L/Lo)	Radius(R/Ro)	Absolute magnitude(Mv)
Temperature (K)	1.00	0.39	0.06	-0.42
Luminosity(L/Lo)	0.39	1.00	0.53	-0.69
Radius(R/Ro)	0.06	0.53	1.00	-0.61
Absolute magnitude(Mv)	-0.42	-0.69	-0.61	1.00



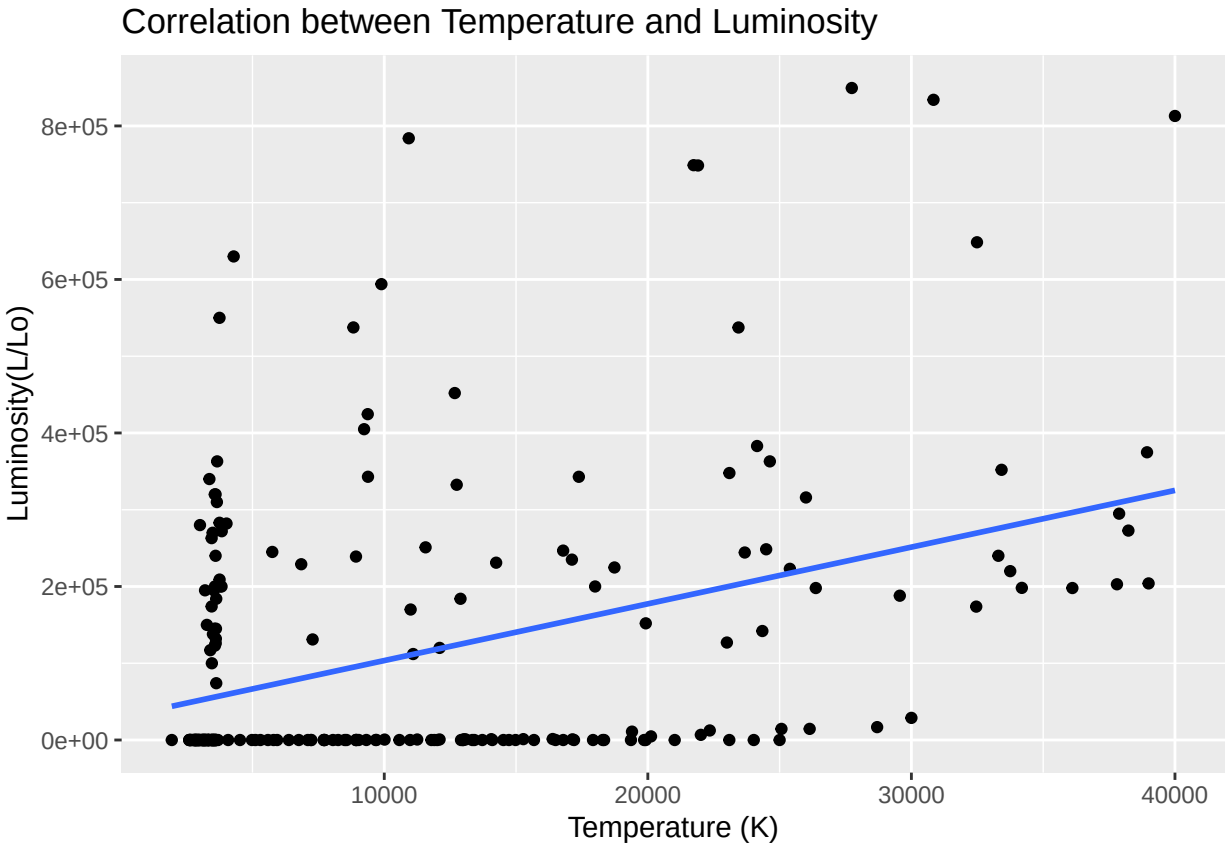

```
##
##  Pearson's product-moment correlation
##
## data:  stardata$`Luminosity(L/Lo)` and stardata$`Radius(R/Ro)`
## t = 9.5542, df = 238, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4284518 0.6123206
## sample estimates:
##       cor
## 0.5265157
```



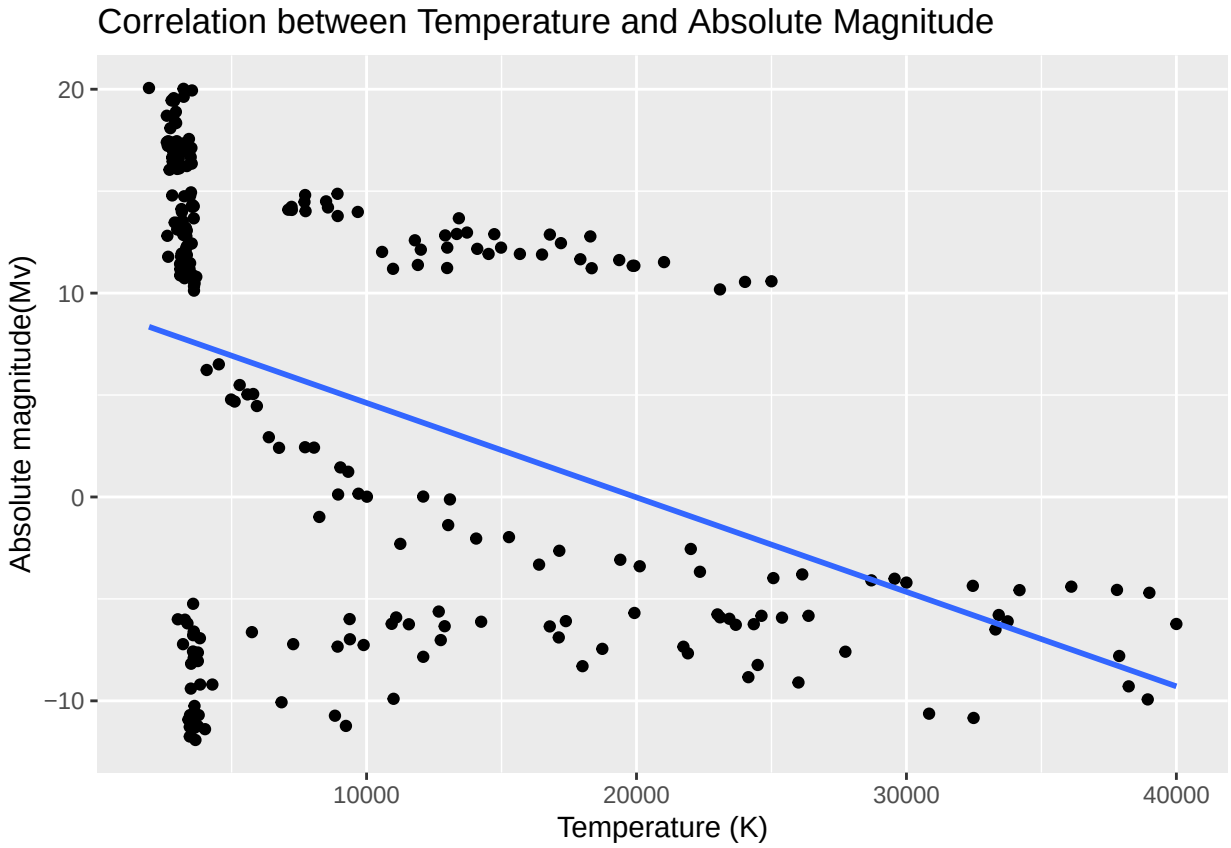

```
##
##  Pearson's product-moment correlation
##
## data:  stardata$`Luminosity(L/Lo)` and stardata$`Absolute magnitude(Mv)`
## t = -14.814, df = 238, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7531897 -0.6204025
## sample estimates:
##        cor
## -0.6926192
```




```
##
##  Pearson's product-moment correlation
##
## data:  stardata$`Radius(R/Ro)` and stardata$`Absolute magnitude(Mv)`
## t = -11.837, df = 238, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6827310 -0.5223638
## sample estimates:
##        cor
## -0.6087282
```




```
##
##  Pearson's product-moment correlation
##
## data:  stardata$`Temperature (K)` and stardata$`Luminosity(L/Lo)`
## t = 6.6014, df = 238, p-value = 2.625e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2807607 0.4953569
## sample estimates:
##       cor
## 0.3934041
```




```
##
##  Pearson's product-moment correlation
##
## data:  stardata$`Temperature (K)` and stardata$`Absolute magnitude(Mv)`
## t = -7.1451, df = 238, p-value = 1.092e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5192569 -0.3101353
## sample estimates:
##        cor
## -0.4202605
```

Above are correlations for the numeric variables, and graphs showing those correlations. There seems to be significant correlation between many of the variables. Pairs which contain significant correlation include: Radius and Luminosity, Luminosity and Absolute Magnitude, Radius and Absolute Magnitude, Temperature and Luminosity, and Temperature and Absolute Magnitude. It would be interesting to see if you could predict some of these variables based on other variables. For example, it might be interesting to attempt to predict Temperature based upon Absolute Magnitude.

Table 13: Proportion of Star Type by Star Color

	Blue	Blue-White	Red	White	Yellow-White
0	0.0	0.0	34.5	0.0	0.0
1	0.0	0.0	34.5	0.0	0.0
2	23.2	34.1	0.9	66.7	26.7
3	8.9	51.2	0.9	16.7	73.3
4	55.4	0.0	7.8	0.0	0.0

	Blue	Blue-White	Red	White	Yellow-White
5	12.5	14.6	21.6	16.7	0.0

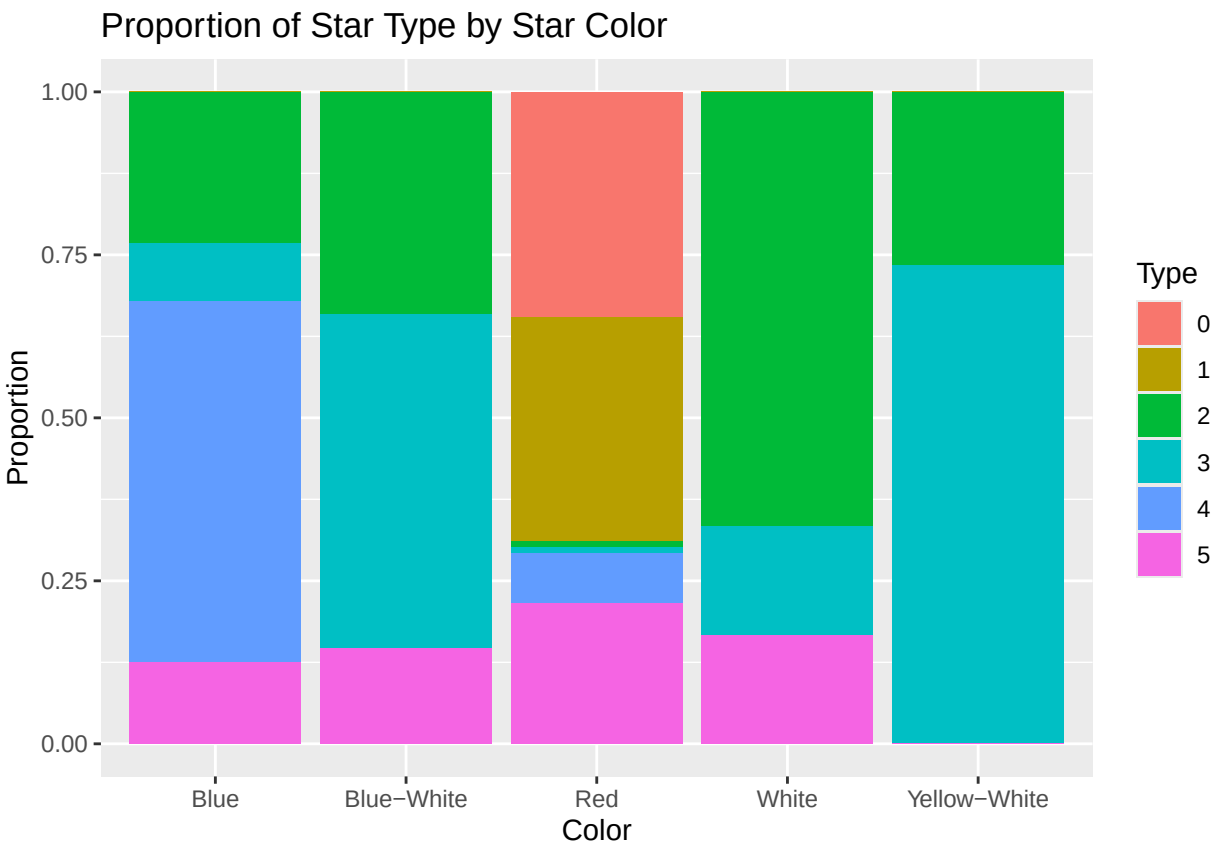


Table 14: Proportion of Star Types by Spectral Class

	A	B	F	G	K	M	O
0	0.0	0.0	0.0	0	0.0	36.0	0.0
1	0.0	0.0	0.0	0	0.0	36.0	0.0
2	36.8	52.2	52.9	0	0.0	0.0	0.0
3	52.6	28.3	47.1	0	66.7	0.0	12.5
4	0.0	4.3	0.0	0	0.0	8.1	72.5
5	10.5	15.2	0.0	100	33.3	19.8	15.0

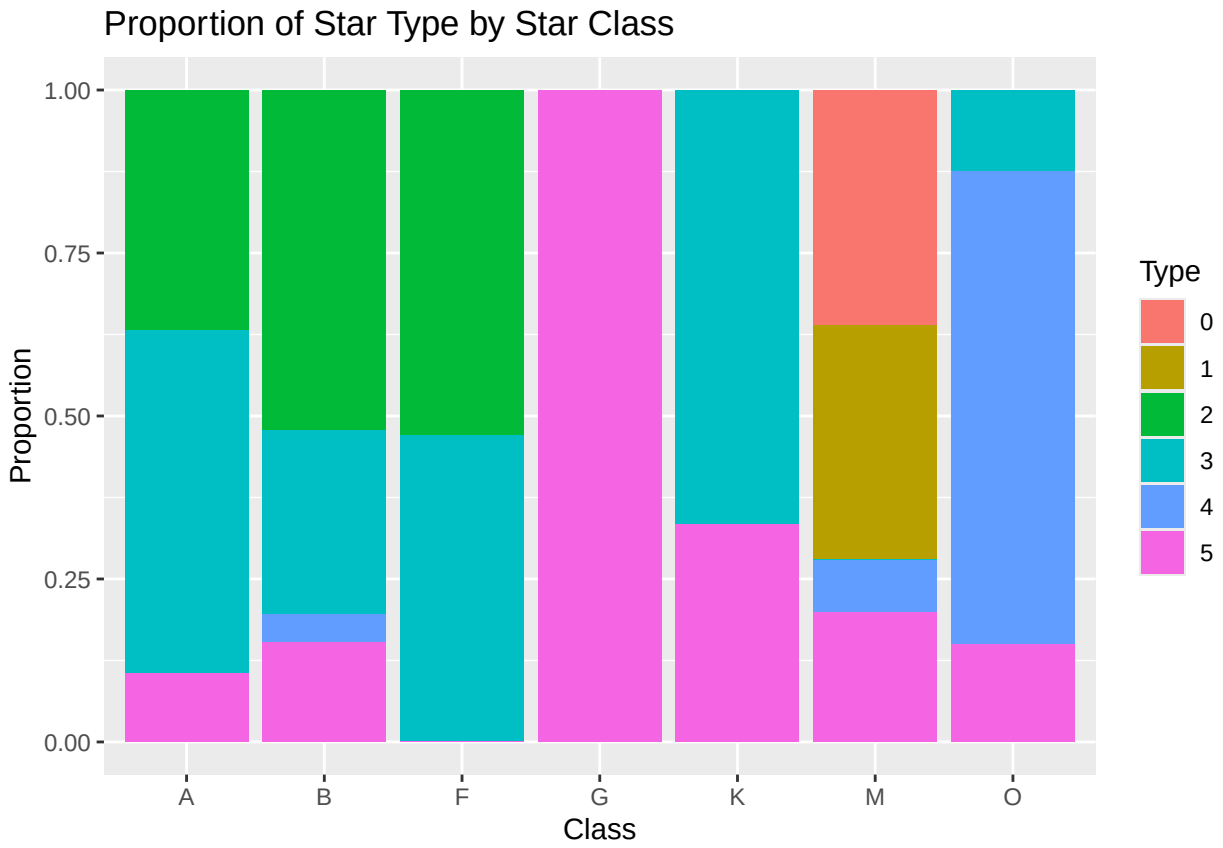
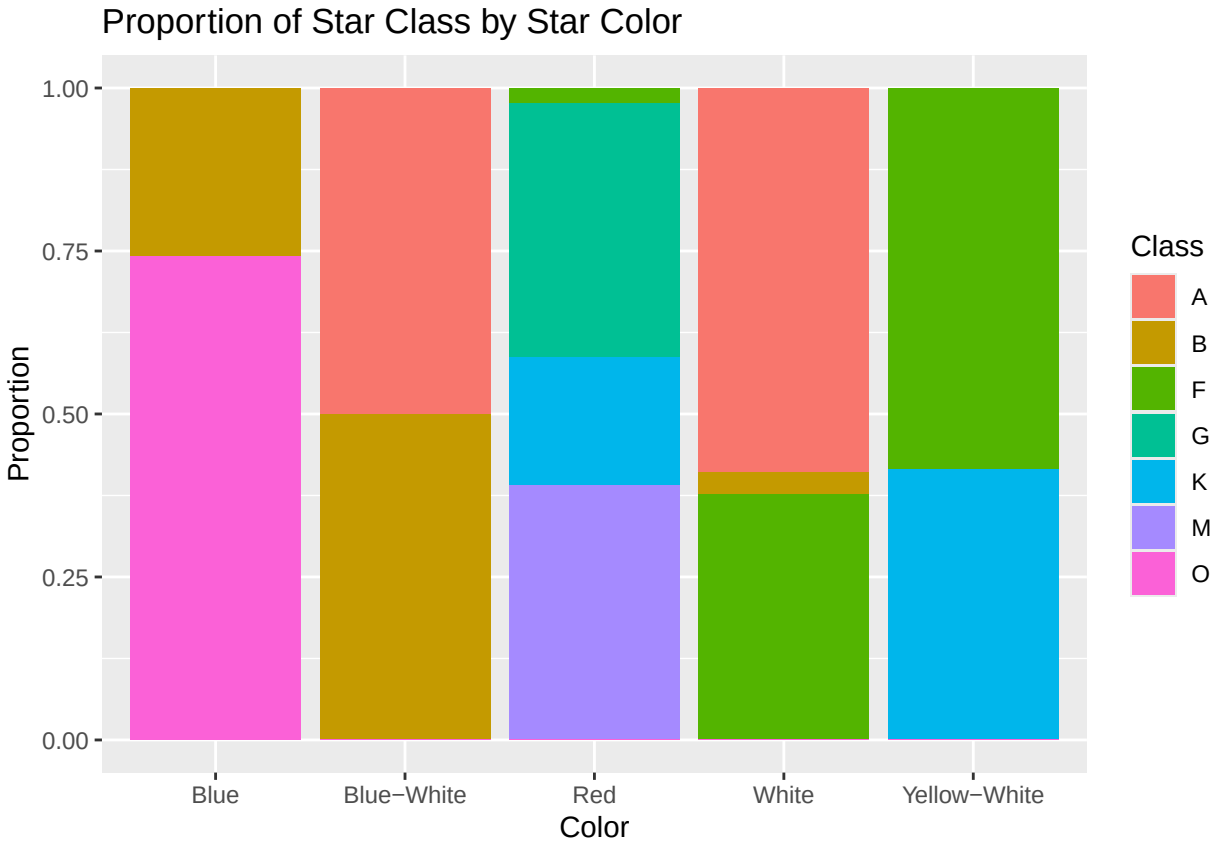


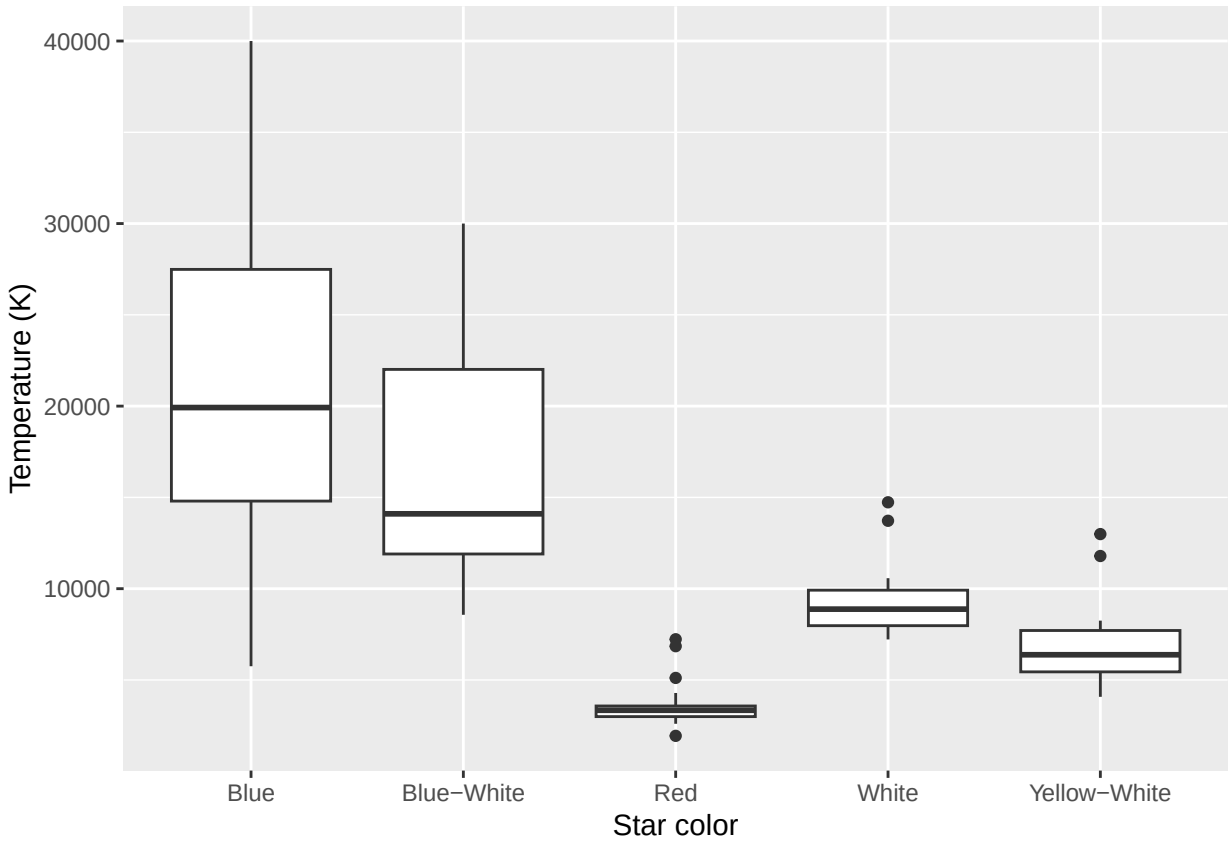
Table 15: Proportion of Star Color by Spectral Class

	A	B	F	G	K	M	O
Blue	0.0	34.8	0.0	0	0	0	100
Blue-White	63.2	63.0	0.0	0	0	0	0
Red	0.0	0.0	5.9	100	50	100	0
White	36.8	2.2	23.5	0	0	0	0
Yellow-White	0.0	0.0	70.6	0	50	0	0

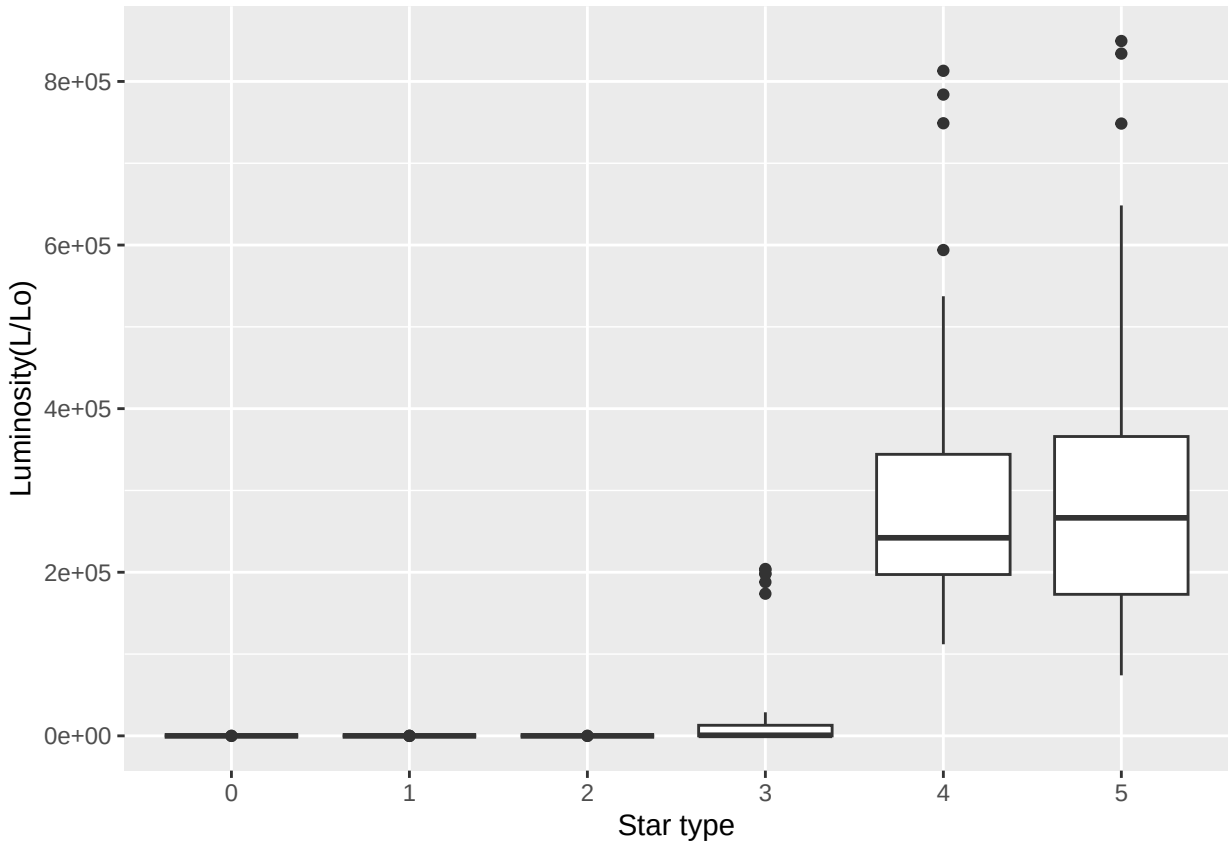


Above are proportions of each of the categorical variables, sorted by each other. It seems that some of the categorical values are associated with each other. For example, it seems a "B" class star can be on a range from Blue to White in color, but a Blue star can be either a "B" or "O" class star. Furthermore, a large portion of the class "O" stars are type 4, but they can be types 3, 4, or 5. One interesting thing to note is that red stars were by far the largest color group of star, but they do not span across all of the spectral classes. Rather, they are only listed as class F, G, K, and M stars. Considering there is only one class "G" star recorded in the dataset, though, it is possible there are class G stars of more colors and types. This is what gives us the 100% type 5 proportion in Class G.

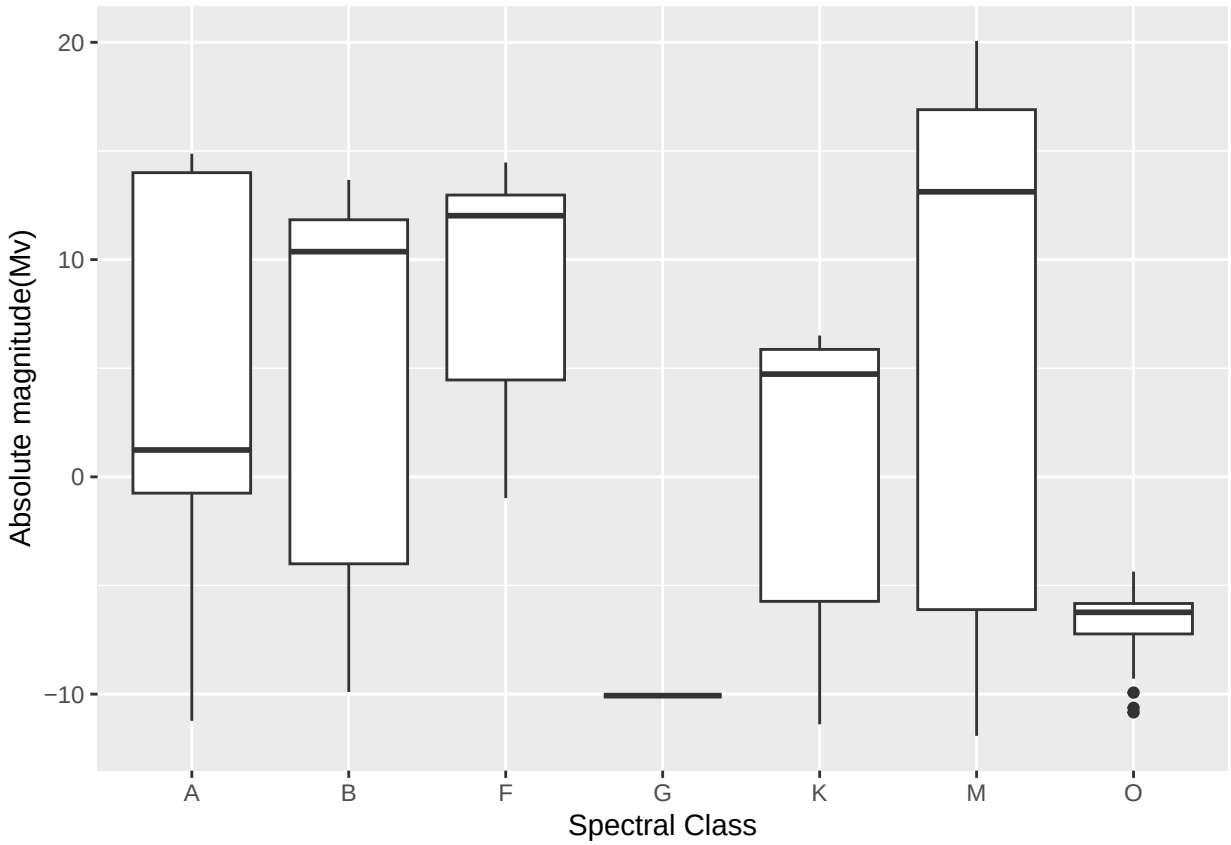
```
##          Color  Avg Temp
## 1         Blue 21918.339
## 2   Blue-White 16659.951
## 3          Red  3353.948
## 4        White  9579.583
## 5 Yellow-White  6992.867
```



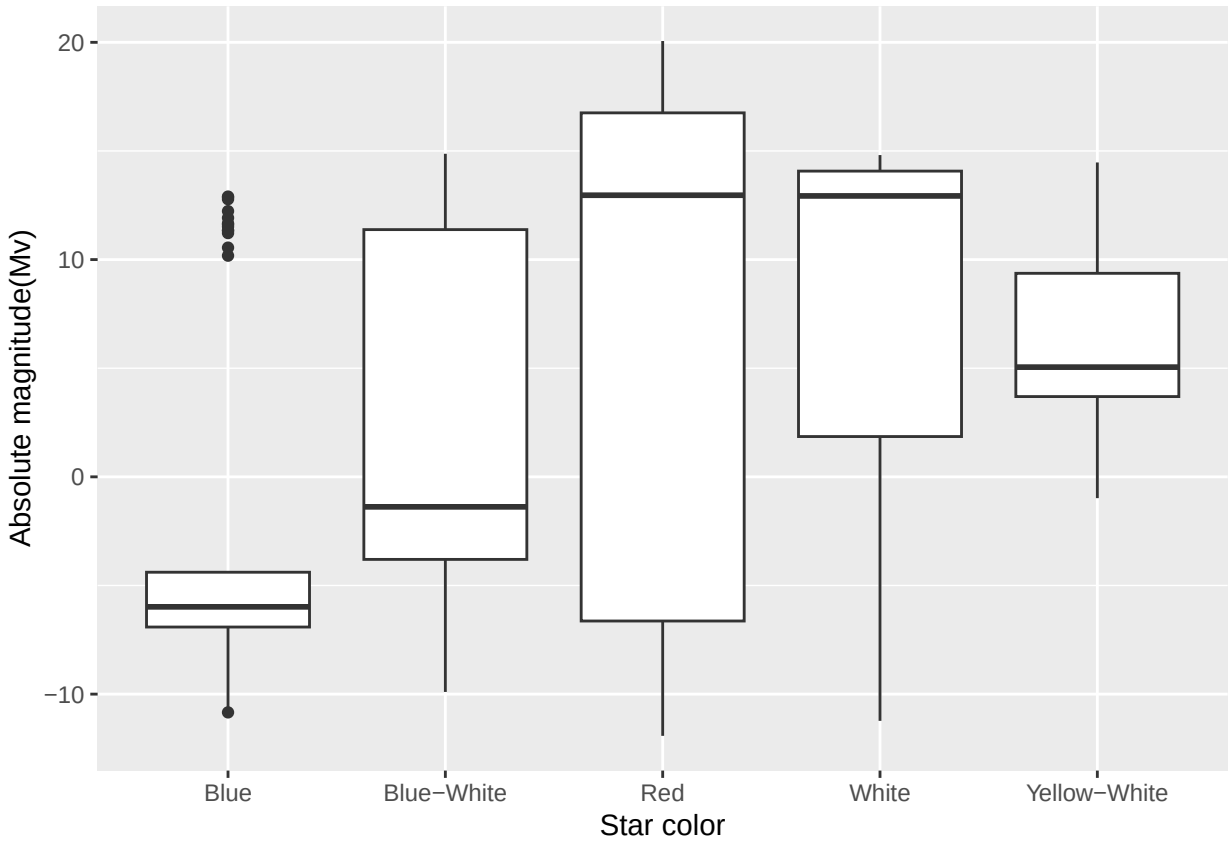

```
##   Star Type Avg Luminosity
## 1         0   6.932750e-04
## 2         1   5.405750e-03
## 3         2   2.433625e-03
## 4         3   3.206739e+04
## 5         4   3.018162e+05
## 6         5   3.092465e+05
```




```
##   Spectral Class Avg Abs Magnitude
## 1              A         4.0852105
## 2              B         3.7226087
## 3              F         8.6117647
## 4              G       -10.0700000
## 5              K         0.2673333
## 6              M         8.3678288
## 7              O        -6.5961750
```




```
##     Star Color Avg Abs Magnitude
## 1         Blue          -2.382446
## 2   Blue-White           1.968268
## 3          Red           7.904560
## 4        White           7.486667
## 5 Yellow-White           6.514933
```



Above are averages and box plots for each of the numeric variables. I am very interested in how these properties of stars affect color. It seems to me that the color of a star correlates with a few of the variables. In this case, it seems that The absolute Magnitude of a star can be based upon color, with the highest variation in the largest group (Red). Interestingly enough, though, the colors of star with the highest absolute magnitude (the Red-Yellow-White spectrum) seem to have the lowest temperature. This is strange, because one would intuitively expect things that are brighter to also be hotter.

Analysis

I have decided to use LDA to predict star color. This is because I have multiple continuous predictor variables, and a variable that I want to predict with multiple (5) levels. An LDA allows me to predict, using a machine learning algorithm, the color of each star without splitting the colors up into multiple binary response variables. At the end, I have also included a decision tree for similar reasons, and because it is interesting to see, coupled with the LDA.

Data Overview

```
## spc_tbl_ [240 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Temperature (K)       : num [1:240] 3068 3042 2600 2800 1939 ...
##  $ Luminosity(L/Lo)      : num [1:240] 0.0024 0.0005 0.0003 0.0002 0.000138 0.00065 0.00073 0.0004 0
##  $ Radius(R/Ro)          : num [1:240] 0.17 0.154 0.102 0.16 0.103 ...
##  $ Absolute magnitude(Mv): num [1:240] 16.1 16.6 18.7 16.6 20.1 ...
##  $ Star type             : num [1:240] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Star color            : chr [1:240] "Red" "Red" "Red" "Red" ...
##  $ Spectral Class        : chr [1:240] "M" "M" "M" "M" ...
##  - attr(*, "spec")=
##   .. cols(
```

```
##   ..   `Temperature (K)` = col_double(),
##   ..   `Luminosity(L/Lo)` = col_double(),
##   ..   `Radius(R/Ro)` = col_double(),
##   ..   `Absolute magnitude(Mv)` = col_double(),
##   ..   `Star type` = col_double(),
##   ..   `Star color` = col_character(),
##   ..   `Spectral Class` = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

As shown by the R output above, the data contains 7 variables, 5 of which are continuous quantitative variables, and 2 of which are categorical variables. There are 240 total observations. I will be using the 5 continuous variables to attempt to predict the color of each star.

Scaling the Data

I will be attempting to use Linear Discriminant Analysis to predict the color of the stars. One of the most important assumptions of an LDA is that the predictor variables have the same variance. In order to meet this assumption, I will be scaling the data, such that each predictor variable has a mean of 0 and a standard deviation of 1.

```
##        Temperature (K)       Luminosity(L/Lo)           Radius(R/Ro)
##           2.518963e-17           1.371967e-17          -2.081668e-17
## Absolute magnitude(Mv)              Star type
##          -4.077233e-17           0.000000e+00
```

```
##        Temperature (K)       Luminosity(L/Lo)           Radius(R/Ro)
##                      1                      1                      1
## Absolute magnitude(Mv)              Star type
##                      1                      1
```

As you can see from the R output above, the means are incredibly close to 0, and the standard deviations are 1 across all 5 predictor variables.

Training and Test Sets

Now, I need to split the dataset into a training set, which will be used to train the model to attempt to predict star color, and a testing set, which will be used to determine if the model can accurately predict star color. I am going to use 70% of the dataset as a training set and 30% of the dataset as a test set.

Fitting the Model

Here, I will be fitting the LDA model using the testing set. The resulting R output is as follows:

```
## Call:
## lda(`Star color` ~ `Temperature (K)` + `Luminosity(L/Lo)` + `Radius(R/Ro)` +
##     `Absolute magnitude(Mv)` + `Star type`, data = train)
##
## Prior probabilities of groups:
##         Blue   Blue-White          Red        White Yellow-White
##    0.2721893    0.1538462    0.4556213    0.0591716    0.0591716
##
## Group means:
##              `Temperature (K)` `Luminosity(L/Lo)` `Radius(R/Ro)`
## Blue                1.2912149         0.77771882     0.02611607
## Blue-White          0.6936256        -0.08863457    -0.11223432
## Red                -0.7458181        -0.20645952     0.16917759
```

```
## White               -0.1604370        -0.07209848     0.03251878
## Yellow-White        -0.3700068        -0.59736390    -0.45706407
##              `Absolute magnitude(Mv)` `Star type`
## Blue                       -0.5963444   0.6097241
## Blue-White                 -0.2180183   0.2696856
## Red                         0.3141325  -0.3983992
## White                       0.2003894   0.1752957
## Yellow-White                0.2336578   0.1168638
##
## Coefficients of linear discriminants:
##                                    LD1        LD2        LD3        LD4
## `Temperature (K)`        -1.59918786  0.4567125 -0.3629854 -0.6133714
## `Luminosity(L/Lo)`       -0.08843192  0.4900012  1.2647083  0.3354316
## `Radius(R/Ro)`            1.19605362  0.5547555 -0.1597165 -1.1499454
## `Absolute magnitude(Mv)` -2.05419989 -2.5609993  1.7496483 -1.3532720
## `Star type`              -3.12048760 -3.3835218  1.1868437 -0.4442658
##
## Proportion of trace:
##    LD1    LD2    LD3    LD4
## 0.9178 0.0697 0.0107 0.0019
```

Here you can see the prior probability that a star would have been a certain color. For example, 6% if the stars in the test set were white, and 46% were red. As well, you can see the group means, or the means of each predictor variable across each color of star. The coefficients of linear discriminants can be read to build the linear combinations of the predictor variables that were used to form the models LD1 and LD2. Lastly, the proportion of trace shows the percentage separation achieved by each subsequent linear discriminant. For example, LD1 alone achieves approximately 92% of the separation in the model. Ld1 separates stars based on a value that is high if the star has low temperature, luminosity, magnitude, and type, with a low radius, and a low value if the reverse is true.

Making Predictions in the Test Set using the Model

```
## [1] "class"     "posterior" "x"
```

```
## [1] Red        Red        Red        Blue       Blue-White Blue-White
## Levels: Blue Blue-White Red White Yellow-White
```

Here, you can see the three variables the prediction has returned. "Class" is the color the model predicted. "Posterior" is the posterior probability that an observation belongs to each class. "x" contains the linear discriminants. Below that are the color predictions for 6 observations.

```
##      Blue Blue-White      Red    White Yellow-White
## 1 0.0e+00   0.000010 0.999955 0.000028     0.000007
## 2 0.0e+00   0.000013 0.999939 0.000039     0.000009
## 3 0.0e+00   0.000013 0.999930 0.000046     0.000011
## 4 4.0e-06   0.000282 0.998652 0.000715     0.000348
## 5 5.9e-05   0.002694 0.977161 0.013633     0.006454
## 6 4.0e-06   0.000262 0.999129 0.000409     0.000196
```

Here are the posterior probabilities for the first 6 observations in the test set.

```
##        LD1        LD2         LD3         LD4
## 1 3.133223  1.0272846 -0.13280076 -0.04393737
## 2 3.062049  0.9489033 -0.07948609 -0.08879475
## 3 3.049265  0.8808594 -0.03190897 -0.10663416
## 4 2.284355  0.2266987 -0.24226579  0.33161183
## 5 1.672794 -0.4975566  0.25175150 -0.06448357
```

```
## 6 2.302199  0.5156470 -0.44509161  0.39014259
```

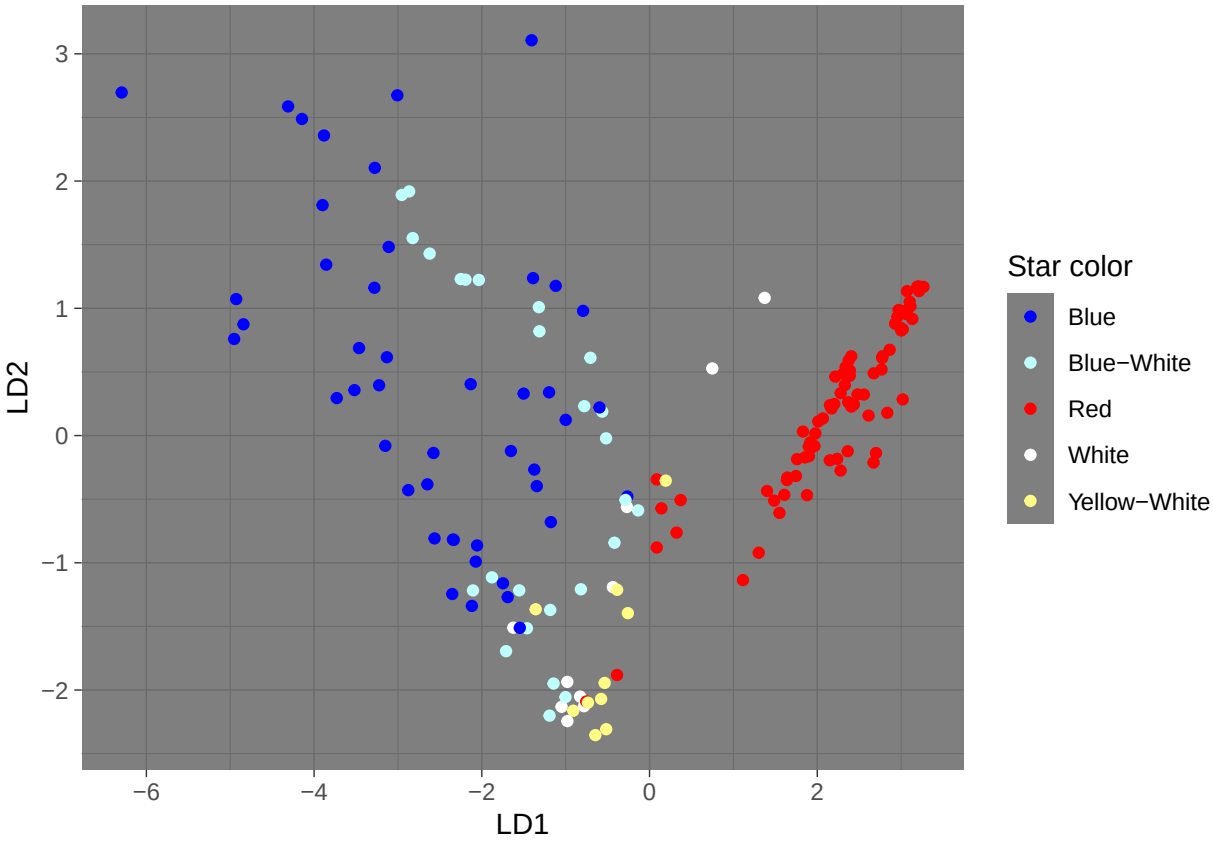
Here are the linear discriminants for the first 6 observations in the test set. As I have said before, LD1 increases value with high radius, but low temperature, luminosity, magnitude, and star type. Luminosity has much less of an effect on these values. LD2, however, increases with temperature, luminosity and radius being high values, and absolute magnitude and star type being low values. I will use LD1 and LD2 to graph separation later in this paper.

Accuracy of the Model

We can use r code to see the percentage of observations in which the LDA model correctly predicted the color.

```
## [1] 0.8873239
```

As you can see, this LDA model correct predicted about 88.7% of the observations in the test dataset. That is not bad, but we would hope for a higher percentage correctly predicted. Perhaps some adjustments in the dataset may help. In order to determine if adjustments may help, we can check the assumptions for an LDA in the dataset. A graph to visualize the separation is shown below.

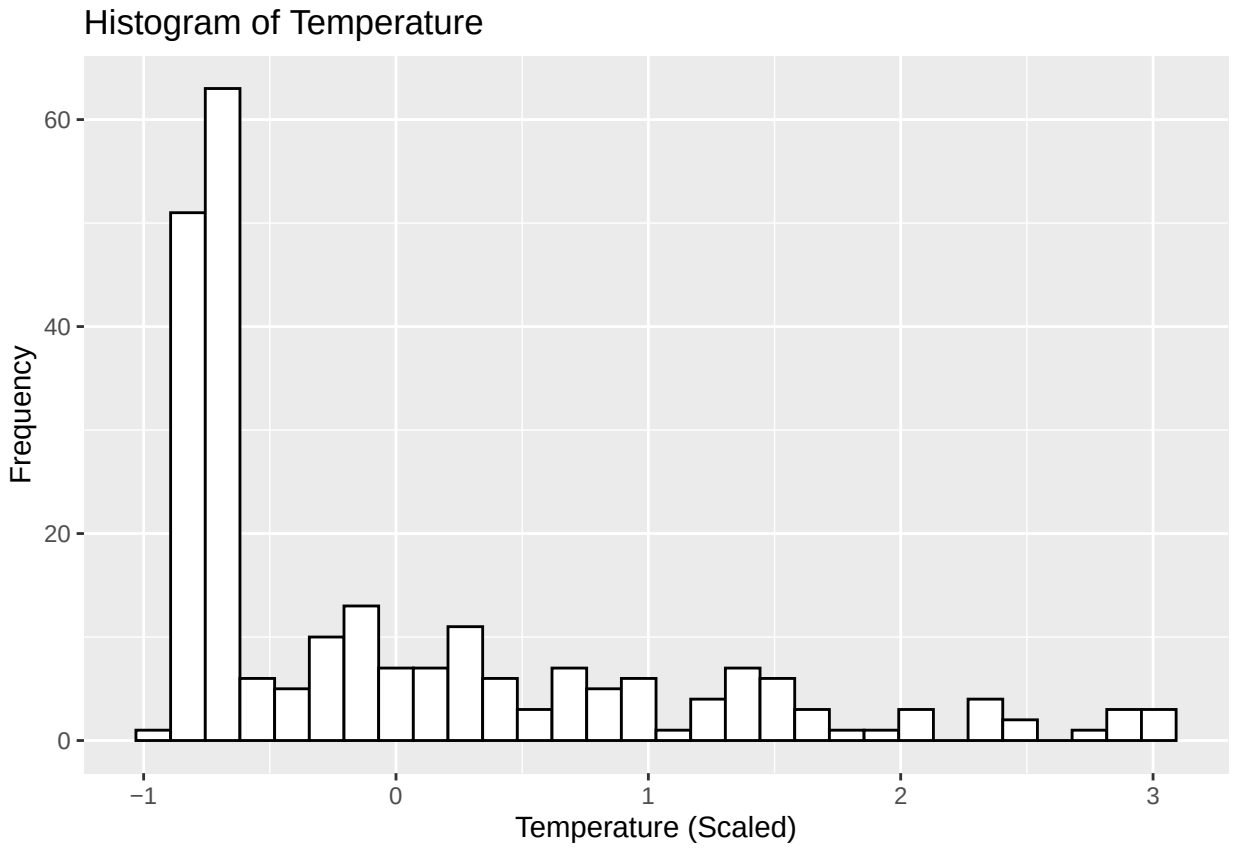


In this graph, the value for LD1 is scaled along the x-axis, and LD2 the y-axis. There seems to be far more separation along the LD1 axis, as was expected. As well, the model does a very good job separating the red stars from other stars, but gets a little muddied somewhere in the middle of the other colors. This could be because machine learning models prioritize the category with the most responses. Perhaps if we had more data of stars for the rest of the colors, the model could do a better job separating out those colors as well.

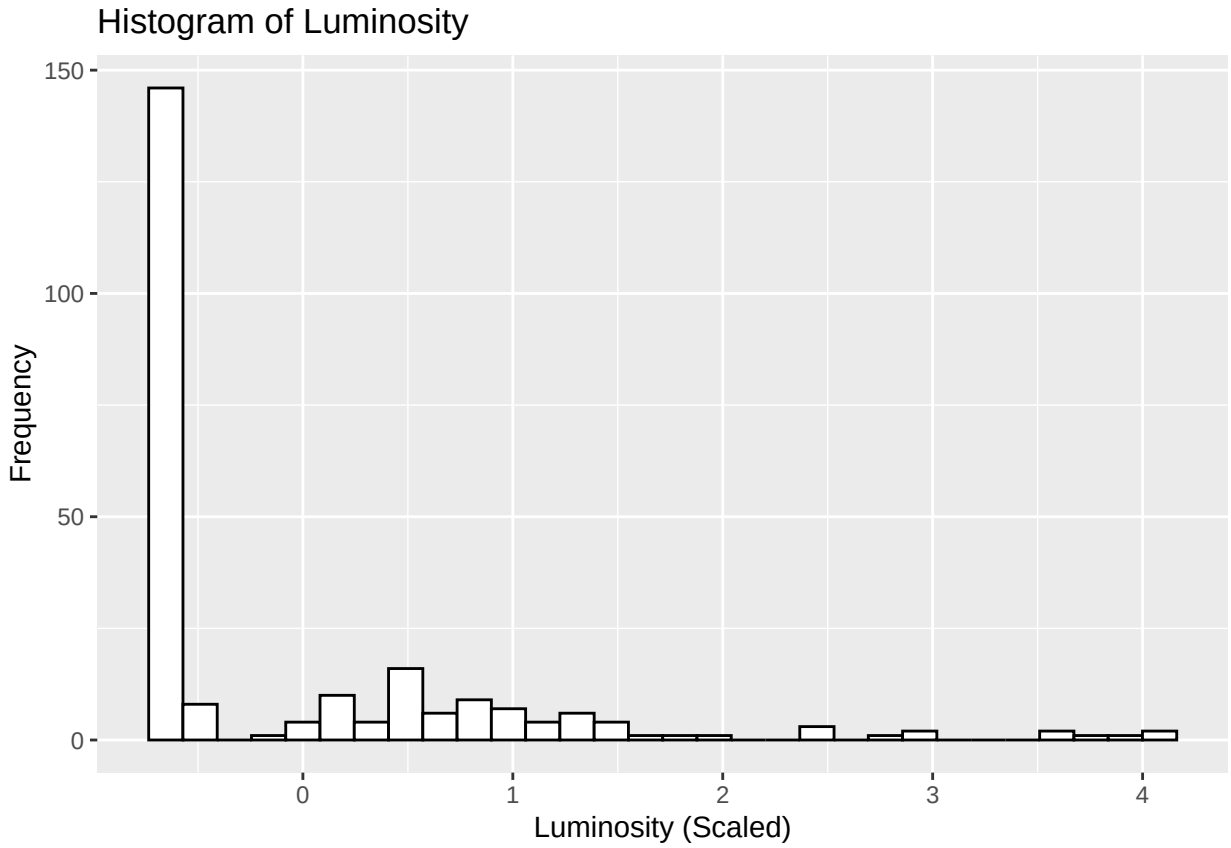
Checking Assumptions

There are three key assumptions in an LDA, which stem from it being a parametric test. It assumes the data measurements are independent from each other, normality within the independent data, and equal variance-covariance matrices. As the measurements are all stars that are independent from each other, I am going to assume that assumption holds. As for the normality assumption, we can use a histogram and a Q-Q plot to visually inspect the data, and the Shapiro-Wilk test to mathematically examine normality. Note that I have already scaled the data in an attempt to correct the equal variance assumption.

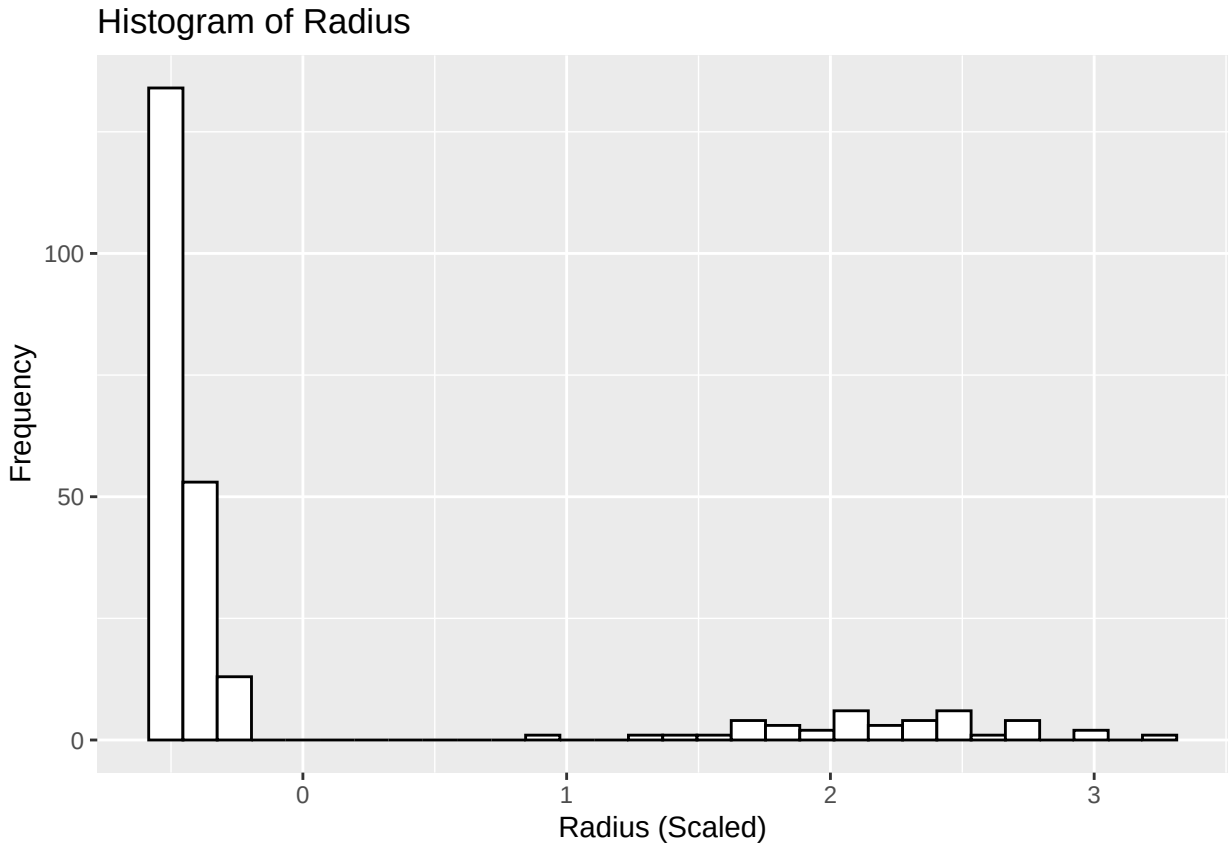
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



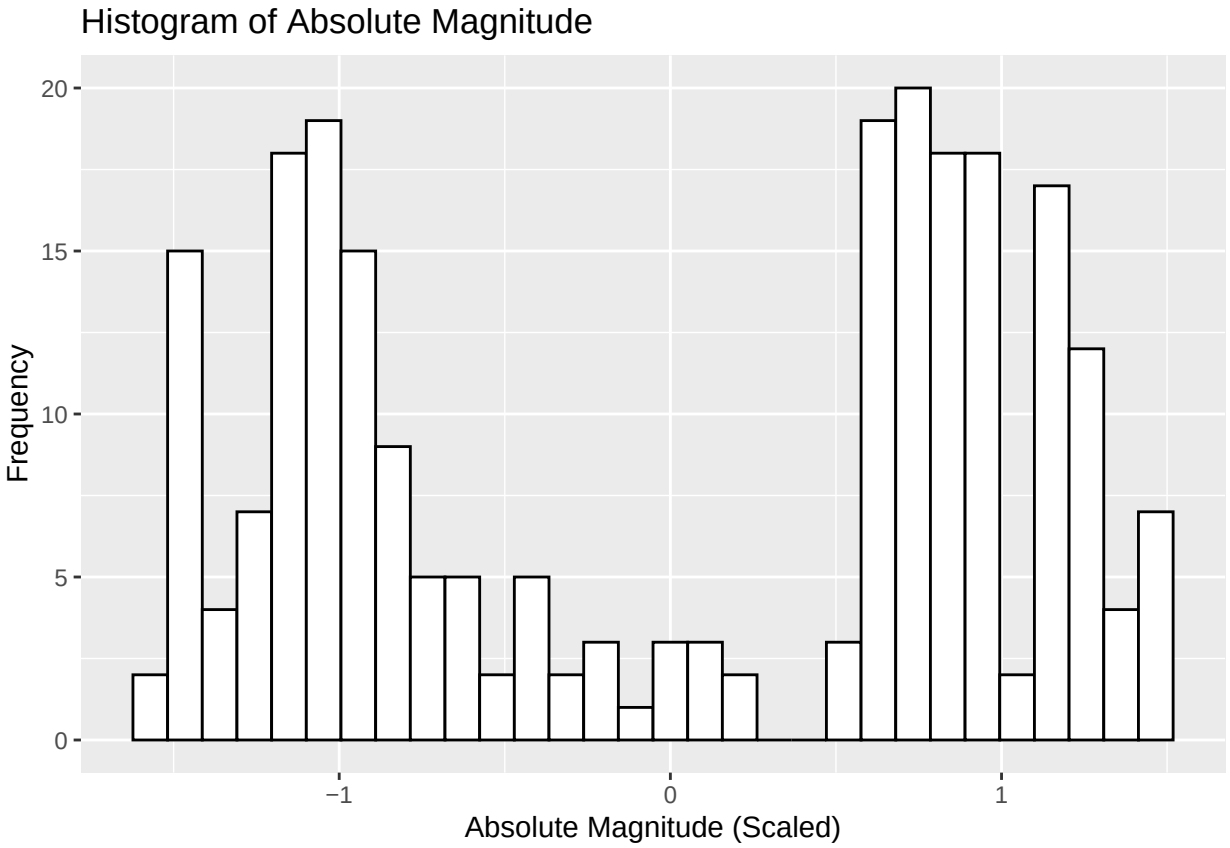

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



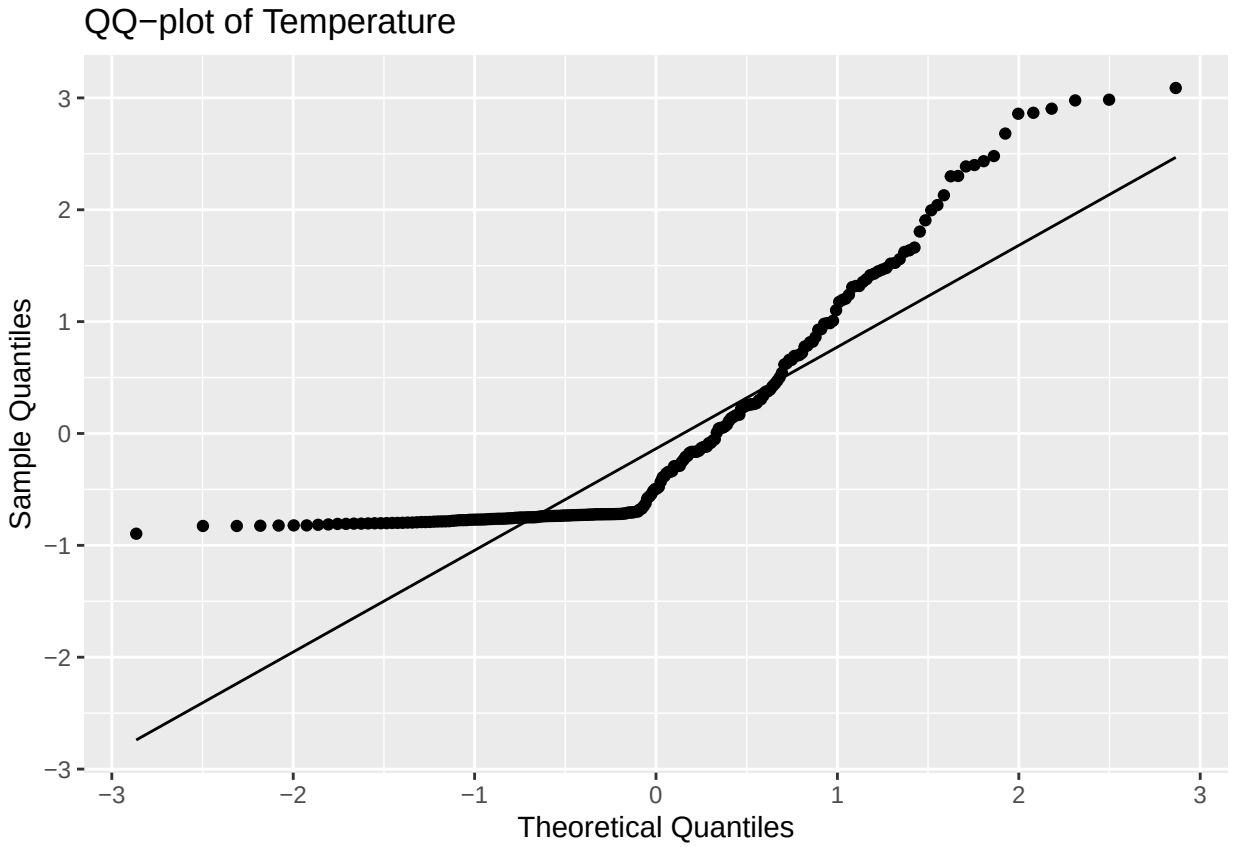

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

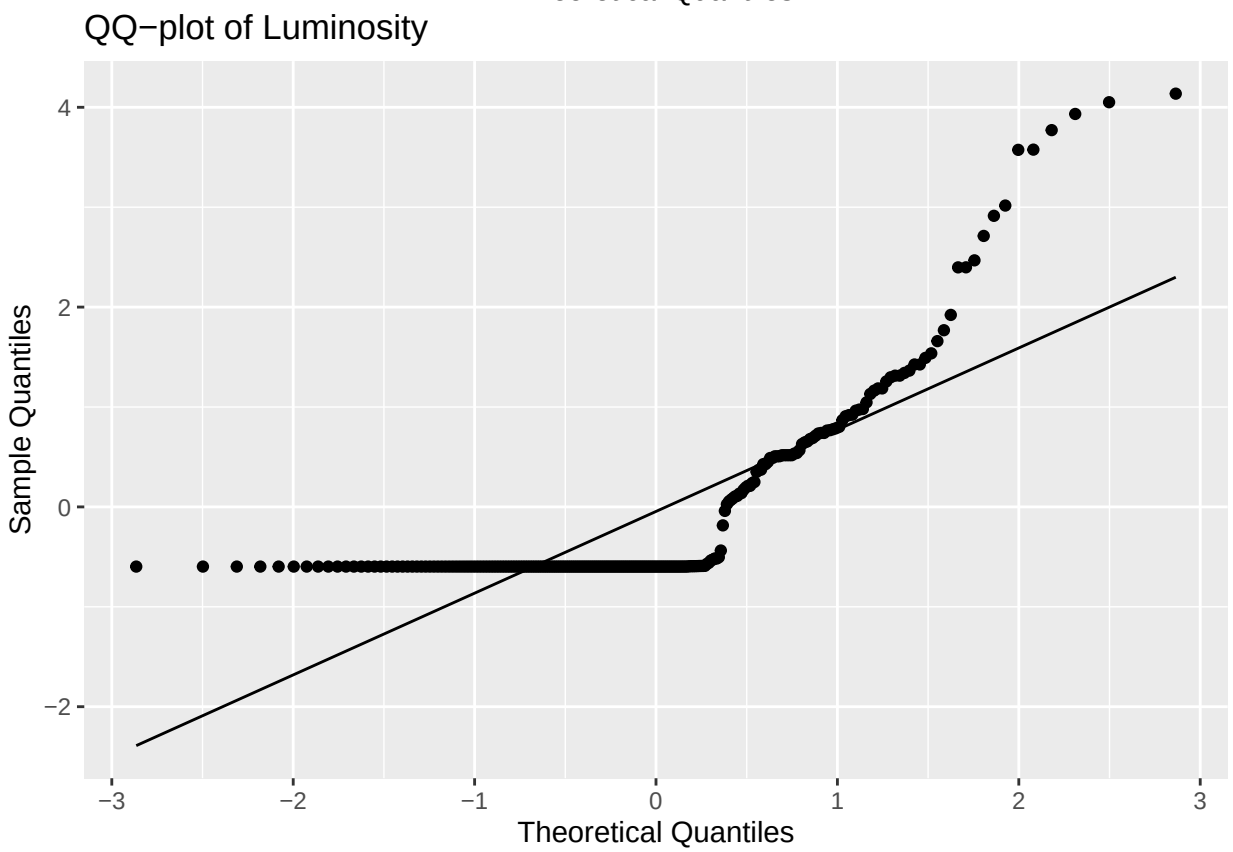


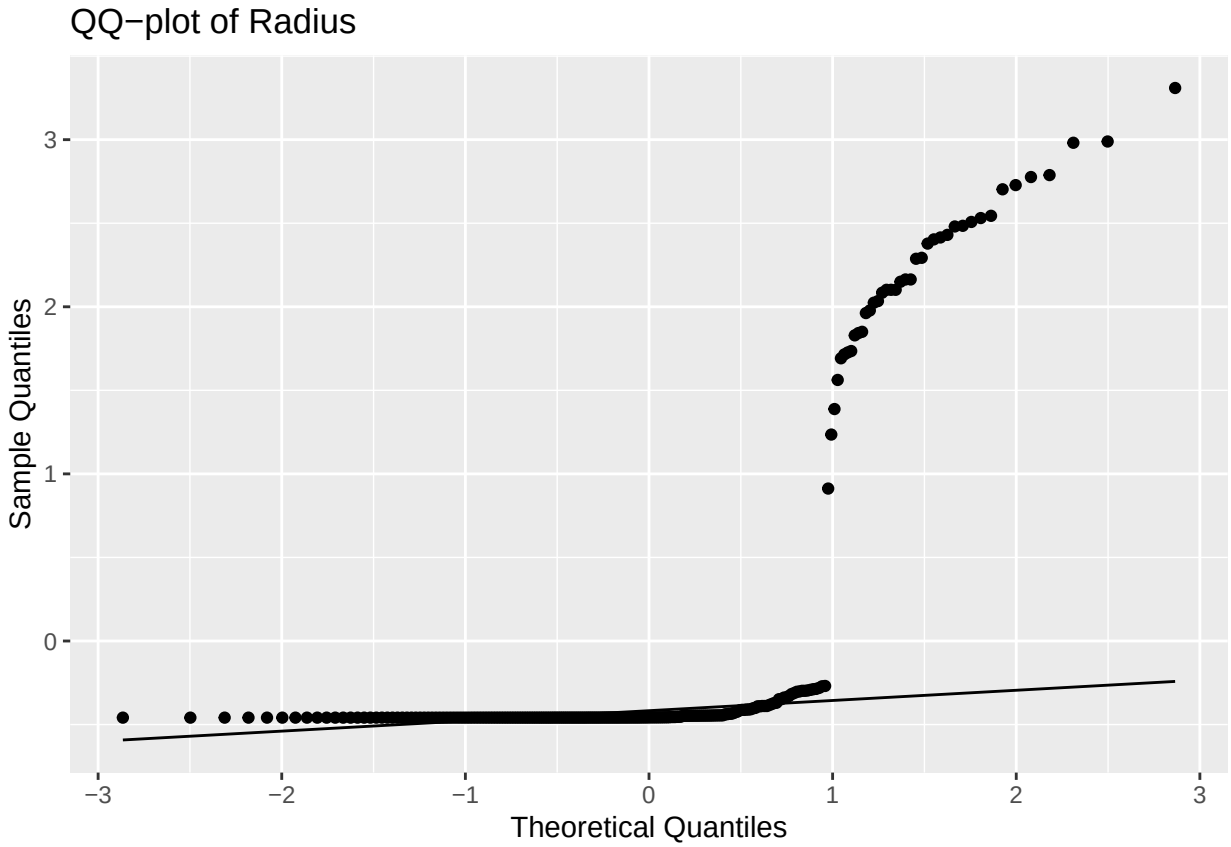

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

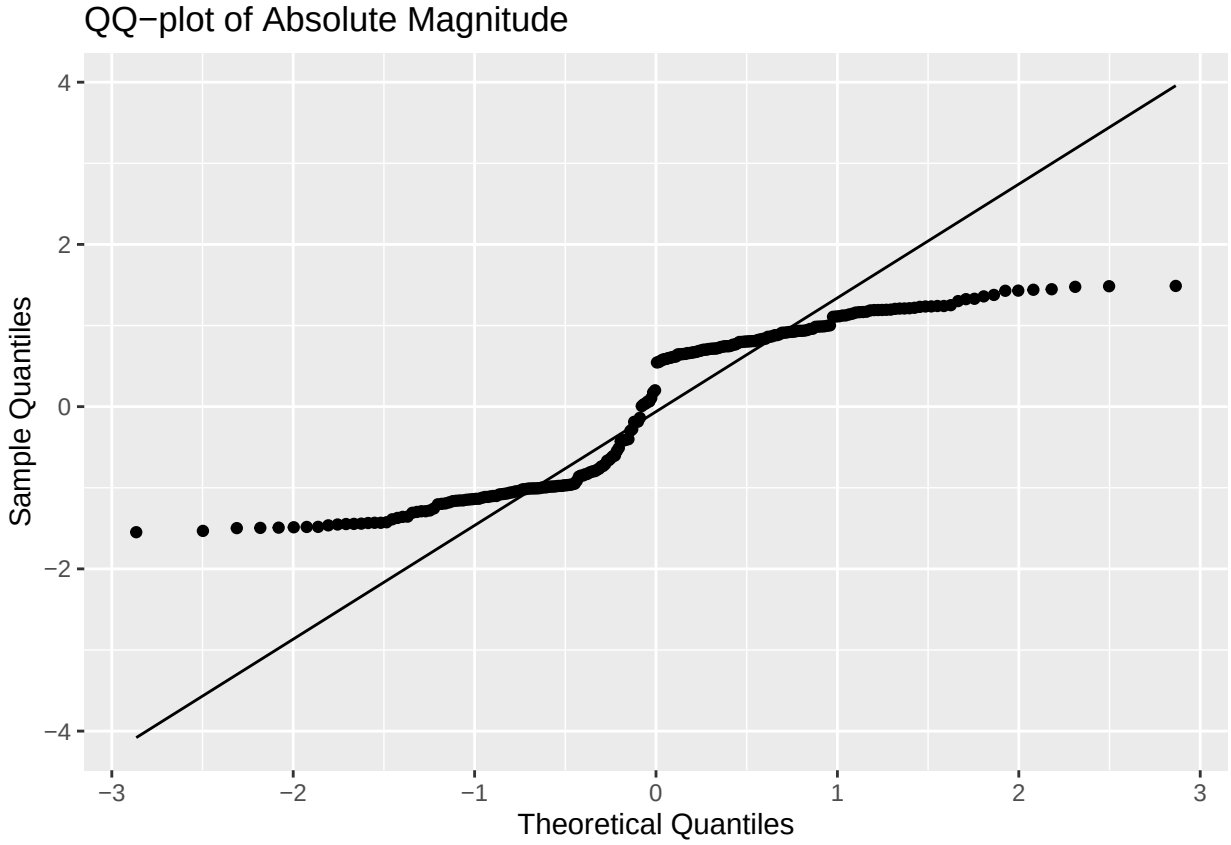


Above are Histograms of each of the scaled numeric variables. As you can see, based on the histograms, none of the data appears to be normal. Note that I did not include Star Type, as it has 6 levels of classification, and the data is evenly distributed across each type. I may attempt to exclude that variable in further examination after checking assumptions. I will now check Q-Q plots and conduct Shapiro-Wilk tests for each of these variables.


QQ−plot of Temperature
Sample Quantiles
Theoretical Quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


QQ−plot of Luminosity
Sample Quantiles
Theoretical Quantiles
4
2
0
−2
−3
−2
−1
0
1
2
3


QQ−plot of Radius
Sample Quantiles
Theoretical Quantiles
3
2
1
0
−3
−2
−1
0
1
2
3


QQ−plot of Absolute Magnitude
Sample Quantiles
Theoretical Quantiles
4
2
0
−2
−4
−3
−2
−1
0
1
2
3

```
##
##  Shapiro-Wilk normality test
##
## data:  cleaned_star_data$`Temperature (K)`
## W = 0.79323, p-value < 2.2e-16
```

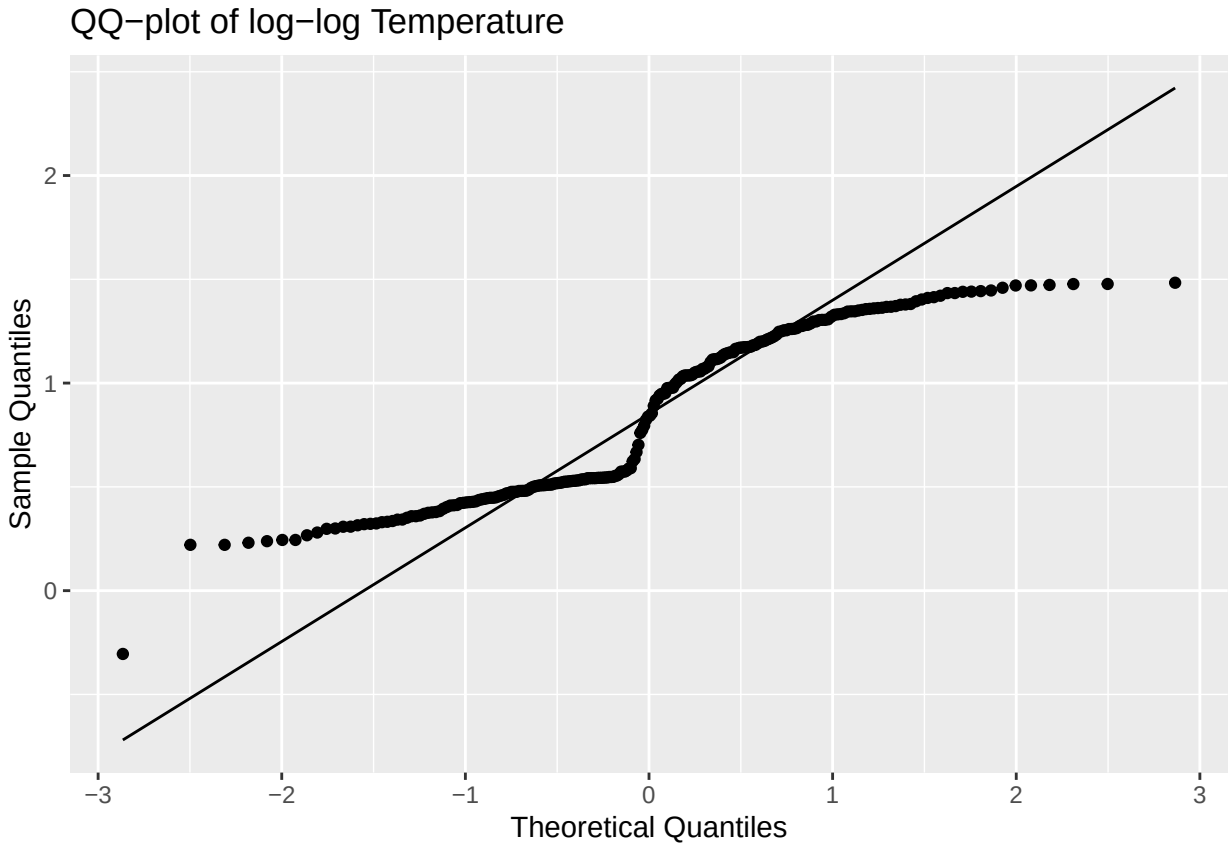
```
##
##  Shapiro-Wilk normality test
##
## data:  cleaned_star_data$`Luminosity(L/Lo)`
## W = 0.65973, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cleaned_star_data$`Radius(R/Ro)`
## W = 0.50215, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cleaned_star_data$`Absolute magnitude(Mv)`
## W = 0.8691, p-value = 1.789e-13
```

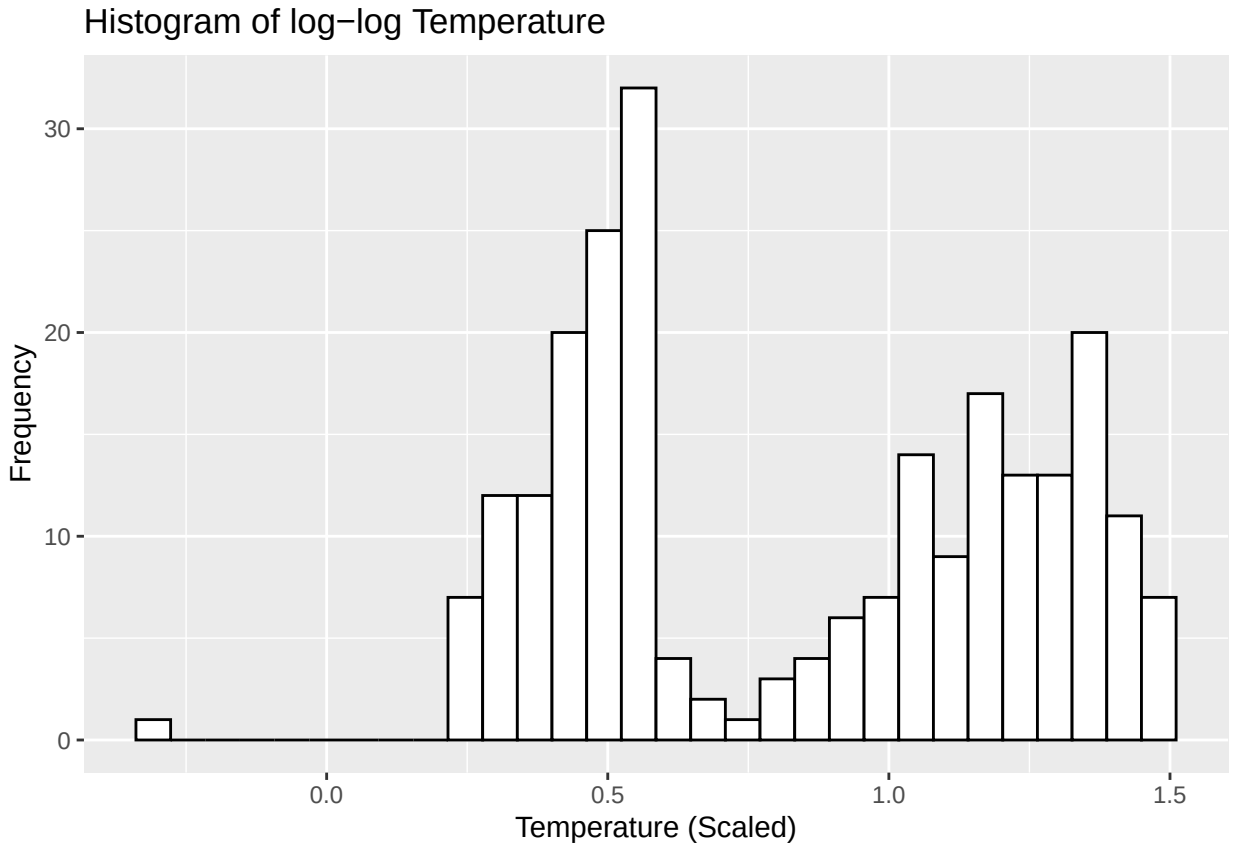
Above are Q-Q plots and Shapiro-Wilk tests for each variable. As you can see, the data does not appear normal on Q-Q plots either. To confirm that, there are related Shapiro-Wilk tests for each of these four variables, all of which reject the null hypothesis that the data is normally distributed.

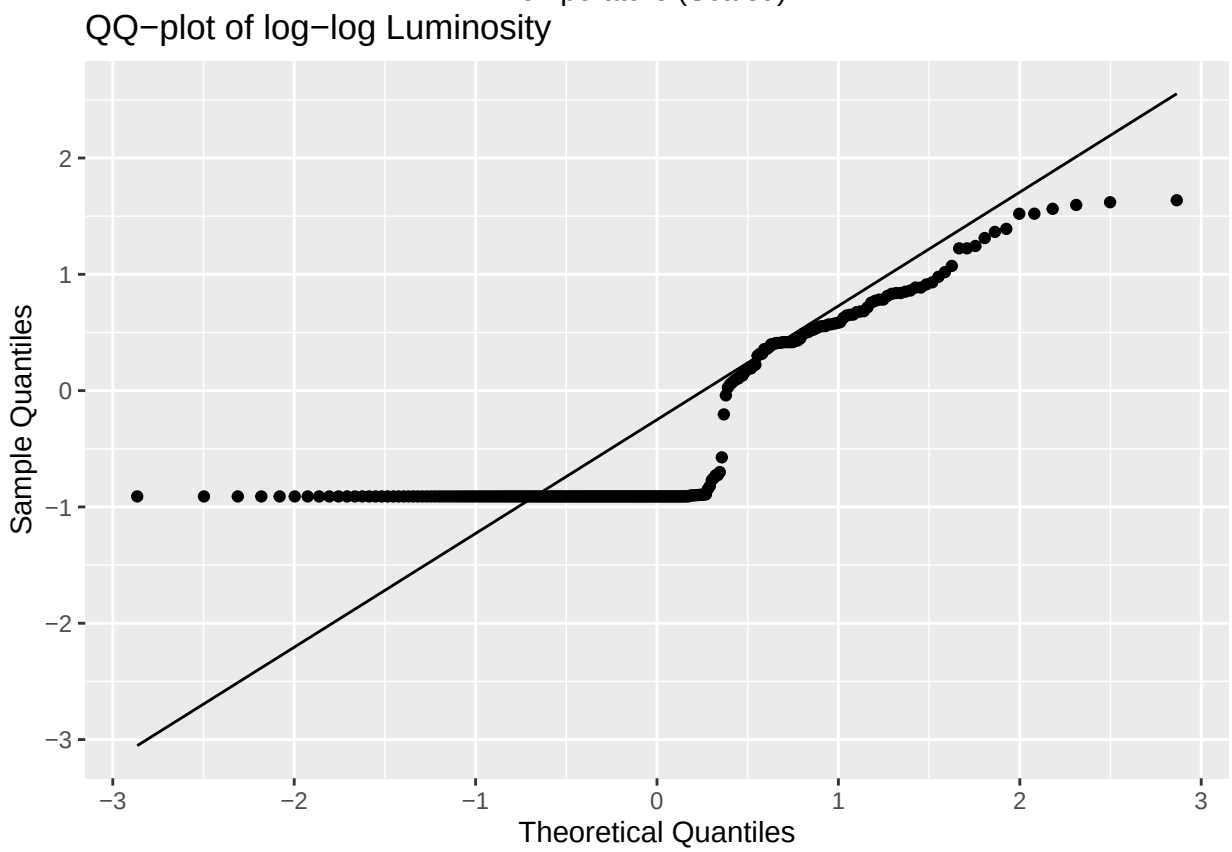
I will now attempt to transform the variables to make them follow a normal distribution. I will be using log-log transformation, since I still need the variable to be continuous. The data was adjusted such that it did not allow any negative variables, in order to take logs. The plots and distributions were not affected. The following is the results:




```
##
##  Shapiro-Wilk normality test
##
## data:  cleaned_star_data$loglogtemp
## W = 0.90219, p-value = 2.14e-11
```

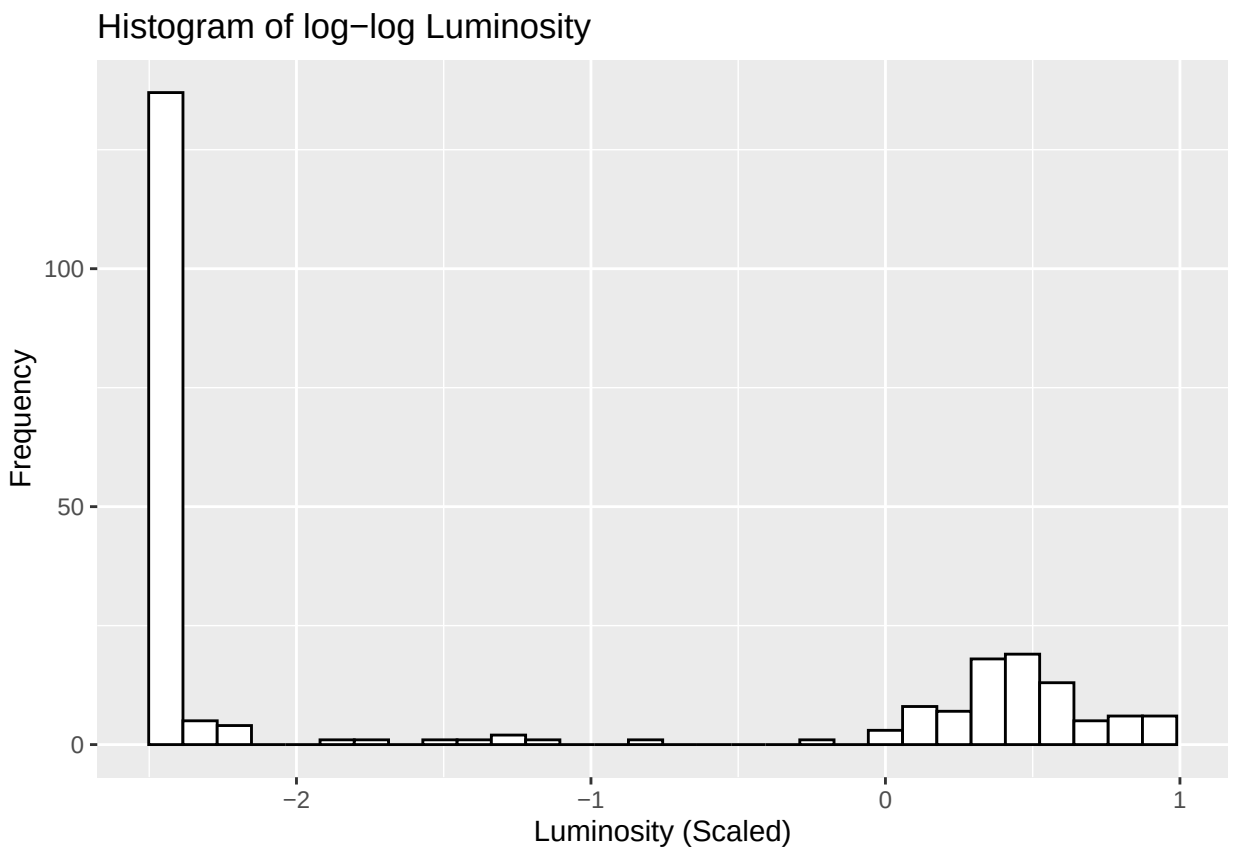
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

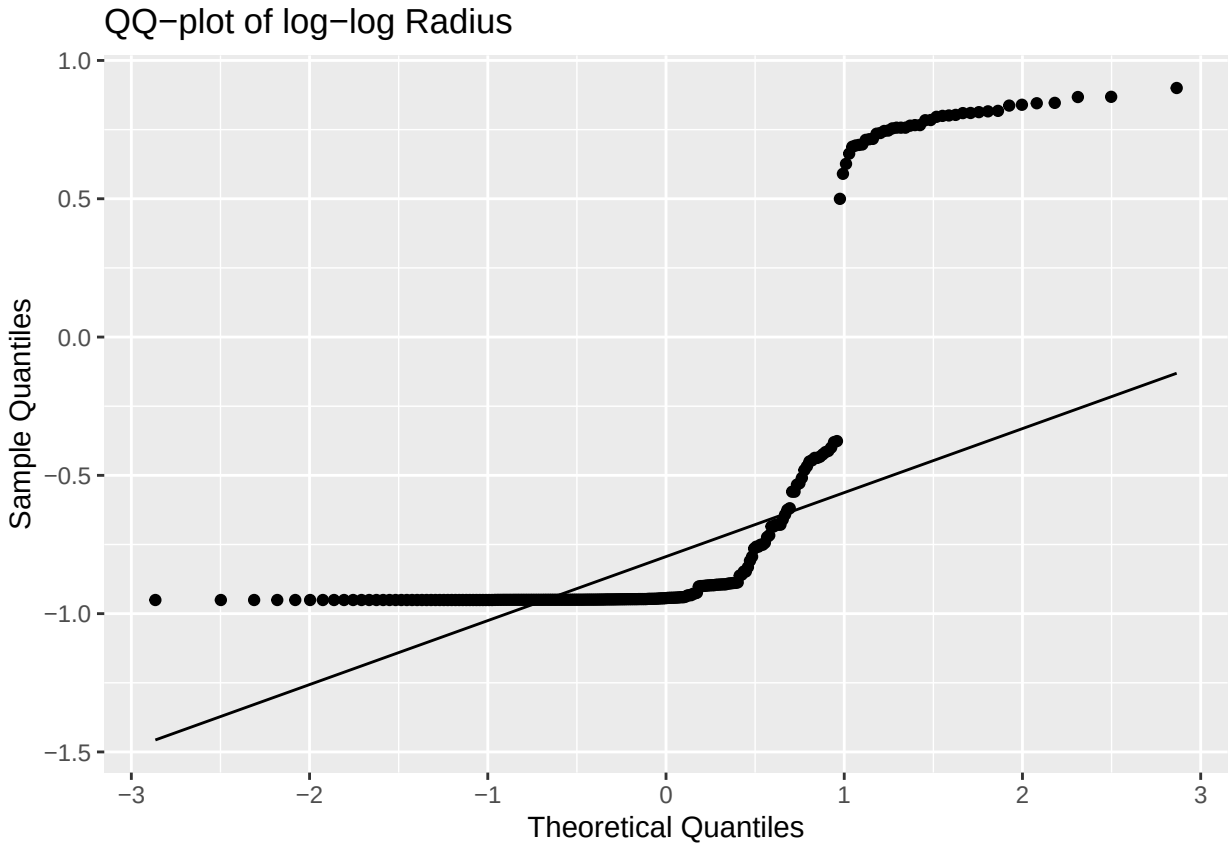

Histogram of log−log Temperature
Frequency
30
20
10
0
0.0
0.5
1.0
1.5
Temperature (Scaled)


QQ−plot of log−log Luminosity
Sample Quantiles
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Theoretical Quantiles

```
##
##  Shapiro-Wilk normality test
##
## data:  cleaned_star_data$logloglum
## W = 0.68234, p-value < 2.2e-16
```

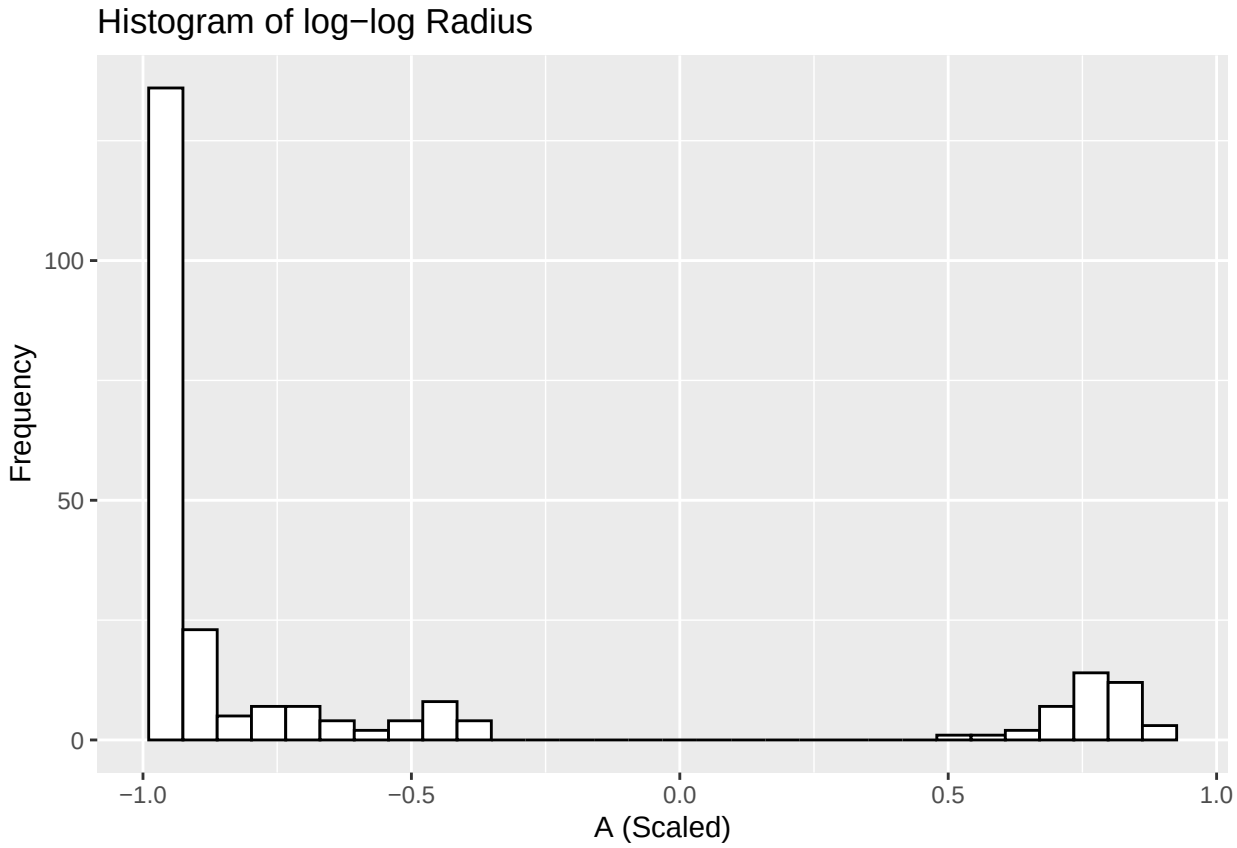
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

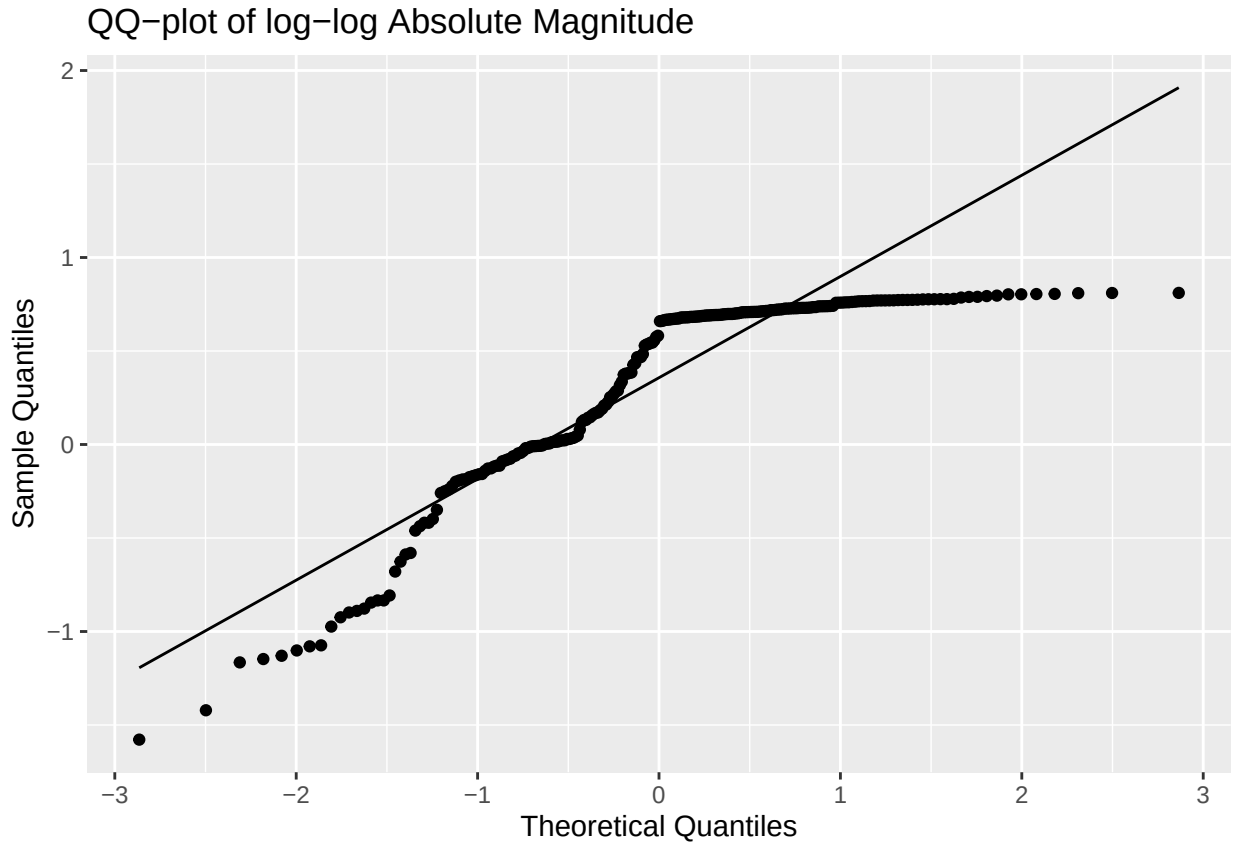





```
##
##  Shapiro-Wilk normality test
##
## data:  cleaned_star_data$loglograd
## W = 0.5824, p-value < 2.2e-16
```

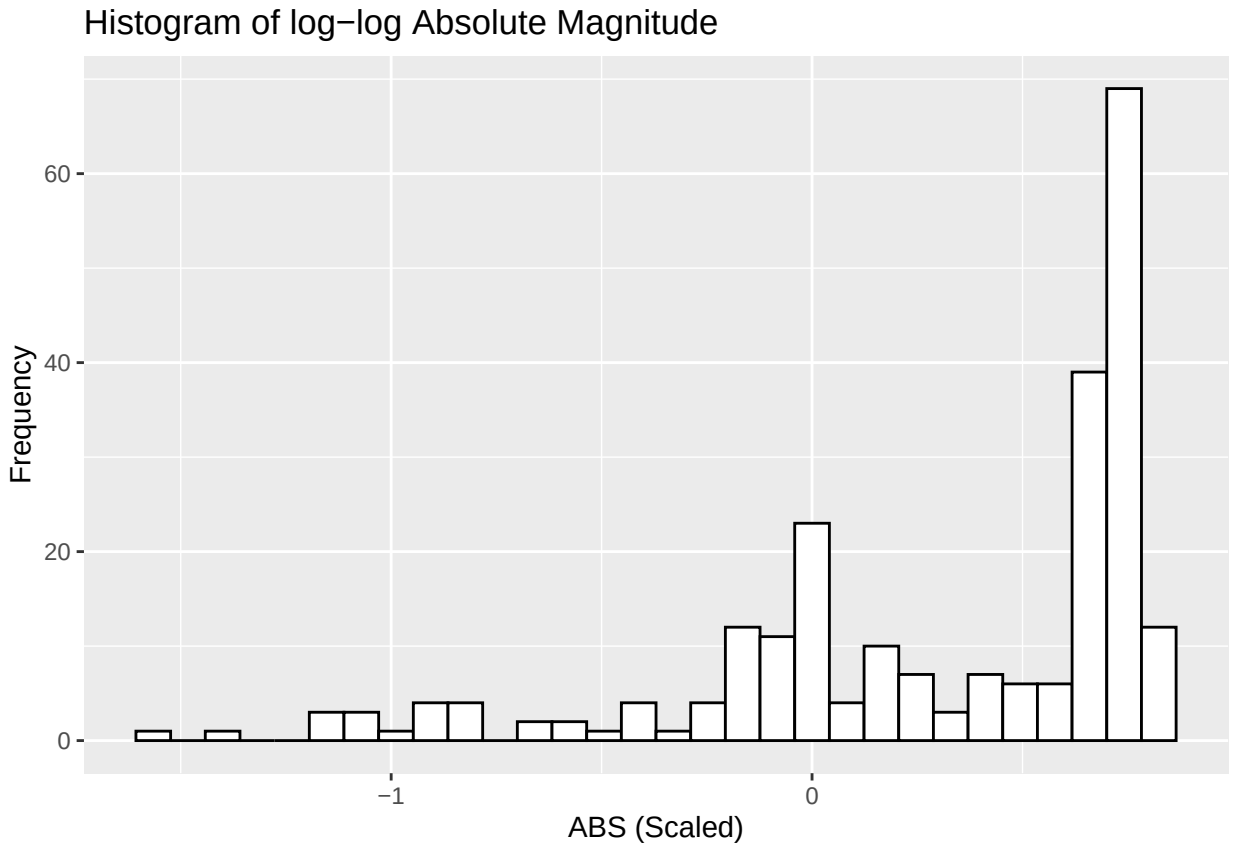
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Histogram of log−log Radius
Frequency
100
50
0
−1.0
−0.5
0.0
0.5
1.0
A (Scaled)


QQ−plot of log−log Absolute Magnitude
Sample Quantiles
2
1
0
−1
−3
−2
−1
0
1
2
3
Theoretical Quantiles

```
##
##  Shapiro-Wilk normality test
##
## data:  cleaned_star_data$loglogabs
## W = 0.82272, p-value = 7.624e-16
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Above are the log-log transformation graphs for each variable. As you can see, while log-log transformations certainly make some of the variables, notably Temperature, appear more normal visually, they do not make them follow a normal distribution, and they certainly do not appear normal in a Shapiro-Wilk test after transformation. An LDA is still robust when the data is not normal, and considering the almost 90% correct prediction rate in the LDA, along with the high separation at LD1, I think it is best to continue without transformation of the variables. I am, however, going to attempt an LDA without the variable star type in it, to see if that model can more accurately predict star color.

Second LDA model (without type)

Since the data was already scaled, and prepared for LDA, we can skip preparation steps in our LDA and fit new testing and training models. I am using the same training models before, as set in set.seed(1).

```
## Call:
## lda(`Star color` ~ `Temperature (K)` + `Luminosity(L/Lo)` + `Radius(R/Ro)` +
##     `Absolute magnitude(Mv)`, data = train)
##
## Prior probabilities of groups:
##         Blue   Blue-White          Red        White Yellow-White
##    0.2721893    0.1538462    0.4556213    0.0591716    0.0591716
##
```

```
## Group means:
##              `Temperature (K)` `Luminosity(L/Lo)` `Radius(R/Ro)`
## Blue                1.2912149         0.77771882     0.02611607
## Blue-White          0.6936256        -0.08863457    -0.11223432
## Red                -0.7458181        -0.20645952     0.16917759
## White              -0.1604370        -0.07209848     0.03251878
## Yellow-White       -0.3700068        -0.59736390    -0.45706407
##              `Absolute magnitude(Mv)`
## Blue                       -0.5963444
## Blue-White                 -0.2180183
## Red                         0.3141325
## White                       0.2003894
## Yellow-White                0.2336578
##
## Coefficients of linear discriminants:
##                                  LD1        LD2        LD3        LD4
## `Temperature (K)`        -1.5359598 -0.1488313 -0.6319491 -0.3809112
## `Luminosity(L/Lo)`       -0.1620140  1.2583883  0.5846884 -0.2660161
## `Radius(R/Ro)`            0.6312301  0.2630331 -1.2373130  0.1200783
## `Absolute magnitude(Mv)`  0.4237922  0.8582315 -0.6525896 -1.1500028
##
## Proportion of trace:
##    LD1    LD2    LD3    LD4
## 0.9651 0.0302 0.0043 0.0004
```

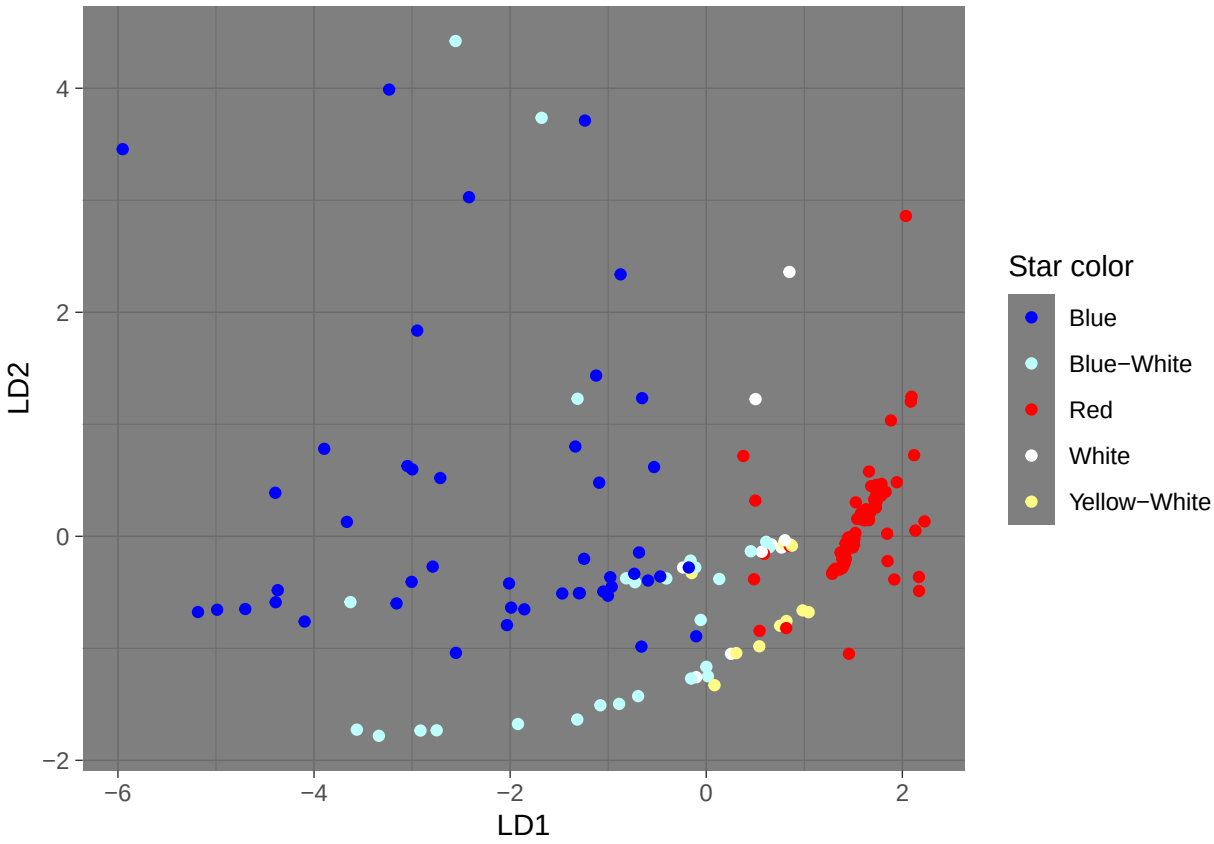
Since we are using the same data set, prior probabilities remain the same, as well as group means. However, the Coefficients of linear discriminants have changed, and we have achieved almost 5 percentage points more separation in LD1, as measured under proportion of trace.

```
## [1] Red Red Red Red Red Red
## Levels: Blue Blue-White Red White Yellow-White
```

```
##           Blue  Blue-White       Red      White Yellow-White
## 1 0.0002848211 0.005350042 0.8852903 0.06513194   0.04394288
## 2 0.0002786856 0.005200495 0.8868687 0.06511906   0.04253302
## 3 0.0002396751 0.004617295 0.8920337 0.06273218   0.04037713
## 4 0.0005034786 0.009711410 0.8460274 0.06941765   0.07434005
## 5 0.0003666402 0.006895700 0.8710423 0.06748915   0.05420623
## 6 0.0010333865 0.016934008 0.8074549 0.08206146   0.09251625
```

```
##        LD1         LD2          LD3        LD4
## 1 1.667897  0.19143536  0.018276091 -0.9145021
## 2 1.674682  0.21767644 -0.004697164 -0.9521402
## 3 1.717001  0.24040409 -0.006178481 -0.9702461
## 4 1.496120 -0.20290196  0.329866360 -0.3770080
## 5 1.593218  0.03989067  0.133586994 -0.7113763
## 6 1.289464 -0.29898626  0.320986666 -0.3112614
```

```
## [1] 0.7464789
```



Above, you can see the predicted colors, the posterior probabilities, and the linear discriminant for the first 6 observations, respectively. Below that, a graph to demonstrate separation within the new model. Despite the higher separation achieved in LD1 for the new model, it actually predicted far less accurately than the last model, from almost 89% to approximately 75% correct predictions. There is also visually much less clumping and separation in the graph comparing LD1 and LD2, indicating that this model likely contributes to much less separation overall.

Third LDA Model (Combining Response Variables)

Since the LDA puts an emphasis on the largest group (red stars), I am going to group two of the less common responses, yellow-white and white, to see if that improves the prediction power of the model.

```
## Call:
## lda(train3$StarColor2 ~ `Temperature (K)` + `Luminosity(L/Lo)` +
##     `Radius(R/Ro)` + `Absolute magnitude(Mv)` + `Star type`,
##     data = train3)
##
## Prior probabilities of groups:
##         Blue   Blue-White          Red Yellow/White
##    0.2071006    0.1893491    0.4852071    0.1183432
##
## Group means:
##              `Temperature (K)` `Luminosity(L/Lo)` `Radius(R/Ro)`
## Blue                 1.3154590          0.9877604    0.002657706
## Blue-White           0.6900872         -0.1700735   -0.107016113
## Red                 -0.7511193         -0.2123531    0.126389539
## Yellow/White        -0.2753764         -0.3347332   -0.212266694
##              `Absolute magnitude(Mv)` `Star type`
```

```
## Blue                        -0.6859274   0.6594456
## Blue-White                  -0.2694563   0.2921594
## Red                          0.2888854  -0.3919212
## Yellow/White                 0.1900120   0.1752957
##
## Coefficients of linear discriminants:
##                                    LD1        LD2       LD3
## `Temperature (K)`        -1.82315675 -0.3547881  0.534394
## `Luminosity(L/Lo)`       -0.06607148 -0.5354311 -1.277473
## `Radius(R/Ro)`            1.18581695 -0.4671854  0.492073
## `Absolute magnitude(Mv)` -2.15603561  2.9480187 -1.548624
## `Star type`              -3.11319118  3.5816975 -1.235813
##
## Proportion of trace:
##    LD1    LD2    LD3
## 0.9029 0.0781 0.0190
```

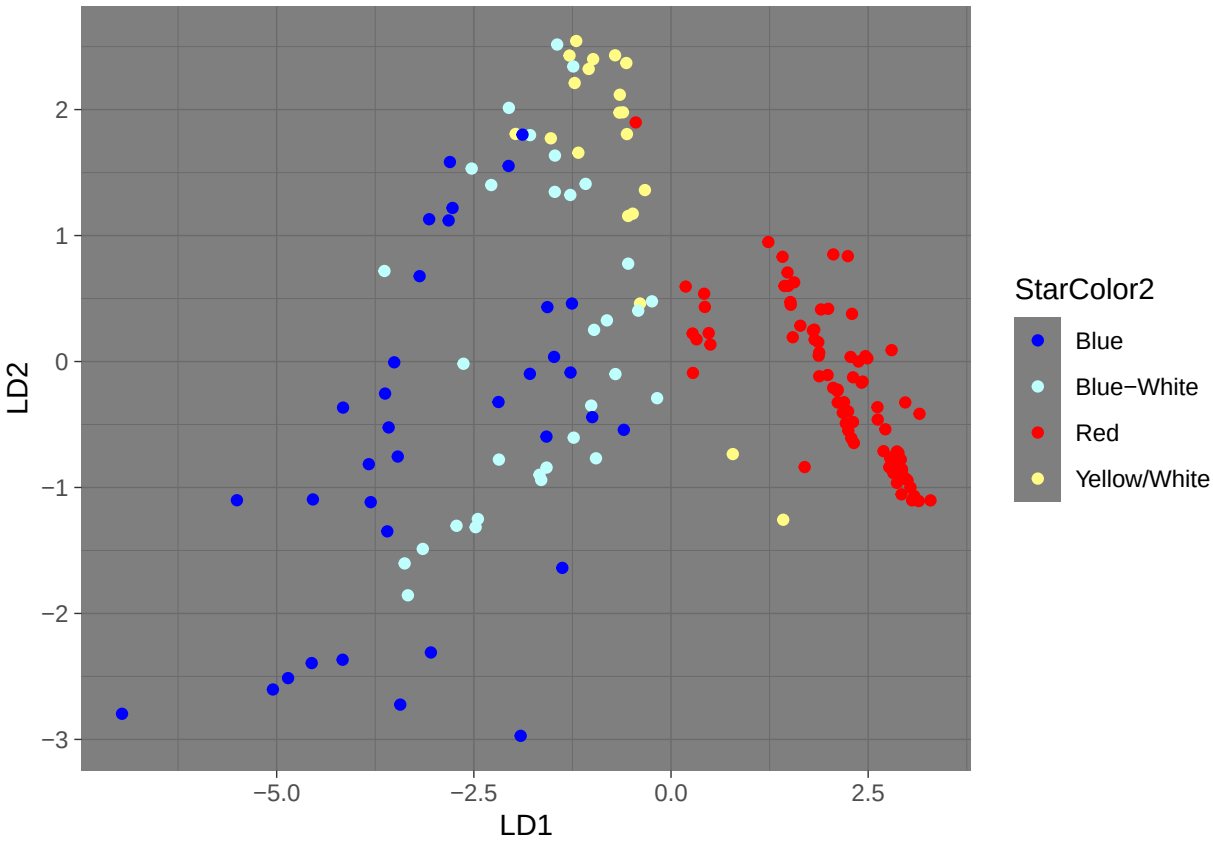
Since the data has one less response option, and new training and testing sets had to be made including the new response option "Yellow/White", the prior probabilities have changed. LD1 is now separating 90.29% of the data based upon low temperature, luminosity, and absolute magnitude, and high radius. LD2, on the other hand, is attributing to 7.81% of the separation based upon low temperature, luminosity, and radius, and high absolute magnitude and star type.

```
## [1] Red          Red          Red          Red          Red
## [6] Yellow/White
## Levels: Blue Blue-White Red Yellow/White
```

```
##        Blue Blue-White       Red Yellow/White
## 1 0.0000000  0.0000065 0.9999733    0.0000202
## 2 0.0000004  0.0001546 0.9991715    0.0006735
## 3 0.0000027  0.0007241 0.9953793    0.0038938
## 4 0.0000004  0.0001540 0.9994879    0.0003577
## 5 0.0000037  0.0009339 0.9929712    0.0060912
## 6 0.0499977  0.4035034 0.0005061    0.5459928
```

```
##         LD1        LD2         LD3
## 1  2.963562 -0.9666484  0.13668854
## 2  2.205508 -0.2077665  0.10104047
## 3  1.810388  0.1270231 -0.04964173
## 4  2.208409 -0.5216246  0.30403351
## 5  1.744372  0.2613118 -0.12559760
## 6 -1.679065  1.6265589 -0.10860909
```

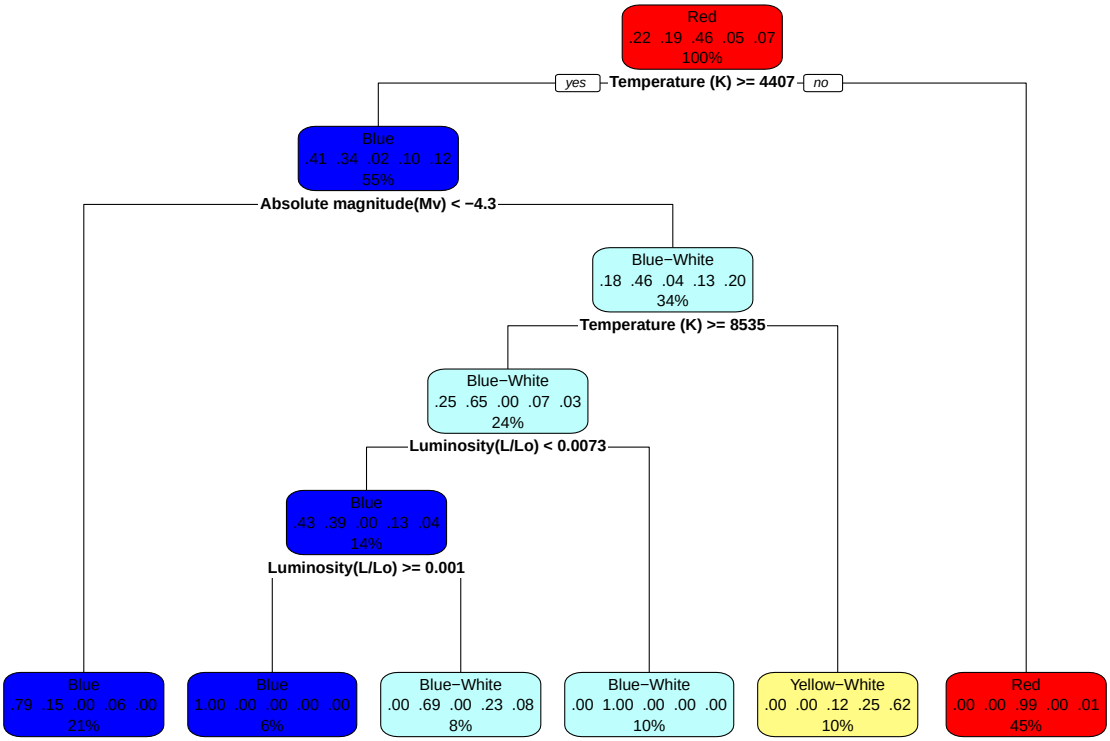
```
## [1] 0.7746479
```



Again, you can see the first 6 predictions, posterior probabilities, and linear discriminants. As with the second model, though, the percentage of correct predictions has dropped significantly, this time to 77.46%. A have arrived at a similar conclusion; I will be keeping the original model that predicted with 88.7% accuracy.

Decision Tree

I thought that it would be interesting to include a decision tree, considering the nature of the research. Below is a decision tree created based on a new train dataset that includes 70% of the data before it was scaled:



I will now use the created decision tree to predict star color within the test set. I have constructed a confusion matrix to examine those predictions as follows:

```
## Confusion Matrix and Statistics
##
##                Reference
## Prediction     Blue Blue-White Red White Yellow-White
##   Blue            0          0   0     0            0
##   Blue-White      0          0   0     0            0
##   Red            10         15  39     2            5
##   White           0          0   0     0            0
##   Yellow-White    0          0   0     0            0
##
## Overall Statistics
##
##                Accuracy : 0.5493
##                  95% CI : (0.4266, 0.6677)
##     No Information Rate : 0.5493
##     P-Value [Acc > NIR] : 0.5489
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Blue Class: Blue-White Class: Red Class: White
## Sensitivity               0.0000            0.0000     1.0000      0.00000
## Specificity               1.0000            1.0000     0.0000      1.00000
## Pos Pred Value               NaN               NaN     0.5493          NaN
## Neg Pred Value            0.8592            0.7887        NaN      0.97183
```

```
## Prevalence                0.1408           0.2113     0.5493      0.02817
## Detection Rate            0.0000           0.0000     0.5493      0.00000
## Detection Prevalence      0.0000           0.0000     1.0000      0.00000
## Balanced Accuracy         0.5000           0.5000     0.5000      0.50000
##                      Class: Yellow-White
## Sensitivity                      0.00000
## Specificity                      1.00000
## Pos Pred Value                       NaN
## Neg Pred Value                   0.92958
## Prevalence                       0.07042
## Detection Rate                   0.00000
## Detection Prevalence             0.00000
## Balanced Accuracy                0.50000
```

```
##               predict_tree
##                Blue Blue-White Red White Yellow-White
##   Blue            0          0  10     0            0
##   Blue-White      0          0  15     0            0
##   Red             0          0  39     0            0
##   White           0          0   2     0            0
##   Yellow-White    0          0   5     0            0
```

As you can see, the decision tree correctly predicted star color in the test set 54.93% of the time. As with the LDA, it prioritized getting the largest group of responses correct, in this case being the red stars.

Conclusions

I have concluded that I should reject using the newer LDA model I have created, in favor of the more accurate model I originally created. In summation, I have modeled a linear discriminant analysis that predicts star color based on its Temperature, Relative Luminosity, Solar Radius, Absolute Magnitude, and star type with 88.7% accuracy. In support of this, I have also modeled a short decision tree, which predicts star color within the same dataset, using the same variables, with an accuracy of 54.93%.

As I, the author, am not an astronomer, there were a few limitations in my research. Primarily because I don't completely understand all of the measurements, It has been difficult for me to identify possible problems with correlation, or variables that predict each other, for example. I believe that I have done a good job in making these decisions using statistical analysis, however a larger, more diverse sample size would have likely fostered better results.

This study was mostly for fun, because I am interested in the workings of our universe. The study could be furthered, though, using different variables related to the stars, or possibly using color to predict other aspects of stars.