StarsCapstone

Saketha Male

9/4/2021

Introduction

A star is a massive ball of plasma that emits light throughout the universe. There are billions upon billions of stars throughout our galaxy and exponentially more in the billions of galaxies in the universe. A star can be defined by many characteristics like Temperature, Radius, Luminosity, Magnitude, Spectral Class, Type, and Color.

Using data science, we can better analyze information on stars and create algorithms to discern patters that can help us classify more stars in the future.

Methods/Exploratory Analysis

In This report, we will begin exploring the stars data set provided by user deepu1109 on kaggle. This is the download link for the data set: https://www.kaggle.com/deepu1109/star-dataset/download.

- Cleaning and Tidying Data

This was a fairly simple process. The tidying of the data consisted mainly of importing the data using the read.csv function and then renaming the column names for conciseness.

However this was a bit of a complex process since the zip and csv link weren't given directly in kaggle. A plan B was given in the case the code failed to execute because the download link was made inaccessible.

- Inferencing and Modeling

The dimensions and different aspects of the stars dataset are first introduced and made familiar through fragments of code.

The data analysis was depicted and supported with use of graphs, plots, and other visual features.

- Exploratory Analysis

Exploring the relationships between variables in the dataset through visuals such as graphs/plots helped outline the conclusions.

- Accuracy and Predictions

We made using a k-nearest neighbors algorithm. The first step to optimize this model was to use the "tune grid" function in order to get the highest accuracy from various k-values.

We also seperated the stars data set into a training and testing sets, which each contained different observations.

Overview of the Stars dataset

The columns of this dataset include the following features of stars:

- Absolute Temperature (Kelvin)
- Relative Luminosity (L/Lo)
- Relative Radius (R/Ro)
- Absolute Magnitude (Megavolt)
- Star Type:
  - 0 = Brown Dwarf
  - 1 = Red Dwarf
  - 2 = White Dwarf
  - 3 = Main Sequence
  - 4 = Super Giants
  - 5 = Hyper Giants
- Star Color
- Spectral Class (M, B, A, F, O, K, G)
- Lo = 3.828 * 10^26 Watts (Average Luminosity Of Sun)
- Ro = 6.9551 * 10^8 Meters (Average Radius Of Sun)

Dimensions and Overview of the Dataset

```
# Number of Rows
nrow(stars)
```

```
## [1] 240
```

```
# Number of Columns
ncol(stars)
```

```
## [1] 7
```

```
head(stars)
```

```
##   Temperature Luminosity Radius Magnitude Type Color Class
## 1        3068   0.002400 0.1700     16.12    0   Red     M
## 2        3042   0.000500 0.1542     16.60    0   Red     M
## 3        2600   0.000300 0.1020     18.70    0   Red     M
## 4        2800   0.000200 0.1600     16.65    0   Red     M
## 5        1939   0.000138 0.1030     20.06    0   Red     M
## 6        2840   0.000650 0.1100     16.98    0   Red     M
```

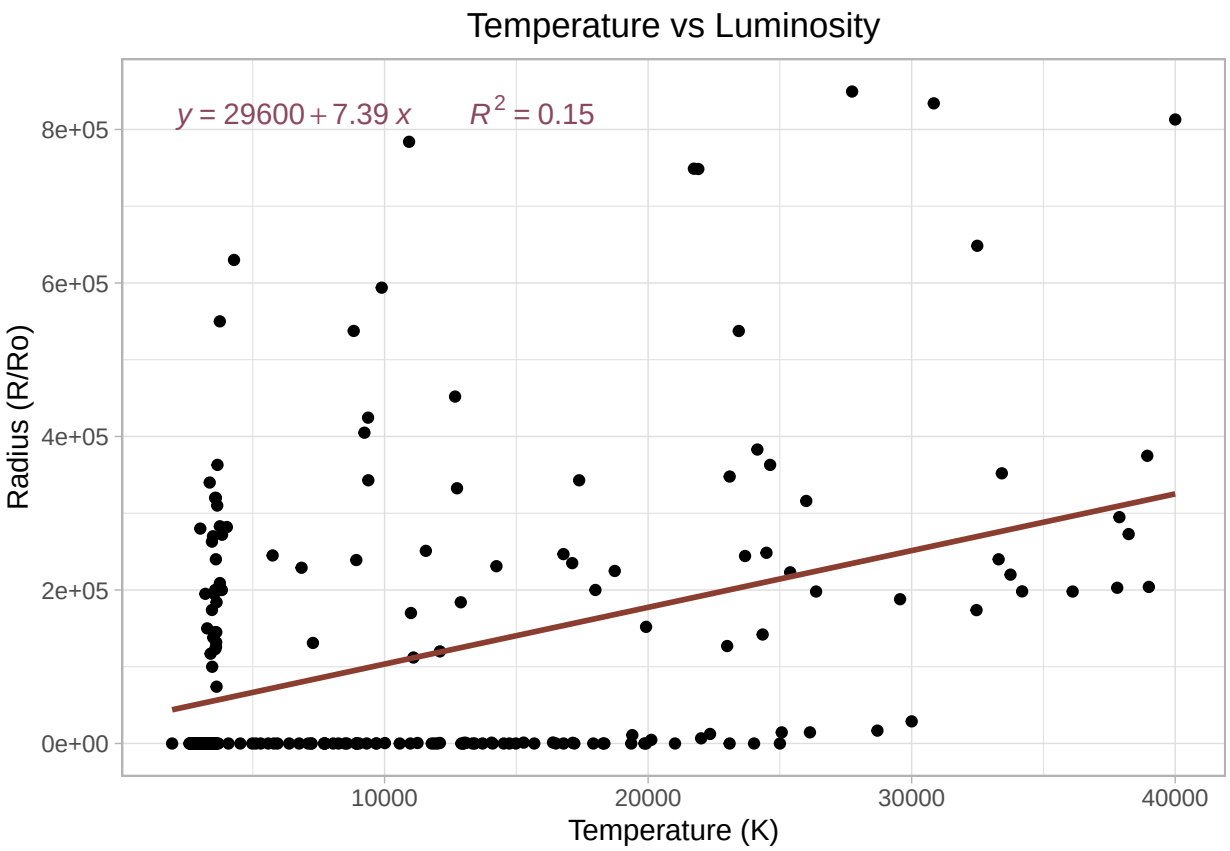
```
summary(stars)
```

```
##   Temperature      Luminosity           Radius           Magnitude
##  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
##       Type        Color              Class
##  Min.   :0.0   Length:240         Length:240
##  1st Qu.:1.0   Class :character   Class :character
##  Median :2.5   Mode  :character   Mode  :character
##  Mean   :2.5
##  3rd Qu.:4.0
##  Max.   :5.0
```

Data Exploration and Visualization

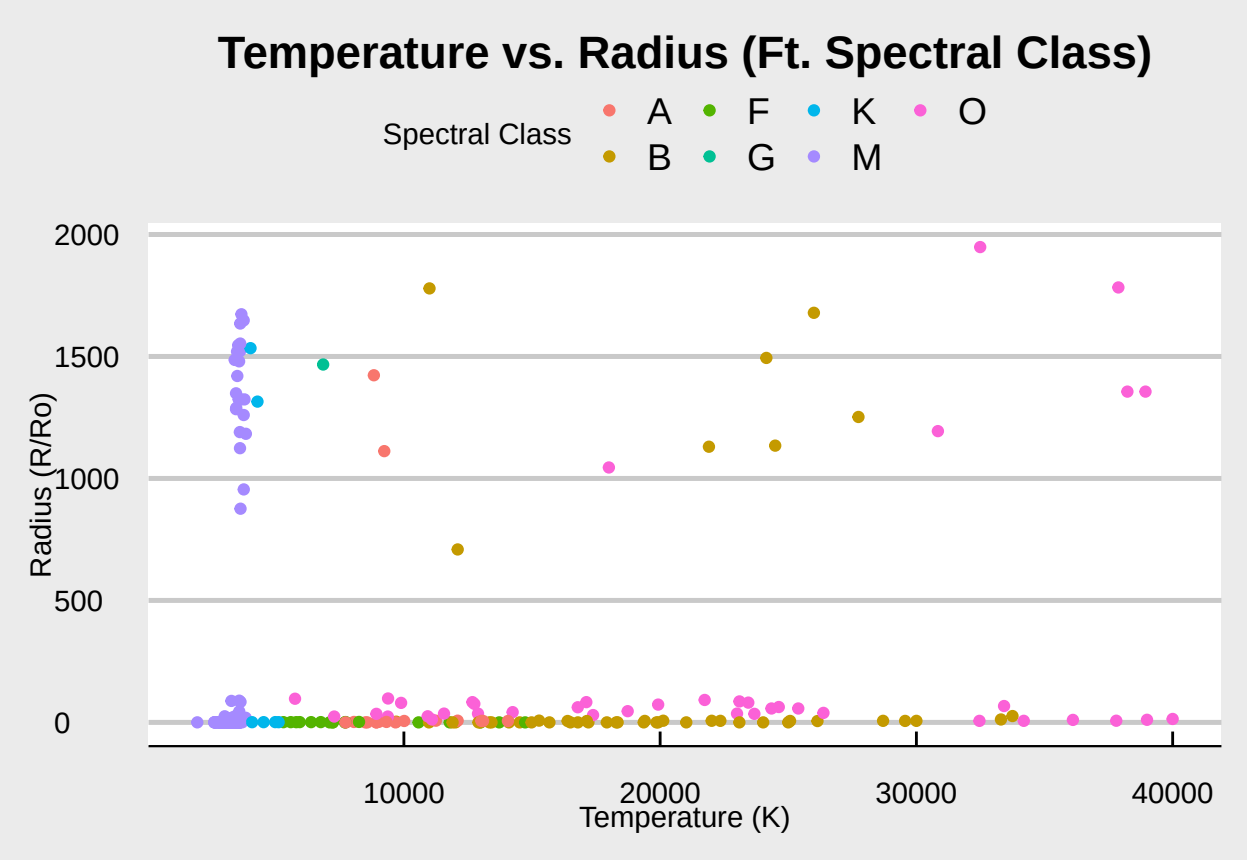
Now lets begin exploring relationships among variables in the dataset.

Temp vs. Radius



We see that there isnt a very good correlation, evidenced by the R^2 value. But maybe if we add a factor/variable like spectral class, there may be a relationship among the varaibles.

Temp vs. Radius (Spectral Class)

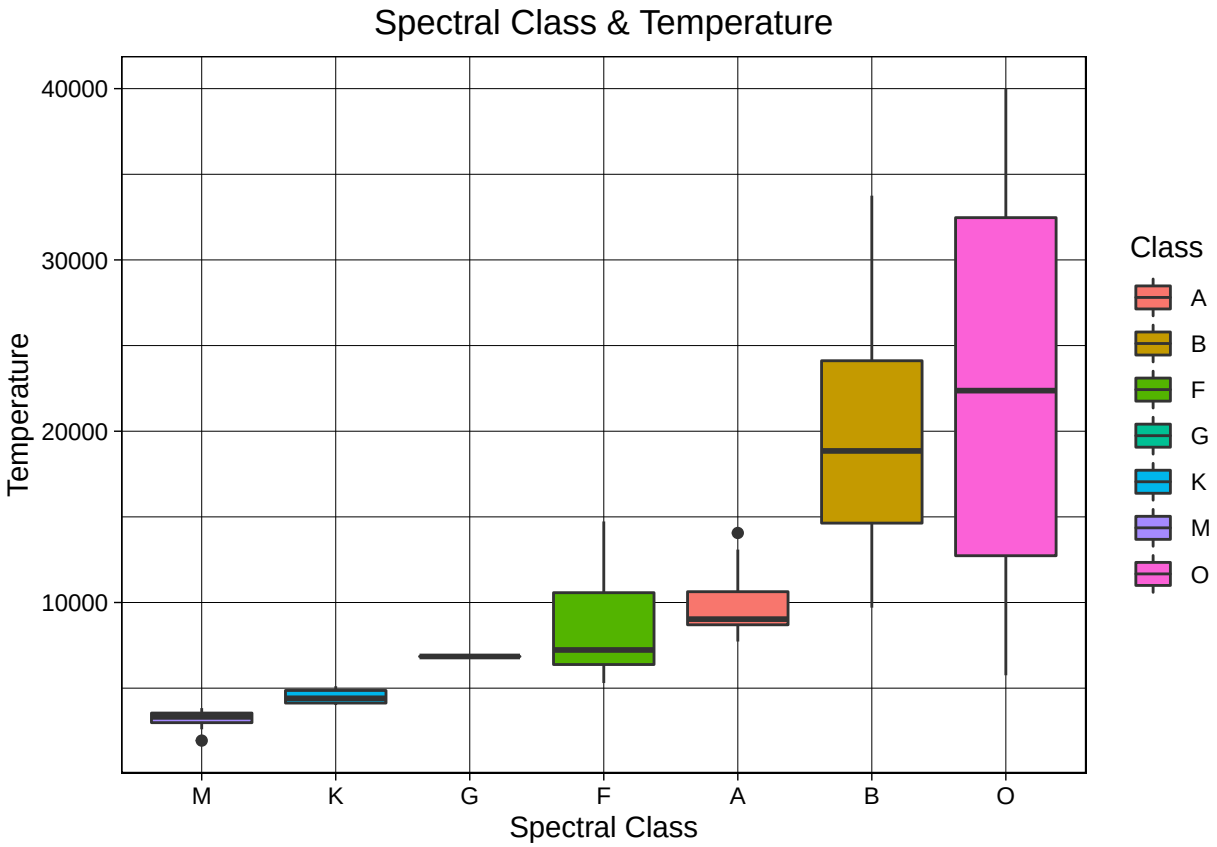


Although the stars in spectral classes O and B demonstrate fluctuating temperatures and values, many correlations are demonstrated in this plot:

- Stars in spectral class M tend to show the lowest temperatures.
- Stars in spectral class K tend to show the 2nd lowest temperatures.
- Stars in spectral classes G, F, and A tend to show temperatures between approximately 5000 to 15000 Kelvin.
- We see that stars in spectral class O don't have a very strong correlation with temperature as the values vary strongly. However, regarding radius, stars of class O tend to have very low radii.
- Overall, stars in spectral classes F and A have the lowest radii.

To demonstrate this more accurately and clearly, we will start by making boxplots for each class in relation to its temperature, radius, magnitude, etc.

Spectral Class & Temperature (Box Plot)

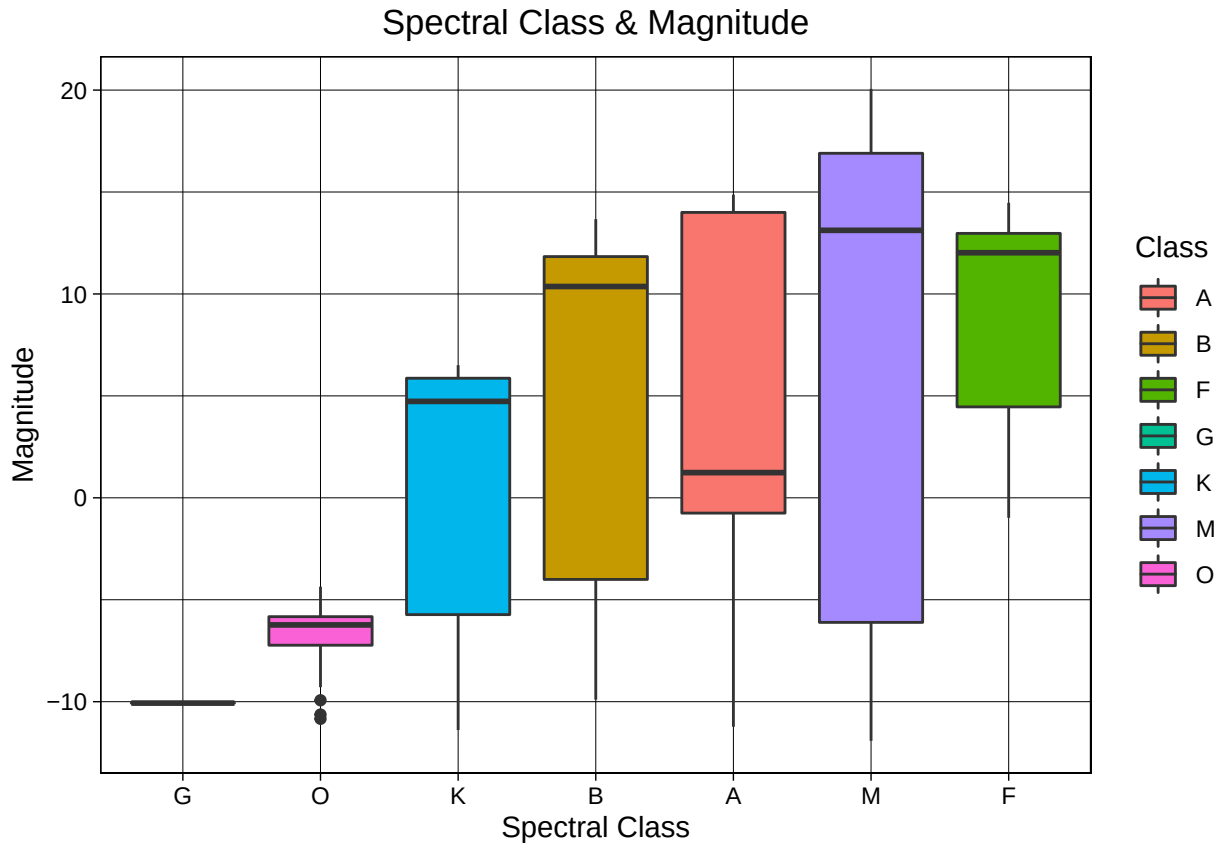


We see that though stars in spectral class O have a higher mean, they also have a higher range, meaning more variability. This also applies to stars in spectral class B.

Stars in spectral classes M, K, and A have lower variability. Spectral Class M has the lowest median.

Note: (Plotting Spectral Class vs. Radius and Spectral Class vs Luminosity didn't have good correlations and weren't good predictors.)

Spectral Class & Magnitude



We see that there is a lot of variability within a stars spectral class when measured for magnitude.

Temperature, Luminosity, Radius, and Magnitude Summary for each Spectral Class

Spectral Class A

```
##   Temperature      Luminosity           Radius            Magnitude
##  Min.   : 7723   Min.   :     0.0   Min.   :   0.0088   Min.   :-11.230
##  1st Qu.: 8700   1st Qu.:     0.0   1st Qu.:   0.0104   1st Qu.: -0.750
##  Median : 9030   Median :    38.0   Median :   2.4870   Median :  1.236
##  Mean   : 9842   Mean   : 49860.2   Mean   : 135.8784   Mean   :  4.085
##  3rd Qu.:10631   3rd Qu.:   738.5   3rd Qu.:   6.1010   3rd Qu.: 14.000
##  Max.   :14060   Max.   :537493.0   Max.   :1423.0000   Max.   : 14.870
```

Spectral Class B

```
##   Temperature      Luminosity         Radius            Magnitude
##  Min.   : 9700   Min.   :     0   Min.   :   0.0084   Min.   :-9.900
##  1st Qu.:14636   1st Qu.:     0   1st Qu.:   0.0104   1st Qu.:-4.003
##  Median :18850   Median :     0   Median :   0.0146   Median :10.365
```

```
##  Mean   :19574   Mean   : 78179   Mean   : 202.0223   Mean   : 3.723
##  3rd Qu.:24114   3rd Qu.: 16222   3rd Qu.:   6.5800   3rd Qu.:11.832
##  Max.   :33750   Max.   :849420   Max.   :1779.0000   Max.   :13.670
```

Spectral Class F

```
##   Temperature      Luminosity          Radius          Magnitude
##  Min.   : 5300   Min.   :0.00008   Min.   :0.00892   Min.   :-0.980
##  1st Qu.: 6380   1st Qu.:0.00014   1st Qu.:0.01100   1st Qu.: 4.460
##  Median : 7230   Median :0.00029   Median :0.01300   Median :12.020
##  Mean   : 8517   Mean   :1.38396   Mean   :0.55133   Mean   : 8.612
##  3rd Qu.:10574   3rd Qu.:1.35000   3rd Qu.:0.99000   3rd Qu.:12.970
##  Max.   :14732   Max.   :9.25000   Max.   :1.93000   Max.   :14.470
```

Spectral Class G

```
##   Temperature     Luminosity         Radius       Magnitude
##  Min.   :6850   Min.   :229000   Min.   :1467   Min.   :-10.07
##  1st Qu.:6850   1st Qu.:229000   1st Qu.:1467   1st Qu.:-10.07
##  Median :6850   Median :229000   Median :1467   Median :-10.07
##  Mean   :6850   Mean   :229000   Mean   :1467   Mean   :-10.07
##  3rd Qu.:6850   3rd Qu.:229000   3rd Qu.:1467   3rd Qu.:-10.07
##  Max.   :6850   Max.   :229000   Max.   :1467   Max.   :-10.07
```

Spectral Class K

```
##   Temperature     Luminosity           Radius            Magnitude
##  Min.   :4015   Min.   :     0.1   Min.   :   0.7950   Min.   :-11.3900
##  1st Qu.:4130   1st Qu.:     0.2   1st Qu.:   0.8678   1st Qu.: -5.7300
##  Median :4406   Median :     0.5   Median :   1.0030   Median :  4.7300
##  Mean   :4500   Mean   :152000.2   Mean   : 475.4443   Mean   :  0.2673
##  3rd Qu.:4866   3rd Qu.:211500.2   3rd Qu.: 986.5325   3rd Qu.:  5.8660
##  Max.   :5112   Max.   :630000.0   Max.   :1534.0000   Max.   :  6.5060
```

Spectral Class M

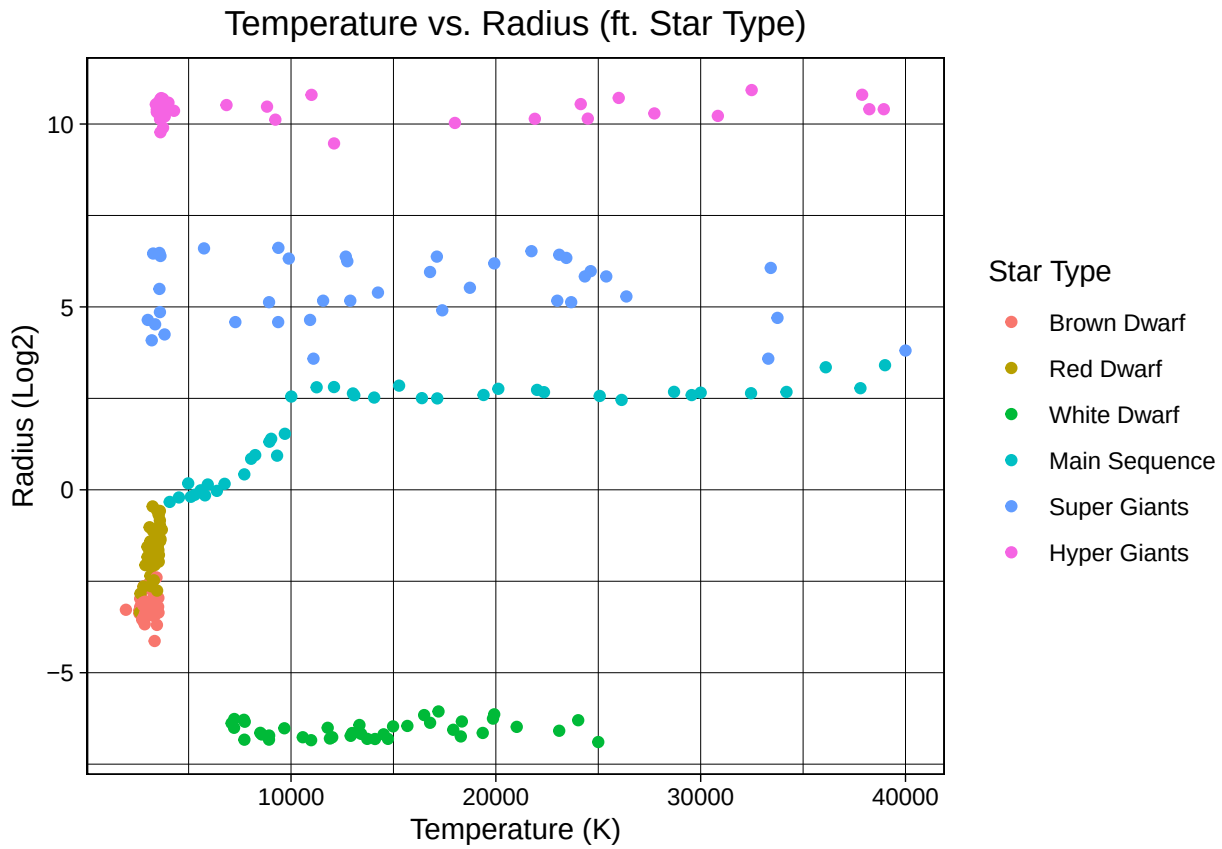
```
##   Temperature     Luminosity         Radius           Magnitude
##  Min.   :1939   Min.   :     0   Min.   :   0.057   Min.   :-11.920
##  1st Qu.:2986   1st Qu.:     0   1st Qu.:   0.117   1st Qu.: -6.110
##  Median :3324   Median :     0   Median :   0.291   Median : 13.120
##  Mean   :3257   Mean   : 61423   Mean   : 273.895   Mean   :  8.368
##  3rd Qu.:3546   3rd Qu.:120000   3rd Qu.:  24.000   3rd Qu.: 16.900
##  Max.   :3834   Max.   :550000   Max.   :1673.000   Max.   : 20.060
```

Spectral Class O

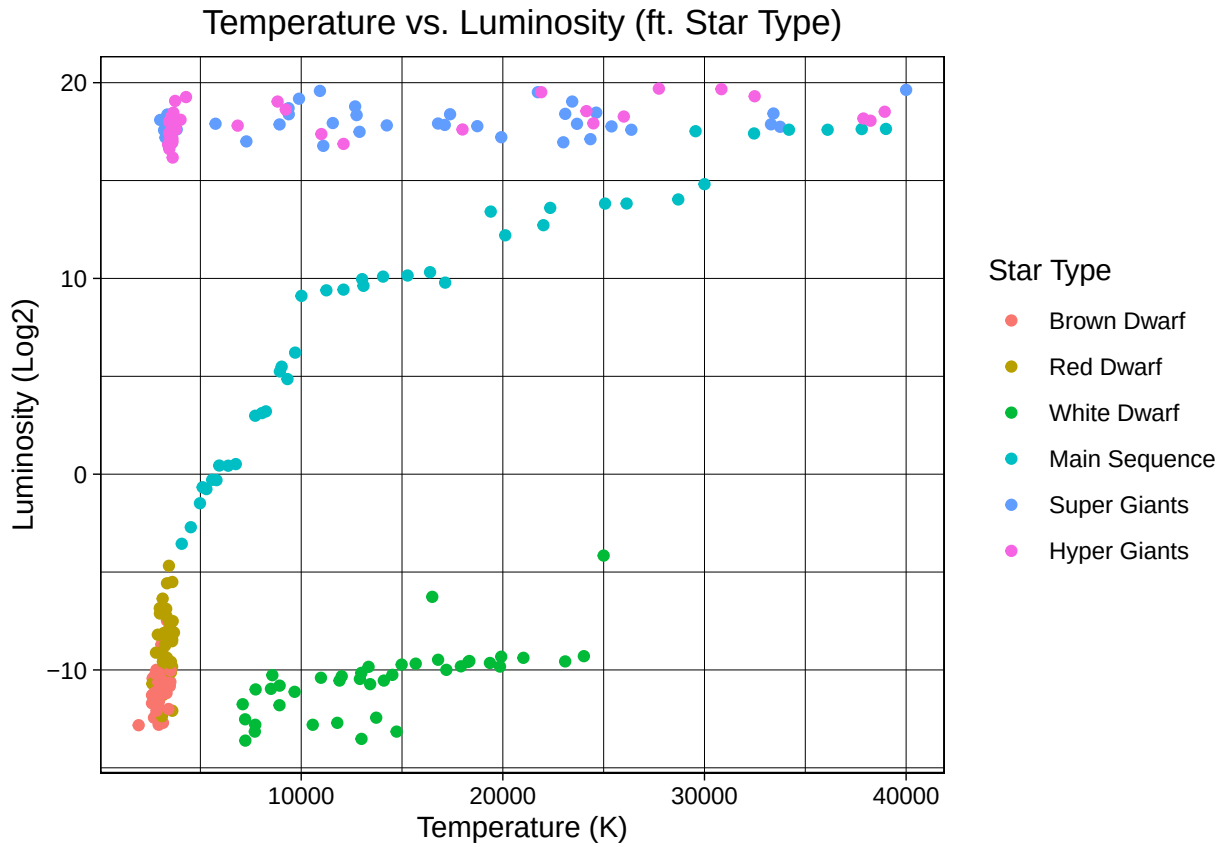
```
##   Temperature      Luminosity         Radius           Magnitude
##  Min.   : 5752   Min.   :112000   Min.   :   6.237   Min.   :-10.840
##  1st Qu.:12730   1st Qu.:199550   1st Qu.:  28.750   1st Qu.: -7.231
##  Median :22369   Median :245865   Median :  57.000   Median : -6.235
##  Mean   :22294   Mean   :330565   Mean   : 257.795   Mean   : -6.596
##  3rd Qu.:32467   3rd Qu.:365958   3rd Qu.:  83.750   3rd Qu.: -5.830
##  Max.   :40000   Max.   :834042   Max.   :1948.500   Max.   : -4.360
```

More Visualization and Models

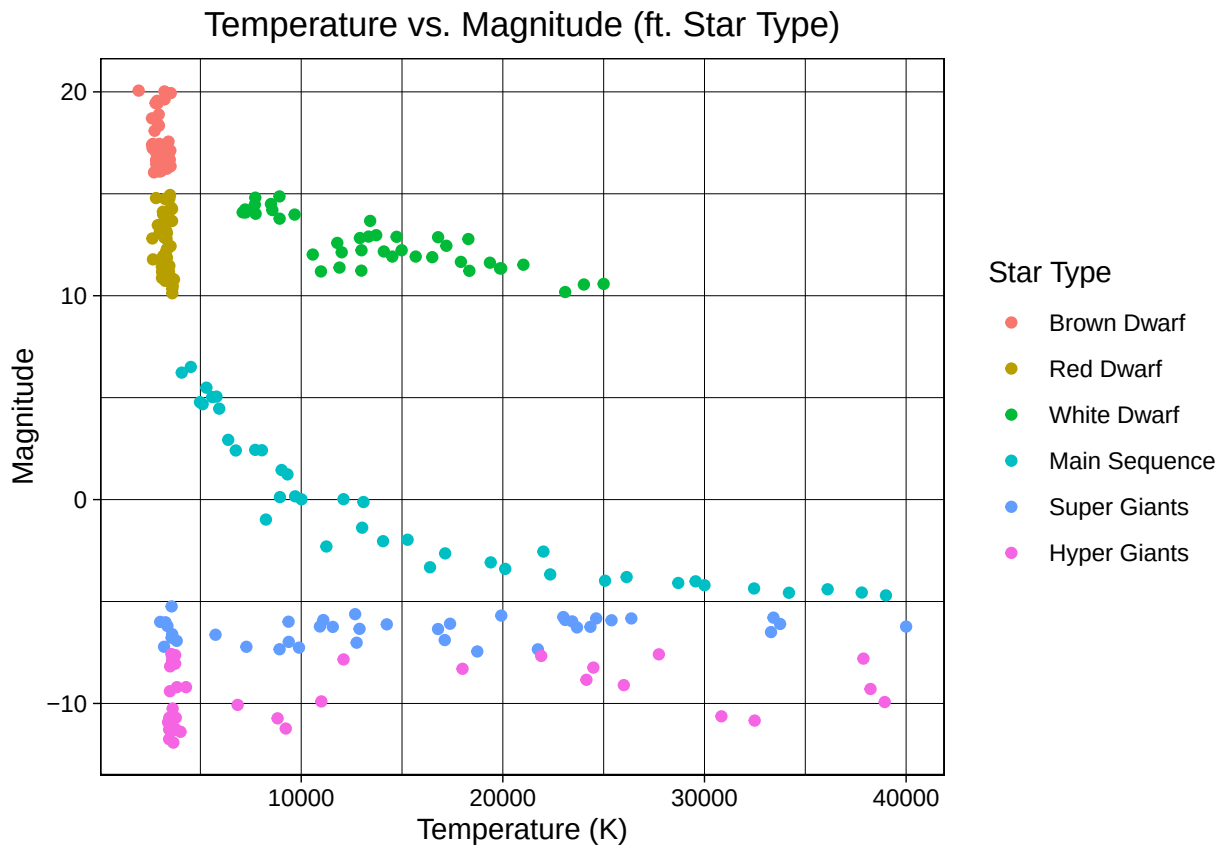
Temp vs Radius (Star Type)



Temp vs Luminosity (Star Type)



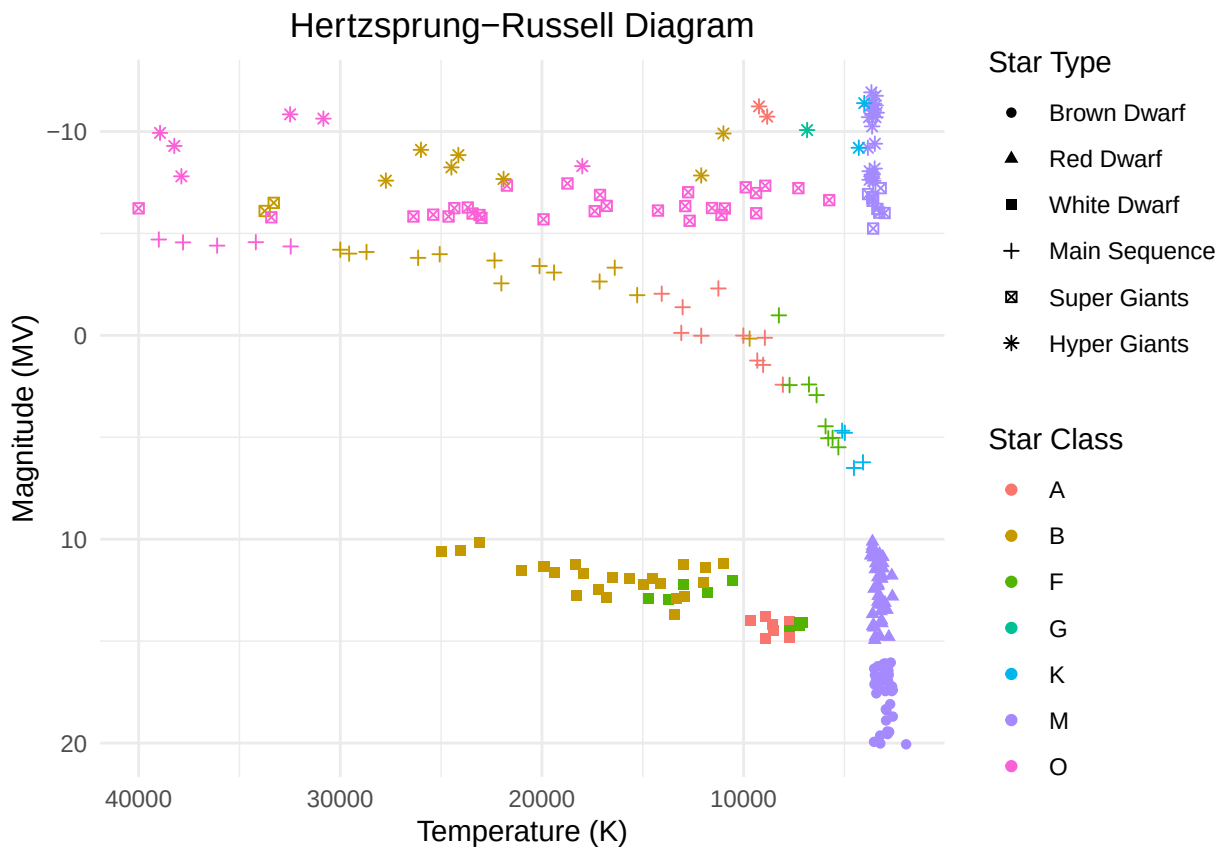
Temp vs Magnitude (Star Type)



Hertzprung-Russell Diagram

The purpose of making graphs and diagrams is to prove that stars follow a certain order and have certain relationships in the Celestial Space.

Specifically, the Hertzsprung-Russell Diagram or HR-Diagram classifies stars by plotting its features based on that graph.

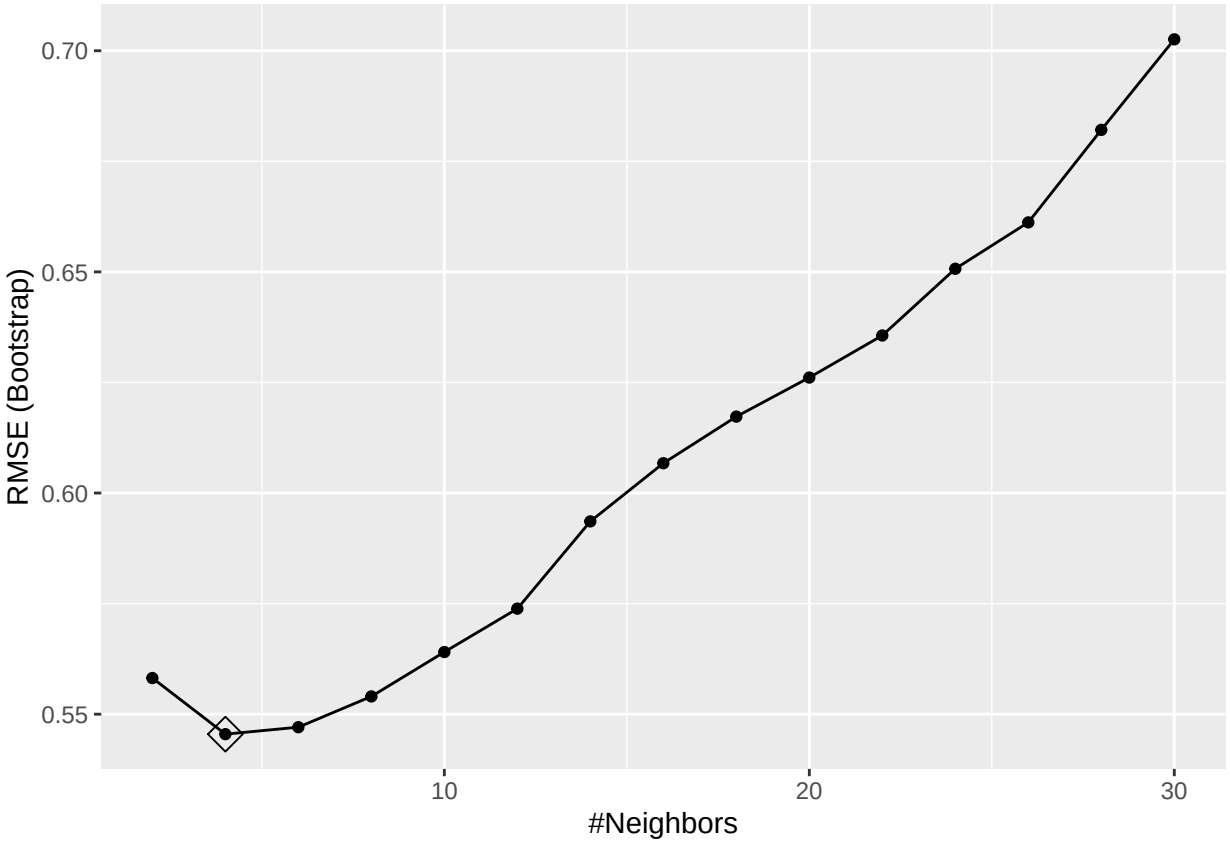


This graph really enables us to see all the correlations and relationships among many different variables.

Modeling and Predicting (Results)

KNN Plot

```
library(nnet)

trainingSamp <- stars$Type %>% createDataPartition(p = 0.7, list = FALSE)

trainData  <- stars[trainingSamp, ]
testData <- stars[-trainingSamp, ]

# We train a k-nearest neighbor algorithm with a tunegrid parameter to optimize for k
train_knn <- train(Type ~ ., method = "knn", data = trainData,tuneGrid = data.frame(k = seq(2, 30, 2)))

# Visualize and save the optimal value for k

knnplot <- ggplot(train_knn, highlight = TRUE)
knnplot
```




```
optim_knn <- train_knn$bestTune[1, 1]
optim_knn
```

```
## [1] 4
```

This graph displays the optimized value for k in relation to accuracy. The value k = 4 is thus chosen to calculate the results for this algorithm. It isn' a very good relation/predictor because of the low RMSE.

Predictions

Multinomial logistic regression model to predict star type:

```
# Fitting model

model <- nnet::multinom(Type ~ Temperature + Magnitude + Radius + Luminosity, data = trainData)
```

```
## # weights:  36 (25 variable)
## initial  value 301.015591
## iter  10 value 237.361879
## iter  20 value 162.346954
## iter  30 value 28.881393
## iter  40 value 5.166452
## iter  50 value 0.016409
## iter  60 value 0.002482
## iter  70 value 0.001247
```

```
## iter  80 value 0.000810
## iter  90 value 0.000769
## iter 100 value 0.000598
## final  value 0.000598
## stopped after 100 iterations
```

```
# Summarizing model

summary(model)
```

```
## Call:
## nnet::multinom(formula = Type ~ Temperature + Magnitude + Radius +
##     Luminosity, data = trainData)
##
## Coefficients:
##   (Intercept) Temperature Magnitude      Radius    Luminosity
## 1   198.94572 0.005312097 -13.90191  0.01211816 -0.0004881757
## 2   -75.50171 0.023822109  -1.33431 -0.03324736 -0.0019396741
## 3   205.81185 0.013876825 -18.75640 -0.27160794 -0.0016165739
## 4   259.74770 0.005040610 -23.83548 -0.15592734 -0.0005526738
## 5   161.18963 0.006958565 -15.72291  0.06996110 -0.0005027724
##
## Std. Errors:
##    (Intercept) Temperature    Magnitude       Radius   Luminosity
## 1 1.677510e-05  0.01428018 6.176874e-04 1.944734e-05 4.045830e-05
## 2 5.183700e-05  0.18202425 8.596847e-04 6.356920e-06 4.261484e-08
## 3 2.477931e-05  0.05108910 2.523352e-04 9.990427e-06 5.756210e-01
## 4 4.334623e-05  0.15451561 2.499853e-04 9.947822e-04 5.751460e-01
## 5 3.765322e-06  0.10916178 2.584994e-05 1.085129e-03 2.031616e+00
##
## Residual Deviance: 0.001195769
## AIC: 50.0012
```

```
# Predicting Class

classPredict <- model %>% predict(testData)
head(classPredict)
```

```
## [1] 0 0 0 0 1 1
## Levels: 0 1 2 3 4 5
```

```
# Accuracy of Model

mean(classPredict == testData$Type)
```

```
## [1] 0.9861111
```

This model has a very high accuracy and is thus a good predictor of star type.

Concluding Remarks

In this report, we have profoundly explored the dataset and identified relationships among variables like temperature, radius, magnitude, luminosity, star type, spectral class, and etc. We used box plots, scatterplots, and even made a Hertzprung-Russell Diagram using the data from the stars dataset.

The aim of this report was to help distinguish algorithims and relationships amon star features.