Lab_03T

36-290 – Statistical Research Methodology

Week 3 Tuesday – Fall 2021

Data

We'll begin by importing some data from the 36-290 GitHub site:

```
rm(list=ls())
file.path = "https://raw.githubusercontent.com/pefreeman/36-290/master/EXAMPLE_DATASETS/DRACO/draco_photometry.Rdat
a"
load(url(file.path))
df = data.frame(ra,dec,velocity.los,log.g,temperature,mag.g,mag.i)
rm(file.path,ra,dec,velocity.los,log.g,temperature,mag.u,mag.g,mag.r,mag.i,mag.z,metallicity,signal.noise)
objects()
```

```
## [1] "df"
```

If everything loaded correctly, you should see one variable in your global environment: `df`. `df` is a data frame with 2778 rows and 7 columns. See this README file (https://github.com/pefreeman/36-290/tree/master/EXAMPLE_DATASETS/DRACO) for a full description of the data and its variables. Note that I have removed `signal.noise`, `metallicity`, and three of the magnitudes from the data frame, to reduce the dimensionality and thus make analyses easier. To be clear: the data do not explicitly include a response variable. It's just a multidimensional set of data.

Exploratory Data Analysis

This lab will be different from most if not all of the others, in that I want you to bring the tools that you've learned to bear by performing an exploratory analysis on the Draco dataset.

There are no "right answers" in this lab. It is more that some answers may be better (or more complete or tell a fuller story) than others.

Some things that you want to keep in mind:

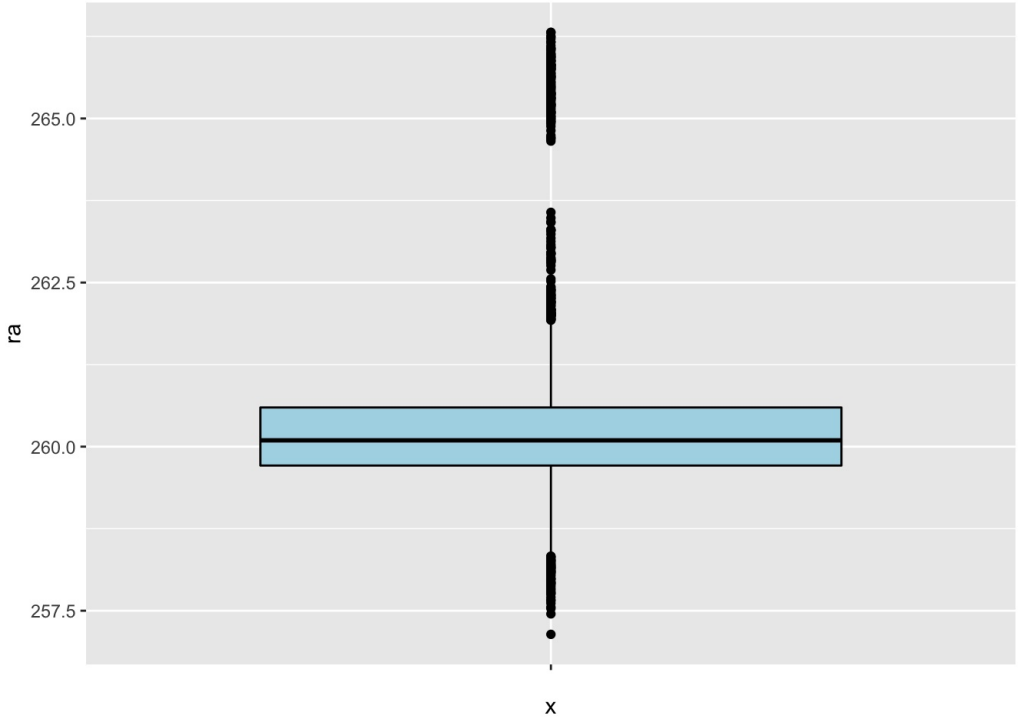
- When doing EDA, you are attempting to construct a story, not a laundry list. You do not need to create every possible plot using every possible combination of variables. If the two magnitudes are distributed similarly relative to `temperature`, say, it is sufficient to show one plot and mention how the other variable is not shown because the behavior is similar. Or something like that.
- Descriptions are good: are distributions unimodal or multimodal? Skew or symmetric? Are there outliers? (If there are outliers, perhaps use tools at your disposal to "zoom in"…see below.) Are two variables correlated? Linearly or is there non-linear dependence?
- Don't assume your first attempt at a plot will be your last attempt. Change limits. Change point sizes. Change labels and titles. Showing a histogram with all the bins smushed to the left because there is one outlier far to the right is no good! Change limits, change the number of bins.
- Faceting is good. It condenses things down for reports and posters.
- To learn how to do more than what you already know how to do with `ggplot`, see this set of notes on correlation plots, pairs plots, etc. (https://github.com/pefreeman/36-290/blob/master/LECTURES/Intro_ExtraViz.Rmd).
- Correlation plots are good. (See my last point above.) They indicate what subset of variables might be the ones to look at more closely, with pairs plots, etc.
- Variable transformations are good! If a distribution is unimodal but skew, explore whether, e.g., a square-root or logarithmic transformation might make the distribution more symmetric. We will talk "more officially" about variable transformations when we get to linear regression analyses. However, if you want to read a small write-up that discusses transformations a bit earlier, go here (https://github.com/pefreeman/36-290/blob/master/LECTURES/) and download `Variable_Transformations.pdf`.

```
summary(df)
```

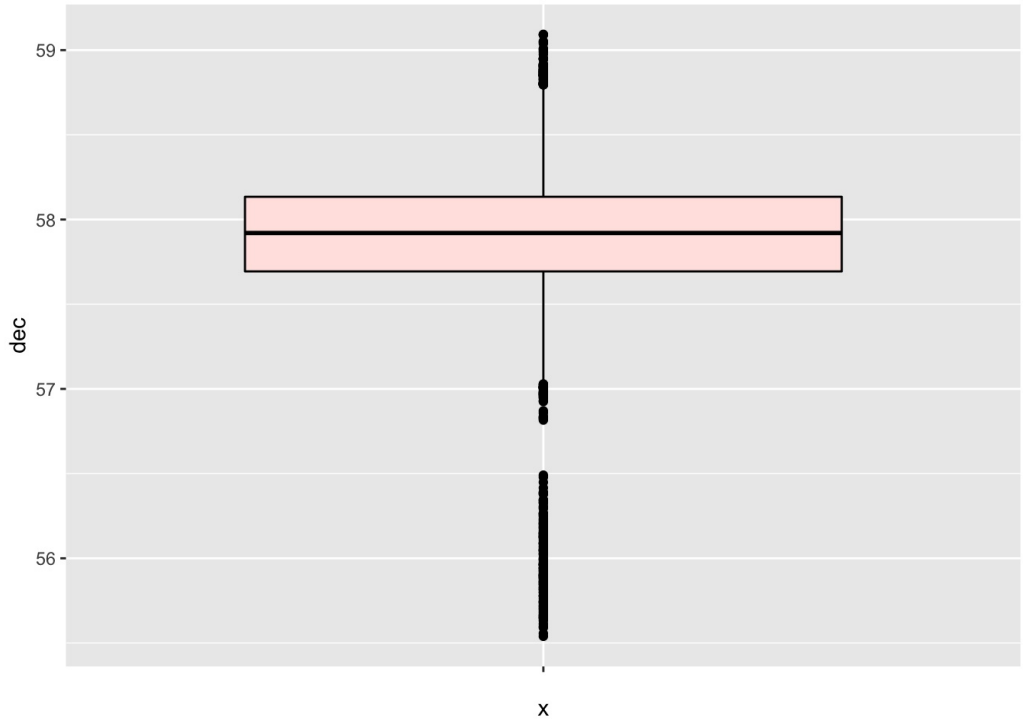
```
##        ra              dec         velocity.los         log.g       
##  Min.   :257.1   Min.   :55.54   Min.   :-496.70   Min.   :0.700  
##  1st Qu.:259.7   1st Qu.:57.69   1st Qu.:-290.90   1st Qu.:1.900  
##  Median :260.1   Median :57.92   Median :-193.45   Median :4.100  
##  Mean   :260.4   Mean   :57.87   Mean   :-175.94   Mean   :3.522  
##  3rd Qu.:260.6   3rd Qu.:58.13   3rd Qu.: -56.33   3rd Qu.:5.000  
##  Max.   :266.3   Max.   :59.09   Max.   : 474.80   Max.   :5.600  
##   temperature       mag.g           mag.i      
##  Min.   :4320   Min.   :15.65   Min.   :14.92  
##  1st Qu.:4789   1st Qu.:18.74   1st Qu.:17.60  
##  Median :4997   Median :19.41   Median :18.43  
##  Mean   :5080   Mean   :19.34   Mean   :18.35  
##  3rd Qu.:5238   3rd Qu.:20.15   3rd Qu.:19.36  
##  Max.   :7464   Max.   :21.46   Max.   :20.72
```

It seems that there is no missing data, and no outliers that are too out of the ordinary. The summary for mag.g and mag.i seem very similar, with all min/1Q/median/mean/3Q/max values within one or two values apart from each other.

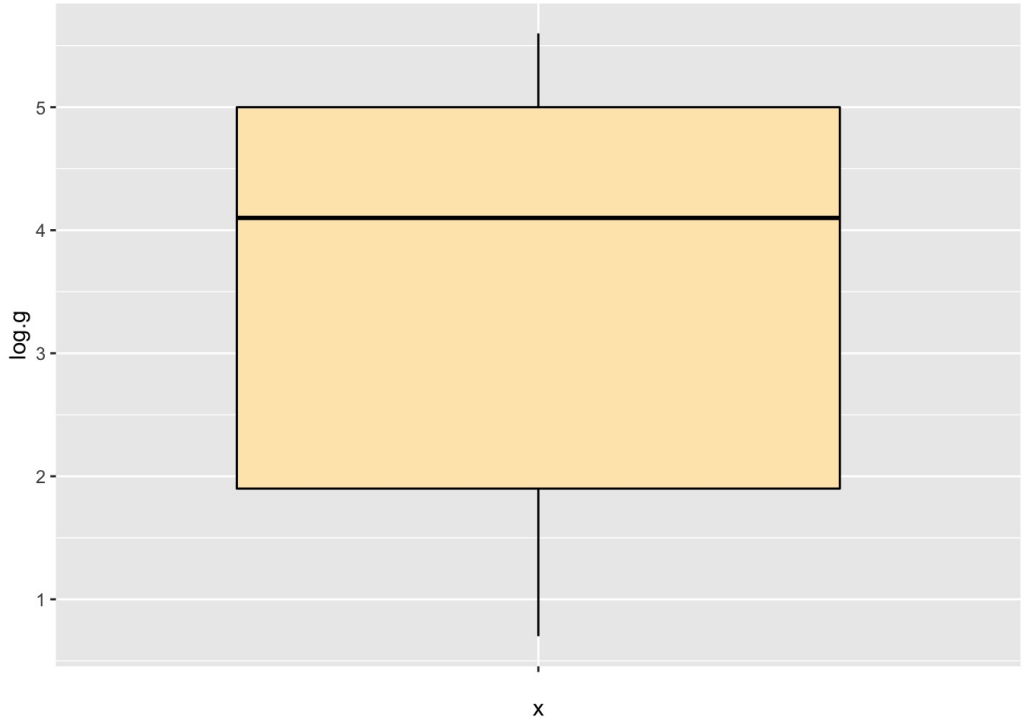
```
library(ggplot2)
ggplot(data=df,mapping=aes(x="",y=ra)) + geom_boxplot(color="black",fill="lightblue")
```



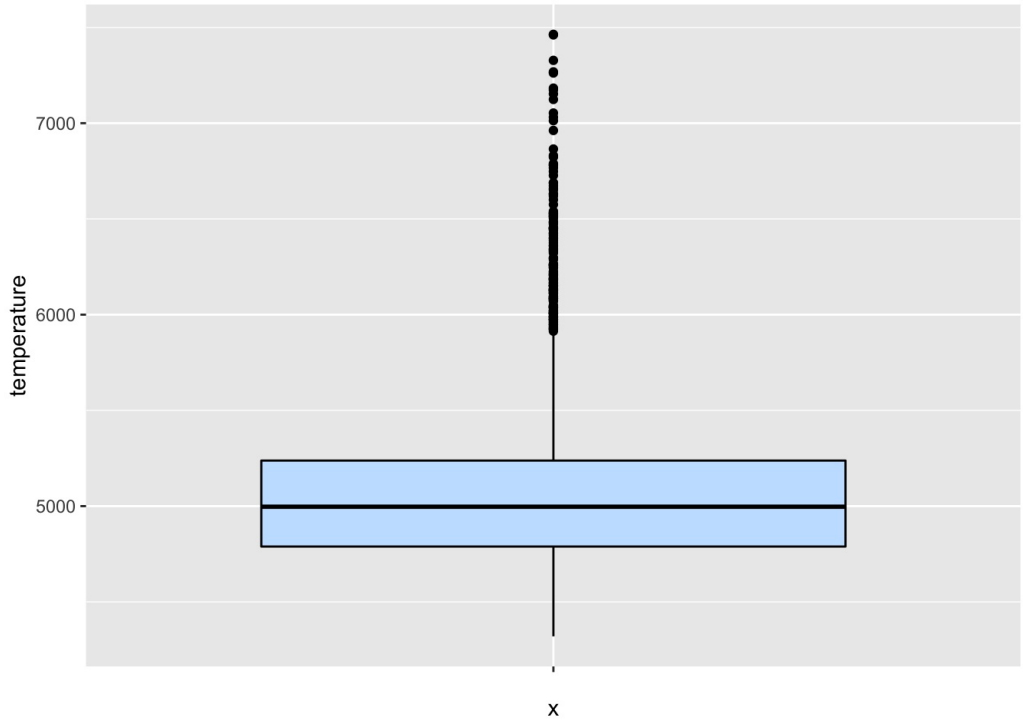

```
ggplot(data=df,mapping=aes(x="",y=dec)) + geom_boxplot(color="black",fill="mistyrose")
```



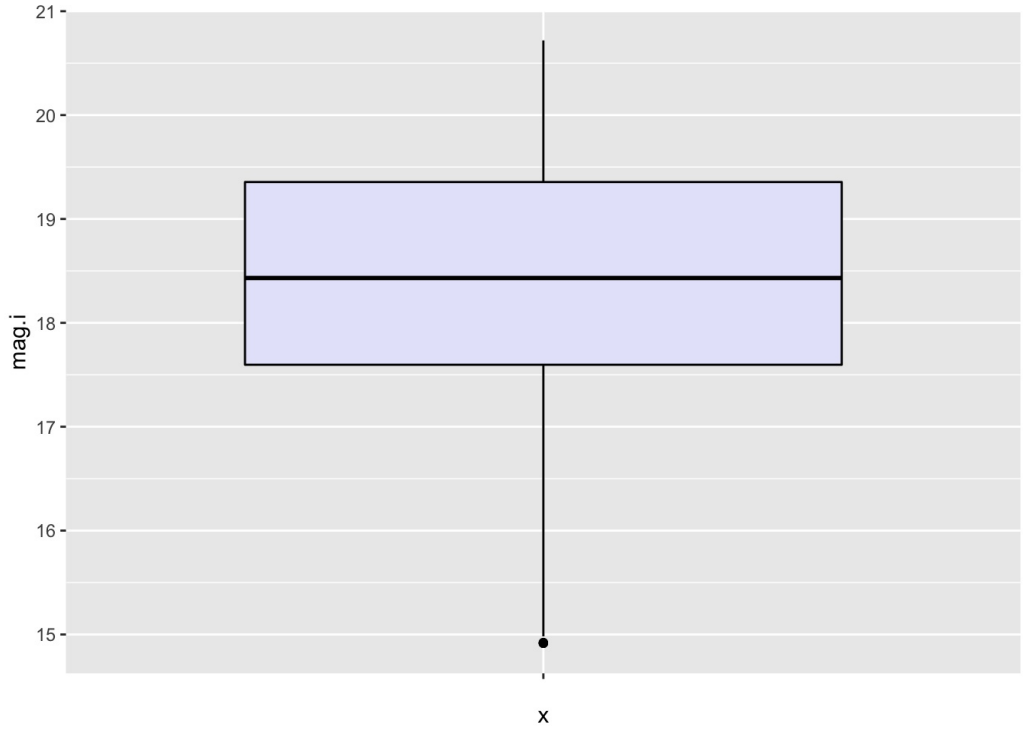

```
ggplot(data=df,mapping=aes(x="",y=log.g)) + geom_boxplot(color="black",fill="wheat1")
```



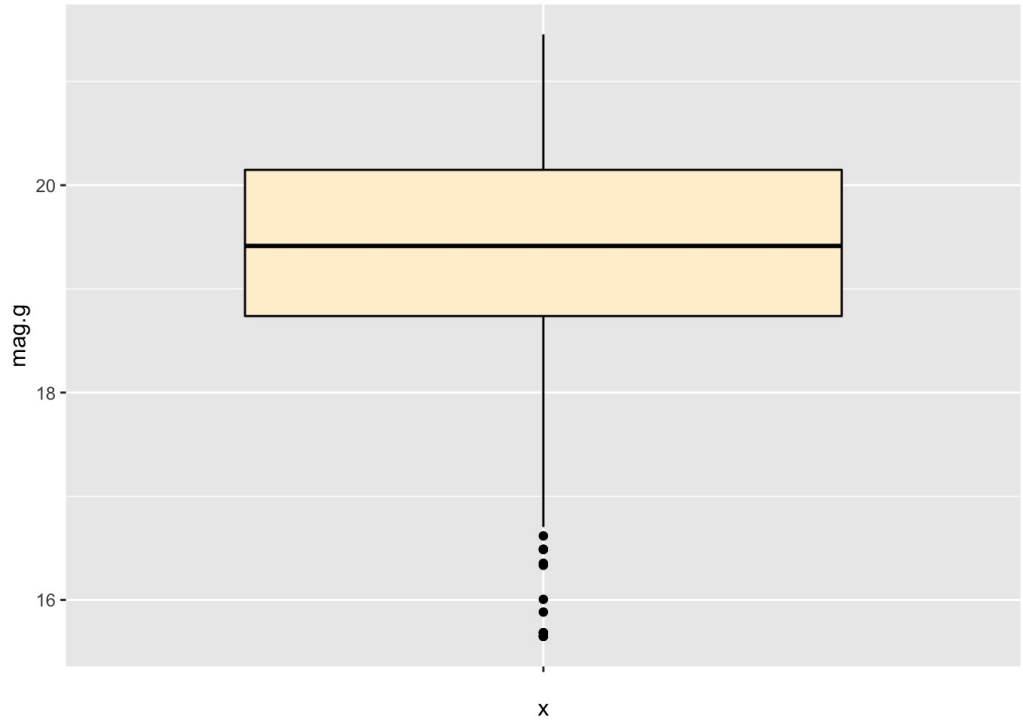

```
ggplot(data=df,mapping=aes(x="",y=temperature)) + geom_boxplot(color="black",fill="slategray1")
```




```
ggplot(data=df,mapping=aes(x="",y=mag.i)) + geom_boxplot(color="black",fill="lavender")
```

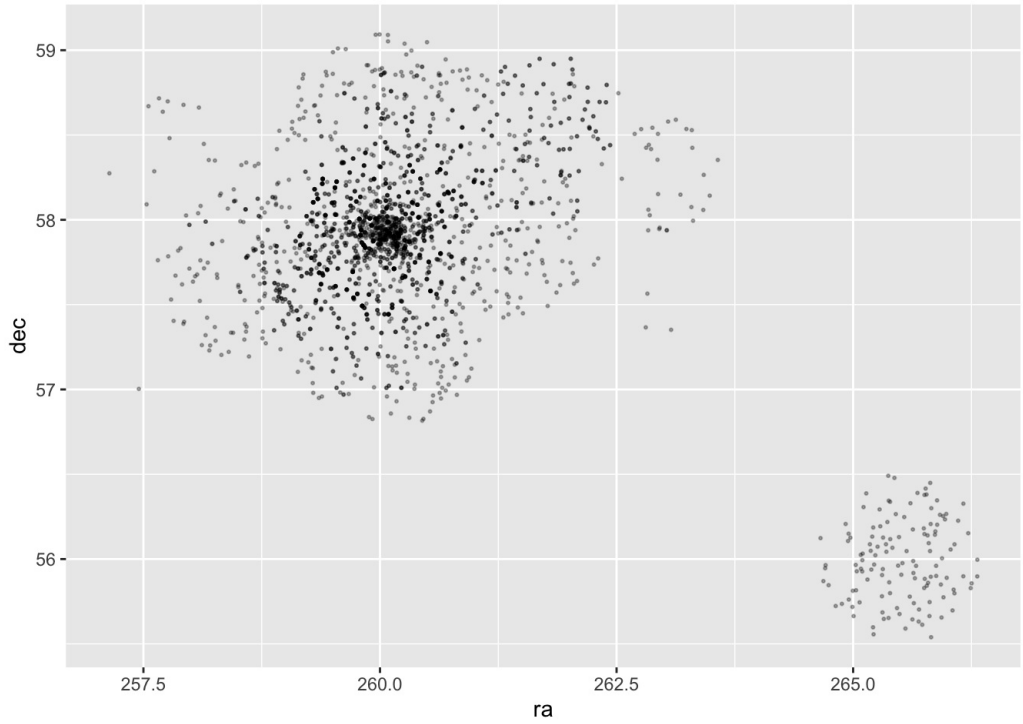



```
ggplot(data=df,mapping=aes(x="",y=mag.g)) + geom_boxplot(color="black",fill="papayawhip")
```



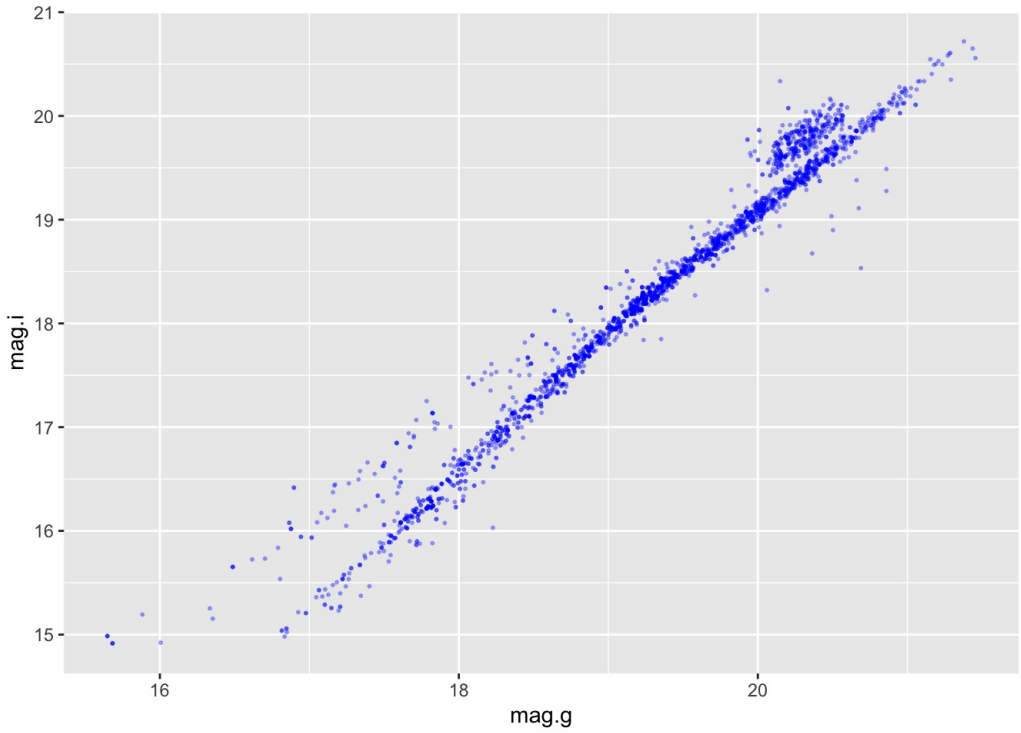
The boxplots above show a visual representation of the information given in the summary. As noted before,it appears that the distribution of the mag.g and mag.i boxplots are very closely related in terms of the IQR/median, although mag.g does have a couple more outliers. The ra and temperature boxplots also seem similar in their distribution, an d they both have a lot of outliers towards the top of the boxplot.

```
library(ggplot2)
ggplot(data=df,mapping=aes(x=ra, y=dec)) + geom_point(color="black", size=.4, alpha=.3)
```



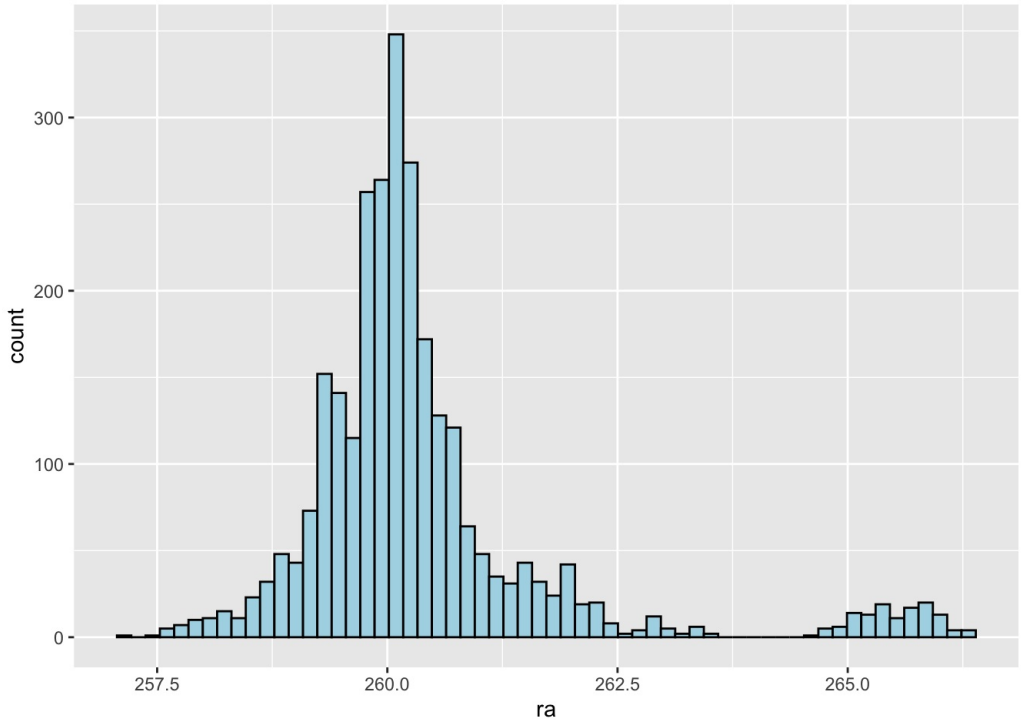
The scatterplot showing ra and dec appears to have two distinct clusters. Given that ra is the measure of celestial longitude and dec is the measure of celestial latitude in degrees, perhaps it does make sense there there is this d istinct clustering in the plot.

```
ggplot(data=df,mapping=aes(x=mag.g, y=mag.i)) + geom_point(color="blue", size=.4, alpha=.3)
```

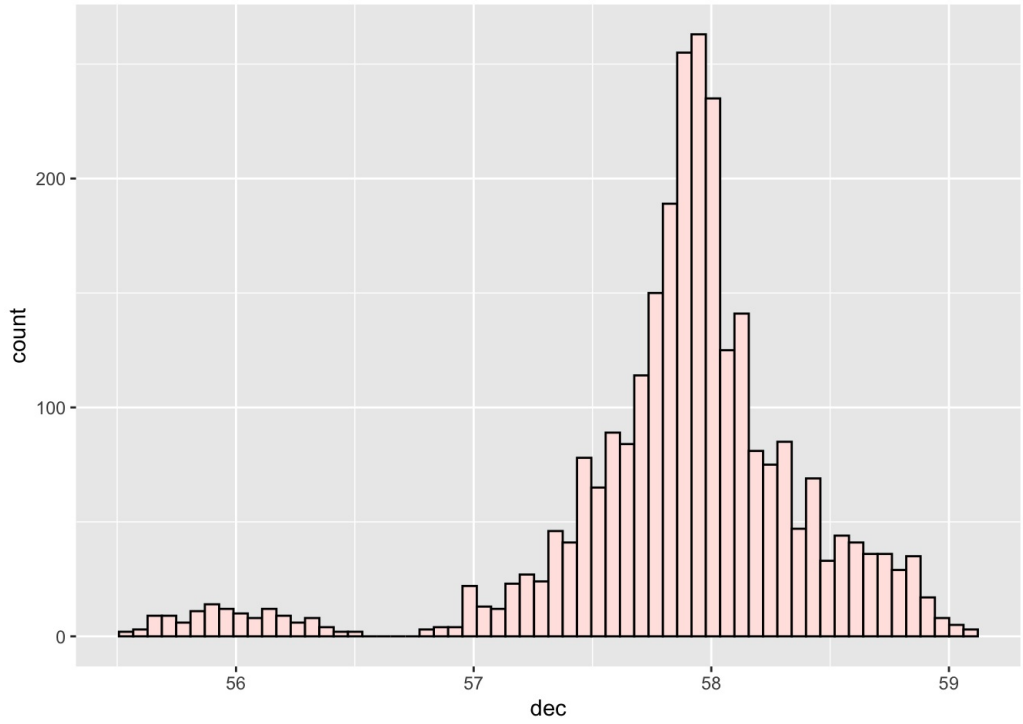


It seems that the two magnitudes are linearly related and there is a storng positive correlation between the mag.g and mag.i

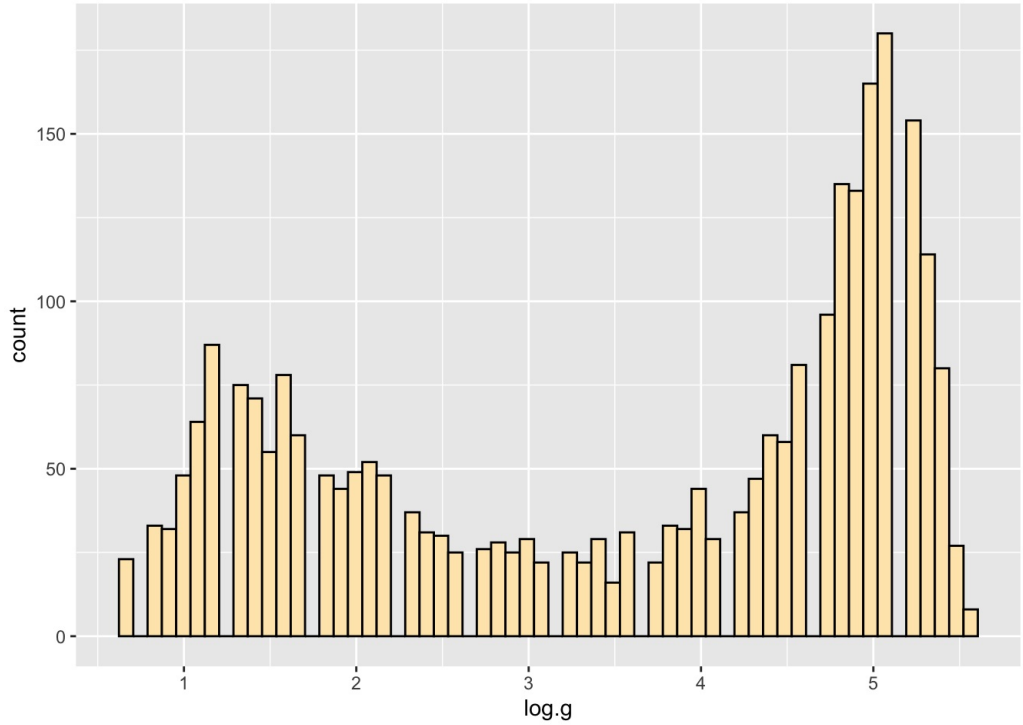
```
ggplot(data=df,mapping=aes(x=ra)) + geom_histogram(color="black", fill="lightblue", bins=60)
```



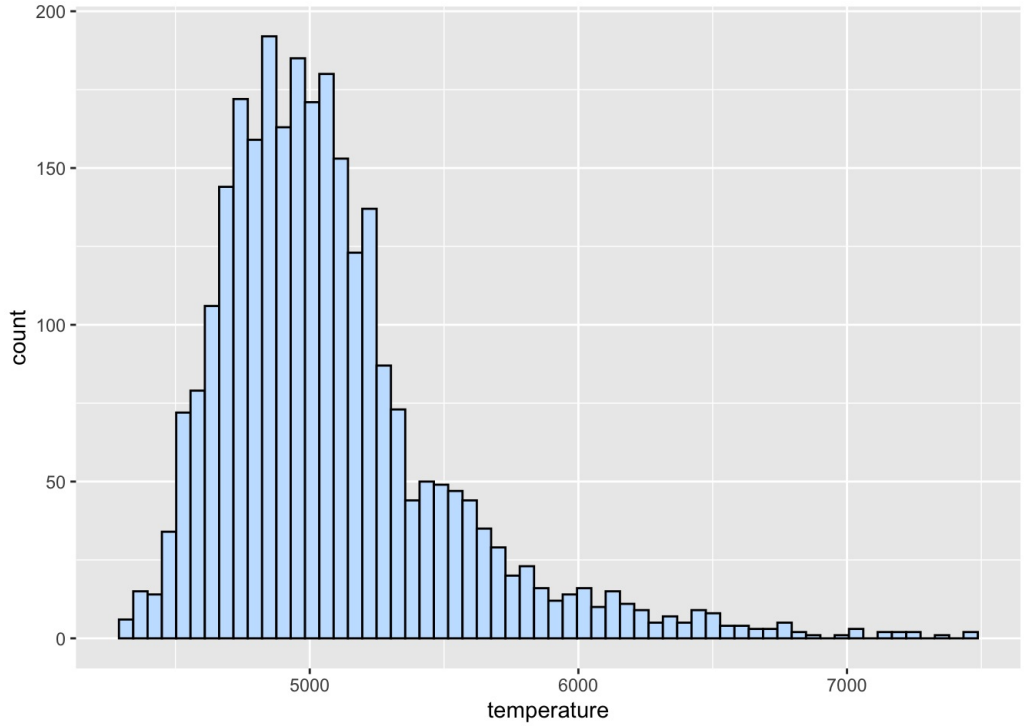

```
ggplot(data=df,mapping=aes(x=dec)) + geom_histogram(color="black", fill="mistyrose", bins=60)
```



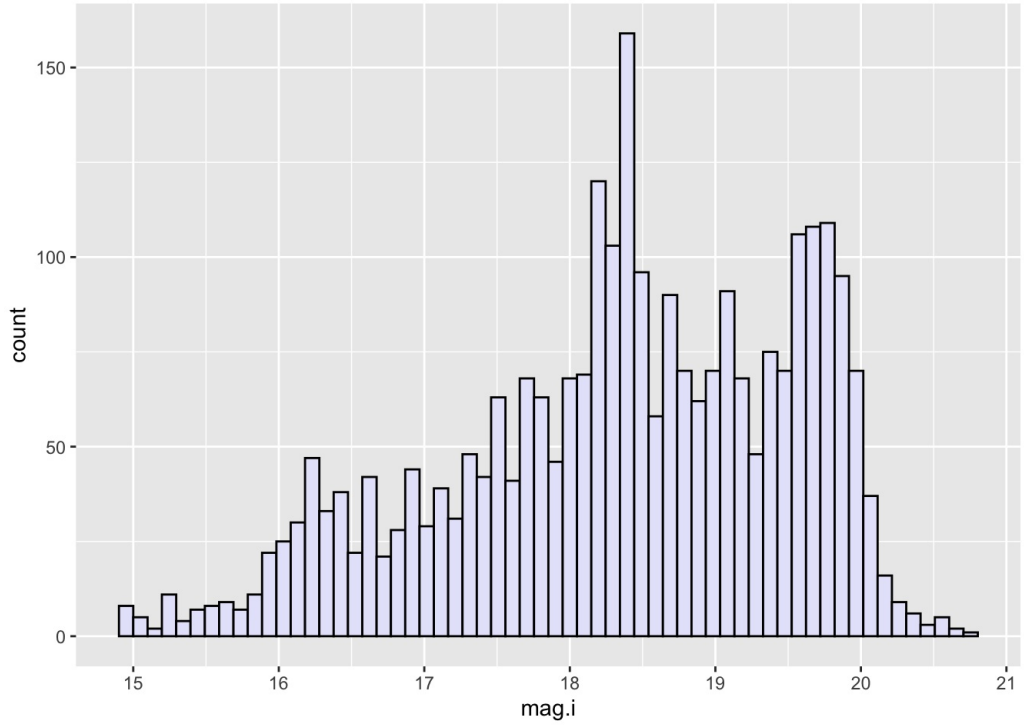

```
ggplot(data=df,mapping=aes(x=log.g)) + geom_histogram(color="black", fill="wheat1", bins=60)
```



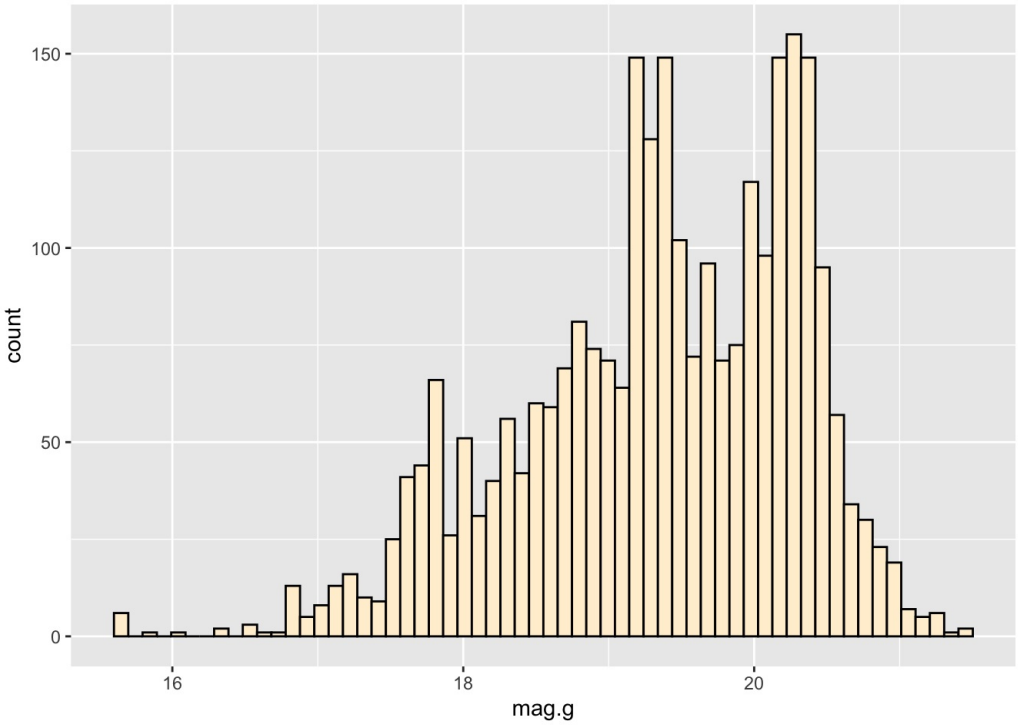

```
ggplot(data=df,mapping=aes(x=temperature)) + geom_histogram(color="black", fill="slategray1", bins=60)
```




```
ggplot(data=df,mapping=aes(x=mag.i)) + geom_histogram(color="black", fill="lavender", bins=60)
```




```
ggplot(data=df,mapping=aes(x=mag.g)) + geom_histogram(color="black", fill="papayawhip", bins=60)
```

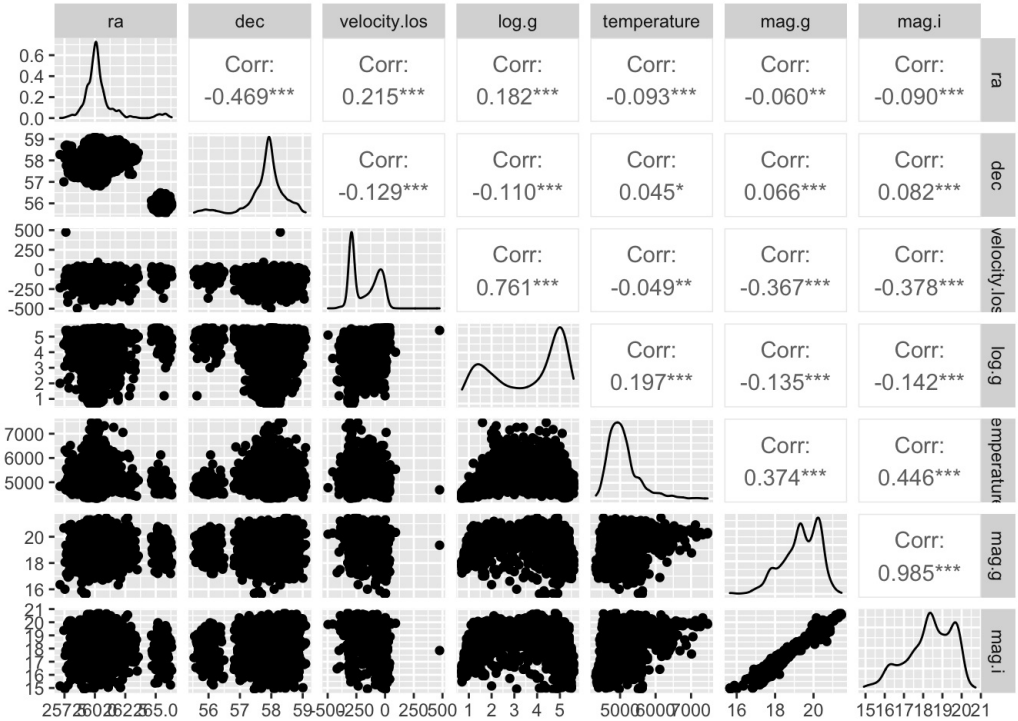


These histograms show the distributions of the following variables in this set of data:

ra= right ascension (or celestial longitude), in degrees
dec= declination (or celestial latitude), in degrees
velocity.los= radial velocity, along the line-of-sight, in km/s
temperature= the star's effective temperature (the Sun's is 5500 K)
log.g= the log of the surface gravity in cgs units
mag.g=the star's apparent magnitude in the SDSS g-band; larger values indicate fainter stars
mag.i=the star's apparent magnitude in the SDSS i-band

It appears that the distribution of dec, mag.i, and mag.g are skewed left while temperature and ra are skewed right. All of the histograms are unimodal, except log.g, which appears to be bimodal. Interestingly, ra and dec seem to be almost symmetric, where their distributions are very similar except that their skewness is directly opposite from one another.

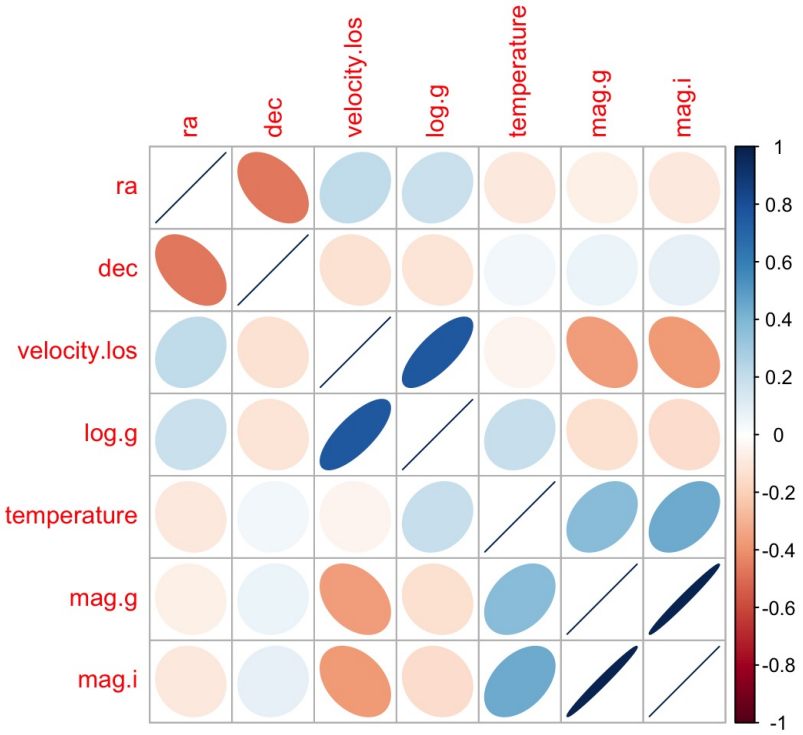
```
suppressMessages(library(GGally))

df %>% dplyr::select(.,ra, dec, velocity.los, log.g, temperature, mag.g, mag.i) %>% ggpairs(.,progress=FALSE, lower=list(combo=wrap("facethist", binwidth=.8)))
```




```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
df %>% dplyr::select(.,ra, dec, velocity.los, log.g, temperature, mag.g, mag.i) %>% cor(.) %>% corrplot(.,method="e
llipse")
```



It appears that mag.g and mag.i have strong linear correlation; mag.i/mag.g are also somewhat positive correlation while dec and ra have somewhat negative correlation.