# | Milestone Project 1: Scenario 1

## Overview

You have been hired by a space research company named AWAY (Aliens Where Are You). One of their biggest projects has been to compile a list of yellow-dwarf stars. Yellow-dwarf stars are important because that is the type of star our sun is. AWAY scientists have theorized that other yellow-dwarf stars may be able to support life the way our sun supports life on earth. AWAY looks for planets around these stars in the search for alien lifeforms

You have been tasked with determining which stars are most likely to have nearby planets that could possibly support life.

**Audience:** Scientists

## About the Dataset

- **Temperature:** The average temperature of the star
- **L:** The L column stands for “luminosity,” which measures the brightness of a star
- **R:** The R column stands for “radius,” which is the distance from the center of a star to its outer edge
- **Is_star:** Displays 1 if the object is a star or 0 if the object is not a star
- **A_M:** The column A_M means “absolute magnitude,” which is the magnitude of a star when measured from a distance of 10 parsecs (1 parsec = 3.26 lightyears)
- **Spectral Class:** The group that a star belongs to depending on its spectrum and luminosity

- **Color:** The color of the star

| CLASS | COLOR | TEMPERATURE |
|---|---|---|
| O | Blue | >= 30,000 K |
| B | Blue-White | 10,000-30,000 K |
| A | White | 7,500-10,000 K |
| F | Yellow-White | 6,000-7,500 K |
| G | Yellow | 5,200-6,000 K |
| K | Orange | 3,700-5,200 K |
| M | Red | 2,400-3,700 K |

- **Type:** Type of star

| NUMBER | TYPE |
|---|---|
| 0 | Red Dwarf |
| 1 | Brown Dwarf |
| 2 | White Dwarf |
| 3 | Main Sequence |
| 4 | Super Giants |
| 5 | Hyper Giants |

# Part 1 - Data Preparation

**Tool: Microsoft Excel**

Step 1: Define the goal.

What is the goal for this data analysis? What questions are you trying to answer?

The goal (hypothesis) of this data analysis is to find all the yellow dwarf stars that are closest to the Sun's attributes (Radius, Luminosity, Absolute Magnitude, Color). The questions I'm trying to answer are whether the yellow dwarf stars can support other lifeforms similar to Earth by their Absolute Magnitude and Radius.

Step 2: Remove irrelevant columns.

Not every column will be useful for analysis. We can remove any columns where every row is the same. *Hint: An easy way to tell if every row is the same is to use the filter columns tool.*

What column(s) did you remove and why?

The type column and spectral class are removed because there's no need to define the star type in this analysis. I deleted the spectral class column after I did Data Validation (explained in Step 4) since it is no longer needed for analysis.

Step 3: Identify typos.

Search for any typos that exist in the dataset in the *Color* column and correct them.

Additionally, ensure that the data in the *Color* column is formatted to be the same. For example, two-worded colors have a hyphen (-) between the words, while others do not. Choose one consistent format.

Step 4: Identify nulls and missing values.

Almost every dataset contains missing values. The goal is to handle such values uniformly throughout the dataset.

How did you choose to handle missing values and why?

There were no null values nor missing values in the dataset. However, I did data validation since some of the stars in the dataset did not have temperatures matching the corresponding color and spectral class. I used an IFS function in Excel to validate the data and ensure the Color column has a consistent format.

Step 5: Remove duplicates.

Remove any duplicates in the dataset. Navigate to the **Data** tab in the Data Tools section and click the **Remove Duplicates** button.

# Part 2 - Data Exploration

**Tool: Excel**

Step 1: Calculate average luminosity.

AWAY is expanding its research on stars. Yellow-white stars have the spectral class "F." Calculate the average luminosity of all the stars with the spectral classification "F." The average luminosity of a yellow-white star is: **51,428.97 Watts**

**SQL query**

```
SELECT AVG(Luminosity) AS average_luminosity
FROM stars.stars
WHERE Color = 'Yellow-White';
```

Step 2: Filter the top 5.

AWAY would like a list of the five hottest stars (by temperature) and their color.

The temperature and color of the hottest star is: **40,000 K and Blue**

SQL Query:

```
-- Used a windows function to number the stars by row
SELECT ROW_NUMBER() OVER() AS Star, Temperature, Color
FROM stars.stars     -- stars is the name of the database and the table
ORDER BY Temp DESC
LIMIT 5;
```

SQL Output:

| STAR | TEMPERATURE (K) | COLOR |
|---|---|---|
| 1 | 40,000 | Blue |
| 2 | 39,000 | Blue |
| 3 | 38,940 | Blue |
| 4 | 38,234 | Blue |
| 5 | 37,882 | Blue |

# Part 3 - Gather Insights with Statistics

**Tool: Excel Data Analysis ToolPak, Excel functions, and visualizations**

Step 1: Calculate and visualize descriptive statistics with the Data Analysis ToolPak.

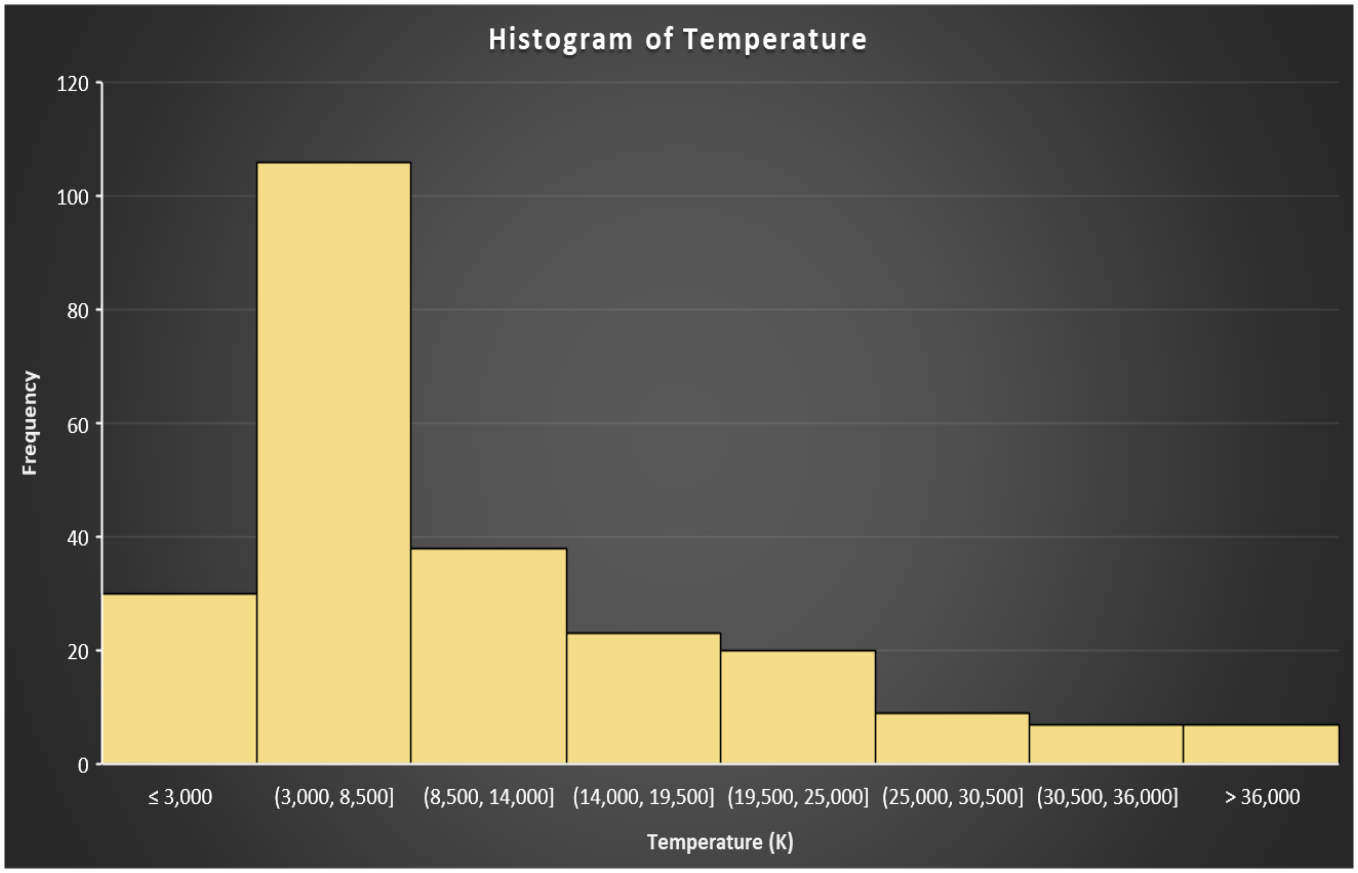
Report the summary statistics for the *Temperature* column.

| Temperature (K) | |
|---|---|
| **Mean** | 10,497 |
| **Standard Error** | 616 |
| **Median** | 5,776 |
| **Mode** | 3,600 |
| **Standard Deviation** | 9,552 |
| **Sample Variance** | 91,244,824 |
| **Kurtosis** | 0.88 |
| **Skewness** | 1.32 |
| **Range** | 38,061 |
| **Minimum** | 1,939 |
| **Maximum** | 40,000 |
| **Sum** | 2,519,391 |
| **Count** | 240 |

Interpret the skewness (**S**) and kurtosis (**K**) for *Temperature*.

The distribution of the Temperate data is right-skewed (positive skew) since mean > median and S > 1. Compared to the normal distribution curve, the curve of the Temperature data displays a platykurtic curve and is lightly tailed (K < 3).

Create and interpret either a histogram or box and whisker plot for *Temperature.* Be sure to paste your visualization below.



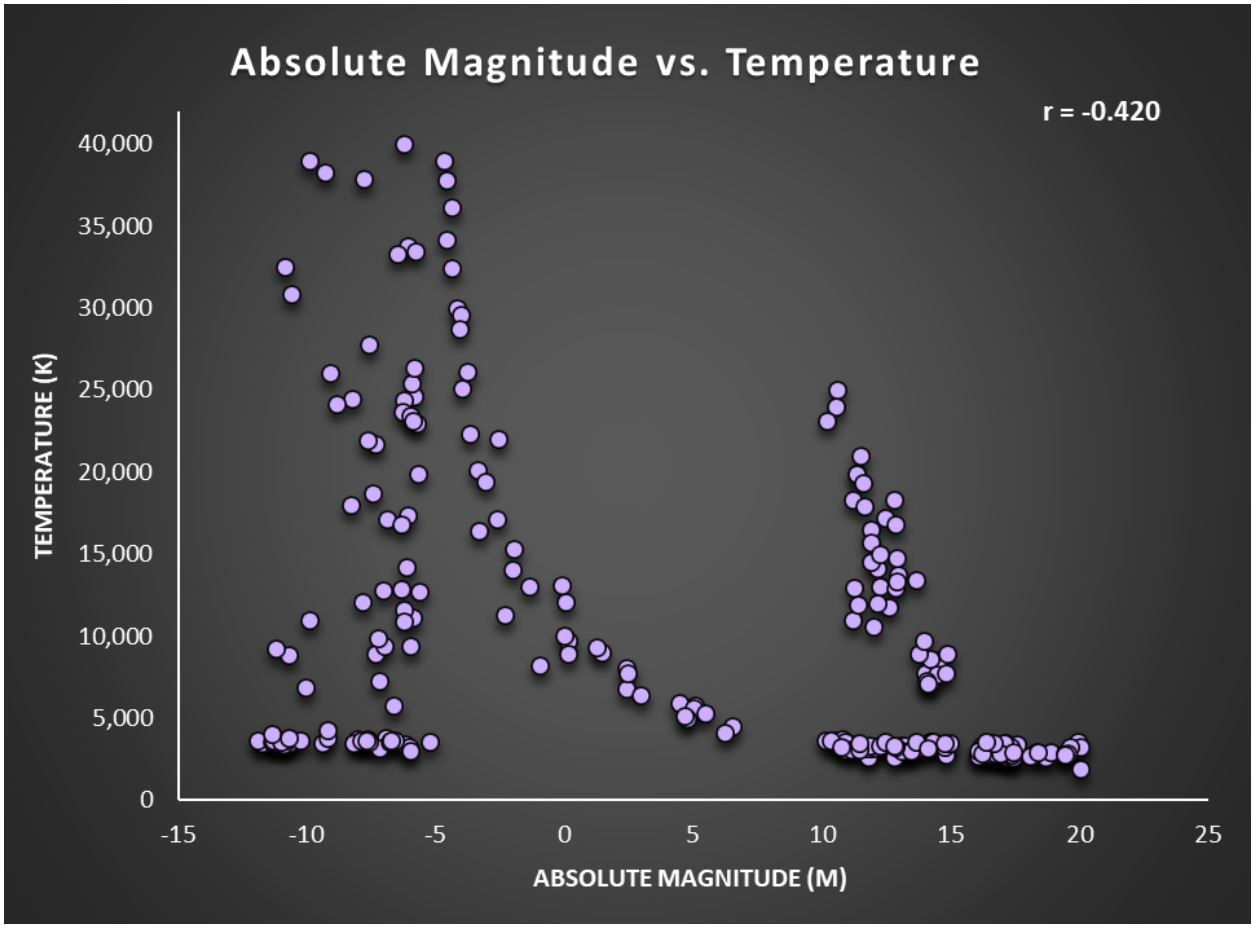

Step 2: Calculate and interpret the correlation of two variables using a scatterplot and the correlation coefficient.

Create and interpret a scatterplot of *Temperature* and *Absolute Magnitude*. Report the correlation coefficient by:

- Displaying the correlation coefficient on the scatterplot
- Using the CORREL function
- Calculating the correlation coefficient in the Data Analysis ToolPak

Be sure to paste your visualization below.

There is a weak, negative association between Absolute Magnitude and Temperature (r = -0.420).

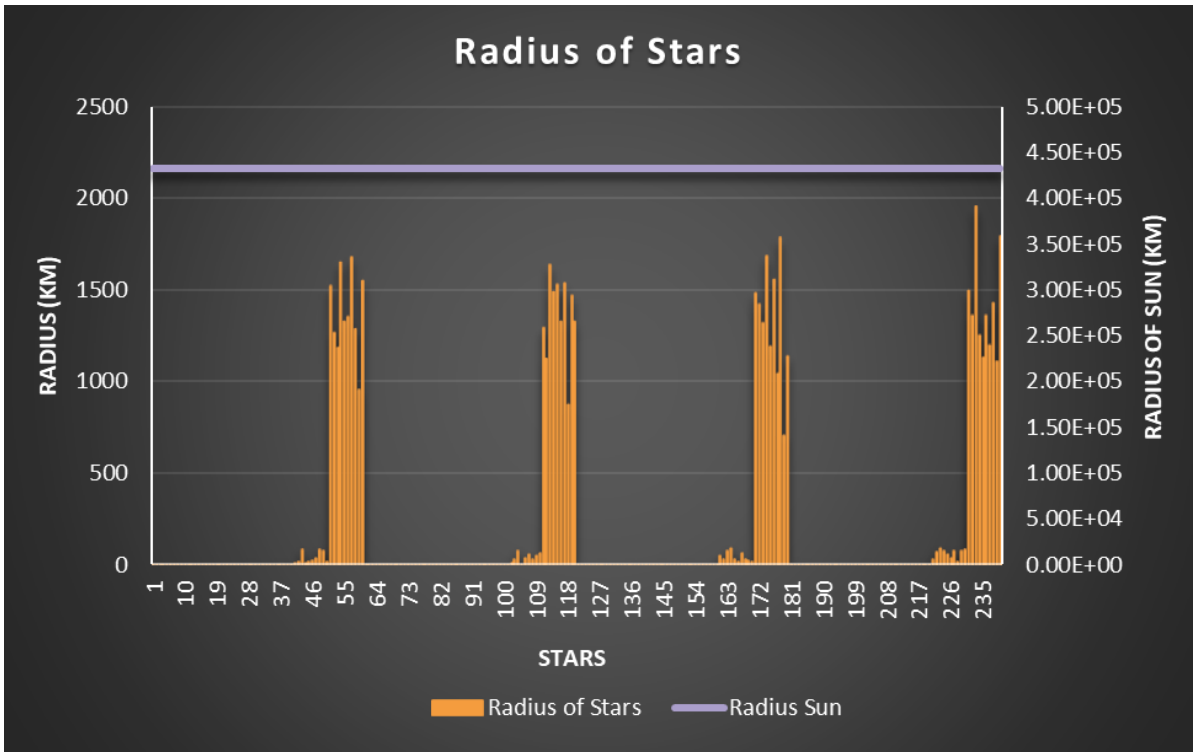



Step 3: Use a combination of bar and line charts to compare groups.

AWAY is looking for stars that have similar properties to the sun. Below is the luminosity, radius, and absolute magnitude of the sun:

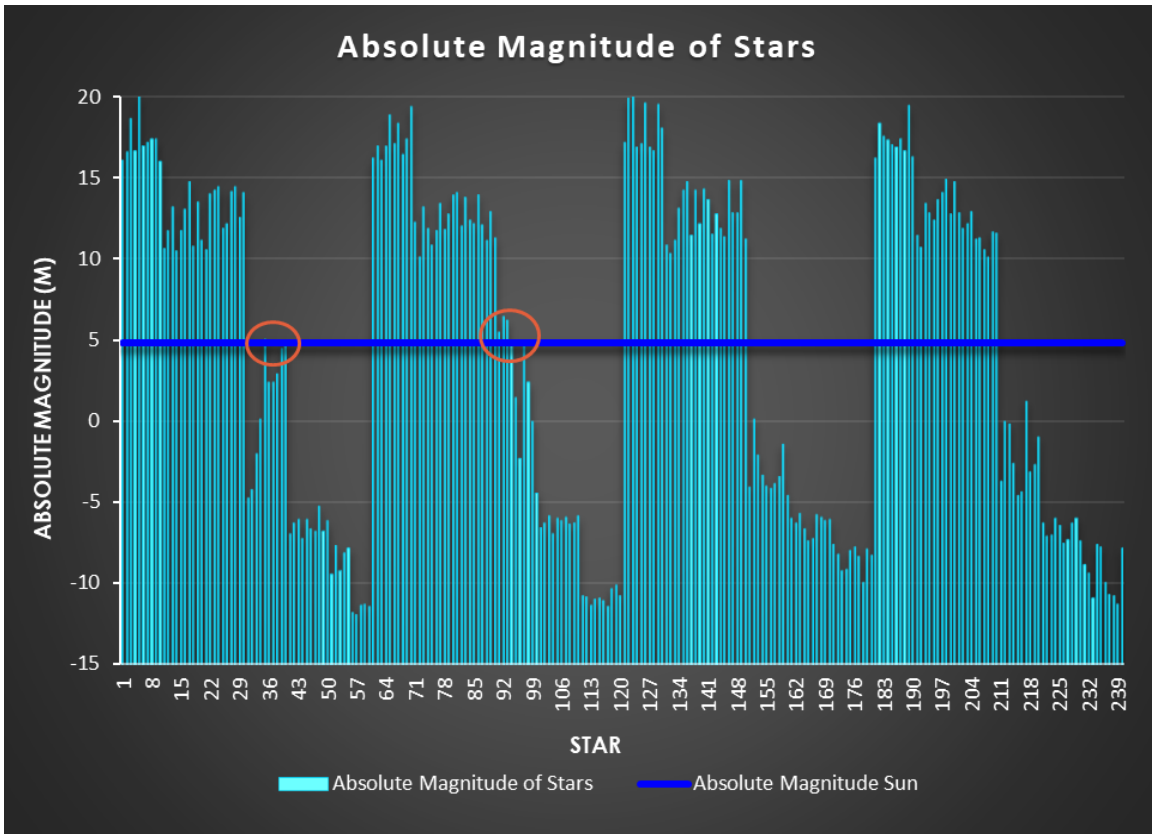
- Luminosity: 3.75E+28
- Radius: 4.33E+05
- Absolute Magnitude: +4.83

Create a combination chart with *Radius and Sun Radius*. Be sure to paste your visualization below. Note the stars that have a similar radius to the sun.




None of the star's radii were similar to the radius of the Sun!

Create a combination chart with *A_M (absolute magnitude) and A_M Sun*. Be sure to paste your visualization below. Note the stars that have a similar absolute magnitude to the sun.




| Stars similar to Absolute Magnitude of Sun | | | | |
|---|---|---|---|---|
| **Star** | **Absolute Magnitude (M)** | **Absolute Magnitude of Sun (M)** | **Temperature (K)** | **Color** |
| 1 | 5.05 | 4.83 | 5,800 | Yellow |
| 2 | 4.46 | 4.83 | 5,936 | Yellow |
| 3 | 5.03 | 4.83 | 5,587 | Yellow |
| 4 | 5.49 | 4.83 | 5,300 | Yellow |
| 5 | 4.78 | 4.83 | 4,980 | Orange |
| 6 | 4.68 | 4.83 | 5,112 | Orange |

Step 4: Create a simple regression equation and interpolate information given new information.

There are many indicators that could predict whether our data contains dwarf stars near planets that might contain life. Choose two:

A. Temperature
B. Luminosity
C. Radius
D. Absolute magnitude

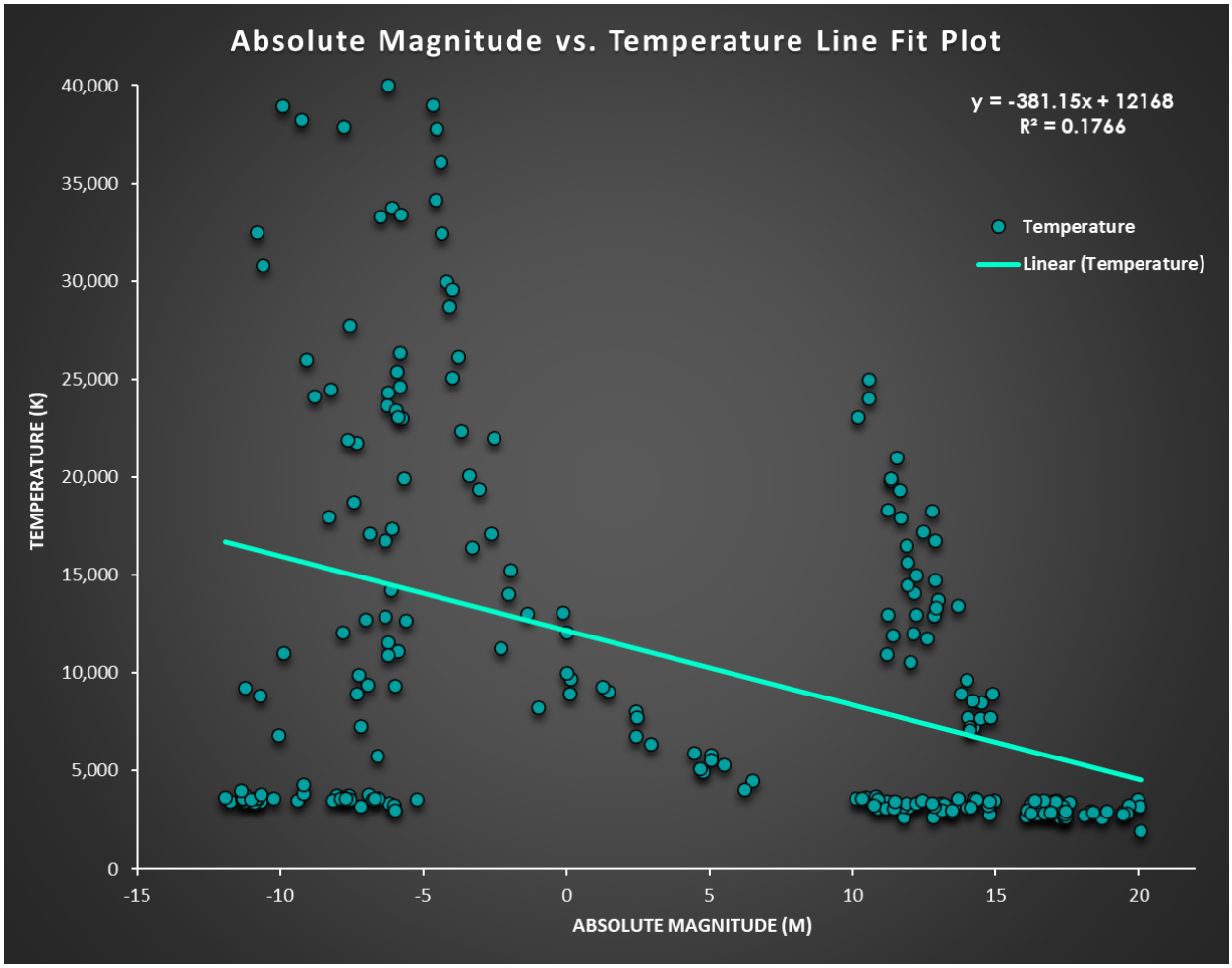
Use the **Data Analysis ToolPak** to create a regression line with the two indicators you have chosen. Use a 95% confidence level. Report your equation below and the value of the correlation coefficient.

AWAY has found a new star with the following characteristics.

- Luminosity: 1.45E+04
- Radius: 3.19E-01
- Absolute Magnitude: -6.12

Use your regression line to find $\hat{y}$. What can be said about this new star? Can you predict the color?

- Go to the next page for the regression line results!




According to the regression line above, the predicted temperature for a new star is 14,500 K with an absolute magnitude of -6.12. Since Temperature is the dependent variable, the color of the new can be predicted. The color of the new star is predicted to have a color of Blue-White.

From the line of best-fit regression plot above, about 18% of the variation in Temperature can be explained by the variation in the Absolute Magnitude of stars.

# Part 4 - Plan a Report

**Tool: Word document, whiteboard application such as Miro**

Step 1: Choose a report style.

Which report style will you use?

A. Annual, quarterly, monthly
B. Compliance
C. Progress
D. Feasibility
E. Operational
F. Strategic
G. Executive
H. Showcase a specific issue
I. Specific sector

Detail why you choose this option.

I chose the progress report style because there's no ending to the story and there needs to be more data collection in order to accurately predict the Temperature of the Stars

Step 2: Gather report details.

Provide a title for your report based on the main goal or key insight.

Analyzing Star's Data for the Elonians

Write a brief description (about 2–3 sentences) on what your report is about.

This report is about stating the current findings of the data needed to conduct the analysis of dwarf stars that have the potential to sustain life on other planets for a specific alien species known as "Elonians". Scientists hypothesized that the planets surrounding those dwarf stars are yellow dwarf stars. However, the data in this report concludes that the dwarf-stars are not actually yellow dwarf stars although there are a few stars that have similar values in Absolute Magnitude to that of the Sun that sustains life on Earth.

Produce a list of everyone on the team and their roles, such as “created visualizations” or “completed data preparation.”

I did all of the work myself!

Step 3: Plan the visualizations.

What will be the main graphic or chart? It should be the most important insight you want to share.

Absolute Magnitude Combo Plot

What will be the supporting graphics or charts? Keep in mind that you might need other visualizations to illustrate the main point and convince your audience.

Include the Line Fit Plot and Histogram for Temperature

Are there any other topic-relevant images that you will add to the report for a visual boost? For instance, you might want to include an image of stars or a yellow dwarf in your report.

I’m adding the radius combo plot to my report to show that none of the dwarf star have similar radius values to the Sun to further explain my conclusion.

Step 4: Report key insights.

List the main insights you found in your data.

- Six stars have an Absolute Magnitude similar to the Absolute Magnitude of the Sun
- 18% of the variation in Temperature can be explained by the variation in Absolute Magnitude
- There is a weak negative association between the Absolute Magntude and Temperature of Stars

What solution or conclusion will you make? List the insights or data you gathered to support this.

The conclusion is to find more yellow dwarf stars that have radii similar to the radius of the Sun. There should also be more data collection to look for yellow-white dwarf stars that have similar Absolute Magnitude and Radius values to that of the Sun.

# Part 5 - Develop a Data Story

**Tool: Word document, whiteboard application such as Miro**

Step 1: Complete the data story checklist.

What do you want to do with your dataset?

A. Inform – summarize findings of a study
B. Classify the data
C. Make a company decision or predict future results
D. Inspire/persuade people to act

Who is your audience?

Scientists interested in Astronomy and Aliens

Step 2: Organize your story points.

Choose some common story points for your data story. Write a few details on how you will illustrate these points.

- Change over time
- Relationship of two metrics
- Intersection (when one metric surpasses another)
- Prediction
- Compare and contrast
- Drill down (general → specific)
- Zoom out (specific → general)
- Cluster (values concentrated in an area)
- Outlier (data that lies outside the norm)

- Prediction: Explain the linear regression results for the Absolute Magnitude vs. Temperature of Stars.
- Relationship: Show the correlation between the Absolute Magnitude and Temperature of stars
- Compare and Contrast: Comparing the Absolute Magnitude of stars to the Absolute Magnitude of the Sun

Step 3: Create a story arc.

What is the setting (context: who, what, where, when)? Include a hook—something to get the audience's attention.

- Setting: Scientists in the Space Research Facility at AWAY in the year 2060 who are searching for a home for a specific alien species named "Elonians".
- Hook: The "Elonians" are running out of resources to sustain themselves on Earth and are on the brink of extinction.

What are the rising insights that support/lead to your goal or main point?

- Four of the six dwarf stars found to have an absolute magnitude similar to the Sun are yellow dwarf stars
- None of the dwarf stars in the dataset have similar radii to that of the Sun

Step 4: Add context to your story.

Is there any background information the audience needs to know to make sense of the data insights?

- The Elonians are gray aliens that need electricity as their energy source in order for their bodies to survive.
- The yellow dwarf stars surrounding those planets with similar Absolute Magnitude values to the Sun were found to be a sustainable energy source to grow crops, power the solar panels and smart technology.

# Part 6 - Build a Report

**Tool: Excel**

Create a one-page report (using the ***Part 6_Report Template.xlsx***) that includes:

- Specific, targeted metrics illustrated with meaningful visualizations
- Storytelling techniques
- The recommendation or solution for the client

Consider the following when structuring your report:

- Report goal
- Color scheme
- Visualizations
- Text and graph balance

The final format must be an Excel document that your team will turn in, in addition to this packet.