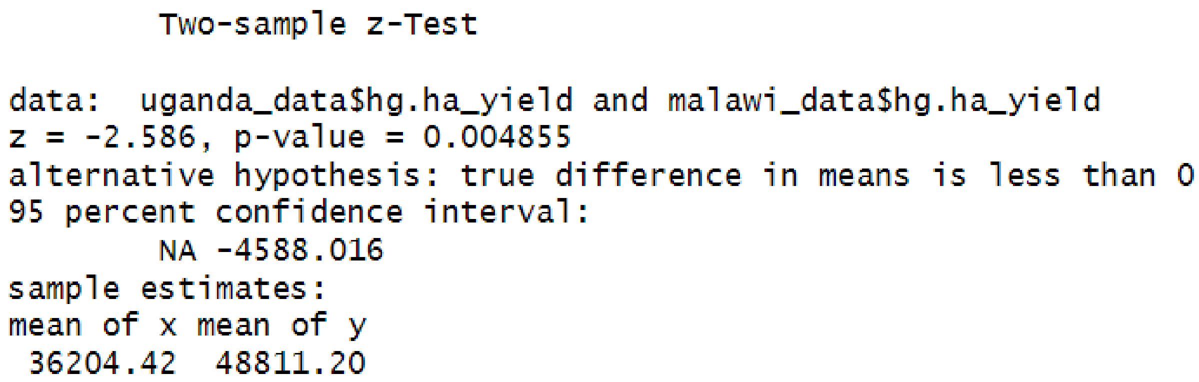
Janine Yanes (jqy2)

Data Science Capstone - HW2 Report

Dataset used: https://www.kaggle.com/datasets/mrigaankjaswal/crop-yield-prediction-dataset

The dataset described crop yields and the various factors associated with each one: the country where the crop was harvested, what crop it was, what year it was, the amount harvested (hg/ha), the average yearly rainfall for the respective country (mm), the amount of pesticides used (tonnes), and the average temperature for the respective country (℃) that year.



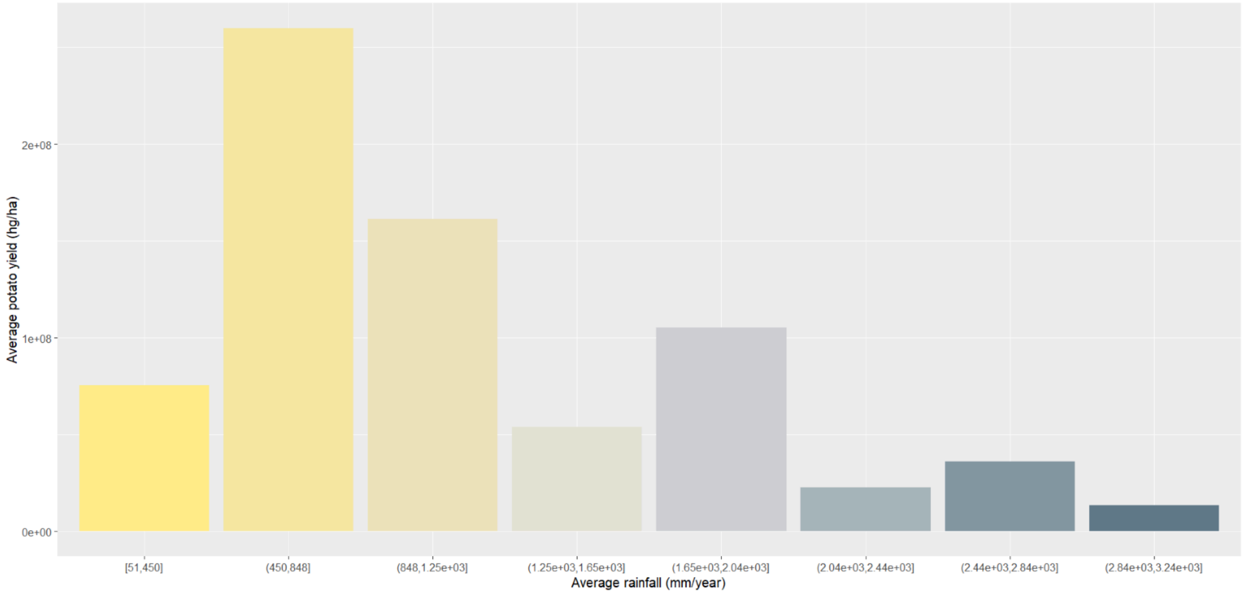
```
        Two-sample z-Test

data:  uganda_data$hg.ha_yield and malawi_data$hg.ha_yield
z = -2.586, p-value = 0.004855
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        NA -4588.016
sample estimates:
mean of x mean of y
 36204.42  48811.20
```

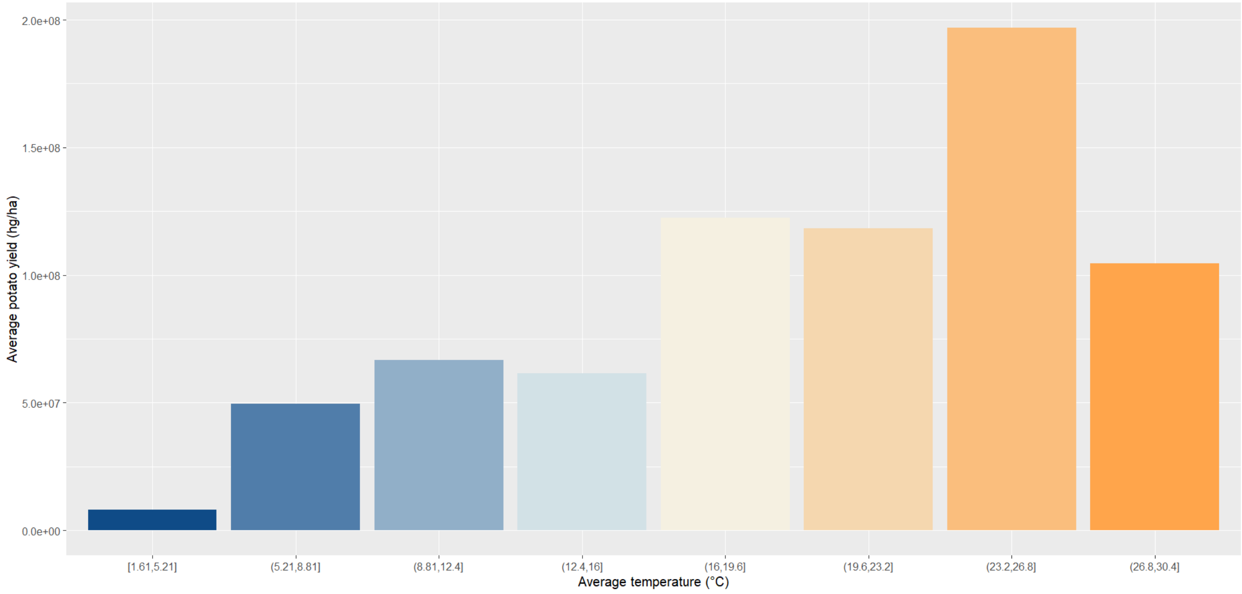
For my hypothesis test, I looked at the crop yields for the countries Uganda and Malawi. Comparing the two highlighted their similarities: both were located in East Africa, with a 1 mm difference in average yearly rainfall (1180 for Uganda, 1181 for Malawi) and a < 4 ℃ difference in mean yearly temperature (23.79478 for Uganda, 20.62953 for Malawi). However, this made their difference in mean crop yield (36204.42 for Uganda, 48811.20 for Malawi) even more prominent.

My null hypothesis was, “The average crop yield for Uganda is equal to the average crop yield for Malawi” (H_0: μ_{Uganda} - $\mu_{Malawi} = 0$), while my alternative hypothesis was, “The average crop yield for Uganda is less than the average crop yield for Malawi” (H_A: μ_{Uganda} - $\mu_{Malawi} < 0$). Conducting a z-test resulted in a p-value of 0.004855; based on this small p-value, we reject the null hypothesis, implying that the average crop yield for Uganda is indeed less than the average crop yield for Malawi, despite their similar environmental conditions.

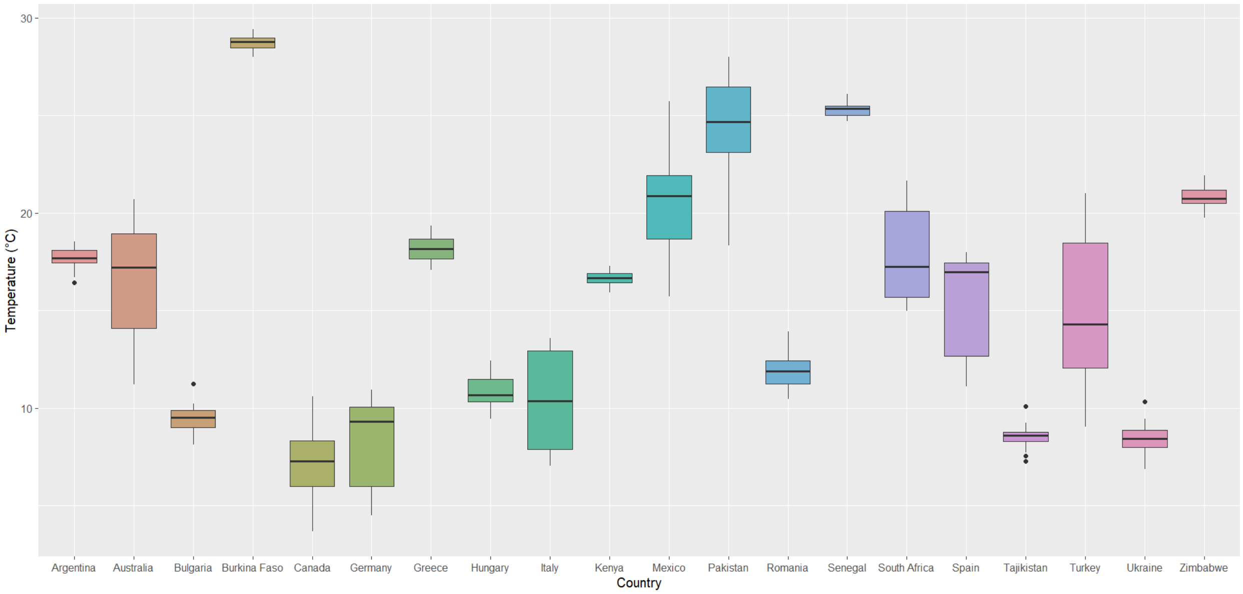
The following visualizations make a collective argument.



Looking at this bar graph, we see that countries with average yearly rainfall > 450 mm and ≤ 848 mm have the largest potato crop yields on average. This suggests that when considering rainfall, the ideal environment for growing potatoes is one with average yearly rainfall > 450 mm and ≤ 848 mm.



Looking at this bar graph, we see that countries with average yearly temperature > 23.20 ℃ and ≤ 26.80 ℃ have the largest potato crop yields on average. This suggests that when considering temperature, the ideal environment for growing potatoes is one with average yearly temperature > 23.20 ℃ and ≤ 26.80 ℃.



Combining our findings from the bar graphs above, we look at the distributions of average yearly temperature for all countries with average yearly rainfall in the (450, 848] mm range. The box plots show that Pakistan's median yearly temperature is ≈ 24.5 ℃, close to the middle of the (23.20, 26.80] ℃ range. Furthermore, it is the only country that completely fits (23.20, 26.80] ℃ in its interquartile range. This implies that Pakistan is the country whose environment is most suited for growing potatoes, since it is the country whose normal conditions are most centered around the presumed ideal conditions for growing potatoes (established based on the previous visualizations).