Matthew Jusino

Lab 9: Modeling Data 2

- **Q1 (1 pt.):** State the null hypothesis of the Chi-square test.
    - Make sure you state the null hypothesis in terms of Brown Creeper presence/absence and edge/interior habitats.

The null hypothesis of the Chi-square test would be that whether a habitat is edge/interior has no effect on the presence/absence of brown creepers. We should see an equal number present/absent in both edge and interior habitats.

- **Q2 (2 pts.):** Consider the results of your test and explain whether you think that Brown Creepers show a significant habitat preference.
    - Make sure your use the output of your statistical test to support your answer.

With the result of a p-value of 1.386e-6, we should reject the null hypothesis. This is evident simply looking at the data; brown creepers show a marked preference for Interior habitats.

- **Q3 (1 pt.):** Show the R-code you can use to create a model fit (call it `fit_species`) of penguin **body mass** as predicted by **penguin species**.

```
fit_species =

  lm(

    formula = body_mass_g ~ species,

    data = penguins)
```

- **Q4 (1 pt.):** Show the R-code you can use to create a model fit (call it `fit_sex`) of penguin **body mass** as predicted by **sex**.

```
fit_sex =

  lm(

    formula = body_mass_g ~ sex,

    data = penguins)
```

- **Q5 (1 pt.):** Show the R-code you can use to create a model fit (call it `fit_both`) of penguin **body mass** as predicted by **species** and **sex**. This should be an *interactive* model, i.e. it should include a sex and species interaction.

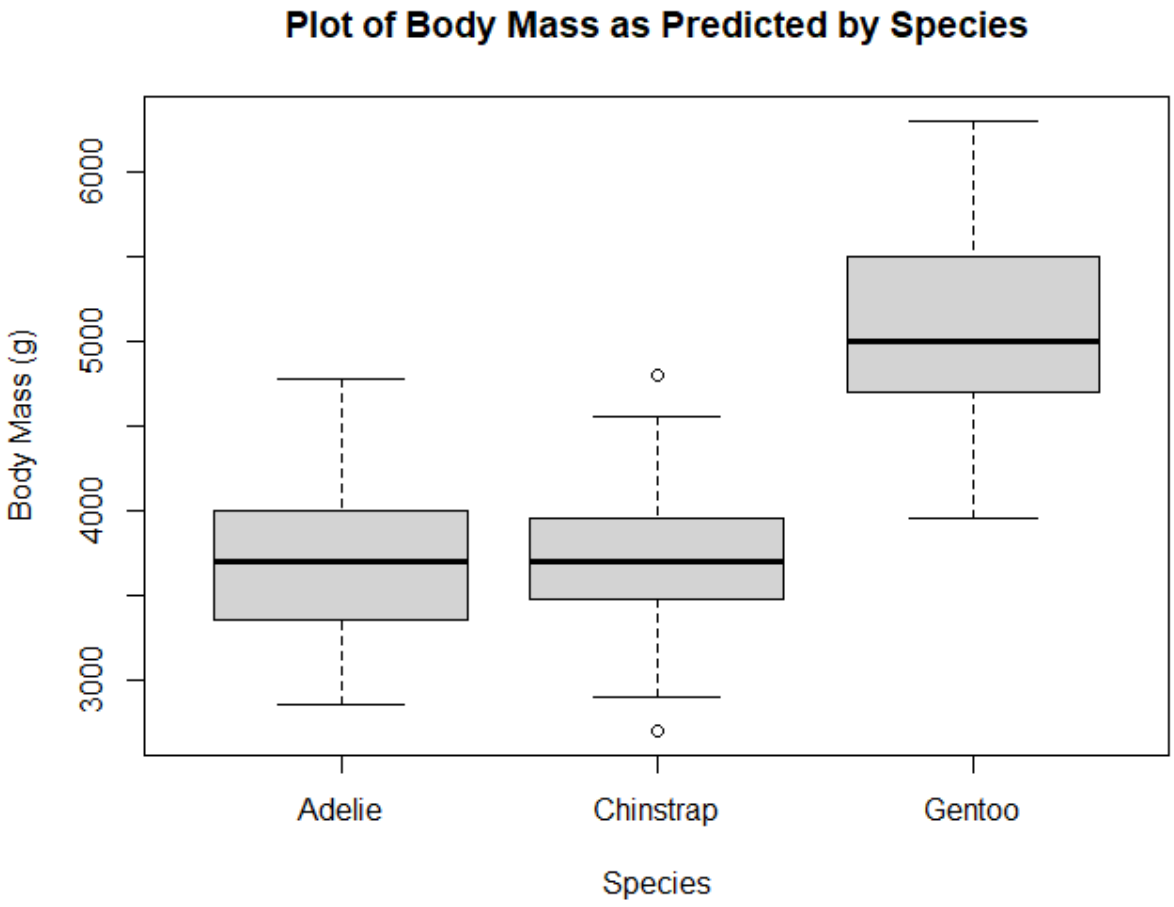
```
fit_both =
```

lm(

formula = body_mass_g ~ species*sex,

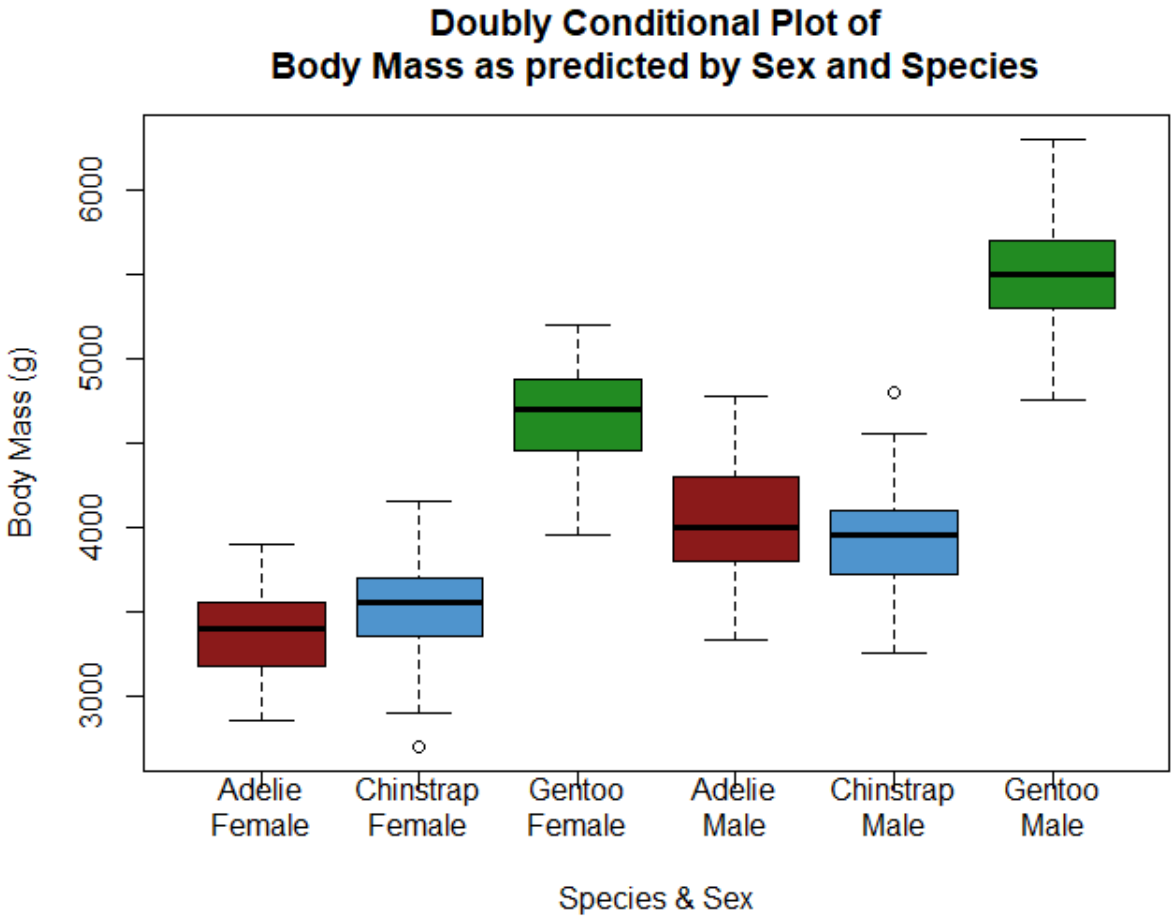
data = penguins)

- **Q6 (1 pt.):** Include a conditional boxplot corresponding to the grouping structure in your `fit_species` model.




- **Q7 (1 pt.):** Include a conditional boxplot corresponding to the grouping structure in your `fit_sex` model.


- **Q8 (3 pts.):** Include a conditional boxplot corresponding to the grouping structure in your `fit_both` model.
    - Your group labels must all correspond to the correct box, be visible, and sensible.




- **Q9 (3 pts.):** Based on the shapes of the boxes, which of the models (if any) do you think may have problems fulfilling the homogeneity assumption?

I think the fit_sex model might have some problems, as the variance looks slightly higher for males. I think the fit_both model will definitely have issues fulfilling the homogeneity assumption, as there is a clear higher variance for Gentoo Male compared to other species and sexes.

- **Q10 (1 pt.):** State the null hypothesis of the Bartlett test.

The null hypothesis for the Bartlett test is that there is no significant difference in variances in the model fit.

- **Q11 (1 pt.):** What was the p-value from the Bartlett test of homogeneity for observations grouped by *species*?
    - You can round your answer to 4 decimal digits.

p-value = 0.0501

- **Q12 (1 pt.):** What was the p-value from the Bartlett test of homogeneity for observations grouped by sex?

    - You can round your answer to 4 decimal digits.

p-value = 0.0319

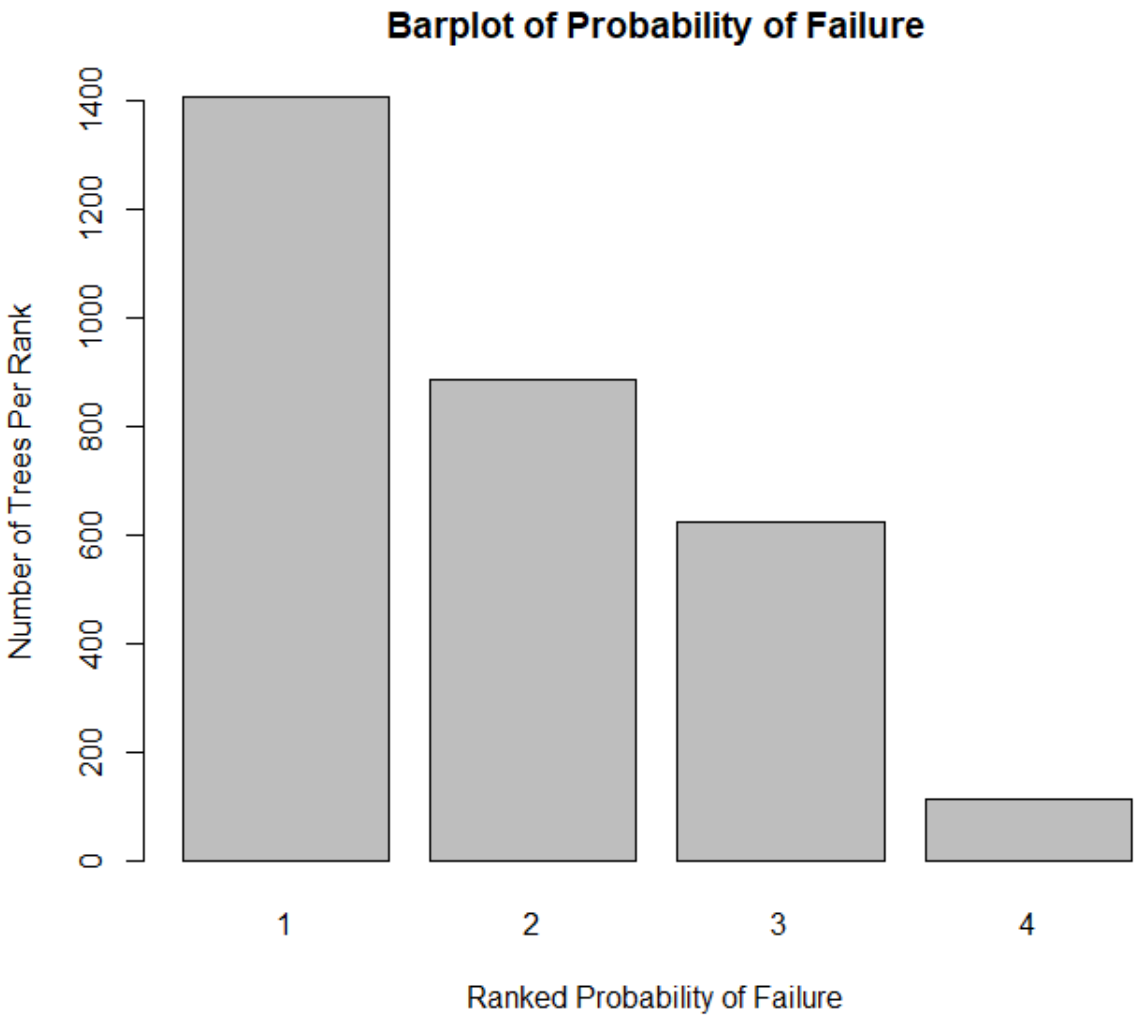
- **Q13 (1 pt.):** What was the p-value from the Bartlett test of homogeneity for observations grouped by both factors?
    - You can round your answer to 4 decimal digits.

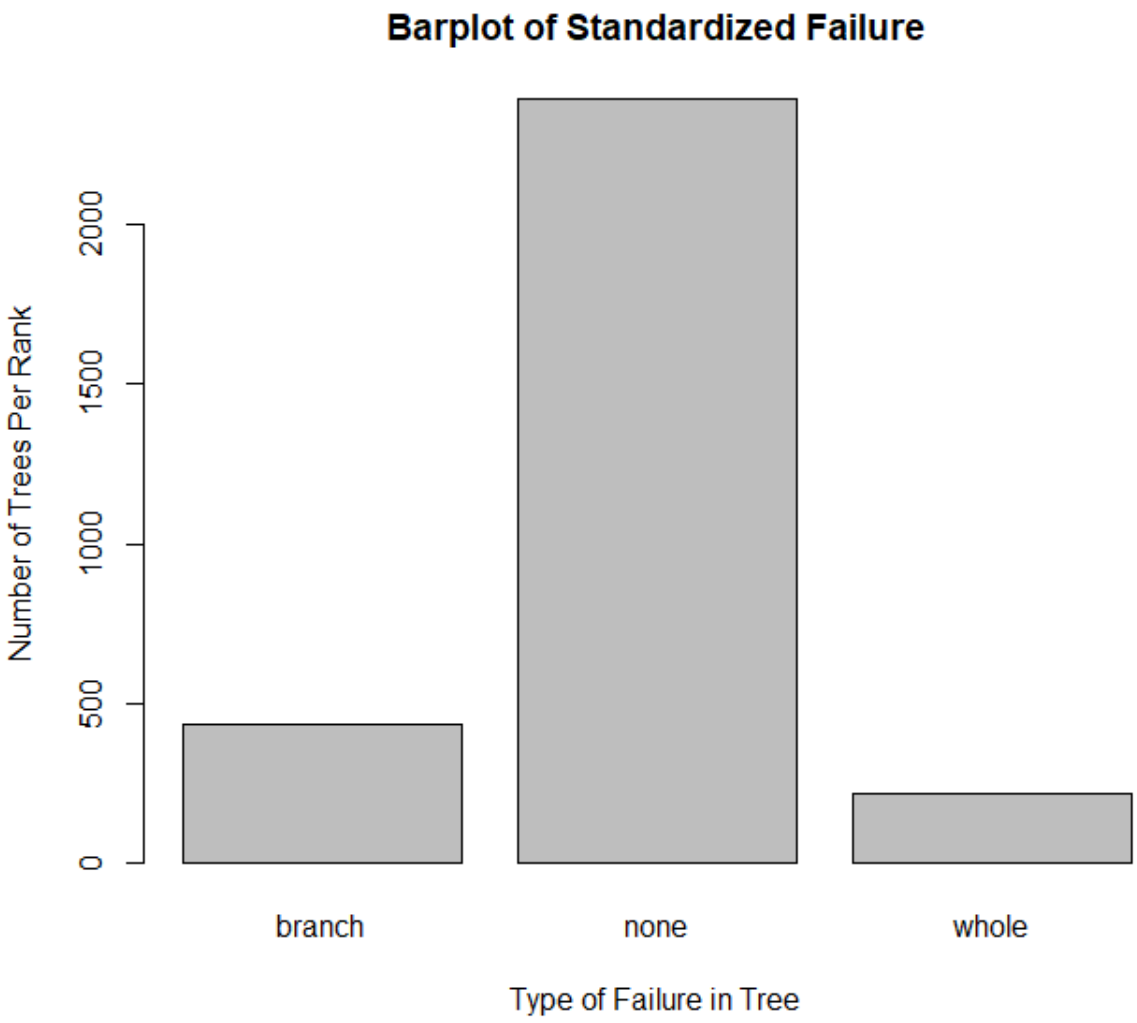
p-value = 0.1741

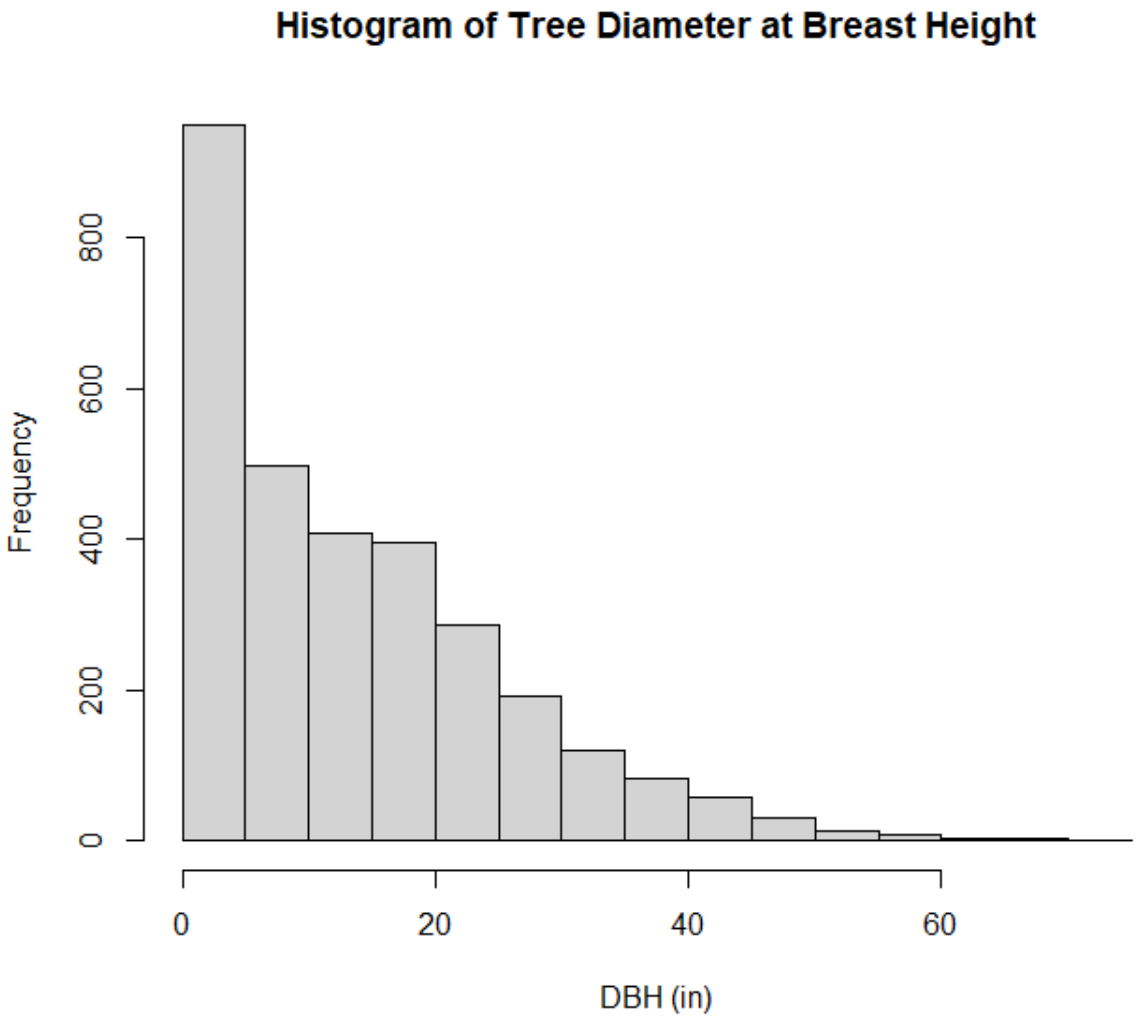
- **Q14 (3 pts.):** Based on the results of the Bartlett tests, do you anticipate any issues with heterogeneity in any of the models?
    - Make sure you justify your response with the results of your tests.

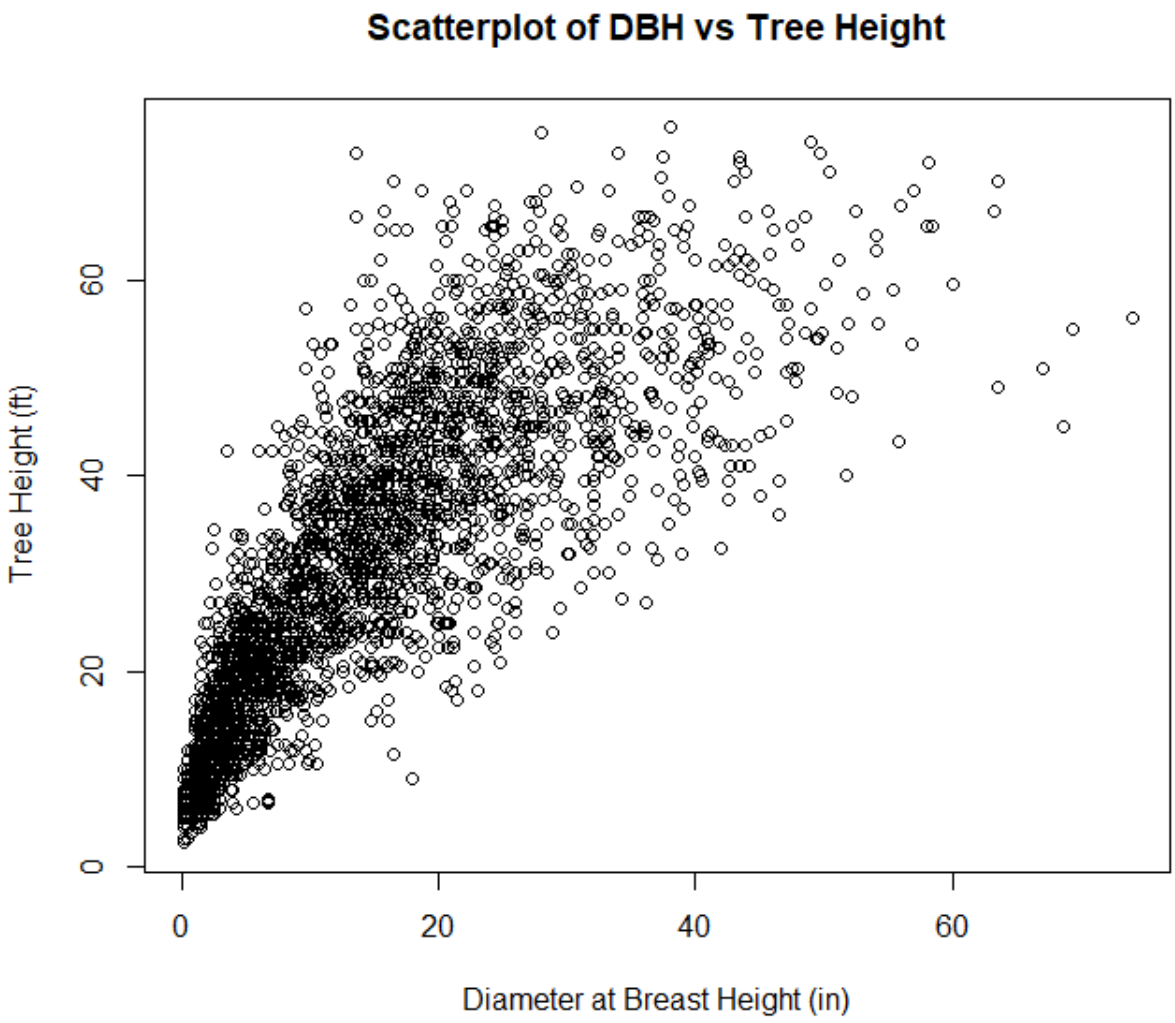
Based on the results of the Bartlett tests, I would expect to have issues with the heterogeneity of the fit_species and fit_both models, as their p-values are large enough that I would not reject the null hypothesis that the the variances are homogeneous, so the only model I would expect to see the heterogeneity in would be the fit_sex model.

- **Q15 (5 pts.):** Perform a graphical exploration of the dataset. Create the following plots and include them in your report. You may create separate figures, or combine them into one multi-panel figure.
    - A barplot of counts of trees in each probability of failure class (column `ProbabilityofFailure`.
    - A barplot of the counts of trees in each of the failure classes (column `Failure_Standardized`)
    - A histogram of DBH
    - A scatterplot of DBH (x-axis) and tree height (y axis)


Barplot of Probability of Failure
Number of Trees Per Rank
0
200
400
600
800
1000
1200
1400
1
2
3
4
Ranked Probability of Failure


Barplot of Standardized Failure
Number of Trees Per Rank
0
500
1000
1500
2000
branch
none
whole
Type of Failure in Tree


Histogram of Tree Diameter at Breast Height
Frequency
0
200
400
600
800
0
20
40
60
DBH (in)




- **Q16 (1 pt.):** State the null hypothesis for the Kolmogorov-Smirnov test. Your answer should be in terms of the DBH of the two groups of trees.

The null hypothesis for the KS test is that there should be no difference in the distribution of DBH of whole and intact trees.

- **Q17 (1 pt.):** What was the p-value of the test? Based on the evidence, do you think the distribution of DBH is the same for the two groups?

p-value = 0.02125

Based on the evidence, I do not think the distribution of DBH between the two groups is the same. I would reject the null hypothesis.

- **Q18 (1 pt.):** Qualitatively describe the shape of the relationship betwen DBH and height. Is it linear? Curved? Monotonic?

Qualitatively speaking, the relationship between DBH and height appears to be a curved relationship.

- **Q19 (1 pt.):** Given your answer to the previous question, which type of correlation coefficient is most appropriate?

I believe the Spearman correlation coefficient is most appropriate for the model.

- **Q20 (1 pt.):** What is the p-value? Do you conclude that the two variables are significantly correlated?

p-value < 2.2e-16

Based on the incredibly low p-value, I would conclude that the two variables are significantly correlated.

- **Q21 (2 pts.):** What was the value of the test statistic (X-squared)? What was the corresponding p-value?

X-squared = 202.65, p-value < 2.2e-16

- **Q22 (1 pt.):** What is the value of the chi-square residual (rounded to the nearest whole number) for the count of failures in probability category 1?

-136

- **Q23 (1 pt.):** Were there more, or fewer, tree failures than expected by chance in failure probability category #1?

There were fewer than expected by chance in failure probability category #1.

- **Q24 (1 pt.):** Were there more, or fewer, tree failures than expected by chance in failure probability category #4?

There were 38 more failures than expected by chance in failure probability category #4.

- **Q25 (2 pts.):** Given your answers to the previous two questions, do you conclude that the probability of failure rating system is effective?

I would say that the probability of failure rating system is likely effective, as the rating system's residuals seem to reflect that the failure rates are not due to random chance. All the categories above category 1, which is the least likely to fail have positive residuals, showing that they indeed fail more often than expected due to random chance.