End of semester:

Using the appropriate tools, answer these basic questions regarding the dataset:

1. How many employees were included in the dataset? How many variables?

307 employees were included in the dataset with 18 variables.

2. Check variable type for each variable and change as needed.

Done in R

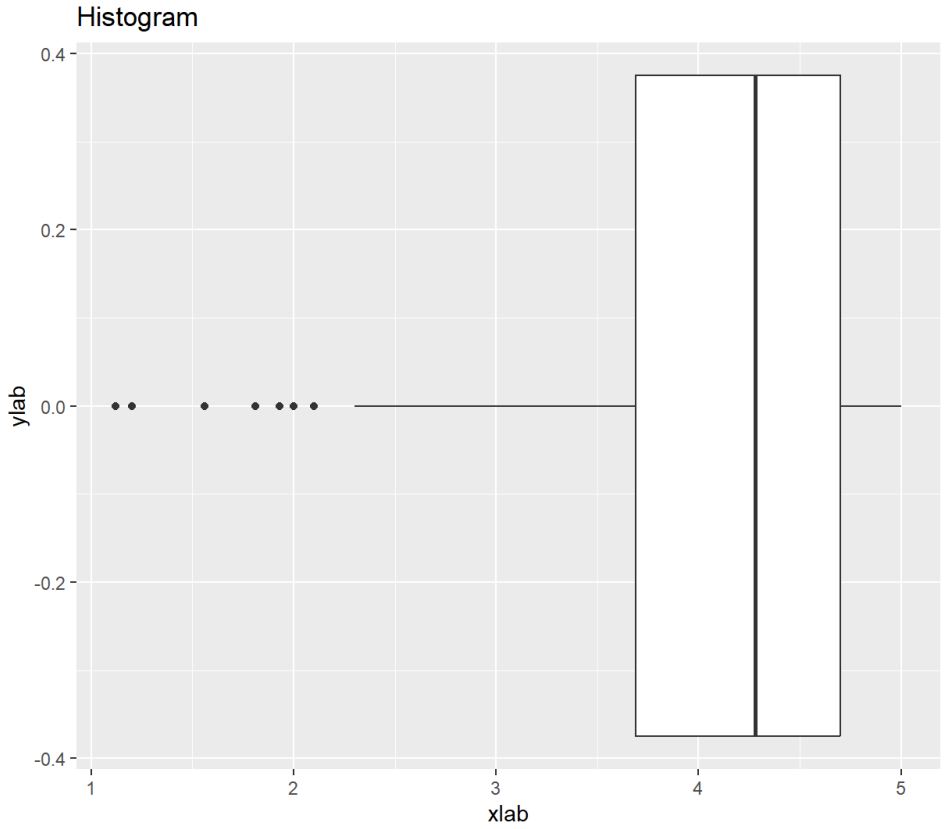
3. What are the different performance ratings for employees? Which is most common?

Exceeds, Fully Meets, Needs Improvement, and PIP

Exceeds	**Fully Meets**
36	**240**
Needs Improvement	**PIP**
18	**13**

→ **Fully Meets is the most common - 240**

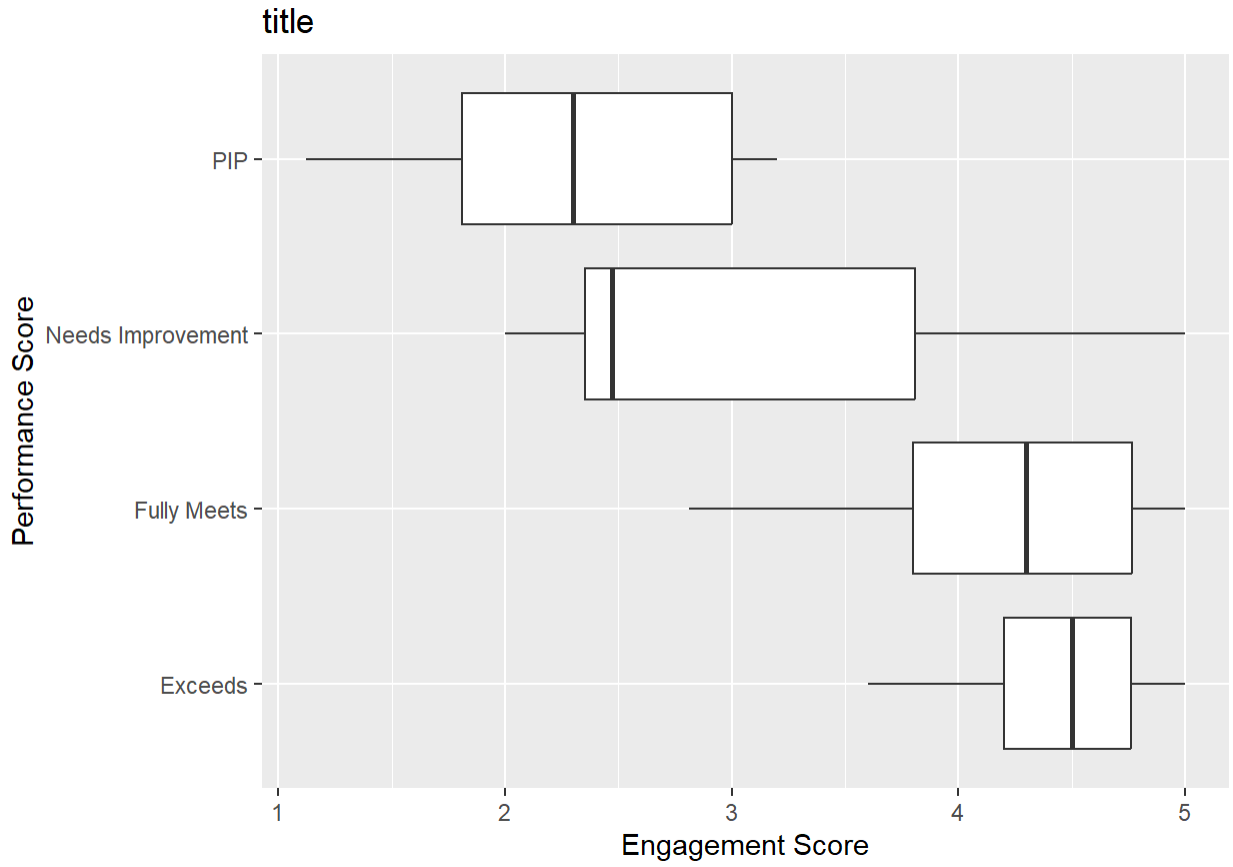
4. Describe the distribution of employee engagement making sure to use an appropriate graphical display and include appropriate summary statistics.



Description: **The distribution of employee engagement is unimodal with a major peak around 4.2 and is skewed to the left due to the presence of 7 outliers with engagement scores from 1-2.3. The median score is 4.28 with an IQR of 1.01.**

min	Q1	median	Q3	max	mean	sd	n
1.12	3.69	4.28	4.7	5	4.108274	0.73741	307

5. Describe the relationship between employee engagement and performance rating. Explain if there appears to be a relationship. Make sure to reference appropriate summary statistics as needed.



There seems to be a relationship between employee engagement and performance rating because the engagement score goes up as an employee's performance level goes up.

6. The above could be used to answer the question: “Does employee engagement have an impact on employee performance?” Based on the data available, come up with a question that could be of interest for the company

Is the relationship between employee engagement and performance consistent with the relationship between employee satisfaction and performance?

QUESTION 1: What are the best recruiting sources if the company wants to ensure a diverse workforce? There are multiple ways to go about answering this question as there are many facets of diversity. We will focus on the more common indicators of diversity but keep in mind there are many ways we could define diversity and answer this question.

1. One aspect is to look at race. What are the different races the employees identify as? Which is most common? (Note: due to a small sample size, employees identifying as Native American or Alaska Native were not included in the dataset.)
White, Black or African American, Two or more races, and Asian.

Asian	**Black or African American**
29	**80**
Two or more races	**White**
11	**187**

→ **White is most common**

2. We are also concerned about how employees were recruited.
a. What are the different recruitment sources? Which are the most common?
LinkedIn, Google Search, Diversity Job Fair, CareerBuilder, Indeed, Employee Referral, On-line Web application, Website, and Other.

CareerBuilder	**Diversity Job Fair**
23	**29**
Employee Referral	**Google Search**
31	**48**
Indeed	**LinkedIn**
85	**75**
On-line Web application	**Other**
1	**2**
Website	**<NA>**
13	**0**

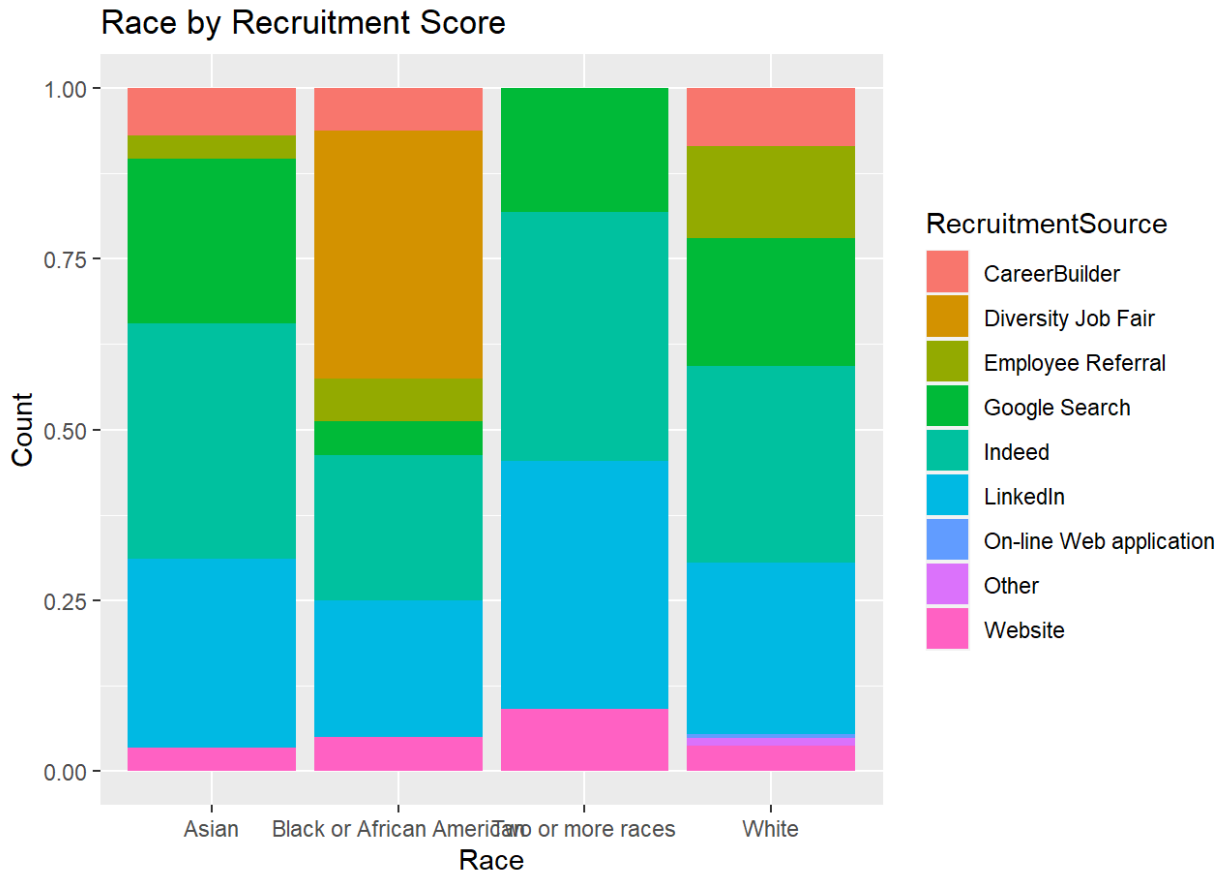
→ **Indeed, LinkedIn are the most common**

b. What percentage of employees were recruited from a diversity job fair?

```
> prop.table(rec_prop)

          CareerBuilder      Diversity Job Fair
             0.074918567             0.094462541
      Employee Referral          Google Search
             0.100977199             0.156351792
                Indeed                 LinkedIn
             0.276872964             0.244299674
On-line Web application                  Other
             0.003257329             0.006514658
               Website
             0.042345277
```

94.45%

3. Let's next look at the relationship between recruitment source and race.
a. Using the appropriate graphical display and summary statistics, describe the relationship between race and recruitment source.



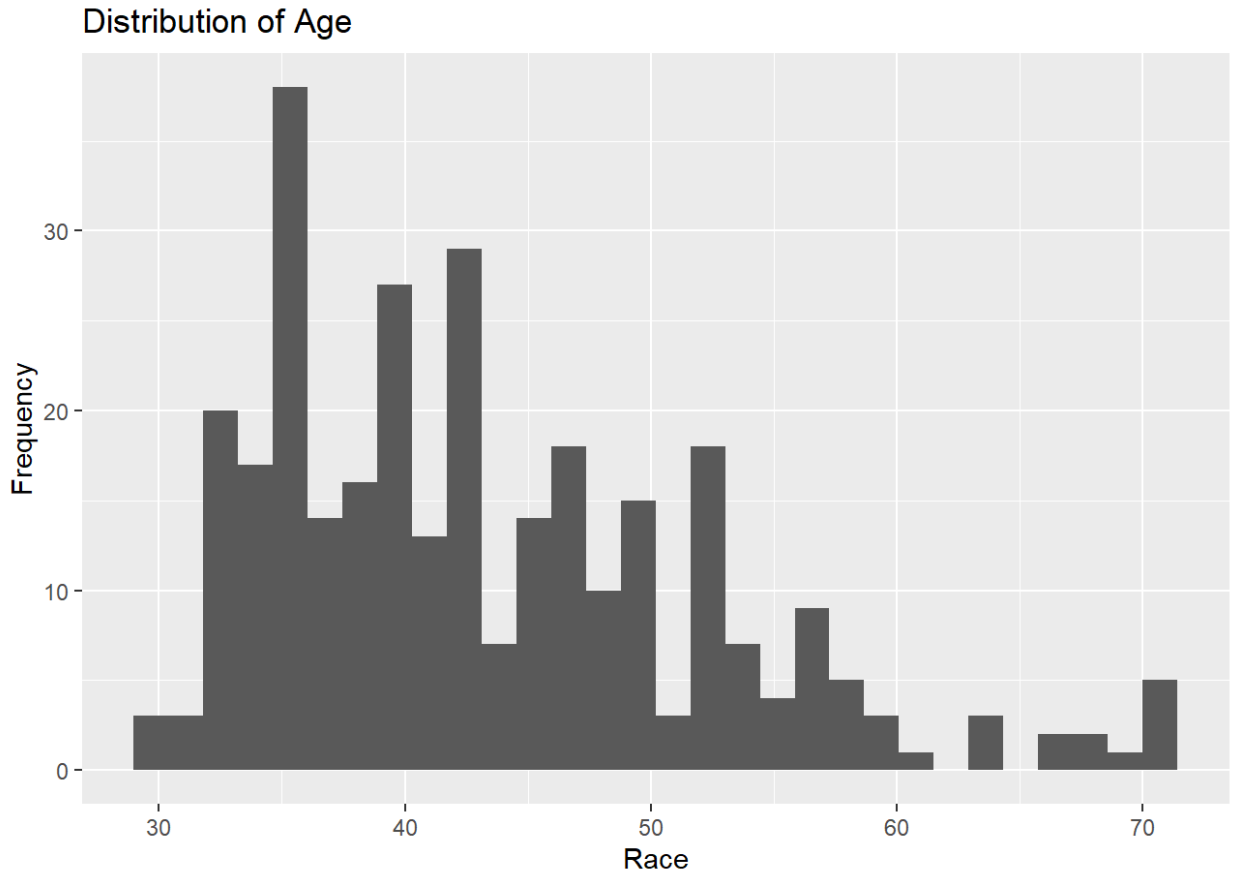
The use of LinkedIn (high), Indeed (high) and Website (low), are consistent amongst races, while Diversity Job Fair (high) is only used by the Black or African American.

b. Based on this, which source do you recommend the company use if they want to ensure a racially diverse workforce?

LinkedIn and Indeed

c. Based on these results, would you recommend the company continue recruiting at diversity job fairs? Why or why not?
Yes, because this is where most Black or African American people search and apply for jobs.

4. Another aspect of diversity is age. Describe the distribution of age for this company making sure to use an appropriate plot and summary statistics.



The histogram is unimodal with its peak at the race value at around 35, is skewed to the right with potential outliers after the race value of 60.

min	Q1	median	Q3	max	mean	sd	n
30	36	42	49	71	43.43	8.89	307

5. Let's now look at the relationship between age and recruitment source.
a. Using the appropriate graphical display and summary statistics, describe the relationship between age and recruitment source.

```
> favstats(hr$Age~hr$RecruitmentSource)
     hr$RecruitmentSource min Q1 median   Q3 max     mean
1           CareerBuilder  32 36     43 49.0  70 44.60870
2      Diversity Job Fair  33 38     48 53.0  67 46.44828
3       Employee Referral  34 39     43 52.0  61 45.19355
4           Google Search  31 38     41 49.0  71 44.00000
5                 Indeed  30 35     42 49.0  68 42.49412
6               LinkedIn  30 36     40 45.5  70 42.33333
7 On-line Web application  45 45     45 45.0  45 45.00000
8                  Other  43 44     45 46.0  47 45.00000
9                Website  33 34     38 47.0  57 40.38462
         sd  n missing
1 11.130381 23       0
2  9.120566 29       0
3  7.560509 31       0
4 10.191778 48       0
5  8.560184 85       0
6  8.248396 75       0
7        NA  1       0
8  2.828427  2       0
9  7.377200 13       0
```

Most sources are approached by pretty much the same range of age. For On-line Web applications and Other, they have a significantly lower number of people. Additionally, there are a few potential outliers that represent people of ages that far exceed the majority.

b. Based on this, if the company would like to recruit younger workers, what recruitment sources do you recommend?
Websites, Indeed, CareerBuilder

Question 2: Do more engaged employees have fewer absences? To address this question, answer the following:

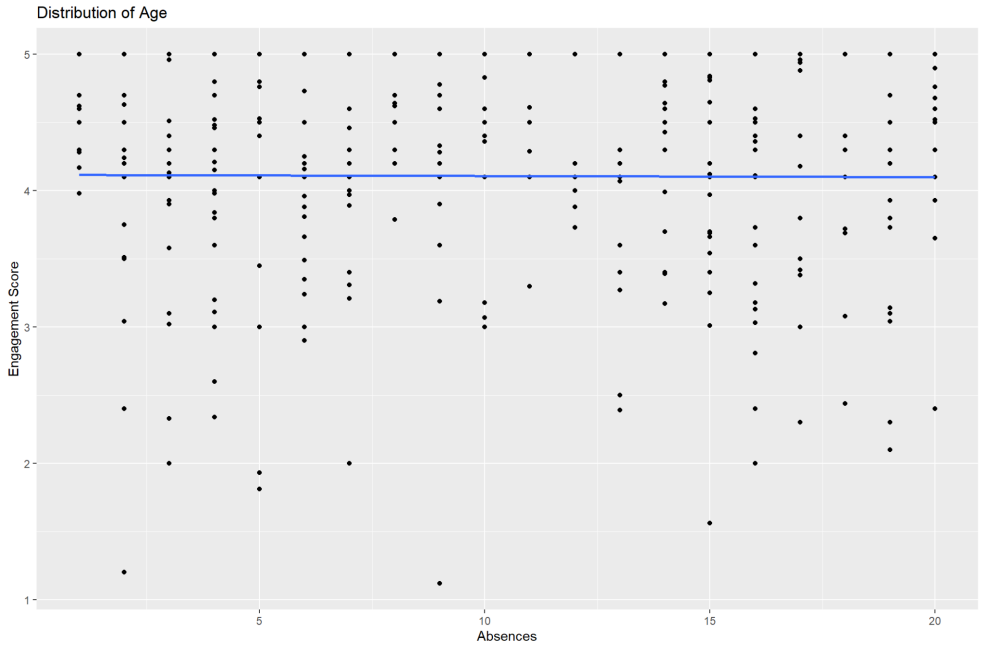
1. First let's look at the number of absences.
a. What is the largest number of absences?
20

b. Report the appropriate measure of center and spread to describe the number of absences.

min	Q1	median	Q3	max	mean	sd	n
1	5	10	15	20	10.29642	5.84719	307

Amongst the 307 samples, the lowest number of absences is 1 and highest is 20 with a median and IQR of 10.

2. Employee engagement is likely related to the number of absences of an employee in that more engaged employees will have fewer absences.
a. Using the appropriate plot, explain if there appears to be a relationship between employee engagement and number of absences.



Based on the plot, there is no relationship between the number of absences and the employees' engagement score.

b. To examine this potential relationship, which tool(s) that we have learned about this semester would be appropriate to use? Why?
We use cor() to calculate the correlation between the two variables. Because we can determine how strong the relationship is based on the r value.

3. Though there are several tools we could use, let's consider using linear regression. Fit the linear regression with number of absences as the response and employee engagement, performance rating, age, and sex as predictors.

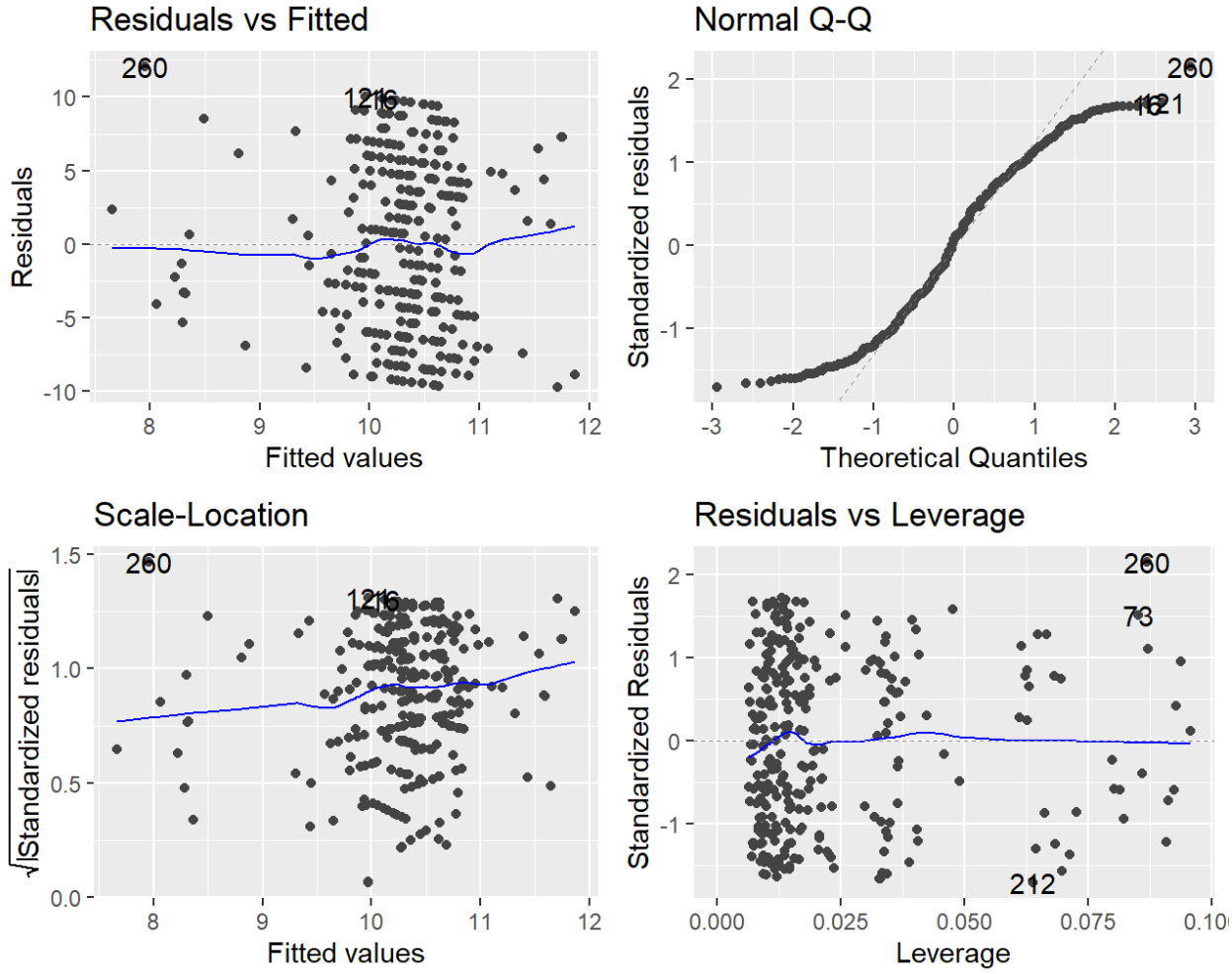
```
Call:
lm(formula = Absences ~ EngagementSurvey + PerformanceScore +
    Age + Sex, data = hr)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7112 -5.1189  0.3827  4.8412 12.0466

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     13.14841    3.11961   4.215 3.31e-05 ***
EngagementSurvey                -0.33031    0.54840  -0.602    0.547
PerformanceScoreFully Meets     -0.28412    1.05989  -0.268    0.789
PerformanceScoreNeeds Improvement  0.24341    1.88820   0.129    0.898
PerformanceScorePIP             -2.94929    2.27159  -1.298    0.195
Age                             -0.02691    0.03800  -0.708    0.479
SexM                             0.01452    0.68006   0.021    0.983
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.876 on 300 degrees of freedom
Multiple R-squared:  0.009803,  Adjusted R-squared:  -0.01
F-statistic: 0.495 on 6 and 300 DF,  p-value: 0.812
```

a. Report the estimated regression line.
Engagement-hat = 13.15 - 0.33EngagementSurvey - 0.28PerformaceMeets - 0.24PerformanceImprov - 2.95PerformancePip - 0.03Age + 0.01SexM

b. What level of performance rating is the baseline level?
PerformanceScoreNeeds Improvement

c. Using the appropriate tool, examine if there is evidence to suggest that more engaged employees have fewer absences.
r = -0.006142051
→ There is no linear relationship between the two variables, therefore, there is no evidence to suggest that more engaged employees have fewer absences.

d. For the results of our analysis to be reliable there are four assumptions that must be reasonably satisfied. Carefully explain if these assumptions are satisfied.



Assumptions:

- Linearity: **Using residuals plot, there does not appear to be a flat smoother so this assumption is not satisfied.**
- Independence: **According to the CodeBook, this is a synthetic dataset created specifically to go along with the case study. Therefore, this assumption is not satisfied.**
- Normality: **Using the normal Q-Q plot, the data points closely follow the dashed line, but the two tails are too astray, therefore, this assumption is not satisfied.**
- Equal Variances: **This assumption is satisfied because the distribution is pretty much parallel, and evenly distributed.**

4. Look at the p-values for each predictor in our model.

a. What do you notice?

The p-value for y-intercept is <0.0001

All other variables' p-values are >0.05

b. This suggests that this model is not helpful in predicting the number of absences. What else should be considered that could predict the number of absences? Make sure to explain your reasoning.
EmpSatisfaction and Absences has a correlation value of r = 0.07890464, which is very close to 1 and suggests that there is a strong correlation between the two variables and it would be helpful in predicting the number of absences based on the employees' Satisfaction.

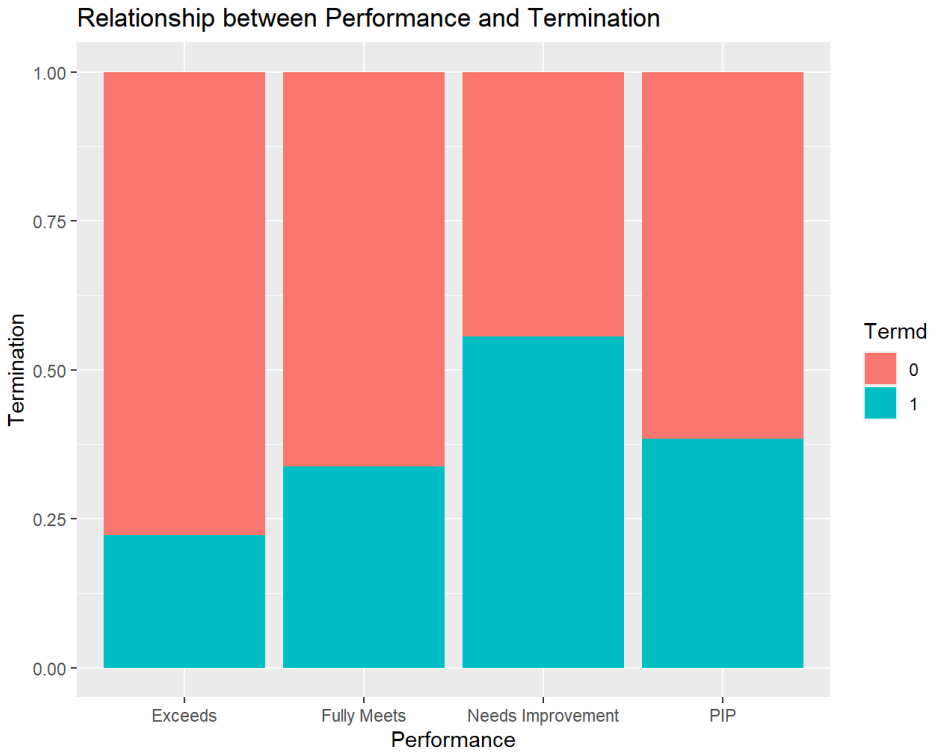
Question 3: Can we predict who will be terminated? To answer this question, complete the following:

1. What percentage of employees in the dataset were terminated?
33.88% of employees were terminated.

2. What are the different reasons for termination?
[1] "N/A-StillEmployed"
[2] "career change"
[3] "hours"
[4] "return to school"
[5] "Another position"
[6] "unhappy"
[7] "attendance"
[8] "performance"
[9] "Learned that he is a gangster"
[10] "retiring"
[11] "relocation out of area"
[12] "more money"
[13] "military"
[14] "no-call, no-show"
[15] "Fatal attraction"
[16] "maternity leave - did not return"
[17] "medical issues"
[18] "gross misconduct"

3. One variable that may impact whether someone is terminated is their performance. Create the appropriate plot to examine the relationship between **performance and termination** (Note: use **Termd** not TermReason) and explain if there is evidence to suggest a relationship.

I think there is evidence to suggest a relationship between performance and termination because most people at the level of PIP or the worst score, get terminated and as the performance requirement gets harsher (from Needs improvement to Exceeds), the amount of employees getting terminated increases.

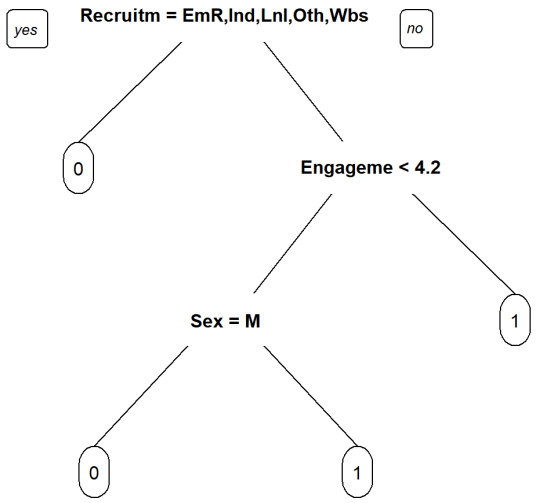


4. In addition to performance, what other variables in the dataset do you think may predict whether an employee is terminated? Why?
SpecialProjectsCount, Absences, and EmpSatisfaction may predict whether an employee is terminated because the employer would look at how many projects got done by the employee, how many absences they had, and their attitude towards their work.

5. Create a new dataset that removes TermReason and EmploymentStatus (as these will not be useful predictor variables).
Done

6. Which tool(s) that we have discussed this semester would be most appropriate to use for this analysis? Why?
Decision tree would be the most appropriate because by following a pruned tree, we can easily figure out the output of many different variables.

7. Though there are multiple tools we could use, let's use a decision tree. Go through the process we have discussed this semester to obtain the best pruned tree. Use this tree to answer the following:



a. Which variables are useful in predicting whether an employee is terminated?
RecruitmentSource, EngagementSurvey, and Sex

b. How well does your model perform (e.g. how accurate it is?)?
It is 69.23% accurate

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 47 15
         1 13 16

               Accuracy : 0.6923
                 95% CI : (0.5868, 0.7849)
    No Information Rate : 0.6593
    P-Value [Acc > NIR] : 0.2931

                  Kappa : 0.3042

 Mcnemar's Test P-Value : 0.8501

            Sensitivity : 0.5161
            Specificity : 0.7833
         Pos Pred Value : 0.5517
         Neg Pred Value : 0.7581
             Prevalence : 0.3407
         Detection Rate : 0.1758
   Detection Prevalence : 0.3187
      Balanced Accuracy : 0.6497

       'Positive' Class : 1
```

c. Would you recommend using this tree to predict who will or will not be terminated? Why or why not? In other words, can we really predict who will be terminated?

I think it is better for this model to predict who will not be terminated because its Specificity is 78.33%, which is 78.33% accurate to predict. On the other hand, its Sensitivity is only 51.61% which is very low to successfully predict who will be terminated.

8. In our analysis we had two outcomes: **terminated or not terminated**. The terminated outcome contains employees that were voluntarily terminated (e.g. **quit, retired**) and those that were terminated with cause (e.g. **fired**). Do you think the results of our analysis would differ if we considered three outcomes (not terminated, voluntarily terminated, terminated with cause)? Why?

Yes, because other predictor variables other than what we considered, will have different effects on the different outcomes. And, since there are more predictor variables involved, the accuracy in predicting different outcomes might fluctuate.