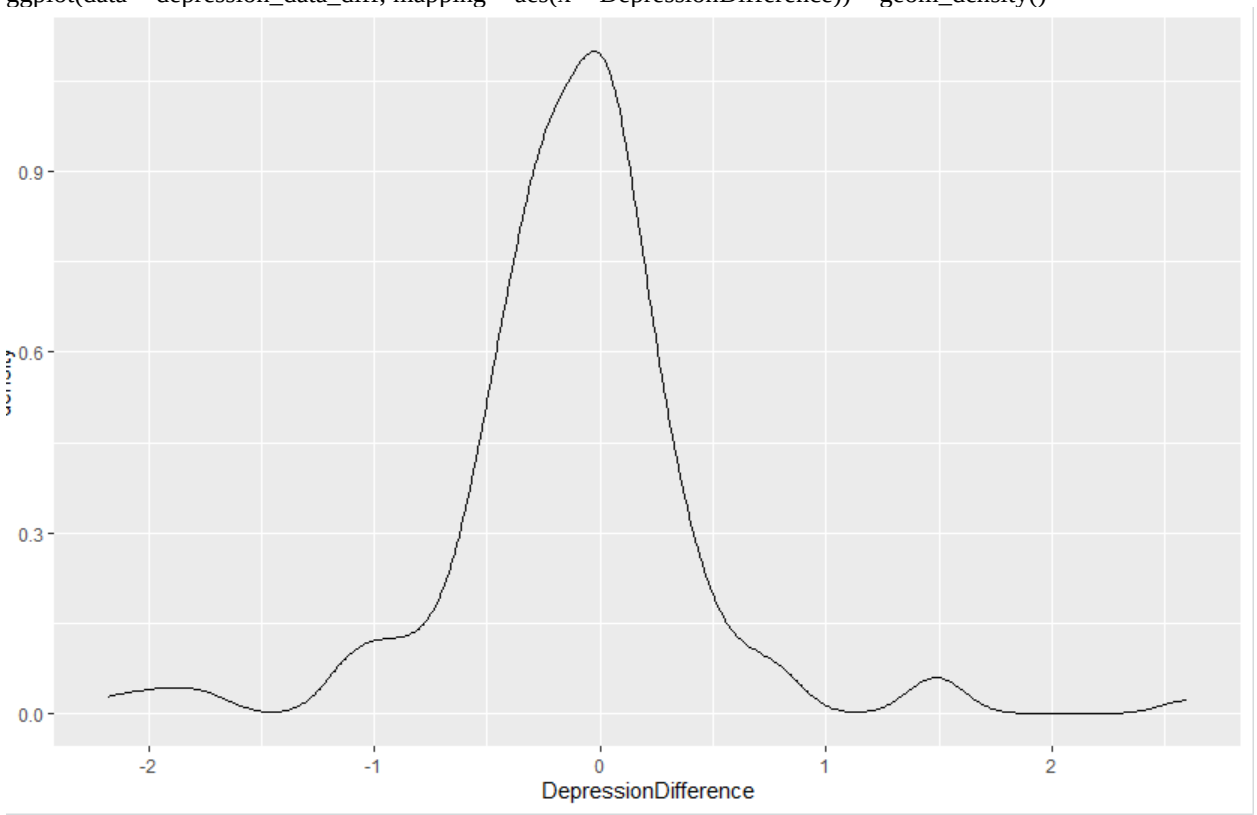
Using this new calculated **difference** variable column you just created, **answer the following questions**:

1. **Plot your data** using the geom_density() ggplot function. Does it look normal?

```
ggplot(data = depression_data_diff, mapping = aes(x = DepressionDifference)) + geom_density()
```



It looks slightly normal as the mean is near to zero and slightly right-skewed but relatively close to zero whi ch suggests that the distribution is near to normally distributed. Meanwhile, it has sharper peaks and heavier tails suggesting that the kurtosis value is higher. So, the over examination shows that it does not follow a pe rfectly normal distribution.

2. Using your descriptive statistics skills in R, **fully describe this "difference" data** (include whatever metrics you think are appropriate to fully describe the nature of this data).

summary(depression_data_diff$DepressionDifference)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-2.17609 -0.35230 -0.09045 -0.11807  0.12378  2.60004       7
```

pivot(depression_data_diff, c(IQR, skew, kurtosis, mean, sd, var), DepressionDifference)

```
 n na   IQR  skew  kurt   mean    sd   var
 162  7 0.476 0.208 5.981 -0.118 0.553 0.305
```

The table including all

Statistic	Value
n	162
NA	7
Min.	-2.17609
1st Qu.	-0.35230
Median	-0.09045
Mean	-0.11807
3rd Qu.	0.12378
Max.	2.60004
IQR	0.476
Skew	0.208
Kurt	5.981
SD	0.553
Var	0.305

The "difference" data, derived from the "after" and "before" depression rating scores, consists of 162 observations. Seven values are missing in this dataset. The differences range from -2.17609 to 2.60004, with a mean of approximately -0.11807 and a median of -0.09045, indicating a slight negative skew. The interquartile range is approximately 0.476, suggesting variability in the middle 50% of the data. The distribution exhibits positive skewness (0.208) and high kurtosis (5.981), indicating a peaked distribution with heavy tails compared to a normal distribution. The standard deviation is approximately 0.553, indicating moderate variability around the mean. The variance is approximately 0.305, providing additional insight into the spread of the data.

3. **Is this data normally distributed**? Make note of the mean and standard deviation for this data. You will need it moving forward. Now let's test some probabilities.

   While the mean (-0.11807) and median (-0.09045) being close to each other may suggest some semblance of normal distribution, the high kurtosis value (5.981) indicates heavy tails and potential outliers, diverging from the typical bell curve shape. Additionally, the positive skewness (0.208) suggests a slight right skew, further deviating from normality. Thus, despite the mean and median proximity, the combination of high kurtosis and skewness, along with the observed variability, suggests that the data is not perfectly normally distributed.

4. Now use your data description results to answer the following questions:
   - **What is the probability** of seeing a person in the general population with an increase of more than 1 depression unit (i.e. measured value (X) for your difference column is p(X>1).

     Zscore are used for normal distribution.

     #z_score <- (value-mean)/sd

     > (1-(-0.118))/0.305

     [1] 3.665574

     #calculate probability for normal distribution using pnorm

     > pnorm(3.665574, mean = 0, sd = 1, lower.tail = TRUE)

     [1] 0.9998766

     #lower tail = F mean less than of X=1, if set to T mean greater than given X)

```
#P(X>=1)
> pnorm(q=1, mean=-0.11807 , sd=0.553, lower.tail = F)
[1] 0.02159714
```

2.17

- Based on this data, **what is the probability** of seeing someone decrease in their depression rating (i.e. measured value (X) for your difference column is p(X< 0). Now, let's consider the power of this test.

```
#P(X<0)

pnorm(q=0, mean=-0.11807 , sd=0.553, lower.tail = T)

[1] 0.5845347
```

41.55%

- If we wanted to flag people whose depression increased significantly (using a 1-tailed 0.05 alpha threshold) **what is the depression value** we would use as the threshold to flag people for additional follow-up?

1-tailed: This refers to the type of hypothesis test being conducted. A one-tailed test focuses on the possibility of the data falling entirely in one direction of the distribution (either above or below a certain value), rather than both directions. It's often used when there is a specific direction of interest in the hypothesis being tested.

0.05 alpha threshold: Alpha (α) is the significance level used in hypothesis testing. It represents the probability of incorrectly rejecting the null hypothesis when it is actually true. In this case, an alpha level of 0.05 (or 5%) is commonly used, indicating a willingness to accept a 5% chance of making a Type I error (incorrectly rejecting a true null hypothesis).

The qnorm function in R is used to calculate quantiles of a normal distribution. In the context of hypothesis testing and determining a threshold value, qnorm is used to find the value in the distribution of the "difference" data column (which represents the change in depression ratings) that corresponds to a specific probability or percentile.

```
# Calculate the threshold value for a one-tailed 0.05 alpha threshold
```

```
> threshold_value <- qnorm(p = 0.95, mean = -0.11807, sd = 0.553, lower.tail
= FALSE)
> threshold_value
[1] -1.027674
```

0.79