The objective of this project is to test for normality of given dataset

I am going to use the `garden` data, with the measurement of ozone concentrations in 3 different areas.

> Objective: the idea is to display the distribution of all three series and look at the shapes.

1.1 Loading the data

```
garden <- read_delim("garden.csv")
```

```
## Rows: 10 Columns: 3
```

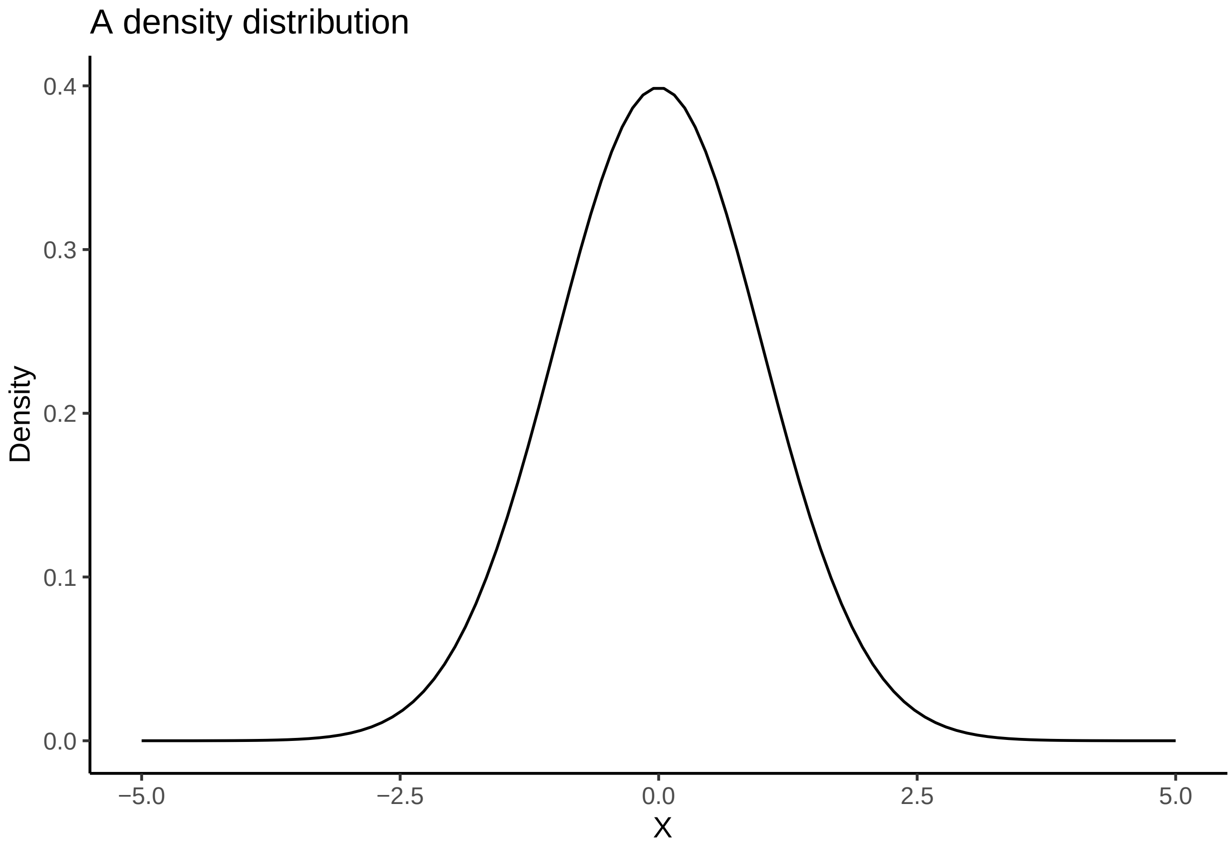
```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): gardenA, gardenB, gardenC
```

```
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
garden
```

```
## # A tibble: 10 x 3
##    gardenA gardenB gardenC
##      <dbl>   <dbl>   <dbl>
##  1       3       5       3
##  2       4       5       3
##  3       4       6       2
##  4       3       7       1
##  5       2       4      10
##  6       3       4       4
##  7       1       3       3
##  8       3       5      11
##  9       5       6       3
## 10       2       5      10
```

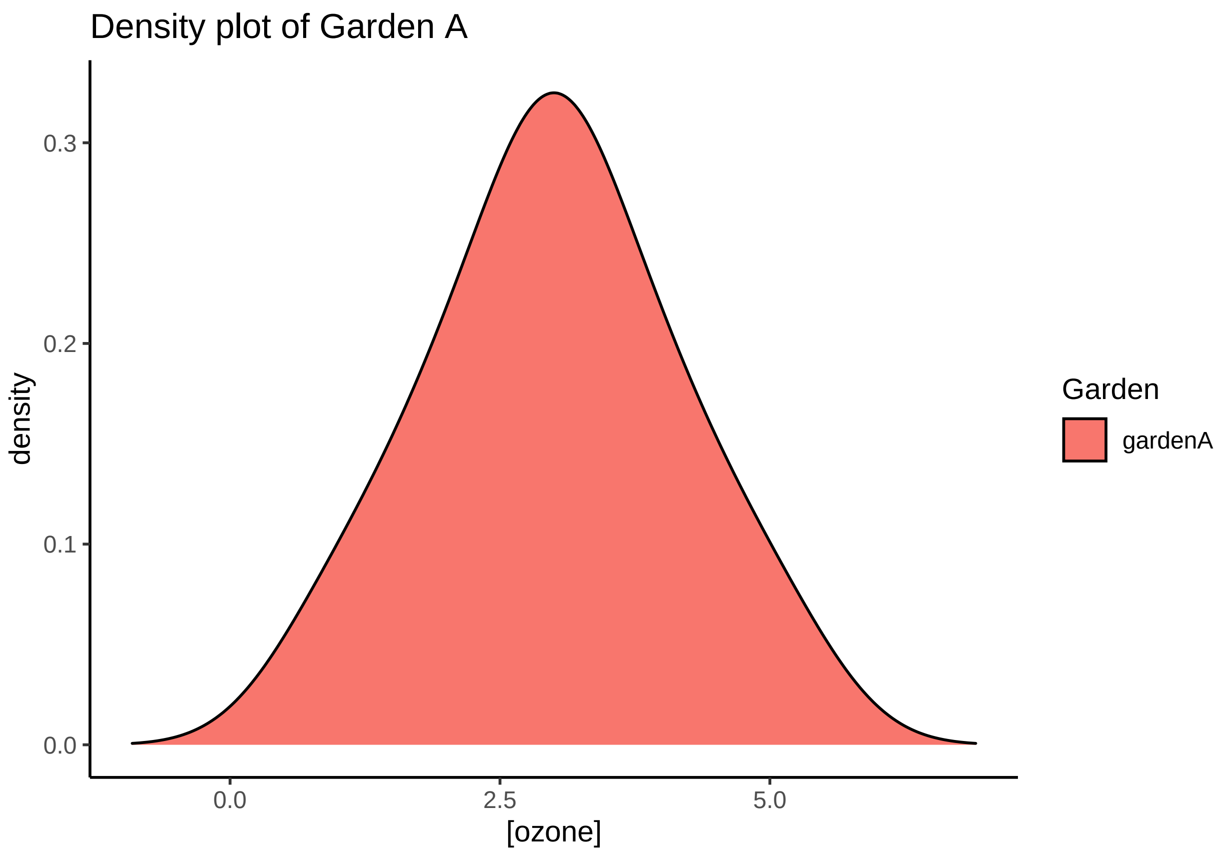
```
# A little of exploration of plotting a normal distribution
x <- seq(from = -5, to = 5, length.out = 100)
dens_x <- dnorm(x)

ggplot() +
  geom_line(aes(x = x,
                y = dens_x)) +
  labs(title = "A density distribution",
       y = "Density",
       x = " X") + theme_classic()
```



It's now time to apply a density function to the provided data.

I will start by drawing the density plot of gardenA using `ggplot2`.

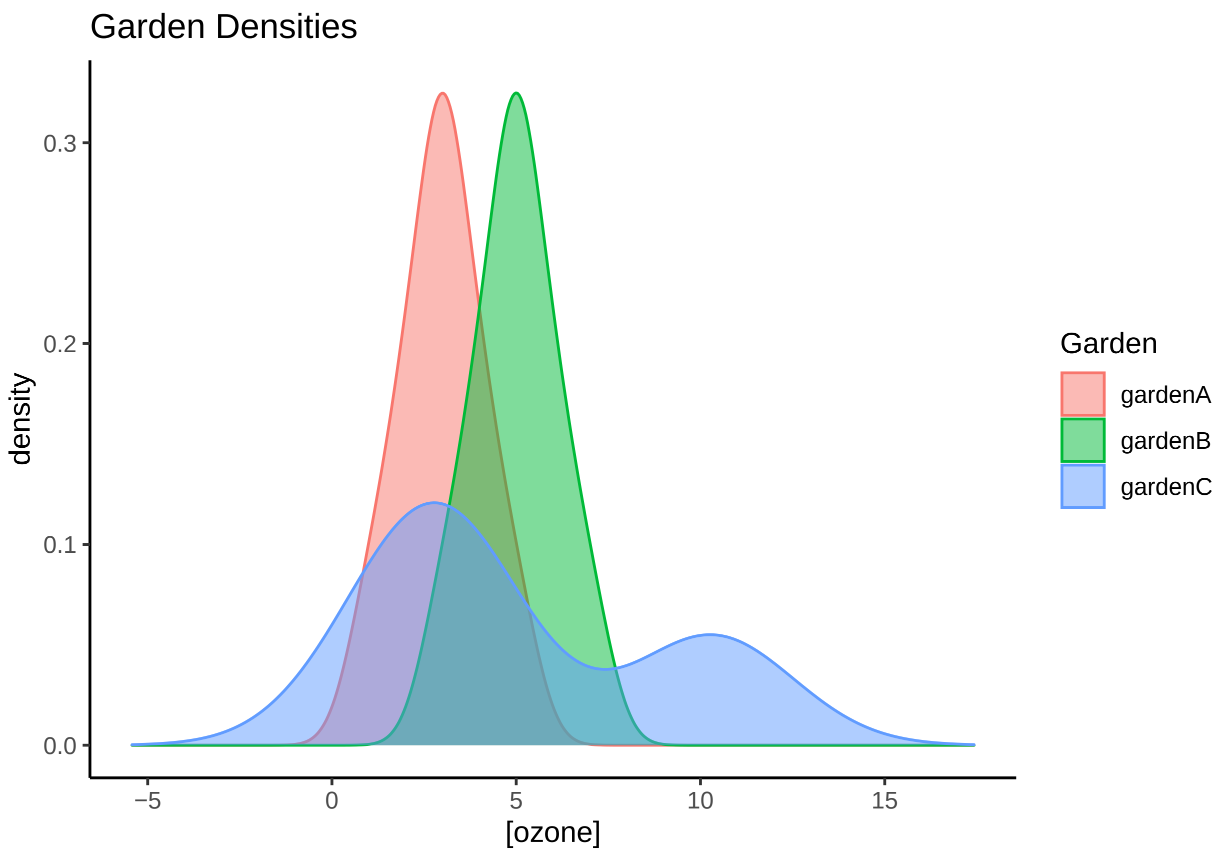
```
# Write your answer here
dens_a <- tidy(density(garden$gardenA))

garden %>%
  pivot_longer(everything(),
               names_to = "Garden",
               values_to = "ozone") -> garden_2
garden_2 %>%
  filter(Garden == "gardenA") %>%
  ggplot() +
  geom_density(aes(x = ozone, fill = Garden)) +
  scale_x_continuous(limits = c(min(dens_a$x), max(dens_a$x))) +
  theme_classic() +
  labs(title = "Density plot of Garden A",
       x = "[ozone]",
       y = "density")
```



```
# I would like to get limits i.e minimum and maximum of the density from each Garden
garden_2 %>%
  group_by(Garden) %>%
  summarise(min = min(density(ozone)$x),
            max = max(density(ozone)$x)) -> limits_x
limits_x
```

```
## # A tibble: 3 x 3
##   Garden     min   max
##   <chr>    <dbl> <dbl>
## 1 gardenA -0.907  6.91
## 2 gardenB  1.09   8.91
## 3 gardenC -5.42  17.4
```

```
# Use the limits from the density, I plot the densities of each garden
garden_2 %>%
  ggplot(aes(x = ozone, color = Garden, fill = Garden)) +
  geom_density(alpha = 0.5) +
  scale_x_continuous(limits = c(min(limits_x$min), max(limits_x$max))) +
  theme_classic() +
  labs(title = "Garden Densities",
       x = "[ozone]",
       y = "density")
```



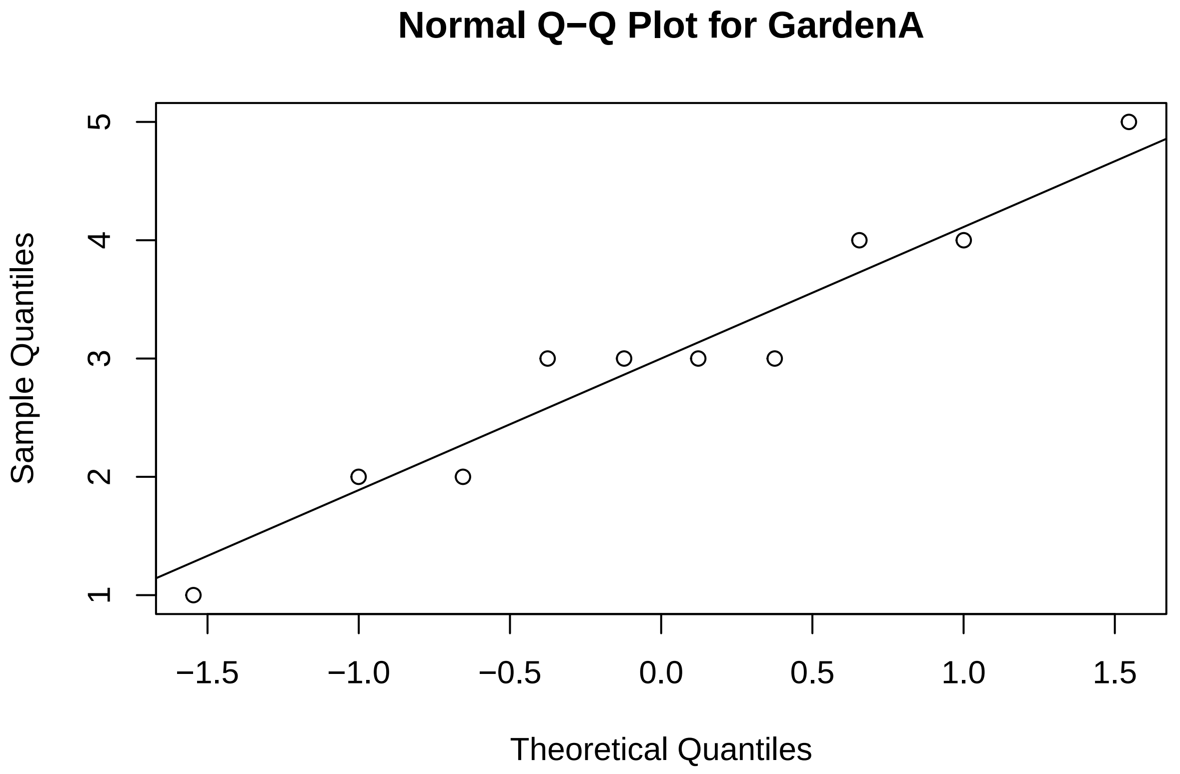
```
# Some observations from the density plot

# The empirical distributions of gardenA and gardenB have a bell-shaped curve which
# suggests it follows a normal distribution.
# That of gardenC is a double bell curve. This suggests it's a bimodal distribution.
```

Plotting Quantile-Quantile comparison to the normal distribution

A Q-Q plot allows you to see how the quantiles of two distributions fit. In our case, we want to compare each of the garden's distribution to the normal distribution.

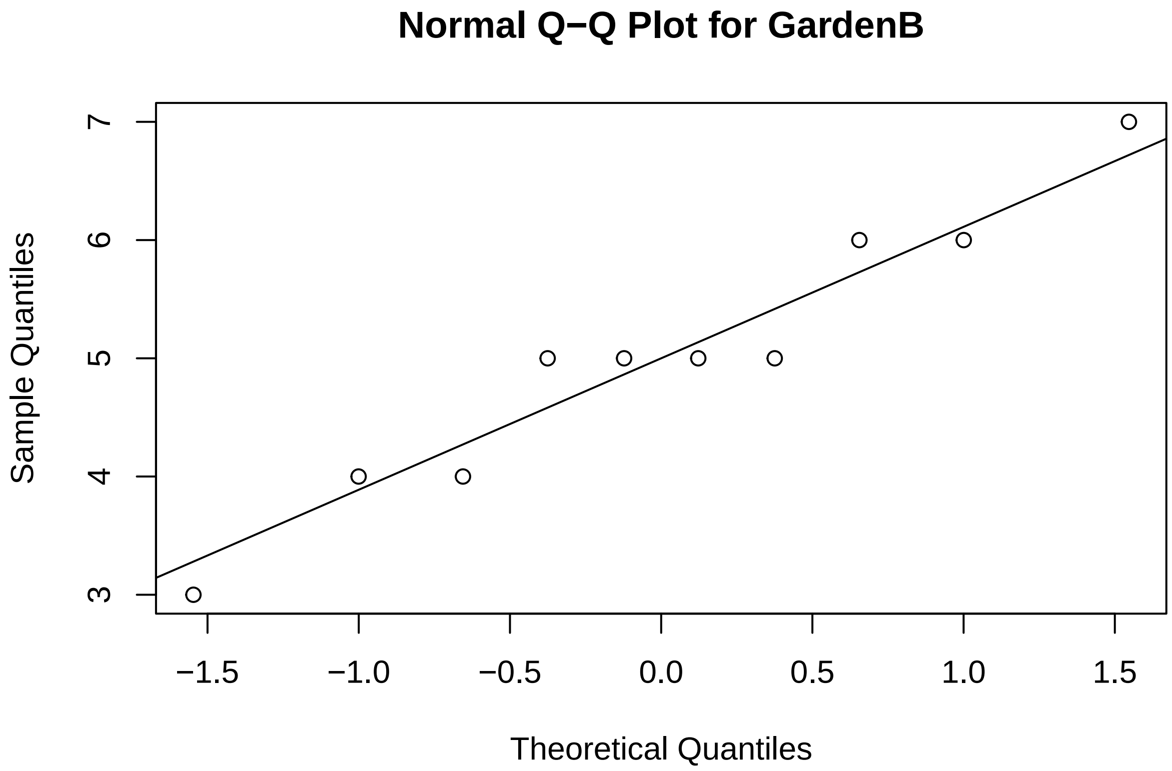
METHOD: I use the function `qqnorm()` and add the ideal line with `qqline()`. When the dots are close to the plain line, then the distribution follows the normal one.

```
# Write your answer here
# Q-Q plot for garden A
qqnorm(garden$gardenA, main = "Normal Q-Q Plot for GardenA")
qqline(garden$gardenA)
```



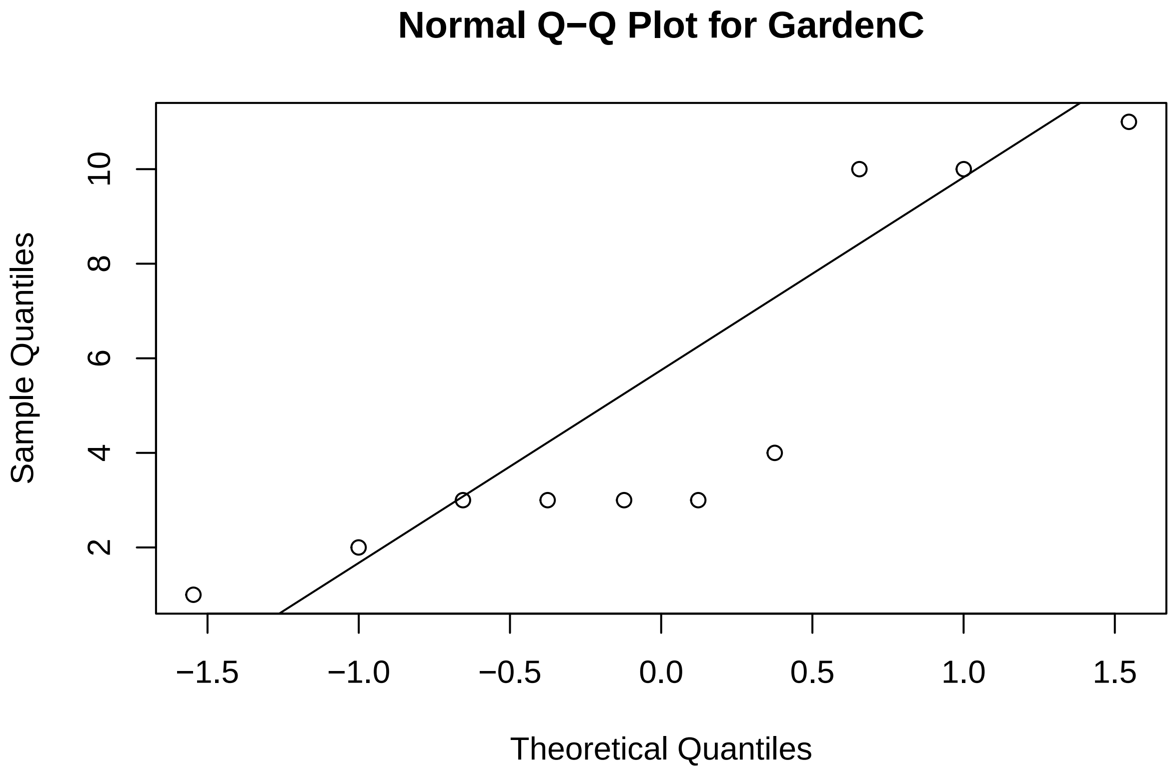
```
# Observation
# The dots are close to the plain line. Means the distribution follows the normal one.
```

```
# Q-Q plot for garden B
qqnorm(garden$gardenB, main = "Normal Q-Q Plot for GardenB")
qqline(garden$gardenB)
```



```
# Observation
# The dots are close to the plain line. Means the distribution follows the normal one.
```

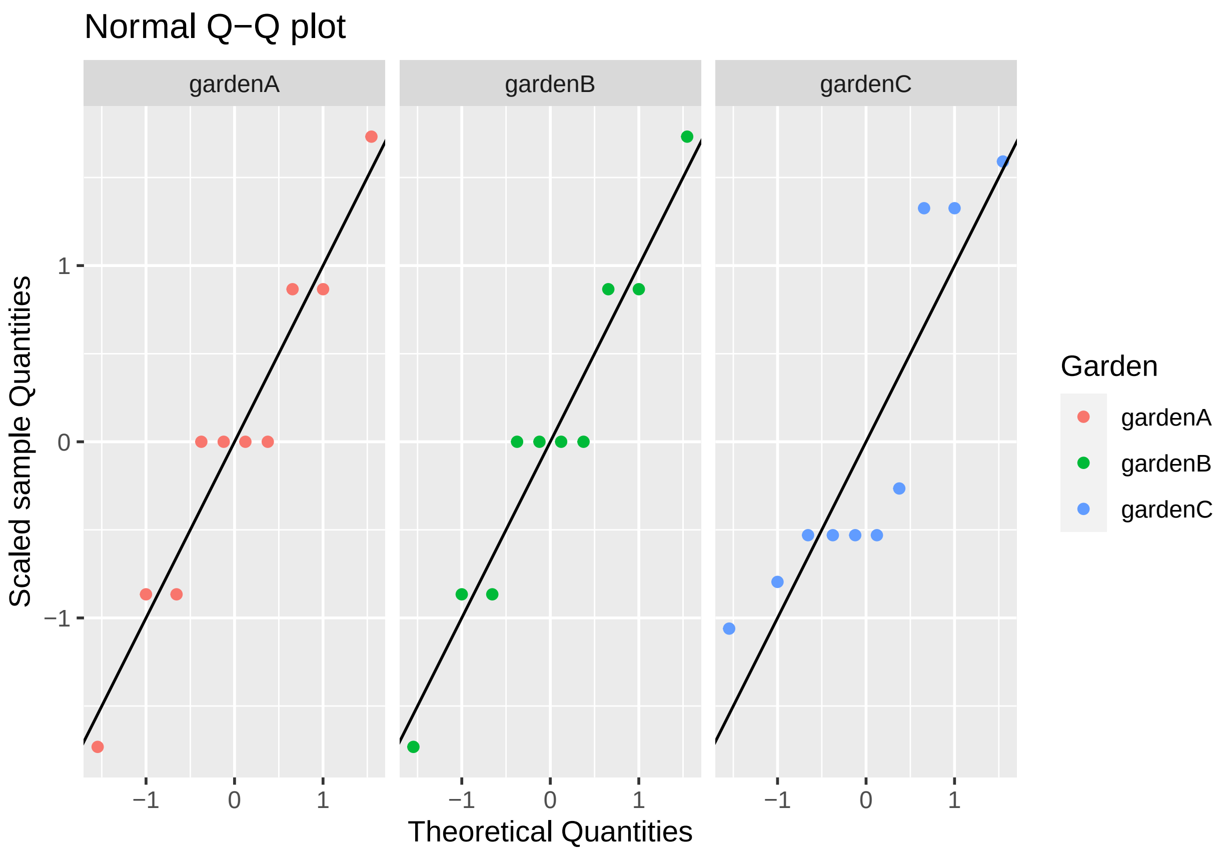
```
# Q-Q plot for garden C
qqnorm(garden$gardenC, main = "Normal Q-Q Plot for GardenC")
qqline(garden$gardenC)
```



```
# Observation
# The dots are not close to the plain line. Means the distribution does not follow the normal one.
```

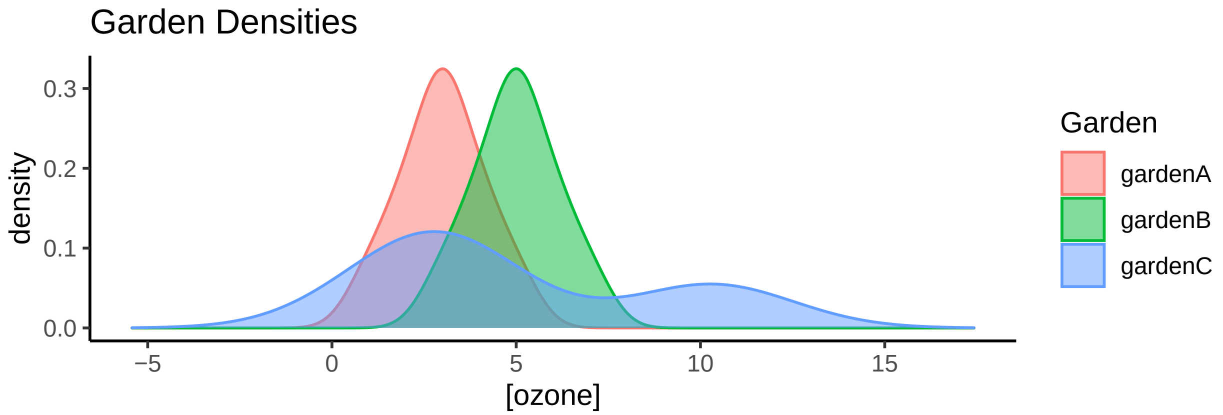
Next step is to draw the Q-Q plot with ggplot.

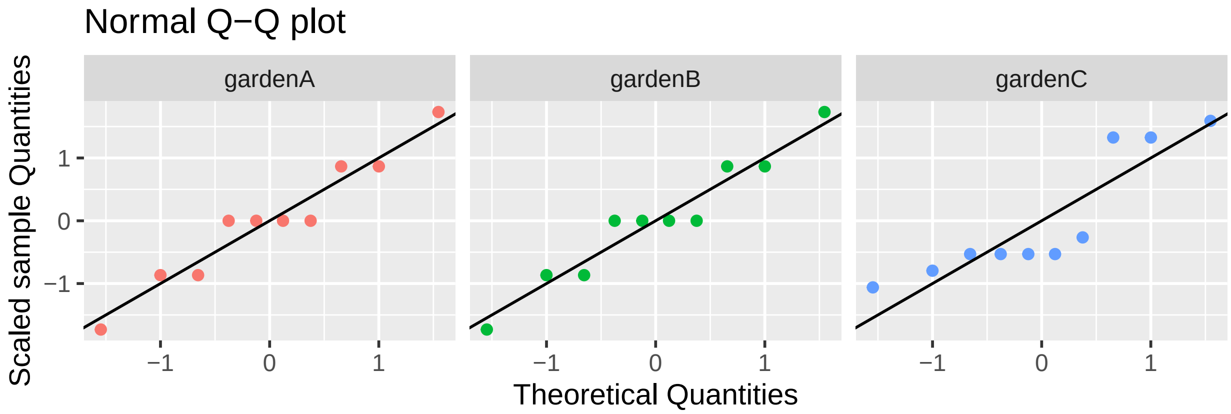
```
garden_2 %>%
  group_by(Garden) %>%
  mutate(scaled_ozone = scale(ozone)) %>%
  ggplot(aes(sample = scaled_ozone, colour = Garden)) +
  stat_qq(distribution = stats::qnorm) +
  geom_abline(slope = 1, intercept = 0) +
  facet_wrap(Garden~.) +
  labs(title = "Normal Q-Q plot",
       x = "Theoretical Quantities",
       y = "Scaled sample Quantities")
```



I now plot everything together

```
library(cowplot)

garden_2 %>%
  ggplot(aes(x = ozone, color = Garden, fill = Garden)) +
  geom_density(alpha = 0.5) +
  scale_x_continuous(limits = c(min(limits_x$min), max(limits_x$max))) +
  theme_classic() +
  labs(title = "Garden Densities",
       x = "[ozone]",
       y = "density") -> p1

garden_2 %>%
  group_by(Garden) %>%
  mutate(scaled_ozone = scale(ozone)) %>%
  ggplot(aes(sample = scaled_ozone, colour = Garden)) +
  stat_qq(distribution = stats::qnorm) +
  geom_abline(slope = 1, intercept = 0) +
  facet_wrap(Garden~.) +
  theme(legend.position = "none") +
  labs(title = "Normal Q-Q plot",
       x = "Theoretical Quantities",
       y = "Scaled sample Quantities") -> p2

plot_grid(p1, p2, nrow = 2)
```





FINAL TEST

I use the Shapiro test for the three datasets, using $\alpha = 0.05$.

```
tidy(shapiro.test(garden$gardenA))
```

```
## # A tibble: 1 x 3
##   statistic p.value method
##       <dbl>   <dbl> <chr>
## 1     0.953   0.703 Shapiro-Wilk normality test
```

```
tidy(shapiro.test(garden$gardenB))
```

```
## # A tibble: 1 x 3
##   statistic p.value method
##       <dbl>   <dbl> <chr>
## 1     0.953   0.703 Shapiro-Wilk normality test
```

```
tidy(shapiro.test(garden$gardenC))
```

```
## # A tibble: 1 x 3
##   statistic p.value method
##       <dbl>   <dbl> <chr>
## 1     0.780 0.00828 Shapiro-Wilk normality test
```

```
# Conclusion
# Since the null hypothesis of this test is that our distribution is normal
# For gardenA and gardenB, the p-values, 0.703 are greater than 0.05. This means we can accept the hypo
# For gardenC, the p-value, 0.0082 is < 0.05. We reject the null hypothesis. gardenC's distribution is
```