Illuminating the Magical Patterns of the Wizarding World

Casey Hutchinson

2023-10-03

Project Patronus

Questions:

What is the average level of Order activity across all locations? What is the median level of of Dark activity across all locations?

Construct a histogram and describe the shape of the Spell Mastery measurements. Use a horizontal boxplot to identify whether there are any outliers present.

Which is more spread out, Order activity or Dark activity?

Construct horizontal boxplots for Order activity and Dark Arts. Which variable appears more symmetric?

Locations where Dark Arts concentrations are above the third quartile may warrant additional attention by the Order. Which locations have Dark Arts concentrations above the third quartile?

Using the filter command from the dplyr library, create a new data set that is comprised of only the locations where a horcrux is present. Call this data set horcrux. Also create a new data set that is comprised of only the locations where a Horcrux is not present. Call this dataset no.horcrux.

Compare the average Order activity in locations with a Horcrux to those without a Horcrux. Do the same for Dark activity. Which difference is greater: the difference in Order activity or the difference in Dark activity between locations with and without a Horcrux.

Create a scatterplot of the Order activity vs Dark activity in locations without a Horcrux. Comment on the type of association.

On average, which locations demonstrate higher Spell Mastery - locations where a Horcrux is present, or locations where a Horcrux is not present?

Create a scatterplot of the Dark Arts vs Dark activity for locations where a Horcrux is not present. There appears to be a data point that is very different from the others (this is called an influential point). Which location does this point correspond to?

Data

The data in this study can be found here: Harry Potter Data (http://datasets.barrymonk.com/MATH3440/HP_Data.csv)

Data Set Names

```
library(readr)
Patronus <- read.csv("http://datasets.barrymonk.com/MATH3440/HP_Data.csv")
```

I will also add data set names *horcrux* and *no.horcrux* when we dive deeper into the data.

Question 1) What is the *average* level of Order activity across all locations?

```
mean(Patronus$order_activity)
```

```
## [1] 60.56424
```

What is the *median* level of of Dark activity across all locations?

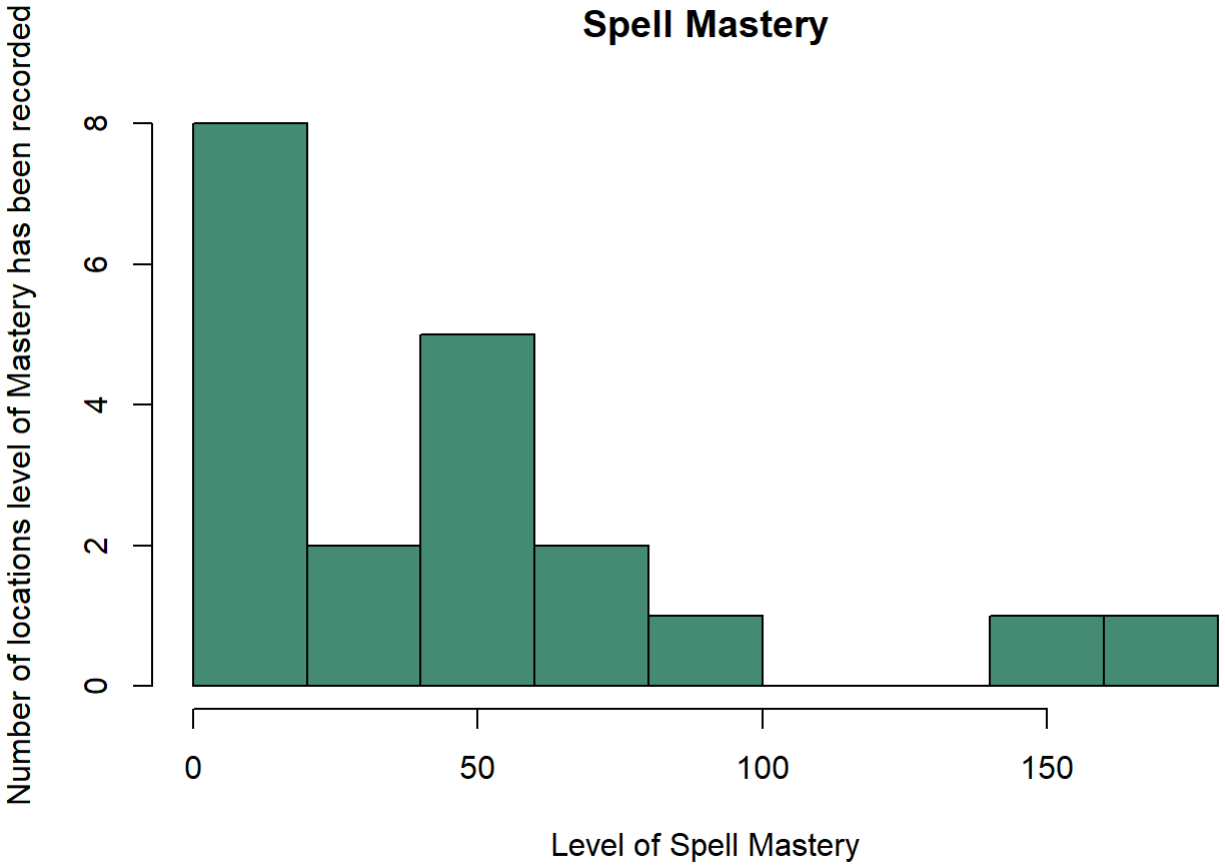
```
median(Patronus$dark_activity)
```

```
## [1] 53.45536
```

The average level of Order activity per year across the locations provided is currently a shade under 61 incidents. The median level of Dark activity per year in those same locations is a shade under 54 incidents.

Question 2) Construct a historgram and describe the

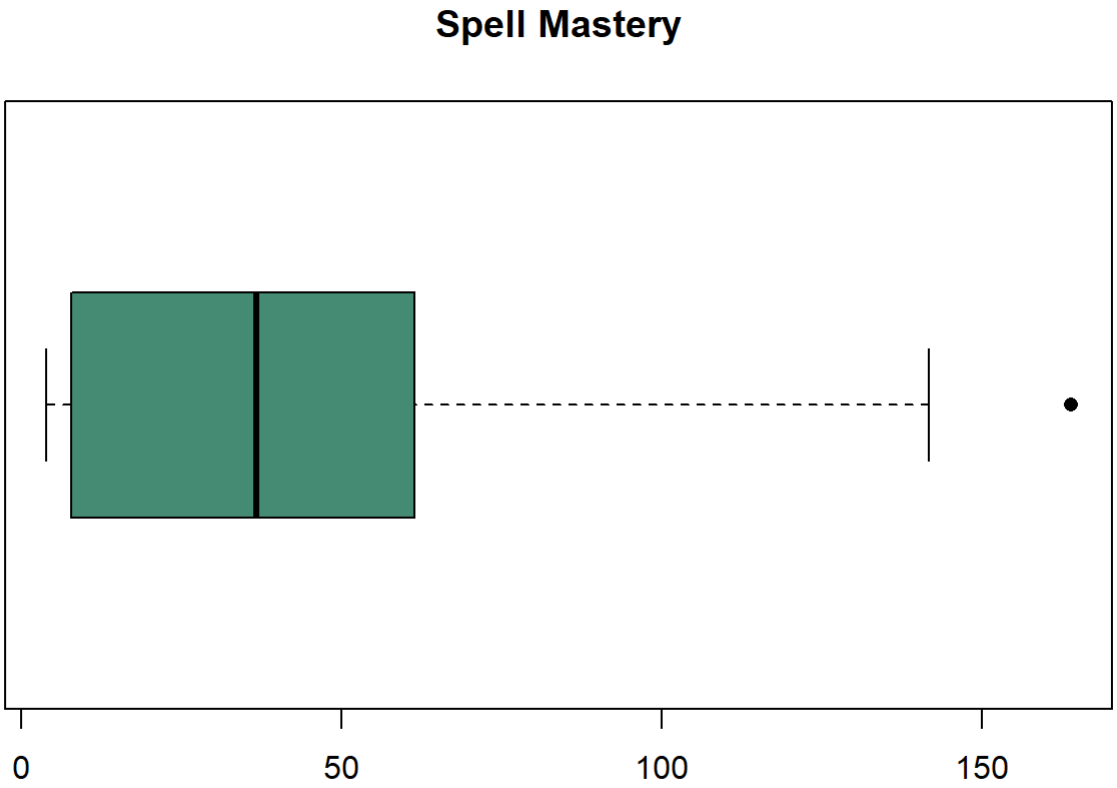
shape of the Spell Mastery measurements.



Visualizing the data highlights how it's skewed to the right, and the numbers back that up. The average level of Spell Mastery is *45.7*, however the median level is only *36.8.* There is also a significant gap between 2 data points and the rest, suggesting there may be multiple outliers.

Use a horizontal boxplot to identify whether there are

any outliers present.



As you can see from the boxplot there is an outlier; this mark corresponds to Hogsmeade. Some might suggest this high level is worth additional attention by the Order, and point to local spell limits to say this level is too high (spell limits have never been officially enforced by the Ministry). However the overall data suggests that there has been little concentration of dark magic recorded in Hogsmeade and combined with special assurances from Professor Dumbledore, we have little reason to believe this should be of concern to the Ministry or the Order.

Question 3) Which is more spread out, Order activity or Dark activity?

We'll start by looking at a few key data points for each: the mean, median, standard deviation, and variance.

Order activity:

```
mean(Patronus$order_activity)
```

```
## [1] 60.56424
```

```
median(Patronus$order_activity)
```

```
## [1] 58.52147
```

```
sd(Patronus$order_activity)
```

```
## [1] 21.93714
```

```
var(Patronus$order_activity)
```

```
## [1] 481.238
```

Dark activity:

```
mean(Patronus$dark_activity)
```

```
## [1] 53.9192
```

```
median(Patronus$dark_activity)
```

```
## [1] 53.45536
```

```
sd(Patronus$dark_activity)
```

```
## [1] 18.35667
```

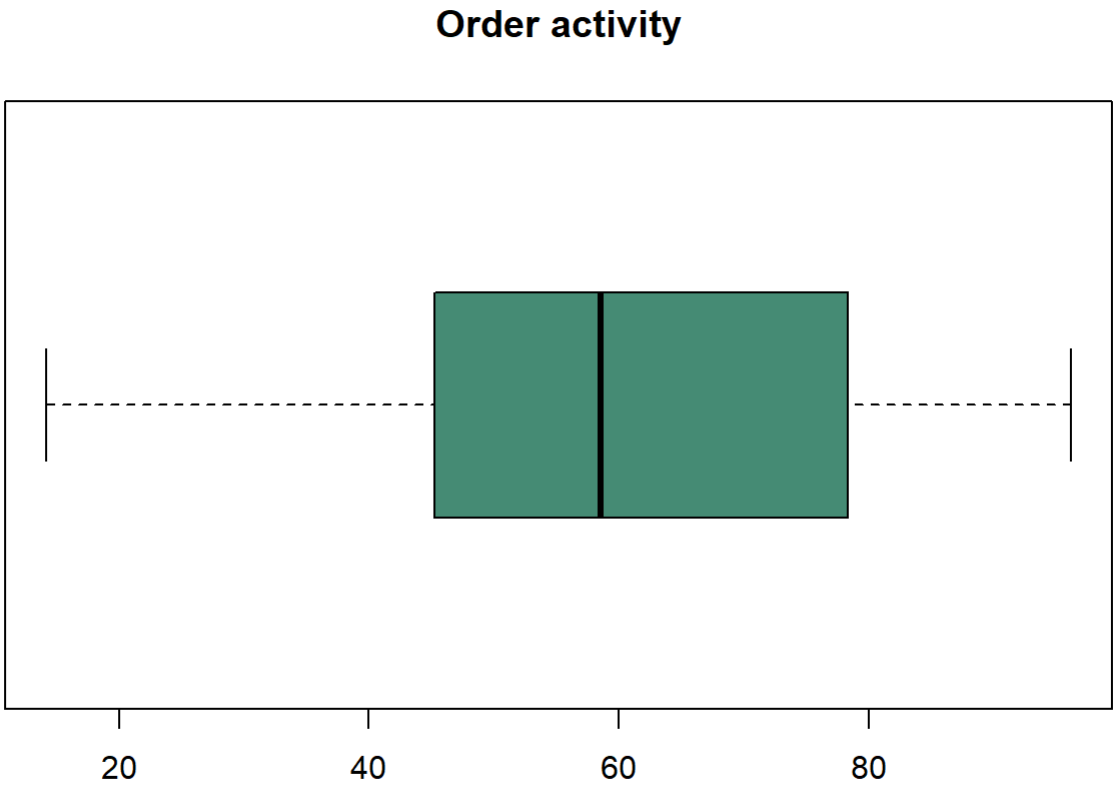
```
var(Patronus$dark_activity)
```

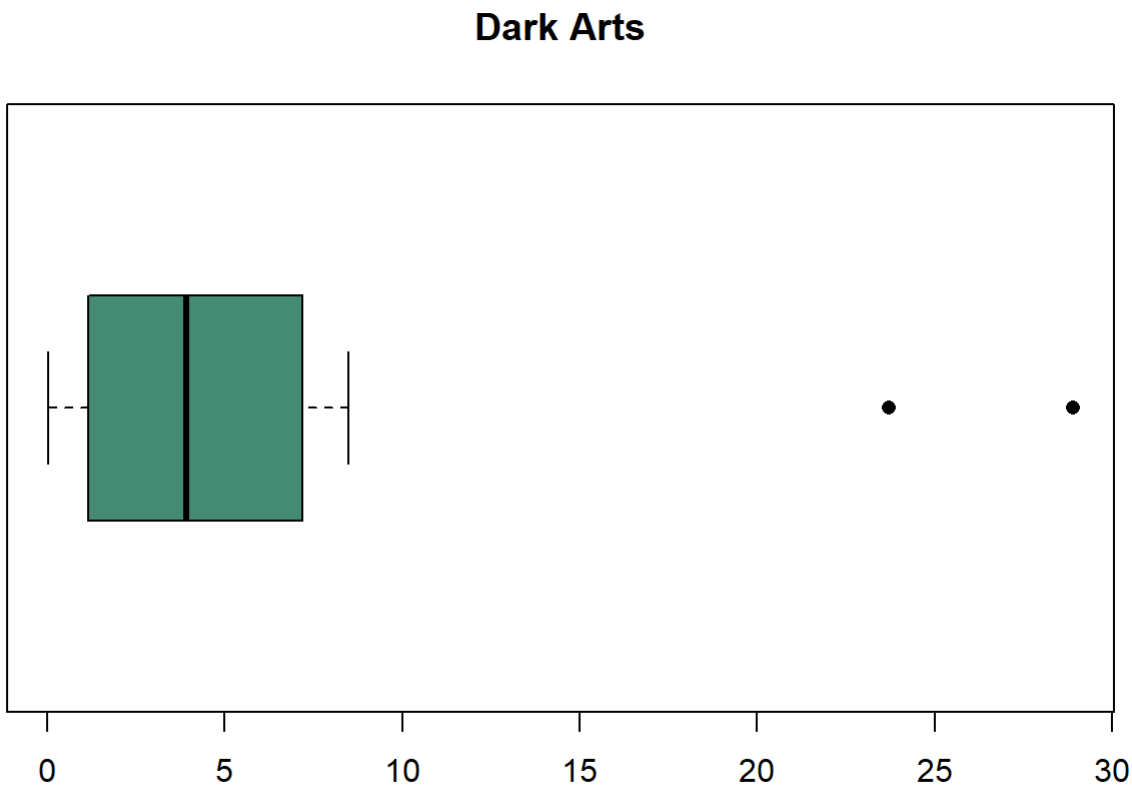
```
## [1] 336.9674
```

We can see the average number of instances of Order activity is higher than the average of Dark activity, but comparing variance tells us that Order activity is much more spread out.

Question 4) Construct horizontal boxplots for Order activity and Dark Arts. Which variable appears more

symmetric?





Comparing the two boxplots we can see that Order activity is much more symmetric. There are two noticeable outliers in Dark Arts: Malfoy Manor and Knockturn Alley. It's always been my understanding that the Malfoy's are well-respected in our world but it's impossible to ignore the concentration of dark spells recorded near their home. The less said about Knockturn Alley the better, but we will comment on this location more later on.

Question 5) Locations where Dark Arts concentrations that are above the third quartile may warrant additional attention by the Order. Which locations have Dark Arts concentrations above the third quartile?

The third quartile for Dark Arts concentration is given below:

```
quantile(Patronus$dark_arts, 0.75)
```

```
##      75%
## 7.164685
```

With the third quartile established as 7.16 we can see from the data that we have 5 locations with concentrations above that level: The previously mentioned Malfoy Manor and Knockturn Alley, as well as The Burrow, Godric's Hollow, and the Chamber of Secrets. It's worth noting that at 7.13 Gringotts just missed the cut.

Hogsmeade, the outlier mentioned in Question 2, has a dark arts reading of 0.91.

Question 6) Using the filter command from the dplyr library, create a new data set that is comprised of only the locations where a horcrux is present. Call this data set horcrux. Also create a new data set that is comprised of only the locations where a Horcrus is not present. Call this dataset no.horcrux.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
horcrux <- filter(Patronus, horcrux == 1)
no.horcrux <- filter(Patronus, horcrux == 0)
```

Question 7) Compare the average Order activity in locations with a Horcrux to those without a Horcrux. Do the same for Dark activity. Which difference is greater: the difference in Order activity or the difference in Dark activity between locations with and without a Horcrux.

Let's start by looking at Order activity:

```
mean(horcrux$order_activity)
```

```
## [1] 84.85709
```

```
mean(no.horcrux$order_activity)
```

```
## [1] 50.15302
```

```
mean(horcrux$order_activity) - mean(no.horcrux$order_activity)
```

```
## [1] 34.70406
```

Dark activity:

```
mean(horcrux$dark_activity)
```

```
## [1] 74.43731
```

```
mean(no.horcrux$dark_activity)
```

```
## [1] 45.12572
```

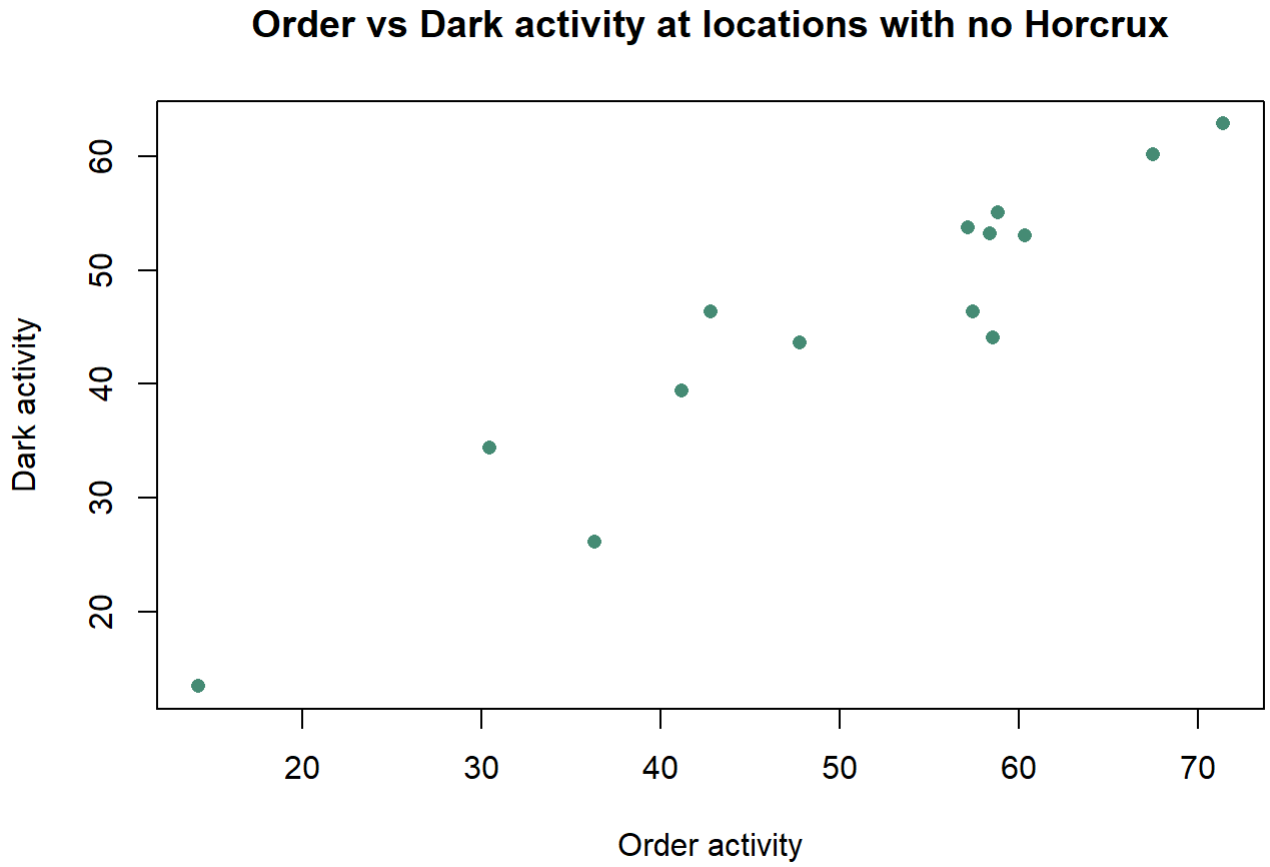
```
mean(horcrux$dark_activity) - mean(no.horcrux$dark_activity)
```

```
## [1] 29.31159
```

As we can see, the difference between Order activity levels at locations with a Horcrux was 34.7 instances higher than at locations without a Horcrux. This is a few points higher than the difference of Dark levels at the same locations which was 29.3 instances

Question 8) Create a scatterplot of the Order activity vs Dark activity in locations without a Horcrux.

Comment on the type of association.




```
## [1] 0.9450731
```

There is a clear positive assocation between Order and Dark activity seen in the Scatterplot. Below the scatterplot is the correlation data between the two, and it's calculated to be 0.95 (the closer you get to 1 when calculating correlation the stronger the linear relationship between the variables).

Question 9) On average, which locations demonstrate higher Spell Mastery - locations where a Horcrux is present, or locations where a Horcrux is not present?

Average at locations where a Horcrux is present:

```
mean(horcrux$spell_mastery)
```

```
## [1] 53.28065
```

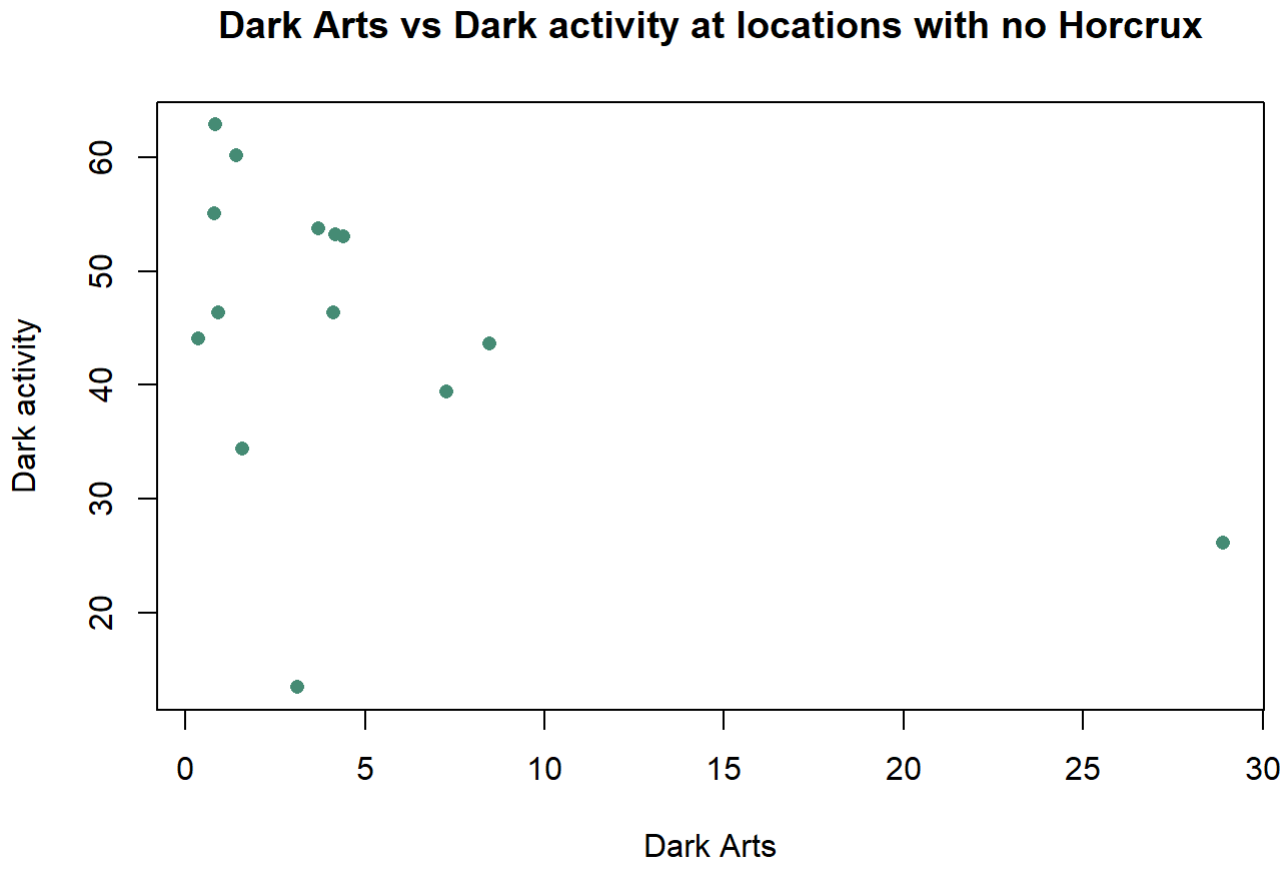
Average at locations where a Horcrux is not present:

```
mean(no.horcrux$spell_mastery)
```

```
## [1] 42.45215
```

Here we can see that Spell Mastery is nearly 10 levels higher at locations with a Horcrux present vs those where a Horcrux is not present.

Question 10) Create a scatterplot of the Dark Arts vs Dark activity for locations where a Horcrux is not present. There appears to be a data point that is very different from the others (this is called an influential point). Which location does this point correspond to?



The influential point corresponds to the location Knockturn Alley (previously mentioned in *Question 4*). Knockturn Alley is noteworthy due to it being the home of Borgin and Burkes and it's proximity to Diagon Alley, but we don't believe it requires the Order's additional attention - for now. While it's true that it registered as an outlier along with Malfoy Manor (a location I do recommend additional attention), this location does not have a high level of annual Dark activity. In fact, Knockturn Alley records significantly less annual Dark activity than Hogwarts, Hogsmeade, or the Ministry of Magic.