KA Contact Tracing

```
dat <- read_csv("KAtrace.csv",  skip = 19)
dat$Date <- as.Date(dat$Date, "%d-%b")
```

Cleaning & summarizing Clusters variable

```
dat <- dat %>% group_by(Cluster) %>% mutate(n_cluster = n()) #Variable for number of
 cases in each cluster
cases_per_cluster <- dat %>% group_by(Cluster) %>% summarize(n_cluster = n()) %>% arr
ange(-n_cluster) #Table for number of cases in each cluster

dat$origin <- ifelse(dat$n_cluster < 20, "Others", ifelse(dat$Cluster == "Influenza l
ike illness", "ILI", ifelse(dat$Cluster == "Severe Acute Respiratory Infection", "SAR
I", as.character(dat$Cluster)))) #Creating a cleaned up variable with information fro
m `Clusters' for diagramming
```

Superspreading behaviour

How many cases confirmed and contact traced till July 7 caused secondary infections within the next two weeks (till July 21)

```
july7_parent_c <- dat %>% filter(Case <= 26815 & C == 1) # Collapsing C variable
nrow(july7_parent_c)
```

```
## [1] 1677
```

```
july7_parent_p <- dat %>% group_by(P) %>% summarize(secondary = n()) %>% filter(P <=
26815 & P != 0) # Collapsing P variable
nrow(july7_parent_p)
```

```
## [1] 1684
```

- Looking at cases that do not overlap in the two approaches

```
setdiff(july7_parent_c$Case, july7_parent_p$P)
```

```
## [1]   133   300   502   503  2091  2092 18248
```

```
setdiff(july7_parent_p$P, july7_parent_c$Case)
```

```
##  [1]   124   421   423   424   426   427   536   848  1724  1852  3857  5823
## [13] 14329 25338
```

- Collapsing on P seems to work better

```
## Joining, by = "P"
```

```
## [1] "1684 cases diagnosed and contact traced till July 7 caused secondary infectio
ns by July 21, 2020"
```

How many secondary infections did these 1684 cases cause

```
sum(july7_parent_p$secondary)
```

```
## [1] 5031
```

How many cases confirmed and contact traced till July 7 did NOT cause secondary infections within the next two weeks (till July 21)

Subset the data to approximate proportion who were contact traced

I assume that cases fulfilling all of the following criteria were *NOT* contact traced at all: * Cluster is Unknown * Reason is NA * C = 0 * P = 0

```
july21_traced <- dat %>% filter(Cluster != "Unknown" | !is.na(Reason) | C != 0 | P !=
0)
```

```
## [1] "38077 out of 71068 cases were contact traced till July 21"
```

- Checking if this makes sense by recalculating number of parents that caused infections by July 5, should be same as above (a lower figure would indicate that I oversubsetted)

```
temp <- july21_traced %>% group_by(P) %>% summarize(secondary = n()) %>% filter(P <=
26815 & P != 0) # Collapsing P variable
nrow(temp) #seems right
```

```
## [1] 1684
```

Number of cases confirmed and traced by July 7 that did not cause other infections

```
july7_traced <- july21_traced %>% filter(Case <= 26815) # Number of cases confirmed a
nd traced by July 7
nrow(july7_traced)-nrow(july7_parent_p) # Number of cases confirmed and traced by Jul
y 7 that did not cause other infections
```

```
## [1] 16211
```

```
## [1] "Of the 17895 cases that were confirmed by July 7 and contact traced 1684 caus
ed secondary infections, while 16211 did not cause any secondary infections at all"
```

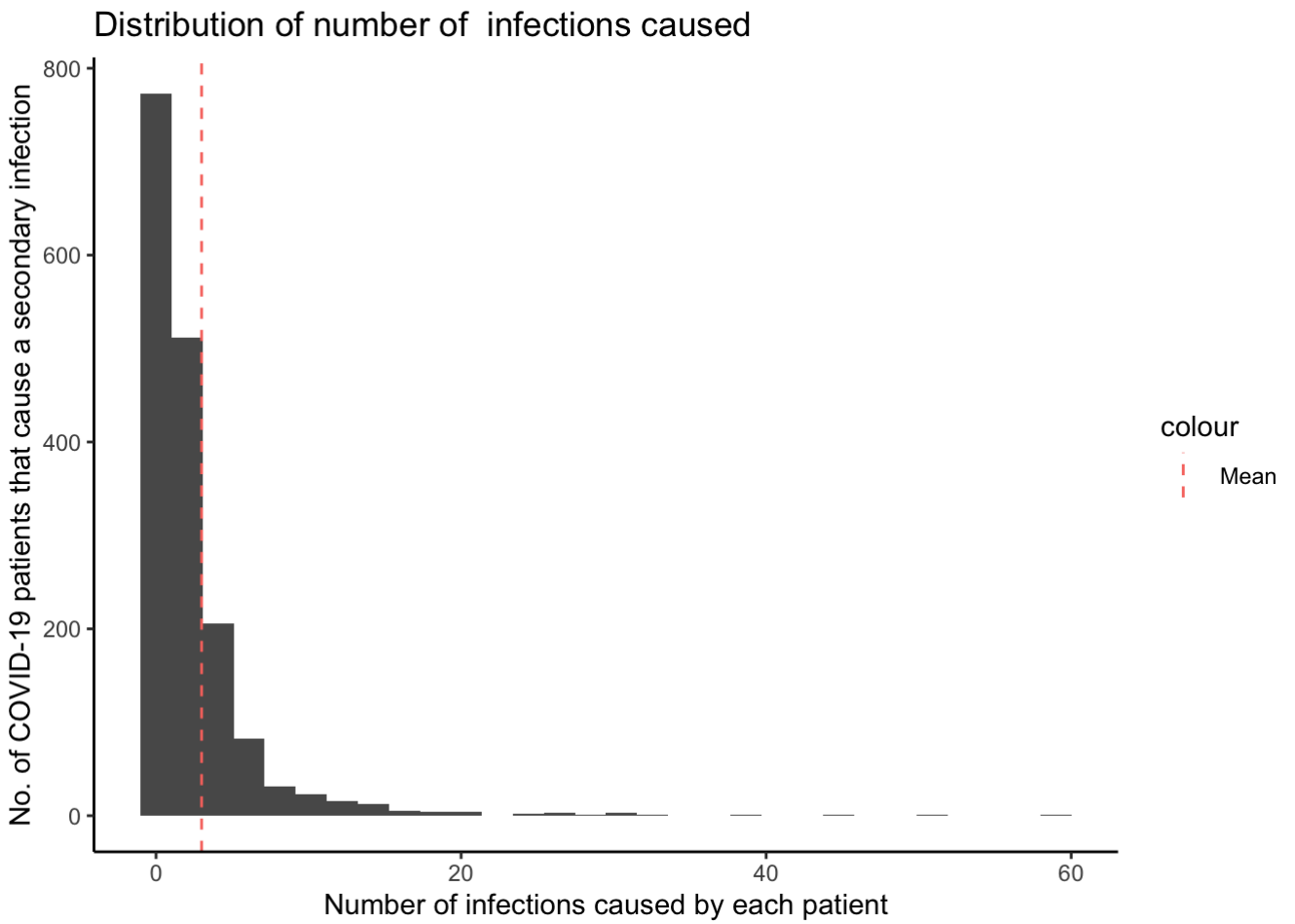
Average number of secondary infections caused by cases that do cause infections

```
mean_infections <- as.numeric(july7_parent_p %>% summarise(mean = mean(secondary)))
july7_parent_p %>% summarize(avg = mean(secondary), med = median(secondary))
```

```
## # A tibble: 1 x 2
##     avg   med
##   <dbl> <dbl>
## 1  2.99     2
```

```
july7_parent_p %>% ggplot(aes(secondary)) + geom_histogram() + geom_vline(aes(xinterc
ept = mean_infections, color = "Mean"), linetype="dashed") + theme_classic() + xlab(
"Number of infections caused by each patient") + ylab("No. of COVID-19 patients that
 cause a secondary infection") + ggtitle("Distribution of number of  infections cause
d")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



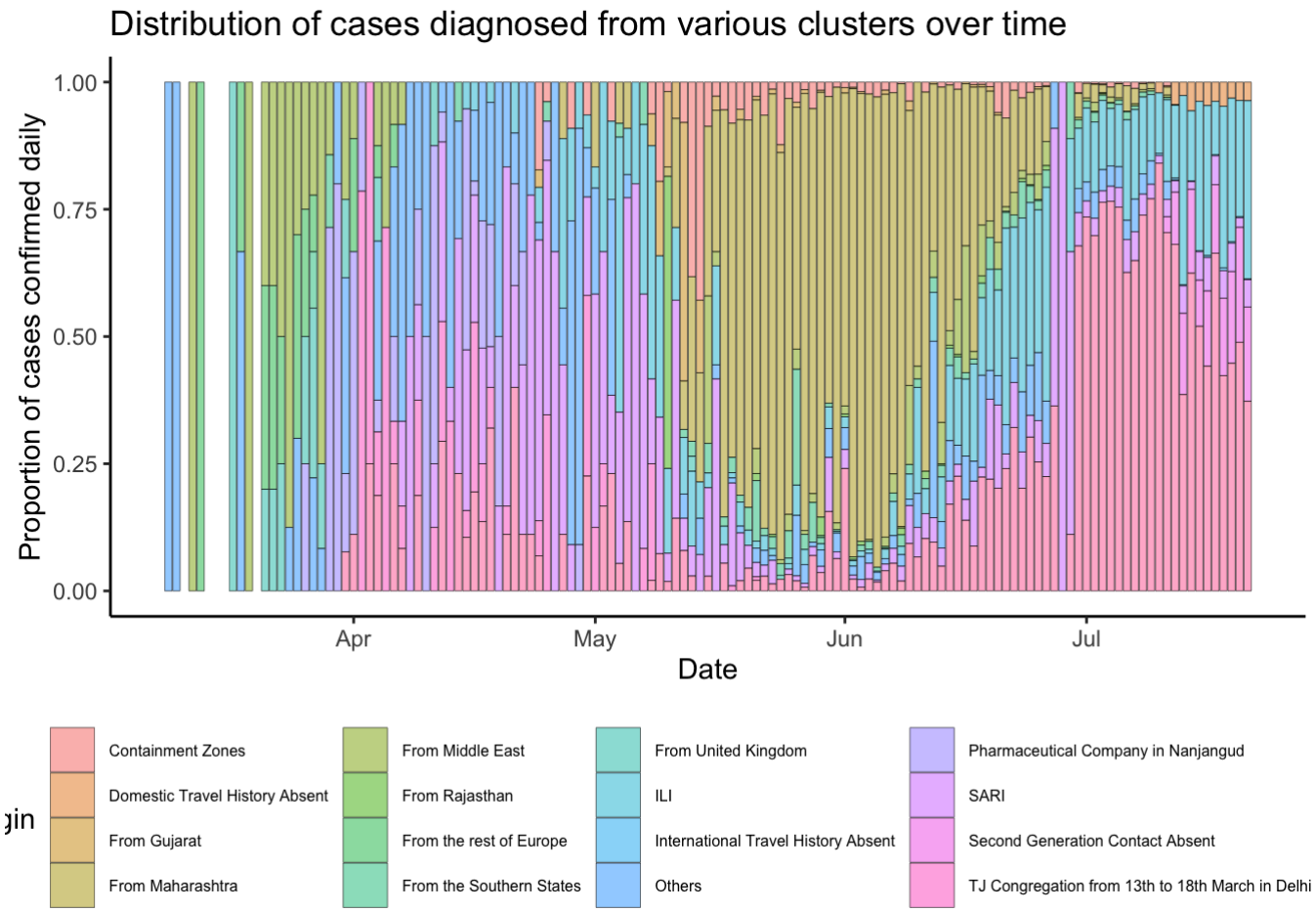

```
ggsave("secondary_hist.png")
```

```
## Saving 7 x 5 in image
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Visualization of change in cluster size over time

```
dat %>% filter(Cluster != "29-June Trace History Absent" & Cluster != "28-June Trace
 History Absent" & Cluster != "27-June Trace History Absent") %>% ggplot(aes(Date)) +
geom_bar(position = "fill", aes(fill=origin), alpha = 0.5, color = "black", size = 0.
1) + theme_classic() + theme(legend.position="bottom", legend.spacing = unit(0.4, "po
ints"), legend.text = element_text(size = 6))  + ggtitle("Distribution of cases diagn
osed from various clusters over time") + ylab("Proportion of cases confirmed daily")
```

```
## Warning: Removed 2 rows containing non-finite values (stat_count).
```




```
ggsave("clusters_stacked.png")
```

```
## Saving 7 x 5 in image
```

```
## Warning: Removed 2 rows containing non-finite values (stat_count).
```