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, "Under 20 cases", ifelse(dat$Cluster == "Inf
luenza like illness", "ILI", ifelse(dat$Cluster == "Severe Acute Respiratory Infectio
n", "SARI", ifelse(dat$Cluster == "TJ Congregation from 13th to 18th March in Delhi",
"TJ Congregation", ifelse(dat$Cluster == "Pharmaceutical Company in Nanjangud", "Phar
ma Co, Nanjangud", as.character(dat$Cluster))))))

dat$origin <- ifelse(dat$Cluster == "From Gujarat", "Domestic Travel", ifelse(dat$Clu
ster == "From Rajasthan", "Domestic Travel", ifelse(dat$Cluster == "From the Southern
States", "Domestic Travel", ifelse(dat$Cluster == "From Middle East", "International
 Travel",ifelse(dat$Cluster == "From United Kingdom", "International Travel",ifelse(d
at$Cluster == "From the rest of Europe", "International Travel", ifelse(dat$Cluster =
= "From USA", "International Travel", ifelse(dat$Cluster == "From South America", "In
ternational Travel",ifelse(dat$Cluster == "Second Generation Contact", "Others", as.c
haracter(dat$origin))))))))))


#Creating a cleaned up variable with information from `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`.
```

Distribution of number of infections caused
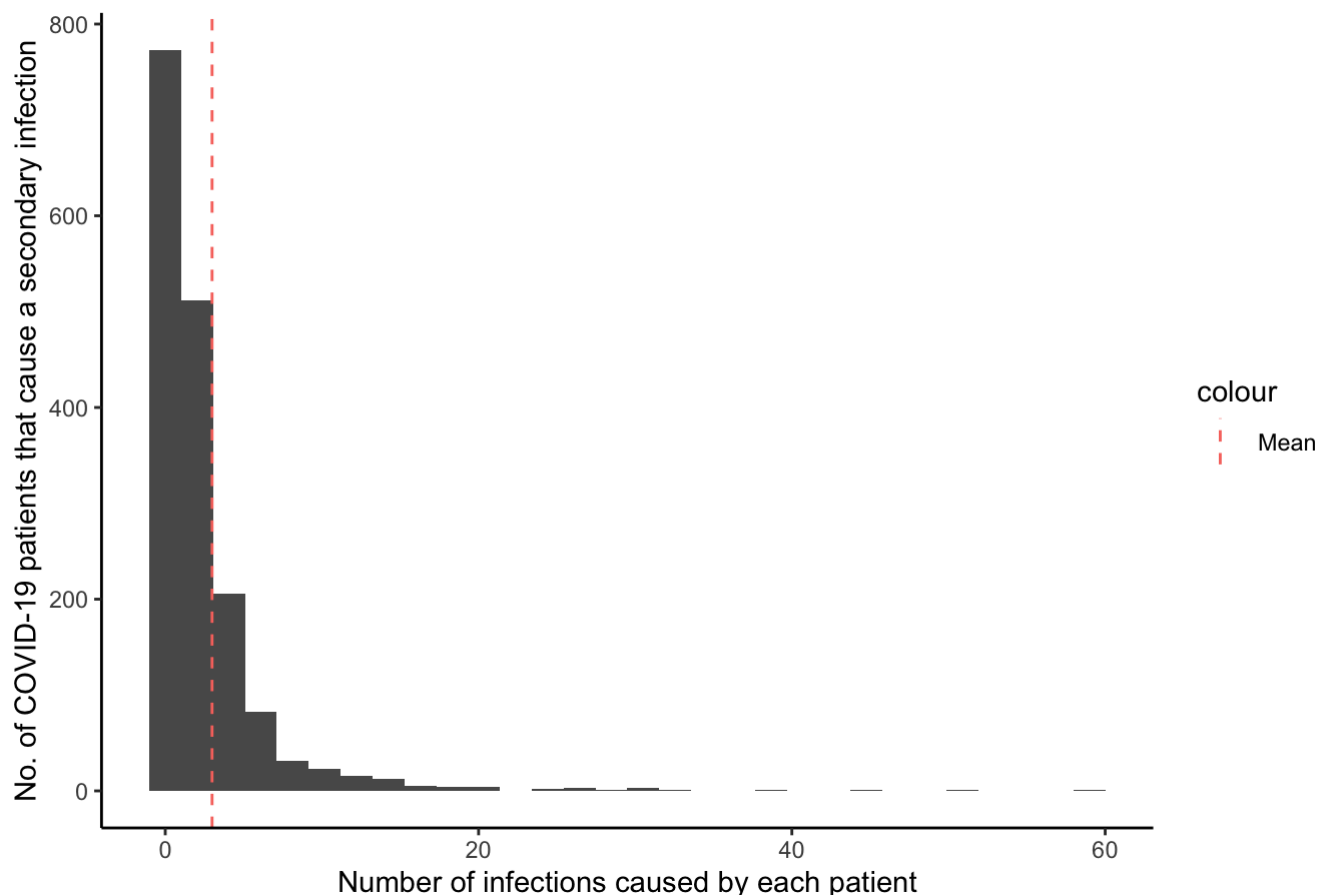

```
ggsave("secondary_hist.png")
```

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

Number of cases that caused more than 10 secondary infections

```
x <- july7_parent_p %>% filter(secondary > 10)
paste(nrow(x), "cases caused 10 or more secondary infections")
```

```
## [1] "69 cases caused 10 or more secondary infections"
```

% of cases with unknown P

March 9 - June 1

```
p <- dat %>% filter(Date < "2020-05-01") %>% group_by(P) %>% summarize(secondary = n
()) # number of secondary infections
x <- dat %>% filter(Date < "2020-05-01") %>% ungroup() %>% summarize(n = n()) # total
cases
```

```
## [1] "Out of 563 cases confirmed by June 1, 219 ( 38.898756660746 ) had no contact
history"
```

July 1 - July 21

```
p <- dat %>% group_by(P) %>% summarize(secondary = n()) # number of secondary infecti
ons
x <- dat %>% ungroup() %>% summarize(n = n()) # total cases
```

```
## [1] "Out of 71068 cases confirmed by June 1, 65977 ( 92.8364383407441 ) had no con
tact history"
```

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" & Cluster != "Second Gen
eration Contact Absent" & Cluster != "Domestic Travel History Absent" & Cluster != "I
nternational Travel History Absent") %>% ggplot(aes(Date)) + geom_bar(position = "fil
l", aes(fill=origin), alpha = 0.8, color = "black", size = 0.1)  + scale_fill_brewer
(palette = "Spectral") + theme_classic() + theme(legend.position="bottom", legend.spa
cing = unit(0.4, "points"), legend.text = element_text(size = 6))  + ggtitle("Distrib
ution of case origins in Karnataka, March 9 - July 21, 2020") + ylab("Proportion of c
ases confirmed daily")
```

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

Distribution of case origins in Karnataka, March 9 - July 21, 2020
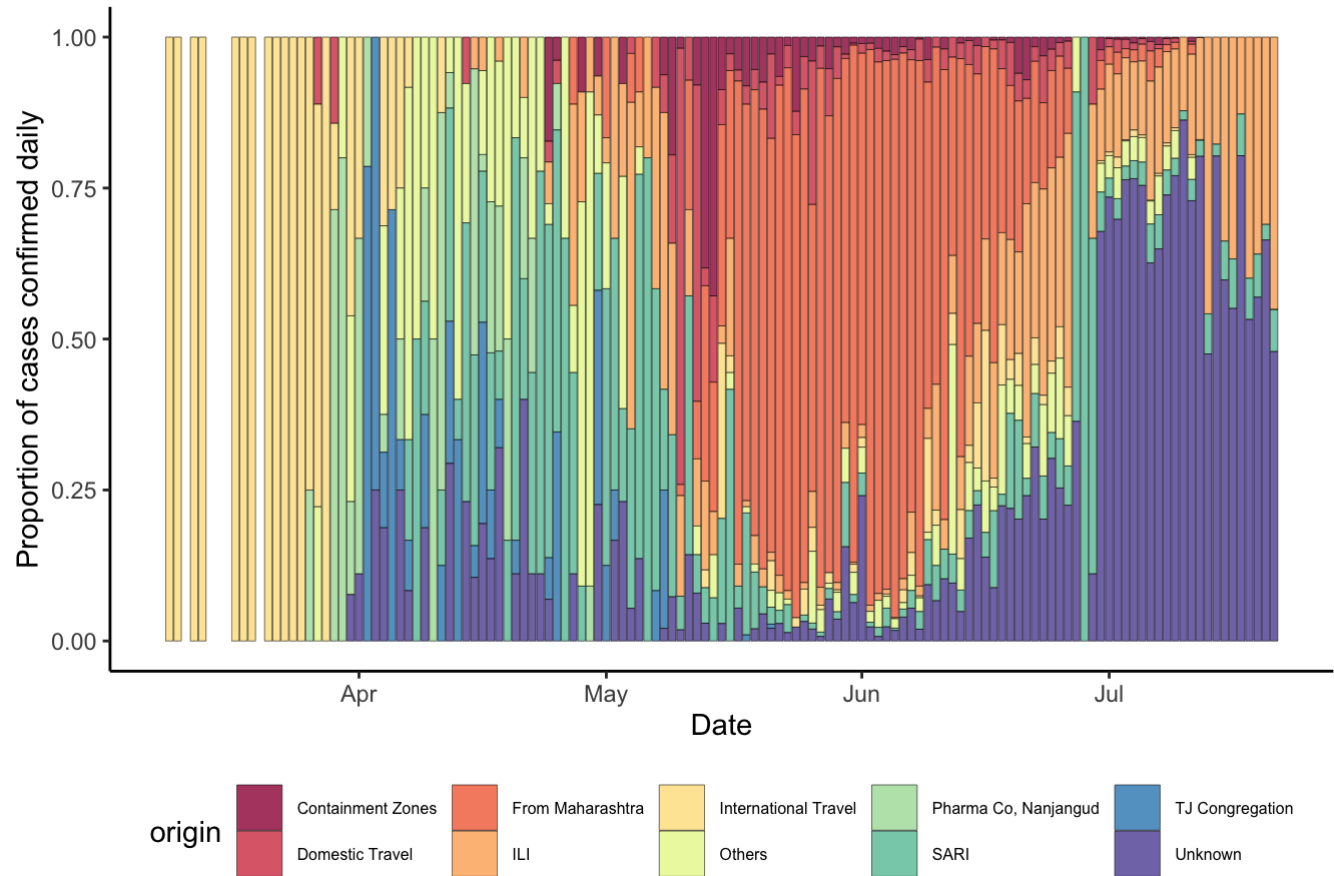

```
ggsave("clusters_stacked.png")
```

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

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