SVAP Assignment

Shashank Shekhar

10/6/2017

Background

I have picked Agriculture Domain. The dataset has been acuired from [here] [https://data.gov.in/catalog/district-wise-and-month-wise-queries-farmers-kisan-call-centre-kcc-during-2017] in **csv** format. The dataset is about **KCC (Kishan Call Centre)**, where Kishan (Farmer) calls to enquire/resolve issue/get information about different queries. I have picked the data fro DELHI KCC and for the month of September 2017

> Sector : Agriculture
>
> Group : Kisan Call Centre (KCC)
>
> Catalogs: District wise and month wise queries of farmers in Kisan Call Centre (KCC) during **September 2017** for **DELHI, INDIA**

Frame

- What are the top 5 **QueryType** asked in each **Category**.
- Based on **Sept** month Data, can we predit the top KCC's queryType for October.

```
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
```

```
library(ggplot2)
library(tidyr)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

Acquire

you can get the dataset from [here] (https://data.gov.in/catalog/district-wise-and-month-wise-queries-farmers-kisan-call-centre-
I have taken only for **Delhi** region and for the month of **Sept'17**.

```
getwd()
```

```
## [1] "/Users/shashankshekhar/PGP-BDA/R-Studio"
```

```r
data <- read.csv("KCC.csv", header = TRUE)
str(data)
```

```
## 'data.frame':    1132 obs. of  11 variables:
##  $ Season      : Factor w/ 3 levels "JAYAD","KHARIF",..: 2 2 2 2 3 2 2 2 2 2 ...
##  $ Sector      : Factor w/ 4 levels "AGRICULTURE",..: 1 1 1 1 4 4 1 1 1 1 ...
##  $ Category    : Factor w/ 16 levels "Animal","Cereals",..: 12 12 12 12 16 16 12 12 12 2 ...
##  $ Crop        : Factor w/ 66 levels "Aloe Vera","Aonla",..: 44 44 44 44 65 65 44 44 44 45 ...
##  $ QueryType   : Factor w/ 40 levels "\tField Preparation\t",..: 39 39 39 39 29 29 35 35 36 31 ...
##  $ QueryText   : Factor w/ 415 levels "APHIDS IN TOMATO",..: 332 142 142 142 370 369 7 7 380 38 ...
##  $ KCCAns      : Factor w/ 375 levels "??? ?? ????? ?????? - ???? ?? ?????-94",..: 226 223 223 223 3
##  $ StateName   : Factor w/ 1 level "DELHI": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DistrictName: Factor w/ 3 levels "CENTRAL","EAST",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ BlockName   : Factor w/ 1 level "DELHI": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CreatedOn   : Factor w/ 1114 levels "00:05.4","00:05.5",..: 354 112 4 273 512 1049 689 80 305 742
```

```r
dim(data)
```

```
## [1] 1132   11
```

Refine/Transform

```r
colnames(data) <- c('season', 'sector', 'category', 'crop', 'queryType', 'queryText', 'kccAnswer', 'sta
data$season <- as.character(data$season)
data$sector <- as.character(data$sector)
data$category <- as.character(data$category)
data$crop <- as.character(data$crop)
data$queryType <- as.character(data$queryType)
data$queryText <- as.character(data$queryText)
data$kccAnswer <- as.character(data$kccAnswer)
data$state <- as.character(data$state)
data$district <- as.character(data$district)
data$block <- as.character(data$block)
data$queryType <- gsub('\t','',data$queryType)
data <- data[, 1:10] # removed createdOn column as the value in the column doesn't make sense. Also, my
str(data)
```

```
## 'data.frame':    1132 obs. of  10 variables:
##  $ season   : chr  "KHARIF" "KHARIF" "KHARIF" "KHARIF" ...
##  $ sector   : chr  "AGRICULTURE" "AGRICULTURE" "AGRICULTURE" "AGRICULTURE" ...
##  $ category : chr  "Others" "Others" "Others" "Others" ...
##  $ crop     : chr  "Others" "Others" "Others" "Others" ...
##  $ queryType: chr  "Weather" "Weather" "Weather" "Weather" ...
##  $ queryText: chr  "TELL ME RAIN FALL INFORMATION." "TELL ME ABOUT WEATHER INFORMATION ?" "TELL ME A
##  $ kccAnswer: chr  "RAIN POSSIBILITY IN NEXT 3-4 DAYS IN YOUR DISTRICT" "RAIN POSSIBILITY IN NEXT 3-
##  $ state    : chr  "DELHI" "DELHI" "DELHI" "DELHI" ...
##  $ district : chr  "EAST" "EAST" "EAST" "EAST" ...
##  $ block    : chr  "DELHI" "DELHI" "DELHI" "DELHI" ...
```

```r
dim(data)
```

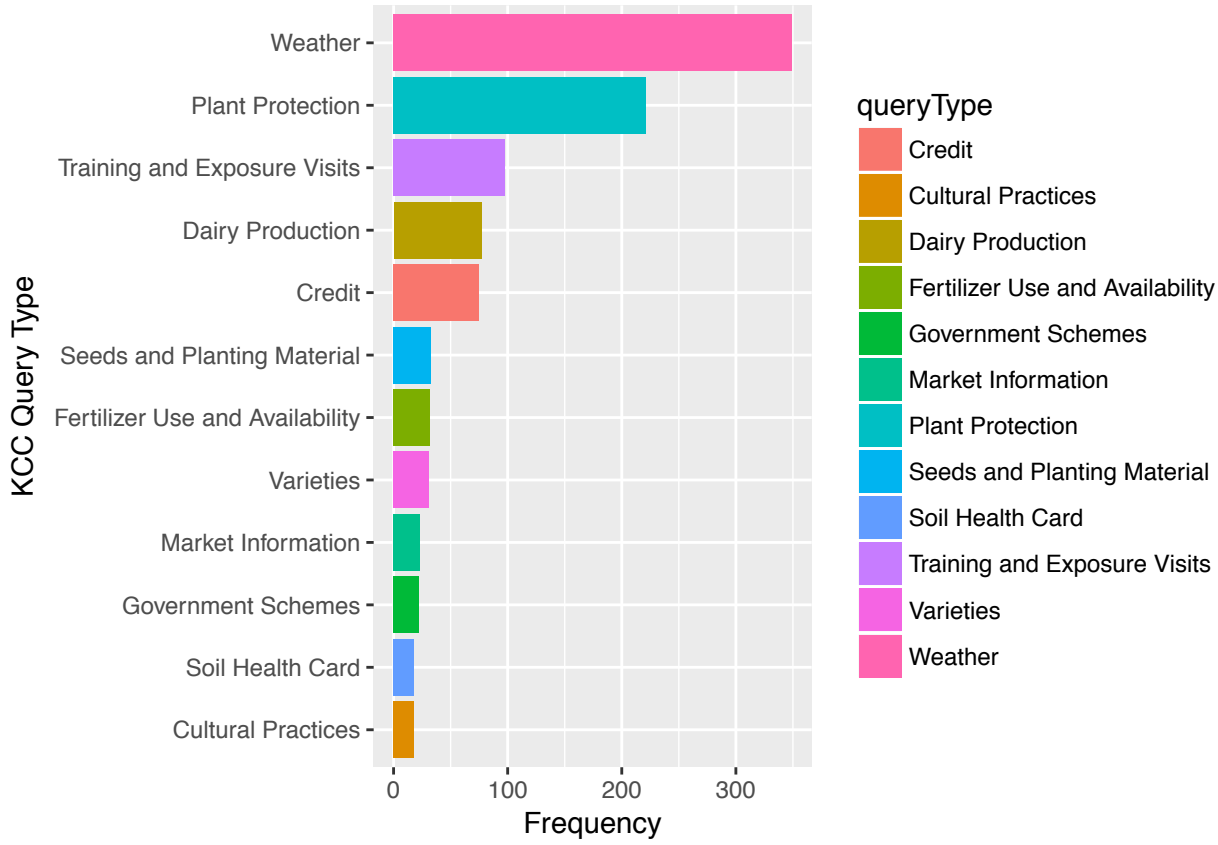
```
## [1] 1132   10
```

Explore

I want to know which are top most query types which KCC gets usually.

- This Graph shows that the **Kisan Call Centre** gets most of the calls enquiring about the `Weather`

```
queryType2 <- as.data.frame.table(cbind(table(data$queryType)))

colnames(queryType2) <- c("queryType", "type", "freq")

plot1 <- queryType2 %>%
  arrange(desc(freq)) %>%
  head(12) %>%
  ggplot(aes(reorder(queryType,freq), freq, fill=queryType)) + geom_col() +
  coord_flip() +
  xlab("KCC Query Type") +
  ylab("Frequency")

plot1
```

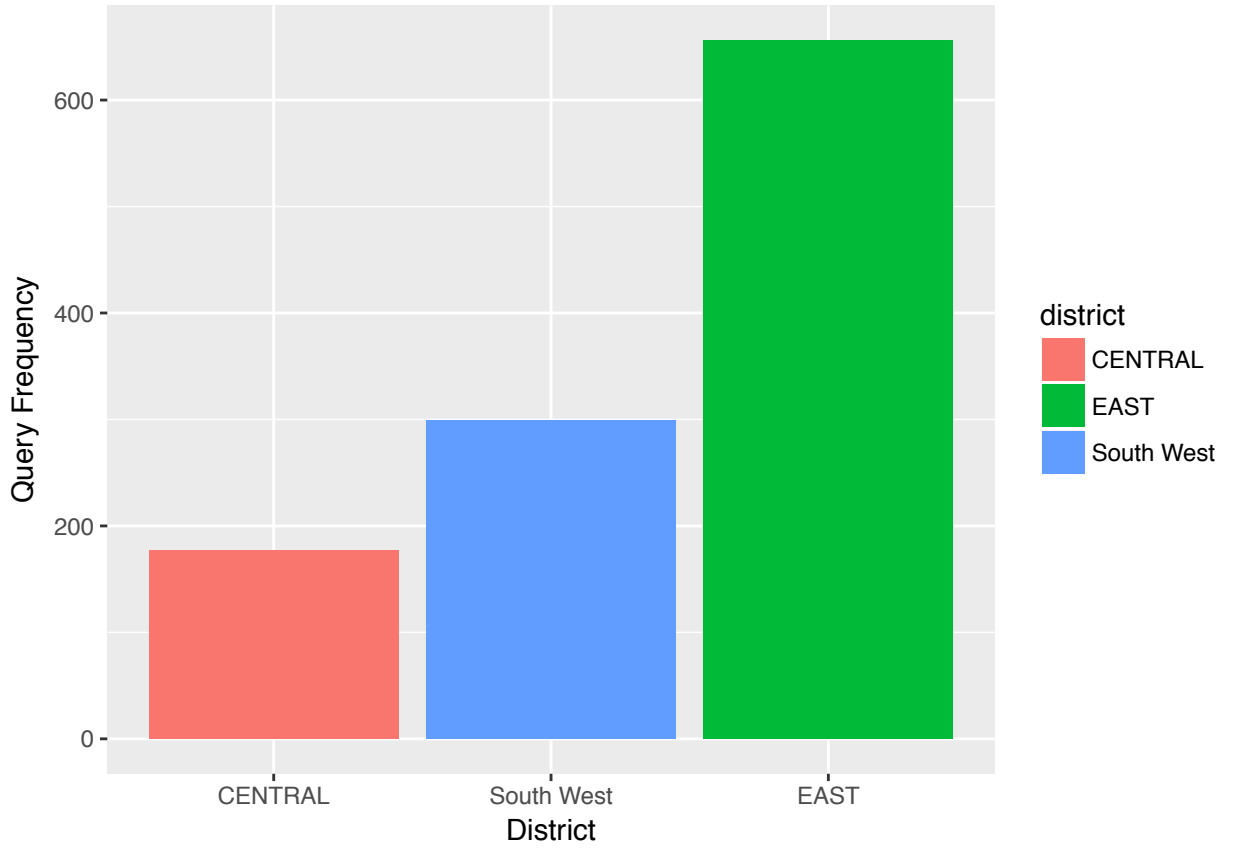


There are 3 districts mentioned under Delhi, Figure out which distric gets most calls

- The Graph shows that the **EAST DELHI** has got the most calls

```
queryType3 <- as.data.frame.table(cbind(table(data$district)))

colnames(queryType3) <- c("district", "type", "freq")
```

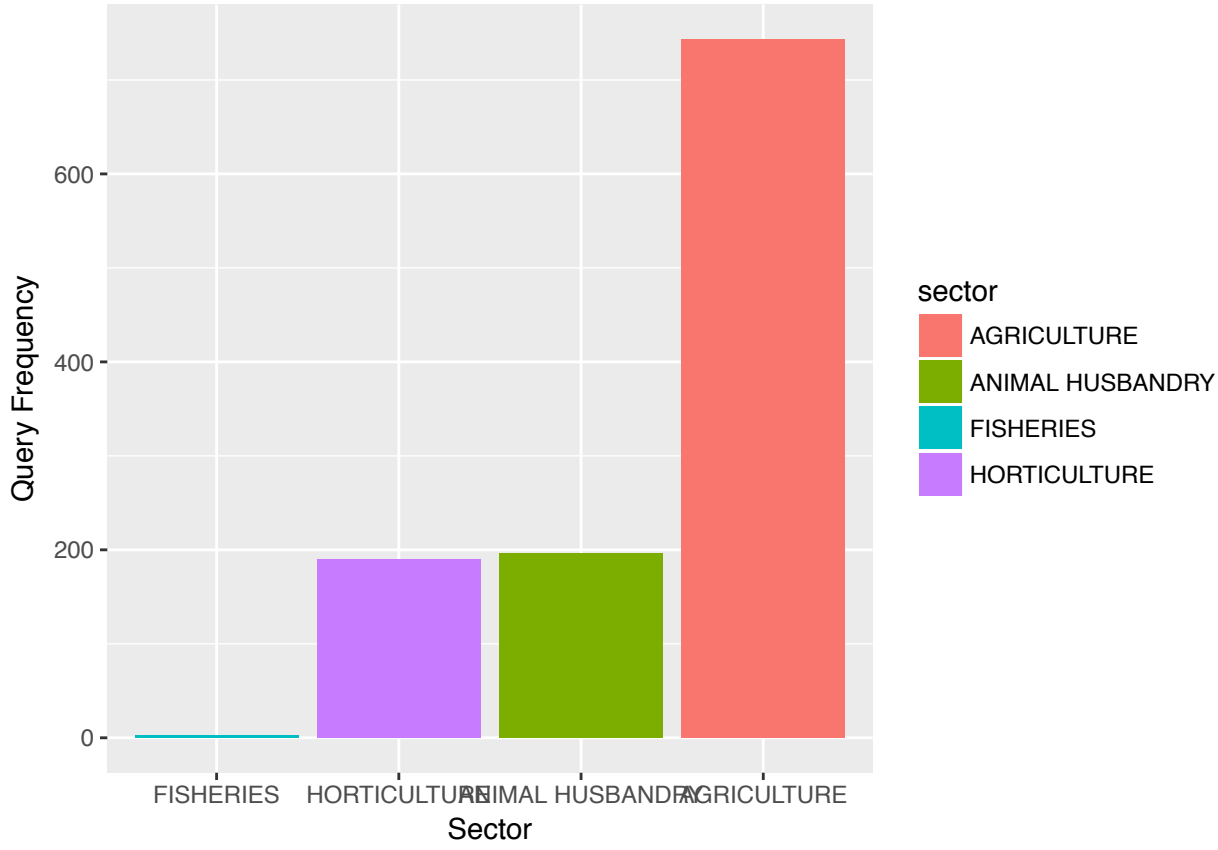
```
plot2 <- queryType3 %>%
  ggplot(aes(reorder(district,freq), freq, fill=district)) + geom_col() +
  xlab("District") +
  ylab("Query Frequency")

plot2
```



** Which Sector gets most queries?**

- AGRICULTURE Sector leads, followed by ANIMAL HUSBANDRY & HORTICULTURE. FISHERIES is the least

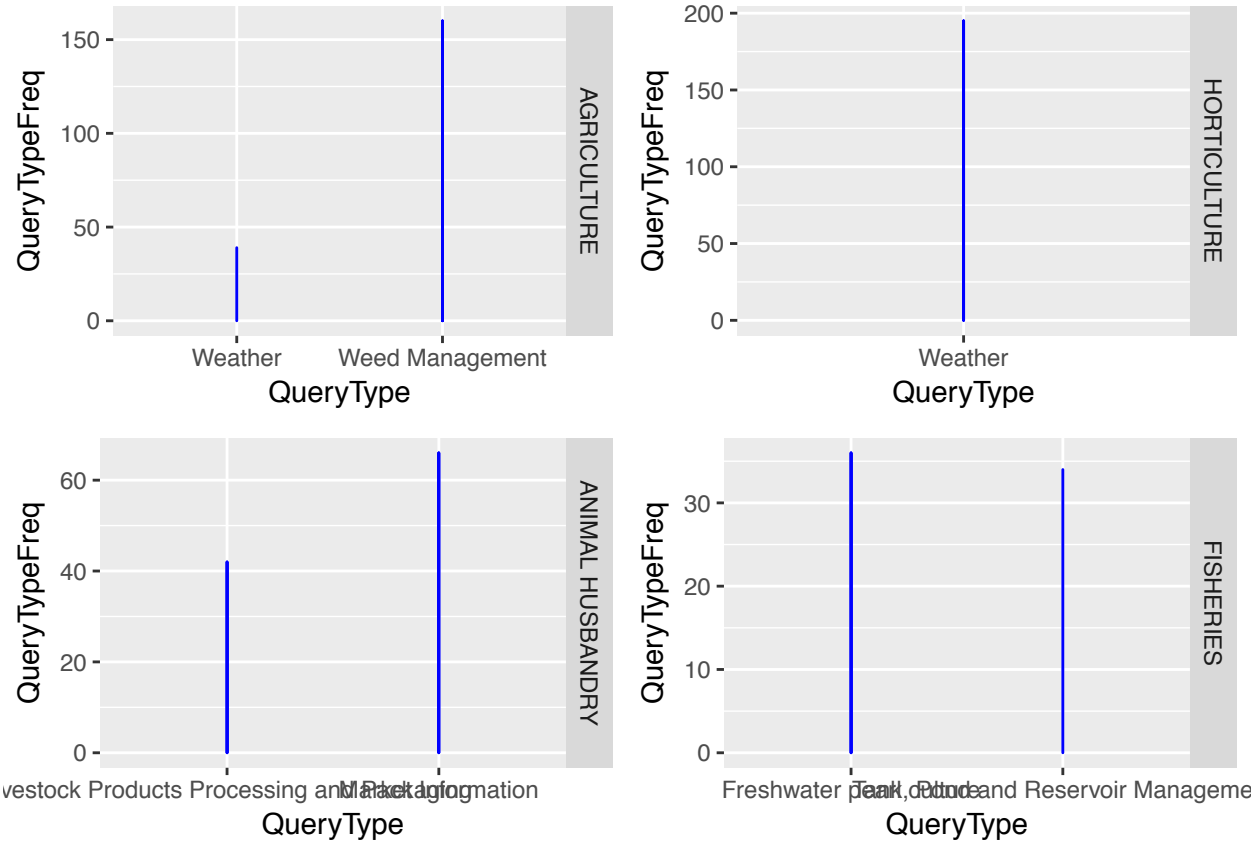
```
queryType4 <- as.data.frame.table(cbind(table(data$sector)))

colnames(queryType4) <- c("sector", "type", "freq")

plot3 <- queryType4 %>%
  ggplot(aes(reorder(sector,freq), freq, fill=sector)) + geom_col() +
  xlab("Sector") +
  ylab("Query Frequency")

plot3
```



Top 5 Query Type in each Sector

```
df<- read.csv("KCC.csv", header = TRUE)

df$QueryTypeFreq<- as.numeric(df$QueryType)

#"AGRICULTURE", "HORTICULTURE", "ANIMAL HUSBANDRY", "FISHERIES"

dfA <- df %>%
  filter(., Sector == c("AGRICULTURE")) %>%
  arrange(desc(QueryTypeFreq)) %>%
  head(5) %>%
  ggplot(aes(QueryType, QueryTypeFreq, color = Sector)) + geom_area(color='blue') +
  facet_grid(Sector~.)

dfB <- df %>%
  filter(., Sector == c("HORTICULTURE")) %>%
  arrange(desc(QueryTypeFreq)) %>%
  head(5) %>%
  ggplot(aes(QueryType, QueryTypeFreq, color = Sector)) + geom_area(color='blue') +
  facet_grid(Sector~.)

dfC <- df %>%
  filter(., Sector == c("ANIMAL HUSBANDRY")) %>%
  arrange(desc(QueryTypeFreq)) %>%
  head(5) %>%
  ggplot(aes(QueryType, QueryTypeFreq, color = Sector)) + geom_area(color='blue') +
```

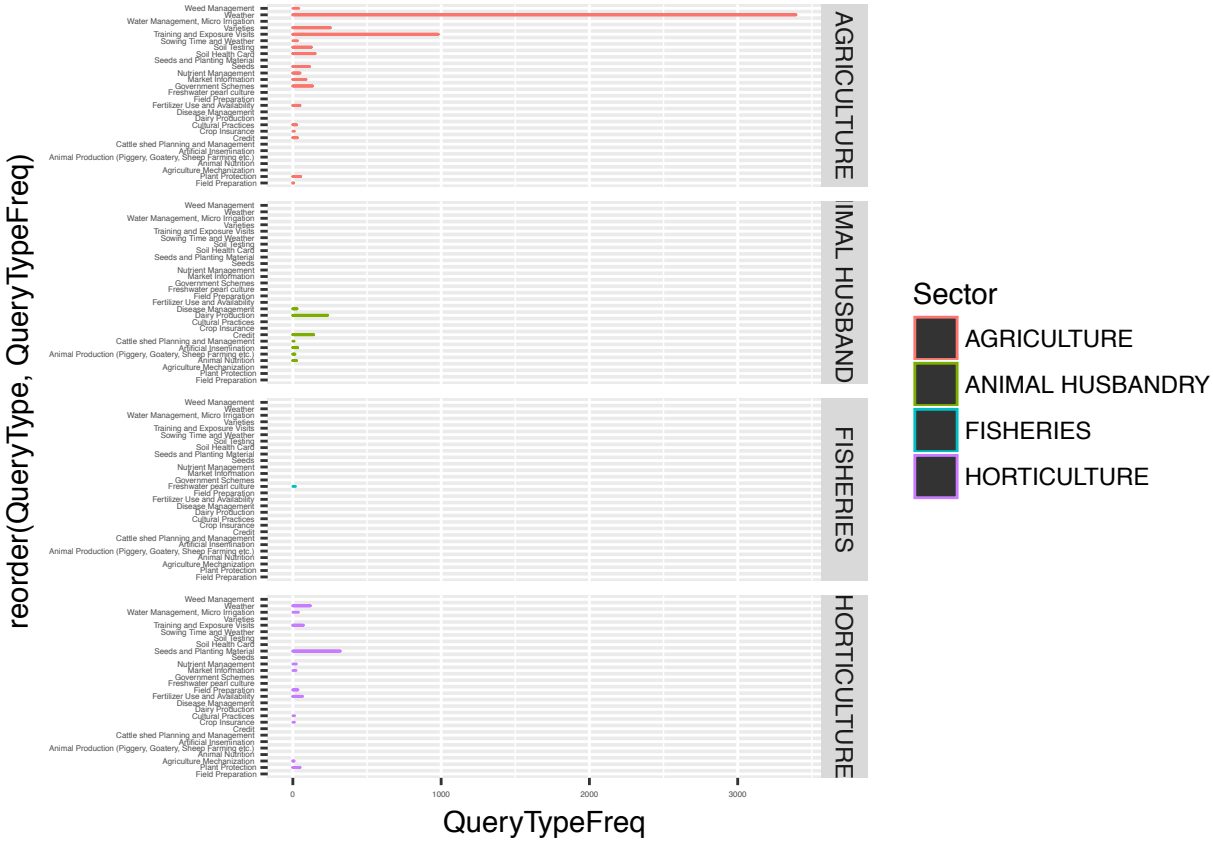
```
  facet_grid(Sector~.)

dfD <- df %>%
  filter(., Sector == c("FISHERIES")) %>%
  arrange(desc(QueryTypeFreq)) %>%
  head(5) %>%
  ggplot(aes(QueryType, QueryTypeFreq, color = Sector)) + geom_area(color='blue') +
  facet_grid(Sector~.)

grid.arrange(dfA, dfB, dfC, dfD)
```




```
df %>%
  filter(Sector == c("AGRICULTURE", "HORTICULTURE", "ANIMAL HUSBANDRY", "FISHERIES")) %>%
  ggplot(aes(reorder(QueryType, QueryTypeFreq), QueryTypeFreq , color = Sector)) + geom_area() +
  facet_grid(Sector~.) +
  coord_flip() +
  theme(axis.text = element_text(size = 3))
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9
```

```
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x
## $y, : font width unknown for character 0x9

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x
## $y, : font width unknown for character 0x9

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x
## $y, : font width unknown for character 0x9

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x
## $y, : font width unknown for character 0x9
```



Model

Well I am trying to predict which Secotr or Query Type would get most query in October 2017 based on Sept 2017 dataset.

```
df$date <- as.Date('2017-09-01')
df2 <- df %>%
  select(date, QueryTypeFreq)

colnames(df2) <- c('ds','y')
library(prophet)
```

```
## Loading required package: Rcpp
```

```
#m<- prophet(df2)
#future <- make_future_dataframe(m, periods = 10, freq = "d")
#forecast <- predict(m, future)
#plot(m, forecast)
#tail(future)
```

Communicate

Based on the graphs derived from the dataset, I can think of following insights

- KCC gets call for these Sectors: “AGRICULTURE”, “HORTICULTURE”, “ANIMAL HUSBANDRY”, “FISHERIES”
- Wheater is the mosted asked query
- East Delhi gets more number of query
- Most of the query is asked for AGRICULTURE sector
- least number of query is noted from FISHERIES sector
- Central Delhi has least query